UNIVERSITÉ DU QUÉBEC EN ABITIBI-TÉMISCAMINGUE

PREDICTION INTELLIGENTE DES DEFAILLANCES DANS LES RESEAUX DE TUYAUTERIE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INGÉNIERIE

PAR

YASSINE KANOUN

AOUT 2023

REMERCIEMENTS

C’est avec un grand plaisir que je réserve cette page en signe de gratitude et de profonde reconnaissance à tous ceux qui m’ont aidé à la réalisation de ce travail.

Tout d'abord, je remercie Dieu de m'avoir donné la santé, la puissance et les conseils pour mon programme de maîtrise et toute ma vie. J'ai une dette de gratitude envers un certain nombre de personnes qui ont contribué à cette recherche. Ce travail n'aurait pas été réalisé sans leurs commentaires perspicaces, leur soutien et leurs encouragements.

Avant tout, je tiens à exprimer ma plus sincère gratitude à mon superviseur, le professeur Hatem Mrad, pour son encadrement, son professionnalisme, ses conseils et ses encouragements tout au long du processus d'élaboration de cette recherche. Je souhaite également reconnaître la contribution du professeur Tikou Belem et Bassem Zouari qui m'ont apporté toute l'aide dont j'avais besoin pour structurer mon étude.

Je tiens à remercier toute l’équipe pédagogique de l’Université de Québec en Abitibi-Témiscamingue, de l’École nationale d’ingénieurs de Sfax et les intervenants professionnels responsables de la formation d’ingénierie, pour avoir assuré la partie théorique de celle-ci.

J'adresse mes remerciements, pour l’équipe de la compagnie Norda-Stelo, à Messieurs Joël Fortin, François Grégoire et madame Alexia Blanchard-Lapierre, pour leur sympathie et soutien continu tout le long de mon travail.

Je suis redevable à mes parents et mes chères sœurs pour leur amour, leurs prières et le don de la vie. Je dois les remerciements à ma fiancée Mariem pour m'avoir donné la force et la motivation de poursuivre mes recherches.

Enfin, je remercie tous mes chers amis Ahmed, Yassine, Oussema, Hassen et Mohamed.

Table des matières

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES SIGLES ET ABRÉVIATIONS

ASTM: American Society for Testing and Materials

ASME: American Society of Mechanical Engineers

API: American Petroleum Institute

AWWA: American Water Works Association

ACG: Association Canadienne Du Gaz

CONCAWE: Association européenne des compagnies pétrolières

UT: Ultrasons

SCADA: Système de surveillance et d'acquisition de données

ILI : Inspection en ligne

CM: Surveillance de l'état (Condition Monitoring)

IA: Intelligence artificielle

RBI : Analyse basée sur les risques

DNV: Det Norske Veritas

CML : Lieux de surveillance de l'état (Condition Monitoring Location)

TML : Lieu de surveillance de l'épaisseur (Thickness Monitoring Location)

CUI: Corrosion sous isolation

NDT: Examen non destructif

RL : Durée de vie restante (remaining life)

VT: Inspection visuelle

ET: Contrôle par courants de Foucault

MFL: Contrôle par particules magnétiques

RT: Radiographie

EMAT: Transducteur acoustique électromagnétique

ANN: Réseau de neurone artificiel

SVM: machine à vecteurs de support

K-NN: K-voisin le plus proche

DOT : Département des Transports des États-Unis

UAV: Véhicules aériens sans pilote

CNN: Réseau neuronal convolutif

BD : Base de données

AA : Apprentissage automatique

LSTM: Long Short-Term Memory

NB: Naïve Bayes

IF: Forêt d’isolement (Isolation Forest)

LOF: Local Outlier Factor

TAN : Indice d'acidité totale

PoF : probabilité de défaillance

CoF : Conséquence de la défaillance.

AUC: Aire sous la courbe (Area Under the Curve)

ROC: Caractéristiques de fonctionnement du récepteur (Receiver Operating Features)

VP: Vrais positifs

FP: Faux Positif

FN: Faux Négatifs

VN: Vrais négatifs

A : Taux de succès

P : Précision

R : Rappel

S : Spécificité

TVP= Sensibilité

TFP =1- Spécificité

MAE : Erreur absolue moyenne

MSE : Erreur quadratique moyenne

R^2 : coefficient de détermination

MPF : "Main Process Flow", Région principale du circuit.

DL : Deadleg, Point mort.

IP : Point d'injection

MX : Point de mélange

LISTE DES ANNEXES

RÉSUMÉ

Les systèmes de tuyauterie sont des composants essentiels dans diverses industries, notamment le pétrole et le gaz, le traitement de l'eau et l'industrie manufacturière. Cependant, ces systèmes sont souvent sujets à des défaillances dues à divers facteurs tels que la corrosion, la fatigue et les facteurs environnementaux. Ce qui peut entraîner des pertes financières considérables, des arrêts de production et des risques pour la sécurité.

La prédiction des défaillances dans les systèmes de tuyauterie est une tâche importante pour assurer la sécurité et la fiabilité des installations industrielles. Les approches traditionnelles de surveillance des systèmes de tuyauterie reposent sur des inspections périodiques et des analyses manuelles des données, ce qui peut prendre du temps et être sujet à des erreurs. Les algorithmes d'apprentissage automatique (AA) se sont révélés prometteurs pour la maintenance prédictive dans divers domaines, y compris les systèmes de tuyauterie. Ils exploitent les données historiques pour prédire les défaillances et permettre une maintenance préventive.

Dans cette étude, nous traitons les différents types de défaillances qui peuvent survenir dans les installations de tuyauterie et les défis associés à la détection et à la prévention de celles-ci. Nous examinons également les différentes techniques d'AA qui peuvent être utilisées pour prédire et prévenir les défaillances, tels que l'apprentissage supervisé et non supervisé. Les données historiques tels que les rapports d'inspection des systèmes de tuyauterie, y compris les conditions opératoires, les données de mesure d'épaisseur et d'autres caractéristiques pertinentes, ont été utilisés pour développer les modèles d'AA. Les modèles de régression et classification, tels que les arbres de décision, les forêts aléatoires et les machines à vecteur support, sont utilisées pour prédire le taux de corrosion, son niveau de sévérité et les sources de défaillances. Les techniques de clustering, tels que K-means et k-prototype, sont utilisées pour développer les boucles de corrosion. D'autres approches d'AA non-supervisé sont implémentées pour détecter les anomalies dans les systèmes de surveillance.

Nous évaluons les performances des différents algorithmes à l'aide de diverses mesures de performance, notamment la métrique R^2 pour les problèmes de régression, et la précision ainsi que la courbe AUC-ROC pour les problèmes de classification. Les résultats montrent que les méthodes basées sur les arbres de décision ont une bonne performance en termes de précision pour réaliser la prédiction et répondre à nos objectifs.

Nous effectuons également une analyse de l'importance des prédicteurs afin d'identifier les paramètres les plus pertinents pour réaliser la prédiction de la sévérité de la corrosion et la source de défaillance. Les résultats montrent que les conditions opératoires, tels que la pression, la température, le débit et le type de fluide circulant, sont les caractéristiques les plus déterminantes pour cette étude.

Notre étude démontre le potentiel des algorithmes d'apprentissage automatique pour la prédiction des défaillances. Cependant, il est crucial d'aborder ce processus avec prudence et en comprenant bien ses limites et ses exigences.

Mots clés : Maintenance, Prédiction, Défaillance, Données, Apprentissage automatique

ABSTRACT

Pipeline systems are vital components in various industries, including oil and gas, water treatment, and manufacturing. However, these systems are often prone to failures due to factors such as corrosion, fatigue, and environmental factors, which can result in significant financial losses, production downtime, and safety risks.

Predicting failures in pipeline systems is an important task to ensure the safety and reliability of industrial facilities. Traditional approaches to pipeline system monitoring rely on periodic inspections and manual data analysis, which can be time-consuming and prone to errors. Machine learning algorithms have shown promise in predictive maintenance in various fields, including pipeline systems. They leverage historical data to predict failures and enable preventive maintenance.

In this study, we address the different types of failures that can occur in pipeline installations and the associated challenges in detecting and preventing them. We also explore various machine learning techniques that can be used for failure prediction and prevention, such as supervised and unsupervised learning. Historical data, such as pipeline system inspection reports including operating conditions, thickness measurement data, and other relevant features, were used to develop the machine learning models. Regression and classification models, such as decision trees, random forests, and support vector machines, are employed to predict corrosion rates, severity levels, and failure sources. Clustering techniques, such as K-means and k-prototype, are utilized to develop corrosion loops. Other unsupervised machine learning approaches are implemented to detect anomalies in monitoring systems.

We evaluate the performance of different algorithms using various performance metrics, including the R2 metric for regression problems and accuracy as well as the AUC-ROC curve for classification problems. The results show that tree-based methods perform well in terms of accuracy to achieve prediction and meet our objectives.

We also conduct an analysis of predictor importance to identify the most relevant parameters for predicting corrosion severity and failure sources. The results demonstrate that operating conditions, such as pressure, temperature, flow rate, and the type of fluid being transported, are the most influential features for this study.

Our study highlights the potential of machine learning algorithms for failure prediction. However, it is crucial to approach this process with caution and a thorough understanding of its limitations and requirements.

Keywords: Maintenance, Prediction, Failure, Data, Machine learning

CHAPITRE 1: INTRODUCTION

1.1 Contexte de l'étude

Le premier pipeline a été construit par Benson en 1869 pour éviter le monopole de Rockefeller sur le transport ferroviaire du pétrole. Des progrès technologiques importants ont été réalisés dans tous les domaines, en particulier depuis les années 1950. Les frais de port ont été réduits partout, mais certains processus sont beaucoup plus chers que d'autres. En effet, l'intérêt croissant pour les sources d'énergie, par exemple le pétrole, le gaz et d'autres hydrocarbures nécessite progressivement le développement de nouvelles lignes de pipelines afin de satisfaire les besoins de l'industrie.

Les États-Unis ont le plus long réseau de pipelines dans le monde, avec 1 984 321 km pour le transport de gaz naturel et 240 711 km pour le transport des produits pétroliers. Le deuxième pays ayant le plus de kilomètres de pipelines est la Russie avec 163 872 km, puis le Canada avec 100 000 km. [1]

Quel que soit l'endroit où le système de tuyauterie est installé, ce dernier risque d'être endommagé. Le comportement des pipelines peut être quantifié par l'interaction interne liée aux paramètres du procédé (température, pression débit, etc) ou bien à d'autres interactions, telles que les contraintes et les réactions externes développées sous des charges appliquées. Les valeurs admissibles pour chacun de ces paramètres sont définies après l'examen des critères de défaillance appropriés pour le système. Les critères de réponse et de défaillance du système dépendent de la nature des charges, qui peuvent être classées selon diverses distinctions, telles que primaire vs secondaire, soutenue vs occasionnelle ou statique vs dynamique.

Cette défaillance est le résultat d'un affaiblissement structurel couplé à des contraintes imposées de l'extérieur et de l'intérieur [2] et entraîne des coûts environnementaux, économiques et sociaux très élevés. Une moyenne de 850 défaillances de conduites d'eau se produit quotidiennement en Amérique du Nord, avec un coût total annuel de réparation de plus de 3 milliards de dollars [3]. Selon les statistiques, plus de 3 000 défaillances se sont produites dans des gazoducs situés aux États-Unis depuis 1986, entraînant des dégâts matériels de plus d'un milliard de dollars. Ces chiffres révèlent l'importance de maintenir ces installations dans de bonnes conditions afin de pouvoir prévenir les défaillances [4].

En raison de ces difficultés, des modèles prédictifs ont été mis au point pour prévoir la probabilité de défaillance des conduites de manière préventive et contribuer aux plans de gestion des actifs.

Le défi consiste à développer des modèles fiables pour prédire les besoins futurs de renouvellement de chaque conduite constituant le réseau. L'expérience a montré qu'un nombre important de réparations est effectuée de manière non programmée [5]. Cette maintenance réactive présente l'inconvénient que les dommages peuvent se produire avant que des mesures soient prises. Avec cette stratégie de maintenance, les conduites réhabilitées sont sélectionnées en fonction de critères d'urgence, tels que le nombre de ruptures sur la conduite actuelle. Une alternative à la stratégie réactive est une stratégie proactive. Dans une stratégie proactive, le service détermine les besoins de maintenance en tenant compte de l'état des conduites et en prévoyant leur dégradation. Les défaillances des canalisations entraînent des coûts et des désagréments considérables [5]. Comme il n'est pas pratique ou économiquement possible de réhabiliter toute la longueur du réseau, un ciblage des ressources de réhabilitation est nécessaire. Avec des ressources limitées, la capacité d'éviter les dommages et d'optimiser l'utilisation des fonds disponibles pour la maintenance préventive en employant des modèles prédictifs est une option privilégiée pour la gestion des réseaux de tuyauterie [5]. La stratégie prédictive nécessite une bonne connaissance des caractéristiques du réseau, y compris les facteurs de détérioration et l'historique des défaillances. Cela implique que l'installation doit avoir une base de données (BD) informatisée, de préférence sous la forme d'un système d'information géographique. À ce jour, les avantages d'une approche proactive par rapport à une approche réactive n'ont pas été démontrés. Toutefois, cela pourrait être dû à l'insuffisance des modèles d'évaluation utilisés jusqu'à présent [5].

1.2 Problématique

L’ingénierie de la tuyauterie dans les installations industrielles offre des solutions à divers problèmes de mise en œuvre des réseaux de transport de fluides. En effet, cette discipline étudie l’efficacité et la fiabilité des transports afin d’éviter l’arrêt d’une ligne de production ou même l’arrêt complet d’une usine. Un autre défi à cette discipline est celui de la sécurité du personnel travaillant en chantier [6]. Il

constitue de nos jours un des objectifs principaux de cette ingénierie puisqu'une seule défaillance au niveau d'un tuyau peut entraîner un sérieux impact autant sur la santé et la sécurité des opérateurs que sur l'environnemental. Chaque système de tuyauterie est soumis à différentes sollicitations et contraintes au cours de sa durée de vie. Les conditions opératoires liées à l'installation, les propriétés des fluides et le matériau utilisé des tuyaux contribuent au processus de vieillissement et de dégradation de ces réseaux.

En fonction de l'état de distribution des contraintes, le système est classé comme un système critique ou non. C'est la raison pour laquelle, il est nécessaire d'assurer une bonne analyse de l'état de la tuyauterie afin de se conformer aux normes applicables. Cette analyse permet de vérifier que les tuyauteries sont bien soutenues et que l'installation résiste bien aux différentes conditions opératoires. De plus, le surdimensionnent des équipements installés, la présence des défauts de soudure, le non-équilibrage des couples de serrage de fixation et d'assemblage des brides provoquent inévitablement un désalignement des composants et des contraintes résiduelles permanentes au niveau du système de tuyauterie [6].

Afin d'améliorer la fonctionnalité du système de tuyauterie et prolonger sa durée de vie utile en fonctionnant sans fuites ni pannes accidentelles, il est nécessaire de faire des interventions sur l'installation industrielle. Traditionnellement, la maintenance était effectuée en réparant ou en remplaçant les pièces de manière réactive après leur panne. De plus, une maintenance préventive a été utilisée en changeant et en entretenant les composants du système avant qu'elles ne tombent en panne. Les développements vers l'industrie 4.0 et l'apprentissage automatique (AA) ont apporté de nouvelles solutions aux modèles de prédiction des défaillances. Il est maintenant possible de planifier la maintenance plus efficacement, d'anticiper les anomalies, de réduire les temps d'arrêt dus à des pièces usées ou à des travaux de maintenance qui ne sont pas réellement nécessaires [6]. Plusieurs industriels ont annoncé ou déjà amorcé le virage vers cette révolution. Plus particulièrement, l'installation d'un nombre considérable de capteurs (vibration, température, pression, gaz, etc.) pour le monitoring des procédés industriels à travers les endroits les plus risqués a permis d'obtenir des bases de données, structurées, massives et hétérogènes. Ce choix stratégique concerne non seulement la sécurité fonctionnelle

des machines et la productivité, mais aussi la protection des individus. De plus, ces bases de données ont été enrichies par des données non structurées issues des opérations régulières d'inspection et de diagnostic. Par conséquent, les stratégies de collecte de données, de leurs traitements et d'intervention doivent être continuellement à jour et fiables afin de garantir le meilleur indice de santé, des machines et la prédiction la plus précise des modes de défaillances. On distingue dans ce cas deux approches principales de prédiction, la première c'est la prédiction par les méthodes théoriques classiques (analytique, expérimentale, mathématique), la deuxième est la prédiction par les méthodes d'intelligence artificielle (IA) [7].

La combinaison de ces méthodes permet, par conséquent, de réduire les coûts et d'augmenter à la fois le niveau de sécurité au travail et l'indice de santé des équipements. La qualité et la richesse des données recueillies/calculées sont à la base d'une maintenance optimale et efficace [8]. Ces données proviennent non seulement de l'équipement lui-même, mais aussi des rapports d'inspection (radiographie, thermographie, analyse vibratoire, etc..), des conditions opérationnelles et de l'historique des maintenances. Le déploiement optimal des instruments de contrôle dans les endroits les plus à risque d'une tuyauterie constitue une étape primordiale afin de relever les indicateurs les plus pertinents de température, pression, débit, déformations, etc. En effet, ces données doivent être collectées à des fréquences, positions et orientations bien précises sur une tuyauterie donnée. Les prélèvements sont discrets ou continus et respectent les normes de formatage des bases de données (BD) massives (BigData) dont la gestion prend en considération les cinq piliers de 5V : volume, vélocité, variété, véracité et valeur.

Une fois l'outil de prédiction par les méthodes théoriques classiques ou encore par apprentissage automatique est réalisée, il sera prêt à être implémenté dans des plateformes digitales. En effet, la numérisation des modèles de méthodes de calcul avec l'avènement de l'usine du futur apporte de multiples solutions face à de nombreuses contraintes et problématiques de production. C'est également l'enjeu de toutes les entreprises industrielles désireuses de rester compétitives. Elles ont de plus en plus recours à des solutions numériques, codes et logiciels pour améliorer la gestion de leur maintenance et de leurs données.

1.3 Objectifs

Les objectifs de ce projet sont subdivisés en deux catégories :

1.3.1 Objectif principal

L'objectif principal de ce projet est de développer des approches de prédiction de défaillances dans les réseaux de tuyauterie.

1.3.2 Objectifs spécifiques

Une approche graduelle nous permettra d'atteindre l'objectif général mentionné ci-dessus :

i. Analyser les phénomènes de dégradation des systèmes de tuyauteries et identifier les paramètres contributifs associés
ii. Classifier les critères de défaillances
iii. Examiner des modèles de prévision pour faire la planification de l'entretien des pipelines
iv. Réussir à établir des modèles pour prédire les défauts

1.4 Hypothèse de la recherche

Les systèmes de tuyauteries sont soumis à des sollicitations, en effet l'ensemble des hypothèses suivantes illustre la recherche :

- La performance et la rentabilité d'un système de tuyauterie augmentent lorsque le nombre de défaillances diminue.
- Le choix des paramètres de procédés d'un tel réseau de tuyauterie doit être dans l'intervalle des valeurs limites
- La défaillance dépend de la géométrie de l'élément et du matériau utilisé
- La résistance à la corrosion dépend de la nature du fluide et les paramètres opératoires
- Le choix de la méthode d'apprentissage automatique dépend de la nature des données des bases disponibles
- La performance des algorithmes dépend de la qualité de la base de données (BD)

1.5 Retombées du projet

Ce projet présente une initiative pour développer une approche de prédiction de défaillances destinée aux industries de pipeline afin d'anticiper les pannes qui

peuvent potentiellement endommager un tel équipement. Cette approche vise à minimiser le coût de la gestion des actifs et par la suite le gain d'argent. D'autre part, la prédiction des défauts concerne non seulement la sécurité fonctionnelle des machines et la productivité, mais aussi la protection des individus.

1.6 Originalité de la recherche

Les systèmes de tuyauterie sont considérés comme des moteurs clés dans l'économie nationale du Canada. Bien que les pipelines soient reconnus comme l'un des dispositifs les plus sûrs pour le transport des produits, un nombre considérable de défaillances se sont produites dans ces installations. Cela souligne l'importance de maintenir ces installations dans de bonnes conditions pour éviter les défaillances. L'originalité de la recherche porte sur la prédiction des défaillances des pipelines en utilisant les différentes techniques de l'apprentissage automatique et les méthodes d'analyse avancées. À la fin de cette étude, il faudrait établir des modèles de prédiction en s'appuyant sur des bases de données d'historiques disponibles et aussi en fonction des paramètres tirés à partir des instruments de mesures permanents installés sur les conduites de tuyauterie.

CHAPITRE 2: REVUE DE LITTÉRATURE

2.1 Généralités

Les pipelines constituent la structure principale de plusieurs secteurs industriels (minier, chimique, pharmaceutique…). Ils sont préfabriqués le plus souvent puis assemblés sur site. Les pipelines sont exposés à des agressions extérieures comme les vibrations et autre en plus de la charge nominale de fonctionnement (pression du fluide à transporter, température, …) pour cela ils doivent être protégés et constitués de matériaux résistant. Plusieurs matériaux sont utilisés tels que l'Acier, cuivre, béton armé, PVC et la fonte.

L'acier au Carbonne est l'un des matériaux les plus répandus dans nombreux service dans l'industrie pétrochimique grâce à sa disponibilité dans le marché international, son coût relativement faible ainsi que sa résistance à la corrosion élevée. La spécification des matériaux est définit par les deux standards: American Society for Testing and Materials (ASTM) et l'American Society of Mechanical Engineers (ASME) [9].

Les produits fluidiques circulant dans les conduites sont divisés selon le code ASME en 3 catégories de fluide, la première catégorie est D, elle concerne les fluides les moins critiques qui ne sont pas toxiques inflammables et non dangereux pour les tissus humains, la catégorie M englobe les fluides les plus critiques, pour cette catégorie une fois une personne exposée à une petite dose de fluide transporté peut avoir des séquelles permanentes même s'il y a une intervention chirurgicale pouvant rétablir ses conditions d'origine (phosgène, H_2S, cynaure d'hydrogène) et la catégorie normale lorsque D et M ne s'appliquent pas [10].

Un système de tuyauterie est composé par des tuyaux pour véhiculer le fluide, des brides munies des joints d'étanchéités assemblées par des boulons, des coudes indispensables pour tout changement de direction des tuyaux et des jonctions en Té nécessaire pour toutes dérivations à partir de la conduite principale, des vannes qui servent à régler le débit (vanne régulatrice), des joints de dilatation pour permettre une flexibilité inhérente de la tuyauterie en compensant les sollicitations mécaniques appliquées sur les équipements mécaniques installés (pompe,

échangeur) et d’autres instruments de mesure pour contrôler le système et faire la régulation de certaines grandeurs (débit, pression, température) selon la nécessité.

Les systèmes de tuyauterie se composent d'une série d'éléments tels que des pompes, des nœuds et des tuyaux répartis sur le réseau géographique. Tous ces éléments sont susceptibles de tomber en panne en raison d'une dégradation progressive ou d'éventuels défauts dans les opérations de fabrication ou d'installation. Le coût de l'entretien/remplacement des conduites constitue une charge énorme pour les sociétés et représente généralement la plupart de ses dépenses.

2.2 Règles et normes de conception

Lors de la conception d’un système de tuyauterie, l'ingénieur doit premièrement vérifier la cohérence réglementaire et la compatibilité des codes avec les besoins et les exigences du client afin d'éviter les conflits et éviter toutes contradictions. Les spécifications doivent également respecter les lois nationales qui exigent des normes spécifiques pour les zones sismiques, par exemple. Les normes varient également en fonction de l'utilisation industrielle, et les normes relatives aux canalisations souterraines ne sont pas les mêmes que celles relatives aux canalisations aériennes. La norme utilisée pour raccorder un tuyau à un réservoir n'est pas nécessairement la même pour le raccorder à une pompe. Par conséquent, plusieurs normes peuvent être appliquées dans un seul système de tuyauterie. Lorsqu'elle est appliquée, chaque norme fournit les informations nécessaires telles que les matériaux qui doit être utilisés, les dimensions des tuyaux, les contrôles et les tests à effectuer. [9]

La combinaison des données de calcul et le respect de toutes les réglementations permettent d'augmenter la fiabilité du modèle analysé et d'augmenter la fiabilité du pipeline en assurant une sécurité suffisante contre les déformations excessives.

Au Canada, les normes les plus importantes sont : The American Society of Mechanical Engineers (ASME), American Petroleum Institute (API), American Water Works Association (AWWA), Association Canadienne Du Gaz (ACG), NACE International standard.

2.3 Exigences liées aux installations

L'interface entre le tuyau et l'équipement est extrêmement importante et doit être correctement gérée tout au long du processus de conception. Pour cela certaines exigences doivent être considérées.

- Pour les assemblages boulonnés, il faut :
  - Vérification des axes et les orientations des brides des équipements.
  - Nettoyage de toutes les pièces, vannes et accessoires et l'élimination des impuretés, lors de l'installation de nouveaux équipements ou des périodes d'arrêt programmé.
  - Application des produits antirouilles (huile, graisse) pour protéger les équipements.
  - Correction du désalignement qui dépasse les tolérances acceptables, par coupe et nouveau soudage, par cintrage ou par serrage des boulons (avec respect des seuils des couples).
- Pour l'installation des supports, il faut :
  - Respect des distances admissibles pour la fixation des supports au voisinage des vannes, des coudes et des brides en se référant à la norme ASME B31.3.
  - S'assurer que la conception du support et sa liaison avec la structure permet de suivre le déplacement/glissement/dilatation prévu de la tuyauterie sans la déformer
- Pour les assemblages soudés :
  - La conception des assemblages soudés, la préparation des chanfreins et des renforts des éléments à souder doivent être conformes à la norme ASME B31.3
  - Il est strictement interdit de chauffer la tuyauterie pour l'amener à un certain alignement, si un désalignement se présente, en effet le préchauffage ou le traitement thermique des métaux doivent être exécutés conformément au paragraphe 331 de ASME B31.3.
  - Après chaque passe de soudure, les surfaces doivent être soigneusement nettoyées afin d'éliminer toutes oxydations, écailles ou défauts (les défauts peuvent être corrigés si nécessaire par meulage).
  - Les soudures de pointage doivent être de la même qualité et du même matériau que la soudure définitive et leur fusion doit être complète.

2.4 Types de défaillances dans les systèmes de tuyauterie

La défaillance des tuyauteries est l'effet cumulatif de plusieurs facteurs agissant sur elles. L'association européenne des compagnies pétrolières, (CONCAWE), classe ces défaillances en différents types tels que : mécanique, opérationnel, corrosion, risque naturel et les activités liées aux tiers [11]. Le CONCAWE a été fondé en 1963 par un groupe de grandes compagnies pétrolières pour mener des recherches sur les enjeux environnementaux liés à l'industrie pétrolière. Le CONCAWE publie des rapports, récolte et analyse les accidents des pipelines en Europe. Cette section présente les principaux types de défaillance selon CONCAWE [12].

2.4.1 Défaillance mécanique

Les défaillances mécaniques englobent toutes les défaillances dues à une mauvaise construction ou à l'utilisation inappropriée du matériau, ces actions provoquent l'éraflure (enlèvement de matière) et l'enfoncement (un changement de la courbure de la paroi sans changement d'épaisseur) du tuyau. Ces défauts surviennent généralement pendant la phase de construction. Les dégradations mécaniques peuvent provoquer une défaillance immédiate, une défaillance différée ou aucune défaillance, selon la gravité et la complexité des défauts. La manière la plus courante pour détecter les défaillances est de procéder à des inspections en ligne ILI (In-line inspection) telle qu'une inspection par ultrasons (UT) ou par flux magnétique (Magnetic Flux Pig) [7].

2.4.2 Défaillance liée à la corrosion

Le mot corrosion provient du mot latin ancien "*corrodere*" ou du latin récent "*corrosio*". Ce phénomène se manifeste par la détérioration du matériau qui est causée principalement par le fluide qui y circule ainsi que l'environnement ou le tuyau est installé. Elle conduit à une diminution de l'épaisseur des canalisations dans les zones affectées. Ce processus est généralement très lent. La défaillance par corrosion est considérée comme la source la plus fréquente de défaillances dans les réseaux de tuyauterie. Le comportement de dégradation de ce phénomène doit être clairement identifié. Il existe trois types de corrosion [13] :

- Corrosion externe

La corrosion externe désigne un type de mécanisme de dégradation électrochimique qui se produit à la surface des métaux. Elle est causée principalement par des

facteurs environnementaux tels que le sol, l'eau (corrosion souterraine) et l'air (corrosion atmosphérique). La corrosion externe est prévalente dans les applications de pipelines, mais elle peut se présenter dans n'importe quelle circonstance. Des mécanismes de protection adéquats doivent être implémentés tels que les revêtements (pipeline coating) et la protection cathodique (CP).

- Corrosion interne

Ce type de corrosion attaque principalement la surface interne du tuyau. Il est moins sévère que la corrosion souterraine, mais plus périlleux que la corrosion atmosphérique. Elle est généralement liée au type de produit transporté.

- Corrosion sous contrainte

Ce type de corrosion est provenant de la combinaison des actions mécaniques (contraintes et déformation) et chimiques (corrosivité de l'environnement).

2.4.3 Défaillance liée aux activités de tiers

Les défaillances liées au tiers englobent les dommages causés par des facteurs non associés au réseau de tuyauterie. Ces opérations intentionnelles ou accidentelles sont la source de défaillance la plus courante des oléoducs, bien qu'elles soient le sujet le moins examiné dans l'évaluation des risques liés aux pipelines.

2.4.4 Défaillance opérationnelle

Les défaillances opérationnelles proviennent d'erreurs d'opérateurs, de perturbations opérationnelles (température, pression, débit), de dysfonctionnement ou l'inadéquation d'un ou plusieurs systèmes de contrôle. Ce type de défaillance est peu fréquent, bien qu'il entraîne des répercussions néfastes. En plus de la surveillance des paramètres opératoires, le déploiement de dispositifs de sécurité, de contrôle, de supervision et d'acquisition de données peut aider à prévenir ces défaillances [14].

2.4.5 Défaillance due aux risques naturels

Les risques naturels entraînent rarement la défaillance d'un pipeline, mais ils doivent être pris en compte dans l'évaluation de la défaillance en raison de leurs implications sur la sécurité publique. Les risques naturels comprennent les inondations, les mouvements de terrain, l'activité volcanique et les tremblements de terre, qui peuvent tous endommager gravement un pipeline et l'environnement.

Dans la plupart des cas, des études géotechniques et hydrotechniques sont réalisées avant la construction du pipeline.

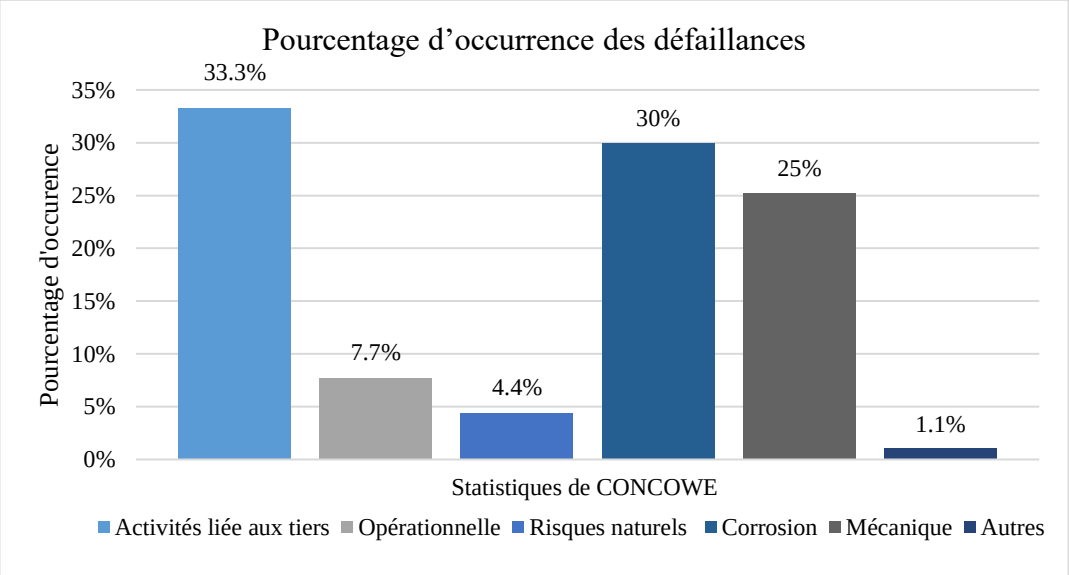


Figure 2.1 Pourcentage d'occurrence des défaillances

La Figure 2.1 ci-dessus présente le pourcentage d'occurrence de chaque type de défaillance au cours des 20 dernières années selon les rapports de l'association européenne des compagnies pétrolières (CONCAWE) [15]. Le graphique montre que 88% des défaillances ont été provoquées soit par une défaillance mécanique, soit par une défaillance due à la corrosion, soit par une défaillance liée au tiers. Chacune de ces catégories de défaillance peut être affectée par un certain nombre de facteurs. Le paragraphe suivant traite les différents facteurs qui contribuent aux types de défaillances susmentionnés.

2.5 Facteurs contribuant aux défaillances dans les systèmes de tuyauterie

La désignation des facteurs qui contribuent aux défaillances des pipelines est une tâche complexe, car à leur tour ils dépendent de plusieurs sous-facteurs, et la majorité de ces facteurs doivent être pris en compte afin de déterminer le poids de leur contribution individuelle à la détérioration du pipeline. Ces facteurs sont les piliers de tout modèle de prédiction ou d'inspection basé sur le risque (*Risk Based inspection model*).

2.5.1 Facteurs contribuant à la défaillance opérationnelle

La défaillance opérationnelle peut être causée par des erreurs humaines et des erreurs liées au système de tuyauterie. Les facteurs suivants doivent être analysés pour évaluer le risque de défaillance [13] :

- Paramètre opérationnel liée au procédé : Ces dernières peuvent être classées en deux groupes soient les sollicitations statiques et les sollicitations dynamiques. Tout particulièrement, la répartition statique hétérogène de la température tout au long d'une tuyauterie conduira à des cycles de dilatation/contraction thermiques localisés des tuyaux. D'autre part les sollicitations dynamiques sont générées par les instabilités de l'écoulement fluidique, les variations cycliques relativement rapides de la température et de la pression, les vibrations et les chocs thermiques/pression.
- Problème au niveau des systèmes de surveillance et d'acquisition de données (SCADA).
- Défaillance au niveau des systèmes de sécurité.
- Défauts mécaniques au niveau des équipements mécaniques (pompes, compresseurs, vanne de sécurité ...)
- Niveau de formation des opérateurs

2.5.2 Facteurs contribuant à la défaillance mécanique

Les défaillances mécaniques sont liées aux erreurs de conception, mauvais choix de matériaux et/ou d'une construction défectueuse (soudure, assemblage) [12].

- Facteurs liés à la défaillance des matériaux et au défaut de construction

La Norme ASME édicte certaines règles et recommandations visant à réduire le risque des défaillances mécaniques. Ces facteurs sont récapitulés comme suit :

- Choix de matériaux : les lignes qui sont constituées de différents matériaux se détériorent et se brisent de différentes manières.
- La procédure de fabrication et de construction appropriée.
- La procédure d'installation appropriée des équipements.
- L'inspection des fuites.

2.5.2.1 Facteurs liés à une défaillance causée par une erreur de conception

L'une des principales causes de défaillance mécanique est l'erreur de conception. Parmi les sujets importants susceptibles de conduire à une erreur de conception [13], on peut citer :

- Le mauvais choix de coefficient de sécurité lors de la conception.
- Rupture par fatigue ; qui dépend principalement des variations cycliques de chargement. La fréquence des cycles de pression interne est considérée comme le facteur le plus important qui affecte la fatigue.
- Coup de Bélier : se produit lorsqu'il y a un changement brusque du débit de fluide circulant, qui peut être dû à la fermeture d'une vanne, déclenchement d'une pompe. L'intensité de cette surtension dépend de la densité, la vitesse d'un fluide, ainsi que de la vitesse d'arrêt. Les dispositifs de protection doivent être installés pour protéger le système de tuyauterie.

2.5.3 Facteurs contribuant à la défaillance due à la corrosion

➢ Corrosion externe [16]

Les mécanismes de corrosion externe sont liés à l'emplacement géographique où le pipeline est installé. Parmi les causes principales de ce phénomène de dégradation, on trouve :

- La composition chimique qui peut être d'origine naturelle, comme le sel et le CO_2, ou d'origine humaine, comme le chlore et le SO_2.
- Les températures élevées et notamment l'humidité élevée augmentent le risque et la vitesse de la corrosion.
- Le choix et le manque du revêtement extérieur.
- L'âge de la conduite.

➢ Corrosion interne [16]

- La corrosion interne est liée directement au niveau de la corrosivité du produit transporté, cette corrosivité dépend de plusieurs facteurs tels que :
- Les réactions chimiques qui se produisent au niveau des surfaces internes des pipelines en raison de la présence d'eau, d'oxygène, d'électrolytes et de certaines molécules, notamment le dioxyde de carbone (CO_2), le sulfure d'hydrogène (H_2S) ainsi que l'existence des impuretés.

- Les paramètres opératoires tels que la température et le débit de fluide (turbulence) ainsi que le PH, la densité et la concentration du fluide.

