RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Nationale Polytechnique

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

Département : Génie Industriel
Entreprise : SLB North Africa

Mémoire de projet de fin d'études

En vue de l'obtention du diplôme d'Ingénieur d'État en Génie Industriel
Option : Data Science and Intelligence Artificielle

Approches d'apprentissage automatique pour la détection et la prédiction du chargement de liquide dans les puits de gaz

cas d'étude de la division Digital & Integration de SLB NAF

Réalisé par

Hafdi Ramy

Encadré par

Mme. Debbi Latifa (ENP)
M. Sourabh Shukla (SLB)

Présenté et soutenu publiquement le 4 juillet 2023

Devant le jury composé de

M. Zouaghi Iskander : ENP - Président
M. Tachi Salah Eddine : ENP - Examinatuer
M. Abbaci Ayoub : ENP - Examinateur

ENP 2023

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Nationale Polytechnique

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

Département : Génie Industriel
Entreprise : SLB North Africa

Mémoire de projet de fin d'études

En vue de l'obtention du diplôme d'Ingénieur d'État en Génie Industriel
Option : Data Science and Intelligence Artificielle

Approches d'apprentissage automatique pour la détection et la prédiction du chargement de liquide dans les puits de gaz

cas d'étude de la division Digital & Integration de SLB NAF

Réalisé par

Hafdi Ramy

Encadré par

Mme. Debbi Latifa (ENP)
M. Sourabh Shukla (SLB)

Présenté et soutenu publiquement le 4 juillet 2023

Devant le jury composé de

M. Zouaghi Iskander : ENP - Président
M. Tachi Salah Eddine : ENP - Examinatuer
M. Abbaci Ayoub : ENP - Examinateur

ENP 2023

Dédicace

Je souhaite dédier ce travail,

À la personne la plus chère à mon cœur, ma mère, cette femme exceptionnelle qui n'a cessé de faire des sacrifices pour me permettre de concrétiser mes ambitions et atteindre mes objectifs,

À mon père, celui qui a toujours été à mes côtés et qui a été le pilier solide de mon éducation, inculquant les valeurs qui ont guidé chaque pas de mon parcours,

À mes chers frères Issam, Mohamed Chakib et Akram, leur souhaitant beaucoup de succès dans leurs parcours et futures études,

À mes tantes et oncles, spécialement Nadia, Soumia et Badis,

À mes précieux amis qui ont toujours été présents, en particulier mes amis du groupe "Mandat s6",

Aux membres de l'IEC, avec lesquels j'ai passé des moments merveilleux et inoubliables,

Et finalement, à tous ceux qui ont vu en moi un potentiel, qui ont cru en mes compétences et en ma réussite.

Ramy

Remerciements

Qu’il me soit permis de remercier et d’exprimer ma profonde gratitude en premier lieu à Dieu le tout-puissant de m’accorder sa bénédiction, de m’avoir donné de la force, de la volonté et de la patience pour mener à bien ce travail.

Ensuite, je tiens à exprimer ma profonde gratitude à mon encadrant en entreprise chez SLB, M. Sourabh Shukla. Sa guidance, son expertise et son soutien ont été inestimables tout au long de ce projet. Son engagement constant a été une source d'inspiration et a grandement contribué à la réussite de ce travail.

Je voudrais également remercier chaleureusement Mme Latifa Debbi, mon encadrante de l'école. Malgré les contraintes, elle a accepté de m'accompagner dans ce parcours. Son aide précieuse, ses conseils éclairés et son soutien indéfectible ont été des atouts indispensables à la réalisation de ce projet.

Un merci particulier à l'ensemble de la division D&I de SLB, qui m'a accueilli et m'a permis de réaliser ce projet dans des conditions optimales. Je tiens à souligner l'apport de M. Ahmed Mustapha, Mme Mounia Chekkai, M. Oussama Rouabeh et mes collègues Zineb Lazib et Yasmine Bouchafa, dont la contribution a été d'une grande aide pour ce projet.

Je souhaite exprimer ma gratitude anticipée au jury qui évaluera ce travail. Je tiens à les remercier pour le temps et l'expertise qu'ils consacreront à la lecture et à l'évaluation de ce projet. Leur contribution à l'amélioration et à l'évaluation de la qualité de mon travail est très appréciée.

Je suis également reconnaissant envers le corps professionnel de l'École Nationale Polytechnique, département de génie industriel, pour la formation de qualité et l'encadrement professionnel qu'ils m'ont fournis.

ملخص:

يهدف هذا المشروع إلى مواجهة التحديات المتعلقة بالكشف والتنبؤ بتحميل السوائل في آبار الغاز التي تعاني من تدفق متعدد العوامل. تعتبر مشكلة تحميل السوائل في هذه الآبار أمرًا هامًا يقيّد إنتاج الغاز ويتطلب اتخاذ إجراءات عملية فعالة. يهدف هذا البحث إلى استكشاف المبادئ الأساسية لإنتاج آبار الغاز والتدفق متعدد العوامل، وذلك باستخدام تقنيات التعلم الآلي. تهدف الدراسة إلى تطوير حلول فعّالة للكشف والتنبؤ بتحميل السوائل في آبار الغاز بدقة وفي الوقت المناسب. وسيتم دراسة أداء نماذج التعلم الآلي المختلفة ومقارنتها مع الارتباطات التجريبية. يهدف هذا المشروع إلى تحسين استراتيجيات مكافحة تحميل السوائل وتعزيز الكفاءة التشغيلية لإنتاج الغاز.

الكلمات الرئيسية : تحميل السوائل، آبار الغاز، تدفق متعدد العوامل، التعلم الآلي، الكشف، التوقع، تحسين إنتاج الغاز.

Abstract :

This project aims to address the challenges related to the detection and prediction of liquid loading in multiphase gas wells. Liquid loading in these wells is a significant issue that limits gas production and requires effective practical actions. This research explores the fundamental principles of gas well production and multiphase flow using machine learning techniques. The objective of this study is to develop effective solutions for accurate and timely detection and prediction of liquid loading in gas wells. The performance of different machine learning models will be investigated and compared with empirical correlations. This project aims to enhance liquid loading mitigation strategies and improve the operational efficiency of gas production.

Keywords : liquid loading, gas wells, multiphase flow, machine learning, detection, prediction, production optimization.

Résumé :

Ce projet vise à relever le défi de la détection et de la prédiction du chargement de liquide dans les puits de gaz en utilisant des techniques d'apprentissage automatique. La détection et la prédiction du chargement de liquide sont essentielles pour maintenir une production de gaz optimale et éviter toute perte de production. En explorant les principes fondamentaux de la production de puits de gaz et de l'écoulement multiphasique, ainsi qu'en exploitant différentes approches d'apprentissage automatique, cette étude cherche à développer une solution efficace pour une identification précise et opportune du chargement de liquide dans les puits de gaz. La recherche examinera les performances de différents modèles d'apprentissage automatique, les comparera aux corrélations empiriques et explorera le potentiel des modèles hybrides. Grâce à une analyse basée sur les données et à des expérimentations, ce projet vise à contribuer à l'avancement des stratégies d'atténuation du chargement de liquide et à améliorer l'efficacité opérationnelle de la production de gaz.

Mots-clés : chargement de liquide, puits de gaz, écoulement multiphasique, apprentissage automatique, détection, prédiction, optimisation de la production.

Table des matières

Liste des figures

Liste des tableaux

Liste des abreviations

A : section transversale du tubage.

Q : débit critique de gaz.

M_{air} : masse molaire du gaz.

P: pression à la tête du puits en bar.

R : constante de gaz.

T : température à la tête du puits.

Z : compressibilité du gaz.

γ_g : gravité spécifique du gaz.

v_{crit} : vitesse critique.

ρ_g : densité du gaz.

ρ_l : densité du liquide.

σ : tension de surface.

ANN : Artificial neural networks / Réseau de neurones artificiels.

AUC : Area Under Curve / Surface sous la courbe.

CART : Classification and Regression Trees.

CV : Cross validation / Validation croisée.

DBSCAN : Density-Based Spatial Clustering of Applications with Noise.

DL : Deep Learning / Apprentissage profond.

EM : Algorithme de maximisation de l'espérance

E-step : Étape d'espérance.

GDBT : Gradient Boosting Trees.

GM : Gaussian mixture / Mélange gaussien.

IA : Intelligence Artificielle.

KNN : K nearest neighbor / K plus proches voisins.

LSTM : Long short-term memory.

MI : Mutual Information / Information mutuelle.

ML : Machine learning / Apprentissage automatique

M-step : Étape de maximisation.

NB : Naive Bayes.

RF : Random forest/ Forêt aléatoire.

RMSE : Racine de l'erreur quadratique moyenne.

ROC : Receiver Operating Characteristics

SS : Spectral Clustering / Clustering spectral.

SVD : Singular value decomposition / Décomposition en valeurs singulières.

SVM : Support Vector Machines / Machine à vecteurs de support.

XGBoost : eXtreme Gradient Boosting.

Introduction générale

L'apprentissage automatique (ou machine learning) est un domaine de l'informatique qui vise à développer des modèles et des algorithmes capables d'apprendre à partir des données et de prendre des décisions ou d'effectuer des prédictions sans être explicitement programmés. Ce domaine connaît une croissance exponentielle et a des applications dans de nombreux secteurs, y compris l'industrie du pétrole et du gaz.

Ce mémoire se concentre sur l'application de techniques d'apprentissage automatique dans le domaine de la production de puits de gaz au sein de l'entreprise de services pétrolier SLB Ltd. La production de gaz est un processus complexe qui implique l'extraction du gaz naturel du sous-sol et son acheminement vers les consommateurs. La compréhension de ce processus est essentielle pour optimiser la production, réduire les coûts et minimiser les risques environnementaux.

Dans la première partie de ce mémoire, nous présenterons les concepts et les approches clés en apprentissage automatique, notamment l'apprentissage supervisé, non-supervisé, semi-supervisé et par renforcement. Nous explorerons également l'apprentissage profond, en mettant l'accent sur les perceptrons et les réseaux multicouches. De plus, nous aborderons l'apprentissage génératif et son application à la synthèse de données.

Ensuite, nous nous pencherons sur les principes fondamentaux de la production de puits de gaz. Nous examinerons les différents aspects de la production, tels que le profil de production, le chargement de liquide et les sources de liquides dans les puits de gaz. Nous effectuerons également une étude bibliographique pour comprendre les modèles existants utilisés pour prédire la production de puits de gaz.

Dans la troisième partie, nous présenterons l'entreprise SLB Ltd. et son secteur d'activité dans l'industrie des hydrocarbures. Nous analyserons le marché mondial des champs pétrolifères et le marché des services pétroliers en Algérie. Nous nous concentrerons également sur SLB Ltd., ses activités, sa structure hiérarchique et ses divisions opérationnelles, notamment la division Digital & Integration.

Dans la quatrième partie, nous décrirons la conception de la solution proposée pour le projet d'apprentissage automatique de bout en bout. Nous discuterons de la collecte et de la compréhension des données, ainsi que du prétraitement des données nécessaires à la mise en œuvre du projet.

Enfin, nous conclurons ce mémoire en résumant les principaux points abordés et en soulignant l'importance de l'application de techniques d'apprentissage automatique dans l'industrie du pétrole et du gaz.

L'objectif de ce mémoire est de développer un projet d'apprentissage automatique pour la prédiction et l'optimisation de la production de puits de gaz chez SLB. Pour cela, nous utiliserons des techniques d'apprentissage supervisé, non-supervisé et profond, ainsi que des modèles génératifs pour synthétiser des données de production.

Ce mémoire vise à fournir une base solide pour la mise en œuvre d'un projet d'apprentissage automatique dans le domaine de la production de puits de gaz qui servira à améliorer et optimiser ce processus chez SLB. Il contribue également à l'avancement des connaissances dans ce domaine et offre des perspectives intéressantes pour l'optimisation des opérations de production de gaz.

Chapitre 1

Techniques et Approches en Apprentissage Automatique

Introduction

L'apprentissage automatique, une sous-discipline influente de l'intelligence artificielle, a catalysé des transformations significatives dans de nombreux secteurs de notre société, y compris l'industrie pétrolière et gazière. Doté de la capacité d'apprendre de manière autonome à partir de données, de générer des prédictions et de prendre des décisions sans avoir été explicitement programmé pour accomplir ces tâches, l'apprentissage automatique s'est avéré être un instrument précieux pour la résolution de problèmes complexes et l'extraction d'informations pertinentes.

Dans ce chapitre, nous évoquerons les diverses formes d'apprentissage automatique - supervisé, non supervisé, semi-supervisé et par renforcement. Nous traiterons également des notions d'apprentissage profond, des réseaux neuronaux, et nous nous intéresserons au rôle de la synthétisation des données dans l'élaboration des modèles d'apprentissage automatique.

En somme, à travers ce chapitre nous allons procurer une compréhension robuste des diverses techniques et approches en apprentissage automatique. Que vous soyez un professionnel expérimenté à la recherche de connaissances plus approfondies ou un débutant avide de savoir, ce chapitre vous fournira les moyens de naviguer dans le paysage complexe de l'apprentissage automatique avec une confiance accrue.

1.1 Apprentissage automatique

L'apprentissage automatique, souvent appelé machine learning, est un sous-domaine de l'Intelligence Artificielle qui utilise des techniques statistiques pour créer des modèles informatiques capables d'apprendre à partir de données. Il vise à modéliser et à comprendre des structures ou des phénomènes complexes à l'aide de ces données, pouvant représenter un éventail d'éléments allant d'un concept à un attribut ou à un résultat particulier.

L'apprentissage automatique cherche à simuler le processus d'apprentissage humain en utilisant des algorithmes d'optimisation mathématique qui s'améliorent progressivement à mesure qu'ils sont exposés à davantage de données. Ces algorithmes apprennent, comme le feraient les humains, à faire des prédictions ou à prendre des décisions sans être explicitement programmés pour effectuer la tâche.

L'objectif principal des algorithmes de Machine Learning est d'estimer une fonction f qui minimise la différence entre les valeurs prédites et les valeurs réellement observées dans les données. Cette fonction peut prendre de nombreuses formes, en fonction du type d'algorithme de Machine Learning utilisé et de la nature des données. (Géron Aurélien | Hands on Machine Learning, 2022)

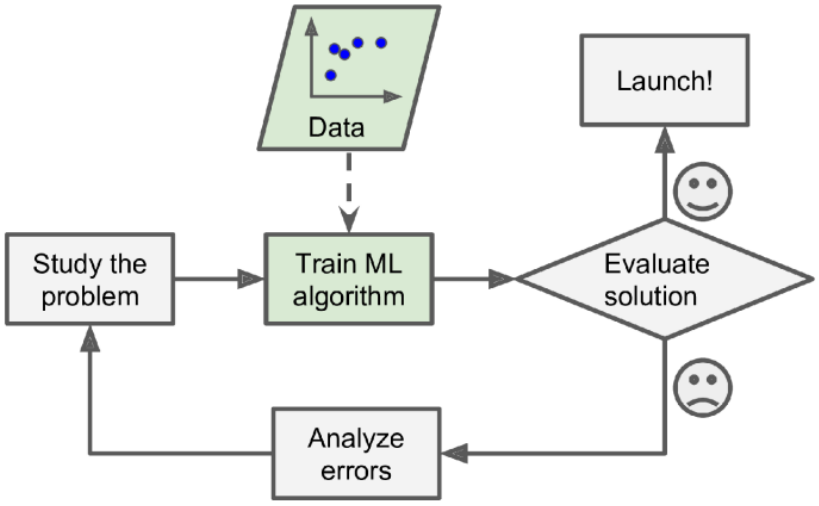


Figure 1.1 : L'approche du Machine Learning (Géron Aurélien, 2022).

1.1.1 Projet d'apprentissage automatique de bout en bout

Dans tout projet concret d'apprentissage automatique, il existe des étapes essentielles qui doivent être suivies pour garantir le succès et la pertinence des résultats. Ces étapes, lorsqu'elles sont bien exécutées, permettent de mener à bien un projet d'apprentissage automatique de bout en bout. Que ce soit pour résoudre des problèmes complexes, optimiser des processus métiers ou prendre des décisions basées sur les données, les étapes suivantes constituent le fondement de tout projet d'apprentissage automatique (Fernando Martínez-Plumed, 2021) (Figure 1.2).

Compréhension des métiers

Au début du projet, il est crucial de bien comprendre les objectifs commerciaux et les enjeux métiers sous-jacents. Cette étape permet de définir clairement ce que l'on cherche à atteindre grâce à l'apprentissage automatique. Il est également important de formuler une problématique précise liée à l'exploration des données afin de répondre aux besoins spécifiques de l'entreprise. En parallèle, un plan préliminaire est établi pour orienter les étapes suivantes du projet.

Compréhension des données

Cette phase est dédiée à la collecte des données provenant de différentes sources pertinentes. Il est primordial d'explorer ces données de manière approfondie afin de comprendre leur structure, leur format, ainsi que les informations qu'elles contiennent. Cette étape inclut également la description des données, en identifiant les caractéristiques clés et en évaluant leur qualité. La vérification de la qualité des données permet de s'assurer de leur fiabilité et de leur adéquation pour la construction d'un modèle de machine learning.

Préparation des données

La préparation des données est une étape cruciale dans laquelle les données collectées sont préparées pour être utilisées dans le modèle final. Cela comprend le nettoyage des données, où les valeurs manquantes ou aberrantes sont traitées de manière adéquate. De plus, la création de nouvelles caractéristiques à partir des données existantes peut être effectuée pour améliorer la performance du modèle. La transformation des données, telle que la normalisation ou l'encodage, est également réalisée afin de les rendre appropriées pour l'apprentissage automatique.

Modélisation

Dans cette phase, plusieurs modèles de machine learning sont développés et testés afin de trouver celui qui répond le mieux aux objectifs du projet. Différentes techniques et algorithmes sont explorés, en ajustant les paramètres de chaque modèle pour améliorer leurs performances. Cette étape implique souvent un processus itératif de construction, d'évaluation et de réglage des modèles afin d'obtenir les résultats les plus optimaux.

Évaluation

Une fois que les modèles ont été entraînés et testés, il est essentiel de les évaluer de manière approfondie. L'évaluation ne se limite pas seulement à des métriques techniques, mais elle doit également prendre en compte les objectifs commerciaux du projet. Il est important d'analyser comment les modèles se traduisent concrètement dans le contexte métier, en évaluant leur pertinence, leur fiabilité et leur capacité à résoudre les problématiques identifiées.

Déploiement

La création du modèle ne marque pas la fin du projet, car il doit être déployé pour être utilisé par les utilisateurs finaux. Cette étape implique la mise en place d'une infrastructure appropriée pour héberger le modèle, ainsi que son intégration dans les systèmes existants. Il est également crucial de fournir une interface conviviale

permettant aux utilisateurs d'interagir facilement avec le modèle et de tirer pleinement parti de ses capacités, tout en assurant sa maintenance et sa mise à jour régulières.

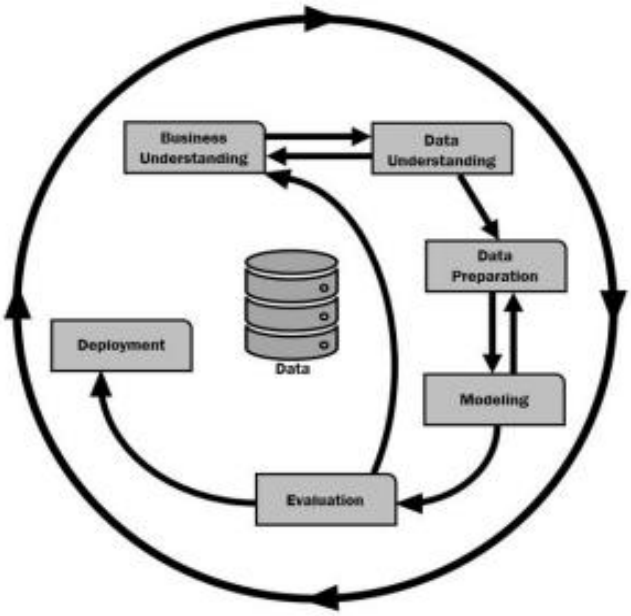


Figure 1.2 : Les étapes du CRISP-DM (F.M.P, 2021).

1.2 Types d'apprentissage automatique

Il existe tant de types différents de systèmes d'apprentissage automatique qu'il est utile de les classer dans des catégories générales, en fonction des critères suivants :

- S'ils sont ou non entraînés avec une supervision humaine (apprentissage supervisé, non supervisé, semi-supervisé et apprentissage par renforcement)
- S'ils peuvent ou non apprendre de manière incrémentale en temps réel (apprentissage en ligne versus apprentissage par lots)
- S'ils fonctionnent en comparant simplement de nouveaux points de données à des points de données connus, ou s'ils détectent des motifs dans les données d'entraînement pour construire un modèle prédictif, à la manière des scientifiques (apprentissage basé sur les instances versus apprentissage basé sur les modèles)

Ces critères permettent de mieux comprendre les différentes catégories d'apprentissage automatique et d'orienter le choix des méthodes appropriées en fonction des besoins spécifiques de chaque projet.

1.2.1 Apprentissage supervisé

L'apprentissage supervisé, en français "Supervised Learning", est une méthode où nous disposons d'un échantillon de données d'entrée, également appelées variables exogènes (X), ainsi que de leur valeur correspondante pour la variable endogène y

(données en sortie). Le but de cette méthode est d'estimer une fonction f qui permet de réduire au maximum l'erreur.

$$\hat{y} = \hat{f}(X) \qquad (1.1)$$

$$y = \hat{y} + \epsilon = \hat{f}(X) + \epsilon \qquad (1.2)$$

Dans ce contexte, f est la fonction estimée. ϵ représente l'écart entre les valeurs estimées avec cette fonction (ŷ) et les valeurs réelles observées dans la variable endogène (variable de sortie) y. L'erreur résultante peut être décomposée en une erreur réductible, que l'on peut minimiser avec davantage de calculs ou un meilleur algorithme d'apprentissage, et une erreur irréductible, qui est inhérente à la nature stochastique du problème.

$$\epsilon = \epsilon_{rd} + \epsilon_{irr} \qquad (1.3)$$

L'apprentissage supervisé utilise donc un ensemble de données pré-étiquetées pour former l'algorithme, et le but est d'optimiser la fonction d'estimation de manière à minimiser l'erreur entre les prédictions de l'algorithme et les résultats réels. L'objectif ultime est de construire un modèle capable de faire des prédictions précises sur de nouvelles données non étiquetées.

Donc, l'apprentissage supervisé permet de traiter deux types de problèmes :

1.2.1.1 Régression

La régression est une tâche qui vise à prédire une valeur numérique cible : la variable endogène (dépendante ou de sortie), y est donc un vecteur comportant des valeurs numériques. Le modèle appris permet d'obtenir un y_i pour une nouvelle entrée X. Un exemple est de prédire le prix d'une voiture, à partir d'un ensemble de caractéristiques (kilométrage, âge, marque, etc.) appelées prédicteurs.
Pour entraîner le système, on devrait lui fournir de nombreux exemples de voitures, incluant à la fois leurs prédicteurs et leurs étiquettes (leurs prix).

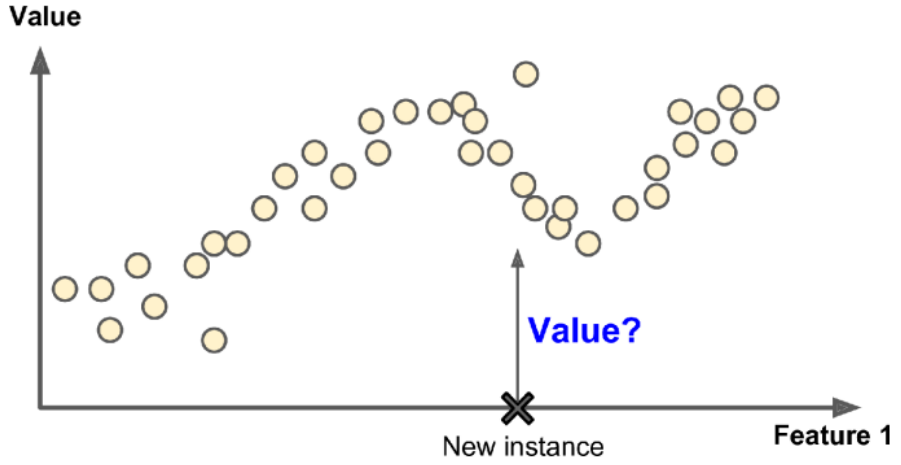


Figure 1.3 : Représentation graphique d'une régression (Géron Aurélien, 2022).

1.2.1.2 Classification

Dans ce type de problème, l'objectif est de construire un modèle capable d'attribuer chaque entrée de données X, tirée de l'ensemble des entrées X, à une classe d'un groupe de classes défini par Ω. Ici, la variable Y est catégorielle (soit nominale ou ordinale) et prend ses valeurs dans l'ensemble Ω.

Le filtre anti-spam est un bon exemple de cela : il est entraîné avec de nombreux exemples d'e-mails accompagnés de leur classe (spam ou non-spam), et il doit apprendre comment classer les nouveaux e-mails. Dans ce cas, la classification est binaire, car Ω est composé de deux résultats possibles : {spam = 1, non-spam= 0}.

Il est également possible de rencontrer des problèmes nécessitant une classification multi-classes.

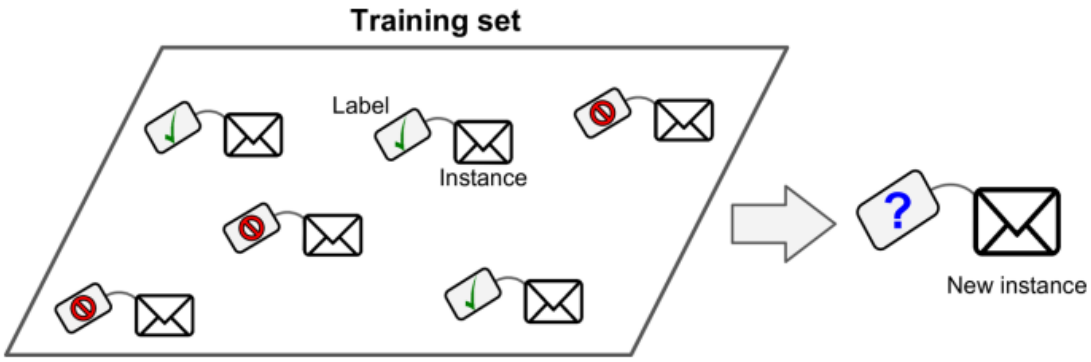


Figure 1.4 : Exemple de l'apprentissage supervisé (Géron Aurélien, 2022).

1.2.2 Apprentissage non-supervisé

L'apprentissage non supervisé est une autre forme d'apprentissage automatique. Dans cette méthode, nous avons seulement accès aux données d'entrée, notées X, sans avoir de variable observée ou cible, notée y. L'essentiel de ce type d'apprentissage repose sur le regroupement de données en différents clusters, une technique connue sous le nom de Clustering (IBM | Unsupervised). Les modèles d'apprentissage non supervisé sont utilisés également pour d'autres tâches, notamment : la détection d'anomalie, l'association et la réduction de dimensionnalité des données pour simplifier leur analyse.

1.2.2.1 Clustering

Est une technique d'exploration de données qui regroupe des données non étiquetées en fonction de leurs similitudes ou de leurs différences. Les algorithmes de clustering sont utilisés pour traiter des objets de données brutes et non classifiées en groupes représentés par des structures ou des motifs dans l'information. Les algorithmes de clustering peuvent être catégorisés en quelques types, spécifiquement exclusifs (k-moyennes), chevauchants (Soft, k-moyennes floues), hiérarchiques (Liaison de Ward, Liaison moyenne, Liaison complète/maximum, Liaison simple/minimum) et probabilistes (Gaussian Mixture Models).

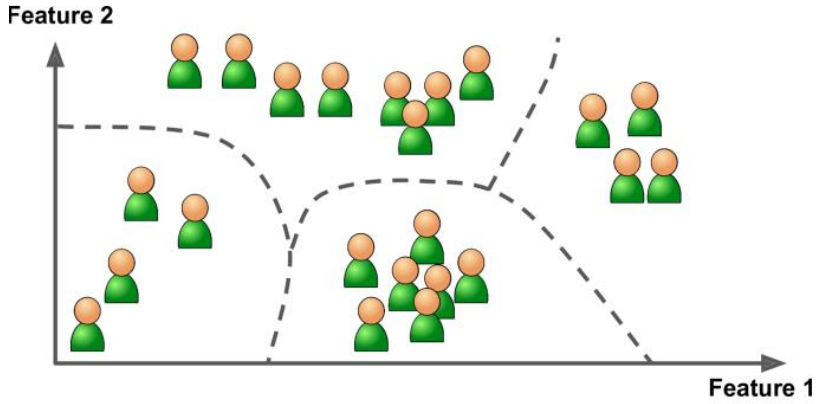


Figure 1.5 : Clustering (Géron Aurélien, 2022).

1.2.2.2 Réduction de dimensionnalité

Bien que davantage de données donnent généralement des résultats plus précis, cela peut également affecter la performance des algorithmes d'apprentissage automatique notamment avec le surapprentissage (annexe) et peut rendre difficile la visualisation des ensembles de données. La réduction de dimensionnalité est une technique utilisée lorsque le nombre de caractéristiques, ou dimensions, dans un ensemble de données donné est trop élevé. Elle réduit le nombre d'entrées de données à une taille gérable tout en préservant autant que possible l'intégrité de l'ensemble de données. Elle est couramment utilisée lors de la préparation des données, et il existe différentes méthodes de réduction de dimensionnalité qui peuvent être utilisées, telles que : Analyse en Composantes Principales, Décomposition en Valeurs Singulières, t-SNE (Figure 1.6) et Autoencoders.

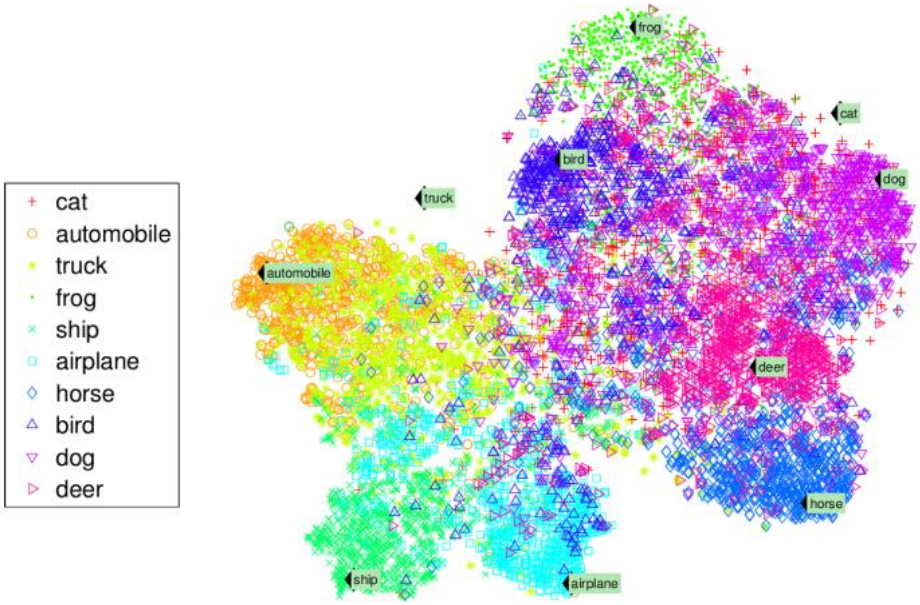


Figure 1.6 : Exemple de l'utilisation d'un t-SNE (Géron Aurélien, 2022).

1.2.2.3 Détection d'anomalies

Par exemple, détecter des transactions par carte de crédit inhabituelles pour prévenir la fraude, détecter des défauts de fabrication, ou éliminer automatiquement les valeurs aberrantes d'un ensemble de données avant de l'alimenter à un autre algorithme d'apprentissage. Le système est exposé à des instances majoritairement normales pendant l'entraînement, il apprend donc à les reconnaître ; puis, lorsqu'il voit une nouvelle instance, il peut déterminer si elle ressemble à une instance normale ou si elle est probablement une anomalie (Figure 1.7). Une tâche très similaire est la détection de nouveauté : elle vise à détecter de nouvelles instances qui semblent différentes de toutes les instances de l'ensemble d'entraînement. Cela nécessite d'avoir un ensemble d'entraînement très "propre", dépourvu de toute instance que vous souhaiteriez que l'algorithme détecte.

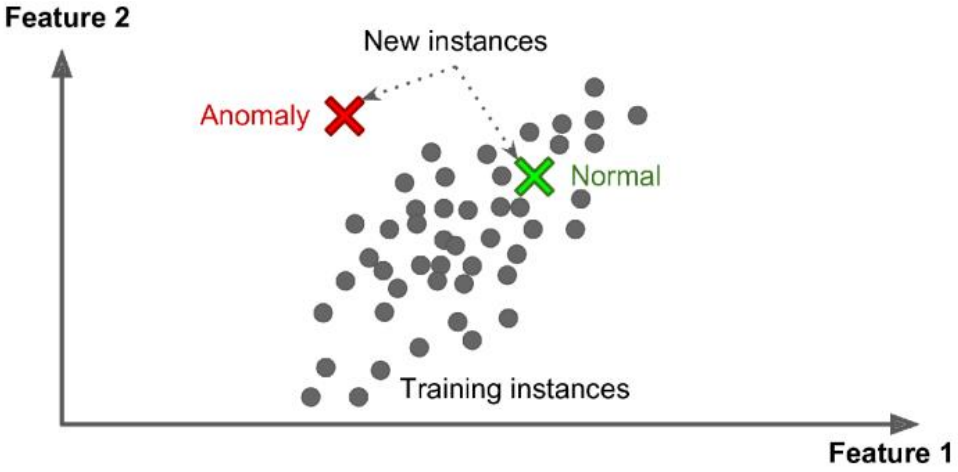


Figure 1.7 : Représentation des typologies d'instances (Géron Aurélien, 2022).

1.2.2.4 Règles d'association

Est une méthode basée sur des règles permettant de trouver des relations entre les variables d'un ensemble de données donné. Ces méthodes sont fréquemment utilisées pour l'analyse des paniers d'achat, permettant aux entreprises de mieux comprendre les relations entre différents produits. La compréhension des habitudes de consommation des clients permet aux entreprises de développer de meilleures stratégies de vente croisée et de recommandation.

