

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA
RECHERCHE SCIENTIFIQUE

ÉCOLE NATIONALE POLYTECHNIQUE



Département : Génie Industriel
Entreprise : ADFC × Sonatrach

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'ingénieur d'état en Génie Industriel
Option : Data Science & Intelligence Artificielle

**Création d'un modèle d'IA pour détecter les
anomalies dans la composition des fluides de forage
en analysant les données en temps réel**

Réalisé par :
Dounia KAMEL

Encadré par :
Mr Oussama ARKI
Mr Hacene MOUHOUB

Présenté et soutenu publiquement le (24/06/2025)

Composition du jury :

Président :	Dr. Hakim FOURAR-LAIDI	MCA	ENP
Examinatrice :	Dr. Samia BELDJOUDI	MCA	ENP
Promoteur :	Dr. Oussama ARKI	MCA	ENP

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA
RECHERCHE SCIENTIFIQUE

ÉCOLE NATIONALE POLYTECHNIQUE



Département : Génie Industriel
Entreprise : ADFC × Sonatrach

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'ingénieur d'état en Génie Industriel
Option : Data Science & Intelligence Artificielle

**Création d'un modèle d'IA pour détecter les
anomalies dans la composition des fluides de forage
en analysant les données en temps réel**

Réalisé par :
Dounia KAMEL

Encadré par :
Mr Oussama ARKI
Mr Hacene MOUHOUB

Présenté et soutenu publiquement le (24/06/2025)

Composition du jury :

Président :	Dr. Hakim FOURAR-LAIDI	MCA	ENP
Examinatrice :	Dr. Samia BELDJOUDI	MCA	ENP
Promoteur :	Dr. Oussama ARKI	MCA	ENP

الهداء

الحمد لله دُجًا وامتنانًا على البدء، وعلى الختام.
في لحظة سعادةٍ، وامتنانٍ، وفخرٍ، أهدي ثمرةً جهدي، وحصادَ سنين مضت في سبيل العلم، بكل ما فيها من تحديات ونجاحات، إلى من يستحقونها
بامتنياز:

إلى من لا ينفصل اسمه عن اسمي، إلى من علّمني أن الدنيا كفاح، وسلاحها العلم والمعرفة.
إلى من غرس في روحي مكارم الأخلاق، وكان ناصري الأول وسندي، وسعى لأجل راحتي ونجاحي.
إلى أعظم رجل في الكون، إلى فخري واعتزازي:
أبي الغالي.

إلى أنيسة الروح، وعطر العمر، إلى من كانت دعواتها النور الذي أنار دربي، والحنان الذي ضمّني، والملجأ الذي احتضنت به.
إلى من أفتقدها كل لحظة، وخاصة في هذا اليوم الغالي:
أمي الحبيبة، رحمك الله وجعل مثواك الجنة.

إلى السند الذي لا يميل، إلى الإخوة الذين كانوا لي قوة و عونًا،
وإلى **أختي أشواق**، نبع الحب والحكمة،
وإلى **أختي الصغيرة توبة**، ابنتي في الحنان، ونبضي في كل إنجاز.
إخوتي عادل و إيمان ، فخر عمري ودعمي الأبدي.

إلى **عمي ياسين وزوجته بدره وبناتهم العزيزات زينب، غزلان، حفصة، مريم و خديجة**، الذين كانوا لي بعد عائلتي السند والملاذ،

وإلى **خالتي الغالية بزة**، التي كانت دومًا بجانبني قلبًا ودعاءً.

إلى عائلتي التي تسكن قلبي، كل أفراد عائلة كامل و بن مبروك،
أنتم من رسمتم في مشواري دفاء الحب وأمان الانتماء.

إلى **صديقاتي**، رفيقات الأيام الثقيلة والخفيفة،
إلى من كنّ الأمان والفرح والكتف المشترك
إلى **زينة، بشرى، بلقيس**

إلى **أصدقائي (Dar djwaz#)**، من شاركتمهم أجمل اللحظات، ومضينا معًا في هذه الرحلة الرائعة. ما كانت لتكون بهذه الروعة لولا وجودكم.
أشكركم على ثلاث سنوات لا تُنسى، وأتمنى لكم مستقبلًا مشرقًا مليئًا بالنجاحات.

إلى **أساتذتي الكرام** الذين لم يخلوا عليّ بعلمهم وتوجيههم،
أقدّر كل لحظة تعلّمت فيها منكم، وسأظل ممتنة لذلك طوال حياتي.

وإلى كل من كان له أثر في مسيرتي، وإن لم يُذكر اسمه هنا،
أنتم جزء من هذا الإنجاز، فشكرًا لكم من القلب.

وأخيرًا...

إلى **نفسي**،

لكل التعب الذي صبرت عليه،
لكل لحظة خذلت فيها التعب وأكملت الطريق،
أهديك هذه الثمرة، فهنيئًا لك النصر الجميل.

Remerciements

Au terme de ce travail, je tiens à remercier Dieu Tout-Puissant de m'avoir donné la force, la patience et la persévérance nécessaires pour mener à bien ce projet de fin d'études.

Je souhaite exprimer ma profonde reconnaissance à mon encadrant, **Mr Oussama ARKI**, pour son accompagnement rigoureux, ses conseils précieux et sa disponibilité tout au long de la réalisation de ce mémoire. Sa bienveillance et son expertise ont été essentielles à l'avancement de ce travail.

Je remercie chaleureusement **l'ensemble de l'équipe de l'ADFC (Algerian Drilling Fluids Company)**, et en particulier **Mr Hacene MOUHOUB**, pour leur accueil, leur encadrement et les ressources mises à disposition. Leur confiance m'a permis de mener une mission stimulante en lien direct avec les enjeux industriels du secteur.

Je tiens à adresser mes sincères remerciements à **tous les enseignants du département de Génie Industriel de l'École Nationale Polytechnique**, et plus particulièrement à **l'ensemble des professeurs du parcours DSIA (Data Science et Intelligence Artificielle)**. Leur engagement, leurs enseignements de qualité et leur encadrement bienveillant ont grandement contribué à mon évolution académique et professionnelle.

Je souhaite également remercier chaleureusement **la grande famille du club IEC (Industrial Engineering Club)**, véritable source d'inspiration, de soutien et d'épanouissement tout au long de ma formation. Merci pour les expériences humaines et professionnelles partagées.

À mes camarades de promotion, et plus particulièrement à **ma classe, les Daita** – sans conteste la meilleure classe – je vous adresse toute mon affection et ma gratitude pour les moments de solidarité, de rires et de dépassement collectif qui ont marqué ces années.

Enfin, je remercie **ma famille** pour son amour inconditionnel, son soutien moral constant et sa confiance en moi à chaque étape de ce parcours.

Merci à toutes et à tous !

ملخص

هدف هذا العمل إلى دمج تقنيات الذكاء الاصطناعي في نظام مراقبة جودة سوائل الحفر البترولية داخل الشركة الوطنية لسوائل الحفر، التابعة لسوناطراك. تعتبر هذه السوائل عنصراً حاسماً لضمان استقرار الآبار وسلامة المعدات أثناء عمليات الحفر، إلا أن أساليب المراقبة الحالية تظل تقليدية، تعتمد على القياسات اليدوية والتحليل المخبرية، ما يجعل الاستجابة للأخطاء بطيئة وغير فعالة في كثير من الأحيان.

قمنا بتطوير نموذج تعلم آلي يعتمد على خوارزمية الغابة العشوائية لرصد الشذوذات تلقائياً باستخدام بيانات متعددة المصادر تشمل قراءات حساسات في الزمن الحقيقي، تقارير مختبرية وسجلات تاريخية. تم دمج هذا النموذج ضمن تطبيق تفاعلي باستخدام منصة ضترمليت، يتيح للمستخدمين تحميل البيانات، عرض النتائج، وإنشاء تقارير مفصلة تلقائياً.

أظهرت نتائج التجارب تحسناً ملحوظاً في دقة اكتشاف الشذوذات وسرعة التفاعل معها، مما يساهم في تحسين اتخاذ القرار الفني وتقليل المخاطر التشغيلية. تمثل هذه الحلول خطوة مهمة نحو التحول الرقمي في قطاع الطاقة وتعزيز المراقبة الذكية للعمليات البترولية

الكلمات المفتاحية: ذكاء صطناعي، اكتشاف الشذوذات، سوائل الحفر، التعلم الآلي، التحليل في الزمن الحقيقي، ضترمليت، الشركة الوطنية لسوائل الحفر، مراقبة الجودة

Abstract

This thesis explores the use of artificial intelligence to enhance the quality control of drilling fluids in the oil and gas sector, with a focus on ADFC, a subsidiary of Sonatrach. Drilling fluids play a crucial role in maintaining well stability, cooling equipment, and ensuring operational safety. However, the traditional monitoring methods are still highly manual and reactive, lacking real-time intelligence and automation.

To address this gap, we designed a machine learning-based anomaly detection model using the Random Forest algorithm, trained on multisource data including real-time sensor readings, historical logs, and laboratory reports. The model was deployed through an interactive Streamlit application that allows engineers to upload data files, monitor fluid parameters, visualize anomalies, and generate detailed reports.

The results show a substantial improvement in early anomaly detection, operational responsiveness, and technical decision-making. This solution provides a tangible step toward digital transformation in petroleum operations and lays the groundwork for smart, automated fluid surveillance systems.

Keywords : *Artificial Intelligence (AI), Anomaly Detection, Drilling Fluids, Random Forest, Machine Learning, Real-Time Analysis, Streamlit, ADFC, Quality Control.*

Résumé

Ce mémoire présente une étude approfondie sur l'intégration de l'intelligence artificielle dans le domaine du forage pétrolier, plus précisément dans le contrôle qualité des fluides de forage au sein de l'ADFC, filiale de Sonatrach. Les fluides de forage, essentiels à la stabilité des puits et à la sécurité des opérations, sont traditionnellement surveillés à l'aide de méthodes manuelles et de tests en laboratoire, souvent limités par le manque d'automatisation et la réactivité tardive.

Dans ce contexte, nous avons conçu et mis en œuvre une solution basée sur un modèle de Machine Learning, utilisant l'algorithme Random Forest, afin de détecter automatiquement les anomalies dans les propriétés des fluides à partir de données multi-sources (capteurs en temps réel, historiques, et analyses en laboratoire). L'approche est complétée par une application interactive développée sous Streamlit, permettant aux ingénieurs de charger des fichiers, visualiser les résultats et générer automatiquement des rapports d'anomalies.

Les résultats obtenus démontrent une amélioration significative en matière de réactivité, de fiabilité des prédictions et de support à la décision technique. Cette solution constitue une avancée concrète vers la digitalisation du secteur pétrolier et ouvre la voie à une surveillance intelligente et autonome des opérations de forage.

Mots clés : *Intelligence Artificielle (IA), Détection d'anomalies, Fluides de forage, Random Forest, Apprentissage automatique, Analyse en temps réel, Streamlit, ADFC, Contrôle qualité.*

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

Introduction générale 15

I. État des lieux 17

1 Présentation de l'Organisme d'Accueil 18

1.1 Introduction 18

1.2 Présentation générale de l'ADFC 18

1.3 Présentation de Sonatrach et de son lien avec l'ADFC . . . 19

1.4 Missions de l'ADFC 20

1.5 Vision de l'ADFC 20

1.6 Objectifs stratégiques de l'ADFC 21

1.7 Termes et jargon professionnels 21

1.8 Digitalisation des processus métiers et intégration technologique 23

1.9 Clients et Partenaires 23

1.10 Engagements qualité, sécurité et environnement 24

1.11	Conclusion	24
2	Étude de l'existant : le système actuel	25
2.1	Introduction	25
2.2	Les rapports et méthodes de contrôle qualité des fluides . .	25
2.2.1	Types de rapports produits	26
2.2.2	Sources des données utilisées	26
2.3	Le processus actuel de surveillance	27
2.4	Limites du système actuel	27
2.4.1	Faiblesses techniques	27
2.4.2	Manque d'automatisation	28
2.5	Identification des opportunités d'amélioration	28
2.5.1	Énoncé de la problématique	28
2.5.2	Axes d'amélioration envisagés	29
2.6	Conclusion	29
II.	État de l'art	30
3	Fluides de forage et anomalies associées	31
3.1	Introduction	31
3.2	Définition et rôle des fluides de forage	31
3.3	Propriétés physico-chimiques à surveiller	32
3.4	Types d'anomalies rencontrées : nature et classification . .	32
3.5	Conséquences des anomalies sur les opérations	33
3.6	Techniques traditionnelles de détection des anomalies . . .	34
3.7	Limites des méthodes conventionnelles	35
3.8	Conclusion	36

4	Intelligence Artificielle pour la détection d'anomalies	37
4.1	Introduction	37
4.2	Principes du Machine Learning appliqué à la détection d'anomalies	37
4.3	Techniques utilisées pour la détection d'anomalies	38
4.3.1	Random Forest	39
4.3.2	Isolation Forest	39
4.3.3	Autoencoders	39
4.3.4	Méthodes statistiques (Z-score, IQR, etc.)	39
4.3.5	Réseaux de neurones récurrents (LSTM)	39
4.3.6	Évaluation des modèles et optimisation	40
4.4	Analyse comparative des approches existantes	40
4.5	Conclusion	41
5	Travaux existants et approches comparatives	42
5.1	Introduction	42
5.2	Travaux sur la qualité des fluides de forage	42
5.3	Travaux sur l'intelligence artificielle appliquée à la détection d'anomalies	43
5.4	Comparaison des approches existantes	44
5.5	Conclusion	44
III.	Conception et réalisation de la solution	46
6	Conception de la solution	47
6.1	Introduction	47
6.2	Rappel sur le besoin	48
6.2.1	Problématique industrielle chez ADFC	48

6.2.2	Besoin d'automatisation et de fiabilité dans la détection	48
6.3	Méthodologie de la solution proposée	48
6.4	L'idée d'application envisagée	50
6.5	Conclusion	51
7	Réalisation de la solution	52
7.1	Introduction	52
7.2	Construction et préparation du dataset	52
7.2.1	Collecte des données (rapports, capteurs...)	52
7.2.2	Extraction des données	53
7.2.3	Nettoyage et normalisation	54
7.2.4	Centralisation et stockage	56
7.2.5	Analyse exploratoire et visualisation des anomalies	59
7.3	Modélisation et entraînement	65
7.3.1	Choix de l'algorithme : Random Forest	65
7.3.2	Validation croisée et optimisation des hyperparamètres	66
7.3.3	Évaluation et résultats du modèle Random Forest .	66
7.4	Développement de l'application interactive	68
7.4.1	Introduction	68
7.4.2	Architecture de l'application	69
7.4.3	Page d'accueil de l'application	70
7.4.4	Fonctionnalités principales de l'application	72
7.4.5	Conclusion de l'application interactive	79
7.5	Conclusion	79
	Conclusion générale	80

Perspectives	82
Bibliographie	85
Annexes	86
Annexe A : Bibliothèques Python utilisées	87
Annexe B : Outils et environnements techniques	89
Annexe C : Notions et concepts avancés	90
Annexe D : Résultats comparatifs des algorithmes testés	92
Annexe E : Codes sources et scripts d'interface Streamlit	95

Liste des tableaux

1.1	Fiche technique de l'ADFC SPA	19
1.2	Glossaire des termes techniques liés aux fluides de forage	22
4.1	Comparaison des techniques de détection d'anomalies	41
5.1	Comparaison des approches de détection d'anomalies appliquées aux fluides de forage	44
7.1	Variables conservées et leur signification	64
7.2	Rapport de classification du modèle Random Forest	67
7.3	Résumé comparatif des performances des algorithmes testés	94

Liste des figures

1.1	Logo de l'ADFC SPA	18
1.2	Logo de Sonatrach	20
1.3	Clients et Partenaires	24
4.1	Typologie des approches d'apprentissage pour la détection d'anomalies dans un contexte industriel.	38
5.1	Architecture typique du système de contrôle classique des fluides de forage [10]	43
6.1	Illustration du processus de forage.	47
6.2	Comparaison des coûts et risques entre la détection précoce et tardive des anomalies dans les fluides de forage. [5]	48
6.3	Étapes de la méthodologie de détection d'anomalies appliquée aux fluides de forage.	50
7.1	Exemple de fiche de formulation et résultats de tests physiques d'un fluide de forage	53
7.2	Exemple d'extraction des données depuis un fichier Excel.	54
7.3	Résultat de l'extraction des données issues des différents fichiers Excel . . .	54
7.4	Extrait du fichier de données nettoyé utilisé pour l'entraînement du modèle.	56
7.5	Tableau des données après centralisation et préparation pour l'entraînement du modèle.	58
7.6	Nombre d'anomalies détectées par type de fluide.	60
7.7	Heatmap des corrélations entre variables.	61
7.8	Distribution de la turbidité pour les anomalies détectées.	62
7.9	Distribution du Oil Water Ratio (%) pour les fluides OBM.	63

7.10	Matrice de confusion du modèle Random Forest	67
7.11	Courbe ROC du modèle Random Forest (AUC = 0.97)	68
7.12	Architecture fonctionnelle de l'application de détection d'anomalies.	70
7.13	Interface de l'application Streamlit pour la détection d'anomalies dans les fluides de forage.	71
7.14	Interface de téléversement de fichiers dans l'application.	73
7.15	Résultats de l'analyse des échantillons avec explication automatique.	74
7.16	Visualisation des distributions des variables Density, pH et Viscosity selon les classes normales (0) et anormales (1) dans la nouvelle base.	75
7.17	Boxplots comparatifs des variables <i>Yield Point</i> , <i>RPM</i> et <i>pH</i> selon la classe IsAnomaly.	75
7.18	Exemple de notification d'alerte dans l'application.	76
7.19	Bouton de téléchargement du rapport PDF généré automatiquement par l'application après l'analyse.	77
7.20	Extrait d'un rapport d'analyse PDF généré par l'application. Chaque échan- tillon est accompagné d'une probabilité d'anomalie et d'une explication automatique basée sur les seuils métiers.	78

Liste des abréviations

- **ADFC** : *Algerian Drilling Fluids Services Company*
- **API** : *American Petroleum Institute*
- **IA / AI** : *Intelligence Artificielle / Artificial Intelligence*
- **OBM** : *Oil-Based Mud* (fluide de forage à base d'huile)
- **WBM** : *Water-Based Mud* (fluide de forage à base d'eau)
- **RPM** : *Revolutions Per Minute* (vitesse de rotation)
- **SMOTE** : *Synthetic Minority Over-sampling Technique*
- **ML** : *Machine Learning* (apprentissage automatique)
- **SCADA** : *Supervisory Control And Data Acquisition*
- **LWD / MWD** : *Logging While Drilling / Measurement While Drilling*
- **LSTM** : *Long Short-Term Memory*
- **PDF** : *Portable Document Format*
- **CSV** : *Comma-Separated Values*
- **ERP** : *Enterprise Resource Planning*
- **ISO** : *International Organization for Standardization*
- **IQR** : *Interquartile Range* (plage interquartile)
- **AUC** : *Area Under Curve* (aire sous la courbe ROC)
- **ROC** : *Receiver Operating Characteristic*
- **UI** : *User Interface*
- **API RP 13B** : *Recommended Practice 13B* (méthodologie normalisée pour les tests de fluides)
- **ENP** : *École Nationale Polytechnique*
- **SSPP** : *Sous-secteur Para-Pétrolier*

Introduction générale

Dans un contexte industriel où la fiabilité des opérations constitue un enjeu stratégique majeur, la surveillance en temps réel des fluides de forage apparaît comme un levier fondamental de performance. Ces fluides remplissent des fonctions critiques tout au long du processus de forage : stabilisation du puits, transport des déblais, lubrification, et maintien de la pression. Toute anomalie dans leurs propriétés physico-chimiques peut engendrer des conséquences graves, allant de la baisse d'efficacité à des accidents opérationnels.

Or, les systèmes actuels de contrôle qualité restent largement fondés sur des procédures manuelles ou semi-automatisées, souvent peu réactives face aux variations soudaines des conditions de terrain. Face à ces limites, l'intégration de l'intelligence artificielle (IA) et de l'apprentissage automatique (machine learning) offre des perspectives nouvelles pour améliorer la détection précoce d'anomalies dans les échantillons de fluides.

Ce projet de fin d'études s'inscrit dans cette dynamique d'innovation, en partenariat avec l'entreprise ADFC, filiale du groupe Sonatrach spécialisée dans les services liés aux fluides de forage. L'objectif est de développer un système intelligent, capable de détecter automatiquement les anomalies à partir de données réelles issues de rapports industriels, et d'en fournir une interprétation explicable via une application interactive.

Problématique

Comment améliorer la fiabilité et la réactivité du processus de contrôle qualité des fluides de forage, en automatisant la détection d'anomalies à partir de données multivariées, tout en garantissant une interprétabilité adaptée aux utilisateurs métier ?

Objectifs

Ce mémoire vise à :

- Analyser le fonctionnement actuel du contrôle qualité au sein d'ADFC et identifier ses limites ;
- Étudier les approches existantes, traditionnelles et basées sur l'IA, pour la détection d'anomalies dans les données industrielles ;
- Concevoir un modèle d'apprentissage automatique, basé sur Random Forest et optimisé par SMOTE, pour classifier les échantillons ;
- Développer une application utilisateur permettant le chargement de fichiers, la visualisation des résultats, et la génération automatique de rapports explicatifs.

Structure du mémoire

Le mémoire est structuré en trois grandes parties cohérentes :

Partie I – État des lieux Cette première partie présente l’organisme d’accueil (ADFC), ses missions, ses partenaires et son fonctionnement actuel en matière de contrôle qualité des fluides. Elle identifie les limites du système en place et propose des axes d’amélioration pour une digitalisation plus efficace.

Partie II – État de l’art Cette partie développe les bases théoriques nécessaires à la compréhension du projet. Elle introduit les caractéristiques des fluides de forage, les types d’anomalies rencontrées, puis passe en revue les méthodes traditionnelles de détection. Un second chapitre explore les techniques modernes issues de l’intelligence artificielle, avec un accent particulier sur les modèles supervisés adaptés au contexte industriel.

Partie III – Conception et réalisation de la solution Cette dernière partie détaille la méthodologie suivie pour construire le modèle de détection d’anomalies, depuis la collecte et la préparation des données jusqu’à l’entraînement du modèle et le développement d’une application interactive. Elle présente également les résultats obtenus, les visualisations produites, ainsi que les bénéfices et perspectives d’évolution de la solution proposée.

Ce travail se veut à la fois scientifique et appliqué, en combinant rigueur méthodologique et application concrète au sein d’un environnement industriel réel. Il démontre le potentiel des solutions IA pour renforcer la qualité, la sécurité et la performance dans le secteur du forage pétrolier.

Première partie

État des lieux

Chapitre 1

Présentation de l'Organisme d'Accueil

1.1 Introduction

Dans cette première partie, nous allons présenter l'entreprise d'accueil **ADFC SPA** (**Algerian Drilling Fluids Services Company**), filiale du **Holding Sonatrach – SSPP**. Cette présentation permettra de mieux comprendre le rôle stratégique de l'entreprise dans l'écosystème pétrolier algérien, ainsi que le contexte professionnel dans lequel s'inscrit le présent projet. Nous introduirons également des concepts clés et un vocabulaire technique liés aux fluides de forage afin de faciliter la compréhension des chapitres à venir.

1.2 Présentation générale de l'ADFC

L'Algerian Drilling Fluids Services Company (ADFC SPA) est une entreprise algérienne spécialisée dans la gestion des fluides de forage pour les opérations pétrolières. Créée dans le cadre d'un partenariat entre **Sonatrach** et **MI-SWACO** (filiale de **Schlumberger**), elle est devenue, depuis avril 2024, une filiale à **100 % du Holding Sonatrach – SSPP**. Cette intégration lui confère un rôle stratégique dans la chaîne de valeur du forage pétrolier national.



FIGURE 1.1 – Logo de l'ADFC SPA

Élément	Détails
Nom complet	Algerian Drilling Fluids Services Company (ADFC SPA)
Statut juridique	Société par actions (SPA) – Filiale du Holding Sonatrach – SSPP
Domaine d’activité	Services de formulation, traitement et contrôle qualité des fluides de forage
Produits/Services	Additifs et produits chimiques, fluides de forage, de workover et de complétion
Clients principaux	Sonatrach et ses filiales (ENAFOR, ENSP...), entreprises pétrolières opérant en Algérie
Adresse	168, rue Hassiba Ben Bouali, Tour B, 6 étage, El Hamma, Alger, Algérie
Base d’opération	Route d’El Borma – P.O. Box 565 – 30500 Hassi Messaoud – Algeria
Téléphone/Fax	+213 29 74 99 14
Nombre d’employés	Environ 200 (objectif à moyen terme : 500)
Transformation digitale	ERP (Odoo), infrastructure IT renforcée, intégration d’IA pour l’analyse des fluides
Normes et certifications	ISO 9001 (Qualité), ISO 14001 (Environnement), API RP 13B, ISO 13500

TABLE 1.1 – Fiche technique de l’ADFC SPA

1.3 Présentation de Sonatrach et de son lien avec l’ADFC

Sonatrach, créée en 1963, est la compagnie nationale algérienne des hydrocarbures et représente le cœur du secteur énergétique du pays. Elle constitue aujourd’hui l’un des plus grands groupes pétroliers intégrés en Afrique, avec des activités couvrant toute la chaîne de valeur : exploration, production, transport, raffinage, commercialisation et distribution. Elle contribue à environ **95 % des recettes d’exportation nationales** et à une part significative du produit intérieur brut (PIB) de l’Algérie, ce qui souligne son poids stratégique dans l’économie nationale.

