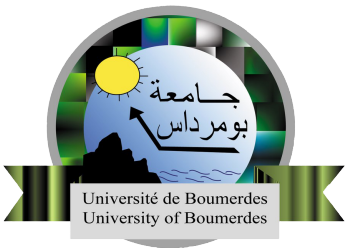
République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université M'hamed Bougara de Boumerdes

Faculté des science

Département de mathématiques



Modélisation de l'influence des paramètres physico-chimique sur la vitesse de corrosion Cas de SONATRACH

Mémoire Présenté Pour l'obtention du Diplôme de Master En Recherche Opérationnelle

Option : Recherche Opérationnelle, Optimisation et Management Stratégique

Réalisé par :	**Encadré par :**
CHABLI Saida	**Mme F. Gatt MAA**
KEBOUR Imane	**Mme D. Zerrouki SONATRACH**

Soutenu le 06 juillet 2023, devant le jury composé de

Mme N. Larbi	**MAA**	**UMBB**	**Présidente**
Mme M. Benmansour	**MAA**	**UMBB**	**Examinatrice**

Année Universitaire : 2022/2023

Remerciements

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui nous voudrions témoigner toute notre gratitude.

Nous tenons tout particulièrement à remercier notre encadreur **Mme.Gatt Fella** our sa patiente, sa disponibilité et ses judicieux conseils, qui ont contribué à alimenter notre réflexion. Merci aux membres du jury **Benmansour Madina et Larbi Noura** pour le grand honneur qu'ils nous font en acceptant de juger ce travail, nous vous remercions de votre enseignement et nous sommes très reconnaissantes de bien vouloir porter intérêt à ce travail.

Nous tenons à exprimer notre sincère gratitude à Monsieur **Laouzai Abdelfattah** pour ses précieux conseils tout au long de ce travail.

Un merci à tous les professeurs de l'Université M'Hamed Bougara de Boumerdes qui nous ont fait grandir durant ce parcours universitaire. Nous remercions nos très chers parents, nos frères et sœurs, qui ont toujours été là pour nous, merci pour leurs soutien constant et leurs encouragements.

Enfin, que toute personne ayant contribué, de près ou de loin, à la réalisation de ce travail soient chaleureusement remerciés.

Dédicaces

Je dédie ce travail à **mes chers parents**, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études.

A mes chers frères **Ibrahim, Madjid, Abdel kader et Soheib** qui ont été une source constante de soutien et d'encouragement tout au long de mon parcours. Leur présence et leur soutien ont été inestimables. Leurs paroles d'encouragement et de confiance en moi ont été une motivation première pour surmonter les défis et persévérer dans mes efforts.

A ma cousine **Nacira** qui m'a soutenu de ses conseils, je la remercie pour sa présence dans ma vie.

A mes amis **Fatma, Narimen, Ouarda, Saida ,Ahlem et Feriel** Avec lesquels j'ai passé de bons moments .

Imane.

Dédicaces

En ce moment précieux où je clôture ce chapitre de ma vie avec la réalisation de mon mémoire, je dédie humblement ce travail à ceux qui, peu importe les termes utilisés, ne pourront jamais comprendre l'amour sincère que j'ai pour eux.

À l'homme de ma vie, **mon cher père**, source inépuisable de force et de sagesse, ainsi qu'à **ma merveilleuse mère**, ma principale source d'amour et de soutien, je vous suis profondément reconnaissant pour tout ce que vous avez fait pour moi. Votre soutien constant, vos sacrifices et votre foi en mes capacités ont été mes plus grandes motivations. Votre amour indéfectible a été ma force dans les moments difficiles.

À mes frères, **Mohamed Farid** et **Karim**, et à mes sœurs, **Ouiza**, Fatima, **Nora**, **Sobhia**, **Nadia** et **Ouerdia**, ainsi qu'à mes beaux-frères et belle-sœur **Hafidha**, vous êtes mes complices, mes alliés et mes meilleurs amis.

Je tiens également à exprimer ma gratitude envers ma sœur **Sobhia** et son mari **Slimane**,qui ont généreusement consacré leur temps et leur expertise pour m'aider à perfectionner ce mémoire et m'encourager à donner le meilleur de moi-même.

À **mes neveux** et **nièces**, vous êtes les rayons de soleil qui illuminent ma vie.

Enfin à mes amies :**Dihia,Fouzia**, **Silia** et **Imane**.Vous avez été présentes à mes côtés tout au long de cette aventure.

Que cette dédicace soit le témoignage de ma gratitude et de mon amour envers chacun d'entre vous. Votre soutien indéfectible a été essentiel dans la réalisation de ce mémoire. Je vous suis profondément reconnaissant pour votre présence dans ma vie.

Saida.

Résumé

Dans ce travail, nous abordons le problème de la corrosion, un phénomène bien connu qui présente un risque pour notre environnement. L'objectif de cette étude est de modéliser l'influence des paramètres physico-chimiques sur la vitesse de corrosion.

Pour atteindre cet objectif, nous avons utilisé l'apprentissage automatique supervisé pour évaluer les performances des modèles existants de prédiction basés sur les données de SONATRACH. Nous avons mis en place deux modèles de régression : la régression linéaire et la régression polynomiale.

En utilisant ces modèles, nous avons pu prédire la vitesse de corrosion en fonction des paramètres physico-chimiques étudiés. Nous avons ensuite comparé les performances des deux modèles à l'aide d'une métrique appropriée et déterminé le meilleur modèle.

De plus, afin de faciliter l'utilisation de notre approche de prédiction, nous avons développé une interface graphique conviviale. Cette interface permet aux utilisateurs d'obtenir les meilleures prédictions pour la vitesse de corrosion en entrant les valeurs des paramètres physico-chimiques.

Mots clés : Apprentissage automatique, Régression linéaire , Régression polynomiale.

Table des matières

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GÉNÉRALE

La corrosion est un phénomène préoccupant dans de nombreux secteurs industriels, notamment dans l'industrie pétrolière. Elle engendre des pertes économiques considérables et représente un risque majeur pour la sécurité et l'intégrité des installations. Afin de prévenir et de gérer efficacement ce problème, il est essentiel de comprendre les mécanismes de corrosion, d'identifier les facteurs qui y contribuent et de développer des modèles prédictifs pour évaluer la vitesse de corrosion.

Ce mémoire de fin d'études se concentre sur l'étude de la corrosion dans l'industrie pétrolière, en mettant particulièrement l'accent sur l'entreprise Sonatrach. Sonatrach est une société algérienne spécialisée dans l'exploration, la production et la commercialisation des hydrocarbures. Elle joue un rôle majeur dans l'économie nationale et régionale, ce qui souligne l'importance de la gestion efficace de la corrosion pour préserver ses installations et ses actifs.

La premier chapitre de ce mémoire est consacré à la présentation de l'entreprise d'accueil Sonatrach, en mettant en évidence son historique, ses missions, ses objectifs et son organisation. Une vue d'ensemble des activités de Sonatrach est également abordée pour comprendre les enjeux et les défis auxquels l'entreprise est confrontée en termes de corrosion.

Le deuxième chapitre est consacré à l'étude bibliographique approfondie sur la corrosion. Nous abordons les différentes définitions de la corrosion, les causes qui la favorisent, ainsi que les différents types de corrosion, tels que la corrosion chimique, la corrosion électrochimique et la corrosion bactérienne. De plus, nous examinons les aspects morphologiques de la corrosion, en distinguant la corrosion généralisée de la corrosion localisée. Enfin, nous explorons les méthodes de mesure de la corrosion, en mettant en évidence les techniques d'immersion et les méthodes électrochimiques.

Le troisième chapitre porte sur les fondements théoriques de la régression et de la modélisation,

en particulier dans le contexte de la vitesse de corrosion. Nous présentons les concepts de base de la régression linéaire, en mettant en évidence la régression linéaire simple et la régression linéaire multiple. Les techniques d'estimation des coefficients de régression et l'évaluation de la qualité du modèle sont également abordées. De plus, nous explorons la régression non linéaire, en mettant en évidence la régression polynomiale univariée et multivariée

Le quatrième chapitre de ce mémoire se concentre sur l'implémentation des modèles de régression pour prédire la vitesse de corrosion dans l'industrie pétrolière. Nous décrivons les outils et l'environnement de travail utilisés, notamment le langage de programmation Python et les bibliothèques associées. Nous présentons ensuite un aperçu des données recueillies, ainsi que la méthode utilisée pour les lire à partir de fichiers Excel. Enfin, nous détaillons les étapes de modélisation et d'évaluation des paramètres physico-chimiques influençant la corrosion.

En conclusion, ce mémoire de fin d'études vise à approfondir la compréhension du phénomène de corrosion dans l'industrie pétrolière, en mettant l'accent sur Sonatrach. Nous explorons les différents aspects de la corrosion , ses mécanismes ainsi que les méthodes théoriques qui répondent à la résolution de la problématique posée à savoir prédire la vitesse de corrosion en tenant en compte des paramètres physico-chimiques. Les résultats obtenus peuvent contribuer au développement de stratégies de prévention et de maintenance prédictive, offrant ainsi des avantages économiques et de sécurité significatifs pour Sonatrach et d'autres acteurs de l'industrie pétrolière

CHAPITRE 1

PRÉSENTATION DE L'ENTREPRISE

Introduction

Dans ce premier chapitre, nous allons présenter l'entreprise qui nous a accueillis, à savoir la SONATRACH. Nous commencerons par décrire l'organisation de l'entreprise, en mettant en avant ses missions, ses fonctions, ses différentes directions ainsi que son organigramme.

1.1 sonatrach

FIGURE 1.1 – Logo de SONATRACH

1.1.1 Historique

SONATRACH est une entreprise algérienne créée en 1963 après l'indépendance pour prendre le contrôle des richesses pétrolières et gazières du pays. Elle a élargi ses activités dans l'industrie pétrolière en prenant des participations dans des concessions étrangères.

En 1971, elle a été chargée de gérer et de développer toutes les branches de l'industrie pétrolière et gazière algérienne.

En 1998, elle est devenue une société par actions et a lancé un emprunt obligataire en 1999.

Aujourd'hui, elle intervient également dans d'autres secteurs tels que la production d'électricité, les énergies renouvelables et le dessalement de l'eau de mer. Avec près de 120 000 employés, elle compte actuellement 16 filiales nationales et 24 filiales internationales dans l'exploitation, le raffinage, la commercialisation, le stockage et les services aux puits.

1.1.2 Les missions

En 1998, SONATRACH a été transformée en une société par actions entièrement détenue par l'état et réglementée par la législation en vigueur. La société s'est donné les missions stratégiques suivantes pour valoriser ses ressources en hydrocarbures :

•Prospection, recherche et exploitation d'hydrocarbures ;

•Développement, exploitation et gestion des réseaux de transport par canalisation, stockage et chargement portuaire d'hydrocarbures ;

•Transformation et raffinage d'hydrocarbures ;

•Liquéfaction de gaz naturel, traitement et valorisation d'hydrocarbures gazeux ;

•Commercialisation d'hydrocarbures ;

•Approvisionnement constant du pays en hydrocarbures ;

•Développement de toutes les formes d'activités conjointes en Algérie et hors d'Algérie avec des sociétés algériennes ou étrangères ;

•Acquisition et détention de tous les portefeuilles d'actions ;

•Prises de participations et autres investissements dans toute société existante ou à créer en Algérie ou à l'étranger ;

•Étude, promotion et valorisation de toute autre forme et source d'énergie ;

•Développement de toute activité ayant un lien direct ou indirect avec l'industrie des hydrocarbures.

Ces nouvelles missions ont exprimé la volonté de SONATRACH de devenir une entreprise mondiale en développant toutes les activités en amont et en aval de la chaîne des hydrocarbures à travers le monde.

1.1.3 Les objectifs

Les objectifs de SONATRACH en tant que groupe pétrolier international sont les suivants :

•Maintenir une maîtrise continue de ses métiers de base.

•Renforcer ses capacités techniques et ses méthodes de gestion.

•Diversifier son portefeuille d'activités.

•Développer des partenariats pour la recherche et la production de pétrole, ainsi qu'étendre les exportations de gaz.

•Optimiser la valorisation des produits pétroliers.

Ces objectifs visent à améliorer la compétitivité de SONATRACH sur le marché mondial en adaptant ses stratégies et en exploitant de manière optimale ses ressources humaines, ses moyens de production, ses moyens financiers et ses investissements planifiés.

1.1.4 L'organisation de SONATRACH

SONATRACH a réorganisé sa structure organisationnelle pour devenir un groupe pétrolier de renommée internationale et faire face aux changements constants du marché mondial des hydrocarbures. Cette réorganisation a permis une meilleure coordination et gestion de ses activités opérationnelles et fonctionnelles afin de maintenir sa position de leader dans l'industrie.

La nouvelle organisation de la macrostructure de SONATRACH a été conçue pour répondre aux évolutions de son environnement interne et externe, et pour atteindre ses objectifs. Elle vise à renforcer le rôle de la Direction Générale dans la conception de la stratégie, l'orientation, la coordination, le pilotage et la gestion de l'entreprise.

Cette réorganisation vise également à concentrer les structures opérationnelles pour une meilleure synergie et une efficacité accrue, tout en permettant une décentralisation accompagnée d'une clarté en matière de responsabilités et d'une maîtrise des pouvoirs.

En outre, la nouvelle organisation vise à assurer la réactivité, la transparence et la fluidité de l'information nécessaire pour la conduite et le pilotage des activités, dans le but d'assurer l'efficacité globale de l'entreprise.

Les structures de la nouvelle organisation de SONATRACH comprennent la Direction Générale, les Structures Fonctionnelles et les Structures Opérationnelles. Cette organisation permettra à SONATRACH de mieux répondre aux défis de l'industrie pétrolière et gazière, de maximiser sa productivité et de maintenir sa position de leader dans le secteur.

Direction générale

Dirigée par un présidant directeur général, assisté du comité exécutif. Il est également assisté de conseillers et de directeurs chargés du traitement et du suivi de dossiers spécifiques et à

caractère stratégique. D'autres comités sont rattachés à la direction générale :

•Le comité d'examen des projets (CEP).

•Le comité de coordination des projets (CIP).

•Le comité d'éthique.

Le président-directeur général peut à tout moment créer d'autres comités chargés d'assurer la coordination de l'étude de problèmes particuliers. Ces comités peuvent être chargés d'une mission permanente ou d'une durée déterminée. La direction relation publique (REP) et le service sûreté interne l'établissement (SIE) relèvent également de la direction générale.

structures opérationnelles

Les activités opérationnelles exercent les métiers du groupe et développent son potentiel d'affaires tant en Algérie qu'à l'étranger.

Les activités opérationnelles, qui sont placées sous l'autorité d'un vice-président sont :

•L'Activité Exploration-Production (EP) : Couvre les activités de recherche, d'exploration, de développement et de production d'hydrocarbures. Elles sont assurées par Sonatrach seule, ou en association avec d' compagnies pétrolières.

•L'Activité Liquéfaction, Raffinage et Pétrochimie (LRP) : Couvre le développement et l'exploitation des complexes de liquéfaction de gaz naturel, de séparation de GPL, de raffinage et des gaz industriels.