Bien qu'il existe plusieurs algorithmes utilisés pour générer des règles d'association, tels que : Apriori, Eclat et FP-Growth, l'algorithme Apriori est le plus largement utilisé.

1.2.3 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est une catégorie de techniques d'apprentissage automatique qui utilise à la fois des données étiquetées et non étiquetées. Il s'agit d'une technique hybride entre l'apprentissage supervisé et non supervisé (Lilian Weng, 2021).

En général, l'idée centrale de la semi-supervision est de traiter un point de données différemment selon qu'il a une étiquette ou non :

- Les points étiquetés : l'algorithme utilisera une supervision traditionnelle afin de mettre à jour les poids du modèle
- Les points non étiquetés : l'algorithme minimise la différence de prédictions entre d'autres exemples d'entraînement similaires.

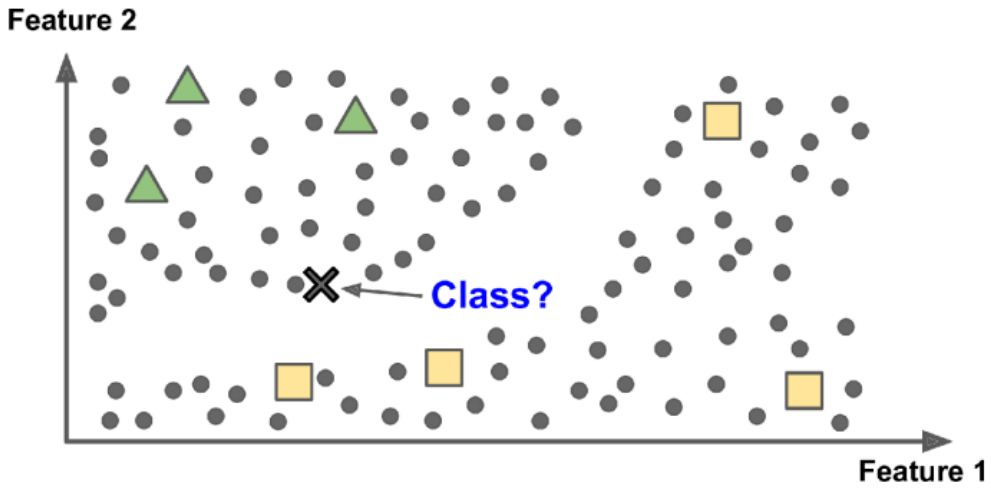


Figure 1.8 : Apprentissage semi supervisé a deux clusters (Géron Aurélien, 2022).

Ces algorithmes sont souvent une combinaison de techniques supervisées et non supervisées. Par exemple, les réseaux de croyance profonds (DBNs)[1] (Hilton G.E., 2006) sont basés sur des composants non supervisés appelés machines de Boltzmann restreintes (RBMs)[2]. Les RBMs sont entraînées séquentiellement de manière non supervisée, puis l'ensemble du système est affiné à l'aide de techniques d'apprentissage supervisé.

1.2.4 Apprentissage par renforcement

L'apprentissage par renforcement est une méthode d'apprentissage automatique où un système d'apprentissage, appelé "agent" apprend à prendre des décisions en réalisant certaines actions dans un environnement afin d'atteindre un certain objectif. Ce type d'apprentissage est assez particulier puisque son paradigme est assez différent. En effet, l'agent observe son environnement puis est capable de sélectionner et de réaliser des actions afin d'obtenir des récompenses ou des pénalités.

[1] (DBN) sont des réseaux de neurones avec de multiples couches cachées, qui peuvent apprendre efficacement à partir de données non étiquetées grâce à une méthode d'apprentissage non supervisé.

[2] (RBM) sont des modèles génératifs stochastiques à deux couches, non supervisés. Ils sont capables de détecter et d'apprendre des modèles et des caractéristiques dans les données d'entrée.

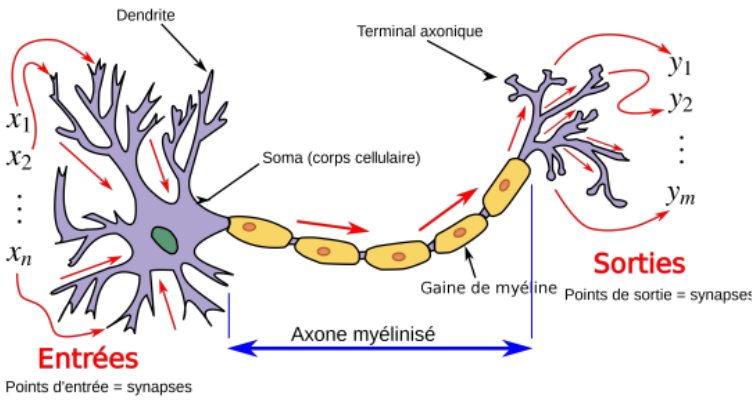
Dans ce processus, l'agent essaie de maximiser la somme totale des récompenses. L'agent interagit avec l'environnement, prend des actions et fait des erreurs. Au fil du temps, l'agent apprend la meilleure stratégie à suivre, appelée "policy", afin de maximiser ses récompenses. Cette "policy" définit les actions que l'agent choisit au cours d'une situation donnée.

L'apprentissage par renforcement est utilisé dans de nombreux domaines tels que la robotique, la logistique, le contrôle autonome de véhicules et notamment la programmation de robots pour leur apprendre à se mouvoir. Un progrès significatif a été réalisé en 2017 lorsque l'agent AlphaGo a pu vaincre un champion du jeu de plateau Go[3] en appliquant une politique apprise suite à l'analyse de millions de parties du jeu. Cette avancée a conduit les développeurs de jeux vidéo, comme EA Sports[4], à utiliser ces algorithmes dans la programmation de ses simulations de sports de tout genre.

1.3 Apprentissage profond

Un type d'apprentissage particulier et très populaire est l'apprentissage profond (deep learning). Cette méthode d'apprentissage s'inspire directement du cerveau humain et repose sur l'utilisation de réseaux de neurones artificiels. (Artificial Neural Networks : ANN)

Figure 1.9 : Exemple d'un neurone biologique.



Un ANN est basé sur un rassemblement de nœuds connectés appelés neurones artificiels, qui tentent de modéliser les neurones d'un cerveau biologique.

3 "Go" est un ancien jeu de plateau stratégique d'origine chinoise où deux joueurs placent alternativement des pierres noires et blanches sur un tableau à grille, visant à contrôler plus de territoire que leur adversaire.

4 "EA Sports" est une division d'Electronic Arts (EA) spécialisée dans les jeux vidéo sportifs tels que FIFA, Madden NFL et NHL. Leurs titres sont reconnus pour offrir des expériences réalistes et immersives dans différents sports.

1.3.1 Perceptron

Le réseau neuronal existant le plus basique est le perceptron, qui est composé d'un seul neurone formel. Il est considéré comme le processus d'apprentissage le plus ancien qui ait été réalisé, pour la classification de deux classes linéairement séparables. Un perceptron est composé d'une ou plusieurs entrées, d'une unité de traitement et d'une sortie.

La Figure 1.10 ci-dessous montre la structure d'un neurone artificiel qui procède par la sommation pondérée de son vecteur d'entrée :

$$A = (a_1; a_2; \ldots; a_M) \in \Re^M \qquad (1.4)$$

Et le vecteur de poids synaptique :

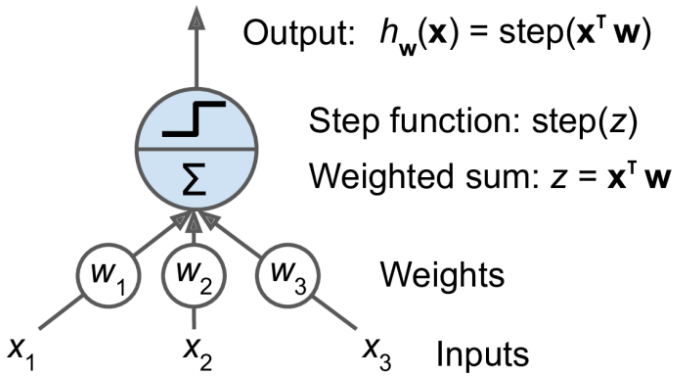
$$W = (w_1; w_2; \ldots; w_M) \in \Re^M \qquad (1.5)$$

La somme obtenue sera sous cette forme :

$$z_j = \sum_{i=1}^{p} w_{ij}^h x_i + b_j^h, \;\; j = 1,2,\ldots,h \qquad (1.6)$$

Il applique ensuite une fonction d'activation qui fournit une sortie y. La fonction d'activation imite le fonctionnement du soma (noyau). Elle effectue la somme des entrées pondérées et la compare à un seuil pour activer ou non la sortie en envoyant un potentiel somatique.

Figure 1.10 : Représentation graphique d'un neurone artificiel.



1.3.2 Perceptron multicouche

Le Perceptron multicouche (PMC) est une extension du perceptron. Composé de trois types de couches, où chaque neurone est lié à la totalité des neurones des couches adjacentes, ce qui donne un réseau entièrement connecté, comme illustré par la Figure 1.11 La dimension des couches d'entrée et de sortie détermine la nature du problème.

- Le nombre de neurones dans la couche d'entrée déterminent la dimension de la donnée traitée.
- Le nombre de neurones dans la couche de sortie détermine le nombre de classes.

- Les couches cachées, leurs nombres et le nombre de neurones qui les constituent est un problème de conception.
- Avec peu de neurones cachés, le modèle sera faible face aux décisions complexes (modèles non linéaires). Au contraire, trop de neurones réduisent la généralisation du modèle dû à un sur-apprentissage.
- Ainsi, un réglage expérimental est nécessaire pour trouver le meilleur compromis entre le nombre de nœuds cachés et la performance de généralisation du réseau.

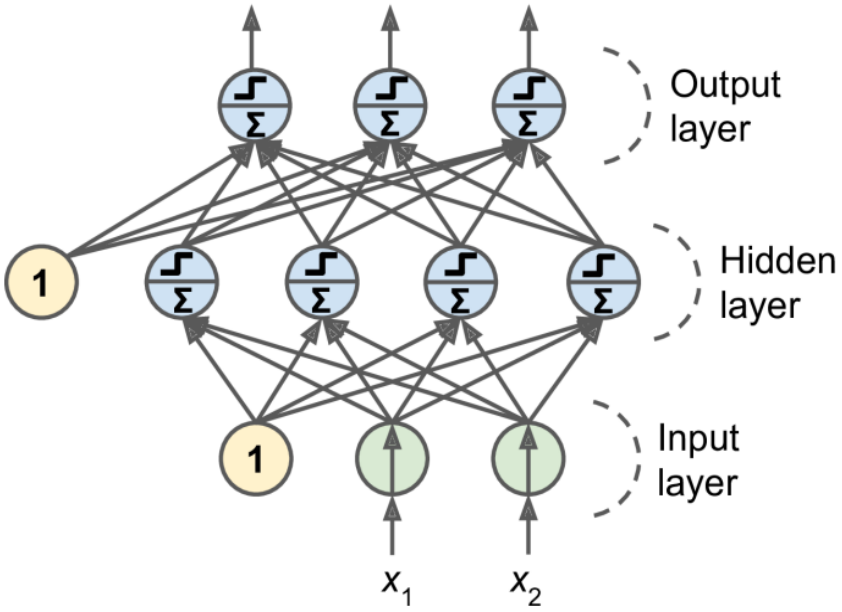


Figure 1.11 : Architecture d'un perceptron multicouche.

Le PMC est entraîné par l'algorithme de rétropropagation [5]du gradient de l'erreur. Il permet aux informations de circuler en sens inverse dans le réseau afin de calculer le gradient. Il consiste à ajuster les poids synaptiques de tous les neurones des différentes couches en calculant l'erreur quadratique moyenne E entre le vecteur de sortie estimé y et le vecteur de sortie réelle y_r par l'Équation 1.7 suivante :

$$E = \frac{1}{2}\sum_{i=1}^{N}(y_r^i - y^i)^2 \qquad (1.7)$$

Les poids synaptiques sont ensuite modifiés tel que :

$$\Delta w(t+1) = -\alpha \frac{\delta E}{\delta w} + \mu \Delta w(t) \qquad (1.8)$$

t : Itération en cours (correspond au passage d'une donnée à travers le réseau).

$\Delta w()$: Changement de poids au fil des itération

$\frac{\delta E}{\delta w}$: Gradient de l'erreur par rapport au poids w

[5] La rétropropagation est une technique qui ajuste les poids dans les réseaux de neurones en calculant l'erreur à partir de la sortie. Cette erreur est ensuite répartie en arrière à travers le réseau pour mettre à jour les poids.

Il est à noter que l'objectif est de minimiser une fonction coût représentée par l'erreur quadratique moyenne E jusqu'à son minimum local ce qui est appelée descente du gradient. Quant au paramètre α, il représente le taux d'apprentissage ou le pas d'apprentissage, qui constitue le pas que fait la descente du gradient dans la direction du minimum local déterminant ainsi la rapidité ou la lenteur avec laquelle le minimum local et donc les poids optimaux sont approchés. La Figure 1.12 montre l'impact et l'importance du pas d'apprentissage.

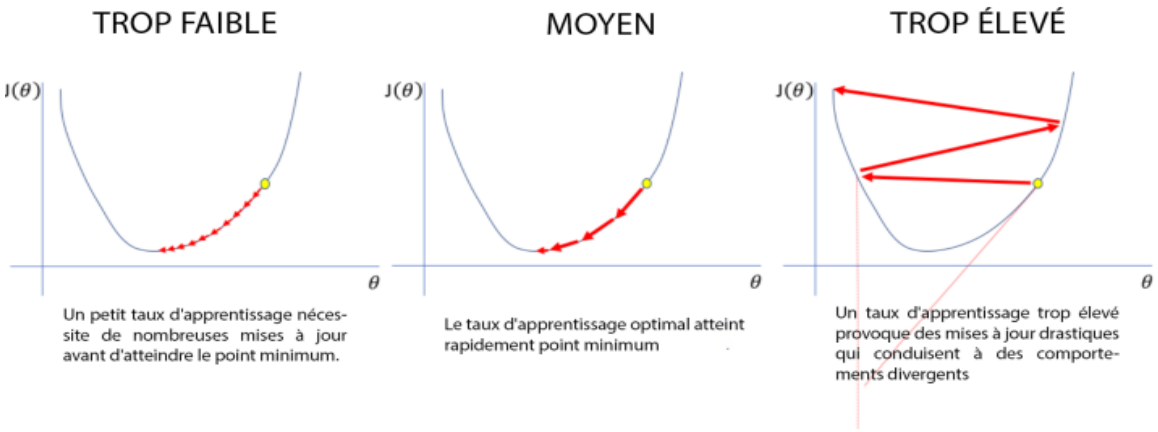


Figure 1.12 : Illustration de l'architecture CNN.

D'un autre côté, le paramètre μ appelé momentum permet de conserver les informations relatives à la dernière mise à jour des poids synaptiques (itération t) pour en tenir compte dans la mise à jour actuelle (itération t+1). Il évite ainsi de rester coincé dans un minimum local ainsi que les effets d'oscillations.

Il existe plusieurs types de réseaux de neurones artificiels comme les réseaux simples (feed forward networks) employé pour la régression ou la classification, les réseaux récurrents (RNNs) utilisé dans le traitement de langage naturel et sur des données de nature séquentiel, les réseaux à convolution (CNNs) utilisé principalement sur les données sous forme d'images ou de vidéos ainsi que les réseaux génératifs (VAEs, GANs, etc) qui permettent de générer de nouvelles données comme des images, des vidéos ou autre.

1.4 Apprentissage génératif

L'apprentissage génératif est une branche de l'apprentissage automatique qui se concentre sur la création de modèles capables de générer de nouvelles données. Les modèles génératifs utilisent des réseaux de neurones pour apprendre la distribution de probabilité des données d'entraînement, ce qui permet d'identifier les motifs et les structures au sein des données existantes afin de produire de nouvelles instances de données qui sont similaires à celles qu'ils ont apprises (Lilian Weng, 2018).

L'une des percées avec les modèles génératifs est la capacité d'exploiter différentes approches d'apprentissage, y compris l'apprentissage non supervisé ou semi-supervisé pour la formation. Cela a donné aux organisations la possibilité d'exploiter plus facilement et rapidement une grande quantité de données non étiquetées pour créer des modèles de base. Comme le suggère le nom, les modèles de base peuvent être

utilisés comme une base pour des systèmes d'IA capables d'effectuer plusieurs tâches, dont la synthétisation de données.

1.4.1 Synthétisation de données

La synthétisation de données représente le processus de création de données synthétiques ou artificielles, qui reproduisent fidèlement les propriétés des données réelles. Ces données synthétiques sont utilisées pour optimiser les résultats de divers algorithmes d'apprentissage automatique.

Cette méthodologie est de plus en plus adoptée dans des contextes où la collecte de données réelles est complexe ou même irréalisable. Son utilisation a gagné en popularité, à tel point que les estimations suggèrent qu'en 2030, les données synthétiques éclipseront totalement les données réelles dans les modèles d'IA, selon Gartner[6] (Gartner, 2023).

La synthétisation de données offre l'opportunité de générer une grande quantité de données qui peuvent améliorer significativement les performances des modèles de Machine Learning. L'un des grands avantages de cette méthode est qu'elle n'introduit pas de biais dans les modèles, étant donné que les données synthétiques sont construites en respectant les propriétés intrinsèques des données réelles.

Par conséquent, la synthétisation de données s'avère être une approche puissante et prometteuse pour relever les défis de la collecte de données, tout en améliorant l'efficacité des algorithmes d'apprentissage automatique.

Notre focus dans ce qui suit sera principalement axé sur les données tabulaires, même s'il est possible de générer des données dans divers formats, tels que des images, des vidéos, de l'audio et d'autres formats.

Données tabulaires : Il y'a deux types des données tabulaires, qui sont :

Les données numériques : ce sont les données qui peuvent être continues comme le poids, la longueur, etc. Ou des données discrètes comme le nombre de conteneurs par jour.

Les données catégoriques : qui eux se divisent en deux : ordinales comme les jours de la semaine ou nominales comme la couleur.

La Figure 1.13 résume les types des données tabulaires :

[6] Gartner est une société de conseil et de recherche dans le domaine des technologies avancées. Fondée en 1979 par Gideon Gartner et basée à Stamford.

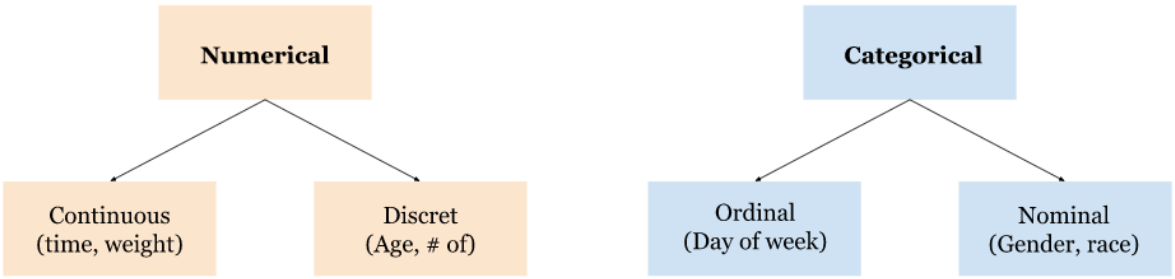


Figure 1.13: Les différents types de Données tabulaires.

1.4.2 Approches de la synthétisation des données

Il existe principalement 3 approches pour synthétiser des données et plusieurs méthodes dans chaque approche :

1.4.2.1 Méthodes statistiques

Bootstrap/Rééchantillonnage : Le bootstrap est une méthode statistique qui génère des données en tirant des échantillons de manière aléatoire avec remplacement à partir de l'ensemble de données original (Hesterberg,T., 2011). Cette méthode est souvent utilisée pour estimer la variabilité d'un estimateur ou pour obtenir des intervalles de confiance pour un paramètre (annexe).

Simulation de Monte Carlo[7] : utilisée pour modéliser la probabilité de différents résultats dans un processus qui ne peut pas être facilement prédit en raison de l'intervention de variables aléatoires (Will, k., 2023). C'est une technique utilisée pour comprendre l'impact du risque et de l'incertitude.

Loi usuelle : Ces méthodes supposent que les données suivent une certaine loi usuelle (loi normale, binomiale, de Poisson, etc.) et génèrent des échantillons à partir de cette loi.

1.4.2.2 Simulation à événements discrets

Cette approche est couramment utilisée lorsque le processus de génération des données est bien maîtrisé, ce qui est le cas dans de nombreuses applications industrielles. Le principe consiste à modéliser l'intégralité du processus étudié à l'aide d'un logiciel de simulation tel que OLGA simulator[8] (annexe) et Pipesim.

[7] La simulation de Monte Carlo a été nommée d'après la destination de jeu à Monaco car le hasard et les résultats aléatoires sont au cœur de cette technique de modélisation, tout comme ils le sont pour des jeux comme la roulette, les dés et les machines à sous.

[8] OLGA est un outil de modélisation pour le transport simultané de pétrole, de gaz naturel et d'eau dans le même pipeline, également appelé transport multiphase.

1.4.2.3 Apprentissage profond

Tabular Variational Auto-Encoder (VAE) : algorithme non supervisé qui peut apprendre la distribution d'un ensemble de données original et générer des données synthétiques via une double transformation, connue sous le nom d'architecture codage-décodage (Kingma & Welling, 2014). Le modèle formule une erreur de reconstruction, qui peut être minimisée grâce à un entraînement itératif. Les VAE proviennent de la famille des autoencodeurs. En tant que modèles génératifs, ils sont très efficaces pour générer des modèles complexes.

Ils fonctionnent en deux étapes. D'abord, un réseau encodeur transforme une distribution complexe originale en une distribution latente. Ensuite, un réseau décodeur transforme la distribution en retour à l'espace original. Cette double transformation, codage-décodage, peut sembler complexe à première vue, mais elle est nécessaire pour formuler une erreur de reconstruction quantifiable. Minimiser cette erreur est l'objectif de l'entraînement des VAE et ce qui en fait la fonction de transformation souhaitée, tandis qu'un objectif de régularisation supplémentaire contrôle la forme de la distribution latente.

Les CTGANs (Conditional Tabular GANs) : sont une variante des Generative Adversarial Networks (GANs) (Xu,L., 2019) spécifiquement adaptée aux données tabulaires. Ils sont utilisés pour générer des données synthétiques réalistes en entraînant un générateur et un discriminateur de manière antagoniste. Le générateur prend des données aléatoires en entrée et les transforme pour ressembler aux données réelles, tandis que le discriminateur essaie de distinguer les données réelles des données générées. Les CTGANs permettent de générer des données tabulaires très similaires à la réalité et sont particulièrement utiles pour l'analyse de données structurées. Cependant, les CTGANs peuvent être plus difficiles à entraîner que les VAEs et peuvent présenter des problèmes tels que le surapprentissage ou le phénomène d'effondrement de mode. Malgré cela, les CTGANs offrent une approche puissante pour la génération de données synthétiques.

Modèles de diffusion : Les modèles de diffusion dans la génération de données tabulaires sont des algorithmes inspirés de la thermodynamique et basés sur l'apprentissage non supervisé (Sohl-Dickstein, 2015). Ils fonctionnent en corrompant les données d'entraînement en ajoutant du bruit gaussien jusqu'à ce que la donnée devienne du bruit pur, puis en entraînant un réseau neuronal à inverser ce processus, en éliminant progressivement le bruit jusqu'à produire une nouvelle donnée. Les modèles de diffusion offrent une grande stabilité d'entraînement et peuvent produire des résultats de haute qualité tant pour les images que pour l'audio. Ils permettent de générer des données réalistes dans différents formats.

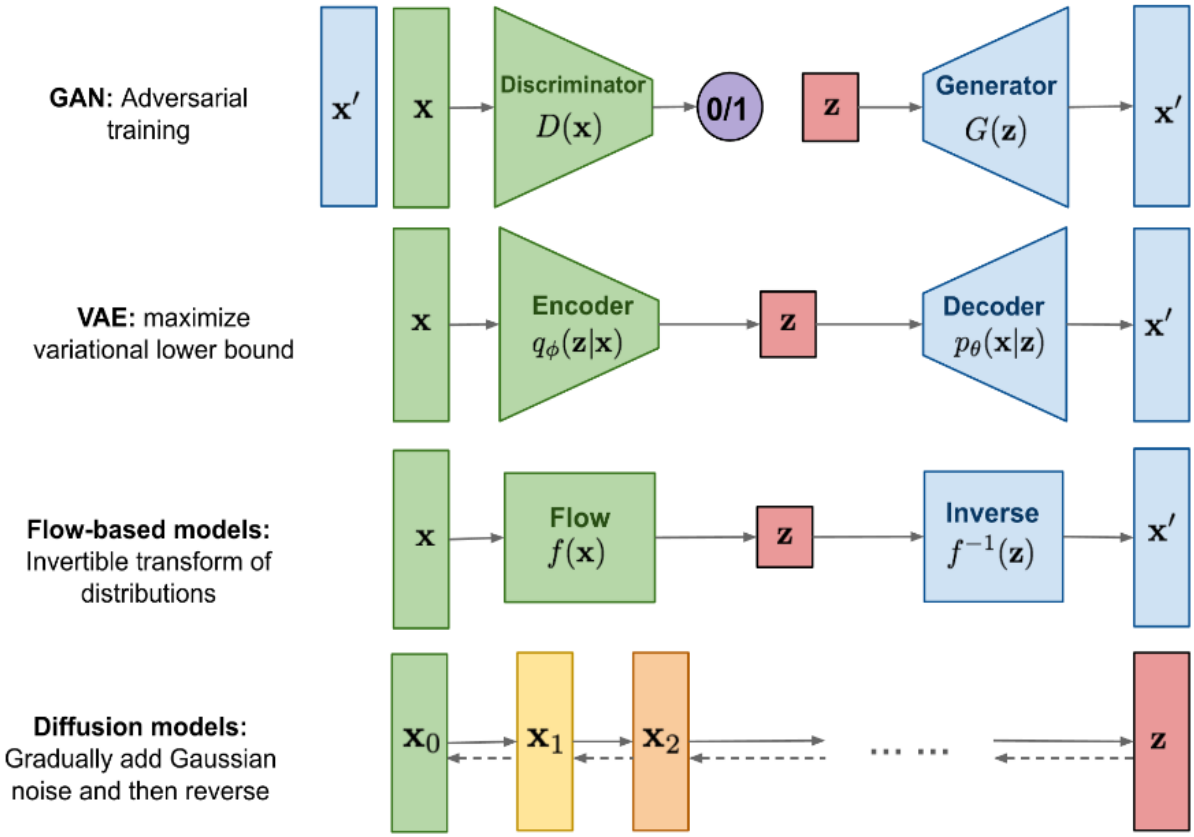


Figure 1.14 : Vue d'ensemble sur les types des modèles génératifs.

1.5 Algorithmes identifiés

Dans cette section, nous allons présenter les algorithmes d'apprentissage automatique identifiés pour notre projet. Ces algorithmes sont utilisés pour résoudre des problèmes de classification et de régression, ainsi que pour des tâches non supervisées telles que le clustering. Les algorithmes que nous avons sélectionnés sont les suivants :

1.5.1 Algorithmes supervisés

Support Vector Machines (SVM)

Les machines à vecteurs de support (SVM) reposent sur des concepts mathématiques solides. L'objectif de SVM est de trouver un hyperplan qui maximise la marge entre les classes dans un espace multidimensionnel. Mathématiquement, cela revient à résoudre un problème d'optimisation quadratique. L'hyperplan est défini par une fonction de décision linéaire, qui peut être exprimée comme :

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b) \qquad (1.9)$$

où α_i sont les coefficients de Lagrange, y_i sont les étiquettes de classe, x_i sont les vecteurs de support et $K(x, x_i)$ est une fonction de noyau qui mesure la similarité entre les points de données. Des fonctions de noyau couramment utilisées incluent le noyau linéaire, le noyau polynomial et le noyau RBF (Radial Basis Function).

En ce qui concerne la régression, on peut utiliser le modèle SVM pour modéliser des relations non linéaires en introduisant l'astuce du noyau (Kernel trick). L'idée est de remplacer xi par une fonction non linéaire φ(x_i) et d'utiliser le dual du programme d'optimisation. Le problème de minimisation quadratique peut être formulé comme suit :

$$min\ \frac{1}{2}\ \|\ w\ \|^2 + C\ \sum_{i=1}^{n} = \xi_i \qquad (1.10)$$
$$s.c. |y_i - wT\varphi(x_i)|\ \leq\ \varepsilon + \xi_i$$
$$\xi_i\ \geq\ 0$$

où w est le vecteur des poids, C'est un coefficient de tolérance, ε est la marge d'erreur et ξ_i sont des variables d'écart. Ce programme mathématique peut être résolu à l'aide de la méthode de Lagrange ou d'autres méthodes d'optimisation.

Passons maintenant aux arbres de décision. Les arbres de décision sont des algorithmes de classification et de régression. L'idée est de construire un arbre binaire T avec une certaine profondeur prédéfinie. Chaque nœud k de l'arbre est divisé selon une dimension k et un seuil tk. Le couple (k, tk) est trouvé en minimisant la fonction objective suivante, connue sous le nom de CART (Classification and Regression Trees) :

$$J(k, tk)\ =\ (\frac{m_{gauche}}{m})\ MSE_{gauche}\ +\ (\frac{m_{droit}}{m})\ MSE_{droit} \qquad (1.11)$$

où mgauche/droit est le nombre de données dans le nœud à gauche/droite, m est le nombre total de données et MSEgauche/droit est l'erreur quadratique moyenne entre les valeurs estimées ŷ et les valeurs réelles yi. L'algorithme itère à travers toutes les combinaisons possibles pour trouver une bonne solution, mais pas nécessairement la solution optimale en termes de l'arbre.

Ces formulations mathématiques illustrent comment les SVM et les arbres de décision utilisent des équations pour décrire leur fonctionnement et comment ils sont optimisés pour trouver les hyperplans ou les divisions qui regroupent efficacement les données.

K-Nearest Neighbors (KNN)

KNN est basé sur la distance entre les points de données. Mathématiquement, pour chaque point de données à classer, l'algorithme calcule les distances entre ce point et ses k voisins les plus proches. La classe attribuée est déterminée par un vote majoritaire parmi les voisins. La distance euclidienne est souvent utilisée, mais d'autres mesures de distance peuvent également être utilisées, telles que la distance de Manhattan ou la distance de Minkowski.

Pour un problème de classification donné, supposons que l'on a deux étiquettes pour classer des points : rouge et bleu. Nous avons un point noir en input, l'algorithme tâchera alors de trouver ses K plus proches voisins et vérifier la couleur de ces voisins. Si la majorité est étiquetée « rouge » alors le point noir sera classé parmi les points rouges. Ces K points sont trouvés par l'algorithme à travers une métrique de distance.

Pour les variables réelles en entrée, la plus populaire est la distance Euclidienne (Équation 1.12), connue sous le nom de Norme 2 entre 2 points p et q :

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} \quad (1.12)$$

D'autres métriques peuvent être utilisées comme la distance de Manhattan[9] qui calcule la distance entre 2 vecteurs en utilisant la somme de leurs différences absolues. On peut également citer la distance de Hamming[10] ou alors la distance de Minkowski qui est une généralisation des distances Euclidienne et de Manhattan.

Un input x à classifier sera assigné à la classe avec la probabilité la plus importante suivante :

$$P(y = j \mid X = x) = \frac{1}{K}\sum_{i \in A} I(y^{(i)} = j) \quad (1.13)$$

Naive Bayes

Naive Bayes est un algorithme probabiliste qui repose sur le théorème de Bayes et l'hypothèse d'indépendance conditionnelle entre les caractéristiques. Mathématiquement, cela peut être formulé comme :

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(y) \times P(x_1|y) \times P(x_2|y) \times \ldots \times P(x_n|y)}{P(x_1, x_2, \ldots, x_n)} \quad (1.14)$$

où y est la classe, x_1 ,x_2 , ..., x_n sont les caractéristiques et P(.) représente les probabilités. Les probabilités conditionnelles $P(x_i|y)$ sont généralement estimées à partir des données d'entraînement en utilisant des méthodes telles que la distribution multinomiale ou la distribution de Bernoulli.

L'hypothèse naïve sous-jacente à Naive Bayes est que les caractéristiques sont indépendantes les unes des autres. Malgré cette simplification, Naive Bayes est souvent utilisé avec succès dans diverses applications, telles que la classification de documents en spam ou non spam. Par exemple, en analysant les fréquences des mots dans les courriels, Naive Bayes peut estimer les probabilités que certains mots apparaissent dans les courriels indésirables, et ainsi classer les nouveaux courriels en conséquence.

Random Forest

Random Forest est un algorithme d'ensemble qui combine plusieurs arbres de décision pour effectuer la classification. Chaque arbre de décision est construit à partir d'un sous-ensemble aléatoire des données d'entraînement et des caractéristiques. L'agrégation des prédictions des arbres individuels permet d'obtenir une prédiction finale plus robuste et moins sujette au surapprentissage. Les Random Forests sont particulièrement adaptées aux ensembles de données complexes et bruités. Par exemple, dans une tâche de prédiction de la qualité du vin en fonction de ses

9 La distance de Manhattan tire son nom de l'agencement en grille des rues de Manhattan, New York. Elle mesure la distance entre deux points en suivant uniquement les trajets parallèles aux axes, semblable à la façon dont un taxi se déplacerait dans la ville.