Afin de développer une expertise nationale dans les services techniques spécialisés, Sonatrach a mis en place plusieurs holdings sectoriels. **L’ADFC SPA** s’inscrit dans cette stratégie à travers son rattachement au **Holding Sonatrach – SSPP (Sous-secteur Para-Pétrolier)**. Initialement créée en partenariat avec MI-SWACO (filiale de Schlumberger), l’ADFC est devenue, depuis avril 2024, **une filiale à 100 % de Sonatrach**. Ce changement structurel a permis de renforcer l’intégration des services liés aux fluides de forage au sein du dispositif national.

Cette relation hiérarchique et fonctionnelle entre Sonatrach et l’ADFC confère à cette

dernière un rôle stratégique dans la chaîne de forage pétrolier, en l’alignant sur les objectifs de modernisation, de réduction des coûts, et de souveraineté technologique nationale. Le projet de détection d’anomalies dans les fluides de forage s’inscrit ainsi pleinement dans cette vision industrielle intégrée, en apportant une réponse technologique innovante à des enjeux opérationnels concrets du groupe Sonatrach.



FIGURE 1.2 – Logo de Sonatrach

1.4 Missions de l’ADFC

L’**Algerian Drilling Fluids Services Company (ADFC SPA)** a pour mission principale d’accompagner les opérations de forage pétrolier à travers la gestion complète des fluides de forage. Cela inclut la formulation de solutions adaptées aux conditions géologiques, le suivi en temps réel des propriétés physico-chimiques des fluides (telles que la rhéologie, la densité ou la filtration), ainsi que la fourniture d’additifs spécifiques.

L’entreprise assure également le traitement et la régénération des fluides usagés dans une optique de réduction des coûts et de respect des normes environnementales. Elle met à disposition des experts pour assister les équipes sur le terrain et garantit un contrôle qualité conforme aux standards internationaux (API RP 13B, ISO 13500). Enfin, l’ADFC s’inscrit dans une dynamique d’innovation en intégrant des outils numériques, notamment pour la détection d’anomalies via l’intelligence artificielle.

1.5 Vision de l’ADFC

La vision stratégique de l’**ADFC SPA** est de devenir un acteur de référence dans le domaine des fluides de forage à l’échelle nationale et régionale. Elle ambitionne d’atteindre l’excellence opérationnelle en s’appuyant sur l’innovation, la maîtrise technologique locale, et le développement continu des compétences internes. Cette vision s’inscrit dans les objectifs plus larges du groupe Sonatrach en matière de digitalisation et de souveraineté industrielle.

1.6 Objectifs stratégiques de l'ADFC

Les objectifs de l'ADFC visent à renforcer la performance, la fiabilité et la durabilité de ses services. Parmi les priorités figurent :

- L'optimisation de la qualité des fluides de forage en conditions réelles.
- La réduction des incidents liés aux anomalies (pertes, éruptions, contamination).
- L'intégration de technologies numériques pour le suivi et l'analyse en temps réel.
- La réduction des coûts opérationnels via le traitement et la réutilisation des fluides.
- Le respect strict des normes de sécurité et d'environnement (API, ISO).

1.7 Termes et jargon professionnels

Avant d'aborder la problématique de ce projet, il est essentiel de se familiariser avec certains termes techniques couramment utilisés dans l'industrie des services pétroliers, et plus particulièrement dans le domaine des fluides de forage. Le tableau suivant regroupe les notions clés nécessaires à une bonne compréhension des chapitres qui suivent.

Terme	Définition
Fluide de forage	Boue industrielle utilisée pour refroidir le trépan, stabiliser les parois du puits, transporter les déblais et contrôler la pression.
Rhéologie	Étude de l'écoulement et du comportement mécanique des fluides sous contrainte (viscosité, seuil d'écoulement, etc.).
Turbidité	Mesure de la concentration en particules solides en suspension dans le fluide ; indicateur de propreté ou de contamination.
Viscosité	Résistance d'un fluide à l'écoulement ; détermine sa capacité à transporter les déblais.
Densité	Masse volumique du fluide, utilisée pour équilibrer les pressions de formation.
Filtration API	Quantité de fluide perdue à travers une barrière poreuse, mesurée selon les standards API (RP 13B).
Perte de circulation	Phénomène où le fluide s'échappe vers la formation, réduisant l'efficacité du forage.
Coup de bélier (Kick)	Entrée soudaine de fluide de formation dans le puits due à une pression insuffisante du fluide de forage.
Additifs	Composés chimiques ajoutés pour ajuster les propriétés du fluide (lubrification, inhibition, stabilité, etc.).
OBM (Oil-Based Mud)	Fluide de forage à base d'huile, performant mais à fort impact environnemental.
WBM (Water-Based Mud)	Fluide de forage à base d'eau, plus écologique, utilisé selon les contraintes du sous-sol.
OWR (Oil Water Ratio)	Rapport volumique huile/eau dans un fluide OBM ; doit rester dans une plage optimale pour assurer la stabilité.
Point de rendement (Yield Point)	Force minimale nécessaire pour initier l'écoulement du fluide.
Solids Content (%)	Pourcentage de solides dans le fluide ; un excès peut engendrer des risques d'obstruction ou de sédimentation.
Trépan	Outil de coupe utilisé pour forer la roche au fond du puits.
MWD / LWD	Technologies de mesure et d'enregistrement pendant le forage (Measurement/Logging While Drilling).
SCADA / WITSML	Systèmes de supervision et d'échange de données en temps réel entre équipements de forage et interfaces de contrôle.

TABLE 1.2 – Glossaire des termes techniques liés aux fluides de forage

1.8 Digitalisation des processus métiers et intégration technologique

L'organisation de l'**ADFC SPA** repose sur une structuration interne claire, divisée en pôles spécialisés couvrant les volets technique, logistique, qualité, R&D, et informatique. Cette organisation permet une gestion efficace des opérations de terrain, du support scientifique, ainsi que de la conformité aux normes internationales en vigueur.

Dans une dynamique d'amélioration continue et de modernisation, l'entreprise a engagé une **transformation numérique progressive**. Celle-ci vise à renforcer la performance opérationnelle, à optimiser le traitement des données, et à sécuriser les processus critiques. Les principaux axes de cette transformation sont les suivants :

- **Implémentation d'un système ERP (Odoo)** : centralisation de la gestion des processus métiers (achats, stock, maintenance, RH, etc.) pour une meilleure coordination entre les départements,, comme recommandé dans les approches modernes de gestion opérationnelle [1]. Cette transition numérique rejoint les standards internationaux adoptés dans les projets de champs pétroliers numériques [2].
- **Renforcement de l'infrastructure informatique** : mise en place de serveurs sécurisés, systèmes de sauvegarde, et solutions de partage de données inter-sites.
- **Digitalisation du contrôle qualité** : automatisation partielle de la collecte de données en laboratoire et standardisation des rapports techniques.
- **Exploration de l'intelligence artificielle** : développement de projets pilotes pour l'analyse prédictive, la détection d'anomalies dans les fluides de forage, et l'optimisation des interventions sur site.

Cette démarche s'inscrit dans la stratégie globale du Holding Sonatrach – SSPP, visant à faire émerger des filiales plus autonomes, innovantes et compétitives dans un contexte énergétique en mutation.

1.9 Clients et Partenaires

L'**ADFC SPA** collabore principalement avec les filiales opérationnelles du groupe Sonatrach, telles que *ENAFOR*, *ENSP* et *ENTP*, en leur fournissant des services de formulation et de contrôle des fluides de forage. Ces collaborations couvrent des projets de forage terrestres répartis sur tout le territoire national, notamment dans les régions du Sud (Hassi Messaoud, In Salah, Illizi, etc.).

Par ailleurs, l'entreprise développe des partenariats techniques avec des prestataires spécialisés, notamment dans les domaines des additifs chimiques, des instruments de mesure, et des systèmes d'acquisition de données (SCADA, WITSML). Cette ouverture favorise l'intégration de solutions technologiques avancées, en phase avec les standards internationaux du secteur.



FIGURE 1.3 – Clients et Partenaires

1.10 Engagements qualité, sécurité et environnement

L'ADFC place la qualité, la sécurité et la protection de l'environnement au cœur de sa stratégie opérationnelle. Consciente des enjeux liés aux risques industriels et à l'impact environnemental de ses activités, l'entreprise s'est engagée dans une démarche structurée conforme aux référentiels internationaux :

- **ISO 9001** pour le management de la qualité, assurant la conformité des produits et services aux exigences client.
- **ISO 14001** pour la gestion environnementale, intégrant la maîtrise des rejets et le traitement des déchets.
- **API RP 13B** pour les essais normalisés des fluides de forage.

Des contrôles rigoureux sont effectués en laboratoire et sur site afin de garantir la performance des fluides et la sécurité des opérations. L'ADFC applique également les règles HSE du groupe Sonatrach et développe une culture d'amélioration continue à tous les niveaux.

1.11 Conclusion

La présentation de l'ADFC SPA permet de situer le projet dans un environnement industriel structuré, tourné vers l'innovation et la maîtrise technique. Grâce à son expertise, ses partenariats stratégiques et sa démarche qualité, l'ADFC offre un cadre favorable à l'expérimentation de solutions d'intelligence artificielle appliquées au domaine du forage. Cette compréhension globale de l'entreprise d'accueil constitue une base solide pour analyser les limites du système existant et justifier la mise en place d'un modèle de détection des anomalies performant.

Chapitre 2

Étude de l'existant : le système actuel

2.1 Introduction

Le contrôle de la qualité des fluides de forage constitue une étape essentielle dans le déroulement des opérations de forage pétrolier. Il permet d'assurer la stabilité des puits, la sécurité des équipements et l'efficacité globale des travaux en cours. Au sein de l'ADFC, ce contrôle repose actuellement sur un ensemble de méthodes techniques et de pratiques établies, impliquant des relevés manuels, des tests normalisés et des interprétations humaines.

Cependant, avec l'évolution des technologies et l'augmentation du volume de données collectées en temps réel, les limites du système actuel apparaissent de plus en plus clairement. Ce chapitre a pour objectif de décrire en détail l'état du système existant de contrôle qualité utilisé par l'ADFC, d'en analyser les principales failles et d'identifier les opportunités d'amélioration. Cela permettra de mieux justifier, dans les chapitres suivants, le recours à une approche basée sur l'intelligence artificielle pour améliorer la détection des anomalies dans les fluides de forage.

2.2 Les rapports et méthodes de contrôle qualité des fluides

Le système actuel de contrôle qualité des fluides de forage mis en œuvre par l'ADFC repose sur une combinaison de rapports techniques, d'analyses de laboratoire et de mesures en temps réel. Cette approche permet de suivre les propriétés des boues de forage tout au long des opérations, et d'ajuster leur composition en fonction des conditions rencontrées sur le terrain.

2.2.1 Types de rapports produits

Les rapports utilisés par l'ADFC peuvent être classés selon leur nature et leur moment de production. On distingue principalement :

- **Rapports de laboratoire** : ils incluent les résultats des tests physico-chimiques réalisés sur les échantillons prélevés (ex. viscosité, densité, pH, filtration API). Ces tests sont réalisés selon des protocoles standards comme l'API RP 13B ou l'ISO 13500.
- **Rapports journaliers de puits** : documents produits sur le terrain, généralement sous forme de fiches Excel ou papier, contenant les mesures de base observées sur site (température, débit, comportement du fluide, incidents).
- **Fiches de suivi** : documents internes à l'ADFC utilisés pour tracer les ajustements réalisés sur les formulations de fluides, les recommandations techniques, et les consommations d'additifs.
- **Bilans hebdomadaires ou mensuels** : récapitulatifs produits pour les chefs de projet ou les coordinateurs techniques, incluant des graphiques de tendance et des analyses de performance.

2.2.2 Sources des données utilisées

Ces rapports sont alimentés par deux grandes catégories de sources de données :

- **Mesures de terrain et données temps réel** : recueillies via des capteurs de surface ou des systèmes automatisés tels que SCADA, WITSML ou MWD/LWD. Elles permettent un suivi continu des paramètres critiques comme la pression, la température, la densité ou la vitesse de circulation.
- **Analyses en laboratoire** : les échantillons sont régulièrement envoyés dans les laboratoires d'ADFC pour des tests plus détaillés, incluant la spectroscopie, la chromatographie, ou des analyses de contamination. Ces données sont fiables mais non continues, avec un certain délai de traitement.
- **Données historiques** : logs de forage, historiques d'incidents, comptes rendus de chantier et archivage des rapports précédents. Elles permettent une contextualisation et une analyse des dérives dans le temps.
- **Données complémentaires** : parfois croisées avec des données géologiques (nature de la formation), climatiques (opérations offshore) ou relatives à l'état de maintenance des équipements.

La diversité des sources rend l'analyse complexe, notamment à cause de la disparité des formats (CSV, Excel, PDF, flux JSON ou temps réel) et de l'absence de structuration centralisée des données. Cette problématique sera abordée plus en détail dans la section suivante.

2.3 Le processus actuel de surveillance

Le processus de surveillance de la qualité des fluides de forage chez ADFC repose actuellement sur une combinaison d'interventions humaines, de tests normalisés et d'analyses périodiques. Ce suivi est réalisé à la fois sur le terrain (au niveau du puits) et en laboratoire, dans une logique réactive plus que proactive.

Sur site, les techniciens boue et les ingénieurs procèdent à des mesures régulières à l'aide d'instruments portables (viscosimètre, balance de densité, kit API de filtration, etc.). Ces mesures sont consignées manuellement dans des fiches de suivi ou dans des fichiers Excel. En parallèle, des échantillons sont envoyés au laboratoire pour des tests plus poussés, notamment en cas de suspicion d'anomalie ou de dérive.

Les résultats sont ensuite analysés par les superviseurs (ingénieurs fluides), qui valident ou recommandent des ajustements sur la base de seuils métiers définis (API RP 13B, normes internes ADFC), de l'expérience terrain, et de la tendance observée dans les tests. Il n'existe toutefois pas de système formel d'aide à la décision, ce qui limite l'uniformité des recommandations.

Bien que ce processus permette une certaine réactivité, il présente plusieurs limites :

- Dépendance à l'expertise humaine pour l'interprétation des résultats.
- Absence d'un système d'alerte automatique en cas d'écart critique.
- Décalage temporel entre l'apparition d'une anomalie et sa détection effective (temps d'analyse).
- Données souvent non intégrées dans une base centralisée ni exploitables en temps réel.

Ce constat met en évidence la nécessité d'un système plus intelligent, capable de détecter automatiquement les écarts et d'assister les équipes dans la prise de décision rapide.

2.4 Limites du système actuel

L'analyse du système en place au sein de l'ADFC pour le suivi et le contrôle qualité des fluides de forage permet de mettre en évidence plusieurs limites structurelles et fonctionnelles. Ces limites freinent l'efficacité globale du processus de détection d'anomalies et justifient l'exploration de solutions innovantes, notamment basées sur l'intelligence artificielle.

2.4.1 Faiblesses techniques

Le premier ensemble de limitations concerne la nature même des outils actuellement utilisés. Les mesures sont souvent réalisées manuellement à l'aide d'appareils de terrain, ce qui introduit une forte dépendance aux opérateurs. L'absence de standardisation complète dans la saisie des données rend leur exploitation difficile pour une analyse à grande échelle.

En outre, les données provenant de capteurs ne sont pas toujours fiables à cause de :

- conditions de forage extrêmes (vibrations, chaleur, humidité),

- bruit dans les signaux mesurés,
- dérive des capteurs non calibrés régulièrement.

Le traitement des données est par ailleurs souvent local et dispersé, comme constaté dans plusieurs environnements pétroliers [3]. De plus, les capteurs sont parfois sujets à des dérives, avec des mesures bruitées en conditions extrêmes [4].

Les documents sont enregistrés sous différents formats (Excel, PDF, papier), ce qui empêche toute consolidation automatique, recherche efficace ou apprentissage basé sur l'historique.

2.4.2 Manque d'automatisation

Le second frein majeur est l'absence de système automatisé de détection et de classification des anomalies. Aujourd'hui, la détection d'un écart repose essentiellement sur l'œil expert des techniciens et ingénieurs. Cela suppose :

- une présence humaine continue,
- un niveau d'expertise élevé,
- une réaction parfois tardive en cas de situation critique.

Par ailleurs, il n'existe pas de système d'alerte dynamique basé sur des seuils intelligents, ni de solution intégrée pour anticiper les anomalies avant qu'elles ne deviennent problématiques. Cela limite les capacités de prévention, augmente les risques d'incidents opérationnels (perte de circulation, contamination), et impacte les coûts de traitement.

Ces limites ouvrent la voie à l'implémentation de solutions intelligentes capables d'exploiter les données en temps réel pour automatiser la détection, la classification et, à terme, la prédiction des anomalies.

2.5 Identification des opportunités d'amélioration

L'analyse du système de contrôle qualité actuellement en place met en évidence un besoin croissant de modernisation. L'exploitation massive des données générées au cours du processus de forage constitue une opportunité stratégique pour passer d'une logique de suivi manuel à une logique de détection automatisée, intelligente et prédictive.

2.5.1 Énoncé de la problématique

Malgré l'existence de procédures bien établies, le processus de surveillance des fluides de forage souffre d'un manque de réactivité et de structuration des données. Les anomalies sont souvent détectées tardivement, après qu'elles ont commencé à produire des effets négatifs sur les opérations (pertes de fluide, instabilités, contamination).

Problématique :

Comment concevoir un système intelligent capable de détecter automatiquement, en temps réel, les anomalies dans les fluides de forage à partir de données multi-sources (capteurs, laboratoire, historiques), afin d'améliorer la réactivité opérationnelle et la fiabilité du processus de forage ?

2.5.2 Axes d'amélioration envisagés

L'étude de l'existant ouvre la voie à plusieurs pistes d'amélioration, tant sur le plan technologique que méthodologique. Les axes prioritaires identifiés sont :

- **Centralisation des données** : création d'un entrepôt de données structuré, intégrant les relevés terrain, les tests de laboratoire et les historiques d'incidents.
- **Surveillance en temps réel** : connexion directe aux capteurs via SCADA/WITSML pour permettre l'analyse continue des paramètres critiques.
- **Détection automatisée des écarts** : utilisation de modèles d'apprentissage automatique (Machine Learning) pour identifier les comportements anormaux à partir de seuils dynamiques ou de patterns appris.[5,6]
- **Visualisation et alertes intelligentes** : développement d'interfaces permettant aux ingénieurs de visualiser les tendances, recevoir des notifications, et explorer les causes probables d'une anomalie.
- **Prévision des dérives** : déploiement d'algorithmes prédictifs pour anticiper les anomalies à partir de l'évolution temporelle des données (LSTM, ARIMA, Prophet...).

La mise en œuvre de ces axes constituera le cœur du travail pratique, présenté dans les chapitres suivants.

2.6 Conclusion

Cette étude de l'existant a permis de dresser un état des lieux du système de contrôle qualité actuellement mis en œuvre par l'ADFC dans la gestion des fluides de forage. Si les méthodes en place reposent sur des fondations solides, notamment en termes de normes et de savoir-faire métier, elles montrent aujourd'hui des limites importantes face aux exigences modernes de réactivité, de fiabilité et d'exploitation des données.

Les rapports, bien que nombreux et détaillés, souffrent d'un manque d'intégration et d'exploitation automatisée. Le processus de surveillance reste majoritairement manuel, avec une forte dépendance à l'expérience humaine et une absence de mécanismes intelligents d'alerte ou de prédiction.

Face à ces constats, plusieurs axes d'amélioration ont été identifiés, ouvrant la voie à l'introduction d'un système d'analyse intelligent, basé sur l'intelligence artificielle. Ce dernier permettra de détecter plus tôt les anomalies, de prédire les dérives futures, et d'accompagner les équipes terrain dans une prise de décision rapide et efficace.

Dans le chapitre suivant, nous décrirons en détail la conception du modèle d'intelligence artificielle proposé, ainsi que la méthodologie adoptée pour le développement, l'entraînement et l'évaluation du système.

Deuxième partie

État de l'art

Chapitre 3

Fluides de forage et anomalies associées

3.1 Introduction

Les fluides de forage jouent un rôle central dans les opérations pétrolières en assurant des fonctions critiques telles que le refroidissement du trépan, le transport des déblais et le contrôle des pressions. Ce chapitre présente leur définition, leurs propriétés physico-chimiques essentielles, ainsi que les principales anomalies observées sur le terrain. Ces éléments sont indispensables pour comprendre les enjeux liés à leur surveillance et poser les bases d'une détection intelligente des anomalies.

3.2 Définition et rôle des fluides de forage

Les fluides de forage, aussi appelés boues de forage, sont des mélanges complexes utilisés pour faciliter et sécuriser les opérations de forage pétrolier. Leur composition comprend généralement une phase liquide (eau ou huile), des agents visqueux, des matériaux pondéreux et divers additifs chimiques[8].

Ils assurent plusieurs fonctions essentielles [9] :

- Refroidir et lubrifier le trépan pendant le forage,
- Transporter les déblais rocheux vers la surface,
- Maintenir la stabilité des parois du puits,
- Équilibrer les pressions des formations géologiques traversées,
- Prévenir les incidents comme les influx ou les pertes de fluide.

La performance d'un fluide dépend de ses propriétés physico-chimiques, qui doivent être ajustées en continu en fonction des conditions de terrain. Toute anomalie dans leur comportement peut compromettre la sécurité, l'efficacité ou le coût des opérations de forage. C'est pourquoi leur suivi en temps réel constitue un enjeu majeur pour les acteurs du secteur.

3.3 Propriétés physico-chimiques à surveiller

Le bon fonctionnement d'un fluide de forage dépend directement de ses propriétés physico-chimiques, qui influencent sa capacité à remplir ses fonctions de manière optimale [16]. Ces propriétés doivent être suivies de manière régulière, que ce soit via des mesures de terrain ou des analyses en laboratoire. Toute variation anormale peut indiquer une anomalie en cours ou un risque opérationnel.

Parmi les principales propriétés surveillées, on distingue :

- **La densité (mud weight)** : elle conditionne la pression hydrostatique exercée dans le puits. Une densité trop faible peut entraîner un influx (entrée de fluide de formation), tandis qu'une densité trop élevée peut provoquer des pertes de circulation.
- **La viscosité** : elle définit la capacité du fluide à transporter les déblais en suspension. Elle est mesurée en utilisant des instruments comme le Marsh Funnel ou le viscosimètre rotatif.
- **Le pH** : un paramètre chimique important qui affecte la stabilité de certains additifs et la réaction du fluide avec la formation. Un pH non maîtrisé peut accélérer la corrosion ou dégrader les performances du fluide.
- **La filtrabilité (test de filtration API)** : elle mesure la quantité de liquide perdue par le fluide à travers une formation poreuse. Une filtrabilité excessive peut conduire à une perte de fluide dans la formation et à une instabilité des parois du puits.
- **Le pouvoir de suspension** : c'est la capacité du fluide à maintenir les déblais en suspension lorsque la circulation est arrêtée. Elle dépend de la viscosité statique et de la gélification.
- **Les solides contenus** : leur concentration influence fortement la densité, la viscosité et l'usure des équipements. Une surcharge en solides peut réduire l'efficacité du fluide.
- **La conductivité électrique et la salinité** : utiles dans les fluides à base aqueuse pour détecter des contaminations ou des déséquilibres ioniques.

Ces paramètres sont conformes aux recommandations de l'API et des fournisseurs industriels majeurs [8, 9].

Le suivi de ces paramètres permet de détecter les premiers signes de dérive du fluide, d'anticiper les ajustements nécessaires, et de prévenir les anomalies majeures. Leur surveillance continue constitue donc un pilier central du contrôle qualité dans les opérations de forage.

3.4 Types d'anomalies rencontrées : nature et classification

Malgré les contrôles réguliers, les fluides de forage peuvent présenter diverses anomalies au cours des opérations. Ces anomalies, souvent détectées trop tardivement [10], peuvent compromettre la sécurité, ralentir les travaux, ou générer des surcoûts importants. Il est donc essentiel de les connaître, de les classer, et de comprendre leurs causes pour améliorer leur détection.

Les principales anomalies rencontrées peuvent être regroupées en plusieurs catégories :

- **Anomalies de densité** : une densité insuffisante peut entraîner un influx de fluide de formation (kick), tandis qu'une densité excessive peut provoquer des pertes de circulation. Ces écarts peuvent être liés à une dilution non maîtrisée, à la présence de gaz dissous ou à des ajouts incorrects d'additifs.
- **Pertes de fluide** : on parle de pertes de circulation lorsque le fluide s'échappe dans la formation au lieu de circuler en boucle. Elles peuvent être totales, partielles ou sélectives selon la nature géologique traversée. Ce phénomène est souvent causé par une pression excessive ou une perméabilité anormalement élevée.
- **Contaminations** : il s'agit de la présence d'éléments indésirables dans le fluide, comme du ciment, du gaz, de l'eau salée, ou des agents chimiques incompatibles. Ces contaminations altèrent les propriétés du fluide (pH, viscosité, filtrabilité) et compliquent son traitement.
- **Sédimentation et perte de suspension** : une mauvaise capacité à maintenir les déblais en suspension peut entraîner leur dépôt au fond du puits, provoquant un bouchage partiel (pack-off) ou un coincement du train de tiges.
- **Problèmes de filtration** : une perte excessive de liquide lors du test API peut signaler une instabilité du fluide. Cela peut être causé par une absence d'agent de colmatage, une dispersion de particules fines ou une contamination.
- **Anomalies rhéologiques** : des variations inhabituelles de la viscosité, du seuil de gel ou du comportement plastique du fluide peuvent signaler une altération de la formulation. Elles se traduisent souvent par une difficulté à pomper, des vibrations anormales ou un mauvais nettoyage du puits.

