جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

République Algérienne Démocratique et Populaire
Université Abderrahmane MIRA de Béjaia
Faculté des Sciences Exactes

Département de Recherche Opérationnelle

Mémoire Présenté pour L'obtention du Diplôme de Master
en Mathématiques Appliquées

Spécialité : Sciences de données et aide à la decision

Évaluation des risques de corrosion de l'oléoduc Haoud El Hamra – Béjaïa par des méthodes d'apprentissage automatique

Présenté par :
IKHLEF Anis
ZIDI Ziane

Défendu le 29/06/2025, devant le jury composé de :

Mlle Z. AOUDIA	M.A. classe A	Présidente du jury	UAMB - Béjaïa
Mr B. BRAHMI	M.A. classe A	Encadrant	UAMB - Béjaïa
Mr R. LAGGOUNE	M.C. classe A	Co-encadrant	UAMB - Béjaïa
Mr R. SAHLI	Doctorant	Examinateur	UAMB - Béjaïa
Mr L. BOUZIDI	M.C. classe A	Examinateur	UAMB - Béjaïa
Mr S. TOUATI	M.C. classe B	Examinateur	UAMB - Béjaïa

Année Universitaire 2024 – 2025

Remerciements

Je tiens à exprimer ma sincère reconnaissance à toutes les personnes qui ont contribué à l'élaboration de ce mémoire.

Je souhaite également remercier mes parents pour les efforts matériels et les sacrifices qu'ils ont consentis au cours de mon parcours.

J'adresse mes remerciements les plus chaleureux à Monsieur BRAHMI, mon encadrant, pour son encadrement rigoureux, ses conseils avisés et son accompagnement tout au long de ce mémoire.

Je remercie tout particulièrement Monsieur ASLI, chef de département, pour son aide précieuse, sa disponibilité et ses conseils tout au long de ce projet.

Mes remerciements vont également à Monsieur AMROUN Kamal, pour ses orientations pertinentes et son appui professionnel qui ont enrichi notre réflexion durant la réalisation de ce travail.

Je tiens aussi à remercier Monsieur YESSAD Abdelhanine, spécialiste en protection cathodique chez Sonatrach, pour sa disponibilité, ses explications claires sur le fonctionnement de la protection cathodique, ainsi que pour les données précieuses qu'il nous a fournies.

Je n'oublie pas mes camarades Hakim, Cylia et Elissa pour les échanges fructueux et leur soutien constant.

Enfin, j'exprime ma gratitude à ma famille et mes proches pour leur présence et leurs encouragements.

A. IKHLEF

Je remercie tout d'abord Dieu Tout-Puissant pour m'avoir accordé la force et la patience nécessaires à la réalisation de ce travail.

Je tiens à exprimer ma plus profonde gratitude à ma famille, qui a été le pilier fondamental de ma réussite. À mes parents, pour leur amour inconditionnel, leurs sacrifices, leur patience et leur soutien moral constant. À mes frères et sœurs, pour leur présence, leurs encouragements et leur confiance, qui m'ont toujours porté dans les moments les plus exigeants.
Ce mémoire leur est dédié, en témoignage de mon attachement et de ma reconnaissance infinie.

Mes sincères remerciements vont à mes amis, ainsi que mes enseignants (B.BRAHMI, L.ASLI, N.khimoum) et camarades de promotion, pour leur aide, leurs échanges enrichissants et leur soutien tout au long de cette formation.

Je tiens également à remercier YESSAD Abdelhanine, DJERROUD Takfarinas, Sid-Ali OUADI et MEZIANE Farid de SONATRACH pour leur accompagnement, leur disponibilité et leurs conseils techniques précieux.

Z. ZIDI

Table des matières

Table des figures

Liste d'abréviations et notations

API	American Petroleum Institute
ML	Machine Learning
IA	Intelligence artificielle
LSTM	Long Short-Term Memory
PVC	Polychlorure de vinyle
ASME	American Society of Mechanical Engineers
DNVGL	Det Norske Veritas – Germanischer Lloyd
UT	Ultrasonic Testing
RT	Radiographic Testing
ILI	In-Line Inspection
PIG	Pipeline Inspection Gauges
SVM	Support Vector Machine
KNN	K-Nearest Neighbors (KNN)
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GMM	Gaussian Mixture Model
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative
MSE	Mean Squared Error
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
DBI	Davies–Bouldin Index
CH	Calinski-Harabasz Index
MFL	Magnetic Flux Leakage
USD	United States Dollar
CNPC	China National Petroleum Corporation.

Introduction générale

Le transport des hydrocarbures à grande échelle repose sur un réseau de pipelines terrestres et sous-marins, essentiels pour assurer l'approvisionnement continu en pétrole brut, gaz naturel et autres produits. Ces infrastructures, efficaces et économiques, sont soumises à de nombreuses défaillances que soient environnementales ou mécaniques. Parmi celles-ci, on trouve la corrosion au sommet des défaillances les plus critiques. La corrosion est responsable de près de 75% des dégradations observées, cela implique des pertes économiques majeures et d'énormes risques sur l'environnement [34].

Parmi les méthodes de contrôle non destructif les plus utilisées dans le secteur, l'inspection par fuite de flux magnétique est largement utilisée pour détecter les pertes de métal et anomalies sur les parois internes et externes des pipelines. Toutefois, malgré sa précision, cette dernière reste limitée par son approche rétrospective, cette méthode est incapable de prédire de manière fiable les risques futurs [8]. Dans ce contexte, les récentes avancées dans le domaine de l'IA, et plus exactement le *machine learning*, offrent des solutions innovantes pour dépasser ces limitations.

Le *clustering*, ou le regroupement non supervisé, constitue une stratégie pertinente pour analyser les données issues des inspections MFL. Cette méthode permet de segmenter automatiquement les défauts observés selon leur niveau de gravité, en exploitant des algorithmes telles que K-means [41], DBSCAN, le clustering hiérarchique agglomératif, et bien d'autres algorithmes. Ces méthodes ont démontré leur efficacité pour extraire des structures cachées dans des données complexes et volumineuses, typiques des systèmes industriels modernes [44].

Le présent travail s'inscrit dans cette dynamique. Il exploite un jeu de données issu d'une inspection MFL réalisée en 2010 par l'entreprise ROSEN Group sur une ligne de pipeline reliant Haoud El Hamra à Béjaïa (actuellement la ligne est abandonnée), pour le compte de SONATRACH. Le rapport d'inspection, basé sur le modèle RoCorr MFL-A, fournit 14 846 observations caractérisées par des mesures précises de profondeur, de longueur et de largeur de corrosion. L'objectif de ce mémoire est triple : classifier les types de défauts observés sur les parois des pipelines, appliquer des techniques de *clustering* pour évaluer automatiquement la sévérité de la corrosion, et cartographier les résultats pour identifier les zones critiques.

En apportant une nouvelle approche prédictive fondée sur l'IA, ce travail ambitionne de transformer les données d'inspection MFL en outils d'aide à la décision, permettant ainsi d'optimiser les interventions de maintenance, de prévenir les éclatements, et de réduire les risques environnementaux.

Le mémoire s'articule autour de trois chapitres principaux, chacun traitant une dimension complémentaire du sujet :

— **Chapitre 1 : Généralités sur les défaillances et les inspections des pipelines**
Ce chapitre présente les différents types de pipelines, les normes de conception, les défaillances majeures (notamment la corrosion), ainsi que les méthodes d'inspection, avec un accent particulier sur la technologie MFL.

— **Chapitre 2 : Apprentissage automatique**
Ce chapitre introduit les concepts fondamentaux du *machine learning*, les principales catégories d'apprentissage (supervisé, non supervisé, par renforcement), les algorithmes de *clustering*, ainsi que les indicateurs d'évaluation.

— **Chapitre 3 : Étude de cas — Oléoduc Haoud El Hamra – Béjaïa**
Cette partie est consacrée à l'analyse du jeu de données MFL fourni par SONATRACH. Elle décrit le prétraitement des données, l'application des algorithmes de *clustering* (K-means, DBSCAN, GMM, etc.) et l'interprétation des résultats.

— **Annexes**
Les annexes détaillent les environnements de développement, les bibliothèques logicielles et les outils informatiques utilisés pour la mise en œuvre des modèles d'analyse.

1

Généralités sur les défaillances et les inspections des pipelines

Sommaire

Introduction

De nos jours, les pipelines sont devenus indispensables pour le monde. Ils sont utilisés pour le transport du pétrole brut, gaz naturel, produits pétroliers raffinés et d'autres liquides dans le domaine industriel. C'est pourquoi une telle évolution a conduit à l'émergence de défis techniques qui se profilent comme de plus en plus problématiques : ces installations doivent être fiables, solides et, surtout sûres.

Presque tous les types de pipelines sont construits à partir de matériaux différents, en fonction de la pression, du produit transporté, des conditions météorologiques et des risques de corrosion. Parmi les principaux matériaux, on trouve l'acier, le cuivre, le béton armé, le PVC et la fonte. Par exemple, l'acier au carbone est le plus utilisé, notamment dans l'industrie pétrolière, car il est abondant, relativement moins cher et résistant à la corrosion. Ce matériau offre un bon équilibre entre fiabilité et durabilité.

Un système de pipeline comprend divers composants indispensables qui collaborent pour garantir le transfert efficace et sûr des fluides. Il s'agit principalement de la conduite elle-même, constituée de tuyaux en acier ou en plastique, selon le produit transporté. On utilise également des pompes ou des compresseurs de stations pour assurer la pression et faire circuler le fluide, des vannes pour réguler ou stopper le courant, des détecteurs pour suivre en direct des mesures comme la pression ou les fuites, ainsi qu'un mécanisme de défense contre l'usure due à la corrosion afin d'éviter la détérioration des matériaux.

1.1 Définition d'un pipeline

Un pipeline est un système de conduites utilisé pour transporter des liquides ou des gaz sur de longues distances, généralement sous terre ou sous l'eau, depuis les sites de production jusqu'aux raffineries, aux centres de distribution ou aux utilisateurs finaux [18].

1.2 Utilisation des pipelines

Les pipelines sont principalement des infrastructures enfouies. Par conséquent, la population n'est pas nécessairement informée que de nombreux fluides sont acheminés par ces moyens. Néanmoins, une grande partie des pays possède un réseau de pipelines très dense.

— Pour le gaz naturel, on parle de *gazoduc*,
— Pour le pétrole, on parle d'*oléoduc*,
— Pour l'eau industrielle, il s'agit de *conduite*. Le terme *aqueduc* est plutôt réservé aux ouvrages maçonnés, avec écoulement libre d'eau,
— Pour l'eau salée, on utilise le mot *saumoduc*,
— Pour l'oxygène, on utilise le terme *oxygénoduc* ou *oxyduc*,
— Pour l'hydrogène, on utilise le terme *hydrogénoduc*.

En général, le suffixe d'origine latine « ductus » signifie « conduite », ce qui permet de caractériser le terme français d'une conduite spécialisée pour le transport d'un certain type de produit.

L'efficacité des pipelines repose principalement sur la simplicité de ce mode de transport. D'une part, ils sont sûrs, rentables et représentent un avantage dans le transport de produits délicats tels que le pétrole ou le gaz. Effectivement, la circulation de fluides dangereux est plus sûre par ce moyen que par transport terrestre. Il y a également un risque réduit de dommages ou de vol [45].

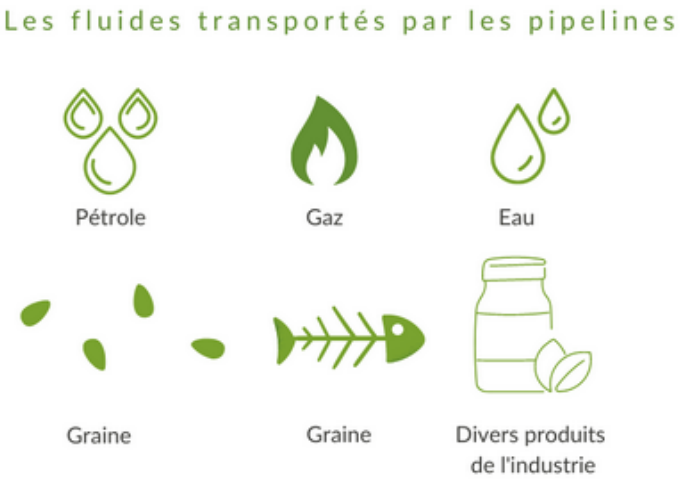

FIGURE 1.1 – Les fluides transportés par les pipelines.

1.3 Comment fonctionne un pipeline ?

Le fonctionnement des pipelines peut varier en fonction du type de fluide qu'ils acheminent.

Par exemple, pour les pipelines qui acheminent le pétrole brut, une conduite de petit diamètre peut servir à extraire le produit. Ensuite, le produit est guidé vers une installation de collecte. Par la suite, le produit est acheminé à travers des conduites d'alimentation de plus grand diamètre, notamment pour le transport sur de longues distances. Afin d'assurer le déplacement fluide du liquide à travers le tuyau, des pompes performantes sont utilisées pour générer le mouvement.

Le trajet du gaz est également jalonné de diverses phases lors de son transport. Initialement, le gaz naturel provenant d'un gisement est acheminé via un pipeline en acier. On utilise par la suite des compresseurs à haute pression, ce qui entraîne des variations de pression au sein du réseau. Ainsi, le gaz peut aisément progresser jusqu'à une usine de distribution. Des canalisations plus petites sont finalement mises en œuvre pour le transport jusqu'aux consommateurs [19].

1.4 Normes de conception

Les normes de conception des pipelines sont cruciales pour assurer la sécurité, l'efficience et la rentabilité économique des réseaux de pipelines. Ces standards offrent des orientations pour la conception, la construction et la maintenance des pipelines, prenant en compte divers éléments comme le choix des matériaux, l'épaisseur des parois et le comportement mécanique sous diverses conditions. Dans la conception des pipelines offshore, les normes américaines et norvégiennes ont une importance considérable, alors que pour les pipelines terrestres, c'est surtout les normes ASME et européennes qui sont largement adoptées. L'adoption des technologies numériques modifie aussi le processus de normalisation dans le secteur des pipelines.

1.4.1 Normes relatives aux pipelines offshore

Les normes américaines ASME B31.8 et norvégiennes DNVGL-ST-F101 sont couramment appliquées dans la construction des pipelines de transport de gaz en mer. Ces standards aident à définir l'épaisseur idéale des parois en tenant compte de la qualité de l'acier et de la profondeur de l'eau, assurant ainsi la solidité du pipeline et sa viabilité économique [2].

Les normes API 1111 et DNV-ST-F101 traitent principalement des problématiques de conception mécanique comme l'effondrement, la diffusion du gauchissement et l'arrêt, qui sont cruciales lors de la mise en place des pipelines en mer [25].

FIGURE 1.2 – Pipeline offshore.

1.4.2 Normes relatives aux pipelines terrestres

Les normes ASME B978-0-323-88663-5.8 et B978-0-323-88663-5.4 sont largement adoptées à l'échelle mondiale pour les pipelines terrestres d'hydrocarbures. La norme européenne EN 1594 est aussi applicable aux gazoducs à haute pression [58].
La série ASME B31, et plus particulièrement les sections B31.1, B31.3, B31.4 et B31.8, englobe différents éléments des systèmes de conduites, allant de la conduite d'alimentation et de traitement des hydrocarbures liquides à la transmission de gaz [56].

FIGURE 1.3 – Pipeline terrestre.

1.4.3 Transformation numérique dans les normes

L'élargissement de l'efficacité du traitement de l'information dans l'industrie des pipelines est favorisé par le progrès des standards de lecture automatique et des jeux d'étiquettes standardisés. Les standards numériques favorisent l'élaboration de graphes de connaissances et stimulent l'innovation [30]. Bien que les standards traditionnels offrent une base robuste pour la création des pipelines, l'incorporation de technologies numériques pave la route vers des méthodes plus performantes et novatrices. Cette transition vers le numérique est essentielle pour l'évolution à long terme de l'industrie des pipelines, puisqu'elle procure de nouvelles possibilités d'optimisation et de gestion du savoir.

