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

**ÉCOLE NATIONALE POLYTECHNIQUE**

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

سوناطراك
sonatrach

**Département Génie Minier**

# Mémoire de projet de fin d'études

**Pour l'obtention du diplôme d'ingénieur d'état en Génie Minier**

---

# Système Automatisé de Classification Lithologique utilisant l'Apprentissage Automatique

---

**BENGUETTAF Bilal Gholameddine**

Sous la direction de **Pr. AKKAL Rezki** ENP

Présenté et soutenu publiquement le (02/11/2025)

**Composition du jury :**

| | | |
|---|---|---|
| Président : | Pr. YAHYAOUI Sami | ENP |
| Promoteur : | Pr. AKKAL Rezki | ENP |
| Examinateur : | Mr. CHANANE Larouci | ENP |
| Représentant de l'incubateur : | Mr. BOUSBAI M'hamed | ENP |

ENP 2025

---

Génie Minier / ENP, 10 Rue des Frères OUDEK, El Harrach 16200 Alger Algérie.

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

**ÉCOLE NATIONALE POLYTECHNIQUE**

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

سوناطراك
sonatrach

**Département Génie Minier**

# Mémoire de projet de fin d'études

**Pour l'obtention du diplôme d'ingénieur d'état en Génie Minier**

---

# Système Automatisé de Classification Lithologique utilisant l'Apprentissage Automatique

---

**BENGUETTAF Bilal Gholameddine**

Sous la direction de **Pr. AKKAL Rezki** ENP

Présenté et soutenu publiquement le (02/11/2025)

**Composition du jury :**

| | | |
|---|---|---|
| Président : | Pr. YAHYAOUI Sami | ENP |
| Promoteur : | Pr. AKKAL Rezki | ENP |
| Examinateur : | Mr. CHANANE Larouci | ENP |
| Représentant de l'incubateur : | Mr. BOUSBAI M'hamed | ENP |

ENP 2025

---

Génie Minier / ENP, 10 Rue des Frères OUDEK, El Harrach 16200 Alger Algérie.

## ملخص :

يقدّم هذا العمل مقاربة مبتكرة تعتمد على الذكاء الاصطناعي وتقنيات التعلّم الآلي بهدف تفسير سجلات الآبار، وذلك من أجل تجاوز القيود المرتبطة بالأساليب التقليدية. من خلال تطبيق خوارزميا Gradient Boostingمثل XGBoost و CatBoost و LightGBM على تسعة آبار تعود إلى المكامن الأوردوفيسية والترياسية في الأطراف الشمالية-الشرقية لحقل حاسي مسعود بمنطقة تقرت، تمكّن النظام المطوَّر من تحقيق أداء متميز في التنبؤ بالخصائص البتروفيسيائية والليثولوجية.

بالنسبة للمتغيرات الشائعة، حقق النموذج عبر الانحدار المباشر معاملات تفسير ($R^2$) بلغت 0.9828 لحجم الطين (VCL) ، و0.9055 للكوارتز، و0.8564 للمسامية الفعّالة (PIGE). أما بالنسبة لليثولوجيات النادرة، فقد أظهرت عملية كشف الصخور النارية والمعادن غير الشائعة دقة تراوحت بين 96% و97% مع قيم F1 بلغت 0.95، في حين سمحت عملية تقدير نسبها بتحقيق معاملات ($R^2$) بلغت 0.9391 و0.8140 على التوالي. كما وصلت عملية التمييز بين الليثولوجيات قليلة التكرار إلى دقة 97% مع قيم F1 تراوحت بين 0.91 و0.98 لالكالسيت، الدولوميت، الأنهيدريت، والهاليت.

وقد أكدت عملية التحقق الخارجي على بئر مستقل متانة النظام، حيث حقق قدرة تمييز بلغت 100% للمكامن المنتجة و80% من التوصيف الدقيق. ويساهم هذا النظام في تقليص زمن المعالجة بشكل كبير من عدة ساعات إلى بضع دقائق، مع توفير استقلالية تامة عن البرمجيات ذات التراخيص المملوكة مثل Techlog وPetrel، مما يمنح سوناطراك ميزة استراتيجية مهمة.



**الكلمات المفتاحية** : الذكاء الاصطناعي، التعلم الآلي، السجلات البترولية، تعزيز التدرج، التفسير البتروفيزيائي، مكامن الأورد و فيشي والترياسي

# Abstract

This work presents an innovative approach based on artificial intelligence and machine learning for well log interpretation, aiming to overcome the limitations of traditional methods. By applying gradient boosting algorithms (XGBoost, CatBoost, LightGBM) to nine wells from Ordovician and Triassic reservoirs located on the periphery of the Hassi Messaoud field in the Touggourt region, the developed system achieved remarkable performance in predicting petrophysical and lithological properties.

For abundant targets, the model obtained through direct regression an $R^2$ of 0.9828 for clay volume (VCL), 0.9055 for Quartz, and 0.8564 for effective porosity (PIGE). For sporadic lithologies, the detection of igneous rocks and other rare minerals showed an accuracy of 96-97% with F1-scores of 0.95, while their quantification achieved $R^2$ values of 0.9391 and 0.8140 respectively. Finally, the discrimination of rarely present lithologies reached an accuracy of 97% with F1-scores between 0.91 and 0.98 for four facies (calcite, dolomite, anhydrite, halite).

External validation on an independent well confirmed the system's robustness with 100% identification of productive reservoirs and 80% perfect characterization. This system significantly reduces processing time from several hours to a few minutes, while providing independence from proprietary licenses (Techlog, Petrel), conferring valuable strategic autonomy to Sonatrach.

**Keywords :** Artificial intelligence, Machine learning, Well logs, Gradient boosting, Petrophysical interpretation, Ordovician and Triassic reservoirs.

# Résumé

Ce travail présente une approche innovante basée sur l'intelligence artificielle et l'apprentissage automatique pour l'interprétation des logs de puits, visant à surmonter les limitations des méthodes traditionnelles. En appliquant des algorithmes de gradient boosting (XGBoost, CatBoost, LightGBM) à neuf puits des réservoirs ordovicien et triasique situés en périphérie du champ de Hassi Messaoud dans la région de Touggourt, le système développé a permis d'atteindre des performances remarquables dans la prédiction des propriétés pétrophysiques et lithologiques.

Pour les cibles abondantes, le modèle a obtenu par régression directe un $R^2$ de 0,9828 pour le volume d'argile (VCL), 0,9055 pour le Quartz et 0,8564 pour la porosité effective (PIGE). Pour les lithologies sporadiques, la détection des roches ignées et autres minéraux rares a montré une précision de 96-97% avec des F1-scores de 0,95, tandis que leur quantification a permis d'atteindre des $R^2$ de 0,9391 et 0,8140 respectivement. Enfin, la discrimination des lithologies rarement présentes a atteint une précision de 97% avec des F1-scores entre 0,91 et 0,98 pour quatre faciès (calcite, dolomite, anhydrite, halite).

La validation externe sur un puits indépendant a confirmé la robustesse du système avec une identification à 100% des réservoirs productifs et 80% de caractérisation parfaite. Ce système réduit significativement le temps de traitement de plusieurs heures à quelques minutes, tout en offrant une indépendance vis-à-vis des licences propriétaires (Techlog, Petrel), conférant à Sonatrach une autonomie stratégique précieuse.

**Mots-clés :** Intelligence artificielle, Apprentissage automatique, Logs de puits, Gradient boosting, Interprétation pétrophysique, Réservoirs ordovicien et triasique.

# Remerciements

Au terme de ce travail, je tiens à exprimer ma profonde gratitude à tous ceux qui ont contribué à sa réalisation.

Je remercie chaleureusement l'ensemble des membres du jury pour l'honneur qu'ils me font en évaluant ce travail. Mes sincères remerciements s'adressent particulièrement à **Pr. YAHYAOUI Sami**, président du jury, pour le temps consacré à l'examen de ce mémoire et pour ses remarques constructives.

J'exprime toute ma gratitude à **Pr. AKKAL Rezki**, promoteur du projet, pour son encadrement bienveillant, ses conseils précieux et son accompagnement scientifique rigoureux tout au long de ce travail. Ses orientations méthodologiques ont été déterminantes dans l'accomplissement de ce projet.

Je remercie également **Mr. CHANANE Larouci**, examinateur, pour ses observations pertinentes et ses échanges enrichissants, ainsi que **Mr. BOUSBAI M'hamed**, représentant de l'incubateur de l'ENP, pour son soutien dans le volet entrepreneurial de ce projet.

Ma reconnaissance s'adresse à **tous les enseignants du département de génie minier de l'École Nationale Polytechnique** pour la qualité de la formation dispensée et leur dévouement constant.

Je remercie sincèrement **Mr. Saoudi Idir** qui m'a offert l'opportunité d'effectuer mon stage au sein de **Sonatrach**, ainsi que **Mr. Taleb Mohamed**, mon encadrant de stage, pour sa disponibilité, ses conseils techniques et son encadrement attentif. Ma gratitude va également à **Mr. Berbachi Mohamed**, chef du département *Reservoirs & Reserves* à la direction exploration de Sonatrach, pour son appui et l'accès aux données nécessaires à ce projet.

Je remercie l'ensemble des ingénieurs et log analystes de Sonatrach qui ont participé à la validation de l'outil développé et enrichi ce travail par leurs retours d'expérience.

*Merci à tous.*

# Table des matières

# Liste des tableaux

# Table des figures

# Liste des acronymes

- **AI** : Artificial Intelligence (Intelligence Artificielle)
- **ML** : Machine Learning (Apprentissage Automatique)
- **DL** : Deep Learning (Apprentissage Profond)
- **IA** : Intelligence Artificielle
- **GR** : Gamma Ray
- **RHOB** : Bulk Density Log (Densité en vrac)
- **NPHI** : Neutron Porosity Log (Porosité Neutronique)
- **DT** : Delta Time (Temps de transit acoustique)
- **CAL** : Caliper (Diamétreur)
- **SP** : Spontaneous Potential (Potentiel Spontané)
- **Vsh** : Volume of Shale (Volume d'argile)
- $\Phi$ : Porosité totale
- $\Phi_e$ : Porosité effective
- **Sw** : Water Saturation (Saturation en eau)
- **Sxo** : Saturation en eau de la zone rincée
- **k** : Perméabilité
- **COT** : Carbone Organique Total
- **LAS** : Log ASCII Standard (format de fichiers de diagraphies)
- **VCL** : Volume de Clay (Argile) – variable prédite
- **PIGE** : Pourcentage d'Igneous (Roches ignées)
- **XGB / XGBoost** : Extreme Gradient Boosting (algorithme de boosting d'arbres)
- **LGBM** : Light Gradient Boosting Machine
- **CatBoost** : Categorical Boosting (algorithme de boosting développé par Yandex)
- $R^2$ : Coefficient de Détermination
- **MSE** : Mean Squared Error (Erreur Quadratique Moyenne)
- **MAE** : Mean Absolute Error (Erreur Absolue Moyenne)
- **RMSE** : Root Mean Squared Error
- **SMAPE** : Symmetric Mean Absolute Percentage Error
- **ENP** : École Nationale Polytechnique
- **IAP** : Institut Algérien du Pétrole
- **SPE** : Society of Petroleum Engineers
- **AAPG** : American Association of Petroleum Geologists
- **R&D** : Recherche et Développement
- **UI** : User Interface (Interface Utilisateur)

# Introduction Générale

Le domaine de la géoscience et de l'ingénierie pétrolière a connu un développement considérable ces dernières années, notamment grâce aux avancées technologiques qui permettent une meilleure compréhension des réservoirs souterrains. Dans ce contexte, l'interprétation des données de puits, notamment à travers l'analyse des logs (diagraphie), joue un rôle clé dans la caractérisation des réservoirs et la gestion des ressources énergétiques.

Bien que largement utilisée, la méthode traditionnelle d'interprétation nécessite du temps et présente des coûts associés, ce qui a conduit à une recherche continue de nouvelles approches visant à accélérer le processus tout en améliorant la précision des résultats. Ce travail s'inscrit dans cette dynamique en explorant l'application de l'intelligence artificielle, plus précisément l'apprentissage automatique (machine learning), pour prédire les propriétés lithologiques et pétrophysiques à partir des logs de puits.

L'importance de ce sujet découle des besoins croissants de l'industrie pétrolière et gazière, où l'optimisation de la production d'hydrocarbures repose sur une compréhension approfondie des propriétés des réservoirs. Cette compréhension nécessite une caractérisation rapide et précise des formations géologiques, permettant ainsi de réaliser des gains de temps et de réduire les coûts. Or, les méthodes classiques d'interprétation sont souvent laborieuses et incapables de traiter efficacement les grandes quantités de données générées par les nouvelles technologies de forages. L'application de l'apprentissage automatique pour automatiser l'interprétation des logs de puits apparaît comme une solution prometteuse pour obtenir des résultats aussi fiables que ceux fournis par les experts humains, tout en réduisant considérablement le temps d'analyse.

La problématique centrale de cette recherche est de concevoir et de mettre en place un système de machine learning capable de prédire les propriétés pétrophysiques continues, telles que le volume de l'argile, la saturation en eau et la porosité effective, ainsi que d'identifier les lithologies dominantes et sporadiques du sous-sol. Ce système doit fonctionner avec une précision équivalente à celle des experts humains tout en étant capable de généraliser les résultats à des puits non utilisés lors de l'entraînement du modèle. Ce défi repose sur la nécessité de développer une approche adaptée à différents types de formations géologiques et capable de gérer la variabilité des données, tout en assurant une grande fiabilité dans l'interprétation.

L'approche méthodologique employée dans cette thèse repose sur l'utilisation de techniques d'apprentissage supervisé et non supervisé adaptées à l'analyse des données géophysiques. Les données utilisées proviendront de plusieurs types de logs de puits, tels que les logs de résistivité, gamma-ray, densité et sonique. Les méthodes d'analyse incluront des techniques d'apprentissage automatique telles que les réseaux de neurones profonds (deep learning) et les arbres de décision. Le cadre expérimental sera composé de l'application de ces algorithmes à des ensembles de

données provenant de réservoirs géologiques variés, avec une validation croisée pour évaluer la robustesse et la capacité de généralisation du modèle.

Ce mémoire se structure en quatre grandes étapes :

- **Fondements théoriques :** Présentation du contexte géologique du bassin d'Amguid Messaoud, des diagraphies et paramètres pétrophysiques, ainsi qu'une introduction au Machine Learning avec un focus sur les algorithmes de gradient boosting (XGBoost, CatBoost, LightGBM).
- **Méthodologie et données :** Description des données utilisées (neuf puits d'entraînement et un puits de validation), analyse du déséquilibre lithologique et présentation de la stratégie multi-cible pour l'adaptation des modèles.
- **Résultats et validation :** Présentation des prédictions obtenues, validation externe sur le puits de contrôle et interprétation pétrophysique et lithologique des réservoirs.
- **Perspectives de valorisation :** Proposition de valorisation du projet sous forme de startup, avec un modèle économique, une étude de marché et des perspectives d'évolution.

# Chapitre 1

# Généralités sur la zone d'étude

## 1.1 Introduction

La plate-forme saharienne, vaste région située au sud de l'Algérie alpine et appartenant au craton nord-africain, repose sur un socle précambrien surmonté en discordance par une épaisse couverture sédimentaire, et se caractérise par une stabilité tectonique notable [8]

Du point de vue pétrolier, le domaine saharien est classiquement subdivisé en trois grandes provinces : la province occidentale (Ahnet, Timimoun, Reggane), la province triasique (Tilrhemt, Talemzane, Amguid–Hassi Messaoud, Oued Mya) et la province orientale (Illizi, Berkine, Mouydir), comme le montre la Figure 1.1. Cette plateforme regroupe également les plus grands gisements du pays, notamment Hassi Messaoud pour le pétrole et Hassi R'mel pour le gaz [3].

L'étude menée dans ce travail porte sur la province tiasique, et plus particulièrement sur le bassin Amguid–Messaoud, qui englobe le champ de Hassi Messaoud ainsi que plusieurs gisements voisins. Dans cette région, les réservoirs sont majoritairement classés comme *simples*, c'est-à-dire argileux-gréseux, par opposition aux réservoirs *complexes* de type carbonaté [9].