Chacune de ces anomalies présente un risque opérationnel significatif. Dans un environnement aussi sensible que le forage pétrolier, leur détection précoce est un enjeu de performance et de sécurité. Cela souligne l'importance d'une surveillance continue, idéalement assistée par des systèmes intelligents capables d'identifier automatiquement ces écarts à partir des données collectées en temps réel.

3.5 Conséquences des anomalies sur les opérations

Les anomalies dans les fluides de forage, lorsqu'elles ne sont pas détectées et corrigées à temps, peuvent avoir des conséquences graves sur le déroulement et la sécurité des opérations de forage. Leur impact peut être technique, économique, environnemental, voire humain.

Les principales conséquences observées sont les suivantes :

- **Perte de contrôle du puits** : une anomalie de densité peut entraîner un déséquilibre de pression. Si la pression du fluide est inférieure à celle de la formation, un influx incontrôlé (kick) peut survenir, menant dans certains cas à une éruption (blowout), avec des risques majeurs pour la sécurité et l'environnement.
- **Pertes de circulation** : lorsque le fluide s'échappe dans les formations poreuses, il n'est plus disponible pour remplir ses fonctions critiques. Cela peut interrompre

le forage, exiger des produits de colmatage coûteux, et allonger significativement la durée des travaux.

- **Coincement ou perte d'outils** : une mauvaise suspension des déblais ou une sédimentation excessive peut bloquer le train de forage, entraînant des opérations de pêche (récupération d'équipements) ou même la perte définitive du puits.
- **Usure prématurée des équipements** : des anomalies dans la concentration des solides ou la viscosité peuvent accélérer l'érosion des pompes, tiges, vannes et autres composants critiques.
- **Altération de la qualité des données** : les systèmes de mesure pendant le forage (MWD/LWD) peuvent être perturbés par des anomalies de composition du fluide, compromettant la précision des données géologiques recueillies.
- **Surcoûts et retards** : chaque incident dû à une anomalie implique des interventions correctives (traitements chimiques, interventions mécaniques, arrêt du forage), qui se traduisent par une augmentation des coûts journaliers et un dépassement des délais prévus.
- **Impacts environnementaux** : certaines anomalies, comme les éruptions non contrôlées ou les pertes de fluide dans des zones sensibles, peuvent provoquer une pollution du sous-sol ou des nappes phréatiques, exposant l'opérateur à des sanctions réglementaires.

Ainsi, la surveillance des fluides de forage n'est pas seulement une exigence de qualité, mais une nécessité stratégique pour garantir la continuité, la sécurité et la rentabilité des opérations. C'est dans cette optique que se justifie l'intégration de systèmes intelligents capables d'anticiper les dérives et de déclencher des alertes précoces.

3.6 Techniques traditionnelles de détection des anomalies

Avant l'intégration de technologies intelligentes, la détection des anomalies dans les fluides de forage repose principalement sur des méthodes classiques combinant observations humaines, tests physico-chimiques standardisés et interprétations techniques.

Ces approches, bien que largement éprouvées, présentent plusieurs limites en matière de réactivité, d'automatisation et d'analyse à grande échelle. On peut distinguer trois grandes catégories de techniques traditionnelles :

- **Mesures manuelles sur site** : réalisées par les techniciens boue à l'aide d'outils comme le Marsh Funnel (viscosité), la balance de densité, ou le kit API (filtrabilité). Ces mesures sont reportées sur des fiches papier ou des fichiers Excel, et interprétées à l'œil humain selon des seuils fixés à l'avance.
- **Analyses en laboratoire** : des échantillons de fluide sont prélevés à intervalles réguliers et analysés en laboratoire pour déterminer des paramètres comme le pH, la composition chimique, la conductivité, ou la présence de contaminants. Ces tests offrent une meilleure précision, mais leur fréquence est limitée et les résultats ne sont pas disponibles en temps réel.

- **Interprétation basée sur l'expérience** : les décisions sont souvent guidées par le jugement des superviseurs et des ingénieurs de forage, en fonction de leur connaissance du terrain et de la formation géologique. Cette approche empirique permet une certaine flexibilité, mais reste sujette à l'erreur humaine ou au manque d'uniformité entre équipes.

En complément, certains sites de forage utilisent des systèmes automatisés comme le SCADA (Supervisory Control And Data Acquisition) ou le protocole WITSML pour centraliser les données collectées par les capteurs (pression, température, débit, etc.). Toutefois, ces systèmes restent le plus souvent limités à de la collecte passive ou à des alertes basées sur des seuils fixes.

Les techniques traditionnelles permettent donc une forme de contrôle de qualité, mais ne permettent pas de détecter des comportements complexes, des anomalies émergentes ou des dérives subtiles difficilement perceptibles. C'est cette limitation qui motive l'exploration de solutions basées sur l'intelligence artificielle pour améliorer la réactivité, la fiabilité et l'autonomie du système de surveillance.

3.7 Limites des méthodes conventionnelles

Les méthodes traditionnelles de détection des anomalies dans les fluides de forage, bien qu'encore largement utilisées dans l'industrie, présentent plusieurs limites structurelles et opérationnelles. Ces limites freinent la capacité des équipes à détecter précocement des dérives critiques et à réagir de manière proactive face à des situations anormales.

Les principales faiblesses identifiées sont les suivantes :

- **Réactivité limitée** : les analyses en laboratoire introduisent un délai entre la mesure et l'interprétation. Ce temps de latence peut être critique lorsque des anomalies évoluent rapidement dans des environnements instables.
- **Manque d'automatisation** : la plupart des mesures nécessitent une intervention humaine, ce qui limite la fréquence des relevés et introduit une dépendance au facteur humain.
- **Analyse locale et fragmentée** : les données collectées sont souvent réparties dans différents formats (papier, Excel, SCADA isolé), sans centralisation ni historisation cohérente. Cela empêche l'exploitation des données à grande échelle ou la détection de patterns complexes.
- **Seuils statiques et règles rigides** : les systèmes de détection classiques se basent généralement sur des seuils fixes (ex. densité $> 1.3 \text{ g/cm}^3$), sans capacité d'adaptation aux conditions spécifiques du puits ou de la formation. Cela peut générer des fausses alertes ou des non-détections.
- **Dépendance à l'expertise humaine** : la qualité de l'interprétation repose largement sur l'expérience des opérateurs. Cela crée des risques d'incohérence entre sites, et rend difficile la standardisation ou la traçabilité des décisions.
- **Absence de capacité prédictive** : les méthodes conventionnelles se limitent à détecter ce qui s'est déjà produit. Elles ne permettent ni d'anticiper une dégradation progressive, ni de simuler des scénarios à venir.

Ces limitations soulignent la nécessité d'un changement de paradigme vers des approches plus intelligentes, capables d'analyser de grands volumes de données en temps réel, d'apprendre des comportements historiques, et de générer des alertes précoces. C'est précisément dans ce contexte que s'inscrit l'intégration de l'intelligence artificielle, qui sera développée dans le chapitre suivant.

3.8 Conclusion

Dans ce chapitre, nous avons présenté les bases techniques nécessaires à la compréhension des problématiques liées aux fluides de forage. Nous avons défini leur rôle fondamental dans les opérations pétrolières, identifié les propriétés physico-chimiques clés à surveiller, et classifié les principales anomalies rencontrées sur le terrain.

Nous avons également passé en revue les méthodes traditionnelles de détection utilisées dans l'industrie, en mettant en évidence leurs limites en termes de réactivité, d'automatisation, et de capacité d'analyse prédictive. Ces constats renforcent la nécessité d'adopter des approches plus robustes et intelligentes pour améliorer le contrôle qualité en temps réel.

Le prochain chapitre sera donc consacré aux techniques d'intelligence artificielle, en particulier au Machine Learning, qui offrent des solutions innovantes pour la détection automatique et anticipée des anomalies dans les fluides de forage.

Chapitre 4

Intelligence Artificielle pour la détection d'anomalies

4.1 Introduction

Ce chapitre explore les apports de l'intelligence artificielle, en particulier du Machine Learning, dans la détection des anomalies au sein des systèmes industriels. Appliquée aux fluides de forage, l'IA permet de dépasser les limites des méthodes traditionnelles en exploitant les données collectées en temps réel pour identifier des comportements anormaux.

Nous présenterons les notions clés du Machine Learning et du Deep Learning, les types d'apprentissage, les algorithmes les plus pertinents pour la détection d'anomalies, ainsi que les principaux critères d'évaluation. Cette base théorique permettra de motiver les choix méthodologiques adoptés dans le cadre de notre solution.

4.2 Principes du Machine Learning appliqué à la détection d'anomalies

Le Machine Learning (ML), ou apprentissage automatique, est une branche de l'intelligence artificielle qui permet à un système d'apprendre des modèles à partir de données [11], sans être explicitement programmé pour chaque tâche. Dans le domaine industriel, le ML est particulièrement utile pour détecter automatiquement des anomalies dans de grands volumes de données hétérogènes, comme ceux issus de capteurs ou d'analyses de laboratoire.

La détection d'anomalies consiste à identifier des comportements, valeurs ou événements qui s'écartent de manière significative du fonctionnement habituel d'un système. Contrairement à la classification classique, où chaque donnée est associée à une étiquette connue, la détection d'anomalies implique souvent une situation déséquilibrée où les données normales sont nombreuses, et les anomalies rares, variables ou inconnues.

Le ML permet de traiter ce type de problème en apprenant les motifs récurrents associés au comportement "normal", et en identifiant toute observation qui s'en éloigne. Les approches

peuvent être :

- **Supervisées** : les données d'entraînement incluent des exemples d'anomalies déjà identifiées. Cela nécessite une phase de labellisation préalable, parfois difficile à obtenir dans les environnements industriels.
- **Non supervisées** : le modèle apprend la structure des données normales et détecte toute observation atypique comme potentiellement anormale. Ces méthodes sont adaptées lorsque les anomalies ne sont pas bien documentées.
- **Semi-supervisées** : seul un ensemble représentatif de données normales est utilisé pour l'entraînement. Le système considère toute déviation comme suspecte. Cette approche est très pertinente pour les systèmes où les anomalies sont rares ou évolutives.

Dans le contexte de la qualité des fluides de forage, le ML permet d'analyser en continu les paramètres physiques et chimiques mesurés sur le terrain, et de déclencher des alertes en cas de dérive non conforme aux modèles appris. Cela renforce la réactivité, réduit la dépendance à l'intervention humaine, et améliore la robustesse du processus de surveillance.

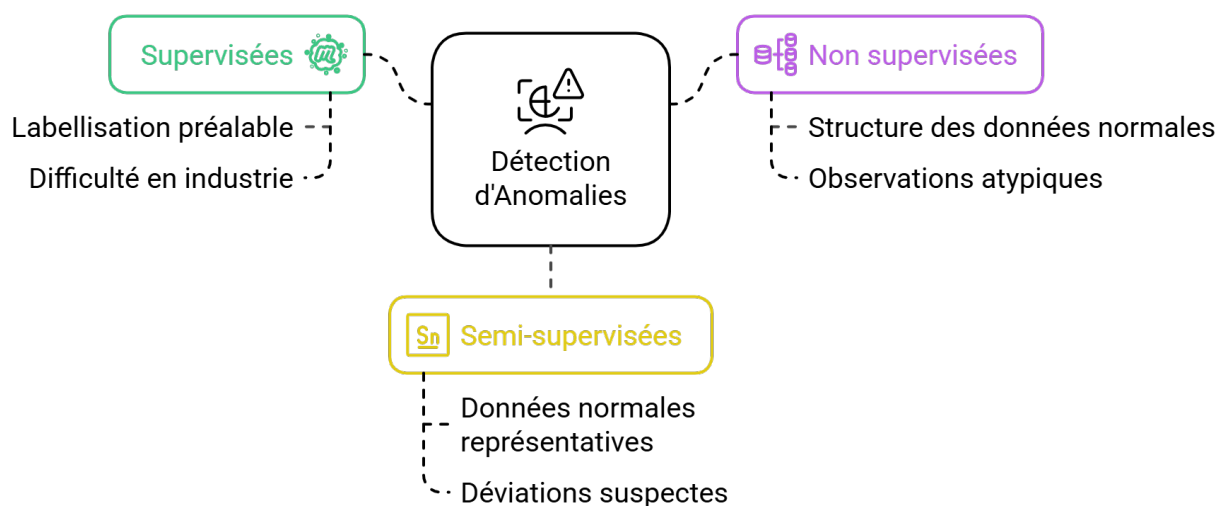


FIGURE 4.1 – Typologie des approches d'apprentissage pour la détection d'anomalies dans un contexte industriel.

4.3 Techniques utilisées pour la détection d'anomalies

Plusieurs algorithmes de Machine Learning sont utilisés pour détecter automatiquement des anomalies dans des systèmes industriels complexes. Le choix de la technique dépend notamment de la nature des données, de la disponibilité de labels (données annotées), de la fréquence des anomalies, et du besoin de transparence ou de vitesse d'exécution.

Voici les principales méthodes utilisées dans les contextes industriels, dont celui du forage pétrolier :

4.3.1 Random Forest

La Random Forest est un algorithme d'apprentissage supervisé basé sur des arbres de décision. Il fonctionne en construisant un ensemble d'arbres indépendants, chacun entraîné sur un sous-échantillon aléatoire des données. L'agrégation des résultats permet une classification robuste.

Pour la détection d'anomalies, il est utilisé lorsque les anomalies sont labellisées. Il offre une bonne précision, gère bien les données bruitées et permet une certaine interprétabilité via l'analyse des variables importantes.

4.3.2 Isolation Forest

L'Isolation Forest est une méthode non supervisée spécifiquement conçue pour la détection d'anomalies. Elle fonctionne en isolant les observations en construisant des arbres de partition [12]. Les points anormaux sont plus rapidement isolés que les points normaux, ce qui permet de les identifier efficacement.

Cette technique est particulièrement adaptée aux grands ensembles de données avec peu d'anomalies connues, comme c'est souvent le cas dans le forage.

4.3.3 Autoencoders

Les autoencodeurs sont des réseaux de neurones utilisés pour l'apprentissage non supervisé. Ils apprennent à reproduire les données en sortie après un passage par une couche de compression (bottleneck) [14]. Lorsqu'un échantillon anormal est introduit, l'erreur de reconstruction est généralement élevée, ce qui permet de le détecter.

Les autoencodeurs sont adaptés aux anomalies complexes et non linéaires, mais nécessitent un volume important de données pour bien fonctionner.

4.3.4 Méthodes statistiques (Z-score, IQR, etc.)

Ces approches reposent sur l'hypothèse que les données normales suivent une certaine distribution (souvent gaussienne). Les points situés en dehors de seuils statistiques (ex. au-delà de 3 écarts-types) sont considérés comme anomalies.

Elles sont simples à mettre en œuvre et rapides, mais peu robustes dans des environnements où les données sont multi-dimensionnelles ou non distribuées normalement.

4.3.5 Réseaux de neurones récurrents (LSTM)

Les LSTM (Long Short-Term Memory) sont une variante des réseaux de neurones récurrents adaptés aux séries temporelles [13]. Ils permettent de modéliser les dépendances temporelles dans les données et sont efficaces pour anticiper des dérives progressives.

Ils sont particulièrement utiles pour analyser les variations continues dans les mesures des fluides de forage, en détectant les tendances inhabituelles sur plusieurs heures ou jours.

4.3.6 Évaluation des modèles et optimisation

Afin d'évaluer correctement les performances d'un modèle d'apprentissage automatique, plusieurs techniques sont utilisées pour garantir sa robustesse et sa capacité de généralisation. Parmi elles, la validation croisée, l'optimisation des hyperparamètres, et les métriques d'évaluation sont essentielles.

Validation croisée La validation croisée (*cross-validation*) consiste à diviser le jeu de données en k sous-ensembles. À chaque itération, l'un des sous-ensembles est utilisé pour le test, les autres pour l'entraînement. Le score final est obtenu par moyenne des performances sur l'ensemble des itérations. Cette méthode permet d'évaluer la capacité du modèle à généraliser et de détecter d'éventuels surapprentissage.

Optimisation des hyperparamètres Les modèles comme Random Forest nécessitent l'ajustement d'hyperparamètres (ex : `n_estimators`, `max_depth`, `min_samples_split`). La recherche de la meilleure configuration peut être réalisée via la méthode de grille (`GridSearchCV`), qui teste plusieurs combinaisons d'hyperparamètres avec validation croisée à chaque essai.

Métriques d'évaluation Les performances du modèle sont mesurées à l'aide de plusieurs indicateurs :

- **Précision** : proportion des prédictions positives correctes.
- **Rappel (Recall)** : proportion d'anomalies correctement détectées.
- **F1-score** : moyenne harmonique entre précision et rappel.
- **Matrice de confusion** : tableau des vrais/faux positifs et négatifs.
- **Courbe ROC** : visualisation du compromis entre sensibilité et spécificité.

Ces outils permettent de comparer plusieurs modèles et de choisir celui le plus adapté à la tâche cible.

4.4 Analyse comparative des approches existantes

Le choix d'une méthode de détection d'anomalies dépend fortement des contraintes du contexte industriel, du type de données disponibles, des exigences de précision et de temps réel, ainsi que de la facilité de déploiement. Dans le domaine du forage, les données sont souvent bruitées, multidimensionnelles, et collectées en continu, ce qui exige des modèles robustes, adaptatifs et capables d'apprendre à partir de données partiellement étiquetées.

Voici une analyse comparative des principales approches abordées :

- **Méthodes statistiques** : Elles sont rapides, faciles à mettre en œuvre, et ne nécessitent pas d'apprentissage. Cependant, elles manquent de flexibilité et ne capturent

pas les relations complexes entre variables. Elles peuvent être utiles comme première couche de filtrage.

- **Isolation Forest** : Méthode légère et efficace, particulièrement adaptée aux cas non supervisés. Elle est performante même sur de grands volumes de données, mais elle peut être sensible à l'échelle des variables et nécessite un prétraitement soigné.
- **Random Forest** : Très précise lorsque des données labellisées sont disponibles. Elle permet aussi une interprétabilité partielle via les scores d'importance des variables. Cependant, elle est moins adaptée en cas d'absence d'anomalies connues.
- **Autoencodeurs** : Très puissants pour détecter des anomalies subtiles dans des données complexes et non linéaires. Ils nécessitent toutefois un grand volume de données d'apprentissage et sont peu interprétables pour les non-experts.
- **LSTM** : Idéal pour les séries temporelles comme les données issues de capteurs en continu. Leur capacité à mémoriser des séquences longues permet de détecter des dérives progressives, mais leur entraînement est plus coûteux en temps et ressources.

Méthode	Supervision	Explicabilité	Adaptée au temps réel
Random Forest	Oui	Moyenne	Oui
Isolation Forest	Non	Moyenne	Oui
Autoencoder	Non	Faible	Partiellement
Méthodes statistiques	Non	Élevée	Oui
LSTM	Oui/Non	Faible	Oui

TABLE 4.1 – Comparaison des techniques de détection d'anomalies

En résumé, il n'existe pas une seule méthode universelle. Une approche hybride combinant plusieurs techniques (ex. détection initiale par Isolation Forest + validation par Random Forest) peut s'avérer plus efficace. Ce constat est confirmé dans plusieurs études récentes sur les méthodes combinées [15]. Le choix final dépendra des caractéristiques des données disponibles au sein de l'ADFC, des contraintes de déploiement en environnement réel, et des besoins en explicabilité pour les opérateurs.

4.5 Conclusion

Ce chapitre a présenté les principales techniques d'intelligence artificielle utilisées pour détecter des anomalies dans des environnements industriels. Les approches étudiées, telles que Random Forest, Isolation Forest, Autoencoders, méthodes statistiques et LSTM, offrent des performances variables selon les données disponibles et les contraintes du terrain.

L'analyse comparative a montré qu'il n'existe pas de solution unique : le choix du modèle dépend du besoin de précision, de rapidité, et surtout d'interprétabilité. Ces éléments théoriques préparent le terrain à la mise en œuvre concrète de la solution de détection, qui sera développée dans le chapitre suivant.

Chapitre 5

Travaux existants et approches comparatives

5.1 Introduction

Dans le contexte des opérations de forage pétrolier, la qualité des fluides joue un rôle déterminant dans la stabilité du puits, la sécurité des équipements, et l'efficacité globale des interventions. L'identification rapide et précise d'anomalies dans la composition ou le comportement de ces fluides constitue donc un enjeu critique pour les industriels du secteur.

Ce chapitre a pour objectif d'examiner les principaux travaux de recherche et initiatives industrielles relatifs à la détection d'anomalies dans les fluides de forage. Il vise à situer notre démarche dans un panorama plus large, en analysant d'une part les approches traditionnelles basées sur le contrôle qualité standardisé, et d'autre part les apports récents de l'intelligence artificielle et du machine learning dans ce domaine.

L'étude s'attarde notamment sur les publications scientifiques, les applications concrètes en environnement industriel, ainsi que sur les performances et les limites des méthodes existantes. Une comparaison structurée des approches permettra d'identifier les lacunes persistantes et de motiver le choix de notre propre solution technique présentée dans les chapitres suivants.

5.2 Travaux sur la qualité des fluides de forage

De nombreux travaux académiques et industriels ont été consacrés à la compréhension, la caractérisation et l'amélioration des fluides de forage. Ces recherches se sont principalement focalisées sur le suivi des propriétés physico-chimiques essentielles comme la densité, la viscosité, la rhéologie et la composition chimique, dans le but de prévention des incidents tels que la perte de fluide, la contamination ou la mauvaise stabilisation du puits.

Les méthodes traditionnelles reposent souvent sur l'utilisation de capteurs sur site et sur

des analyses en laboratoire, combinées à des règles empiriques basées sur l'expérience des opérateurs. Par exemple, Zhou et al. [10] ont proposé un aperçu des techniques de "mud logging" qui permettent de collecter des données critiques sur les fluides pendant le forage.

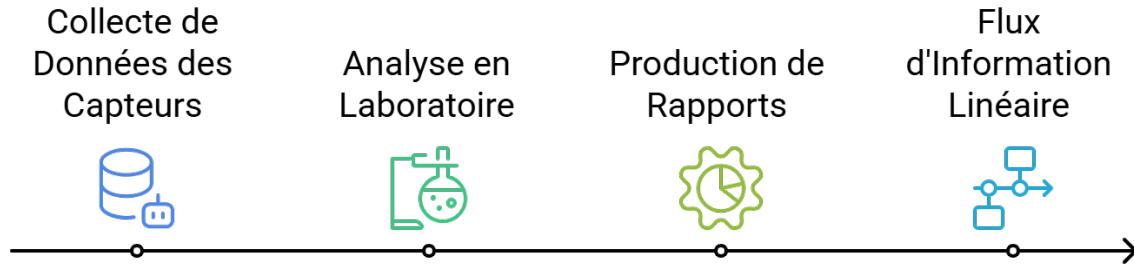


FIGURE 5.1 – Architecture typique du système de contrôle classique des fluides de forage [10]

Cependant, ces approches présentent plusieurs limitations :

- forte dépendance à l'expertise humaine,
- délais dans la collecte et l'analyse des données,
- difficulté à détecter des dérives progressives ou non-linéaires,
- absence d'automatisation et de traitement prédictif.

Ces limites ont été discutées dans plusieurs revues de la littérature, notamment Khosravi et al. [4], qui soulignent le besoin d'intégrer des solutions d'analyse en temps réel plus intelligentes.

5.3 Travaux sur l'intelligence artificielle appliquée à la détection d'anomalies

La détection d'anomalies par intelligence artificielle (IA) a fait l'objet de nombreuses recherches ces dernières années, notamment dans des secteurs à forte variabilité comme l'industrie pétrolière, la maintenance prédictive ou les systèmes de surveillance en temps réel. Ces approches exploitent la capacité des algorithmes de Machine Learning à apprendre les comportements normaux d'un système et à identifier automatiquement les observations qui s'en écartent.

Plusieurs travaux ont porté sur des algorithmes comme Random Forest, Isolation Forest, Autoencodeurs ou LSTM. Chandola et al. [11] ont proposé une revue de référence sur la détection d'anomalies par apprentissage, tandis que Liu et al. [12] ont introduit l'algorithme Isolation Forest spécifiquement conçu pour ce type de tâche.

Dans le domaine de la surveillance de processus industriels, Malhotra et al. [13] ont démontré l'efficacité des réseaux LSTM pour la détection de dérives dans des séries temporelles. D'autres travaux, comme ceux de Ahmad et al. [15], ont comparé l'efficacité des approches non supervisées pour l'identification de points atypiques dans des données bruitées.

Ces contributions montrent que l'IA permet non seulement de détecter plus précisément les anomalies, mais aussi de s'adapter dynamiquement à l'évolution du comportement

normal, ce qui est essentiel dans le contexte des fluides de forage, où les conditions varient en fonction de la profondeur, de la géologie et des ajustements opérationnels.

5.4 Comparaison des approches existantes

Pour synthétiser les différentes contributions présentées, nous proposons une comparaison des approches classiques et intelligentes de détection d'anomalies, en mettant en regard leurs domaines d'application, leurs forces et leurs limites.

Référence	Méthode	Domaine	Type	Forces	Limites
Zhou et al. (2017)	Règles empiriques	Forage	Classique	Simple, connues	Peu adaptatives
Khosravi et al. (2021)	Capteurs + seuils	Forage	Classique	Temps réel	Faible précision
Chandola et al. (2009)	Revue ML	Multi-domaines	IA	Couverture large	Pas de cas concret
Liu et al. (2008)	Isolation Forest	Industrie	IA non sup.	Léger, rapide	Interprétabilité moyenne
Malhotra et al. (2015)	LSTM	Temps réel	IA sup.	Modèle temporel	Données lourdes

TABLE 5.1 – Comparaison des approches de détection d'anomalies appliquées aux fluides de forage

Cette comparaison montre que les approches traditionnelles, bien qu'encore largement utilisées, manquent de souplesse et de capacité prédictive. Les techniques d'intelligence artificielle offrent des perspectives plus avancées, en particulier pour le traitement dynamique des données et l'automatisation de la détection. Toutefois, leur mise en oeuvre suppose des préalables importants : qualité des données, capacité de calcul, expertise technique, et explicabilité des résultats.