•L'Activité Transport par Canalisations (TRC) : Assure l'acheminement des hydrocarbures (pétrole brut, condensat, GPL et gaz naturel) et dispose d'un réseau de canalisations de près de 19 623 km en 2015 contre 14 915 en 2005, soit une augmentation de 4 708 km.

•L'Activité Commercialisation (COM) : A pour missions l'élaboration et l'application de la stratégie de Sonatrach en matière de commercialisation des hydrocarbures sur le marché intérieur et à l'international par les opérations de trading et de hipping. Ces opérations sont menées en coopération avec les filiales NAFTAL pour l'approvisionnement du marché national en produits pétroliers et gaziers (GPL), HYPROCSC pour le transport maritime de ces produits et COGIZ pour la commercialisation des gaz industriels.

Structures fonctionnelles

- Direction Corporate :

— Stratégie, Planification Économie (SPE)

— Finances (FIN)

— Ressources Humaines (RHU)

- Direction centrale :

— Filiales participations (FIP)

— Activités Centrales (ACT)

— Juridique (JUR)

— Informatique Système d'Information (ISI)

— Marchés et Logistique (MLG)

— Santé, sécurité environnement (HSE)

— Business Développent (BSD)

— Recherche Développement (RDT)

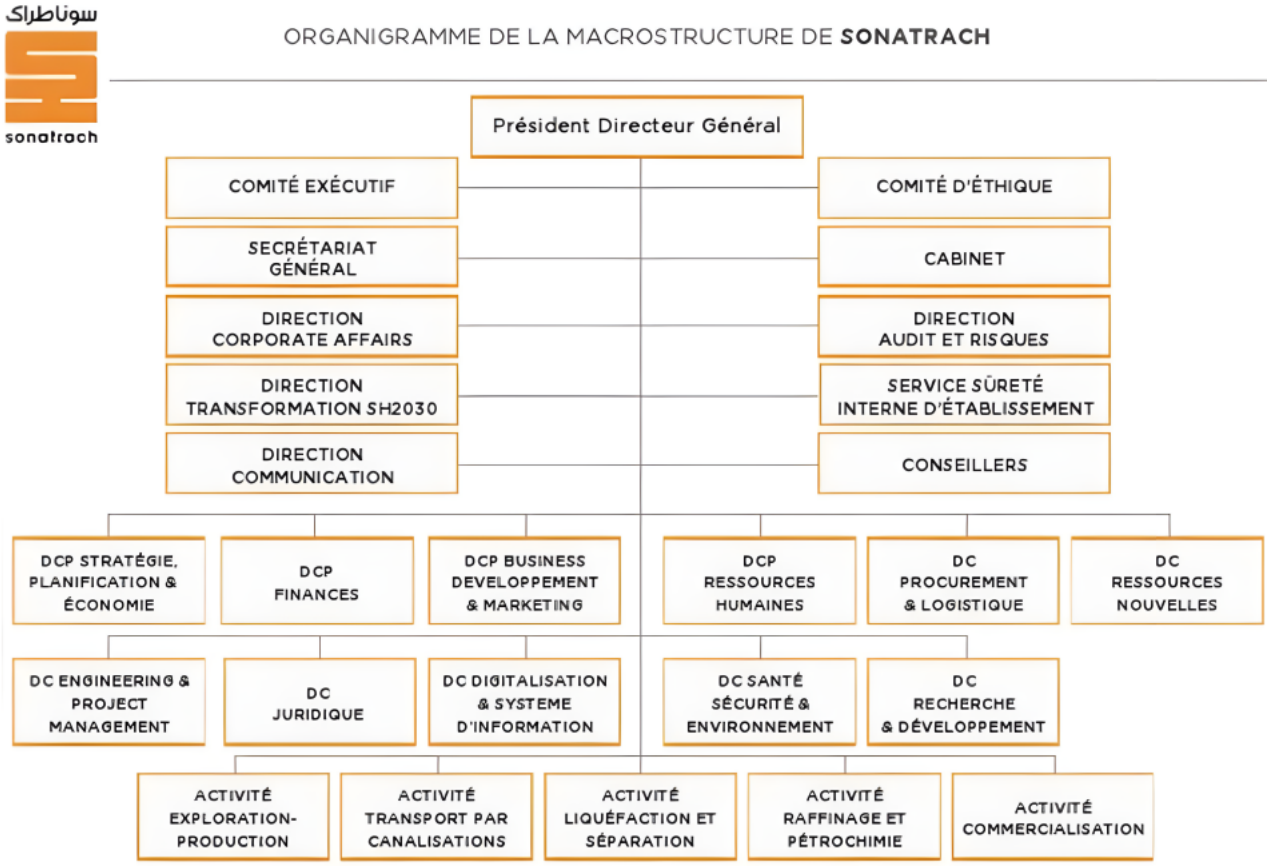


FIGURE 1.2 – Schéma organisationnel et fonctionnel de Sonatrach.

1.1.5 Activités de SONATRACH

Activités de SONATRACH	
Exploitation et production	SONATRACH a toujours mis l'accent sur la prospection de nouveaux gisements de pétrole et de gaz naturel sur le territoire national afin de renouveler ses réserves et d'accroître ses capacités de production.
Transport par canalisations	Le réseau de canalisations de SONATRACH en Algérie, qui s'étend sur environ 22 000 kilomètres, est utilisé pour transporter les hydrocarbures liquides et gazeux produits par l'activité d'exploration-production.
Liquéfaction et Séparation	Pionnier dans le GNL, SONATRACH s'est hissée parmi les tous premiers acteurs mondiaux dans la production et la commercialisation de produits à forte valeur ajoutée.
Raffinage et Pétrochimie	La mission de l'activité Raffinage-Pétrochimie consiste à transformer les matières premières en carburants pour répondre aux besoins du marché intérieur.
Commercialisation	Depuis plus de 50 ans, SONATRACH est un fournisseur clé de référence sur la scène européenne et internationale.

TABLE 1.1 – LES Activité SONATRACH

Conclusion

Sonatrach est la seule entreprise qui extrait et importe du pétrole et du gaz en Algérie, c'est donc une entreprise majeure dans l'économie algérienne puisque le pays dépend fortement des revenus du pétrole, d'ou les programmes de travail et les divisions bien organisés et structurés.

CHAPITRE 2

ETUDE BIBLIOGRAPHIQUE

Introduction

Dans ce chapitre, nous allons fournir des définitions simples de la corrosion, de ses différents types et formes, ainsi que des méthodes pour mesurer la vitesse de corrosion. Nous allons également expliquer le type de données sur lesquelles nous allons travailler dans le chapitre suivant.

Problématique

La corrosion est un processus électrochimique qui peut compromettre la résistance et la durabilité des matériaux, notamment des métaux. Dans le cas spécifique de l'acier au carbone, la corrosion peut être influencée par un certain nombre de paramètres physico-chimiques dans le milieu aqueux environnant. Parmi ces facteurs clés, nous pouvons citer le pH, la salinité, la température et la présence de CO2.

Dans le contexte du champ hassi R'mel, il est crucial de prendre en compte la composition chimique spécifique de l'eau et des fluides présents, ainsi que les conditions de pression et de température. Une approche de modélisation numérique peut être utilisée pour étudier l'impact de ces paramètres physico-chimiques sur la corrosion de l'acier au carbone dans ce champ. Cette modélisation prend en compte les interactions complexes entre les différentes espèces chimiques présentes dans la solution.

La modélisation numérique peut fournir des informations précieuses pour prévenir la corrosion de l'acier dans le champ hassi R'mel et prolonger ainsi la durée de vie des installations

pétrolières et gazières. En comprenant les mécanismes sous-jacents et en identifiant les conditions les plus favorables, des stratégies de protection et de maintenance appropriées peuvent être mises en place pour assurer l'intégrité des équipements et minimiser les risques de défaillance dus à la corrosion.

2.1 Définition de corrosion

La corrosion est un processus de dégradation des matériaux qui résulte de leur interaction chimique avec leur environnement. Ce phénomène est généralement considéré comme nuisible car il altère les propriétés physico- chimiques et mécaniques du matériau, le rendant ainsi inutilisable pour sa fonction d'origine.

Toutefois, il peut être considéré comme bénéfique dans certains cas, notamment lorsqu'il s'agit d'éliminer des objets abandonnés dans la nature.

De plus, la corrosion peut être utilisée de manière intentionnelle dans certains procédés industriels tels que l'anodisation de l'aluminium ou le polissage électrochimique [3].

2.2 Les types de corrosion

Selon le processus et le mécanisme de corrosion, il existe trois types de corrosions, la corrosion électrochimique, la corrosion chimique. et La corrosion bactérienne.

2.2.1 Corrosion chimique

La corrosion chimique correspond à une réaction entre un métal et un milieu liquide ou gazeux. Si cette corrosion survient à des températures élevées, elle est alors appelée "corrosion à haute température" ou "corrosion sèche". Au cours de la corrosion chimique, l'oxydation du métal et la réduction de l'oxydant se produisent simultanément. En d'autres termes, les atomes du métal réagissent directement avec l'oxydant, qui arrache les électrons de valence des atomes métalliques pour former des liaisons chimiques[8].

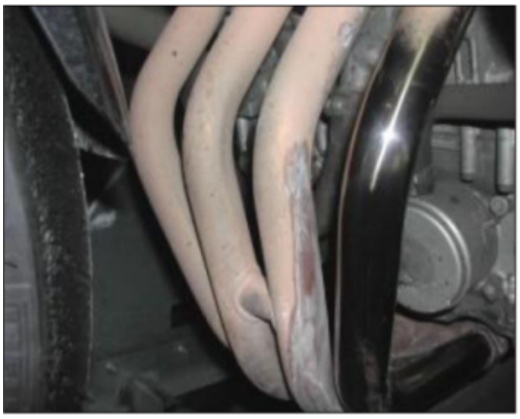


FIGURE 2.1 – exemple de la corrosion d'un collecteur d'échappement

[8]

2.2.2 Corrosion électrochimique

La corrosion électrochimique est le type de corrosion le plus fréquent et important. Elle est due à l'oxydation du métal en ions ou en oxydes et nécessite la présence d'un agent réducteur tel que H2O, O2 ou H2. Elle implique à la fois une réaction chimique et un transfert de charges électriques, ce qui entraîne la circulation d'un courant électrique. La corrosion électrochimique est une réaction d'oxydo-réduction où la réaction d'oxydation du métal est appelée réaction "anodique" et la réaction de réduction de l'agent oxydant est appelée réaction "cathodique". Ces deux réactions sont interdépendantes et ne peuvent pas se produire sans l'autre.[8]

2.2.3 Corrosion bactérienne

La corrosion bactérienne, également appelée biocorrosion, regroupe toutes les formes de corrosion impliquant l'action directe ou indirecte des bactéries et de leur métabolisme. Les bactéries jouent un rôle clé dans ce processus de corrosion en accélérant les réactions chimiques déjà établies ou en créant des conditions favorables à leur développement,
par exemple en produisant de l'acide sulfurique (H2SO4) dans certains cas [8].

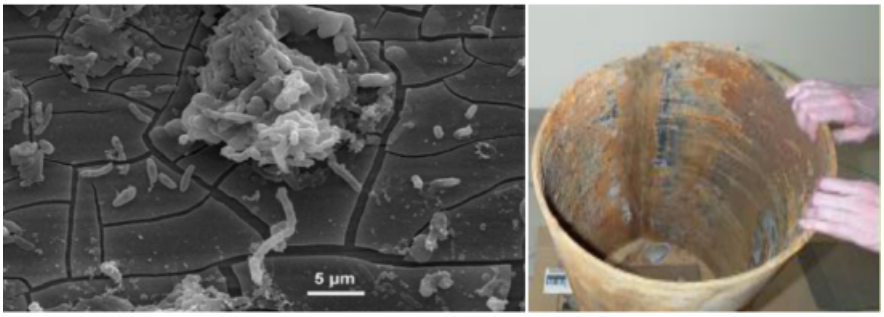


FIGURE 2.2 – Exemple de bactéries responsable de la corrosion bactérienne ; action des bactéries sulfato-rédutrices.

[8]

Dans le cadre de notre sujet, nous étudions le phénomène de corrosion électrochimique, qui est le type de corrosion le plus fréquent et important.

2.3 Aspect morphologique de la corrosion

En termes de détection, la corrosion est souvent évaluée visuellement, et selon son apparence morphologique, elle peut être catégorisée en deux grandes classes : la corrosion généralisée et la corrosion localisée, qui est également connue sous le nom de corrosion "zonale"[3].

2.3.1 Corrosion généralisée

La corrosion généralisée ou uniforme se produit lorsqu'un métal est en contact avec un environnement corrosif, et que la corrosion se propage à peu près à la même vitesse sur toute la surface du métal. Cette forme de corrosion est considérée comme la plus simple car elle se produit de manière homogène et régulière sur la surface du métal[3].

FIGURE 2.3 – Corrosion généralisée : exemple d'une d'une porte et d'un véhicule corrodes.

[3]

2.3.2 Corrosion localisée (zonale)

La corrosion localisée est l'un des modes de corrosion les plus courants et problématiques, car elle ne se produit que dans certaines zones spécifiques de la surface du matériau. Il existe plusieurs types de corrosion localisée, notamment :

Corrosion galvanique (bimétallique)

La corrosion galvanique, également connue sous le nom de corrosion bimétallique [3], est un phénomène qui survient lorsque deux métaux différents sont placés dans un même environnement. La phase métallique la moins noble est alors préférentiellement attaquée, formant ainsi une pile électrochimique. L'anode est constituée de la partie la moins noble, tandis que la cathode est constituée de la partie la plus noble. La surface relative de l'anode par rapport à celle de la cathode est un facteur important dans la vitesse de corrosion. Dans les alliages dentaires, le choix des métaux utilisés pour les restaurations prothétiques doit tenir compte de ce phénomène pour éviter la corrosion galvanique et ses conséquences néfastes [6].

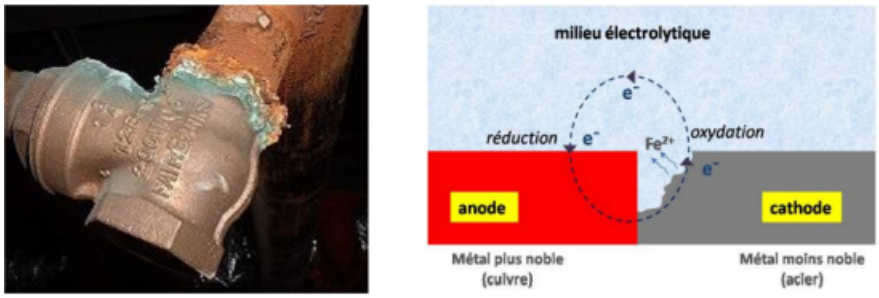


FIGURE 2.4 – Corrosion galvanique résultante d'un assemblage de deux métaux différents : robinet en cuivre et conduite en acier galvanisé.

