جامعة ابن طفيل
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉ ⵎⵓⵃⴰⵎⵎⴷ
Ibn Tofaïl University
Faculté des Sciences

Université Ibn Tofail
Faculté des Sciences, Kénitra

Mémoire de Projet de Fin d'Etudes

Master Intelligence Artificielle et Réalité Virtuelle

CONCEPTION D'UN MODÈLE DE DÉTECTION D'ANOMALIES DANS LES SÉRIES TEMPORELLES

Établissement d'accueil : **S E R L I**

Elaboré par : Mme : Lyna HAFID

Encadré par : Mr Anass Nouri (Faculté des sciences -UIT)

Mr Cédric JOULAIN (SERLI)

Soutenu le 11/09/2024, devant le jury composé de :

- Mme Rajae TOUAHNI (Faculté des sciences -UIT)
- Mr Rochdi MESSOUSSI (Faculté des sciences -UIT)
- Mr Clency PERRINE (Faculté des sciences -UP)
- Mr David HELBERT (Faculté des sciences -UP)
- Mr Christian CHATELLIER (Faculté des sciences -UP)
- Mr Cédric JOULAIN (SERLI)

Année universitaire 2023/2024

REMERCIEMENTS

Je tiens à exprimer ma profonde gratitude envers toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire et à l'achèvement de mon parcours scolaire.

En premier lieu, je souhaite remercier chaleureusement mes professeurs de Master 1 à l'université Ibn Tofail à Kénitra. Leur encadrement rigoureux et leur passion pour l'enseignement m'ont permis de poser des bases solides pour mes études supérieures. Merci pour votre patience, votre soutien et les connaissances précieuses que vous m'avez transmises.

Je tiens également à exprimer ma reconnaissance à mes professeurs de Master 2 à l'université de Poitiers. Votre expertise, votre dévouement et votre ouverture d'esprit ont été essentiels pour approfondir mes compétences et m'épanouir académiquement. Vous avez su m'inspirer et m'encourager tout au long de cette année déterminante.

Je souhaite adresser mes sincères remerciements à toute l'équipe de SERLI où j'ai eu la chance de réaliser mon stage. Merci à mes collègues pour leur accueil chaleureux et leur soutien quotidien. Un merci particulier à mon tuteur en entreprise, Cédric Joulain, pour son encadrement précieux, ses conseils avisés et pour m'avoir permis d'acquérir une expérience professionnelle enrichissante.

Enfin, je remercie mon tuteur universitaire, David Helbert, pour son accompagnement tout au long de ce projet de fin d'études. Votre disponibilité, vos retours constructifs et votre guidance m'ont été d'une grande aide pour mener à bien ce travail.

À tous, merci de m'avoir aidé à franchir cette étape importante de mon parcours académique et professionnel.

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION

Dans un monde en perpétuelle transformation, où les avancées technologiques redéfinissent constamment les attentes des consommateurs, l'industrie des piscines résidentielles se trouve à un carrefour crucial. Les propriétaires de piscines ne se contentent plus d'une simple expérience aquatique, ils recherchent désormais des solutions intelligentes et automatisées qui non seulement simplifient la gestion quotidienne de leur bassin, mais aussi minimisent leur impact environnemental. Face à cette demande croissante, l'innovation devient non seulement une nécessité, mais aussi une opportunité pour transformer un secteur traditionnellement manuel et énergivore.

C'est dans ce contexte dynamique que s'inscrit notre projet de fin d'études, qui se concentre sur la détection d'anomalies dans les séries temporelles recueillies par les capteurs des piscines. En intégrant ces technologies de pointe, notre projet ambitionne de créer des solutions novatrices pour offrir aux utilisateurs une expérience de piscine qui allie confort, praticité et durabilité. Notre démarche, loin d'être une simple réponse aux attentes actuelles, vise à anticiper les besoins futurs, en posant les bases d'une gestion de piscine résolument moderne et proactive.

Ce mémoire est structuré de manière à vous guider à travers l'ensemble du processus de conception et d'implémentation d'un modèle de détection d'anomalies dans les séries temporelles. Chaque section du document est dédiée à un aspect spécifique du projet, offrant ainsi une compréhension complète et détaillée de notre démarche.

Nous commençons par une vue d'ensemble de l'entreprise dans laquelle le projet a été réalisé, en détaillant son histoire, ses activités, son expertise, et sa structure organisationnelle. Cette introduction permet de comprendre le contexte dans lequel notre travail s'inscrit.

Nous poursuivons avec un état de l'art qui explore les différentes approches et modèles utilisés dans le domaine des séries temporelles, y compris les théories traditionnelles et les avancées modernes. Cette section situe notre projet par rapport aux connaissances actuelles et aux défis rencontrés.

L'analyse du projet examine les besoins spécifiques auxquels notre modèle répond, en mettant en lumière les exigences fonctionnelles et techniques, ainsi que les contraintes rencontrées. Cette partie nous aide à comprendre les objectifs du projet et les défis associés.

Nous détaillons ensuite les outils et les données utilisés, en décrivant les outils de traitement et les caractéristiques des données collectées. Cette section offre un aperçu des ressources disponibles pour la réalisation du projet.

La réalisation du projet est ensuite détaillée, en abordant le prétraitement des données, le développement et l'entraînement des modèles, ainsi que les techniques d'optimisation pour la détection d'anomalies. Les résultats obtenus seront analysés pour évaluer l'efficacité des solutions mises en œuvre.

Par la suite, la gestion du projet est examinée, incluant l'établissement du cahier des charges, l'application de la méthode Scrum, ainsi que la gestion des risques et du budget. Cette partie met en lumière l'organisation et le suivi du projet tout au long de son déroulement.

Enfin, la réflexion méthodologique évalue les choix méthodologiques adoptés, les difficultés rencontrées, et les perspectives futures. Cette section fournit un bilan des méthodes utilisées et des leçons apprises, tout en explorant les évolutions possibles pour les travaux à venir.

En somme, ce mémoire se veut une contribution significative à l'évolution de la gestion des piscines, en répondant aux défis majeurs que sont la maintenance inefficace, la consommation énergétique élevée, et la qualité fluctuante de l'eau. Grâce à des algorithmes avancés, notre projet offre une réponse innovante aux besoins des propriétaires modernes, en leur permettant de vivre une expérience de piscine plus simple, efficace, et respectueuse de l'environnement.

Chapitre 1

PRÉSENTATION DE LA STRUCTURE DE L'ENTREPRISE

1.1 Historique de l'entreprise

Fondée en 1981, SERLI SAS est une société par actions simplifiée (SAS) spécialisée dans le conseil en systèmes et logiciels informatiques. Immatriculée sous le numéro SIREN 322 770 850, la société est active depuis plus de 42 ans. Son siège social est situé à Chasseneuil-du-Poitou, dans la Vienne (86), à l'adresse suivante : Avenue Thomas Edison, 86360 Chasseneuil-du-Poitou.

Depuis sa création, SERLI SAS a évolué pour devenir un acteur reconnu dans le domaine des technologies de l'information, en particulier dans le conseil en systèmes et logiciels informatiques. L'entreprise a su s'adapter aux évolutions technologiques et aux besoins changeants de ses clients, en développant des solutions innovantes et sur mesure.

Table 1.1: Fiche Technique de l'Entreprise SERLI SAS

Caractéristique	**Détails**
Capital social	99 020,00 €
PDG	Jérôme Petit
Effectif	50 salariés
Siège social	Avenue Thomas Edison, 86360 Chasseneuil-du-Poitou 2 Rue de la Tête de Boeuf, 79000 Niort

1.2 Activités et expertise

SERLI SAS se distingue par son expertise dans l'élaboration et la commercialisation de matériels et programmes informatiques. L'entreprise intervient principalement dans le conseil en systèmes et logiciels informatiques, ce qui inclut le développement de logiciels sur mesure, l'intégration de systèmes informatiques, ainsi que la gestion et l'optimisation des infrastructures IT pour ses clients.

L'entreprise est inscrite sous le code NAF 6202A, qui correspond au secteur d'activité du conseil en systèmes et logiciels informatiques. Grâce à une équipe d'experts qualifiés, SERLI SAS propose des solutions adaptées aux besoins spécifiques de chaque client, en s'appuyant sur une méthodologie rigoureuse et des outils à la pointe de la technologie.

1.3 Structure et organisation

SERLI SAS est dirigée par l'entreprise GERANIUM, représentée par Jérôme Petit, qui en est le président depuis le 29 décembre 2016. L'entreprise a su maintenir une structure organisationnelle efficace, lui permettant de répondre rapidement aux demandes de ses clients tout en garantissant un haut niveau de qualité dans les services rendus.

L'effectif de SERLI SAS est compris entre 20 et 49 salariés. Ce groupe de professionnels est composé d'ingénieurs, de développeurs, d'analystes et de chefs de projet, tous spécialisés dans différents domaines de l'informatique. L'entreprise encourage la formation continue de ses employés pour rester à la pointe des nouvelles technologies et des meilleures pratiques du secteur.

1.4 Performances financières

Sur l'année 2015, SERLI SAS a réalisé un chiffre d'affaires de 4 707 100,00 €, témoignant de sa solidité financière et de sa capacité à attirer et fidéliser une clientèle diversifiée. Bien que les comptes récents ne soient pas disponibles, ces chiffres reflètent l'importance de l'entreprise dans son secteur et son impact sur le marché local.

1.5 Clients et projets réalisés

Au fil des ans, SERLI SAS a collaboré avec une variété de clients, allant de petites et moyennes entreprises à de grandes corporations. Ses services couvrent plusieurs secteurs d'activité, incluant l'industrie, les services financiers, et les administrations publiques. L'entreprise se distingue par sa capacité à gérer des projets complexes, depuis la conception initiale jusqu'à la mise en œuvre finale, tout en garantissant la satisfaction de ses clients.

Chapitre 2

ÉTAT DE L'ART

2.1 Introduction

La détection d'anomalies dans les séries temporelles est une tâche cruciale pour de nombreuses applications industrielles, financières et scientifiques. Le traitement de séries temporelles à pas irréguliers et multivariées pose des défis spécifiques, nécessitant des modèles capables de gérer des dépendances complexes à travers différentes échelles de temps et de variables. Ce chapitre examine les approches actuelles en mettant un accent particulier sur les *Temporal Convolutional Networks* (TCN) et les *Long Short-Term Memory* (LSTM), qui ont été sélectionnés pour notre travail en raison de leur capacité à répondre à ces défis.

2.2 Méthodes Classiques de Détection d'Anomalies

Les méthodes classiques de détection d'anomalies dans les séries temporelles reposent souvent sur des techniques statistiques et des modèles linéaires. Elles sont généralement plus simples à mettre en œuvre mais peuvent avoir des limitations en termes de capacité à capturer des structures complexes.

2.2.1 Méthodes Statistiques Univariées

Analyse de la Moyenne et de l'Écart-Type

- **Méthode :** Consiste à calculer la moyenne et l'écart-type d'une série temporelle. Les points qui se trouvent au-delà d'un certain seuil (souvent un multiple de l'écart-type) sont considérés comme des anomalies.