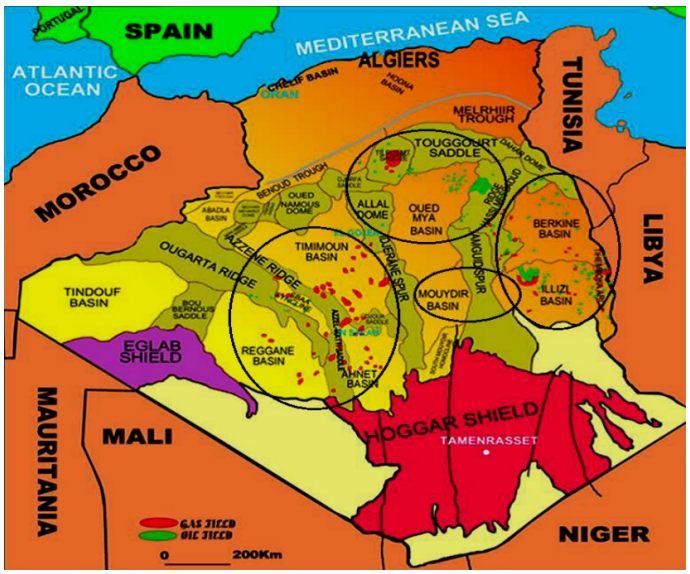


FIGURE 1.1 – Répartition des principaux bassins sédimentaires de la plateforme saharienne algérienne.

## 1.2 Cadre géologique

### 1.2.1 Géologie régionale

Dans la partie septentrionale de la plate-forme saharienne, le môle Amguid–Messaoud est recouvert par une succession sédimentaire pouvant atteindre 6 000 m d'épaisseur. Celle-ci débute par des dépôts paléozoïques réduits jusqu'à l'Ordovicien et le Cambrien, puis se poursuit par une couverture mésozoïque discordante, du Trias au Crétacé, surmontée d'une mince couverture cénozoïque, représentée par des dépôts détritiques d'âge éocène et mio-pliocène. La Figure 1.2 illustre la colonne stratigraphique type de la région, en mettant en évidence les principales unités lithostratigraphiques et leurs âges et épaisseurs ainsi que les différentes discordances [8].

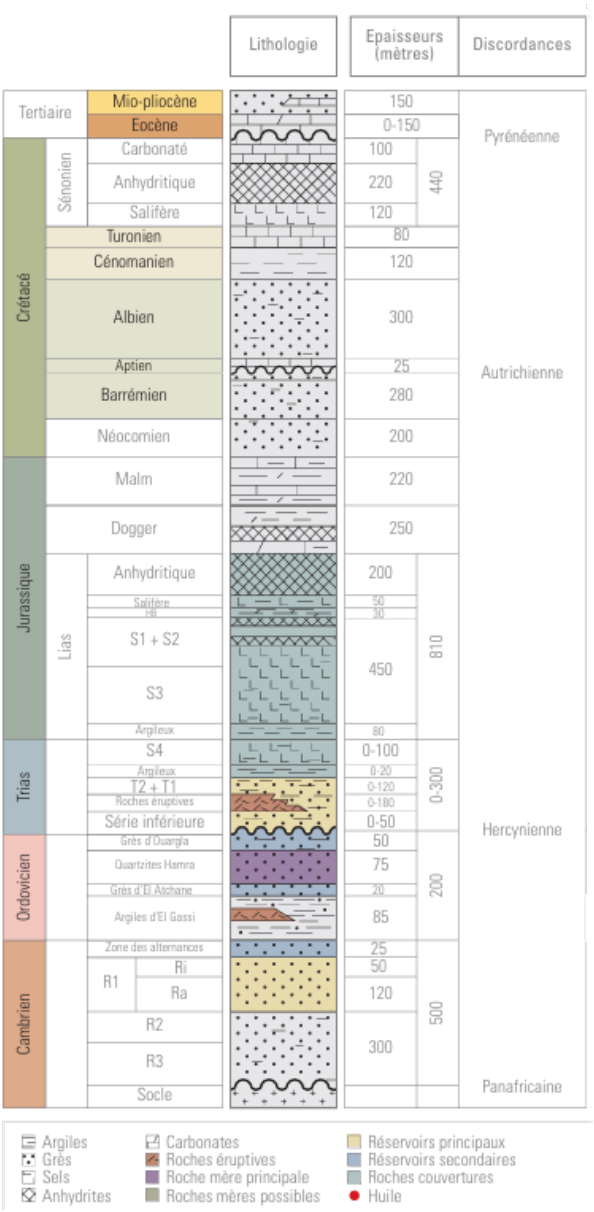


FIGURE 1.2 – Coupe lithostratigraphique type du bassin Amguid Messaoud [3].

Ce domaine structural renferme d'importantes accumulations d'hydrocarbures mises en évidence dans les niveaux cambrien, ordovicien et triasique.

Les secteurs central et méridional de la dorsale se distinguent par une forte concentration de pièges anticlinaux. Par endroits, l'érosion a mis en contact direct les réservoirs paléozoïques et triasiques, favorisant la migration et le remplissage des structures. Les découvertes les plus significatives concernent les zones périphériques de Hassi Messaoud, où les réservoirs cambrien et ordovicien constituent les principaux horizons producteurs, comme le montre la coupe transversale présentée en Figure 1.3 [10].

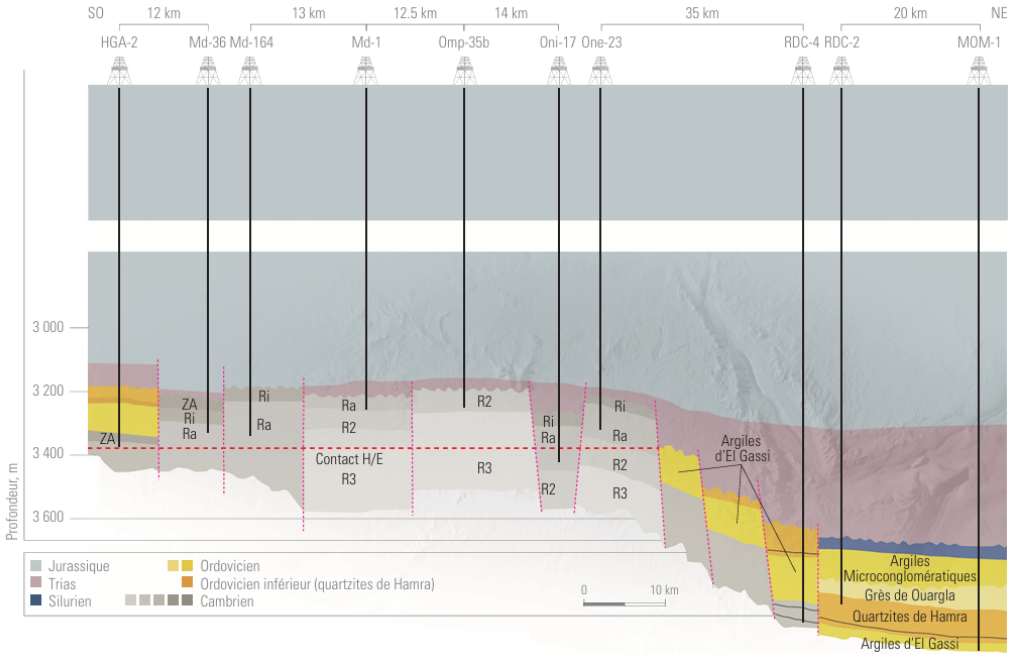


FIGURE 1.3 – Coupe transversale schématique du champ de Hassi Messaoud [3].

La région a été marquée par plusieurs épisodes tectoniques majeurs. Le premier correspond à l'*orogenèse hercynienne*, responsable d'un soulèvement régional et d'une érosion intense des dépôts paléozoïques, plus marquée dans les zones centrales que sur les marges. Au Trias, une phase de *rifting* réactive les failles hercyniennes d'orientation NE–SW, accompagnée localement par la mise en place de roches ignées au sommet des séries érodées. Enfin, au Crétacé, une inversion tectonique provoque un réajustement des failles selon une direction NNE–SSW et un basculement des bassins [3].

### 1.2.2 Géologie locale

Le champ de Hassi Messaoud, situé à environ 800 km au sud-est d'Alger, constitue un vaste réservoir gréseux d'environ 2 000 km². Il se développe au sein d'une structure anticlinale aplatie, localisée au nord de l'élévation El-Biod–Hassi Messaoud (Figure 1.4), elle-même constituée d'un enchaînement de *Horsts* et *Grabens* orientés globalement SSW–NNE.

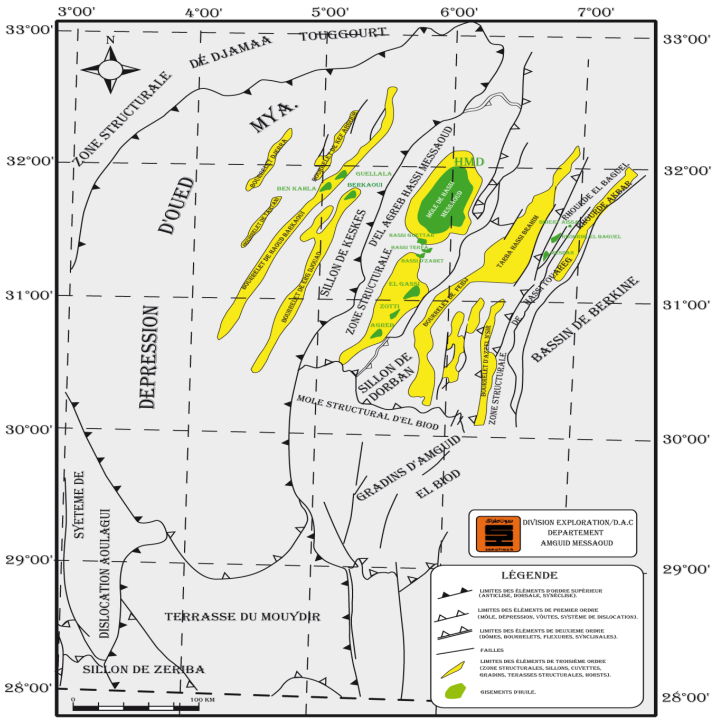


FIGURE 1.4 – Schéma structural de la région de Hassi Messaoud [2].

Le réservoir producteur, d'âge cambrien–ordovicien, se situe vers 3 300 m de profondeur. Il est subdivisé en quatre formations principales (Ri, Ra, R2, R3) auxquelles s'ajoute une zone d'alternance (ZA), l'ensemble étant compartimenté en drains producteurs distincts selon leurs propriétés pétrophysiques [11].

À environ 35 km au nord-est du champ de Hassi Messaoud se situe la région de Touggourt Est (I, II, III) (Figure 1.5), où se concentre la majorité des puits utilisés dans ce travail pour l'entraînement des modèles de *Machine Learning*. Cette zone renferme des formations paléozoïques (pré-hercyniennes) et triasiques (post-hercyniennes), constituant un système multizone.

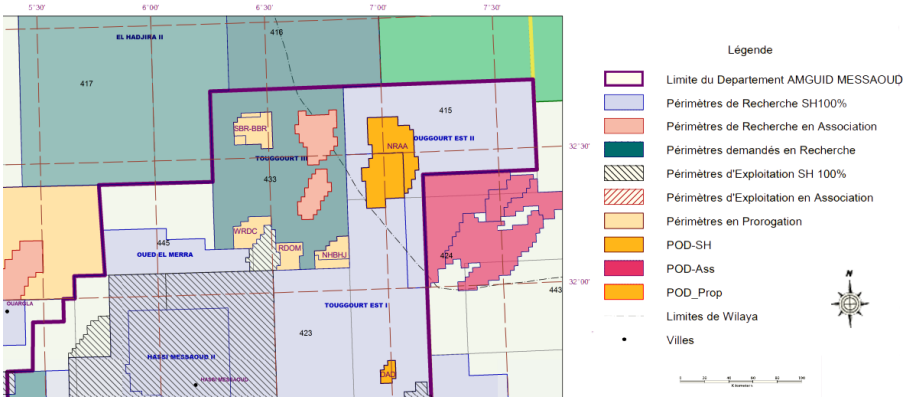


FIGURE 1.5 – Zoom cartographique sur la région de Touggourt [2].

La séquence réservoir est structurée en deux horizons principaux : le triasique (Trias inférieur, T1) et l'ordovicien (Quartzites de Hamra, QH). Des réservoirs secondaires peuvent toutefois apparaître localement, par exemple la série inférieure ou, dans certains secteurs, les grès de Rhourd Chegga (RDC), leur présence et leur continuité variant selon la position des puits et la variabilité latérale des dépôts [12].

## 1.3 Système pétrolier

Un système pétrolier est dit fonctionnel lorsqu'il existe une continuité entre la génération des hydrocarbures dans la *roche-mère*, leur migration et leur accumulation dans un réservoir, où ils sont piégés et préservés grâce à une couverture imperméable [13].

### 1.3.1 Roches mères et migration

Le Silurien constitue la principale roche-mère des bassins de l'Oued Mya et d'Amguid–Hassi Messaoud. Il est représenté par un niveau basal d'argiles radioactives noires, riches en matière organique, dont l'épaisseur varie généralement entre 50 et 85 m. Son développement est particulièrement marqué vers l'est de Tinhert et vers le sud-ouest de la plate-forme. Les teneurs en *carbone organique total* (COT) y atteignent des valeurs remarquables, comprises entre 9 et 11%, ce qui en fait une roche-mère de très bonne qualité, comme l'illustre la distribution spatiale présentée en Figure 1.6 [3].

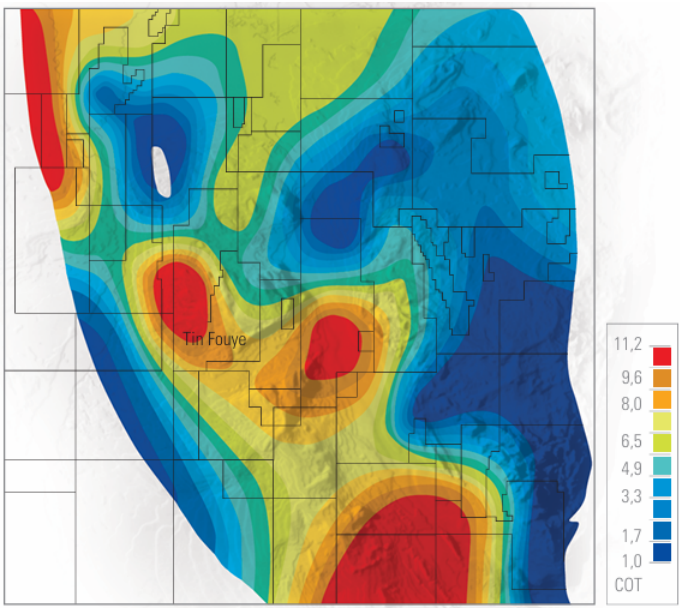


FIGURE 1.6 – Distribution spatiale du COT dans le Silurien [3].

L'évolution thermique de cette formation a conduit à plusieurs phases de génération d'hydrocarbures. Une première phase ancienne a favorisé principalement la production d'huiles, tandis qu'une seconde, postérieure à l'orogenèse hercynienne, a généré successivement du gaz humide puis du gaz sec durant le Crétacé et le Tertiaire [3].

Deux périodes principales d'expulsion ont été mises en évidence. La première, précoce, s'est produite entre le Carbonifère et le Jurassique, alimentant en priorité les régions nord-est et ouest de la plate-forme d'Illizi. La seconde, plus tardive, correspond au Crétacé supérieur et au Tertiaire, et a concerné les zones situées au nord-ouest et au sud de la plate-forme, contribuant ainsi au remplissage des pièges structuraux et stratigraphiques régionaux [3].

### 1.3.2 Réservoirs

Les réservoirs cambrien et cambro-ordovicien sont constitués d'une puissante série détritique reposant directement sur le socle précambrien. Leur épaisseur atteint environ 150 m et se compose principalement de grès, quartzites et conglomérats [10].

A. **Cambrien**

- **R3 :** grès grossiers à conglomératiques, mal classés, feldspathiques et argileux, interprétés comme un comblement de pédiplaine infra-tassilienne, avec un milieu de sédimentation probablement *deltaïque* [10].
- **R2 :** grès moyens à grossiers, mal classés, très argileux, avec de fréquentes passées d'argiles [10].
- **Ra :** grès quartzitiques, de granulométrie variable, associés à des intercalations argileuses [10].

B. **Cambro-Ordovicien**

- **RI :** grès bien classés, glauconieux, riches en fossiles (*Tigillites*, *Lingulidae*), traduisant un environnement littoral à marin peu profond et calme [10].
- **Zone des alternances :** alternance de grès et d'argiles, reflétant des conditions de dépôts variables [10].

C. **Ordovicien**

Le réservoir ordovicien est représenté par les Quartzites de Hamra (QH), formation massive et compacte d'une épaisseur moyenne de 90 m. Elle est constituée de grès quartzitiques à quartzites, de teinte blanche à gris-blanc, généralement fins à moyens, parfois localement grossiers, avec des intercalations d'argiles noires silteuses et feuilletées [3].