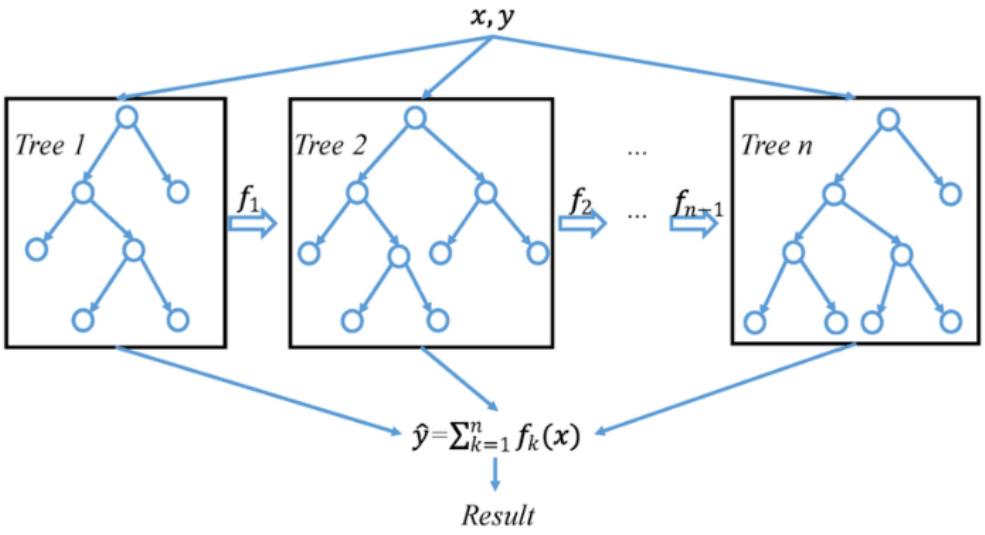
10 La distance de Hamming est une mesure utilisée pour comparer deux chaînes de même longueur. Elle est définie comme le nombre de positions où les symboles correspondants sont différents, souvent utilisée dans la correction d'erreurs et la cryptographie.

caractéristiques chimiques, Random Forest peut apprendre à partir des différentes caractéristiques (par exemple, l'acidité, le pH, la teneur en alcool, etc.) pour prédire la qualité du vin avec précision.

Figure 1.15 : Schématisation de l'algorithme Random Forest.

1. On génère B échantillons de données à travers le bootstrapping.

2. On entraîne B arbres de décisions avec un nombre aléatoire des variables exogènes disponibles (en général si on a p variables exogènes



(dimension/facteurs), chaque arbre sera entraîné sur $\sqrt{p}$ de variables).

3. On agrège le résultat par la moyenne de ces arbres :

$$\hat{f}_{moy} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b \qquad (1.15)$$

Cette procédure permet d'entraîner un modèle fiable et robuste avec une variance réduite.

Gradient Boosted Trees

L'algorithme du Gradient Boosted Trees (GBDT) est un autre algorithme puissant basé sur l'approche d'apprentissage par ensemble pour la régression. Il entraîne un ensemble d'arbres de décision de manière séquentielle en se concentrant sur les résidus (le gradient d'erreur) de chaque arbre. Cette approche séquentielle permet une réduction continue de l'erreur.

L'algorithme du GBDT fonctionne selon les étapes suivantes :

1. Pour la première itération, les résidus (r_i) sont initialisés avec les valeurs réelles (y_i) pour toutes les données.
2. Un nombre B d'arbres est choisi pour être entraîné.
3. Pour chaque itération b de 1 à B :
    a. Un arbre $\hat{f}^b$ avec découpes (d + 1 feuilles) est entraîné.
    b. Le modèle est mis à jour en ajoutant cet arbre multiplié par un coefficient $\lambda < 1$:

$$\hat{f}(x) = \hat{f}(x) + \lambda \hat{f}^b(x) \qquad (1.16)$$

    c. Les résidus sont mis à jour :

$$r_i = r_i - \lambda \hat{f}^b(x_i) \qquad (1.17)$$

4. Le modèle final est calculé en combinant les prédictions de tous les arbres :

$$\hat{f}(x) = \Sigma(\lambda \hat{f}^b(x)), pour\ b\ allant\ de\ 1 \text{ à } B. \qquad (1.18)$$

L'algorithme GBDT est utilisé pour améliorer progressivement les prédictions en se concentrant sur les erreurs résiduelles. Les arbres sont entraînés de manière itérative pour capturer les relations résiduelles non expliquées par les arbres précédents. Les coefficients λ contrôlent l'importance de chaque arbre dans le modèle final.

Pour représenter visuellement le processus d'entraînement des arbres et l'amélioration des prédictions, vous pouvez utiliser des graphiques montrant les prédictions successives à chaque étape de l'algorithme, ainsi que les résidus mis à jour. Ces graphiques peuvent être générés en utilisant des bibliothèques graphiques telles que Matplotlib, en traçant les valeurs prédites par le modèle à chaque étape.

Il est important de noter que la mise en œuvre détaillée de l'algorithme GBDT peut varier selon la bibliothèque ou le langage de programmation utilisé. Vous pouvez consulter la documentation spécifique de la bibliothèque GBDT que vous utilisez pour obtenir des exemples concrets et des instructions détaillées sur la visualisation du processus d'apprentissage par ensemble des arbres.

eXtreme Gradient Boosting

XGBoost est le principal modèle ensembliste (annexe) pour travailler avec des données tabulaires standard.

Pour atteindre une précision maximale, les modèles XGBoost nécessitent plus de connaissances et de mise au point des modèles que des techniques comme RF. XGBoost est une implémentation de l'algorithme des arbres de décision (DT) renforcés par gradient.

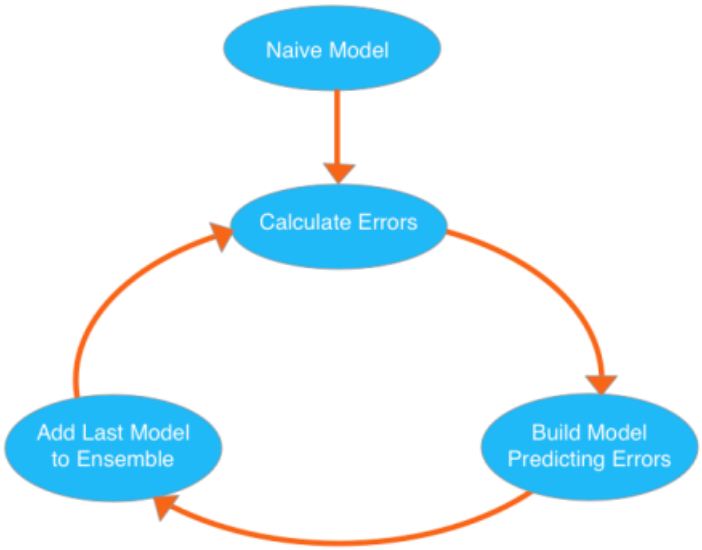


Figure 1.16 : Cycle pour entraîner un modèle XGBoost.

L'algorithme passe par des cycles qui construisent sans cesse de nouveaux modèles et les combinent en un modèle d'ensemble. Le cycle débute en calculant les erreurs pour chaque observation dans l'ensemble de données, pour ainsi construire ensuite un nouveau modèle pour les prévoir, enfin, celui-ci ajoute les prédictions de ce modèle de prédiction des erreurs à l'ensemble des modèles.

Pour faire une prédiction, XGBoost ajoute les prédictions de tous les modèles précédents et peut utiliser ces prédictions pour calculer de nouvelles erreurs, construire le modèle suivant et l'ajouter à l'ensemble.

1.5.2 Algorithmes non-supervisé

K-means (K-moyennes)

K-means est un algorithme de clustering largement utilisé qui vise à partitionner les données en k clusters en minimisant la variance intra-cluster. L'algorithme fonctionne en itérant entre l'affectation des points de données aux clusters les plus proches et la mise à jour des centres de ces clusters. Cette méthode est basée sur la notion d'une distance euclidienne entre les points de données dans l'espace des caractéristiques.

K-means est basé sur la minimisation de la variance intra-cluster. Mathématiquement, l'algorithme cherche à minimiser la somme des carrés des distances entre les points de données et les centres de cluster attribués. La fonction objectif à minimiser est définie comme :

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \qquad (1.19)$$

où k est le nombre de clusters, C_i représente le ième cluster, x est un point de données et μ_i est le centre du cluster *i*.

Par exemple, dans le domaine de la segmentation des clients en fonction de leurs habitudes d'achat, K-means peut être utilisé pour regrouper les clients en différents segments, tels que les acheteurs fréquents, les acheteurs occasionnels, etc.

Gaussian Mixture Models (GMM)

GMM est basé sur la modélisation probabiliste à partir de distributions gaussiennes. Mathématiquement, GMM cherche à maximiser la log-vraisemblance des données en utilisant l'estimation des paramètres des gaussiennes. Supposons que nous avons un ensemble de données $X = \{x_1, x_2,..., x_n\}$, où chaque x_i est un vecteur de caractéristiques. Nous supposons que les données sont générées à partir d'un mélange de k distributions gaussiennes. La densité de probabilité d'un point xi est donnée par :

$$P(x_i) = \sum_{j=1}^{k} \pi_j N(x_i|\mu_i, \sum_j' k) \qquad (1.20)$$

où π_j est le poids de la j-ème composante gaussienne, μj est le vecteur de moyenne de la j-ème composante gaussienne et Σ_j est la matrice de covariance de la j-ème composante gaussienne. Les poids πj doivent être positifs et sommer à 1 ($\Sigma^k_{j=1}\ \pi_{j=1}$).

L'estimation des paramètres des gaussiennes dans GMM est généralement réalisée en utilisant l'algorithme de maximisation de l'espérance (EM). L'algorithme EM itère entre deux étapes : l'étape d'espérance (E-step) et l'étape de maximisation (M-step). Dans l'E-step, les responsabilités des points de données pour chaque composante gaussienne sont calculées à l'aide du théorème de Bayes. Dans le M-step, les paramètres des gaussiennes sont mis à jour en maximisant la log-vraisemblance pondérée des données. L'algorithme EM continue à itérer jusqu'à ce que la convergence soit atteinte.

Density-Based Spatial Clustering of Applications with Noise

L'algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifie les clusters en se basant sur les zones de haute densité dans l'espace des caractéristiques. Mathématiquement, la densité d'un point xi est définie comme le nombre de points situés dans un rayon ε autour de xi, noté $N_\varepsilon(x_i)$:

$$N_\varepsilon(x_i) = \{x_j \in X | dist(x_i, x_j) \leq \varepsilon\} \qquad (1.21)$$

Le paramètre minPts représente le nombre minimum de points requis dans le voisinage ε pour qu'un point soit considéré comme central :

$$|N_\varepsilon(x_i)| \geq minPts \qquad (1.22)$$

Les points centraux sont les points clés pour former des clusters. En attribuant les points à des clusters, les règles suivantes sont appliquées :

1. Un point x_i est attribué au même cluster que ses points centraux voisins.
2. Les points dans le voisinage ε d'un point central sont également attribués au même cluster.
3. Les points isolés qui ne sont pas des points centraux ni des voisins de points centraux sont considérés comme du bruit et ne sont pas attribués à un cluster spécifique.

Cela permet de former des clusters de points densément connectés. Les clusters détectés par DBSCAN peuvent être représentés par un ensemble de clusters $C = \{C_1, C_2, ..., C_k\}$, où chaque Ci est un cluster contenant un sous-ensemble de points.

L'algorithme DBSCAN est non paramétrique, ce qui signifie qu'il peut détecter automatiquement le nombre de clusters à partir des données, en fonction des critères de densité et des paramètres ε et minPts.

Cette formulation mathématique de DBSCAN permet de mieux comprendre son fonctionnement en identifiant les points centraux, en définissant les règles d'attribution des clusters et en représentant les clusters détectés.

Spectral Clustering

Le clustering spectral (SC) est une méthode de clustering qui utilise la théorie des graphes et l'analyse spectrale pour identifier des structures de regroupement dans un ensemble de données. Mathématiquement, l'algorithme commence par construire une matrice de similarité S qui capture les relations entre les points de données. Cette matrice peut être calculée à l'aide de mesures de similarité, telles que la similarité cosinus ou la similarité basée sur un noyau.

Ensuite, une décomposition spectrale est appliquée à la matrice de similarité S pour obtenir les vecteurs propres dominants. Cette décomposition spectrale peut être réalisée en utilisant des techniques telles que la décomposition en valeurs singulières (SVD) ou la décomposition de Laplacien. Les vecteurs propres dominants correspondent aux nouvelles représentations des points de données dans un espace de dimension réduite.

Les vecteurs propres dominants sont ensuite utilisés pour attribuer les points aux clusters correspondants en utilisant des méthodes de partitionnement, telles que la k-means. Cette étape consiste à regrouper les points de données similaires en clusters distincts en fonction de leurs coordonnées dans l'espace transformé.

Le SC est apprécié pour sa capacité à détecter des structures complexes et non linéaires dans les données. Il est particulièrement efficace pour le regroupement de données où les frontières de décision ne sont pas linéaires dans l'espace d'origine.

Un exemple d'application du SC est la détection de communautés dans un réseau social. En utilisant les relations d'amitié ou d'interaction entre les utilisateurs, la matrice de similarité peut être construite pour représenter la proximité entre les individus. Le SC peut alors être appliqué pour regrouper les utilisateurs qui interagissent fréquemment et partagent des intérêts communs, en se basant sur leurs connexions mutuelles dans le réseau social.

En somme, le SC est une méthode de clustering basée sur la théorie des graphes et l'analyse spectrale. Il utilise une matrice de similarité et la décomposition spectrale pour identifier des structures de regroupement dans les données. Sa capacité à détecter des structures complexes en fait une approche précieuse pour l'analyse de clusters dans divers domaines d'application.

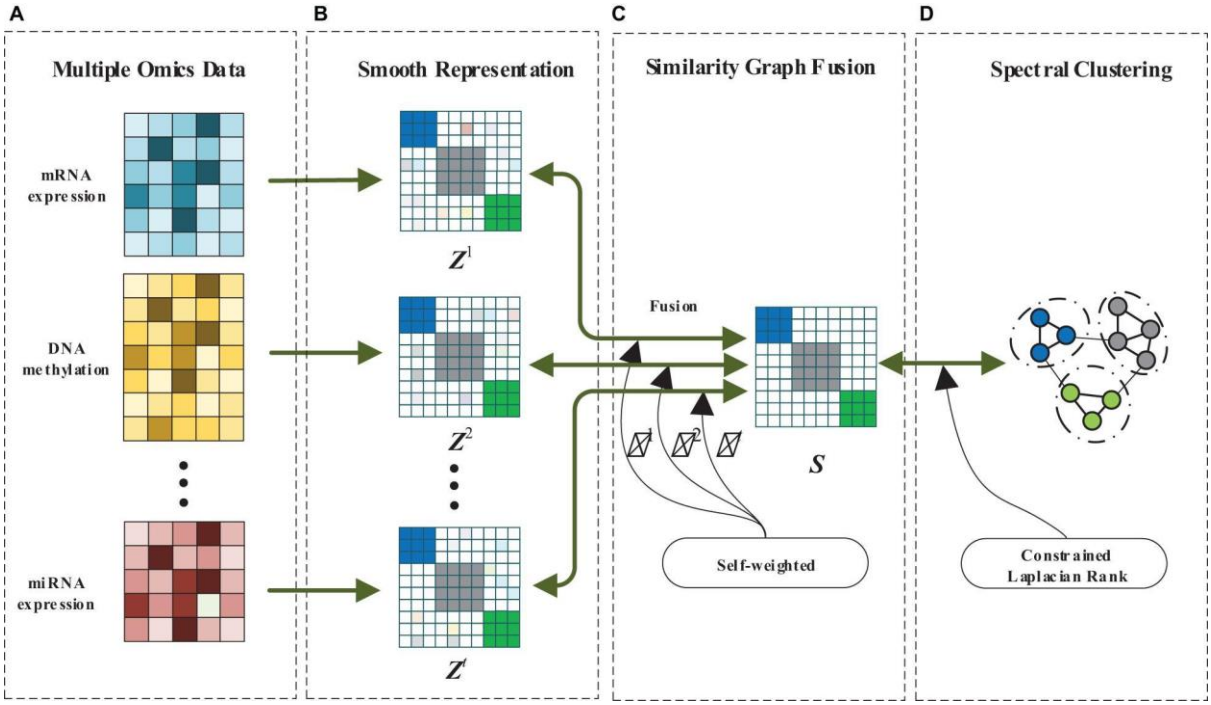


Figure 1.17 : Illustration d'un clustering spectral.

Conclusion

Dans ce chapitre, nous avons exploré différents aspects de l'apprentissage automatique. Nous avons commencé par discuter du projet d'apprentissage automatique de bout en bout. Ensuite, nous avons examiné les types d'apprentissage automatique, en nous concentrant sur les types d'apprentissage, surtout l'apprentissage supervisé et l'apprentissage non supervisé. De plus, nous avons abordé l'apprentissage profond et l'apprentissage génératif et des approches de synthétisation des données.

Ce chapitre nous a permis de mieux comprendre les différents domaines de l'apprentissage automatique, ce qui nous permettra de mieux cerner notre problématique et de choisir les algorithmes les plus appropriés pour la résoudre. Nous avons réalisé l'importance de comprendre les principes fondamentaux de chaque type d'apprentissage et de sélectionner les méthodes appropriées en fonction des caractéristiques de nos données et de nos objectifs spécifiques.

Chapitre 2

État des lieux

Partie 1 : SLB Ltd. et son secteur d'activité

La première partie du deuxième chapitre aspire à offrir une perspective panoramique sur l'industrie des hydrocarbures et le marché international des champs pétrolifères. L'augmentation incessante de la demande énergétique couplée à la finitude des réserves de pétrole et de gaz naturel font de la prospection et de la production une priorité majeure pour un grand nombre d'entreprises opérant dans ce secteur. Dans ce paysage, la compréhension approfondie du marché des services pétroliers et de sa contribution significative à la réussite de l'industrie est primordiale.

Cette partie consacre une attention spécifique au marché mondial des services pétroliers, mettant en lumière les tendances contemporaines et les défis persistants. De surcroît, une étude analytique détaillée du marché des services pétroliers en Algérie sera exposée, mettant en relief les opportunités uniques et les entraves spécifiques à cette nation.

Par ailleurs, une présentation de l'entreprise d'accueil **SLB Ltd.**, sera réalisée. Cela englobe une illustration de ses diverses opérations, de son organigramme et de sa structure organisationnelle. Une emphase particulière sera placée sur **SLB NAF** et **SLB** Algérie, soulignant leur rôle crucial et leur contribution substantielle dans le secteur pétrolier.

En somme, cette partie met en place le contexte indispensable à la compréhension de l'industrie des hydrocarbures, en mettant un accent particulier sur le marché international des champs pétroliers, le marché des services pétroliers en Algérie, et le rôle pivot de **SLB Ltd.**

2.1.1 L'industrie des hydrocarbures

Le secteur des hydrocarbures est une industrie qui englobe l'exploration, l'extraction, la raffinerie, le transport, la distribution et la vente de produits d'hydrocarbures, y compris le pétrole brut, le gaz naturel et les produits dérivés tels que l'essence, le diesel, le kérosène, et une variété de produits chimiques et de plastiques. Ce secteur a généré un revenu mondial de 5 billions de dollars au cours des années 2017-2022 (IBIS WORLD, 2022) et 2,8 milliards de dollars par jour au cours des 50 dernières années (The Guardian, 2022). Donc, en termes de valeur en dollars, c'est le plus grand secteur au monde.

Les restes d'animaux et de plantes en décomposition se déposent et s'accumulent au fil du temps dans le calcaire et le grès profondément dans les océans (Sarah El Shatby, 2023). Avec la bonne pression et température, ils forment ce que l'on appelle des "hydrocarbures", qui sont la matière première pour la fabrication du pétrole et du gaz. Les hydrocarbures sont simplement des composés organiques qui sont essentiellement constitués d'atomes de carbone et d'hydrogène. Même si le processus semble simple, il est très compliqué, coûteux et prend beaucoup de temps. Pour simplifier les choses, l'industrie du pétrole et du gaz est divisée en 3 étapes :

Exploration & Production (Amont)

Comme le titre le suggère, les entreprises énergétiques explorent des lieux autour du monde à la recherche de matières premières. Cette étape nécessite beaucoup de temps et de ressources car la recherche de réserves et de puits est difficile et parfois une entreprise investit d'énormes sommes d'argent et utilise des machines coûteuses et exigeantes en main-d'œuvre et peut toujours ne pas trouver ce qu'elle recherche. Les organisations abordent ce problème en passant des contrats avec des entrepreneurs de forage au lieu d'utiliser leur propre équipement.

Transport (intermediaire)

Dans la deuxième étape, l'entreprise transporte les matériaux extraits vers les raffineries et les usines pour commencer leur traitement.

Conversion & Vente (Aval)

Enfin, les raffineries éliminent les impuretés et transforment les matériaux extraits en produits à base de pétrole et dérivés et les libèrent sur le marché.

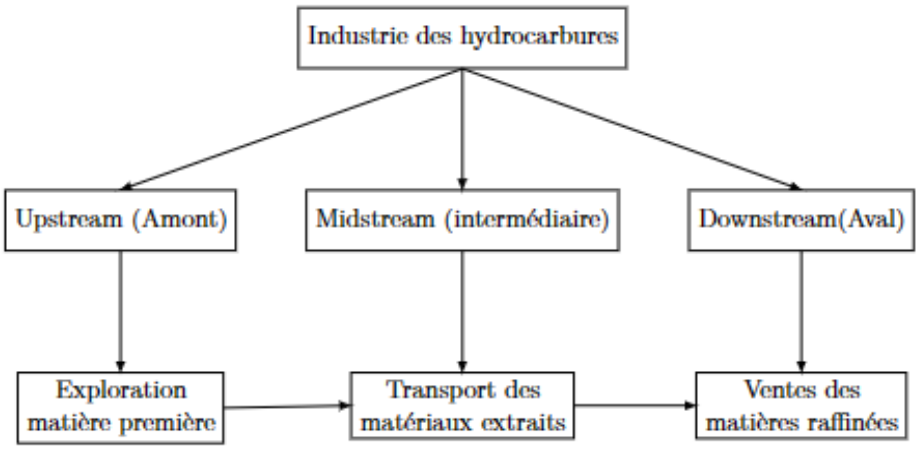


Figure 2.1 : L'industrie des hydrocarbures.

Ces sous-secteurs travaillent ensemble pour permettre l'extraction et la production efficaces de pétrole et de gaz, qui sont ensuite raffinés et distribués pour une variété d'utilisations dans l'économie mondiale.

2.1.1.1 Le marché mondial des champs pétrolifères

Le marché mondial des champs pétrolifères, également connu sous le nom de marché de l'exploration et de la production (E&P), est un secteur crucial de l'industrie énergétique. Il englobe les activités d'identification des gisements de pétrole et de gaz, leur forage et leur exploitation pour la production d'hydrocarbures.

La taille et l'évolution de ce marché sont fortement influencées par plusieurs facteurs, notamment les prix du pétrole et du gaz, les politiques gouvernementales, les progrès technologiques, et la découverte de nouveaux gisements. Par exemple, les avancées dans les technologies de forage, comme le forage horizontal et la fracturation hydraulique, ont permis d'accéder à des réserves de pétrole et de gaz qui étaient auparavant inaccessibles.

En termes de pays, des acteurs majeurs sur ce marché comprennent les États-Unis, la Russie, l'Arabie Saoudite, l'Iran, et le Canada, qui sont parmi les plus grands producteurs de pétrole au monde.

L'Organisation des pays exportateurs de pétrole, plus connue sous son acronyme OPEP, joue un rôle significatif dans le marché mondial des champs pétrolifères. L'OPEP est une organisation intergouvernementale qui compte 13 membres, principalement situés au Moyen-Orient, en Afrique et en Amérique du Sud. L'organisation a été créée dans le but de coordonner et d'unifier les politiques pétrolières de ses pays membres, afin de garantir des prix stables et un approvisionnement régulier de pétrole pour les consommateurs, tout en assurant des revenus stables pour les producteurs. (OPEP, brief history)

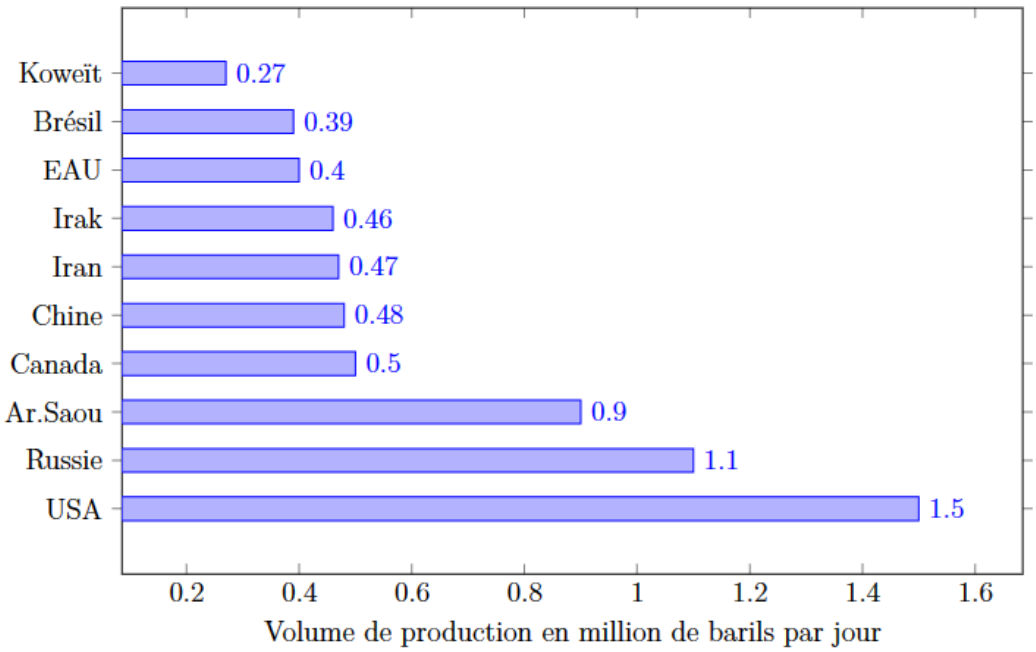


Figure 2.2 : Les plus gros producteurs mondiaux de pétrole en 2023 (Statista, 2023).

Du côté des entreprises, le marché des champs pétrolifères est dominé par de grandes multinationales, souvent appelées "supermajors", qui comprennent BP, Shell, ExxonMobil, Chevron, et TotalEnergies. Ces entreprises ont des activités dans tous les segments de l'industrie pétrolière, de l'exploration à la distribution.

La plupart des grands pays producteurs de pétrole possèdent aujourd'hui leur propre compagnie pétrolière et gazière qui gère la production et défend les intérêts nationaux dans le domaine des hydrocarbures. Ces compagnies nationales sont contrôlées à plus de 50 % par l'État, comme on peut le retrouver dans les pays membres de l'OPEP. On peut citer, par exemple, le groupe Sonatrach[11] en Algérie et Aramco en Arabie Saoudite.

Il convient de noter que, bien que le marché des champs pétrolifères reste une composante essentielle de l'économie mondiale, il fait face à des défis croissants en raison des préoccupations liées au changement climatique et de la transition vers des sources d'énergie plus propres.

2.1.1.2 Présentation du marché des services pétroliers

Marché mondial des services pétroliers

Le marché mondial des services parapétroliers, aussi appelé le marché des services pétroliers et gaziers, est une industrie vaste et complexe qui fournit des services essentiels à l'industrie pétrolière et gazière. Ces services comprennent l'exploration, le forage, la production, l'entretien, la logistique, la gestion des déchets, et bien d'autres services qui permettent l'extraction, le raffinage et la distribution de pétrole et de gaz.

La demande mondiale de pétrole devrait atteindre un record de 101,9 millions de barils par jour en 2023 (IEA[12] | rapport d'avril 2023 sur le marché des hydrocarbures). Cela représente une croissance de la demande de 2 millions de barils par jour par rapport à l'année précédente. Cette croissance de la demande de pétrole suggère une croissance correspondante du marché des services parapétroliers, car ces services sont essentiels pour répondre à la demande de pétrole.

Le marché des services parapétroliers est dominé par un certain nombre de grandes entreprises multinationales, notamment SLB, Halliburton, Baker Hughes (une entreprise de GE), et Weatherford. Ces entreprises fournissent une gamme de services allant du forage à l'équipement, en passant par les services logistiques et environnementaux.

L'évolution technologique joue un rôle de plus en plus important dans la transformation de la manière dont les services pétroliers sont fournis, avec des progrès notables dans des domaines tels que l'intelligence artificielle, l'automatisation, le forage avancé et l'Internet des objets. Les réglementations environnementales, de plus en plus strictes dans de nombreux pays, encouragent les entreprises de services pétroliers à adopter des pratiques plus durables et à réduire leur empreinte environnementale.

[11] Sonatrach est la compagnie nationale de pétrole et de gaz de l'Algérie. C'est la plus grande entreprise d'Afrique, impliquée dans toute la chaîne de production pétrolière et gazière.

[12] L'Agence Internationale de l'Énergie (IEA) est une organisation internationale fondée en 1974 par l'organisation de coopération et de développement économiques, visant à favoriser la sécurité énergétique de ses pays membres et favoriser la protection de l'environnement.

C'est un marché qui voit les prix fluctuer sous l'effet de ces diverses forces, notamment les conséquences géopolitiques comme la guerre en Ukraine[13] qui a perturbé l'offre et la demande de pétrole à l'échelle mondiale.

Il est important de noter que, malgré la transition mondiale vers des sources d'énergie plus propres, le pétrole devrait continuer à jouer un rôle clé dans le mix énergétique mondial dans un avenir prévisible. Par conséquent, le marché des services pétroliers reste un secteur d'activité important et pertinent.

L'industrie des services pétroliers peut être assimilée à un oligopole composé de trois leaders bien connus illustrées comme suit sur la Figure 2.3 :

Figure 2.3 : Acteurs du marché pétrolier dans le monde.

Marché des services pétroliers en Algérie

L'industrie parapétrolière en Algérie revêt une importance fondamentale pour l'économie nationale. En 2021, l'Algérie s'est classée quatrième producteur de pétrole en Afrique avec une production moyenne de 969 mille barils par jour (Radio Algérienne | Les Hydrocarbures en Algérie par Les Chiffres, 2021). Ce secteur englobe un large éventail d'activités qui accompagnent l'extraction et la production de pétrole et de gaz. Cela comprend la mise en place et l'entretien des infrastructures, le transport, le raffinage, la distribution et autres services associés. Plusieurs entreprises parapétrolières opèrent dans ce domaine, offrant une palette complète de services destinés à soutenir et optimiser ces activités.

Le gouvernement algérien a mis en place des politiques pour encourager le développement du secteur parapétrolier, notamment des incitations fiscales pour les entreprises et des initiatives pour le développement des compétences. Cependant, le secteur parapétrolier en Algérie fait face à des défis, notamment la volatilité des prix du pétrole sur les marchés internationaux et la nécessité d'investir dans de nouvelles technologies pour améliorer l'efficacité et minimiser l'impact environnemental.

Les principales compagnies parapétrolières du marché algerien, sont illustrées sur la Figure 2.4 suivante :

[13] La guerre en Ukraine, qui a commencé en 2014, est un conflit armé complexe et en cours entre le gouvernement ukrainien et les séparatistes soutenus par la Russie dans l'est du pays, spécifiquement dans les régions de Donetsk et de Louhansk.

Figure 2.4 : Acteurs du marché pétrolier en Algérie (hesp.com).

Les hydrocarbures jouent un rôle central dans l'économie de l'Algérie, représentant environ 65 % des recettes publiques, 26 % du produit intérieur brut PIB et 98 % des recettes d'exportation (Radio Algérienne | Les Hydrocarbures en Algérie par Les Chiffres, 2021).

En 2019, l'Algérie figurait parmi les 16 plus grands producteurs de pétrole, les 10 plus grands producteurs de gaz naturel et les 7 plus grands exportateurs de gaz naturel au monde (Statista, 2022). Cependant, en 2020, en raison de la crise sanitaire liée à la pandémie de Covid-19, les exportations d'hydrocarbures ont considérablement diminué, menaçant ainsi la situation économique du pays (Pétrole : Alger a Le Blues - Le Point, 2021).

Pour les entreprises parapétrolières les plus importantes, on retrouve Expro, Halliburton, ENSP filiale de SONATRACH, Weatherford, NPS et le leader mondial SLB.

2.1.2 Présentation de SLB. Ltd

Mon travail s'inscrit dans la division D&I de SLB, il est donc important de présenter cette entreprise ainsi que son organisation et son histoire au niveau globale, par la suite présenter SLB Algérie en détail.

2.1.2.1 SLB. Ltd

SLB est une multinationale franco-américaine fondée en 1926 par les frères Conrad et Marcel SLB (Our History | SLB, 2023). Avec son siège principal réparti entre Paris, Houston, Londres et La Haye, l'entreprise est reconnue comme l'un des leaders mondiaux dans le secteur des services pétroliers et gaziers. Elle opère dans plus de 120 pays et emploie environ 100 000 professionnels issus de plus de 160 nationalités différentes.

SLB propose une gamme diversifiée de services et de technologies destinés à l'industrie de l'exploration et de l'exploitation pétrolière et gazière. Elle se distingue par son engagement envers l'innovation et la recherche et le développement, ayant contribué à de nombreuses avancées technologiques dans le domaine (Who We Are | SLB, 2023).

Face à l'évolution des préoccupations environnementales et à l'impératif d'une transition vers une économie à faible émission de carbone, SLB s'engage également dans le développement de solutions plus durables. Cette orientation démontre la

volonté de l'entreprise de répondre aux défis actuels tout en continuant à fournir des services de qualité à ses clients.

Tableau 2.1 : Carte d'identité de SLB.

Date de création	1926
Fondateurs	Conrad & Marcel SLB
Forme juridique	Société anonyme avec appel public à l'épargne
Cotée à la bourse	New York Stock Exchange et Euronext
Siége social	Huston, Texas (USA)
Direction	CEO : Oliviers Le Peuch | EVP & CFO : Simon Ayat
Secteur d'activité	Prestation de services pétroliers
Effectif	98 000 employés en 2022
Capitalisation	74,5 milliards USD (2022)
Chiffre d'affaire	28,1 milliards USD (2022)

2.1.2.2 Les activités de SLB Ltd.

SLB comprend quatre principaux groupes d'activités qui couvrent toute la durée de vie d'un réservoir. Chaque groupe comporte des segments ou des Product Line (PL) qui, a leur tour, se composent de sous-segments, et ces premiers sont :

• **Reservoir Characterization Group** (Groupe de caractérisation du reservoir) : Premier intervenant, il définit les caractéristiques des gisements pétroliers lors de la découverte et la prospection des sites potentiellement favorables.