[3]

Corrosion caverneuse (par crevasse)

La corrosion par crevasse, également connue sous le nom de corrosion caverneuse, se produit lorsque deux zones d'une structure métallique présentent une différence d'accessibilité à l'oxygène. Les parties métalliques les moins accessibles à l'oxygène sont alors attaquées de manière sélective [6].

Ce phénomène peut affecter tous les types de matériaux, y compris les joints souples, poreux ou fibreux tels que le bois, le plastique, le caoutchouc, le ciment, l'amiante, les tissus, etc [3].

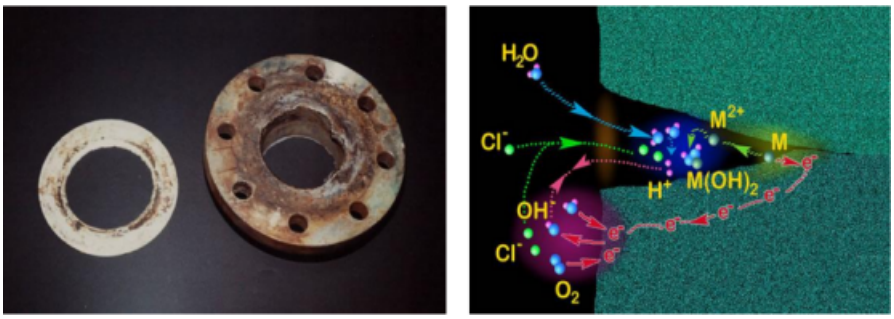


FIGURE 2.5 – Aspect et mécanisme d'attaque de la corrosion caverneuse.
[3]

Corrosion par piqûres

Dans certaines conditions environnementales, les métaux et alliages qui sont protégés par un film passif peuvent subir une attaque de piqûre. Ce phénomène survient lorsque le film protecteur se rompt localement, ce qui provoque la formation de piqûres localisées sur la surface métallique. Ces piqûres ont tendance à se propager progressivement et de manière sournoise, car l'hydrolyse des ions métalliques dissous à l'intérieur de la cavité créée augmente le degré d'acidité et favorise la corrosion [6].

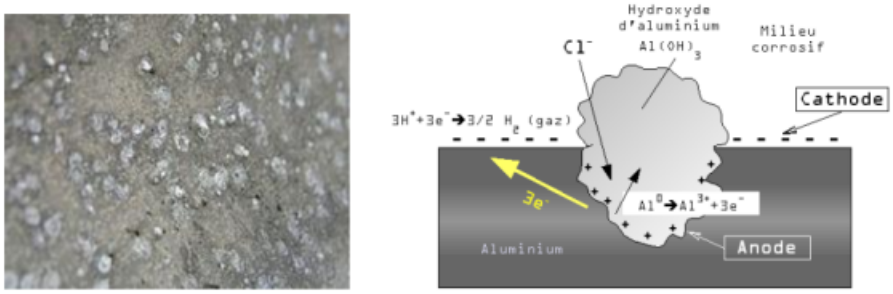


FIGURE 2.6 – Corrosion par piqûre de l'aluminium.
[3]

Corrosion intergranulaire

Est un phénomène d'attaque sélective qui affecte les joints de grains des matériaux métalliques. Ce phénomène est dû à des hétérogénéités locales qui provoquent l'appauvrissement ou l'enrichissement de certains constituants, tels que des précipités formés lors d'un traitement thermique. La présence de ces hétérogénéités crée des piles locales et entraîne la dissolution des zones anodiques, ce qui conduit finalement à la corrosion intergranulaire [6].

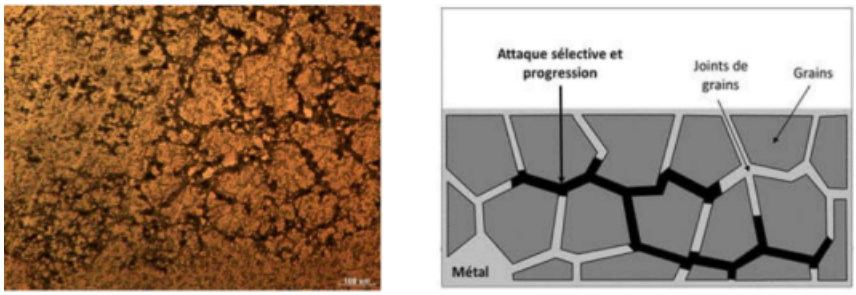


FIGURE 2.7 – Corrosion au niveau des joints de grains d'une structure métallique. [3]

Corrosion sélective

Est un type de corrosion où un élément ou une phase spécifique d'un alliage subit une dissolution sélective, ce qui peut entraîner la formation d'une structure métallique poreuse. Dans le cas d'un alliage homogène, un élément spécifique de l'alliage peut être dissous, tandis que dans un alliage polyphasé, l'une des phases peut subir une dissolution sélective. L'un des exemples les plus courants de corrosion sélective est la dézincification, où le zinc est sélectivement dissous dans un alliage de laiton [3].

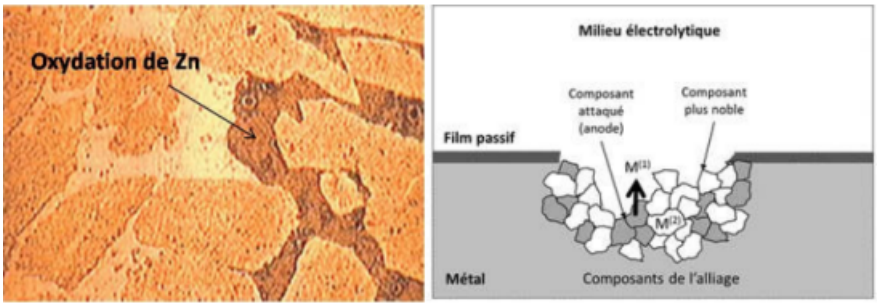


FIGURE 2.8 – Mécanisme de la corrosion sélective d'un laiton (alliage cuivre- zinc). [3]

Corrosion érosion

Est un type de corrosion qui résulte de l'interaction entre une réaction électrochimique et l'enlèvement mécanique de la matière. Elle se produit généralement sur des métaux exposés à un écoulement rapide de fluides tels que l'air ou l'eau. Bien que la plupart des métaux et alliages puissent être affectés par la corrosion érosion, les métaux plus tendres tels que le cuivre et le plomb sont particulièrement vulnérables, ainsi que ceux qui dépendent de la présence d'un film protecteur en surface tels que l'aluminium et les aciers inoxydables [3].

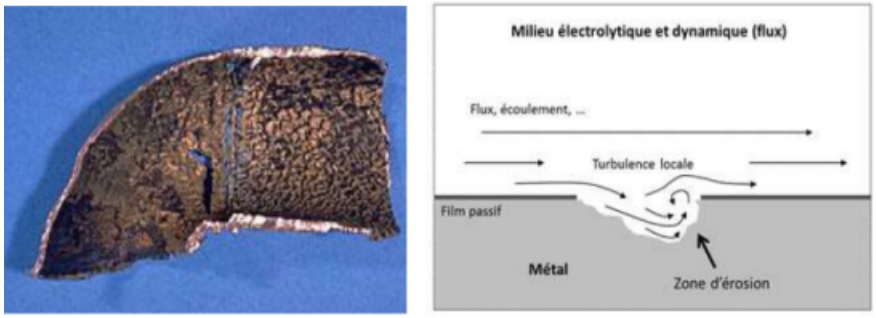


FIGURE 2.9 – Aspect et mécanisme de la corrosion érosion.

[3]

Corrosion frottement (tribocorrosion)

La corrosion frottement fait référence aux dommages causés par la corrosion lorsqu'il y a contact entre deux surfaces métalliques en mouvement relatif l'une par rapport à l'autre. Ce type de corrosion seproduit généralement lorsque l'interface est soumise à des vibrations et des charges de compression répétées. Si le mouvement de frottement est continu en présence d'un milieu corrosif, on parle plutôt de tribocorrosion [3].

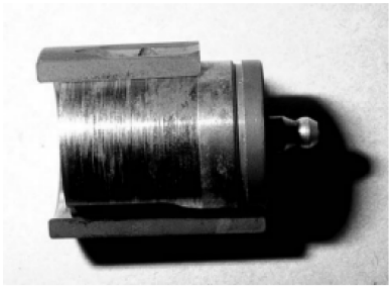


FIGURE 2.10 – La tribocorosion.

[3]

Corrosion sous contrainte

La corrosion sous contrainte et la fatigue-corrosion sont des phénomènes de fissuration des métaux qui résultent de l'action combinée d'une contrainte mécanique, généralement une force de traction, et d'une réaction électrochimique. La fatigue-corrosion se produit en raison d'une application répétée de contraintes mécaniques [6].

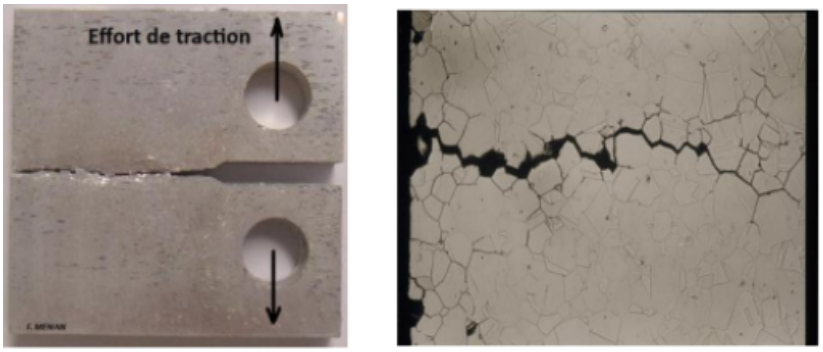


FIGURE 2.11 – La corrosion sous contrainte.

[3]

Fragilisation par hydrogène

Lorsqu'il y a de l'hydrogène dans un matériau métallique, cela peut causer une fragilisation due à une augmentation de la pression à l'intérieur du métal. Cette fragilisation peut éventuellement conduire à une rupture du matériau. L'hydrogène peut provenir de différentes sources, notamment de l'environnement atmosphérique, des procédés d'électrolyse et de la corrosion électrochimique [3].

FIGURE 2.12 – La fragilisation par hydrogène d'une pièce métallique.

[3]

2.4 La vitesse de corrosion

La vitesse de corrosion est la mesure de la quantité de métal qui est perdue ou consommée par unité de temps en raison de la corrosion.

La vitesse de corrosion est un phénomène complexe qui dépend de plusieurs facteurs interdépendants. La composition chimique du matériau joue un rôle important dans la vitesse de corrosion, car certains métaux et alliages peuvent être plus résistants à la corrosion que d'autres en raison de leur structure cristalline ou de leur comportement électrochimique.En outre, les facteurs environnementaux jouent un rôle important dans la vitesse de corrosion de l'acier au carbone.

•La température de l'environnement peut avoir un effet significatif sur la vitesse de corrosion, car elle peut affecter la cinétique de la réaction chimique. Une température plus élevée peut accélérer la réaction de corrosion, tandis qu'une température plus basse peut la ralentir.

•Le pH a un impact significatif sur la vitesse de corrosion. Dans les solutions acides, avec un pH inférieur à 7, la vitesse de corrosion de l'acier peut augmenter en raison de la concentration élevée d'ions hydrogène (H+), qui favorise la réaction de corrosion. En revanche, dans les solutions alcalines, avec un PH supérieur à 7, la vitesse de corrosion tend à diminuer en raison de la formation d'une couche protectrice à la surface de l'acier. Ainsi, le PH joue un rôle clé dans la régulation de la corrosion des matériaux métalliques.

•La salinité de l'eau , car elle peut affecter la concentration des ions chlorure dans la solution. Les ions chlorure sont connus pour être particulièrement corrosifs pour les métaux et peuvent accélérer la réaction de corrosion de l'acier au carbone.

•la présence de CO2 peut accélérer la vitesse de corrosion de l'acier au carbone, tandis que son absence peut avoir un effet bénéfique sur la corrosion de l'acier. Toutefois, les effets de la présence ou de l'absence de CO2 sur la corrosion de l'acier dépendent de plusieurs facteurs environnementaux et chimiques, tels que la salinité, la température et le pH.

En somme, la vitesse de corrosion dépend de plusieurs facteurs interdépendants. Une compréhension approfondie de ces facteurs est nécessaire pour développer des stratégies efficaces pour prévenir la corrosion de l'acier et prolonger sa durée de vie.

2.4.1 Mesure de la vitesse de corrosion

Pour toutes les réactions électrochimiques, la cinétique d'une réaction de corrosion peut être contrôlée, voire limitée par la vitesse d'une des étapes réactionnelles.

Corrosion contrôlée par le transfert de charge (électrons) à l'interface métal/solution. C'est le cas par exemple l'électrode du fer se corrodant dans un milieu acide.

Corrosion limitée par le transfert de matière (l'oxydant). C'est le cas de l'acier se corrodant dans un milieu aqueux neutre et aéré [9].

Pour déterminer la vitesse de corrosion en milieu liquide, Il existe deux types de méthode expérimentale[9] :

— Méthode par immersion

— Méthode électrochimique

Au niveau de SONATRACH la méthode électrochimique qui est utilisé.

Méthode par immersion

Dans cette méthode L'échantillon est placé sur un support non métallique, pesé au préalable (P_1), puis plongé dans une solution corrosive maintenue à température constante. Après une certaine période d exposition bien définie,l'échantillon est retiré puis nettoyé afin d éliminer les produits de corrosion. Une fois l'échantillon est bien nettoyé et pesé (P_2), on procède à la mesure de la perte du poids (m) de l'échantillon qui est égale à : $m = P_1 - P_2$. Le taux de corrosion (vitesse de corrosion) se définit comme une perte de poids par unité de surface et de temps [1].

Méthodes électrochimiques

Les essais par immersion ne donnent pas d'indication sur les mécanismes réactionnels, de plus ils sont extrêmement long dans les milieux peux corrosifs. Les essais électrochimiques n'ont pas ces inconvénients., on distingue deux méthodes expérimentales[1] :

•L'extrapolation des droites de Tafel.

•Mesure de la résistance de polarisation .

Extrapolation des droites de Tafel

Cette méthode consiste à utiliser les parties anodiques et cathodiques de la courbe de Tafel dans le but d'obtenir la valeur de potentiel de corrosion et la valeur correspondante de la densité de courant de corrosion[1]

L'équation de Butler-Volmer pour un système ne comporte qu'une seule réaction anodique et cathodique est comme suit :

$$i = i_{corr} exp(\frac{\eta}{\beta_\alpha}) - i_{corr} exp(\frac{-\eta}{\beta_c}) \tag{2.1}$$

L'extrapolation de la droite de Tafel vers le potentiel, fournit alors la valeur de la densité de courant de corrosion i_{corr}. Puis par la loi de Faraday on détermine la vitesse de corrosion de l'échantillon au repos dans la solution corrosive [1].

Par exemple : la région de Tafel cathodique de l'équation de Butler-Volmer correspond à l'équation suivant :

$$i = i_{corr} exp(\frac{-\eta}{\beta_c}) \tag{2.2}$$

Au potentiel de corrosion (η) , le courant de corrosion vaut $i = i_{corr}$ et la vitesse de corrosion :

$$v_{corr} = (\frac{i_{corr}}{nFA}) \tag{2.3}$$

$$i_{corr} = (\frac{I_{corr}}{A}) \tag{2.4}$$

i :densité du courante.