L'analyse des niveaux carottés met en évidence plusieurs *lithofaciès*, principalement :

- Grès quartzitiques fins à très fins, à litage horizontal ou oblique ;
- Grès fins quartzitiques bioturbés ;
- Grès à litage entrecroisé et granoclassé ;
- Grès fins à très fins contenant des copeaux, galets ou films argileux.

D. **Trias**

Le Trias se caractérise par des dépôts volcano-détritiques et lagunaires, constituant l'un des principaux réservoirs de la région. Le bassin triasique, vaste dépression d'environ 200 000 km², a été comblé par une succession de faciès variés où la puissance actuelle des sédiments varie de 0 à 500 m, contrôlés à la fois par les environnements de sédimentation et la proximité des zones d'apport [8] [10].

Les principaux réservoirs correspondent aux niveaux T1 et T2, formés de séquences gréso-argileuses où la granulométrie évolue d'un matériel grossier à la base vers des niveaux plus fins au sommet. Ces dépôts, typiques d'un milieu *fluviatile*, sont constitués de grès brun-rouge, fins à moyens, associés à des argiles silteuses brun-rouge, localement dolomitiques [10].

### 1.3.3 Types de pièges

L'exploration repose principalement sur des pièges structuraux et mixtes, avec une attention particulière aux pièges stratigraphiques liés aux lentilles gréseuses du Trias. Les plis anticlinaux, associés aux phases tectoniques de la dorsale Amguid–Hassi Messaoud, ainsi que les corps gréseux fermés par biseautage latéral, constituent les principaux pièges identifiés, comme le schématise la figure 1.7 [3].

Autour de Hassi Messaoud, le Silurien radioactif aurait généré environ 1 080 milliards de barils d'huile et 730 trillions de pieds cubes de gaz, dont 850 milliards de barils d'huile et la totalité du gaz auraient été expulsés. Avec un coefficient moyen de piégeage de 12%, les volumes effectivement piégés s'élèveraient à près de 102 milliards de barils d'huile, illustrant le potentiel considérable de ces pièges [3].

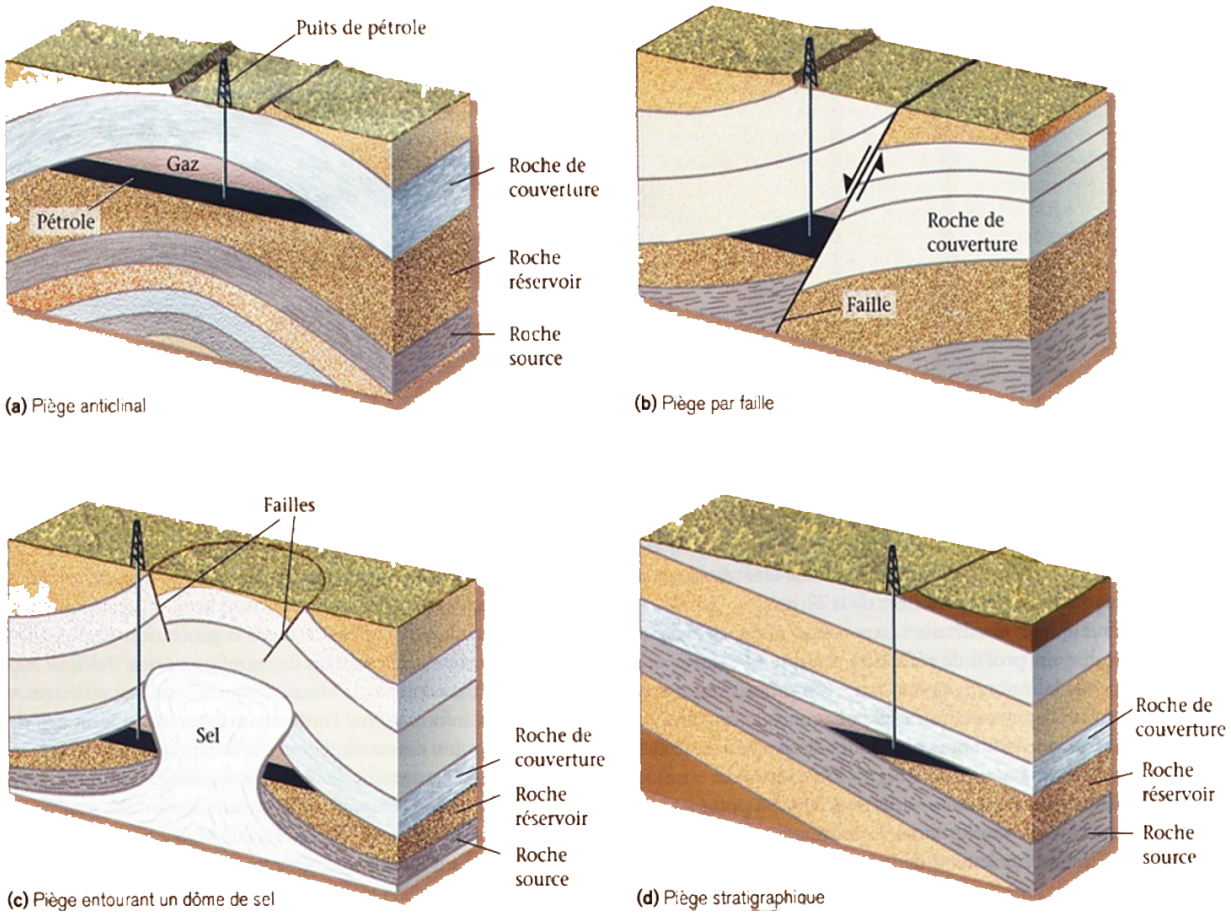


FIGURE 1.7 – Schéma caractérisant les différents types de pièges [3].

### 1.3.4 Roches couvertures

Le secteur Amguid–Messaoud présente plusieurs niveaux de couverture assurant l'étanchéité des réservoirs. Les argiles d'El Gassi couvrent les grès cambriens, celles d'Azzel protègent le quartzite de Hamra (Ordovicien), tandis que les dépôts argilo-évaporitiques du Trias et du Lias forment les couvertures les plus efficaces [3].

## 1.4 Conclusion

Ce chapitre a présenté le cadre géologique régional et local de la zone d'étude ainsi que les principaux éléments du système pétrolier. L'ensemble des informations sur la stratigraphie, la lithologie, la structuration tectonique, les réservoirs, les pièges et les couvertures permet de mieux comprendre l'organisation géologique et les conditions de mise en place des hydrocarbures. Ces éléments constituent une base indispensable pour l'analyse des données de diagraphies et l'application des approches de modélisation développées dans le projet.

# Chapitre 2

# Caractérisation pétrophysique et lithologique à partir des diagraphies

## 2.1 Introduction

Les diagraphies constituent un ensemble de mesures physiques réalisées le long des puits forés, permettant de caractériser les formations géologiques traversées. L'objectif principal du géologue est de déterminer, à partir de l'interprétation de ces diagraphies, d'une part le **contenant** — c'est-à-dire la nature minéralogique et le pourcentage des éléments solides constituant la roche — et d'autre part le **contenu** — soit la nature et le pourcentage des fluides (porosité, saturation) remplissant les interstices entre ces éléments solides [9].

Les diagraphistes distinguent généralement deux catégories d'éléments solides : la matrice et l'argile. Cette distinction s'impose pour deux raisons : premièrement, ces deux types de solides présentent des comportements différents vis-à-vis des phénomènes physiques exploités en diagraphie ; deuxièmement, l'argile exerce une influence particulière sur les propriétés pétrophysiques des réservoirs, notamment sur la porosité et la saturation en eau. [9]

Dans le cadre de cette étude, nous nous concentrons principalement sur les diagraphies en trou ouvert (*Openhole Logging*), réalisées immédiatement après le forage et avant la cimentation du tubage. Ce chapitre est structuré en deux parties principales : la première présente les différents types de diagraphies et détaille les principes physiques des outils utilisés dans cette étude ; la seconde partie aborde les méthodes d'interprétation pétrophysique, notamment l'approche Quanti-ELAN mise en œuvre par Sonatrach Exploration.

## 2.2 Diagraphies

Les diagraphies, également appelées *logs*, désignent l'enregistrement en continu, le long d'un puits, de paramètres physiques du sous-sol (résistivité, radioactivité, vitesse acoustique, densité, etc.) et leur interprétation en termes de caractéristiques géologiques (porosité, saturation en eau, argilosité, lithologie, épaisseur, etc.). Ces mesures sont effectuées à l'aide d'outils spécialisés descendus au bout d'un câble électrique (*wireline*) dans le trou de forage [1].

De nombreux paramètres couvrant pratiquement tous les domaines de la physique sont aujourd'hui mesurés, fournissant une représentation verticale détaillée des propriétés des formations géologiques traversées. Les données diagraphiques constituent une source d'information fondamentale pour l'exploration et l'exploitation des réservoirs pétroliers, permettant la caractérisation des formations sans nécessiter de prélèvement physique systématique [4].

### 2.2.1 Classification des diagraphies

Les paramètres enregistrés lors des diagraphies sont des propriétés physiques des formations traversées, mesurées par des outils appropriés descendus au bout d'un câble dans le trou de forage. On distingue deux catégories de paramètres selon leur nature :

- **Paramètres naturels** : engendrés spontanément par la formation et détectés à l'aide de capteurs passifs (exemple : radioactivité naturelle gamma, potentiel spontané) [1] ;
- **Paramètres induits** : nécessitant l'émission d'un signal (électrique, acoustique ou nucléaire) et la mesure de la réponse de la formation (exemple : résistivité, densité, porosité neutron) [1].

### 2.2.2 Les diagraphies utilisées

A. **Diagraphie de rayonnement gamma naturel (Gamma Ray — GR)**

La diagraphie de rayonnement gamma naturel mesure la radioactivité naturelle des formations géologiques, principalement issue de la désintégration des isotopes radioactifs du potassium ($^{40}$K), du thorium (Th) et de l'uranium (U). L'isotope radioactif du potassium ne représente qu'environ 0,0118 % du potassium total, mais son abondance moyenne dans la croûte terrestre assure une contribution notable à la radioactivité naturelle. Le thorium et l'uranium, bien que présents à des teneurs plus faibles (quelques ppm), possèdent une activité $\gamma$ massique nettement plus élevée, respectivement environ 1300 et 3600 fois celle du potassium (Tableau 2.1) [1].

TABLE 2.1 – Contribution à la radioactivité $\gamma$ de trois éléments radioactifs [1]

| | **K** | **Th** | **U** |
|---|---|---|---|
| Abondance relative dans la croûte terrestre (%) | 2,35 | $10^{-4}$ à $10^{-5}$ | $10^{-5}$ à $10^{-6}$ |
| | | $\sim$ 12 ppm | $\sim$ 3 ppm |
| Activité $\gamma$ relative par unité de poids | 1 | 1.300 | 3.600 |

Cette différence explique pourquoi, dans les mesures spectrométriques, le potassium est généralement exprimé en pourcentage, tandis que le thorium et l'uranium le sont en ppm [1].

On enregistre soit une courbe globale incluant les trois rayonnements K, Th et U (Gamma Ray total, $GR_{total}$), soit trois courbes distinctes obtenues par spectrométrie gamma naturelle. La présence localisée d'uranium, notamment dans certains niveaux sableux enrichis,

peut augmenter artificiellement la radioactivité mesurée et fausser l'interprétation lithologique (Figure 2.1). Pour pallier cet effet, un Gamma Ray corrigé ($GR_{corrigé}$ ou KTH) est calculé en excluant la contribution de l'uranium, offrant ainsi une lecture plus fiable de la nature lithologique des formations [4].

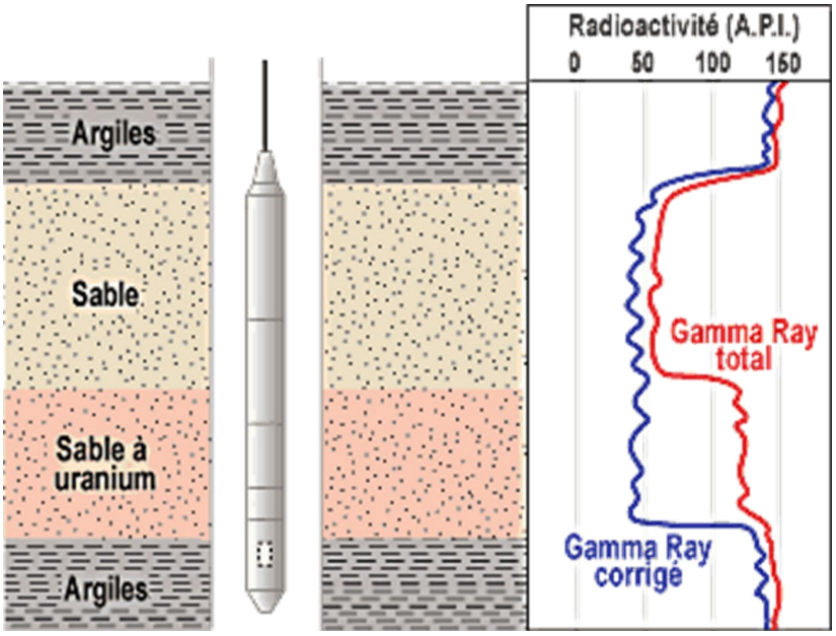


FIGURE 2.1 – Principe de mesure du Gamma Ray et distinction entre GR total et GR corrigé [4].

**Identification lithologique par spectrométrie gamma**

L'analyse différentielle des trois composantes radioactives (K, Th, U) permet d'affiner l'identification lithologique :

- **Évaporites potassiques** : caractérisées par une forte teneur en potassium associée à une absence de thorium et d'uranium, permettant leur identification directe dans les séries évaporitiques [1] ;
- **Argiles** : le rapport Th/K constitue un bon indicateur de la nature du minéral argileux, bien que les concentrations en thorium et potassium puissent varier selon le type d'argile [1] ;
- **Roches organiques** : une teneur élevée en uranium caractérise généralement les roches riches en matière organique [1] ;
- **Sables et grès** : la teneur en thorium est directement liée au pourcentage d'argile présent dans la formation [1].

B. **Diagraphie Gamma-Gamma ou Densité (RHOB)**

La diagraphie de densité repose sur le principe du bombardement de la formation par une source de rayons gamma ($\gamma$) à énergie comprise entre 0,1 et 1 MeV. Ces rayons gamma entrent en collision avec les atomes de la formation et perdent de leur énergie selon trois effets principaux :

- **Effet photoélectrique** : Le photon gamma est complètement absorbé par un électron qui est éjecté de l'atome [4] ;

- **Effet Compton** : L'effet le plus fréquent et le principe utilisé en diagraphie de densité. Le rayon gamma perd une partie de son énergie, éjecte un électron et continue sa trajectoire sous forme de photon diffusé à énergie réduite [1] ;
- **Effet de production de paires** : À très haute énergie, le photon gamma se transforme en une paire électron-positron [4].

Un ou plusieurs détecteurs placés sur l'outil reçoivent les rayons gamma diffusés par la formation vers le puits. Le signal reçu est fonction du nombre d'électrons par $cm^3$ de volume de la formation, c'est-à-dire de la densité électronique de la formation.

**Relation entre densité électronique et densité globale**

L'objectif est de déterminer la densité globale de la formation. Le nombre d'électrons par unité de volume ($\rho_e$) est relié à la densité globale ($\rho_b$) par la relation :

$$\rho_e = \rho_b \left(\frac{Z}{A}\right) N \tag{2.1}$$

avec :

- $\rho_e$ : densité électronique
- $\rho_b$ : densité globale (bulk density)
- $Z$ : numéro atomique
- $A$ : masse atomique
- $N$ : nombre d'Avogadro ($6{,}02 \times 10^{23}$ $mol^{-1}$)

Le rapport $Z/A$ est très proche de 0,5 pour la plupart des éléments et composés constituant les roches, à l'exception de l'hydrogène pour lequel il est proche de 1. Cette relation permet de convertir la mesure de densité électronique en densité globale de la formation, paramètre essentiel pour l'estimation de la porosité [1].

C. **Diagraphie Neutron (NPHI)**

La diagraphie neutron repose sur le bombardement des formations par des neutrons rapides émis par une source radioactive. Ces neutrons possèdent une énergie initiale comprise entre 4 et 6 MeV et une vitesse initiale élevée (environ 10 000 km/s), leur conférant un grand pouvoir de pénétration dans la formation [4].