1.5 Types de défaillances

Les vulnérabilités des infrastructures d'ingénierie, notamment celles liées aux plateformes en mer et aux pipelines, peuvent être réparties en plusieurs catégories principales : mécaniques, de corrosion, opérationnelles, dues à des opérations de tiers ou encore liées à des aléas naturels. Chacune de ces causes pose des problèmes particuliers liés à la sécurité, la fiabilité et l'efficacité des installations. En revanche, les pipelines peuvent être sujets à divers types de défaillances dont les origines et les effets peuvent varier considérablement. Il est essentiel de comprendre ces vulnérabilités en détail pour établir une gestion des risques adéquate et élaborer des stratégies de maintenance évolutives. Cela contribue à éviter les défaillances, prolonger la longévité des équipements et diminuer les dépenses opérationnelles.

1.5.1 Défaillances mécaniques

Les pannes mécaniques se produisent généralement en raison de vices matériels ou de dégâts causés par l'apparition de fissures, souvent exacerbées par des conditions environnementales adverses telles que les fluctuations de température ou les surcharges. Ces pannes pourraient sérieusement perturber le fonctionnement des équipements, entraînant des arrêts de production, des pertes financières et des dangers pour la sécurité. Il est donc primordial de mener une étude détaillée des raisons de ces pannes afin d'identifier leur source précise et d'instaurer des actions correctives et préventives adéquates [57].

1.5.2 Défaillances de corrosion

La corrosion est l'un des facteurs majeurs de défaillance des structures techniques, en particulier dans les milieux marins où elle est à l'origine de près de 75% des dommages constatés. Les installations en mer, telles que les plates-formes pétrolières et les conduites sous-marines, sont particulièrement exposées aux risques de corrosion en raison de leur contact permanent avec l'eau salée, l'humidité et des conditions météorologiques extrêmes. Outre ses répercussions techniques, la corrosion a aussi des conséquences économiques considérables : dans les nations industrialisées, on estime que les pertes dues à la corrosion représentent entre 4 et 5% du produit national brut. Cela met en évidence la nécessité primordiale d'établir des stratégies de prévention, de contrôle et d'entretien spécifiques afin de minimiser ses impacts et d'allonger la longévité des infrastructures essentielles [48, 46].

1.5.3 Défaillances opérationnelles

Les défaillances qui se produisent pendant l'utilisation sont généralement associées à une détérioration graduelle de la solidité structurelle des appareils, en particulier dues à la fatigue, aux sollicitations mécaniques récurrentes et à l'exposition durable à des conditions environnementales hostiles. Si ces éléments ne sont pas adéquatement contrôlés, ils risquent d'affecter la sûreté et l'efficacité des structures, notamment dans les contextes offshore où les contraintes sont permanentes. Pour gérer ces risques, il est essentiel de mettre en place des stratégies de maintenance efficaces et préventives pour augmenter la longévité des installations, renforcer la fiabilité opérationnelle et minimiser les dépenses liées aux interruptions imprévues [46].

1.5.4 Défaillances liées aux activités tierces

Les pipelines subissent souvent des défaillances dues à des facteurs externes, notamment les dommages infligés par des actions de tiers comme le creusement de tranchées, l'ancrage de navires ou les travaux de construction à proximité. Ces actions, fréquemment menées sans concertation avec les gestionnaires des infrastructures, risquent de provoquer des cassures ou des dégradations majeures des tuyaux, mettant par conséquent en péril leur solidité et leur sûreté. Ces événements soulignent le besoin crucial d'une gestion stricte des risques externes, y compris l'instauration de dispositifs de protection plus robustes, un contrôle intensifié des régions sensibles et une communication améliorée entre les acteurs intervenant dans les mêmes zones géographiques [5].

1.5.5 Défaillance due aux risques naturels

Les risques naturels représentent également une source possible de défaillances structurelles, touchant tant les pipelines que les installations offshore. Même si les informations exactes sur leur occurrence et leur influence demeurent restreintes dans les publications existantes, on ne saurait sous-estimer leur impact. Des événements comme les tremblements de terre, les coulées de boue, les tempêtes en mer ou les inondations ont le potentiel de menacer la solidité des infrastructures, entraîner des mouvements, des cassures ou des affaissements, et par conséquent provoquer des arrêts de fonctionnement ou des incidents majeurs. Il est donc primordial d'intégrer ces phénomènes naturels dans les plans de conception, de contrôle et d'entretien pour améliorer la robustesse des infrastructures en réponse à des incidents inattendus mais susceptibles d'être dévastateurs.

Bien qu'il soit essentiel de se concentrer sur les défaillances mécaniques et de corrosion, il est également essentiel de prendre en compte le contexte plus large des pratiques opérationnelles et des influences externes qui peuvent exacerber ces problèmes. Relever ces défis multidimensionnels nécessite une approche globale de la conception technique et de la maintenance.

1.6 Les Causes de défaillances

Les défaillances des pipelines peuvent survenir pour de nombreuses raisons, souvent liées à l'environnement, aux matériaux utilisés, aux méthodes de construction ou encore aux conditions

dans lesquelles ils sont exploités. Voici un aperçu des principales causes identifiées :

1.6.1 Corrosion

La corrosion est une interaction physico-chimique entre un matériau métallique et son milieu environnant entraînant des modifications dans les propriétés du métal et pouvant conduire à une dégradation significative de la fonction du métal, du milieu environnant ou du système technique dont ils font partie.

FIGURE 1.4 – Corrosion interne

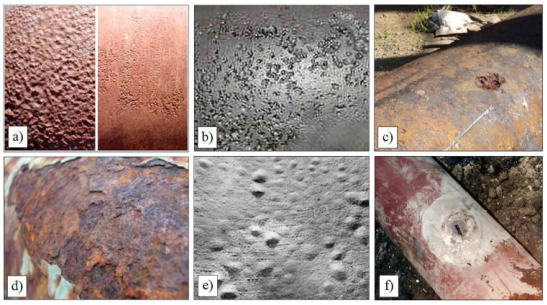

FIGURE 1.5 – Corrosion externe

Corrosion interne

La corrosion interne est plus fréquente dans les pipelines de taille moyenne et grande, en particulier dans les environnements sous-marins tels que le golfe du Mexique, où la présence d'eau et de gaz acides favorise le processus de corrosion [42]. Dans les gazoducs, la corrosion interne est responsable d'un grand nombre d'incidents, ce qui indique qu'il est nécessaire d'avoir un contrôle efficace de la corrosion interne [7].

Corrosion externe

L'un des principaux motifs de défaillance des conduites est la corrosion externe, en particulier dans les conduites de petite taille et celles situées à proximité des plateformes. Des éléments environnementaux comme la qualité du sol, l'humidité et les courants errants jouent un rôle notable dans la corrosion externe des systèmes de distribution d'eau [21]. La corrosion externe dans les pipelines et gazoducs est à l'origine d'un nombre significatif d'accidents, mettant en évidence le rôle crucial des revêtements de protection et de la gestion environnementale.

1.6.2 Défauts de fabrication

Les défauts de fabrication et de soudure sont des problèmes critiques en construction des structures métalliques parce qu'ils peuvent se résumer à l'échec significatif due aux défauts et à la fatigue. Les défauts sont généralement dus à des techniques de soudure incorrectes, emploi de matières non-conformes standards et un contrôl de qualité inadéquat et ceci peut sacrifier l'intégrité des joints retenus par soudure. Les conséquences de tels défauts sont importantes et se reflètent en termes de durée de vie et de sécurité des structures d'art. Les sections ci-dessous

font le point sur les types spécifiques de défauts, leurs origines et leur incidence sur la fatigue métallique.

Types de défauts de soudage

Matériaux défectueux : De fortes concentrations d'azote, d'hydrogène et d'oxygène dans les matériaux de soudage peuvent causer des imperfections comme la porosité et les fissures dans les soudures [62].
Défauts de surface : Les cavités, l'insuffisance de fusion et les impuretés résiduelles sont des irrégularités superficielles courantes qui affectent significativement la qualité des soudures.

1.6.3 Défaillance mécaniques externes

Les dommages mécaniques externes des pipelines, surtout les dommages humains, comme les dommages provenant d'activités telles que des travaux publics et des excavations incontrôlées, offrent des dangers importants pour l'intégrité opérationnelle. Ces dommages généralement tirent leur origine des impacts en raison d'appareils, d'objets chutant et d'ancrages de navires, ce qui rend l'appel à des stratégies solides d'évaluation et de gestion. Les secteurs ci-dessous englobent les clefs de ce problème.

Types de dommages mécaniques externes

Dommages causés par impact : ils sont généralement causés par des ancrages et des objets tombés, ce qui entraîne la formation de bosses et d'entailles qui compromettent l'intégrité du pipeline.

Interférence de tiers : des activités telles que les engins de terrassement peuvent provoquer des dommages mécaniques importants, qui ont toujours été l'une des principales causes de défaillance des pipelines.

Déformation du sol : les événements naturels tels que le mouvement des pentes peuvent aggraver les dommages, en particulier dans les environnements éloignés où l'accès pour les réparations est limité.

1.6.4 Pression excessive

Des risques significatifs pour l'intégrité des infrastructures sont présentés par une pression excessive et des modifications brusques dans les systèmes de tuyauterie, principalement dues aux effets de coups de bélier. Les coups de bélier se produisent lorsque la vitesse du liquide change brusquement, généralement en raison d'une fermeture rapide des vannes ou de l'arrêt de la pompe. Cela entraîne des pics de pression qui peuvent endommager les conduites. Pour des stratégies de gestion et d'atténuation efficaces, il est crucial de saisir ces dynamiques.

Causes des coups de bélier

Fermetures soudaines de valves : La fermeture rapide des vannes génère des ondes de haute pression, qui peuvent dépasser les limites de sécurité, entraînant une défaillance potentielle du pipeline [63].

Arrêt de la pompe : L'arrêt brutal du fonctionnement de la pompe peut créer des fluctuations de pression extrêmes, entraînant des pics de pression positifs et négatifs qui menacent l'intégrité du pipeline [67].

Effets de la surpression et de la dépression

Surpression : les surpressions peuvent entraîner des défaillances catastrophiques, notamment des ruptures de canalisations et des dommages structuraux [61].
Sous-pression : À l'inverse, une chute de pression peut provoquer des conditions de vide, entraînant l'effondrement des canalisations en raison de la formation de cavités de vapeur, puis leur implosion [36].

1.6.5 Causes environnementales

Les conditions environnementales ont un impact considérable sur les déplacements du sol, pouvant provoquer divers dangers tels que des coulées de boue, des tassements ou encore des phénomènes météorologiques extrêmes qui peuvent nuire aux infrastructures comme les pipelines. Pour assurer la sécurité et l'intégrité des réseaux de canalisations, il est crucial de saisir ces éléments. Les parties suivantes détaillent les facteurs environnementaux majeurs et leurs effets sur les déplacements du sol.

Les mouvements du sol et leurs types

Glissements de sol : à cause de fortes pluies ou d'une sismicité, les glissements de sol peuvent provoquer un grand mouvement du sol, compromettant l'équilibre du gazoduc.
Affaissement : Il s'agit d'un affaissement graduel du sol susceptible d'être le résultat d'action naturelle ou d'action de l'homme, entraînant les défaillances structurelles des canalisations équipées dans le sol.
Tremblement de terre : les séismes peuvent provoquer des mouvements de faille et des secousses du sol, engendrant de graves déformations susceptibles d'affaiblir l'intégrité du pipeline.

1.6.6 Conditions météorologiques extrêmes

Cycles gel/dégel : les cycles peuvent entraîner une extension et une retraite du sol, changer l'interaction sol-tuyau et potentiellement provoquer du préjudice.
Inondations : Un excès d'eau peut dégrader le sol en entourant les canalisations, augmentant ainsi la probabilité d'exposition et de défaillance.
Ouragans : les pluies fortes et les vents turbulents contribueront probablement à rendre les mouvements du sol pires, notament dans les régions littorales où les oléoducs le sont encore plus [4].

1.7 Méthodes d'inspection

Les méthodes d'inspection des pipelines sont essentielles au maintien de l'intégrité des pipelines dans l'industrie pétrolière et gazière. Plusieurs techniques de contrôle non destructif

sont utilisées, notamment le contrôle par ultrasons (UT), qui utilise des ondes sonores à haute fréquence pour détecter les défauts et mesurer l'épaisseur des parois . L'inspection par fuite de flux magnétique (MFL) est une autre méthode qui permet d'identifier la corrosion et les piqûres en mesurant le champ magnétique autour d'un tuyau magnétisé . Les tests radiographiques (RT) utilisent des rayons X ou des rayons gamma pour inspecter la structure interne pour détecter des défauts tels que des fissures et une porosité . De plus, des jauges d'inspection des canalisations (PIG) sont insérées dans les conduites pour détecter les défauts internes et nettoyer les conduites . Les outils d'inspection en ligne (ILI) fournissent des évaluations détaillées de l'état du pipeline, y compris la géométrie et les défauts internes, et sont essentiels pour déterminer les besoins de maintenance . Ensemble, ces méthodes garantissent le fonctionnement sûr et efficace des réseaux de canalisations [55].

Nous allons nous focaliser particulièrement sur la méthode d'inspection par fuite de flux magnétique (MFL), en raison de son efficacité reconnue pour la détection de la corrosion et des piqûres dans les pipelines.

1.7.1 Méthode MFL

La technologie MFL repose sur la magnétisation des parois métalliques des pipelines et la détection des fuites de flux magnétique causées par les défauts. Voici les étapes clés du processus :

— **Magnétisation du Pipeline :**
Un champ magnétique puissant est généré à l'intérieur du pipeline à l'aide d'aimants permanents ou d'électroaimants. Ce champ traverse les parois métalliques, créant un flux magnétique uniforme.

— **Détection des Fuites de Flux Magnétique (MFL) :**
Lorsque le pipeline présente des défauts (corrosion, érosion, piqûres, etc.), le flux magnétique est perturbé. Ces perturbations, appelées fuites de flux magnétique, sont détectées par des capteurs magnétiques haute résolution.

— **Analyse des Données :**
Les données collectées sont transmises à un système d'acquisition et analysées par des logiciels sophistiqués. Cela permet de déterminer la localisation, la taille, la profondeur et la nature des défauts avec une grande précision.

1.7.2 Applications du RoCorr MFL-A Service

— **Inspection des pipelines terrestres et sous-marins :**
Le service RoCorr MFL-A est utilisé pour inspecter les pipelines terrestres et sous-marins, qu'ils transportent du pétrole, du gaz ou d'autres fluides. Il est particulièrement efficace pour détecter et caractériser les pertes de métal internes et externes, ainsi que les dommages mécaniques.