➢ Intervention pour prévenir la corrosion interne [16]

- Le contrôle peut être effectué à l'aide des sondes électroniques qui mesurent le taux de corrosion ou avec des coupons qui sont susceptibles à se corroder en présence d'une substance corrosive qui permet à la suite de donner une indication sur le taux de corrosion.
- Injection d'inhibiteur : certains produits chimiques pourraient être injectés dans un pipeline pour réduire la réaction qui favorise la corrosion.
- Appliquer un revêtement interne
- La séparation des impuretés des produits
- Instrument de raclage (*Pigging*) : Il s'agit d'un instrument cylindrique qui est utilisé pour nettoyer les parois intérieures des pipelines, en éliminant les résidus.

Les experts en matériaux et en corrosion ont proposé une série d'approches basée sur la surveillance de l'état du réseau de tuyauterie qui a pour but de limiter et gérer les risques de mécanisme de dégradation cités ci-dessus. La surveillance de l'état consiste à mesurer et contrôler les paramètres de l'équipement qui indiquent une défaillance. La surveillance de l'état, bien sûr, chevauche la maintenance prédictive. Le suivi du comportement des actifs est une partie importante de la maintenance prédictive. Les données collectées fournissent la base pour découvrir les tendances et perfectionner les algorithmes. Cependant, avec l'implémentation des techniques de l'intelligence artificielle, il est évident de parler de surveillance en temps réel, inspection en ligne (ILI), de l'état même sans programme de maintenance prédictive.

2.6 Surveillance de l'état des systèmes industriels (*Condition Monitoring*)

Les applications de la surveillance de l'état (CM) dans les industries et le développement de nouvelles techniques de CM sont devenues l'une des tâches les plus importantes [17]. Ce besoin peut être vu de deux côtés. Tout d'abord, la santé et la sécurité du fonctionnement des équipements dans les systèmes industriels sont si importantes qu'une défaillance et un arrêt inattendus peuvent provoquer un accident grave et entraîner une pénalité élevée en termes de coûts de production perdus. D'autre part, les équipements mécaniques et les pipelines eux-mêmes sont

les actifs les plus coûteux et leur entretien coûte cher [16]. Il ne fait aucun doute que les industriels doivent trouver des moyens d'éviter les pannes soudaines, de minimiser les temps d'arrêt, de réduire les coûts de maintenance et d'allonger la durée de vie des actifs. La CM est juste la réponse à ces problèmes avec la capacité de fournir des informations utiles pour utiliser les machines d'une manière optimale. Deuxièmement, le développement des technologies informatiques, des techniques de traitement des signaux ainsi que des techniques d'intelligence artificielle (IA) a permis de mettre en œuvre plus efficacement la CM sur les différents équipements constituant le réseau de tuyauterie. On s'attend à ce que les techniques de CM soient plus fiables, plus intelligents et moins chers, de sorte qu'ils puissent être largement utilisés. L'objectif de l'inspection est d'évaluer l'intégrité des actifs d'une installation, de quantifier le risque et de prévoir les taux de dégradation (fatigue, corrosion…) afin de mieux gérer la fiabilité. Dans un monde idéal, une installation devrait effectuer des inspections complètes de tous ses actifs à chaque occasion. La norme API 570 [18] propose des intervalles d'inspection allant de 6 mois jusqu'à 10 ans, tout dépend du type du mécanisme de dégradation. Cependant, ces intervalles vont être déterminés après avoir procédé à une analyse basée sur les risques (RBI). Cette analyse est généralement menée par des consultants tels que la compagnie Det Norske Veritas (DNV) [19], mais certaines firmes disposent de leurs propres départements de recherche qui réalisent ces études en identifiant le risque en effectuant une ILI puis en étudiant les conséquences de la défaillance.

Les zones spécifiques du circuit de tuyauterie où les inspections doivent être effectuées (CML) doivent être déterminé pour chaque système étudié [20]. Ces CML comprennent les endroits destinés à la mesure de l'épaisseur, les examens de fissuration sous contrainte, les emplacements pour les CUI (corrosion sous isolation), les examens d'attaque par l'hydrogène à haute température. Les circuits de tuyauterie présentant des conséquences potentielles assez fortes de défaillance auront le nombre de CML le plus élevé et seront surveillés plus fréquemment [21]. Ces zones doivent posséder l'une des critères suivants :

- Risque qui peut provoquer une urgence en matière de sécurité ou d'environnement.
- Taux de corrosion élevés (prévus ou expérimentés).

- Risque élevé de corrosion localisée.
- Risque élevé de CUI (corrosion sous l'isolation, y compris la corrosion sous contrainte fissuration sous l'isolation)

La distribution des CML doit être de manière appropriée dans chaque circuit de tuyauterie afin d'assurer une couverture de surveillance adéquate des principaux composants. Le nombre de CML doit tenir compte des résultats des inspections précédentes, les schémas de corrosion et de dommage attendus. L'ajout ou l'élimination d'un certain CML doit être approuvé par les experts et les consultants [21]. Généralement un nombre limité de CML peut être adopté dans les cas suivant :

- Faible potentiel de risque de créer une urgence sécuritaire ou environnementale en cas de fuite.
- Des circuits de tuyauterie relativement non corrosifs.
- Des systèmes de tuyauterie longs et droits

D'autre part, chaque CML devrait avoir au minimum un ou plusieurs points d'examen identifiés. Les exemples incluent :

- Un emplacement repéré sur les tuyaux non isolés
- Des trous creusés au niveau de l'isolant et bouchés par des couvercles.
- Couvercle d'isolant temporaire pour les raccords de tuyauterie.
- Dessins isométriques qui définissent la position des CML ainsi que les lieu de surveillance de l'épaisseur (TML :*Thickness Monitoring Location*)

Un certain nombre de processus de corrosion communs aux unités de raffinage et de pétrochimie sont relativement uniformes en nature, ce qui entraîne un taux assez constant de l'amincissement de la paroi du tuyau, indépendamment de l'emplacement dans le circuit de tuyauterie (axialement ou circonférentielle). Parmi les exemples de tels phénomènes de corrosion, on peut citer la corrosion par le soufre à haute température et la corrosion par l'eau acide (à condition que les vitesses ne soient pas élevées au point de provoquer une corrosion/érosion locale des coudes, des tés et d'autres éléments similaires). Dans ces situations, le nombre de CML requis pour surveiller un circuit sera inférieur à celui requis pour surveiller des circuits qui subissent des pertes métalliques plus localisées. En théorie, un

circuit soumis à une corrosion uniforme pourrait être surveillé de manière adéquate avec un seul CML.

En réalité, la corrosion n'est jamais vraiment uniforme et peut même être très localisée, de sorte que des CML supplémentaires peuvent être nécessaires. Les pratiques de surveillance de la corrosion plaçaient un emplacement de contrôle de l'épaisseur TML sur chaque raccord (par exemple, coude, té, réducteur). Ces TML faisaient référence à un point d'examen unique où des mesures d'épaisseur seront prises afin d'établir divers calculs, y compris les taux de corrosion. Généralement, ils sont placés aux points médians des sections droites ainsi qu'au niveau des quatre quadrants (0°, 90°, 180°, 270°), avec une attention particulière au niveau des zones intrados et extrados des coudes où la corrosion/érosion pourrait augmenter le taux de dégradation. Les mesures d'épaisseur sont réalisées à l'aide des techniques d'examen non destructif (NDT) [20]. API 510 [22] et API 570 [18] recommandent de noter soit la valeur minimale, soit la moyenne de plusieurs lectures dans la zone du point d'examen. Ces données sont utiles pour établir la RBI. L'ensemble des CML peuvent être regroupés à la suite en des boucles de corrosion (corrosion loop (CL)). Elles sont définies comme le groupement des circuits de tuyauterie qui sont soumis à des mécanismes de corrosion, conditions opératoires et conception similaires, ce qui permet par la suite la réduction et l'élimination des activités d'évaluation qui n'ont aucune valeur ajoutée. Par le biais des CL, les inspecteurs peuvent comprendre plus clairement les mécanismes de corrosion, cela améliorera l'efficacité de l'inspection ainsi que la sécurité de l'usine. D'autre part, le développement de ces boucles est une partie intrinsèque de la méthodologie d'inspection basée sur le risque (RBI) [23]. Cependant, ce processus est une tâche à forte intensité de connaissances qui présente une grande variabilité de résultats. Cette variation peut impliquer le jugement et l'intuition des ingénieurs, car elle peut entraîner soit une sous-inspection des équipements à haut risque soit une surinspection des équipements à faible risque et la fin elle peut menacer l'intégrité globale du système.

2.7 Analyse basée sur le risque (RBI)

Au cours des 20 dernières années, les industriels se sont fiés à l'analyse des risques pour hiérarchiser les actions d'inspections et de la maintenance [23]. Avant cela, les

intervalles d'inspection sont définis en se basant sur les classifications des services de tuyauterie cités dans les réglementations telles que l'API, l'ASME, et en fonction de l'expérience et des connaissances des inspecteurs. Cela peut entraîner des risques inacceptables, ainsi que des pertes de ressources coûteuses. Les organismes de réglementation ont défini les lignes directrices pour la mise en œuvre des inspections basées sur le risque. La RBI est une méthodologie qui permet d'élaborer un plan d'inspection rentable et de garantir la conformité à la réglementation et aux règles de l'entreprise. C'est une analyse systématique qui permet d'établir et de classer les niveaux de risque associés à l'exploitation de chaque composant industriel (tuyauterie, échangeur, les appareils sous pression, etc.). Cette méthode sert à déterminer les calendriers d'inspections des équipements sur la base des niveaux de risque. Cet indicateur est défini comme la probabilité de défaillance (PoF) multipliée par la conséquence de la défaillance (CoF).

$$Niveau\ de\ risque = PoF * CoF \quad (1)$$

2.7.1 Probabilité de défaillance

La PoF est estimée sur la base des facteurs qui influencent le taux de défaillance de l'équipement, telle que le type et le taux de dégradation, les paramètres opérationnels, la conception de l'équipement, l'efficacité et les résultats des inspections précédentes ainsi que l'âge de l'équipement. La valeur de la PoF, équation (2), est obtenue par la multiplication entre la valeur de la fréquence des défaillances générales g_{ff}, les facteurs de dommages D_f et les facteurs du système de gestion F_{MS}. La détermination du g_{ff} et F_{MS} ainsi que les étapes nécessaires pour calculer le D_f sont décrits dans l'API 581 pour chaque mode de défaillances.

$$Pf(t) = g_{ff} * Df(t) * F_{MS} \quad (2)$$

2.7.2 Conséquence de la défaillance

Le CoF est l'ensemble des événements qui se produisent à la suite de l'endommagement d'un tel équipement. Le RBI prend en compte quatre catégories de conséquences : l'impact sur la santé et la sécurité du personnel, l'impact environnemental, les pertes de production et les coûts de réparation des installations.

2.7.3 Niveau de risque

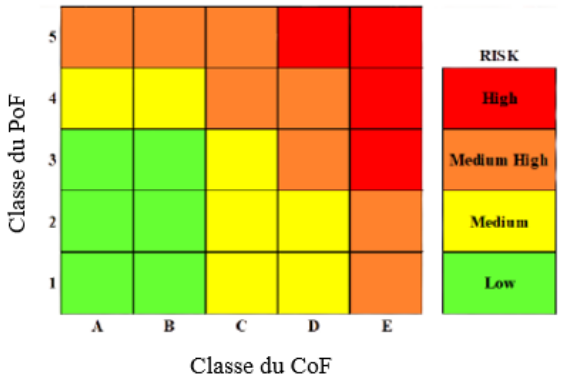


Figure 2.2 Matrice de risque

Les catégories de PoF et CoF sont présentées dans la matrice de risque Figure 2.2 afin de déterminer le niveau de risque (élevé, moyen et faible) des équipements inspectés. Le plan d'inspection sera à la suite établie en fonction du niveau de risque de chaque équipement. Les étapes suivantes sont recommandées pour l'application du RBI pour les réseaux de tuyauterie :

- Collecter des données d'inspection à partir des méthodes NDT
- Identifier les zones qui ont des taux de corrosion élevés.
- Identifier les facteurs qui provoquent les défauts.
- Détermination du Pof et Cof
- Recommander le programme de maintenance le plus adapté et la date de la prochaine inspection.

2.7.4 Prédiction de la durée de vie restante

Par le biais des inspections, il est aussi possible de prédire la durée de vie restante pour qu'un équipement fonctionne sans défaillance. La prédiction de RL est une approche efficace qui permet d'établir la durée maximale pendant laquelle les équipements peuvent fonctionner au-delà de leur durée de vie initiale. L'équation (3) présente le calcul de RL pour les réseaux de tuyauterie mentionnés par l'API.

$$RL = \frac{t_{actuel} - t_{minimum}}{taux\ de\ corrosion} \quad (3)$$

t_{actuel} : Valeur de l'épaisseur mesurée

$t_{minimum}$: Valeur de l'épaisseur minimal

$taux\ de\ corrosion$: Valeur de la perte de matière par corrosion

2.8 Intégrité des réseaux de tuyauterie et méthodes d’inspections

L'intégrité des réseaux de tuyauterie fait référence à l'état des actifs tel que les pipelines. La gestion de l'intégrité garantit qu'un pipeline et ses composants sont en bon état de fonctionnement. Les essais non destructifs (NDT) [24] sont utilisés pour garantir l'intégrité des actifs, ce sont des méthodes fiables de détection des défauts et de la corrosion. Cela permet non seulement de préserver la fiabilité, l'efficacité et la durabilité des circuits, mais aussi la santé et la sécurité de l'environnement. En utilisant ces techniques, des mesures proactives peuvent être adoptées pour atténuer tout risque et assurer la sécurité du pipeline. D’autre part, les inspections fournissent aux opérateurs un aperçu sur les différents paramètres susceptibles d’engendrer les défaillances. La liste suivante présente les paramètres les plus importants qui doivent être inspectés : l’état du revêtement, l’état de la protection cathodique, l’existence des éraflures et des enfoncements, les dimensions de la géométrie des éléments, l’épaisseur des composants et la perte du métal causés par la corrosion interne et externe.

Au cours de la dernière décennie, plusieurs méthodes d’inspection basées sur les NDT ont été développées pour évaluer en ligne (ILI) l'état des pipelines, détecter, quantifier et localiser avec précision les défauts dans les réseaux de tuyauterie. Les méthodes NDT sont des techniques communes pour la détection des défaillances et l’évaluation de la sécurité. Elle fait référence aux tests de contrôle effectués sans affecter et endommager le fonctionnement de l’objet testé ces techniques peuvent être regroupées en trois catégories comme montre la Figure 2.3:

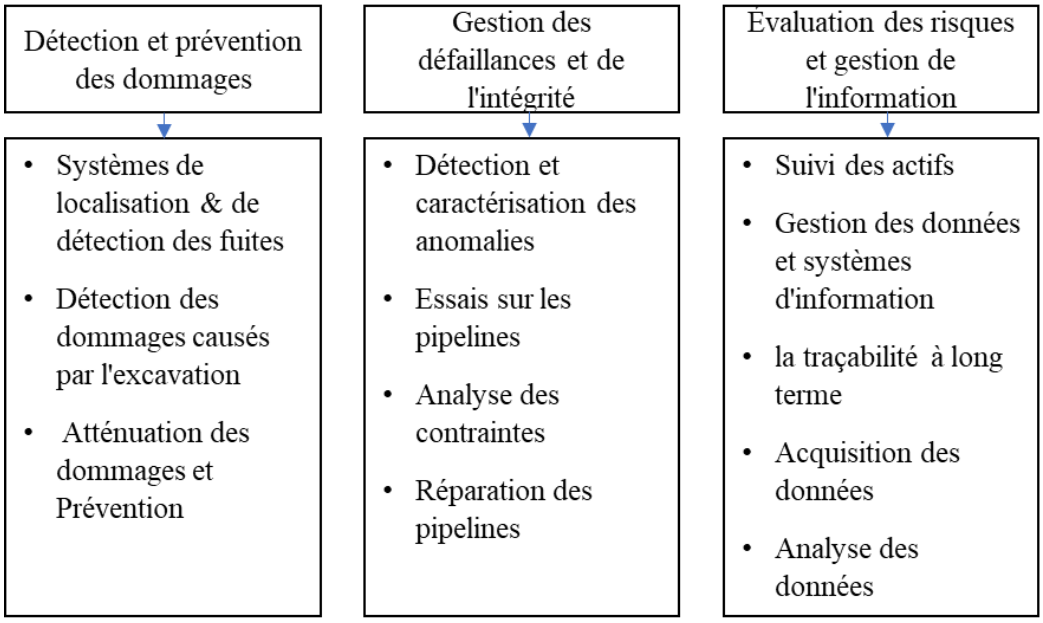


Figure 2.3 Catégories des méthodes d'inspections

Les technologies de la détection et la prévention des défauts permettent la localisation des pipelines et des installations souterraines, la détermination des dommages dus aux excavations et des empiètements sur le terrain, la détection des fuites et l'atténuation des dommages. Par conséquent, il est nécessaire de développer un équipement de localisation, notamment des localisateurs électromagnétiques, pour identifier la position du défaut. Par exemple, les méthodes acoustiques ou de radar à pénétration de sol sont des avancées récentes. D'autre part, les technologies d'atténuation et de prévention des accidents visent à réduire au minimum le temps et la charge de travail nécessaires pour prendre des précautions afin d'éviter l'expansion des accidents. Cependant, ces technologies semblent difficiles à être implantées et protégées à l'intérieur du pipeline. Les technologies de gestion des menaces et de l'intégrité comprennent l'inspection des défauts existants (tels que les fissures, la perte de métal, la rouille et les bosses), l'utilisation de plusieurs dispositifs pour l'évaluation directe de la corrosion externe, l'inspection interne, la détection et la surveillance des pipelines, ainsi que l'analyse des contraintes. Quant à l'évaluation des risques et à la gestion de l'information, elle concerne diverses technologies pour la visualisation des données, le suivi et la traçabilité des actifs, le système d'information géographique, l'évaluation des risques, la sensibilisation à la réponse, ainsi que la sécurité de l'environnement.

Parmi les méthodes NDT on trouve notamment la radiographie, l'inspection visuelle (VT), les ultrasons (UT), le contrôle par courants de Foucault (ET) et le contrôle par particules magnétiques (MFL), sont les outils les plus réputés. Les méthodes NDT sont basées sur différentes approches qui ont leurs caractéristiques et leurs utilisations pour l'inspection en ligne des pipelines. De plus, les différentes origines de défauts engendrent différents types de dommages. Par conséquent, il faut bien étudier le choix de la méthode appropriée en fonction des exigences spécifiques pour la détection des défauts. Les différentes techniques couramment utilisées sont :

- Contrôle par particules magnétiques (MFL)

Les outils MFL, Figure 2.4, utilisent des aimants puissants pour créer un champ magnétique saturant dans le matériau du tuyau. La perturbation ou la "fuite" du flux magnétique est détectée par des capteurs situés sur la circonférence de l'outil. La technologie utilise un outil de raclage multi-segments avec un magnétiseur et des capteurs dans un segment, et des composants de stockage de données, une alimentation électrique et d'autres capteurs dans des segments supplémentaires. Cependant, elle nécessite une gestion importante des données et génère une magnétisation à long terme des conduites. La MFL est capable de détecter les dommages mécaniques et aussi elle peut dimensionner efficacement la perte de métal à environ 5% de l'épaisseur de la paroi. Elle est efficace pour détecter les fissures circonférentielles, mais ne convient pas pour les fissures axiales. Cela dépend de l'alignement du champ magnétique. Les pipelines doivent être nettoyés avant d'utiliser la MFL, car les capteurs doivent être très proches de la paroi et ils pourraient être endommagés si le pipeline n'est pas nettoyé [25].

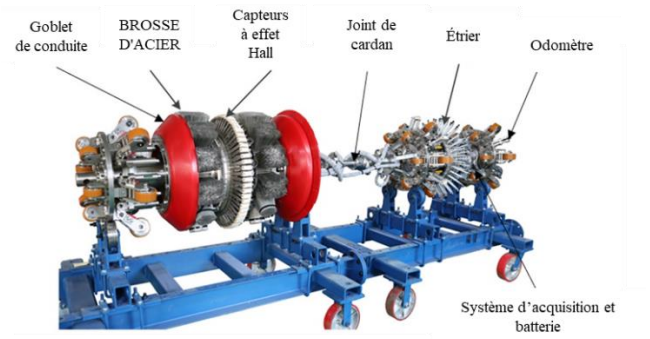


Figure 2.4 Composition du MFL [26]

- Courant de Foucault (ET)

Le courant électrique est induit dans des bobines situées près de la surface. Le courant crée des champs magnétiques dans la paroi qui s'opposent aux champs originaux. L'impédance des bobines est affectée par le courant induit qui est déformé par la présence de défauts. Cette technologie est sans contact, sans effet résiduel, mais sensible aux variations de distance. Elle a une réponse relativement lente qui limite son application. Les ET sont capables d'analyser l'épaisseur de la paroi et détecter les piqûres, les fissures ainsi que les défauts laminaires.

- Ultrason (UT)

L'UT consiste à appliquer des ondes acoustiques à haute fréquence qui permettent de détecter les défauts et de mesurer l'épaisseur des parois [27]. L'installation de l'UT nécessite l'utilisation d'un couplant entre le transducteur et la surface de la conduite, ce qui limite ses applications dans les conduites de gaz. Cette méthode permet de quantifier l'amincissement de la paroi, la dégradation, la perte de métal et la détection des fissures.

- Radiographie (RT)

Cette méthode se base sur les rayons X qui permettent d'analyser intuitivement les défauts. La RT ne nécessite pas un traitement de surface ni le retrait de l'isolant. La RT est la méthode plus répandue pour identifier des défauts dans les soudures. Cependant, elle comporte toujours des inconvénients majeurs liés au risque sanitaire associé aux radiations.

- Transducteur acoustique électromagnétique (EMAT)

Des bobines induisent un courant alternatif à travers la paroi du tuyau, Figure 2.5 [28]. L'interaction avec les champs magnétiques appliqués produit des champs chargés (forces de Lorentz) et génère des ondes acoustiques dans le matériau. Le type et la configuration du transducteur utilisé caractérisent la propagation des ondes à travers la paroi du tuyau. Cette technologie ne nécessite pas l'installation d'un couplant et peut produire des ondes de cisaillement pour l'inspection des zones telles que les soudures. C'est une méthode de contrôle relativement rapide dont elle permet de détecter les pertes de métal, inspecter les soudures, inspecter le laminage

des plaques et elle peut être utilisée pour recueillir des informations sur le décollement du revêtement.

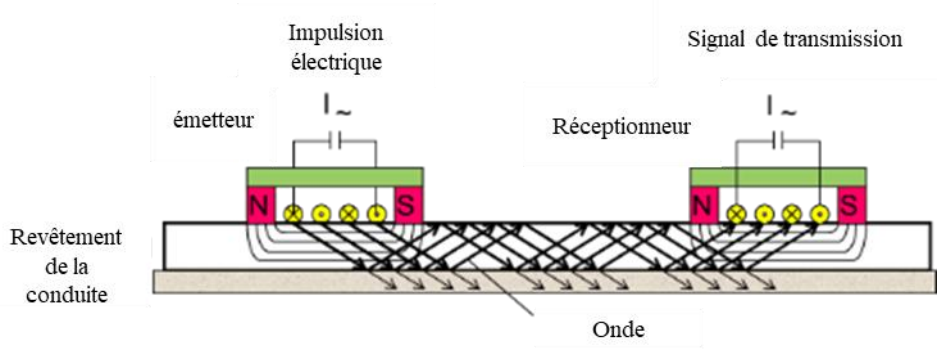


Figure 2.5 Transducteur acoustique électromagnétique [29]

Jusqu'à présent, seules les méthodes d'essai basées sur les ultrasons et les ultrasons électromagnétiques, les fuites de flux magnétique et les courants de Foucault ont permis de concevoir des outils de contrôle interne pour l'inspection en ligne des pipelines (ILI). Cependant, ces méthodes peuvent peut-être affectées par la sensibilité des capteurs à divers facteurs : la température, les facteurs environnementaux, la performance du transducteur et le couplage ultrasonique [30]. Ces erreurs peuvent entraîner une déviation inattendue des données enregistrées, ce qui se traduit par des valeurs aberrantes dans l'ensemble des données, ce qui nécessite leur détection pour bien analyser les défauts à la suite.

L'association de l'ensemble des résultats obtenus avec les données historiques des instruments installés et en exploitant les techniques de l'intelligence artificielle permet de prédire, quantifier et identifier le type de défaut ainsi que son niveau de risque.

2.9 Apprentissage automatique

L'une des orientations majeures de la quatrième révolution industrielle est de rendre autonomes et intelligents les systèmes de contrôle des processus industriels, par l'intégration des systèmes d'acquisition de données à travers des capteurs et des instruments installés au cœur du procédé. Ces données massives issues des instruments sont souvent stockées sur des serveurs locaux ou bien des bases de données afin d'être traitées à partir des modèles d'apprentissage : Apprentissage supervisé et non supervisé pour prédire l'état du système.

Les modèles prédictifs basés sur l'AA sont des algorithmes permettant d'apprendre automatiquement à partir des données, des rapports et des expériences réelles traitées et validées. Ces algorithmes souvent appelés boite noire, car leurs représentations des connaissances ne sont pas intuitives. En effet, il existe plusieurs méthodes d'AA divisées principalement en deux principales catégories comme indique la Figure 2.6.

La première concerne la méthode d'apprentissage supervisé basée essentiellement sur la classification et régression [31]. Fondamentalement, la classification consiste à prédire une variable discrète (qualitative) dont a chaque entrée une étiquette doit être attribuée. C'est le cas de prédiction du niveau de risque d'un tel actif (faible, moyen, élevé). Alors que la régression consiste à prédire une variable continue (quantitative). Par exemple : prédire le taux de corrosion, coût de maintenance de l'actif. Le succès de ces algorithmes repose sur une étape fondamentale qui consiste à préparer et nettoyer la BD en analysant les valeurs manquantes, les données aberrantes et la normalisation des données recueillies (voir chapitre suivant). Les méthodes d'apprentissage supervisées permettent de déterminer les corrélations qui existent entre les données d'entrée (attribue) et les indicateurs cibles de la défaillance à prédire.

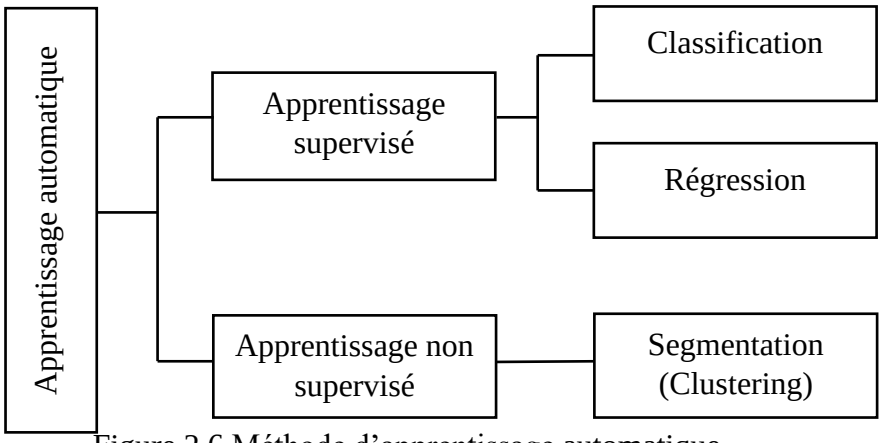


Figure 2.6 Méthode d'apprentissage automatique

La seconde catégorie concerne l'apprentissage non supervisé [32]. Ce type d'apprentissage n'intègre pas le concept d'entrée-sortie (Il n'y a que des données d'entrée). Cette méthode permet de partitionner les données en groupes d'éléments homogènes (*Clustering)* tout en se basant sur leurs degrés d'homogénéité. Pour cet algorithme l'apprentissage se base sur les variations et les fluctuations dans les

données. Chaque groupe (*cluster*) devrait disposer de données semblables et les données différentes devraient être regroupées en groupes distincts. La distance entre les éléments est souvent utilisée comme mesure de similarité. L'applicabilité de cette méthode dans le contexte de prédiction des défaillances devrait être vérifiée. En effet, l'algorithme doit prédire les défaillances par lui-même et les résultats en fonction des données sans utiliser des indicateurs de performances explicites.

De nombreuses recherches ont été menées pour examiner l'application des techniques d'apprentissage automatique dans le domaine de la prédiction des défaillances. El-Abbasy, et al. [33] ont développé des modèles basés sur l'analyse de régression pour permettre l'évaluation de l'état des oléoducs et soutenir les compagnies à planifier l'inspection et les dates de maintenance. Les données d'entrée de ces modèles sont les caractéristiques générale et opérationnelle des pipelines, tel que l'âge, le diamètre, l'état de revêtement, la pression de fonctionnement et la protection cathodique sont inclus. La sortie du modèle est l'évaluation de l'état du pipeline sur la base d'une échelle de 0 à 10, 0 étant la condition critique alors que 10 signifie que l'état du pipeline est parfait. Kimiya et al. [34] ont élaboré un modèle de réseau de neurone artificiel (ANN) pour prédire les défaillances dans les oléoducs.

Liao et al. [35] ont prédit le taux de corrosion dans les gazoducs en utilisant un algorithme de réseau neuronal. Sumayah et al. [36] ont développé 5 algorithmes d'AA à savoir la forêt aléatoire, la machine à vecteurs de support (SVM), le k-voisin le plus proche (K-NN), le boosting de gradient et l'arbre de décision, pour prédire les fuites. Bersani et al. [14] ont élaboré un modèle d'évaluation des risques en utilisant les données historiques du ministère des Transports des États-Unis (DOT) pour prévoir les défaillances causées par les activités de tiers. De Kerf et al. [37] ont proposé un modèle de détection des fuites de pétrole dans un port à l'aide de caméras thermiques IR et de véhicules aériens sans pilote (UAV). Le modèle proposé fait appel au traitement d'image. Pour cela, les images IR étaient nécessaires pour détecter les fuites d'hydrocarbures pendant la nuit. Les chercheurs ont présenté une méthode pour attribuer des étiquettes aux images rouges, vertes et bleues (RVB) et les faire correspondre aux images IR pour construire la BD. Les images collectées ont été redimensionnées et utilisées pour entraîner un réseau neuronal convolutif (CNN). Une fois le réseau formé, il a permis l'inspection

fréquente des fuites d'huile sur l'eau à un faible coût. Lors de la phase de test, les chercheurs ont réussi à détecter les fuites d'huile sur l'eau avec une précision de 89 %. La solution mise en œuvre permet de réduire le coût du nettoyage des fuites d'huile dans l'eau, de minimiser l'interaction humaine pendant le processus et d'augmenter le taux de détection. Andika et al. [38] ont développé un algorithme basé sur la technique de clustering afin de regrouper les systèmes de tuyauterie en des boucles de corrosion. Les données du modèle sont les paramètres opératoires, le type du revêtement et l'isolant ainsi que les caractéristiques des services (type de fluide et phase). Song Fu et al. [39] ont utilisé la forêt d'isolation, technique de AA non supervisée, pour détecter les anomalies dans les systèmes de turbine à gaz. Mariam et al. [40] et Di Hu [41] ont employé la technique Long Short-Term Memory (LSTM), respectivement, pour détecter les anomalies dans les systèmes d'alimentation photovoltaïques (PV) et dans les systèmes de surveillance des centrales électriques.

Le domaine de la santé, outre les secteurs industriels, a utilisé de manière significative l'apprentissage automatique pour aider à diagnostiquer diverses maladies. Chen et al. [42] ont développé des modèles CNN pour la prédiction du risque d'infarctus cérébral, la précision de prédiction des modèles proposés atteignant 94 %. Asri et al. [43] ont utilisé les méthodes SVM, un arbre de décision, Naïve Bayes (NB) et K-NN pour prédire le risque de cancer du sein. Leurs résultats ont permis d'obtenir une précision de prédiction de 97,13 %.

2.10 Synthèse

Sur la base de cette revue de la littérature, on peut conclure qu'il est indispensable de mener des recherches sur la prédiction des défaillances dans les systèmes industriels. Ce chapitre s'est concentré sur les techniques proposées pour la détection et la prédiction des défaillances, tout en présentant les différents types de défauts ainsi que les paramètres qui peuvent contribuer à ces différents modes de dégradation dans les systèmes de tuyauterie. Cette recherche propose des modèles de prédiction de défaillances qui repose sur les attributs du pipeline pour prédire objectivement la défaillance qui menace un pipeline, sur la base des données historiques. Les techniques d'AA sont exploitées pour développer ces modèles en raison de leurs capacités à analyser des variables mixtes.

CHAPITRE 3: MÉTHODOLOGIE

Comme mentionné dans le CHAPITRE 1:, l'un des principaux objectifs de cette étude est de développer un modèle d'analyse prédictive qui correspond à la prédiction et l'anticipation des défauts dans les installations de tuyauterie. La méthodologie retenue pour la réalisation des objectifs fixés de ce projet est présentée au niveau de la Figure 3.1. La première étape consiste à faire une revue de littérature sur les différents modes de défaillance dans les réseaux de tuyauterie et les méthodes d'apprentissage automatique qui permettent de développer des modèles de prédiction pour identifier les signatures et les symptômes des modes de dégradation dans les systèmes de tuyauteries. La deuxième étape consiste à recueillir les données à travers des rapports et des fiches d'inspection fournies par le partenaire industriel ou bien les données par des capteurs placés sur un élément du système de tuyauterie. La collecte des données va être suivie par un processus de prétraitement, qui est une phase vitale pour le développement d'un modèle prédictif. Cette phase permet d'extraire, d'intégrer les données provenant de différentes sources, de les transformer en un format lisible et compréhensible qui facilite l'analyse, le nettoyage et l'élimination des données aberrantes. Après cette phase, nous avons commencé le développent du modèle AA, pour répondre aux objectifs fixés pour notre projet, nous avons utilisé les modèles d'apprentissage automatique supervisés et non supervisés. Pour l'apprentissage supervisé, les algorithmes de régression et de classification ont été développés. Les régressions ont été utilisées pour prédire le taux de corrosion dans les systèmes de tuyauterie alors que la classification est employée pour anticiper la source de défaillance qui peut impacter le réseau de tuyauterie ainsi que le niveau de sévérité de la corrosion. Pour l'apprentissage non supervisé, nous avons testé plusieurs techniques d'une part pour détecter les anomalies dans les systèmes de monitoring, d'autre part pour regrouper le système de tuyauterie en des boucles de corrosion. La structure de la phase de développement du modèle de prédiction sera présentée en détail dans les prochaines sections. Une fois que le modèle est développé, une phase de validation est nécessaire pour approuver la performance de l'algorithme, la validation se base sur la comparaison des résultats réels avec les prédits. Nous avons utilisé le logiciel

Python 3.9 avec les paquets nécessaires tels que *NumPy, Pandas et Scikit-learn* [44].

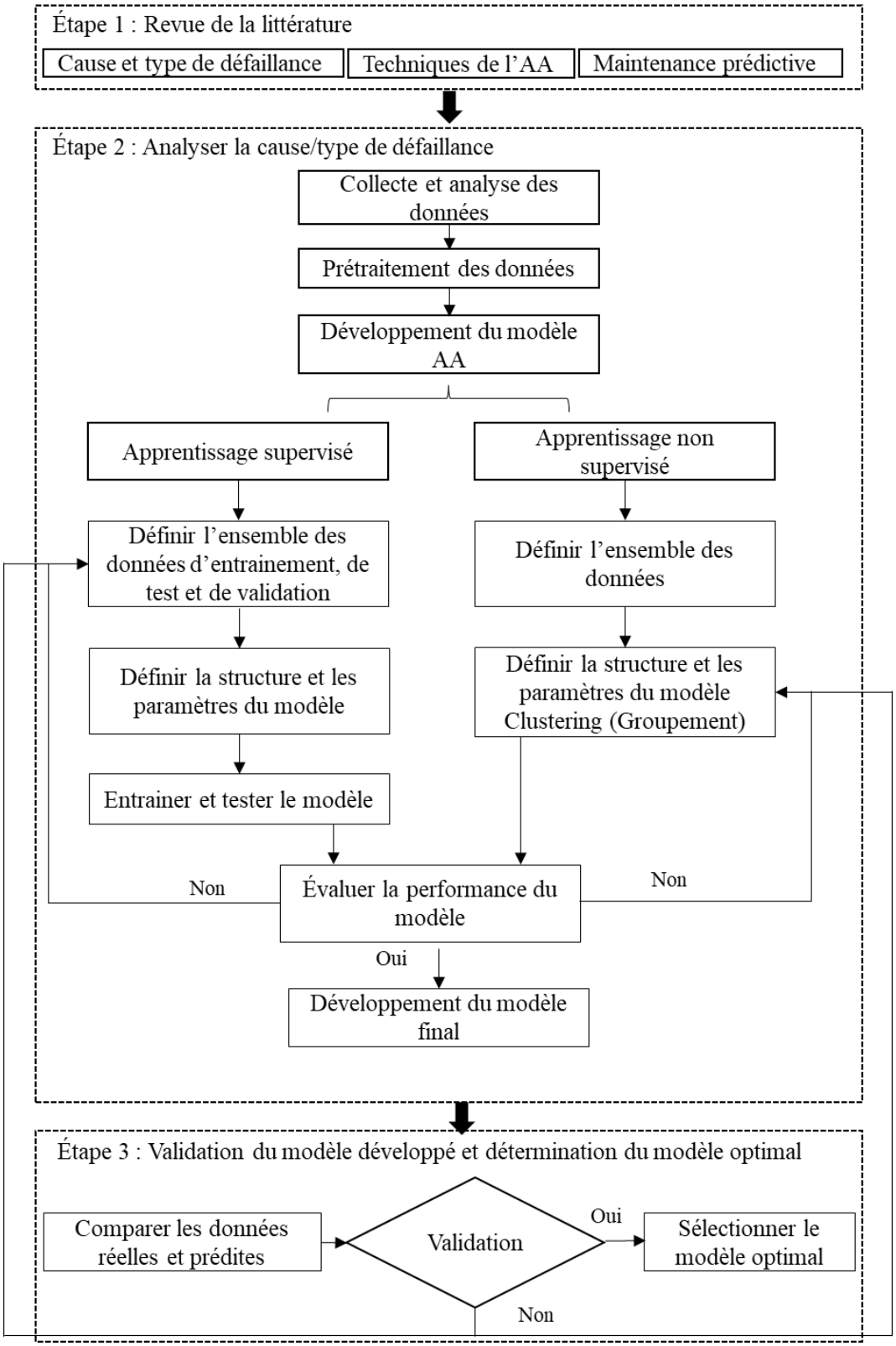


Figure 3.1 Méthodologie de recherche adaptée

3.1 Collecte des données

Pour ce projet, nous avons utilisé deux types de données, le premier type est issu des rapports d'inspection d'une raffinerie et d'une usine d'acide sulfurique situées au Québec et le deuxième est issu des données des capteurs ultrasoniques installés sur une conduite d'acide sulfurique. Les données peuvent être numériques ainsi que catégoriques. Le contenu de chaque base est présenté au niveau de 4.1. L'ensemble des deux bases peut être exploité pour des problèmes d'AA supervisé et non supervisé. Les paramètres nécessaires pour construire un modèle de prédiction de défaillance sont présentés au niveau du Tableau 3.1.