**Principe physique**

Les neutrons rapides entrent en collision avec les noyaux des atomes des formations qu'ils traversent, perdant progressivement leur énergie au cours de ce processus de ralentissement [1]. Les outils neutron peuvent mesurer :

- **Les neutrons ralentis** (neutrons épithermiques) – Outil CNL (*Compensated Neutron Log*)
- **Les photons gamma ($\gamma$) de capture** (neutrons thermiques) – Outil SNP (*Sidewall Neutron Porosity*)

Les outils actuels mesurent principalement les neutrons épithermiques, car leur ralentissement est généralement contrôlé par les atomes d'hydrogène, directement liés à la teneur en eau et donc à la porosité de la formation [1].

**Interprétation de la mesure**

Le nombre de neutrons arrivant au détecteur varie en fonction de l'indice d'hydrogène de la formation :

- **Diminue dans les grandes porosités** : Un fort indice d'hydrogène (porosité élevée saturée en fluides) ralentit davantage les neutrons, réduisant le nombre de neutrons détectés
- **Augmente dans les faibles porosités** : Un faible indice d'hydrogène (porosité faible) ralentit moins les neutrons, augmentant le nombre de neutrons détectés

La mesure NPHI (Neutron Porosity Hydrogen Index) est ainsi directement corrélée à la porosité de la formation et constitue, avec la diagraphie de densité, un outil fondamental pour l'estimation de la porosité des réservoirs [4].

La combinaison des diagraphies Densité et Neutron permet d'identifier la nature des fluides présents dans les réservoirs. Lorsque les échelles sont compatibles, l'analyse comparative des deux courbes fournit des informations précieuses (Figure 2.2) :

- **Réservoir à eau** : Les courbes Densité et Neutron coïncident, indiquant une cohérence entre les mesures de porosité
- **Réservoir à gaz** : Une grande séparation des courbes Neutron-Densité est observée. Le gaz, ayant un faible indice d'hydrogène et une faible densité, crée un écart important entre les deux mesures (effet crossover)
- **Réservoir à huile** : Une petite séparation des courbes est notée, intermédiaire entre l'eau et le gaz

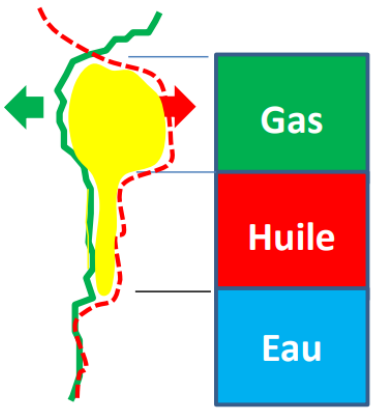


FIGURE 2.2 – Identification des fluides par combinaison des diagraphies Densité-Neutron [4]

D. **Diagraphies acoustiques (Soniques)**

À l'aide d'un générateur à magnétostriction excité depuis la surface par envoi d'une impulsion électrique, l'outil émet un train d'ondes acoustiques dont la fréquence moyenne est de l'ordre de 20 à 40 kHz. Cette émission, de très courte durée (moins de 1 ms), est répétée plusieurs fois par seconde (10 à 60 fois suivant le type d'outil) [1].

Le train d'ondes émis se propage dans tout l'espace à partir du générateur (émetteur E),

suivant des fronts d'ondes sphériques. Il traverse la colonne de boue et atteint la paroi du trou en des temps et sous des angles d'incidence croissants [1].

À l'origine, les principes acoustiques étaient développés pour aider l'interprétation des données sismiques, mais il a été constaté que ces mesures apportent une contribution majeure pour :

- L'estimation de la porosité des roches
- Les corrélations entre sondages
- La définition de la lithologie
- La détection de fractures et l'évaluation des propriétés mécaniques des formations

Il existe deux types d'ondes acoustiques principales :

**Ondes longitudinales (ondes P - Primaires ou de compression)**

Les ondes P constituent le type d'onde principalement utilisé en diagraphie sonique. Leurs caractéristiques sont :

- Le déplacement des particules s'effectue par dilatations et compressions successives, parallèlement à la direction de propagation de l'onde
- Ces ondes se propagent dans les solides, les liquides et les gaz
- Elles sont les plus rapides (vitesse typique d'environ 6 km/s dans les roches) et arrivent les premières au récepteur
- La mesure du temps de transit de ces ondes permet d'estimer la porosité de la formation

**Ondes transversales (ondes S - Secondaires ou de cisaillement)**

Les ondes S présentent des caractéristiques différentes :

- Le mouvement des particules s'effectue perpendiculairement au sens de propagation de l'onde
- Ces ondes ne se propagent pas dans les liquides, uniquement dans les solides
- Leur vitesse est plus faible (environ 4 km/s dans les roches)
- Elles apparaissent en second sur les récepteurs, après les ondes P
- Elles sont utilisées pour l'évaluation des propriétés mécaniques des roches

Le paramètre mesuré en diagraphie sonique est le temps de transit ($\Delta t$), exprimé en microsecondes par pied (µs/ft), qui représente le temps mis par l'onde pour parcourir une distance donnée dans la formation. Ce temps de transit est inversement proportionnel à la vitesse de propagation et dépend directement de la porosité de la roche [1].

E. **Diamétreur (Caliper - CALX)**

Le diamétreur est une diagraphie auxiliaire mesurant le diamètre du puits. Bien que n'apportant pas directement d'informations sur les propriétés pétrophysiques, cette mesure est effectuée avec pratiquement tous les outils de diagraphie et s'avère indispensable dans les interprétations diagraphiques car elle permet d'appliquer les corrections nécessaires liées aux effets du trou sur les autres mesures [4].

Le diamétreur fournit d'importantes informations sur l'état du puits :

- **Caves** : Zones d'élargissement du trou indiquant des formations friables ou fracturées
- **Mud-cake** : Dépôt de filtrat de boue sur la paroi du puits, indiquant des zones perméables
- **Argiles gonflantes** : Zones de rétrécissement du trou dues au gonflement des argiles au contact de la boue de forage
- **Estimation du volume de ciment** : Calcul du volume nécessaire pour la cimentation du puits
- **Forme de la section du puits** : Identification de l'ovalisation ou des irrégularités du trou

La mesure du caliper est exprimée en pouces (inches) et constitue un log de contrôle qualité essentiel pour valider les autres mesures diagraphiques et identifier les zones nécessitant des corrections spécifiques [4].

## 2.3 Paramètres pétrophysiques

Les paramètres pétrophysiques désignent les propriétés physiques caractérisant les roches réservoirs et leur capacité à contenir et à laisser circuler les fluides. Les principaux paramètres pétrophysiques déterminés à partir des diagraphies sont les suivants :

### 2.3.1 Porosité ($\Phi$)

La porosité est la fraction du volume d'une roche non occupée par des éléments solides. Elle est définie comme le rapport du volume total des espaces vides (pores, canalicules, vacuoles, géodes, etc.) existant entre les éléments minéraux de la roche, au volume total de la roche :

$$\Phi = \frac{V_p}{V} = \frac{V - V_s}{V} \tag{2.2}$$

où $V_p$ représente le volume des espaces vides (généralement occupés par des fluides : eau, gaz, huile), $V_s$ le volume occupé par les éléments solides, et $V$ le volume total de la roche [14].

On distingue la **porosité totale** ($\Phi_{tot}$) qui englobe :

- La **porosité primaire** ($\Phi_1$) : porosité intergranulaire ou intercristalline, dépendant de la forme, de la taille et du classement des éléments solides, rencontrée surtout dans les roches clastiques.
- La **porosité secondaire** ($\Phi_2$) : porosité vacuolaire (acquise par dissolution) et porosité de fissures et fractures (acquise mécaniquement), rencontrée le plus souvent dans les roches chimiques ou biochimiques [1].

Ainsi, la porosité totale peut s'écrire : $\Phi_{tot} = \Phi_1 + \Phi_2$.

### 2.3.2 Perméabilité ($k$)

Un milieu poreux ne permet le déplacement des fluides que dans la mesure où ses pores sont reliés entre eux. On dit alors qu'il est perméable. La perméabilité, désignée par la lettre $k$, mesure la facilité avec laquelle une formation permet à un fluide de viscosité donnée de la traverser. Si le fluide est homogène et n'a aucune action chimique importante sur le milieu encaissant, la perméabilité est dite *absolue* ($k$). Elle est généralement exprimée en millidarcy (mD) [1].

### 2.3.3 Volume d'argile ($V_{sh}$)

Le volume d'argile représente la fraction volumique d'argile présente dans la formation. La diagraphie Gamma Ray (GR) est utilisée comme log de base pour le calcul du volume d'argile, qui constitue un indicateur notable de la qualité du réservoir. La détermination de $V_{sh}$ est une étape importante car elle impacte directement le calcul de la porosité effective [15]. La méthode linéaire est donnée par :

$$V_{sh} = \frac{GR - GR_{matrix}}{GR_{shale} - GR_{matrix}} \tag{2.3}$$

où $GR_{matrix}$ représente les valeurs du log GR dans une roche 100 % matrice (grès propre), $GR_{shale}$ indique les valeurs GR dans 100 % d'argile, et $GR$ correspond à la valeur à une profondeur donnée [15].

### 2.3.4 Saturation en fluides ($S$)

Il est essentiel de connaître la nature des fluides qui occupent les pores d'une roche réservoir. La saturation d'une roche en fluide est le rapport du volume de ce fluide sur le volume des pores, exprimé en pourcentage :

$$\text{Saturation} = \frac{\text{Volume de Fluide}}{\text{Volume des Pores}} \tag{2.4}$$

Dans le cas d'un gisement à hydrocarbures, les pores contiennent de l'eau, de l'huile ou du gaz. On définit ainsi une saturation en eau ($S_w$), une saturation en huile ($S_o$) et une saturation en gaz ($S_g$) [14].

Dans le cas d'une roche saturée, Archie a établi une relation expérimentale liant la résistivité de la roche à la porosité et à la résistivité de l'eau d'imbibition :

$$S_w^n = \frac{a \cdot \rho_w}{\Phi^m \cdot \rho_t} \tag{2.5}$$

avec :

- $\rho_w$ : résistivité de l'eau d'imbibition ;
- $\Phi$ : porosité ;

- $\rho_t$ : résistivité vraie de la formation ;
- $a$ : facteur dépendant de la lithologie (entre 0,6 et 2) — $a < 1$ pour les roches à porosité intergranulaire et $a > 1$ pour les roches à porosité de fracture ;
- $m$ : facteur de cimentation (varie entre 1,3 pour les sables non consolidés à 2,2 pour les calcaires cimentés) ;
- $n$ : exposant de saturation (généralement proche de 2).

### 2.3.5 Porosité effective ($\Phi_e$)

La porosité effective représente la porosité accessible aux fluides mobiles, excluant la porosité liée aux argiles. Elle correspond au volume poreux réellement disponible pour le stockage des hydrocarbures et dépend directement du volume d'argile présent dans la formation [15].

## 2.4 Interprétation des diagraphies

L'interprétation des diagraphies se divise en deux approches complémentaires permettant d'exploiter les données mesurées pour caractériser les formations géologiques et évaluer leur potentiel pétrolier.

### 2.4.1 Interprétation qualitative (Quick Look)

L'interprétation qualitative consiste en une analyse visuelle rapide des diagraphies permettant de :

- **Découpage en zones homogènes** : Identification des différentes unités lithologiques et limites de formations par corrélation des signatures diagraphiques
- **Identification des fluides en place** : Détection de la présence d'hydrocarbures (gaz, huile) ou d'eau par lecture des signatures caractéristiques des diagraphies (séparation Densité-Neutron, résistivité élevée, etc.)

Cette approche rapide permet une première évaluation du potentiel des formations avant de procéder à une analyse quantitative détaillée [4].

### 2.4.2 Interprétation quantitative

L'interprétation quantitative vise à calculer précisément les paramètres pétrophysiques mentionnés précédemment (porosité, saturation, volume d'argile, etc.) en tenant compte des effets correctifs, notamment l'effet de l'argile sur la porosité pour obtenir la porosité effective [4].

#### 2.4.2.1 Méthode déterministe

La méthode déterministe repose sur l'utilisation de relations mathématiques directes (équations linéaires) pour calculer les paramètres pétrophysiques à partir des mesures diagraphiques. Les formules présentées dans les sections précédentes (volume d'argile à partir du Gamma Ray, saturation en eau via la formule d'Archie, etc.) constituent des exemples de cette approche déterministe [4].

#### 2.4.2.2 Méthode probabiliste : Quanti-ELAN

Pour la détermination précise de la lithologie et de la composition minéralogique des formations, une approche probabiliste est nécessaire. Le module Quanti-ELAN de Techlog (Schlumberger) met en œuvre cette méthode d'évaluation quantitative des formations [16].

**Principe de la méthode Quanti-ELAN**

Quanti-ELAN résout le problème inverse en optimisant simultanément des équations décrivant un ou plusieurs modèles d'interprétation. La méthode repose sur un modèle pétrophysique triangulaire (Figure 2.3) représentant les relations entre trois vecteurs fondamentaux :

- **t** : Le vecteur outil, comprenant toutes les mesures diagraphiques et courbes synthétiques
- **v** : Le vecteur volume, représentant les volumes des composants de la formation
- **R** : La matrice de réponse, contenant les valeurs des paramètres que chaque outil mesure pour 100% de chaque composant de formation

La relation fondamentale s'exprime par : $t = R \cdot v$

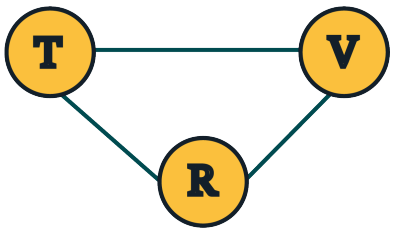


FIGURE 2.3 – Modèle pétrophysique utilisé par Quanti-ELAN

**Composants de formation**

Lors de la configuration du modèle d'interprétation, l'utilisateur définit les composants susceptibles d'être présents dans la formation :

- **Minéraux** : Solides à formule chimique relativement constante ($SiO_2$, $CaCO_3$, $CaSO_4$, etc.) avec des paramètres par défaut généralement fiables
- **Roches** : Combinaisons définies par l'utilisateur de plusieurs minéraux (silt, carbonate, roches ignées, etc.)
- **Fluides** : Substances occupant l'espace poreux (eau, huile, gaz)

Le nombre de composants à résoudre ne peut jamais excéder le nombre total d'équations disponibles (nombre de diagraphies utilisées). Il est recommandé de limiter les composants

sélectionnés uniquement à ceux attendus en quantité appréciable ou ayant un impact significatif sur les mesures diagraphiques [16].

**Résolution du problème**

Le problème inverse résout uniquement les volumes des composants de formation. Les autres résultats d'interprétation traditionnels (saturation en eau, densité de matrice, etc.) sont fournis par une étape de post-traitement, permettant ainsi un contrôle flexible des définitions et des calculs supplémentaires [16].

## 2.5 Conclusion

Ce chapitre a présenté les fondements des diagraphies différées et des méthodes conventionnelles d'interprétation pétrophysique. Les différentes mesures (Gamma Ray, Densité, Neutron, Acoustique) ont été expliquées dans leur principe physique et leur utilité pour caractériser les formations réservoirs. L'interprétation combine une approche qualitative pour l'identification rapide des zones d'intérêt et une approche quantitative via le module Quanti-ELAN, qui résout le problème inverse pour déterminer la composition volumétrique des formations à partir des mesures de diagraphies.

Ces résultats d'interprétation conventionnelle constituent précisément la base de données qui sera exploitée dans le chapitre suivant consacré aux approches par apprentissage automatique. Le chapitre 3 explorera ainsi comment les techniques de régression et de classification peuvent être mises en œuvre pour automatiser et optimiser les processus d'interprétation pétrophysique, en utilisant les données issues des diagraphies comme variables d'entrée pour entraîner les modèles de machine learning.

# Chapitre 3

# Apprentissage automatique (machine learning)

## 3.1 Introduction

L'intelligence artificielle (IA) désigne l'ensemble des théories, méthodes et techniques permettant à une machine de simuler, étendre ou renforcer les capacités cognitives humaines pour accomplir des tâches complexes [17]. Ce domaine englobe plusieurs disciplines interconnectées, dont le machine learning (apprentissage automatique), qui permet aux systèmes d'apprendre à partir de données en identifiant des régularités pour construire des modèles prédictifs. Le deep learning (apprentissage profond), sous-ensemble du machine learning, exploite des réseaux de neurones multicouches pour traiter des informations volumineuses et complexes, comme l'illustre la Figure 3.1 [18].