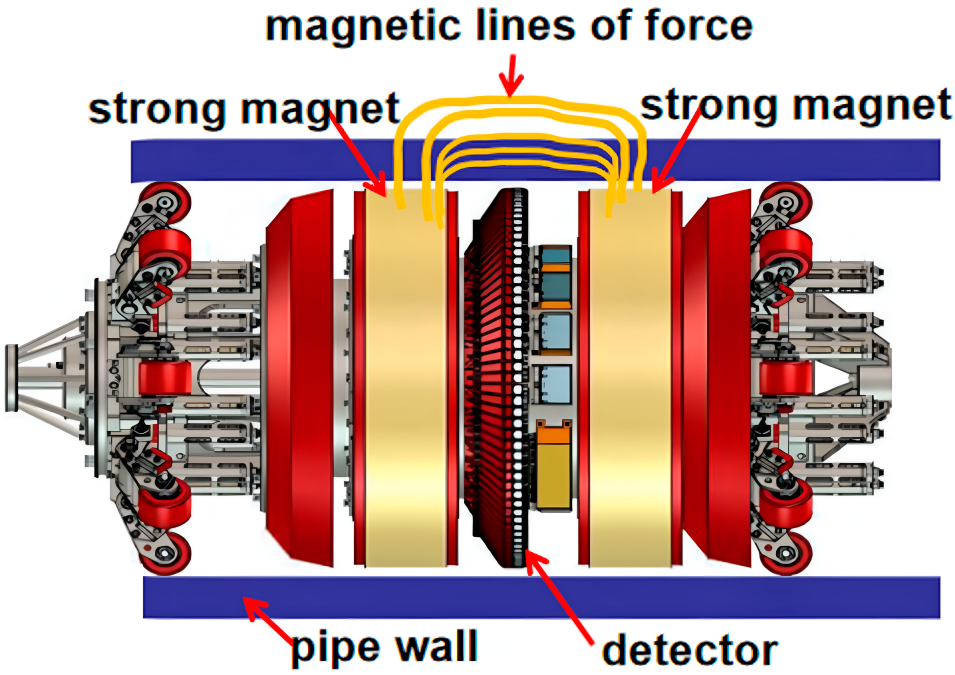

FIGURE 1.6 – Representation de MFL

— **Surveillance des pipelines vieillissants :**
Pour les pipelines anciens ou exposés à des environnements corrosifs, ce service permet une surveillance régulière et proactive, aidant à prévenir les fuites ou les ruptures.
— **Inspection des réservoirs de stockage :**
En plus des pipelines, le RoCorr MFL-A Service peut être adapté pour inspecter les réservoirs de stockage, notamment leurs fonds et parois internes.

1.8 SONATRACH

SONATRACH, la Société nationale algérienne pour le transport et la commercialisation des hydrocarbures, a été créée en décembre 1963. Elle est aujourd'hui considérée comme un pilier de l'économie algérienne, représentant environ 90 % des exportations du pays et près de 60 % des recettes budgétaires. Acteur clé de la sécurité énergétique de l'Europe, SONATRACH s'oriente également vers des solutions d'énergie plus propres, notamment à travers le gaz naturel et l'hydrogène. Un leadership affirmé, mais encore en phase d'adaptation face aux exigences climatiques actuelles.

En 2021, la compagnie produisait environ 2,9 millions de barils équivalent pétrole par jour, dont 59 % sous forme de gaz naturel. Ses deux principaux champs d'exploitation sont Hassi Messaoud (pétrole) et Hassi R'Mel (gaz), représentant ensemble près de 32 % de sa production totale [26]. En 2022, Sonatrach a réalisé un chiffre d'affaires à l'exportation de 60 milliards USD, dégageant un bénéfice net record de plus de 10 milliards USD. Elle figure aujourd'hui parmi les leaders de son secteur en Afrique.

SONATRACH intervient dans l'ensemble de la chaîne pétrolière et gazière : exploration, production, transport, raffinage et commercialisation, aussi bien sur le marché national qu'inter-

national. Elle assure la totalité du raffinage domestique, avec notamment la raffinerie de Skikda, d'une capacité annuelle de 16,5 millions de tonnes, ainsi que d'autres installations comme celle d'Adrar (en co-gestion avec la CNPC) [11].

La compagnie gère également un vaste réseau d'infrastructures, dont des pipelines terrestres reliant les zones de production aux raffineries et aux ports d'exportation. Parmi ces infrastructures stratégiques figure l'oléoduc Haoud El Hamra–Béjaïa, essentiel pour le transport du brut saharien jusqu'au port pétrolier de Béjaïa, l'un des terminaux d'exportation les plus importants du pays.

La direction SONATRACH de Béjaïa joue un rôle central dans l'exploitation, la supervision et la sécurisation de ce terminal. Elle est notamment chargée de garantir l'intégrité des pipelines arrivant au port grâce à des techniques avancées de surveillance et de maintenance. Un des aspects les plus critiques de cette maintenance est la protection cathodique, un procédé électrochimique qui permet de prévenir la corrosion des conduites métalliques enterrées ou immergées [15, 47].

Ce système repose sur l'application d'un courant imposé ou l'utilisation d'anodes sacrificielles afin de polariser les surfaces métalliques à protéger [6]. À Béjaïa, des équipes techniques spécialisées assurent la gestion de ces dispositifs : contrôle des potentiels, inspection des postes de redressement, relevés de terrain, et diagnostic des niveaux de protection. Ces données, indispensables pour évaluer l'état de santé des conduites, ont été gracieusement fournies par la division protection cathodique de la direction SONATRACH Béjaïa.

Mon stage s'est déroulé au sein de la Direction Transport par Canalisation (TRC) de SONATRACH, rattachée à l'activité "Transport par canalisation" .

La Protection cathodique, dans laquelle j'ai effectué mon stage, est une division technique dépendant de la Direction Exploitation de la TRC. Elle joue un rôle essentiel dans la prévention de la corrosion externe des canalisations enterrées

La figure ci-dessous représente l'organigramme global de l'entreprise SONATRACH avec ces différentes branches sur le terain national.

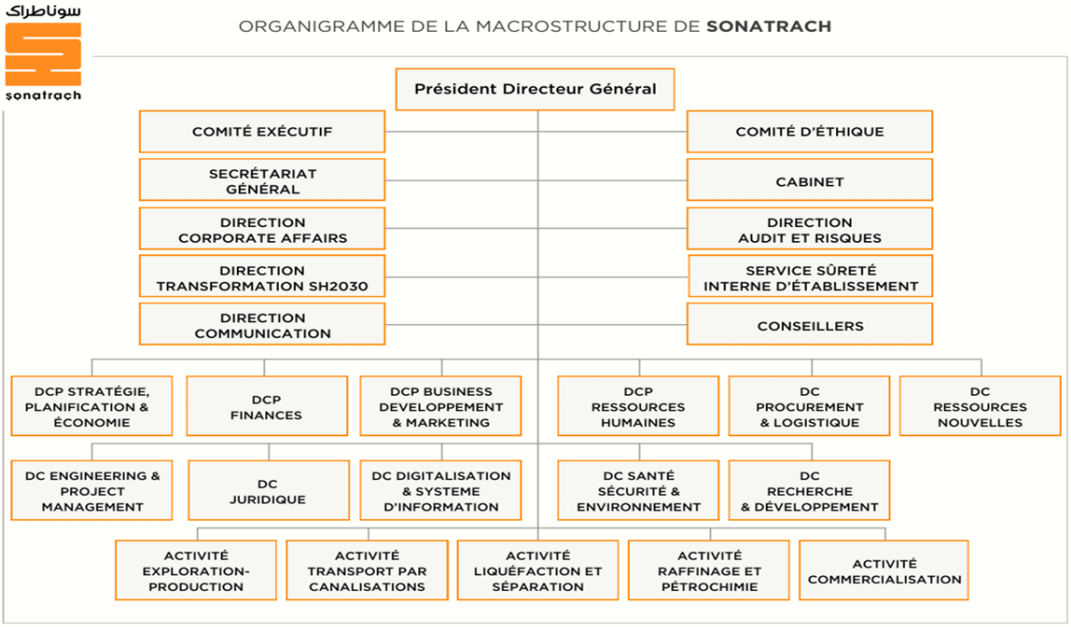

FIGURE 1.7 – Organisme de la SONATRACH

Conclusion

Ce chapitre a permis de poser les bases nécessaires pour comprendre le fonctionnement des pipelines, les principaux types de défaillances auxquelles ils sont exposés, ainsi que les techniques employées pour leur inspection. Notamment, la corrosion s'est avérée être un défi significatif pour la pérennité de ces infrastructures. Nous avons également mis en lumière l'importance cruciale de Sonatrach Béjaïa dans la supervision et la sauvegarde des pipelines, en particulier par le biais de la protection cathodique. Ces aspects constitueront la base pour présenter, dans le chapitre suivant, les techniques d'analyse fondées sur l'apprentissage automatique.

2
Apprentissage automatique

Sommaire

Introduction

Le machine learning est une branche de l'intelligence artificielle qui cherche à élaborer des algorithmes et modèles statistiques aptes à apprendre des données, sans avoir besoin de programmation explicite. Ce processus implique l'interaction et l'analyse des données pour repérer des schémas, réaliser des prévisions ou effectuer des choix et d'améliorer la performance à mesure que de nouvelles informations sont traitées. Actuellement, l'apprentissage automatique joue un rôle central dans diverses applications comme la reconnaissance d'images et de sons,

le traitement du langage naturel ou l'analyse prédictive. Ces domaines mettent en évidence les limites des méthodes traditionnelles de programmation [32].

2.1 Prétraitement des données

Le prétraitement des données est une étape essentielle de l'analyse des données, servant de base à une utilisation efficace des données dans divers domaines, notamment l'apprentissage automatique et les systèmes d'information géographique. Ce processus implique le nettoyage, la transformation et l'organisation des données brutes pour améliorer leur qualité et leur facilité d'utilisation, afin d'aboutir à des informations et décisions précises [27].

Nettoyage des Données

Le nettoyage des données implique l'identification et la correction des inexactitude , la suppression des doublons, et des valeurs aberrantes ainsi que le traitement des données manquantes. ce processus nous permet de réduire le biais ce qui nous offre des résultats plus fiable [54].

Normalisation des données

La normalisation des données est une technique essentielle pour le ML ainsi que pour l'exploration de données. il vise à redimensionner les données pour garantir la comparabilité, en réduisant ainsi les biais. On trouve différentes méthodes de normalisation en fonction du type de données qu'on a, telles que la normalisation par mise à l'échelle Min-Maxe, et Standardisation par score Z [60].

*** Normalisation Min-Max :** La normalisation min-max permet de transformer les valeurs d'une variable pour les ramener dans un intervalle défini, généralement [0, 1]. La formule utilisée est la suivante [27] :

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

où :
— x est la valeur initiale,
— x_{min} est la plus petite valeur de la variable,
— x_{max} est la plus grande valeur de la variable.
Cette méthode conserve les proportions entre les valeurs, mais elle est très sensible aux valeurs aberrantes, qui peuvent fortement affecter la transformation.

*** Standardisation par score Z :** La standardisation transforme les données pour qu'elles aient une moyenne nulle et un écart type égal à un. Cela est particulièrement utile lorsque les données suivent une distribution normale. La formule est la suivante [69] :

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

où :

— x est la valeur initiale.
— μ est la moyenne de la variable.
— σ est l'écart type de la variable.

Contrairement à la normalisation min-max, la standardisation ne ramène pas les valeurs dans une plage spécifique, mais elle est moins sensible aux valeurs extrêmes.

2.2 Notions sur le machine learning

Variables d'entrée :

En apprentissage automatique (ML), une variable d'entrée est une caractéristique spécifique d'un dataset, utilisée comme entrée pour un modèle de ML ou un algorithme d'analyse de données. Elles sont essentielles parce qu'elles fournissent l'information pertinente au modèle pour prendre des décisions précises [65].

Variables de sorties :

La variable cible (variable de sortie) désigne la variable que l'on cherche à prédire ou à classer à partir des variables d'entrée d'un dataset en ML. Son type (catégoriel ou continu) détermine le type d'approche de ML à utiliser (classification ou régression) [13].

Sur-apprentissage (overfitting) :

Le sur-apprentissage est un problème fréquent en ML. Il se produit lorsqu'un modèle performe très bien sur les données d'entraînement mais de manière médiocre sur des données de test, ce qui révèle une mauvaise capacité de généralisation [49].

Techniques pour éviter le sur-apprentissage :

— **Régularisation** : cette technique ajoute une pénalité de complexité à la fonction de perte, empêchant ainsi d'incorporer du bruit dans les données d'entraînement. Elle est notamment plus efficace si le nombre des caractéristiques d'entrée est grand par rapport au nombre d'échantillons d'entraînement [9].
— **Arrêt anticipé** : en surveillant les performances du modèle sur un ensemble de validation, l'entraînement peut être interrompu lorsque les performances commencent à se dégrader, évitant ainsi un sur-apprentissage pendant le processus d'entraînement [12].
— **Élagage** : il est possible d'implémenter des méthodes de pré-élagage et de post-élagage afin de simplifier les modèles. Le pré-élagage prend en charge la croissance du modèle pendant l'entraînement, et le post-élagage supprime toute complexité inutile après l'entraînement [23].
— **Méthodes d'ensemble** : plusieurs techniques telles que l'ensachage et le boosting combinent différents modèles pour améliorer la généralisation et réduire le risque de sur-apprentissage en faisant la moyenne des prédictions [52].

Sous-apprentissage (Underfitting) :
Un sous-apprentissage se produit lorsqu'un modèle d'apprentissage automatique est trop simpliste pour capturer les modèles sous-jacents des données, ce qui entraîne de mauvaises performances à la fois pour l'entraînement et pour les ensembles de données invisibles. Ce phénomène est souvent caractérisé par un biais élevé et une faible variance, ce qui fait que le modèle ne parvient pas à tirer des enseignements adéquats des données d'entraînement [24].

Causes du sous-apprentissage

- **Complexité du modèle :** les modèles simples, tels que la régression linéaire, peuvent ne pas saisir les relations complexes dans les données, ce qui entraîne un sous-apprentissage [29].
- **Entraînement insuffisant :** un temps d'entraînement ou des itérations inadéquats peuvent empêcher le modèle d'apprendre efficacement, comme le montrent les scénarios d'entraînement contradictoire où un entraînement prolongé entraîne une perturbation et un sous-adaptation [39].

Techniques pour éviter le sous-apprentissage :

- **Sélection du modèle :** le choix des modèles ou des méthodes d'ensemble plus complexes peut aider à capturer des modèles complexes dans les données [66].
- **Techniques de régularisation :** L'application de la régularisation permet d'équilibrer la complexité du modèle et d'éviter le sous-apprentissage tout en maintenant les capacités de généralisation [66].

2.3 Pourquoi l'apprentissage automatique ?

L'apprentissage automatique (ML) et la programmation traditionnelle représentent deux approches différentes pour traiter et analyser les données. Contrairement à la programmation traditionnelle qui s'appuie sur des directives manuelles pour accomplir une tâche, l'apprentissage automatique permet aux systèmes d'apprendre des données et de progresser au fil du temps sans avoir besoin d'être spécifiquement codés pour chaque tâche distincte. Cette distinction essentielle donne lieu à diverses applications et méthodes qui soulignent les avantages et les restrictions de chaque démarche. Les trois types de l'apprentissage automatique (apprentissage supervisé, non supervisé et par renforcement) sont destinés à atteindre des objectifs et des méthodologies diverses [32].