Tableau 3.1 Type des données d'entrée du modèle de prédiction

No	Facteur	Description	Type de variable
1	Service	Le fluide qui circule dans les conduites	Qualitative
2	Température	La température opératoire du fluide (°F)	Numérique
3	Diamètre	Diamètre de la conduite (pouce (po))	Numérique
4	Type de circuit	Type de la section de la tuyauterie	Qualitative
5	Isolation	Type d'isolant	Qualitative

3.2 Prétraitement des données

Pour exploiter les données dans les algorithmes d'apprentissage automatique, plusieurs étapes de prétraitement des données sont nécessaires [45]. Ce qui suit explique toutes les étapes requises pour le modèle proposé dans l'ordre souhaité :

3.2.1 Nettoyage des données

Dans tous les projets, la première étape du prétraitement des données consiste à examiner les valeurs manquantes, redondantes et aberrantes [45]. Il existe plusieurs approches pour traiter ces valeurs : à condition que le nombre total de ces valeurs soit insignifiant par rapport à la quantité de données disponibles, la meilleure pratique consiste simplement de les supprimer de la base. Sinon, si la perte de tous ces échantillons entraîne la perte d'une quantité considérable de données, le remplacement de ces valeurs par la moyenne ou la médiane de la même caractéristique dans les mêmes conditions est une autre piste.

3.2.2 Transformation des données

Les algorithmes d'AA ne travaillent qu'avec des nombres, c'est la raison pour laquelle il faut convertir les variables catégoriques en numériques. Ce processus est appelé encodage catégoriel [46]. Il existe deux types d'approche :

- Encodage d'étiquette où *Label Encoding* : Pour cette technique, chaque caractéristique se voit assignée à un entier unique basé sur l'ordre alphabétique. Généralement ce type d'encodage est utilisé lorsque les données catégoriques sont ordinales.
- Encodage à un coup où *One-hot Encoder* : Cette technique convertit les données catégorielles en divisant la colonne qui contient des variables catégoriques en plusieurs colonnes numériques binaires. Comme montre la Figure 3.2 ci-dessous.

Service
ACIDGAS
KEROSENE
BRUT

Service_ACIDGAS	Service_KEROSENE	Service_BRUT
1	0	0
0	1	0
0	0	1

Figure 3.2 Encodage de type *One Hot*

3.2.3 Normalisation des données

La normalisation des données est une étape essentielle, car une base de données peut contenir des valeurs avec des plages et des unités différentes. Par exemple, si une caractéristique a une très large gamme de valeurs par rapport aux autres, cette variable dominera le processus de décision de l'algorithme. Par conséquent, la mise à l'échelle/la normalisation doit être effectuée pour toutes les caractéristiques de sorte que chacune d'entre elles contribue de la même manière à la décision finale. Nous examinerons à la suite les méthodes les plus populaires :

- *Min-Max Scaler* [47] : cette méthode consiste à mettre à l'échelle les variables à un intervalle entre 0 et 1, ce qui peut être réalisé en utilisant l'équation suivante :

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (4)$$

Où : X’ : est la valeur normalisée. Min(X), Max(X) : le minimum et le maximum de la variable.

- *Standard Scaler* [48] : Il est également une méthode largement employée dans laquelle la distribution des données est modifiée pour avoir une moyenne nulle et une variance unitaire. La nouvelle variable est obtenue à partir de l’équation suivante :

$$X' = \frac{X - moyenne(X)}{\sigma(X)} \quad (5)$$

Où : X′ est la valeur normalisée, la moyenne (X) est la moyenne du vecteur de la variable X, et σ est l'écart type.

- *Robust Scaler* [49] : cette technique met à l'échelle les fonctionnalités qui sont robustes aux valeurs aberrantes. La méthode qu'il suit est presque similaire au Min-Max Scaler, mais il utilise la plage interquartile. La nouvelle variable est obtenue à partir de l’équation suivante :

$$X' = \frac{X - Q_1(X)}{Q_3(X) - Q_1(X)} \quad (6)$$

Où : Q_1 est le premier quartile, et Q_3 est le troisième quartile des données.

Il convient de noter qu'une méthode n'est pas toujours plus performante que les autres. C'est toujours intéressant de les tester pour les comparer. Pour ce projet, les résultats obtenus peuvent confirmer que le *Robust Scaler* produit les meilleurs résultats en termes de prédiction.

3.3 Développement de modèles de prédiction

Une fois la base de données est préparée, le modèle de prédiction peut être développé. Pour le cas d’apprentissage supervisé, la BD doit être divisée aléatoirement en 3 ensembles :

- L’ensemble d’entraînement : c’est la portion de données utilisée pour entraîner le modèle. Le modèle est destiné à observer et à apprendre à partir de l'ensemble des données. Cette portion représente 60% de l’ensemble de données.
- L’ensemble de tests : c’est la portion de données qui est testée dans le modèle final et qui est comparée aux ensembles de données précédents. L'ensemble de

tests sert à évaluer le modèle développé. Cette portion constitue 20% de l'ensemble de données.

- L'ensemble de validation : c'est la portion de données qui est utilisée pour valider la performance du modèle final. Cette portion constitue 20% de l'ensemble de données.

Dans ce qui suit, les différentes techniques adoptées dans cette recherche seront présentées.

3.3.1 Méthode d'apprentissage supervisé

La plupart du temps, l'application de l'apprentissage automatique repose sur l'apprentissage supervisé. Selon Guikema [50] l'apprentissage supervisé est une approche pour les applications dans lesquelles les données de résultat y_i sont enregistrées simultanément avec les données d'entrées x_i, qui peuvent toutes deux être obtenues à partir des données historiques. L'apprentissage automatique supervisé formule des hypothèses pour estimer la relation $y_i = f(x_i)$ sur la base de données. Deux techniques peuvent être adoptées pour développer un modèle prédictif :

- Les techniques de classification anticipent des réponses discrètes (par exemple, si la corrosion des pipelines est "mineure" ou "sévère"). Dans cette méthode, les données d'entrée et les résultats souhaités doivent être définis, collectés et organisés avant d'exécuter l'apprentissage automatique supervisé. Ainsi, les sorties de classification seront faites sur la base de ces données.
- Les techniques de régression permettent de prévoir des réponses continues. Elles sont généralement adaptées à la prédiction du nombre réel de fluctuations de variables telles que le taux de contrainte et/ou l'évolution de l'épaisseur d'une conduite.

Pour cette étude les algorithmes adoptés pour effectuer la prédiction sont : les méthodes basées sur les arbres de décision, machine à vecteur de support SVM (*Support vector machines*), et bien d'autres.

3.3.1.1 Arbre de décision

Les arbres de décision [51] comptent parmi les modèles d'apprentissage automatique supervisé non paramétriques les plus répandus dans le contexte de la

classification et de la régression. D'une part, il utilise une structure de type arbre ce qui améliore leur simplicité algorithmique et, d'autre part, en raison de la facilité d'interprétation et d'explication des résultats générés. Comme toutes les méthodes d'apprentissage, ces méthodes sont conçues à faire des corrélations entre les données d'entrée et sortie de chaque problème. Ces arbres possèdent une structure arborescente hiérarchique, comprenant un nœud racine, des branches, des nœuds internes (nœuds de décision) et des nœuds de feuilles (nœuds terminaux). Un arbre de décision repose sur un nœud racine, qui n'a pas de branches entrantes. Les branches sortantes du nœud racine alimentent ensuite les nœuds de décision. Sur la base des caractéristiques entrainées, les deux types de nœuds mènent des réponses pour produire des sous-ensembles homogènes, qui sont désignés par les nœuds feuilles ou les nœuds terminaux. Chaque feuille est caractérisée par un chemin spécifique à travers l'arbre appelé règle. L'ensemble de ces règles sert à la construction du modèle. Un arbre de décision est destiné à concevoir des règles à partir des données qui permettent de les ordonner et d'accroître leur homogénéité. Ces règles sont basées sur plusieurs notions telles que l'index (critère de *split*), l'entropie, la Gini (indice de concentration), le nombre et la profondeur d'arbres. L'objectif principal de cette technique est d'apprendre ces règles afin de prédire la valeur d'une variable cible.

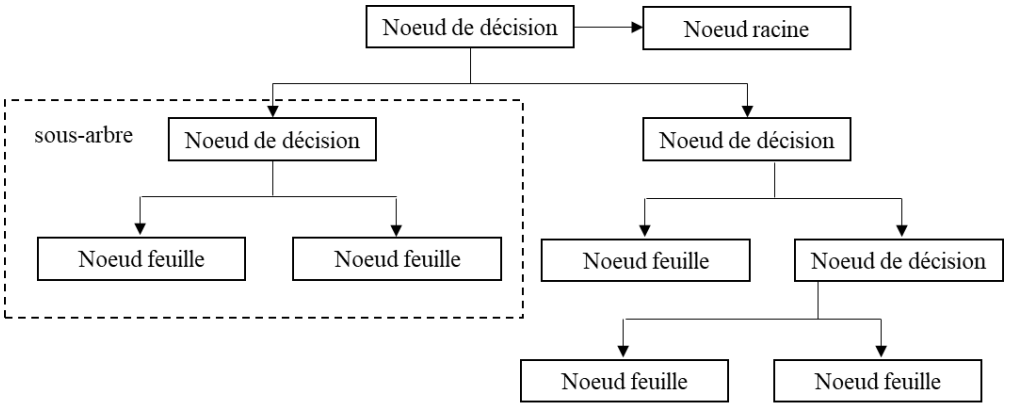


Figure 3.3 Structure de l'arbre de Décision

3.3.1.2 Les méthodes d'ensemble

L'apprentissage d'ensemble est une méta-approche générale de l'apprentissage automatique qui cherche à obtenir de meilleures performances prédictives en combinant les prédictions de plusieurs modèles uniques (comme les arbres de décision). Ces techniques servent à éliminer les erreurs d'un simple modèle

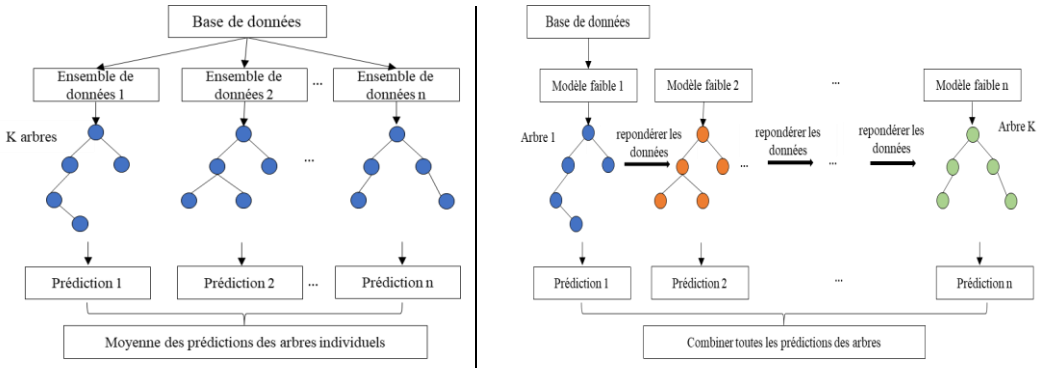
prédictif, en particulier les problèmes de sur et sous-apprentissage et elles peuvent être classifiées en deux sous-catégories : les méthodes d'ensemble parallèles (*Bagging*) et les méthodes d'ensemble séquentielles (*Boosting*) [52].

- Les méthodes d'ensemble parallèles (*Bagging*)

Le Bagging, aussi appelé *Bootstrap Aggregating*, est l'une des techniques d'ensemble dans laquelle les prédicteurs (arbre de décision) sont générés de façon indépendante et en parallèle. Cette méthode consiste à sous-échantillonner les données, en créant une BD pour chaque modèle de sorte que tous les modèles sont un peu différents les uns des autres. Le résultat final est déterminé par un vote des résultats de chaque modèle pour la classification, ou par une moyenne des résultats pour la régression. Un exemple de ces méthodes est le modèle forêt aléatoire (*Random forest*). La Figure 3.4 présente le diagramme simplifié de ce type d'algorithme [52].

- Méthodes d'ensemble séquentielles (*Boosting*)

Contrairement aux techniques d'ensemble parallèles, les modèles pour ces méthodes sont générés d'une façon séquentielle (en série) et dépendante. Cette méthode tente de construire un modèle fort à partir d'un modèle faible (régression, arbre de décision) puis l'améliorer. À chaque itération, cette technique identifie les instances mal classées, en augmentant leur poids afin que le prochain modèle accorde une attention particulière pour les corriger. *AdaBoost*, qui signifie "algorithme de boost adaptatif", est l'un des algorithmes séquentiels les plus répandus, car il était l'un des premiers du genre. D'autres types d'algorithmes de *boost* incluent *XGBoost* et *GradientBoost* [52,53].



a) Méthodes d'ensemble parallèles b) Méthodes d'ensemble séquentielles

Figure 3.4 Structure des méthodes d'ensemble

Certaines règles (paramètres) doivent être implémentées dans l'algorithme afin d'atteindre les meilleures performances, ces paramètres sont souvent appelés 'hyperparamètre'[54,55]. Parmi ces derniers on trouve :

- Nombre d'arbres (*n_estimators*) : ce paramètre contrôle le nombre d'arbres à l'intérieur du modèle.
- Profondeur d'arbre (*max-depth*) : ce paramètre est l'un des paramètres les plus importants. Il régit la profondeur maximale jusqu'à laquelle les arbres à l'intérieur de la forêt peuvent pousser.
- Nombre de caractéristiques (*max_features*) : la forêt prend des sous-ensembles aléatoires de caractéristiques et tente de déterminer la meilleure répartition. *max_features* peut prendre quatre valeurs : "auto", "sqrt", "log2" et None.
- Flottant du taux d'apprentissage (*learning_rate*) : Pondération appliquée à chaque classifieur à chaque itération de boosting. (Pour *AdaBoost*)
- Algorithme(algorithm) : c'est la technique utilisée pour la pondération (pour *AdaBoost*)

3.3.1.3 Machine à vecteurs de support (SVM)

Le SVM, une approche d'apprentissage supervisé, introduit par Boser (1992), [56] est l'une des techniques d'AA les plus répandues et les plus simples, car ses résultats sont souvent parfaits et uniques. Le SVM peut être utilisé pour la prédiction de régression et de classification, car il maximise le taux de précision prédictive grâce à l'utilisation de la théorie AA et évite le surajustement des données. En outre, elle présente une forte capacité de généralisation en raison du principe de minimisation du risque structure [57]. Ce principe réduit l'intervalle de confiance tout en gardant les valeurs de l'erreur d'apprentissage constantes. Lors de l'utilisation du SVM, il faut considérer qu'il s'agit d'une technique non paramétrique (technique de dispersion), l'exécution de l'algorithme nécessite le stockage en mémoire de toutes les données pendant la phase d'apprentissage afin de déterminer les paramètres du modèle. Pour la prévision des futurs résultats, l'algorithme s'appuie sur des vecteurs supports qui sont un sous-ensemble de processus d'apprentissage. Comme illustré à la Figure 3.5, les vecteurs de support sont des points de données éparpillés autour d'une ligne droite appelée hyperplan, une ligne unique servant à séparer et classer les données. Les vecteurs de support influencent la position et l'orientation de

l’hyperplan et leur utilisation sert à maximiser la marge de l’algorithme. L'idée du SVM est de trouver un hyperplan qui réalise une séparation optimale [57]. En outre, la représentation de cet hyperplan varie selon la facilité avec laquelle les données peuvent être séparées, ce qui donne lieu à deux types de SVM, linéaires et non linéaires. Plusieurs hyperplans sont représentés sur la Figure 3.5, et un SVM sélectionnera le meilleur d'entre eux.

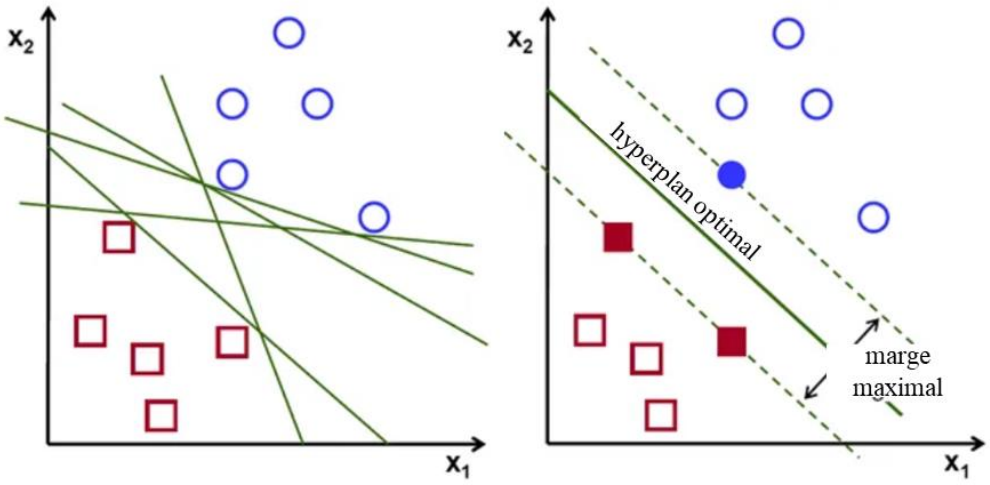


Figure 3.5 Méthode SVM

3.3.1.4 Réglage des paramètres

Les hyperparamètres sont les paramètres optimaux qui définissent la structure du modèle. Le réglage des hyperparamètres fait référence au processus de recherche et de sélection du paramètre optimal. La valeur de l'hyperparamètre ne peut pas être estimée à partir des données et doit être définie avant d'initier le processus d'apprentissage [54].

- Recherche par grille (*Gridsearch CV*)

La recherche par grille [54] est une méthode de base pour le choix des hyperparamètres. *Gridsearch CV* de *sklearn* permet la recherche par grille, où il génère des candidats à partir d'une grille de paramètres prédéfinis. À la suite de la mise en œuvre de la recherche par grille sur l'ensemble de données, la meilleure combinaison est retenue après l'évaluation de toutes les combinaisons possibles de valeurs de paramètres. Un exemple des paramètres à chercher pour les méthodes basées sur l’arbre de décision est le nombre ainsi que la profondeur d’arbre.

3.3.2 Méthode d'apprentissage non supervisé

Les techniques d'apprentissage non supervisées sont plus complexes que les autres stratégies en raison de l'absence d'étiquettes (résultats). Cependant, ils sont pertinents dans le domaine de l'apprentissage automatique, car ils permettent de réaliser efficacement des missions sophistiquées. Les principaux domaines d'application sont le clustering, la détection d'anomalies et la réduction de la dimensionnalité [58]. Dans ce projet on s'intéresse aux deux premiers domaines cités.

3.3.2.1 Clustering

Le clustering constitue la pierre angulaire de l'analyse intelligente des données, c'est le processus qui consiste à rassembler n objets en k groupes avec k inférieur à n (*cluster*) [59].

En général, les algorithmes non supervisés font des inférences à partir d'ensembles de données en utilisant les variables d'entrée d'une base de données (*features*), sans référence à des résultats connus ou étiquetés (*Target*). Les variables de données sont regroupées en clusters distincts par le processus de clustering, de sorte que les variables à l'intérieur d'un cluster sont similaires les unes aux autres sur la base d'une variété de critères de similarités prédéterminés. Pour cette technique trois paramètres spécifiques doivent être définis :

- Le nombre de clusters : La compréhension des données est nécessaire pour l'estimation de k.

pour la détermination il existe plusieurs approches, pour ce projet, la méthode de coude '*Elbow Method*' est utilisée pour déterminer le nombre optimal de cluster K en se basant sur la fonction coût présentée dans l'équation (7), K c'est le nombre de cluster qui atteignent la valeur de la fonction de coût le plus bas suivante [60] ;

$$Coût = \sum_{j=1}^{k} \sum_{x_t \epsilon B_j} dist(x_t, b_j) \tag{7}$$

- Initialisation des clusters : Initialement l'emplacement de chaque centroïde (centre d'un cluster) est arbitraire. Par conséquent, une initialisation différente des clusters peut donner lieu à un clustering différent.
- Définir le critère d'évaluation ; généralement la distance euclidienne (8)

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (8)$$

Où : x,y = Deux points de données
x_i,y_i = Vecteur euclidien
n= Espace n

En tenant compte du type de variables contenues dans l'ensemble de données, on peut identifier trois méthodes principales de clustering : Kmeans [60], Kmodes [61], et Kprotoypes [62].

Les k-modes sont utilisés pour le regroupement de variables catégorielles. Il définit les clusters sur la base du nombre de catégories concordantes entre les points de données. Contrairement à l'algorithme, K-means, qui classe les données numériques en fonction de la distance euclidienne. Cependant, dans le monde réel, les données collectées ont souvent des attributs à la fois numériques et catégoriels (données mixtes). Il est donc difficile d'appliquer les deux méthodes citées précédemment. Haung[59] propose l'algorithme k-prototypes qui combine les deux méthodes K-modes et K-means et qui est capable de regrouper des données mixtes.

- K-means

L'algorithme K-Means commence par des estimations initiales de K centroïdes qui sont choisis au hasard dans l'ensemble de données. Chaque cluster est représenté par un centroïde, à la suite la distance entre les centroïdes et les points de données est calculée. Ensuite, les points de données sont attribués au cluster qui a le centroïde le plus proche. La Figure 3.6 présente les étapes de l'algorithme de clustering K-Means :

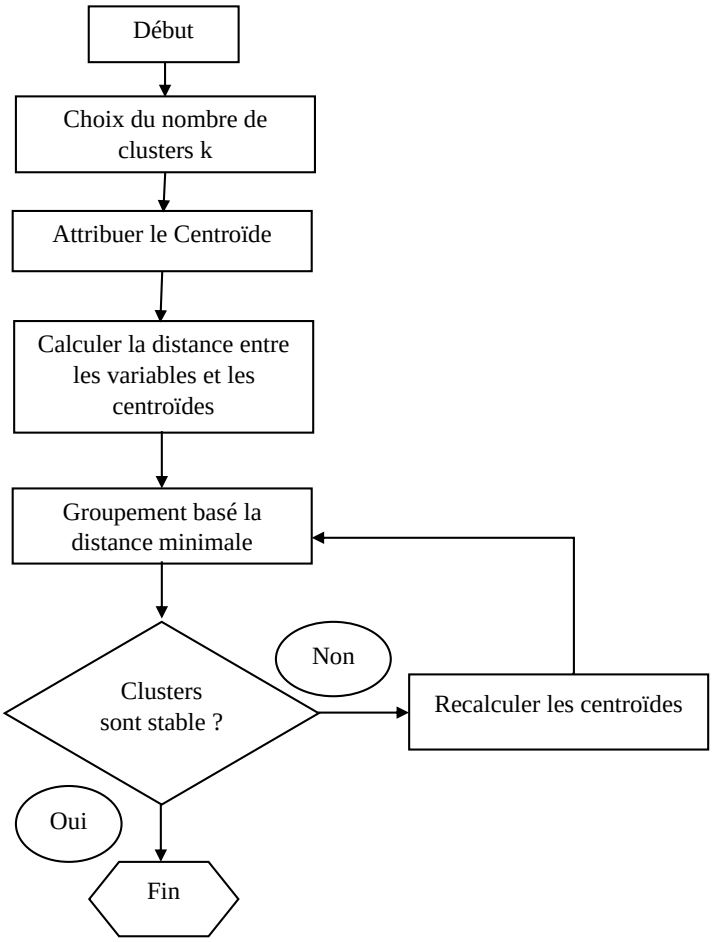


Figure 3.6 Algorithme de la méthode K-means

- K-modes

Le clustering en mode K est une technique d'apprentissage automatique non supervisée utilisée pour regrouper un ensemble d'objets de données dans un nombre spécifié de clusters[63], sur la base de leurs attributs catégoriels. Cette approche modifie le processus standard K-Means pour le regroupement de données catégorielles en remplaçant la fonction de distance euclidienne par une mesure de dissimilarité simple. L'algorithme est appelé "K-Mode", car il utilise les modes (les valeurs les plus fréquentes) au lieu des moyennes ou des médianes pour représenter les clusters. La distance entre les données est calculée à l'aide d'une mesure de dissimilarité appelée distance de Hamming qui est le nombre d'attributs catégoriels qui diffèrent entre les deux objets.

L'algorithme de clustering K-modes comprend les étapes suivantes :

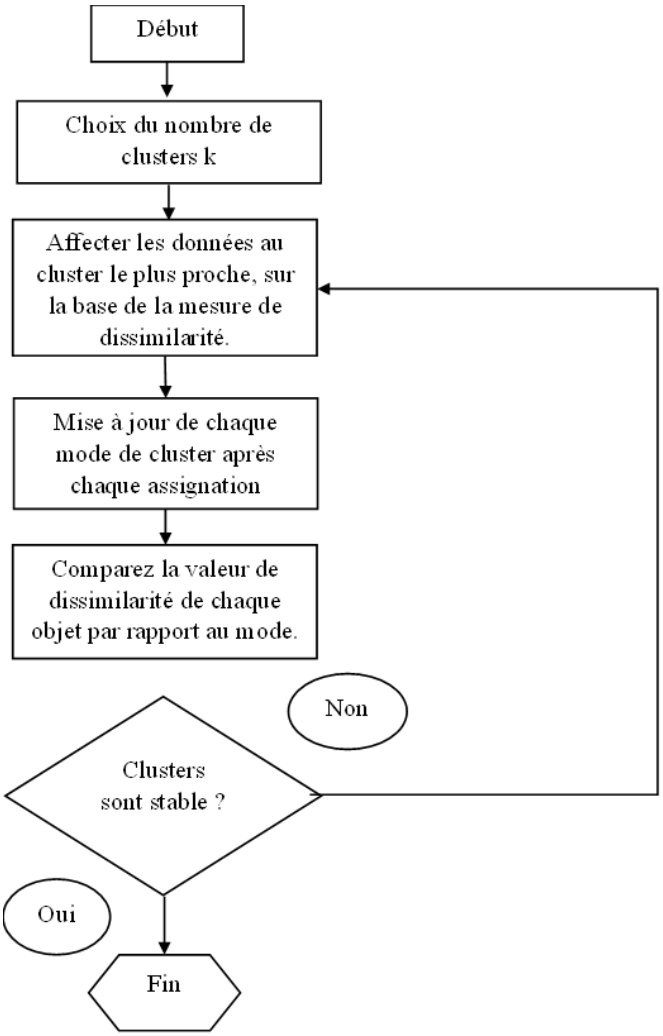


Figure 3.7 Algorithme de la méthode K-Mode

- K- prototype

K-prototype est une technique d'apprentissage automatique non supervisée [62,63]. Cet algorithme combine les « moyennes » de la partie numérique et les « modes » de la partie catégorielle pour créer un nouveau centre de cluster hybride appelé « prototype ». Sur la base du « prototype », il construit une formule de coefficient de dissimilarité équation (10) et une fonction de coût applicable aux données de type mixte d'équation (10). Le paramètre γ est introduit pour contrôler l'influence de la donnée catégorielle et numérique sur le processus de clustering. On suppose que l'ensemble de données de type mixte possède p données numériques et m – p données catégorielles. Pour tout x_i, $q_j \in D$.

La fonction coût est :

$$Coût = \sum_{j=1}^{k} \sum_{i=1}^{n} \mu_{ij} dist(x_i, q_j) \quad (9)$$

Coefficient de dissimilarité

$$d(x_i, q_j) = \gamma \sum_{s=1}^{p} \delta(x_{i,s}^{C} - q_{j,s}^{C}) + \sum_{s=p+1}^{m} \sqrt{(x_{i,s}^{N} - q_{j,s}^{N})^{2}} \quad (10)$$

Où

$$\delta(x_{i,s}, q_{l,s}) = \begin{cases} 0, x_{i,s} = q_{j,s} \\ 1, x_{i,s} \neq q_{j,s} \end{cases} \quad (11)$$

$q_{j,s}$: Centre du cluster
$x_{i,s}$: les données de la base
$x_{i,s}^{C}$: les données catégoriques de la base
$x_{i,s}^{N}$: les données numériques de la base
μ_{il} : est l'appartenance de i^{ème} observation de données au cluster j (valeur binaire)

Les étapes fondamentales de l'algorithme des k prototypes sont décrites comme suit :

1- Choix aléatoire du nombre de clusters k.
2- Utiliser l'équation (11) pour calculer la dissimilarité entre x_i et q_l. Selon le résultat du calcul, x_i est affecté au cluster le plus proche.
3- Selon les centres de clusters actuels, la dissimilarité de l'objet de données est recalculée. Réaffecter les objets de données au sous-cluster le plus proche, les valeurs avec la fréquence la plus élevée sont utilisées dans la partie catégorielle, et la partie numérique utilise la méthode de la valeur moyenne pour déterminer. Mettez à jour les centres de clusters.
4- Répéter les étapes 2 et 3 jusqu'à ce que la fonction de coût ne change plus. Si la fonction de coût ne change plus, l'algorithme se termine. Sinon, passez à l'étape 2 pour continuer.

3.3.2.2 Détection d'anomalie

La surveillance est une tâche difficile dans le paradigme de l'industrie 4.0. Elle est étudiée non seulement dans le secteur des pipelines, des bâtiments, etc., mais aussi dans les véhicules, les avions et les appareils biomédicaux. La détection d'anomalies dans les systèmes de surveillance est d'une grande importance pour la maintenance prédictive. Elle s'agit d'identifier les évènements et les

comportements inhabituels des actifs. L'application de l'apprentissage non supervisée dans ce cas s'agit de rassembler des données ayant des caractéristiques similaires qui permettent d'éliminer les données anormales (anomalies). Pour ce projet, nous avons appliqué différentes techniques pour détecter les anomalies dans le cas des capteurs ultrasoniques. Parmi ces algorithmes, citons le clustering basé sur la densité, le clustering basé sur la distance [64] et la forêt d'isolement [65] aussi on a utilisé la méthode interquartile qui est une méthode statistique[66]

- Méthode d'écart interquartile (IQR)

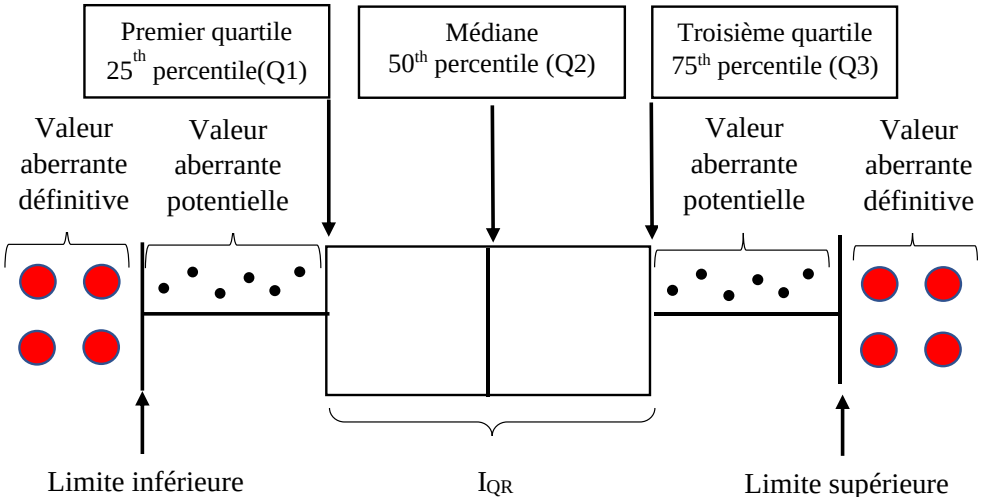


Figure 3.8 Description de la méthode IQR

Traditionnellement, l'ensemble de données d'une base est représenté à l'aide d'un résumé à cinq chiffres, qui comprend la valeur minimale et maximale, la médiane, et le premier et le troisième quartile, qui sont les valeurs qui séparent le premier et les trois premiers quarts des données du reste des données, respectivement [66]. Ces valeurs apportent plus d'informations sur la dispersion des données que les simples lignes et colonnes. La Figure 3.8 présente une distribution d'un ensemble de données. La différence entre Q_1 et Q_3 est l'écart interquartile (I_{QR}), qui reflète l'écart de l'ensemble de données par rapport à la médiane.

$$I_{QR} = Q_3 - Q_1 \tag{12}$$

Les limites inférieure et supérieure sont représentées comme suit :

$$F_L = Q_1 - 1.5 I_{QR} \tag{13}$$

$$F_U = Q_3 + 1.5 I_{QR} \tag{14}$$

L'ensemble des données qui se trouvent au-delà des limites F_L et F_U représentent les valeurs aberrantes (Outliers). Le 1.5 préserve la sensibilité de l'ensemble des données.

Pour une base de données, il existe deux types de valeurs aberrantes : Valeur aberrante potentielle (O_P) et valeur aberrante définitive (O_D) :

$$F_L < O_P < Q_1 \; ou \; F_U < O_P < Q_3 \tag{15}$$

$$O_D < F_L \; ou \; F_U < O_D \tag{16}$$

- Méthode basée sur la distance

Les méthodes basées sur la distance sont des algorithmes de l'apprentissage automatique qui classifient les données selon la distance entre les différentes variables constituant la base [67]. Cette méthode est basée sur l'approche des K-plus proches voisins (K-NN), elle peut être utilisée pour des problèmes d'apprentissage supervisé ainsi que les non supervisés. Cette technique a pour objectif de localiser tous les voisins les plus proches autour d'un nouveau point de donnée afin de déterminer la classe qui lui appartient. Pour ce projet, la K-NN est proposée pour la détection des valeurs aberrantes, ceci exploite la relation entre les divers voisins, où la distance entre un objet et son K-plus proche voisin fournit des informations si la valeur cible est considérée ou non comme une valeur aberrante dans les points de données, ce qui suppose que les valeurs aberrantes sont éparpillées ou éloignées de leurs proches voisins. Dang et al. [67] ont présenté les étapes fondamentales de cette méthode comme indiques au dans la Figure 3.9 :

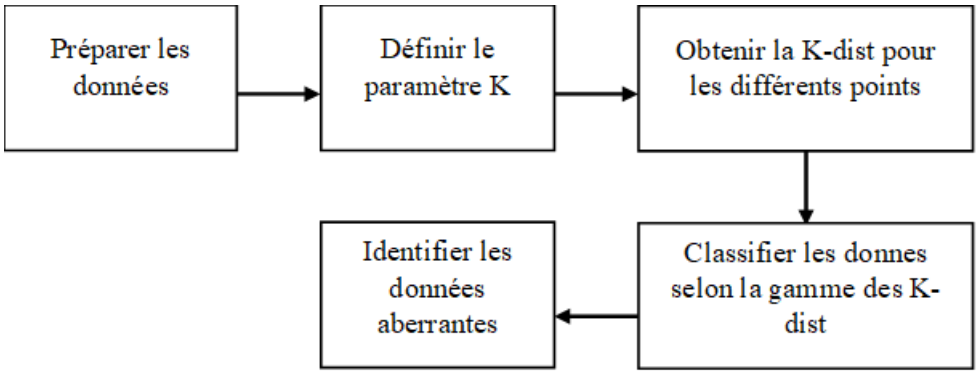


Figure 3.9 Méthode basée sur la distance

- Méthode basée sur la densité

Étant donné que le traitement des bases de données ayant des densités différentes peut s'avérer difficile pour les approches basées sur la distance, les techniques basées sur la densité pourraient être exploitées pour traiter ce type de données. Elles visent à attribuer à chaque exemple de données un degré d'aberration, appelé *Local Outlier Factor* (LOF) [68]. Pour ce type d'approche, une donnée aberrante est reconnue si sa densité locale est significativement différente de celle de ses voisins. Plus précisément, la localité est déterminée par les K-NN, dont la distance est un critère permettant d'évaluer la densité locale. Les étapes suivantes peuvent être considérées pour la détermination du LOF :

1) Calculer la distance entre les différents points de données à l'aide d'une des fonctions de distance telles que la fonction euclidienne ou la fonction de Manhattan.
2) Trouver le k (k-plus proche voisin) point le plus proche
3) Trouvez les k points les plus proches.
4) Trouvez la densité d'accessibilité locale à l'aide de l'équation suivante :

$$lrd_k(O) = \frac{\|N_k(O)\|}{\sum_{O' \in N_k(O)} reachdist_k(O' \leftarrow O)} \tag{17}$$

Où $reachdist_k$ est égal à :

$$reachdist_k(O' \leftarrow O) = max\{dist_k(O), dist(O, O')\} \tag{18}$$

Avec $N_k(O)$ désigne le nombre de voisins

5) Calculer le facteur de valeur aberrante (LOF) :

$$LOF_k(O) = \frac{\sum_{O' \in N_k(O)} \frac{lrd_k(O')}{lrd_k(O)}}{\|N_k(O)\|} \tag{19}$$

- Forêt d'isolement

L'algorithme de forêt d'isolement (IF) a été introduit par Liu et al. [65] et est basé sur la construction des arbres de décision. Comme son nom l'indique, cet algorithme isole les variables en sélectionnant aléatoirement une donnée dans un échantillon et en choisissant aléatoirement une valeur de division entre les valeurs min et max de cette donnée. Ensuite, si la valeur choisie maintient le point au-dessus, il faut

associer la valeur minimale de la plage de donnée à cette valeur, sinon, si la valeur choisie maintient le point au-dessous, modifiez la valeur maximale de la plage de donnée à cette valeur. Ces deux dernières étapes doivent être répétées jusqu'à ce que les points soient isolés. Le nombre de fois que ces étapes sont répétées et appelées "nombre d'isolement". Pour chaque point de données, un score d'anomalie, équation (20), est déterminé en calculant la longueur du chemin moyen entre la racine de l'arbre et le nœud entourant le point :

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{20}$$

Où

$$c(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \tag{21}$$

n c'est le nombre de points de données

c(n) est une métrique de référence qui normalise le score entre 0 et 1

E(h(x)) désigne la moyenne des longueurs

H est le nombre harmonique et peut être estimé comme suit :

$$H(i) = \ln(i) + \gamma \quad (\gamma: constant\ d'euler = 0.5772156649) \tag{22}$$

La forêt d'isolement a le même concept de classification que la forêt aléatoire. Si un point s'enfonce plus profondément dans l'arbre, on ne s'attend pas à ce qu'il soit aberrant. En revanche, si un point se trouve dans des branches plus courtes, il est plus probable qu'il soit aberrant, car les valeurs attribuées s'écartent fortement de la norme comme indique la Figure 3.10.