- **Avantages :** Simple et facile à comprendre.
- **Inconvénients :** Ne prend pas en compte la structure temporelle des données et peut être sensible aux fluctuations saisonnières.

Tests de Déviation

- **Méthode :** Utilisation de tests statistiques pour détecter des écarts significatifs par rapport à une distribution attendue.
- **Avantages :** Basé sur des principes statistiques rigoureux.
- **Inconvénients :** Moins efficace pour des séries temporelles avec des tendances ou une saisonnalité marquée.

2.2.2 Méthodes de Décomposition

Décomposition en Composantes Multiples (STL)

- **Méthode :** Décompose une série temporelle en trois composants : tendance, saisonnalité et résidu. Les anomalies sont détectées en analysant le composant résiduel.
- **Avantages :** Permet de traiter les données avec des tendances et des cycles saisonniers.
- **Inconvénients :** Peut être complexe à paramétrer et sensible aux choix de paramètres.

Décomposition de la Série Temporelle

- **Méthode :** Séparation de la série temporelle en composants individuels pour isoler les anomalies.
- **Avantages :** Facilite l'analyse des anomalies dans des contextes variés.
- **Inconvénients :** Ne capture pas toujours les relations complexes entre les composants.

2.2.3 Méthodes Basées sur les Modèles

ARIMA (AutoRegressive Integrated Moving Average)

- **Méthode :** Modèle basé sur l'autocorrélation des données et les différences pour stationnariser les séries temporelles.
- **Avantages :** Efficace pour modéliser des séries temporelles stationnaires ou transformées pour être stationnaires.

- **Inconvénients :** Moins performant pour des séries temporelles non linéaires ou fortement non stationnaires [7].

Modèles de Lissage Exponentiel

- **Méthode :** Utilisation de moyennes mobiles exponentielles pour modéliser les tendances et les niveaux saisonniers.
- **Avantages :** Simple et rapide à mettre en œuvre.
- **Inconvénients :** Peut ne pas bien capturer les anomalies dans les séries temporelles avec des comportements complexes.

2.3 Méthodes Modernes de Détection d'Anomalies

Les méthodes modernes de détection d'anomalies exploitent des techniques d'apprentissage automatique et d'apprentissage profond pour capturer des motifs complexes et améliorer la précision de la détection.

2.3.1 Méthodes Basées sur l'Apprentissage Automatique

- **Méthodes :** Utilisation de modèles tels que les forêts aléatoires (Random Forests), les machines à vecteurs de support (SVM) et les réseaux neuronaux pour classer les points comme normaux ou anormaux.
- **Avantages :** Capacité à apprendre des patterns complexes et à s'adapter à des données variées.
- **Inconvénients :** Nécessite des échantillons étiquetés, ce qui peut être difficile à obtenir.

2.3.2 Méthodes Basées sur l'Apprentissage Profond

Réseaux de Neurones Récurrents (RNN)

- **Méthodes :** Utilisation de réseaux RNN comme les Long Short-Term Memory (LSTM) et Gated Recurrent Units (GRU) pour modéliser les dépendances temporelles à long terme dans les séries temporelles.
- **Avantages :** Très efficace pour capturer des relations complexes et des séquences longues [2].
- **Inconvénients :** Entraînement complexe et nécessite une grande quantité de données.

Temporal Convolutional Networks (TCN)

- **Méthodes :** Utilisation de réseaux de convolution temporelle pour modéliser les dépendances temporelles dans les séries temporelles. Les TCN utilisent des couches de convolution 1D avec des filtres causals pour maintenir l'ordre temporel et capturer des motifs à différentes échelles.
- **Avantages :** Permet la modélisation de séquences longues avec une gestion efficace des dépendances temporelles. Moins sujet au problème de gradient vanishing comparé aux RNN. De plus, les TCN sont parallélisables, ce qui permet un entraînement plus rapide et une meilleure utilisation des ressources matérielles modernes [3].
- **Inconvénients :** Peut nécessiter un ajustement minutieux des hyperparamètres et une gestion des ressources computationnelles pour des séries temporelles de grande longueur.

Transformers

- **Méthodes :** Utilisation de modèles Transformers, comme BERT ou GPT adaptés aux séries temporelles, pour modéliser les dépendances temporelles et les relations complexes entre les points de données.
- **Avantages :** Excellente capacité à gérer des séquences longues et à capturer des dépendances complexes.
- **Inconvénients :** Besoin de beaucoup de ressources computationnelles et de données pour un bon entraînement.

Auto-encodeurs

- **Méthodes :** Réseaux neuronaux utilisés pour apprendre une représentation compacte des données. Les anomalies sont détectées par la reconstruction d'erreurs.
- **Avantages :** Adapté pour les anomalies rares et les données avec des dimensions élevées [5].
- **Inconvénients :** Sensible aux choix de l'architecture et des hyperparamètres.

2.4 Conclusion

L'analyse de l'état de l'art démontre que les *Long Short-Term Memory* (LSTM) et les *Temporal Convolutional Networks* (TCN) sont des choix pertinents pour la détection d'anomalies dans des séries temporelles à pas irréguliers et multivariées. Leur complémentarité permet de capturer à la fois des dépendances à long terme et des motifs locaux complexes, rendant notre approche à la fois robuste et efficace pour les applications envisagées dans ce mémoire.

Chapitre 3

ANALYSE ET ÉTUDE DU PROJET

3.1 Introduction

Dans le domaine des piscines résidentielles, les propriétaires sont souvent confrontés à des défis liés à la maintenance, à la consommation énergétique, et à la qualité de l'eau. La gestion traditionnelle des piscines nécessite des interventions manuelles fréquentes pour le nettoyage, le réglage des paramètres chimiques, et la surveillance des équipements. Cette gestion laborieuse n'est pas seulement chronophage, mais elle est également sujette à des erreurs humaines, pouvant conduire à des problèmes de qualité de l'eau, à une dégradation des équipements, et à des coûts de maintenance élevés. De plus, elle exige une expertise spécifique, car une connaissance approfondie des produits chimiques et des systèmes de filtration est indispensable pour maintenir la piscine en bon état.

En parallèle, la consommation énergétique associée aux systèmes de filtration, de chauffage, et de traitement de l'eau constitue une part importante des dépenses des propriétaires de piscines. Les méthodes de gestion énergétique actuelles manquent souvent d'optimisation, entraînant un gaspillage d'énergie significatif.

De plus, la qualité de l'eau reste une préoccupation majeure, car un mauvais équilibre chimique peut rendre la piscine inutilisable et causer des dommages matériels. Les propriétaires doivent surveiller constamment des paramètres tels que le pH et les niveaux de chlore pour éviter la prolifération d'algues ou d'autres problèmes similaires.

Face à ces défis, il est impératif de développer des solutions qui automatisent la gestion des piscines, réduisent les interventions manuelles, optimisent la consommation énergétique, et garantissent une qualité d'eau optimale. C'est dans ce contexte que le projet de piscine intelligente a été

initié, avec pour objectif de surmonter ces limitations par l'intégration de technologies avancées.

3.2 Étude du besoin

Afin de répondre aux problématiques identifiées, le projet de piscine intelligente se concentre sur plusieurs axes d'innovation technologique pour améliorer l'expérience des propriétaires de piscines.

3.2.1 Surveillance et gestion de la qualité de l'eau

La qualité de l'eau est essentielle pour garantir la sécurité et le confort des utilisateurs. Un système capable de surveiller en temps réel les paramètres de l'eau et de maintenir un équilibre chimique optimal est nécessaire. Ce système doit également pouvoir anticiper les variations de qualité de l'eau, comme les fluctuations de pH, en utilisant des algorithmes de prévision.

3.2.2 Optimisation de la consommation énergétique

La réduction de la consommation énergétique est un besoin crucial pour minimiser les coûts d'exploitation. Un système qui analyse les données en temps réel pour ajuster automatiquement les paramètres énergétiques (comme le chauffage et la filtration) en fonction des besoins précis permettrait de réduire le gaspillage énergétique.

3.3 Exigences

Pour répondre aux besoins identifiés, le système de piscine intelligente doit satisfaire les exigences suivantes :

3.3.1 Exigences fonctionnelles

A) Détection d'Anomalies :

- Le système doit être capable de détecter les anomalies liées aux actions des utilisateurs, telles que l'ajout d'une pastille dans le skimmer.
- En cas de détection d'une augmentation significative d'un paramètre de l'eau suite à une action utilisateur, le système doit générer une alerte.

B) Prévision (Forecasting) :

- Le système doit surveiller en continu les performances des équipements de la piscine.
- Il doit permettre la détection précoce des défaillances potentielles pour planifier la maintenance préventive.

3.3.2 Exigences techniques

A) Traitement de Données :

- Le système doit utiliser des algorithmes avancés pour analyser et traiter les données collectées.
- Ces algorithmes doivent être capables de détecter les schémas et les anomalies, ainsi que de générer des prévisions précises.

B) Évaluation et Validation :

- Un suivi régulier de l'avancement du projet est nécessaire pour garantir que le projet progresse conformément aux objectifs fixés.

3.4 Contraintes

Le développement du projet de piscine intelligente doit également prendre en compte plusieurs contraintes, qui peuvent influencer la conception et la mise en œuvre du système.

3.4.1 Contraintes technologiques

- Le choix des technologies doit assurer la compatibilité avec les infrastructures existantes et offrir une évolutivité future. L'utilisation de LispTick pour le calcul des séries temporelles et de Python pour l'implémentation des modèles de prévision doit être optimisée pour tirer pleinement parti des ressources matérielles disponibles.
- Les algorithmes doivent être suffisamment performants pour fonctionner en temps réel, tout en minimisant la charge sur les systèmes de traitement.

3.4.2 Contraintes matérielles

- Le projet nécessite un matériel informatique puissant, notamment pour l'entraînement des modèles d'intelligence artificielle. Un ordinateur équipé d'une carte graphique performante est essentiel, ainsi qu'une infrastructure serveur capable de gérer des calculs intensifs sans risque de surchauffe.
- Les ressources matérielles doivent être disponibles en quantité suffisante pour garantir la fiabilité du système, avec des coûts maîtrisés.

3.4.3 Contraintes de données

- La qualité et la diversité des données collectées sont cruciales pour le succès du projet. Les données doivent être fiables, précises et représentatives des différentes conditions d'utilisation des piscines.
- La fréquence élevée de collecte des données nécessite une infrastructure de stockage et de traitement capable de gérer de grandes quantités de données en temps réel.

3.4.4 Contraintes de temps

- Le projet doit respecter un calendrier précis, avec des jalons intermédiaires pour évaluer l'avancement et ajuster la stratégie si nécessaire.
- Des délais peuvent survenir en raison de défis imprévus liés au développement des modèles ou à l'intégration des systèmes, ce qui nécessite une planification proactive et flexible.