• **Reservoir Drilling Group** (Groupe de forage de réservoirs) : Il offre les principales technologies de mise en œuvre du forage des puits de pétrole ou de gaz, comme les Geoservices, PathFinder, outils de forage et reconditionnement. . . Etc.

• **Reservoir Production Group** (Groupe de production de réservoirs) : Il intervient après le forage, et offre les technologies nécessaires à la production des réservoirs tout au long de leur cycle de vie.

• **Cameron :** Spécialisé dans la fabrication d'équipements de contrôle de pression dans le secteur.

2.1.2.3 Structure hiérarchique de SLB

En 2020, SLB a revu la structure de son Top Management. Elle se constitue d'un CEO qui coordonne les activités de la Leadership Team et des Corporate Functions.

La Leadership Team est constituée des managers de cinq entités qui sont : Performance Management, Technology, Services & Equipment, Geographies et New Energy.

Les Corporate Functions quant à elles représentent les six fonctions support suivantes : Finance, Legal, HR, HSE, Strategy and Sustainability et Sales & Commercial. L'ensemble de la structure est représenté sur la Figure 2.5.

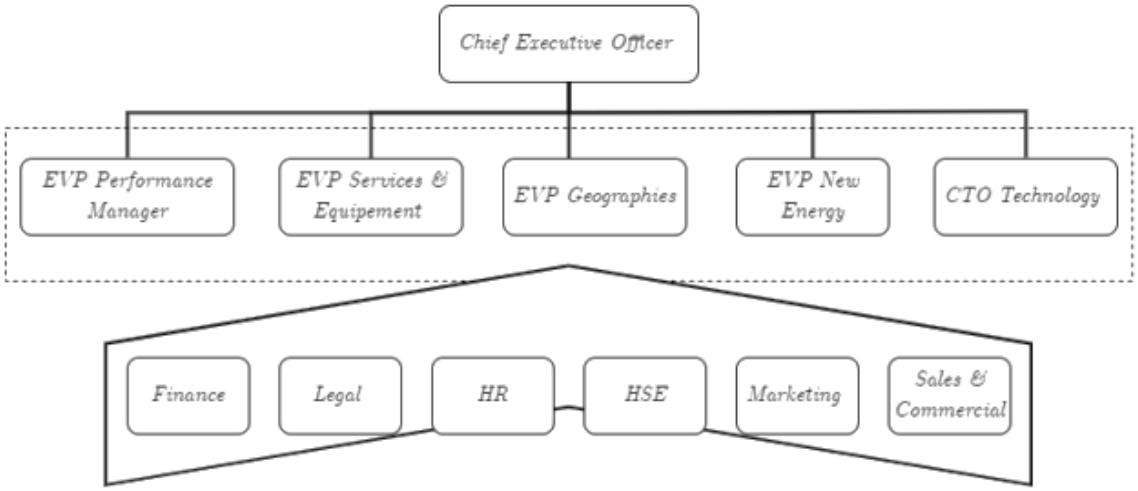


Figure 2.5 : Structure hiérarchique de SLB.

2.1.2.4 Valeurs de SLB

SLB s'engage à fournir des technologies et des services qui améliorent et optimisent les performances de ses clients tout en tirant le meilleur parti de ses atouts uniques (Our values | SLB, 2023). À cette fin, elle compte sur trois valeurs établies de longue date pour guider ses décisions dans la poursuite de ses ambitions :

- Le profit : L'entreprise est déterminée à produire des profits supérieurs pour garantir sa croissance.

- La ressource humaine : L'entreprise considère qu'un potentiel humain épanoui, ambitieux d'exceller dans n'importe quel environnement et dévoué à la sécurité et au service de la clientèle dans le monde entier est sa plus grande force et la clé de son succès ;

- La technologie : Depuis 1926, SLB n'a cessé de développer des technologies de pointe et est convaincue que le secret du succès de l'entreprise est de réinvestir les profits dans la recherche et le développement pour assurer une qualité inégalable et maintenir son avantage concurrentiel ;

Pour atteindre leurs objectifs communs qui sont la création de valeur, la satisfaction des besoins des clients et l'amélioration de l'image de l'entreprise, les employés doivent tenir compte de quatre pratiques essentielles qui sont : l'engagement, l'intégrité, le travail d'équipe et l'entraînement.

SLB a élaboré un code de conduite appelé « The blue print », représenté sur la Figure 2.6, qui dicte toute la culture de l'entreprise et caractérise l'attitude que chaque employé doit adopter pour protéger l'identité de son entreprise et sa position de leader.

Figure 2.6 : The Blueprint.

2.1.2.5 Divisions opérationnelles

Par rapport à ses activités, SLB s'est organisée en quatre divisions regroupant plusieurs Business Lines (BL). Ces divisions sont : Digital & Integration, Reservoir Performance, Production Systems et Well Construction. Les divisions ont amélioré les portefeuilles de fonctionnalités alignées sur les flux de travail des clients. Chaque division offre des possibilités de croissance grâce à la transition des clients vers l'efficacité du capital, l'amélioration de la production et de la récupération et la réduction de l'empreinte carbone (Le Peuch, 2020).

Digital & Integration

La division Digital & Integration (D&I) comprend les technologies numériques et l'intégration des données, la technologie et les processus pour améliorer les performances des actifs et de l'entreprise. Elle a un potentiel de croissance élevé grâce à la transformation numérique en cours. Cela prend en charge l'adoption rapide du cloud computing et la croissance de l'informatique de pointe et de l'automatisation dans le secteur de l'énergie. Elle regroupe les Business Lines suivantes :

Digital Subsurface Solutions : Geoscience and reservoir engineering ;
Exploration Data : Multiclient seismic and associated processing ;
Digital Operations Solutions : Drilling and production automation ;
Integrated Well Construction : Integrated well construction project management ;
Integrated Reservoir Performance : Production, recovery, and asset performance management.

Production Systems

La division Production Systems (PS) stimule l'innovation technologique et l'intégration totale du système, de l'interface réservoir-puits à mi-chemin. En prévision des besoins de l'industrie, des progrès technologiques significatifs dans les achèvements, l'ascenseur artificiel, l'équipement de surface, le traitement et le sous-marin ont été réalisés.

Les Business Lines suivantes font partie de cette division :

- **Well Production Systems :** Completions and downhole artificial lift systems
- **Surface Production Systems** : Wellheads frac services and surface production pumps
- **Subsea Production Systems :** Subsea production and processing systems*
- **Midstream Production Systems :** Valves, process systems, production chemistry and facilities.

Well Construction

La division Well Construction (WC) combine la gamme complète de produits et de services pour maximiser l'efficacité du forage et le contact avec le réservoir. Alors que les clients s'efforcent d'améliorer les rendements des actifs, cette division bénéficiera d'une échelle, d'une exposition au marché et d'une approche holistique de la construction de puits.

Elle abrite les Business Lines ci-dessous :

- **Well Construction Measurement :** Drilling data acquisition
- **Well Construction Drilling :** Directional drilling and bits
- **Well Construction Fluids :** Drilling fluids and well cementing
- **Well Construction Equipment :** Drilling rigs and equipment, pressure control equipment.

Reservoir Performance

La division Reservoir Performance (RP) comprend des technologies et des services centrés sur les réservoirs qui sont essentiels à l'optimisation de la productivité et de la performance des réservoirs. Elle capitalise sur la croissance des initiatives d'exploration de proximité, de réaménagement des friches industrielles et d'amélioration de la récupération dans les puits étroits ou matures.

Elle est constituée des Business Lines suivantes :

- **Reservoire Performance Evaluation :** Wireline, downhole testing services and fluids & rock sampling & analysis ;
- **Reservoire Performance Intervention :** Coiled tubing, surface testing, slickline,perforating and wireline intervention ;
- **Reservoire Performance Simulation :** Sand management and simulation

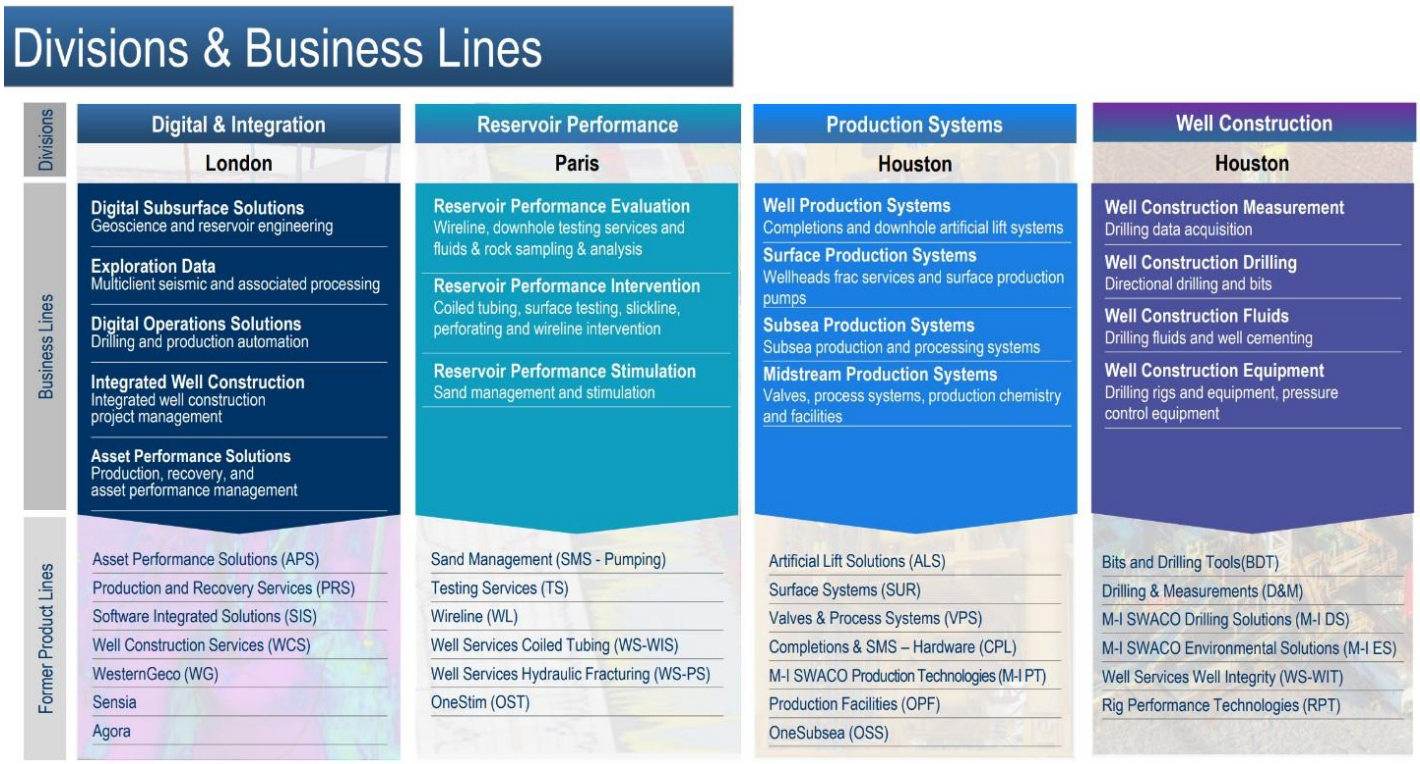


Figure 2.7 : Schématisation des divisions et Business Lines SLB.

2.1.2.6 Organisation de SLB

Aujourd'hui, SLB opère sur 120 pays et afin de maintenir ses performances, elle est divisée en cinq bassins géographiques :

- les Amériques
- l'Asie
- la Russie et l'Asie centrale
- le Nord d'Afrique et le Moyen-Orient
- L'Offshore de l'Atlantique

Ces bassins géographiques partagent des besoins technologiques similaires, chaque bassin géographique est divisé en GeoUnit, il existe au total 30 GeoUnits. Une GeoUnits est un pays ou un groupe de pays gérés dans l'un des cinq bassins (Le Peuch, 2020). La carte ci-dessous (Figure 2.8) représente la répartition des bassins ainsi que les GeoUnits de SLB :

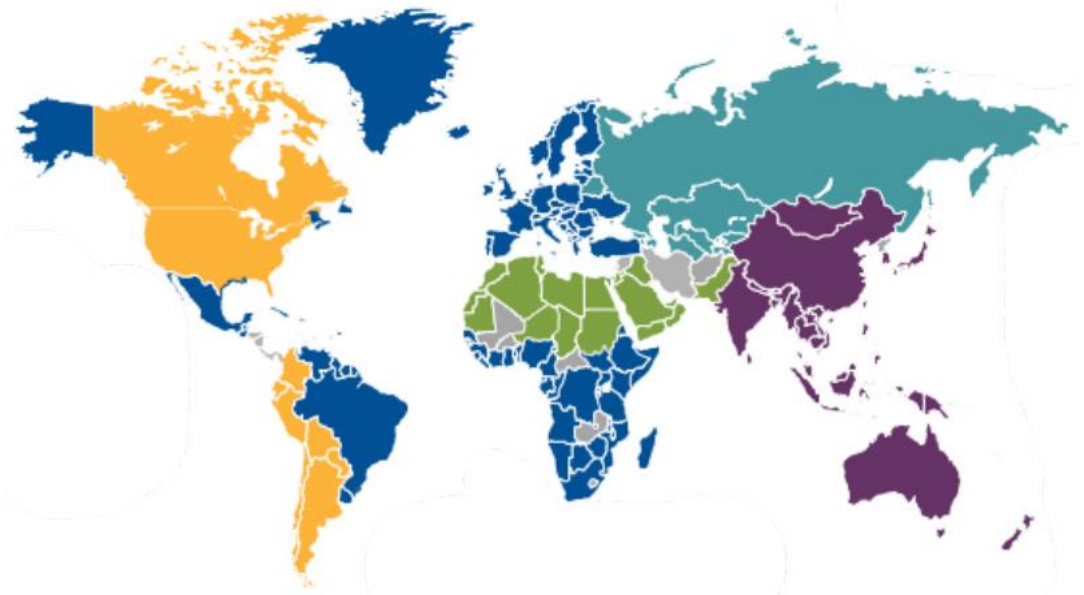


Figure 2.8 : Carte des bassins et des GeoUnits de SLB.

Tableau 2.2 : Répartition des GeoUnits dans les 5 bassins de SLB.

Bassins	Americas Land	Offshore Atlantic	Russia and Central Asia	Asia	Middle East & North Africa
Capitale	Houston	London	Moscow	Kuala Lumpur	Dubai
GeoUnits	Canada Land	Angola Central & East Africa	Russia Land	East Asia	North Africa
	US Land	Brazil	Arctic & South Offshore	India	Egypt, Sudan & East Mediterranean
	Ecuador, Colombia & Peru	Europe	Azerbaijan & Turkmenistan	Australia, New Zealand & Papua New Guinea	Iraq
	Argentina, Bolivia & Chile	Guyana, Trinidad & Caribbean	Kazakhstan	Indonesia	Kuwait
		Mexico & Central America	Sakhalin	China	Saudi & Bahrain
		North America Offshore			Qatar
		Nigeria & West Africa			Emirates
		Scandinavia			Oman, Yemen & Pakistan

2.1.3 SLB NAF

La NAF GeoUnit regroupe les pays de l'Afrique du Nord dans lesquels SLB est présente. Ces pays sont l'Algérie, la Tunisie, le Maroc, la Libye et le Tchad, comme on peut le voir sur la Figure 2.9.

Figure 2.9 : La carte de la GeoUnit NAF.

Sa structure hiérarchique est constituée du manager de la GeoUnit qui coordonne les activités des managers des pays les plus importants, entre autres l'Algérie, la Libye et le Tchad, ainsi que des managers des fonctions support : Sales & Marketing, Supply Chain (SC), Human ressources (HR), Health Safety and Environment (HSE), Finance et Legal (se référer à la Figure 2.10).

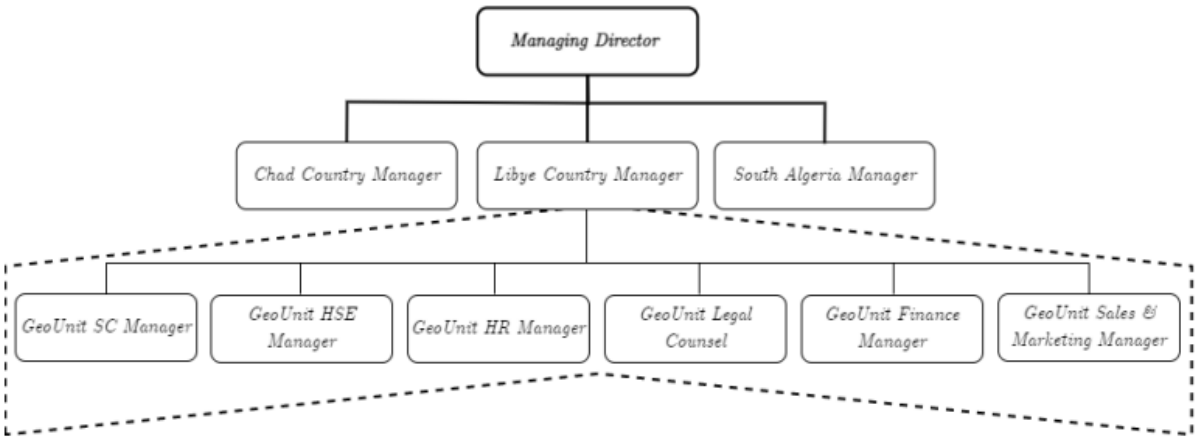


Figure 2.10 : Structure hiérarchique de la GeoUnit NAF.

2.1.3.1 SLB Algerie

En Algérie, l'entreprise est installée en 1955 et possède un siège au niveau de la zone d'activités de Chéraga, route d'Ouled-Fayet à Alger qui représente aussi le siège social du "North Africa GeoMarket", elle possède aussi un ensemble de 11 bases opérationnelles au niveau des 4 zones d'activités : Hassi Messaoud, Ain Amenas, Hassi Berkine et Ain Salah.

En plus de ces bases opérationnelles, l'entreprise possède aussi des bases logistiques, des bunkers d'explosifs et des Guest House (maison d'hôtes).

Elle est présente sur le pays à travers deux entités légales : COPS (Compagnie des Opérations Pétrolières SLB) et SPS (Services Pétroliers SLB) et elle fournit des services pétroliers principalement à l'entreprise nationale SONATRACH ainsi qu'aux entreprises : Total, Anadarko, British Petroleum, AGIP et autres. Parmi ses services en Algérie :

— L'installation des bases opérationnelles.

— Les études géologiques et sismiques.

— La construction des puits.

— Le test des puits.

— L'importation des équipements requis.

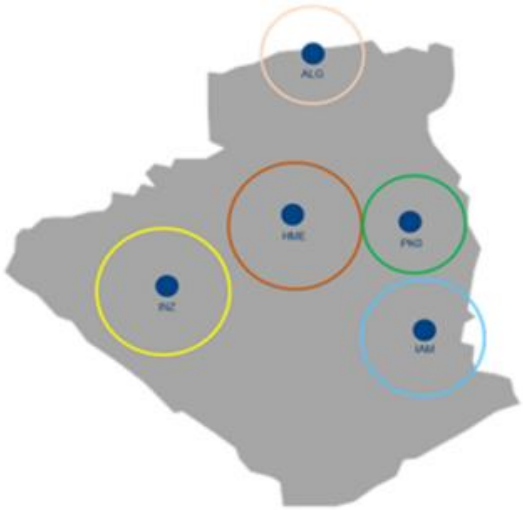


Figure 2.11 : Présence SLB en Algérie.

2.1.4 Divisions Digital & Integration

La division Digital & Integration (D&I) joue un rôle crucial dans la transformation numérique de l'industrie pétrolière et gazière. Avec son expertise dans les technologies numériques et l'intégration des données, elle vise à améliorer les performances des actifs et de l'entreprise grâce à l'utilisation efficace de la technologie et des processus.

2.1.4.1 Les objectifs de la division D&I

- **Intégration des données :** La division D&I travaille à la consolidation et à l'intégration des données provenant de différentes sources, telles que la géoscience, l'ingénierie des réservoirs, les données d'exploration sismique, l'automatisation des opérations de forage et de production, et la gestion de la performance des actifs. L'objectif est de permettre une meilleure compréhension des réservoirs, une prise de décision plus éclairée et une optimisation des performances globales.

- **Amélioration des processus :** Grâce à l'utilisation des technologies numériques, la division D&I vise à automatiser et à optimiser les processus liés à la construction de puits, à la gestion de projets intégrés, à la performance des réservoirs et à l'exploitation des actifs. Cela permet d'accroître l'efficacité opérationnelle, de réduire les coûts et d'améliorer la rentabilité des activités pétrolières et gazières.

- **Innovation et recherche :** La division D&I s'engage dans la recherche et le développement de nouvelles technologies et solutions numériques pour répondre aux défis et aux besoins de l'industrie. Cela inclut l'utilisation de l'informatique de pointe, de l'automatisation, de l'intelligence artificielle et de l'apprentissage automatique pour améliorer les capacités de prévision, d'optimisation et de prise de décisions

En résumé, la division Digital & Integration de SLB se concentre sur l'intégration des données, l'optimisation des processus et l'innovation technologique pour améliorer les performances des actifs et de l'entreprise dans le contexte de la transformation numérique de l'industrie pétrolière et gazière. Ses objectifs sont d'améliorer la compréhension des réservoirs, d'optimiser les opérations et de stimuler l'innovation pour une meilleure rentabilité et compétitivité sur le marché.

2.1.4.2 Types de Données dans l'Industrie du Pétrole et du Gaz

Comme nous l'avons mentionné précédemment, la science des données est devenue une partie intégrale du succès de presque toutes les industries et cela s'applique particulièrement au pétrole et au gaz. Dans ce domaine, la demande pour l'analyse des grands volumes de données (big data) augmente pour améliorer les techniques et les processus impliqués dans l'exploration et la production de pétrole.

Avec l'avancement des méthodes technologiques en exploration et production, d'énormes quantités de données sont générées chaque jour (Mehdi Mohammadpoor et Farshid Torabi, 2020). Ainsi, le besoin pour l'analyse des grands volumes de données dans l'industrie du pétrole et du gaz a énormément augmenté. Les entreprises qui opèrent dans le domaine collectent des données provenant de deux types principaux de sources - structurées et non structurées.

Sources de données structurées :

- Rapports de gestion des risques et de projet
- Installations de surface et de subsurface
- Données de forage
- Données de production
- Prix du marché
- Données météorologiques

Sources de données non structurées :

- Journaux de puits
- Rapports quotidiens écrits de forage
- Dessins CAD

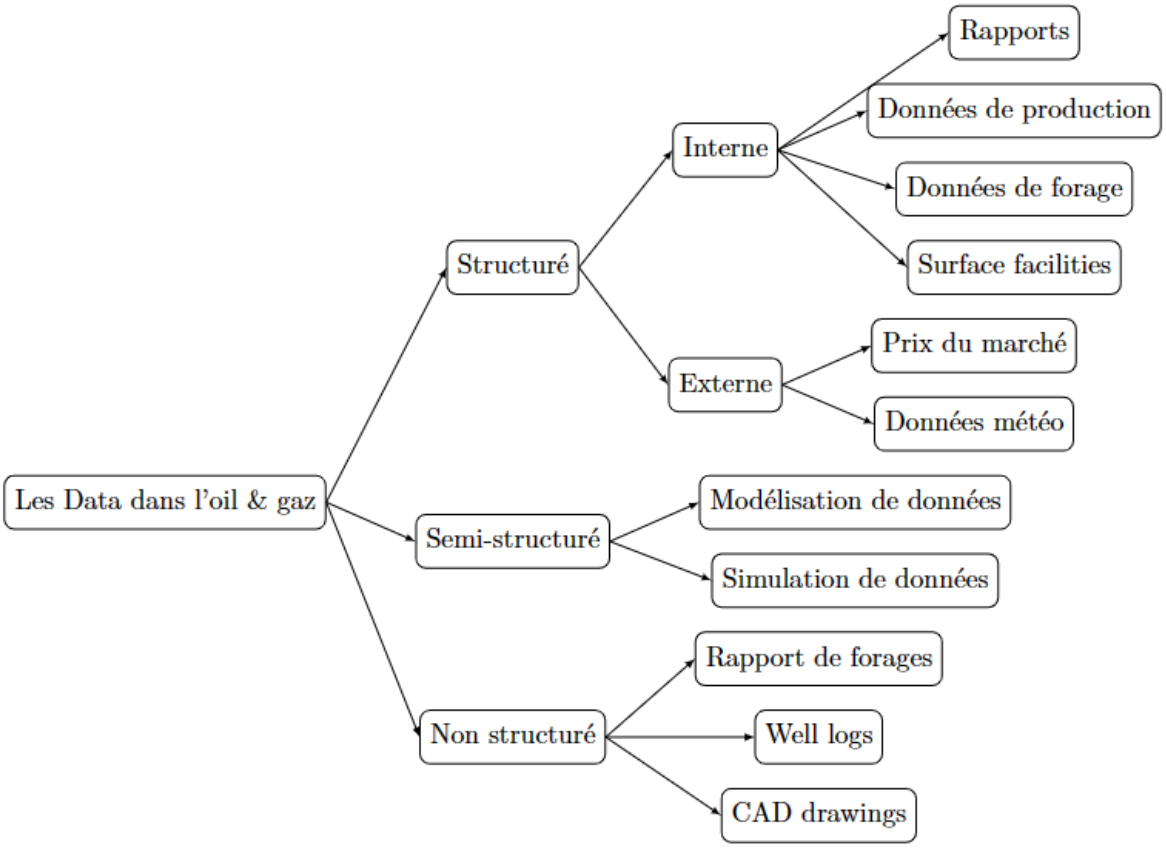


Figure 2.12 : Typologie des données dans le secteur pétro-gazier.

Partie 2 : Principes fondamentaux de la production de puits de gaz

La production mondiale de gaz naturel, un facteur crucial de la matrice énergétique mondiale, a connu une croissance considérable au cours des dernières années. Cette croissance est largement attribuable à l'exploitation de nouvelles réserves et à l'évolution des technologies d'extraction. Le gaz naturel, considéré comme une ressource énergétique plus propre par rapport aux combustibles fossiles traditionnels, a joué un rôle essentiel dans la transition mondiale vers une production d'énergie plus durable et respectueuse de l'environnement.

Cependant, l'industrie du gaz naturel n'est pas sans défis. En particulier, le chargement de liquide est un phénomène technique complexe qui peut limiter sérieusement la production de gaz (Rao, 1999). Le problème se réfère à l'accumulation de liquides tels que l'eau et les condensats dans un puits de gaz, ce qui peut entraver la production de gaz en créant une contre-pression supplémentaire.

Dans cette section, nous découvrons plus profondément ce phénomène et les mécanismes de base du chargement de liquide, l'impact du chargement de liquide sur la production de gaz et les techniques actuellement disponibles pour atténuer ce problème.

Enfin, nous explorerons également les recherches existantes qui ont développé des solutions basées sur des corrélations empiriques pour prédire et gérer efficacement le chargement de liquide. Ainsi, nous pourrons mieux comprendre notre problématique et les défis liés à la mise en œuvre de solutions basées sur l'apprentissage machine pour le développement de l'industrie du gaz.

2.2.1 Profil de Production

La production d'un champ de pétrole ou de gaz tend à passer par un certain nombre d'étapes. Cela peut être décrit par la courbe de production montrée pour le pétrole dans la Figure 2.13.

Après la découverte d'un champ de pétrole, le nouveau champ est évalué pour déterminer le potentiel de développement du réservoir. S'il répond aux volumes et aux taux de production requis pour la viabilité commerciale, un développement supplémentaire suit et la première production de pétrole marque le début de la phase de montée en puissance. La production augmente progressivement jusqu'à la phase de plateau, où la capacité d'extraction entièrement installée est utilisée, avant d'arriver finalement au début du déclin, car les conditions souterraines ne pourront plus soutenir ce taux d'extraction.

La phase de déclin se termine par l'abandon une fois que la limite économique est atteinte (Höök, 2009). Le moment de l'abandon est de préférence prolongé jusqu'à la dernière goutte de pétrole. Le profil du gaz est similaire, mais montre souvent une phase de plateau plus courte. Afin de prolonger l'abandon, des techniques de fin de vie du champ sont mises en œuvre. L'objectif de cette étude est de prédire le moment du chargement de liquide afin d'éviter un abandon à un stade précoce.

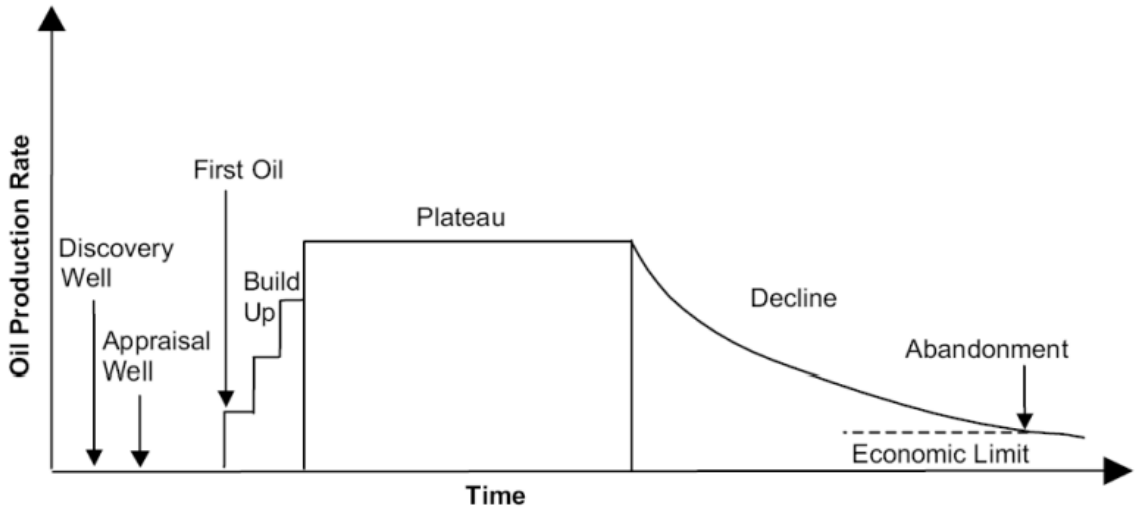


Figure 2.13 : Courbe de production théorique décrivant les différents stages de maturité.

2.2.2 Le chargement de liquide

Le chargement de liquide d'un puits de gaz est l'incapacité du gaz produit à éliminer les liquides produits du puits. Lorsqu'un puits de gaz est en production, la pression dans le réservoir de gaz est élevée et la vitesse du gaz dans le tubage est suffisante pour entraîner le liquide vers la surface. Cependant, après plusieurs années, vers la fin de la vie du champ, la pression dans le réservoir est devenue si basse que le gaz n'atteint pas la vitesse critique et une accumulation de liquide se produit en fond de puits. L'accumulation de liquide imposera une contre-pression supplémentaire sur la formation qui peut affecter significativement la capacité de production du puits (Turner et al., 1969). Le liquide peut provenir de l'eau interstitielle dans la matrice du réservoir ou il peut être formé en raison de la condensation de la vapeur d'eau et du gaz d'hydrocarbures à mesure que la pression et la température diminuent le long de la trajectoire du tubage du puits (Van Nimwegen, 2015). La production cessera totalement au point que le puits doit être fermé, même s'il reste encore du gaz naturel dans le réservoir.

2.2.2.1 Écoulement Multiphasique

Pour comprendre les effets des liquides dans le puits de gaz, il est important de comprendre comment les liquides et les gaz se comportent lorsqu'ils s'écoulent ensemble vers le haut dans le tubage de production du puits (Eissa.M, 2017). L'écoulement multiphasique dans un conduit vertical est généralement représenté par quatre régimes d'écoulement de base, comme illustré dans les Figures : 2.14 et 2.15 (Adriana Molinari, 2019). À un moment donné de l'histoire du puits, un ou plusieurs de ces régimes seront présents. Un régime d'écoulement est déterminé par la vitesse des phases de gaz et de liquide ainsi que les quantités relatives de gaz et de liquide à un point donné dans le flux d'écoulement.

- Écoulement Annulaire/Nébulisé : Il se produit à une vitesse élevée du gaz, où le gaz constitue la phase continue et le liquide est présent sous forme de gouttelettes dispersées (brouillard) dans le gaz, ainsi qu'un film mince (annulaire) le long de la paroi du tuyau.
- Écoulement de Transition : Lorsque la vitesse du gaz diminue, l'écoulement commence à passer de nébuliser à en bourrelet, ce qui entraîne un changement de phase continue du gaz au liquide, et le liquide peut encore être présent sous forme de brouillard dans le gaz. Au lieu de se déplacer vers le haut, le film liquide atteint un certain point où il commence à se déplacer vers le bas, et le chargement de liquide est lié à cette transition.
- Écoulement en Bourrelet : À mesure que le débit de gaz diminue encore davantage, le gaz apparaît sous forme de gros bourrelets dans le liquide, mais la phase continue est liquide.
- Écoulement en Bulles : À de très faibles débits de gaz, l'écoulement en bulles se produit, où le tubage est presque rempli de liquide (phase continue). Le gaz libre est présent sous forme de petites bulles qui s'élèvent dans le liquide.

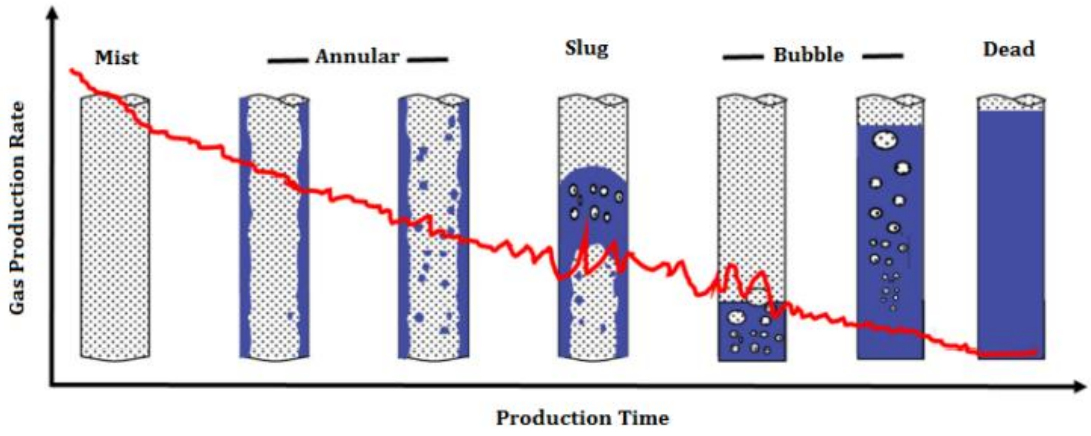


Figure 2.14 : Schématisation d'un pattern de flux typique.