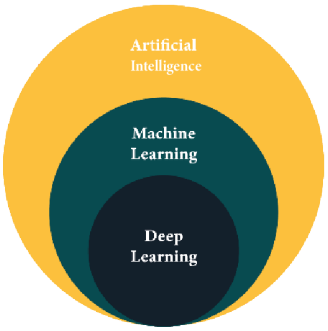


FIGURE 3.1 – Structure hiérarchique des sous-domaines de l'intelligence artificielle [5].

Le Machine Learning s'impose aujourd'hui comme une approche incontournable dans les domaines académique et industriel, avec des applications majeures en santé, finance, cybersécurité, gestion de l'information et prévision [19].

Cette discipline se subdivise en trois branches principales : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. Dans le cadre de cette étude, l'accent est mis sur l'apprentissage supervisé, et plus particulièrement sur deux de ses techniques fondamentales : la classification et la régression, largement utilisées pour résoudre des problèmes prédictifs [20].

## 3.2 Notion de machine learning

Le machine learning, défini par Samuel (1959) comme la capacité d'une machine à apprendre sans programmation explicite, constitue une réponse aux défis posés par l'analyse de données complexes et volumineuses. Face à la difficulté d'extraire manuellement du sens ou d'identifier des corrélations significatives dans de vastes ensembles de données, l'apprentissage automatique permet aux systèmes d'analyser les informations et d'en dégager des relations utiles en s'appuyant sur l'expérience acquise lors de la phase d'entraînement [21].

Comme évoqué précédemment, l'apprentissage supervisé constitue l'une des trois branches principales du *machine learning*, aux côtés de l'apprentissage non supervisé et de l'apprentissage par renforcement. Au sein de l'apprentissage supervisé, une distinction fondamentale est établie entre les problèmes de classification, visant à prédire des catégories discrètes, et les problèmes de régression, portant sur la prédiction de valeurs continues comme l'illustre la Figure 3.2.

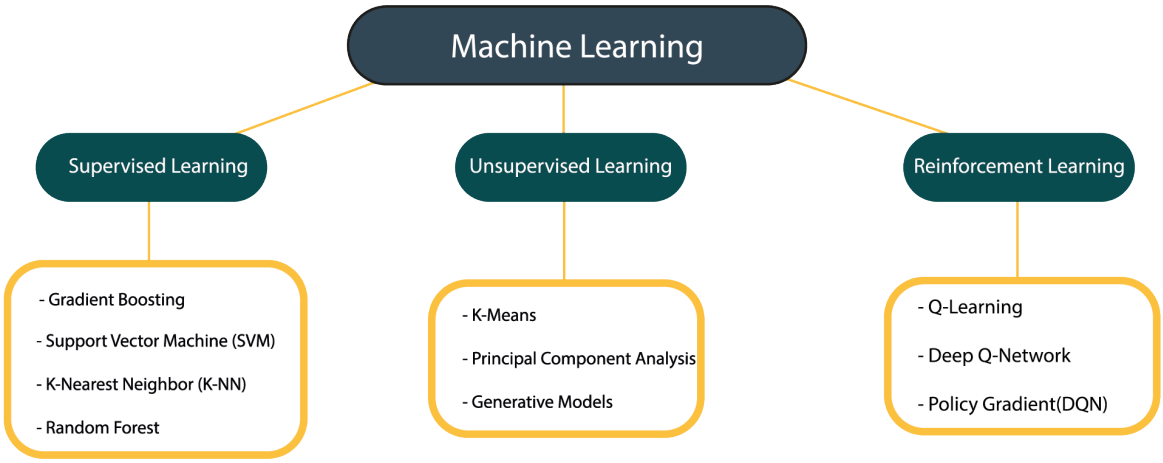


FIGURE 3.2 – Classification des types d'apprentissage automatique (*machine learning*) [6].

## 3.3 Algorithmes d'apprentissage supervisé

### 3.3.1 Principe de la régression

Le principe de la régression consiste à estimer une relation entre une variable dépendante $y$ et une ou plusieurs variables explicatives $x$ [7]. Pour illustrer ce processus, considérons le cas le plus simple : la régression linéaire à une seule variable.

**Étape 1 : Modélisation linéaire**

Le modèle s'écrit :

$$f(x) = wx + b \tag{3.1}$$

où $w$ représente la pente et $b$ l'ordonnée à l'origine. La figure 3.3 illustre un modèle linéaire ajusté à des points expérimentaux. Les lignes rouges représentent les *résidus*, c'est-à-dire les écarts entre les valeurs réelles (points noirs) et les prédictions du modèle (ligne jaune).

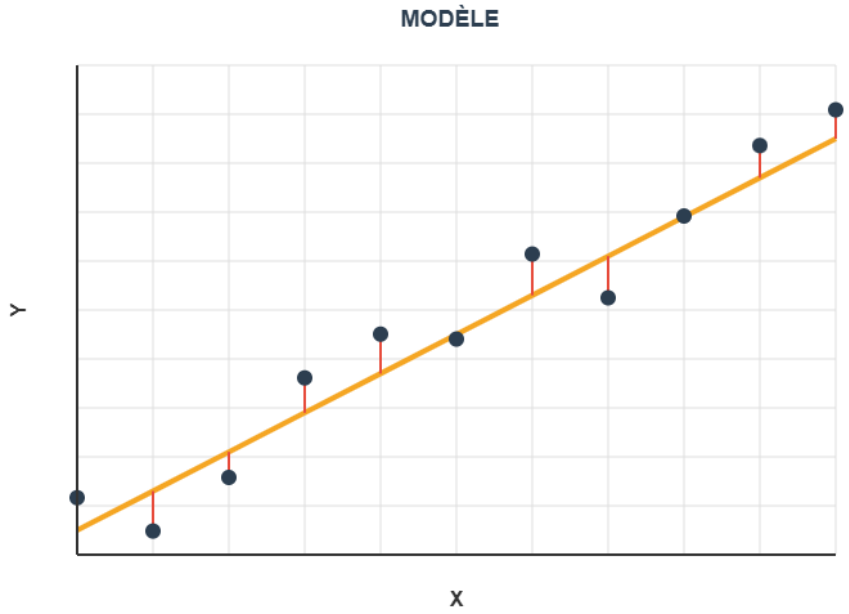


FIGURE 3.3 – Modèle de régression linéaire $f(x) = wx + b$ [7].

**Étape 2 : Définition de la fonction de coût**

Pour mesurer la qualité de l'ajustement, on définit une fonction de coût :

$$\mathcal{L}(b) = \frac{1}{2m} \sum_{i=1}^{m} (y^{(i)} - f(x^{(i)}))^2 \tag{3.2}$$

où $m$ est le nombre d'observations. Cette fonction est minimale lorsque le modèle reproduit le mieux les données expérimentales.

La figure 3.4 illustre l'évolution de la fonction de coût en fonction du paramètre $b$. On observe une courbe parabolique dont le minimum correspond à la valeur optimale $b^*$ minimisant l'erreur du modèle.

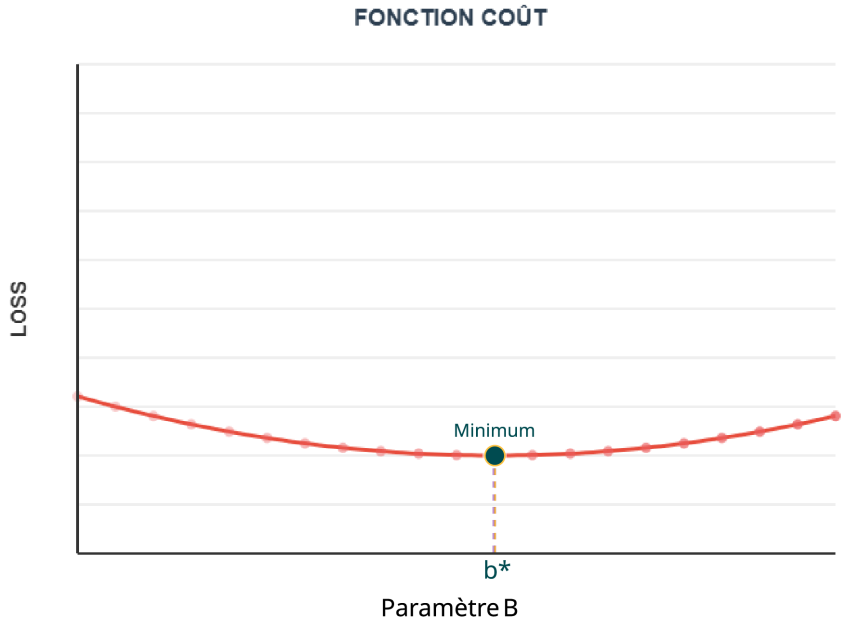


FIGURE 3.4 – Fonction de coût $\mathcal{L}(b)$ en fonction du paramètre $b$ [7].

**Étape 3 : Descente de gradient**

L'algorithme de descente de gradient permet de trouver automatiquement la valeur optimale du paramètre ($b$) minimisant la fonction de coût. La règle de mise à jour du paramètre est donnée par :

$$b_{t+1} = b_t - \eta \frac{\partial \mathcal{L}}{\partial b_t} \tag{3.3}$$

où $\eta$ est le taux d'apprentissage.

**Étape 4 : Processus itératif**

L'algorithme se déroule selon les étapes suivantes :

a. **Initialisation** des paramètres ($b_0$, éventuellement $w_0$) ;
b. **Calcul du gradient** de la fonction de coût ;
c. **Mise à jour** des paramètres dans la direction opposée au gradient ;
d. **Convergence** vers le minimum de la fonction de coût.

La figure 3.5 illustre ce processus de convergence vers le minimum global.

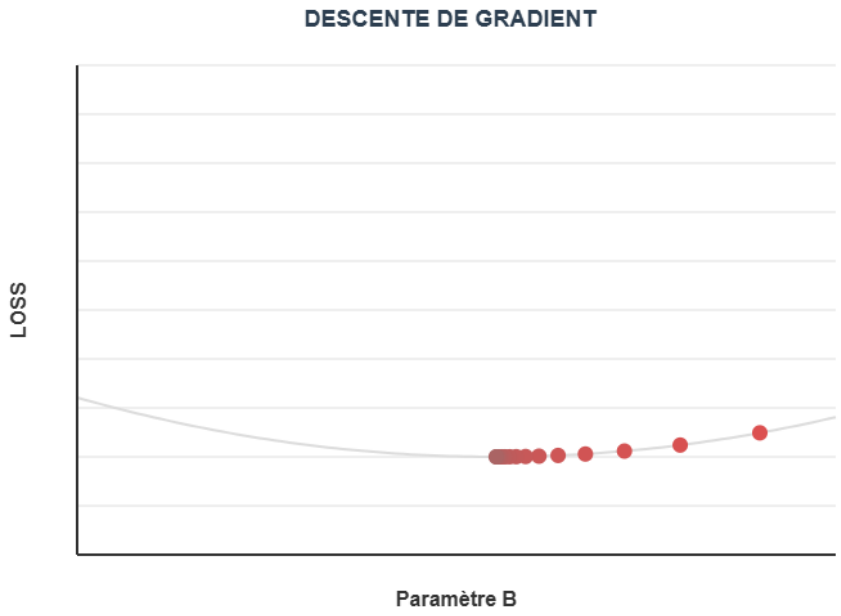


FIGURE 3.5 – Trajectoire de la descente de gradient sur la fonction de coût [7].

### 3.3.2 Principes de la classification

#### Définition et objectifs

La classification est une technique d'apprentissage automatique utilisée pour prédire l'appartenance d'instances de données à des catégories spécifiques. Elle se distingue de la régression par la nature de la variable cible : alors que la régression produit des valeurs continues, la classification génère des étiquettes de classe discrètes [22].

La classification constitue l'une des tâches les plus étudiées en apprentissage automatique et en fouille de données, en raison de son rôle clé dans la prise de décision et l'extraction de

connaissances. Parmi les différentes approches d'apprentissage automatique, elle demeure l'une des plus utilisées dans les applications industrielles et scientifiques [23].

**Principes fondamentaux**

L'apprentissage supervisé, dont la classification fait partie, vise à établir une relation entre des variables d'entrée (indépendantes) et une variable cible (dépendante) [24]. Le processus de classification comprend généralement deux étapes principales :

- **Phase d'apprentissage** : construction du modèle à partir de données étiquetées ;
- **Phase de prédiction** : utilisation du modèle pour classer de nouvelles observations.

**Principaux modèles de classification**

Les méthodes de classification supervisée se déclinent en plusieurs familles. Les modèles les plus courants sont présentés ci-dessous :

A. **Arbres de décision (ID3 et C4.5)**

Les arbres de décision figurent parmi les algorithmes les plus populaires en classification. Ils offrent une représentation intuitive du processus décisionnel et facilitent la compréhension du modèle. L'algorithme ID3 (Iterative Dichotomiser 3), introduit en 1986, repose sur le concept de gain d'information. Il est simple à interpréter mais souffre de certaines limites : absence de retour arrière, incapacité à gérer les valeurs manquantes et manque d'optimisation globale [25].

L'algorithme C4.5 constitue une amélioration d'ID3, notamment par son mécanisme d'élagage qui supprime les branches peu fiables. Il gère les données manquantes, accepte des attributs continus et discrets, et permet un élagage pré- et post-traitement. Ses limites concernent principalement les petits jeux de données et le coût de calcul élevé [25].

B. **Réseaux bayésiens**

Un réseau bayésien (BN) est un modèle probabiliste graphique représentant les relations de dépendance entre un ensemble de variables. Il est structuré sous forme d'un graphe orienté acyclique (DAG), où les nœuds représentent les variables et les arcs traduisent leurs dépendances conditionnelles.

Les réseaux bayésiens présentent plusieurs avantages : robustesse face aux variations mineures des données, flexibilité d'utilisation (classification et régression), et capacité à traiter les valeurs manquantes en intégrant leurs distributions probables. Le principal inconvénient réside dans la nécessité de discrétiser les variables continues, ce qui peut introduire du bruit et de la sensibilité aux variations [26].

C. **K-plus proches voisins (KNN)**

L'algorithme des K plus proches voisins (KNN) classe une observation en fonction de la majorité de ses voisins les plus proches, déterminés par une distance métrique . On distingue deux variantes : le KNN structuré (tenant compte de la distribution des données) et le KNN non structuré (fonctionnant directement sur les points observés) [27].

Ses principaux avantages sont sa simplicité, sa robustesse au bruit et son efficacité sur de

grands ensembles de données. En revanche, il est coûteux en calculs, exigeant en mémoire et sensible à la présence d'attributs non pertinents [27].

D. **Machines à vecteurs de support (SVM)**

Les Machines à Vecteurs de Support (SVM), proposées par Vapnik, sont fondées sur la théorie de l'apprentissage statistique. Elles reposent sur la recherche d'un hyperplan séparant les données de manière optimale dans un espace de grande dimension. Le classificateur à marge maximale vise à maximiser la distance entre les classes pour une meilleure généralisation.

Les SVM sont puissantes pour les problèmes complexes, notamment ceux présentant des frontières non linéaires ou de haute dimension. Leur principal inconvénient réside dans la nécessité d'un réglage précis des paramètres (noyau, pénalité, etc.) pour obtenir des performances optimales [28].

#### 3.3.2.1 Synthèse comparative

Chaque méthode présente des avantages et des limites. Le choix du modèle dépend du contexte d'application, des caractéristiques du jeu de données et des objectifs de l'étude. Les arbres de décision sont très interprétables, les réseaux bayésiens gèrent bien l'incertitude, KNN est simple mais coûteux en calcul, et les SVM offrent d'excellents résultats sur les problèmes complexes [22].

### 3.3.3 Principe du gradient boosting

Dans le cadre de cette étude, trois algorithmes de *gradient boosting* de référence ont été sélectionnés et comparés pour la classification lithologique : XGBoost (*Extreme Gradient Boosting*), CatBoost (*Categorical Boosting*) et LightGBM (*Light Gradient Boosting Machine*). Ces algorithmes, bien que partageant un fondement théorique commun, se distinguent par leurs stratégies d'optimisation, leur gestion des variables catégorielles et leur efficacité computationnelle.

Le *gradient boosting* constitue une famille d'algorithmes d'apprentissage automatique puissants, conçus pour traiter des données tabulaires structurées composées d'un ensemble de variables explicatives (X) et d'une variable cible (y). Comme pour tout algorithme d'apprentissage supervisé, l'objectif est d'apprendre suffisamment à partir des données d'entraînement pour généraliser efficacement sur des données non vues.

Le principe fondamental du *gradient boosting* repose sur la construction séquentielle d'un ensemble de modèles faibles (généralement des arbres de décision peu profonds) qui sont combinés pour former un modèle fort. À chaque itération, un nouveau modèle est ajouté pour corriger les erreurs résiduelles des prédictions précédentes, en minimisant une fonction de perte par descente de gradient.