η :surtention.

v_{corr} :vitesse de corrosion(mm/ans).

A : surface de l'électrode de travail [cm^2].

F :nombre de Faraday = 96500 coulombs.

n :nombre d'électrons mise en jeux.

β_α : Coefficient anodique de Tafel.

β_c :Coefficient cathodique de Tafel.

Mesure de la résistance de polarisation

Cette technique de résistance électrique de polarisation évite certaines difficultés liées à la méthode d'extrapolation des droites de Tafel. La méthode de la résistance électrique de polarisation est utilisée dans le cas d'une surtension relativement faible, il est admissible de remplacer les exponentielles de la relation (2.1) par leurs développements limités au premier ordre ($e^x = 1 + x$ et $e^{-x} = 1 - x$)[9]. On obtient alors :

$$i = i_{corr} = [(1 + \frac{\eta}{\beta_a}) - (1 - \frac{\eta}{\beta_c})] \tag{2.5}$$

soit $i = i_0 \times \eta(\frac{\eta}{\beta_\alpha} + \frac{\eta}{\beta_c})$

On a donc une relation linéaire entre le courant et le potentiel et par analogie avec la loi d'Ohm on définit une résistance de polarisation R_p :

$$R_{p,corr} = \frac{1}{i_{corr}} \times \frac{\beta_a \times \beta_c}{\beta_a + \beta_c} \tag{2.6}$$

La mesure de la résistance de polarisation R_p en Ωm^2 au potentiel de corrosion permet de déterminer i_{corr} , puis V_{corr} [1].

2.5 Fondamentaux de la Prédiction

La vitesse de corrosion est influencée par des paramètres physico-chimiques , la collecte de données sur ces paramètres peut être affectée par des erreurs expérimentales, qui peut entraîner une mauvaise qualité des données et compromettre la capacité à extraire des informations fiables et significatives à partir de ces données. Pour remédier à ce problème, l'utilisation de l'intelligence artificielle et de l'apprentissage automatique est de plus en plus courante pour identifier et éliminer les valeurs aberrantes dans les données de corrosion.

2.6 L'intelligence artificielle (IA)

L'intelligence artificielle (IA) englobe la science et l'ingénierie qui permettent de développer des machines intelligentes, notamment des programmes informatiques intelligents. Bien qu'elle soit liée à l'objectif similaire d'utiliser des ordinateurs pour comprendre l'intelligence humaine,[13]

2.7 Apprentissage automatique

L'apprentissage automatique (en anglais : machine learning) apprentissage artificiell ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'apprendre à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune.

Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentations de telles méthodes.

2.8 Principe de l'apprentissage automatique

Le principe de l'apprentissage automatique est d'utiliser des algorithmes pour permettre à un ordinateur d'apprendre à partir de données sans être explicitement programmé. Le but de l'apprentissage automatique est de trouver des modèles dans les données qui peuvent être utilisés pour faire des prédictions ou prendre des décisions.

2.9 Types d'apprentissage automatique

Il existe plusieurs types de système d'apprentissage automatique .Danse ce qui suit,nous les classons selon qu'ils nécessitent ou non une supervision humaine (supervisés,non supervisés,semi supervisés et apprentissage par renforcement)

2.9.1 Apprentissage supervisé

L'apprentissage supervisé est un type d'apprentissage automatique dans lequel la machine apprend à partir d'ensembles de données connus (ensemble d'exemples d'apprentissage), puis prédit la sortie à partir d'ensembles de données d'entrées connus (ensemble d'exemples d'apprentissage). Un agent d'apprentissage supervisé doit trouver la fonction qui correspond à un ensemble d'échantillons donnés.

L'objectif de l'apprentissage supervisé est de construire un systéme artificiel capable d'apprendre la correspondance entre l'entrée et la sortie, et de prédire la sortie du systéme en fonction de nouvelles entrées.

Il est divisé en deux méthodes :

— Les algorithmes de **classification** , qui cherchent à prédire une classe/catégorie.

— Les algorithmes de **régression** , qui cherchent à prédire une valeur continue, une quantité.

Par exemple, Une tâches typique consiste à prédire une valeur numérique cible, tel que le prix d'une voiture, en fonction d'un ensemble de caractéristiques (kilométrage, âge, marque, etc.) appelées prédicteurs.

Le schéma ci-dessous qui illustre ce concept utilise uniquement une donnée en entrée et une en sortie.Dans la pratique ,les régression utilisent plusieurs paramètres en entrée.

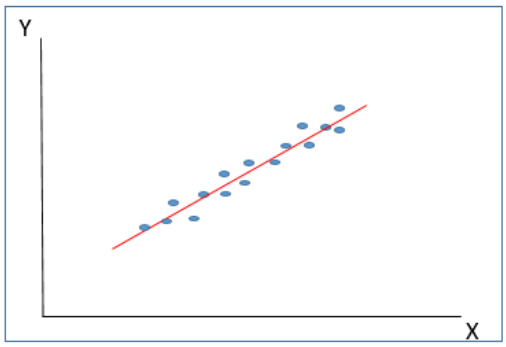


FIGURE 2.13 – Régression linéaire

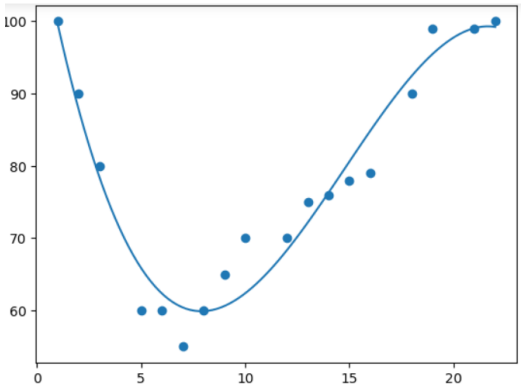


FIGURE 2.14 – Régression non linéaire

2.9.2 Apprentissage non supervisé

Dans l'apprentissage non supervisé, les données d'apprentissage ne sont pas étiquetées. Le système tente d'apprendre sans professeur.

Le modèle n'a pas de « réponses » dont il peut tirer des enseignements ; il doit donner un sens aux données en fonction des observations elles-mêmes. L'apprentissage non supervisé nous permet d'aborder les problèmes avec peu ou pas d'idée de ce à quoi nos résultats devraient ressembler. Nous pouvons obtenir une structure à partir de données dont nous ne connaissons pas nécessairement l'effet des variables[2].

2.9.3 Apprentissage semi-supervisé

C'est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non étiquetées.

Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées. Il a été démontré que l'utilisation de données non étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage.

2.9.4 Apprentissage par renforcement

Avec l'apprentissage par renforcement la machine n'a pas besoin de l'aide de l'être humain, ni en termes de supervision, ni en termes de fourniture de données.

L'apprentissage par renforcement est une branche très différente. Le système d'apprentissage, appelé un agent dans ce contexte , peut observer l'environnement, sélectionner et effectuer des actions, et enfin obtenir des récompenses ou des pénalités (des récompenses négatives). La machine peut apprendre toute seule la meilleure stratégie à suivre, appelée une politique, pour obtenir plusieurs récompenses au fil du temps. Une politique définit l'action que l'agent devrait choisir lorsqu'il est dans une situation donnée [4].

2.10 Les techniques pour développer des modèles prédicatifs

L'apprentissage automatique a différents techniques pour développer des modèles prédicatifs. Dans ce travail on utilise la technique de régression, qui permet de trouver un modèle en fonction des données d'entraînements.

Le modèle calculé permettra de donner une estimation sur une nouvelle donnée non encoure vue par l'algorithme.

Conclusion

Le transport du pétrole est une activité cruciale pour Sonatrach, mais l'existence des produits chimiques dans les pipelines peut accélérer la corrosion et entraîner des pertes si les dommages ne sont pas détectés à temps.

Le prochain chapitre expliquera en détail l'utilisation de la modélisation de régression pour prédire la vitesse de corrosion, ce qui permettra de prévenir les dommages et de découvre la durée de vie des pipelines.

CHAPITRE 3

MODÈLE DE RÉGRESSION ET FORMULATION

Introduction

Ce chapitre se concentre sur l'utilisation de la régression comme méthode d'analyse dans notre étude. Nous explorerons les techniques de régression linéaire simple et multiple, qui sont couramment utilisées pour estimer les valeurs de sortie à partir des caractéristiques d'entrée. De plus, nous aborderons la régression non linéaire et la régression polynomiale comme alternatives pour modéliser la vitesse de corrosion. Ces techniques nous permettront de découvrir la relation entre les variables et d'obtenir des estimations précises.

3.1 régression linéaire

3.1.1 Principe du modèle

Le modèle de régression linéaire est un modèle de machine learning dont la variable cible (Y) est quantitative tandis que la variable X peut être quantitative ou qualitative.

L'objectif est de trouver une fonction de prédiction ou une fonction coût qui décrit la relation entre X et Y . Autrement dit,donner une prédiction des valeurs de Y à partir des valeurs connues de X.

3.2 Régression linéaire simple

3.2.1 Présentation du modèle

Le modèle de régression linéaire simple décrit la liaison entre une variable expliquer Y, dite variable dépendante, et une variable explicative X,dite variable indépendante , cette liaison se formule comme suit :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \forall i = 1, ..., n \tag{3.1}$$

où

•Y_i : variable à expliquer ;

•X_i : variable explicative ;

•β_0, β_1 : paramètres du modèle estimer ;

•ε_i : erreur de spécification (différence entre le modèle vrai et le modèle spécifié), cette erreur est inconnue et restera inconnue ;

•n : nombre d'observations.

3.2.2 Méthode des moindres carrés Ordinaires

La méthode des moindres carrés Ordinaires est une technique d'estimation utilisée pour rechercher une fonction affine entre X et Y ce qui revient à chercher une droite qui s'ajuste le mieux possible au nuage de points.parmi toutes les droites possibles on retient celle qui minimise la somme des carrés des écarts de valeurs observée y_i à la droit de régression $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

•$\hat{y}_i$:représentent les valeurs prédites de y_i obtenues à partir de l'équation de la droite de régression .

Donc,le principe de la méthode MCO est de trouver β_0 et β_1 qui minimisent la quantité :

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{3.2}$$

Le minimum s'obtient par la résolution du système :

$$\begin{cases} \frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_1} = 0 \end{cases} \tag{3.3}$$

On obtient

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases} \quad (3.4)$$

Une formule équivalente à $\hat{\beta}_1$ est :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3.5)$$

•$\bar{y}$:c'est la moyenne des y_i,tel-que

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Donc, la droite de régression est donnée par :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

3.2.3 Équation d'analyse de la variance

variation expliquée et inexpliquée :

Le but d'un modèle de régression linéaire est d'expliquer une partie de la variation de la variable Y du fait de sa dépendance de la variable X.

•si X varie, Y varie en conséquence, il s'agit de la variation expliquée par le modèle.

•si X est fixe, Y varie encore, on dira qu'il s'agit de la variation inexpliquée par le modèle.

On a donc la situation suivante :

(variation totale de Y)=(variation expliquée par le modèle)+(variation inexpliquée par le modèle)

L'équation fondamentale d'analyse de la variance est :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y})^2 \quad (3.6)$$

ou encore SCT = SCE + SCR

•SCT : somme des carrées totale.

•SCE : somme des carrées expliquée.

•SCR : somme des carrées résiduelle.

3.2.4 Coefficient de détermination

Le coefficient de détermination R^2 est un indice qui mesure la qualité de l'ajustement réalisé par le modèle de régression.

Le R^2 se calcule à partir de la formule suivante :

$$R^2 = \frac{SCE}{SCT} = \frac{SCT - SCR}{SCT} = 1 - \frac{SCR}{SCT} \tag{3.7}$$

Le coefficient de détermination se situe entre 0 et 1 :

•Une valeur de R^2 proche de 1 montre que l'ajustement est bon.

•Une valeur de R^2 proche de 0 montre que l'ajustement est mauvais.

3.2.5 Erreur absolue moyenne (EAM)

L'erreur absolue moyenne (EAM), également connue sous le nom de Mean Absolute Error (MAE) en anglais, est une mesure de l'erreur moyenne entre les valeurs prédites et les valeurs réelles. Elle est utilisée pour évaluer la précision d'un modèle de prédiction ou d'estimation.

La formule pour calculer l'EAM est la suivante :

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.8}$$

Erreur quadratique moyenne (EQM)

L'erreur quadratique moyenne (EQM), ou Mean Squared Error (MSE) en anglais, est une mesure couramment utilisée pour évaluer la performance d'un modèle de régression. Elle est définie comme la moyenne des carrés des différences entre les valeurs prédites par le modèle et les valeurs réelles.

Mathématiquement, l'EQM est calculée comme suit :

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.9}$$

L'EQM mesure l'écart quadratique moyen entre les prédictions du modèle et les vraies valeurs. Une valeur d'EQM plus faible indique une meilleure adéquation du modèle aux données.

Droite de régression

Afin d'analyser l'influence de la variable X sue la variable Y. Un modèle mathématique doit être établi pour décrire la relation entre ces deux variables.

Considérons un tableau de données de un échantillon de taille "n", représentant la relation entre la variable indépendante X et la variable dépendante Y. Le nuage de points relatif à ces données est représenté graphiquement comme suit :

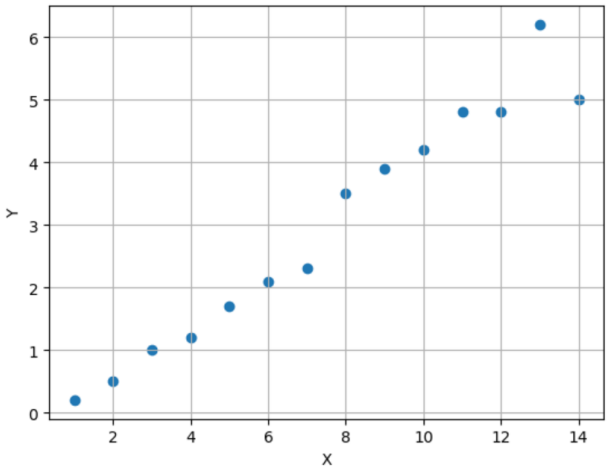


FIGURE 3.1 – Nuage de points

En utilisant la méthode des moindres carrés ordinaires, nous avons pu estimer les valeurs de l'ordonnée à l'origine (β_0) et de la pente (β_1) de la droite de régression linéaire.
Ces valeurs sont importantes car elles déterminent la position et l'inclinaison de la droite de régression, qui est utilisée pour prédire les valeurs de la variable dépendante en fonction des valeurs de la variable indépendante.