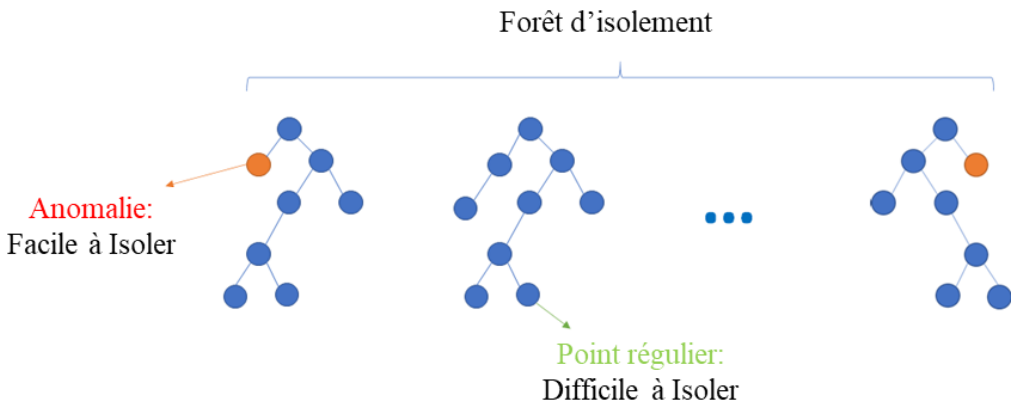


Figure 3.10 Structure de la forêt d'isolement

3.3.3 Évaluation de la performance des modèles

Après avoir implémenté un algorithme d'apprentissage automatique, l'étape suivante consiste à surveiller la pertinence du modèle prédictif en se basant sur des ensembles de mesure de performance [69]. Ces métriques dépendent de la tâche de modélisation (une régression ou classification) ainsi que la dispersion des données (déséquilibré ou non). L'objectif de ces métriques est de sélectionner le modèle qui a les meilleures performances.

3.3.3.1 Métriques pour la classification

L'évaluation de performance d'un modèle de classification se base sur la matrice de confusion, un tableau de taille NxN [70], a pour objet de visualiser et comparer les données prédites et les données observées comme indiqué au niveau duTableau 3.2 . 'n'est le nombre de classes.

		Classe actuelle	
		Positif	Négatif
Classes prédites	Positif	VP	FP
	Négatif	FN	VN

Tableau 3.2 Matrice de confusion

VP, (vrais positifs) : nombre de résultats initialement positifs et prédits comme positifs.

FP, (Faux Positif) : nombre de résultats initialement négatifs, mais prédits positifs.

FN, (Faux Négatifs) : nombre de résultats initialement positifs, mais prédits négatifs.

VN, (vrais négatifs) : nombre de résultats initialement négatifs et prédits négatifs.

La matrice de confusion sert à calculer les métriques suivantes :

- Taux de sucées (Accuracy) : C'est la proportion des données qui sont correctement classifiées.

$$A = \frac{VP + VN}{VP + FP + FN + VN} \quad (23)$$

  C'est la métrique la plus utilisée pour évaluer la performance d'un modèle. Cependant, elle peut nous amener en erreur dans le cas d'un ensemble de données déséquilibré (ensembles de données avec une répartition inégale des classes).

- Précision : s'agit d'une mesure de l'exactitude qui est atteinte dans la vraie prédiction. En termes simples, il s'agit de la vraie prédiction positive divisée par le nombre total de prédictions positives. La précision est une meilleure mesure pour les données équilibrées.

$$P = \frac{VP}{VP + FP} \tag{24}$$

- Rappel : c'est le pourcentage des données positives correctement classifiées, appelé aussi sensibilité.

$$R = \frac{VP}{VP + FN} \tag{25}$$

- Spécificité : c'est le pourcentage des données négatives correctement classifiées.

$$S = \frac{FP}{FP + VN} \tag{26}$$

- Score F1 : une mesure qui combine à la fois la précision et le rappel est égale à la moyenne harmonique de la précision et du rappel. Plus que la valeur de F1 s'approche de 1, plus que le modèle a la capabilité de faire une bonne décision.

$$F_1 = 2 * \frac{P * R}{P + R} \tag{27}$$

- La courbe AUC (*Area Under the Curve*) - ROC (*Receiver Operating Features*) : (également écrit comme AUROC) est l'une des mesures d'évaluation les plus importantes pour vérifier les performances de tout modèle de classification [71].
    - ROC est une courbe de probabilité : est un tracée entre les taux des faux positifs (TFP) (axe X) et celui des taux des vrais positifs (TVP) (axe Y) comme indique la Figure 3.11 ou TVP= sensibilité et la TFP =1-Spécificité. Il faut évaluer l'espace vide en haut à gauche de la courbe pour sélectionner le meilleur modèle. Plus l'espace est petit (proche du coin supérieur), le meilleur est le résultat.
    - AUC représente le degré ou la mesure de séparabilité, c'est un indicateur quantifier numériquement. Un excellent modèle a une AUC proche de 1, ce qui signifie qu'il a une bonne mesure de séparabilité. Un modèle médiocre a une AUC proche de 0, ce qui signifie qu'il a la pire mesure de séparabilité. En fait, cela signifie qu'il rend la pareille au résultat. Il prédit les 0 comme des 1 et les 1 comme des 0. Et lorsque AUC est de

0,5, cela signifie que le modèle n'a aucune capacité de séparation de classe. Le modèle qui a l'AUC la plus élevée c'est celui qui est le plus performant.

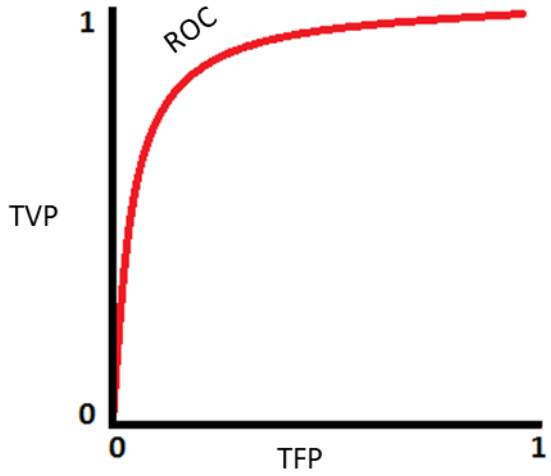


Figure 3.11 Courbe AUC-ROC

3.3.3.2 Métriques pour la régression

Les métriques utilisées pour évaluer la performance d'un modèle de régression sont :

- Erreur absolue moyenne (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{28}$$

- Erreur quadratique moyenne (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{29}$$

- R^2 également appelé coefficient de détermination :

$$R2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{30}$$

Avec N : c'est le nombre des données entrainées

y_i c'est la variable cible réelle

$\hat{y}_i$ est la variable prédite du modèle

$\bar{y}$ c'est la valeur moyenne des y_i

3.4 Récapitulatif

La méthodologie de la présente recherche inclut une revue extensive de la littérature, la collecte et la préparation des données, le développement de modèles de prédiction de défaillance dans les systèmes de tuyauterie à l'aide des techniques d'AA. De plus, la méthodologie décrit la procédure suivie dans cette recherche pour développer des modèles de prédiction de défaillance pour les oléoducs en démontrant chaque étape, de la collecte des données à la validation du modèle.

CHAPITRE 4: RÉSULTATS

4.1 Applications industrielles

Ce chapitre présente une introduction aux données historiques qui ont été collectées afin de développer des modèles d'AA de prédiction de l'état des actifs. Pour les industries pétrolières et gazières, les données sont récoltées à partir des sources non structurées telles que ; les rapports d'inspections visuelles et les données brutes de capteurs. L'étape de collecte est une étape clé qui consiste à extraire, transformer et charger des données dans un format utile et gérable.

La préparation des données constitue une étape fondamentale de l'AA, cette étape comprend la détection des valeurs aberrantes, l'exclusion ou le remplacement des informations manquantes provenant de sources tierces. La visualisation des données est souvent l'outil le plus pertinent pour l'analyse des données. Des visualisations convaincantes peuvent aider à décrire l'histoire des données, ce qui peut aider à mieux comprendre l'état des actifs.

Pour ce mandat trois bases de données ont été élaborées, la première contient des données historiques d'inspection d'une raffinerie installée au Québec[1]. La deuxième contient des données issues à partir des rapports d'inspection visuels et des fiches d'inspection des épaisseurs des conduites d'une usine d'acide sulfurique. Alors que la dernière est basée sur des données instrumentales issues d'un dispositif expérimental qui est conçu à contrôler par UT l'amincissement d'épaisseur d'une conduite d'acide sulfurique.

Les données fondamentales constituant chaque BD comprennent les dates d'inspection, les relevés de mesure d'épaisseur, les paramètres opératoires (température, pression, débit et concentration), type de fluide et le type de défaut.

4.1.1 Étude de cas n◦ 1 : analyse des données d'une raffinerie

4.1.1.1 Collecte et préparation des données

Les données utilisées dans cette section sont provenant des rapports d'inspection de mesure d'épaisseur d'une raffinerie installée au Québec. Les données recueillies, qui consistent en 80 289 relevés d'inspection, sont composées d'un ensemble de

[1] Le partenaire industriel veut conserver le nom de la raffinerie confidentiel

caractéristiques telles que les fluides circulant dans les pipelines, la date d'inspection, les mesures d'épaisseur, la localisation des points d'inspection, et les méthodes utilisées pour la surveillance. Par conséquent les données collectées sont de type mixte (numériques et catégoriques). Cette base sera exploitée pour des problèmes de classification et de régression au niveau de le deuxième paragraphe de ce chapitre. Le Tableau 4.1 décrit les caractéristiques principales constituant la base.

Tableau 4.1 Données de la raffinerie

Caractéristique	**Explication**
Unité de production	4 unités (01,04,10,32)
Equip_id	Référence de la ligne inspectée
Type de circuit	Environnement corrosif similaire plus précise que faisant partie de l'EQUIP ID - MPF = "Main Process Flow". Région principale de l'EQUIP ID qui n'est pas couvert pas un autre type de circuit ID (ex : point d'injection ou deadleg). - DL = "Deadleg". Point mort selon les critères établis. - IP = "Point d'injection". Sont des endroits où des produits chimiques, ou des additifs de processus sont introduits dans un flux de processus (inhibiteurs de corrosion, neutralisants, les piégeurs d'oxygène et d'hydrogène.). - MX = "Point de mélange". Régions où 2 produits avec gradient de température au-delà d'un certain écart se rencontrent et se mélangent.
Service	Les fluides circulant Annexe 1
fluid_state	État du fluide
Class_code	Classe du service selon API 571.
Read_dt	Date d'inspection
Operating temperature	Température opératoire du fluide (F)
Operating pressure	Pression opératoire du fluide (psig)
Diameter	Diamètre du tuyau
Thickness	Mesure d'épaisseur (pouce (po))
Température	Température de la paroi lors de la prise de mesure (F)
Material	Matériau du tuyau

Nominal Thickness	Épaisseur nominale
Access_method	Méthode d'accès pour faire l'inspection
Tml_id	Point précis d'évaluation de la condition de la tuyauterie (CML)
Read_method_code	Méthode d'inspection : - N = Épaisseur nominale (n'est pas une mesure chantier) - U = UT - C = Phased Array - R = Radiography - V = Épaisseur évaluée par Pit gage (read method "V") à partir de la dernière lecture précédente réelle (peut être le nominal). - D = Radiographie digitale. - O = Pour "OTHER" (peut être le nominal)
Key_words	Indice de l'élément de tuyauterie sur lequel la CML est située.
Insulation	Type de l'isolant utilisé

D'autre part, on a ajouté à partir des documents fournis deux autres caractéristiques :

- La cause de défaillance liée à chaque élément inspecté, cette information est extraite à partir des dessins isométriques pour chaque circuit de tuyauterie.
- Le taux de corrosion (mpy : millième de po/année) pour chaque Tml en se référant à l'équation extraite de l'API 570 :

$$CR = \frac{t_{ini} - t_{actu}}{T} \quad (31)$$

CR : Taux de corrosion.

t_{ini} : Épaisseur initiale.

t_{act} : Épaisseur mesurée après une période T.

T : Intervalle du temps qui sépare deux relevés successifs d'épaisseur.

4.1.1.2 Analyse des données

Pour les raffineries, il est question d'identifier les zones qui ont des taux de dégradation importante. Afin de comprendre les problèmes de corrosion et les solutions dans l'industrie du pétrole, du gaz et du raffinage, nous allons décrire les caractéristiques des fluides circulant, les matériaux utilisés, les géométries des éléments de tuyauterie (type et dimension) et leur corrosivité.

D’après l’analyse effectuée avec le logiciel Python, on a pu remarquer qu’il y a des taux de corrosion négatifs Figure 4.1. Plusieurs raisons expliquent ces valeurs négatives telles que :

- Changement de l’élément corrodé par un neuf ayant une épaisseur plus grande.
- Changements des points de mesure surtout pour les lignes non isolées.
- Imprécision au niveau des méthodes d’inspection utilisées

Il est à noter que les taux de corrosion positifs eux-mêmes sont sujets à des imprécisions, et ce pour cette même dernière raison.

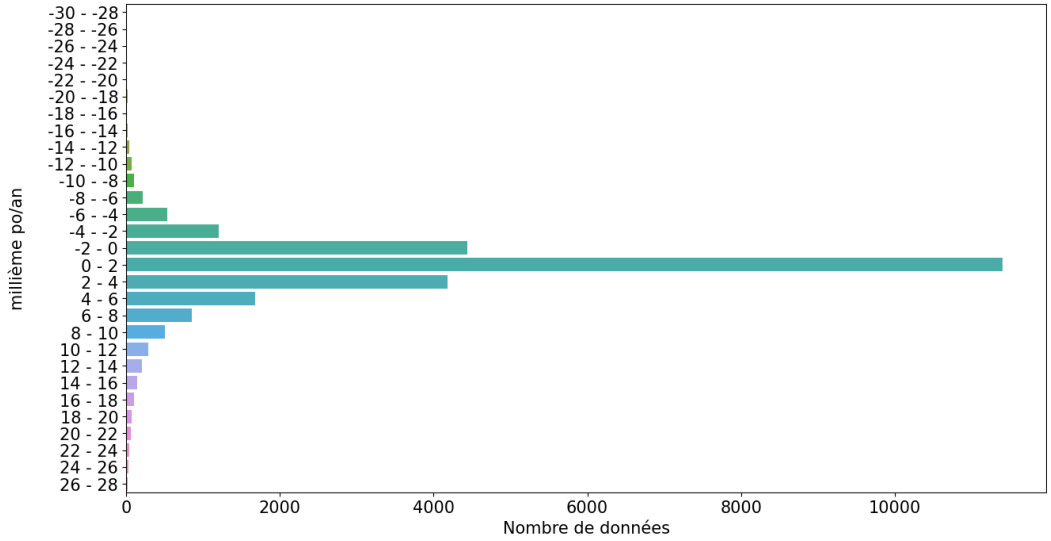
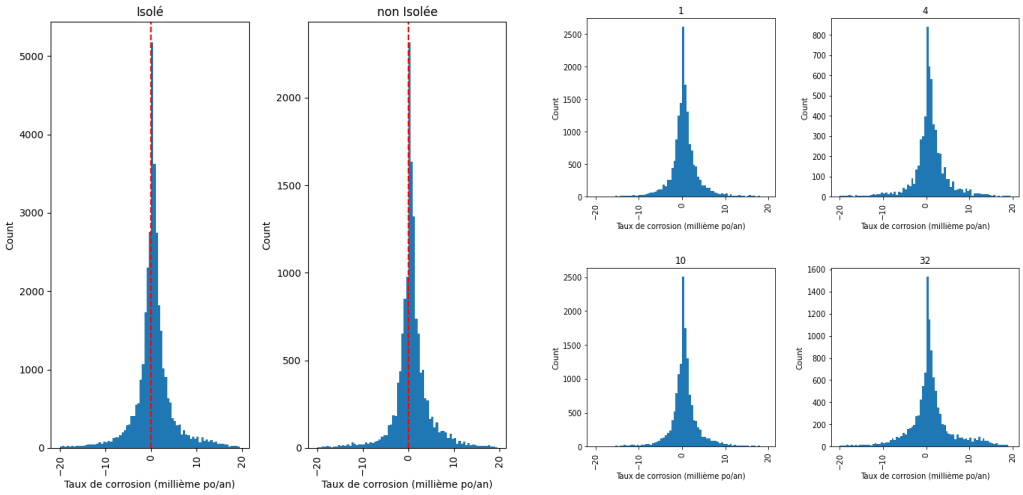


Figure 4.1 Distribution des taux de corrosion

D’après les résultats présentés dans les Figure 4.2 et Figure 4.3, on remarque que la distribution des taux de corrosion est similaire pour les lignes isolées et non isolées ainsi que pour les différentes unités de production. D’autre part, il y a beaucoup d’incertitude au niveau des intervalles d’inspection présentés précédemment.



a-Distribution des taux de corrosion pour les lignes isolées et non isolées

b- Distribution des taux de corrosion par unité de production

Figure 4.2 Distributions des taux de corrosion par a) type d'isolant et b) unité de production

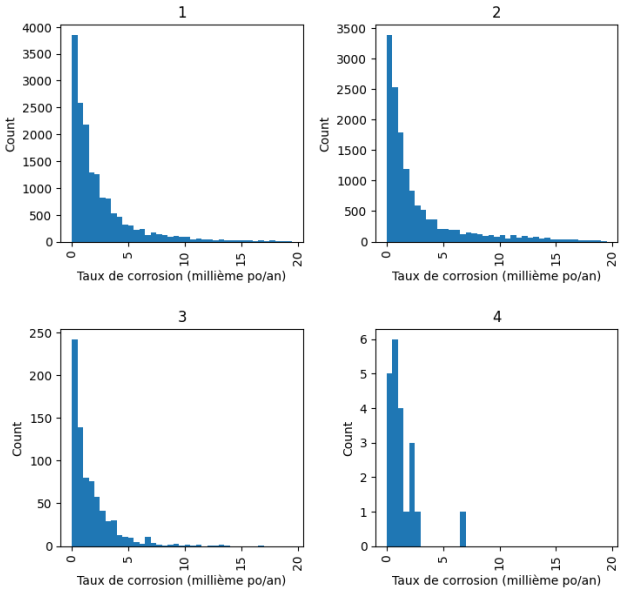


Figure 4.3 Variation de l'intervalle d'inspection par classe de fluide

Pour continuer notre analyse, on a adopté la démarche suivante pour minimiser les enjeux d'incertitude et aussi pour avoir une BD utile pour les prochains travaux de l'apprentissage automatique :

- Supprimer les mesures d'inspection non-chantiers (Épaisseur nominale).
- La durée d'inspection qui sépare deux relevées de mesure doit être plus qu'un 1 an et maximum 15 ans
- Supprimer les taux de corrosion négative.

En adoptant cette démarche, on a pu remarquer que le taux de corrosion est généralement faible. On a utilisé les box-plot du logiciel python pour visualiser la dispersion du CR, la Figure 4.4 montre que la majorité des taux sont situées dans l'intervalle $[0.5mpy - 3mpy]$

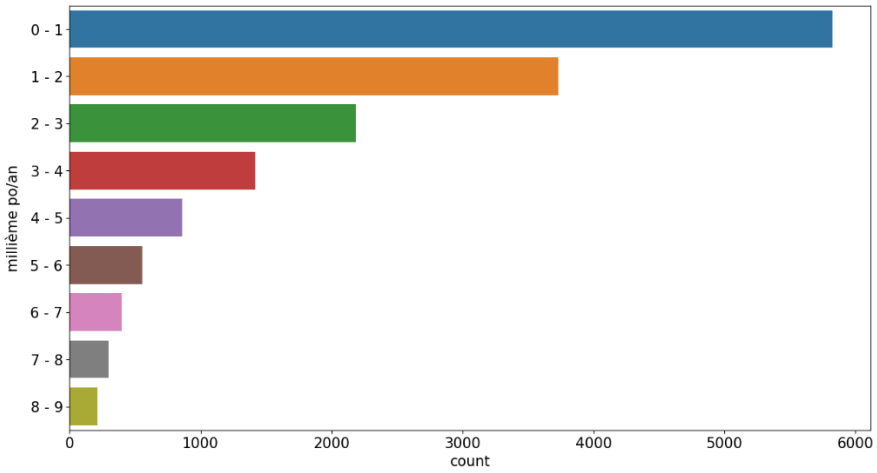


Figure 4.4 Distribution des taux de corrosion positive

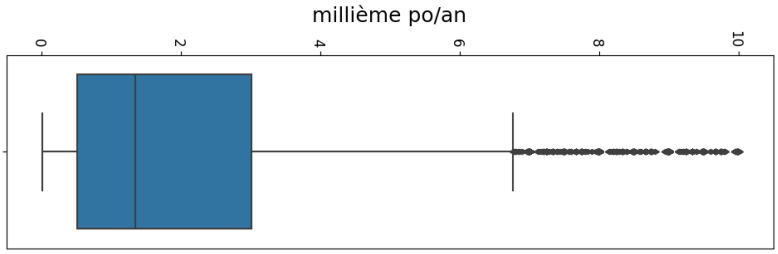


Figure 4.5 Dispersion des taux de corrosion

- Variation des taux de corrosion durant les dernières années

On a commencé notre analyse, par la visualisation de la distribution des taux de corrosion durant les dernières années, d'après la Figure 4.6 , Il est remarquable que le taux de corrosion ait diminué pendant les 10 dernières années. Il y a plusieurs explications à cela telles que, l'injection des inhibiteurs de corrosion au niveau des pipelines, l'amélioration des actions de maintenance, l'addition d'une épaisseur de corrosion pour les nouveaux composants de tuyauterie, application de la protection cathodique et d'autre méthode de revêtement, le nettoyage périodique de l'intérieur des pipelines en utilisant les racleurs et le bon contrôle des paramètres opératoires ainsi que les caractéristiques chimiques des fluides.

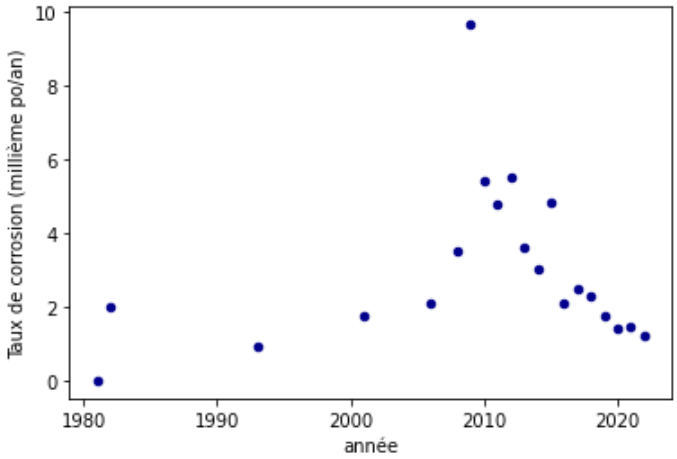


Figure 4.6 Variation des taux de corrosion durant les dernières années

- Variation des taux de corrosion par type de service

Le type de service a un impact direct sur la détérioration de la paroi intérieure des lignes de tuyauterie. Cette détérioration dépend des caractéristiques physico-chimiques de chaque fluide. Le Tableau 4.2 présente un extrait des taux de corrosion des différents services qui peuvent être présents dans les raffineries, on peut remarquer que le taux de corrosion varie selon le type de service.

Tableau 4.2 Variation des taux de corrosion par fluide

Service	**CR(mpy)**	**Service**	**CR(mpy)**	**Service**	**CR(mpy)**
FRACOVHD	3.97	GPL	2.74	CATFEED	1.93
CRUDOVHD	3.95	BRUTNAOH	2.58	EAUACIDE	1.92
GASOILEG	3.81	GASOILOU	2.15	BRUT	1.79
FLUEGCAT	3.20	KEROSENE	2.13	NAPHTA	1.98
LIGHTHC	2.77	C3C4H2S	2.08	CATALYST	1.19
ACIDGAS	2.74	H2H2S	2.07	UTILIH2O	1.17

Les résultats du Tableau 4.3 montrent qu'un service peut adopter différentes classes de code. Il semble que les fluides circulant à haute température appartiennent aux classes de code 1 selon API570, ils ont les taux de corrosion les plus élevés.

Tableau 4.3 Variation du taux de corrosion par état de fluide et température opératoire

Service	État du fluide	T opératoire	Classe code	CR (mpy)
CATFEED	LIQUID	150	2	2.17
			3	1.44
		200	2	5.48
			3	0.67
		250	2	4.91
		350	2	2.09
		400	2	2.37
		550	1	1.86
		600	1	1.85
		650	1	2.10
		700	1	2.13
GASOILEG (GASOIL léger)	LIQUID	500	1	1.02
		100	2	0.77
		150	2	1.09
		200	2	1.08
		300	2	4.90
		350	2	2.60
		550	2	3.01
		600	1	4.33
			2	2.35
	VAPEUR	600	1	4.39
H2H2S (Hydrogène/sulfure d'hydrogène)	VAPEUR	100	2	1.08
		150	1	2.15
			2	2.52
		200	1	1.35
		300	1	2.14

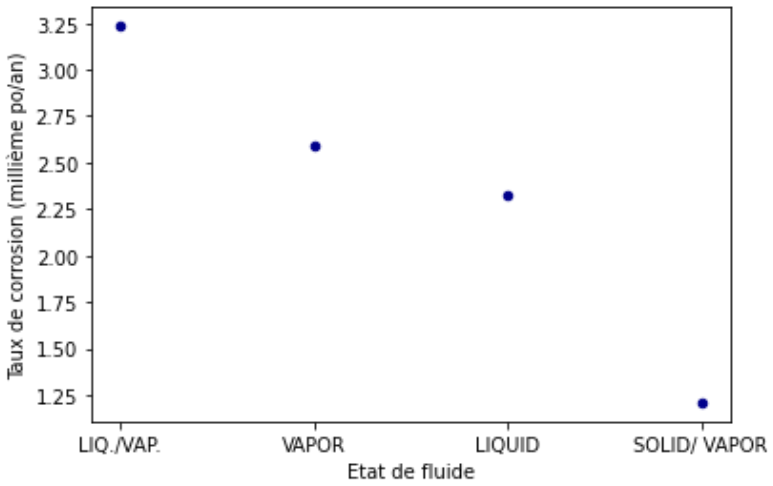
- Variation des taux de corrosion en fonction des paramètres opératoires

La température et la pression influent sur le nombre de phases (liquide, gaz, solide) qui, à leur tour, peuvent provoquer une corrosion considérable [72], Figure 4.7. La phase liquide vapeur est considérée comme la phase la plus corrosive. En effet, Pietro et al. [73] ont montré que les produits multiphase sont plus corrosifs à cause de la présence de plusieurs agents corrosive comme le dioxyde de Carbon CO_2 et le sulfure d'hydrogène H_2S qui a leur présence le taux de détérioration de la paroi interne des conduites augmente.

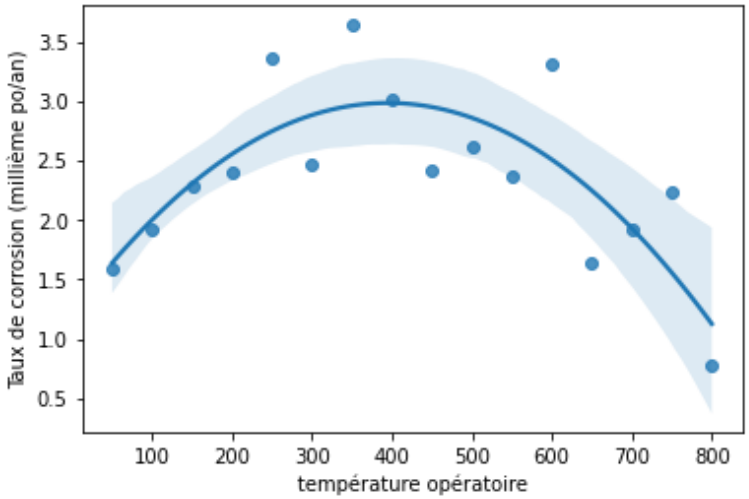
La Figure 4.7 montre que la température a un effet compliqué sur les taux de corrosion. En effet, l'augmentation de la température accélère la vitesse des réactions de corrosion électrochimique et le transfert des participants : les substances agressives vers la surface des métaux et les substances de corrosion de la surface vers l'environnement [72]. L'augmentation de la température entraîne une diminution du pH de l'eau ce qui implique que la vitesse de corrosion devrait augmenter avec l'augmentation de la température. D'un autre côté, la solubilité des gaz corrosifs (H_2S, O_2, CO_2) diminue avec l'augmentation de la température[72]. Jusqu'à une certaine température, la vitesse de corrosion augmente avec l'augmentation de la température. Après une certaine température, la vitesse de corrosion diminue en raison de la réduction de la solubilité des gaz corrosifs dans les solutions aqueuses. En plus de cela, la température augmente la vitesse de sédimentation et la formation d'un film protecteur de $FeCO_3$.

En ce qui concerne la pression, ce paramètre peut avoir un effet sur la vitesse de corrosion d'un matériau. Dans certains cas, une pression élevée peut augmenter la vitesse de corrosion en raison de l’effet des contraintes appliquées qui empêche la formation de la couche protectrice d'oxyde à la surface du matériau. Dans d'autres cas, une pression élevée peut réduire la vitesse de corrosion en diminuant la quantité d'oxygène et d'eau disponibles pour réagir avec le matériau [72].

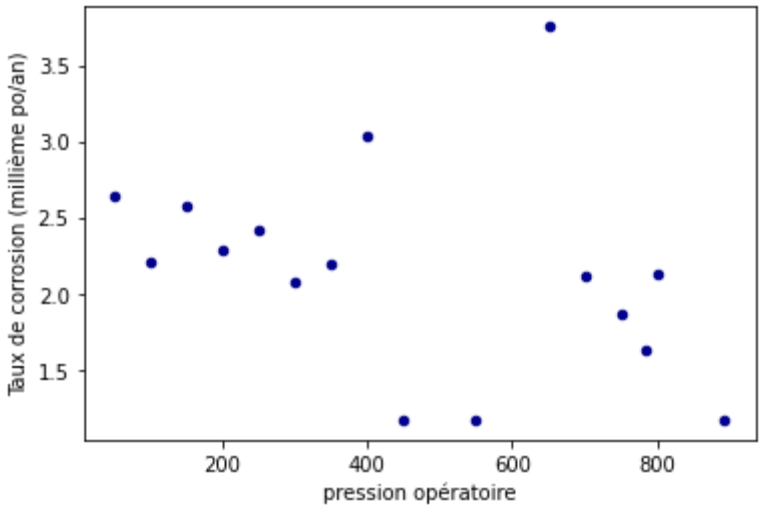
Il convient de noter que l'effet de la pression et la température sur la vitesse de corrosion sont complexes et dépend de nombreux facteurs, tels que le type de matériau, le type d'environnement (par exemple, le service corrosif, l'humidité) et la présence d'autres éléments qui peuvent contribuer à la corrosion (par exemple, des impuretés corrosives). Par conséquent, l'effet de paramètres opératoires sur la vitesse de corrosion ne peut être décrit comme une tendance générale, mais doit être évalué au cas par cas.



a-Variation du taux de corrosion en fonction de l'état du fluide



b-Variation du taux de corrosion en fonction de la température opératoire



c-Variation du taux de corrosion en fonction de la pression opératoire

Figure 4.7 Variation du taux de corrosion en fonction de l'état de fluide et paramètres opératoires

- Variation des taux de corrosion par type de matériaux

La sélection des matériaux marque l'utilisation de métaux et d'alliages, de matériaux polymères et composites adaptés à différents environnements dans la technologie du gaz naturel. Il n'existe pas de matériau idéal qui résiste à tous les milieux dans toutes les conditions. Il est acceptable de qualifier les aciers inoxydables et les alliages à base de nickel de matériaux résistant à la corrosion. Pour notre mandat, les résultats obtenus, Figure 4.8, montrent que les matériaux en acier carbone représentent plus de 90% de tous les matériaux utilisés et possèdent le taux de corrosion le plus élevé avec un taux moyen de 2.7 mpy. Le taux de corrosion moyenne pour l'acier inoxydable et les différents alliages ne dépassent pas 2 mpy.

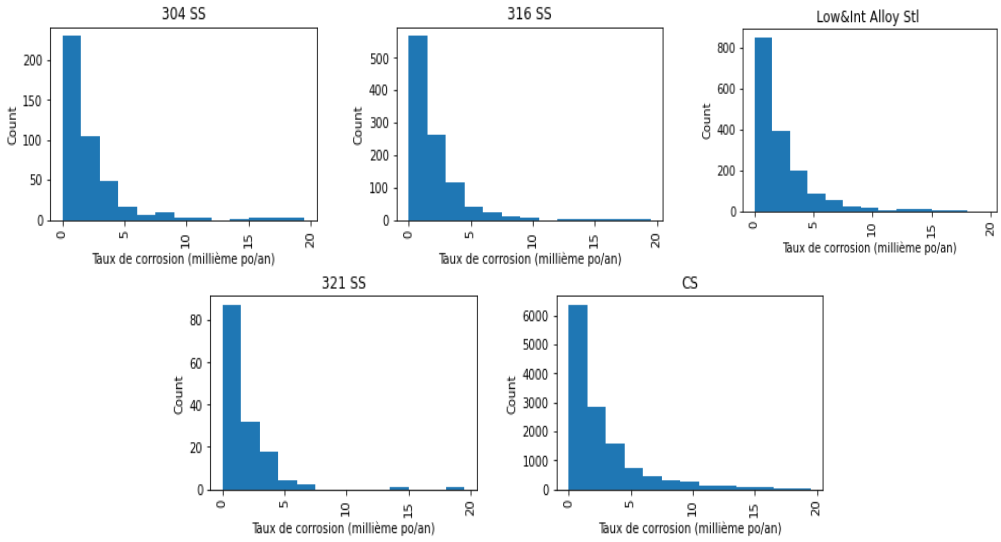


Figure 4.8 Variation des taux de corrosion en fonction du type de matériau

- Variation des taux de corrosion par type de circuit

Les MPF et les DL dominent la majorité des circuits pour cette raffinerie, on rappelle que les DL sont des composants d'un système de tuyauterie qui n'ont pas un débit significatif (point de drainage, des évents de point haut, les lignes de bypass.). Contrairement au MPF, le fluide reste stagnant au niveau des DL et peut provoquer une corrosion accélérée.

La Figure 4.9 montre que la distribution du taux de corrosion est similaire pour le DL et MPF avec un taux de corrosion moyen similaire de 2.5 mpy. Les MX et les IP sont des endroits fréquemment utilisés dans les raffineries, ils présentent un risque potentiel d'augmentation du taux de dégradation par rapport à la conduite

principale en raison des changements de température, de pH, des changements de phase et de la concentration d'espèces corrosive. Le taux de corrosion moyen pour les MX est 5.2 mpy et 2.7 mpy pour les IP. Afin de préserver les IP et MX de toute défaillance surprenante, une attention minutieuse doit être accordée, lors de l'ingénierie, à l'application des configurations de conception appropriées, aux matériaux de construction, aux paramètres opératoires (débit, pression, température, ph..), aux types et aux positions de l'emplacement sur la conduite. En plus, il faut préparer des plans d'inspection appropriés pour ces types de circuits.