Pour illustrer le processus sous-jacent du *gradient boosting* dans le cadre de cette étude, considérons un échantillon représentatif des données de diagraphie utilisées. Le tableau 3.1 présente un extrait du jeu de données, où les variables explicatives (X) correspondent aux

paramètres physiques enregistrés par les outils de diagraphie, et la variable cible (y) représente le volume d'argile (*VCL*), l'une des propriétés pétrophysiques à prédire.

TABLE 3.1 – Extrait d'un jeu de données utilisé pour l'entraînement des modèles de *gradient boosting*.

| **Profondeur (m)** | **Gamma Ray (API)** | **Sonic (µs/ft)** | **... (v/v)** | **VCL** |
|---|---|---|---|---|
| 3766.718 | 113.19 | 77.14 | ... | 0.50525 |
| 3766.871 | 122.46 | 77.43 | ... | 0.51170 |
| 3767.023 | 127.34 | 77.79 | ... | 0.51781 |
| 3767.176 | 131.05 | 77.86 | ... | 0.53236 |
| 3767.328 | 128.55 | 77.08 | ... | 0.57077 |
| 3767.480 | 119.81 | 76.24 | ... | 0.63711 |

### 3.3.4 Processus d'entraînement du gradient boosting

**Étape 1 : Prédiction initiale par la moyenne**

Le processus du gradient boosting débute par une prédiction initiale basée sur la valeur moyenne de la variable cible. Cette prédiction constitue le point de départ du modèle itératif [29]. Pour notre dataset pétrophysique, la prédiction initiale est la moyenne des valeurs de VCL (volume d'argile) observées sur l'ensemble des échantillons d'entraînement.

Mathématiquement, cette prédiction initiale $F_0(x)$ est définie comme :

$$F_0(x) = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.4}$$

où $y_i$ représente la valeur cible (VCL) pour l'échantillon $i$ et $n$ est le nombre total d'échantillons.

Le choix de la moyenne comme prédiction initiale est justifié théoriquement : lorsque l'on dérive la fonction de perte (loss function) par rapport à chaque valeur observée et que l'on somme les dérivées, on obtient précisément la moyenne de la variable cible. Cette propriété garantit que la prédiction initiale minimise la fonction de perte au premier ordre [29].

**Étape 2 : Calcul des pseudo-résidus**

À chaque itération $m$, après avoir généré une prédiction $F_{m-1}(x)$, on calcule les pseudo-résidus représentant l'écart entre les valeurs réelles observées et les prédictions actuelles :

$$r_i^{(m)} = y_i - F_{m-1}(x_i) \tag{3.5}$$

Ces pseudo-résidus contiennent l'information sur les erreurs non encore corrigées par le

modèle. Ils quantifient précisément ce que le modèle actuel ne capture pas encore. Contrairement à la régression linéaire classique où l'on parle simplement de résidus, on utilise le terme pseudo-résidus en gradient boosting pour souligner leur rôle dans l'optimisation itérative. L'objectif des itérations suivantes est de réduire progressivement ces pseudo-résidus en les minimisant [29].

**Exemple appliqué au dataset pétrophysique :**

Supposons que la prédiction initiale soit la moyenne du VCL : $F_0(x) = 0.549$ (calculée sur les 21 724 échantillons). Les pseudo-résidus pour les premiers échantillons sont calculés comme suit :

TABLE 3.2 – Calcul des pseudo-résidus pour le dataset pétrophysique.

| **Profondeur (m)** | **Gamma Ray (API)** | **Sonic (µs/ft)** | **VCL Réel (v/v)** | **Pseudo-résidus (v/v)** |
|---|---|---|---|---|
| 3766.718 | 113.19 | 77.14 | 0.50525 | -0.04375 |
| 3766.871 | 122.46 | 77.43 | 0.51170 | -0.03730 |
| 3767.023 | 127.34 | 77.79 | 0.51781 | -0.03119 |
| 3767.176 | 131.05 | 77.86 | 0.53236 | -0.01664 |
| 3767.328 | 128.55 | 77.08 | 0.57077 | 0.02177 |
| 3767.480 | 119.81 | 76.24 | 0.63711 | 0.08811 |

Ces pseudo-résidus indiquent les écarts que le modèle doit apprendre à corriger. Les valeurs négatives indiquent que la prédiction est supérieure à la valeur réelle, tandis que les valeurs positives indiquent l'inverse.

**Étape 3 : Entraînement d'un apprenant faible**

À chaque itération, on entraîne un nouvel apprenant faible (généralement un arbre de décision peu profond) pour prédire les résidus calculés à l'étape précédente. Cet arbre, noté $h_m(x)$, apprend à modéliser les patterns d'erreur que le modèle précédent n'a pas captés [29].

L'entraînement s'effectue sur les données d'entrée $(x_i, r_i^{(m)})$, où les résidus jouent le rôle de nouvelle variable cible. L'arbre fragmenté légèrement le domaine d'entrée et produit des prédictions constantes dans chaque région.

**Étape 4 : Mise à jour itérative du modèle**

Le modèle global est mis à jour en ajoutant la prédiction de l'apprenant faible, pondérée par un taux d'apprentissage $\eta$ (learning rate), généralement compris entre 0 et 1 :

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{3.6}$$

Le taux d'apprentissage contrôle la contribution de chaque nouvel apprenant. Un taux faible permet une convergence plus progressive et robuste, tandis qu'un taux élevé accélère l'entraînement mais risque d'introduire de l'instabilité. Cette mise à jour incrémentale est au cœur du mécanisme d'amélioration graduelle du gradient boosting [29].

**Étape 5 : Itérations successives**

Le processus se répète itérativement : à chaque nouvelle itération, on recalcule les résidus en fonction du modèle mis à jour, on entraîne un nouvel apprenant faible sur ces résidus, et on enrichit le modèle global. Cette boucle continue jusqu'à atteindre un critère d'arrêt, tel qu'un nombre maximal d'itérations ou une stagnation de la performance.

À chaque itération $m$, le modèle final après $M$ itérations est :

$$F_M(x) = F_0(x) + \eta \sum_{m=1}^{M} h_m(x) \tag{3.7}$$

Cette construction cumulative permet au gradient boosting d'améliorer progressivement ses prédictions en se concentrant sur les erreurs les plus persistantes.

**Étape 6 : Validation et régularisation**

Parallèlement à l'entraînement, on valide les performances du modèle sur un ensemble de validation distinct. Des techniques de régularisation (limitation de la profondeur des arbres, réduction du taux d'apprentissage, et contrôle du nombre d'itérations) sont appliquées pour éviter le surapprentissage et assurer une bonne généralisation sur des données non vues.

**Principaux hyperparamètres :**

- **Learning rate (taux d'apprentissage)** : Contrôle la contribution de chaque apprenant faible en ajustant le facteur de réduction. Des valeurs plus faibles réduisent l'influence de chaque arbre faible, nécessitant plus d'arbres à construire mais produisant un modèle plus robuste et moins sujet au surapprentissage.
- **Nombre d'arbres (n_estimators)** : Contrôle le nombre d'itérations de boosting. Plus d'arbres renforcent l'ensemble et augmentent les performances, mais risquent d'accroître le surapprentissage. Une combinaison de taux d'apprentissage faible et d'early stopping atténue ce risque.
- **Max Depth (profondeur maximale)** : Limite le nombre de niveaux dans chaque arbre de décision. Une profondeur proche de 3 prévient le surapprentissage, avec un maximum recommandé de 10 pour éviter une complexité excessive.
- **Minimum samples per leaf** : Définit le nombre minimal d'échantillons dans les nœuds terminaux. Une valeur plus élevée prévient le surapprentissage en empêchant les arbres de créer des divisions basées sur trop peu de données.
- **Subsampling rate** : Contrôle la proportion des données utilisées pour entraîner chaque arbre. Un taux inférieur à 1 (par exemple 0.7) accélère l'entraînement mais peut augmenter le risque de surapprentissage.
- **Feature Sampling Rate** : Similaire au subsampling, mais appliqué aux caractéristiques. Pour les datasets avec de nombreuses variables, un taux entre 0.5 et 1 est recommandé pour réduire le surapprentissage.

L'efficacité du gradient boosting repose largement sur l'équilibre entre la puissance du mo-

dèle et sa régularisation. Les hyperparamètres présentés permettent de contrôler cette balance pour éviter le surapprentissage tout en maximisant les performances.

#### 3.3.4.1 Implémentation dans le projet

Le gradient boosting est bien établi en Python via plusieurs bibliothèques principales :

- **XGBoost** : eXtreme Gradient Boosting - optimisé pour la vitesse et la performance
- **LightGBM** : Light Gradient Boosting Machine - efficace pour les grands ensembles de données
- **CatBoost** : Categorical Boosting - spécialisé dans la gestion des variables catégoriques

## 3.4 Critères d'évaluation

L'évaluation des performances des modèles de régression et de classification repose sur l'utilisation de critères qui permettent de quantifier l'écart entre les prédictions du modèle et les valeurs observées. Ces métriques jouent un rôle fondamental dans la sélection, la comparaison et l'optimisation des modèles. Leur choix dépend des objectifs de l'analyse, de la nature des données et de la sensibilité souhaitée aux erreurs extrêmes [30]. Une compréhension approfondie de ces critères est donc essentielle pour garantir une évaluation rigoureuse et pertinente des modèles prédictifs.

### 3.4.1 Critères pour la régression

- **Erreur quadratique moyenne (MSE)** : Mesure la moyenne des carrés des écarts entre les valeurs prédites et les valeurs réelles. Elle est très sensible aux grandes erreurs, ce qui en fait une métrique utile lorsque les déviations importantes doivent être pénalisées [30].

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.8}$$

- **Erreur absolue moyenne (MAE)** : Représente la moyenne des écarts absolus entre les valeurs prédites et les valeurs observées. Moins sensible que la MSE aux valeurs extrêmes, elle fournit une évaluation plus robuste des erreurs de prédiction [30].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.9}$$

- **Racine de l'erreur quadratique moyenne (RMSE)** : Racine carrée de la MSE. Elle a l'avantage d'être exprimée dans la même unité que la variable cible, facilitant l'interprétation des résultats [30].

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{3.10}$$

- **Coefficient de détermination ($R^2$)** : Mesure la proportion de la variance totale des données expliquée par le modèle. Il s'agit d'un indicateur global et informatif de la qualité de l'ajustement [30].

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3.11}$$

- **Erreur absolue en pourcentage symétrique (SMAPE)** : Métrique en pourcentage qui prend en compte la symétrie des erreurs relatives, évitant notamment la division par zéro [30].

$$\text{SMAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \tag{3.12}$$

- **Biais en pourcentage (PBIAS)** : Mesure la tendance systématique d'un modèle à surestimer ou sous-estimer les valeurs. Il quantifie le biais moyen en pourcentage, permettant d'identifier une tendance systématique.

$$\text{PBIAS} = \frac{100}{n} \sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i} \tag{3.13}$$

### 3.4.2 Critères pour la classification

- **Précision (precision)** : Proportion des prédictions positives correctes parmi toutes les prédictions positives. Elle mesure la fiabilité des prédictions positives du modèle [30].
- **Rappel (recall)** : Proportion des cas positifs correctement identifiés parmi tous les cas positifs réels. Elle mesure la capacité du modèle à détecter tous les cas positifs [30].
- **F1-Score** : Moyenne harmonique entre la précision et le rappel. Elle synthétise la balance entre ces deux métriques et offre une mesure équilibrée de performance [30].
- **Accuracy (exactitude)** : Proportion globale des prédictions correctes parmi toutes les prédictions. Elle fournit une mesure générale de la performance du modèle de classification [30].

Ces métriques permettent une évaluation équilibrée des performances en classification binaire ou multi-classe, particulièrement utile pour les lithologies minoritaires présentant des données déséquilibrées.

### 3.4.3 Intervalles d'appréciation

Les métriques d'évaluation utilisées dans cette étude sont interprétées selon les seuils de performance présentés ci-dessous. Ces intervalles permettent une interprétation objective et comparative des résultats des modèles.

#### 3.4.3.1 Régression

TABLE 3.3 – Intervalles d'appréciation des critères de validation pour les modèles de régression (variables normalisées 0-1).

| **Paramètre** | **Très bon** | **Bon** | **Satisfaisant** | **Insatisfaisant** |
|---|---|---|---|---|
| MSE | 0-0.005 | 0.005-0.02 | 0.02-0.05 | $> 0.05$ |
| MAE | 0-0.05 | 0.05-0.10 | 0.10-0.15 | $> 0.15$ |
| $R^2$ | $> 0.95$ | 0.90-0.95 | 0.85-0.90 | $< 0.85$ |
| RMSE | 0-0.07 | 0.07-0.14 | 0.14-0.22 | $> 0.22$ |
| SMAPE | 0%-5% | 5%-15% | 15%-25% | $> 25\%$ |
| PBIAS | -2%-2% | -5%-5% | -10%-10% | $> 10\%$ |

#### 3.4.3.2 Classification

TABLE 3.4 – Intervalles d'appréciation des critères de validation pour les modèles de classification.

| **Métrique** | **Très bon** | **Bon** | **Satisfaisant** | **Insatisfaisant** |
|---|---|---|---|---|
| Précision | $> 0.95$ | 0.90-0.95 | 0.85-0.90 | $< 0.85$ |
| Rappel | $> 0.95$ | 0.90-0.95 | 0.85-0.90 | $< 0.85$ |
| F1-Score | $> 0.95$ | 0.90-0.95 | 0.85-0.90 | $< 0.85$ |
| Accuracy | $> 0.95$ | 0.90-0.95 | 0.85-0.90 | $< 0.85$ |

## 3.5 Conclusion

Ce chapitre a présenté les bases de l'apprentissage automatique appliqué à la prédiction pétrophysique. L'apprentissage supervisé, à travers la régression et la classification, permet respectivement d'estimer des valeurs continues et d'identifier des classes lithologiques. Parmi les approches testées, le *gradient boosting* – notamment ses implémentations *XGBoost*, *LightGBM* et *CatBoost* – s'est révélé particulièrement adapté aux données tabulaires. L'évaluation des modèles repose sur des métriques spécialisées et un réglage précis des hyperparamètres afin d'assurer un bon compromis entre performance et généralisation. Enfin, la stratégie hybride adoptée, combinant régression et classification, offre une modélisation robuste des propriétés pétrophysiques du bassin de Touggourt.

# Chapitre 4

# Présentation des données et stratégie de modélisation

## 4.1 Introduction

Ce chapitre est consacré à la présentation du jeu de données et à l'exploration de ses caractéristiques structurelles et distributives.

Nous commencerons par décrire les données provenant des interprétations pétrophysiques réalisées par Sonatrach Exploration, le format des fichiers d'entrée et de sortie, ainsi que l'organisation générale du dataset. Par la suite, nous analyserons la distribution des lithologies, révélant un déséquilibre caractéristique des réservoirs argileux-gréseux de la région de Touggourt. Cette exploration justifiera l'adoption d'une stratégie hybride combinant classification et régression pour améliorer la prédiction des lithologies minoritaires.

L'objectif global de ce chapitre est de fournir une compréhension complète de la base de données avant de procéder à l'entraînement des modèles.

## 4.2 Présentation du jeu de données

### 4.2.1 Source et nature des données

Le jeu de données utilisé dans cette étude provient d'interprétations pétrophysiques réalisées par les log analystes de Sonatrach Exploration sur des puits réels du bassin de Touggourt. Ces interprétations ont été effectuées à l'aide du logiciel Techlog (Schlumberger), et plus particulièrement du module Quanti-ELAN, qui permet de générer les paramètres pétrophysiques et d'identifier la lithologie présente dans chaque intervalle.

Le processus d'interprétation repose sur la configuration des paramètres physiques représentatifs de chaque unité lithostratigraphique (ou zone). Il débute par la pondération des différentes mesures diagraphiques en fonction de leur fiabilité pour chaque zone. À titre d'exemple, une pondération de 0,65 est généralement attribuée à la résistivité profonde (*deep resistivity*), tandis que la résistivité superficielle (*shallow resistivity*) reçoit une pondération de 0,5. Ces coefficients

de pondération sont issus de l'expérience opérationnelle accumulée par les log analystes et sont ajustés selon les caractéristiques lithologiques propres à chaque intervalle.

Dans une deuxième étape, les lithologies présentes sont identifiées à travers les minéraux activés dans le module *Quanti Elan*, en fonction des signatures diagraphiques observées dans chaque zone. Par exemple, l'anhydrite présente typiquement une densité de 2,98 g/cm$^3$, une lecture sonique de 50 µs/ft et une porosité neutron de 0,02 m$^3$/m$^3$. Le tableau 4.1 présente les réponses diagraphiques théoriques des principaux minéraux sédimentaires.