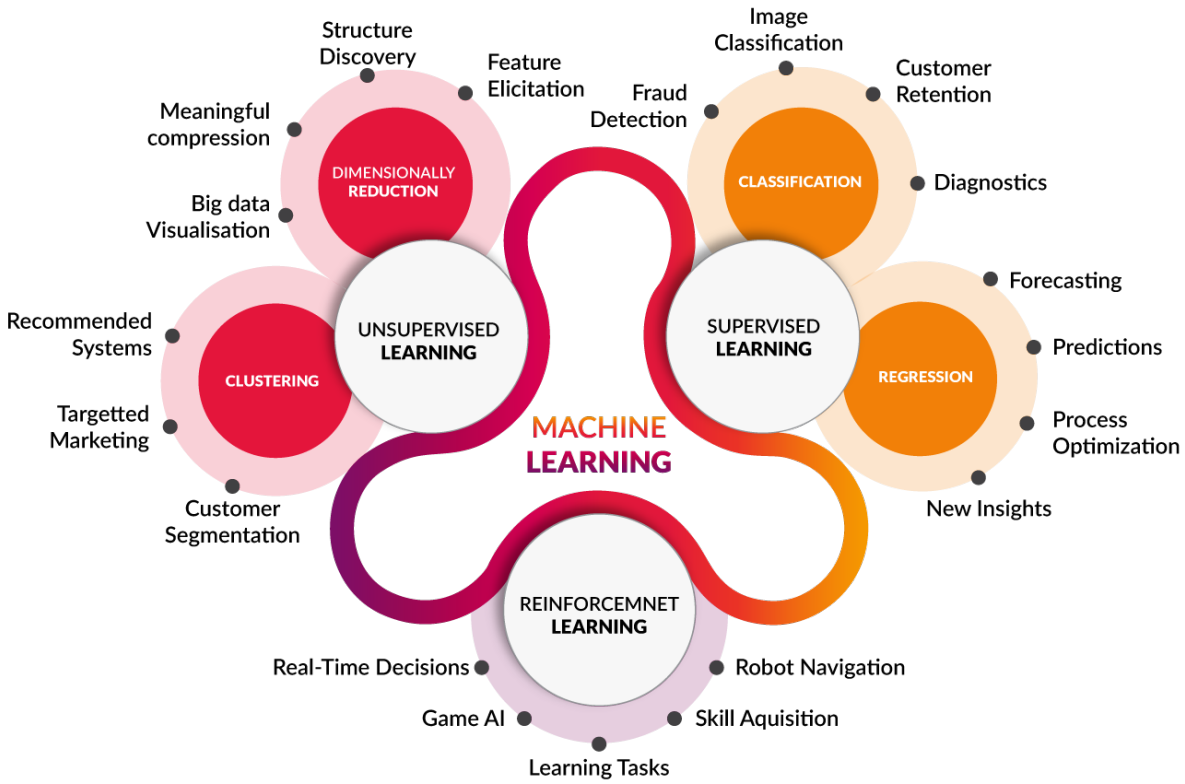

FIGURE 2.1 – Types du machine learning.

2.4 Apprentissage supervisé

On parle d'apprentissage supervisé lorsque l'on dispose de données d'entraînement étiquetées, c'est-à-dire dont on connaît la sortie voulue. En notant les N entrées x_i et les sorties cibles associées y_i, on dispose de l'ensemble de données suivant :

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

L'objectif est d'entraîner le modèle choisi pour qu'il puisse prédire correctement la sortie pour de nouvelles entrées non étiquetées [14].

Les algorithmes d'apprentissage supervisé sont des piliers de l'apprentissage automatique, dans la mesure où ils se basent sur des jeux de données étiquetés pour apprendre à prédire les sorties ou classer les informations. Il s'agit d'une composante essentielle dans des diverses utilisations.

2.4.1 Algorithmes de classification

Les algorithmes de classification sont des méthodes essentiels de l'apprentissage automatique, utilisés pour prédire des étiquettes discrètes à partir des données d'entrée. Les algorithmes

de classification peuvent être divisés en une variété de catégories, chacune avec des méthodologies et des applications distinctes [35].

— **Arbres de décision** : Modèles qui partagent les données récursivement à partir des valeurs des attributs pour classer ou prédire.
— **Forêts aléatoires (Random Forest)** : Ensemble d'arbres de décision agrégés pour augmenter la robustesse [69].
— **Support Vector Machines (SVM)** : Sert à récupèrer la frontière optimale séparant les classes avec la marge maximale, efficace pour les données non linéairement séparables [69].
— **K-Nearest Neighbors (KNN)** : Classe un point selon la majorité des classes de ses k plus proches voisins dans l'espace des caractéristiques [69].
— **Naive Bayes** : Classificateur probabiliste basé sur le théorème de Bayes, rapide et efficace pour les données catégorielles ou textuelles.

2.4.2 Algorithmes de régression

Est une technique statistique de base, utilisée pour modéliser les liens entre les variables, par laquelle les chercheurs sont capables de comprendre et prévoir les résultats sur la base de facteurs explicatifs. Il est appliqué en différentes formes, comme la régression linéaire et logistique, chacune servant à différents buts relativement à l'analyse des données [27].

— **Régression linéaire :** Ce genre de méthode reflète le rapport entre une variable dépendante et une ou plusieurs variables indépendantes en posant qu'elle soit linéaire.Il peut être est simple (un prédicteur unique) ou multiple (plusieurs prédicteurs) [3].
— **Régression logistique :** La régression logistique estime la probabilité qu'un événement se produise sur la base d'une ou de plusieurs variables prédictives, utilisée principalement pour les résultats binaires. Il est essentiel pour identifier les associations dans les études épidémiologiques [43].

2.5 Apprentissage par renforcement

Le concept d'apprentissage par renforcement repose sur l'idée d'apprendre en interagissant avec un environnement. Le modèle prend des décisions et, par conséquent, reçoit des récompenses ou des punitions en fonction de la pertinence de ses actions. Initialement, le modèle fonctionne de façon aléatoire, puis modifie progressivement sa stratégie en se basant sur les retours d'expérience obtenus pendant la phase d'apprentissage. C'est un processus itératif : à chaque étape, le modèle accomplit une tâche, obtient une évaluation pour mesurer son efficacité, puis ajuste ses paramètres en fonction des performances observées. Les modifications apportées aux paramètres sont guidées par les résultats obtenus, dans le but de promouvoir les actions menant aux performances optimales. Il est donc essentiel de sélectionner avec soin les fonctions de récompense et de pénalité, car cela a un impact direct sur la qualité de l'apprentissage et les performances finales du modèle [14].

2.6 Apprentissage non supervisé

On parle cette fois d'apprentissage non supervisé lorsque les données ne sont pas étiquetées. On dispose donc uniquement de données d'entrée dont on ne connaît pas les sorties associées. L'ensemble de données est défini comme :

$$\mathcal{D} = \{x_i\}_{i=1}^{N}$$

L'objectif du système est d'identifier des structures ou des caractéristiques communes présentes dans les données d'entraînement [14].

2.6.1 Algorithmes de clustering

Les algorithmes de clustering sont des outils de base de l'exploration de données, car ils permettent de classer des points de données semblables en groupes appelés *clusters*, en fonction de critères de similarité préétablis. Différents algorithmes ont été développés, chacun ayant des forces et des faiblesses spécifiques influençant leurs performances selon les jeux de données [69].

K-Means :

K-Means est une méthode de **clustering partitionnel** basée sur les centroïdes, qui sépare les données en k groupes. Il est reconnu pour son efficacité et sa précision, notamment sur de grands jeux de données [37].

L'objectif de *K-Means* est de minimiser la somme des distances au carré entre chaque point et le centroïde de son cluster [69]. Chaque point x_j est affecté au cluster dont le centroïde est le plus proche, noté λ_j.

Algorithme 1 : Clustering K-means

Entrée : Ensemble de données $D = \{x_1, x_2, \ldots, x_m\}$;
Nombre de clusters k.
Procédure :

1. Sélectionner aléatoirement k échantillons comme vecteurs moyens initiaux $\{\mu_1, \mu_2, \ldots, \mu_k\}$;
2. **répéter**
   (a) $C_i = \varnothing$ pour $1 \leq i \leq k$;
   (b) **pour** $j = 1, 2, \ldots, m$ **faire**
      i. Calculer la distance entre l'exemple x_j et chaque vecteur moyen μ_i :

$$d_{ij} = \|x_j - \mu_i\|_2^2, \quad \forall i = 1, \ldots, k;$$

      ii. Selon le vecteur moyen le plus proche, déterminer le label de cluster de x_j : $\lambda_j = \arg \min_{i \in \{1,2,\ldots,k\}} d_{ij}$;
      iii. Ajouter x_j au cluster correspondant : $C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$;
   (c) **pour** $i = 1, 2, \ldots, k$ **faire**
      i. Calculer les vecteurs moyens mis à jour : $\mu'_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$;
      ii. **si** $\mu'_i \neq \mu_i$ **alors**
         A. Mettre à jour le vecteur moyen courant $\mu_i \leftarrow \mu'_i$;
      iii. **sinon**
         A. Laisser le vecteur moyen courant inchangé ;
      iv. **fin si**
   (d) **fin pour**
3. **jusqu'à** ce que tous les vecteurs moyens restent inchangés

Sortie : Clusters $C = \{C_1, C_2, \ldots, C_k\}$.

DBSCAN :

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est une méthode de **clustering basé sur la densité**. Elle est connue pour sa capacité à identifier des clusters de forme arbitraire et à filtrer le bruit dans les ensembles de données. Elle est particulièrement efficace lorsque les clusters ont une densité relativement homogène [20, 69].

Cette méthode repose sur deux paramètres :

— ε : la distance maximale pour être considéré comme voisin ;
— `MinPts` : le nombre minimal de voisins requis pour qu'un point soit considéré comme un *objet cœur*.

On définit le ε-voisinage d'un point x_j comme :

$$N_\varepsilon(x_j) = \{x_i \in D \mid \|x_i - x_j\| \leq \varepsilon\}$$

Un point est dit *centrale ou coeur* s'il possède au moins `MinPts` voisins dans ce voisinage.

Algorithme 2 : DBSCAN

Entrée : Ensemble de données $D = \{x_1, x_2, \ldots, x_m\}$;
Paramètres de voisinage (ε, MinPts).
Procédure :

1. Initialiser l'ensemble des centres des exemples : $\Omega = \varnothing$;
2. **pour** $j = 1, 2, \ldots, m$ **faire**
   (a) Déterminer le ε-voisinage $N_\varepsilon(x_j)$ de l'échantillon x_j ;
   (b) **si** $|N_\varepsilon(x_j)| \geq$ MinPts **alors**
      i. Ajouter x_j à l'ensemble des objets centraux : $\Omega = \Omega \cup \{x_j\}$;
   (c) **fin si**
3. **fin pour**
4. Initialiser le nombre de clusters : $k = 0$;
5. Initialiser l'ensemble des échantillons non traités : $\Gamma = D$;
6. **tant que** $\Omega \neq \varnothing$ **faire**
   (a) Garder une copie de l'ensemble courant des données non traitées : $\Gamma_{\text{old}} = \Gamma$;
   (b) Sélectionner aléatoirement un objet cœur $o \in \Omega$, et initialiser la file $Q = \{o\}$;
   (c) $\Gamma = \Gamma \setminus \{o\}$;
   (d) **tant que** $Q \neq \varnothing$ **faire**
      i. Retirer le premier échantillon q de Q ;
      ii. **si** $|N_\varepsilon(q)| \geq$ MinPts **alors**
         A. Soit $\Delta = N_\varepsilon(q) \cap \Gamma$;
         B. Ajouter les échantillons de Δ à Q ;
         C. $\Gamma = \Gamma \setminus \Delta$;
      iii. **fin si**
   (e) **fin tant que**
   (f) $k = k + 1$, générer le cluster $C_k = \Gamma_{\text{old}} \setminus \Gamma$;
   (g) $\Omega = \Omega \setminus C_k$;
7. **fin tant que**

Sortie : Clusters $C = \{C_1, C_2, \ldots, C_k\}$.

Gaussian Mixture Models (GMM) :

GMM (Gaussian Mixture Models) est une méthode de **clustering probabiliste** basée sur la densité. Elle suppose que les données proviennent d'un mélange de plusieurs distributions gaussiennes, chacune représentant un cluster. Ce modèle est particulièrement adapté pour représenter des données multimodales où les sous-populations peuvent se chevaucher [10, 69].

Chaque composante du mélange est caractérisée par trois paramètres :

— α_i : le poids de la composante i, tel que $\sum_{i=1}^{k} \alpha_i = 1$;

— μ_i : le vecteur moyenne de la composante i ;

— Σ_i : la matrice de covariance de la composante i.

Algorithme 3 : Clustering par Mélange de Gaussiennes

Entrée : Ensemble de données $D = \{x_1, x_2, \ldots, x_m\}$;
Nombre de composantes du mélange gaussien k.

Procédure :

1. Initialiser les paramètres $\{(\alpha_i, \mu_i, \Sigma_i) \mid 1 \leq i \leq k\}$ du mélange de distributions gaussiennes ;
2. **répéter**
   (a) **pour** $j = 1, 2, \ldots, m$ **faire**
   — Selon (9.30), calculer les probabilités a posteriori que x_j soit généré par chaque composante gaussienne, c'est-à-dire,
   $\gamma_{ji} = P(M_i \mid x_j) \quad (1 \leq i \leq k)$;
   (b) **fin pour**
   (c) **pour** $i = 1, 2, \ldots, k$ **faire**
   — Calculer la nouvelle moyenne mise à jour :
   $\mu_i' = \frac{\sum_{j=1}^{m} \gamma_{ji} x_j}{\sum_{j=1}^{m} \gamma_{ji}}$;
   — Calculer la nouvelle matrice de covariance mise à jour :
   $\Sigma_i' = \frac{\sum_{j=1}^{m} \gamma_{ji} (x_j - \mu_i')(x_j - \mu_i')^T}{\sum_{j=1}^{m} \gamma_{ji}}$;
   — Calculer le nouveau coefficient du mélange :
   $\alpha_i' = \frac{1}{m} \sum_{j=1}^{m} \gamma_{ji}$;
   (d) **fin pour**
   (e) Mettre à jour les paramètres du modèle $\left\{(\alpha_i', \mu_i', \Sigma_i') \mid 1 \leq i \leq k\right\}$;
3. **jusqu'à** ce que la condition de terminaison soit satisfaite ;
4. **pour** $j = 1, 2, \ldots, m$ **faire**
   — Déterminer le label de cluster de x_j selon (9.31) ;
   Attribuer x_j au cluster correspondant : $C_{y_j} = C_{y_j} \cup \{x_j\}$;
5. **fin pour**

Sortie : Clusters $C = \{C_1, C_2, \ldots, C_k\}$.

Agglomerative Hierarchical Clustering :

L'*agglomerative hierarchical clustering* fait partie des méthodes de **clustering hiérarchique**. C'est une technique largement utilisée en analyse de données, fondée sur une approche ascendante (*bottom-up*) qui regroupe les points de données en fonction de leur similarité.

Au départ, chaque point de données est considéré comme un cluster individuel. À chaque itération, les deux clusters les plus proches sont fusionnés selon une **métrique de distance** (souvent la distance euclidienne ou d'autres mesures comme linkage simple, complet ou moyen), jusqu'à ce qu'un seul cluster regroupe l'ensemble des points [59, 69].

Algorithme 4 : AGNES (Clustering hiérarchique ascendant)

Entrée : Ensemble de données $D = \{x_1, x_2, \ldots, x_m\}$;
Fonction de distance entre clusters d ;
Nombre de clusters final k.

Procédure :

1. **Pour** $j = 1, 2, \ldots, m$ **faire**
   - Initialiser le cluster $C_j = \{x_j\}$;
2. **Fin pour**
3. **Pour** $i = 1, 2, \ldots, m$ **faire**
   - **Pour** $j = i + 1, \ldots, m$ **faire**
     - Calculer $M(i, j) = d(C_i, C_j)$;
     - Affecter $M(j, i) = M(i, j)$;
   - **Fin pour**
4. **Fin pour**
5. Initialiser le nombre courant de clusters : $q = m$;
6. **Tant que** $q > k$ **faire**
   - Trouver deux clusters C_r et C_s ayant la plus petite distance ;
   - Fusionner les clusters : $C_r = C_r \cup C_s$;
   - **Pour** $j = s + 1, s + 2, \ldots, q$ **faire**
     - Changer l'indice $j = j - 1$;
   - **Fin pour**
   - Supprimer la $s^{\text{ème}}$ ligne et colonne de la matrice de distances M ;
   - **Pour** $j = 1, 2, \ldots, q - 1$ **faire**
     - Calculer $M(r, j) = d(C_r, C_j)$;
     - Affecter $M(j, r) = M(r, j)$;
   - **Fin pour**
   - Mettre à jour $q = q - 1$;
7. **Fin tant que**

Sortie : Clusters $C = \{C_1, C_2, \ldots, C_k\}$.