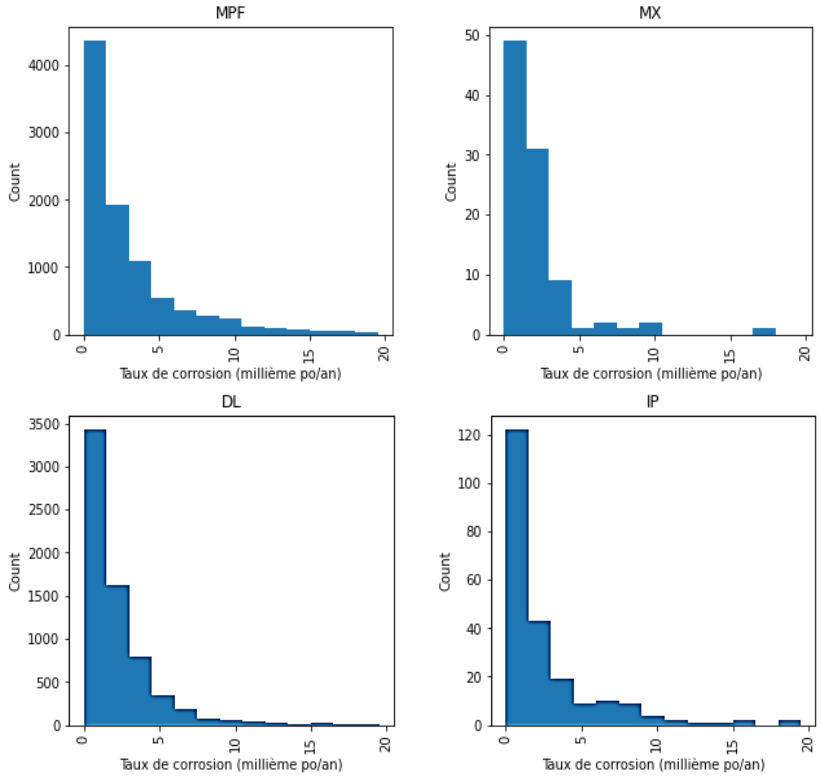


Figure 4.9 Variation des taux de corrosion en fonction du type de circuit

- Variation des taux de corrosion par géométrie de l'élément

➢ Diamètre de la conduite

Il semble y avoir un consensus total dans la littérature sur le fait que le plus grand nombre de défaillances est observé dans les tuyaux de petits diamètres [74], cela peut être expliqué par la résistance réduite de tuyau ainsi que l'épaisseur réduit des parois. Pour notre cas, les conduites de grand diamètre ont le taux de corrosion le plus élevé, en effet, les analyses montrent que les fluides CRUDOVHD et FRACOVHD sont transportés par des conduites de grands diamètres et la majorité

des fluides circulants sont de type classe 1 et 2. Alors que les tuyaux de petits diamètres sont dédiés à transporter les différentes classes de service.

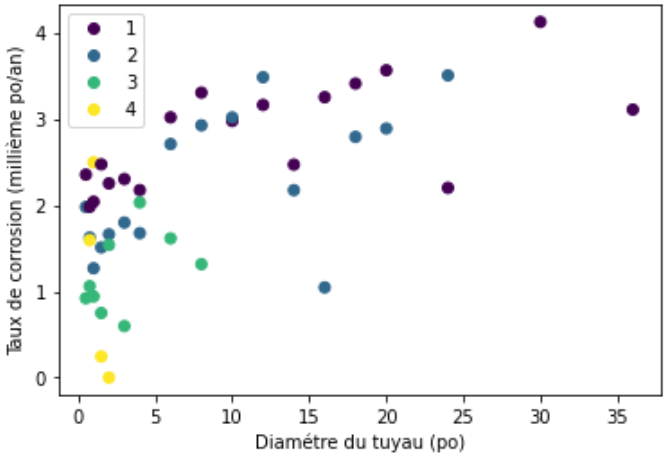
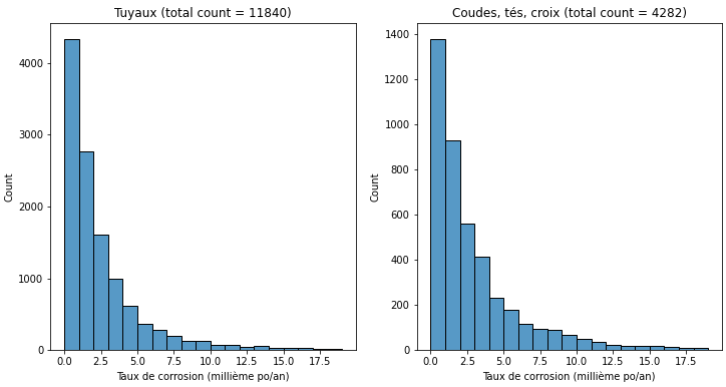


Figure 4.10 Variation des taux de corrosion par diamètre et classe de service

- Type de l'élément

Le taux de corrosion dépend nécessairement de la géométrie de l'élément de tuyauterie. En effet, la géométrie influe directement sur le régime d'écoulement du fluide, ce dernier affecte la contrainte de cisaillement exercée par le fluide sur la paroi du tuyau. Les coudes et les Tés sont les éléments qui dominent les taux de corrosion la plus élevée, d'autre part au niveau de ces zones le risque de défaillances en rapport avec la fuite de fluide croît considérablement.

La Figure 4.11 montre que la distribution du taux de corrosion est similaire pour les éléments droits et non-droits (Coude, Réduction, croix.).



Élément	**CR moyen (mpy)**
Tuyau droit	2.5
Coude	3.2
Réduction	2
Tee	3.1
Cap	3.9

Figure 4.11 Variation des taux de corrosion en fonction de la géométrie

- Variation des taux de corrosion en fonction de la source de corrosion

Les résultats montrent qu'il existe de nombreuses sources de corrosion dans les opérations de raffinage, pour distinguer entre les différentes sources, il faut bien comprendre les propriétés physico-chimiques de chaque fluide. Le Tableau 4.4 montre que le HCl et le H_2S sont les produits chimiques les plus destructeurs dans cette raffinerie, cela a été approuvé par les travaux de littérature. Le taux de corrosion moyen est de 4.37 mpy pour la corrosion par HCl. D'un autre côté, la corrosion par H2S à haute température est plus sévère que la corrosion par H_2S humide. Selon la NACE [72], l'indice d'acidité totale (TAN), la teneur en soufre total, l'eau, la teneur en sel et les micro-organismes sont autant de facteurs qui influencent la corrosivité. En effet, la combinaison de ces paramètres contribue à corroder les équipements métalliques à divers endroits de la raffinerie (corrosion locale, généralisée.), entraînant divers types de corrosion comme indiqué au niveau du Tableau 4.4.

Tableau 4.4 Taux de corrosion en fonction de la source de défaillance

Source de corrosion	Service	CR (mpy)
Corrosion par acide chlorhydrique (HCl)	CRUDE OVHD	4.37
Corrosion à haute température par sulfure d'hydrogène (H_2S)	Kérosène	3.8
Corrosion par sulfure d'hydrogène (H_2S) humide	ACIDGAS hydrogène/sulfure d'hydrogène FRACOVHD LCN (naphta catalytique léger) GPL (gaz de pétrole liquéfié)	3.06
Corrosion par chlorure d'ammonium	Naphta catalytique lourd	2.65
Corrosion caustique et SCC	BRUTNAOH	2.62
Érosion	Slurry Gaz de cheminée	2.36
Corrosion par l'eau d'acide	Reflux Naphta Kérosène Brut	2.26
Corrosion sous contrainte de chlorure, acide polythionique et érosion	Slurry	2.25

Corrosion à haute température par sulfure d'hydrogène et corrosion par acide naphténique	GASOIL BRUTNAOH CATFEED	1.8
Corrosion sous isolation	ACIDGAS	1.45

- Application de la RBI

En ce qui concerne l'application de la RBI, en se basant sur les taux de corrosion calculés, on a choisi le service CRUDEOVHD pour déterminer le niveau de risque approprié pour chaque Tml. Cette méthode prend en compte les résultats d'inspection tels que le taux de corrosion ainsi que l'efficacité des méthodes d'inspection.

On a commencé par le calcul du Pof pour chaque Tml suivi par l'identification des Tml qui semble critique pour chaque circuit de tuyauterie. À partir de la valeur du Pof limite proposée par le standard DNV on a prédit les dates cibles des prochaines inspections. Le Pof limite dépend du type de la défaillance, pour notre cas le Pof_{limite} est fixé 10^{-4} pour les différents circuits. On a classifié la probabilité de défaillance selon les intervalles proposés par l'API581.

La Figure 4.12 présente un extrait des résultats obtenus pour les Tml qui appartient au même circuit de tuyauterie, on remarque la grande variabilité au niveau des dates cibles pour les prochaines inspections (Target date) pour les différents Tml même s'ils appartiennent aux mêmes circuits, cela approuve que dans certaines circonstances, la méthode puisse atteindre des divergences dans l'évaluation de la prochaine inspection qui peuvent être induites par des jugements subjectifs d'un membre de l'équipe de maintenance.

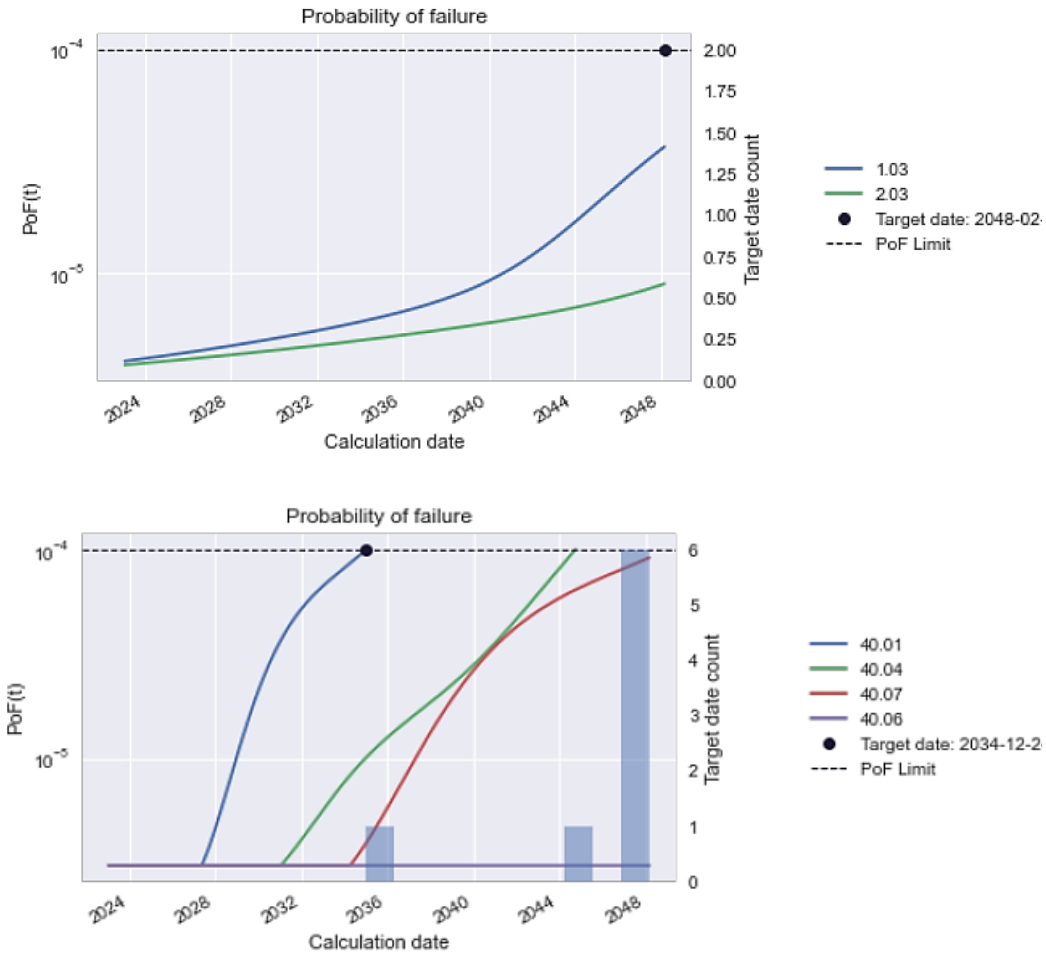


Figure 4.12 Résultats de la RBI

La méthodologie de la RBI comprend un processus de classement pour le Pof et le Cof des réseaux de tuyauterie. En effet, le programme d'inspection, Tableau 4.5, ne peut influer que sur la valeur de la probabilité d'une défaillance et non sur la conséquence. Quelle que soit la fréquence d'inspection effectuée, la conséquence reste inchangée. Pour notre cas, on a relié le niveau de conséquence Cof à la classe de service, pour le CRUDOVHD c'est un service de classe 1, on lui a associé un Cof de niveau VI, c'est-à-dire ce sont des circuits où le niveau conséquence est élevé. Finalement, avec le Pof et la valeur du Cof le niveau de risque est déterminé en se basant sur la matrice de risque fournie (Figure 4.13).

Figure 4.13 Matrice de risque de la raffinerie

CoF \ Likelihood	RARE	UNLIKELY	POSSIBLE	LIKELY	ALMOST_CERTAIN
IV	LOW	MODERATE	HIGH	CRITICAL	EXTREME
III	NONE	LOW	MODERATE	HIGH	CRITICAL
II	NONE	NONE	LOW	MODERATE	MODERATE
I	NONE	NONE	NONE	NONE	NONE

Tableau 4.5 Résultats de la RBI

Type de circuit	Tml	Dernière inspection	Prochaine inspection	Pof	Catégorie de Probabilité	Niveau de risque
DL01.0-DVI	800.03	2021-04-13	2043-06-27	3.06 E-06	RARE	Faible
DL01.0-DVI	805.03	2018-03-06	2030-12-27	3.06 E-06	RARE	Faible
DL01.0-DVI	808.03	2018-03-02	2024-01-26	3.03 E-05	RARE	Faible
DL03.0-CAP	40.01	2020-02-13	2034-12-26	3.06 E-06	RARE	Faible
DL03.0-CAP	40.03	2020-02-13	2048-02-25	3.06 E-06	RARE	Faible
DL03.0-CAP	40.04	2020-02-13	2044-09-25	3.06 E-06	RARE	Faible
MPF	29.03	2021-11-16	2035-03-28	5.68 E-06	RARE	Faible
MPF	30.03	2019-07-25	2044-09-25	5.86 E-06	RARE	Faible
MPF	45.04	2019-07-25	2027-01-26	3.19 E-06	RARE	Faible

4.1.2 Étude de cas n◦ 2 : analyse des données d'une usine d'acide sulfurique

Cette section présente les premiers résultats obtenus après l'étude de l'état des lignes de tuyauterie d'acide sulfurique appartenant à l'usine CEZinc de la compagnie Glencore. Cette section va comporter trois paragraphes. Premièrement, une analyse des rapports visuels effectués dans les différentes zones de l'usine suivit par l'analyse du taux de corrosion des différentes fiches d'inspection et enfin la dernière section est consacrée à présenter les critères de choix de l'élément à inspecter. L'objectif principal à travers cette étude étant de relever les états critiques menant à des fuites d'acide et aussi d'identifier les endroits optimaux (TmL) pour installer des instruments ultrasoniques permanents permettant de contrôler l'amincissement d'épaisseur provoqué par l'acide sulfurique.

4.1.2.1 Collecte et préparation des données

Autres que les relevés de mesure d'épaisseur qui sont faits généralement chaque 18 mois, des rapports d'inspections visuelles sont produits régulièrement pour contrôler s'il y a des défauts qui peuvent apparaitre (autre que la corrosion) afin de

protéger le système de toutes anomalies. Ces rapports contiennent différentes données telles que la zone de l'usine, le numéro de fiche, les paramètres de procédés : Concentration C, Température T, etc. On y trouve aussi les origines des défauts s'ils existent et quelques commentaires (valeur de l'épaisseur si elle est proche de la valeur minimale, manque d'isolation…).

Au total, 273 inspections ont été faites dernièrement dans les différents endroits de l'usine.

La Figure 4.14 et le Tableau 4.6 montrent que la distribution des inspections est inégale entre les différentes zones. En effet, les zones de production (UA1, UA2, UA3) et l'entreposage sont les plus inspectés dont on va étudier à la suite l'état de conformité de chaque zone.

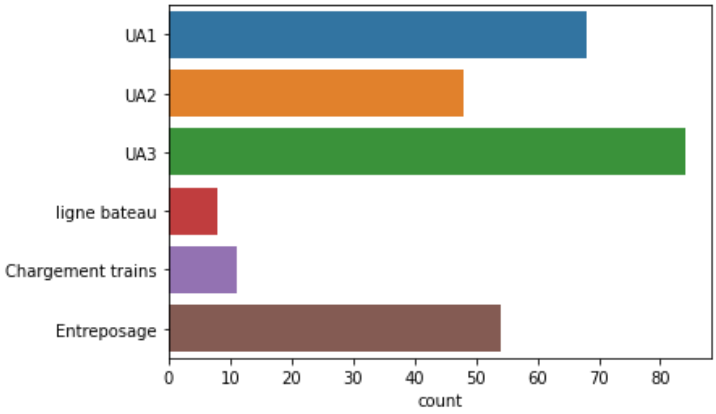


Figure 4.14 Distribution des rapports par zone

Tableau 4.6 Nombre d'inspections par zone

Zone	**UA1**	**UA2**	**UA3**	**Entreposage**	**Ligne bateau**	**Chargement train**	**Total**
Nbre d'inspection	68	48	84	54	8	11	**273**

4.1.2.2 Analyse des données

- Analyse des rapports d'inspection visuelles

Comme le nombre d'inspections n'est pas le même, on doit se référer aux pourcentages d'état de conformité de chaque zone. Ce pourcentage est égal au rapport de nombre des conformités et non-conformités par le nombre total des fiches par zone. D'après le Tableau 4.7

Zone	% conformité	% non-conformité	% inspection repoussée	% hors service
UA3	21.43	**78.57**	0	0
UA1	25	**54.41**	17.65	2.94
Entreposage	59.26	40.74	0	0
UA2	79.17	20.83	0	0
Chargement train	81.82	18.18	0	0
Ligne bateau	100	0	0	0

, on peut noter que les zones UA1 et UA3 ont le taux de non-conformité le plus élevé. Les inspections repoussées concernent les fiches dont la mesure d'épaisseur n'est pas faite et elle doit être faite prochainement.

Tableau 4.7 Distribution des résultats d'inspection par zone

Zone	% conformité	% non-conformité	% inspection repoussée	% hors service
UA3	21.43	**78.57**	0	0
UA1	25	**54.41**	17.65	2.94
Entreposage	59.26	40.74	0	0
UA2	79.17	20.83	0	0
Chargement train	81.82	18.18	0	0
Ligne bateau	100	0	0	0

- Distribution des résultats d'inspection par matériau

Les matériaux des lignes pour l'usine d'acide sulfurique sont généralement la Fonte Mondi, l'acier inoxydable, les alliages (Alloy SX, Alloy 20). Le Tableau 4.8 montre que la fonte Mondi domine davantage le taux de non-conformité.

Tableau 4.8 Distribution des résultats d'inspection par matériau

Matériau	Nombre d'éléments inspectés	% conformité	% non-conformité
Saramet 35	4	75	25
Alloy SX	16	75	25
Stainless Steel	74	60	40
Fonte	131	25	**75**

- Distribution des résultats d'inspection selon les paramètres de procédé

Les paramètres de procédé tels que : la température, la concentration et le type d'écoulement du fluide circulant influent sur le taux de corrosion ainsi que la propagation des fuites au niveau des brides. Il est nécessaire de déterminer la dépendance des défaillances (non-conformité) et ces paramètres. En effet, pour CEZinc, la concentration de l'acide est de 93% à 98%, la plage de température de l'acide est de l'ambiante à 100 °C et l'écoulement est soit sous pression ou bien gravitaire, or d'après les fiches d'inspection et les PFD on peut noter que l'écoulement est sous pression lorsqu'il est à côté des équipements centrifuges tels que les pompes alors qu'il est gravitaire lorsque le fluide circule de haut vers le bas ou bien l'existence d'une pente. D'après les résultats (Figure 4.15, Figure 4.16, Figure 4.17) obtenus de l'analyse des données, on a remarqué qu'il y a un manque de mesures de température et de concentration dans plusieurs zones. De plus, il n'y a pas de tendance claire pour déterminer la valeur dominante correspondant à la non-conformité.

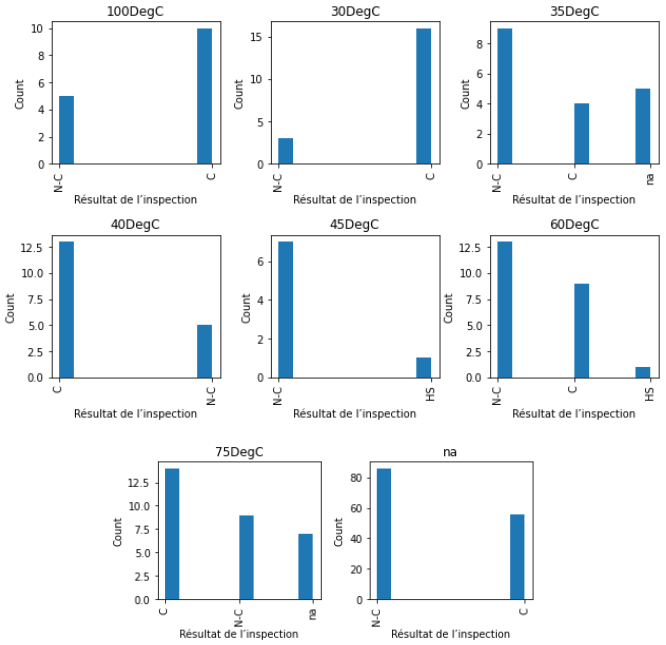


Figure 4.15 Distribution des résultats d'inspections par température

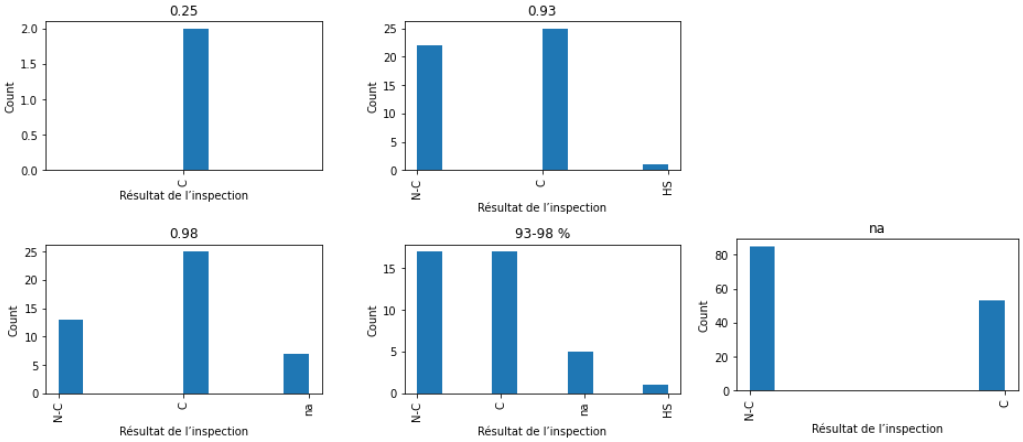


Figure 4.16 Distribution des résultats d'inspections par concentration

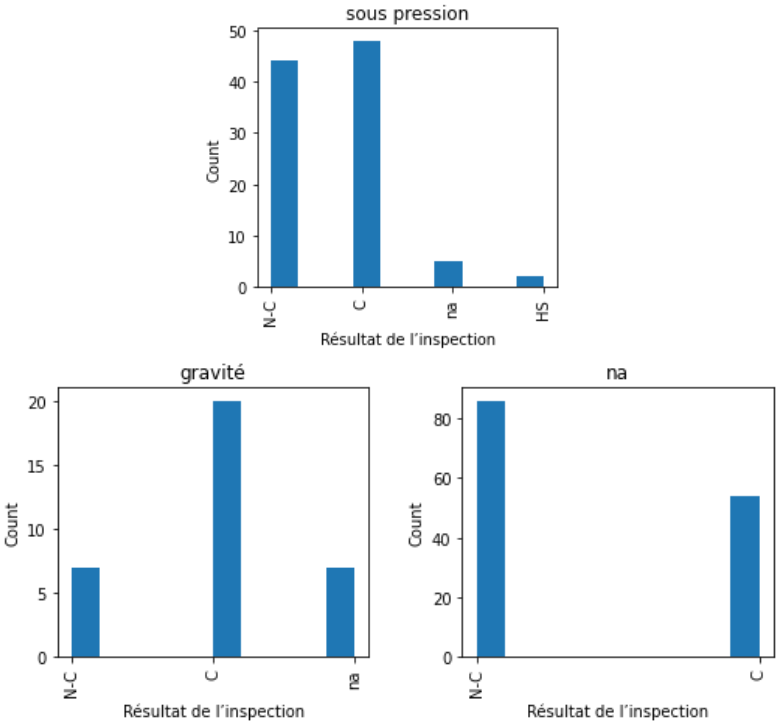


Figure 4.17 Distribution des résultats d'inspections par type d'écoulement

- Distribution des résultats d'inspection par géométrie d'élément

La géométrie des éléments du système de tuyauterie est l'un des paramètres qui doivent être analysés pour faire la prédiction des fuites dans les systèmes de tuyauterie. Généralement, les coudes, les Tés, les réductions et les brides sont les principaux éléments à étudier parce qu'ils sont trop sollicités par des coups de bélier, couples de serrage des brides, etc.

Le Tableau 4.9 montre que les coudes ont le taux de non-conformité le plus élevé.

Tableau 4.9 Distribution des résultats d'inspection par géométrie d'élément

Élément	Nombre d'éléments inspectés	% conformité	% non-conformité
Soupape de contrôle	1	0	100
Té Égale	10	0	100
Té Réduite	9	0	100
Vanne	8	0	100
Bride d'obturation	3	0	100
Croix	1	0	100
Coude 30	3	0	100
Coude 90	38	5.26	94.74
Tuyau	60	10	90
Réduction	6	16.67	83.33
Coude 45	4	25	75

Les petits diamètres 3''et 6'' dominent les non-conformités.

Tableau 4.10 Distribution des résultats d'inspection par diamètre d'élément

Diamètre (Pouce)	Nombre d'éléments Inspectés	% conformité	% non-conformité
3	36	8.33	91.67
6	43	8.51	91.49
8	6	0	100
12	15	0	100
14	8	0	100

La Figure 4.18 présente une étude comparative entre les sections droites (tuyau) et les autres éléments afin de savoir si la distribution de la non-conformité dépend de la nature de l'élément. En effet, on peut noter que l'état de conformité entre les tuyaux et les coudes/Tés/croix est similaire.

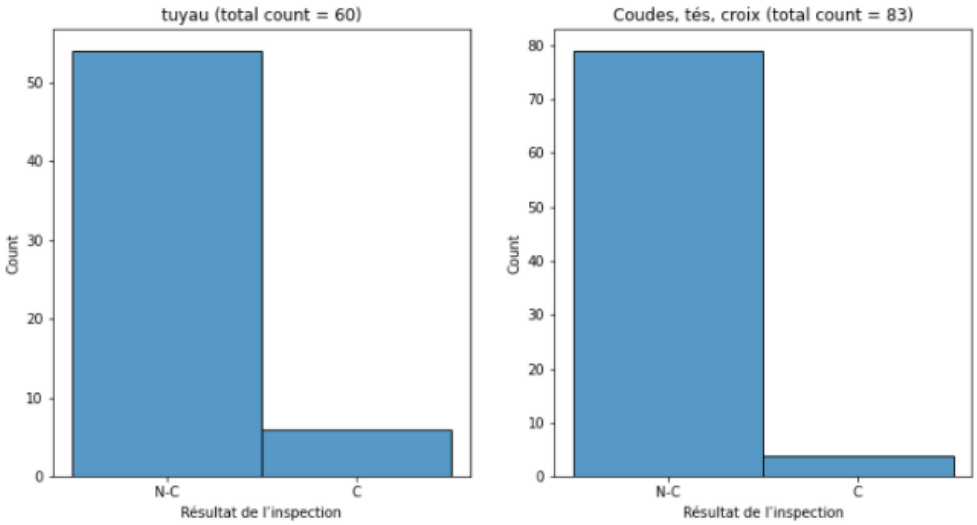


Figure 4.18 Comparaison entre les tuyaux et les Raccords

- Origines de la non-conformité

Pour entamer notre analyse, il est nécessaire de savoir les origines des fuites, d'après les fiches d'inspection visuelles, les principales causes des fuites sont : L'accumulation de sulfate humide, l'amincissement de l'épaisseur, la manque d'isolation et l'existence de support brisé.

- Analyse des taux de corrosion

D'après l'analyse effectuée avec les deux logiciels Python et Excel, on a pu remarquer qu'il y a des taux de corrosion négatifs pour certains éléments (Figure 4.22). Les raisons qui expliquent ces valeurs négatives sont quasiment les mêmes présentées précédemment. Les résultats obtenus ont affirmé que l'incertitude de mesure pour les lignes d'acide 98% sont moins que les lignes 93% puisque les lignes 98% sont isolées.

Le Tableau 4.11, montre que l'acide 93% sont plus corrosif.

Tableau 4.11 Variation du taux de corrosion entre 2015 et 2018

	2015	**2017**	**2018**
Acide 93%	16 mpy	11 mpy	33 mpy
Acide 98%	16 mpy	14 mpy	22 mpy

L'analyse des fiches d'inspection a montré que les zones UA1, UA3 et l'hydrométallurgie sont les zones qui ont les taux de corrosion les plus élevés (Figure 4.19).

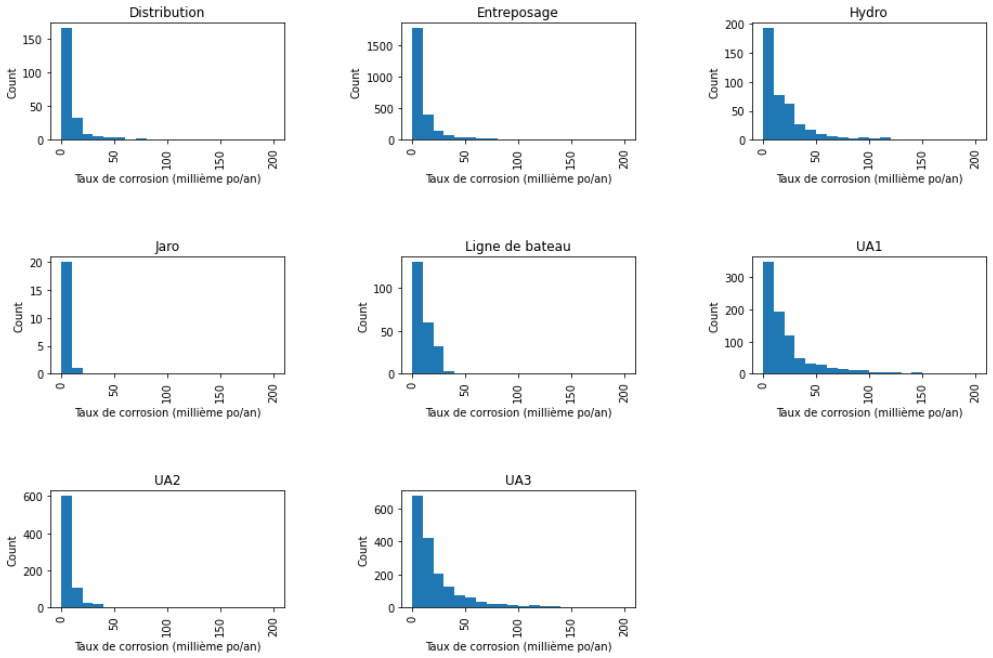


Figure 4.19 Distribution du taux de corrosion par zone

Les Figure 4.20 et Figure 4.21 montrent que la distribution du taux de corrosion moyenne de ces zones pour les matériaux de fonte Mondi et l'acier inoxydable des différents éléments est similaire. En effet, cette vitesse est à l'ordre de 20 mpy pour la fonte, alors qu'il ne dépasse pas 10 mpy pour l'acier inoxydable. De plus les coudes et les Tés dominent le taux de corrosion le plus élevé.

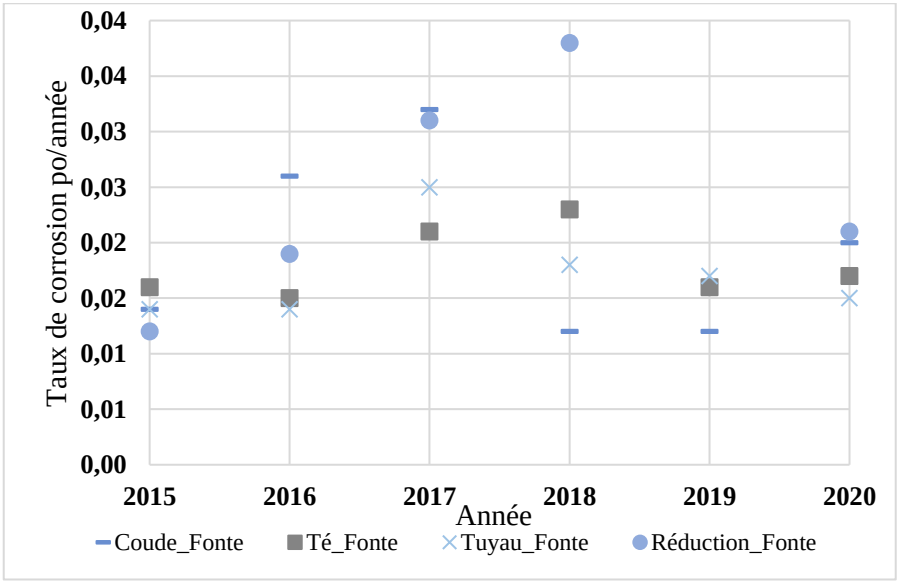


Figure 4.20 Variation du taux de corrosion pour la fonte Mondi

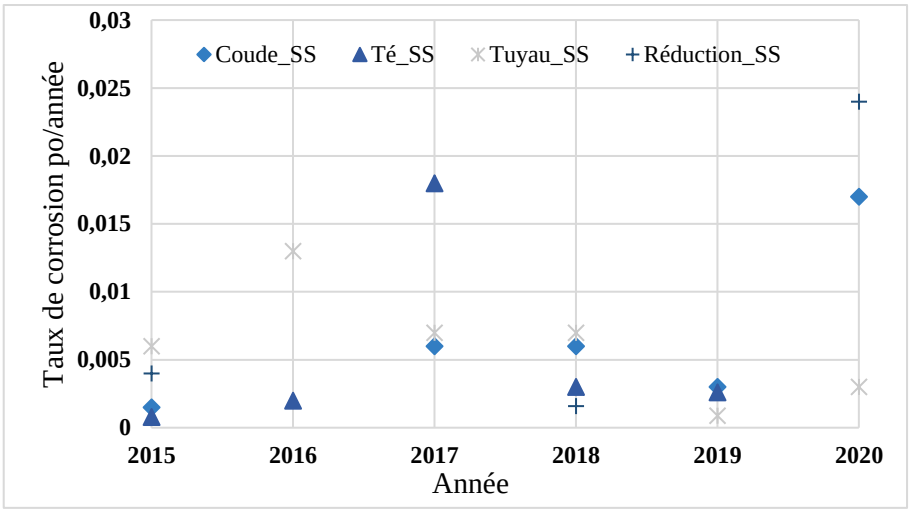


Figure 4.21 Variation du taux de corrosion pour l'acier inoxydable

4.1.2.3 Choix du composant

Une décision importante dans la mise en place d'un système de surveillance est la détermination des points de surveillance où doivent être placés les capteurs. Comme un nombre limité de points peut être envisagé pour des raisons économiques évidentes, il est généralement désirable de surveiller les conditions les plus défavorables, aux points où l'on s'attend à ce que les dommages soient les plus sévères.

Selon les analyses des fiches d'inspection et des rapports visuels, on a choisi à faire installer les capteurs UT sur les deux premiers coudes à 90 situés au niveau du tronçon de tuyauterie situé à la zone UA1. La Figure 4.23 présente la disposition de la ligne à étudier. En effet, ce tronçon est situé entre deux équipements qui sont la pompe PA 009 et le refroidisseur circuit d'absorption 23-104. Le Tableau 4.12 présente les différents éléments du tronçon choisi :

Tableau 4.12 les composants de la ligne d'étude

Item	Description	Dimensions	Matériaux
1	Coude réduit 12" dia.x 10" dia. R.L.		FONTE
2	Coude 12" dia.x 90 R.C.		FONTE
3	Coude 12" dia.x 90 R.C.		FONTE
4	Robinet à papillon 12" dia., Garlock		/
5	Tuyau 12" dia.	3'-7"	FONTE
6	Tuyau 12" dia.	2'-0 9/16"	FONTE
7	Tuyau 12" dia.	2'-3 3/8"	FONTE
8	Tuyau 12" dia.	16'-11 15/16"	FONTE
9	Té 12" x12"x10" dia.		FONTE
10	Tuyau 12" dia.	3'-1 1/4"	FONTE
11	Réduit 18" x12" dia.		FONTE
12	Tuyau 10" dia.	16 11/16"	FONTE
13	Coude 10" dia.x 90 R.C.		FONTE
14	Tuyau 10" dia.	14'-3 3/16"	FONTE
15	Coude 10" dia.x 90 R.C.		FONTE
16	Tuyau 10" dia.	4'-9"	FONTE
17	Coude 10" dia.x 90 R.C.		FONTE

La Figure 4.22 présente la variation du taux de corrosion de cette ligne entre les années 2011 et 2018. Le taux de corrosion moyen de cette ligne est 20 mpy.