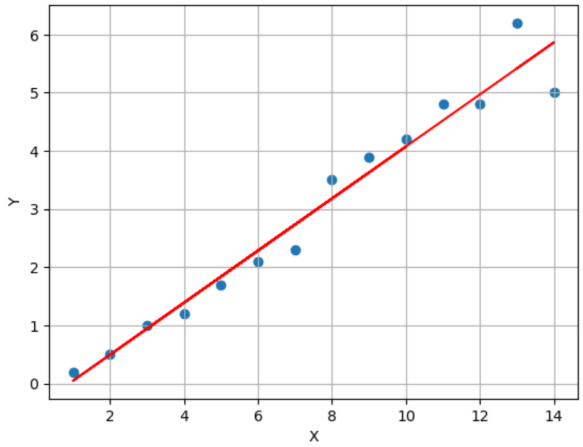


FIGURE 3.2 – Nuage de points et droite de régression.

La droite de régression linéaire (Figure 3.2) est un outil important pour analyser la relation entre deux variables. Une fois que la droite de régression est tracée, il est possible de projeter les points de données sur la droite et d'évaluer la qualité de l'ajustement du modèle. Cela peut aider à identifier les points qui s'éloignent considérablement de la droite et à évaluer leur influence sur l'analyse.

La figure 3.3 montre la projection des points de données sur la droite de régression .

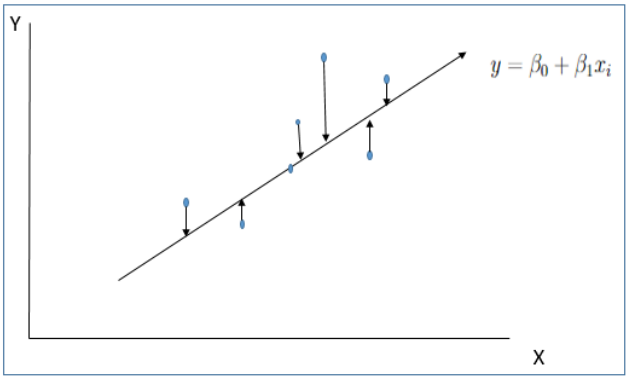


FIGURE 3.3 – Illustration des résidus ordinaires ajoutés à la droite de régression.

Une fois que les coefficients ont été estimés, on évalue la qualité du modèle de régression linéaire. Cela peut être fait en examinant des mesures telles que le coefficient de détermination (R^2), l'erreur quadratique moyenne (EQM) ou en effectuant des tests statistiques pour évaluer l'adéquation du modèle.

3.3 Régression linéaire multiple

3.3.1 Présentation du modèle

Le modèle linéaire multiple est une généralisation du modèle de régression simple dans lequel figurent plusieurs variables explicatives :

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon_i \tag{3.10}$$

Avec $i = 1, ..., n$

Y_i : Variable à expliquer ;

X_1 : $1^{ère}$ variable explicative 1 ;

X_2 : $2^{ème}$ variable explicative 2 ;

$\vdots$

X_k : $k^{ème}$ variable explicative k ;

$\beta_1, \beta_2, , , , \beta_k$: Paramètres du modèle à estimer ;

ε_i : erreur de spécification (Différence entre le modèle vrai et le modèle spécifié), cette erreur est inconnue et restera inconnue ;

n : Nombre d'observation.

L'écriture matricielle du modèle linéaire général :

$$\mathrm{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} ; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} ; \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} ; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

donc :

$$Y = X\beta + \varepsilon$$

3.3.2 Estimation des coefficients de régression

Afin d'estimer le vecteur des paramètres beta, nous appliquons la méthode des Moindres Carrés Ordinaires MCO, qui consiste à minimiser

$\sum_{i=1}^{n} \varepsilon_i^2$,soit :

$$Min \sum_{i=1}^{n} \varepsilon_i^2 = Min\varepsilon'\varepsilon = Min(Y - X\beta)'(Y - X\beta) = MinS \tag{3.11}$$

Avec ε' : transposé du vecteur ε

Pour minimiser cette fonction par rapport à β, nous différencions S par rapport à β .

$$\frac{\partial S}{\partial \beta} = -2X'X\hat{\beta} = 0 \Longrightarrow \hat{\beta} = (X'X)^{-1}X'Y \tag{3.12}$$

•X' : Matrice transposée de la matrice des variables explicatives.

•$(X'X)^{-1}$: Matrice inverse de la matrice.

Le modèle estimé s'écrit :

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} + \hat{\varepsilon}_i$$

Avec $\hat{\varepsilon}_i = y_i - \hat{y_i}$ où $\hat{\varepsilon}_i$ est le résidus, c'est-à-dire l'écart entre la valeur observée de la variable

à expliquer et sa valeur estimée (ajustée).

•Il convient de bien distinguer entre l'erreur de spécification du modèle (noté ε_i) qui est et restera inconnue et le résidu ($\hat{\varepsilon_i}$) qui lui est connu.

3.3.3 Équation d'analyse de la variance

Nous avons les relations suivantes :

$\sum_i y_i = \sum_i \hat{y}_i \Longrightarrow \bar{y} = \bar{\hat{y}}$

$\sum_i \hat{\varepsilon_i} = 0$

De ces deux relations, nous en déduisons l'équation fondamentale d'analyse de la variance :

$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i \varepsilon_i^2$

SCT=SCE+SCR

La variabilité totale (SCT) est égale à la variabilité expliquée (SCE) + la variabilité des résidus (SCR).

3.3.4 Coefficient de détermination

Il indique la quantité de l'ajustement y par $\hat{y}$, avec :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{\sum_{i=1}^{n}(\varepsilon_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{SCE}{SCT} \tag{3.13}$$

•Une valeur de R^2 proche de 1, montre que l'ajustement est bon.

•Une valeur de R^2 proche de 0, montre que l'ajustement est mauvais.

•Si R^2=1, alors $y_i = \hat{y}_i$, donc l'ajustement est parfait.

Erreur Quadratique Moyenne

L'erreur quadratique moyenne (EQM) est la fonction coût la plus couramment utilisée dans la régression multiple, elle mesure la moyenne des carrés des différences entre les valeurs prédites et les valeurs réelles. Sa formule est la suivante :

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

3.4 Formulation des paramètres de régression pour la vitesse de corrosion

3.4.1 Énonce du problème

Comme souligné dans le chapitre 2 le problème consiste à déterminer une description analytique et numérique pour la vitesse de corrosion(V-corr(mm/ans)) dans les pipes pétrolier en se basant sur les paramètres physico-chimiques suivants :

•Température :T(°C)

•Salinité :S(g/l)

•Potentiel d'hydrogène :PH

•Dioxyde de carbone :CO_2(bar)

Il s'agit d'expliquer la vitesse de corrosion par les quatre variables explicatives :T,S,PH et CO2. Cette relation est traduit mathématiquement par le modèle suivant :

$$V - corr = \beta_0 + \beta_1 T + \beta_2 S + \beta_3 PH + \beta_4 CO_2 + \varepsilon_i$$

tels-que les β_i , i=0,...,4 sont les paramètres à estimer inconnus du modèle et epsilon le résidu (partie inexpliquée).

3.5 Régression non linéaire

La régression non linéaire a pour but d'ajuster un modèle non linéaire pour un ensemble de valeurs afin de déterminer la courbe qui se rapproche le plus de celle des données de Y en fonction de X.

3.5.1 Forme générique

La régression non linéaire est utilisée lorsque l'on envisage une relation non linéaire entre les variables. Son modèle générique peut être exprimé par l'équation :

$$Y = f(X_1, \ldots, X_p, \beta_0, \ldots, \beta_q) + \varepsilon \tag{3.14}$$

où

•f :est la fonction de régression, la plupart du temps non linéaire. Elle dépend d'une variable

réelle X et du vecteur des paramètres β ;

•$\beta_0, ..., \beta_q$:sont q + 1 coefficients réels inconnus ;

•ε :est une variable quantitative de valeur moyenne nulle, indépendante de $X_1, ..., X_p$;

3.5.2 Régression polynomiale

La régression polynomiale est une forme d'analyse de régression dans laquelle la relation entre les variables indépendantes et les variables dépendantes est modélisée par un polynôme de degrés "n" .

Modèle de régression polynomiale

Un modèle de régression polynomiale est un modèle de régression non linéaire où la fonction $f(X_1, ..., X_p, \beta_0, ..., \beta_q)$ est un polynôme en $X_1, ..., X_p$ avec des coefficients $\beta_0, ..., \beta_q$. On utilise cette représentation polynomiale en se basant sur l'intuition, le contexte de l'expérience ou des critères statistiques spécifiques tels que on peut vous poser une question sur ça : qu'est que les résidus partiels ?.[5] Le modèle de régression polynomiale à une variable peut s'exprimer comme suit :

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k \quad (\text{pour } i = 1, 2, \ldots, n) \tag{3.15}$$

où k est le degré du polynôme. [14]

3.5.3 Régression polynomiale multivariable

La régression polynomiale peut être utilisée soit avec une seule variable prédictive, ce qu'on appelle la régression polynomiale simple, soit avec plusieurs variables prédictives, ce qu'on appelle la régression polynomiale multiple. Pour un modèle de régression polynomiale multiple d'ordre deux et k=2, l'équation peut être exprimée de la manière suivante :[15]

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2$$

Pour un modèle de régression polynomiale d'ordre m et de degré k, l'équation peut être

exprimée comme suit :

$$\hat{y} = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{m} \beta_i x_i^2 + \sum_{i=1}^{m} \sum_{j=1}^{m} \beta_i x_i x_j + \sum_{i=1}^{m} \beta_i x_i^3 + \sum_{i=1}^{m} \sum_{j=1}^{m} \beta_i x_i x_j^2 + ... + \sum_{i=1}^{m} \sum_{j=1}^{m} ... \sum_{l=1}^{m} \beta_i ... x_i x_j ... x_l + ... \tag{3.16}$$

Méthode d'estimation des paramètres

Il y a deux approches qui permettent d'estimer des modèles non linéaires, à savoir :

•Approche intuitives :Pour le cas du modèle logistique.

•Approche purement mathématiques de résolution d'équations non linéaires . [12]

Dans ce mémoire, nous avons choisi d'appliquer l'algorithme de Gauss-Newton pour l'estimation des paramètres du modèle.

3.5.4 Algorithme de Gauss-Newton

L'algorithme de Gauss-Newton est une méthode de résolution des problèmes de moindres carrés non linéaires.

Comme en régression linéaire, les paramètres d'un modèle de régression polynomiale sont estimés en minimisant la somme des carrés des résidus du modèle. C'est-à-dire qu'on cherche à minimiser l'expression suivante [12] :

$$SCR(\beta) = \sum_i (Y_i - f(X_i, \beta))^2$$

Pour cela, il faut dériver cette somme par rapport à chacun de ses paramètres et chercher les solutions qui annulent les dérivés. On peut réécrire ceci sous forme vectorielle :

$$SCR(\beta) = (Y - f(x, \beta))(Y - f(x, \beta))' = YY' - 2Yf(x, \beta)' + f(x, \beta)f(x, \beta)'$$

On dérive cette expression par rapport à toutes les composantes du vecteur β à p paramètres, et on annule toutes les dérivées partielles[12] :

$$\frac{\partial S}{\partial \beta} = -2\frac{\partial f(X, \beta)}{\partial \beta}(Y_i - f(X, \beta)) \tag{3.17}$$

Avec :

$$\frac{\partial \mathbf{f}(\mathbf{x}, \beta)}{\partial \beta} = Z(\beta) = \begin{bmatrix} \frac{\partial f(x_1, \beta)}{\partial \beta_0} & \dots & \frac{\partial f(x_1, \beta)}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(x_n, \beta)}{\partial \beta_0} & \dots & \frac{\partial f(x_n, \beta)}{\partial \beta_k} \end{bmatrix}$$

Notons : $Z(\beta^*)$ = matrice calculée pour les valeurs particulières de a ($\beta = \beta^*$). Grâce au développement limité de taylor au voisinage de β^*, nous obtenons l'approximation suivante à la ième observation :

$$f(x_i, \beta) \approx f(x_i, \beta^*) + \left\{ \frac{\partial f(x_i, \beta)}{\partial \beta_0}\bigg|_{\beta=\beta^*} \cdots\cdots \frac{\partial f(x_i, \beta)}{\partial \beta_k} \right\} \cdot (\beta - \beta^*) \quad (3.18)$$

Autrement, en notation matricielle :

$$f(X, \beta) \approx f(X, \beta^*) + Z(\beta^*)(\beta - \beta^*).....(1)$$

ou encore :

$$y = f(X, \beta^*) + Z(\beta^*)(\beta - \beta^*) + e....(2)$$

En posant : $\bar{y}(\beta^*) = y - f(X, \beta^*) + Z(\beta^*)\beta^*$ alors (2) se réduit en :

$$\bar{y}(\beta^*) = Z(\beta^*)\beta + e....(3)$$

L'estimateur du modèle linéaire (3) par MCO est : :

$$\beta^2 = [Z(\beta^*)'Z(\beta^*)]^{-1}Z(\beta^*)'\bar{y}(\beta^*) = \beta^* + [Z(\beta^*)'Z(\beta^*)]^{-1}Z(\beta^*)'[y - f(X, \beta)]$$

On aura k+1 nouvelles valeurs pour le vecteur ($\beta = \beta^2$). La convergence est atteinte lorsque $\hat{\beta} = \beta^p \approx \beta^{p-1}$ (stabilité des coefficients à la p-ième itération).

3.6 Avantages et Inconvénients de l'algorithme

Avantages

- La modélisation des prévisions sous forme de somme pondérée rend transparente la manière dont les prévisions sont produites.
- Il est simple d'estimer les poids et on a la garantie de trouver les poids optimaux (toutes les hypothèses du modèle de régression linéaire sont respectées par les données).

Inconvénients

- Les modèles de régression linéaire ne peuvent représenter que des relations linéaires, c'est-à-dire une somme pondérée des entités en entrée. Chaque non-linéarité ou interaction doit être créer à la main et explicitement attribuée au modèle en tant quantité en entrée.
- Les modèles linéaires ne sont souvent pas très utiles en ce qui concerne les performances prédicatives , car les relations pouvant être apprises sont tellement restreintes et simplifient généralement à l'extrême la complexité de la réalité.

3.7 Prévision

Un des buts de la régression est de faire de la prévision, c'est-à-dire de prévoir la variable à expliquer Y en présence d'une nouvelle valeur de la variable explicative X[7].
Soit donc X_{n+1} une nouvelle valeur, pour laquelle nous voulons prédire Y_{n+1}. Le modèle s'écrit comme suit :

$$Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}$$

Conclusion

Nous avons examiné différents modèles de régression pour prédire la vitesse de corrosion, à savoir : la régression linéaire (simple et multiple), la régression non linéaire (polynomiale et polynomiale multivariée). Le prochain chapitre présentera la mise en œuvre réussie du modèle sélectionné, ainsi que les résultats obtenus à partir de données réelles .

CHAPITRE 4

IMPLÉMENTATION ET RÉSULTATS

Introduction

Dans ce chapitre, nous nous engageons dans l'implémentation des algorithmes et les comparons de manière équitable. Nous fournissons également des définitions et des explications succinctes nécessaires pour suivre le processus et la présentation du langage que nous avons utilisé.

4.1 Outils et environnement de travail

Dans cette section, nous allons aborder les outils qui ont été utilisés tout au long du processus de travail.

4.1.1 Langage de programmation

Pour la réalisation de notre modèle de prédiction nous avons utilisé le langage de programmation **python**.