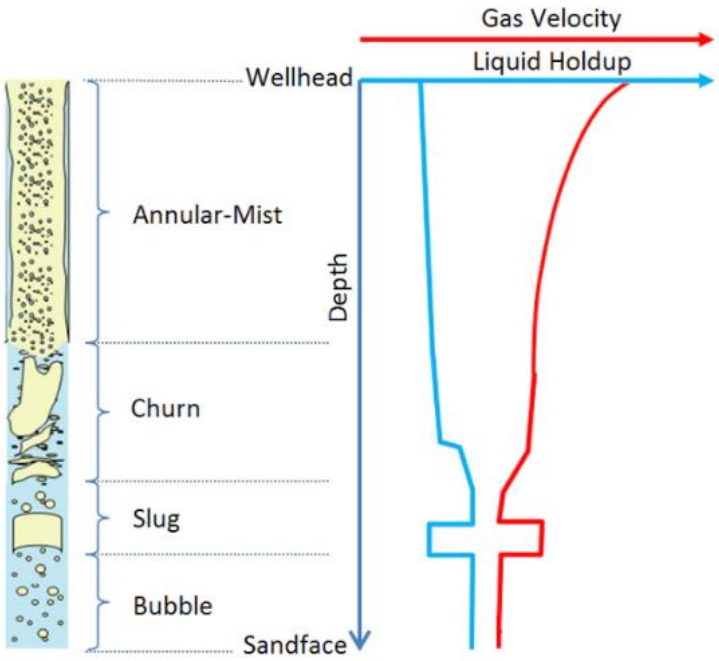


Figure 2.15 : Schématisation d'un pattern de flux typique (Adriana Molinari, 2019).

Un puits de gaz peut traverser un ou plusieurs de ces régimes d'écoulement tout au long de sa durée de vie. Initialement, il peut avoir une vitesse élevée du gaz, ce qui entraîne un écoulement en nébulisation dans le tubage, mais peut se trouver en écoulement en bulles, en transition ou en bourrelets en dessous de l'extrémité du tubage jusqu'aux perforations médianes. À mesure que le temps passe et que la production diminue, les régimes d'écoulement des perforations à la surface changeront à mesure que la vitesse du gaz diminuera. La production de liquide peut également augmenter à mesure que la production de gaz diminue.

2.2.2.2 Sources de Liquides dans les puits de Gaz en Production

De nombreux puits de gaz produisent non seulement du gaz, mais aussi des liquides. Ces liquides peuvent être de l'eau libre, de la condensation d'eau et/ou de la condensation d'hydrocarbures. (Nallaparaju, 2012)

- Si la pression du réservoir est inférieure au point de rosée, la condensation est produite avec le gaz sous forme liquide
- Si la pression du réservoir est supérieure au point de rosée, la condensation entre dans le tubage sous forme de vapeur avec le gaz et se transforme en liquide lorsque la pression diminue.

Les liquides produits, avec le gaz, peuvent avoir plusieurs sources en fonction des conditions et des types de réservoir à partir desquels le gaz est produit :

Effet de Coning d'Eau : Si le débit de gaz du puits est suffisamment élevé, cela peut entraîner une pression de déclin élevée, suffisante pour tirer la production d'eau d'une

zone sous-jacente, même si les perforations ne s'étendent pas jusqu'à la zone sous-jacente. Les puits horizontaux réduisent généralement les effets du coning d'eau.

Eau de Nappe Aquifère : L'aquifère qui soutient la pression du gaz produit finit par atteindre les perforations et pénètre dans le tubage, provoquant le chargement de liquide.

Formation d'Eau Libre : L'eau peut pénétrer dans le puits à travers les perforations avec le gaz produit. Cela peut résulter de couches minces de gaz et de liquide.

Production d'Eau à partir d'une Autre Zone : Il est possible de produire des liquides à partir d'une autre zone, soit avec une complétion à trou ouvert, soit dans un puits comportant plusieurs sections perforées.

Eau de Condensation : L'eau de condensation est présente lorsque le gaz naturel extrait du réservoir est saturé d'eau. Au fur et à mesure que la solution de gaz s'écoule à travers la colonne de production, l'eau se condense si les conditions de température et de pression chutent en dessous du point de rosée. Cela conduit à une accumulation progressive d'eau condensée au fond du puits.

Condensats d'Hydrocarbures : Tout comme l'eau, les hydrocarbures peuvent également pénétrer dans le puits avec le gaz produit sous forme de vapeur. Lorsque la solution de gaz s'écoule vers la surface, les hydrocarbures à l'état de vapeur peuvent commencer à se condenser lorsque les conditions chutent en dessous du point de rosée, entraînant finalement un chargement du puits, tout comme l'eau.

2.2.2.3 Vitesse critique

Dans cette section, la méthode pour prédire le début du chargement de liquide est présentée. Le critère de Turner (Turner et al., 1969) est le plus couramment utilisé pour prédire le chargement de liquide. Cette technique a été développée à partir d'une accumulation substantielle de données de puits et s'est avérée raisonnablement précise pour les puits verticaux. La méthode est applicable à n'importe quel point du puits et doit être utilisée en conjonction avec des méthodes d'analyse nodale si possible (Lea et al., 2008). Il existe deux possibilités lorsque le gaz s'écoule vers le haut à travers le tubage. Soit la vitesse du gaz est suffisante pour entraîner le liquide vers le haut, de sorte que la vitesse moyenne est dirigée vers le haut, soit la vitesse du gaz est insuffisante et le liquide se déplace en moyenne vers le bas, ce qui entraîne une accumulation de liquide au fond du puits. La vitesse à laquelle une gouttelette de liquide peut s'élever contre la gravité, et donc la vitesse minimale à laquelle le débit de gaz doit se produire avec des liquides, est appelée vitesse critique du gaz (Binli, 2009). Comme la relation de Turner a été développée à partir de données de pression de tubage principalement supérieures à 1000 psi, le débit critique de Turner peut être utilisé pour les pressions élevées, mais il n'est pas aussi précis pour les pressions de tête de tubage basses. Par conséquent, le critère de Coleman (Coleman et al., 1991) est introduit. Des relations similaires décrivent le débit critique minimum pour les pressions de tubage de surface plus basses, mais sans l'ajustement de Turner que Turner a utilisé pour ajuster ses données. Les puits en fin de vie de production sont examinés, donc Coleman est considéré comme le plus approprié. La vitesse de Turner est donnée dans l'Équation 2.1 et le critère de Coleman dans l'Équation 2.2

$$v = 1.92 \frac{\sigma^{\frac{1}{4}}(\rho_l - \rho_g)^{\frac{1}{4}}}{\rho_g^{\frac{1}{4}}} \qquad (2.1)$$

$$v = 1.59 \frac{\sigma^{\frac{1}{4}}(\rho_l - \rho_g)^{\frac{1}{4}}}{\rho_g^{\frac{1}{4}}} \qquad (2.2)$$

où v est la vitesse critique, σ est la tension de surface, et ρ_l et ρ_g sont respectivement la densité du liquide et du gaz. Pour l'eau, la tension de surface est de 60 dyne/cm et la densité est de 67 lbm/ft. La densité du gaz en lbm/ft est donnée par l'Équation 2.3.

$$\rho_g = \frac{M_{air}\gamma_g P}{R(T+460)Z} = 2.715\gamma_g \frac{P}{(460+T)Z} \qquad (2.3)$$

où M_{air} est la masse molaire du gaz, P est la pression à la tête du puits en bar, γ_g est la gravité spécifique du gaz, R est la constante du gaz, T est la température à la tête du puits et Z est la compressibilité du gaz. Bien que la vitesse critique soit le facteur déterminant, on pense généralement aux puits de gaz en termes de débit de production plutôt que de vitesse dans le tubage. À partir de la loi des gaz parfaits, le débit peut être calculé et est montré dans l'Équation 2.4.

$$Q = \frac{PT_{sc}AU}{P_{sc}TZ} = \frac{3.067PAU}{(460+T)Z} \qquad (2.4)$$

où A est la section transversale du tubage. La deuxième formule donne le débit critique de gaz en unités de terrain, à savoir MMscf/jour.

2.2.3 Etude bibliographique

Un certain nombre de chercheurs ont proposé des méthodes pour déterminer le début du chargement de liquide. Ces méthodes sont généralement basées sur la corrélation des données de terrain et expérimentales, sur des équations dérivées de modèles mécanistiques ou une combinaison des deux. Presque toutes ces études s'accordent cependant sur deux modèles physiques pour l'élimination du liquide dans les puits de gaz (voir les Figures 2.16 et 2.17) :

- Le liquide est transporté sous forme de gouttelettes entraînées dans le noyau de gaz à haute vitesse.

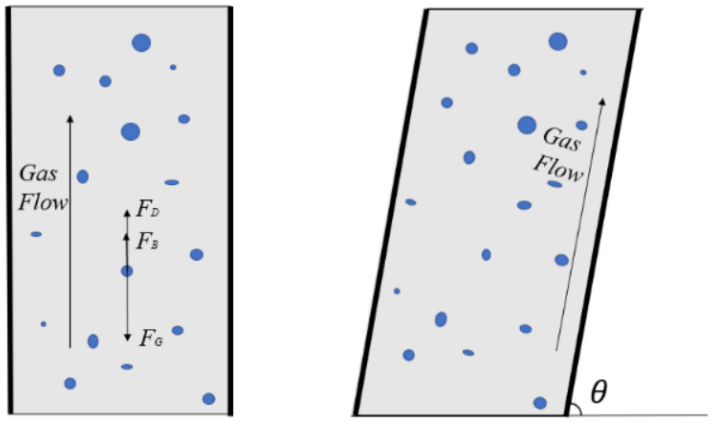


Figure 2.16 : Forces sur une gouttelette de liquide.

- Le liquide est transporté sous forme de film qui se déplace le long des parois du tuyau.

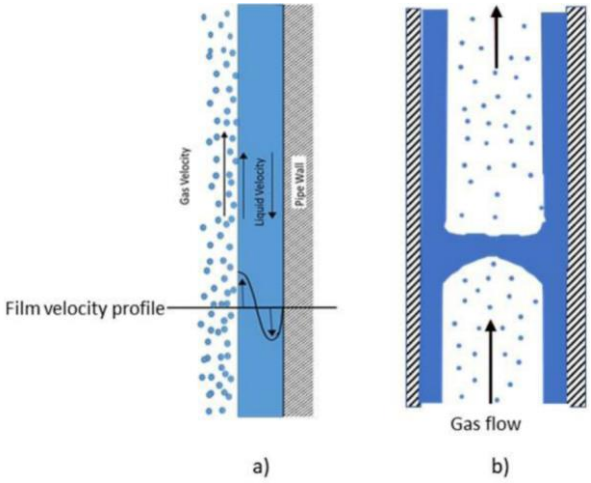


Figure 2.17 : Modèle d'Inversion de film. a) Instabilité du film. b) Pontage du liquide au centre du puits.

Tableau comparatif

2.2.3.1 Modèles de l'Élimination des Gouttelettes de Liquide

Tableau 2.3 : Corrélations empiriques - élimination des gouttelettes de liquide.

Titre	Auteur (Année)	Équation
Analysis and Prediction of Minimum Flow Rate for the Continuous Removal of Liquids from Gas Wells	Turner et al. (1969)	$v_{crit} = 6.494\left(\frac{\sigma(\rho_l-\rho_g)}{\rho_g^2}\right)^{\frac{1}{4}}$
A New Look at Predicting Gas-Well Load-Up	Coleman et al. (1991)	$v_{crit} = 5.464\left(\frac{\sigma(\rho_l-\rho_g)}{\rho_g^2}\right)^{\frac{1}{4}}$
Prediction of the Critical Gas Velocity of Liquid Unloading in a Horizontal Gas Well	Wang et Liu (2007)	$v_{crit} = C\left(\frac{sin^{0.38}(1.7\,\theta)}{0.74}\right)\left(\frac{\sigma(\rho_l-\rho_g)}{\rho_g^2}\right)^{\frac{1}{4}}$
Prediction Onset and Dynamic Behaviour of Liquid Loading Gas Wells	Belfroid et al. (2008)	$v_{crit} = \frac{sin^{0.38}(1.7\theta)}{0.74}\,5.464\left(\frac{\sigma(\rho_l-\rho_g)}{\rho_g^2}\right)^{\frac{1}{4}}$
A Simple Critical Gas Velocity Equation as Direct Functions of Diameter and Inclination for Horizontal Well Liquid Loading Prediction: Theory and Extensive Field Validation	Nagoo (2018)	$v_{crit} = \left(\frac{\rho_l-\rho_g}{40(\rho_g\sigma)^{0.5}}\right) g\,cos(90-\theta) D^{\frac{3}{2}}$

2.2.3.2 Modèles de la Dynamique/Inversion du Film de Liquide

Tableau 2.4 : Corrélations empiriques - méthode d'inversion du film de liquide.

Titre	Auteur (Année)	Équation
Transition from annular flow and from dispersed bubble flow—unified models for the whole range of pipe inclinations	Barnea (1986)	$\delta_c = \frac{1}{2}[\delta(0,\theta) + \delta(\pi,\theta)]$
Effects of thickness on the nanocrystalline structure and semiconductor-metal transition characteristics of vanadium dioxide thin films. Thin Solid Films.	Luo et al. (2014)	$\delta(\phi,\theta) = (1 - \alpha\theta\, cos\phi)\delta_c$ $\alpha = \begin{matrix} 0.0287 & 0 \le \theta \le 30 \\ 0.055\theta^{-0.868} & 30 \le \theta \le 90 \end{matrix}$
Improved Prediction of Liquid Loading In Gas Wells	Shekhar et al. (2017)	$\delta(\phi,\theta) = [1 - (\frac{1 - e^{-0.0880}}{1 + e^{-0.0880}})cos\theta]\underline{\delta}_{avgL}$

2.2.4 Formulation de la problématique

La production de gaz naturel revêt une importance économique majeure dans l'industrie énergétique mondiale. Toutefois, cette industrie est confrontée à des défis techniques significatifs, parmi lesquels le phénomène du chargement de liquide dans les puits de gaz. Ce phénomène se manifeste lorsque le débit de gaz n'est pas suffisant pour éliminer efficacement les liquides présents dans le puits, entraînant ainsi une diminution, voire un arrêt de la production de gaz. Par conséquent, la détection et la prédiction précoces du chargement de liquide revêtent une importance cruciale.

Cependant, la détection et la prédiction du chargement de liquide sont des tâches complexes, nécessitant l'analyse de multiples paramètres et une prise en compte approfondie des conditions du réservoir, des propriétés des fluides et des caractéristiques du puits. De plus, la modélisation de l'écoulement multiphasique dans les puits de gaz présente un défi en raison de la complexité des interactions entre le gaz et les liquides.

Dans ce contexte, l'apprentissage automatique, avec ses différentes approches telles que l'apprentissage supervisé et non supervisé, se présente comme une solution potentiellement prometteuse. Cette discipline a démontré sa capacité à traiter des problèmes complexes et à extraire des informations précieuses à partir de grandes quantités de données. Toutefois, l'application de l'apprentissage automatique à la détection et à la prédiction du chargement de liquide demeure un défi de taille. Cette application requiert une solide compréhension des principes fondamentaux de la production de puits de gaz ainsi que de l'écoulement multiphasique, conjuguée à une connaissance approfondie des différentes techniques et approches en apprentissage automatique.

Par conséquent, la problématique centrale de cette étude peut être formulée de la manière suivante :

Comment utiliser de manière efficace et optimale les différentes approches de l'apprentissage automatique pour détecter et prédire le chargement de liquide dans les puits de gaz, en tenant compte de la complexité inhérente de l'écoulement multiphasique et des conditions spécifiques du réservoir ?

Cette problématique peut être décomposée en plusieurs sous-questions :

1. L'apprentissage automatique peut-il surpasser l'efficacité des corrélations empiriques traditionnelles ?
2. Est-ce qu'un modèle hybride combinant à la fois des approches d'apprentissage automatique et des corrélations empiriques peut offrir une meilleure efficacité pour la détection et la prédiction du chargement de liquide ?
3. Parmi les différentes variables cibles telles que la vitesse du gaz, le débit du gaz, etc., laquelle revêt une importance primordiale pour la détection et la prédiction du chargement de liquide dans les puits de gaz ? et quels sont les attributs (features) les plus influents et informatifs pour l'étude du chargement de liquide dans les puits de gaz ?

En répondant à ces questions, nous serons en mesure d'évaluer de manière plus approfondie les avantages et les limites des différentes approches dans le contexte spécifique de la détection et de la prédiction du chargement de liquide. Cette démarche nous permettra d'identifier les stratégies et les pratiques optimales afin d'améliorer l'efficacité des techniques d'apprentissage automatique dans ce domaine.

Chapitre 3

Conception de la solution

Introduction

Pour résoudre notre problématique, au cœur d'une industrie en perpétuel évolution, l'exploration et la mise en œuvre de solutions de pointe sont primordiales. Elles nous permettent d'améliorer notre compréhension et notre maîtrise de ses complexités. Le chargement de liquide, l'un des phénomènes les plus difficiles à anticiper et à gérer, nécessite une solution à la fois novatrice et efficace. Dans ce chapitre, nous répondrons à ce besoin en introduisant deux approches distinctes, basées sur l'apprentissage automatique, pour la détection et la prédiction du chargement de liquide dans les puits de gaz.

Ce chapitre fournit un aperçu complet de l'application des techniques d'apprentissage automatique pour la détection et la prédiction du chargement de liquide. Il offre une vision détaillée du processus de développement de solutions, des premières étapes jusqu'à la mise en œuvre finale.

- Dans la première partie de ce chapitre, nous utilisons le simulateur OLGA (annexe), un outil de computation réputé, pour créer un modèle virtuel de l'écoulement multiphasique dans les puits de gaz. Cette démarche génère un ensemble de données complet et réaliste, imitant les conditions opérationnelles rencontrées dans les véritables puits de gaz. Cet ensemble de données, basé sur des scénarios conformes aux standards de l'industrie, sert de source principale d'input pour nos modèles d'apprentissage automatique, nous assurant ainsi que nos solutions sont adaptées aux dynamiques et aux complexités du monde réel de l'industrie pétrolière et gazière.

- La deuxième partie est centrée sur le prétraitement des données. Les données brutes, même soigneusement simulées, nécessitent souvent une préparation minutieuse pour être exploitées efficacement. Nous explorerons diverses techniques de transformation des données, notamment la standardisation, afin d'assurer que notre ensemble de données est dans le format le plus approprié pour l'apprentissage du modèle. Nous étudierons également la distribution des caractéristiques et définirons les métriques ciblées.

- La troisième partie s'immergera dans le processus de modélisation et l'ajustement des hyperparamètres, en concluant les features les plus influentes sur l'accumulation de liquide. Nous explorerons deux approches différentes de l'apprentissage automatique:

  Approche supervisée hybride :

  en tenant compte de la corrélation empirique de Turner et en utilisant une méthode de prédiction basée sur l'apprentissage supervisé, à savoir la classification.

  Approche non supervisée indépendante :

  en se basant sur l'apprentissage non supervisé pour segmenter les états des puits en deux groupes, chargé et non chargé.

  Approche supervisée comparative :

  en comparant les résultats d'un modèle d'apprentissage automatique avec celles des corrélations empiriques.

- La quatrième partie évaluera ces techniques. Chacune offre des avantages et des perspectives uniques sur le problème à résoudre, nous permettant de construire une compréhension globale des schémas et des comportements de chargement de liquide.

La Figure 3.1 schématise notre solution :

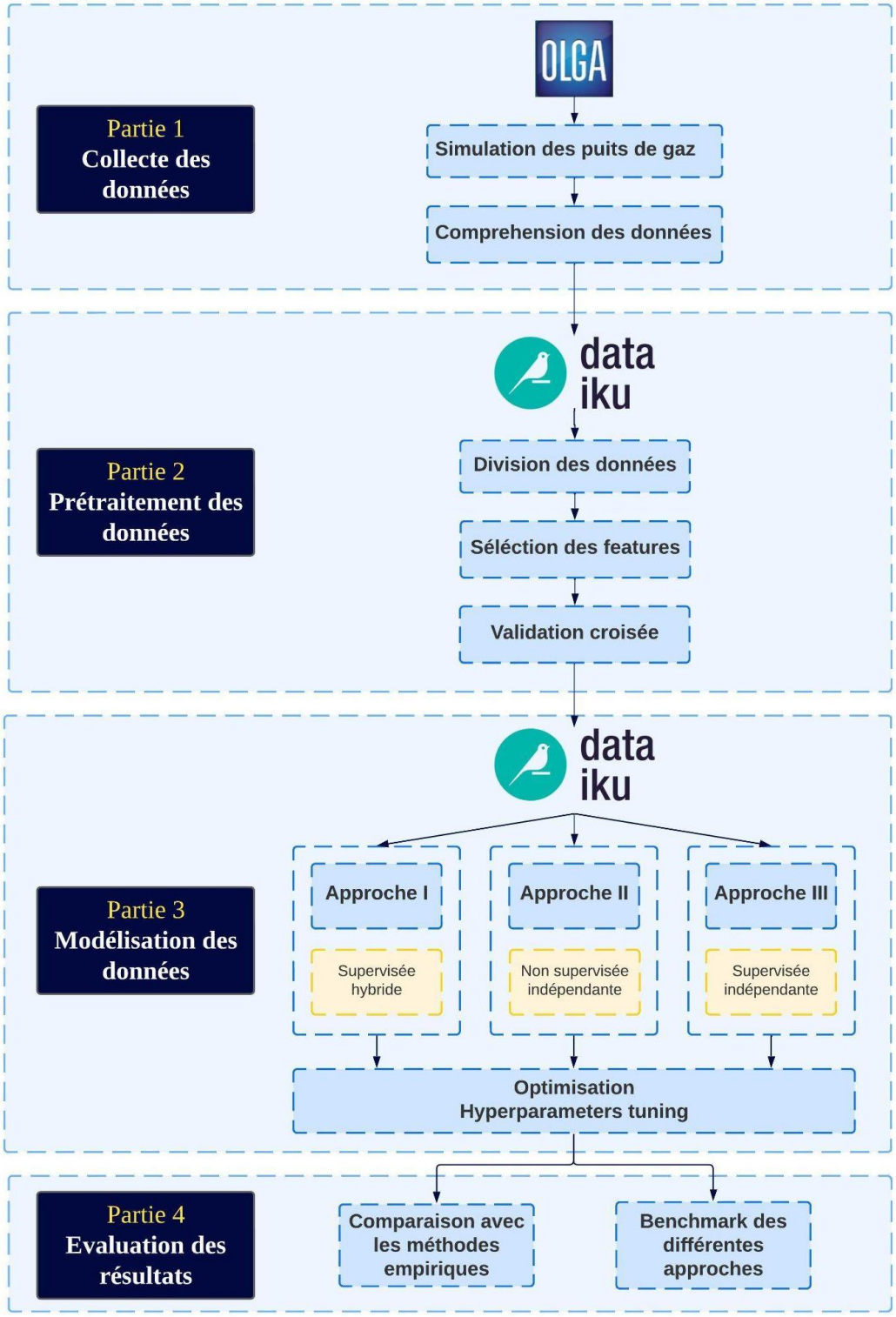


Figure 3.1 : Plan de la solution.

En fin de compte, ce chapitre vise à montrer le potentiel de l'apprentissage automatique pour révolutionner la détection et la prédiction du chargement de liquide. Grâce à une simulation diligente, un prétraitement des données méticuleux et une modélisation réfléchie, nous visons à concevoir une solution qui atténue non seulement les défis actuels posés par le chargement de liquide, mais ouvre également la voie à de futures avancées dans ce domaine crucial de l'industrie du pétrole et du gaz. Nous conclurons en discutant des résultats, en réfléchissant à l'efficacité de notre solution et en considérant ses implications pour les travaux futurs.

3.1 Collecte et compréhension des données

La première étape pour détecter et prédire le chargement de liquide dans les puits de gaz à l'aide de l'apprentissage automatique commence par la création d'un ensemble de données complet et de haute fidélité. Cet ensemble de données est la pierre angulaire de tout notre projet, informant et guidant nos modèles d'apprentissage automatique. Pour développer un tel ensemble de données, nous utilisons des mélange de données provenant des puits réels et du simulateur OLGA (annexe) pour augmenter les données. Ce dernier est un outil de calcul avancé reconnu pour sa capacité à simuler avec précision l'écoulement multiphasique dans les puits de gaz.

Composition des données d'entraînement

- Les données historiques sont utilisées pour évaluer nos différents modèles sur des puits réels. Elles représentent 10% du nombre total de puits (16 puits).
- Les puits simulés avec OLGA sont utilisés pour renforcer nos données d'entraînement. Ils représentent 90% du nombre total de puits (144 puits).

3.1.1 Simulation des puits de gaz

Dans cette partie du chapitre, nous nous concentrons sur la création d'un modèle virtuel de l'écoulement multiphasique dans les puits de gaz à l'aide du simulateur OLGA. En imitant avec précision les conditions opérationnelles complexes que l'on trouve dans les puits de gaz réels, nous sommes en mesure de générer un ensemble de données qui capture les dynamiques détaillées des scénarios standard de l'industrie.

Cette approche basée sur la simulation nous permet de modéliser les divers facteurs qui peuvent influencer le chargement de liquide, y compris les changements de pression, de température, et de composition du flux, entre autres. L'ensemble de données résultant nous fournit la profondeur et l'étendue des données nécessaires pour entraîner des modèles d'apprentissage automatique robustes et efficaces.

La Figure 3.2 démontre une simulation du phénomène de chargement de liquid sur le simulateur OLGA :

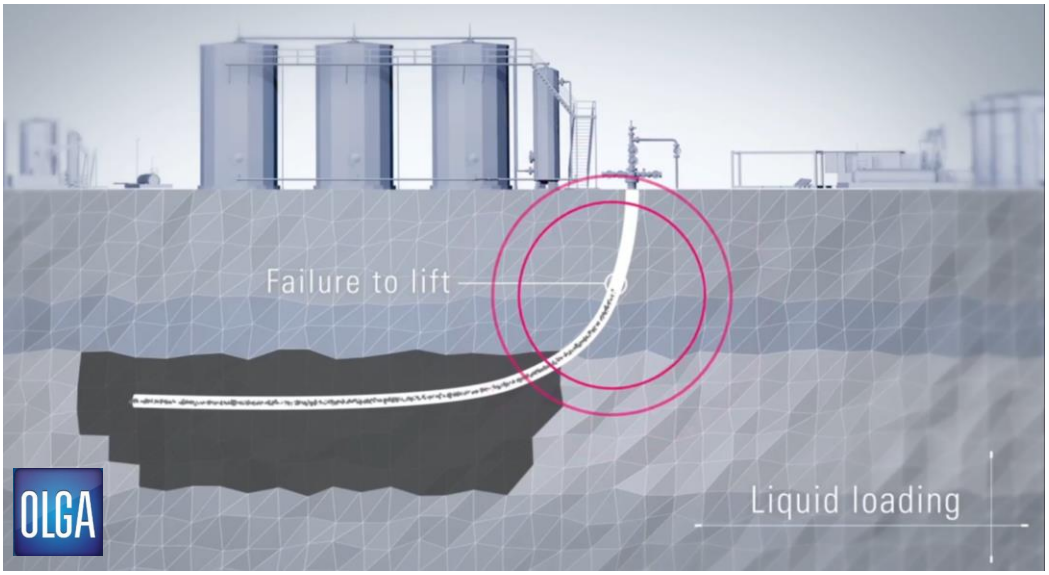


Figure 3.2 : Simulation d'un chargement de liquide sur OLGA.

3.1.2 Les raisons pour choisir OLGA

Lorsqu'il s'agit de simuler l'écoulement multiphasique dans les puits de gaz, le simulateur OLGA se démarque pour plusieurs raisons, ce qui en fait un choix idéal pour ce projet :

- OLGA est reconnu comme un simulateur de premier plan dans l'industrie en raison de sa capacité à représenter avec précision et réalisme une large gamme de scénarios opérationnels complexes. Sa capacité dans l'imitation des dynamiques du monde réel en fait un outil précieux pour générer des ensembles de données représentatifs pour les modèles d'apprentissage automatique.

- Lors de l'utilisation de données réelles, il y a un risque de biais inhérent dû aux conditions spécifiques ou aux limitations des données collectées. En simulant nos propres données avec OLGA, nous pouvons garantir une représentation large et complète des scénarios. Cela réduit le risque de biais et permet à nos modèles de mieux généraliser à une variété de conditions.

- Il offre un ensemble complet de capacités, y compris des simulations de forage sous-équilibrées transitoires, stationnaires et dynamiques. Cette flexibilité nous permet de modéliser un large spectre de scénarios, augmentant la richesse et la diversité de notre ensemble de données.

- Le niveau de détail fourni par OLGA est exceptionnel. Il nous permet de prendre en compte de nombreux facteurs influents tels que les changements de pression, les variations de température, et les changements de composition du flux. Une telle granularité améliore la qualité de notre ensemble de données, contribuant à la création de modèles d'apprentissage automatique plus précis et fiables.

- Dans de nombreux scénarios réels, obtenir une quantité suffisante de données de haute qualité et pertinentes peut être un défi. En utilisant le simulateur OLGA, nous pouvons générer autant de données que nécessaire, fournissant un ensemble de données robuste pour nos modèles d'apprentissage automatique. Cela assure une formation adéquate de nos modèles, conduisant à des prédictions plus fiables.

- L'utilisation d'OLGA nous permet de contrôler les paramètres exacts et les conditions sous lesquelles les données sont générées. Nous pouvons simuler une grande variété de scénarios, y compris certains qui peuvent être rares ou dangereux dans des situations réelles. Cela ajoute à la richesse et à la diversité de notre ensemble de données.

- Efficacité en termes de coûts et de temps : La collecte de données réelles à partir de puits de gaz peut être à la fois longue et coûteuse. La simulation des données en utilisant OLGA nous permet de contourner ces problèmes, économisant un temps et des ressources significatifs.

En somme, la reconnaissance industrielle d'OLGA, sa polyvalence, son niveau de détail, et son acceptation généralisée font de ce simulateur un excellent choix pour la phase de simulation de ce projet. Son utilisation nous permet de générer de manière rentable et efficace un ensemble de données complet, sans biais et riche. Ces données, que nous obtenons grâce à cet outil, serviront de base solide à nos modèles d'apprentissage automatique pour détecter et prédire efficacement le chargement de

liquide dans les puits de gaz, et elles seront précieuses lors des étapes ultérieures de prétraitement des données et de développement des modèles.

La Tableau 3.1 montre caractéristique d'un puits de gaz A

Tableau 3.1 : Caractéristique d'un puits de gaz A.

Pression (psia)	Temperature (C)	B (psi2/(scf/d))	C (psi2/(scf2/d2))
7200	134.2	30	0

Les Figures 3.3 et 3.4 montrent la précision d'une simulation d'un puits de gaz A avec OLGA, en comparant avec des mesures réelles.

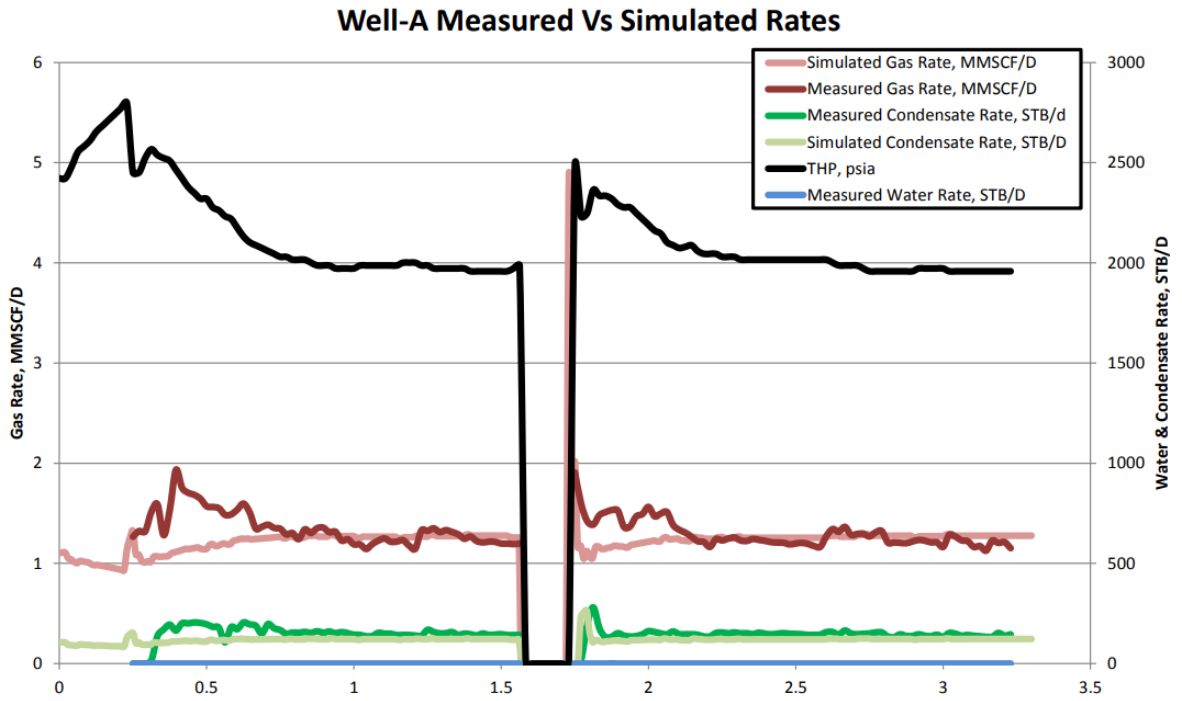


Figure 3.3 :Comparaison des résultats mesurées d'un puits A avec celles de la simulation sur OLGA

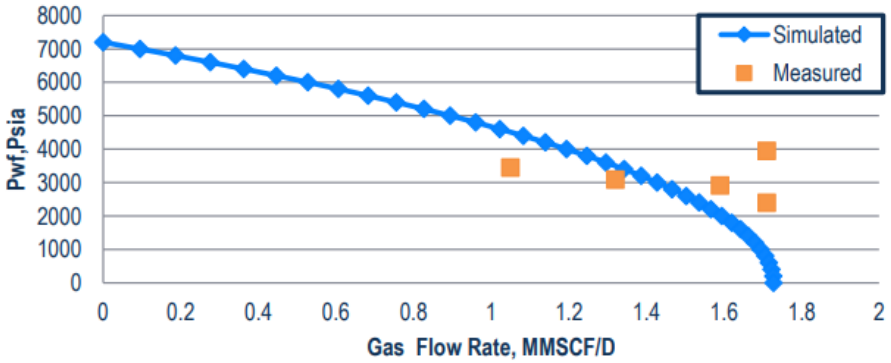


Figure 3.4 : Comparaison des résultats mesurées d'un puits A avec celles de la simulation sur OLGA.