3.5 Livrables

Le principal livrable de ce projet sera une étude détaillée sur les différents algorithmes utilisés pour détection d'anomalies , accompagnée d'un benchmark comparatif pour évaluer leurs performances. Ce livrable comprendra les éléments suivants :

- Étude sur les Différents Algorithmes de Détection d'anomalies :
  Une analyse approfondie des principaux algorithmes de détection d'anomalies appliqués aux time séries.
- Benchmark comparatif :
  Nous avons entrepris un Benchmark comparatif des différents modèles disponibles pour évaluer leurs performances et leurs capacités. L'objectif principal de cette démarche est de déterminer quel modèle offre le meilleur compromis entre efficacité, précision et coût, et de concentrer nos efforts sur celui qui se distingue clairement des autres.
- Rapport d'Étude :
  Un rapport exhaustif décrivant les résultats de l'étude sur les différents algorithmes de détection d'anomalies, ainsi que les conclusions et les recommandations découlant de l'analyse comparative. Ce rapport servira de référence pour les décisions futures concernant le choix des algorithmes pour la détection d'anomalies sur des time séries. progrès et réajuster si nécessaire.

3.6 Conclusion

Le développement d'une piscine intelligente doit non seulement répondre aux besoins des propriétaires en termes de réduction des coûts énergétiques, et de qualité de l'eau, mais aussi satisfaire à des exigences fonctionnelles et techniques précises, tout en tenant compte des contraintes technologiques, matérielles, de données, et de temps. L'intégration réussie de ces aspects est essentielle pour transformer la gestion des piscines résidentielles et offrir une solution innovante et durable.

Chapitre 4

OUTILS ET DONNÉES

4.1 Introduction

Dans ce chapitre, nous passons en revue les dernières avancées dans notre domaine d'étude, nous discuterons également des outils, des données et des méthodes utilisés pour mener à bien notre projet. Nous expliquerons les raisons qui ont motivé nos choix d'outils et de méthodes, et nous décrirons les données que nous avons recueillies ou utilisées. Nous expliquerons également en quoi ces données sont pertinentes et importantes pour la réalisation de nos objectifs.

4.2 Description des outils

Afin de bien gérer nos séries temporelles, nous avions choisi d'utiliser LispTick qui est un puissant moteur de calcul de séries temporelles construit autour du streaming pur développé par Cédric Joulain, mon tuteur. Implémenté dans Go, un langage moderne destiné à la programmation simultanée, LispTick peut utiliser pleinement tous les cœurs des machines et peut être disponible sur n'importe quelle plate-forme et système d'exploitation sur lesquels Go est disponible.

Pour créer des requêtes compactes mais puissantes, un dialecte de LISP est utilisé. Les types intégrés tels que Time, Duration et Timeseries, une liste de paires (Time, Value), permettent des calculs empilés à la volée.

En tant que langage de prédilection en intelligence artificielle, Python s'est imposé comme un choix logique pour notre équipe pour l'implémentation de nos modèles de forecasting et anomaly detection, assurant une cohérence et une efficacité accrues dans le développement de notre projet.

4.3 Description des données

Le succès de notre projet repose sur la diversité et la qualité des données, afin de mettre en place des modèles capables de capturer les tendances réelles et les interactions complexes entre les phénomènes.

Le client nous a fourni un ensemble de données relatives à environ 3000 bassins, couvrant 1276 variables, allant d'avant le 3 août 2023 jusqu'à aujourd'hui. Ces données incluent à la fois des mesures métriques et des événements indiquant des changements.

Ayant accès aux données brutes du client, nous avons effectué l'ingénierie des données pour les rendre exploitables. Avant le 3 août, nos données étaient enregistrées toutes les 15 minutes. Après cette date, la fréquence d'enregistrement des données est passée à environ toutes les 10 secondes, devenant encore plus fréquente lorsqu'un événement se produit.

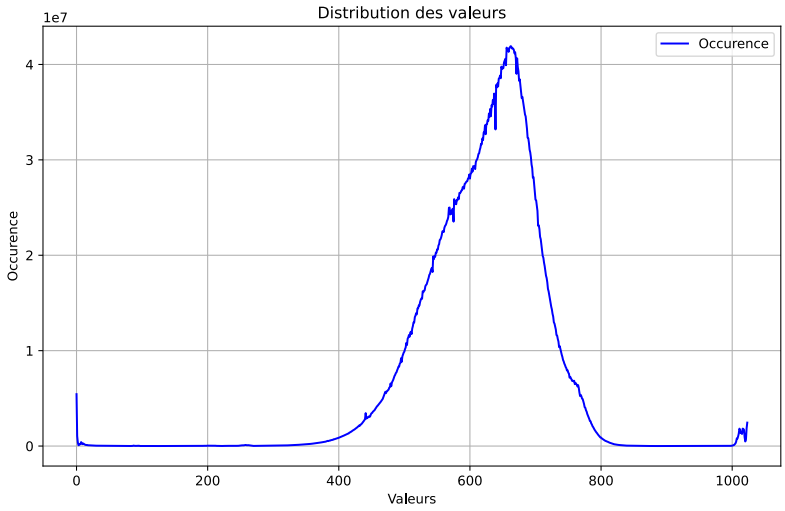

FIGURE 4.1 – Valeurs de température de l'air

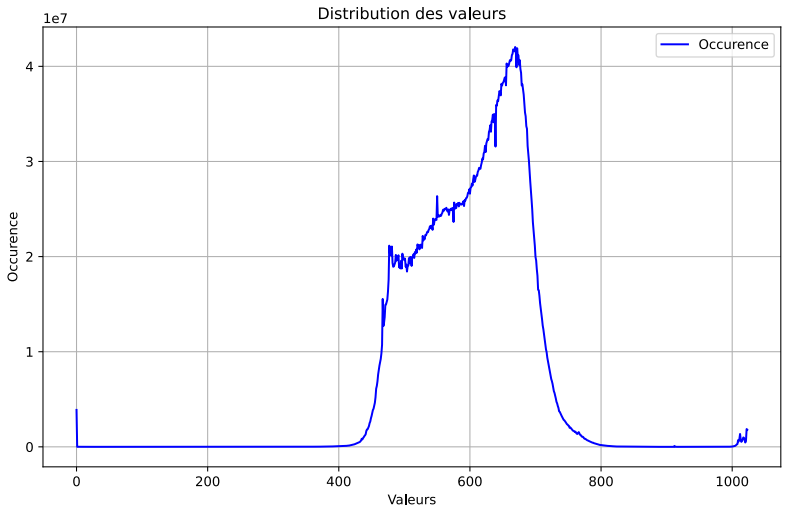

FIGURE 4.2 – Valeurs de température de l'eau

4.4 Conclusion

En conclusion, ce chapitre a passé en revue les outils, les données et les méthodes utilisés dans notre projet de détection d'anomalies.

Chapitre 5

RÉALISATION

5.1 Introduction

Dans ce chapitre, nous présentons l'ensemble des étapes réalisées au cours du projet, allant du traitement des données brutes à l'optimisation des modèles de détection d'anomalies. Le projet a été structuré de manière à assurer une progression méthodique, avec une attention particulière portée à la qualité des données, à la sélection des modèles, à leur entraînement et à leur migration vers des frameworks plus performants. Nous aborderons également les différentes approches explorées pour l'optimisation des performances, notamment en ce qui concerne la détection d'anomalies, un aspect central du projet.

5.2 Données brutes

Comme nous avons accès aux données brutes, nous avons choisi de les utiliser afin de minimiser les erreurs potentielles.

Prenons l'exemple de la mesure de la température de l'eau, une thermistance NTC est utilisée. Il s'agit d'un composant électronique avec une résistance électrique qui diminue lorsque la température augmente. La thermistance est immergée dans l'eau de la piscine et connectée en série avec une résistance fixe. Cette configuration forme un diviseur de tension, ce qui permet de traduire les variations de résistance en variations de tension. Le cœur du système réside dans le microcontrôleur, tel qu'un Arduino ou un ESP8266, qui mesure la résistance de la thermistance en convertissant la tension électrique qu'elle produit en une valeur numérique. Cette valeur numérique est souvent représentée sur 10 bits, donc elle peut varier de 0 à 1023. Chaque valeur numérique correspond à une plage de tensions spécifique.

La température correspondante est ensuite calculée en utilisant une formule mathématique basée

sur les caractéristiques spécifiques de la thermistance, telles que sa résistance nominale et son coefficient Beta.

5.3 Prétraitement

Normalisation

Comme les attributs sont sur différentes échelles, il est crucial d'appliquer une normalisation à nos données. Dans ce contexte, les histogrammes s'avèrent particulièrement utiles comme ils offrent une méthode pratique pour calculer des moyennes et des écarts types plus représentatifs, évitant ainsi les biais potentiels causés par les valeurs aberrantes (outliers). De plus, les histogrammes fournissent une visualisation claire de la distribution des données, permettant de détecter toute similitude avec des distributions gaussiennes. Cette analyse approfondie renforcera la clarté et la pertinence de nos présentations.

Fenêtres glissantes

Afin de simplifier l'apprentissage des modèles, améliorer la détection des anomalies, nous utilisons des fenêtres glissantes.

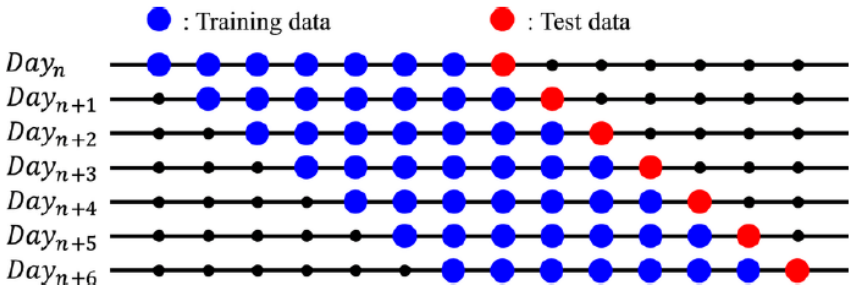

FIGURE 5.1 – Les fenêtres glissantes

L'idée est de traiter les séries temporelles par petits morceaux ou "fenêtres" qui se chevauchent partiellement au lieu de traiter l'ensemble des données en une seule fois, cela permet de lisser ou de simplifier le traitement en réduisant la quantité de données à manipuler à chaque étape.
Ayant des séries temporelles à des intervalles irréguliers, nous avons, dans un premier temps, moyenné les valeurs pour chaque intervalle de 15 minutes de telle façon à avoir 96 valeurs par jour avec intervalles de temps réguliers de 15 minutes. La taille de de la fenêtre a, donc, été 96 avec un pas de 1.

Attributs d'entrée

Parmi les 1276 variables dont on a accès, nous avons pu, avec l'aide du client, soigneusement sélectionné huit variables pour l'entrainnement de notre modèle. Nous avons gardé les valeurs de ph, chlore, température d'air, température d'eau, état de la pompe parmi d'autres.