Python"Créé par Guido van Rossum et sorti en 1991",est un langage de programmation interprété, orienté objet et de haut niveau avec une sémantique dynamique. Ses structures de données intégrées de haut niveau, combinées au typage dynamique et à la liaison dynamique, le rendent très attrayant pour le développement rapide d'applications, ainsi que pour une utilisation en tant que langage de script ou de collage pour connecter des composants existants entre eux [16].

La syntaxe simple et facile à apprendre de Python met l'accent sur la lisibilité et réduit donc le coût de maintenance du programme. Python prend en charge les modules et les packages, ce qui encourage la modularité du programme et la réutilisation du code. L'interpréteur Python et la vaste bibliothèque standard sont disponibles gratuitement sous forme source ou binaire pour toutes les principales plates-formes et peuvent être librement distribués [16].

Pourquoi utiliser Python ?

La data sciences consiste à extraire des informations utiles à partir de vastes ensembles de données, de statistiques et de registres. Ces données sont généralement non triées et difficiles à corréler avec précision.

Python répond à ce besoin en étant un langage de programmation polyvalent. Il permet de créer des sorties au format CSV pour faciliter la lecture des données dans un tableur. Il est également possible d'utiliser des formats de fichiers plus complexes qui peuvent être utilisés par des clusters d'apprentissage automatique pour des calculs plus avancés.

Python est le langage de programmation préféré des scientifiques des données. Ils ont besoin d'un langage facile à utiliser, avec une disponibilité décente des bibliothèques et une communauté importante. Les projets avec des communautés inactives ont généralement moins de chances de mettre à jour leurs plateformes.

4.1.2 Environnement de programmation

Anaconda

Anaconda est une distribution de logiciels open source pour la science des données et le calcul scientifique,et à l'apprentissage automatique.
Anaconda fourni avec Python et R, ainsi qu'une variété d'autres packages scientifiques, tels que NumPy, SciPy...etc [17].

FIGURE 4.1 – Logo anaconda

Jupyter Notebook

Jupyter Notebook est une application web qui permet aux utilisateurs de créer et de partager des documents interactifs appelés "notebooks". Ces notebooks combinent des cellules de code, de texte, de visualisations et d'autres éléments interactifs [18].

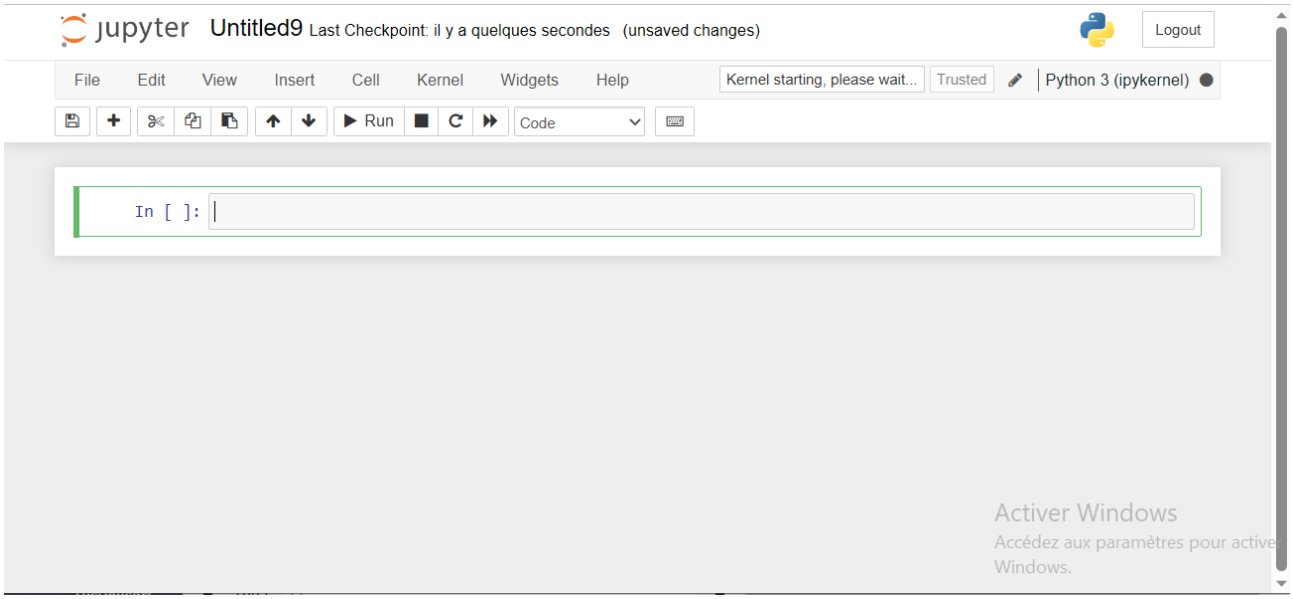


FIGURE 4.2 – Interface de Jupyter.

4.1.3 Bibliothèque utilisées

Numpy

NumPy est le package fondamental pour le calcul scientifique en Python. Il a été créé en 2005. C'est une bibliothèque permettant d'effectuer des calculs numériques avec Python. Il implémente des calculs sur des tableaux multidimensionnels et matrices [19].

SciPy

SciPy est une bibliothèque open source pour les mathématiques, les sciences et l'ingénierie.

C'est une une bibliothèque de calcul scientifique pour Python. Elle fournit des fonctionnalités pour l'optimisation, l'algèbre linéaire, l'intégration, l'interpolation,la résolution d'équations différentielles ordinaires, l'optimisation de fonction,...etc [20]. SciPy est construit sur NumPy.

Pandas

Pandas est une bibliothèque open source fournissant des structures de données et des outils d'analyse de données hautes performances et faciles à utiliser pour le langage de programmation Python [21].

Matplotlib

Matplotlib est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python. nous avons utilisé cette bibliothèque pour visualiser note images sous formes de graphiques [22].

Scikit-learn

C'est l'une des bibliothèques les plus utiles pour l'apprentissage automatique en Python. La bibliothèque Scikit-learn met en œuvre la régression, la classification, et les algorithmes de clustering ainsi que certains prétraitements d'opérations telles que le nettoyage de données [23].

4.2 Aperçu des données reçues

Fichier Accueil Insertion Mise en page Formules Données Révision Affichage

E14 | 0.677

	A	B	C	D	E	F	G
1	T(C°)	PH	CO2(bar)	Sali(g\L)	VITESSE DE CORROSION (mm /an)		
2	20	6	0	1	0.184		
3	40	6	0	1	1.211		
4	20	6	5	1	0.175		
5	40	6	5	1	2.256		
6	20	6	5	1	1.360		
7	20	3	0	1	1.176		
8	20	3	5	1	3.795		
9	40	3	5	1	7.897		
10	40	3	0	1	1.374		
11	20	9.38	5	1	0.997		
12	40	9.38	5	1	1.397		
13	40	9.38	0	1	0.991		
14	20	6	0	3	0.677		
15	40	6	0	3	2.650		
16	20	6	5	3	1.570		
17	40	6	5	3	3.082		
18	20	3	0	3	1.307		
19	20	3	5	3	1.617		
20	40	3	5	3	1.937		
21	40	3	0	3	1.415		
22	20	9.38	5	3	1.455		
23	20	9.38	0	3	0.414		

FIGURE 4.3 – Tableau des données réelles.

4.3 Lecture des fichier EXCEL sur python

4.3.1 Utilisation de Python pour la lecture

•Nous procéderons maintenant à l'importation des fichiers sur Python en utilisant les bibliothèques suivantes :

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn import linear_model
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression
from scipy.optimize import curve_fit
```

FIGURE 4.4 – Chargement des bibliothèques en python

•Utilisez la fonction $pd.read-excel()$ pour lire le fichier Excel . Voici la syntaxe de lecture d'un fichier Excel :

```
df = pd.read_excel('C:\\Users\\Admin\\Documents\\JUUUUST.xlsx')
```

FIGURE 4.5 – Interface de Jupyter.

•L'instruction print(df) permet d'afficher le contenu du fichier importé à l'aide de la variable df.

•Après avoir importé le fichier et créé le DataFrame df à l'aide de la bibliothèque pandas, vous pouvez extraire les variables X et Y , en utilisant les instructions suivantes :

```
X = df.iloc[:,:-1]
Y = df.iloc[:,-1]
```

FIGURE 4.6 – Séparation des features et labels.

•Une fois que vous avez les variables X et Y, vous pouvez entraîner votre modèle de régression et obtenir les prédictions.Supposons que vous utilisez une régression linéaire avec la bibliothèque scikit-learn.

```
model = LinearRegression()
model.fit(X, Y)
```

FIGURE 4.7 – Entrainement du modèle de régression linéaire.

•Nous allons maintenant afficher les coefficients (betas) en utilisant la formulation suivante :

```
coefficients = model.coef_
print(coefficients)

interception = model.intercept_
print(interception)
```

FIGURE 4.8 – Affichage des résultats.

•Maintenant, pour calculer le coefficient de détermination (R^2) et l'erreur quadratique moyenne (MSE) , vous pouvez utiliser des fonctions fournies par la bibliothèque $scikit - learn$. Voici comment vous pouvez le faire :

```
Y_pred = model.predict(X)
r2 = r2_score(Y, Y_pred)
mse = mean_squared_error(Y, model.predict(X))
print(r2)
print(mse)
```

FIGURE 4.9 – Résultats de la régression linéaire.

4.4 Modélisation des paramètres physico-chimiques sur la vitesse de corrosion

4.4.1 Évaluation du modèle de régression linéaire simple

Dans cette section, nous procédons l'estimation des paramètres de plusieures modeles de régression simple en considérons pour chaque modèle une variable explicative différente. La variable expliquer tant la vitesse de corrosion .Le tableau suivant fournit les résultats trouvés :

	$\beta 0$	$\beta 1$	R^2	MSE
PH	3.1678	-0.2462	0.1390	2.1655
CO2	1.1399	0.2310	0.1315	2.1843
T	0.0335	0.0597	0.1405	2.1617
S	2.0454	-0.1443	0.0082	2.4943

TABLE 4.1 – Analyse des résultats de la régression linéaire simple.

Interprétation des résultats :

En analysant les résultats, nous constatons que la variable PH et T (température) ont un coefficient de détermination R^2 de 14%. En revanche, la variable CO2 a un coefficient de détermination R^2 de 13%, et la salinité présente le coefficient de détermination R^2 le plus faible, soit $0,008\%$. Ces résultats suggèrent que les variables T et PH et CO2 ont une influence plus forte sur la vitesse de corrosion par rapport à la variable S (la salinité).

4.4.2 Évaluation du modèle de régression linéaire Multiple

Scénario 1 :

Le tableau ci-dessous présente les résultats de l'évaluation du modèle de régression linéaire multiple pour deux configurations différentes. La première configuration comprend les paramètres PH, CO2 et la température (T), tandis que la deuxième configuration ajoute la variable de salinité (S).

Interprétation des résultats :

D'après les résultats du tableau 4.2, nous constatons que l'ajout de la variable (S) à la configuration existante n'a eu qu'un impact minime sur la performance du modèle. En effet, les

	R^2	MSE
(T,PH,CO2)	0.4365	1.4171
(T,PH,CO2,S)	0.4387	1.4116

TABLE 4.2 – Comparaison des configurations de variables dans le modèle de régression linéaire multiple

valeurs du coefficient de détermination R^2 pour les deux configurations, avec et sans la variable "S", sont très proches (0.4365 et 0.4387 respectivement). Cela indique que l'ajout de la variable "S" n'a pas conduit à une amélioration significative de l'adéquation du modèle aux données.

Scénario 1 :

Le tableau ci-dessous illustre l'évaluation du modèle de régression linéaire multiple pour deux configurations différentes. La première configuration comprend uniquement les paramètres CO2, PH et la salinité (S), tandis que la deuxième configuration ajoute la variable de température (T).

	R^2	MSE
(CO2,PH,S)	0.3008	1.7584
(T,PH,CO2,S)	0.4387	1.4116

TABLE 4.3 – Comparaison des configurations de variables dans le modèle de régression linéaire multiple.

Interprétation des résultats :

Les résultats du tableau 4.3 montrent que la configuration (CO2, PH, S, T) donne de meilleurs résultats en termes de performance du modèle par rapport à la configuration (CO2, PH, S) seule. L'ajout de la variable de température a permis d'améliorer la qualité des prédictions du modèle (R2 meilleur et MSE plus faible). Cependant, il est important de noter que le MSE reste élevé dans les deux configurations, ce qui suggère qu'il peut y avoir d'autres variables ou facteurs à prendre en compte pour améliorer la modélisation de l'influence des paramètres physico-chimiques sur la vitesse de corrosion.

Scénario 3 :

Le tableau ci-dessous présente les résultats de l'évaluation du modèle de régression linéaire multiple pour deux configurations différentes. La première configuration comprend les paramètres température (T), salinité (S) et CO2, tandis que la deuxième configuration ajoute également le paramètre de (PH).

	R^2	MSE
(T,S,CO2)	0.2970	1.7681
(T,S,CO2, PH)	0.4387	1.4116

TABLE 4.4 – Comparaison des configurations de variables dans le modèle de régression linéaire multiple.

Interprétation des résultats :

Les résultats du tableau 4.5 suggèrent que la configuration (T,S,CO2, PH) donne de meilleurs résultats en termes de performance du modèle par rapport à la configuration (T, S, CO2) seule. L'ajout du paramètre PH a permis d'améliorer la qualité des prédictions du modèle (R2 meilleur et MSE plus faible).

Scénario 4 :

	R^2	MSE
(T,S,PH)	0.2665	1.8448
(T,S,CO2, PH)	0.4387	1.4116

TABLE 4.5 – Comparaison des configurations de variables dans le modèle de régression linéaire multiple.

Interprétation des résultats :

En conclusion, les résultats du tableau 4.5 suggèrent que la configuration (T,S,CO2, PH) donne de meilleurs résultats en termes de performance du modèle par rapport à la configuration (T, S, PH) seule. L'ajout du paramètre CO2 à permis d'améliorer la qualité des prédictions du modèle (R2 meilleur et MSE plus faible).

4.4.3 Conclusion

Dans l'objectif d'expliquer la variable "vitesse de corrosion" en fonction des paramètres physico-chimique comme variables explicatives. Les différentes combinaison étudier montrent que le modèle de régression linéaire multiple quatre variable explicatives qui donne de meilleurs résultats , il s'écrit sous forme :

Vitesse de corosion=$0.8673865421381856 + 0.05987049T - 0.25079335PH + 0.26694776CO_2 - 0.07512843S$

avec un coefficient de détermination : R^2 =0.438709787344266

et une erreur quadratique moyenne : MSE =1.4116486758574875

Validation du modèle de régression linéaire multiple (a 4 variable

Tableau d'analyse de la variance : Pour avoir la table d'ANOVA, il faut d'abord calculer SCT, SCR et SCE.

SCT=53.3483

SCR=24.4308

SCE=28.9174

Le tableau d'analyse de la variance est donné par :

Source de variation	Sommes des carrés	Degrés de liberté	Carrés moyennes	F
Expliqués	24.4308	4	6.1077	3.5906
Résiduelle	28.9174	17	1.7010	
Total	53.3483	21		

TABLE 4.6 – Table d'ANOVA.

Test de signification globale de Fisher :

Ce test permet de voir la relation entre Y et les variables x_1, x_2, x_3, x_4 dans leur ensemble, ce Qui équivalent à l'hypothèse selon laquelle tous les coefficients sont nuls :

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (Tous les paramètres nuls)

H_1 : il existe au moins un des paramètres différent de 0

La statistique utilisée est :

$$F = \frac{\frac{SCR}{p}}{\frac{SCE}{n-(p+1)}}$$

P : degrés de liberté du modèle.

Pour interpréter le test de Fisher, on compare la valeur de la statistique F calculée à une valeur

critique de la distribution de Fisher correspondant à un niveau de signification donné :

$$f_{p,n-(p+1)}^{1-\alpha}$$

avec $\alpha = 0.05$

Comme la statistique F qui de 3.5906 est supérieure à la valeur critique qui égale à 0.02684.

Par conséquent, on peut conclure qu'il y a des preuves statistiquement significatives pour rejeter l'hypothèse nulle. Cela suggère qu'il existe au moins un coefficient non nul, ce qui indique que le modèle est globalement explicatif.

La significativité des X_i :

X_i	T	PH	CO_2	S
p-value	0.0627	0.0456	0.0244	0.6997

TABLE 4.7 – La significativité de chaque variable explicative .

D'aprés le tableau 4.7, On a détecté deux variables explicatives qui sont signifcatives tels que PH et CO_2, car leurs p-value est inférieure au seuil de signification 5%

4.5 Régression polynomial

•Nous avons importé les bibliothèques requises :