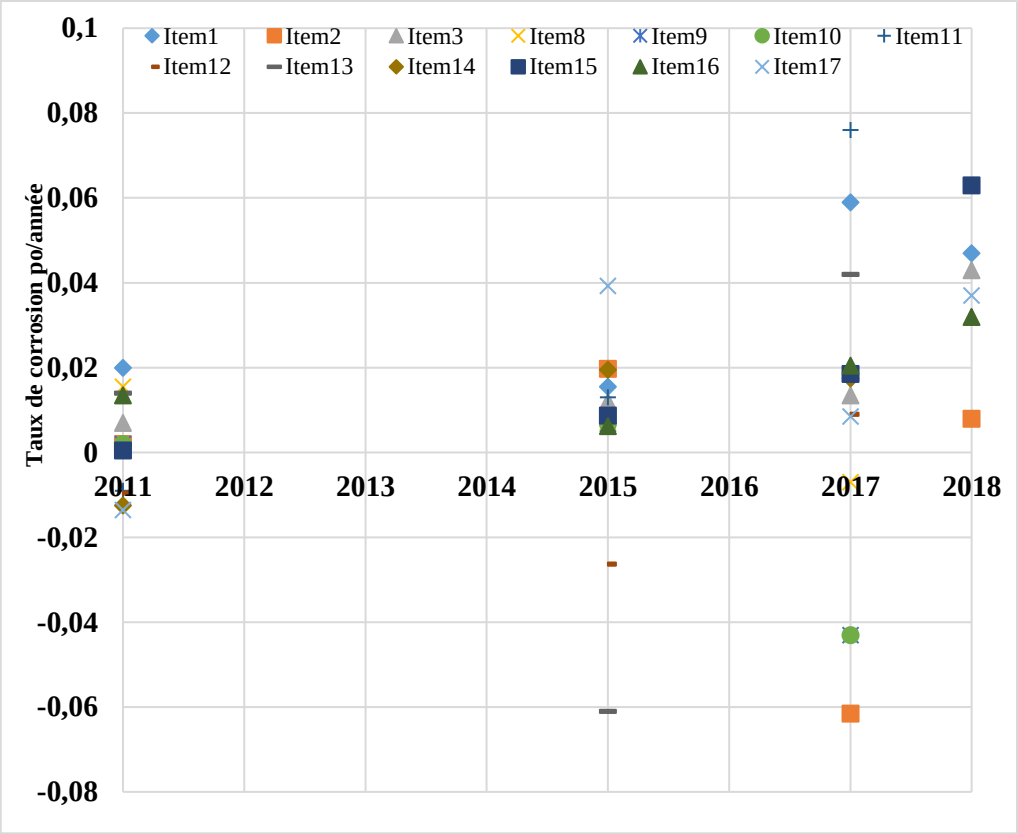


Figure 4.22 Variation du taux de corrosion de la ligne 27-139-A-1001

On a choisi ce tronçon en se basant sur les résultats obtenus et aussi vu l'accessibilité au niveau chantier pour installer les instruments pour les prochains travaux. D'autre part, dans la revue de littérature [75,76], il a été démontré que le coude 90 est une partie importante de la configuration des tuyaux dans les systèmes industriels. Des changements brusques dans la configuration de l'écoulement (direction et vitesse de l'écoulement) se produisent dans un coude 90, ce qui entraîne une différence significative dans le comportement de corrosion à différents endroits. En raison du changement soudain de la configuration de l'écoulement, la réduction de l'épaisseur de la paroi par le phénomène de la corrosion accélérée par l'écoulement (FAC) sera exacerbée au niveau du coude [76], en particulier par l'érosion-corrosion. Le comportement de la corrosion à différents endroits du coude devrait apparemment être corrélé au modèle d'écoulement. Cependant, il n'y a que quelques travaux pour étudier les différents comportements de corrosion à différents endroits du coude. Pour ces raisons, on a choisi d'installer 2 capteurs sur chacun des deux coudes.

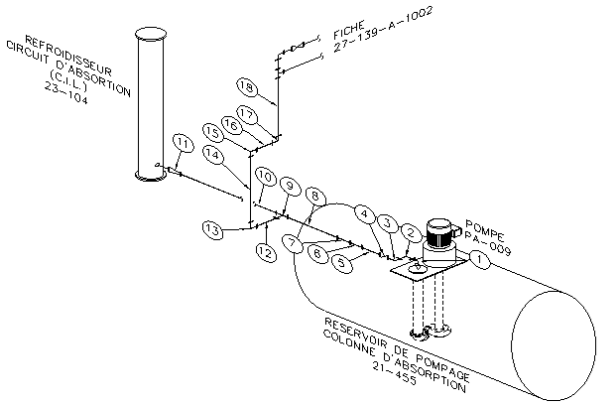


Figure 4.23 Position du tronçon 27-139-A-1001

4.1.3 Étude de cas n◦ 3 : installation des capteurs ultrasoniques

Après avoir choisi les TmL au niveau du paragraphe précédent, 4 capteurs ont été installés sur les deux coudes à différents endroits selon la complexité de l'élément, comme montre la Figure 4.24. On rappelle que les deux coudes sont en fonte ductile, l'acide sulfurique circule sous pression à température opératoire de 75°C, sa concentration est considérée à 98%, et le débit volumique moyen est 520 m^3/h.

4.1.3.1 Collecte et préparation des données des capteurs

La BD récoltée est basée sur des données réelles provenant d'un système de contrôle d'épaisseur UT et des instruments de mesure des paramètres opératoires. Les capteurs UT utilisés dans cette expérience ont été appliqués pour mesurer une gamme d'épaisseur entre 1 mm et 150 mm et ils ont été utilisés pour une gamme de température de [-30 °C, -132 °C]. Les capteurs ont été étalonnés en présence d'un inspecteur à l'aide d'une jauge d'étalonnage en fonte ductile. Un agent de couplage est utilisé pour la connexion entre le capteur et la surface du tuyau. Un transducteur de contact à double élément a été utilisé, avec une fréquence de 5 MHz. Un acquisiteur de données Smart (PIMS) avec batterie et mémoire intégrées, capable de stocker jusqu'à 3000 lectures d'épaisseur, a été utilisé pour le stockage des données. Il enregistre les mesures à chaque intervalle de temps spécifié par l'utilisateur.

a-Phase de calibrage b-Installation des capteurs

Figure 4.24 Dispositif expérimental pour le système de surveillance de l'épaisseur par UT

Les données des paramètres opératoires ; température, débit et concentration de l'acide sulfurique ont été extraits à partir des instruments installés au voisinage des coudes. Le non-contrôle de ces paramètres peut entraîner une détérioration inattendue des composants du système de tuyauterie, comme la rupture fragile, la corrosion et les fuites dans les équipements de tuyauterie (bride, vannes, joint, etc.).

La BD finale obtenue contient ; le numéro du capteur, la date d'acquisition, Épaisseur (mm), Température du fluide (°C), Concentration du fluide (%), débit du fluide (m^3/h). Le tableau ci-dessous présente un extrait de la base.

Tableau 4.13 Base de données à étudier

Date	Numéro	Concentration	Débit	Température	Épaisseur
2022-05-05 14:00:00	1	98.75	520	75	16.1192
2022-05-05 14:00:00	2	98.75	520	75	12.5878
2022-05-05 14:00:00	4	98.75	520	75	12.7303
2022-05-05 14:00:00	5	98.75	520	75	19.5563

4.1.3.2 Analyse des données

La surveillance de l'épaisseur dans ce projet utilisera toutes les mesures d'épaisseur actives pour effectuer les calculs d'analyse de la corrosion. La Figure 4.25, présente la variation d'épaisseur durant toute la période d'étude pour les différents capteurs installés, comme indiqué, les épaisseurs mesurées varient de 12 mm à 20 mm, l'épaisseur critique est définie t_{cr} = 7.9 mm, ce qui confirme que le système doit être

surveillé de près pour éviter toute défaillance, en particulier lorsque l'épaisseur semble être proche de la valeur critique. D'autre part, il est clair que la tendance de la mesure d'épaisseur provenant des différentes sondes montre un comportement similaire en regard de l'évolution de la série chronologique.

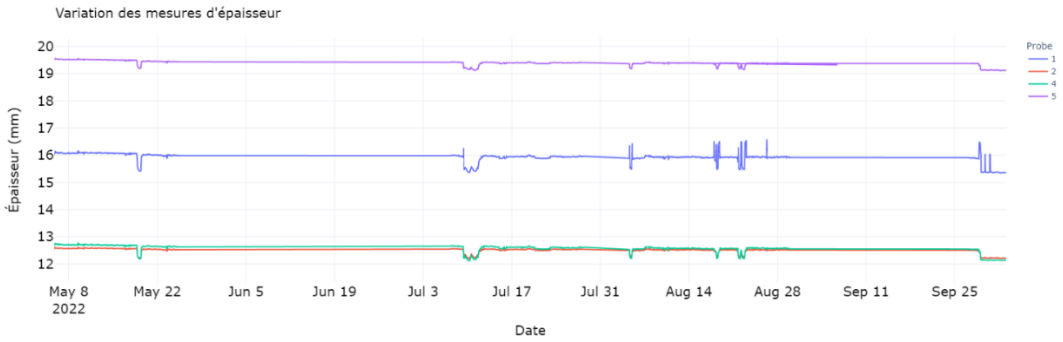


Figure 4.25 Variation des épaisseurs mesurées

Le système de monitoring a vécu plusieurs diverses formes d'anomalies durant la période d'étude telle que : l'inaptitude de stockage des données pour les périodes : 25 mai à 7 juillet et de 30 août à 27 septembre et aussi les sondes ont été affectées par toute perturbation au niveau des caractéristiques opératoires de l'acide (période de 18-19mai, 9-11juillet). Les outils de détection d'anomalie seront exploités au niveau du chapitre 4.

Pour le reste du travail, on s'intéresse à étudier la période de 6 mai à 25 mai pour le capteur 5, dont la Figure 4.26 présente l'évolution de l'épaisseur pendant cette période.

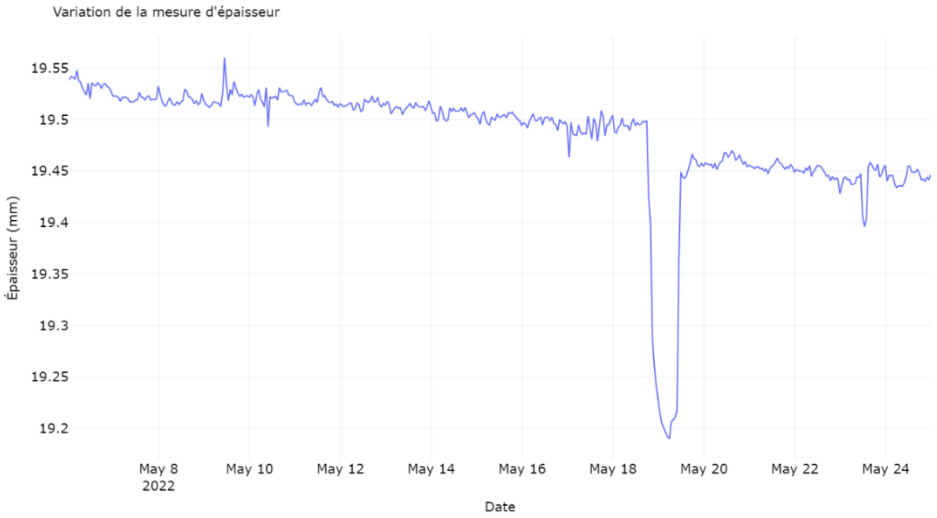


Figure 4.26 Variation des données d'épaisseur mesurées par le capteur UT

Pour étudier l'effet de corrélation entre les paramètres opératoires et les mesures d'épaisseur, nous avons utilisé la bibliothèque Python Seaborn pour générer et visualiser la matrice de corrélation, Figure 4.27. Les coefficients de corrélation entre les différentes variables collectées sont déterminés en utilisant la corrélation de rang de Spearman où -1 signifie que les variables corrélées ont une forte corrélation négative et 1 indique une forte corrélation positive. Les valeurs diagonales dénotent la dépendance d'une variable par rapport à elle-même (également appelée autocorrélation). La Figure 4.27 montre que la température et le débit sont fortement corrélés avec les mesures d'épaisseur, alors que la concentration du fluide est moins significative.

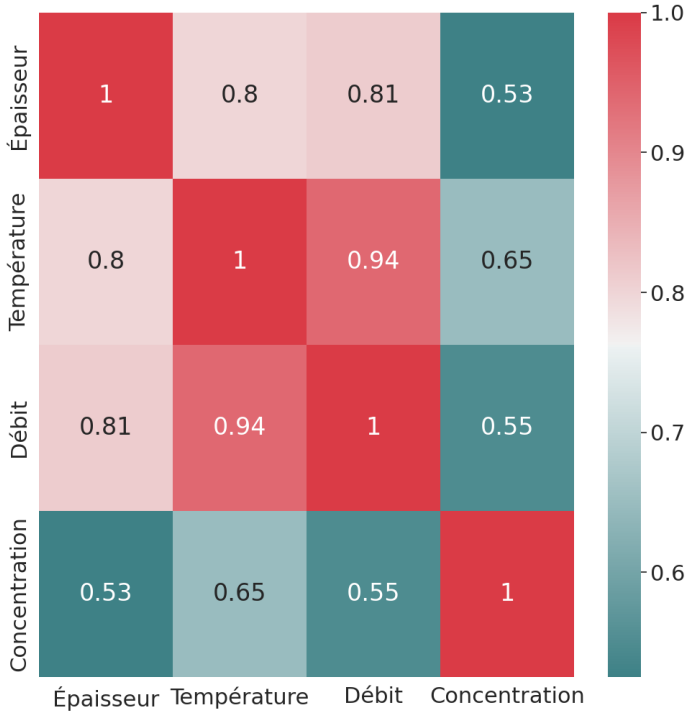
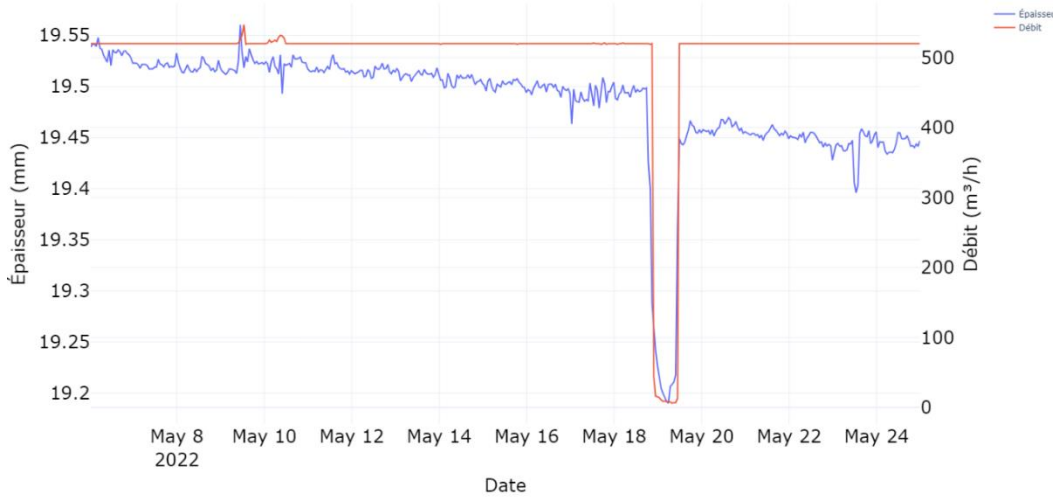


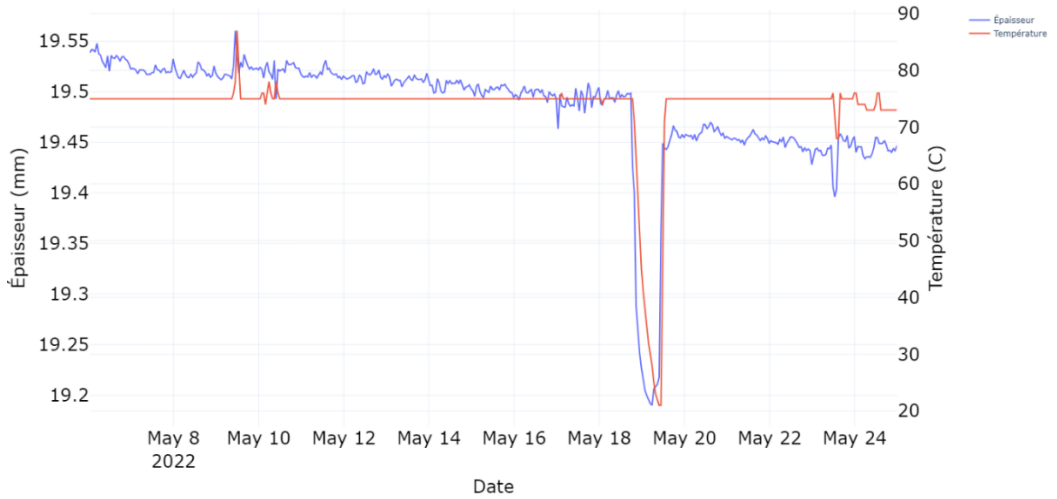
Figure 4.27 Matrice de corrélation des différents paramètres

En exploitant la Figure 4.27, l'erreur de mesure de l'épaisseur a été confirmée en fonction de la variation de la température et du débit en termes de sensibilité et de précision du système, c'est-à-dire qu'une variation de la température du fluide, l'expansion thermique ($\Delta T=54$), peut affecter les échos des capteurs UT. Pour la dégradation de l'épaisseur pendant l'arrêt, la Figure 4.28 montre que l'augmentation du débit de 0 (m^3/h) à 520 (m^3/h) et l'élévation de la température de 20 °C à 75 °C contribuent à la réduction de l'épaisseur de la paroi. De plus, pendant l'arrêt du système, l'acide sulfurique est dilué en raison de la diminution de la concentration

en acide. En outre, la baisse de la température, le pourcentage d'impuretés dans l'acide et les variations de l'humidité relative à l'intérieur du pipeline peuvent favoriser un amincissement drastique du matériau du tuyau [77]. Pour ces raisons, la surveillance de l'épaisseur est nécessaire dans les industries pétrochimiques pour déterminer comment les changements de processus affectent le comportement de la corrosion.



a- Évolution des données d'épaisseur mesurées en fonction du débit d'écoulement



b- Évolution des données d'épaisseur mesurées en fonction de la température

Figure 4.28 Évolutions des données d'épaisseur mesurées en fonction de : a) débit d'écoulement b) température

4.1.4 Synthèse

Ce paragraphe a présenté la collecte et la préparation des trois ensembles de donnes nécessaires pour le développement des modèles de prédiction de défaillance. De plus, les résultats des analyses des données d'entrée ont été illustrés dans ce chapitre en précisant l'effet de chaque paramètre dans les différents modes de dégradation pour les deux études de cas présentés.

4.2 Application des modèles d'apprentissage automatique

4.2.1 Introduction

Dans le monde réel, les données historiques sont généralement très bruitées. Il est particulièrement difficile de créer un modèle de prédiction fiable à partir de données historiques. L'un des principaux intérêts des modelés d'apprentissage automatique est leur capacité à traiter les données historiques, car ils miment le cerveau humain dans sa capacité à faire des modèles de prédiction.

Dans cette section, les méthodes mises en œuvre pour la prédiction de défaillances seront discutées. Ces modèles considèrent comme entrés (attributs) les différents paramètres décrivant la géométrie et le fonctionnement des pipelines : Service, Diamètre, Température opératoire, pression opératoire, matériau, classe de service et le type de circuit. Tout d'abord, les méthodes basées sur l'apprentissage supervisé seront expliquées. Deux modèles de classification seront présentés, le premier modèle est un modèle de classification, qui prédit le niveau de la sévérité de corrosion alors que le deuxième tend à prédire la source de défaillance. Ensuite, les méthodes d'apprentissage non supervisées seront exploitées. Cette technique permet de faire la segmentation des circuits de tuyauterie en des boucles de corrosion en se basant sur les paramètres d'entrée de la base sans savoir le taux de corrosion. D'un autre côté, cette méthode sert aussi à détecter les anomalies au niveau des systèmes de surveillance UT lorsque l'état des capteurs est non connu. Finalement, la performance des modèles de prédiction sera discutée en utilisant les métriques et des données de validations.

4.2.2 Apprentissage automatique supervisé

Une variété de facteurs physiques, environnementaux et opérationnels contribuent à la défaillance des pipelines. Les facteurs pris en compte dans ce modèle sont sélectionnés en fonction de la disponibilité des données historiques fournies.

4.2.2.1 Prédiction du niveau de sévérité de corrosion

- Développement des modèles

Il existe une différence notable entre les problèmes de régression et de classification. Dans notre cas, on a essayé premièrement de développer des modèles de régression qui a pour but de prédire le taux de corrosion moyen pour un tel service. Pour cela on a utilisé la fonction '*groupby*' de python qui calcule le taux de corrosion moyen en respectant un ordre prédéfini de paramètres.

Après l'application des modèles de régression, les résultats obtenus montrent que la métrique R^2 pour les différents modèles d'AA exécutés avec leur paramètre optimal ne dépasse pas 20%. Ceci est dû aux multiples facteurs tels que les modèles de prédiction appliquée n'ont pas capable de capturer des relations implicites entre les variables d'entrées et la variable cible et aussi le manque d'autres paramètres influents (pH, débit d'écoulement).

Pour cela on a converti l'attribut cible de la régression (taux de corrosion) à un problème de classification pour construire lesdits modèles. Les taux de corrosion inférieurs à 2mpy ont été considérés comme "sévérité mineure" et le reste est considéré comme "sévérité sévère". La Figure 4.29 montre la répartition des deux classes de sévérité. On remarque que le nombre d'instances dans chaque classe, la classe "mineure" contient 1481 échantillons, tandis que la classe "sévère" contient 1009 échantillons.

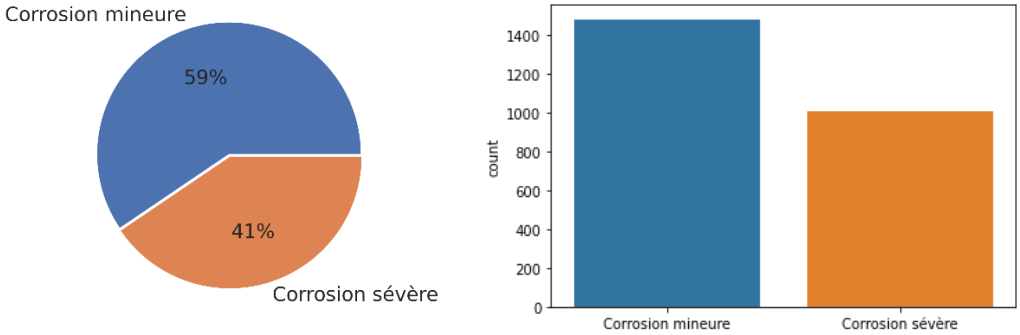


Figure 4.29 Répartition des niveaux de sévérité

Ce modèle est conçu pour prédire objectivement le niveau de sévérité de la corrosion sur la base de données historiques. 8 facteurs sont sélectionnés comme entrées du modèle : Service, diamètre, matériau, type de circuit, classe du service, géométrie, température et pression opératoire. Ces variables sont soit quantitatives,

soit qualitatives. Certaines données de service ont été supprimées à cause de leur faible densité.

Les données qualitatives sont transformées en valeurs quantitatives à l'aide de la technique '*One_hot_encoder*' pour faciliter leur introduction dans le modèle. À la suite l'ensemble des données ont été normalisées, cette étape permet d'améliorer les modèles de l'AA, car les techniques d'apprentissage préfèrent travailler avec des données normalisées. Après la phase de prétraitement, on a commencé le développement des modèles. Pour cela, l'ensemble des données a été divisé en échantillons pour entraîner, tester et valider la performance du modèle. Pour ce faire, nous avons divisé aléatoirement la BD en trois ensembles d'échantillons. L'échantillon d'entraînement représente 60% de l'ensemble de données, et les deux autres échantillons représentent chacun 20 % des données restants. Nous avons utilisé un lot de données de 2490 points, 1494 ont été utilisés comme données d'entraînement.

Après avoir sélectionné les ensembles des données, on a commencé l'entraînement des différents modèles présentés au niveau du CHAPITRE 3:. Avant de tester les algorithmes, l'optimiseur de recherche de paramètres de grille, pris en charge par la librairie *Scikit learn*, a été utilisé pour régler les hyperparamètres de chaque algorithme. Le Tableau 4.14 montre les paramètres optimaux pour chaque modèle.

Tableau 4.14 Description des paramètres optimaux

Modèle de prédiction	Paramètre	Valeur optimale
Random Forest Classifier	bootstrap	True
	Profondeur de l'arbre	300
	Max_features : Nombre de fonctionnalité	sqrt
	Nombre d'arbres	100
Support vector machine	C	10
	Gamma	0,1
	Kernal	Rbf
K-nearest Neighbour	N_neighbors	49
Gradient boosting Classifier	Profondeur de l'arbre	90
	Nombre de fonctionnalité	sqrt
	Nombre d'arbres	300
Decision Tree Classifier	Critère	Entropy
	Profondeur de l'arbre	10
AdaBoost Classifier	Nombre d'arbres	50
	Algorithme de boost	SAME.R
	Coefficient de pondération	1,04
XGBoost Classifier	Profondeur de l'arbre	200
	Max_features : Nombre de fonctionnalité	sqrt
	Nombre d'arbres	100

Après avoir déterminé les valeurs optimales, on a évalué la performance de chaque modèle en utilisant les métriques de classification, la Tableau 4.15 présente les résultats obtenus pour chaque modèle.

Tableau 4.15 Description des performances des modèles développés

Modèle de prédiction	Taux de succès	Précision	Rappel	ROC-AUC
Random Forest	69%	69%	72%	71%
Support vector machine	66.1%	67.2%	64.3%	59.6%
K-nearest Neighbour	67%	69%	69%	69%
Gradient boosting	69%	66.1%	69.3%	68.8%
Decision Tree	69%	65.6%	68.8%	69.2%
XGBClassifier	68%	65,8%	70%	70%
AdaBoost Classifier	63%	64%	72%	69,5%

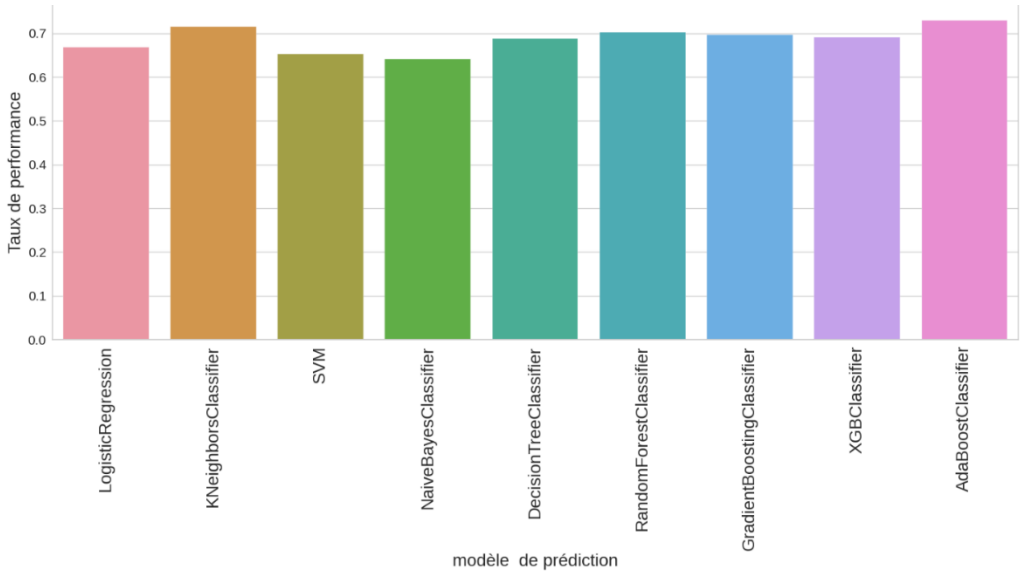
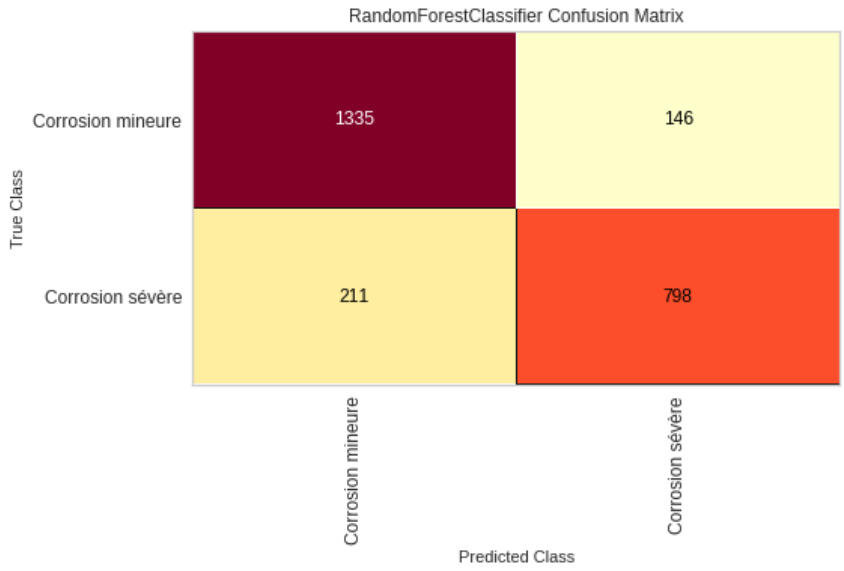
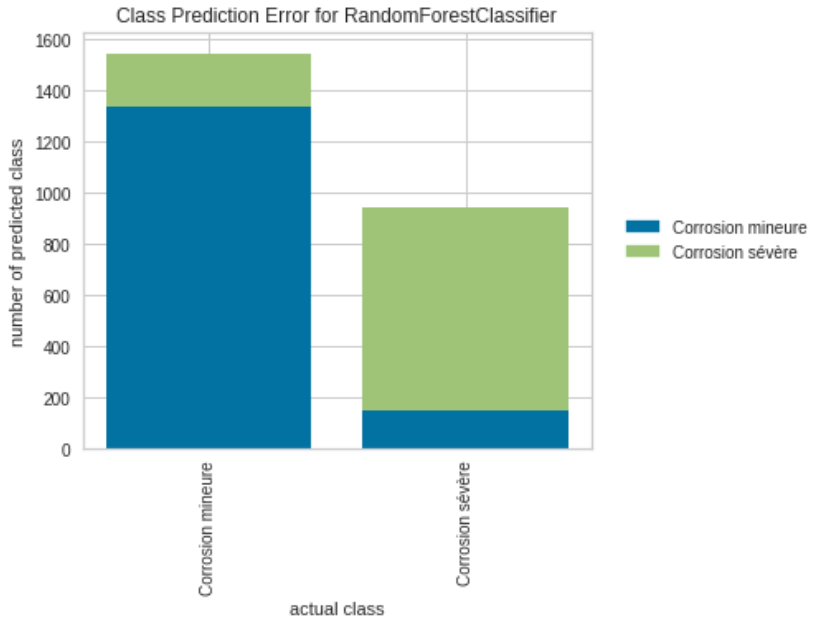


Figure 4.30 Performance des modèles développés

Le Tableau 4.15 et la Figure 4.30 montrent la comparaison des taux de sucées et la métrique AUC-ROC entre les différents modèles de classification que nous avons entraînée, testé et validé. Nous pouvons constater que la performance des modèles ne dépasse pas 72% qui est considérée comme une performance sous-optimale. Cela peut être approuvé aussi par la matrice de confusion et l'erreur de prédiction, Figure 4.31. Les résultats présentés permettent aussi d'évaluer la performance d'un modèle AA en comparant les valeurs prédites et les valeurs réelles du modèle '*Random forest classifier*' qui a la performance la plus élevée.



a) Matrice de confusion



b) Erreur de prédiction

Figure 4.31 Matrice de confusion de l'algorithme ‘Random forest’ et son erreur de prédiction

Malgré la performance sous optimale, la Figure 4.31 montre que le modèle a mal classé 146 échantillons dans la classe "sévérité élevée" et 211 échantillons dans la classe "sévérité faible". Cela signifie que la capacité du modèle à identifier une sévérité élevée est très élevée, ce qui est nécessaire pour prédire les fuites de pipelines.

La Figure 4.32 illustre la courbe ROC-AUC du modèle ‘Random forest classifier’, le AUC-ROC estimé est de 0.71 est proche de 0.5 que de 1, cela signifie que le modèle n'a pas une bonne capacité de séparation de classe.

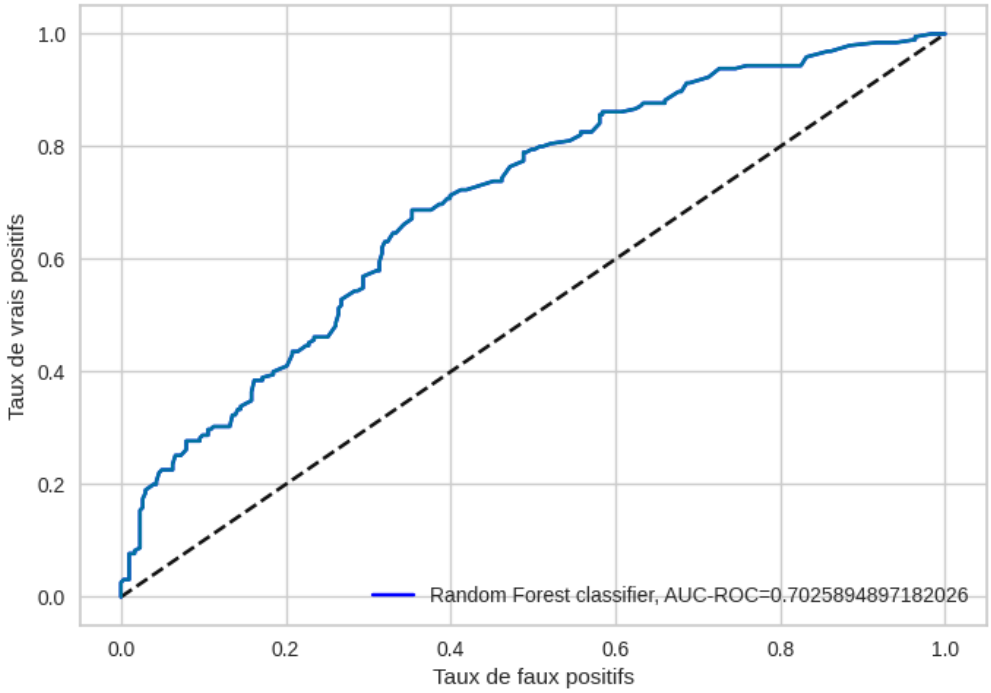


Figure 4.32 Courbe AUC-ROC de l'algorithme Random Forest

Un exemple des résultats de la prédiction par apprentissage automatique supervisé est présenté dans le Tableau 4.16, au niveau de la colonne ‘Niveau de sévérité prédit’. En décrivant la classe de sévérité prédite, nous pouvons davantage comprendre ce qui peut mal tourner à l'avenir dans diverses conditions, par exemple la géométrie du composant, les paramètres opératoires et le service ainsi que sa classe. Le Tableau 4.16 montre l'incertitude de l'algorithme de prédire avec précision. La technique de l'algorithme, qui consiste à trouver une reconnaissance des formes entre les entrées et les sorties pour générer une prédiction, n'est pas transparente. C'est pour cette raison, la validité des résultats prédits sera incertaine dans certains cas. D’autre part, la faible performance est due aussi à la faible corrélation entre les variables d’entrée et de sortie.

Dans la pratique, la corrosion peut être prédite sur la base de plus de 8 caractéristiques. Il y aura des données importantes et complexes ainsi que plusieurs hypothèses que nous devrons traiter pour générer une telle prédiction. En effet, il serait compliqué et frustrant de convertir toutes les connaissances de base disponibles (données et hypothèses) en informations sur la corrosion dans le futur. Nos connaissances sont limitées dans la compréhension et l'intégration des aspects généraux liés à la corrosion des pipelines. Par conséquent, la prédiction peut ne pas

être fiable et des aspects importants liés à l'événement futur peuvent être négligés. De plus, le traitement de cette prédiction peut prendre beaucoup de temps. Dans la pratique, les résultats de l'évaluation doivent être produits rapidement et précisément, car les décisions doivent être prises immédiatement pour résoudre les problèmes rencontrés.

Tableau 4.16 Comparaison des niveaux de sévérité réels et prédits

Service	Diamètre	Matériau	Type de circuit	Classe de service	Température opératoire	Pression opératoire	Géométrie	Niveau de sévérité réel	Niveau de sévérité prédit
EAUDESAL	3	CS	DL	2	275	130	TEE	Faible	Élevé
EAUDESAL	2	CS	DL	2	275	345	COUDE	Élevé	Élevé
EAUDESAL	2	CS	DL	2	275	345	TUYAU	Élevé	Faible
BRUDESAL	12	CS	MPF	2	275	374	COUDE	Élevé	Faible
ACIDGAS	6	CS	MPF	1	110	25	TEE	Faible	Faible
ACIDGAS	0.75	CS	DL	1	101	26	REDUCER	Élevé	Faible
BRUTNAOH	3	CS	DL	1	496	325	COUDE	Faible	Élevé
BRUTNAOH	1	CS	DL	1	374	400	TUYAU	Faible	Faible
KÉROSÈNE	4	CS	MPF	2	375	173	COUDE	Élevé	Faible
GASOILEG	1	CS	DL	1	557	159	TUYAU	Faible	Élevé

La prise en compte des informations de base telles que les données d’inspection, la saisie des données du processus, les paramètres de l'algorithme et les hypothèses est susceptible de collaborer avec l'incertitude. Par conséquent, les résultats de la classification doivent être utilisés avec prudence, car les aspects de l'incertitude ne sont pas reflétés de manière exhaustive. Dans cette partie, ce qui a été prédit comme étant une corrosion mineure peut être une corrosion sévère dans les situations réelles et vice versa. Il est donc primordial de ne pas négliger l'incertitude, car elle peut conduire à l'apparition de résultats surprenants qui peut causer des désastres plus graves pour les systèmes industriels. C'est pourquoi l'incertitude est considérée comme un facteur dominant du risque de défaillance. Pour cela dans notre cas, les résultats de l’apprentissage supervisé ne sont pas suffisamment solides pour servir de base à la prise de décision en matière de prévention des fuites de pipelines.