2.7 Métriques d'évaluation

Les paramètres d'évaluation des algorithmes d'apprentissage automatique sont essentiels pour évaluer leurs performances dans diverses tâches, notamment la classification, la régression et le clustering. Ces indicateurs fournissent des informations sur l'efficacité des modèles, orientent les améliorations et garantissent la fiabilité des applications du monde réel. Il est essentiel de bien comprendre ces indicateurs, car ils peuvent varier de manière significative en fonction du contexte et des exigences spécifiques de la tâche à accomplir.

2.7.1 Pour les algorithmes de classification

Matrice de confusion :

La matrice de confusion est un outil fondamental pour évaluer les modèles de classification, permettant de calculer l'exactitude, la précision, le rappel et la spécificité.

TP, (True positif) : nombre de résultats initialement positifs et prédits comme positifs.
FP, (False positif) : nombre de résultats initialement négatifs, mais prédits positifs.
FN, (False negatif) : nombre de résultats initialement positifs, mais prédits négatifs.
TN, (True negatif) : nombre de résultats initialement négatifs et prédits négatifs.

Critère d'évaluation obtenu à partir de la matrice de confusion [27] :

		Classe actuelle	
		Positif	**Négatif**
Classes prédites	**Positif**	TP	FP
	Négatif	FN	TN

TABLE 2.1 – Matrice de confusion

1.Précision :

Mesure la proportion de prédictions correctes par rapport au total des prédictions. Bien qu'il soit largement utilisé, il peut être trompeur dans les ensembles de données déséquilibrés [27].

$$\text{Précision} = \frac{TP}{TP+FP}$$

2.F1-Score :

moyenne harmonique de précision et de rappel, fournissant un équilibre entre les deux métriques, particulièrement utile dans les ensembles de données déséquilibrés [27].

$$F1 = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

utile quand on veut un bon compromis entre précision et rappel.

3.Rappel :

C'est le pourcentage des données positives correctement classifiées, appelé aussi sensibilité [27].

$$\text{Rappel} = \frac{TP}{TP+FN}$$

2.7.2 Pour les algorithmes de régression

1.MSE (Mean Squared Error) :

Le MSE est défini comme l'estimation de l'écart carré d'un estimateur par rapport à la valeur réelle du paramètre, reflétant à la fois le biais et la variance [27].

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

2.MAE (Mean Absolute Error) :

Moyenne des écarts absolus entre les valeurs réelles et les valeurs prédites [27].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

3.R^2 Score (Coefficient de détermination) :

Il mesure la proportion de la variance totale de la variable dépendante (la sortie réelle) qui est expliquée par le modèle de régression.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

4.RMSE (Erreur Quadratique Moyenne Racine) :

Il représente la racine carrée de la moyenne des carrés des écarts entre les valeurs prédites et les valeurs réelles [27].

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2.7.3 Pour les algorithmes de clustering

2.7.3.1. Métriques d'évaluation interne en Clustering

1.Score de la Silhouette

Le score de la silhouette est une métrique utilisée pour la mesure de qualité d'un clustering, il quantifie la similitude d'un point avec son propre cluster et par rapport à d'autres clusters [53].

Pour un point i, son score de silhouette est

$$s(i) = \frac{b(i) - a(i)}{\max(a(i),\ b(i))}$$

avec :
— $a(i)$: distance moyenne entre i et les autres points du même cluster,
— $b(i)$: distance moyenne entre i et les points du cluster le plus proche.
Score global :

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

2.Indice de Davies-Bouldin

L'indice de Davies-Bouldin est une métrique d'évaluation interne du clustering qui cherche à estimer, pour chaque cluster, le pire cas de similarité avec les autres clusters [33].

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right)$$

où :
— $S_i = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - \mu_i\|$ est la dispersion du cluster C_i,
— $M_{ij} = \|\mu_i - \mu_j\|$ est la distance entre les centroïdes des clusters i et j.

3.Indice de Dunn

Il vise à identifier des clusters qui sont bien séparés les uns des autres et internement cohérents

$$D = \frac{\min\limits_{1 \leq i < j \leq k} d(C_i, C_j)}{\max\limits_{1 \leq l \leq k} \delta(C_l)}$$

avec :
— $d(C_i, C_j)$: distance minimale entre deux clusters C_i et C_j,
— $\delta(C_l) = \max_{x,y \in C_l} \|x - y\|$: diamètre du cluster C_l.

4. Indice de Calinski-Harabasz

Sert mesure le rapport entre la variance entre les clusters et la variance intra-cluster, plus les clusters sont denses et bien séparés, meilleure est la partition [40].

$$CH = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \cdot \frac{n-k}{k-1}$$

où :
— n : nombre total de points,
— k : nombre de clusters,
— $\mathrm{Tr}(B_k) = \sum_{i=1}^{k} n_i \cdot \|\mu_i - \mu\|^2$: dispersion inter-cluster,
— $\mathrm{Tr}(W_k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$: dispersion intra-cluster.

2.7.3.2 Métriques d'évaluation externe en clustering

1.Adjusted Rand Index (ARI)

L'ARI mesure la similarité entre deux partitions en corrigeant l'effet du hasard [1] :

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\Big/\binom{n}{2}\right]}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\Big/\binom{n}{2}\right]}$$

où :
— n_{ij} : nombre d'éléments dans l'intersection du cluster i avec la classe j,
— $a_i = \sum_j n_{ij}$,
— $b_j = \sum_i n_{ij}$,
— n : nombre total d'éléments.

2.Adjusted Mutual Information (AMI)

L'information mutuelle (MI) est donnée par [64] :

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|}\frac{n_{ij}}{n}\log\left(\frac{n_{ij}\cdot n}{n_i \cdot n_j}\right)$$

L'AMI est alors définie comme :

$$\text{AMI}(U,V) = \frac{MI(U,V) - \mathbb{E}[MI(U,V)]}{\max(H(U),H(V)) - \mathbb{E}[MI(U,V)]}$$

3.V-Measure

La V-Measure est la moyenne harmonique de l'homogénéité (Homogeneity) et de la complétude (Completeness) [51] :

$$\text{Homogeneity} = 1 - \frac{H(C|K)}{H(C)}$$

$$\text{Completeness} = 1 - \frac{H(K|C)}{H(K)}$$

$$\text{V-Measure} = 2 \cdot \frac{\text{Homogeneity}\cdot\text{Completeness}}{\text{Homogeneity}+\text{Completeness}}$$

4.Fowlkes-Mallows Index (FMI)

Le FMI est calculé comme [22] :

$$\text{FMI} = \sqrt{\frac{TP}{TP+FP}\cdot\frac{TP}{TP+FN}}$$

où :
— TP : vraies paires positives (même cluster et même classe),
— FP : fausses paires positives (même cluster mais classes différentes),
— FN : fausses paires négatives (clusters différents mais même classe).

5.Pureté

La mesure de pureté est définie comme [1] :

$$\text{Purity} = \frac{1}{n} \sum_k \max_j \left| C_k \cap L_j \right|$$

où :
— C_k : le cluster k,
— L_j : la classe réelle j,
— n : nombre total de points.

Conclusion

Ce chapitre a présenté les concepts clés de l'apprentissage automatique, en expliquant son utilité pour résoudre des problèmes complexes. Les trois types (supervisé, non-supervisé et par renforcement) ont été introduits, ainsi que les métriques d'évaluation essentielles pour mesurer les performances des modèles. Ces bases théoriques permettent d'envisager des applications pratiques, notamment dans la détection et la prévention des défaillances des pipelines.

3

Étude de cas : OLÉODUC HAOUD EL HAMRA- BEJAIA

Sommaire

Introduction

Dans ce chapitre, un jeu de données issu d'une inspection MFL (Magnetic Flux Leakage) de pipelines réalisée par la société ROSEN est analysé. Les données décrivent des anomalies telles que la corrosion, en précisant leurs dimensions et leur position géographique. Après un prétraitement rigoureux (nettoyage, gestion des valeurs manquantes et normalisation), des modèles de machine learning sont appliqués pour segmenter les défauts et soutenir une stratégie de maintenance prédictive.

3.1 Collecte de données

La base de données exploitées dans notre étude provienent d'une inspection par la méthode MFL relient une ligne abandonnée entre Haoud El Hamra et Béjaia , réalisé en 2010 par l'entreprise allemande ROSEN qui est spécialisé dans l'inspection et le modèle RoCorr MFL-A, reconnu par sa haute précision dans la détection des pertes de métal et anomalies internes sur les pipelines. L'inspection couvre la totalité de la circonférence du pipeline. Le jeu de données issue de cette inspection contient 14846 observaions. Les variables clés dans ce dataset sont la profondeur qui est exprimée en pourcentage de l'épaisseur de paroi ainsi que les dimensions des défauts, à savoir la longueur (mm) et la largeur (mm).

No	Facteur	Description	Type de variable
1	Profondeur [%]	Profondeur de la corrosion sur la paroi du pipeline	Numérique
2	Longueur [mm]	Longueur de la corrosion sur la paroi du pipeline	Numérique
3	Largeur [mm]	Largeur de la corrosion sur la paroi du pipeline	Numérique
4	Surface [mm²]	Surface de la corrosion sur la paroi du pipeline	Numérique

TABLE 3.1 – Description des facteurs liés à la corrosion du pipeline

3.2 Préparation des données pour le clustering

Avant l'application des algorithmes de clustering, un prétraitement rigoureux des données a été effectué, conformément à la méthodologie décrite dans le chapitre 2. Ces étapes ont permis d'assurer la qualité et la comparabilité des données, condition essentielle pour garantir la fiabilité des résultats produits par les algorithmes de machine learning utilisés dans cette étude.

3.2.1 Algorithmes de clustering et normalisation recommandée

1. Agglomerative Hierarchical Clustering

- **Principe** : approche fondée sur des mesures de distance (euclidienne par défaut) pour regrouper récursivement les points.
- **Normalisation recommandée** : `StandardScaler` (Z-score).
- **Justification** : met toutes les variables sur une échelle comparable, assurant une contribution équitable au calcul des distances. Même si les variables sont homogènes (ex : toutes en mètres), cette étape reste nécessaire.

2. K-Means

- **Principe** : partitionne les données en clusters en minimisant les distances euclidiennes entre les points et les centroïdes.
- **Normalisation recommandée** :
  - `StandardScaler` : recommandé pour des variables continues comme profondeur, largeur, longueur.
- **Justification** : des échelles différentes biaisent le calcul des distances, faussant la formation des clusters.

3. DBSCAN

- **Principe** : identifie les régions de forte densité à partir de distances locales pour former des clusters.
- **Normalisation recommandée** :
  - `StandardScaler` : équilibre les dimensions lors du calcul des distances.
- **Justification** : des échelles incohérentes perturbent la détection fiable des régions denses.

4. Gaussian Mixture Models (GMM)

- **Principe** : approche probabiliste utilisant une combinaison de gaussiennes multivariées pour représenter les clusters.
- **Normalisation recommandée** : `StandardScaler`.
- **Justification** : si les variables n'ont pas la même échelle, certaines gaussiennes peuvent dominer certaines dimensions, faussant la modélisation des groupes.

3.3 Implementation des algorithmes

3.3.1 Application de K-means

Avant l'application de K-Means, il convient d'optimiser les hyperparamètres, notamment le nombre optimal des clusters. Parmi les méthodes utilisées, on trouve :

1. Coefficient de silhouette

Le coefficient de silhouette est une méthode qui permet d'évaluer la qualité d'un clustering. Pour chaque point, il mesure à quel point il est bien regroupé avec les points de son propre cluster par rapport à ceux des autres clusters. Ce score varie entre -1 et 1 : plus il est proche de 1, mieux le point est placé. Pour déterminer le bon nombre de clusters, on calcule le score moyen de silhouette pour différentes valeurs de ‘k‘. Le nombre de clusters qui donne le score le plus élevé est généralement considéré comme le plus adapté, car il reflète des groupes à la fois compacts et bien séparés.

2. Résultats des meilleurs hyperparamètres pour K-Means

```
Meilleure combinaison trouvée:
Nombre de clusters (k): 2
Nombre d'initialisations (n_init): 5
Nombre maximum d'itérations (max_iter): 100
Méthode d'initialisation (init): k-means++
Score de silhouette: 0.724
```

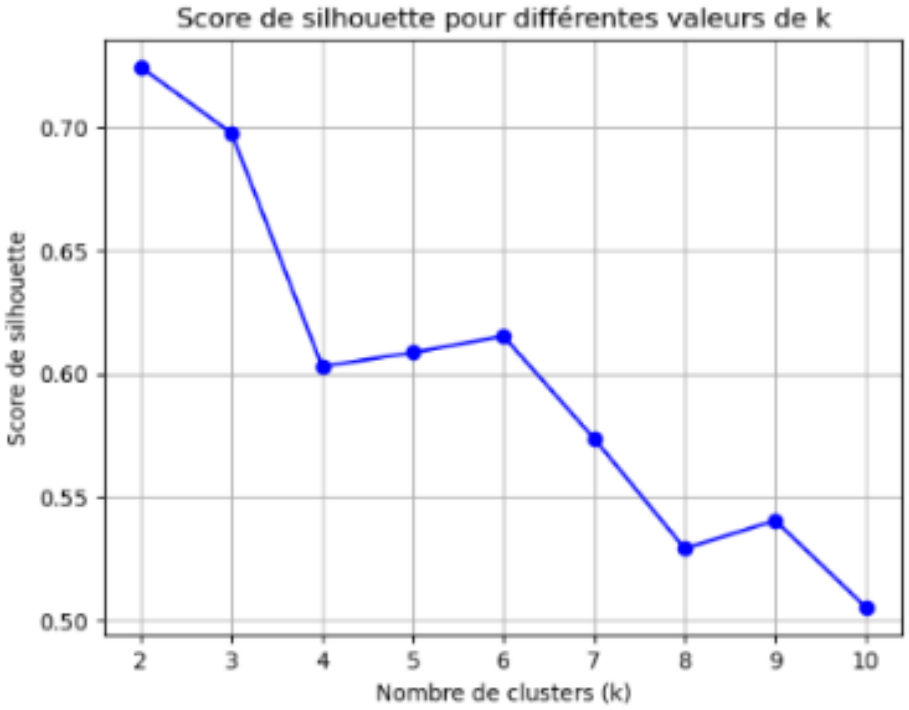

FIGURE 3.1 – Score de silhouette pour K-means

Le graphique du score de silhouette montre clairement que le meilleur nombre de clusters est 2, avec un score élevé de 0.724, indiquant une bonne séparation entre les groupes. L'ajout de clusters supplémentaires dégrade la qualité du partitionnement, confirmant que 2 groupes naturels sont présents dans les données.