5.5 Conclusion

L'état de l'art présenté dans ce chapitre met en évidence l'évolution des méthodes de contrôle qualité des fluides de forage, allant des approches empiriques aux modèles d'intelligence artificielle de plus en plus sophistiqués. Les solutions traditionnelles, bien que simples et largement utilisées, ne permettent pas une surveillance proactive ni une anticipation efficace des anomalies.

Les travaux récents montrent que l'application des algorithmes de machine learning, en particulier ceux basés sur les séries temporelles et le deep learning, constitue une alternative prometteuse. Ils permettent une détection plus précise, adaptative, et automatisée, même dans des contextes opérationnels complexes.

Cependant, ces approches requièrent une infrastructure de données robuste, une grande qualité des jeux d'entraînement, et des mécanismes d'interprétabilité adaptés aux enjeux industriels. Ces constats renforcent la pertinence de notre projet, qui vise à développer une solution hybride combinant précision algorithmique et faisabilité opérationnelle dans un contexte réel.

Troisième partie

Conception et réalisation de la solution

Chapitre 6

Conception de la solution

6.1 Introduction

Dans ce projet, l'objectif principal est de développer une solution de détection d'anomalies dans les fluides de forage, en exploitant les avancées de l'intelligence artificielle (IA) pour améliorer l'efficacité des processus de surveillance et de prise de décision. Le but est de concevoir un système automatisé, capable de détecter en temps réel les anomalies dans les propriétés des fluides, afin de permettre une intervention rapide et réduire les risques associés à des erreurs humaines ou des délais dans la détection des anomalies. Le rôle central de ces fluides dans le processus de forage est illustré à la Figure 6.1, qui montre le circuit complet de circulation de la boue, de l'injection jusqu'à la récupération.

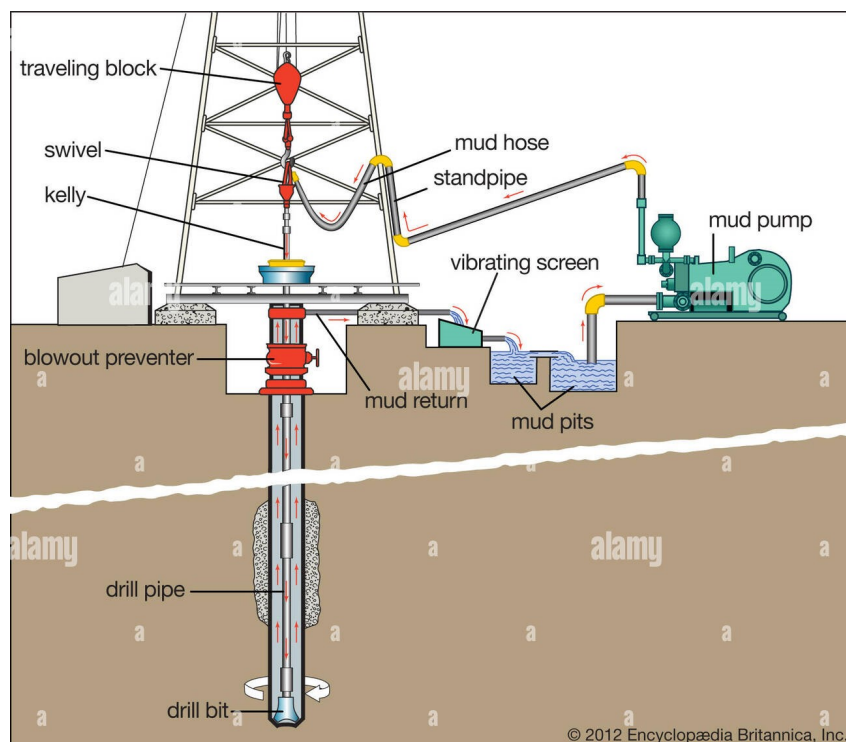


FIGURE 6.1 – Illustration du processus de forage.

6.2 Rappel sur le besoin

6.2.1 Problématique industrielle chez ADFC

L'ADFC (Algerian Drilling Fluids Services Company) est responsable de la gestion des fluides de forage utilisés dans les opérations de forage. Ces fluides sont essentiels à la sécurité et à l'efficacité des forages, car ils régulent la pression dans les puits, refroidissent les équipements et éliminent les débris de forage. La qualité des fluides doit donc être surveillée de manière continue pour éviter tout incident.

6.2.2 Besoin d'automatisation et de fiabilité dans la détection

La solution actuelle manque de mécanismes d'automatisation pour la détection rapide des anomalies. Il devient donc crucial de concevoir un système capable de **surveiller en temps réel** les propriétés des fluides et de **détecter automatiquement les anomalies**. Un tel système permettra de réduire les coûts associés aux erreurs humaines, d'augmenter la réactivité en cas de détection d'anomalies, et de minimiser les risques liés aux fluides de mauvaise qualité.

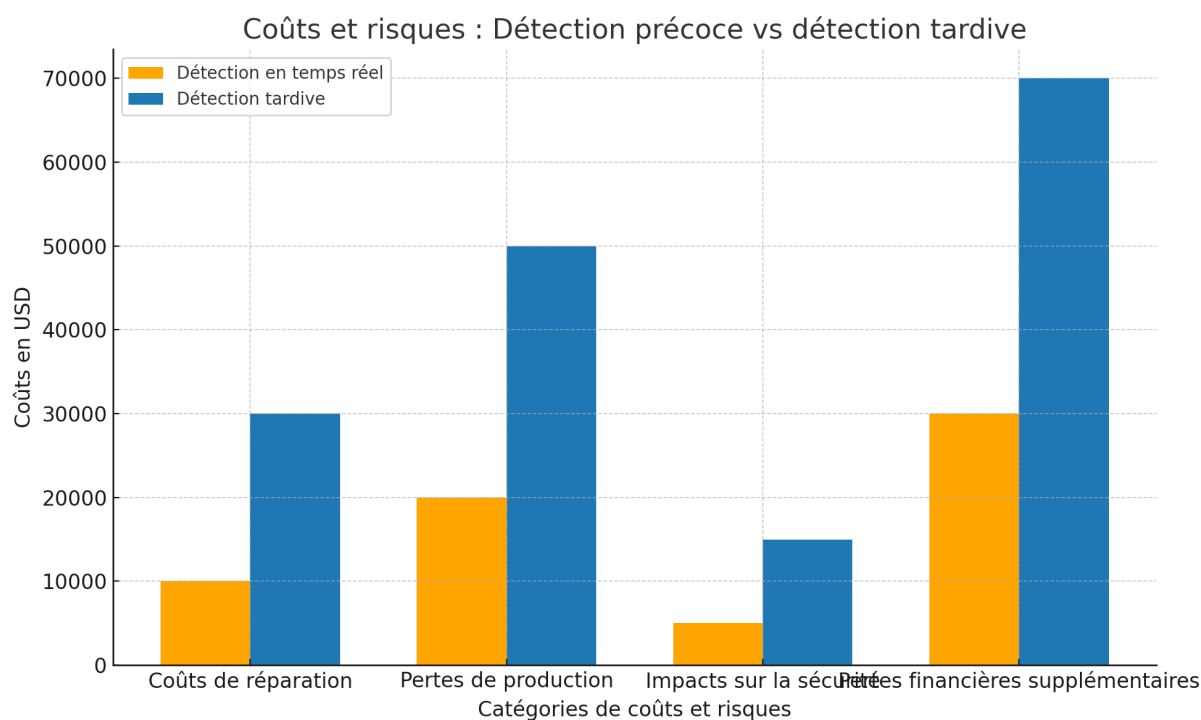


FIGURE 6.2 – Comparaison des coûts et risques entre la détection précoce et tardive des anomalies dans les fluides de forage. [5]

6.3 Méthodologie de la solution proposée

La solution proposée repose sur un **modèle d'intelligence artificielle**, basé sur l'algorithme *Random Forest*, particulièrement adapté à des jeux de données *complexes et dés-*

équilibrés. Ce modèle est entraîné à partir des données historiques des fluides de forage, étiquetées comme “*normales*” ou “*anomalies*”. Une fois entraîné, il est capable d’analyser de nouvelles données en temps réel et de détecter automatiquement des comportements anormaux susceptibles de générer des incidents sur le terrain.

La mise en œuvre de cette solution suit une approche structurée, basée sur la méthodologie **CRISP-DM** (Cross Industry Standard Process for Data Mining), composée de six étapes principales allant de la compréhension métier jusqu’au déploiement. Dans le cadre de ce projet, cette méthode a été adaptée et consolidée en **cinq grandes phases opérationnelles**, représentées à la Figure 6.3.

- **Collecte des données** : les données sont rassemblées à partir de fichiers fournis par l’entreprise ADFC ou récupérées en temps réel depuis les capteurs installés sur le site de forage.
- **Prétraitement** : cette étape inclut le nettoyage des données, la normalisation, la suppression des valeurs aberrantes, la transformation de variables et l’équilibrage des classes à l’aide de la technique *SMOTE*.
- **Modélisation IA** : l’algorithme Random Forest est entraîné pour classifier les enregistrements en deux classes (normal / anomalie) en exploitant les relations non linéaires entre les paramètres physico-chimiques des fluides.
- **Génération de rapports** : chaque détection d’anomalie est automatiquement documentée via un rapport PDF détaillant les paramètres à l’origine de l’alerte.
- **Visualisation des résultats** : l’ensemble de la solution est intégré dans une application *Streamlit*, où les utilisateurs peuvent charger des fichiers, déclencher les prédictions, et consulter les résultats de manière claire et interactive. Une alerte automatique s’affiche à l’écran dès qu’une anomalie est détectée, accompagnée d’un résumé explicatif des paramètres à l’origine de l’alerte.

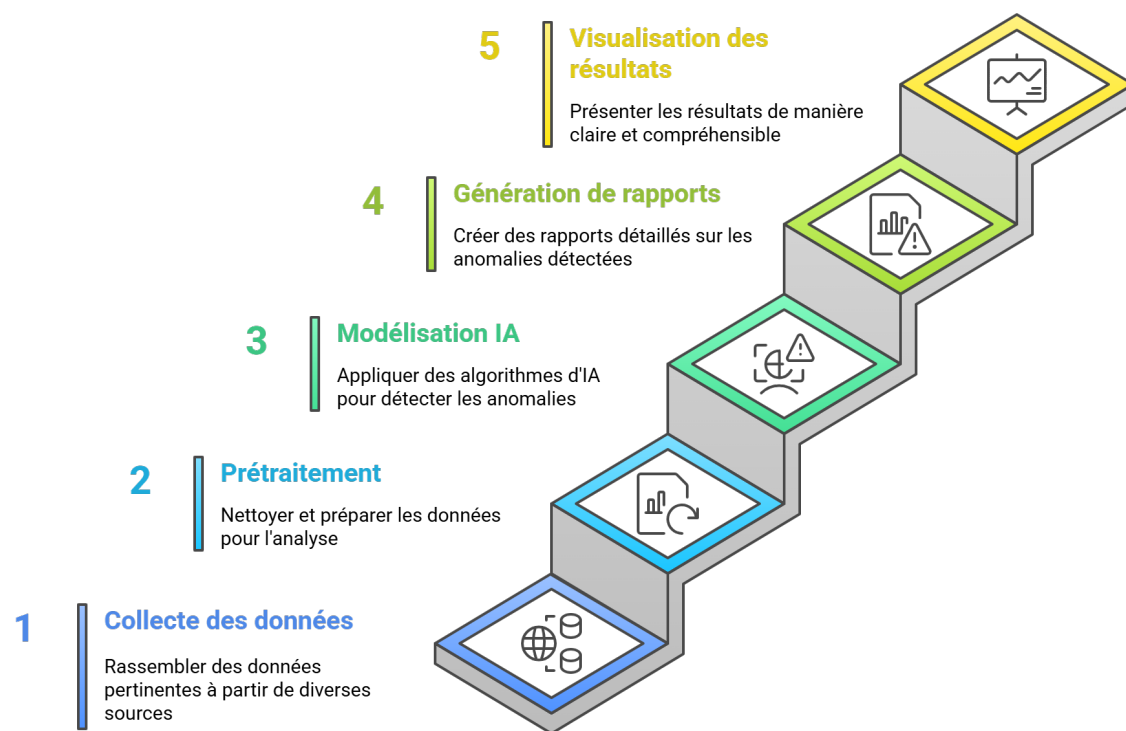


FIGURE 6.3 – Étapes de la méthodologie de détection d’anomalies appliquée aux fluides de forage.

6.4 L’idée d’application envisagée

L’application interactive a été développée à l’aide de **Streamlit**, un framework Python permettant de concevoir des interfaces web dynamiques et accessibles, adaptées aux projets de machine learning.

Cette application vise à fournir un outil pratique et intuitif pour les ingénieurs de terrain. Elle offre les fonctionnalités suivantes :

- **Téléversement de fichiers** : possibilité de charger des fichiers de données (CSV ou Excel) contenant les propriétés physico-chimiques des fluides de forage ;
- **Prédiction automatique** : exécution du modèle d’IA en temps réel pour détecter d’éventuelles anomalies ;
- **Affichage des résultats** : présentation des prédictions sous forme de tableaux, graphiques interactifs et résumés explicatifs des anomalies identifiées ;
- **Génération de rapports** : création de rapports PDF automatisés, listant les anomalies détectées et les paramètres impliqués ;
- **Système d’alerte visuelle** : une alerte s’affiche automatiquement à l’écran lorsque

des anomalies sont détectées, afin de signaler immédiatement aux utilisateurs la présence de comportements anormaux dans les fluides analysés.

Cette interface renforce la réactivité des équipes techniques, tout en facilitant la traçabilité et l'interprétation des résultats.

6.5 Conclusion

Cette solution offre une réponse complète aux besoins de détection d'anomalies dans les fluides de forage chez ADFC. Elle intègre un modèle d'intelligence artificielle avancé avec une interface utilisateur interactive pour automatiser le processus de surveillance et fournir des rapports détaillés. La prochaine étape consiste à mettre en œuvre cette solution dans un environnement de production réel, avec une intégration dans les systèmes existants chez ADFC.

Chapitre 7

Réalisation de la solution

7.1 Introduction

Ce chapitre décrit l'ensemble du processus de réalisation de la solution de détection d'anomalies, depuis la constitution du jeu de données jusqu'au développement de l'application interactive. Il s'articule en trois parties principales : la préparation des données, la modélisation et l'évaluation du modèle d'intelligence artificielle, et enfin l'implémentation technique de l'application.

L'objectif est de montrer comment les choix méthodologiques précédemment présentés ont été concrètement mis en œuvre, en mettant l'accent sur la rigueur du traitement des données, la performance du modèle Random Forest, ainsi que sur l'aspect fonctionnel et ergonomique de l'interface destinée aux utilisateurs finaux.

7.2 Construction et préparation du dataset

L'objectif principal de cette section est de décrire en détail comment les données ont été extraites, nettoyées, normalisées, et préparées pour l'entraînement du modèle de machine learning. Chaque étape est cruciale pour garantir que les données sont de qualité suffisante pour permettre au modèle de détecter les anomalies dans les fluides de forage de manière fiable et précise.

7.2.1 Collecte des données (rapports, capteurs...)

Les données utilisées pour entraîner le modèle proviennent de plusieurs sources, principalement des **fichiers Excel** fournis par ADFC. Ces fichiers contiennent des mesures réelles sur les propriétés des fluides de forage, telles que la densité, la viscosité, le pH, et d'autres caractéristiques mesurées en temps réel par des capteurs.

Les fichiers contiennent plusieurs feuilles. Certaines feuilles sont pertinentes pour l'analyse, tandis que d'autres ne contiennent pas de données utilisables. Une étape d'extraction a donc été réalisée pour ne garder que les feuilles contenant des informations utiles. Cette

extraction a permis de centraliser les données pertinentes, facilitant ainsi leur utilisation pour la détection des anomalies.

The screenshot shows an Excel spreadsheet titled "Hamilton Beach Mixer". The main content is a table titled "DRILLING FLUID FORMULATIONS AND TEST RESULTS".

Header Section:

Lab Job Reference :	12/22
Fluid System :	WBM
Fluid/Sample Description :	WBM #1.1. KCl - Bentonite Spud Mud for Riserless Phases
Date :	14-May-2022
Issued by :	M.ZOUAOUI/Y.ANCER

Mixing work instructions: Refer to OM-LAB-002 Chapter II

Sample composition for 350 ml

Products used	Density SG	Units	1.05 sg	1.10 sg	1.20 sg	1.25 sg	1.40 sg			Min	Max	Design
Water		m3/m3	0,96	0,95	0,92	0,90	0,85					Hamilton Beach Mixer
Soda Ash		Kg/m3	0,71	0,71	0,71	0,71	0,71					
Caustic Soda		Kg/m3	0,71	0,71	0,71	0,71	0,71					
Bentonite		Kg/m3	45,60	45,60	45,60	45,60	45,60					
KCl		Kg/m3	50,66	49,83	48,17	47,31	44,77					
Duo-Vis		Kg/m3	2,50	2,50	2,10	2,10	2,10					
Tannathin		Kg/m3	0,00	0,00	0,00	0,28	1,42					
Barite		Kg/m3	0,00	56,31	190,40	257,17	457,26					
Carbosane 1516		Kg/m3	0,34	0,34	0,34	0,34	0,34					

Testing work instr: Mud weight / PH / Funnel Refer to MI Manual Chapter III

Data Target

FIGURE 7.1 – Exemple de fiche de formulation et résultats de tests physiques d'un fluide de forage

7.2.2 Extraction des données

Une fois les fichiers collectés, l'extraction des données pertinentes a été effectuée à l'aide d'un script Python. Ce script permet de lire les différents fichiers Excel, d'extraire les informations nécessaires et de les organiser sous un format structuré. L'objectif principal était de préparer un fichier CSV final qui contiendrait toutes les informations requises pour l'entraînement du modèle.

Le script a procédé comme suit :

- **Identification des feuilles pertinentes** : Les feuilles du fichier Excel ont été parcourues pour identifier celles qui contiennent des données utiles, tout en excluant celles qui ne sont pas pertinentes pour l'analyse.
- **Extraction des métadonnées** : Pour chaque feuille, les informations importantes telles que la date, le type de fluide, les instructions de mélange, et la référence de travail ont été extraites et structurées dans un dictionnaire.
- **Extraction des mesures de fluides** : Les mesures des fluides (comme la densité, la viscosité, etc.) ont été extraites à partir des lignes et colonnes correspondantes dans les fichiers Excel.
- **Création d'un tableau final** : Les données extraites ont été combinées pour créer un tableau avec toutes les informations nécessaires, y compris les identifiants uniques pour chaque échantillon de fluide.

Cela a permis de générer un fichier CSV final contenant des informations complètes et prêtes pour l'analyse et l'entraînement du modèle de machine learning.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	Sample ID	Sample Type	Date	Fluid System	Brine Type	With Inhibitor	Formulation L	Description	Mixing Instruct	Job Ref	Source	CaCl2	Density	NaCl	Products used	Turbidity	Water
1	NaClBr_01	1st sample	01/02/2022	Brine	NaCl	No	NaCl Brine	Brine	Refer to OM-I	févr-22	MI-02-22.xlsx	-	1.08	42.73	1 st sample	74.7	335.6
2	NaClBr_02	2nd sample	01/02/2022	Brine	NaCl	No	NaCl Brine	Brine	Refer to OM-I	févr-22	MI-02-22.xlsx	37.1	1.08	-	2 st sample	34.7	341.25

FIGURE 7.2 – Exemple d'extraction des données depuis un fichier Excel.

Le fichier CSV généré contient les caractéristiques des fluides, telles que la densité, la viscosité, le pH, et d'autres paramètres essentiels, ainsi que des informations contextuelles sur chaque échantillon.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	Sample ID	Sample Type	Date	Fluid System	Brine Type	With Inhibitor	Formulation L	Description	Mixing Instruct	Job Ref	Source	Ca ++	CaCl2	Cl-	Density	KCl	NaCl	Products used
1	CaCl2B_01	1st sample	08/01/2022	OBM	CaCl2	No	CaCl2 Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	26400.0	26.0	52000.0	1.06			Sample 1
2	CaCl2B_01	1st sample	08/01/2022	OBM	CaCl2	No	CaCl2 Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	26400.0	26.0	52000.0	1.06			Sample 1
3	CaCl2B_01	1st sample	08/01/2022	OBM	CaCl2	Yes	CaCl2 Brine w OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	26400.0	26.0	52000.0	1.06			Sample 1
4	CaCl2B_01	1st sample	08/01/2022	OBM	CaCl2	Yes	CaCl2 Brine w OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	26400.0	26.0	52000.0	1.06			Sample 1
5	CaCl2B_02	2nd sample	08/01/2022	OBM	CaCl2	No	CaCl2 Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	48400.0	49.0	84000.0	1.1			Sample 2
6	CaCl2B_02	2nd sample	08/01/2022	OBM	CaCl2	No	CaCl2 Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	48400.0	49.0	84000.0	1.1			Sample 2
7	CaCl2B_02	2nd sample	08/01/2022	OBM	CaCl2	Yes	CaCl2 Brine w OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	48400.0	49.0	84000.0	1.1			Sample 2
8	CaCl2B_02	2nd sample	08/01/2022	OBM	CaCl2	Yes	CaCl2 Brine w OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	48400.0	49.0	84000.0	1.1			Sample 2
9	CaCl2B_03	3rd sample	08/01/2022	OBM	CaCl2	No	CaCl2 Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	100000.0	101.3	176000.0	1.2			Sample 3
10	CaCl2B_03	3rd sample	08/01/2022	OBM	CaCl2	No	CaCl2 Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	100000.0	101.3	176000.0	1.2			Sample 3
11	CaCl2B_03	3rd sample	08/01/2022	OBM	CaCl2	Yes	CaCl2 Brine w OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	100000.0	101.3	176000.0	1.2			Sample 3
12	CaCl2B_03	3rd sample	08/01/2022	OBM	CaCl2	Yes	CaCl2 Brine w OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	100000.0	101.3	176000.0	1.2			Sample 3
13	KClBr_01	1st sample	08/01/2022	OBM	KCl	No	KCl Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	3200.0	48000.0	1.06	33.4			1 st sample
14	KClBr_01	1st sample	08/01/2022	OBM	KCl	No	KCl Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	3200.0	48000.0	1.06	33.4			1 st sample
15	KClBr_01	1st sample	08/01/2022	OBM	KCl	Yes	KCl Brine with OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	3200.0	48000.0	1.06	33.4			1 st sample
16	KClBr_01	1st sample	08/01/2022	OBM	KCl	Yes	KCl Brine with OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	3200.0	48000.0	1.06	33.4			1 st sample
17	KClBr_02	2nd sample	08/01/2022	OBM	KCl	No	KCl Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	3200.0	83000.0	1.11	62.7			2 st sample
18	KClBr_02	2nd sample	08/01/2022	OBM	KCl	No	KCl Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	3200.0	83000.0	1.11	62.7			2 st sample
19	KClBr_02	2nd sample	08/01/2022	OBM	KCl	Yes	KCl Brine with OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	3200.0	83000.0	1.11	62.7			2 st sample
20	KClBr_02	2nd sample	08/01/2022	OBM	KCl	Yes	KCl Brine with OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	3200.0	83000.0	1.11	62.7			2 st sample
21	KClBr_03	3rd sample	08/01/2022	OBM	KCl	No	KCl Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	4000.0	134000.0	1.168	104.0			3 st sample
22	KClBr_03	3rd sample	08/01/2022	OBM	KCl	No	KCl Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	4000.0	134000.0	1.168	104.0			3 st sample
23	KClBr_03	3rd sample	08/01/2022	OBM	KCl	Yes	KCl Brine with OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	4000.0	134000.0	1.168	104.0			3 st sample
24	KClBr_03	3rd sample	08/01/2022	OBM	KCl	Yes	KCl Brine with OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	4000.0	134000.0	1.168	104.0			3 st sample
25	NaClBr_01	1st sample	08/01/2022	OBM	NaCl	No	NaCl Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	2000.0	50000.0	1.06		28.1		1 st sample
26	NaClBr_01	1st sample	08/01/2022	OBM	NaCl	No	NaCl Brine	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	2000.0	50000.0	1.06		28.1		1 st sample
27	NaClBr_01	1st sample	08/01/2022	OBM	NaCl	Yes	NaCl Brine wif OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	2000.0	50000.0	1.06		28.1		1 st sample
28	NaClBr_01	1st sample	08/01/2022	OBM	NaCl	Yes	NaCl Brine wif OBM	OBM	Refer to OM-I041/2021		MI-01-22.xlsx	2000.0	50000.0	1.06		28.1		1 st sample

FIGURE 7.3 – Résultat de l'extraction des données issues des différents fichiers Excel

7.2.3 Nettoyage et normalisation

Une fois les données extraites, elles ont été nettoyées et normalisées afin de les rendre prêtes à l'analyse. Le nettoyage des données est une étape cruciale pour garantir leur qualité et leur pertinence pour l'entraînement du modèle. Voici les étapes clés du nettoyage et de la normalisation :

- **Suppression des colonnes inutiles** : Les colonnes non pertinentes ont été éliminées. Ces colonnes ne contenaient pas d'informations utiles pour la détection des anomalies et étaient donc superflues. Par exemple, la colonne **Formulation Label** a été supprimée, car elle ne contenait pas de données pertinentes pour le modèle.