- Importance des paramètres

Le logiciel Python à travers sa librairie ‘scikit learn’ permet de déterminer l’importance de chaque prédicteur contribuant aux modèles d’AA développées. L'importance des caractéristiques dans l'apprentissage automatique fait référence au processus de détermination de l'importance relative de chaque variable d'entrée (c.-à-d. caractéristique) utilisée dans un modèle d'apprentissage automatique pour faire des prédictions. La compréhension des données et la construction de modèles prédictifs précis en apprentissage automatique dépendent fortement de l'identification de l'importance des caractéristiques. Plusieurs facteurs clés tels que la qualité et la quantité des données, le développement et l'interprétabilité du modèle peuvent influencer l'importance de ces caractéristiques en AA. La Figure 4.33 présente l’importance de chaque prédicteur. Les paramètres opératoires (pression et température) sont considérés comme les paramètres les plus importants.

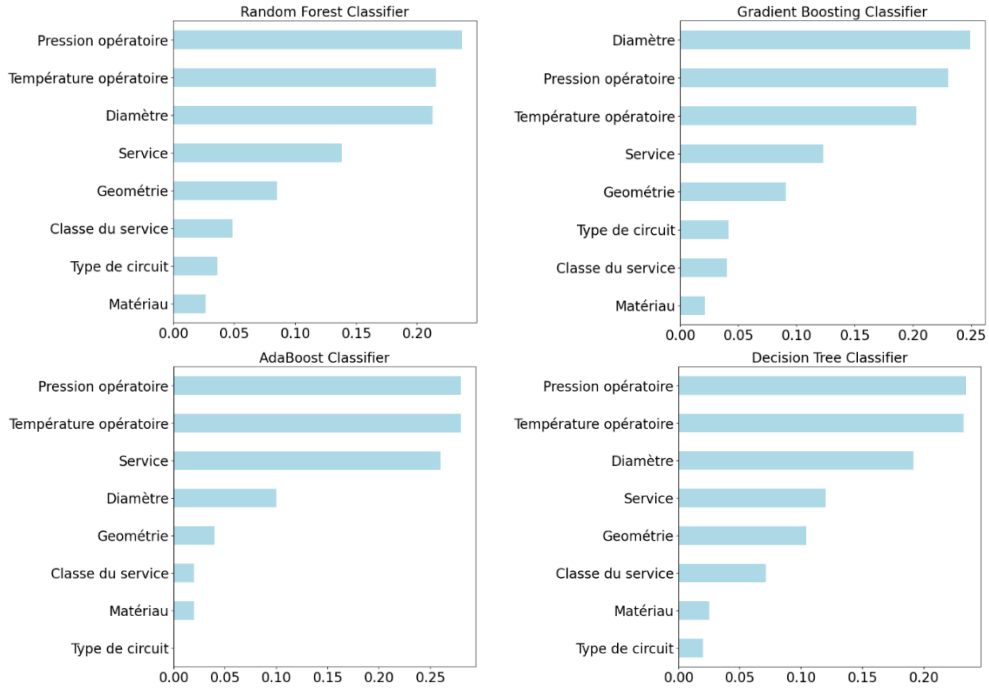


Figure 4.33 Détermination des paramètres dominants

4.2.2.2 Prédiction de la source de défaillance dans les systèmes de tuyauterie

- Développement des modèles

Ce modèle est conçu pour prédire objectivement la source de défaillance qui pourrait menacer le réseau de tuyauterie. La BD collectée comprend 3843 relevés pour les différentes sources de défaillance, y compris les sources de corrosion,

l’érosion et les défaillances mécaniques ainsi que thermiques. Il est à noter qu'au total 1397 d'événements manquent de données pour certains facteurs. Par conséquent, ils sont supprimés de la BD. Les défaillances causées par des risques, opérationnelles et naturelles ne sont pas enregistrées dans l'ensemble de données en raison de leur faible probabilité d'occurrence. La Figure 4.34 présente un échantillon de la BD utilisée pour construire les modèles de prédiction. Chaque source de déversement représente une instance unique et se distingue par cinq caractéristiques distinctes. Le diamètre de la conduite, le type de service, le type de circuit, et les paramètres opératoires sont tous considérés comme des variables explicatives. Le modèle exclut les méthodes de détection des fuites et de l’emplacement de l’élément dans d'installation. Cela peut être attribué à l'impossibilité de déterminer ces variables avant qu'une défaillance ne se produise, alors que le modèle établi est conçu pour anticiper la cause de la défaillance avant son apparition.

Service	Température opératoire	Pression opératoire	Diamètre	Matériau	État de fluide	Type de circuit	Classe de service	Type d'Isolation	Source de défaillance
BRUT	255	345	0,75	CS	LIQUID	DL	2	INSULATED	Corrosion par l'eau d'acide (SW)
BRUDESAL	275	150	6	CS	LIQUID	DL	2	M-MATTRESS	Corrosion par l'eau d'acide (SW)
BRUTNAOH	275	150	1,5	CS	LIQUID	IP	1	INSULATED	Corrosion caustique (CSC)
BRUTNAOH	439	280	14	CS	LIQUID	MPF	1	ASBESTOS	corrosion par acide naphténique et sulfure d'hyhdrogéne à haute température (NA&HTH2S)
GASOILOU	663	50	8	Alloy Stl	LIQUID	MPF	1	BARE	corrosion par acide naphténique et sulfure d'hyhdrogéne à haute température (NA&HTH2S)
GASOILOU	685	110	6	SS	LIQUID	MPF	1	INSULATED	Corrosion par acide polythionique et corrosion sous contrainte de chlorure (PACS&CLS)
KEROSENE	415	60	8	CS	LIQUID	MPF	1	INSULATED	Corrosion par l'eau d'acide (SW)
CRUDOVHD	261	32	36	CS	VAPOR	IP	1	BARE	Corrosion par acide chlorohydrique (HlC)

Figure 4.34 Description de la base de données utilisée

La Figure 4.35 représente le pourcentage d’occurrence de chaque mode de défaillance dans la BD. La corrosion a été identifiée comme le type de défaillance le plus courant pour ces actifs, la corrosion par l'eau acide correspondant à environ un quart de la source totale de défaillance. Toutefois, comme l'indique le titre, ces types de défaillance peuvent être dus à diverses causes sous-jacentes telles que le type de service et les conditions opératoires (température et état des fluides) [72].

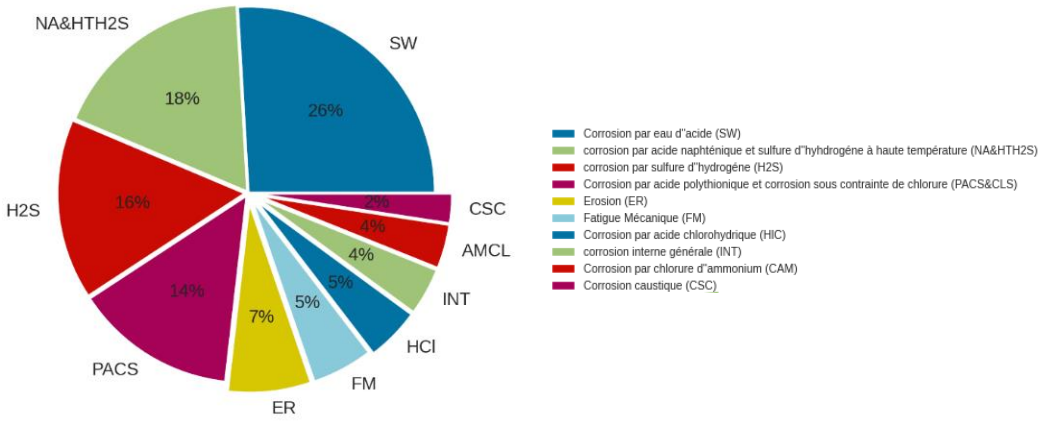


Figure 4.35 Répartition des sources de défaillances

Comme mentionné précédemment, ce modèle est capable de prédire les sources de défaillances dans les systèmes de raffinerie causée par les différents types de corrosion et d'autres défaillances dues par l'érosion et la fatigue mécanique. Ce modèle suit la même procédure que celle présentée auparavant en utilisant 60% des données pour l'entraînement et les 40% sont divisés similairement pour l'ensemble de tests (20%) et l'ensemble de validation (20%). (Division aléatoire). Les hyperparamètres optimaux pour chaque modèle sont décrits au niveau du Tableau 4.17.

Tableau 4.17 Description des paramètres optimaux des modèles développés

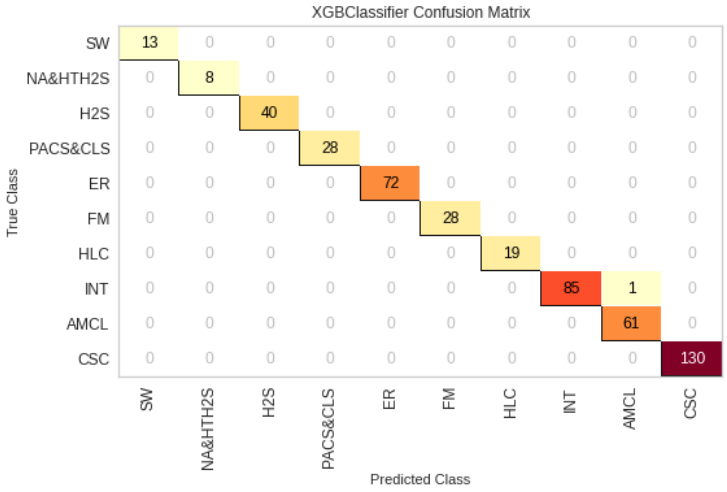
Modèle de prédiction	**Paramètre**	**Valeur optimale**
Random Forest Classifier	bootstrap	False
	Profondeur de l'arbre	50
	Max_features : Nombre de fonctionnalité	Auto
	Nombre d'arbres	100
Support vector machine	C	10
	Gamma	0,1
	Kernel	Rbf
Gradient boosting Classifier	Profondeur de l'arbre	Non spécifiée
	Nombre de fonctionnalité	Tous
	Nombre d'arbres	10
Decision Tree Classifier	Critère	Entropy
	Profondeur de l'arbre	15
AdaBoost Classifier	Nombre d'arbres	400
	Algorithme de boost	SAME.R
	Coefficient de pondération	1,04
XGBoost Classifier	Profondeur de l'arbre	90
	Max_features : Nombre de fonctionnalité	Auto
	Nombre d'arbres	100

Comme il ressort du Tableau 4.18, tous les modèles ont bien performé avec un taux de succès supérieur à 96%, sauf AdaBoost qui a le plus faible taux de succès, égale à 73%. XGB a les meilleures performances en termes de métriques : taux de succès (99.7%), rappel (100%), AUC (100%) et une précision de (99.6%). AdB a le plus faible taux de succès (72.3%), avec des scores de précision (72%), de rappel (72%) et d'AUC (69.5%) les plus bas de tous les classificateurs. Les algorithmes de classification en ensemble (XGB, GDBT et RF) donnent généralement de meilleurs résultats que les classificateurs simples (SVM). Selon la recherche menée par Zhang et al. [78], les techniques de classification ensemblistes surpassent régulièrement les techniques non ensemble pour toutes les métriques d'évaluation. Plus précisément, les classificateurs RF, GBDT et XGB démontrent la plus haute précision de classification sur plusieurs ensembles de données. Cette supériorité s'explique par le fait que l'apprentissage en ensemble améliore la généralisation et la robustesse du classificateur [79]. En outre, la préférence pour les techniques d'ensemble peut être due à la nature des caractéristiques d'entrée, dont cinq sur neuf sont de nature catégorielle. Les classificateurs d'ensemble utilisés dans cette recherche sont construits sur la base d'un apprentissage basé sur l'information, mieux adapté à la manipulation des caractéristiques catégorielles [80]. Cependant, la technique d'ensemble AdB a les moins bonnes performances dans toutes les catégories de métriques, ce qui la rend la moins préférée. Entre GBDT, RF et XGB, il est difficile de sélectionner un classificateur préféré en raison de leur disparité de performance marginale, de l'absence de robustesse dans les différences de performance et de la dispersion de performance analogue. Cependant, XGB est légèrement préféré, car il donne de meilleurs résultats que RF dans toutes les mesures d'évaluation et présente des temps d'exécution d'entraînement et de test plus rapides [78].

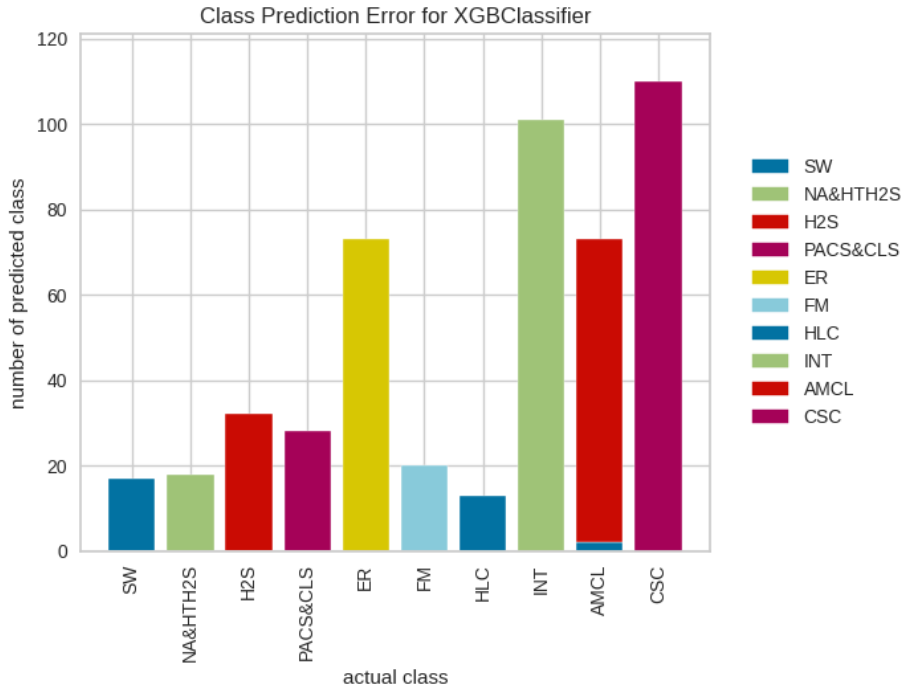
Tableau 4.18 Description des performances des modèles développés

Modèle de prédiction	Taux de sucées	Précision	Rappel	ROC-AUC
Random Forest	99.5%	99.2%	99.4%	99.5%
Support vector machine	96.4%	97.3%	98.3%	98.6%
Gradient Boosting	99.1%	98.9%	99.3%	99.2%
Decision Tree	99.4%	99.1%	99.8%	99.2%
XGB Classifier	99.7%	99.6%	100%	100%
AdaBoost Classifier	72.3%	73.8%	72%	69.5%

La matrice de confusion ainsi que l'erreur de prédiction affirment la performance des algorithmes à prédire la source de défaillance. La Figure 4.36 montre la capacité du modèle à distinguer les différents modes de dégradation. Le nombre d'échantillons mal classé a été considéré négligeable (maximum 4 par classe).



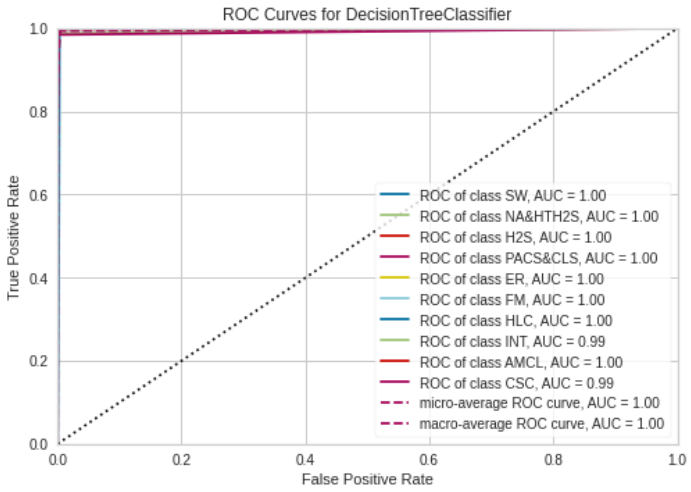
a) Matrice de confusion



b) Erreur de prédiction

Figure 4.36 Matrice de confusion de l'algorithme 'XGB Classifier' et son erreur de prédiction

D'autre part, la courbe AUC-ROC montre aussi que la capacité des modèles à différencier une classe de défaillance d'une autre classe est très élevée (0.99 et souvent 1). La Figure 4.37 illustre un exemple de la courbe AUC-ROC des différents modèles développés.



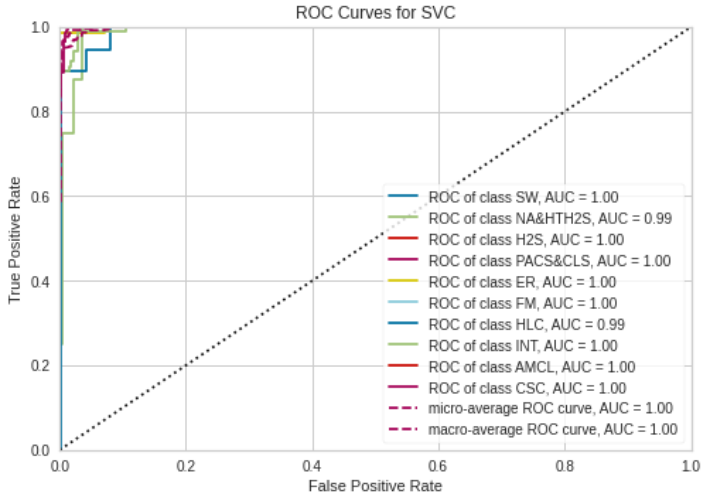


Figure 4.37 Courbe AUC-ROC des différents modèles développés

Il est important de noter qu'aucun classificateur commun n'est particulièrement efficace pour tous les ensembles de données [81]. Même lors de l'utilisation du même ensemble de données, la performance de chaque classificateur peut varier en fonction du prétraitement et des paramètres de données spécifiques appliqués, tels que la sélection des caractéristiques d'entrée et la façon dont ces caractéristiques sont transformées. Pour prédire la source de défaillance dans un système de tuyauterie, le choix du type de service (H_2S, GASOIL, NAPHTA…), la température de fonctionnement et l'état du fluide peuvent tous avoir un impact sur la sélection et la performance des classificateurs.

- Importance des paramètres

Comme illustré dans le modèle précédent, le logiciel python a la capacité de déterminer l'importance de chaque prédicteur contribuant aux algorithmes d'AA. D'après la Figure 4.38, on constate que les facteurs les plus importants sont le service (comme c'est prévu) et la température opératoire, tandis que le prédicteur le moins important est le type de circuit.

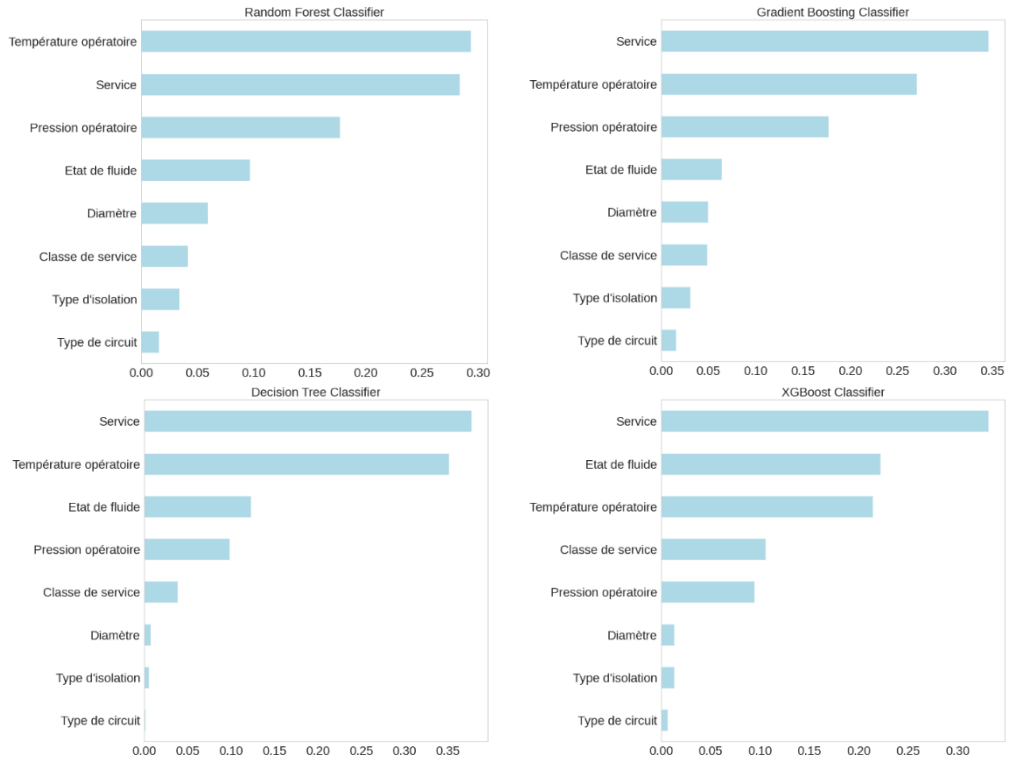


Figure 4.38 Détermination des paramètres dominants

4.2.3 Apprentissage automatique non supervisé

Deux Applications des méthodes d'apprentissage non supervisées seront présentées dans ce paragraphe. La première application sert à utiliser la méthode de clustering pour diviser les sections de tuyauterie en des boucles de corrosion alors que la deuxième sert à exploiter les méthodes AA non supervisée pour faire la détection d'anomalies dans les systèmes de monitoring.

4.2.3.1 Développement des boucles de corrosion

Le développement d'une boucle de corrosion permet d'analyser simultanément un groupe de tuyauteries, réduisant ainsi les activités sans valeur ajoutée en éliminant l'évaluation répétitive des mécanismes de dégradation pour des tuyauteries ayant des caractéristiques opérationnelles et de conception similaires. Ce développement est considéré comme une partie intégrante de la méthodologie d'inspection basée sur le risque. Cependant, le processus de développement de la boucle de corrosion est un type de travail à forte intensité de connaissances qui fait appel au jugement et à l'intuition des experts de corrosion. Cela entraîne une grande variabilité des résultats. Par conséquent, l'objectif de ce paragraphe est d'exploiter les algorithmes d'AA pour développer ladite boucle de corrosion.

Par définition, une boucle de corrosion [38] est définie comme suit :

- La boucle doit être composée de section de tuyauterie dont le type et la phase du fluide, les matériaux de construction, le type de l'isolation et de revêtement sont identiques.
- Les sections de tuyauterie doivent appartenir à la même unité de processus
- Les conditions opératoires (température et pression) doivent être similaires

Sur la base de ces définitions, 8 paramètres sont identifiés à partir de la base de données décrites au début de ce chapitre, le Tableau 4.19 présente les différentes caractéristiques choisies. Les autres données de la base ne sont pas prises en compte en raison de leur faible contribution au processus de développement de la boucle de corrosion.

Tableau 4.19 Types des paramètres utilisés

Donnée	Type de donnée
Service	Qualitative
Phase du fluide	Qualitative
Unité de processus	Qualitative
Matériau	Qualitative
Type d'isolation	Qualitative
Température opératoire	Numérique
Pression opératoire	Numérique
Type de circuit	Qualitative

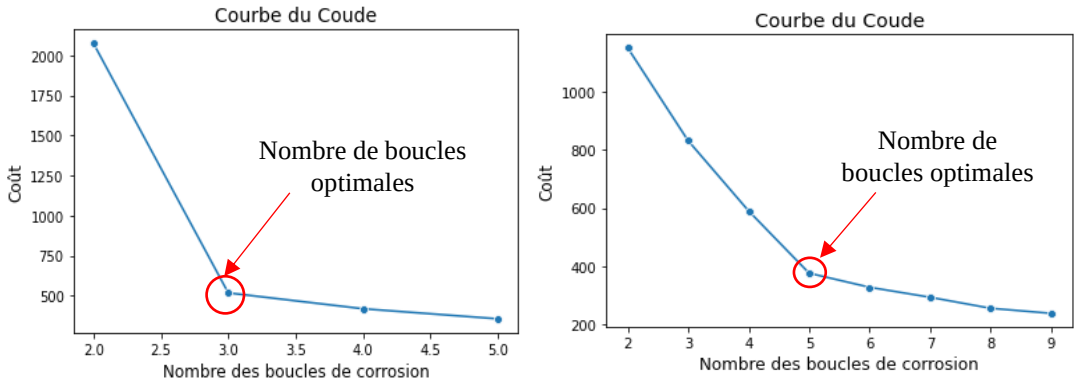
Le tableau ci-dessus montre l'existence de deux types de variable : catégorique et numérique. Pour cela, la méthode K-prototype décrit au niveau chapitre 3 est employée pour répondre à nos objectifs. En outre, une comparaison est effectuée entre les boucles de corrosion générées par l'algorithme K-prototype et l'intelligence humaine (c'est-à-dire le travail fait par des experts) pour déterminer si l'algorithme peut générer des 'K' boucles de corrosion avec une grande cohérence et une faible variabilité que celle du travail manuel. Le processus de traitement des données pour la méthode K-prototype est similaire à ceux qui est présenté au niveau des sections précédentes. En effet, les données avec faible densité seront éliminées à condition qu'elles n'affectent pas le développement de la boucle. À la suite les données numériques préparées seront standardisées. Pour notre mandat, on a déterminé le nombre de boucles pour quelques services et à la suite on a fait la comparaison avec les boucles développées par les experts de la raffinerie.

Le nombre des boucles ‘K’ est déterminé à partir de la méthode ‘Elbow method’. La valeur ‘K’, Figure 4.39, peut être déterminée graphiquement à partir de la courbe (x=nombre de cluster, y=fonction de coût). Le nombre optimal de clusters est celui qui a la plus forte inclinaison au niveau de la courbe.

- Discussion des résultats

Le processus de clustering est réalisé à l’aide du langage de programmation Python. L’algorithme utilisé est le K-prototype que Huang a précédemment étudié [63]. Cet algorithme est adapté au regroupement de données mixtes (catégoriques et numériques).

Après avoir accompli la phase de préparation des données, nous avons implémenté la méthode ‘Elbow method’ de l’algorithme K-prototype pour déterminer le nombre de clusters optimal. La Figure 4.39 présente le nombre de clusters obtenus pour les deux cas de service : ACIDGAS et EAUACIDE.



a) Produit ACIDGAS b) Produit EAUACIDE

Figure 4.39 Courbe de la ‘Elbow method’

Après avoir obtenu les résultats du clustering à l'aide de l'algorithme K-prototype, l'étape suivante consiste à identifier les caractéristiques de chaque cluster. Il est clair à partir des résultats générés que le nombre de boucles dépend du service.

Pour faciliter l'identification, les résultats du clustering peuvent être visualisés à l'avance sous forme graphique afin d'en faciliter l'analyse et la compréhension. Dans ce qui suit, les résultats de l'identification de chaque cluster pour le cas du KÉROSÈNE seront présentés.

On peut déterminer à partir de la Figure 4.40, le nombre optimal des boucles de corrosion, on peut considérer que K=6.

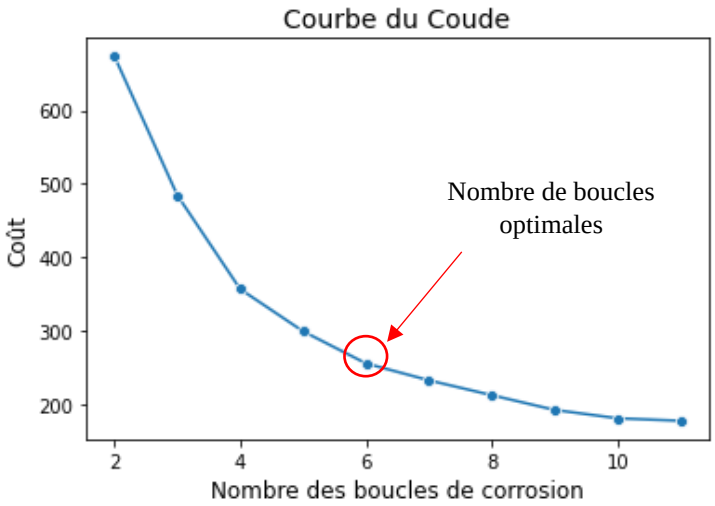


Figure 4.40 Détermination du nombre de boucles optimales pour le cas du KÉROSÈNE

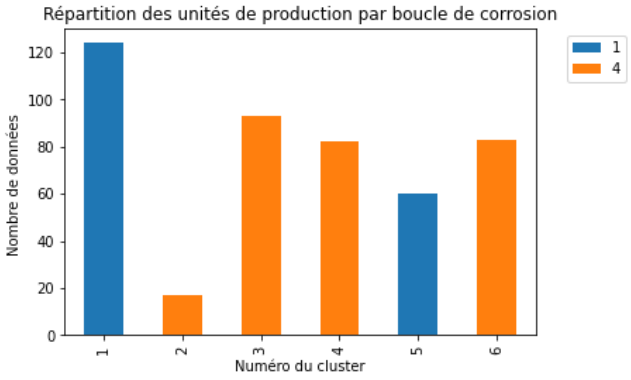
Le nombre total des données récoltées est 459, après le regroupement par K-prototype, les résultats suivants sont obtenus comme suit :

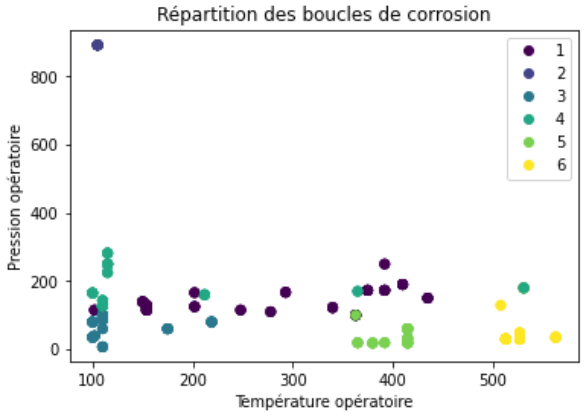
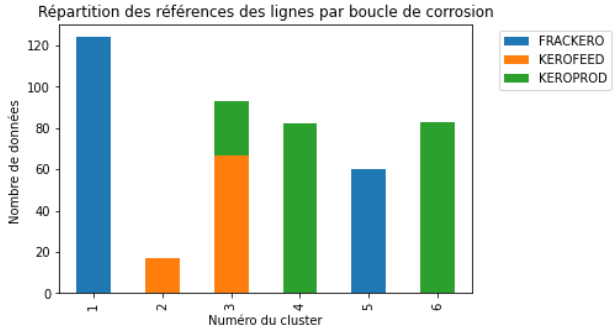
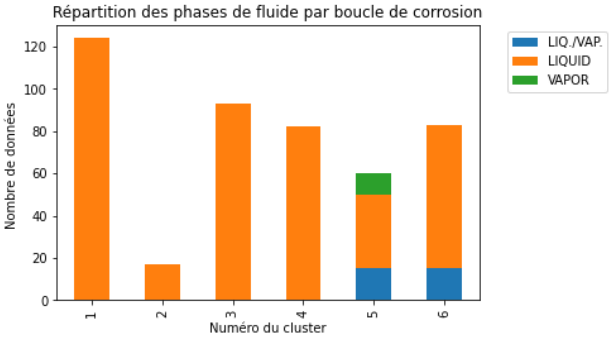
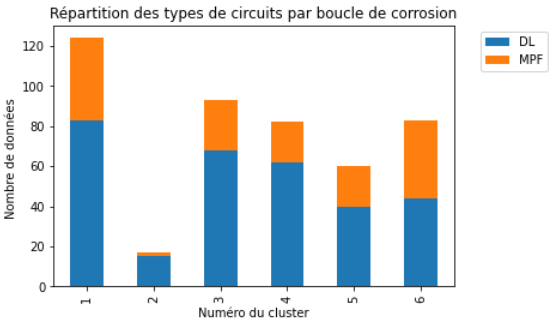
Tableau 4.20 Répartition des données pour chaque cluster

Numéro du cluster	1	2	3	4	5	6
Nombre de données	124	93	83	82	60	17

La distribution des données pour chaque cluster est illustrée au niveau du Tableau 4.20.

Tableau 4.21 Identification des paramètres pour chaque cluster

	L'algorithme a la capacité de distinguer entre les différentes unités. En effet, chaque cluster est caractérisé par une portion de données d'une unité.

	D'après la figure la majorité des clusters sont caractérisés par un intervalle de pression et de température bien déterminée. Par exemple la température opératoire pour le cluster 5 est dans l'intervalle de 400F alors que le cluster 6 est dans l'intervalle dont la température est plus que 500F.
	D'après la figure chaque ligne de tuyauterie peut être divisée en 2 boucles de corrosion.
	D'après la figure le Kérosène à l'état liquide domine la majorité des lignes. Le cluster 5 contient les 3 phases qui peuvent être considérées comme une erreur.
	La distribution des types de circuits est similaire pour les différentes boucles de corrosion.

Le nombre de clusters développés par les experts de cette raffinerie est égal à 6 dont 4 boucles pour l'unité 04 qui est le même résultat trouvé après l'application de l'algorithme. Reste à comparer la répartition des données constituant chaque cluster à partir des deux méthodes utilisées : Manuel (par les experts) et par K-prototype.

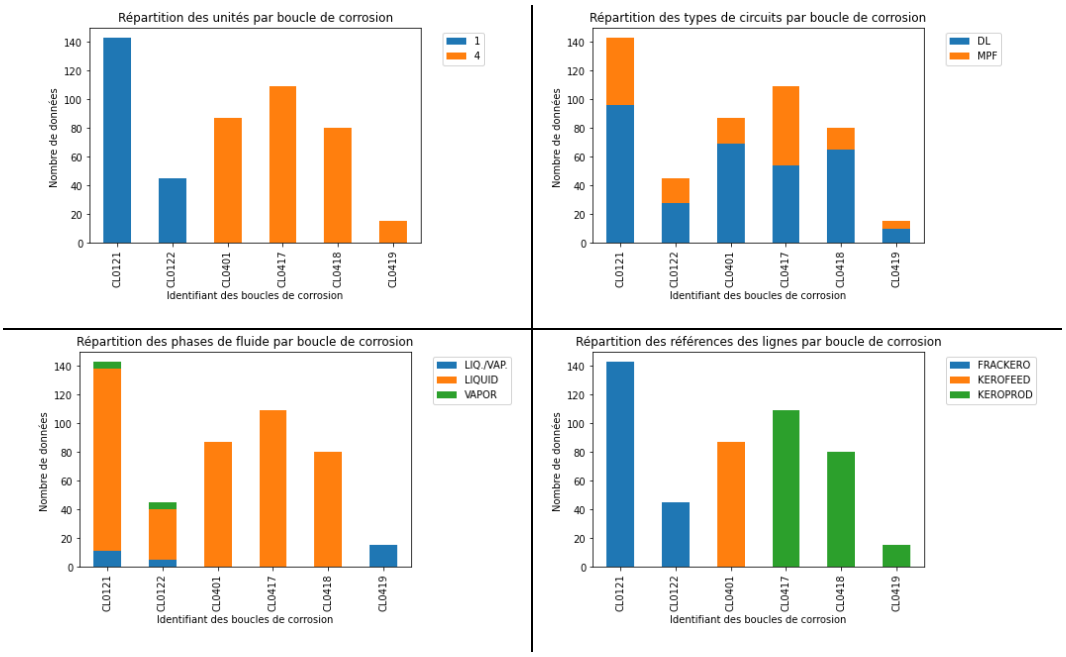


Figure 4.41 descriptions des boucles de corrosion développée par des experts

Les résultats Figure 4.41 montrent qu'il y a une différence au niveau de la répartition des différentes données (température, pression, phase du service…). Mais cela n'empêche pas de confirmer que les méthodes AA aident à connaitre le nombre des boucles de corrosion pour un système de tuyauterie.

Par conséquent, la modification et la révision des résultats obtenus à partir des algorithmes de clustering constituent une étape nécessaire pour permettre l'ajustement des sorties de l'algorithme. En raison de son intensité de connaissances [38], le travail de développement des boucles de corrosion peut encore dépendre des connaissances implicites des ingénieurs et des experts. Alors que la capacité des algorithmes d'apprentissage automatique à traiter les connaissances explicites a déjà été prouvée, leur capacité à traiter les jugements et les intuitions basés sur les connaissances tacites est encore discutable. Par conséquent, il est jugé nécessaire de compléter les résultats des algorithmes par l'intelligence humaine afin de capturer les connaissances tacites [82] et d'affiner les idées générées par l'intelligence des algorithmes [83]. Dans notre cas d'étude, les résultats générés par

l'algorithme peuvent être considérés comme une base pour faciliter l'intégration des règles générales de développement des boucles de corrosion avec des informations et des connaissances supplémentaires provenant des experts.

4.2.3.2 Détection des Anomalies dans les systèmes de surveillance

La détection des anomalies dans les séries temporelles joue un rôle essentiel dans les systèmes de surveillance. Il s'agit d'un sujet de plus en plus important aujourd'hui, en raison de son application élargie au contexte du contrôle des systèmes industriels. Dans notre cas d'étude, nous visons à identifier les anomalies dans les capteurs ultrasoniques qui servent à contrôler l'amincissement d'épaisseur des conduites de transport d'acide sulfurique. Pour cela, le capteur présenté au niveau du chapitre 4 est considéré comme un exemple d'étude, avec un ensemble de données historiques décrivant la variation de l'épaisseur acquis par le capteur UT. L'ensemble de données obtenu couvre la période du 5 mai 2022 au 25 mai 2022. Le choix de cette période est basé sur le fait que des périodes d'arrêt ont eu lieu pendant cette époque.