5.4 Modèles

D'après les recherches récentes [1], les algorithmes se distinguent par leur performance en matière de détection d'anomalies dans les séries temporelles. Les LSTM (Long Short-Term Memory) sont particulièrement efficaces grâce à leur capacité à capturer les dépendances à long terme, ce qui est essentiel pour identifier des motifs anormaux dans des données séquentielles. Les TCN (Temporal Convolutional Networks), quant à eux, utilisent des convolutions causales et dilatées pour traiter les séquences temporelles, offrant une grande stabilité du gradient et un parallélisme efficace lors de l'entraînement. Les auto-encodeurs[6], notamment lorsqu'ils sont combinés avec des LSTM ou des TCN, sont également couramment utilisés pour la détection d'anomalies en apprenant à reconstruire les données normales, ce qui permet de détecter les anomalies par l'erreur de reconstruction élevée. D'ou notre choix d'utiliser des auto-encodeurs avec les deux algorithmes mentionnés.

5.4.1 Premier modèle

Le premier modèle est un auto-encodeur LSTM qui a été conçu pour compresser et reconstruire les séquences temporelles. Le processus commence par l'encodage des données grâce à deux couches LSTM successives.
La dimensionnalité de la première couche LSTM est de 32 unités LSTM, tandis que la deuxième couche LSTM, qui agrège les informations produites par la couche d'avant pour former une représentation latente de dimension réduite, comporte 16 unités LSTM. Cette représentation latente est ensuite répétée pour chaque pas de temps, puis décodée par deux couches LSTM dans le sens inverse. La couche LSTM de décodage reconstitue la structure temporelle des données à partir du vecteur latent répété, tandis que la dernière couche LSTM finalise la reconstruction de la séquence temporelle.

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 96, 1)	0
lstm (LSTM)	(None, 96, 32)	4,352
lstm_1 (LSTM)	(None, 16)	3,136
repeat_vector (RepeatVector)	(None, 96, 16)	0
lstm_2 (LSTM)	(None, 96, 16)	2,112
lstm_3 (LSTM)	(None, 96, 32)	6,272
time_distributed (TimeDistributed)	(None, 96, 1)	33

FIGURE 5.2 – Architecture du premier modèle

Une fois le modèle construit, il est compilé avec l'optimiseur Adam et la perte MSE pour évaluer la différence entre les données d'entrée et les données reconstruites. Cette architecture permet au modèle d'apprendre une représentation compacte des données temporelles tout en conservant les caractéristiques importantes pour la reconstruction précise des séquences. Pour évaluer la performance du modèle, nous avions choisi le paramètre "PH" comme il est plus facile à interpréter et par conséquent détecter quand la valeur est absurde. Nous avons remarqué par la suite les points suivants : Le modèle détecte, de toute façon, les pics qu'ils s'agissent des anomalies ou pas. Plus d'époques n'entraîne pas forcément le surajustement.

5.4.2 Deuxième modèle

Il s'agit d'un modèle comportant des couches de convolution qui extraient des caractéristiques importantes des données temporelles en appliquant des filtres sur la séquence d'entrée. Ensuite, des couches de déconvolution sont utilisées pour reconstruire les données à partir des caractéristiques extraites. Ces couches permettent de remonter progressivement des caractéristiques abstraites aux données reconstruites. Deux couches de dropout sont introduites pour régulariser le modèle et réduire le surajustement. Nous avons gardé les mêmes dimensions, les couches extérieures comportent 32 unités et les couches intérieures comportent 16 unités.

Pour introduire de la non-linéarité dans le modèle, nous avons utilisé la fonction d'activation ReLU. La fonction de perte utilisée est l'erreur quadratique moyenne (MSE), minimisée par l'optimiseur Adam. Le modèle est entraîné sur l'ensemble des données et peut détecter les anomalies lorsque la reconstruction des données dépasse un seuil prédéterminé. En détectant les écarts significatifs entre les données reconstruites et les données d'origine, le modèle peut identifier les points anormaux dans les séries temporelles.

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 48, 32)	256
dropout (Dropout)	(None, 48, 32)	0
conv1d_1 (Conv1D)	(None, 24, 16)	3,600
conv1d_transpose (Conv1DTranspose)	(None, 48, 16)	1,808
dropout_1 (Dropout)	(None, 48, 16)	0
conv1d_transpose_1 (Conv1DTranspose)	(None, 96, 32)	3,616
conv1d_transpose_2 (Conv1DTranspose)	(None, 96, 1)	225

FIGURE 5.3 – Architecture du deuxième modèle

5.4.3 Choix du modèle

En analysant les graphes suivants :

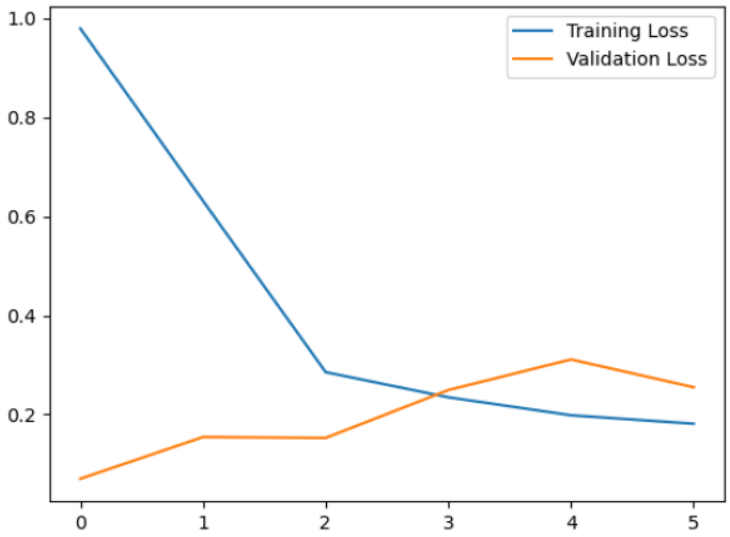

FIGURE 5.4 – Courbe des pertes d'entrainement et de validation du premier modèle

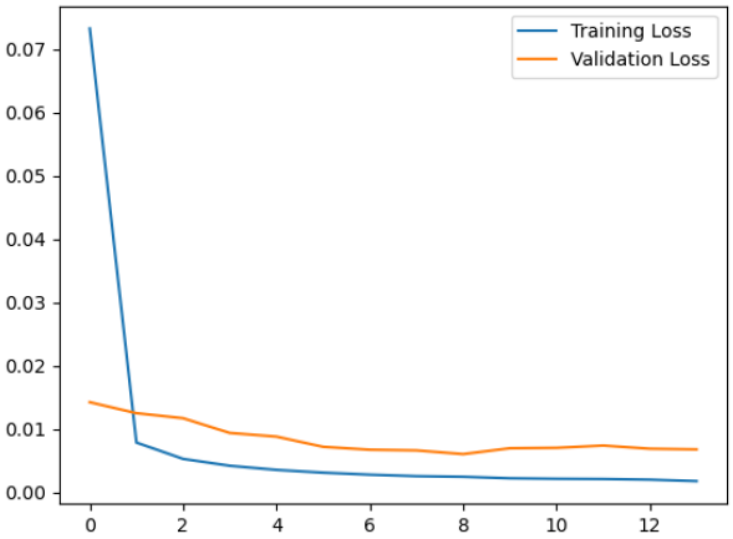

FIGURE 5.5 – Courbe des pertes d'entrainement et de validation du deuxième modèle

Nous remarquons que le deuxième modèle apprend plus vite et converge donc plus rapidement.