- **Remplissage des valeurs manquantes** : Les valeurs manquantes dans les colonnes numériques ont été remplacées par la moyenne des valeurs existantes dans la colonne correspondante. Cela a permis de conserver la continuité des données sans introduire de biais importants. Cette méthode a été appliquée à toutes les colonnes numériques et a été réalisée en fonction du **Fluid System** pour prendre en compte les spécificités de chaque type de fluide.
- **Traitement des valeurs aberrantes** : Les valeurs aberrantes (outliers), qui peuvent fausser les résultats du modèle, ont été identifiées et ajustées. Ces valeurs extrêmes ont été détectées et ajustées ou supprimées si elles étaient jugées incorrectes ou impossibles dans le contexte des fluides de forage. Cela a permis d'éliminer les données anormales qui auraient pu perturber l'entraînement du modèle.

Les étapes précises du nettoyage des données ont été réalisées en utilisant un script Python. Voici les opérations importantes du code qui ont été exécutées pour chaque étape :

- **Suppression des colonnes inutiles** : La colonne **Formulation Label** a été supprimée du **DataFrame** car elle ne contenait pas de données utiles pour l'analyse des anomalies dans les fluides de forage.
- **Nettoyage de la colonne Description** : Les descriptions des fluides ont été extraites pour créer des tags indiquant le type de fluide (OBM, WBM, KCl, NaCl, Polymer, etc.), ce qui permet d'étiqueter chaque échantillon de manière plus précise. Ce processus a été réalisé par une fonction `extract_tags`, qui a été appliquée à la colonne **Description** pour identifier les tags pertinents dans chaque description.
- **Nettoyage de la colonne Job Ref** : Un format spécifique a été extrait dans la colonne **Job Ref** pour conserver uniquement l'identifiant pertinent (par exemple, "MI-07-23"). Ce nettoyage a été réalisé à l'aide d'une fonction `clean_job_ref` qui a utilisé des expressions régulières pour extraire le format souhaité.
- **Ajout de la colonne Année** : La colonne **Date** a été utilisée pour extraire l'année de chaque échantillon et créer une nouvelle colonne **Année** pour faciliter les analyses temporelles. Une fonction `extract_year` a été utilisée pour extraire l'année à partir de la date.
- **Remplissage des valeurs manquantes** : Les valeurs manquantes dans les colonnes numériques ont été remplies avec la moyenne des valeurs du **Fluid System** correspondant. Pour les colonnes textuelles, les valeurs manquantes ont été remplacées par la valeur "None".

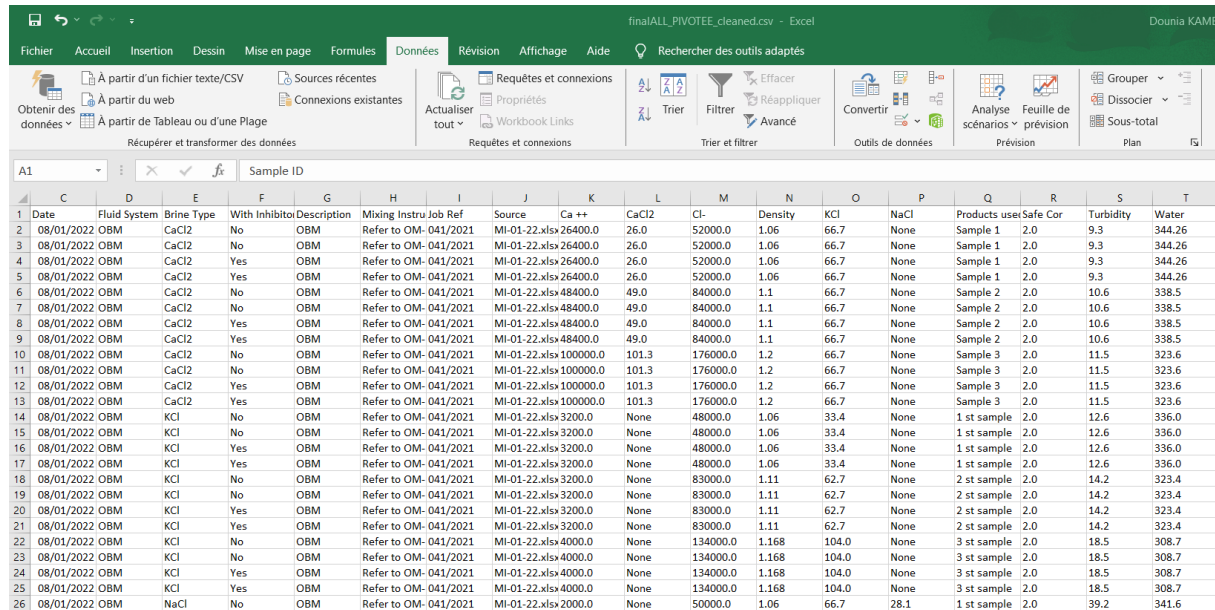
Avant de procéder à l'analyse, un nettoyage complet des données a été réalisé à l'aide d'un script Python. L'algorithme suivant résume les principales étapes de prétraitement appliquées au fichier brut.

Algorithm 1 Prétraitement des données brutes (fichier `finalALL_PIVOTEE.csv`)

Require: Fichier brut au format CSV contenant les données forage

Ensure: Fichier nettoyé et enrichi, prêt pour la modélisation

- 1: Charger le fichier CSV dans un DataFrame `df`
 - 2: **if** la colonne `Formulation Label` existe **then**
 - 3: La supprimer
 - 4: **end if**
 - 5: Créer une nouvelle colonne `Tags` à partir de la colonne `Description` :
 - 6: Détection des mots-clés : OBM, SBM, WBM, KCl, NaCl, Polymer
 - 7: Nettoyer la colonne `Job Ref` en extrayant les identifiants au format `MI-XX-XX`
 - 8: Extraire l'année à partir de la colonne `Date` et l'ajouter comme colonne `Année`
 - 9: Pour chaque colonne numérique :
 - 10: Remplir les valeurs manquantes par la moyenne du groupe `Fluid System`
 - 11: Remplir les valeurs restantes par 0 si nécessaire
 - 12: Pour chaque colonne textuelle :
 - 13: Remplacer les valeurs manquantes par "None"
 - 14: Enregistrer le DataFrame nettoyé dans un nouveau fichier `finalALL_PIVOTEE_cleaned.csv`
-



	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Date	Fluid System	Brine Type	With Inhibitor	Description	Mixing Instru	Job Ref	Source	Ca ++	CaCl2	Cl-	Density	KCl	NaCl	Products use	Safe Cor	Turbidity	Water
2	08/01/2022	OBM	CaCl2	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx26400.0	26.0	52000.0	1.06	66.7	None	Sample 1	2.0	9.3	344.26	
3	08/01/2022	OBM	CaCl2	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx26400.0	26.0	52000.0	1.06	66.7	None	Sample 1	2.0	9.3	344.26	
4	08/01/2022	OBM	CaCl2	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx26400.0	26.0	52000.0	1.06	66.7	None	Sample 1	2.0	9.3	344.26	
5	08/01/2022	OBM	CaCl2	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx26400.0	26.0	52000.0	1.06	66.7	None	Sample 1	2.0	9.3	344.26	
6	08/01/2022	OBM	CaCl2	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx48400.0	49.0	84000.0	1.1	66.7	None	Sample 2	2.0	10.6	338.5	
7	08/01/2022	OBM	CaCl2	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx48400.0	49.0	84000.0	1.1	66.7	None	Sample 2	2.0	10.6	338.5	
8	08/01/2022	OBM	CaCl2	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx48400.0	49.0	84000.0	1.1	66.7	None	Sample 2	2.0	10.6	338.5	
9	08/01/2022	OBM	CaCl2	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx48400.0	49.0	84000.0	1.1	66.7	None	Sample 2	2.0	10.6	338.5	
10	08/01/2022	OBM	CaCl2	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx100000.0	101.3	176000.0	1.2	66.7	None	Sample 3	2.0	11.5	323.6	
11	08/01/2022	OBM	CaCl2	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx100000.0	101.3	176000.0	1.2	66.7	None	Sample 3	2.0	11.5	323.6	
12	08/01/2022	OBM	CaCl2	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx100000.0	101.3	176000.0	1.2	66.7	None	Sample 3	2.0	11.5	323.6	
13	08/01/2022	OBM	CaCl2	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx100000.0	101.3	176000.0	1.2	66.7	None	Sample 3	2.0	11.5	323.6	
14	08/01/2022	OBM	KCl	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx3200.0	None	48000.0	1.06	33.4	None	1 st sample	2.0	12.6	336.0	
15	08/01/2022	OBM	KCl	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx3200.0	None	48000.0	1.06	33.4	None	1 st sample	2.0	12.6	336.0	
16	08/01/2022	OBM	KCl	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx3200.0	None	48000.0	1.06	33.4	None	1 st sample	2.0	12.6	336.0	
17	08/01/2022	OBM	KCl	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx3200.0	None	48000.0	1.06	33.4	None	1 st sample	2.0	12.6	336.0	
18	08/01/2022	OBM	KCl	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx3200.0	None	83000.0	1.11	62.7	None	2 st sample	2.0	14.2	323.4	
19	08/01/2022	OBM	KCl	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx3200.0	None	83000.0	1.11	62.7	None	2 st sample	2.0	14.2	323.4	
20	08/01/2022	OBM	KCl	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx3200.0	None	83000.0	1.11	62.7	None	2 st sample	2.0	14.2	323.4	
21	08/01/2022	OBM	KCl	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx3200.0	None	83000.0	1.11	62.7	None	2 st sample	2.0	14.2	323.4	
22	08/01/2022	OBM	KCl	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx4000.0	None	134000.0	1.168	104.0	None	3 st sample	2.0	18.5	308.7	
23	08/01/2022	OBM	KCl	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx4000.0	None	134000.0	1.168	104.0	None	3 st sample	2.0	18.5	308.7	
24	08/01/2022	OBM	KCl	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx4000.0	None	134000.0	1.168	104.0	None	3 st sample	2.0	18.5	308.7	
25	08/01/2022	OBM	KCl	Yes	OBM	Refer to OM-	041/2021	MI-01-22.xlsx4000.0	None	134000.0	1.168	104.0	None	3 st sample	2.0	18.5	308.7	
26	08/01/2022	OBM	NaCl	No	OBM	Refer to OM-	041/2021	MI-01-22.xlsx2000.0	None	50000.0	1.06	66.7	28.1	1 st sample	2.0	39.2	341.6	

FIGURE 7.4 – Extrait du fichier de données nettoyé utilisé pour l'entraînement du modèle.

7.2.4 Centralisation et stockage

Après le nettoyage et la normalisation des données, celles-ci ont été centralisées et stockées sous forme de fichier CSV. Cette étape est cruciale pour garantir une gestion efficace des données avant qu'elles ne soient utilisées pour l'entraînement du modèle de machine learning. Le fichier final contient toutes les caractéristiques pertinentes des fluides de forage, comme la densité, la viscosité, le pH, etc. Ces données, une fois centralisées et stockées, peuvent être facilement accessibles pour l'analyse et l'entraînement du modèle.

L'objectif principal de cette étape est d'assurer une organisation optimale des données, ce qui facilite leur exploitation et garantit qu'elles sont prêtes à être utilisées de manière cohérente et efficace tout au long du processus d'analyse.

Le fichier final `synthetic_drilling_fluid_data_enriched.csv` contient ainsi l'ensemble des variables critiques, prêtes pour l'entraînement du modèle.

Objectifs de la centralisation :

- **Accessibilité facile** : toutes les données sont stockées dans un format CSV structuré, permettant un accès rapide et une exploitation optimale.
- **Organisation cohérente** : l'utilisation d'un fichier unique centralisé garantit une structure uniforme et lisible.
- **Préparation au modèle** : les données sont directement prêtes pour l'analyse prédictive et la détection d'anomalies.

Étapes du processus de centralisation et de stockage :

La centralisation des données suit plusieurs étapes afin de garantir que les informations sont organisées et prêtes à être utilisées. Voici les principales étapes :

- **Fusion des jeux de données** : Les données extraites des différents fichiers Excel ont été fusionnées en un seul tableau centralisé. Chaque échantillon de fluide a été enrichi avec des informations contextuelles, telles que le type de fluide, la date de prélèvement, et d'autres métadonnées essentielles.
- **Création d'un tableau pivot** : Après la fusion, un tableau pivot a été créé pour structurer les données sous une forme lisible et organisée. Les différentes mesures des fluides (comme la densité, la viscosité, etc.) ont été associées à chaque échantillon de fluide, avec des colonnes spécifiques pour chaque test effectué.
- **Stockage dans un fichier CSV** : Le fichier final a été sauvegardé sous le format CSV afin d'assurer une compatibilité optimale avec les outils d'analyse et d'entraînement de modèles, tout en permettant une gestion facile des données.

La phase suivante consiste à transformer les données longues en un tableau centralisé, où chaque ligne correspond à un échantillon unique et chaque colonne à un paramètre mesuré. L'algorithme ci-dessous décrit ce processus.

Algorithm 2 Construction d'un tableau centralisé des échantillons

Require: Fichier `finalALL_PIVOTEE_cleaned.csv` contenant les données nettoyées

Ensure: Fichier enrichi sous forme de tableau pivoté, avec un échantillon par ligne

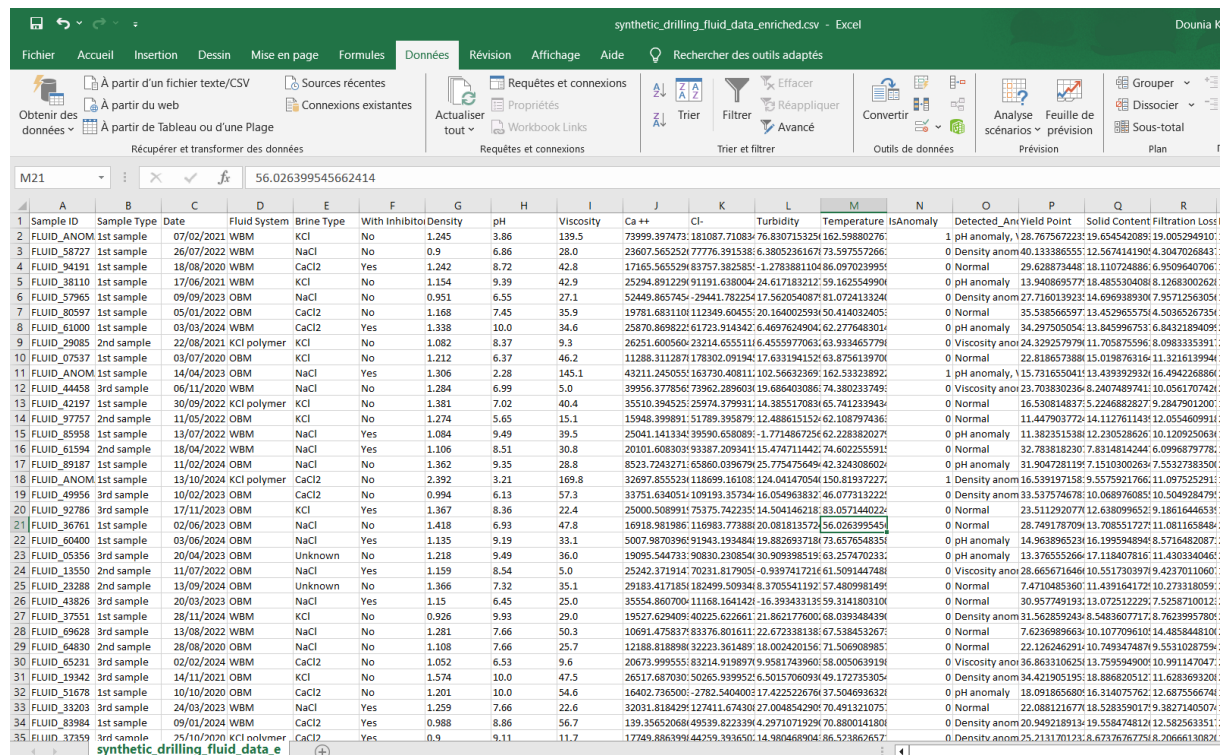
- 1: Charger les données nettoyées dans un DataFrame `df`
 - 2: Créer un identifiant unique pour chaque échantillon :
 - 3: `Sample ID = Formulation Label + Sample Type`
 - 4: Construire un tableau pivoté avec :
 - 5: Index : `Sample ID, Date, Fluid System`
 - 6: Colonnes : types de tests (`Test`)
 - 7: Valeurs : résultats numériques (`Value`)
 - 8: Fonction d'agrégation : `first` (valeur unique par combinaison)
 - 9: Réinitialiser les index pour obtenir un tableau plat
 - 10: Enregistrer le résultat dans le fichier `synthetic_drilling_fluid_data_enriched.csv`
-

Explication du code :

Le code présenté ci-dessus décrit les étapes utilisées pour centraliser et stocker les données :

1. Le fichier `finalALL_PIVOTEE_cleaned.csv` est chargé dans un `DataFrame` `pandas`.
2. Une colonne `Sample ID` est générée en combinant les informations de `Formulation Label` et `Sample Type`, ce qui permet d'identifier chaque échantillon de fluide de manière unique.
3. Un tableau pivot est créé en utilisant la fonction `pivot_table` de `pandas`, afin d'organiser les mesures sous une forme plus lisible et structurée.
4. Enfin, le fichier final est exporté sous le format `CSV` avec la commande `to_csv`, ce qui permet de stocker toutes les données centralisées dans un fichier compatible avec les étapes suivantes du processus de modélisation.

Visualisation des données après centralisation : Voici un aperçu des données centralisées, montrant la structure après la création du tableau pivot et avant l'exportation en fichier `CSV`.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Sample ID	Sample Type	Date	Fluid System	Brine Type	With Inhibitor	Density	pH	Viscosity	Ca ++	Cl-	Turbidity	Temperature	IsAnomaly	Detected	AnYield Point	Solid Content	Filtration Loss
2	FLUID_ANOM	1st sample	07/02/2021	WBM	KCl	No	1.245	3.86	139.5	73999.397473	181087.71083	76.830715325	162.59880276	1	pH anomaly	128.76756223	19.654542089	19.005294910
3	FLUID_58727	1st sample	26/07/2022	WBM	NaCl	No	0.9	6.86	28.0	23607.565252	77776.391538	6.3805236167	73.597557266	0	Density anom	40.133386555	12.567414190	4.3047026843
4	FLUID_94191	1st sample	18/08/2020	WBM	CaCl2	Yes	1.242	8.72	42.8	17165.565529	83757.382585	-1.2783881104	86.097023995	0	Normal	29.628873448	18.110724886	6.9509640706
5	FLUID_38110	1st sample	17/06/2021	WBM	KCl	No	1.154	9.39	42.9	25294.891229	91191.638004	24.617183212	59.162554990	0	pH anomaly	13.940868957	18.485530408	1.268300262
6	FLUID_57965	1st sample	09/09/2023	OBM	NaCl	No	0.951	6.55	27.1	52449.865745	-29441.782254	17.562054087	81.072413324	0	Density anom	27.716013923	14.696938990	3.7571256305
7	FLUID_80597	1st sample	05/01/2022	OBM	CaCl2	No	1.168	7.45	35.9	19781.683108	112349.69425	20.164002939	50.414032405	0	Normal	35.538566597	13.452965575	4.5036526754
8	FLUID_61000	1st sample	03/03/2024	WBM	CaCl2	Yes	1.338	10.0	34.6	23870.869822	61723.914342	6.4697624904	62.277648301	0	pH anomaly	34.297505054	13.845986753	6.8432189409
9	FLUID_29085	2nd sample	22/08/2021	KCl polymer	KCl	No	1.082	8.37	9.3	26251.600569	23214.655511	6.4555977063	63.933465779	0	Viscosity ano	24.329257979	11.705875596	8.098333591
10	FLUID_07537	1st sample	03/07/2020	OBM	KCl	No	1.212	6.37	46.2	11288.311287	178302.09194	17.633194152	63.875613970	0	Normal	22.818657388	15.01987316	11.321613994
11	FLUID_ANOM	1st sample	14/04/2023	OBM	NaCl	Yes	1.306	2.28	145.1	43211.245055	163730.40811	102.56632369	162.5323892	1	pH anomaly	15.731655041	13.439392932	16.494226886
12	FLUID_44458	3rd sample	06/11/2020	WBM	NaCl	No	1.284	6.99	5.0	39956.377856	73962.289630	19.686403086	74.380233749	0	Viscosity ano	23.703830236	8.2407489741	10.056170742
13	FLUID_42197	1st sample	30/09/2022	KCl polymer	KCl	No	1.381	7.02	40.4	35510.394525	25974.379931	14.385517083	65.741233943	0	Normal	16.530814837	5.2246882827	9.2847901200
14	FLUID_97757	2nd sample	11/05/2022	OBM	KCl	No	1.274	5.65	15.1	15948.399891	51789.395879	12.488615152	62.108797436	0	Normal	11.447903772	14.112761143	12.055460991
15	FLUID_85958	1st sample	13/07/2022	WBM	NaCl	Yes	1.084	9.49	39.5	25041.141334	39590.658089	-1.7714867256	62.228382027	0	pH anomaly	11.382351538	12.230528626	10.102925063
16	FLUID_61594	2nd sample	18/04/2022	WBM	NaCl	Yes	1.106	8.51	30.8	20101.608303	93387.209341	15.474714442	74.60225591	0	Normal	32.783818230	7.8314814244	6.0996879778
17	FLUID_89187	1st sample	11/02/2024	OBM	NaCl	No	1.362	9.35	28.8	8523.724327	165860.039679	25.775475649	42.324308602	0	pH anomaly	31.904728119	7.1510300263	7.5532738350
18	FLUID_ANOM	1st sample	13/10/2024	KCl polymer	CaCl2	No	2.392	3.21	169.8	32697.855523	118699.16108	124.04147054	150.81937227	1	Density anom	16.539197158	9.5575921766	11.097525291
19	FLUID_49956	3rd sample	10/02/2023	OBM	CaCl2	No	0.994	6.13	57.3	33751.634051	109193.35734	16.054963832	46.077313222	0	Density anom	33.537574678	10.068976085	10.504928479
20	FLUID_92786	3rd sample	17/11/2023	OBM	KCl	Yes	1.367	8.36	22.4	25000.508991	75375.742235	14.504146218	83.057144022	0	Normal	23.511292077	12.638099652	9.1861644653
21	FLUID_36761	1st sample	02/06/2023	OBM	NaCl	No	1.418	6.93	47.8	16918.981986	116983.77388	20.081813572	56.026399545	0	Normal	28.749178709	13.708551727	11.081165848
22	FLUID_60400	1st sample	03/06/2024	OBM	NaCl	Yes	1.135	9.19	33.1	5007.9870396	91943.193484	19.882693718	73.657645835	0	pH anomaly	14.963895523	16.199594884	8.5716432087
23	FLUID_05356	3rd sample	20/04/2023	OBM	Unknown	No	1.218	9.49	36.0	19095.544733	90830.230854	30.909398519	63.25740233	0	pH anomaly	13.76555266	17.118407816	11.430334046
24	FLUID_13550	2nd sample	11/07/2022	OBM	NaCl	Yes	1.159	8.54	5.0	25242.371914	70231.817905	-0.9397417216	61.509144748	0	Viscosity ano	28.665671648	10.5517303971	9.4237011060
25	FLUID_23288	2nd sample	13/09/2024	OBM	Unknown	No	1.366	7.32	35.1	29183.417185	162499.50934	8.3705541192	57.480998149	0	Normal	7.471048580	11.439164172	10.273318059
26	FLUID_43826	3rd sample	20/03/2023	OBM	NaCl	Yes	1.15	6.45	25.0	35554.960709	11186.164142	-16.39343315	59.314180310	0	Normal	30.93749193	13.072512229	7.528701002
27	FLUID_37551	1st sample	28/11/2024	WBM	KCl	No	0.926	9.93	29.0	1957.629409	40225.622651	21.863177600	68.039348439	0	Density anom	31.562859249	8.548360771	8.782395780
28	FLUID_69628	3rd sample	13/08/2022	WBM	NaCl	No	1.281	7.66	50.3	10691.475837	93376.801611	22.672338138	67.538453267	0	Normal	7.623689662	10.107709610	14.485844510
29	FLUID_64830	2nd sample	28/08/2020	OBM	NaCl	No	1.108	7.66	25.7	12188.818988	32223.361489	18.002401556	71.50690895	0	Normal	22.126346291	10.749347487	9.5531028759
30	FLUID_65231	3rd sample	02/02/2024	WBM	CaCl2	No	1.052	6.53	9.6	20673.399555	93214.519897	9.9581743960	58.005063919	0	Viscosity ano	36.863310625	13.795949001	10.991147047
31	FLUID_19342	3rd sample	14/11/2021	OBM	KCl	No	1.574	10.0	47.5	26517.687030	90265.39952	6.5015706093	49.172735305	0	Density anom	34.421905195	13.886820512	11.628369201
32	FLUID_51678	1st sample	10/10/2020	OBM	CaCl2	No	1.201	10.0	54.6	16402.736500	-2782.540400	17.422522676	37.504693632	0	pH anomaly	18.09186580	16.314075762	12.687556674
33	FLUID_33203	3rd sample	24/03/2023	WBM	NaCl	Yes	1.259	7.66	22.6	32031.818429	127411.674301	27.048542905	70.491321075	0	Normal	22.088216717	18.528390179	3.827140507
34	FLUID_83984	1st sample	09/01/2024	OBM	CaCl2	Yes	0.988	8.86	56.7	139.35652068	49539.822339	4.2971071929	70.880014180	0	Density anom	20.949218913	19.558474812	12.582563351
35	FLUID_37359	3rd sample	25/10/2020	KCl polymer	CaCl2	Yes	0.9	9.11	11.7	17749.886399	44259.393650	14.980468904	86.523862657	0	Density anom	25.213170123	8.6737676775	8.2066613082

FIGURE 7.5 – Tableau des données après centralisation et préparation pour l'entraînement du modèle.