TABLE 4.1 – Réponses diagraphiques des principaux minéraux sédimentaires [2].

| **Minéral** | **Formule** | **RHOB** (g/cm³) | **NPHI** (v/v) | **DT** (µs/ft) | **PE** (barns/e) | **U** (ppm) | **THOR** (ppm) | **POTA** (%) |
|---|---|---|---|---|---|---|---|---|
| **Fluides** | | | | | | | | |
| Eau | $H_2O$ | 1.0–1.05 | 1.00 | 189 | – | – | – | – |
| Pétrole | $(C_nH_{2n})_x$ | 0.7–0.9 | 0.95 | 215 | – | – | – | – |
| Condensat | $(C_nH_{2n})_4$ | 0.5–0.7 | 0.90 | 222 | – | – | – | – |
| Gaz | $(C_nH_{2n})_2$ | 0.15–0.4 | 0.50 | – | – | – | – | – |
| **Silicates et Feldspaths** | | | | | | | | |
| Quartz | $SiO_2$ | 2.64 | –0.02 | 56 | 1.8 | 4.8 | – | – |
| Orthoclase | $KAlSi_3O_8$ | 2.52 | –0.03 | 69 | 2.9 | 7.2 | – | – |
| Albite | $NaAlSi_3O_8$ | 2.59 | –0.02 | 49 | 1.7 | 4.4 | – | – |
| **Carbonates** | | | | | | | | |
| Calcite | $CaCO_3$ | 2.71 | 0.00 | 49 | 5.1 | 13.8 | – | – |
| Dolomite | $CaMg(CO_3)_2$ | 2.85 | 0.01 | 44 | 3.1 | 9.0 | – | – |
| Ankerite | $Ca(Mg,Fe)(CO_3)_2$ | 2.86 | 0.01 | – | 9.3 | 27 | – | – |
| Sidérite | $FeCO_3$ | 3.89 | 0.12 | 47 | 15 | 57 | – | – |
| **Argiles** | | | | | | | | |
| Kaolinite | $Al_2Si_2O_5(OH)_4$ | 2.41 | 0.37 | – | 1.8 | 4.4 | 80–130 | 0.2 |
| Chlorite | $(Mg,Fe)_5Al_2Si_3O_{10}(OH)_8$ | 2.76 | 0.52 | – | 6.3 | 17 | 180–250 | 0.2 |
| Illite | $K_{1.5}Al_4(Si_{6.5}Al_{1.5})O_{20}(OH)_4$ | 2.52 | 0.30 | – | 3.5 | 8.7 | 250–300 | 6 |
| Montmorillonite | $(Ca,Na)_{0.3}(Al,Mg)_2Si_4O_{10}(OH)_2$ | 2.12 | 0.60 | – | 2.0 | 4.0 | 150–200 | 2 |
| **Évaporites** | | | | | | | | |
| Halite | $NaCl$ | 2.04 | 0.03 | 67 | 4.7 | 9.5 | 5 | 0.5 |
| Anhydrite | $CaSO_4$ | 2.98 | 0.02 | 50 | 5.1 | 15 | 5 | 0.5 |
| Gypsum | $CaSO_4{\cdot}(H_2O)_2$ | 2.35 | 0.60 | 52.5 | 4.0 | 9.4 | 5 | 7.8 |

### 4.2.2 Constitution du dataset

Le dataset a été constitué progressivement à partir de 10 puits situés dans la région de Touggourt (voir Chapitre 1, Section 1.3). La Figure 4.1 présente la distribution spatiale de ces puits, identifiés par les lettres A à J, ainsi que les distances inter-puits.

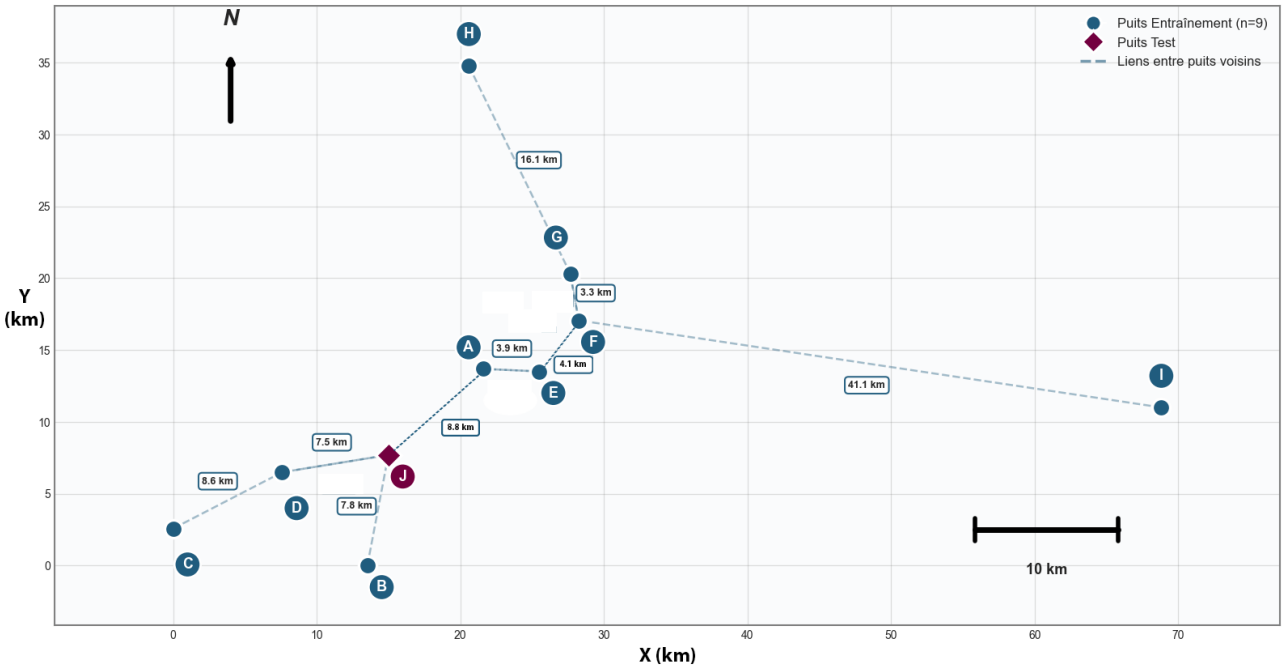


FIGURE 4.1 – Carte de distribution spatiale des puits utilisés dans l'étude.

Après concaténation verticale des données des 9 puits d'entraînement, le dataset complet présente les dimensions suivantes :

- **9 puits d'entraînement et de validation** (A, B, C, D, E, F, G, H, I)
- **1 puits de test indépendant** (J), utilisé comme puits historique pour la comparaison et la validation externe
- **21 724 échantillons** (lignes de mesures), correspondant à une profondeur cumulée d'environ 3 200 m
- **Nombre de variables** : 25 colonnes

#### 4.2.2.1 Structure détaillée des variables

Le tableau 4.2 présente la structure complète du dataset, incluant les variables de référence, les diagraphies utilisées comme features pour l'entraînement des modèles, ainsi que les paramètres pétrophysiques et lithologiques déterminés par l'interprétation Quanti-ELAN utilisés comme variables cibles.

TABLE 4.2 – Structure complète des variables du dataset

| Catégorie | Variable | Symbole | Unité | Nombre |
|---|---|---|---|---|
| ***Variables de référence*** | | | | **2** |
| | Nom du puits | Well-Name | – | |
| | Profondeur | DEPTH | m | |
| ***Variables diagraphiques (Features)*** | | | | **15** |
| | Gamma Ray | GR | gAPI | |
| | Gamma Ray corrigé | KTH | gAPI | |
| | Porosité Neutron | CNC | v/v | |
| | Sonic | DTCQI | µs/ft | |
| | Potassium | K | % | |
| | Uranium | U | ppm | |
| | Thorium | TH | ppm | |
| | Photo-électrique | PE | b/elec | |
| | Caliper | CALX | inches | |
| | Résistivité (10 inches) | M2R1 | Ω·m | |
| | Résistivité (30 inches) | M2R2 | Ω·m | |
| | Résistivité (50 inches) | M2R3 | Ω·m | |
| | Résistivité (70 inches) | M2R4 | Ω·m | |
| | Résistivité (90 inches) | M2R5 | Ω·m | |
| | Densité | ZDEN | $g/cm^3$ | |
| ***Variables cibles - Paramètres pétrophysiques*** | | | | **3** |
| | Volume d'argile | VCL | v/v | |
| | Saturation en eau | SW | v/v | |
| | Porosité effective | PIGE | v/v | |
| ***Variables cibles - Composition lithologique*** | | | | **6** |
| | Quartz | QUARTZ | v/v | |
| | Roches ignées | IGNEOUS | v/v | |
| | Anhydrite | ANHYDRITE | v/v | |
| | Halite | HALITE | v/v | |
| | Dolomite | DOLOMITE | v/v | |
| | Calcite | CALCITE | v/v | |
| **Total des variables** | | | | **25** |

**Notes importantes :**

- **KTH** : Gamma Ray corrigé sans l'effet de l'Uranium
- **M2R1**–**M2R5** : Résistivités correspondant à différentes profondeurs d'investigation
- **Variables cibles** : Toutes normalisées entre 0 et 1, représentant des fractions volumétriques
- **CNC** : Peut aussi être exprimée en $m^3/m^3$ (équivalent à v/v)

Cette structure permet une validation croisée robuste tout en conservant un puits totalement indépendant pour évaluer la capacité de généralisation des modèles sur des données non vues durant l'entraînement.

## 4.3 Exploration et préparation des données

### 4.3.1 Distribution lithologique

L'analyse de la distribution des lithologies révèle un déséquilibre marqué, caractéristique des réservoirs argileux-gréseux de la région de Touggourt. La figure 4.2 et le tableau 4.3 présentent la fréquence d'occurrence de chaque lithologie dans le dataset.

TABLE 4.3 – Distribution des lithologies dans le dataset (21 724 échantillons)

| **Lithologie** | **Présence (%)** | **Nombre d'échantillons** |
|---|---|---|
| VCL | 99,08 | 21 525 |
| Quartz | 69,13 | 15 017 |
| Igneous | 33,44 | 7 265 |
| PIGE | 24,35 | 5 290 |
| Dolomite | 4,11 | 892 |
| Anhydrite | 2,84 | 618 |
| Calcite | 1,74 | 377 |
| Halite | 1,50 | 325 |

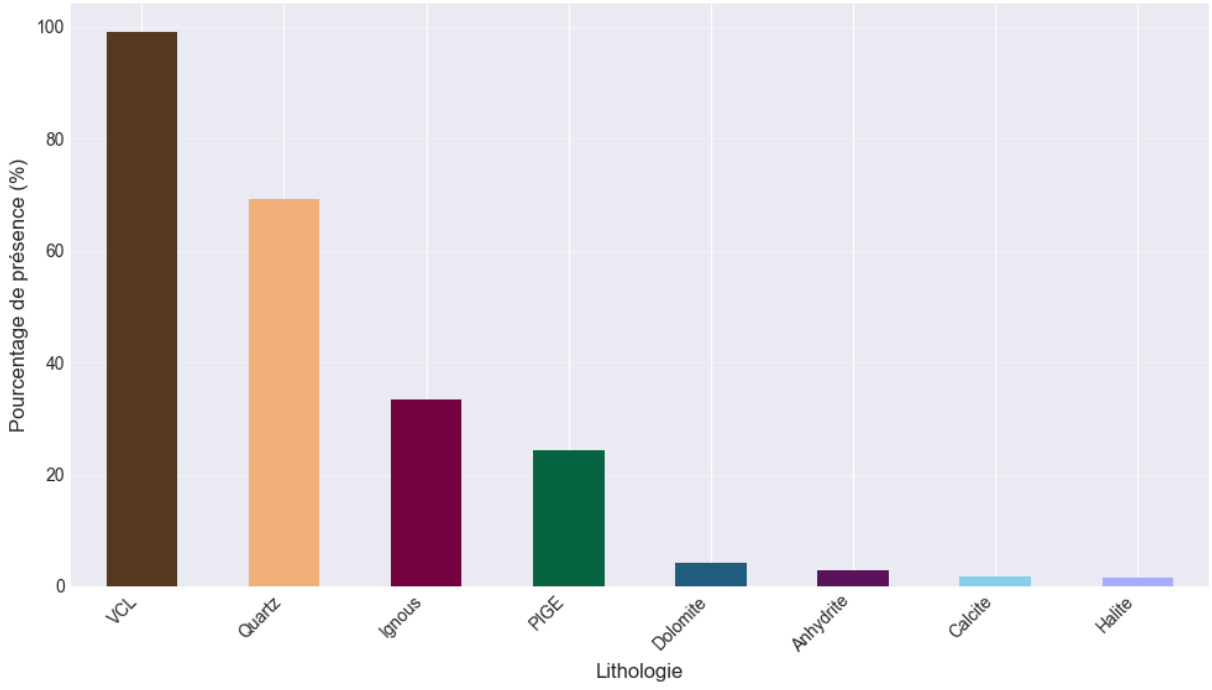


FIGURE 4.2 – Pourcentage de présence des différentes lithologies dans le dataset

### 4.3.2 Interprétation géologique du déséquilibre lithologique

Le fort déséquilibre observé dans la distribution lithologique s'explique par le contexte géologique spécifique de la région de Touggourt. Les variables cibles peuvent être regroupées en deux catégories selon leur distribution et leur interdépendance :

#### 4.3.2.1 Proportions continues majoritaires (VCL, Quartz, PIGE)

Les réservoirs ciblés dans cette étude correspondent principalement à des faciès argileux-gréseux. Ces trois composantes présentent une forte interdépendance caractéristique des systèmes argileux-gréseux :

A. **VCL (Volume d'argile)** : Présent dans 99,08 % des échantillons, constituant la matrice argileuse omniprésente
B. **Quartz** : Présent dans 69,13 % des échantillons, correspondant aux niveaux gréseux réservoirs
C. **PIGE (Porosité effective)** : Observée dans 24,35 % des échantillons, principalement associée aux grès

Cette interdépendance se manifeste par une relation inverse entre VCL et Quartz : lorsque le volume d'argile augmente, la proportion de quartz diminue automatiquement, entraînant une réduction ou une absence de porosité effective. Ce comportement reflète la nature du réservoir argileux-gréseux étudié.

#### 4.3.2.2 Proportions discontinues minoritaires : contexte stratigraphique

Les lithologies minoritaires se caractérisent par leur présence occasionnelle et localisée, déterminée par des conditions géologiques spécifiques.

A. **Roches ignées (Igneous) : héritage tectonique**

   Les roches éruptives (33,44 % de présence) sont observées localement dans certains puits, en particulier au sein de la Série Inférieure du Trias et de l'Ordovicien, plus précisément dans les Argiles d'El Gassi. Cet héritage géologique explique leur présence significative, bien que discontinue, dans le jeu de données.

B. **Carbonates et évaporites : contrôle stratigraphique**

   La faible représentation des évaporites (halite, anhydrite) et des carbonates (calcite, dolomite) résulte de la position stratigraphique des intervalles étudiés. Afin d'illustrer la correspondance entre les lithologies identifiées et leur contexte stratigraphique, la description suivante reprend, à titre indicatif, la succession triasique issue d'un rapport d'implantation de puits appartenant à la région d'étude [31] :

   - **Trias S4** : Sel massif blanc à translucide avec passées d'argile brun-rouge [31]. *Cet intervalle est généralement exclu de l'interprétation pétrophysique car il correspond à un dôme salifère stérile.*

- **Trias argileux G30** : Argile brun-rouge tendre à pâteuse salifère, fines passées de sel massif translucide [31].
- **Trias T2 + T1** : Alternance de dolomie blanche et grise cristalline et d'argile brun-rouge légèrement silteuse [31]. *Cet intervalle explique la présence ponctuelle de dolomite (4,11 %) et de calcite (1,74 %) dans le dataset.*
- **Trias Série Inférieure** : Alternance d'argile brun-rouge et verte, tendre à indurée, légèrement carbonatée, avec présence de roches éruptives gris sombre à gris brun (*Igneous*, 33,44 %), ainsi que de grès brun-rouge, fin à moyen, moyennement dur [31].

La série sédimentaire triasique de la région est caractérisée par une importante formation évaporitique s'étalant du Trias terminal au Dogger. Cependant, les réservoirs productifs se situent stratigraphiquement sous le dôme salifère S4, d'où l'absence quasi-totale de sel (halite : 1,50 %) dans les intervalles interprétés [3].