3.1.3 Génération des données

Dans le cadre de cette étude, nous avons généré les données requises pour nourrir les algorithmes d'apprentissage automatique associés à diverses approches. Notre jeu de données se compose de 144 puits de gaz simulés présentant des caractéristiques distinctes, totalisant plus de 30 000 points. En effet, pour chaque puits, nous avons recueilli plusieurs points de données, chacun correspondant à des moments précis et équidistants. Ainsi, nous avons pu constituer un ensemble de données complet et riche pour l'entraînement et l'évaluation de nos modèles d'apprentissage automatique.

La Figure 3.5 présente une observation des 144 puits simulés.

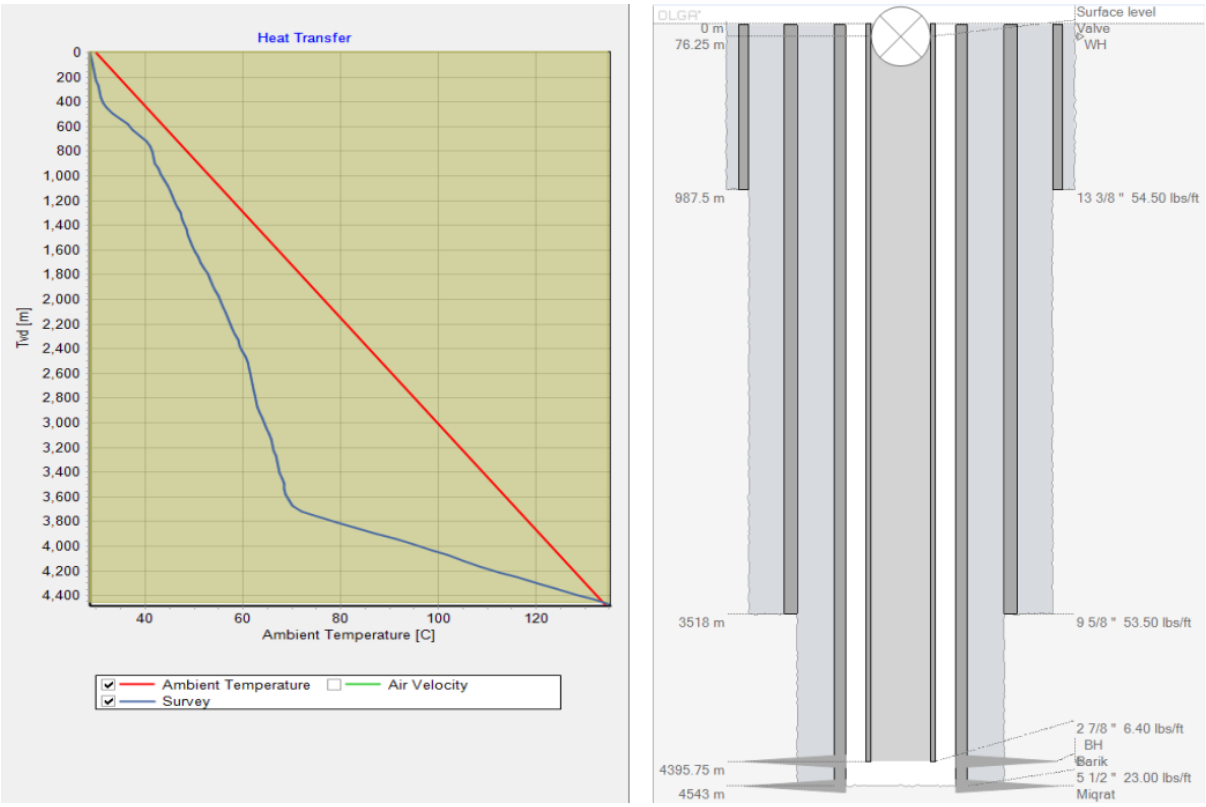


Figure 3.5 : Exemple des puits simulés.

3.1.4 Compréhension des données

Pour mieux comprendre nos données et les attributs, le Tableau 3.2 résume la description des 18 features (attributs) :

Tableau 3.2 : Description des attributs.

Attribut	Description
Date	Il s'agit de la marque temporelle associée à chaque observation ou mesure.
Débit de gaz	Il indique le volume de gaz produit par un puits pendant une période de 24 heures. Semblable au débit de liquide, il est essentiel pour évaluer la performance d'un puits.
Débit de liquide	Il s'agit du volume de liquide produit par un puits pendant une période de 24 heures. Ce chiffre est souvent utilisé pour évaluer la performance d'un puits.
Densité in situ du gaz	C'est la masse du gaz par unité de volume, mesurée dans les conditions de pression et de température du réservoir.
Densité in situ du liquide	Semblable à la densité in situ du gaz, il s'agit de la masse du liquide par unité de volume, mesurée dans les conditions de pression et de température du réservoir.
Déviation de l'assemblage de fond de trou	Cette mesure indique l'angle de déviation de l'assemblage au fond du puits par rapport à la verticale. Elle peut affecter l'écoulement des fluides dans le puits.
Diamètre du cuvelage	Il s'agit de la dimension transversale interne du cuvelage d'un puits. Elle influence le volume de fluide que le puits peut produire.
Diamètre du tubage	C'est la dimension transversale interne du tubage d'un puits, qui a également un impact sur le volume de fluide que le puits peut produire.
Gravité spécifique du gaz	C'est le rapport de la densité du gaz à celle de l'air, ce qui a un impact sur le comportement du gaz.
Gravité spécifique du liquide	C'est le rapport de la densité du liquide à celle de l'eau. Elle affecte les caractéristiques d'écoulement du liquide.
Pression de cuvelage	Cette caractéristique indique la pression à l'intérieur du cuvelage (ou chemisage) d'un puits. Elle est souvent surveillée pour assurer l'intégrité du puits.
Pression de ligne	Cette mesure indique la pression à l'intérieur d'une conduite ou d'un pipeline. Elle est cruciale pour garantir le bon fonctionnement des systèmes de transport de fluide.
Pression de tubage	Il s'agit de la pression à l'intérieur du tubage d'un puits. Cette mesure est importante pour la sécurité et le contrôle des opérations de puits.
Profondeur de l'assemblage de fond de trou	C'est la distance entre la surface et l'assemblage au fond du puits. Cette mesure est importante pour la gestion des opérations de puits et la modélisation de la performance du puits.
Température de fond de trou	Il s'agit de la température au fond du puits, qui peut affecter les caractéristiques d'écoulement des fluides.

Température de surface	C'est la température à la surface du puits. Elle peut avoir un impact sur la performance du puits et la gestion de l'équipement.
Tension de surface	C'est une mesure de l'attraction entre les molécules d'un liquide. Elle a un impact sur des phénomènes comme la capillarité et l'écoulement des fluides dans un puits.
Viscosité in situ du liquide	Il s'agit de la mesure de la résistance d'un fluide à l'écoulement dans les conditions de pression et de température du réservoir.
Vitesse superficielle du liquide	Cette caractéristique représente la vitesse à laquelle un fluide se déplace à travers une section transversale spécifique d'un tuyau ou d'un autre conduit.

On peut également visualiser la distribution de plusieurs attributs :

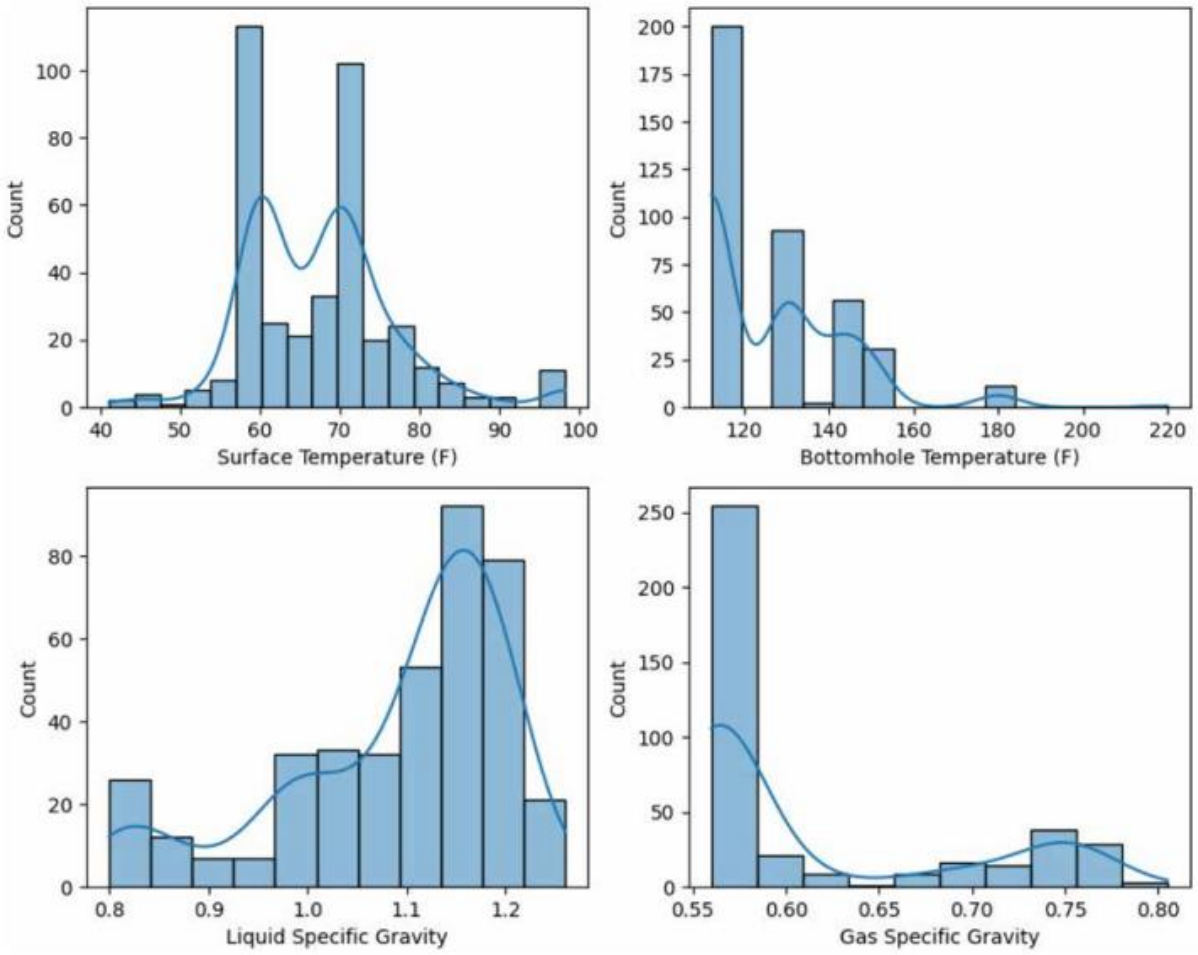


Figure 3.6 : Les histogrammes de la température de surface/températures de fond de trou et des gravités spécifiques du liquide/du gaz pour les puits étudiés.

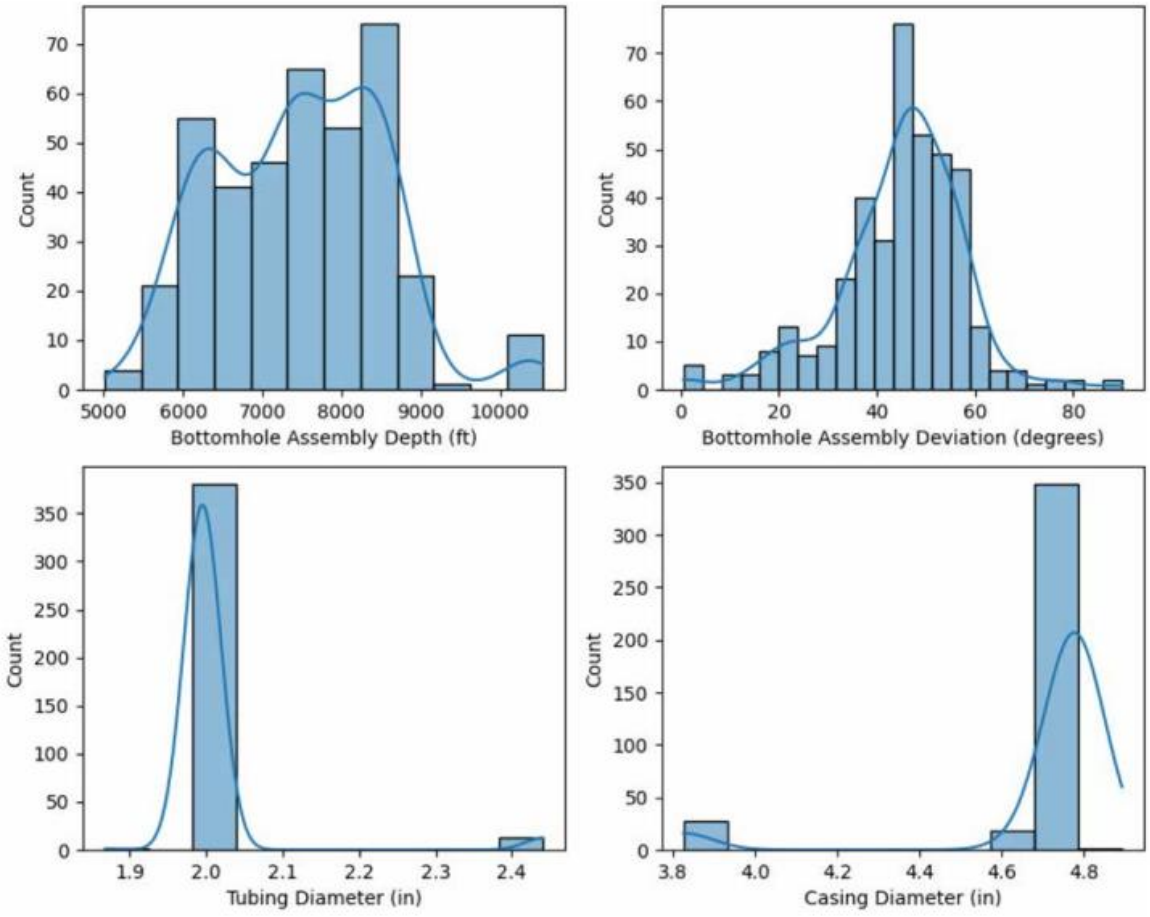


Figure 3.7 : Les histogrammes de la profondeur/déviation de l'assemblage de fond de trou, et du diamètre du tubage/du cuvelage pour les puits étudiés.

Et finalement on peut faire une analyse statistique sur nos données :

Tableau 3.3 : Aperçu sur les données numériques de notre jeu de données.

Attribut	moy	std	min	25%	50%	75%	max
Débit de gaz (Mscf/j)	585,7	216,44	0.05	434,37	542,45	739,44	2161,1
Débit de liquide (barils/j)	4,98	10,63	0	0,80	1,50	3,00	98,56
Densité in situ du gaz (lbm/pieds³)	0,95	0,47	0,16	0,51	1,00	1,31	2,70
Densité in situ du liquide (lbm/pieds³)	69,82	5,29	49,89	67,89	70,99	73,34	78,75
Déviation de l'assemblage de fond de trou (°)	47,57	11,88	0,62	45,30	47,90	55,01	89,9
Diamètre du cuvelage (pouces)	4,74	0,18	3,83	4,78	4,78	4,78	4,89
Diamètre du tubage (pouces)	2,01	0,07	1,87	1,99	1,99	1,99	2,44
Gravité spécifique du gaz (m²/s)	0,58	0,05	0,56	0,56	0,56	0,57	0,81
Gravité spécifique du liquide (m²/s)	1,12	0,09	0,80	1,09	1,14	1,18	1,26
Pression de cuvelage (Psig)	368,19	151,03	74,04	229,23	391,45	479,33	1059,8
Pression de ligne (Psig)	294,77	157,81	50,48	147,62	330,03	420,37	680,93
Pression de tubage (Psig)	308,29	154,63	54,25	161,07	340,97	429,50	808,54
Profondeur de l'assemblage de fond (pieds)	7864,7	909,53	5011	7385	7991	8483	10545
Température de fond de trou (°F)	119,76	14,97	112,2	112,20	112,20	112,20	220
Température de surface (°F)	70,65	9,02	41,10	65,30	70,00	77,00	98,20
Tension de surface (dynes/cm)	63,99	9,30	19,09	65,40	65,97	68,28	70,53
Viscosité in situ du liquide (cP)	2,47	2,08	0,76	1,42	1,73	2,42	15,41
Vitesse superficielle du liquide (pieds/s)	0,01	0,02	0,00	0,00	0,00	0,01	0,18

Index :

La "Date" peut être utilisée comme index pour les séries temporelles.

Importance des variables :

- On remarque que la variable "Vitesse superficielle du liquide" présente une variance très faible, ce qui suggère qu'elle pourrait avoir peu d'importance pour l'entraînement des différents modèles, étant donné qu'elle ne fournit pas beaucoup d'informations.
- On pourrait également envisager d'éliminer d'autres variables présentant un faible écart type, telles que :
    - Gravité spécifique du liquide
    - Gravité spécifique du gaz

- Diamètre du tubage
- Densité in situ du gaz
- Diamètre du cuvelage

Ces variables auront probablement un impact minime sur l'entraînement du modèle.

Distributions des données :

- **Déviation de l'assemblage de fond de trou :** presque elle suit la loi normale, ce type de distribution pourrait faciliter l'entraînement du modèle.

- **Les diamètres :** ont un fort pourcentage des données aberrantes, une transformation non linéaire telle que la transformation Robuste ou celle de Box Cox serait nécessaire.

- **Températures de fond de trou :** une distribution asymétrique négative (asymétrique à gauche). Une transformation logarithmique pourrait être une solution envisageable.

- **Gravité :** une distribution asymétrique positive (asymétrique à droite). Une transformation racine carrée est une solution envisageable.

3.2 Prétraitement des données

Notre travail sera implémenté sur Dataiku (annexe), ou les étapes suivantes seront réalisées.

Cette plateforme prend en charge le cycle complet de l'analyse de données, y compris la collecte, le nettoyage, l'analyse et la visualisation des données, jusqu'à la création et le déploiement de modèles d'apprentissage automatique. Elle permet aux utilisateurs de différentes compétences de travailler ensemble sur des projets de données, grâce à son interface utilisateur intuitive et à sa prise en charge de nombreux langages de programmation tels que Python (annexe), R et SQL.

3.2.1 Methode d'entrainement

Pour cette étude, les données de 160 puits sont collectées. Tout d'abord, l'étude est effectuée sur un seul puits où 80% des informations sont utilisées à des fins d'entraînement, et 20% pour la validation du modèle. Les résultats montrent que le modèle intelligent est capable de prédire précisément le début du chargement de liquide dans le puits et de lever un drapeau d'avertissement lorsque la possibilité de chargement de liquide est élevée.

Ensuite, une série de puits est choisie et un modèle intelligent est construit sur la base de 80% de formation et 20% de validation. Ce modèle est ensuite utilisé pour prédire le chargement de liquide dans un puits différent comme un puits complètement aveugle.

La méthode de la division des données

En plus de la division classique des données, cette étude fait usage de la validation croisée (annexe), une technique robuste qui permet d'estimer l'erreur de généralisation d'un modèle sur l'ensemble de données complet. Parmi les diverses méthodes de validation croisée, la méthode Stratified K-Fold a été choisie.

La méthode Stratified K-Fold est une amélioration de la méthode de validation croisée K-Fold classique. Dans la validation croisée K-Fold, l'ensemble des données est divisé en 'K' sous-groupes ou "folds". Le modèle est alors entraîné sur K-1 folds et testé sur le fold restant, répétant ce processus K fois pour garantir que chaque fold a servi une fois comme ensemble de test. (annexe)

Cependant, cette méthode peut conduire à une représentation inégale des différentes classes de l'ensemble de données dans les folds d'entraînement et de test, en particulier lorsque l'ensemble de données est déséquilibré. Pour résoudre ce problème, la méthode Stratified K-Fold a été utilisée. Dans cette approche, les données sont divisées de manière à maintenir le même ratio de classes dans chaque fold que celui présent dans l'ensemble de données original.

Dans le cadre de cette recherche, l'emploi de la méthode Stratified K-Fold a permis d'assurer une représentation équitable des différentes conditions de chargement de liquide dans les puits à travers les ensembles d'entraînement et de validation, optimisant ainsi la performance et la fiabilité de nos modèles prédictifs.

On va fixer le nombre K = 5, pour l'entraînement des modèles et pour l'ajustement des hyperparamètres, avec l'AutoML (annexe).

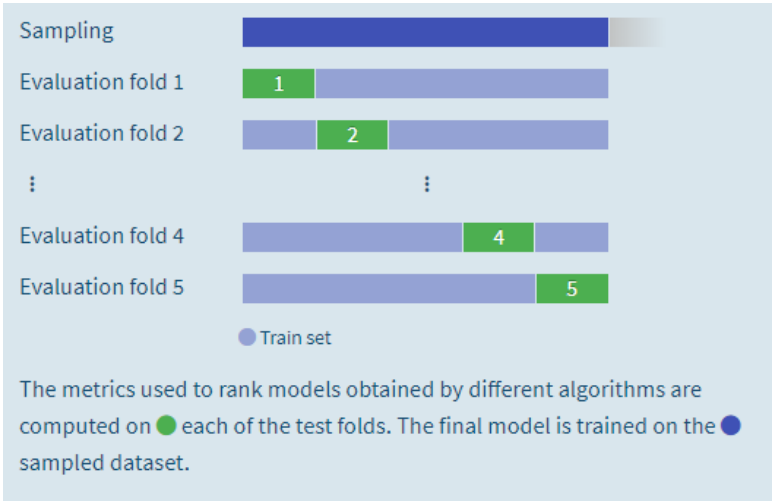


Figure 3.8 : Stratified K-fold avec 5 iterations sur Dataiku.

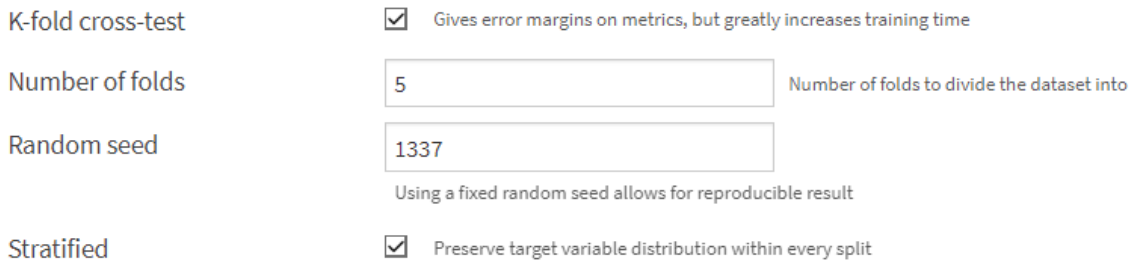


Figure 3.9 : Paramétrage du Stratified K-fold avec 5 itérations sur Dataiku.

3.2.2 Transformation des données

Les transformations qui vont être utilisées :

Normalisation : C'est une technique de mise à l'échelle qui ajuste les valeurs de plusieurs variables afin qu'elles aient une norme unitaire (longueur égale à 1). En d'autres termes, chaque valeur d'une variable est divisée par la norme (ou la longueur) de toutes les valeurs de cette variable.

$$x' = \frac{x}{||x||} \qquad (3.1)$$

Z-Score Normalisation : Aussi appelée standardisation, cette technique transforme une variable pour qu'elle ait une moyenne de 0 et un écart-type de 1. Cela est réalisé en soustrayant la moyenne de chaque valeur, puis en divisant par l'écart-type.

$$x' = \frac{(x - \mu)}{\sigma} \qquad (3.2)$$

MinMax Scaler (mise à l'échelle MinMax) : Cette technique réduit l'échelle des données dans un intervalle spécifié, généralement [0, 1]. Elle soustrait la valeur minimale de chaque valeur de la variable, puis divisé par la plage de cette variable

$$x' = \frac{x - min(x)}{min(x) - max(x)} \qquad (3.3)$$

Transformation logarithmique : Cette transformation applique le logarithme naturel à chaque valeur de la variable. Elle est souvent utilisée pour réduire l'asymétrie à gauche des données .

$$x' = log(x) \qquad (3.4)$$

Transformation racine carrée : Cette transformation applique la racine carrée à chaque valeur de la variable. Elle est souvent utilisée pour réduire l'asymétrie modérée des données à droite.

$$x' = \sqrt{x} \qquad (3.5)$$

Robust Scaler (mise à l'échelle robuste) : Cette technique est similaire à la mise à l'échelle MinMax, mais elle utilise les quartiles (Q1, Q3) plutôt que le min et le max pour calculer la plage. Cela la rend résistante aux valeurs aberrantes.

$$x' = \frac{x - Q_1(x)}{Q_3(x) - Q_1(x)} \qquad (3.6)$$

Transformation Box-Cox : La transformation Box-Cox est utilisée pour rendre une variable plus normalement distribuée à travers une transformation exponentielle. Cette transformation nécessite que toutes les données soient strictement positives. Le paramètre lambda (λ) de la transformation est souvent choisi de manière à maximiser la vraisemblance.

$$\begin{aligned} &\text{Si } \lambda \neq 0 : \ x' = \frac{x^{\lambda} - 1}{\lambda} \\ &\text{Sinon} \ : \ x' = log(x) \end{aligned} \qquad (3.7)$$

3.2.3 Sélection des features

Pour générer un modèle d'apprentissage automatique pour la prédiction du débit critique, une sélection de features est effectuée en utilisant la régression de l'information mutuelle (MI) afin de déterminer quelles variables d'entrée sont les plus pertinentes pour le débit critique du gaz. Le score MI est une quantité non négative sans unité qui mesure la réduction de l'incertitude d'une variable donnée une valeur connue d'une autre variable.

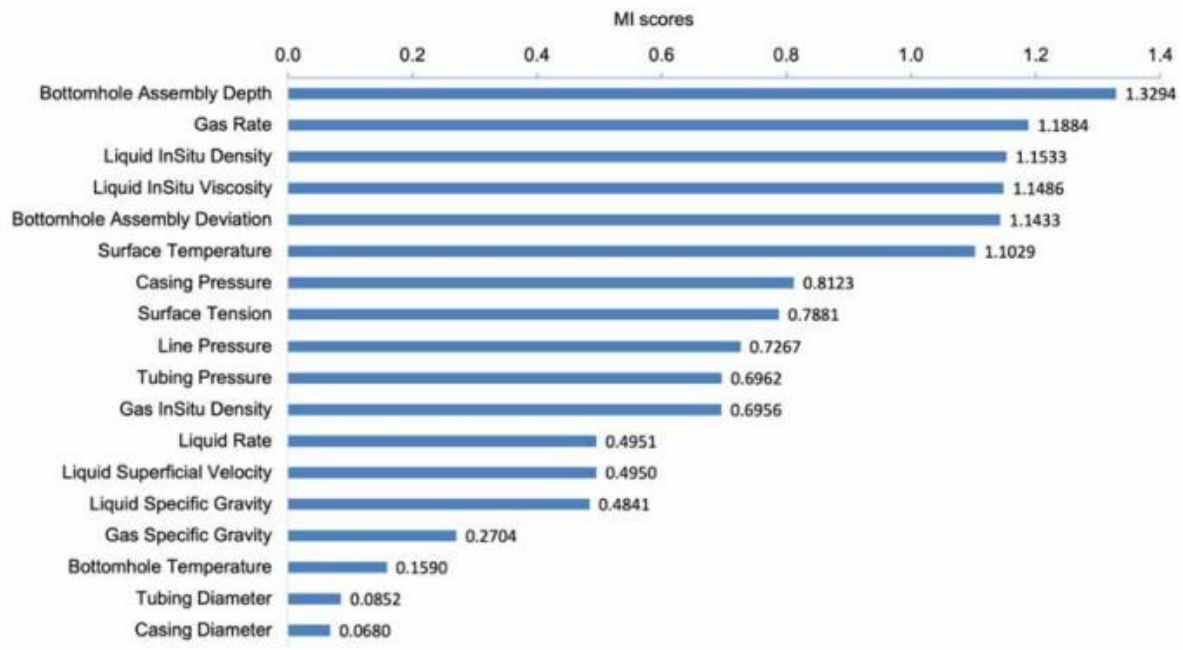
$$I(X,Y) = \int_y^{\cdot} \int_x^{\cdot} P_{(X,Y)}(x,y).\, log\left(\frac{P_{(X,Y)}(x,y)}{P_{(X)}(x) P_{(Y)}(y)}\right) dxdy \qquad (3.8)$$

où P_x et P_y sont les densités de probabilité marginales, et $P_{(X,Y)}$ est la densité de probabilité conjointe.

La régression MI est utilisée pour trouver la dépendance de la variable cible par rapport aux caractéristiques. Ici, la variable cible est le débit du gaz et les caractéristiques sont toutes les autres variables du Tableau 3.2.

La Figure 3.10 montre les scores MI des différentes caractéristiques. Un score MI plus élevé indique une dépendance plus forte de la variable cible par rapport à la caractéristique correspondante.

Figure 3.10 : La dépendance de la variable cible sur les caractéristiques disponibles basée sur la régression de l'information mutuelle.



Le Tableau 3.4 répertorie les caractéristiques éliminées et les raisons de leur élimination. Toutes les caractéristiques éliminées présentent de faibles scores MI, et la variable cible montre une faible dépendance à leur égard. Par conséquent, 12 caractéristiques sont sélectionnées pour l'entraînement du modèle d'apprentissage automatique. Le débit de liquide est conservé dans la liste des caractéristiques car il est considéré comme ayant un impact sur le débit critique, bien qu'il présente un faible score MI ici.

Tableau 3.4 : Caractéristiques éliminées et raisons de les éliminer.

Feature	La raison
Vitesse superficielle du liquide	Faible score MI, existence de “débit de liquide” avec un score MI plus élevé.
Gravité spécifique du liquide	Faible score MI, existence de “densité in situ du liquide” avec un score MI plus élevé.
Gravité spécifique du gaz	Faible score MI, existence de “densité in situ du gaz” avec un score MI plus élevé.
Température de fond de puits	Faible score MI, faibles variations dans le jeu de données (voir Figure 3.6).
Diamètre du tubage	Faible score MI, faibles variations dans le jeu de données (voir Figure 3.7).
Diamètre du cuvelage	Faible score MI, pas de dépendance de la cible par rapport au diamètre du cuvelage basée sur la physique du problème, faibles variations dans le jeu de données (voir Figure 3.7).

3.3 Modélisation

3.3.1 Approche I : Classification hybride

Pour prouver le concept de l'utilisation d'un modèle basé sur les données comme outil prédictif pour détecter le chargement de liquide, 6 algorithmes d'apprentissage automatique sont utilisés :

- XGBoost
- RF
- DT
- SVM
- GDBT
- KNN

Étant donné que ces algorithmes sont des algorithmes d'apprentissage supervisé, chaque échantillon du jeu de données doit être étiqueté comme "Chargé" ou "Non chargé". Pour étiqueter les données, les critères de Turner pour le chargement de liquide sont utilisés.

Il est important de mentionner que nous avons utilisé le taux de Turner comme point de contrôle à ce stade. Ces modèles hybrides combinent à la fois la corrélation empirique classique et une méthodologie prédictive grâce à l'apprentissage automatique.

Si le taux est supérieur au taux de Turner, il n'y a pas de chargement dans le puits, mais si le taux tombe en dessous du taux de Turner, le puits est chargé. Une troisième classe pourrait également être ajoutée en tant que classe "Alerte" où le taux est proche du taux de Turner mais pas inférieur. Trois colonnes binaires sont ajoutées au jeu de données pour indiquer l'état de chargement. Si le puits n'est pas chargé, la valeur dans la colonne "Non chargé" devient "1", et si le puits est chargé, la valeur dans la colonne "Chargé" devient "1", et de même pour la classe "Alerte" comme indiqué dans le Tableau 3.5.

Tableau 3.5 : Statut de classification des puits.

Non chargé	Alerte	Chargé
1	0	0
0	1	0
0	0	1

- Il faut noter que la métrique principale est ROC/AUC score (annexe)

Les résultats d'entraînement montrés dans les Figures 3.11 et 3.12 :

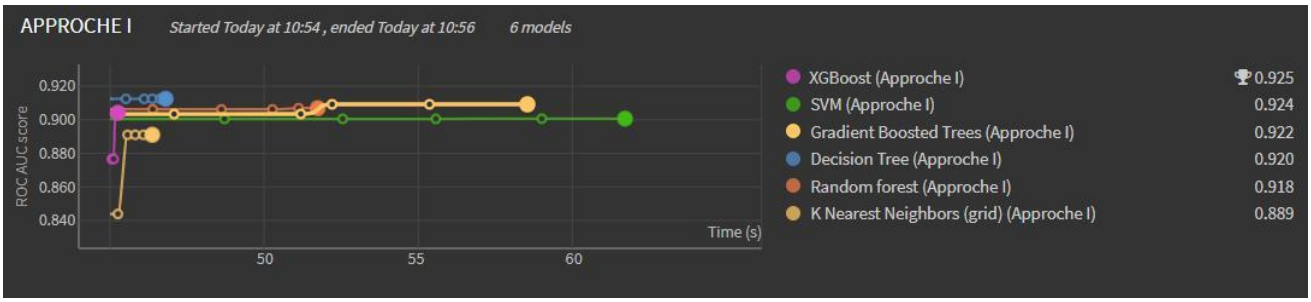


Figure 3.11 : Résultats d'entraînement 1 approche I sur Dataiku.