Conclusion sur la centralisation et le stockage :

La centralisation et le stockage des données dans un fichier unique au format `CSV` ont permis d'organiser l'ensemble des informations de manière cohérente, structurée et facilement exploitable pour les étapes suivantes du projet. Cette étape garantit que toutes les variables pertinentes – telles que la densité, la viscosité, le pH, la turbidité ou encore le Oil Water Ratio – sont consolidées dans un format standardisé, prêt pour l'analyse.

Grâce à ce jeu de données enrichi et centralisé, il a été possible d'entreprendre une **analyse**

exploratoire approfondie, présentée dans la section suivante, afin d'identifier les profils d'anomalies par type de fluide, d'examiner les corrélations entre variables, et de visualiser les distributions de paramètres critiques.

7.2.5 Analyse exploratoire et visualisation des anomalies

Avant d'entraîner le modèle de détection, une phase d'analyse exploratoire des données a été réalisée à l'aide de boxplots, heatmaps de corrélation, et statistiques par classes, afin de mieux comprendre la répartition des anomalies selon différents critères, d'identifier les variables influentes, et de détecter d'éventuels patterns liés à certains types de fluides. Les visualisations suivantes ont permis de guider les choix de modélisation et de valider les hypothèses métier.

a) Répartition des anomalies par type de fluide

La Figure 7.6 présente le nombre d'anomalies détectées pour chaque type de fluide. Cette analyse permet d'identifier les fluides les plus sujets aux anomalies.

- **OBM (Oil-Based Mud)** : représente la majorité des anomalies détectées.
- **WBM (Water-Based Mud)** : arrive en seconde position avec un volume significatif.
- **KCl Polymer et Brine** : présentent peu d'anomalies.

Ainsi, les fluides OBM et WBM nécessitent une attention particulière, notamment en lien avec des paramètres comme la turbidité et le Oil Water Ratio.

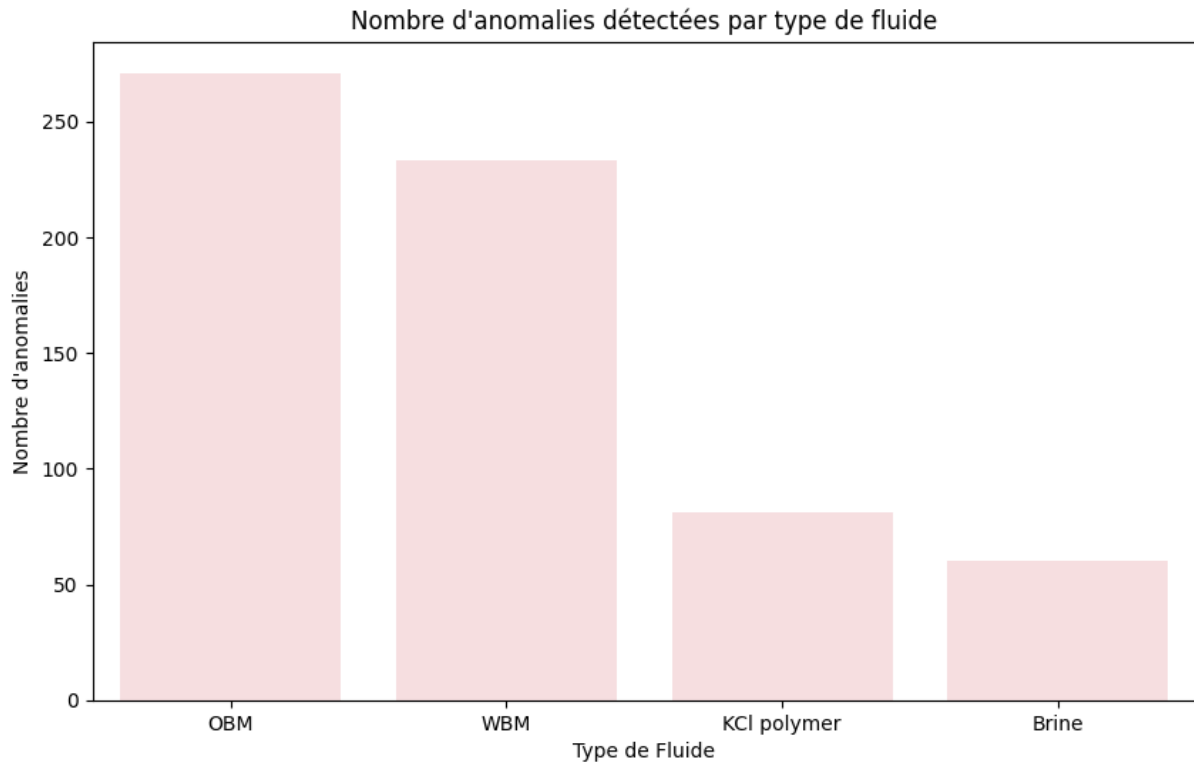


FIGURE 7.6 – Nombre d’anomalies détectées par type de fluide.

Une anomalie correspond ici à un échantillon identifié comme non conforme par les règles métier ou les experts. Le graphique montre le nombre total d’anomalies détectées par catégorie de fluide.

b) Corrélation entre variables

La figure 7.7 présente la matrice de corrélation entre les différentes variables mesurées dans les échantillons de fluides de forage. Quelques observations clés :

- **Viscosité – Turbidité (0,56)** : Corrélation positive attendue. Une plus forte turbidité (présence de particules solides) augmente la viscosité du fluide.
- **Viscosité – Température (0,50)** : Corrélation modérée, traduisant l’effet de la température sur le comportement rhéologique du fluide.
- **Turbidité – Température (0,58)** : Corrélation significative, indiquant que les hautes températures favorisent la mise en suspension de particules.
- **IsAnomaly – autres variables (< 0,16)** : Faibles corrélations, ce qui confirme que les anomalies sont le résultat de plusieurs facteurs combinés plutôt que d’une seule variable.

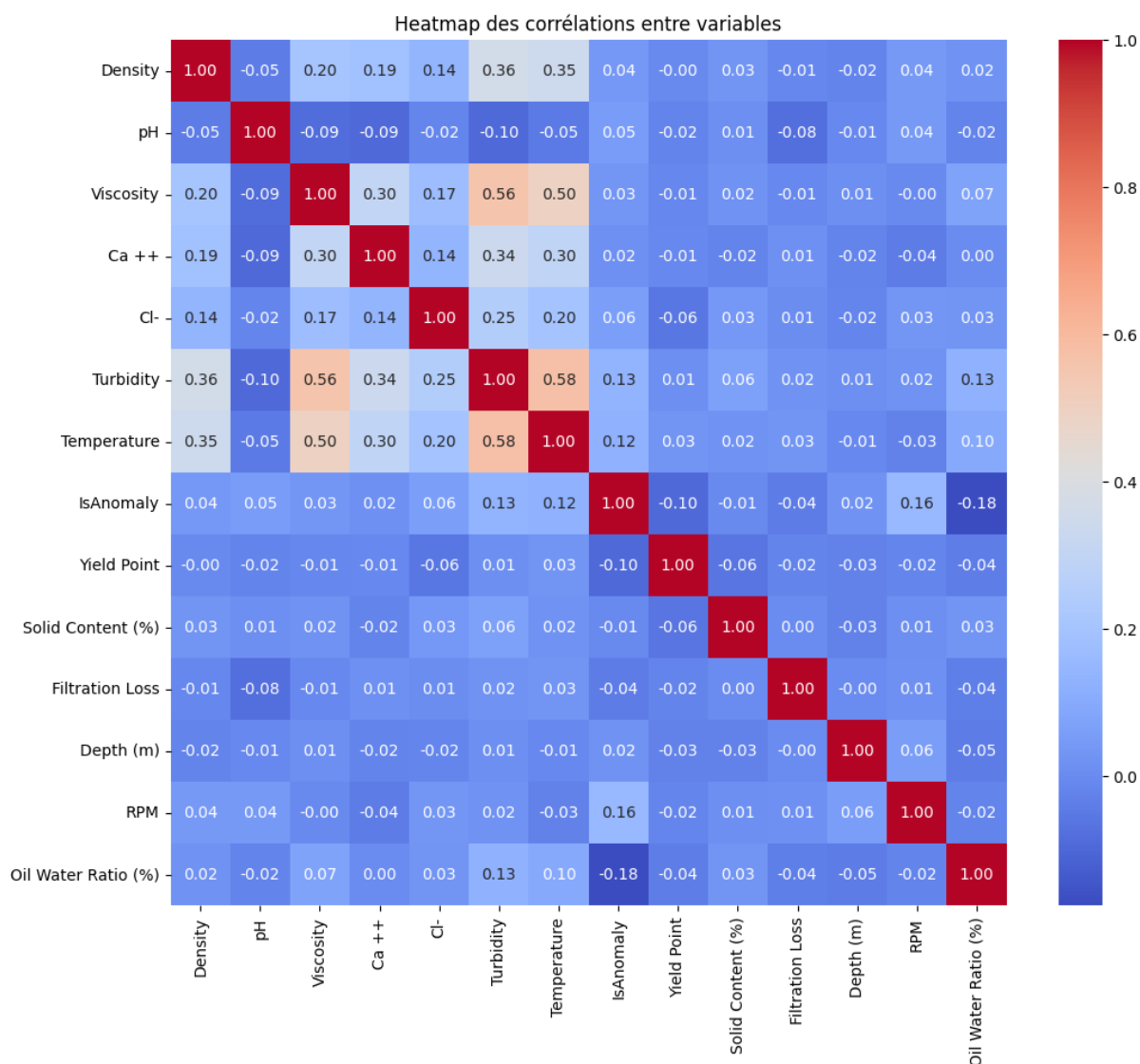


FIGURE 7.7 – Heatmap des corrélations entre variables.

Cette heatmap permet de visualiser les corrélations linéaires entre variables (Pearson). Elle aide à identifier les redondances (corrélations > 0.8) et à guider la sélection de variables pour éviter la multicolinéarité lors de l'entraînement.

c) Distribution de la turbidité dans les anomalies

La Figure 7.8 présente la distribution de la turbidité pour les cas considérés comme anomalies.

- La majorité des anomalies se situent entre **10 et 40 NTU**.
- Quelques cas extrêmes sont présents, signalant une contamination critique.

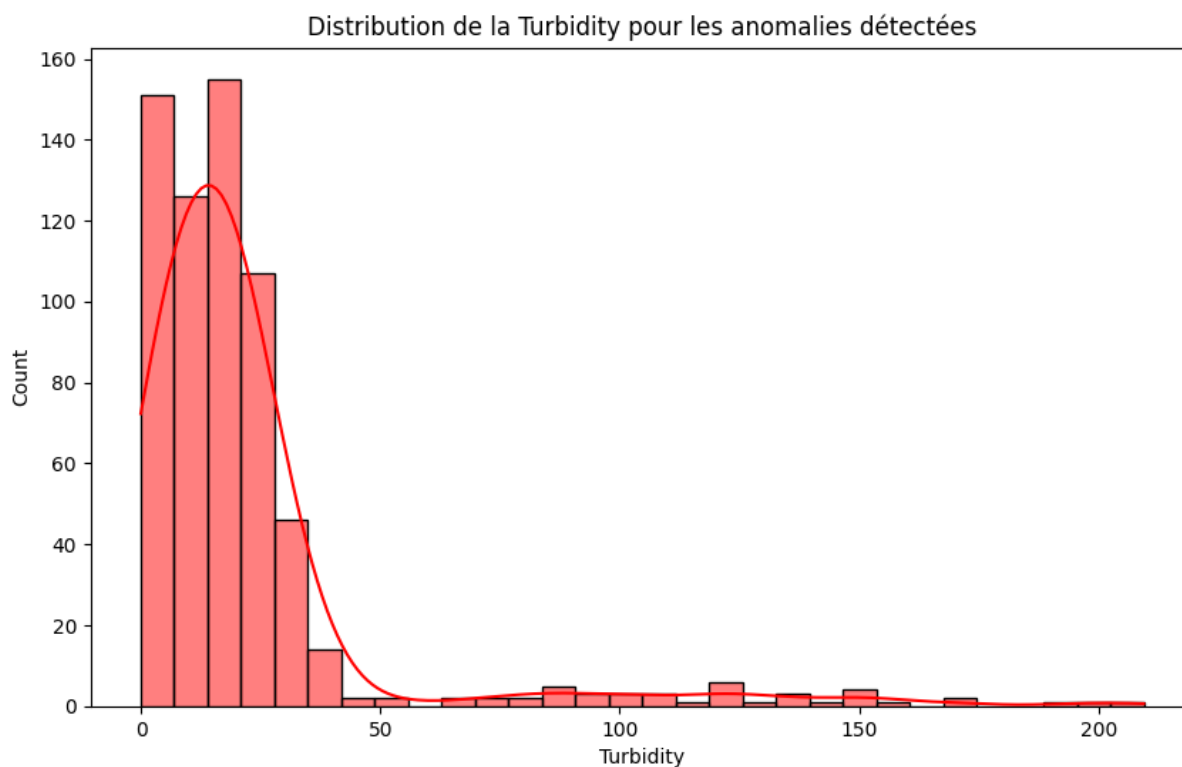


FIGURE 7.8 – Distribution de la turbidité pour les anomalies détectées.

d) Répartition du Oil Water Ratio (OWR) pour OBM

La Figure 7.9 illustre la répartition du OWR, variable critique pour les fluides OBM.

- Le OWR est principalement compris entre **60% et 90%**, avec une concentration autour de **70–80%**.
- Un OWR trop bas ($< 65\%$) indique un excès d'eau, un OWR trop élevé ($> 90\%$) un excès d'huile.

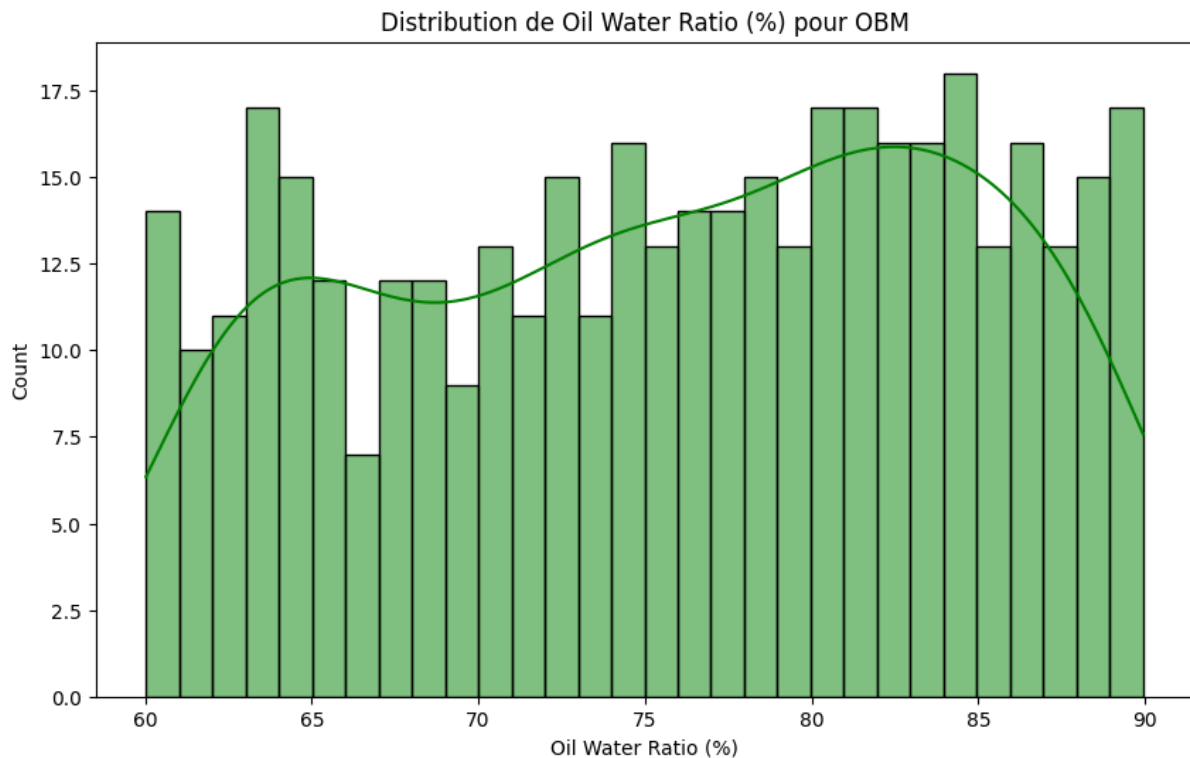


FIGURE 7.9 – Distribution du Oil Water Ratio (%) pour les fluides OBM.

e) Synthèse des visualisations

Cette analyse exploratoire a permis de :

- Identifier les types de fluides les plus affectés (OBM, WBM).
- Mettre en évidence les variables les plus critiques (turbidité, OWR).
- Confirmer la nature multi-dimensionnelle des anomalies, justifiant l'utilisation d'un algorithme comme *Random Forest*.

Colonnes finales conservées et justification

Après les étapes de nettoyage, d'enrichissement et de normalisation des données, et à la lumière de l'analyse exploratoire effectuée, les colonnes suivantes ont été conservées dans le fichier final centralisé :

- **Identification et contexte** : *Sample ID, Sample Type, Date, Fluid System, Brine Type, With Inhibitor*
- **Propriétés physico-chimiques du fluide** : *Density, pH, Viscosity, Turbidity, Temperature, Yield Point, Ca^{++} , Cl^- , Solid Content (%), Filtration Loss, Oil Water Ratio (%)*
- **Paramètres opérationnels** : *RPM, Depth (m)*
- **Supervision** : *IsAnomaly, DetectedAnomalies*

Justification du choix : Ces variables ont été retenues pour leur importance dans le comportement des fluides de forage et leur pertinence dans le processus de surveillance.

Les caractéristiques physico-chimiques (comme la densité, la turbidité ou la viscosité) permettent d'évaluer la stabilité et la performance du fluide. Des paramètres comme le Yield Point ou le Oil Water Ratio (%) sont cruciaux pour détecter les déséquilibres de formulation dans les systèmes OBM/WBM.

Les colonnes *RPM* et *Depth (m)* apportent un contexte d'analyse opérationnel, en reliant chaque mesure aux conditions spécifiques du puits. Enfin, les champs *IsAnomaly* et *DetectedAnomalies* servent de base pour la supervision des futures étapes d'analyse. Ce choix de variables a été validé par un ingénieur fluides de l'ADFC, afin de garantir la cohérence technique avec les pratiques industrielles et les priorités métier en matière de détection d'anomalies.

Variable	Description
Sample ID	Identifiant unique généré pour chaque échantillon de fluide.
Sample Type	Type d'échantillon analysé (ex. : 1 ^{er} passage, contrôle, etc.).
Date	Date de prélèvement ou d'analyse.
Fluid System	Système de fluide utilisé (OBM, WBM, Brine, etc.).
Brine Type	Type de solution saline utilisée (NaCl, KCl, CaCl ₂ , etc.).
With Inhibitor	Présence ou non d'un inhibiteur dans le fluide.
Density	Masse volumique du fluide, déterminant la pression hydrostatique.
pH	Niveau d'acidité ou de basicité du fluide.
Viscosity	Capacité du fluide à transporter les déblais.
Turbidity	Degré de contamination par des particules en suspension.
Temperature	Température de l'échantillon au moment de la mesure.
Yield Point	Capacité du fluide à initier le mouvement de suspension des solides.
Ca++, Cl-	Concentrations ioniques indicatives de contaminations.
Solid Content (%)	Taux de matière solide présente dans le fluide.
Filtration Loss	Quantité de fluide perdue à travers une formation poreuse.
Oil Water Ratio (%)	Rapport huile/eau (particulièrement important pour les OBM).
Depth (m)	Profondeur de prélèvement de l'échantillon.
RPM	Vitesse de rotation au moment de l'échantillonnage.
IsAnomaly	Indicateur binaire : 1 = anomalie, 0 = normal.
DetectedAnomalies	Description textuelle des anomalies détectées.

TABLE 7.1 – Variables conservées et leur signification

7.3 Modélisation et entraînement

7.3.1 Choix de l'algorithme : Random Forest

Dans le cadre de ce projet, le choix de l'algorithme de machine learning s'est révélé déterminant pour garantir une détection précise et fiable des anomalies dans les fluides de forage. Plusieurs modèles ont été testés, notamment des algorithmes de type *Isolation Forest*, *régression logistique*, et *réseaux de neurones (LSTM)*. Toutefois, leurs performances ont été jugées insuffisantes pour notre problématique (voir résultats comparatifs en **Annexe**).

À la suite de cette évaluation, l'algorithme **Random Forest** s'est imposé comme le plus performant. Il s'agit d'un modèle d'ensemble basé sur la combinaison de plusieurs arbres de décision, offrant robustesse, flexibilité et interprétabilité.

Justification du choix de Random Forest Le modèle Random Forest présente plusieurs avantages notables dans notre contexte :

- **Robustesse** : il est résistant au bruit et à l'incomplétude des données. L'agrégation de plusieurs arbres réduit le risque de surapprentissage (*overfitting*) tout en améliorant la généralisation ;
- **Gestion du déséquilibre** : notre jeu de données présente un fort déséquilibre entre les classes. Le modèle a été renforcé avec la technique de rééchantillonnage *SMOTE* (*Synthetic Minority Over-sampling Technique*) pour améliorer la détection des cas rares ;
- **Flexibilité** : Random Forest prend en charge un grand nombre de variables et capture les interactions non linéaires complexes entre les caractéristiques physico-chimiques ;
- **Interprétabilité** : il permet d'évaluer l'importance des variables dans les décisions prises, facilitant ainsi la compréhension du comportement du modèle.

Implémentation et environnement technique L'implémentation a été réalisée en **Python**, à l'aide des bibliothèques suivantes :

- **pandas** : gestion et manipulation des données ;
- **scikit-learn** : modélisation Random Forest, validation croisée, évaluation ;
- **imblearn** : rééquilibrage des classes avec **SMOTE** ;
- **joblib** : utilisée pour sauvegarder le modèle entraîné sous forme de fichier .pkl, afin de l'importer rapidement dans l'application Streamlit lors de l'exécution.

L'entraînement du modèle inclut une normalisation des variables, un rééquilibrage des classes via la méthode SMOTE, puis l'apprentissage supervisé à l'aide d'un Random Forest. L'algorithme ci-dessous résume cette procédure.

Algorithm 3 Entraînement du modèle Random Forest avec rééquilibrage des classes

Require: Données d'entraînement X_{train} , étiquettes y_{train}

Ensure: Modèle entraîné et sauvegardé dans un fichier

- 1: Normaliser les données à l'aide d'un `StandardScaler`
 - 2: Appliquer le suréchantillonnage avec `SMOTE` pour équilibrer les classes
 - 3: Initialiser un classifieur `RandomForest` avec 100 arbres
 - 4: Entraîner le modèle sur les données rééchantillonnées
 - 5: Sauvegarder le modèle entraîné dans un fichier `rf_model_final.pkl` avec `joblib`
-

Ce pipeline complet a été intégré dans l'application web développée avec Streamlit, afin d'offrir une solution interactive de détection d'anomalies adaptée au contexte industriel.

7.3.2 Validation croisée et optimisation des hyperparamètres

Pour garantir la robustesse du modèle Random Forest et éviter le surapprentissage, une **validation croisée à 5 plis** (*5-fold cross-validation*) a été appliquée. Cette méthode consiste à diviser le jeu de données en cinq sous-ensembles, puis à entraîner le modèle successivement sur quatre d'entre eux et à l'évaluer sur le cinquième. Ce processus est répété cinq fois, en changeant le pli de validation à chaque itération, ce qui permet d'**estimer plus finement la capacité de généralisation du modèle**.

Parallèlement, une optimisation des hyperparamètres a été menée à l'aide de `GridSearchCV`, une méthode d'exploration systématique testant plusieurs combinaisons d'hyperparamètres. L'objectif est de **maximiser une métrique de performance** (ici, le **F1-score macro**, plus adapté aux classes déséquilibrées).

Les hyperparamètres explorés comprenaient :

- `n_estimators` : nombre d'arbres dans la forêt,
- `max_depth` : profondeur maximale des arbres,
- `min_samples_split` : nombre minimal d'échantillons requis pour diviser un nœud.

La configuration optimale obtenue est :

```
n_estimators = 200, max_depth = 20, min_samples_split = 5
```

Cette combinaison a permis d'améliorer la capacité de détection du modèle tout en assurant une **bonne généralisation sur les données de test**. La validation croisée a donc joué un rôle essentiel pour **éviter une évaluation biaisée** liée à une simple séparation *train/test*.

7.3.3 Évaluation et résultats du modèle Random Forest

Après l'entraînement du modèle Random Forest avec validation croisée et optimisation des hyperparamètres, une évaluation a été menée sur les données de test afin de mesurer ses performances réelles.

Matrice de confusion La Figure 7.10 présente la matrice de confusion obtenue. Le modèle a correctement classé la quasi-totalité des échantillons, avec seulement quelques erreurs mineures :

- 2 faux positifs (échantillons normaux classés à tort comme anomalies);
- 1 faux négatif (anomalie non détectée).