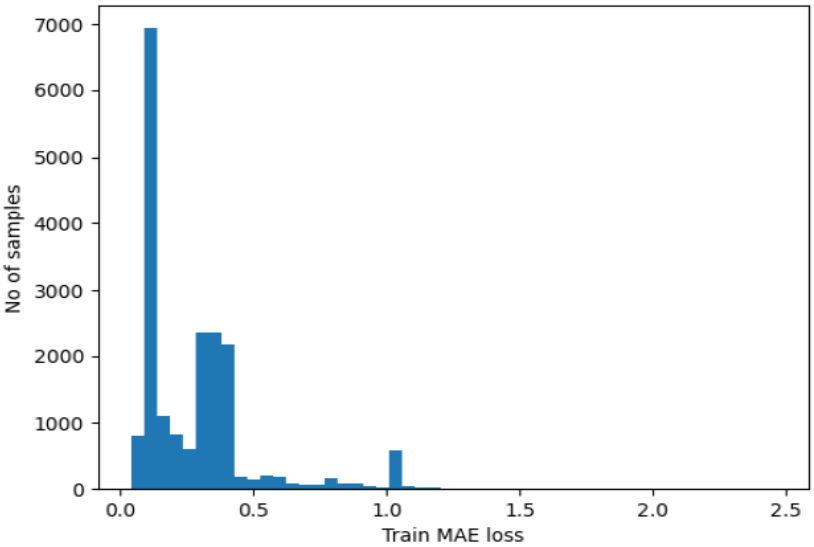

FIGURE 5.6 – Histogramme de la perte Train MAE du premier modèle

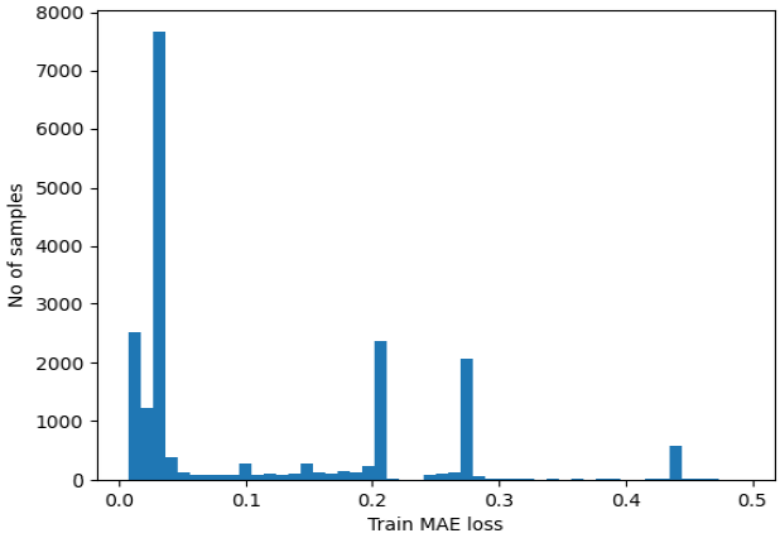

FIGURE 5.7 – Histogramme de la perte Train MAE du deuxième modèle

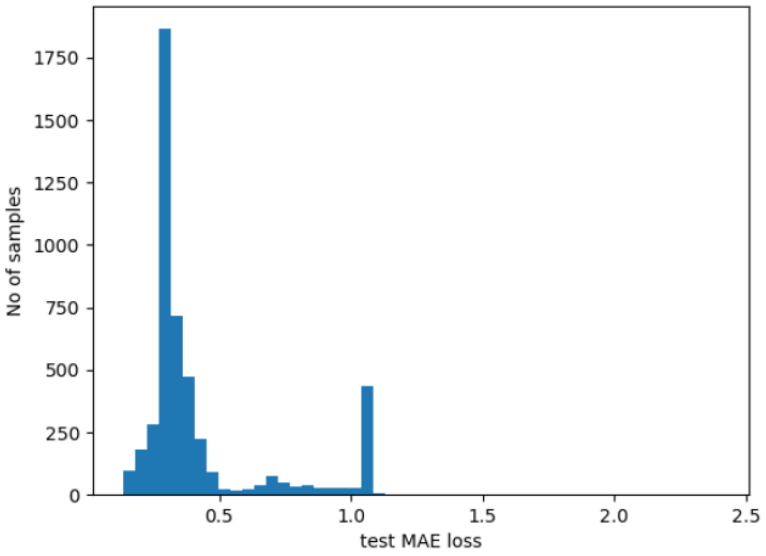

FIGURE 5.8 – Histogramme de la perte Test MAE du premier modèle

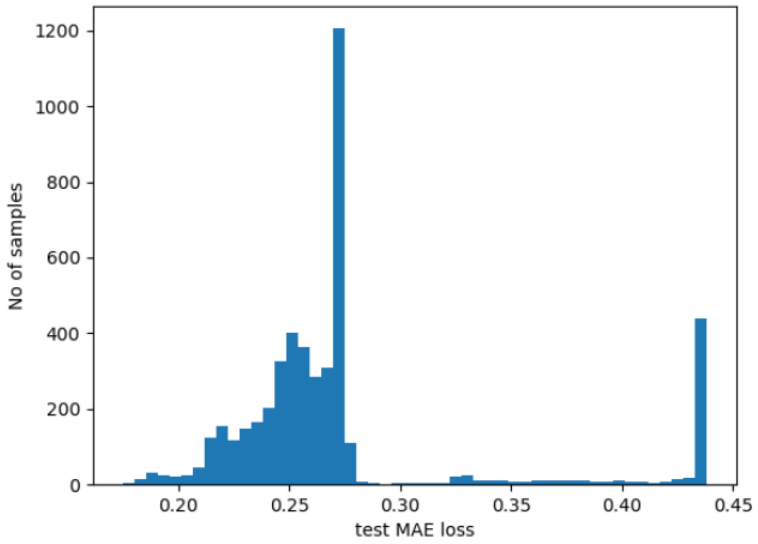

FIGURE 5.9 – Histogramme de la perte Test MAE du deuxième modèle

Et en comparant les histogrammes des pertes MAE des deux modèles, nous constatons qu'ils ne sont pas sur la même échelle. Les pertes du deuxième modèle sont bien plus petites que celles du premier modèle. Nous avons ensuite ajouté à notre deuxième modèle une couche linéaire qui réduit la dimension de la sortie convolutive aplatie en une représentation plus compacte dans l'espace latent, facilitant ainsi la tâche de reconstruction par le décodeur.

5.5 Entraînement

Pour l'entraînement, Lisptick fournit directement trois ensemble de données sur tous les bassins sous forme de mini-batchs avec une clé de hachage pour assurer l'unicité des données. Nous disposons, donc, d'un ensemble d'entraînement, un ensemble de validation et un dernier pour le test. L'apprentissage se fait sur 80% des données, un 10% est réservé pour la validation et le 10% restant pour le test.

5.6 Migration vers pytorch

Après avoir testé nos algorithmes sur des données univariées (les données de pH) d'un seul bassin, nous avons décidé de passer aux données sur plusieurs bassins pour l'entraînement. Cependant, lors de l'implémentation, nous avons découvert que nos codes en Keras rendaient la tâche plus difficile.

Pour y remédier, nous avons décidé de migrer notre code de Keras vers PyTorch, un autre framework d'apprentissage profond largement utilisé.

5.7 Vers données à plusieurs variables

Auparavant, lorsque nous utilisions uniquement la variable pH pour l'entraînement, nous prenions les données moyennes toutes les 15 minutes en incluant uniquement les moments où la pompe filtrait. En passant aux données multivariées, nous utilisons désormais la dernière valeur toutes les 15 minutes, en incluant toutes les données, que la pompe filtre ou non. Cependant, nous fournissons également à notre modèle une autre variable binaire qui indique cette information.

5.8 Optimisation de la détection d'anomalies

5.8.1 Première approche

Initialement, nous avons utilisé l'erreur quadratique moyenne (MSE) comme fonction de perte lors de l'entraînement de notre modèle. La MSE est couramment utilisée pour minimiser les écarts quadratiques entre les prédictions et les valeurs réelles, favorisant ainsi des prédictions précises en pénalisant fortement les erreurs importantes.

Cependant, pour mesurer la perte de reconstruction, nous avons décidé d'utiliser une autre métrique, l'erreur absolue moyenne (MAE), pour définir un seuil d'anomalie. La MAE, contrairement à la MSE, mesure les écarts absolus entre les prédictions et les valeurs réelles, ce qui peut être plus approprié pour la détection des anomalies car elle est moins sensible aux outliers.

Nous avons analysé des séquences de $k = 96$ points en utilisant une fenêtre glissante. Pour chaque fenêtre glissante, nous avons calculé la perte moyenne, ce qui nous a permis de produire un vecteur de pertes de dimension $n - k + 1$ pour n points de données. Cela signifie que pour chaque fenêtre, nous obtenons une valeur de perte moyenne, et en faisant glisser cette fenêtre sur l'ensemble des données, nous obtenons un vecteur qui représente la perte moyenne à chaque étape. Si toutes séquences contenant un point i sont détectés anormales, nous considérons alors que ce point est anormal.

5.8.2 Deuxième approche

Pour cette approche, nous avons adopté une méthode différente pour la perte lors de l'entrainement en héritant de la classe MSELoss pour définir une nouvelle fonction de perte.

Nous avons décidé de calculer la moyenne des pertes pour chaque point de données en utilisant les fenêtres glissantes qui le contiennent et nous l'avons intégrée dans le processus d'entraînement. Cette modification permet d'entraîner le modèle en utilisant une mesure de perte plus détaillée, ce

qui peut potentiellement améliorer la performance globale du modèle et sa capacité à détecter les anomalies.

5.8.3 Troisième approche

Cette fois-ci, nous avons également hérité de la classe MSELoss pour définir une perte contractive. L'idée derrière la perte contractive est d'ajouter un terme de régularisation à la perte afin d'améliorer la robustesse du modèle.

La perte contractive se compose de deux termes : le terme de perte MSE traditionnel et un terme de régularisation qui pénalise les grandes dérivées du modèle par rapport aux entrées. Ce terme de régularisation contient la matrice Jacobienne $\frac{\partial x}{\partial f}$, qui représente les dérivées partielles de la sortie (représentation encodée) par rapport à l'entrée (données d'entrée), capture, alors, la sensibilité de la représentation encodée aux variations des données d'entrée.

L'objectif de la régularisation étant de minimiser la norme de Frobenius de cette matrice Jacobienne, le modèle est encouragé à apprendre des représentations plus stables et plus robustes aux variations dans les données d'entrée , ce qui peut être particulièrement utile pour la détection des anomalies dans des données bruitées ou non stationnaires.

En ajoutant ce terme de régularisation, nous espérons améliorer la capacité du modèle à généraliser à de nouvelles données et à détecter les anomalies de manière plus fiable. Cette approche permet également de contrôler la complexité du modèle et de réduire le risque de surapprentissage, ce qui est crucial pour des applications en détection d'anomalies où les données anormales peuvent être rares et variées.

5.8.4 Quatrième approche

Alors que pour les deux dernières approches, nous avons uniquement modifié la fonction de perte pendant l'entraînement, cette fois-ci, nous changeons également la manière de mesurer l'erreur de reconstruction. Pour l'entraînement, nous utilisons la perte contractive[4], mais pour la reconstruction, il n'est plus nécessaire de parcourir toutes les séquences contenant le point i pour détecter qu'il s'agit d'une anomalie. Cette fois-ci, nous calculons le maximum d'erreur sur chaque fenêtre pendant les sept derniers jours pour déterminer notre seuil d'anomalies. Tout point dépassant ce seuil est considéré comme anormal. Pour l'instant, notre seuil est calculé en se basant uniquement sur l'attribut pH.

$$J_{CAE}(\theta) = \sum_{x \in S}(L(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) + \lambda\Omega(\boldsymbol{h}))$$

$$\Omega(\boldsymbol{h}) = \Omega(f(\boldsymbol{x})) = \left\| \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \right\|_F^2$$

FIGURE 5.10 – Perte contractive

5.9 Résultats

Après avoir testé nos modèles sur une seule variable, le pH, car il est le plus facile à interpréter, nous sommes passés à des données multivariées. Nous avons alors pris en compte les valeurs de pH, chlore, température de l'air, température de l'eau, état de la pompe, ainsi que d'autres paramètres. Après avoir effectué un benchmark entre les deux modèles, nous avons conclu que le TCN convient mieux à notre dataset.

Pour l'entraînement, LispTick fournit directement trois ensembles de données pour tous les bassins sous forme de mini-batchs, avec une clé de hachage pour assurer l'unicité des données. Nous disposons donc d'un ensemble d'entraînement, d'un ensemble de validation et d'un ensemble de tests. L'apprentissage se fait sur 80% des données, 10% sont réservés pour la validation et les 10% restants pour le test.

Pour tester notre modèle, nous avons déterminé le seuil d'anomalies en nous basant uniquement sur les valeurs de pH des sept derniers jours. Le seuil était simplement le maximum des valeurs maximales de chaque fenêtre.

Bien que notre méthode de détection d'anomalies soit plutôt simple, nous avons obtenu des résultats satisfaisants d'après la confirmation avec le client.

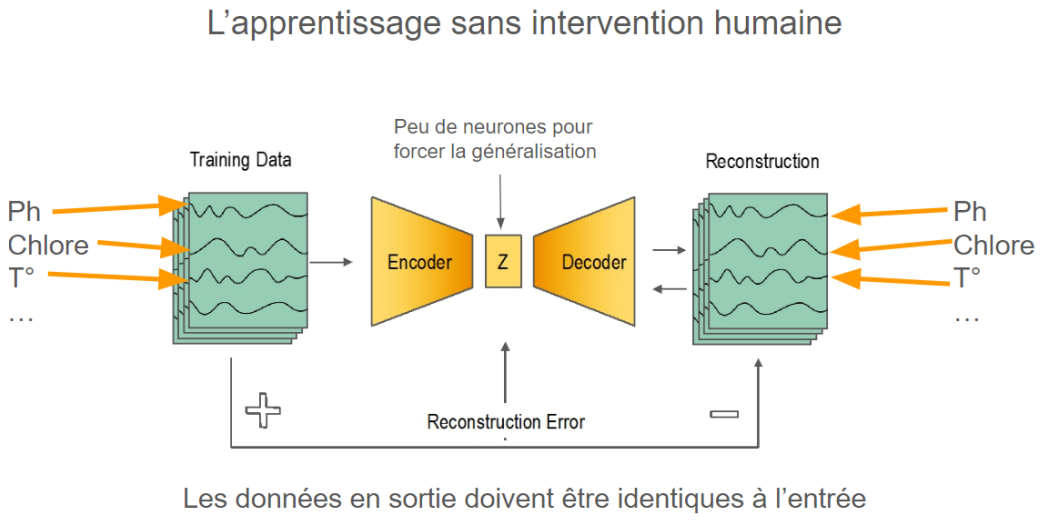

FIGURE 5.11 – Notre modèle

5.9.1 Détection d'anomalies a priori

Nous avons fait tourner notre modèle sur une semaine sur l'ensemble des bassins et nous les avons triés par valeur d'erreurs.