	Name	Trained	Accuracy	Precision	Recall	F1 Score	Cost Matrix Gain	Log Loss	ROC AUC	Calibration Loss	Lift
☐ ✔	XGBoost (Approche I)	2023-06-30 10:54:07	0.88	0.95	0.91	0.93	0.77	0.33	0.92	0.15	1.16
☐ ✔	K Nearest Neighbors (grid)	2023-06-30 10:54:03	0.86	0.86	1.00	0.92	0.82	0.49	0.89	0.06	1.14
☐ ✔	Decision Tree (Approche I)	2023-06-30 10:54:02	0.88	0.95	0.91	0.93	0.77	0.29	0.92	0.10	1.16
☐ ✔	Gradient Boosted Trees	2023-06-30 10:53:47	0.87	0.88	0.99	0.93	0.82	0.22	0.92	0.03	1.16
☐ ✔	SVM (Approche I)	2023-06-30 10:53:37	0.87	0.87	0.99	0.93	0.82	0.22	0.92	0.02	1.16

Figure 3.12 : Résultats d'entraînement 2 approche I sur Dataiku.

Analyse des resultats :

- Le modèle XGBoost était le plus performant, avec une optimisation des hyperparamètres.

On va donc utiliser ce modèle pour prédire le débit de production d'un puits réel et complètement aveugle, Figure 3.13 :

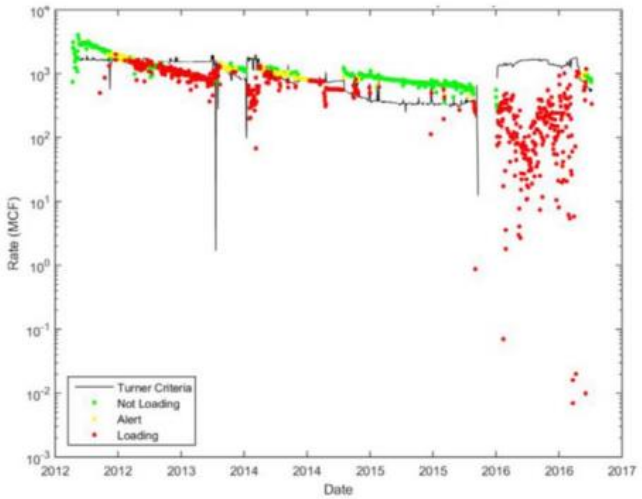


Figure 3.13 : Débit de production du puits aveugle comparant avec l'approche I.

Figure 3.13 compare les trois classes de chargé, déchargé et état d'avertissement du puits par rapport aux critères de taux de Turner. Le profil de production codé par couleur est en très bon accord avec les critères de taux de Turner.

On peut remarquer que notre modèle est plus performant face au modèle du Turner classique en comparant l'état réel du puits.

Dans la formation du modèle dans cette technique pour développer le modèle intelligent en utilisant XGBoost, nous avons encore besoin d'étiqueter nos données avec les statuts chargé, déchargé ou d'avertissement basés sur les critères de vitesse critique de Turner et al. Cette approche considère implicitement que les critères de Turner sont valides dans les réservoirs non conventionnels.

3.3.2 Approche II : Clustering indépendant

Pour éliminer cette dépendance implicite aux critères de Turner dans la section suivante, nous avons essayé un algorithme non supervisé qui ne nécessite aucune information préalable sur l'état du puits. Par conséquent, cela libère la dépendance du modèle à toutes les corrélations théoriques ou empiriques.

Dans les algorithmes non supervisés, il est important d'initialiser le nombre des clustres bien avant l'entraînement. Des différentes méthodes pour déterminer le nombre des clusters, notamment la méthode de coude.

Pour notre cas, le nombre de clusters doit être deux, vu la nature de notre problème. Donc les 4 algorithmes non supervisés auront le même paramétrage de nombre des clusters :

- K-means
- SC
- GM
- DBSCAN

- Il faut noter que la métrique principale est le coefficient de silhouette (annexe)

Pour garantir la reproductibilité de cette approche, les centroïdes initiaux doivent être fournis aux algorithmes

Les résultats d'entraînement montrés dans la Figures 3.14 :

Figure 3.14 : Résultats d'entraînement approche II sur Dataiku.

Analyse des resultats :

- K-means était le modèle le plus performant vu qu'il a la valeur la plus proche à 1.

Determiner les centroides corrects :

L'objectif des algorithmes non supervisés est seulement de pouvoir segmenter les données, donc même après avoir ces résultats, on devrait savoir quel segment représente les charges et les puits non charges.

Les centroïdes initiaux corrects peuvent être obtenus en utilisant un raisonnement logique. À cette fin, nous avons revisité les signes courants de chargement de liquide discutés précédemment et les avons utilisés comme lignes directrices pour choisir les

centroïdes initiaux. Par exemple, dans le problème de chargement de liquide, entre un débit élevé et un débit faible, il est évident qu'un débit élevé est moins susceptible d'avoir un problème de chargement de liquide et qu'un débit faible est un bon candidat pour avoir des problèmes de chargement. De même, entre 2 ensembles de pressions de tubage et de cuvelage, le cas avec une différence plus élevée entre la pression de cuvelage et la pression de tubage est plus susceptible d'être chargé.

Le Tableau 3.6 ci-dessous résume les caractéristiques des deux centroïdes :

Tableau 3.6 : Caractéristiques des centroïdes du K-means.

Centroid	Pression de cuvelage (psi)	Pression de tubage (psi)	Debit de gaz (MMCF)
1	550	450	1160
2	850	650	320

- Grâce à l'analyse précédente, on peut déduire le statut de chaque centroid :

Tableau 3.7 : Statuts des centroïdes du K-means.

Centroid	Statut
1	Non chargé
2	Chargé

Les valeurs du Tableau 3.6 ont été utilisées comme centroids initiaux et une classification réussie a été réalisée en utilisant l'algorithme k-means. Les points verts représentent une situation non chargée et les points rouges représentent l'état de chargement de liquide. Le taux de Turner est également illustré dans le Figure 3.15 pour montrer la différence entre cette approche et la méthode classique de Turner. Il est évident qu'après l'année 2013, le puits est chargé et des mesures ont été prises pour résoudre ce problème. Ensuite, le puits a été produit normalement pendant une courte période de temps, suivie d'une diminution du débit de gaz et de problèmes de chargement de liquide.

La même analyse a été réalisée en utilisant le même centroïde initial pour un autre puits réel #2. Les résultats montrent une grande précision du modèle pour prédire le chargement de liquide, comme présenté dans la Figure 3.15 et la Figure 3.16. La Figure 3.15 et la Figure 3.16 montrent que le même modèle peut être utilisé pour identifier le début du chargement de liquide dans deux puits réels complètement aveugles. En général, les prédictions de l'état du puits (points rouges et verts) sont en accord étroit avec les prédictions de Turner, cependant, il y a des cas où le débit minimal de gaz de Turner est sous-estimé, comme dans la deuxième moitié de 2015 de la Figure 3.15, où le débit de gaz est supérieur au débit de Turner mais notre modèle a quand même identifié le chargement de liquide. Des observations similaires peuvent être trouvées dans la Figure 3.16.

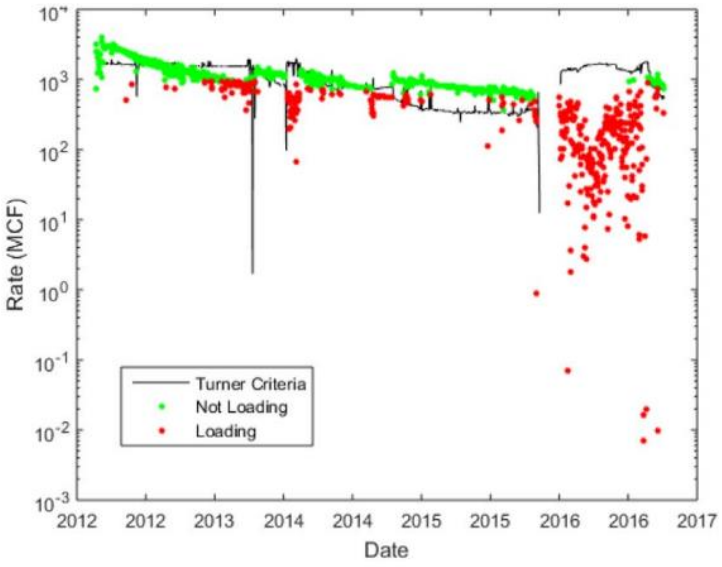


Figure 3.15 : Débit de production du puits #1 aveugle comparant avec l'approche II.

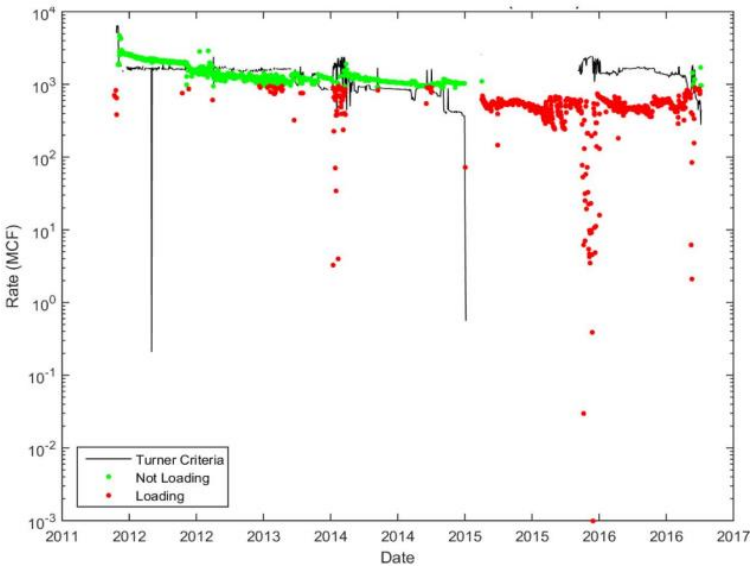


Figure 3.16 : Débit de production du puits #2 aveugle comparant avec l'approche II.

3.3.3 Approche III : Régression comparative

Pour cette approche, on va utiliser juste les données provenant des puits réels. Elle traite uniquement les points ou le chargement de liquide est confirmé.

Le débit de gaz sera notre variable cible pour la modélisation. Notre modèle essayera de prédire le débit de gaz, par contre on va comparer sa précision seulement avec les points du chargement de liquide. L'erreur sera calculée donc que pour les points de chargement de liquide.

Le but de cette approche est de comparer la précision de l'apprentissage automatique face aux différentes corrélations empiriques. Le début du chargement de liquide analytiquement pour cette approche, il est détecté au point où la pression du cuvelage atteint un minimum, en raison de la colonne hydrostatique induisant une contre-pression dans le puits, tandis que la baisse du débit augmente de manière significative, et la pression du tubage et de la conduite reste presque constante.

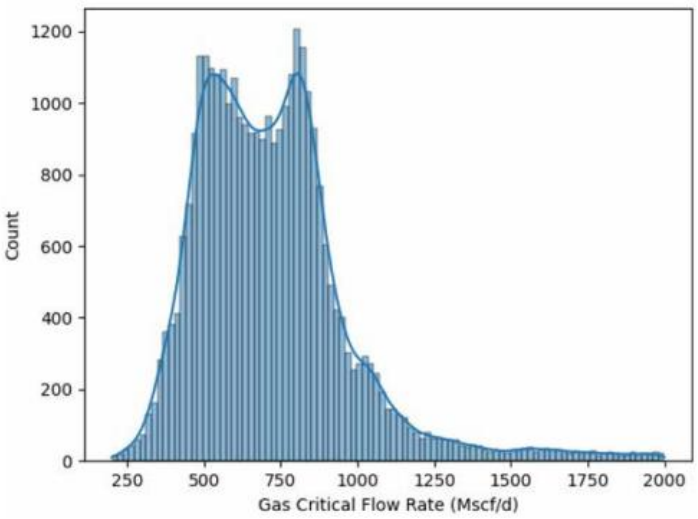


Figure 3.17 : Distribution de débit de gaz critique (au moment du chargement de liquide).

Comparaison entre les débits critiques détectés et les débits critiques basés sur les corrélations empiriques

Le débit critique obtenu à partir des données sur le terrain est comparé aux corrélations du Tableau 2.3 et Tableau 2.4 pour calculer le débit critique. La Figure 3.18 montre l'histogramme des différences entre les débits critiques obtenus et les corrélations empiriques. Les résultats montrent que les équations de Nagoo, Coleman, Turner et Belfroid ont tendance à sous-estimer le débit critique, tandis que les équations de Wang et Shekhar ont tendance à surestimer le débit critique pour les puits de l'ensemble de données.

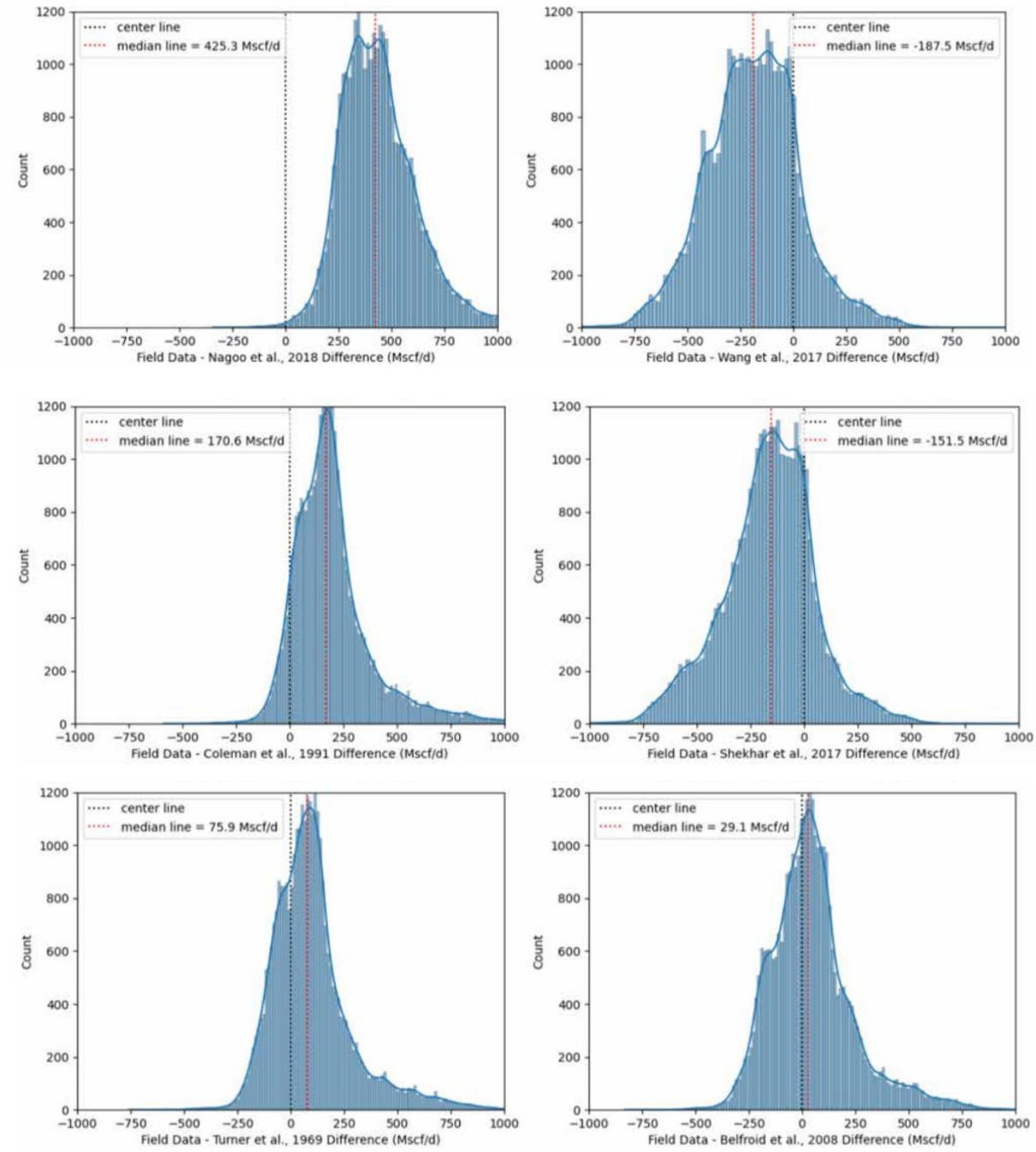


Figure 3.18 : L'histogramme des différences entre les débits critiques trouvés et les équations empiriques.

Les résultats confirment que Belfroid est la corrélation la plus précise parmi les corrélations étudiées, en termes de différences médianes et absolues moyennes. Par conséquent, Belfroid est utilisé dans cette étude pour des analyses ultérieures dans les sections suivantes.

On va choisir le modèle le plus performant de la section précédente, qui est XGBoost pour la partie apprentissage automatique.

Modelisation XGBoost

La Figure 3.19 montre la comparaison entre les prédictions du modèle XGBoost et les points de débit critique détectés. Les résultats montrent que le score R^2 (annexe) pour les données d'entraînement et de test est respectivement de 0,973 et 0,90.

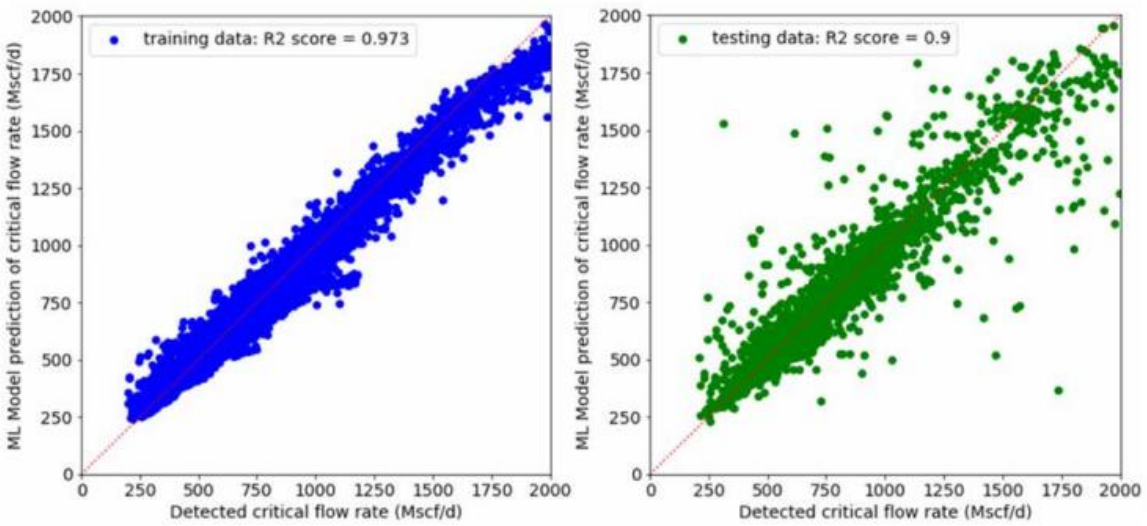


Figure 3.19 : Comparaison des prédictions du modèle XGBoost (données d'entraînement et de test) avec la vérité terrain.

Comparaison avec la corrélation Belfroid

La Figure 3.20 (gauche) montre les prédictions du modèle de XGBoost par rapport à la réalité (points de débit critique détectés), tandis que la Figure 3.20 (droite) montre les résultats de l'équation de Belfroid. XGBoost et Belfroid ont respectivement une erreur quadratique moyenne (MSE) de 2750,7 et 41291,7. Comme on peut le constater, l'exactitude du modèle XGBoost est nettement supérieure à celle du modèle Belfroid (qui s'est avéré être la corrélation la plus précise parmi les équations empiriques mentionnées dans cette étude).

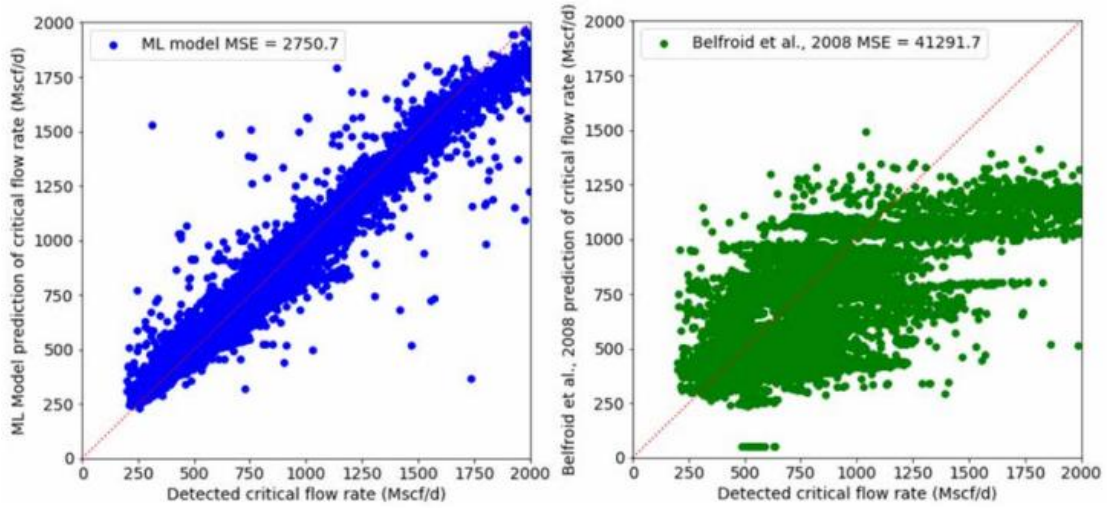


Figure 3.20 : Comparaison des prédictions du modèle XGBoost et de Belfroid avec la réalité.

Les métriques de précision du modèle d'apprentissage automatique XGBoost et de Belfroid sont répertoriées dans le Tableau 3.8. La Figure 3.21 montre l'histogramme des différences entre les débits critiques détectés et les résultats du modèle d'apprentissage automatique/Belfroid. Les résultats montrent que les résultats du modèle d'apprentissage automatique présentent une corrélation beaucoup plus forte avec les débits critiques trouvés.

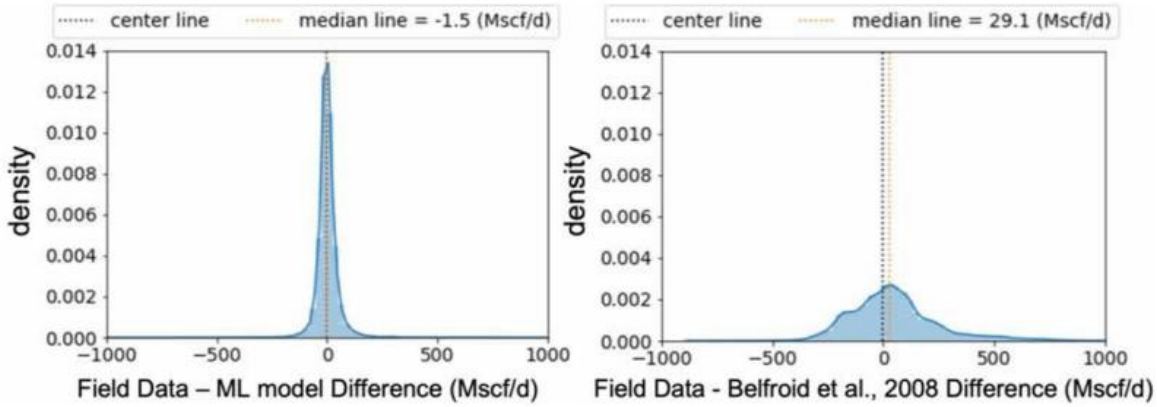


Figure 3.21 : L'histogramme des différences entre les débits critiques trouvés et les modèles ML/Belfroid.

Tableau 3.8 : Précision de corrélation du modèle ML et de Belfroid.

Modèle	MSE	RMSE	R^2
Belfroid	41291.7	203.2	0.3684
XGBoost	2750.7	52.45	0.9579

3.4 Évaluation

En utilisant l'apprentissage automatique, on a pu avoir plus de précision en comparant avec les méthodes classiques. D'où l'importance de cette discipline pour la résolution de ce problème.

Chaque approche a un concept différent, on peut résumer le concept de chaque approche dans le Tableau 3.9 ci-dessous :

Tableau 3.9 : Concept de chaque approche.

	Approche I	Approche II	Approche III
Supervision	✔		✔
Hybride	✔		

Caractéristiques des approches :

Le Tableau 3.10 résume les caractéristiques de chaque approche :

Tableau 3.10 : Caractéristiques de chaque approche.

Approche	Avantages	Inconvenients
Approche I	- Implémentation simple, rapide en temps d'exécution et grande puissance de calcul. - Efficacité par rapport à la corrélation du Turner, malgré la dépendance.	- Considération implicite que les critères de Turner sont valides dans les réservoirs non conventionnels.
Approche II	- Elimination de la dépendance aux critères de Turner. - Plus précise que l'approche I.	- Possibilité d'avoir des résultats très différents, vu la nature des algorithmes non supervisés. - Nécessité d'une réduction de dimensionnalité pour avoir de meilleurs résultats, ce qui mène à une perte d'information.
Approche III	- Haute précision. - Indépendance des corrélation empiriques.	- Exigence en termes de données d'entraînement. - Implementation difficile.

Conclusion

Ce chapitre a permis d'explorer en profondeur notre problématique du chargement de liquide et d'établir les bases d'une solution basée sur l'apprentissage automatique. La simulation des données grâce au OLGA, suivi d'un prétraitement rigoureux des données et d'une modélisation appropriée ont servi à l'élaboration de cette solution.

Les approches supervisée et non supervisée employées ont révélé des facettes différentes de la problématique, enrichissant notre compréhension et éclairant des opportunités pour l'amélioration future.

En conclusion, ce chapitre a permis d'établir une comparaison constructive entre les différentes approches appliquées pour la détection et la prédiction du chargement de liquide. L'analyse comparative a permis de mettre en évidence les forces et les potentialités de chaque approche, enrichissant ainsi notre compréhension du problème. Les résultats obtenus soulignent l'importance de la recherche continue dans ce domaine, tout en illustrant le potentiel de l'apprentissage automatique comme outil essentiel pour résoudre des problèmes complexes dans l'industrie du pétrole et du gaz.

Conclusion générale

En conclusion, ce mémoire a mis en évidence l'application prometteuse des techniques d'apprentissage automatique dans le domaine de la production de puits de gaz, en se concentrant sur l'entreprise SLB Ltd. L'objectif principal était de développer un outil prédictif basé sur l'apprentissage automatique pour prédire le chargement de liquide dans les puits de gaz, afin d'optimiser la production et de réduire les coûts de maintenance.

Nous avons exploré différentes approches de l'apprentissage automatique, telles que l'apprentissage supervisé, non supervisé, semi-supervisé et par renforcement, ainsi que l'apprentissage profond et génératif. Ces approches ont démontré leur capacité à traiter des problèmes complexes et à extraire des informations précieuses à partir des données disponibles.

En comprenant les principes fondamentaux de la production de puits de gaz, nous avons identifié les facteurs clés tels que le profil de production et les sources de liquides pouvant contribuer au chargement de liquide. Une revue bibliographique nous a permis de prendre connaissance des modèles existants utilisés pour prédire la production de puits de gaz et d'en évaluer l'efficacité.

Nous avons également présenté l'entreprise SLB Ltd., son domaine d'activité dans l'industrie pétrolière, ainsi que sa structure organisationnelle et sa division Digital & Integration. Ces informations nous ont permis de contextualiser notre projet dans le cadre des objectifs de digitalisation de SLB Ltd. et de l'industrie 4.0.

La conception de la solution proposée a été détaillée, en mettant l'accent sur les étapes essentielles telles que la collecte, la simulation, la compréhension et le prétraitement des données. Trois approches différentes ont été proposées, chacune visant à atteindre notre objectif principal de développement d'un projet d'apprentissage automatique pour la prédiction du chargement de liquide chez SLB Ltd.

En intégrant l'apprentissage automatique dans l'industrie pétrolière et gazière, nous avons constaté son potentiel pour améliorer l'efficacité opérationnelle, optimiser la production et réduire les coûts. Cependant, il est important de continuer à développer des approches plus avancées, telles que les modèles de deep learning et les modèles ensemblistes, et d'intégrer davantage de données pour des prédictions encore plus précises.

En conclusion, ce mémoire a démontré que l'application des techniques d'apprentissage automatique dans la détection et la prédiction du chargement de liquide dans les puits de gaz présente un potentiel significatif pour améliorer l'efficacité de l'industrie pétrolière et gazière. Ces avancées s'inscrivent dans la vision de SLB Ltd. de digitalisation et d'adoption des technologies de l'industrie 4.0. En exploitant pleinement ces opportunités, SLB Ltd. peut accélérer sa transformation numérique, optimiser ses opérations de production de puits de gaz et renforcer sa compétitivité sur le marché mondial des services pétroliers.

Bibliographie

Adriana Molinari. Understanding basics liquid loading, [en ligne], 13/05/2019. understanding-basics-liquid-loading-adriana-hernandez-1e.

Alexander Linden. Is Synthetic Data the Future of AI? . gartner.com. [En Ligne]. 22/06/2022. Is Synthetic Data the Future of AI? (gartner.com).

Avi Bewtra. The Ultimate Guide to Semi-Supervised Learning. V7labs.com. [En Ligne]. 01/07/2021. Semi-Supervised Learning: Techniques & Examples [2023] (v7labs.com).

Binli, Özmen. Overview of solutions to prevent liquid loading problems in gas wells. MS thesis. Middle East Technical University, 2009.

Coleman, Steve B., et al. "A new look at predicting gas-well load-up." Journal of petroleum technology 43.03 (1991): 329-333.

Eissa, M. Al-Safran, and P. Brill James. Applied Multiphase Flow in Pipes and Flow Assurance : Oil and Gas Production, SPE, 2017. ProQuest Ebook Central. Pages 1-13.

F. Martínez-Plumed et al. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories." in IEEE Transactions on Knowledge and Data Engineering, vol. 33. no. 8. pp. 3048-3061. 1 Aug. 2021. doi: 10.1109/TKDE.2019.2962680. Trajectories | IEEE Journals & Magazine | IEEE Xplore.

Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.". 2022. ISBN: 9781098125974.

Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006 Jul;18(7):1527-54. doi: 10.1162/neco.2006.18.7.1527. PMID: 16764513.cs.toronto.edu/~hinton/absps/ncfast.pdf.

Hesterberg, Tim. "Bootstrap." Wiley Interdisciplinary Reviews: Computational Statistics 3.6 (2011): 497-526.

Höök, Mikael. Depletion and decline curve analysis in crude oil production. Diss. Global Energy Systems, Department for Physics and Astronomy, Uppsala University, 2009.

Ibis World. Global Oil & Gas Exploration & Production Industry - Market Research Report.ibisworld.com.[En Ligne].30/03/2020 https://www.ibisworld.com/global/market-research-reports/global-oil-gas-exploration-production-industry/.

IBM 2023. What is unsupervised learning?. Ibm.com. [En Ligne]. 2023. What is Unsupervised Learning? | IBM.

International Energy Agency. Oil Market Report - April 2023. Iea.org. [En Ligne]. 04/2023. https://www.iea.org/reports/oil-market-report-april-2023.

Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

Lea, James-F., and Henry-V. Nickens. "Use of foam to deliquify gas wells, Gas Well Deliquification." (2008): 193-240.

Le Point Afrique. Pétrole : Alger a le blues. lepoint.fr. [En Ligne].19/01/2021. Pétrole : Alger a le blues (lepoint.fr).

Lilian Weng. Learning with not Enough Data Part 1: Semi-Supervised Learning. Github.io. [En Ligne]. 05/12/2021. Learning with not Enough Data Part 1: Semi-Supervised Learning | Lil'Log (lilianweng.github.io).

Lilian Weng. Flow-based Deep Generative Models. Github.io. 13/10/2018. Flow-based Deep Generative Models | Lil'Log (lilianweng.github.io).

Mehdi Mohammadpoor et Farshid Torabi. Big Data analytics in oil and gas industry: An emerging trend. Petroleum. Volume 6. Issue 4,2020. Pages 321-328. ISSN 2405-6561. https://doi.org/10.1016/j.petlm.2018.11.001.

Nallaparaju, Yashaswini Devi. "Prediction of liquid loading in gas wells." SPE Annual Technical Conference and Exhibition. OnePetro, 2012.

Olivier Le Peuch. Le Peuch Speaks at Cowen 2020 Energy Conference.slb.com.[En Ligne].12/03/2020. Le Peuch Speaks at Cowen 2020 Energy Conference | SLB.

Organization of the Petroleum Exporting Countries. Brief History. OPEC.org. [En Ligne]. 2020. https://www.opec.org/opec_web/en/about_us/24.html.

Radio Algerie. Les hydrocarbures en Algérie par les chiffres. radioalgerie.dz. [En Ligne]. 23/02/2021. Les hydrocarbures en Algérie par les chiffres | Radio Algérienne (radioalgerie.dz).

Rao, Bharath. "Designing coiled tubing velocity strings." CTES, LC www. ctes. com (1999).

Sarah El Shatby. How to Become a Data Scientist in the Oil and Gas Industry. 365datascience.com. [En Ligne]. 02/06/2023. https://365datascience.com/career-advice/how-to-become-a-data-scientist-in-the-oil-and-gas-industry/#1

SLB 2023. Our values. Slb.com. [En Ligne]. 2023. https://www.slb.com/about/who-we-are/our-values.

SLB 2023. Our history. Slb.com. [En Ligne]. 2023. https://www.slb.com/about/who-we-are/our-history.

Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." International Conference on Machine Learning. PMLR, 2015.

The Guardian.Revealed: oil sector's 'staggering' $3bn-a-day profits for last 50 years. The guardian.com. [En Ligne]. 21/07/2022. https://www.theguardian.com/environment/2022/jul/21/revealed-oil-sectors-staggering-profits-last-50-years.

Tristan Gaudiaut. Pétrole : trois pays assurent plus de 40 % de la production mondiale. Statista.com. [En Ligne]. 08/12/2022. https://fr.statista.com/infographie/19382/plus-gros-producteurs-de-petrole-brut-dans-le-monde/.