```
import numpy as np
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
```

FIGURE 4.10 – Chargement des bibliothèques en python.

•Entraînez le modèle en utilisant les données transformées et la variable dépendante Y,et créez une instance de la classe "*PolynomialFeatures*" pour spécifier le degré du polynôme souhaité . Par exemple, si vous souhaitez un polynôme de degré 2, vous pouvez utiliser :

```
poly_features = PolynomialFeatures(degree=2)
X_poly = poly_features.fit_transform(X)
```

FIGURE 4.11 – Entraînement du modèle .

•Vous pouvez maintenant afficher les coefficients "*betas*" du modèle en utilisant :

```
print(regression.intercept_, regression.coef_)
```

FIGURE 4.12 – Affichage des résultats.

•Une fois que vous avez entraîné votre modèle et obtenu les prédictions vous pouvez calculer le coefficient de détermination (R^2) vous pouvez utiliser la méthode "*score*()" du modèle entraîné ,et l'erreur quadratique moyenne (MSE) vous pouvez utiliser la fonction "*mean*" de la bibliothèque scikit-learn :

```
Y_pred = regression.predict(X_poly)
r2 = r2_score(Y, Y_pred)
mse = np.mean((y_pred - Y) ** 2)
print(r2)
print(mse)
```

FIGURE 4.13 – Résultats de la régression.

4.5.1 Choix de degrée du polynôme

Surajustement et sous-ajustement

Le sur-ajustement (overfitting) et le sous-ajustement (underfitting) sont des problèmes courants dans l'apprentissage supervisé. Ils se produisent lorsque le modèle ne parvient pas à généraliser correctement à partir des données d'entraînement vers de nouvelles données.[11] [10]

Le surajustement (overfitting) et le sous-ajustement (underfitting) apparaissent lorsque nous discutons du degré polynomial. Le degré représente le niveau de flexibilité du modèle, où un degré plus élevé permet au modèle d'avoir plus de liberté pour s'adapter à un plus grand nombre de points de données. Cependant, un modèle insuffisamment ajusté sera moins flexible et ne pourra pas tenir compte de l'ensemble des données. La meilleure façon de comprendre ce problème est d'examiner des modèles illustrant les deux situations[11] [10].

Le surajustement se produit lorsqu'un modèle devient trop complexe et s'adapte de manière excessive aux données d'entraînement, capturant ainsi le bruit et les variations aléatoires présentes dans les données. Cela se traduit par une très faible erreur d'entraînement, mais une mauvaise performance sur de nouvelles données (ensemble de test), car le modèle n'a pas généralisé les motifs sous-jacents[11] [10].

D'autre part, le sous-ajustement se produit lorsque le modèle est trop simple et ne parvient pas à capturer les relations complexes présentes dans les données. Cela se traduit par une erreur d'entraînement plus élevée et une performance médiocre à la fois sur l'ensemble d'entraînement et l'ensemble de test[11] [10].

4.5.2 Analyse de l'impact du degré du polynôme sur le surajustement et le sous-ajustement : Étude des courbes de la régression polynomiale

Analyse du paramètre "PH" :

Ci-dessous quatre graphiques représentant les courbes de régression polynomiale de la vitesse de corrosion (variable Y) sur le PH (variable X) pour différents degrés : degré 1, quadratique, degré 4,...etc. Le R2 et le MSE nous permettent de choisir le modèle le plus adéquat pour représenter la liaison entre la vitesse de corosion et le PH.

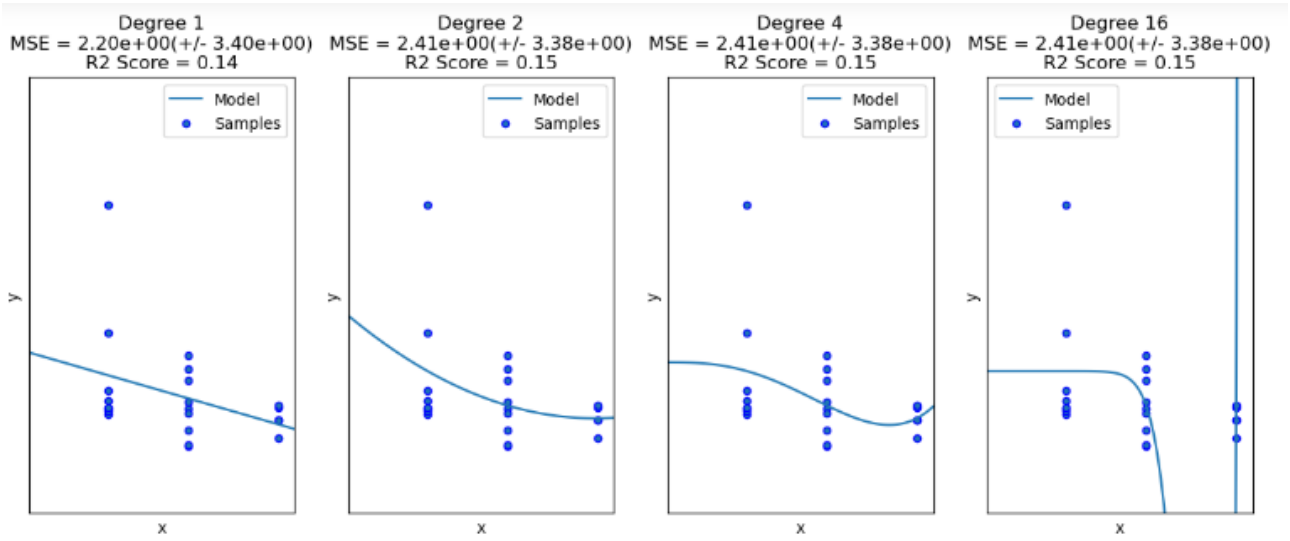


FIGURE 4.14 – Chargement des bibliothèques en python.

•Pour le premier graphe, le coefficient de détermination R^2 est de 14% .

•Pour les autres graphes (degré 2, 4 et 16), le coefficient de détermination R^2 est de 15%.

Cela indique clairement que ces modèles, que ce soit le modèle de degré 2, 4 ou 16, présentent tous une performance supérieure au modèle de degré 1 en termes d'ajustement aux données. Ainsi, nous jugeons utile et plus judicieux de choisir le degré 2 . C'est la propriété de parcimonie.

Analyse du paramètre "Températeur" :

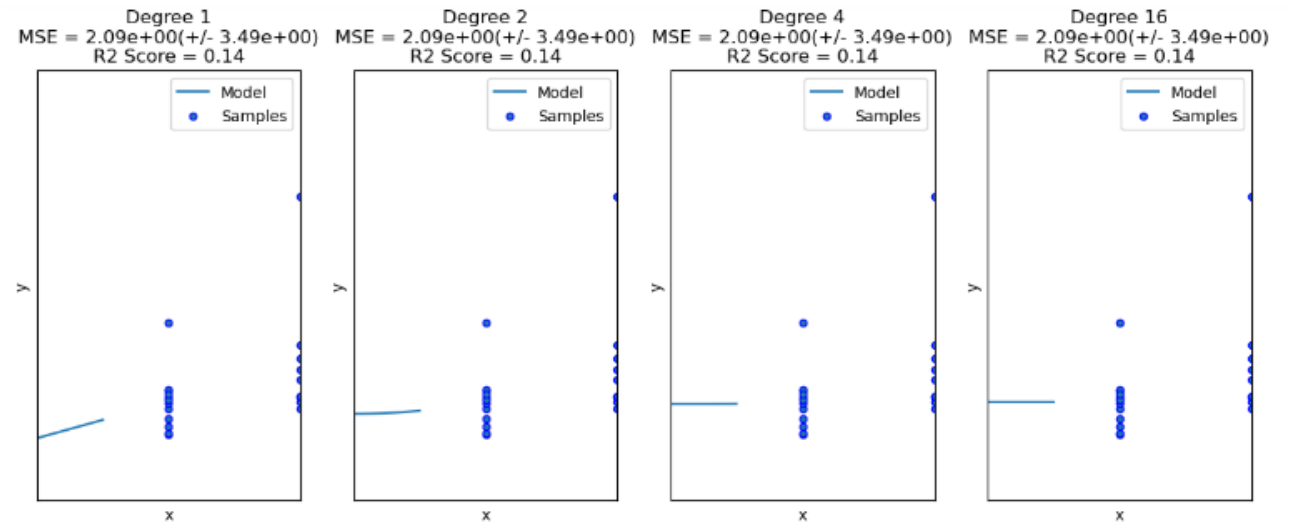


FIGURE 4.15 – Chargement des bibliothèques en python.

Le coefficient de détermination R^2 est de 14% pour tous les degrés du polynôme (1, 2, 4 et 16). Ces résultats indiquent que tous les modèles, qu'ils soient de degré 1, 2, 4 ou 16, présentent une performance similaire en termes d'ajustement aux données.Donc, en respectant la propriété de parcimonie nous choisissons le polynôme de degré 1 .

Analyse du paramètre "Salinité" :

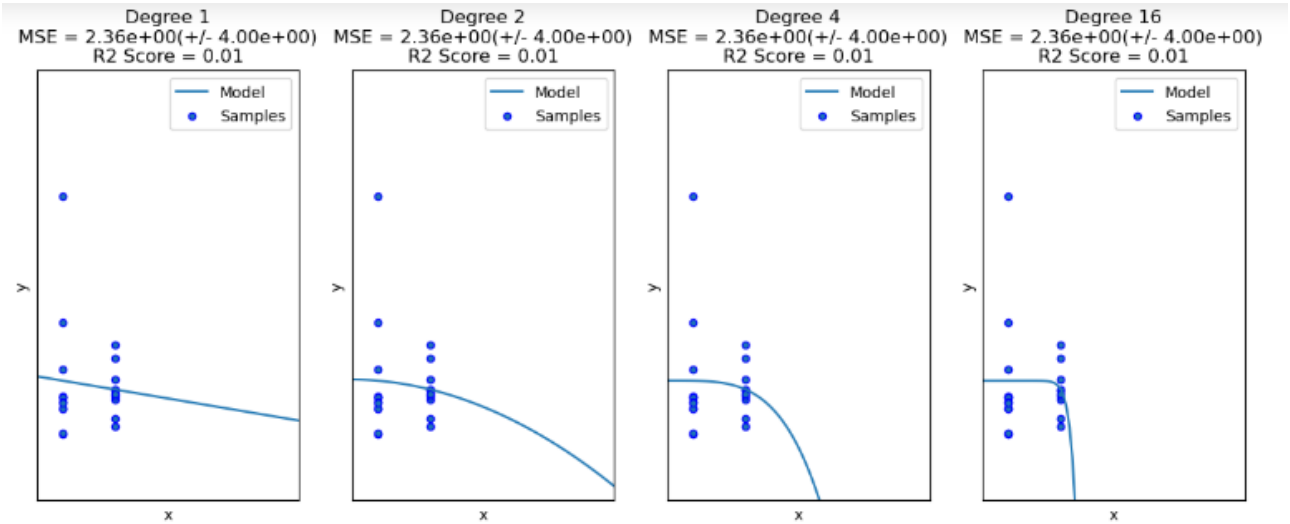


FIGURE 4.16 – Chargement des bibliothèques en python.

Pour la salinité, tous les modèles de degré du polynôme (1, 2, 4 et 16) présentent un coefficient de détermination R^2 de 1%, ce qui indique qu'aucun de ces modèles ne parvient à ajuster les

données de manière significative. Cela signifie que la variable "salinité" n'a pas vraiment un impact sur la vitesse de corosion en comparaison avec les autres variables.

Analyse du paramètre "CO2" :

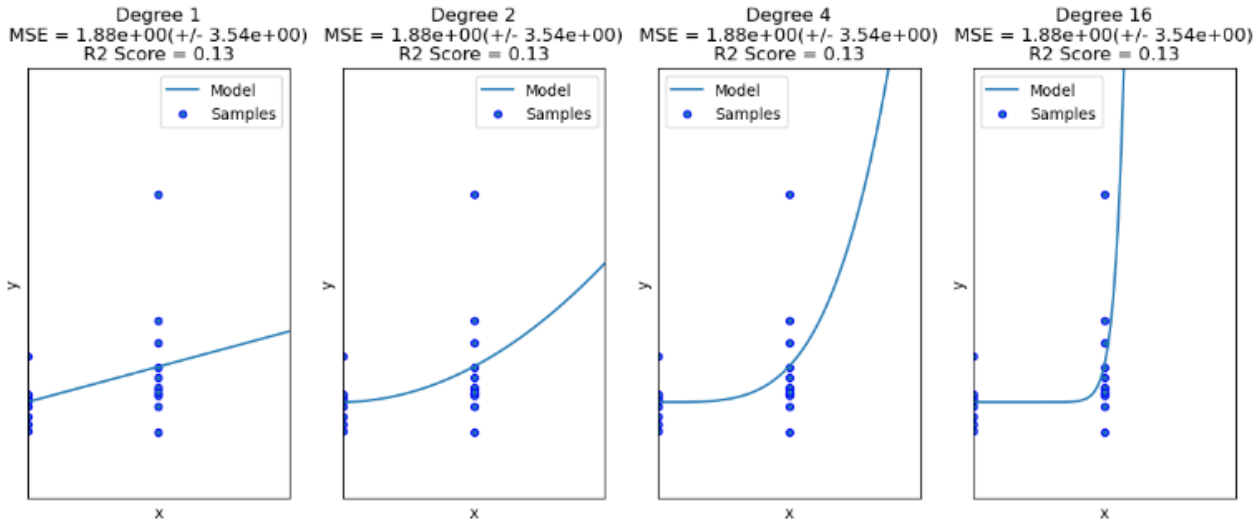


FIGURE 4.17 – Chargement des bibliothèques en python.

Pour le CO2, tous les modèles de degré du polynôme (1, 2, 4 et 16) présentent un coefficient de détermination R^2 de 13% .Donc, en respectant la propriété de parcimonie nous choisissons le polynôme de degré 1 .

4.5.3 Évaluation et validation du modèle de régression polynomial multivariable

Dans cette section, nous avons opté pour la modélisation de notre problématique par un modèle de régression polynomial multivariable. Nous avons utilisé Python pour simuler l'EQM (Erreur Quadratique Moyenne), l'EAM (Erreur Absolue Moyenne) et le coefficient de détermination R^2 pour différents degrés, à savoir le degré 2(forme quadratique), le degré 3(forme cubique) et le degré 4(forme Quadratique).

Les résultats sont représentés dans le tableau suivant :

Interprétation des résultats

En conclusion, la modélisation polynomiale multivariable avec des degrés plus élevés conduit à de meilleurs résultats en termes d'ajustement aux données. Parmi les formes étudiées, la forme

La forme	EQM	EAM	R^2 (%)
Quadratique	0.765	0.676	69.7%
Cubique	0.115	0.282	95.4%
Quadratique	0.032	0.053	98.7%

TABLE 4.8 – Comparaison des configurations de variables dans le modèle de régression polynomial multivariable.

quadratique qui présente les meilleures performances . Par conséquent, nous recommandons d'utiliser la forme quadratique pour le modèle de prédiction, car elle offre la meilleure précision et la plus faible variabilité résiduelle.