3. Résultat de K-Means avec les meilleurs hyperparamètres

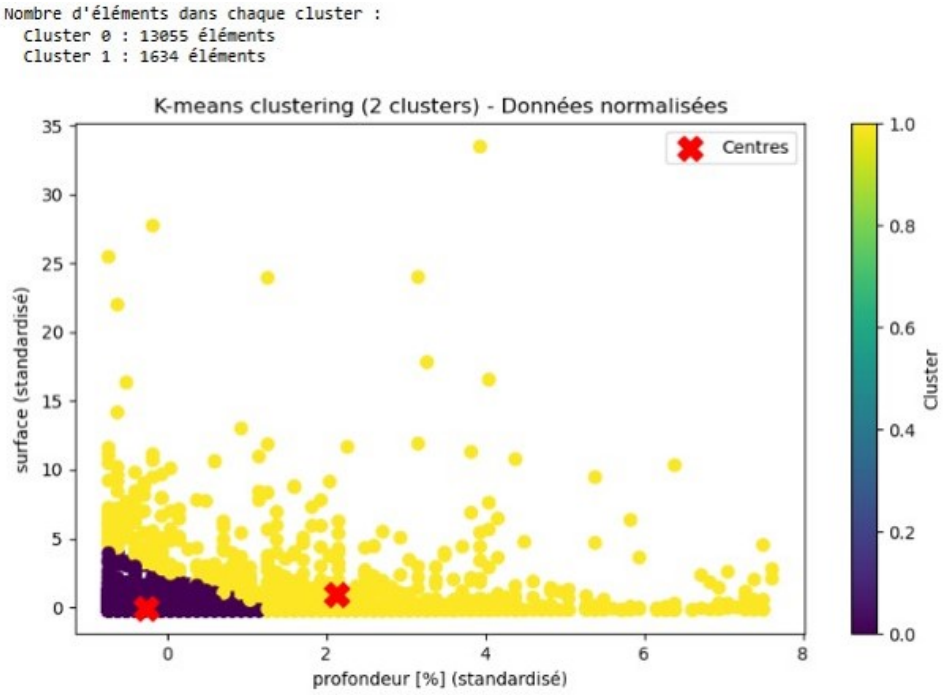

FIGURE 3.2 – Résultat obtenu avec K-means

Ce résultat révèle la présence de deux groupes bien différenciés :

— **Cluster 0** : caractérisé par de faibles profondeurs et de petites surfaces ;

— **Cluster 1** : associé à des profondeurs intermédiaires et des surfaces nettement plus étendues.

Cette analyse confirme l'impact significatif de la profondeur sur la surface. Toutefois, l'hétérogénéité observée au sein du Cluster 1 pourrait justifier un affinement, par exemple via un sous-clustering, pour optimiser la pertinence des conclusions.

3.Résultat K-means sur MAPS

FIGURE 3.3 – Représentation cartographique des clusters K-Means le long du pipeline

La carte montre les résultats du clustering K-means appliqué aux données d'un pipeline, distinguant deux groupes de défauts : le Cluster 0 (en vert) pour les défauts peu sévères et le Cluster 1 (en orange) pour les défauts sévères localisés. L'algorithme permet de différencier les niveaux de gravité et d'orienter la maintenance. Un regroupement important de défauts peu sévères peut également révéler une agressivité du sol, favorisant l'apparition de défauts légers sur certaines zones du tracé.

3.3.2 Application de DBSCAN

Avant d'appliquer l'algorithme DBSCAN on doit trouver les meilleurs hyperparametres. On a utilisé un algorithme qui teste plusieurs combinaisons (eps, minsamples) pour trouver les valeurs les plus optimales. Par conséquent a choisi les hyperparamétres avec le score silhouette plus élevé.

1. Résultat des meilleurs hyperparamètres

```
eps=0.20, min_samples=4, clusters=18, silhouette=0.596
eps=0.20, min_samples=6, clusters=11, silhouette=0.616
eps=0.20, min_samples=8, clusters=9, silhouette=0.520
eps=0.20, min_samples=10, clusters=4, silhouette=0.528
eps=0.20, min_samples=12, clusters=3, silhouette=0.594
eps=0.20, min_samples=14, clusters=3, silhouette=0.602
eps=0.25, min_samples=4, clusters=7, silhouette=0.726
eps=0.25, min_samples=6, clusters=6, silhouette=0.605
eps=0.25, min_samples=8, clusters=6, silhouette=0.644
eps=0.25, min_samples=10, clusters=3, silhouette=0.716
eps=0.25, min_samples=12, clusters=6, silhouette=0.590
eps=0.25, min_samples=14, clusters=3, silhouette=0.817
eps=0.30, min_samples=4, clusters=6, silhouette=0.731
eps=0.30, min_samples=6, clusters=7, silhouette=0.727
eps=0.30, min_samples=8, clusters=6, silhouette=0.618
eps=0.30, min_samples=10, clusters=6, silhouette=0.646
eps=0.30, min_samples=12, clusters=4, silhouette=0.815
eps=0.30, min_samples=14, clusters=4, silhouette=0.646
eps=0.35, min_samples=4, clusters=6, silhouette=0.717
eps=0.35, min_samples=6, clusters=3, silhouette=0.678
eps=0.35, min_samples=8, clusters=2, silhouette=0.829
eps=0.35, min_samples=10, clusters=3, silhouette=0.814
eps=0.35, min_samples=12, clusters=3, silhouette=0.718
eps=0.35, min_samples=14, clusters=3, silhouette=0.714
eps=0.40, min_samples=4, clusters=8, silhouette=0.716
eps=0.40, min_samples=6, clusters=3, silhouette=0.723
eps=0.40, min_samples=10, clusters=2, silhouette=0.826
eps=0.40, min_samples=12, clusters=3, silhouette=0.814
eps=0.40, min_samples=14, clusters=2, silhouette=0.818
eps=0.45, min_samples=4, clusters=2, silhouette=0.754
eps=0.45, min_samples=6, clusters=3, silhouette=0.738
eps=0.45, min_samples=8, clusters=3, silhouette=0.724
eps=0.45, min_samples=12, clusters=2, silhouette=0.821
eps=0.45, min_samples=14, clusters=2, silhouette=0.814
eps=0.50, min_samples=4, clusters=3, silhouette=0.789
eps=0.50, min_samples=6, clusters=5, silhouette=0.718
eps=0.50, min_samples=8, clusters=3, silhouette=0.722
eps=0.50, min_samples=14, clusters=3, silhouette=0.817
eps=0.55, min_samples=4, clusters=5, silhouette=0.781
eps=0.55, min_samples=6, clusters=3, silhouette=0.733
eps=0.55, min_samples=8, clusters=2, silhouette=0.736

✅ Meilleurs hyperparamètres trouvés :
eps = 0.35, min_samples = 8
Silhouette score = 0.829
```

FIGURE 3.4 – Les meilleurs hyperparamétres pour DBSCAN

Le meilleur compromis entre nombre de clusters et qualité de clustering est obtenu pour eps = 0.35 et minsamples = 8 avec un score silhouette de 0.829.

Cela indique une structure de données bien capturée avec un nombre modéré de clusters et des groupes bien définis.

Il est toujours utile de visualiser les clusters pour confirmer l'interprétation, mais les scores numériques indiquent une segmentation pertinente.

2. Résultat de DBSCAN avec les meilleurs hyperparamètres

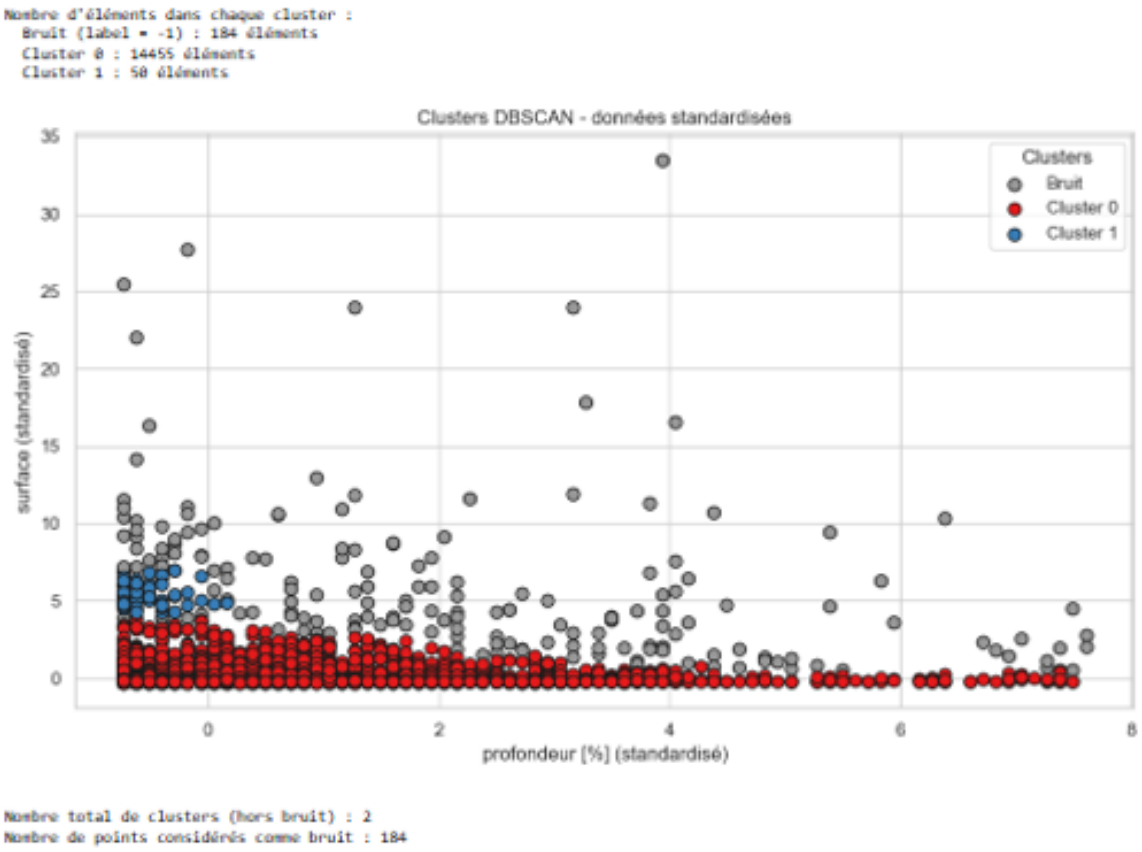

FIGURE 3.5 – Résultat obtenu avec DBSCAN

L'algorithme de clustering DBSCAN a bien séparé les groupes en deux classes : Un grand groupe principal (Cluster 0), un petit groupe secondaire avec des caractéristiques distinctes (Cluster 1). Des valeurs atypiques considérées comme bruit et La valeur élevée du score silhouette (0.829) observée précédemment est cohérente avec cette visualisation : les clusters sont bien séparés et internement compacts. DBSCAN a l'avantage ici de détecter les bruit sans les forcer dans un groupe.

3.3.3 Application de Agglomerative Hierarchical Clustering

Le nombre de clusters est déterminé en maximisant le coefficient de silhouette.

1. Résultat du nombre optimal de clusters

```
--- Clustering hiérarchique automatique : 'profondeur [%]' vs 'surface' ---
n_clusters = 2 → Silhouette = 0.741
n_clusters = 3 → Silhouette = 0.722
n_clusters = 4 → Silhouette = 0.539
n_clusters = 5 → Silhouette = 0.540
n_clusters = 6 → Silhouette = 0.545
n_clusters = 7 → Silhouette = 0.565
n_clusters = 8 → Silhouette = 0.566
n_clusters = 9 → Silhouette = 0.553
n_clusters = 10 → Silhouette = 0.503

✅ Meilleur n_clusters : 2 (Silhouette = 0.741)
```

FIGURE 3.6 – Nombre de cluster avec la méthode silhouette

Le nombre optimal de clusters est 2, car il maximise le coefficient de silhouette. Cela suggère qu'il y a deux groupes naturels dans les données, bien séparés sur la base de la profondeur et de la surface.

2. Résultat avec les meilleurs hyperparamètres

```
📄 Répartition des éléments par cluster :
Cluster 0 : 13381 éléments
Cluster 1 : 1308 éléments
```

FIGURE 3.7 – Nombre d'éléments pour chaque cluster

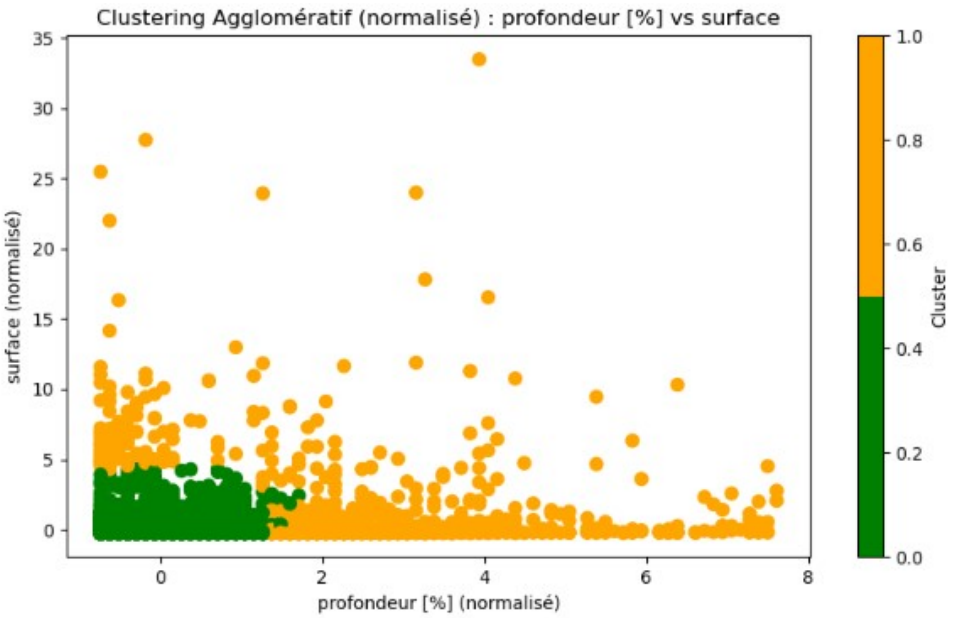

FIGURE 3.8 – Résultat obtenu avec agglomérative heirarchical clustering

Le clustering a identifié deux grands groupes.

Cluster jaune (valeur proche de 1) regroupe les points ayant à la fois une faible profondeur et une faible surface.

Cluster vert (valeur proche de 0) : contient le reste des points, qui sont plus dispersés, avec des valeurs parfois élevées de surface ou profondeur.

3.3.4 Application du modèle de mélanges gaussiens (GMM)

Le nombre optimal de composantes est déterminé à l'aide des critères AIC et BIC.