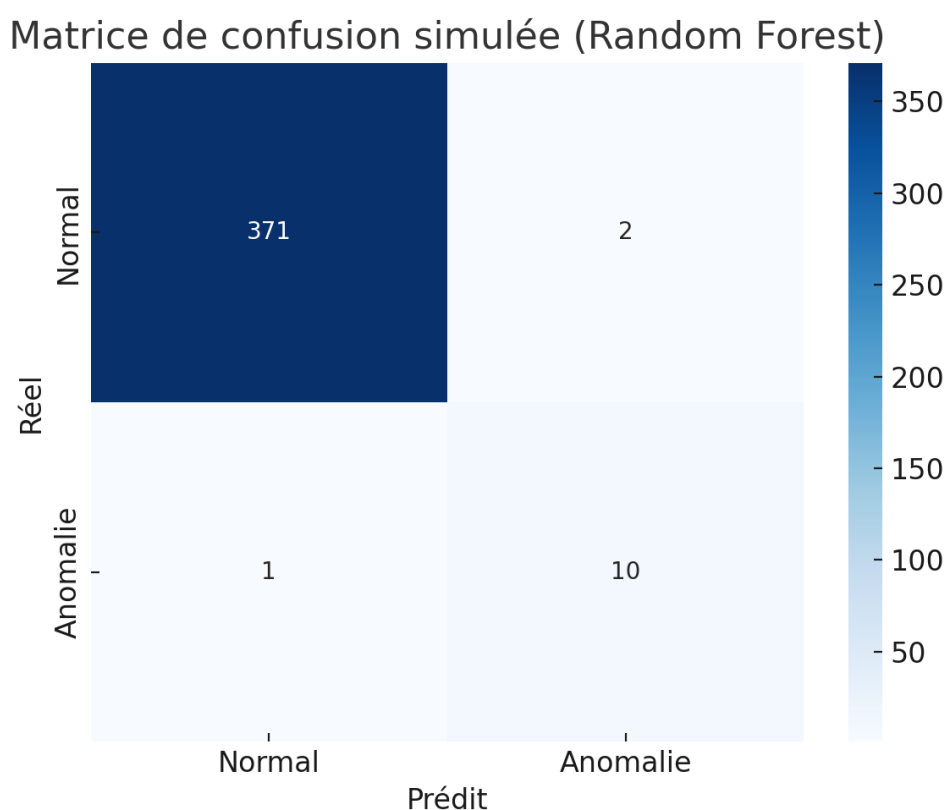


FIGURE 7.10 – Matrice de confusion du modèle Random Forest

Rapport de classification Le tableau suivant présente les principales métriques obtenues :

Classe	Précision	Rappel	F1-score	Support
0 – Normal	0.99	0.99	0.99	373
1 – Anomalie	0.83	0.91	0.87	11
Accuracy		0.992 (384 échantillons)		

TABLE 7.2 – Rapport de classification du modèle Random Forest

Ces résultats témoignent d’une **excellente capacité de détection**, tout en conservant un taux d’erreurs très faible.

Courbe ROC La courbe ROC affichée en Figure 7.11 permet de visualiser la capacité discriminante du modèle. L’aire sous la courbe (*AUC*) atteint **0.97**, ce qui traduit une séparation efficace entre les classes.

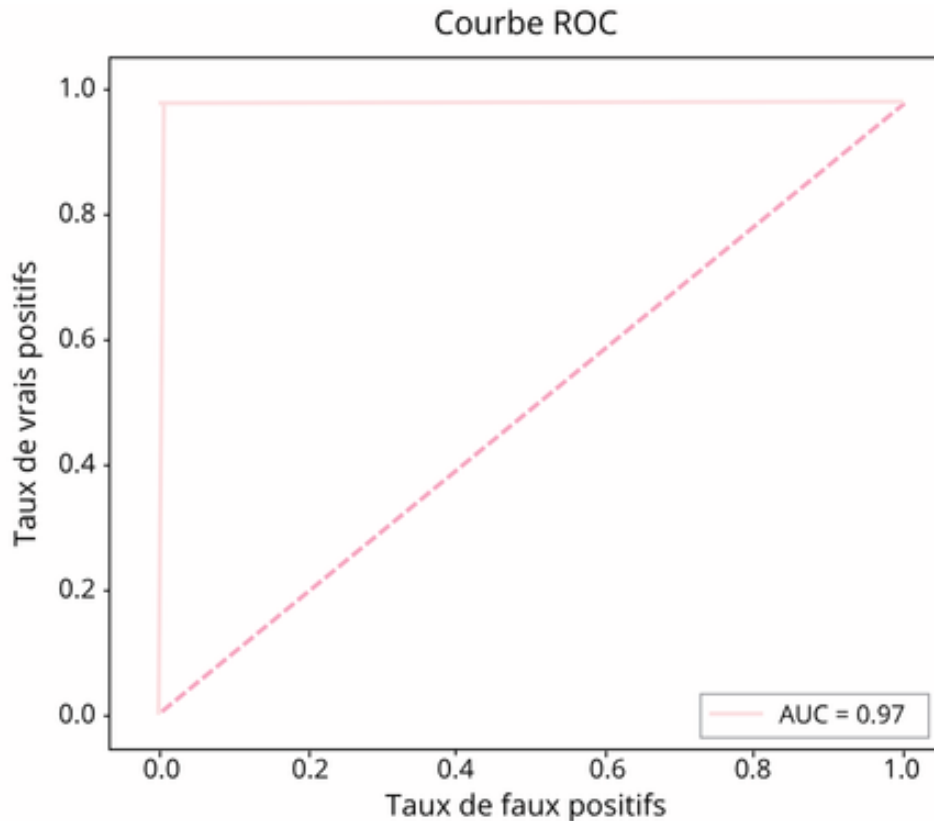


FIGURE 7.11 – Courbe ROC du modèle Random Forest (AUC = 0.97)

Conclusion Le modèle Random Forest présente des performances très satisfaisantes sur l'ensemble de test. La présence de quelques erreurs de classification (faux positifs et faux négatifs) démontre qu'il **ne souffre pas de surapprentissage**, tout en restant suffisamment robuste pour être intégré dans une solution industrielle de détection d'anomalies dans les fluides de forage.

7.4 Développement de l'application interactive

7.4.1 Introduction

L'objectif de cette section est de détailler le processus de développement de l'application interactive permettant de charger des fichiers de données, de prédire des anomalies dans les fluides de forage et de visualiser les résultats en temps réel. Cette application a été réalisée en utilisant la bibliothèque Python **Streamlit**, qui permet de créer des interfaces web interactives pour les projets de machine learning.

L'application permet aux utilisateurs de télécharger des fichiers de données au format **CSV** ou **Excel**, d'extraire les données pertinentes, puis de prédire et visualiser les anomalies détectées dans ces données. En outre, un système de notifications d'alertes a été intégré pour informer les utilisateurs des anomalies détectées en temps réel. Les résultats peuvent être accompagnés de graphiques interactifs et de rapports générés automatiquement. Cette

section décrira en détail l'architecture de l'application, les principales fonctionnalités mises en place et le processus de traitement des données.

7.4.2 Architecture de l'application

L'architecture de l'application peut être décrite comme suit :

- **Interface Utilisateur (UI)** : L'interface permet à l'utilisateur de télécharger un fichier de données (soit en format CSV, soit en format Excel), d'afficher des visualisations des anomalies détectées et d'interagir avec les résultats du modèle.
- **Modèle Prédictif** : Le modèle Random Forest préalablement entraîné est chargé via la bibliothèque `joblib` pour effectuer des prédictions en temps réel sur les données téléchargées.
- **Visualisations Interactives** : L'application génère des graphiques interactifs (comme des matrices de confusion, des graphiques en barres, et des courbes de distribution) pour aider à l'analyse des résultats.
- **Génération de Rapports** : L'application permet également de générer des rapports détaillés, au format PDF ou Markdown, qui résument les anomalies détectées et les résultats des prédictions.
- **Système de Notifications d'Alertes** : Un système de notifications a été intégré pour avertir l'utilisateur dès qu'une anomalie est détectée. Cela inclut des notifications en temps réel affichées dans l'interface utilisateur et des alertes envoyées par email ou SMS.

Cette architecture permet une grande modularité et facilite la gestion des données ainsi que l'interaction avec le modèle. Chaque fonctionnalité a été pensée pour offrir une interface simple et intuitive à l'utilisateur tout en permettant une analyse approfondie des résultats.

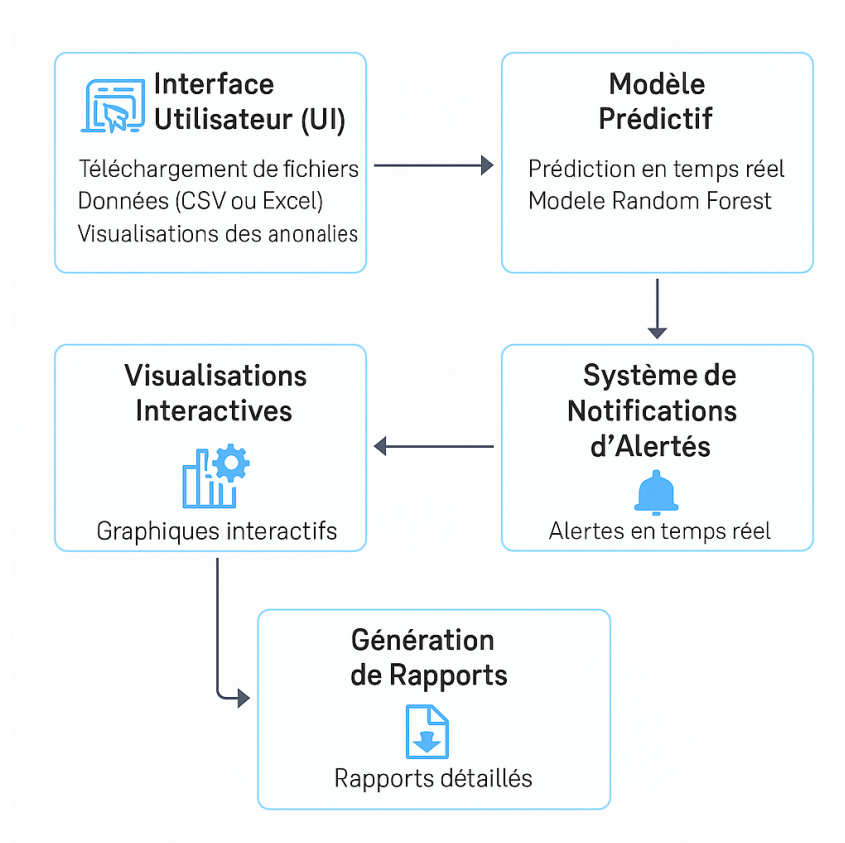


FIGURE 7.12 – Architecture fonctionnelle de l'application de détection d'anomalies.

7.4.3 Page d'accueil de l'application

La page d'accueil de l'application constitue le point de départ pour l'utilisateur. Elle a été conçue pour combiner lisibilité, simplicité d'utilisation et accès rapide aux fonctionnalités essentielles. Dès cette interface, l'utilisateur peut téléverser un fichier de données (au format CSV ou Excel), visualiser les premières analyses statistiques sur la base d'entraînement, et accéder aux outils interactifs de diagnostic comme les histogrammes et boxplots.

La Figure 7.16 présente une capture d'écran de cette interface.



Détection d'Anomalies dans les Fluides de Forage

 Téléverser un fichier CSV ou Excel

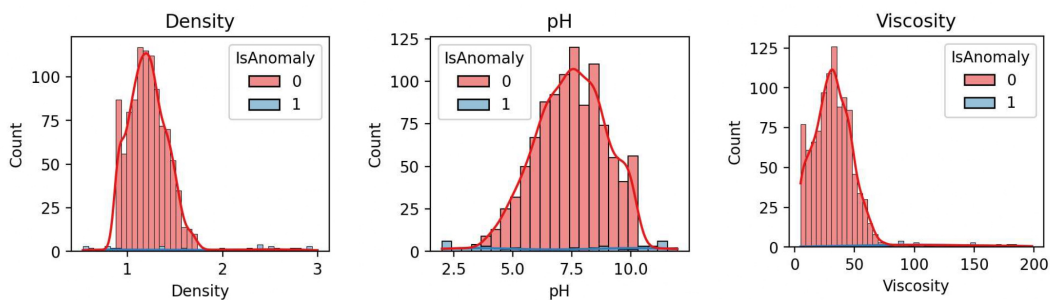


Drag and drop file here

Limit 200MB per file • CSV, XLSX

[Browse files](#)

Visualisations (base d'entraînement)



Boxplots détaillés

Choisissez des colonnes :

[Yield Point](#) × [RPM](#) × [pH](#) ×

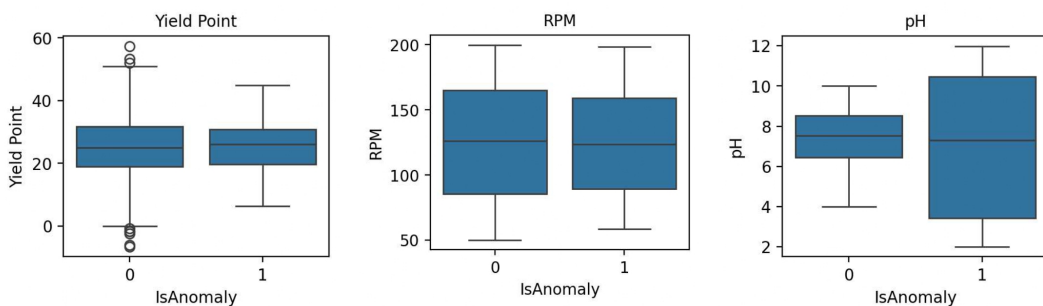


FIGURE 7.13 – Interface de l'application Streamlit pour la détection d'anomalies dans les fluides de forage.

7.4.4 Fonctionnalités principales de l'application

L'application offre plusieurs fonctionnalités interactives permettant à l'utilisateur de charger un fichier de données, d'afficher des visualisations, de prédire les anomalies et de générer des rapports détaillés. Ci-dessous, nous détaillons chaque fonctionnalité.

Chargement de fichiers de données (CSV ou Excel)

L'application permet à l'utilisateur de télécharger un fichier de données au format **CSV** ou **Excel**. Le fichier téléchargé est ensuite automatiquement analysé et les données pertinentes sont extraites pour la prédiction des anomalies.

Une fois le fichier téléchargé, l'application effectue les étapes suivantes :

- Chargement du fichier en fonction du format choisi (CSV ou Excel).
- Extraction des colonnes pertinentes pour l'analyse des anomalies (telles que la densité, la viscosité, le pH, etc.).
- Sauvegarde du fichier extrait dans un nouveau fichier CSV prêt pour l'analyse.

L'interface utilisateur permet à l'opérateur de charger un fichier CSV ou Excel contenant les mesures. Après affichage d'un aperçu, seules les colonnes pertinentes sont extraites et sauvegardées pour une analyse ultérieure.

Algorithm 4 Chargement et extraction des données via l'interface Streamlit

Require: Fichier utilisateur `uploaded_file` (CSV ou Excel)

Ensure: Données extraites sauvegardées dans un fichier CSV

```
1: if un fichier est téléchargé then
2:   if le fichier est au format CSV then
3:     Charger les données avec pd.read_csv
4:   else if le fichier est au format Excel then
5:     Charger les données avec pd.read_excel
6:   end if
7:   Afficher un aperçu des données chargées
8:   Extraire les colonnes pertinentes : Density, pH, Viscosity, Turbidity, Yield Point
9:   Sauvegarder les données extraites dans extracted_data.csv
10:  Afficher un aperçu des données extraites
11: end if
```

Explication :

- La fonction `file_uploader` permet à l'utilisateur de télécharger un fichier de type CSV ou Excel.
- Une fois le fichier téléchargé, il est chargé dans un `DataFrame` avec `pandas`.
- Les colonnes pertinentes sont extraites et un fichier CSV contenant les données extraites est créé pour une utilisation ultérieure.



FIGURE 7.14 – Interface de téléversement de fichiers dans l'application.

Prédiction des anomalies

Après avoir extrait les données pertinentes du fichier téléchargé, le modèle Random Forest est chargé et utilisé pour prédire les anomalies dans les fluides de forage. Le modèle est chargé via la fonction `joblib.load()` et les résultats des prédictions sont affichés dans l'interface de l'application. La prédiction est réalisée en chargeant le modèle Random Forest pré-entraîné, puis en appliquant ce modèle aux données extraites. Les résultats sont affichés en temps réel dans l'interface utilisateur Streamlit, permettant la détection des anomalies.

Algorithm 5 Prédiction des anomalies avec le modèle Random Forest dans Streamlit

Require: Données extraites `df_extracted`, modèle Random Forest pré-entraîné `model`

Ensure: Affichage des résultats de prédiction

- 1: Charger le modèle Random Forest depuis le fichier `rf_model_9374a09d.pkl`
 - 2: Sélectionner les colonnes `Density`, `pH`, `Viscosity`, `Turbidity`, `Yield Point` pour la prédiction
 - 3: Effectuer la prédiction sur les données extraites
 - 4: Afficher les résultats dans l'interface Streamlit
 - 5: Indiquer les étiquettes prédites : 1 pour anomalie, 0 pour normal
-

Explication :

- Le modèle Random Forest est chargé avec `joblib.load()`.
- Les prédictions sont effectuées en utilisant les données extraites (X).
- Les résultats (`y_pred`) indiquent si chaque échantillon est normal ou anomalie, et sont affichés pour l'utilisateur.

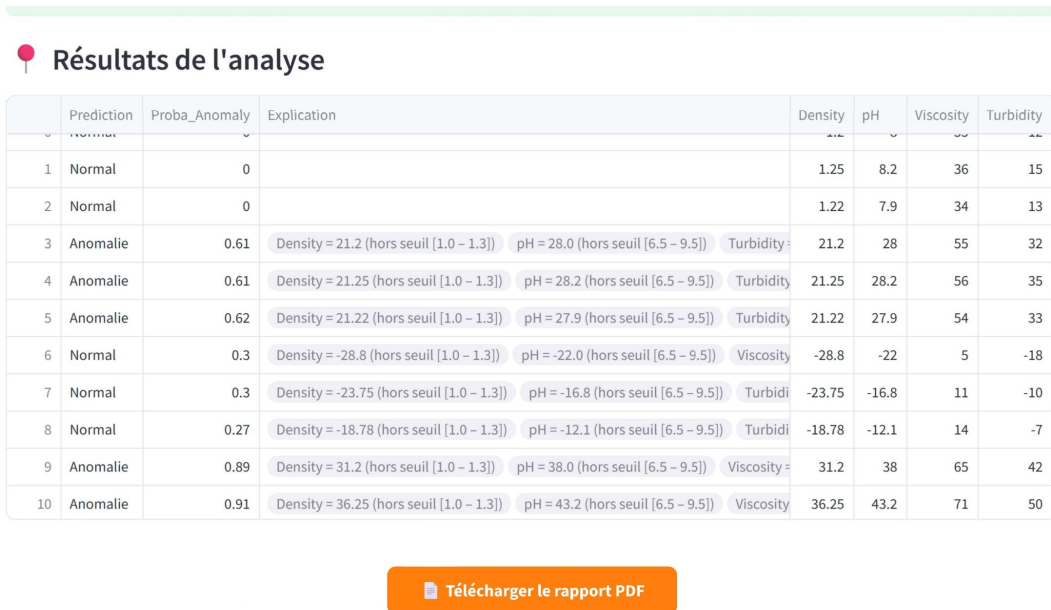


FIGURE 7.15 – Résultats de l'analyse des échantillons avec explication automatique.

Visualisation des résultats de la prédiction

L'application génère plusieurs visualisations pour aider l'utilisateur à mieux comprendre les résultats du modèle de détection des anomalies. Ces visualisations comprennent des matrices de confusion, des graphiques en barres montrant la distribution des anomalies, et des courbes de distribution pour les variables critiques des fluides de forage.

1. Distribution des anomalies par type de fluide : Ce graphique en barres montre la répartition des anomalies détectées pour chaque type de fluide (OBM, WBM, etc.).

Pour mieux comprendre la répartition des anomalies détectées, une visualisation graphique par type de fluide est générée. Seuls les échantillons identifiés comme anormaux sont pris en compte dans ce diagramme.

Algorithm 6 Visualisation des anomalies détectées par type de fluide dans l'interface Streamlit

Require: Données complètes `df` avec colonne `IsAnomaly`

Ensure: Affichage d'un graphique de distribution des anomalies par type de fluide

- 1: Afficher un titre d'introduction dans l'interface Streamlit
 - 2: Filtrer les échantillons où `IsAnomaly = 1`
 - 3: Créer une figure de taille 8x5 pouces
 - 4: Tracer un diagramme en barres (`countplot`) des anomalies par `Fluid System`
 - 5: Ajouter un titre, ainsi que les labels des axes X et Y
 - 6: Intégrer et afficher la figure dans l'interface Streamlit avec `st.pyplot`
-

Visualisations (base d'entraînement)

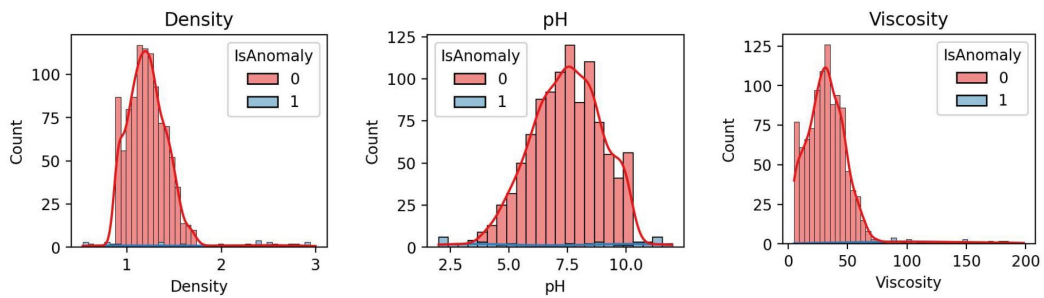


FIGURE 7.16 – Visualisation des distributions des variables Density, pH et Viscosity selon les classes normales (0) et anormales (1) dans la nouvelle base.

2. Graphiques de distribution des variables : Ces graphiques montrent la distribution des anomalies détectées sur différentes variables, comme la densité, la viscosité, et la turbidité, pour mieux comprendre quelles variables sont les plus influentes dans la détection des anomalies.

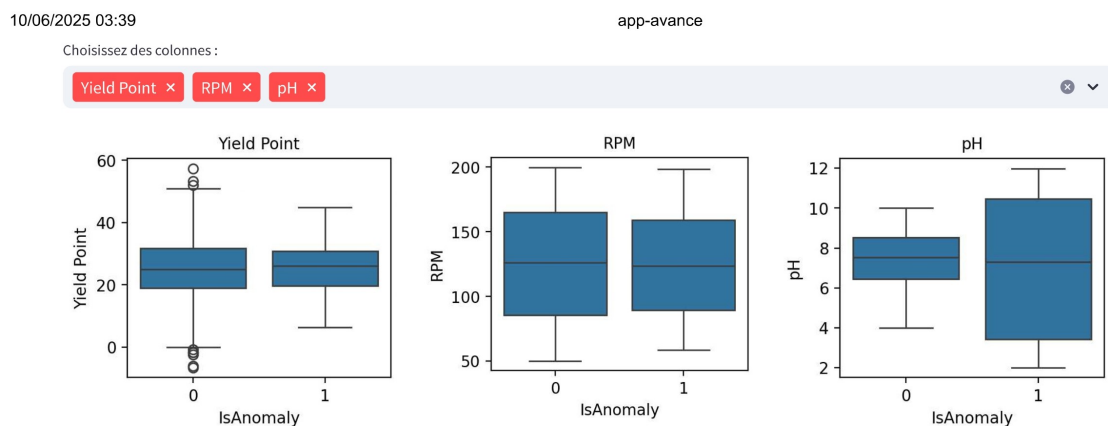


FIGURE 7.17 – Boxplots comparatifs des variables *Yield Point*, *RPM* et *pH* selon la classe IsAnomaly.

Système de Notifications d'Alertes

Le système de notifications d'alertes a été intégré pour fournir des alertes en temps réel lorsque des anomalies critiques sont détectées dans les données des fluides de forage. Lorsqu'une anomalie est détectée, une notification instantanée est générée et peut être affichée dans l'interface de l'application. Les utilisateurs peuvent recevoir ces alertes sous forme de messages dans l'interface ou par e-mail/SMS, en fonction des préférences configurées.

Message d'alerte : “Anomalie détectée! Densité du fluide hors des seuils acceptables (21.2 g/cm^3).”

Visualisation complémentaire : Les résultats de l'analyse sont affichés sous forme de

tableaux et graphiques interactifs, avec une explication détaillée de l'anomalie.

Cette fonctionnalité permet de renforcer la réactivité et d'aider les opérateurs à prendre des décisions rapides pour prévenir les incidents.



FIGURE 7.18 – Exemple de notification d’alerte dans l’application.

Génération automatique de rapports

Une fonctionnalité importante de l’application est la possibilité de générer des rapports détaillés sur les anomalies détectées. Ces rapports peuvent inclure un résumé des anomalies, des visualisations, et des statistiques détaillées sur les résultats des prédictions. L’application permet d’exporter ces rapports sous forme de fichiers **PDF** ou **Markdown**. Afin de documenter les résultats, un rapport PDF est généré automatiquement, synthétisant le bilan des anomalies détectées via le modèle. Ce rapport peut être téléchargé et partagé facilement.