Nous avons communiqué le jour et le bassin avec l'erreur la plus importante et avons eu confirmation par le client qu'il y avait bien des problèmes sur ce bassin qui n'avaient pas été repérés jusque-là.

5.9.2 Détection d'anomalies a posteriori

Inversement le client nous a communiqué un bassin et une journée où il y a eu un problème. Nous avons fait tourner notre modèle sur l'ensemble des bassins sur cette journée et l'erreur était la 11ème plus importante avec une valeur clairement élevée.

Les autres bassins avec des erreurs supérieures venait soit d'être remis en route soit avec effectivement des anomalies.

5.10 Conclusion

Ce chapitre a détaillé les diverses étapes de réalisation du projet, depuis le traitement initial des données brutes jusqu'à l'optimisation finale des modèles de détection d'anomalies. L'approche adoptée a permis une amélioration continue, tant en termes de précision des modèles que d'efficacité des processus. Les résultats finaux montrent une nette amélioration par rapport aux modèles

initiaux, démontrant ainsi la pertinence des choix techniques effectués tout au long du projet.

Chapitre 6

PROJET CONNEXE

6.1 Introduction

Dans le cadre de ce stage, un autre projet connexe a été exploré en raison d'une mise en pause du projet principal. Ce chapitre présente le contexte dans lequel ce projet parallèle a été développé, ainsi que les principales contributions apportées.

6.2 Contexte

En raison des vacances et de la nécessité de gérer le financement du projet principal, celui-ci a été temporairement mis en pause. Nous avons donc décidé de concentrer nos efforts sur un autre projet développé par mon tuteur, qui concerne la classification de documents. Ce projet implique une classification multi-labels, où chaque document peut être associé à plusieurs étiquettes.

6.3 Contributions

Le client n'utilisant que Java, j'étais amenée à adapter notre code pour la mise en production en effectuant les tâches suivantes :

- **Analyse des Modules et du Code :** J'ai procédé à une analyse approfondie du code Python existant. J'ai également étudié les dépendances du projet, tant internes (modules personnalisés) qu'externes (bibliothèques Python), pour m'assurer que leur conversion en Java serait réalisable et efficace.
- **Configuration de l'Environnement de Développement Java :** Étant donné que je n'avais pas travaillé avec Java auparavant à l'entreprise, j'ai configuré un environnement de

développement complet pour ce langage dans Visual Studio Code (VSCode).

- **Utilisation d'ONNX pour l'Interopérabilité :** Afin de faciliter l'intégration du modèle de classification dans le code Java, j'ai utilisé ONNX pour exporter le modèle formé en Python. Cette approche a permis d'assurer la compatibilité entre le modèle existant et le nouveau code Java, sans devoir réimplémenter entièrement le modèle en Java.
- **Conversion partielle du Code Python en Java :** J'ai principalement travaillé sur la conversion du code Python lié à l'inférence vers Java. J'ai veillé à ce que la fonctionnalité d'inférence soit entièrement opérationnelle dans l'environnement Java.
- **Optimisation du Modèle pour les Classifications Multi-Labels :** Étant donné que le modèle a encore du mal à gérer efficacement les classifications multi-labels, je suis en cours d'entreprendre des optimisations spécifiques pour améliorer ses performances dans ce contexte.

Ces contributions ont permis de progresser significativement dans la transition du projet vers un environnement Java, en particulier en ce qui concerne l'inférence des modèles, tout en maintenant la compatibilité avec les systèmes existants utilisés par le client.

6.4 Conclusion

Le projet connexe a joué un rôle complémentaire au projet principal, offrant des perspectives nouvelles et permettant d'aborder des problématiques spécifiques que je n'ai pas eu la chance de faire avec le projet initial. Les contributions effectuées dans ce contexte ont non seulement répondu aux besoins immédiats du client, mais ont également permis de renforcer l'expertise technique sur des aspects cruciaux, tels que l'interopérabilité entre différents environnements de développement. Cette expérience a enrichi le projet global et a ouvert la voie à des développements futurs.

Chapitre 7

GESTION DU PROJET

7.1 Introduction

Dans cette section, nous nous concentrerons sur la gestion du projet de recherche, en examinant la planification, l'organisation et la surveillance. Cette partie du rapport fournira un regard complet sur chaque étape impliquée dans notre projet, depuis sa création jusqu'à sa réalisation. Nous examinerons également les outils, les méthodes et les stratégies que nous avons utilisées pour exécuter avec succès notre projet et atteindre les résultats escomptés. Nous discuterons en profondeur des défis rencontrés au cours du processus et des solutions adoptées pour les surmonter. Ce chapitre contribuera à une compréhension approfondie de la gestion réussie de projets scientifiques et à l'apprentissage de leçons précieuses de notre propre expérience.

7.2 Établissement du cahier des charges

L'établissement du cahier des charges fut une étape cruciale dans le processus de planification du projet. Après discussions avec les clients, nous avons compilé toutes les exigences et spécifications. Ce document détaillé clarifie les objectifs du projet, les fonctionnalités attendues, les contraintes techniques et les délais. Il servit également de référence tout au long du projet pour garantir la conformité aux attentes.

Lors de l'élaboration du cahier des charges, nous avons identifié des éléments cruciaux pour garantir la clarté et la cohérence du projet. Voici les étapes clés du processus :

A) Définition des objectifs du projet :

   Nous avons défini clairement les résultats escomptés et les objectifs à atteindre.

B) Identification des parties prenantes :

Nous avons listé les personnes et les entités impliquées dans le projet, y compris notre client.

C) Délimitation du périmètre :

Nous avons établi les frontières du projet pour déterminer ce qui est inclus et ce qui ne l'est pas. Ensuite nous avons procédé à la rédaction du document.

D) Description des fonctionnalités et des exigences :

Nous avons détaillé les fonctionnalités souhaitées et les exigences nécessaires pour répondre aux besoins de notre client.

E) Établissement d'un planning :

Nous avons créé un calendrier détaillé avec des étapes clés, des jalons et des échéances pour guider l'avancement du projet, aussi bien qu'un diagramme de Gantt estimé.

F) Identification des contraintes :

Nous avons analysé les contraintes potentielles qui pourraient affecter le projet.

7.3 Méthode scrum

Nous avons décidé d'adopter la méthode Scrum pour notre projet qui requiert une forte adaptabilité, collaboration interdisciplinaire, et une capacité à intégrer rapidement les Feedbacks et les découvertes scientifiques. La flexibilité que permet Scrum est essentielle dans un domaine en évolution comme la détection d'anomalies des séries temporelles par l'IA.

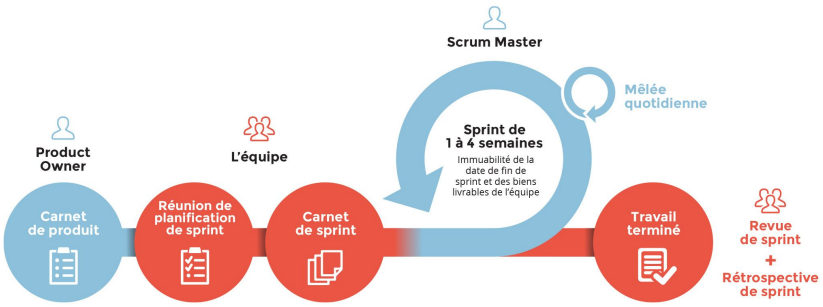

FIGURE 7.1 – Méthode Scrum

7.3.1 Description de la méthode scrum

Notre utilisation de Scrum s'est articulée autour des principaux éléments suivants : Sprints. Le projet a été divisé en périodes de travail intense appelées sprints, d'une durée fixe de deux

semaines. Chaque sprint se concentrait sur un ensemble précis de fonctionnalités ou de tâches de haute priorité.

- Backlog du produit : Nous avons maintenu une liste de tâches en évolution, comprenant toutes les fonctionnalités et exigences du projet, classées par ordre d'importance. Cette liste a été régulièrement révisée et modifiée pour tenir compte des évolutions des besoins du projet.
- Organisation des sprints : À chaque début de sprint, nous avons organisé une séance de planification de sprint pour décider quelles tâches du backlog de produit seraient réalisées pendant le sprint et quels objectifs devaient être atteints.
- Réunions quotidiennes : Chaque jour du sprint, l'équipe organisait une courte réunion appelée Daily Scrum. Lors de cette réunion, l'équipe se synchronise, partage les progrès réalisés, les défis rencontrés et planifier les tâches pour la journée.
- Rétrospectives de sprint : Après la démonstration des tâches terminées lors du sprint, l'équipe a analysé sa performance globale lors d'une réunion de rétrospective de sprint. L'objectif était d'évaluer le déroulement du travail, de dégager les points positifs et les axes d'amélioration, et de déterminer des mesures à mettre en œuvre pour les prochains sprints.

7.3.2 Impact de méthode Scrum sur le progrès de notre projet

L'adaptation de la méthode Scrum nous a permis de mener une bonne planification de notre projet, parmi les biens qu'elle a apportés :

- La flexibilité : Scrum peut facilement s'adapter aux changements et intégrer de nouvelles demandes tout au long du processus de développement.
- La transparence : Non seulement les parties prenantes, mais aussi tous les membres de l'équipe de projet ont une vue complète sur l'état d'avancement du projet à tout moment grâce aux tableaux de bord et aux réunions quotidiennes.

7.4 Équipe du projet

L'équipe du projet est constituée de deux membres principaux, travaillant en étroite collaboration selon la méthodologie Scrum pour garantir une gestion efficace.

Data Engineer (Mon tuteur) : Responsable de l'ingénierie des données, incluant l'intégration, le nettoyage et la préparation des données pour l'analyse.

Data Scientist (Moi) : Chargée de la conception et de l'implémentation des modèles de détection d'anomalies, ainsi que du prétraitement et de l'analyse des données.

7.4.1 Rôles adaptés

Scrum Master : Dans notre petite équipe, le rôle de Scrum Master a été assumé par le Data Engineer (mon tuteur). Il a organisé la planification des sprints, les réunions quotidiennes et les revues de sprint.

Product Owner : Ce rôle a également été pris en charge par le Data Engineer. En tant que Product Owner, il a défini les exigences du projet, priorisé les tâches, et validé les résultats. Cette double fonction a permis une prise de décision rapide et un alignement étroit entre les attentes du projet et le travail réalisé.

Développeur : Nous avions les deux été responsables de l'exécution des tâches techniques.

7.4.2 Pratiques Scrum

Réunions Quotidiennes : Même avec une petite équipe, les réunions quotidiennes se sont révélées cruciales. Elles ont permis de faire le point sur les progrès, d'identifier les obstacles, et de planifier les tâches de la journée. Ces réunions ont été limitées à 15 minutes pour rester efficaces et concentrées.

Revue de Sprint : À la fin de chaque sprint, nous avons organisé une revue pour évaluer les livrables, discuter des réussites et ajuster les priorités en fonction des retours. Cela a assuré que le travail répondait aux attentes du Product Owner et aux objectifs du projet.