1. Résultats des critères AIC et BIC

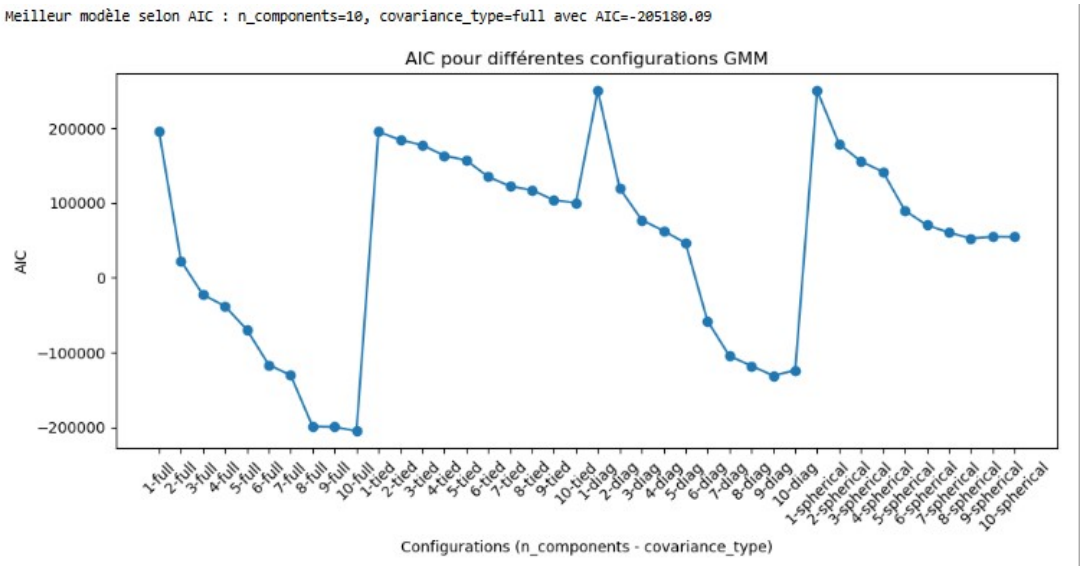

FIGURE 3.9 – Résultat des meilleurs hyperparamétres avec AIC

— Le meilleur modèle selon le critère AIC est exactement le même :
  — Nombre de composantes ($n_{\text{components}}$) : 10
  — Type de covariance : `full`
  — Valeur de l'AIC : −205180.09

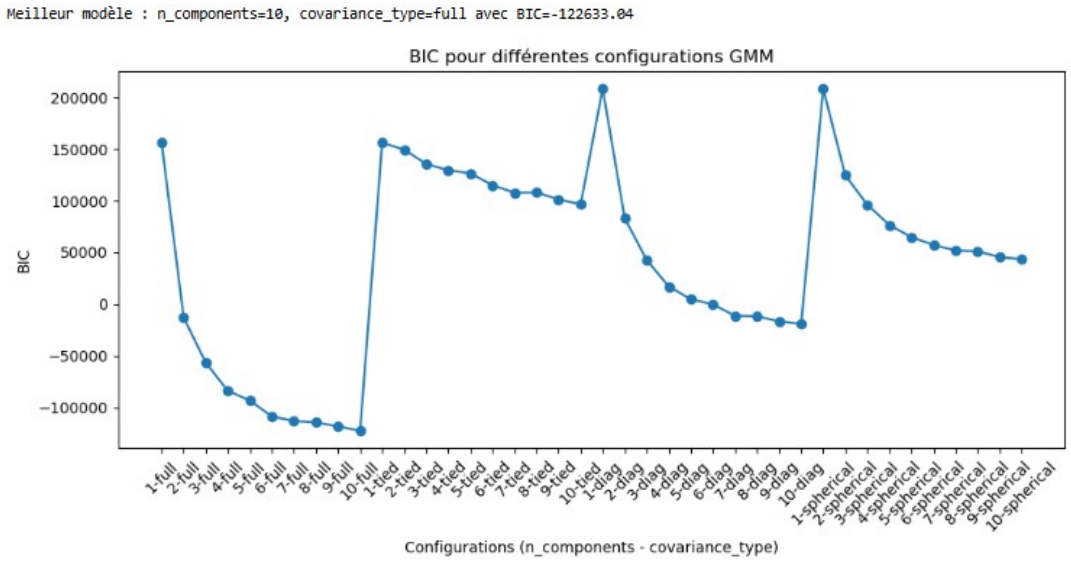

FIGURE 3.10 – Résultat des meilleurs hyperparamétres avec BIC

— Le meilleur modèle selon le critère BIC est :
  — Nombre de composantes ($n_{\text{components}}$) : 10
  — Type de covariance : `full`
  — Valeur du BIC : −122633.04

On observe que le BIC diminue fortement jusqu'à 10 composantes pour le type full, puis remonte pour d'autres configurations.

Cela suggère qu'un GMM avec 10 composantes et des matrices de covariance complètes est le plus efficace en termes d'ajustement et de parcimonie.

L'AIC, moins pénalisant que le BIC pour les modèles complexes, confirme ce choix.

2. Résultat de Gaussian Mixture Models avec les meilleurs hyperparamètres

```
Nombre d'éléments dans chaque cluster :
  Cluster 0 : 2851 éléments
  Cluster 1 : 1929 éléments
  Cluster 2 : 202 éléments
  Cluster 3 : 223 éléments
  Cluster 4 : 59 éléments
  Cluster 5 : 3655 éléments
  Cluster 6 : 1476 éléments
  Cluster 7 : 3025 éléments
  Cluster 8 : 494 éléments
  Cluster 9 : 775 éléments
```

FIGURE 3.11 – Nombre d'éléments pour chaque cluster

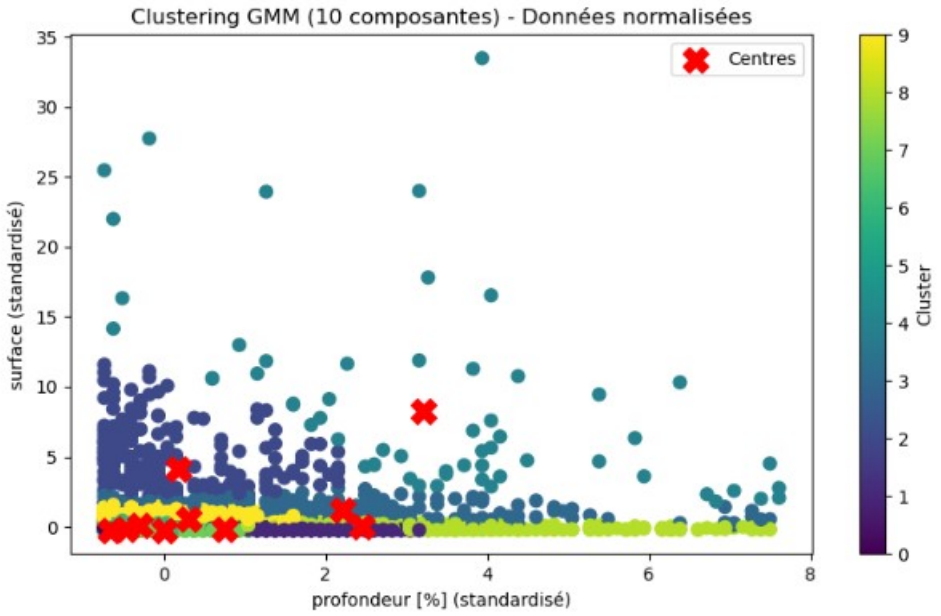

FIGURE 3.12 – Résultat obtenu avec GMM

Le modèle GMM a identifié 10 groupes distincts dans des données normalisées selon deux dimensions (profondeur, surface).

Il semble que plusieurs clusters aient été identifiés dans des zones de forte densité, tandis que d'autres couvrent des régions plus dispersées.

3.4 Développement de modèles de prédiction

Dans le cadre de ce travail, quatre algorithmes de clustering non supervisé ont été appliqués aux données issues de l'inspection MFL d'un pipeline de SONATRACH. L'objectif était de regrouper automatiquement les défauts observés selon leurs caractéristiques géométriques (profondeur, surface) afin d'évaluer leur sévérité. Les performances ont été comparées à l'aide de trois métriques internes : le Silhouette Score, l'indice de Calinski-Harabasz, et l'indice de Davies-Bouldin, dont les résultats sont récapitulés dans le tableau suivant :

Algorithme	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means (k=2)	0.7274	7549.929	0.951
DBSCAN	0.829	3351.251	1.445
Agglomerative Hierarchical Clustering	0.741	7163.8	0.936
Gaussian Mixture Models (GMM)	0.101	3735.397	1.539

TABLE 3.2 – Comparaison des scores de différents algorithmes de clustering

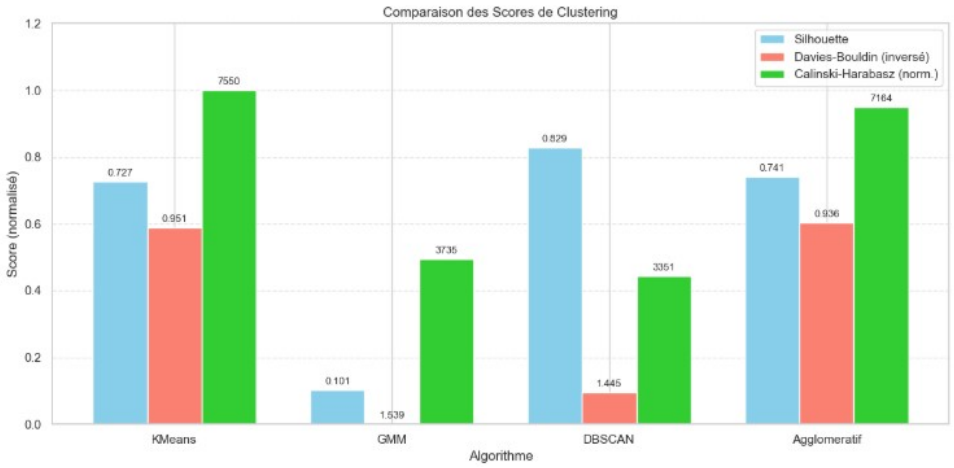

FIGURE 3.13 – Résultat obtenu avec les métriques d'évaluation interne

3.4.1 Interprétation des indicateurs

Silhouette Score, qui mesure la cohésion intra-cluster et la séparation inter-cluster, indique que DBSCAN (0.829) présente les groupes les plus nets et les plus cohérents. Cela est cohérent avec la capacité de cet algorithme à détecter des clusters de forme irrégulière, typiques des défauts corrosifs réels.

L'indice de **Calinski-Harabasz** attribue la meilleure valeur à K-Means (7549.929), ce qui montre que cet algorithme est performant lorsque les clusters sont sphériques et bien équilibrés. Le clustering hiérarchique s'en approche avec une valeur également élevée (7163.8).

L'indice de **Davies-Bouldin**, où une valeur plus faible indique une meilleure séparation, est le plus bas pour le clustering hiérarchique agglomératif (0.936), ce qui confirme la qualité des groupes formés. K-Means obtient également une valeur correcte (0.951).

En revanche, le modèle GMM a donné des performances très faibles avec un Silhouette Score proche de zéro (0.101), ce qui indique une mauvaise délimitation des clusters. De plus, son indice Davies-Bouldin élevé (1.539) montre que les groupes sont peu distincts et se chevauchent fortement.

Conclusion

Ce chapitre a présenté la démarche d'analyse des données issues de l'inspection MFL d'un pipeline exploité par SONATRACH, à travers l'application de plusieurs algorithmes de clustering non supervisé. Les méthodes K-Means, DBSCAN, Agglomerative Clustering et GMM ont été implémentées, puis comparées à l'aide de trois métriques internes : le Silhouette Score, l'indice de Calinski-Harabasz et l'indice de Davies-Bouldin.

Les résultats montrent que K-Means obtient les meilleures performances globales, notamment avec un indice de Calinski-Harabasz très élevé (7549.929) et un Silhouette Score satisfaisant (0.7274), traduisant une bonne séparation entre les groupes, surtout lorsque les clusters sont compacts et équilibrés. De plus, ses résultats sont cohérents et interprétables sur le plan physique, notamment dans la distinction entre défauts superficiels et sévères.

L'algorithme DBSCAN, quant à lui, s'est démarqué par sa capacité à détecter automatiquement le bruit et à identifier des clusters de forme irrégulière, avec un Silhouette Score légèrement supérieur (0.829). Cependant, sa performance est pénalisée par une mauvaise compacité globale (indice de Calinski-Harabasz plus faible) et une séparation moindre (Davies-Bouldin de 1.445), ce qui limite sa robustesse dans le contexte de données techniques standardisées.

Le clustering hiérarchique agglomératif offre un bon compromis, avec des performances équilibrées et une bonne lisibilité des regroupements, tandis que le modèle GMM s'est montré le moins adapté, en raison d'une forte sensibilité au bruit et d'un chevauchement important entre les groupes, comme en témoigne un Silhouette Score très faible (0.101).

En conclusion, K-Means apparaît comme l'algorithme le plus fiable et pertinent pour la classification automatique des défauts de corrosion sur pipeline dans ce cas d'étude, combinant performance, simplicité et lisibilité des résultats.

Conclusion générale

Notre étude s'est inscrite dans une démarche d'exploration de nouvelles approches pour améliorer la détection et l'évaluation des défauts de corrosion sur les pipelines, en s'appuyant sur les données d'inspection MFL fournies par SONATRACH. Dans un contexte où la corrosion reste l'un des facteurs majeurs de défaillance des infrastructures pétrolières, l'intégration des techniques de machine learning, et plus particulièrement des méthodes de clustering non supervisé, offre une alternative puissante aux approches traditionnelles, souvent limitées à des analyses descriptives ou rétrospectives.

À travers une étude de cas sur la ligne Haoud El Hamra – Béjaïa, plusieurs algorithmes de clustering ont été implémentés : K-Means, DBSCAN, Agglomerative Clustering et GMM. Leur performance a été comparée à l'aide de métriques internes (Silhouette, Calinski-Harabasz, Davies-Bouldin), permettant une évaluation objective de la qualité des regroupements. Les résultats obtenus ont mis en évidence la supériorité de K-Means, qui s'est révélé être le plus performant pour segmenter les défauts selon leur gravité, avec une bonne cohérence entre les clusters et les caractéristiques physiques des défauts observés.

Cette démarche a permis d'identifier des zones critiques sur le pipeline, de différencier automatiquement les niveaux de sévérité des corrosions, et de visualiser spatialement les regroupements. Ces apports constituent une avancée significative vers une maintenance prédictive, plus ciblée et plus efficace, réduisant ainsi les risques environnementaux et les pertes économiques liés à une rupture ou une intervention tardive.

En conclusion, ce travail démontre l'intérêt d'exploiter les algorithmes de machine learning pour valoriser les données issues d'inspections industrielles. Il ouvre également la voie à de futures recherches intégrant des modèles plus complexes (tels que les réseaux de neurones ou les séries temporelles), ou des données complémentaires (température, pression, géochimie du sol), afin d'aboutir à une surveillance intelligente, en temps réel, des pipelines.

A

Annexe A : Environnement de développement

Cette section décrit l'environnement de développement utilisé pour la mise en œuvre des modèles de prévision des ventes. Nous aborderons les outils, les bibliothèques et les technologies employés pour développer, tester et valider les modèles.

A.1 Environnement de développement intégré IDE

Dans le cadre de ce projet, l'environnement de développement utilisé est Jupyter Notebook, accessible via la distribution Anaconda. Ce choix s'explique par la facilité d'utilisation, la compatibilité avec les bibliothèques Python scientifiques, et la capacité à combiner code, résultats, visualisations et commentaires dans un même document interactif.

Anaconda est une distribution libre de Python qui intègre un gestionnaire de paquets (conda) et de nombreux outils dédiés à la science des données, dont Jupyter Notebook. L'installation et le lancement de l'environnement se font simplement à travers l'interface graphique "Anaconda Navigator" ou en ligne de commande avec jupyter notebook.

L'utilisation de Jupyter dans ce projet a permis :

l'exécution pas-à-pas des algorithmes (K-Means, DBSCAN, GMM et Clustering Hiérarchique Ascendant),

la visualisation des résultats via les bibliothèques matplotlib et seaborn,

le calcul et l'analyse de métriques comme l'AIC et le BIC,

l'insertion de texte explicatif et de formules mathématiques (grâce à la prise en charge de LaTeX en Markdown),

la reproductibilité des expérimentations.

Grâce à cet outil interactif, il a été possible de documenter efficacement les étapes du projet, d'interpréter les résultats en temps réel, et de produire des visualisations claires et adaptées à l'analyse des données.

FIGURE A.1 – Anaconda

FIGURE A.2 – Jupyter

FIGURE A.3 – Interfaces de Anaconda et Jupyter

A.2 Langage de programmation

Malgré l'existence de nombreux langages de programmation aux fonctionnalités variées, nous avons opté pour Python grâce à sa simplicité d'installation et à sa syntaxe limpide. Python figure actuellement parmi les langages de programmation les plus prisés et couramment employés, surtout dans les secteurs de la science des données et de l'apprentissage automatique.

FIGURE A.4 – Phyton

A.3 Bibliothèques et frameworks utilisées

Dans ce mémoire, plusieurs bibliothèques et frameworks Python ont été utilisés pour la manipulation des données, le clustering et la visualisation. Ces outils offrent des implémentations optimisées et des interfaces simplifiées pour les tâches de data science.

A.3.1 Bibliothèques de manipulation des données

Pandas (`pandas`)

- **Rôle** : Manipulation efficace de données structurées via les `DataFrames`.
- **Utilisation** :
  - Chargement et nettoyage des données (ex. : suppression des valeurs manquantes).
  - Agrégation et filtrage pour préparer l'analyse.