Algorithm 7 Génération automatique d’un rapport PDF des anomalies détectées

Require: Données réelles avec étiquettes `df["IsAnomaly"]`, prédictions `y_pred`

Ensure: Fichier PDF résumé des résultats de classification

- 1: Initialiser un document PDF avec `FPDF`
 - 2: Ajouter une nouvelle page
 - 3: Définir la police à Arial, taille 12
 - 4: Ajouter un titre de rapport
 - 5: Ajouter un texte d’introduction pour le résumé des anomalies détectées
 - 6: Calculer le rapport de classification (`classification_report`)
 - 7: Insérer le rapport dans le PDF en texte multi-lignes
 - 8: Sauvegarder le fichier PDF sous `rapport_anomalies.pdf`
 - 9: Afficher un message de confirmation dans l’interface Streamlit
-

Explication :

- Le module **FPDF** est utilisé pour générer un rapport PDF contenant un résumé des anomalies détectées et un rapport détaillé sur les résultats du modèle.
- Le rapport inclut des informations sur les performances du modèle (précision, rappel, etc.) et peut être téléchargé par l'utilisateur.



FIGURE 7.19 – Bouton de téléchargement du rapport PDF généré automatiquement par l'application après l'analyse.



Rapport d'Analyse des Fluides de Forage



Date : 26/05/2025
Nombre d'échantillons analysés : 5

■ Échantillon 1

➤ Probabilité : **0.00%**

■ Échantillon 2

➤ Probabilité : **0.00%**

■ pH = 5.2 (hors seuil [6.5 – 9.5])

■ Échantillon 3

➤ Probabilité : **71.00%**

■ Density = 0.96 (hors seuil [1.0 – 1.3])

■ Viscosity = 120 (hors seuil [10.0 – 60.0])

■ Turbidity = 45 (hors seuil [5.0 – 20.0])

■ Yield Point = 8 (hors seuil [10.0 – 35.0])

■ Solid Content (%) = 4 (hors seuil [5.0 – 20.0])

■ Filtration Loss = 22 (hors seuil [0.0 – 15.0])

■ Oil Water Ratio (%) = 0 (hors seuil [60.0 – 85.0])

■ Échantillon 4

➤ Probabilité : **0.00%**

■ Oil Water Ratio (%) = 0 (hors seuil [60.0 – 85.0])

■ Échantillon 5

➤ Probabilité : **1.00%**

■ Depth (m) = 950 (hors seuil [1000 – 4000])

■ Oil Water Ratio (%) = 90 (hors seuil [60.0 – 85.0])

FIGURE 7.20 – Extrait d'un rapport d'analyse PDF généré par l'application. Chaque échantillon est accompagné d'une probabilité d'anomalie et d'une explication automatique basée sur les seuils métiers.

7.4.5 Conclusion de l'application interactive

L'application développée avec **Streamlit** permet aux utilisateurs de charger des fichiers de données, de visualiser les anomalies détectées, et de générer des rapports détaillés sur les fluides de forage. Cette interface interactive offre une manière simple et intuitive d'exploiter un modèle de machine learning préalablement entraîné pour la détection d'anomalies en temps réel.

Un système de notifications d'alerte a été intégré pour informer les utilisateurs dès qu'une anomalie critique est identifiée, renforçant ainsi la réactivité du processus de gestion. En complément, chaque prédiction est accompagnée d'une **explication automatique** mettant en évidence les variables à l'origine de l'anomalie (par exemple : pH en dehors des seuils, turbidité excessive).

Ces résultats sont **directement exploitables par les ingénieurs de terrain** pour déterminer les actions correctives à mettre en œuvre, telles que l'ajustement de la formulation (ajout d'additifs, dilution), la validation par des tests complémentaires, ou la remise en observation du puits concerné. L'application s'inscrit ainsi dans une logique d'**aide à la décision opérationnelle**, au service du contrôle qualité en temps réel.

Grâce à **Streamlit**, la solution est facilement déployable, accessible à différents profils utilisateurs, et constitue une avancée concrète vers la digitalisation du suivi des fluides de forage.

7.5 Conclusion

Ce chapitre a permis de valider la faisabilité de la solution proposée en intégrant les différentes composantes du projet : traitement des données, construction du modèle prédictif, et développement d'une application interactive. L'algorithme Random Forest, associé à une démarche rigoureuse de validation croisée et d'optimisation, a montré des performances très satisfaisantes sur le jeu de données étudié.

L'application Streamlit développée permet une exploitation intuitive des résultats, avec des fonctionnalités de visualisation, de rapport automatique, et de notification d'alerte en cas d'anomalie. Cette solution constitue un outil concret d'aide à la décision pour les ingénieurs terrain, renforçant ainsi la capacité de détection et de prévention des anomalies dans les fluides de forage.

Conclusion générale

Dans ce mémoire, nous avons exploré les possibilités offertes par l'intelligence artificielle appliquée au contrôle qualité des fluides de forage, en particulier dans un environnement industriel exigeant comme celui d'ADFC, filiale du groupe Sonatrach. En combinant les capacités de classification des modèles d'apprentissage automatique avec une approche métier orientée terrain, nous avons cherché à apporter une réponse innovante à une problématique réelle : la détection automatique d'anomalies dans des échantillons critiques.

Nous avons commencé par dresser un état des lieux détaillé de l'organisme d'accueil et des méthodes de contrôle existantes, mettant en évidence les limites des approches traditionnelles. Celles-ci, souvent manuelles ou basées sur des seuils fixes, peinent à s'adapter à la complexité et à la variabilité des données réelles collectées sur le terrain.

Dans un second temps, nous avons étudié les fondements théoriques et les outils disponibles en matière de détection d'anomalies, en passant en revue aussi bien les méthodes conventionnelles que les approches modernes basées sur l'IA. Le modèle Random Forest, enrichi par la technique SMOTE pour pallier le déséquilibre des classes, s'est avéré particulièrement performant, obtenant des résultats remarquables sur les jeux de données traités.

La solution que nous avons développée ne s'arrête pas à un modèle théorique. Elle se matérialise sous la forme d'une application interactive développée avec Streamlit, permettant à l'utilisateur de charger des fichiers, d'obtenir une prédiction en temps réel, de visualiser les anomalies détectées et d'éditer automatiquement un rapport explicatif. Cette interface rend l'IA accessible, compréhensible et exploitable dans un environnement de production.

Ce travail marque ainsi une première étape concrète vers l'industrialisation intelligente du contrôle qualité dans le domaine des fluides de forage. Les résultats obtenus montrent que l'IA peut non seulement améliorer la fiabilité et la rapidité de détection, mais aussi renforcer la prise de décision grâce à une meilleure compréhension des causes d'anomalie.

Il reste néanmoins plusieurs perspectives d'amélioration : enrichissement du jeu de données, extension à des données en temps réel, intégration de recommandations automatiques, ou encore utilisation de modèles plus avancés comme les réseaux neuronaux ou les approches hybrides. Ces pistes ouvrent la voie à une optimisation continue du système, au service de la performance industrielle.

Ce mémoire illustre, à son échelle, comment les technologies de l'IA peuvent répondre à des besoins métiers concrets, tout en s'inscrivant dans une démarche de digitalisation plus large des processus industriels.

Perspectives

Malgré les résultats très encourageants obtenus, plusieurs pistes d'amélioration restent ouvertes pour renforcer la portée et la robustesse du système développé.

- **Extension de la base de données** : Il serait bénéfique d'intégrer un volume plus large d'échantillons, provenant de différents sites, profondeurs et types de fluides, pour améliorer la généralisation du modèle.
- **Traitement en temps réel** : L'une des évolutions majeures consisterait à connecter l'application directement aux capteurs de terrain pour permettre une détection continue et immédiate des anomalies.
- **Système de recommandations automatiques** : À partir de l'anomalie détectée, le système pourrait suggérer automatiquement des actions correctives adaptées, en s'appuyant sur des règles métier ou des historiques.
- **Évolution vers des modèles plus avancés** : Le recours à des architectures neuronales plus complexes (Autoencoders, LSTM, etc.) permettrait de détecter des schémas d'anomalie encore plus subtils.
- **Intégration complète dans l'environnement métier** : Le système pourrait être intégré dans une infrastructure cloud sécurisée, avec authentification multi-utilisateurs, journalisation des anomalies, et compatibilité avec les outils internes d'ADFC.
- **Amélioration de l'explicabilité** : L'ajout de modules d'analyse causale ou de visualisation des poids du modèle renforcerait la confiance des utilisateurs dans les prédictions générées.

Ces perspectives ouvrent la voie à un système intelligent plus complet, réactif et fiable, capable de s'adapter à l'évolution des besoins industriels et aux enjeux de digitalisation des opérations de forage.

Bibliographie

Bibliographie

- [1] R. G. Schroeder, S. M. Goldstein, and M. J. Rungtusanatham, *Operations Management : Contemporary Concepts and Cases*. McGraw-Hill, 2010.
- [2] Accenture, “Digital Oilfield Outlook Report,” 2019. [Online]. Available : https://www.accenture.com/_acnmedia/PDF-92/Accenture-Digital-Oilfield-Outlook-Report.pdf
- [3] Y. Zhou, X. Li, and S. Wang, “A review of mud logging techniques in oil and gas drilling,” *Journal of Petroleum Science and Engineering*, vol. 159, pp. 696–709, 2017.
- [4] A. Khosravi, M. Jabbari, and M. Amiri, “A comprehensive review on anomaly detection for real-time oilfield monitoring,” *Petroleum*, vol. 7, no. 1, pp. 78–89, 2021.
- [5] S. Kaplan, “Maintenance, Reliability and Risk,” Elsevier, 2011.
- [6] J. Cheng, Y. Wang, H. Liu, and G. Zhang, “Machine learning techniques for drilling operations : A review,” *Journal of Natural Gas Science and Engineering*, vol. 65, pp. 289–305, 2019.
- [7] S. D. Mohaghegh, “Data-driven analytics machine learning in oil gas,” *Journal of Petroleum Technology*, vol. 70, no. 12, pp. 46–52, 2018.
- [8] API Standard 13B-1, “Recommended Practice for Field Testing Water-Based Drilling Fluids,” American Petroleum Institute, 2009.
- [9] Schlumberger, “Drilling Fluids System Overview,” Oilfield Glossary, [Online]. Available : https://www.glossary.oilfield.slb.com/Terms/d/drilling_fluid.aspx
- [10] Y. Zhou, X. Li, and S. Wang, “A review of mud logging techniques in oil and gas drilling,” *J. Pet. Sci. Eng.*, vol. 159, pp. 696–709, 2017.
- [11] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection : A survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.
- [12] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation Forest,” in *IEEE ICDM*, 2008.
- [13] P. Malhotra et al., “Long short term memory networks for anomaly detection in time series,” *ESANN*, 2015.
- [14] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [15] M. Ahmad et al., “Unsupervised anomaly detection techniques : A review,” *Artificial Intelligence Review*, vol. 52, no. 2, pp. 219–243, 2019.
- [16] R. Caenn, H. C. H. Darley, and G. R. Gray, *Composition and Properties of Drilling and Completion Fluids*, 6th ed., Gulf Professional Publishing, 2011.

Annexes

Annexe A : Bibliothèques Python utilisées

Python est un langage de programmation à haut niveau, largement utilisé pour l'analyse de données, l'intelligence artificielle, et le développement d'applications. Dans le cadre de ce projet, plusieurs bibliothèques ont été utilisées à des étapes clés, allant du prétraitement des données à la visualisation des résultats.

A.1 NumPy

NumPy est une bibliothèque fondamentale pour le calcul scientifique avec Python. Elle permet la création de tableaux multidimensionnels (ndarray), fournit des fonctions mathématiques performantes, et est optimisée pour les opérations vectorielles rapides.

A.2 Pandas

Pandas permet la manipulation efficace de structures de données tabulaires (DataFrame). Elle a été utilisée pour charger les données, effectuer des opérations de nettoyage, transformer les colonnes et analyser statistiquement les paramètres physiques des fluides.

A.3 Scikit-learn

Cette bibliothèque fournit des algorithmes d'apprentissage automatique, des outils de sélection de modèle, de validation croisée, et d'évaluation. Dans notre projet, elle a permis d'entraîner le modèle Random Forest, de normaliser les données, et de calculer des métriques.

A.4 Matplotlib et Seaborn

Ces bibliothèques de visualisation ont permis de générer des graphiques (courbes d'apprentissage, heatmaps de corrélation, boxplots...) facilitant l'interprétation des résultats. Seaborn repose sur Matplotlib mais offre une syntaxe plus intuitive et des styles préconfigurés.

A.5 Imbalanced-learn

Utilisée pour corriger le déséquilibre du jeu de données via SMOTE (Synthetic Minority Oversampling Technique), cette bibliothèque permet de créer artificiellement des exemples d'anomalies, améliorant ainsi l'équilibre du modèle.

A.6 Streamlit

Streamlit a été utilisé pour développer l'interface interactive. Elle permet de transformer un script Python en une application web conviviale, utilisée ici pour charger les fichiers, effectuer les prédictions et générer des rapports.

Annexe B : Outils et environnements techniques

B.1 Jupyter Notebook

Environnement interactif dans lequel les cellules de code, texte et visualisations peuvent être combinées. Utilisé pour expérimenter, visualiser les courbes et documenter le projet de bout en bout.

B.2 Google Sheets et Excel

Sources principales de données. Les fichiers Excel fournis par ADFC contiennent les propriétés physico-chimiques mesurées. Google Sheets a parfois été utilisé pour vérifier manuellement les extraits ou corriger des incohérences.

B.3 Google Drive

Utilisé comme espace de stockage et de synchronisation pour partager les données entre les membres du binôme et l'encadrant.

B.4 VSCode

Éditeur de code principal du projet. Il a permis l'intégration de tous les modules Python, la gestion des environnements virtuels, et le lancement de l'application Streamlit.

B.5 Dropbox

Service cloud secondaire utilisé pour les sauvegardes, les transferts de fichiers d'expérimentation et les documents PDF générés.

Annexe C : Notions et concepts avancés

C.1 La méthode CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) est une méthodologie en six étapes :

- **Compréhension métier** : définir les objectifs de détection d'anomalies dans les fluides de forage.
- **Compréhension des données** : identification des paramètres disponibles et de leurs distributions.
- **Préparation des données** : nettoyage, transformation, normalisation et équilibrage via SMOTE.
- **Modélisation** : application de l'algorithme Random Forest et ajustement des hyperparamètres.
- **Évaluation** : interprétation des performances via les métriques standard.
- **Déploiement** : intégration du modèle dans une application Streamlit.

C.2 L'Overfitting

Le surapprentissage se produit lorsqu'un modèle est trop ajusté aux données d'entraînement. Il perd alors en capacité de généralisation. Pour y remédier :

- Validation croisée à 5 plis.
- Limitation de la profondeur des arbres.
- Réduction du nombre d'estimateurs.

C.3 SMOTE

Technique de rééchantillonnage utilisée pour générer des exemples synthétiques dans la classe minoritaire. Elle aide à prévenir le biais du modèle vers la classe dominante ("normal").

C.4 Métriques d'évaluation

- **Accuracy** : taux global de bonnes prédictions.
- **Précision** : proportion d'anomalies prédites qui sont correctes.
- **Rappel** : proportion d'anomalies réelles détectées.
- **F1-score** : moyenne harmonique entre précision et rappel.

C.5 Analyse exploratoire

L'EDA a permis d'identifier les paramètres les plus liés aux anomalies : Turbidity, Oil Water Ratio, Viscosity. La heatmap de corrélation a confirmé plusieurs dépendances entre variables.

Annexe D : Résultats comparatifs des algorithmes testés

Dans le cadre de ce projet, plusieurs algorithmes de détection d'anomalies ont été testés avant de sélectionner le modèle final. Chaque approche a été évaluée en termes de précision, rappel, f1-score, et robustesse sur un jeu de données déséquilibré. Cette section présente une synthèse des performances observées et des justifications du choix retenu.

D.1 Random Forest (avec SMOTE)

Ce modèle a montré des performances solides et équilibrées sur l'ensemble des données. Contrairement à d'autres modèles, il a su détecter les anomalies avec un excellent compromis entre précision et rappel, tout en conservant un très faible taux de faux positifs.

- **Accuracy** : 0.992 (sur 384 échantillons)
- **Rappel (anomalie)** : 0.91
- **F1-score (anomalie)** : 0.87

Il constitue un choix pertinent pour une mise en production industrielle. Sa capacité à gérer les jeux de données déséquilibrés grâce à SMOTE et à fournir des résultats interprétables le rend supérieur aux autres méthodes testées.

D.2 K-Nearest Neighbors (KNN)

- Accuracy : 0.98
- Rappel (anomalie) : 1.00
- F1-score (anomalie) : 0.71

Performant mais sensible à la structure des données. Produit parfois des faux positifs.

D.3 Isolation Forest (de base)

- Accuracy : 0.89

- Rappel (anomalie) : 1.00
- F1-score (anomalie) : 0.33

Détecte les anomalies mais avec beaucoup de fausses alertes (précision très faible).

D.4 Isolation Forest (tuning + validation croisée)

- Accuracy : 0.71
- F1-score : 0.06

Très mauvaise précision malgré un rappel élevé. Trop de faux positifs.

D.5 Autoencoder

- Accuracy : 0.96
- F1-score : 0.69

Performant et prometteur, mais nécessite un ajustement précis du seuil.

D.6 Autoencoder (seuil ajusté)

- Accuracy : 0.71
- F1-score : 0.08

Faible précision, trop d'échantillons normaux sont considérés à tort comme anomalies.

D.7 One-Class SVM

- Accuracy : 0.92
- F1-score : 0.42

Bonne base mais difficile à calibrer. Sensible au déséquilibre des classes.

D.8 One-Class SVM + SMOTE

- Accuracy : 0.90
- F1-score : 0.35

Légère amélioration, mais pas suffisant pour des performances industrielles fiables.

Modèle	Accuracy	Recall	F1-score	Commentaires
Random Forest + SMOTE	0.992	0.91	0.87	Meilleur compromis global. Fiable, robuste, précis.
KNN	0.98	1.00	0.71	Bonne performance, mais sensible aux données bruitées.
Isolation Forest (base)	0.89	1.00	0.33	Trop de faux positifs malgré un bon rappel.
Isolation Forest (tuning)	0.71	1.00	0.06	Mauvaise précision, non utilisable en l'état.
Autoencoder	0.96	1.00	0.69	Prometteur, mais dépend du calibrage du seuil.
Autoencoder (seuil ajusté)	0.71	1.00	0.08	Surdétection excessive.
One-Class SVM	0.92	1.00	0.42	Résultats moyens, sensible au déséquilibre.
One-Class SVM + SMOTE	0.90	1.00	0.35	Amélioration légère, mais toujours trop de fausses alertes.

TABLE 7.3 – Résumé comparatif des performances des algorithmes testés

Conclusion : Au vu des résultats présentés, le modèle **Random Forest avec SMOTE** est celui qui offre la meilleure balance entre performance, précision, et robustesse. Il a donc été choisi comme solution finale pour l'intégration dans l'application.

Annexe E : Codes sources et scripts d'interface Streamlit

E.1 – Préparation et nettoyage des données

```
# 1. Charger les données
df = pd.read_csv("finalALL_PIVOTEE.csv")

# 2. Supprimer la colonne redondante
if "Formulation Label" in df.columns:
    df.drop(columns=["Formulation Label"], inplace=True)

# 3. Extraire les Tags depuis la colonne "Description"
def extract_tags(text):
    if pd.isna(text):
        return "None"
    tags = []
    text = text.lower()
    if "obm" in text: tags.append("OBM")
    if "sbm" in text: tags.append("SBM")
    if "wbm" in text: tags.append("WBM")
    if "kcl" in text: tags.append("KCl")
    if "nacl" in text: tags.append("NaCl")
    if "polymer" in text: tags.append("Polymer")
    return ", ".join(tags) if tags else "Other"

df["Tags"] = df["Description"].apply(extract_tags)

# 4. Nettoyer "Job Ref" en extrayant le format "MI-XX-XX"
def clean_job_ref(text):
    if isinstance(text, str):
        match = re.search(r"(MI-\d{2}-\d{2})", text)
        return match.group(1) if match else text.strip()
    return "None"

df["Job Ref"] = df["Job Ref"].apply(clean_job_ref)

# 5. Extraire l'année depuis la colonne "Date"
def extract_year(date):
    try:
        return pd.to_datetime(date).year
    except:
        return None

df["Année"] = df["Date"].apply(extract_year)

# 6. Remplir les valeurs manquantes
# a) Colonnes numériques : moyenne par "Fluid System"
num_cols = df.select_dtypes(include=[np.number]).columns
for col in num_cols:
    df[col] = df.groupby("Fluid System")[col].transform(lambda x: x.fillna(x.mean()))
    df[col].fillna(0, inplace=True)

# b) Colonnes textuelles : "None"
obj_cols = df.select_dtypes(include="object").columns
for col in obj_cols:
    df[col].fillna("None", inplace=True)

# 7. Exporter les données nettoyées
df.to_csv("finalALL_PIVOTEE_cleaned.csv", index=False, encoding="utf-8-sig")
print("✅ Fichier nettoyé enregistré : finalALL_PIVOTEE_cleaned.csv")
```

Que puis-je vous aider à créer ?

Variables Terminal

E.2 – Centralisation et création du tableau pivoté

```
# 1. Charger les données nettoyées
df = pd.read_csv("finalALL_PIVOTEE_cleaned.csv")

# 2. Créer un identifiant unique pour chaque échantillon
df["Sample ID"] = df["Formulation Label"] + "_" + df["Sample Type"]

# 3. Créer un tableau pivoté : une ligne = un échantillon, colonnes = types de tests
pivot_df = df.pivot_table(
    index=["Sample ID", "Date", "Fluid System"],
    columns="Test",
    values="Value",
    aggfunc="first"
).reset_index()

# 4. Exporter le tableau centralisé dans un fichier CSV
pivot_df.to_csv("synthetic_drilling_fluid_data_enriched.csv", index=False, encoding="utf-8-sig")
print("✅ Données centralisées enregistrées : synthetic_drilling_fluid_data_enriched.csv")
```

E.3 – Entraînement du modèle Random Forest avec normalisation et SMOTE

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
import joblib

# Exemple de données fictives pour test rapide (à remplacer par les vraies)
from sklearn.datasets import make_classification
X_train, y_train = make_classification(n_samples=100, n_features=10,
                                     n_informative=5, n_redundant=2,
                                     weights=[0.9], random_state=42)

# Prétraitement : normalisation
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)

# Rééquilibrage des classes
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_train_scaled, y_train)

# Initialisation et entraînement du modèle
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_resampled, y_resampled)

# Sauvegarde du modèle entraîné
joblib.dump(model, 'rf_model_final.pkl')

print("✅ Modèle entraîné et sauvegardé avec succès.")
```

E.4 – Interface Streamlit : Chargement et extraction des données utilisateur

```
import streamlit as st
import pandas as pd

# Interface utilisateur : Téléchargement du fichier CSV ou Excel
uploaded_file = st.file_uploader("Choisir un fichier CSV ou Excel", type=["csv", "xlsx"])

if uploaded_file is not None:
    # Charger les données selon le type de fichier
    if uploaded_file.name.endswith(".csv"):
        df = pd.read_csv(uploaded_file)
    elif uploaded_file.name.endswith(".xlsx"):
        df = pd.read_excel(uploaded_file)

    st.write("Aperçu des données :")
    st.write(df.head())

    # Extraire les colonnes pertinentes pour l'analyse
    features = ['Density', 'pH', 'Viscosity', 'Turbidity', 'Yield Point']
    df_extracted = df[features]

    # Sauvegarder les données extraites
    df_extracted.to_csv('extracted_data.csv', index=False)
    st.write("Fichier extrait enregistré : extracted_data.csv")
    st.write(df_extracted.head())
```

E.5 – Interface Streamlit : Prédiction en temps réel avec le modèle entraîné

```
import joblib

# Charger le modèle Random Forest pré-entraîné
model = joblib.load("rf_model_9374a09d.pkl")

# Sélectionner les colonnes pertinentes pour la prédiction
X = df_extracted[['Density', 'pH', 'Viscosity', 'Turbidity', 'Yield Point']]

# Effectuer la prédiction
y_pred = model.predict(X)

# Afficher les résultats dans l'interface Streamlit
st.write("Résultats de la prédiction :")
st.write("Anomalies détectées (1 = Anomalie, 0 = Normal) :")
st.write(y_pred)
```

E.6 – Interface Streamlit : Visualisation des anomalies par type de fluide

```
import seaborn as sns
import matplotlib.pyplot as plt

# Affichage dans l'interface Streamlit
st.write("Distribution des anomalies par type de fluide :")

# Filtrer uniquement les échantillons avec anomalies
anomalies_df = df[df["IsAnomaly"] == 1]

# Créer la figure
plt.figure(figsize=(8, 5))
sns.countplot(data=anomalies_df, x="Fluid System", order=df["Fluid System"].value_counts().index)
plt.title("Nombre d'anomalies détectées par type de fluide")
plt.xlabel("Type de fluide")
plt.ylabel("Nombre d'anomalies")

# Intégrer le graphique dans Streamlit
st.pyplot(plt)
```

E.7 – Interface Streamlit : Génération automatique d'un rapport PDF

```
import seaborn as sns
import matplotlib.pyplot as plt

# Affichage dans l'interface Streamlit
st.write("Distribution des anomalies par type de fluide :")

# Filtrer uniquement les échantillons avec anomalies
anomalies_df = df[df["IsAnomaly"] == 1]

# Créer la figure
plt.figure(figsize=(8, 5))
sns.countplot(data=anomalies_df, x="Fluid System", order=df["Fluid System"].value_counts().index)
plt.title("Nombre d'anomalies détectées par type de fluide")
plt.xlabel("Type de fluide")
plt.ylabel("Nombre d'anomalies")

# Intégrer le graphique dans Streamlit
st.pyplot(plt)
```