Rétrospective de Sprint : Les rétrospectives ont été un moment clé pour analyser ce qui a bien fonctionné et ce qui pouvait être amélioré. Elles ont permis une réflexion continue sur le processus de travail, même avec une petite équipe.

Backlog et Planification des Sprints : Nous avons maintenu un backlog de projet où toutes les tâches et les exigences étaient listées et priorisées. La planification des sprints a été utilisée pour définir les objectifs de chaque sprint et répartir les tâches, facilitant ainsi une gestion agile du projet.

7.5 Outil de planification

Partage de ressources : Dans ce projet, Zotero, un logiciel de gestion de références bibliographiques multiplateforme qui permet de gérer des données bibliographiques et des documents de recherche, était le logiciel qui nous permettait se stocker les papiers de recherche qui nous semblaient pertinents pour notre projet.

Communication : En matière de communication, les courriers étaient notre méthode principale de partage d'informations, de coordination des activités de l'équipe et de communication avec les parties prenantes externes. Cependant, nous avons également utilisé Slack pour des échanges rapides et informels, ce qui a facilité la collaboration en temps réel. De plus, des échanges en personne ont eu

lieu régulièrement, comme mon tuteur était présent dans le bureau, ce qui a permis des discussions directes et une coordination plus efficace.

Planification et gestion des tâches : Nous avons utilisé Jira comme outil central pour planifier, suivre et attribuer les tâches tout au long du projet. Avec Jira, nous avons créé des tableaux de bord de projet, défini des sprints pour chaque semaine, suivi l'avancement des tâches et assigné des responsabilités à chaque membre de l'équipe. Cette plateforme nous a permis d'avoir une vue d'ensemble de l'avancement du projet et de nous assurer que les tâches étaient correctement suivies et gérées.

Diagramme de gantt : Afin de suivre l'avancement du projet et coordonner les tâches d'une façon efficace, nous avons utilisé le digramme de gantt pour s'organiser.

7.6 Gestion des risques

- Risque technologique :
  - Description : Les technologies utilisées peuvent présenter des défis imprévus, y compris des problèmes de surchauffe de l'infrastructure serveur en raison de calculs intensifs.
  - Mitigation : Effectuer une analyse approfondie des technologies avant le début du projet. Mettre en place des mesures de gestion de la température pour prévenir la surchauffe de l'infrastructure serveur.
  - Note **:** Après avoir installé la mise à jour du BIOS datant du 14 août, le serveur fonctionne correctement et les problèmes de surchauffe ont été résolus. Auparavant, nous utilisions une version de mars 2023 qui présentait un problème spécifique lié à Intel, mais cette version a été corrigée avec la mise à jour récente.

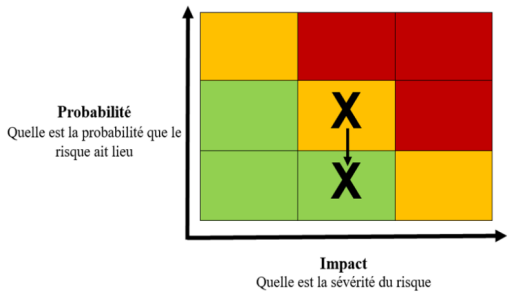

FIGURE 7.2 – Matrice de risque 1

- Risque de délais :

- Description : Les retards dans le développement et l'optimisation de la méthode peuvent compromettre les délais du projet.
- Mitigation : Élaborer un calendrier réaliste en tenant compte des éventuels retards. Planifier des jalons intermédiaires pour évaluer les progrès et réajuster si nécessaire.

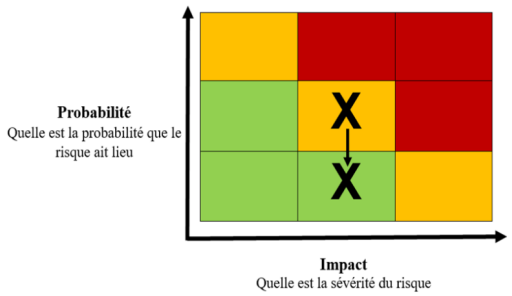

FIGURE 7.3 – Matrice de risque 2

- Risque de données :
  - Description : La qualité des données utilisées peut influencer directement la précision de la méthode.
  - Mitigation : Assurer la qualité et la diversité des données dès le départ.

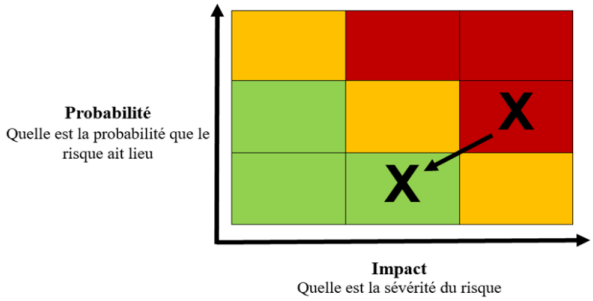

FIGURE 7.4 – Matrice de risque 3

7.7 Coûts et budget

La réalisation du projet de piscine intelligente implique différents éléments de coûts qui doivent être pris en compte. Les principaux éléments de coûts sont les suivants :

Élément de coût	Estimation
Salaires	Salaires brut + 40% des charges sociales
Coûts des infrastructures technologiques	Entre 1500€ et 2000€ pour chaque ordinateur portable 4000€ pour le serveur IA booster 300€ par mois pour Lisptick
Frais de déplacement et d'hébergement	Variables en fonction des déplacements nécessaires

TABLE 7.1 – Estimation des coûts du projet

7.8 Planification et suivi du Projet

7.8.1 Diagramme de Gantt prévu

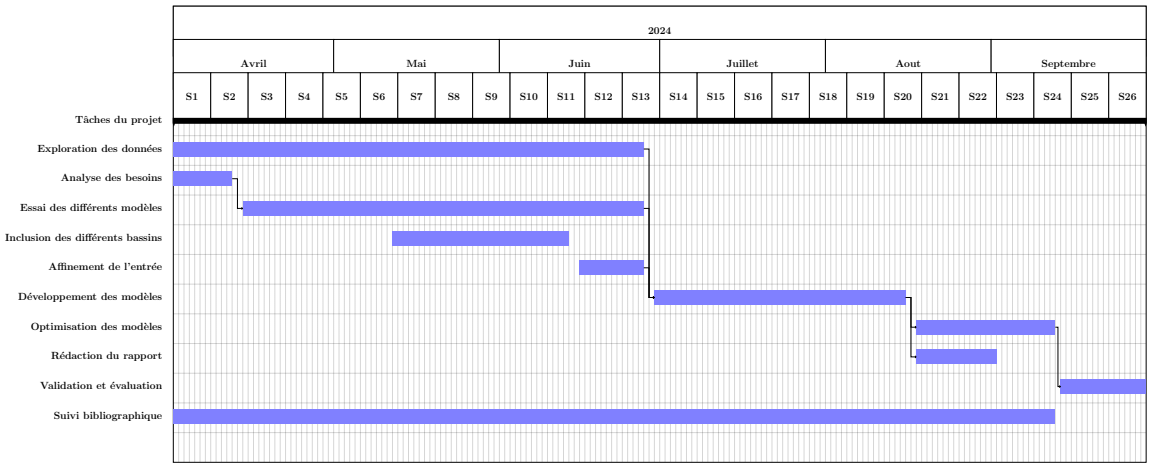

FIGURE 7.5 – Diagramme de Gantt prévu

7.8.2 Diagramme de Gantt réel

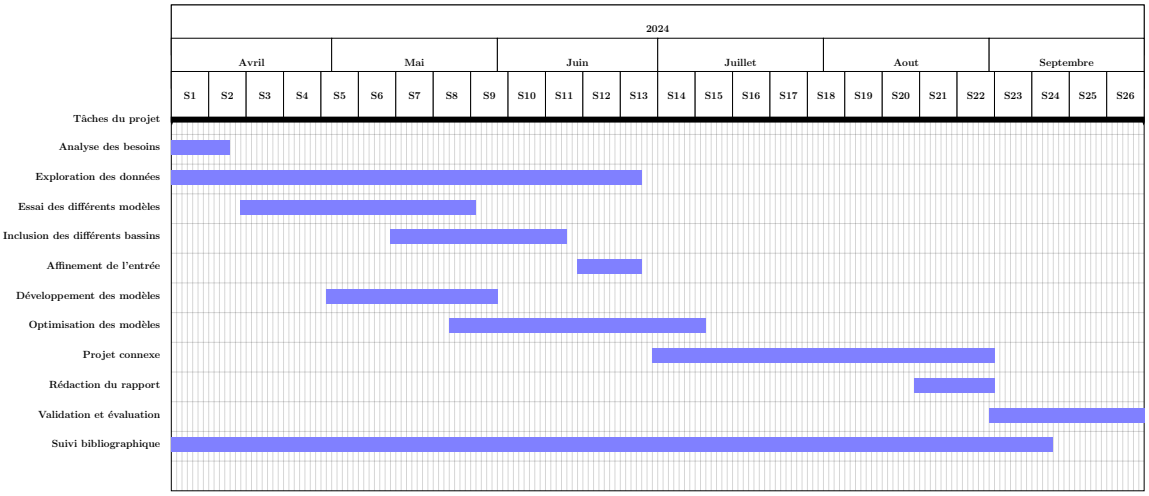

FIGURE 7.6 – Diagramme de Gantt réel

7.9 Conclusion

La gestion du projet a été structurée autour de la méthodologie Agile, et plus particulièrement du cadre Scrum. Cette approche a permis une organisation flexible et réactive, adaptée aux besoins évolutifs du projet. Grâce aux sprints réguliers, nous avons pu segmenter le travail en phases gérables, tout en assurant une communication continue avec l'équipe et les parties prenantes.

Les réunions quotidiennes (daily stand-ups) ont favorisé une collaboration étroite et une résolution rapide des problèmes, tandis que les rétrospectives de sprint ont offert des opportunités précieuses pour affiner les processus et améliorer notre efficacité collective. L'accent mis sur l'incrémentalité a permis de livrer des versions fonctionnelles du projet à intervalles réguliers, offrant ainsi une visibilité constante sur l'avancement du travail et la possibilité de réajuster les priorités en fonction des retours du client.

En somme, l'application de la méthode Scrum dans ce projet a non seulement structuré notre approche, mais a également renforcé l'engagement de l'équipe et la qualité des livrables, assurant ainsi une gestion de projet efficace et orientée vers les résultats.

Chapitre 8

RÉFLEXION MÉTHODOLOGIQUE

8.1 Introduction

La méthodologie constitue la colonne vertébrale de tout projet scientifique et technique. Dans le cadre de notre étude sur la détection d'anomalies dans les séries temporelles des paramètres de piscines, il est crucial d'examiner et de justifier les choix méthodologiques réalisés jusqu'à présent. Bien que le projet soit toujours en cours, cette réflexion vise à éclairer les décisions prises, à évaluer les résultats obtenus à ce stade, et à identifier les forces et les faiblesses de notre approche. Enfin, nous aborderons les leçons apprises et les perspectives futures qui guideront les prochaines étapes du projet.