La Figure 4.42 montre l'existence des valeurs aberrante au niveau de la série temporelle de la mesure d'épaisseur même en dehors de la période d'arrêt.

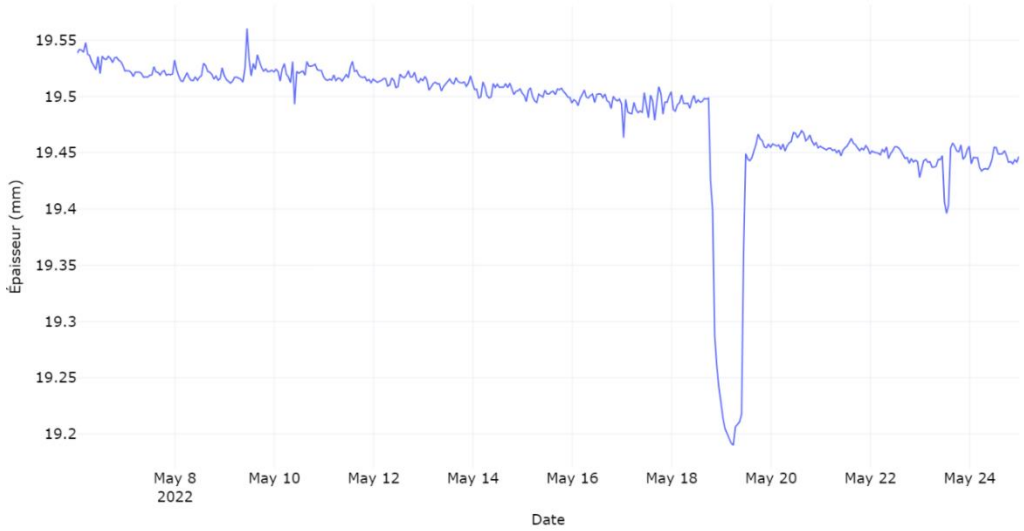


Figure 4.42 Variation de l'épaisseur mesurée

- Description des algorithmes de détection d'anomalie adoptés

Les différentes techniques et méthodes mises en œuvre dans ce travail seront discutées dans cette section. Les algorithmes d'AA non supervisé sont parfaitement

adaptés à l'identification des anomalies dans les grands ensembles de données qui n'ont pas besoin d'être étiquetés. En particulier, nous mettons en valeur les algorithmes suivants : *Interquartile Range, K-Nearest Neighbors (K-NN), Local Outlier Factor (LOF), et Isolation Forest (IF).*

Pour étudier et explorer les différentes approches décrites dans le chapitre 3, nous avons utilisé la bibliothèque Python pour la détection des valeurs aberrantes (*PyOD)* et *l'Anomaly Detection Toolkit (ATDK)*, sont deux bibliothèques qui contiennent plusieurs outils d'AA non supervisés pour détecter les anomalies dans les séries temporelles. La librairie *Matplotlib* est utilisée pour visualiser la dispersion de l'ensemble des données d'épaisseur à l'aide de *boxplots*, Figure 4.43. Cette méthode permet d'identifier la plage des valeurs aberrantes.

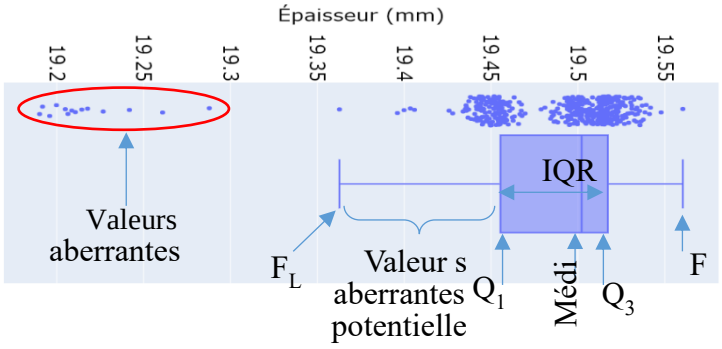


Figure 4.43 Identification des valeurs aberrantes par les *boxplots*

- **Discussion des résultats**

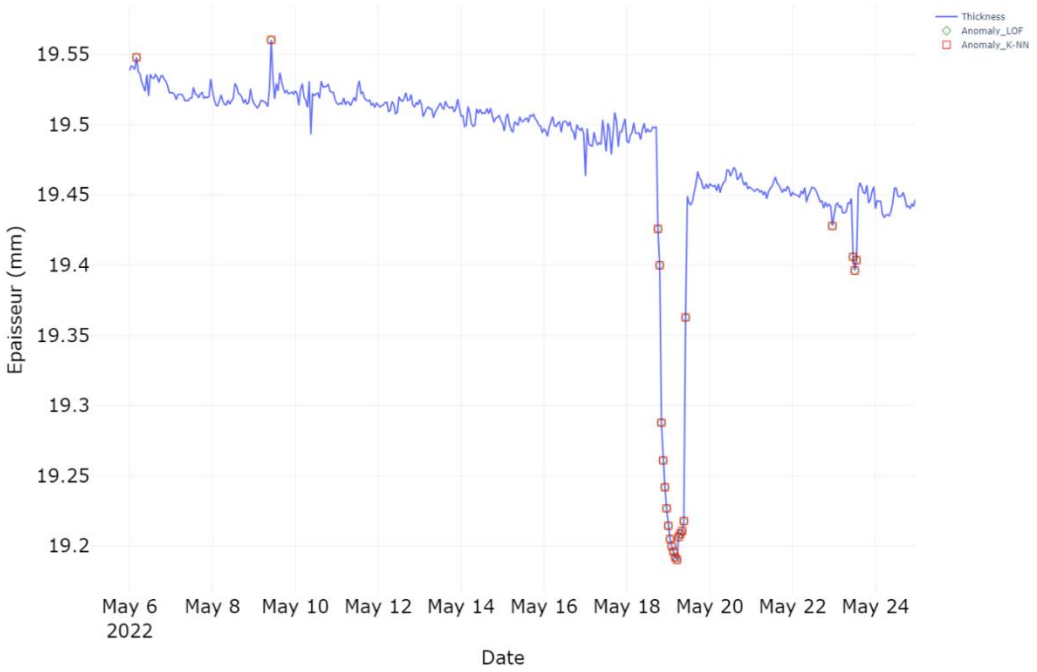
Avant de pouvoir tester la capacité à prédire les valeurs aberrantes dans les ensembles de séries temporelles, il faut définir un ensemble de paramètres essentiels pour la méthode proposée. Pour K-NN et LOF, le nombre de voisins et le degré de contamination de l'ensemble de données sont des paramètres fondamentaux. Pour simplifier, la contamination est définie comme la proportion attendue de valeurs aberrantes dans les données. Pour chaque paramètre, plusieurs valeurs ont été testées dans ce travail pour étudier leurs effets, puis les valeurs optimales seront déterminées. D'autres paramètres standard sont utilisés : la distance euclidienne standard est utilisée comme un paramètre métrique. Pour IF, le nombre d'estimateurs est un paramètre de base lié au nombre d'arbres utilisés pour construire la forêt. IF a la même valeur de contamination que les autres méthodes. Pour IQR, un seul paramètre doit être défini, c=1,5. Les méthodes ci-

dessus visent à retenir les points aberrants et à éliminer les points normaux. En d'autres termes, le point anormal prédit sera étiqueté avec une valeur de 1. Ces résultats sont utilisés à la suite pour construire la matrice de confusion.

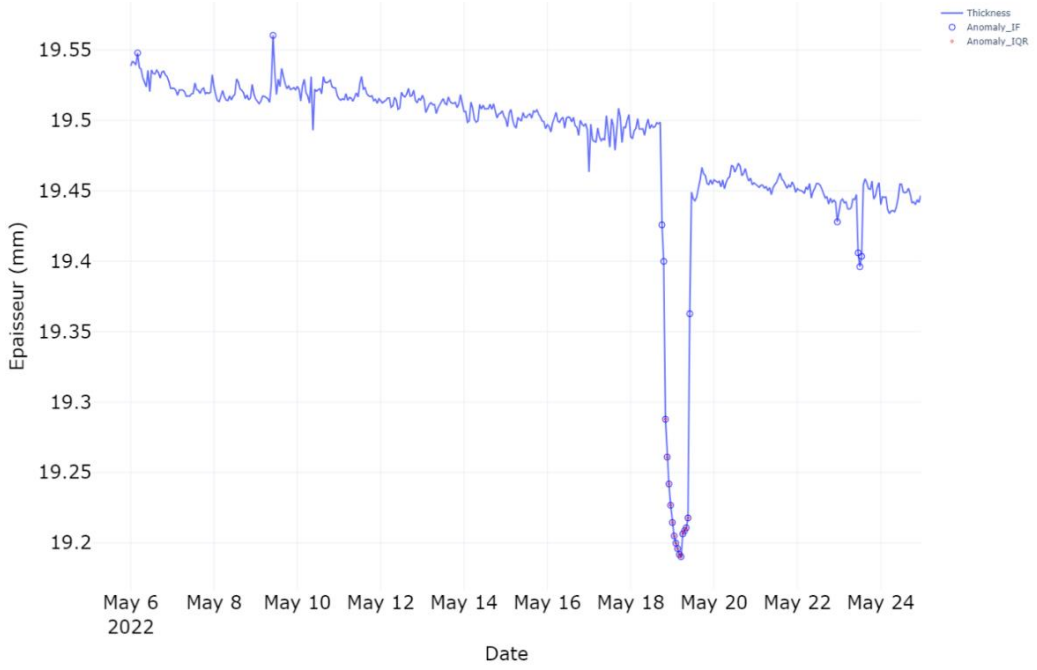
La mise en œuvre des paramètres optimaux pour chaque modèle est étudiée pour évaluer leurs performances, comme la montre le Tableau 4.22. Les résultats de chaque méthode sont présentés dans la Figure 4.44. Nous pouvons remarquer que IQR n'a pas réussi à détecter toutes les irrégularités avec un total de 9 fausses anomalies. Les méthodes K-NN et LOF ont détecté les mêmes points, ils ont réussi à détecter les anomalies correctes (21 vraies anomalies positives), tandis que 3 fausses anomalies ont été identifiées. La forêt d'isolation a également détecté les vraies anomalies positives. En outre, lorsque chaque modèle a été exécuté, les résultats ont montré que ces modèles sont sensibles aux signaux bruyants et qu'il nécessite davantage de données pour sélectionner le meilleur modèle. En d'autres termes, la performance de chaque algorithme pour distinguer une fausse anomalie d'une vraie anomalie.

Tableau 4.22 Déterminations des valeurs optimales et la performance des algorithmes utilisés

Modèle	Paramètre optimal	P	R	F1
K-NN	Nombre de voisins = 2, contamination =0.05	0.88	1	0.93
LOF	Nombre de voisins = 20, contamination =0.05	0.88	1	0.93
IF	Nombre d'estimateurs =20, contamination =0.05	0.88	1	0.93



a- Détection des anomalies par les méthodes LOF et K-NN



b- Détection des anomalies par les méthodes IF et IQR

Figure 4.44 Détection des anomalies

Pour l'étape suivante, la méthode IF est utilisée pour nettoyer les valeurs aberrantes sur la période d'étude. Le choix de l'IF est basé sur sa haute performance et sa simplicité, car il utilise une séquence d'arbres et chaque arbre essaie de corriger les erreurs de celui qui le précède en termes de prédiction. La Figure 4.45 présente la

distribution de mesures d'épaisseur (points verts) après le nettoyage des valeurs aberrantes. Comme mentionné ci-dessus, la principale fonction des systèmes de surveillance en ligne est de contrôler l'amincissement d'épaisseur pendant les différentes périodes de fonctionnement. Par conséquent, on a employé 3 différents modèles polynomiaux (linéaire, quadratique et cubique) pour prédire la variation de l'épaisseur sur cette période en fonction du temps, ce qui permet à la suite d'estimer le taux de corrosion. Sur la base de la fonction R^2 calculée, la Figure 4.45 montre que le modèle appliqué peut performer pour prédire les mesures d'épaisseur avec $R^2 \geq 0.87$. Le taux de corrosion estimé sur cette période est de 60 mpy avant et après les arrêts, ce qui représente une dégradation importante par rapport au taux normal ≈ 6 mpy [77]. Cette importante dégradation est due à la contribution de plusieurs facteurs discutés précédemment (arrêts, corrosion accélérée, dilatation thermique…). Pour cela il est nécessaire de disposer des capteurs UT permanents dans les zones où les taux de corrosion semblent à être élevé afin de mieux contrôler l'amincissement d'épaisseur et éviter tout risque de défaillance due à la corrosion.

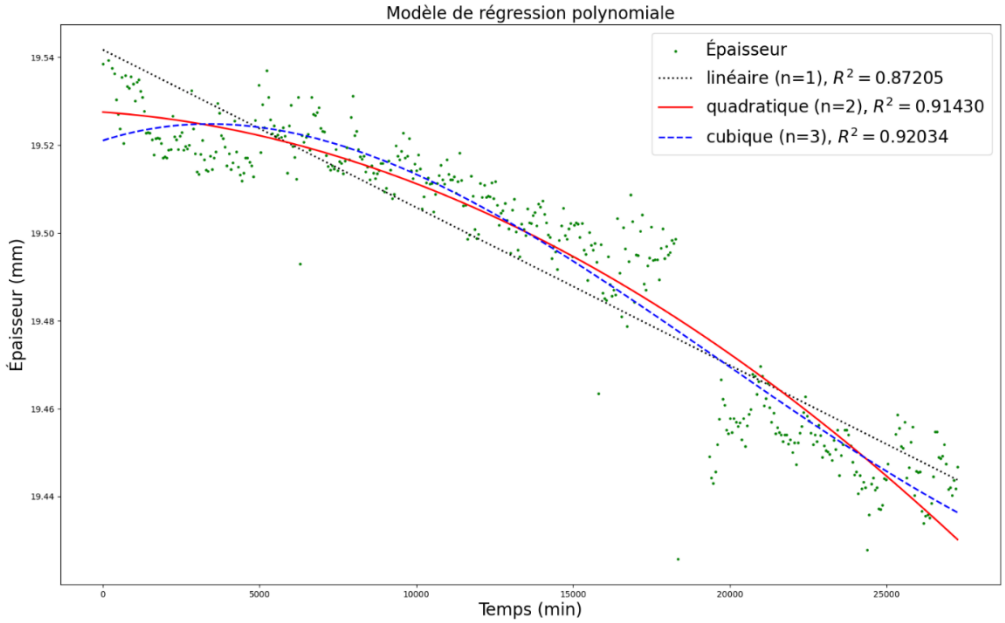


Figure 4.45 Estimation du taux de corrosion

4.2.4 Synthèse

Deux approches d'apprentissage automatiques ont été développées dans ce paragraphe. La première est la technique d'AA supervisée basée sur les algorithmes de classification. La deuxième approche a été dédiée à la technique d'AA non supervisée. On s'est attardé, plus particulièrement, sur la présentation des

algorithmes développés et leur performance dans le contexte de la prédiction des défaillances en présentant la performance de chaque technique et en identifiant la cause de défaillance qui serait la plus susceptible de menacer le fonctionnement des systèmes de tuyauterie.

CHAPITRE 5: CONCLUSION GÉNÉRALE ET PERSPECTIVES

La présente recherche a pour but d'analyser la pertinence des techniques l'apprentissage automatique en tant qu'outils de prédiction des défaillances dans les réseaux de tuyauterie. Sur la base de la revue de la littérature et des résultats obtenus durant ce mandat, les principales conclusions de ce travail sont résumées comme suit :

- La réussite de la prédiction des défaillances à l'aide d'algorithmes d'apprentissage automatique dépend de divers facteurs, notamment de la qualité et de la quantité de données, de la sélection et du réglage des paramètres de l'algorithme approprié.

- Il est essentiel de reconnaître que la prédiction des défaillances n'est pas un événement ponctuel, mais un processus continu qui nécessite une surveillance et une amélioration permanentes. À mesure que de nouvelles données sont disponibles et que les conditions opérationnelles changent, les modèles d'apprentissage automatique doivent être mis à jour et réentraînés pour garantir leur précision et leur efficacité.

- Plusieurs bénéfices peuvent être acquis en effectuant les algorithmes d’apprentissage automatique pour prédire les défauts. Des données détaillées concernant le type, le degré de gravité et les facteurs qui peuvent conduire à la défaillance peuvent être détectées simultanément. En outre, la performance de l'outil en matière de classification et regroupement (clustering) des données permet aux analystes des risques d'identifier, de hiérarchiser et de surveiller la corrosion ainsi que les autres modes de dégradations des pipelines.

- L'utilisation d'algorithmes d'apprentissage automatique pour la prédiction des défaillances est un outil puissant qui peut améliorer l'efficacité opérationnelle, réduire les coûts de maintenance et améliorer la sécurité. Cependant, il est crucial d'aborder le processus avec prudence et en comprenant bien ses limites et ses exigences.

- Les limites de l'apprentissage automatique sont que les aspects importants de l'incertitude et du risque n’est pas reflété de manière exhaustive dans les résultats prédits. Par conséquent, les prévisions concernant un tel défaut peuvent être fausses

(par exemple, une corrosion mineure peut s'avérer grave en réalité). En outre, la prise de décision dans la définition des interventions préventives pour atténuer les fuites des pipelines peut être difficile en raison du manque d'informations sur le risque auquel on est confronté. En bref, une interprétation erronée dans la description des phénomènes futurs et dans le choix des mesures de sécurité peut coûter cher aux entreprises à grande échelle.

Par conséquent, plusieurs suggestions de travaux futurs sont fournies afin d'améliorer et de renforcer l'adéquation de l'approche d'apprentissage automatique pour prévoir les défauts et ses résultats pour servir de base à la décision d'éviter les fuites de pipelines :

- Développer un modèle d'évaluation des systèmes de tuyauterie. Ce modèle permettra de définir l'état d'un pipeline, ce qui aidera les responsables à évaluer l'état d'un pipeline et à effectuer les actions de maintenance appropriées.
- Concevoir une échelle d'état pour des systèmes de tuyauterie, qui reflète l'état de détérioration d'un pipeline en utilisant une échelle de 0 à 10 où 0 est le pire état et 10 le meilleur état. Cette échelle pourrait être développée en utilisant des avis d'experts pour évaluer l'état des pipelines.
- Incorporer davantage de sources de données : actuellement, la plupart des modèles de prédiction des défaillances s'appuient sur les données des capteurs pour prédire les défaillances des équipements. À l'avenir, les chercheurs pourraient intégrer d'autres sources de données, telles que les journaux de maintenance, les données environnementales et les commentaires des opérateurs, afin d'améliorer la précision des modèles.
- Améliorer l'interprétabilité des modèles : à mesure que les modèles d'apprentissage automatique deviennent plus complexes, il peut être difficile de comprendre comment ils font des prédictions. Les recherches futures se concentreront sur le développement de méthodes d'interprétation des prédictions des modèles et sur l'identification des caractéristiques les plus déterminantes pour la prédiction des défaillances.
- Développer des modèles d'ensemble qui combinent les prédictions de plusieurs modèles d'apprentissage automatique (ANN, SVM, Gradient Boosting, …)pour améliorer la précision globale des prédictions. Les recherches futures se

concentreront sur le développement de modèles d'ensemble qui peuvent être facilement déployés dans des environnements industriels et qui peuvent traiter des données en continu.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Alliance F. An Introduction to Oil and Gas Pipelines. FracTracker Alliance 2016. https://www.fractracker.org/2016/06/introduction-oil-gas-pipelines/ (accessed March 4, 2023).

[2] Sadiq R, Rajani B, Kleiner Y. Probabilistic risk analyis of corrosion associated failures in cast iron water mines. Reliab Eng Syst Saf 2004;86:1–10. https://doi.org/10.1016/j.ress.2003.12.007.

[3] M36 Water Audits and Loss Control Programs, Fourth Edition n.d. https://engage.awwa.org/PersonifyEbusiness/Bookstore/Product-Details/productId/51439782 (accessed March 4, 2023).

[4] Zakikhani K. Failure prediction and availability-based maintenance planning of gas transmission pipelines n.d.

[5] Big Data et maintenance prédictive : Du réactif au proactif ! - Toute l'actualité B2B sur manageo.fr n.d. https://blog.manageo.fr/data/big-data-et-maintenance-predictive-du-reactif-au-proactif-_bar13855 (accessed March 4, 2023).

[6] Pipeline Risk Management Manual - 3rd Edition n.d. https://www.elsevier.com/books/pipeline-risk-management-manual/muhlbauer/978-0-7506-7579-6 (accessed March 4, 2023).

[7] Adegboye MA, Fung W-K, Karnik A. Recent Advances in Pipeline Monitoring and Oil Leakage Detection Technologies: Principles and Approaches. Sensors 2019;19:2548. https://doi.org/10.3390/s19112548.

[8] Achouch M, Dimitrova M, Ziane K, Sattarpanah Karganroudi S, Dhouib R, Ibrahim H, et al. On Predictive Maintenance in Industry 4.0: Overview, Models, and Challenges. Appl Sci 2022;12:8081. https://doi.org/10.3390/app12168081.

[9] American Society for Testing Materials (ASTM): Proceedings, Vol. 57 1957 by American Society for Testing Materials: Good Hard Cover (1958) | Top Notch Books n.d. https://www.abebooks.com/American-Society-Testing-Materials-ASTM-Proceedings/60528349/bd (accessed March 4, 2023).

[10] B31.1 - Power Piping - ASME n.d. https://www.asme.org/codes-standards/find-codes-standards/b31-1-power-piping (accessed March 4, 2023).

[11] Concawe Reports. Concawe n.d. https://www.concawe.eu/publications/concawe-reports/ (accessed March 4, 2023).

[12] Davis PM, Dubois J, Gambardella F, Sanchez-Garcia E, Uhlig F, Haan K, et al. Performance of European crosscountry oil pipelines - Statistical summary of reported spillages in 2010 and since 1971 2011.

[13] Muhlbauer WK. Pipeline Risk Management Manual: Ideas, Techniques, and Resources. Elsevier; 2004.

[14] Chiara Bersani, Lucia Citro, Roberta Valentina Gagliardi, Roberto Sacile, Angela Maria Tomasoni. Accident occurrance evaluation in the pipeline transport of dangerous goods. Chem Eng Trans 2010;19:249–54. https://doi.org/10.3303/CET1019041.

[15] PHMSA: Stakeholder Communications - Enforcement Activity n.d. https://primis.phmsa.dot.gov/comm/reports/enforce/enforcement.html (accessed March 4, 2023).

[16] Pipeline Corrosion - Final Report | PHMSA n.d. https://www.phmsa.dot.gov/pipeline/hazardous-liquid-integrity-management/pipeline-corrosion-final-report (accessed March 4, 2023).
[17] Condition Monitoring: An Overview | Reliable Plant n.d. https://www.reliableplant.com/condition-monitoring-31760 (accessed March 4, 2023).
[18] API | API 570 - Piping Inspector n.d. https://www.api.org/products-and-services/individual-certification-programs/certifications/api570 (accessed March 4, 2023).
[19] DNV.com - When trust matters. DNV n.d. https://172.26.30.239/Default (accessed May 21, 2023).
[20] Piping Thickness Management: CML Placement | Inspectioneering n.d. https://inspectioneering.com/journal/2012-11-01/2995/a-discussion-on-the-piping-thi (accessed March 4, 2023).
[21] Condition Monitoring Location Optimization Helps Facility Reduce Risk. Pinnacle n.d. https://pinnaclereliability.com/learn/case-studies/cml-optimization-pilot-project-helps-refinery-reduce-risk-and-identify-minimum-reduced-inspection-spend-of-384k/ (accessed March 4, 2023).
[22] API | API 510 - Pressure Vessel Inspector n.d. https://www.api.org/products-and-services/individual-certification-programs/certifications/api510 (accessed March 4, 2023).
[23] API 580 - Risk Based Inspection n.d. https://www.api.org/products-and-services/individual-certification-programs/certifications/api580 (accessed March 4, 2023).
[24] Torngats. Torngats n.d. https://torngats.ca/ (accessed March 4, 2023).
[25] Beuker T, Palmer J, Quack M. In-Line Inspection using Combined Technologies - Magnetic Flux Leakage and Ultrasonic Testing and their Advantages 2009.
[26] Liu S, Wang H, Li R. Attention Module Magnetic Flux Leakage Linked Deep Residual Network for Pipeline In-Line Inspection. Sensors 2022;22:2230. https://doi.org/10.3390/s22062230.
[27] Measurement uncertainty evaluation of ultrasonic wall thickness measurement - ScienceDirect n.d. https://www.sciencedirect.com/science/article/pii/S0263224119300375?via%3Dihub (accessed March 4, 2023).
[28] Optimizing Performance Assurance | TEAM n.d. https://www.teaminc.fr/solutions-de-services/inspection/cnd-evolue/transducteur-acoustique-electromagnetique-emat (accessed March 4, 2023).
[29] ROSEN - EMAT - Electro Magnetic Acoustic Transducer - ROSEN Group n.d. https://www.rosen-group.com/global/company/explore/we-can/technologies/measurement/emat.html (accessed May 21, 2023).
[30] Lebowitz CA, Brown LM. Ultrasonic Measurement of Pipe Thickness. In: Thompson DO, Chimenti DE, editors. Rev. Prog. Quant. Nondestruct. Eval. Vol. 12A 12B, Boston, MA: Springer US; 1993, p. 1987–94. https://doi.org/10.1007/978-1-4615-2848-7_255.
[31] Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom, 2016, p. 1310–5.

[32] Schneider P, Xhafa F. Chapter 8 - Machine learning: ML for eHealth systems. In: Schneider P, Xhafa F, editors. Anom. Detect. Complex Event Process. IoT Data Streams, Academic Press; 2022, p. 149–91. https://doi.org/10.1016/B978-0-12-823818-9.00019-5.
[33] Senouci A, Elabbasy M, Elwakil E, Abdrabou B, Zayed T. A model for predicting failure of oil pipelines 2014.
[34] Zakikhani K, Nasiri F, Zayed T. Availability-based reliability-centered maintenance planning for gas transmission pipelines. Int J Press Vessels Pip 2020;183:104105. https://doi.org/10.1016/j.ijpvp.2020.104105.
[35] Liao K, Yao Q, Wu X, Jia W. A Numerical Corrosion Rate Prediction Method for Direct Assessment of Wet Gas Gathering Pipelines Internal Corrosion. Energies 2012;5:3892–907. https://doi.org/10.3390/en5103892.
[36] Aljameel SS, Alomari DM, Alismail S, Khawaher F, Alkhudhair AA, Aljubran F, et al. An Anomaly Detection Model for Oil and Gas Pipelines Using Machine Learning. Computation 2022;10:138. https://doi.org/10.3390/computation10080138.
[37] De Kerf T, Gladines J, Sels S, Vanlanduit S. Oil Spill Detection Using Machine Learning and Infrared Images. Remote Sens 2020;12:4090. https://doi.org/10.3390/rs12244090.
[38] Corrosion loop development of oil and gas piping system based on machine learning and group technology method | Emerald Insight n.d. https://www.emerald.com/insight/content/doi/10.1108/JQME-07-2018-0058/full/html (accessed March 4, 2023).
[39] Zhong S, Fu S, Lin L, Fu X, Cui Z, Wang R. A novel unsupervised anomaly detection for gas turbine using Isolation Forest, 2019, p. 1–6. https://doi.org/10.1109/ICPHM.2019.8819409.
[40] Ibrahim M, Alsheikh A, Awaysheh FM, Alshehri MD. Machine Learning Schemes for Anomaly Detection in Solar Power Plants. Energies 2022;15:1082. https://doi.org/10.3390/en15031082.
[41] Hu D, Zhang C, Yang T, Chen G. Anomaly Detection of Power Plant Equipment Using Long Short-Term Memory Based Autoencoder Neural Network. Sensors 2020;20:6164. https://doi.org/10.3390/s20216164.
[42] Chen J, Li J, Xu Z, Zhang L, Qi S, Yang B, et al. Prediction model of early biomarkers of massive cerebral infarction caused by anterior circulation occlusion: Establishment and evaluation. Front Neurol 2022;13:903730. https://doi.org/10.3389/fneur.2022.903730.
[43] Asri H, Mousannif H, Moatassime HA, Noel T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Comput Sci 2016;83:1064–9. https://doi.org/10.1016/j.procs.2016.04.224.
[44] scikit-learn: machine learning in Python — scikit-learn 1.2.1 documentation n.d. https://scikit-learn.org/stable/ (accessed March 4, 2023).
[45] So A, Hooshyar D, Park KW, Lim HS. Early Diagnosis of Dementia from Clinical Data by Machine Learning Techniques. Appl Sci 2017;7:651. https://doi.org/10.3390/app7070651.
[46] sklearn.preprocessing.LabelEncoder. Scikit-Learn n.d. https://scikit-learn/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html (accessed March 4, 2023).

[47] sklearn.preprocessing.MinMaxScaler. Scikit-Learn n.d. https://scikit-learn/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html (accessed March 4, 2023).
[48] sklearn.preprocessing.StandardScaler. Scikit-Learn n.d. https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (accessed March 4, 2023).
[49] sklearn.preprocessing.RobustScaler. Scikit-Learn n.d. https://scikit-learn/stable/modules/generated/sklearn.preprocessing.RobustScaler.html (accessed March 4, 2023).
[50] Guikema SD. Natural disaster risk analysis for critical infrastructure systems: An approach based on statistical learning theory. Reliab Eng Syst Saf 2009;94:855–60. https://doi.org/10.1016/j.ress.2008.09.003.
[51] 1.10. Decision Trees. Scikit-Learn n.d. https://scikit-learn/stable/modules/tree.html (accessed March 4, 2023).
[52] Ben Seghier MEA, Höche D, Zheludkevich M. Prediction of the internal corrosion rate for oil and gas pipeline: Implementation of ensemble learning techniques. J Nat Gas Sci Eng 2022;99:104425. https://doi.org/10.1016/j.jngse.2022.104425.
[53] A Beginner's Guide to Supervised Machine Learning Algorithms | by Soner Yıldırım | Towards Data Science n.d. https://towardsdatascience.com/a-beginners-guide-to-supervised-machine-learning-algorithms-6e7cd9f177d5 (accessed March 4, 2023).
[54] Hyperparameter tuning for machine learning models. Jeremy Jordan 2017. https://www.jeremyjordan.me/hyperparameter-tuning/ (accessed March 4, 2023).
[55] Gaikwad C. Hyperparameter Tuning for Tree Models. ChiGa 2021. https://medium.com/chinmaygaikwad/hyperparameter-tuning-for-tree-models-f99a66446742 (accessed March 4, 2023).
[56] Jakkula V. Tutorial on Support Vector Machine (SVM) n.d.
[57] Gandhi R. Support Vector Machine — Introduction to Machine Learning Algorithms. Medium 2018. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (accessed March 4, 2023).
[58] Dridi S. Unsupervised Learning - A Systematic Literature Review. 2021. https://doi.org/10.13140/RG.2.2.16963.12323.
[59] Ji J, Pang W, Zhou C, Han X, Wang Z. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowl-Based Syst 2012;30:129–35. https://doi.org/10.1016/j.knosys.2012.01.006.
[60] Xie J, Jiang S, Xie W, Gao X. An Efficient Global K-means Clustering Algorithm. JCP 2011;6:271–9. https://doi.org/10.4304/jcp.6.2.271-279.
[61] Huang Z, Ng MK. A Note on K-modes Clustering. J Classif 2003;20:257–61. https://doi.org/10.1007/s00357-003-0014-4.
[62] Jia Z, Song L. Weighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient. Math Probl Eng 2020;2020:e5143797. https://doi.org/10.1155/2020/5143797.
[63] Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Min Knowl Discov 1998;2:283–304. https://doi.org/10.1023/A:1009769707641.

[64] Chen S, Wang W, van Zuylen H. A comparison of outlier detection algorithms for ITS data. Expert Syst Appl 2010;37:1169–78. https://doi.org/10.1016/j.eswa.2009.06.008.
[65] Liu FT, Ting KM, Zhou Z-H. Isolation Forest. 2008 Eighth IEEE Int. Conf. Data Min., 2008, p. 413–22. https://doi.org/10.1109/ICDM.2008.17.
[66] Chatterjee I, Zhou M, Abusorrah A, Sedraoui K, Alabdulwahab A. Statistics-Based Outlier Detection and Correction Method for Amazon Customer Reviews. Entropy 2021;23:1645. https://doi.org/10.3390/e23121645.
[67] Dang TT, Ngan HYT, Liu W. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. 2015 IEEE Int. Conf. Digit. Signal Process. DSP, 2015, p. 507–10. https://doi.org/10.1109/ICDSP.2015.7251924.
[68] Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: identifying density-based local outliers. ACM SIGMOD Rec 2000;29:93–104. https://doi.org/10.1145/335191.335388.
[69] Performance Metrics in Machine Learning [Complete Guide] - neptune.ai n.d. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide (accessed March 4, 2023).
[70] Performance Metrics for Classification problems in Machine Learning | by Mohammed Sunasra | Medium n.d. https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b (accessed March 4, 2023).
[71] Classification: courbe ROC et AUC | Machine Learning. Google Dev n.d. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=fr (accessed March 4, 2023).
[72] Al-Moubaraki AH, Obot IB. Corrosion challenges in petroleum refinery operations: Sources, mechanisms, mitigation, and future outlook. J Saudi Chem Soc 2021;25:101370. https://doi.org/10.1016/j.jscs.2021.101370.
[73] Pedeferri (Deceased) P. Corrosion in Petrochemical Plant. In: Pedeferri P, editor. Corros. Sci. Eng., Cham: Springer International Publishing; 2018, p. 549–74. https://doi.org/10.1007/978-3-319-97625-9_24.
[74] Farmani R, Kakoudakis K, Behzadian K, Butler D. Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors. Procedia Eng 2017;186:117–26. https://doi.org/10.1016/j.proeng.2017.03.217.
[75] Huang H, Tian J, Zhang G, Pan Z. The corrosion of X52 steel at an elbow of loop system based on array electrode technology. Mater Chem Phys 2016;181:312–20. https://doi.org/10.1016/j.matchemphys.2016.06.064.
[76] Si X, Zhou K, Zhang R, Lin T, Xu Q. Experimental and numerical investigation of flow- accelerated corrosion in 90° elbow. Mater Res Express 2018;5:066536. https://doi.org/10.1088/2053-1591/aacc34.
[77] Louie DK. Handbook of sulfuric acid manufacturing / by Douglas K. Louie. Thornhill, Ont.: DKL Engineering; 2005.
[78] Zhang C, Liu C, Zhang X, Almpanidis G. An up-to-date comparison of state-of-the-art classification algorithms. Expert Syst Appl 2017;82:128.
[79] Schapire RE. The Boosting Approach to Machine Learning: An Overview. In: Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B, editors.

Nonlinear Estim. Classif., vol. 171, New York, NY: Springer New York; 2003, p. 149–71. https://doi.org/10.1007/978-0-387-21579-2_9.
[80] Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artif Intell Rev 2006;26:159–90. https://doi.org/10.1007/s10462-007-9052-3.
[81] Salzberg SL. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data Min Knowl Discov 1997;1:317–28. https://doi.org/10.1023/A:1009752403260.
[82] Günther WA, Rezazade Mehrizi MH, Huysman M, Feldberg F. Debating big data: A literature review on realizing value from big data. J Strateg Inf Syst 2017;26:191–209. https://doi.org/10.1016/j.jsis.2017.07.003.
[83] Sharma R, Mithas S, Kankanhalli A. Transforming decision-making processes: A research agenda for understanding the impact of business analytics on organisations. Eur J Inf Syst 2014;23:433–41. https://doi.org/10.1057/ejis.2014.17.

ANNEXES

Un exemple de fluide circulant dans le système de tuyauterie de la raffinerie est illustré au niveau du tableau ci-dessous :

Annexe 1 Exemple de service dans la raffinerie

Service	Description
FRACOVHD	Frac Overhead
CRUDOVHD	Crude Overhead
GASOILEG	Gasoil léger
FLUEGCAT	Gaz de chemine
LIGHTHC	GPL/Naphta
ACIDGAS	Gaz acid
GPL	Gaz de pétrole liquéfié
GASOILOU	Gasoil lourd
KÉROSÈNE	Kerosene
C3C4H2S	Propane/butane/sulfure d'hydrogène
H2H2S	Hydrogène/sulfure d'hydrogène
EAUACIDE	Eau acide
BRUT	Brut
NAPHTA	Naphta
CATALYST	Catalyst
UTILIH2O	Eau d'utilité
FUELGAS	Réseau de gaz combustible

Les principales librairies utilisées dans ce projet pour le développement des modelés d'AA sont présentées au niveau de l'Annexe 2.

```
# Imported Libraries
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import tensorflow as tf
import matplotlib.pyplot as plt
import seaborn as sns

# Classifier Libraries
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

# Other Libraries
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline
from imblearn.metrics import classification_report_imbalanced
from sklearn.metrics import precision_score, recall_score, f1_score, roc_auc_score, accuracy_score, classification_report
from sklearn.preprocessing import StandardScaler,RobustScaler

# clustering
from sklearn.cluster import KMeans
from kmodes.kmodes import KModes
from kmodes.kprototypes import KPrototypes

# outlier Detection
from pyod.models.knn import KNN
from pyod.models.lof import LOF
from sklearn.ensemble import IsolationForest
```

Annexe 2 Exemple des librairies (Python)