NumPy (`numpy`)

- **Rôle** : Calcul numérique haute performance (opérations matricielles).
- **Utilisation** :
  - Conversion des données en formats adaptés aux algorithmes (ex. : tableaux `ndarray`).
  - Calculs de distances entre points pour les métriques de clustering.

A.3.2 Bibliothèques de visualisation

Matplotlib (`matplotlib.pyplot`)

- **Rôle** : Création de graphiques statiques (2D/3D).
- **Utilisation** :
  - Visualisation des clusters (nuages de points, *elbow plots* pour K-means).
  - Comparaison visuelle des résultats entre algorithmes.

Seaborn (`seaborn`)

- **Rôle** : Visualisations statistiques avancées basées sur Matplotlib.
- **Utilisation** :
  - Heatmaps pour analyser les corrélations.
  - Distributions des données avant/après clustering.

A.3.3 Frameworks de machine learning

Scikit-learn (`sklearn`)

- **Rôle** : Bibliothèque complète pour le machine learning.
- **Modules utilisés** :
  - `cluster` : Implémentation de K-means, DBSCAN, clustering hiérarchique.
  - `preprocessing` : Normalisation via `StandardScaler`.
  - `metrics` : Scores de silhouette, Davies-Bouldin, etc.
- **Utilisation** :
  - Comparaison des performances des algorithmes.
  - Optimisation des hyperparamètres (ex. : `eps` pour DBSCAN).

SciPy (`scipy.cluster.hierarchy`)

- **Rôle** : Méthodes scientifiques (ex. : clustering hiérarchique).
- **Utilisation** :
  - Génération de dendrogrammes (`linkage`, `dendrogram`).

A.3.4 Bibliothèques spécialisées

Folium (`folium`)

- **Rôle** : Cartographie interactive (basée sur Leaflet.js).
- **Utilisation** :
  - Représentation géographique des clusters (si applicable).

PyProj (`pyproj`)

- **Rôle** : Transformations de systèmes de coordonnées.
- **Utilisation** :
  - Conversion de coordonnées (ex. : WGS84 vers UTM pour des calculs de distance).

A.3.5 Conclusion

L'écosystème Python offre des outils puissants et complémentaires pour les projets de data science. Dans ce mémoire, leur interopérabilité (ex. : Pandas → Scikit-learn → Matplotlib) a permis une analyse robuste et reproductible. Certaines limites ont pu être rencontrées, comme la sensibilité de DBSCAN au choix des hyperparamètres, atténuée par l'utilisation systématique de métriques d'évaluation.

Bibliographie

[1] Aggarwal, C. C. Cluster validation. In *Data Classification : Algorithms and Applications*, C. C. Aggarwal and C. K. Reddy, Eds. CRC Press, 2014, ch. 13, pp. 517–544.

[2] Ahamad, M. A., Rahman, H. A., and Osman, S. A. Pipeline wall thickness assessment of various material grades and water depths using american and norwegian standards. *Jurnal Kejuruteraan 34*, 6 (2022), 1135–1147.

[3] Aityan, S. K. Linear regression. *Journal of Machine Learning Foundations* (2022), 915–978.

[4] Amirian, E., Moy, T. X., Ahmad, M. N. A., Abdollahzadeh, A., Rohani, M. J., Abdullah, M. S. K., and Hidzir, H. B. Integrated meteorological and geohazard system advisory (imgesa) for pipeline integrity. https://doi.org/10.3997/2214-4609.202377021.

[5] Andersen, T., and Misund, A. Pipeline reliability : An investigation of pipeline failure characteristics and analysis of pipeline failure rates for submarine and cross-country pipelines. *Journal of Petroleum Technology 35*, 04 (1983), 709–717.

[6] Baboian, R. *Corrosion Tests and Standards : Application and Interpretation*. ASTM International, 1995.

[7] Beavers, J. A., and Thompson, N. G. *External Corrosion of Oil and Natural Gas Pipelines*. ASM International, 2006.

[8] Berger, T., et al. Digital transformation in pipeline inspection technologies. *Journal of Pipeline Integrity 14*, 3 (2021), 145–160.

[9] Bilbao, I., Bilbao, J., and Feniser, C. Adopting some good practices to avoid overfitting in the use of machine learning. *WSEAS Transactions on Mathematics Archive 17* (2018).

[10] Bishop, C. M. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 2006. https://www.microsoft.com/en-us/research/wp-content/uploads/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

[11] Boukhari, R. Raffinage : les capacités de production «renforcées ». *Le Jeune Indépendant* (03 juil. 2024 à 15 :54).

[12] Cawley, G. C. Over-fitting in model selection and its avoidance. In *Model Selection and Evaluation in Machine Learning*. Springer, 2012, pp. 1–25.

[13] CNAM. Cours sur les arbres de décision en machine learning, 2023.

[14] De Matteis, L., Janny, S., Nathan, S., and Shu-Quartier, W. Introduction à l'apprentissage automatique. *Culture Sciences de l'Ingénieur* (2022).

[15] Det Norske Veritas. *DNVGL-RP-B401 : Cathodic Protection Design*. Det Norske Veritas, 2018. Recommended Practice.

[16] Doulah, M. S. U., and Islam, M. N. Performance evaluation of machine learning algorithm in various datasets. *Journal of Artificial Intelligence, Machine Learning and Neural Network 32* (2023), 14–32.

[17] Duperrex, M. Le pipeline et la clôture de la frontière. In *Du sillon à la skyline. Des lignes et des paysages*, P.-H. Frangne, Ed. Presses universitaires de Rennes, Rennes, 2020, pp. 159–179.

[18] Encyclopædia Britannica. Pipeline, 2025. Consulté en juin 2025.

[19] Esser, A. Pipeline for the hydraulic or pneumatic transport of solids. https://patents.google.com/patent/US20080298908A1/en, 2006. Patent.

[20] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)* (Portland, OR, USA, 1996), AAAI Press, pp. 226–231.

[21] Farh, H. M. H., Ben Seghier, M. E. A., Taiwo, R., and Zayed, T. Analysis and ranking of corrosion causes for water pipelines : a critical review. *NPJ Clean Water* (2023).

[22] Fowlkes, E. B., and Mallows, C. L. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association 78*, 383 (1983), 553–569.

[23] Fürnkranz, J., Gamberger, D., and Lavrač, N. Pruning of rules and rule sets. In *Foundations of Rule Learning*. Springer, 2012, pp. 187–216.

[24] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[25] Guo, J. Offshore pipeline mechanical design. In *Elsevier eBooks*. Elsevier, 2023, pp. 251–286.

[26] Gupta, G., and Lala, R. Sonatrach : Upstream strategy assessment. *SP Global Commodity Insights* (Jun 09, 2023).

[27] Hossain, E. *Machine Learning Crash Course for Engineers*. Springer, 2024.

[28] Hsieh, W. W. Decision trees, random forests and boosting. In *Advances in Machine Learning*. Cambridge University Press, 2023, pp. 473–493.

[29] Hu, Y. Analysis and research on overfitting and underfitting issues in heart disease prediction models. *Highlights in Science Engineering and Technology* (2024). Open access, peer-reviewed journal.

[30] Hui, Q., Liu, B., Fan, J., Guo, D., Tan, X., and Peng, Z. Design and application of the standard tag set in the oil and gas pipeline domain. *Proceedings of SPIE 118* (2024).

[31] I. A. School. Rôle du feature engineering en machine learning. https://www.example.com/role-feature-engineering-machine-learning, 2023. Consulté le 12 juillet 2025.

[32] Jain, V., and Tiwari, S. K. Overview : machine learning, 2024. Accessed via DOI.

[33] Jollyta, D., Efendi, S., Zarlis, M., and Mawengkang, H. Optimasi cluster pada data stunting : Teknik evaluasi cluster sum of square error dan davies bouldin index. *SENARIS 1* (2019), 918–926.

[34] Kerman, A. Corrosion management in offshore structures. In *Offshore Technology Conference* (2019).

[35] Khujaev, O., Nurmetova, B. B., and Urazmatov, T. K. Algorithms for selecting the most efficient method for solving classification problems. *IEEE Conference on Applied Electronics and Information Engineering* (2023), 1740–1743.

[36] Kim, K.-Y., and Kim, J.-B. Waterhammer caused by startup and stoppage of a centrifugal pump. *Journal of Fluid Machinery 7*, 1 (2004), 51–57.

[37] Kumar, R., Pati, P. B., and Deepa, K. Clustering the various categorical data : An exploration of algorithms and performance analysis. In *2023 IEEE International Conference on Next-Generation Computing (INCET)* (2023), pp. 1–6.

[38] Lee, M.-C., Lin, J.-C., and Stolz, V. Evaluation of k-means time series clustering based on z-normalization and np-free. *arXiv.org abs/2401.15773* (2024).

[39] Li, Z., Liu, L., Dong, C., and Shang, J. Overfitting or underfitting ? understand robustness drop in adversarial training. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence* (2020).

[40] Lima, S. P., and Cruz, M. D. A genetic algorithm using calinski-harabasz index for automatic clustering problem. *RBCA 12*, 3 (2020), 97–106.

[41] MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (1967).

[42] Mandke, J. S. Corrosion causes most pipeline failures in gulf of mexico. *Oil & Gas Journal 88*, 44 (1990).

[43] Moore, D. K., Jerrett, M., Mack, W. J., and Künzli, N. A land use regression model for predicting ambient fine particulate matter across los angeles, ca. *Journal of Environmental Monitoring 9*, 3 (2007), 246–252.

[44] Murphy, K. P. *Machine Learning : A Probabilistic Perspective*. MIT Press, 2012.

[45] Naïte. Pipeline : les éléments clés à retenir, 2025.

[46] Nengjun, B., Vytyaz, O., Jin, X.-M., Hrabovs'kyi, R. S., and , Failure analysis during the operation of offshore oil and gas structures. *Nafta Gaz 79*, 8 (2023), 529–536.

[47] Peabody, A. *Peabody's Control of Pipeline Corrosion*, 2nd ed. NACE International, 2001.

[48] Petrović, Z. C. Catastrophes caused by corrosion. *Vojnotehnički Glasnik 64*, 4 (2016), 1048–1064.

[49] Rabin, Z., Davis, J., Lewis, B., and Scherreik, M. Overfitting in contrastive learning ? *arXiv preprint arXiv :2407.15863* (2024).

[50] ROSEN Group. Inspection report – rocorr mfl-a line haoud el hamra – béjaïa. Rapport technique interne, ROSEN Group, 2010.

[51] Rosenberg, A., and Hirschberg, J. V-measure : A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (Prague, Czech Republic, 2007), Association for Computational Linguistics, pp. 410–420.

[52] Salman, S., and Liu, X. Overfitting mechanism and avoidance in deep neural networks, 2019.

[53] Shahapure, K. R., and Nicholas, C. Cluster quality analysis using silhouette score. In *IEEE International Conference on Data Science and Advanced Analytics* (2020), pp. 747–748.

[54] Singh, N., and Gaur, B. Data preprocessing : A step-by-step guide for clean and usable data. *Turkish Journal of Computer and Mathematics Education 10*, 2 (2019).

[55] Singh, R. *Pipeline Inspection and Health Monitoring*. Elsevier, Oxford, UK, 2021.

[56] Stewart, M. Piping standards, codes, and recommended practices. In *Handbook of Pipeline Engineering*. Gulf Professional Publishing, 2016, pp. 17–158.

[57] Sun, J. Analysis and countermeasures of mechanical equipment failure mode. *Coal Technology* (2012).

[58] Taghavi, S. F. Basic onshore pipeline mechanical design. In *Elsevier eBooks*. Elsevier, 2023, pp. 233–250.

[59] Tan, P.-N., Steinbach, M., Karpatne, A., and Kumar, V. *Introduction to Data Mining*, 2nd ed. Pearson, 2018.

[60] Tokareva, A. O., Chagovets, V., Kononikhin, A. S., Starodubtseva, N. L., Nikolaev, E. N., and Frankevich, V. Normalization methods for reducing interbatch effect without quality control samples in liquid chromatography-mass spectrometry-based studies. *Analytical and Bioanalytical Chemistry 413*, 13 (2021), 3479–3486.

[61] Tripathy, A. K., Nambiar, P., Pereira, A., D'souza, S., Rodrigues, L., D'souza, A., D'souza, B., and D'mello, B. Pressure surge analysis in pump systems. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)* (2015), pp. 1–5.

[62] Turyk, E. Manufacturing defects in welding consumables influencing the quality of welded joints. *The Paton Welding Journal 2014*, 6 (2014), 103–106.

[63] Twyman, J. Golpe de ariete en una red de tuberías debido al cierre rápido de una válvula. *Revista Ingeniería de Construcción 33*, 2 (2018), 193–200.

[64] Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research 11* (2010), 2837–2854.

[65] Wang, L., Yan, Y., Wang, X., and Wang, T. Input variable selection for data-driven models of coriolis flowmeters for two-phase flow measurement. *Measurement Science and Technology 28*, 3 (2017), 035305.

[66] Zhang, H., Zhang, L., and Jiang, Y. Overfitting and underfitting analysis for deep learning based end-to-end communication systems. *International Conference on Wireless Communications and Signal Processing (WCSP)* (2019), 1–6.

[67] Zhang, Y., Liu, M., Liu, Z., Wu, Y., Mei, J., Lin, P., and Xue, F. Pump-stoppage-induced water hammer in a long-distance pipe : a case from the yellow river in china. *Water Science & Technology : Water Supply 19*, 1 (2019), 216–221.

[68] Zheng, H. A review of evaluation metrics in machine learning algorithms. *Lecture Notes in Networks and Systems* (2023), 15–25.

[69] Zhou, Z.-H. *Machine Learning*. Springer, 2021.

Résumé

La fiabilité des pipelines constitue un enjeu stratégique dans l'industrie énergétique, notamment pour le transport du pétrole et du gaz. Face aux nombreuses défaillances possibles — en particulier la corrosion, responsable de près de 75% des cas — il devient essentiel de développer des méthodes de surveillance avancées. Ce mémoire propose une approche innovante fondée sur l'apprentissage automatique (machine learning) pour prédire les défaillances des pipelines. À partir de données issues d'une inspection MFL (Magnetic Flux Leakage) réalisée par ROSEN Group pour SONATRACH, plusieurs algorithmes de clustering non supervisé (K-Means, DBSCAN, GMM, et agglomératif) sont appliqués pour détecter automatiquement les zones critiques de corrosion. Le prétraitement des données, la sélection des métriques d'évaluation, ainsi que la cartographie des défauts permettent d'améliorer la compréhension et l'exploitation des données d'inspection. Les résultats démontrent l'efficacité de ces méthodes pour l'analyse prédictive des risques, contribuant ainsi à une meilleure planification de la maintenance et à la réduction des risques environnementaux et économiques.

Mots-clés : Corrosion, Pipeline, Inspection MFL, Apprentissage automatique, Clustering, DBSCAN, Maintenance prédictive, Risques, Classification des défauts, Sonatrach.

Abstract

Pipeline reliability is a strategic issue in the energy industry, particularly for oil and gas transport. In view of the numerous possible failures - in particular corrosion, responsible for almost 75% of cases - it is becoming essential to develop advanced monitoring methods. This thesis proposes an innovative approach based on machine learning to predict pipeline failures. Using data from a MFL (Magnetic Flux Leakage) inspection carried out by ROSEN Group for SONATRACH, several unsupervised clustering algorithms (K-Means, DBSCAN, GMM, and agglomerative) are applied to automatically detect critical areas of corrosion. Data pre-processing, selection of evaluation metrics and defect mapping improve understanding and exploitation of inspection data. The results demonstrate the effectiveness of these methods for predictive risk analysis, helping to improve maintenance planning and reduce environmental and economic risks.

Keywords : Corrosion, Pipeline, MFL Inspection, Machine Learning, Clustering, DBSCAN, Predictive Maintenance, Risk Assessment, Defect Classification, Sonatrach.