Turner, R. G., M. G. Hubbard, and A. E. Dukler. "Analysis and prediction of minimum flow rate for the continuous removal of liquids from gas wells." Journal of Petroleum technology 21.11 (1969): 1475-1482.

Van Nimwegen, Andreas Teunis. "The effect of surfactants on gas-liquid pipe flows." (2015).

Will Kenton. Monte Carlo Simulation: History, How it Works, and 4 Key Steps. investopedia.com.[En Ligne]. 26/03/2023. Monte Carlo Simulation: History, How it Works, and 4 Key Steps (investopedia.com).

Xu, L. Skoularidou M. Cuesta-Infante, A. & Veeramachaneni, K. . Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems. 2019. vol. 32. 1907.00503.pdf (arxiv.org).

Thèses

Hamour, M. A., & Benhamdine, N. M. Prédiction du Churn Rate Par le Machine Learning dans le secteur des M&A : application : KPMG . ENP. 2020.

Lamdani, W., & Hadjaz, M. R. Mise en place d'un outil d'aide à la décision pour l'optimisation de la configuration du réseau des bases logistiques dans l'industrie pétrolière : application : SLB NAF . ENP. 2021

Souames, M.A., & Mohammedi, L. A. Estimation des lead times liés à l'importation à travers l'apprentissage machine dans le cadre de la méthodologie CRISP-DM: application: SLB NAF . ENP. 2022

Yousfi, M., & BAKHOUCHE, Y. Contribution a l'optimisation de la planification de la demande en pièces de rechange au sein de la division Well Construction: application: SLB NAF . ENP. 2022

Annexes

Présentation des outils utilisés

Python est un langage de programmation interprété, orienté objet, de haut niveau et doté d'une sémantique dynamique. Ses structures de données intégrées de haut niveau, combinées au typage dynamique et à la liaison dynamique, le rendent très attrayant et polyvalent qui peut être utilisé dans de nombreux domaines, tels que le développement Web, l'analyse de données, l'apprentissage profond et de l'intelligence artificielle. Python supporte les modules et les packages, ce qui encourage la modularité des programmes et la réutilisation du code. L'interpréteur Python et sa vaste bibliothèque standard sont disponibles gratuitement pour toutes les principales plates-formes et peuvent être distribués librement.

NumPy, qui signifie "Numerical Python", est une bibliothèque open-source pour le langage de programmation Python. Elle fournit un support pour des tableaux multidimensionnels, des matrices de grande taille, ainsi que pour un large éventail de fonctions mathématiques de haut niveau pour opérer sur ces tableaux. Essentiellement utilisée pour la manipulation numérique des données. Elle offre des opérations de calcul de base telles que l'addition, la soustraction, la multiplication et la division, ainsi que des opérations plus complexes comme les transformations de Fourier, l'algèbre linéaire, et la génération de nombres aléatoires. Cette bibliothèque est à la base de nombreuses autres bibliothèques de data science et d'apprentissage automatique, comme Pandas pour la manipulation des données et Scikit-Learn pour l'apprentissage automatique. NumPy est souvent utilisée dans les domaines de l'apprentissage automatique, de la science des données, de l'analyse des données, et de la visualisation des données.

Pandas est une bibliothèque logicielle open source pour le langage de programmation Python qui fournit des structures de données et des outils d'analyse de données flexibles, rapides et efficaces. Le nom Pandas est dérivé du terme "panel data", un terme économique pour les ensembles de données structurés.La bibliothèque Pandas introduit deux nouvelles structures de données à Python - DataFrame et Series, qui permettent de manipuler les données avec une facilité d'opération qui n'était pas disponible auparavant dans Python. Pandas est particulièrement bien adapté pour de nombreuses tâches de données différentes et est souvent utilisé avec d'autres bibliothèques comme NumPy et Matplotlib pour une analyse de données plus complète. Elle est largement utilisée en science des données et en apprentissage automatique.

Scikit-learn est une bibliothèque open source pour le langage de programmation Python qui fournit une gamme d'outils d'apprentissage automatique supervisé et non supervisé. Elle est construite sur deux autres bibliothèques Python, NumPy et SciPy, et s'intègre bien avec d'autres bibliothèques de la pile scientifique Python, comme Pandas et Matplotlib. Scikit-learn offre une interface uniforme pour de nombreux algorithmes d'apprentissage automatique, ce qui facilite leur utilisation et leur comparaison. Les types d'algorithmes incluent la classification, la régression, le clustering, la réduction de dimensionnalité, l'estimation de densité, et bien d'autres. De plus, Scikit-learn comprend des outils pour le prétraitement des données, la sélection et l'évaluation des modèles, et le tuning des hyperparamètres. La bibliothèque est largement utilisée en science des données et en apprentissage automatique, en raison de sa flexibilité, de sa facilité d'utilisation, et du fait qu'elle est soutenue par une vaste communauté de développeurs et d'utilisateurs.

Dataiku est une plateforme d'analyse de données et de science des données qui aide les organisations à construire et déployer leurs propres solutions d'analyse et d'apprentissage automatique. La plateforme est conçue pour faciliter la collaboration entre les analystes de données, les ingénieurs et les scientifiques de données, et pour aider les entreprises à passer de la découverte de données à la production de modèles prédictifs. Dataiku propose un environnement de travail basé sur l'interface utilisateur graphique (GUI) ainsi que la possibilité de coder en Python, R, SQL et d'autres langages populaires. Il permet le nettoyage des données, l'exploration de données, la visualisation, le développement de modèles, le déploiement de modèles, et le suivi de la performance des modèles. En outre, Dataiku offre une variété de fonctionnalités pour le travail en équipe, comme la possibilité de travailler sur des projets en parallèle, de partager des projets et des modèles, et de contrôler les versions des projets. La plateforme est largement utilisée dans une variété d'industries pour développer et déployer rapidement des solutions d'apprentissage automatique et d'IA.

OLGA est un logiciel de simulation avancé utilisé dans l'industrie du pétrole et du gaz pour modéliser et simuler les flux multiphasiques (c'est-à-dire des flux composés de plusieurs phases distinctes, comme le pétrole, le gaz et l'eau) dans les pipelines, développé par SLB. Le simulateur OLGA est utilisé pour analyser à la fois les opérations régulières et transitoires, telles que le démarrage et l'arrêt des puits, les changements de régime, les dépressurisations et les coupures de courant. Cela aide les ingénieurs à optimiser la production et le transport des hydrocarbures, à augmenter la sécurité opérationnelle, et à réduire les coûts de production et de maintenance. OLGA offre des fonctionnalités pour modéliser divers phénomènes dynamiques, tels que le slugging (les fluctuations de débit dues à l'accumulation et à l'expulsion de liquide), la formation et le transport de dépôts, et les effets thermiques.

Concepts liés à la Data Science et à l'apprentissage automatique

Validation croisée

La validation croisée, également connue sous le nom de cross-validation en anglais, est une technique utilisée en apprentissage automatique pour évaluer la capacité d'un modèle à généraliser à partir de données d'entraînement à des données inédites. Elle est particulièrement utile pour éviter le surapprentissage, qui se produit lorsque le modèle s'adapte trop bien aux données d'entraînement et fonctionne mal sur de nouvelles données.

La validation croisée implique de diviser l'ensemble de données en deux sous-ensembles : l'un pour l'entraînement du modèle et l'autre pour le tester. Cette opération est répétée plusieurs fois en sélectionnant différentes parties de l'ensemble de données pour l'entraînement et le test à chaque fois. La performance moyenne du modèle sur l'ensemble de test au cours de ces itérations donne une estimation de la capacité du modèle à généraliser à partir de nouvelles données.

Il existe plusieurs méthodes de validation croisée, y compris :

La validation croisée de K-Fold : L'ensemble de données est divisé en K sous-ensembles. À chaque fois, un des sous-ensembles est utilisé comme ensemble de test et les autres sous-ensembles sont utilisés comme ensemble d'entraînement. Ce processus est répété K fois, chaque sous-ensemble étant utilisé une fois comme ensemble de test.

La validation croisée Leave-One-Out (LOO) : C'est un cas particulier de la validation croisée de K-Fold où K est égal au nombre d'observations dans l'ensemble de données. C'est-à-dire qu'à chaque itération, une seule observation est utilisée comme ensemble de test et toutes les autres observations sont utilisées comme ensemble d'entraînement.

La validation croisée stratifiée : Elle est similaire à la validation croisée de K-Fold, mais elle est utilisée lorsque les classes de la variable cible sont déséquilibrées. Elle garantit que chaque pli a la même proportion de classes de la variable cible que l'ensemble de données original.

La validation croisée aide à sélectionner le modèle qui performe le mieux en termes de capacité de prédiction sur de nouvelles données.

Figure B.1 : Concept de la validation croisée.

Bootstrap

Le bootstrap est une technique statistique puissante qui consiste à créer des échantillons de taille "n" à partir d'un ensemble de données initial également de taille "n", mais avec remise. C'est-à-dire que chaque fois qu'un échantillon est sélectionné, il est remis dans l'ensemble de données initial, ce qui lui donne une chance d'être sélectionné à nouveau.

Ces échantillons de données bootstrappés sont ensuite utilisés pour estimer divers paramètres statistiques (moyenne, variance, intervalles de confiance, etc.) du modèle ou de la population. En effet, le bootstrap offre une méthode pour quantifier l'incertitude associée à une estimation particulière.

Cette technique est particulièrement utile lorsque la distribution théorique de l'estimation est difficile à déterminer. Elle est largement utilisée dans de nombreux domaines, dont la biostatistique, l'économétrie, la science des données, l'apprentissage automatique et bien d'autres.

Il existe plusieurs variantes du bootstrap, telles que le bootstrap paramétrique, non-paramétrique, bootstrap lisse, bootstrap sauvage, etc., qui sont adaptées à différents scénarios ou types de données.

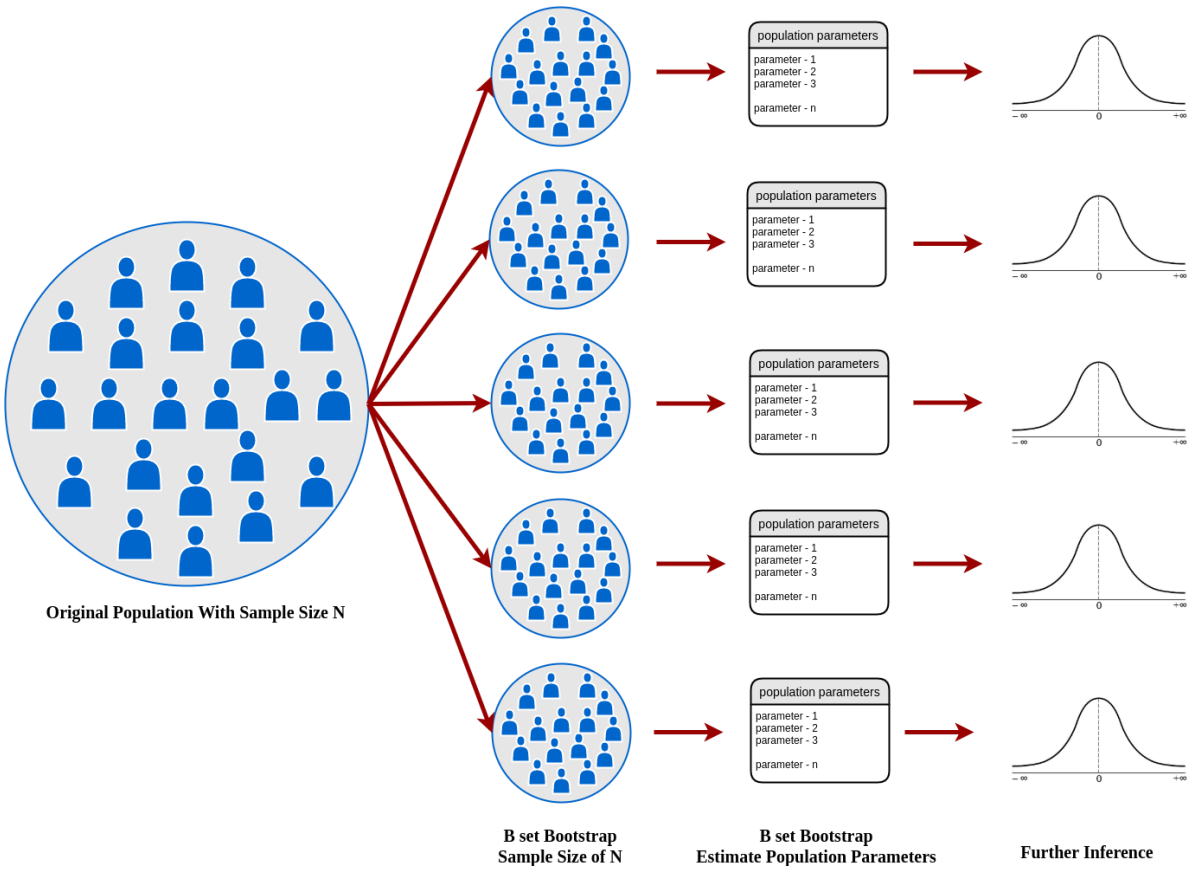


Figure B.2 : Concept de la technique bootstrap.

Apprentissage ensembliste

L'apprentissage ensembliste, aussi appelé "Ensemble Learning" en anglais, est une stratégie d'apprentissage automatique où plusieurs modèles d'apprentissage (appelés "apprenants de base") sont formés pour résoudre le même problème et combinés de manière à obtenir de meilleures performances.

Le principe fondamental de l'apprentissage ensembliste est que plusieurs modèles faibles, combinés ensemble, peuvent former un modèle fort. Les techniques ensemblistes visent à réduire le biais (la sous-estimation des vrais paramètres de modèle) et la variance (l'erreur due à la sensibilité aux petites fluctuations dans l'ensemble de formation) des prédictions.

Les méthodes ensemblistes populaires comprennent :

Bagging : Chaque apprenant de base est formé sur un sous-ensemble distinct de l'ensemble de données d'origine, généralement sélectionné avec remplacement (c'est-à-dire un bootstrap). Les prédictions de tous les apprenants sont ensuite moyennées (pour les problèmes de régression) ou votées (pour les problèmes de classification). Un exemple de technique de bagging est le Random Forest.

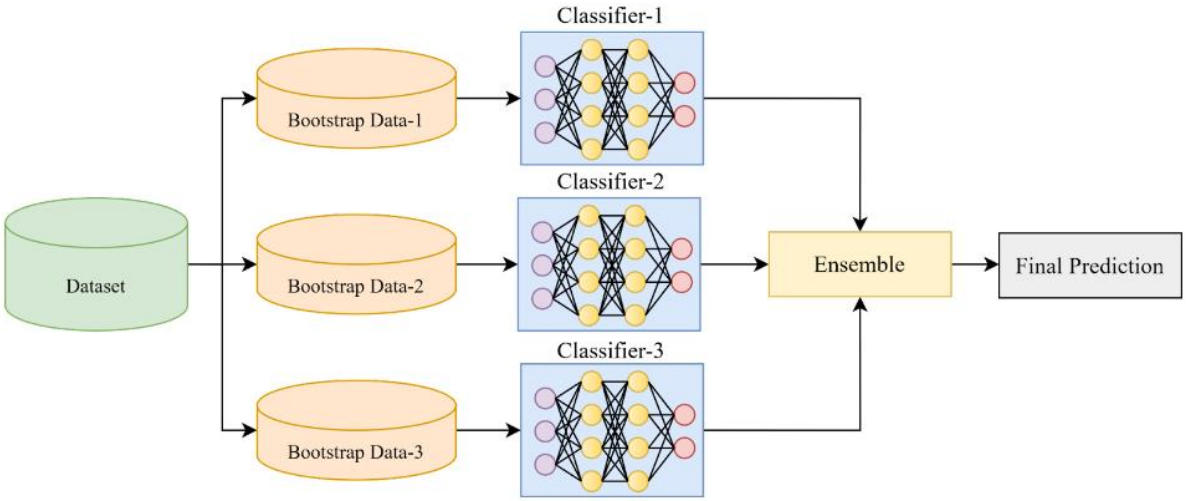


Figure B.3 : Bagging.

Boosting : Les apprenants sont formés de manière séquentielle, chaque nouvel apprenant étant conçu pour corriger les erreurs de l'ensemble précédent. Les prédictions de tous les apprenants sont ensuite combinées par un vote pondéré pour produire la prédiction finale. Des exemples de techniques de boosting comprennent AdaBoost et Gradient Boosting.

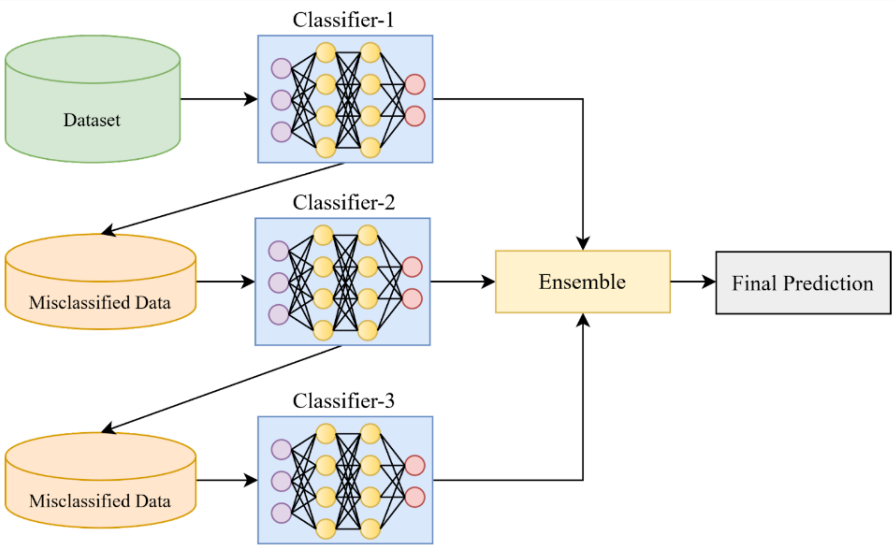


Figure B.4 : Boosting.

Stacking : Les apprenants de base sont formés parallèlement (comme dans le bagging) mais les prédictions sont combinées par un autre modèle d'apprentissage (appelé le métamodèle ou modèle de second niveau), qui est formé pour prédire la sortie finale à partir des prédictions des apprenants de base.

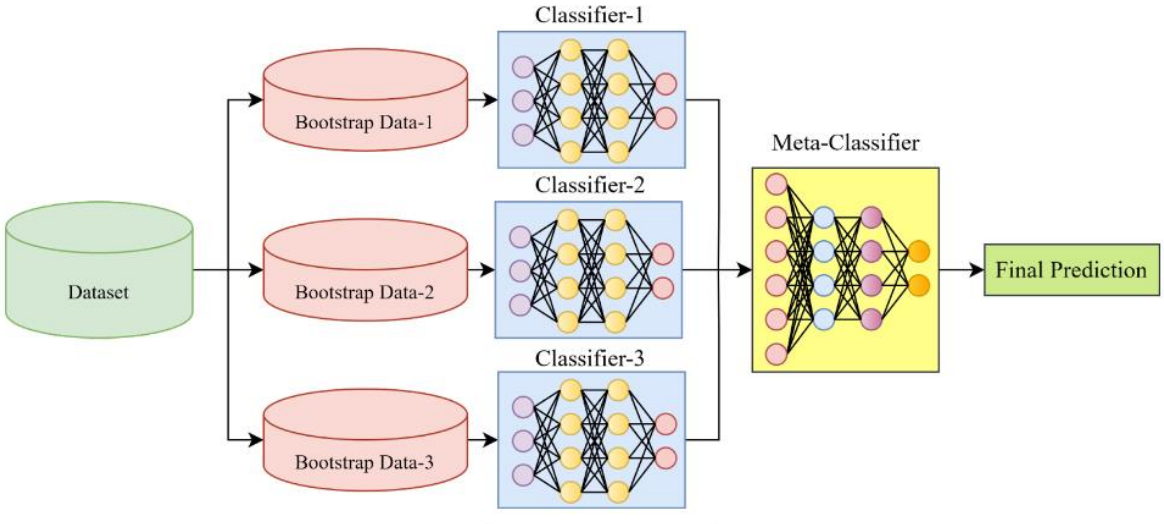


Figure B.5 : Stacking.

Ces méthodes aident à améliorer la précision et la robustesse du modèle, ainsi qu'à prévenir le surapprentissage.

Surapprentissage

Le surapprentissage, ou "overfitting" en anglais, est un concept clé en apprentissage automatique et en statistiques. Cela se produit lorsqu'un modèle apprend trop bien les données d'entraînement, à tel point qu'il capture non seulement les tendances générales, mais aussi le bruit et les anomalies spécifiques à cet ensemble de données.

En d'autres termes, un modèle surajusté est un modèle qui est trop complexe pour les données disponibles. Il a appris le "bruit" des données d'entraînement plutôt que les relations sous-jacentes. Par conséquent, il performe très bien sur les données d'entraînement mais mal sur les nouvelles données inconnues ou les données de test, car il ne généralise pas bien à partir de ce qu'il a appris.

Un modèle surajusté peut être le résultat de :

- Une durée d'entraînement trop longue
- Un modèle trop complexe (par exemple, un réseau de neurones avec trop de couches ou de neurones)
- Un manque de données d'entraînement
- Des données d'entraînement trop bruitées ou présentant trop d'anomalies

Pour éviter le surapprentissage, plusieurs techniques peuvent être utilisées, telles que:

- La validation croisée (Cross-validation) pour estimer l'erreur de généralisation du modèle
- Le dropout ou la régularisation (L1, L2) pour contraindre la complexité du modèle
- L'augmentation des données (Data augmentation) pour fournir plus de variations dans les données d'entraînement
- L'arrêt précoce (Early stopping) pour arrêter l'entraînement lorsque la performance commence à se dégrader sur un ensemble de validation.

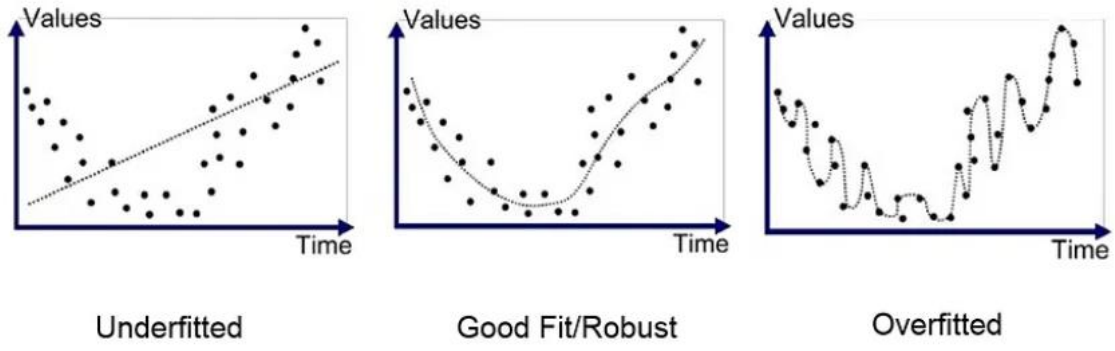


Figure B.6 : Surapprentissage.

Métriques d’évaluation

Métriques des courbes

Receiver Operating Characteristics Curve (ROC Curve)

La courbe ROC est une manière visuelle de représenter les performances d'un classificateur. On trace donc une courbe qui représente l'évolution du rappel (taux de vrais positifs) aussi appelé True Positive Rate (TPR) en fonction de 1 - spécificité (taux de faux positifs) aussi appelé False Positive Rate (FPR) qu'on définit par la loi suivante :

$$FPR = \frac{FP}{FP+TN}$$

- **FP:** False Positives.
- **TN:** True Negatives.

La courbe ROC trace les valeurs du TPR et du FPR pour différents seuils S de classification.
Diminuer la valeur du seuil de classification permet de classer plus d'éléments comme positifs, ce qui augmente le nombre de faux positifs et de vrais positifs. On peut visualiser une courbe ROC par l'illustration suivante :

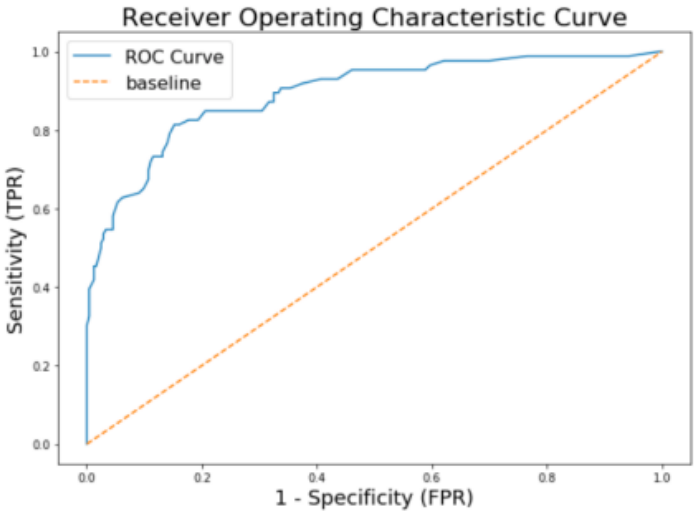


Figure C.1 : Courbe ROC.

Aire sous la courbe ROC (Area Under Curve AUC)

AUC signifie la courbe sous ROC. Cette valeur mesure l'intégralité de l'aire, à deux dimensions, située sous l'ensemble de la courbe ROC (par calcul d'intégrales) du point (0,0) à (1,1) défini par la fonction $f(x) = x$.

Cette mesure permet de quantifier le degré de séparabilité, en indiquant à quel point le modèle est capable de faire la distinction entre les classes. Plus l'AUC est importante, plus le modèle est prompt à prédire les positifs étant positifs et négatifs étant négatifs.

Concrètement parlant, l'AUC présente les avantages suivants :

• L'AUC est invariante d'échelle, elle mesure donc la qualité de la classification des prédictions, plutôt que leurs valeurs absolues

• L'AUC est indépendante des seuils de classification

Nous pouvons voir dans la Figure .2 une représentation de différentes classifications et leurs résultats en comparant leurs courbes ROC respectives, on peut s'apercevoir que plus la courbe de ROC est proche du coin en haut à gauche, plus l'AUC est important et la séparation des classes est fiable. On s'aperçoit également qu'une classification aléatoire donne des résultats qui sont sur la courbe de la première bissectrice du plan : $f(x) = x$.

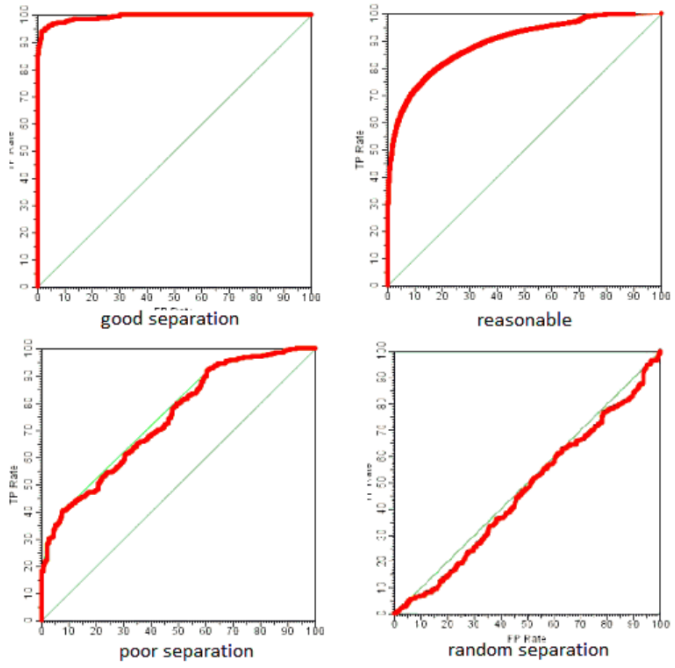


Figure C.2 : Visualisation de la disparité des modèles de classifications selon leur courbe ROC.

Métriques supervisées

Le coefficient de détermination R²

Le coefficient de détermination, noté R², est une statistique qui indique la proportion de la variance dans la variable dépendante qui est prévisible à partir des variables indépendantes dans un modèle de régression. R² est une mesure de la qualité de l'ajustement d'un modèle de régression et varie entre 0 et 1.

L'équation pour le calcul du coefficient R² est :

$$R^2 = 1 - \frac{SSres}{SStot}$$

où:

- SSres est la somme des carrés des résidus, c'est-à-dire la somme des carrés des différences entre les valeurs observées et les valeurs prédites par le modèle.
- SStot est la somme totale des carrés, qui est la somme des carrés des différences entre les valeurs observées et la moyenne des valeurs observées.

Un R² de 1 indique que le modèle de régression prédit parfaitement la variable dépendante. Un R² de 0 indique que le modèle ne prédit pas du tout la variable dépendante.

Métriques non supervisées

Coefficient de silhouette

Le coefficient de silhouette est une mesure utilisée pour évaluer la qualité de l'assignation d'un point à un cluster dans un algorithme de clustering. La valeur du coefficient de silhouette varie de -1 à 1. Une valeur proche de 1 indique que le point est bien placé dans son cluster, tandis qu'une valeur proche de -1 indique que le point est mal placé et devrait probablement être placé dans un autre cluster. Une valeur proche de 0 signifie que le point est à la frontière entre deux clusters.

Le coefficient de silhouette pour un échantillon est calculé comme suit :

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

où:

- a(i) est la distance moyenne entre l'échantillon i et tous les autres points du même cluster.
- b(i) est la distance moyenne entre l'échantillon i et tous les points du cluster le plus proche (autre que le cluster auquel i appartient).

C'est-à-dire que le coefficient de silhouette mesure à quel point un objet est similaire à son propre cluster (cohésion) comparé à d'autres clusters (séparation). Les valeurs élevées du coefficient de silhouette indiquent que l'objet est bien assorti à son propre cluster et mal assorti aux clusters voisins.

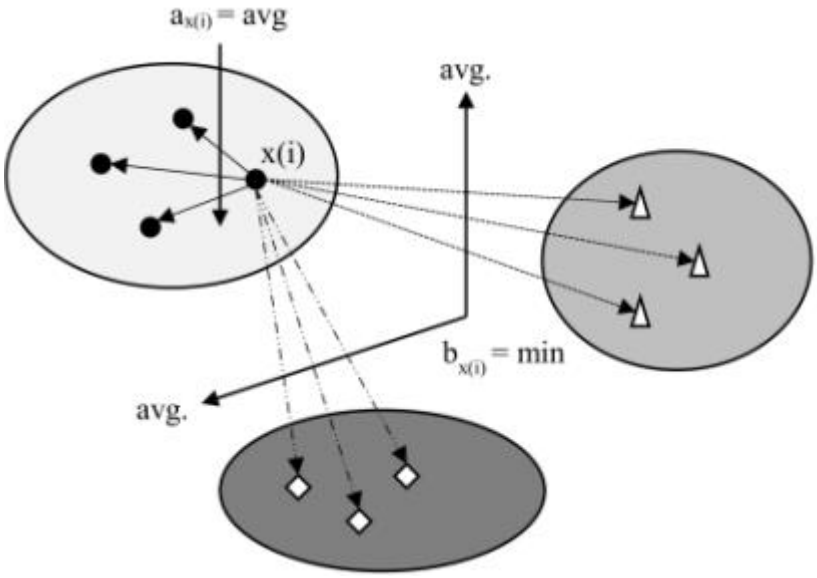


Figure C.3 : Concept du coefficient de silhouette.

Apprentissage automatique automatisé (Auto Machine Learning)

AutoML ou **Apprentissage Automatique Automatisé**, est un domaine d'étude de l'apprentissage automatique qui vise à automatiser les processus complexes et chronophages associés à la conception et l'implémentation de modèles d'apprentissage automatique.

AutoML couvre un certain nombre de tâches importantes, notamment :

Prétraitement des données : AutoML peut aider à nettoyer et à normaliser les données, à gérer les valeurs manquantes et à effectuer une ingénierie des caractéristiques automatisée.

Sélection de modèles : AutoML peut comparer différents types de modèles d'apprentissage automatique pour déterminer lequel est le plus performant pour un ensemble de données donné.

Optimisation des hyperparamètres : AutoML utilise des techniques comme la recherche par grille et la recherche bayésienne pour optimiser automatiquement les hyperparamètres des modèles, ce qui peut améliorer considérablement les performances des modèles.

Analyse et interprétation de modèles : Certains outils AutoML offrent également des fonctionnalités pour aider à comprendre et interpréter les modèles d'apprentissage automatique, ce qui est crucial pour la validation de modèle et l'acceptation par les parties prenantes.

AutoML vise à rendre l'apprentissage automatique plus accessible à ceux qui ne sont pas des experts en la matière, et à augmenter l'efficacité des experts en apprentissage automatique en leur permettant de se concentrer sur des problèmes plus complexes.

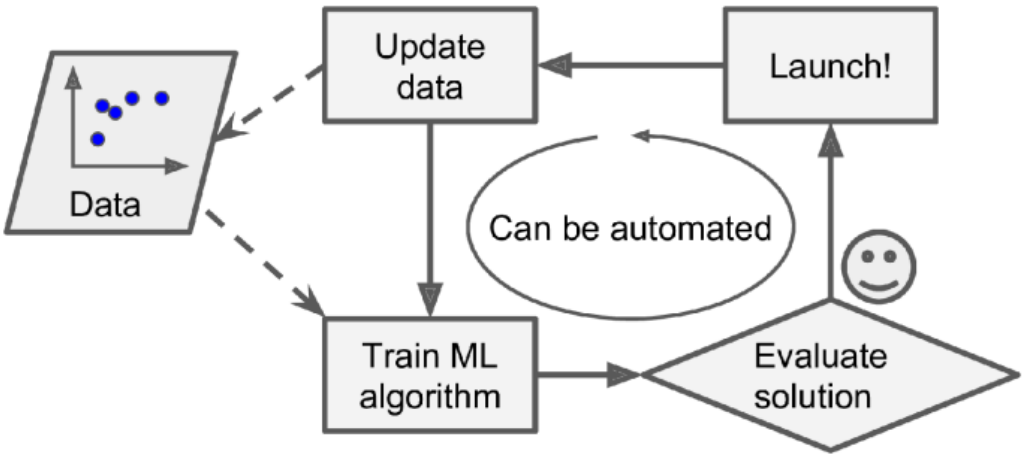


Figure D.1 : Concept du Auto ML (Géron Aurélien, 2022).