4.6 Comparaison des modèles étudiés

Le tableau ci-dessous présente la comparaison des performances des modèles étudiés :

	EQM	EAM	R^2 (%)
Régression linéaire multiple	1.411	0.817	43.8 (%)
Régression polynomiale multivariée	0.032	0.053	98.7 (%)

TABLE 4.9 – Comparaison des modèles étudiés.

En conclusion, le modèle polynomial présente une performance nettement supérieure au modèle linéaire.ce qui indique une excellente adéquation du modèle aux données. Ainsi, **le modèle polynomial** est recommandé pour une meilleure précision et ajustement dans la prédiction des données étudiées.

4.6.1 Interface de l'application

L'objectif de cette interface est de prédire la vitesse de corrosion (mm/ans) en fonction des informations saisies dans le formulaire ci-dessous. Le formulaire comprend les champs suivants : Température, pH, CO2, Salinité et vitesse de corrosion actuelle.
L'interface offre les fonctionnalités suivantes :

- •Des champs de saisie pour remplir les informations requises.
- •Un bouton de calcul pour effectuer la prédiction.
- •Une zone d'affichage pour afficher le résultat de la prédiction.
- •Un bouton pour quitter l'application.

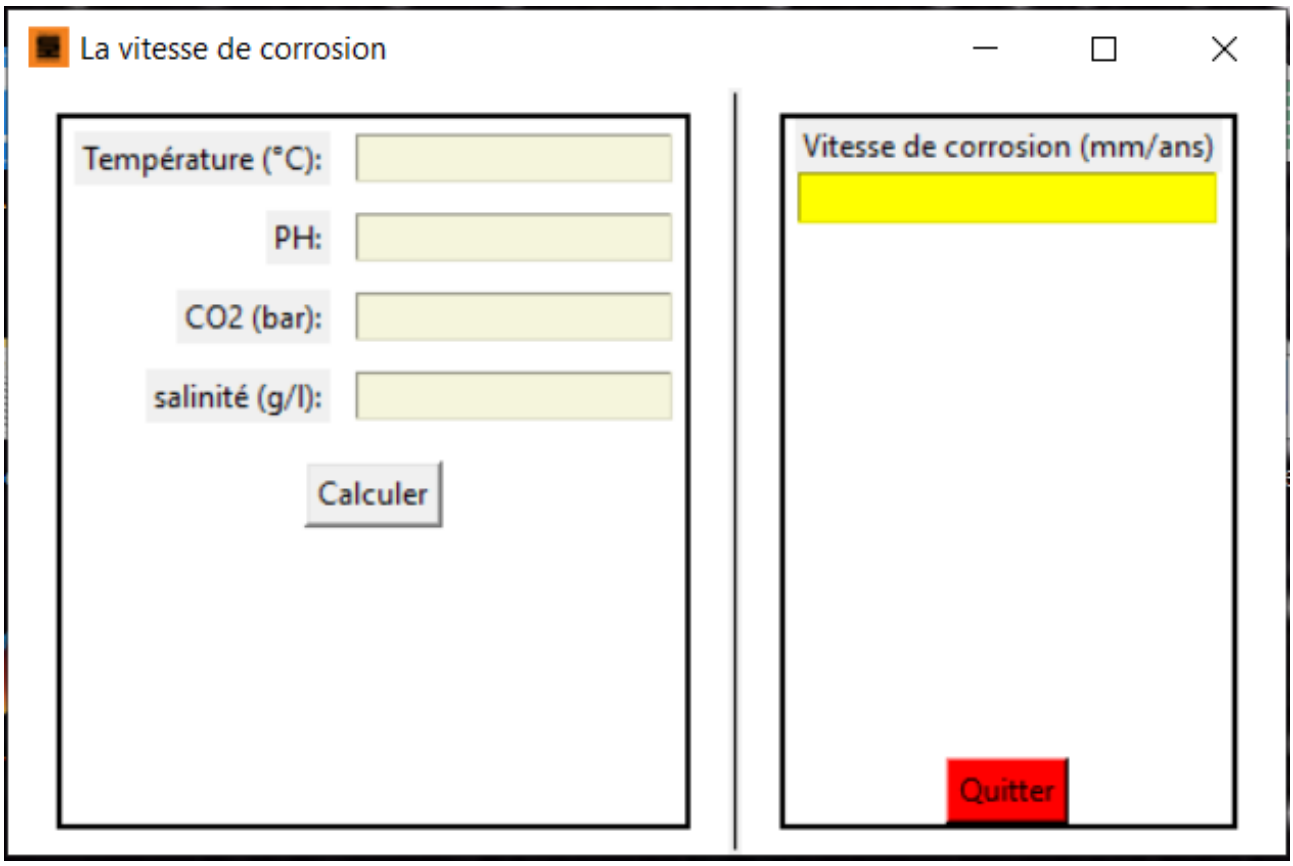


FIGURE 4.18 – Interface de l'application.

Exemple d'interface déjà remplie

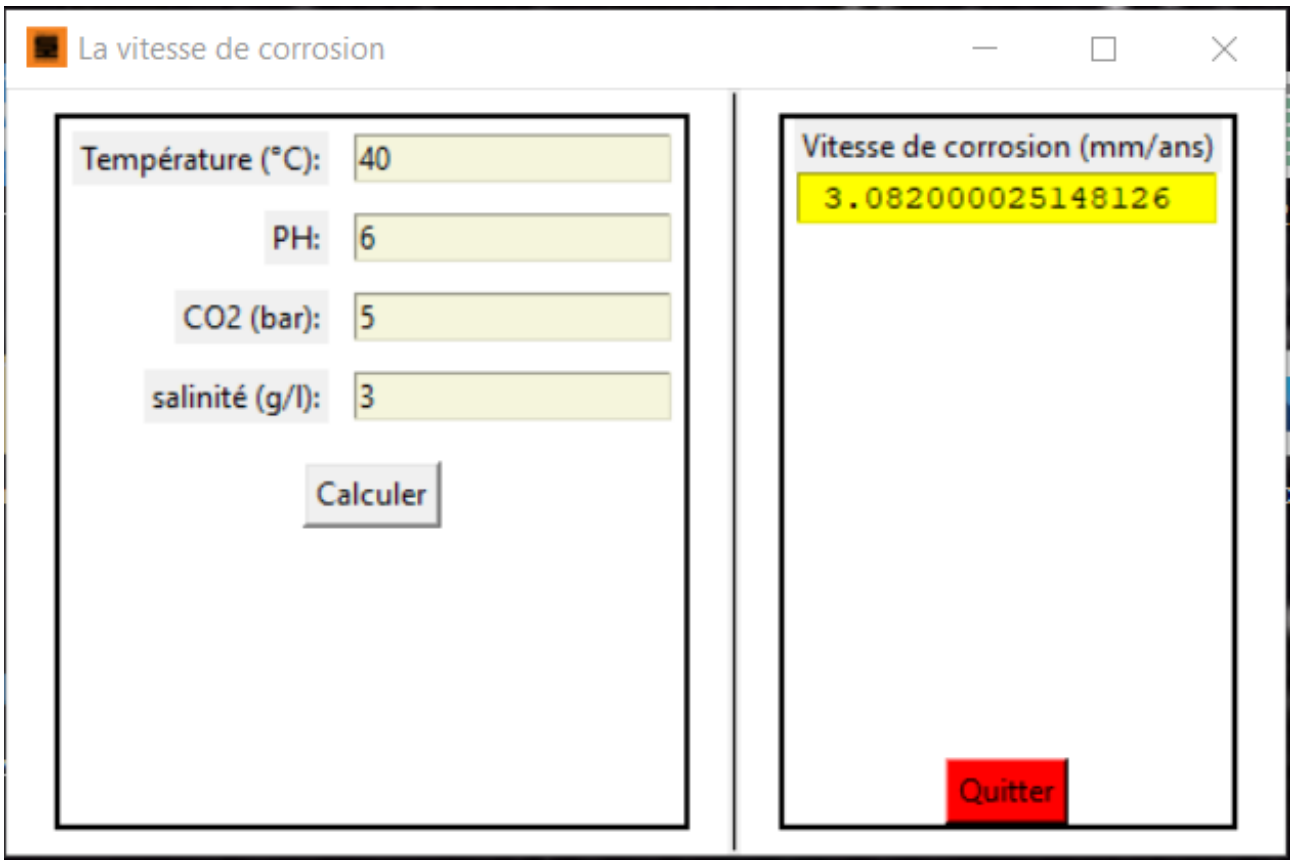


FIGURE 4.19 – Interface remplie.

Conclusion

Ce présent chapitre a été consacré à l'application, sous python, de la théorie de la régression linéaire (simple et multiple) et de la régression non linéaire (polynômiale et polynomiale multivariée). Le choix du meilleur modèle a été basé sur le coefficient de détermination R2 (maximum) et sur l'erreur moyenne quadratique MSE (minimum). Et pour un accès facile et rapide à cette application une interface a été conçu et présentée à la fin du chapitre.

CONCLUSION GÉNÉRALE

En conclusion, cette étude a permis d'approfondir la compréhension de la corrosion et son impact sur les équipements industriels. Nous avons présenté une analyse détaillée de l'entreprise Sonatrach, en mettant en évidence son historique, ses missions, ses objectifs et son organisation. Ensuite, nous avons réalisé une étude bibliographique sur la corrosion, en examinant les causes, les types et les aspects morphologiques de ce phénomène.

Une partie importante de cette recherche a été consacrée à l'exploration des techniques de prédiction de la corrosion, en se concentrant spécifiquement sur l'utilisation de la régression comme outil d'analyse. Nous avons examiné les concepts fondamentaux de la régression linéaire, en mettant en évidence ses principes et ses différentes méthodes, notamment la régression linéaire simple et la régression linéaire multiple. De plus, nous avons abordé la régression non linéaire et la régression polynomiale comme alternatives pour modéliser la vitesse de corrosion.

L'implémentation et les résultats obtenus ont démontré l'efficacité de l'approche de régression pour prédire la vitesse de corrosion en fonction des paramètres physico-chimiques. Nous avons utilisé des outils de programmation tels que Python et des bibliothèques spécialisées pour analyser les données et évaluer les performances des modèles de régression.

L'interprétation des résultats a permis de mettre en évidence l'influence des différents paramètres sur la vitesse de corrosion, ainsi que l'importance de la sélection du modèle adéquat pour obtenir des prédictions précises. Nous avons également réalisé une comparaison entre les différents modèles étudiés, mettant en évidence leurs avantages et leurs inconvénients respectifs.

En conclusion, cette étude offre une contribution significative à la compréhension de la corrosion et à l'application des techniques de régression pour prédire sa vitesse. Les résultats obtenus peuvent servir de base pour le développement de stratégies de prévention et de maintenance prédictive dans le domaine industriel. Cependant, il convient de noter que la corrosion est un

phénomène complexe et multifactoriel, et que des recherches supplémentaires sont nécessaires pour affiner les modèles et prendre en compte d'autres variables pertinentes.

BIBLIOGRAPHIE

[1] Nacer-eddine Arbaoui and N SETTOU T LANEZ. *Effet du gradient de température sur la vitesse de corrosion des concentriques de traitement des puits producteurs d eau Albien.* PhD thesis, 2004.

[2] Hacene Bellahmer. *Implémentation et évaluation d'un modèle d'apprentissage automatique pour l'estimation de la valeur marchande de propriétés immobilières.* PhD thesis, Université Mouloud Mammeri, 2020.

[3] Hakim Bensabra. Cours de corrosion et protection des métaux. *Université de JIJEL*, pages 3–4, 2016.

[4] Oussama Akram Bensiah and Mohamed Berkane. La proposition d'une nouvelle approche basée deep learning pour la prédiction du cancer du sain. 2020.

[5] Christophe Chesneau. Modèles de régression. 2017.

[6] B Grosgogeat and P Colon. La corrosion. *Société Francophone de Biomatériaux Dentaires*, 2009.

[7] Arnaud Guyader. Régression linéaire. *Université Rennes*, 2 :60–61, 2011.

[8] Lilia Ihaddadene. *Etude de la corrosion et la protection des pipelines de réseau de collecte d'huile de la region tft–secteur nord.* PhD thesis, université Akli Mouhand Oulhadje-Bouira, 2018.

[9] KIR Iman. Étude de l'influence du traitement thermique sur la dissolution anodique d'un acier au carbone en milieu aqueux. 2014.

[10] H Jabbar and Rafiqul Zaman Khan. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, 70 :163–172, 2015.

[11] Will Koehrsen. Overfitting vs. underfitting : A complete example. *Towards Data Science*, pages 1–12, 2018.

[12] Jonas Kibala Kuma. Modèles de régression non linéaires : Éléments de théorie et pratiques sur logiciel. 2019.

[13] John McCarthy. What is artificial intelligence. 2007.

[14] Eva Ostertagová. Modelling using polynomial regression. *Procedia Engineering*, 48 :500–506, 2012.

[15] Priyanka Sinha. Multivariate polynomial regression in data mining : methodology, problems and solutions. *Int. J. Sci. Eng. Res*, 4(12) :962–965, 2013.

Webographie :

[16] https ://www.python.org/doc/essays/blurb/

[17] https ://www.anaconda.com/products/distribution

[18] https ://jupyter.org/

[19] https ://numpy.org/

[20] https ://scipy.org/

[21] https ://pandas.pydata.org/

[22] https ://matplotlib.org/

[23] https ://scikit-learn.org/