8.2 Choix méthodologiques

8.2.1 Utilisation des données brutes

L'accès aux données brutes a été privilégié pour travailler directement avec des informations non filtrées, minimisant ainsi les risques d'altérations ou de pertes de données. Cette approche vise à garantir la qualité des modèles développés en conservant l'intégrité des données dès le départ.

8.2.2 Normalisation et fenêtres glissantes

La normalisation des attributs a été nécessaire pour harmoniser les échelles des différentes variables et éviter que certaines dominent lors de l'entraînement des modèles. Par ailleurs, l'utilisation

des fenêtres glissantes a permis de segmenter les séries temporelles en portions plus petites et plus cohérentes, facilitant ainsi la détection de motifs récurrents et d'anomalies. Le choix de la taille de la fenêtre et du pas reste un compromis délicat entre la granularité des données et la complexité computationnelle, et pourrait nécessiter des ajustements futurs en fonction des résultats obtenus.

8.2.3 Sélection des modèles

Le choix entre les architectures LSTM et TCN a été guidé à la fois par des considérations théoriques et empiriques. Bien que les LSTM soient réputés pour capturer les dépendances à long terme, notre analyse a montré que les TCN, grâce à leur architecture convolutionnelle, offraient une convergence plus rapide et des performances supérieures pour notre jeu de données spécifique. Cette observation souligne l'importance de valider empiriquement les modèles sélectionnés.

8.2.4 Intégration des auto-encodeurs

L'intégration d'auto-encodeurs dans les architectures LSTM et TCN a permis de capturer et de reconstruire les caractéristiques essentielles des séquences temporelles, ce qui a amélioré la détection des anomalies. Un des avantages majeurs des auto-encodeurs est qu'ils ne nécessitent pas de labels, ce qui représente un atout considérable, étant donné que l'étiquetage des données peut être très chronophage. Cependant, l'optimisation des hyperparamètres, tels que la dimensionnalité des couches et les fonctions d'activation, demeure un défi itératif crucial pour améliorer les performances des modèles.

8.3 Difficultés rencontrées

8.3.1 Difficultés matérielles

L'une des principales difficultés matérielles a été la gestion et le traitement des données brutes. Les données en grande quantité nécessitent des ressources informatiques importantes pour le prétraitement et l'analyse, ce qui a parfois conduit à des ralentissements et des limitations en termes de capacité de traitement. De plus, les capteurs utilisés pour collecter les données avaient une précision variable, entraînant des écarts qui ont compliqué l'étape de prétraitement.

8.3.2 Difficultés liées à la validation des résultats

La validation des résultats de la détection d'anomalies est une tâche complexe en l'absence d'une base de données spécifique aux anomalies, ce qui rend difficile la corrélation entre les anomalies détectées et les incidents réels ou les dysfonctionnements. Cette difficulté a conduit à la suspension du projet en attendant l'obtention de données de validation adéquates.

8.4 Bilan humain

L'une des principales leçons tirées de ce projet est l'importance d'une approche itérative et flexible dans la conception des modèles. La capacité à adapter les méthodes en fonction des résultats intermédiaires et des besoins spécifiques du projet s'est avérée cruciale pour parvenir à des résultats robustes et pertinents. De plus, la collaboration étroite avec le client, en s'assurant que les anomalies détectées correspondent à des événements réels, a été un facteur clé de validation de notre approche.

Avec du recul, je constate que certaines actions auraient pu être mieux gérées. Par exemple, j'aurais pu accorder plus de temps à l'anticipation des besoins en données de validation dès les premières étapes du projet. Cela m'aurait permis de mieux préparer le terrain pour les phases de validation, évitant ainsi des retards. De plus, j'aurais pu organiser mon temps de manière plus efficace, en évitant de passer trop de temps sur certains aspects techniques au détriment d'autres tâches importantes. Une révision plus fréquente des objectifs et des résultats intermédiaires aurait également été bénéfique pour détecter et corriger rapidement les éventuelles erreurs ou inefficacités, ce qui aurait accéléré le projet dans son ensemble.

8.5 Perspectives futures

Pour les travaux futurs, nous prévoyons de mettre en place une "base" des anomalies qui seront à la fois exclues de l'apprentissage et serviront de tests De plus, l'intégration de nouvelles variables et l'extension à des données provenant de différentes sources ouvriront de nouvelles perspectives pour enrichir notre modèle et accroître sa robustesse.

8.6 Conclusion

Cette réflexion méthodologique a permis d'évaluer les choix stratégiques et les défis rencontrés jusqu'à présent dans notre projet de détection d'anomalies dans les séries temporelles. Bien que le projet soit en cours, les leçons apprises guideront les développements futurs et contribueront à l'amélioration continue de nos méthodes et modèles. Le travail restant se concentrera sur l'affinement des techniques existantes et l'exploration de nouvelles approches pour atteindre les objectifs fixés.

CONCLUSION

Ce projet de détection d'anomalies dans les séries temporelles a représenté un défi complexe, tant sur le plan technique qu'organisationnel. En travaillant sur ce projet, j'ai pu approfondir mes connaissances en traitement des séries temporelles, en recherche scientifique et en gestion de projet. Malgré les difficultés rencontrées, notamment l'absence d'une base de données spécifique pour valider les anomalies détectées, cette expérience a été extrêmement enrichissante.

La suspension temporaire du projet principal en raison des vacances et des contraintes de financement du client m'a permis de découvrir que des pauses ou interruptions de projet peuvent survenir, même lorsque tout semble bien planifié. Cette expérience m'a offert une nouvelle perspective sur l'importance de la flexibilité et de l'adaptabilité dans la gestion de projets, ainsi que sur la nécessité d'être préparé à gérer de tels imprévus de manière proactive.

Cependant, cette situation a également offert une opportunité précieuse : la réalisation d'un projet connexe qui m'a permis d'aborder la mise en production et d'appliquer les compétences acquises dans un contexte pratique. Ce projet secondaire a non seulement complété mon apprentissage mais a également jeté les bases pour une transition fluide vers la phase de mise en production du projet principal. Cela a été une expérience précieuse qui a renforcé mes compétences en gestion de projets et en développement pratique.

En rétrospective, l'un des aspects les plus instructifs a été l'importance d'une gestion efficace du temps et d'une évaluation régulière des progrès. J'aurais dû organiser mon emploi du temps de manière plus équilibrée, en allouant du temps suffisant à chaque phase du projet et en intégrant des mécanismes d'auto-évaluation plus fréquents.

En conclusion, ce projet m'a permis de développer une compréhension approfondie des défis liés aux séries temporelles et à la gestion de projets scientifiques. Les leçons tirées, tant des succès que des difficultés, me guideront dans mes futurs travaux. L'expérience acquise, notamment par le biais du projet connexe, constitue un atout majeur pour la mise en production de projets similaires à l'avenir. Je suis convaincue que les compétences et les connaissances développées au cours de ce projet seront précieuses dans ma carrière professionnelle et pour la réalisation de projets futurs.

RÉSUMÉ

Dans le contexte de l'évolution rapide des technologies et des attentes accrues des consommateurs, l'industrie des piscines résidentielles cherche à adopter des solutions intelligentes pour simplifier la gestion des bassins. Ce projet de fin d'études se concentre sur le développement d'un système de détection d'anomalies dans les données de capteurs des piscines, en utilisant des algorithmes d'intelligence artificielle.

Le projet a été mené en quatre étapes principales : l'analyse des besoins des utilisateurs, le choix des outils technologiques, le développement d'algorithmes de détection d'anomalies, et l'évaluation de leur performance. Les méthodes employées incluent le prétraitement des séries temporelles issues des capteurs et l'utilisation de modèles de machine learning pour identifier les anomalies.

Les résultats obtenus montrent que les algorithmes développés permettent de détecter efficacement les anomalies à partir des paramètres de l'eau et les performances des équipements, avec une précision élevée. Ces résultats offrent des perspectives prometteuses pour une gestion plus proactive et automatisée des piscines, réduisant ainsi la consommation énergétique.

En conclusion, ce projet propose une solution innovante pour l'industrie des piscines résidentielles, alliant confort, efficacité, et durabilité grâce à l'intégration de technologies avancées.

ABSTRACT

In the context of rapidly evolving technologies and increasing consumer expectations, the residential swimming pool industry is seeking to adopt smart solutions to simplify pool management. This final-year project focuses on the development of a system for detecting anomalies in swimming pool sensor data, using artificial intelligence algorithms.

The project was conducted in four main steps : analyzing user needs, choosing technological tools, developing anomaly detection algorithms, and evaluating their performance. The methods used include preprocessing time series from sensors and using machine learning models to identify anomalies.

The results obtained show that the developed algorithms can effectively detect anomalies based on water parameters and equipment performance, with high precision. These results offer promising prospects for more proactive and automated management of swimming pools, thus reducing energy consumption.

In conclusion, this project offers an innovative solution for the residential swimming pool industry, combining comfort, efficiency, and durability through the integration of advanced technologies.

Bibliographie

[1] Saroj Gopali et al. *A Comparative Study of Detecting Anomalies in Time Series Data Using LSTM and TCN Models*. 2021. arXiv : 2112.09293 [cs.LG].

[2] Yangdong He et Jiabao Zhao. "Temporal Convolutional Networks for Anomaly Detection in Time Series". In : *Journal of Physics: Conference Series* 1213.4 (juin 2019), p. 042050. doi : 10.1088/1742-6596/1213/4/042050. url : https://dx.doi.org/10.1088/1742-6596/1213/4/042050.

[3] Pankaj Malhotra et al. "Long Short Term Memory Networks for Anomaly Detection in Time Series." In : *ESANN*. 2015. url : http://dblp.uni-trier.de/db/conf/esann/esann2015.html#MalhotraVSA15.

[4] Salah Rifai et al. "Higher Order Contractive Auto-Encoder". In : *Machine Learning and Knowledge Discovery in Databases*. Sous la dir. de Dimitrios Gunopulos et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2011, p. 645-660. isbn : 978-3-642-23783-6.

[5] Mayu Sakurada et Takehisa Yairi. "Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction". In : *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. MLSDA'14. Gold Coast, Australia QLD, Australia : Association for Computing Machinery, 2014, p. 4-11. isbn : 9781450331593. doi : 10.1145/2689746.2689747. url : https://doi.org/10.1145/2689746.2689747.

[6] Markus Thill, Wolfgang Konen et Thomas Bäck. "Time Series Encodings with Temporal Convolutional Networks". In : *Bioinspired Optimization Methods and Their Applications*. Sous la dir. de Bogdan Filipič, Edmondo Minisci et Massimiliano Vasile. Cham : Springer International Publishing, 2020, p. 161-173. isbn : 978-3-030-63710-1.

[7] Guoqiang Peter Zhang. "Time series forecasting using a hybrid ARIMA and neural network model." In : *Neurocomputing* 50 (2 juin 2003), p. 159-175. url : http://dblp.uni-trier.de/db/journals/ijon/ijon50.html#Zhang03.