### 4.3.3 Stratégie adoptée pour la prédiction face au déséquilibre des données

Face au déséquilibre marqué observé dans la distribution des lithologies, une stratégie de prédiction hybride a été élaborée, combinant les approches de classification et de régression. Cette méthodologie vise à améliorer la précision globale en traitant différemment les lithologies majoritaires et minoritaires.

La première étape consiste à diviser les lithologies cibles en deux groupes principaux selon leur fréquence d'apparition :

- **Groupe 1 :** Proportions continues - **argile (VCL)**, **quartz** et **porosité effective (PIGE)**
- **Groupe 2 :** Proportions discontinues - Présentes de manière irrégulière et localisée - **Igneous**, **calcite**, **halite**, **anhydrite** et **dolomite**

A. **Traitement du groupe 1 : Approche par régression directe**

- Pour les proportions du groupe 1, une régression directe est appliquée
- Des modèles de boosting avancés sont utilisés : XGBoost, LightGBM et CatBoost
- La porosité effective (PIGE) fait exception, étant modélisée exclusivement avec XGBoost

B. **Traitement du groupe 2 : Approche hybride classification-régression**

Le groupe 2 est lui-même subdivisé en deux sous-groupes :

**Sous-groupe 2.1 : Igneous**

- Classification binaire pour identifier la présence ou l'absence d'Igneous
- Régression appliquée uniquement sur les profondeurs où la présence est détectée

**Sous-groupe 2.2 : Carbonates et évaporites**

- Création d'une classe composite « Autres Lithologies » regroupant calcite, halite, dolomite et anhydrite
- Classification binaire pour détecter la présence de cette classe composite
- Régression sur les valeurs positives détectées
- Classification multi-classes finale pour attribuer chaque prédiction à sa lithologie spécifique

L'architecture complète de cette stratégie est résumée dans la Figure 4.3 ci-dessous.

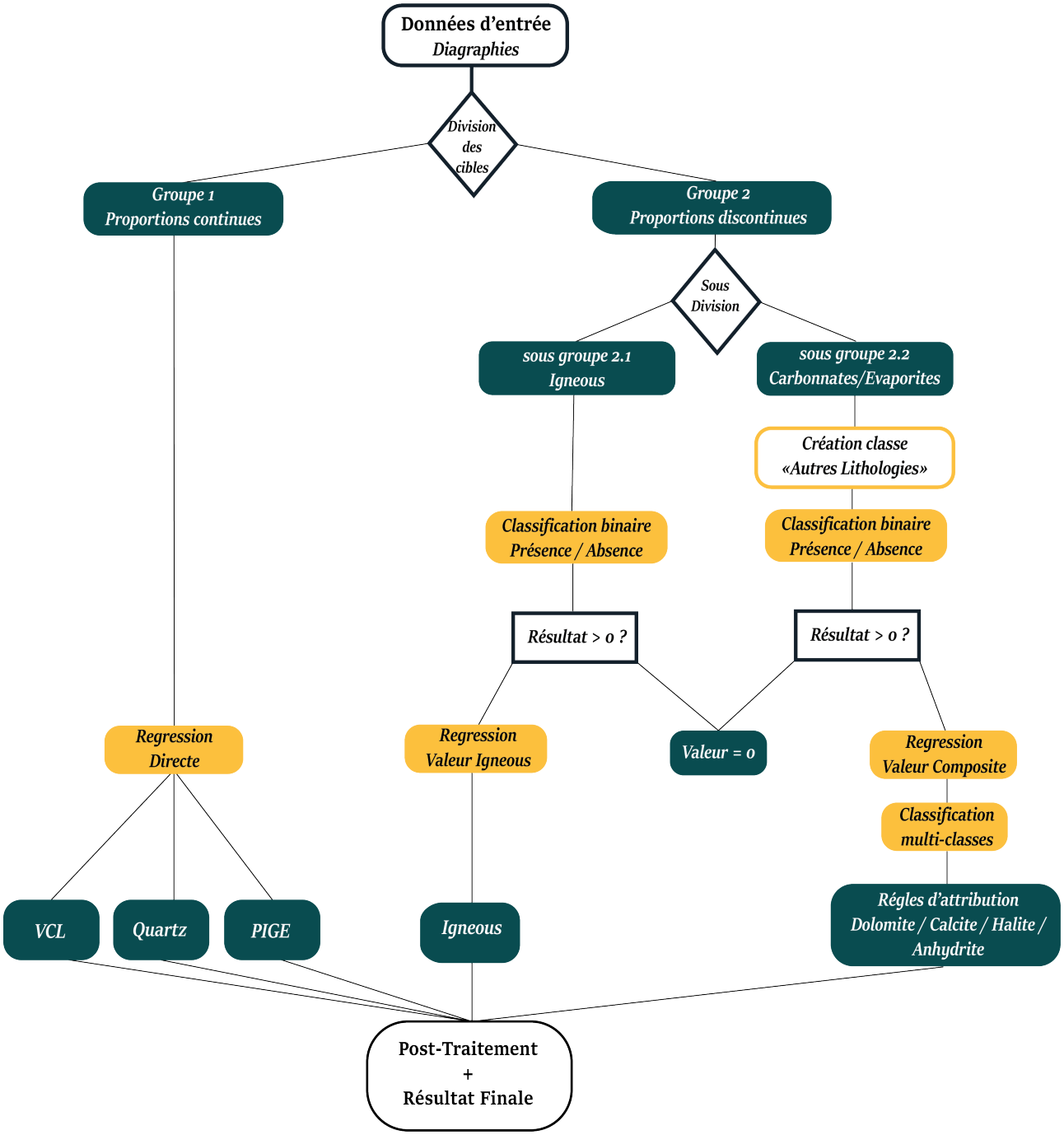


FIGURE 4.3 – Schéma architectural de la stratégie hybride de prédiction des lithologies

## 4.4 Conclusion

Ce chapitre a présenté la structure et les caractéristiques du jeu de données composé de 21 724 échantillons provenant de 10 puits du bassin de Touggourt. L'analyse de la distribution lithologique révèle un déséquilibre marqué, avec une dominance des argiles et du quartz, tandis que les lithologies minoritaires représentent moins de 5 % des occurrences. Cette configuration a conduit à l'adoption d'une stratégie hybride combinant classification et régression pour optimiser la prédiction de toutes les lithologies. Les bases ainsi établies permettent de procéder à l'entraînement et à l'évaluation des modèles d'apprentissage automatique présentés dans le chapitre suivant.

# Chapitre 5

# Modélisation, validation et interprétation des résultats

## 5.1 Introduction

Ce chapitre présente la mise en œuvre des modèles de gradient boosting pour la prédiction des paramètres pétrophysiques et de la composition lithologique. Après une justification rigoureuse du choix des algorithmes, une procédure de prétraitement des données, et une analyse corrélationnelle des variables, nous détaillons les résultats de prédiction pour le groupe 1 (lithologies continues) et le groupe 2 (lithologies sporadiques), avec comparaison du modèle optimal sur un puits de test historiquement indépendant.

## 5.2 Justification du choix des modèles

### 5.2.1 Critères de sélection

Le choix des algorithmes de gradient boosting - XGBoost, CatBoost et LightGBM - repose sur leur capacité à gérer nativement les valeurs manquantes, contrairement aux méthodes classiques. Cette propriété est déterminante dans notre contexte pétrophysique.

**Gestion des valeurs manquantes par chaque modèle :**

- **XGBoost** : Attribue automatiquement les instances avec valeurs manquantes (-9999 dans notre dataset) à une direction de division dans chaque arbre. L'algorithme apprend la meilleure direction pendant l'entraînement, traitant les valeurs manquantes comme une catégorie implicite [32].
- **CatBoost** : Gère les valeurs manquantes de manière native par son mécanisme d'ordered boosting. Il traite les données manquantes sans suppression ni imputation préalable, préservant l'intégrité des informations disponibles [33].
- **LightGBM** : Utilise une stratégie de division basée sur le gain d'information, où les valeurs manquantes sont assignées dynamiquement au nœud enfant maximisant la réduction

de perte [34].

Les méthodes classiques (forêts aléatoires, régressions linéaires) nécessitent de supprimer ou d'imputer les valeurs manquantes. L'imputation par moyenne, médiane ou autres statistiques pose problème pour deux raisons principales :

1. **Perte d'information** : Une suite de valeurs manquantes consécutives (ex : -9999 répétés) reflète souvent une défaillance de l'instrument sur tout un intervalle stratigraphique. Remplacer ces valeurs par des statistiques globales introduit du bruit et masque les signatures géologiques réelles.
2. **Réduction du volume de données** : Supprimer les lignes avec des valeurs manquantes réduit fortement la taille du dataset. Ce problème est amplifié par le déséquilibre naturel des lithologies rares (ex : halite 1,50%), rendant l'apprentissage des lithologies minoritaires difficile.

### 5.2.2 Impact de la suppression : analyse quantitative

Si toutes les lignes contenant des valeurs manquantes sont supprimées (`dropna`) :

- Dataset initial : 21 724 lignes
- Dataset après suppression : 18 844 lignes
- Perte : 2 880 lignes (13,3%)
- Valeurs manquantes restantes : 0

Pour limiter la perte de données, un critère minimal conservateur peut être appliqué : ne retenir que les lignes présentant au moins 3 valeurs valides sur 15 diagraphies. Ce critère garantit que le modèle dispose d'une quantité d'information suffisante tout en tolérant des dysfonctionnements partiels des instruments.

**Résultats du filtrage conservateur :**

- Dataset initial : 21 724 lignes
- Dataset après filtrage : 21 511 lignes
- Perte : 213 lignes (1%)
- Valeurs manquantes restantes : 3 845 (traitées par les modèles)

Cette méthode permet de conserver la majorité du dataset tout en maintenant la cohérence géologique. L'approche choisie, combinant *gradient boosting* natif et prétraitement minimal, préserve les informations essentielles et limite la perte de données.

## 5.3 Optimisation des hyperparamètres

### 5.3.1 Approche hybride : Optuna + optimisation manuelle

L'ajustement des hyperparamètres est crucial pour obtenir un compromis optimal entre biais et variance. Nous avons adopté une stratégie hybride combinant recherche bayésienne automatisée et itérations manuelles.

A. **Optuna : recherche bayésienne automatisée**

Optuna est une bibliothèque d'optimisation hyperparamètrique utilisant la recherche bayésienne. Contrairement aux approches exhaustives (GridSearchCV, RandomizedSearchCV) qui évaluent tous les hyperparamètres selon un produit cartésien, Optuna construit un modèle probabiliste de la fonction objectif pour guider la recherche vers les régions prometteurs, réduisant significativement le coût computationnel, particulièrement sur notre dataset de 21 511 lignes.

Lors d'une première phase, nous avons exécuté Optuna sur des intervalles larges pour chaque hyperparamètre :

- Learning rate : [0,001 ; 0,3]
- Nombre d'arbres : [100 ; 1100]
- Max depth : [3 ; 13]
- Min child weight / min samples leaf : [1 ; 10]
- Subsampling rate : [0,5 ; 1,0]

Cette phase identifie les régions optimales de l'espace hyperparamètrique sans rechercher précision excessive.

B. **Optimisation manuelle : affinement et contrôle du surapprentissage**

Bien que Optuna propose des solutions de haute qualité en termes d'erreur (MSE, RMSE minimisés), elle présente une limitation majeure : négliger le surapprentissage. Les hyperparamètres recommandés par Optuna maximisent souvent les performances d'entraînement au détriment de la généralisation. De plus, les résultats d'Optuna ne sont pas reproductibles à travers plusieurs exécutions sur le même dataset, en raison de la nature stochastique de la recherche bayésienne.

Pour pallier ces limitations, nous avons adopté une optimisation manuelle itérative basée sur les intervalles fournis par Optuna. Cette approche teste systématiquement différentes valeurs d'hyperparamètres (détaillées au Chapitre 3 : Learning Rate, Nombre d'Arbres, Max Depth, Min Samples per Leaf, Subsampling Rate, Feature Sampling Rate) en privilégiant :

a. **Équilibre train-validation** : Écart MSE/RMSE minimal entre ensemble d'entraînement et validation

b. **Minimisation de l'overfitting** : Écart $R^2$ train-validation $< 2\%$

c. **Robustesse en test** : Performances cohérentes sur l'ensemble indépendant

d. **Validation croisée** : Stabilité à travers 5-fold cross-validation

Cette approche garantit des modèles généralisant correctement sur des données non vues, contrairement à une optimisation automatique non supervisée.

## 5.4 Prétraitement et préparation des données

Nettoyage et Filtrage Intelligent

Avant la modélisation, le dataset brut a subi une étape de prétraitement ciblée. Contrairement à une suppression agressive, notre stratégie conserve le maximum d'information géologique :

- **Critères de conservation :**

  Une ligne est conservée si elle contient **au minimum 3 valeurs valides** parmi les 15 diagraphies. Ce critère permet :

  - Au modèle de disposer d'informations minimales pour apprentissage
  - De tolérer les dysfonctionnements partiels d'outils (un à deux logs défaillants)
  - De rejeter uniquement les lignes où l'information est quasi-absente (information inutilisable)

- **Résultats du nettoyage :**

  Le nettoyage a préservé 99% des données originelles (21 511 lignes retenues sur 21 724), avec une perte négligeable de 1% et l'identification de 3 845 valeurs manquantes traitées par les modèles.

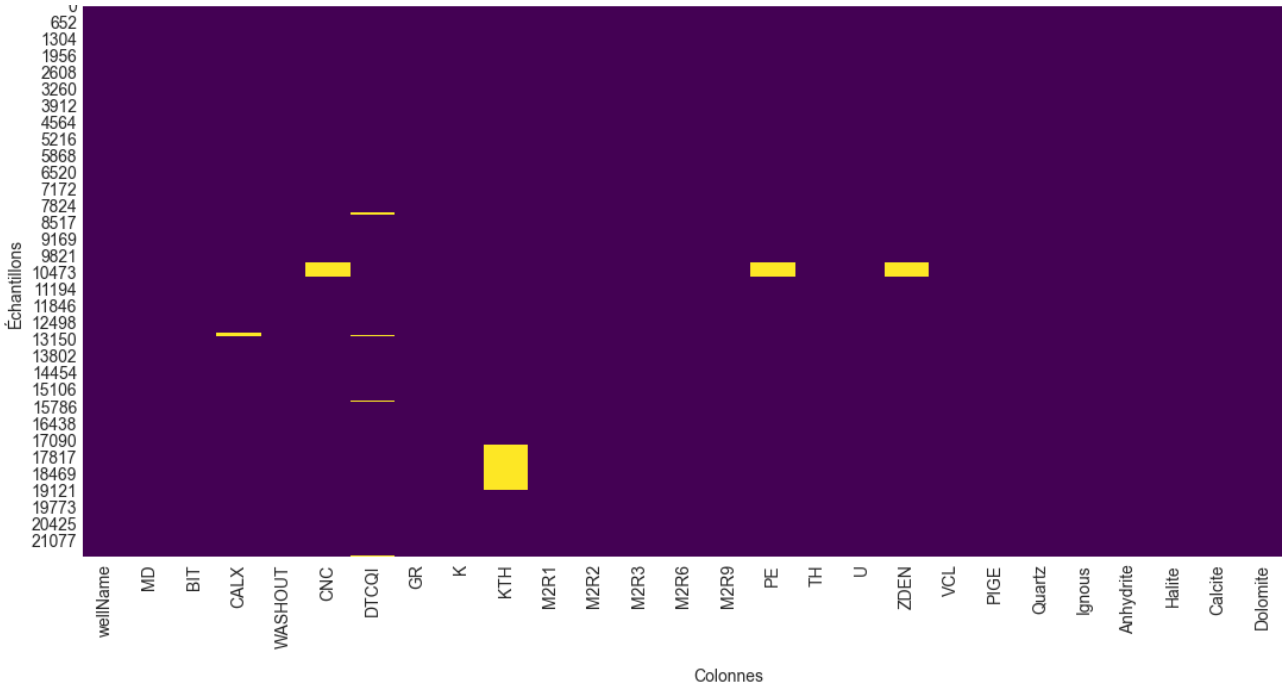


FIGURE 5.1 – Heatmap des valeurs manquantes après prétraitement des données.

## 5.5 Analyse corrélationnelle

### 5.5.1 Matrice de corrélation : logs vs lithologies

La Figure 5.2 montre la matrice de corrélation de Pearson entre les 15 diagraphies et les 5 cibles principales révèle les forces des associations linéaires tel que Les valeurs rouges indiquent corrélation positive forte, bleu indique corrélation négative forte, blanc indique absence de corrélation.

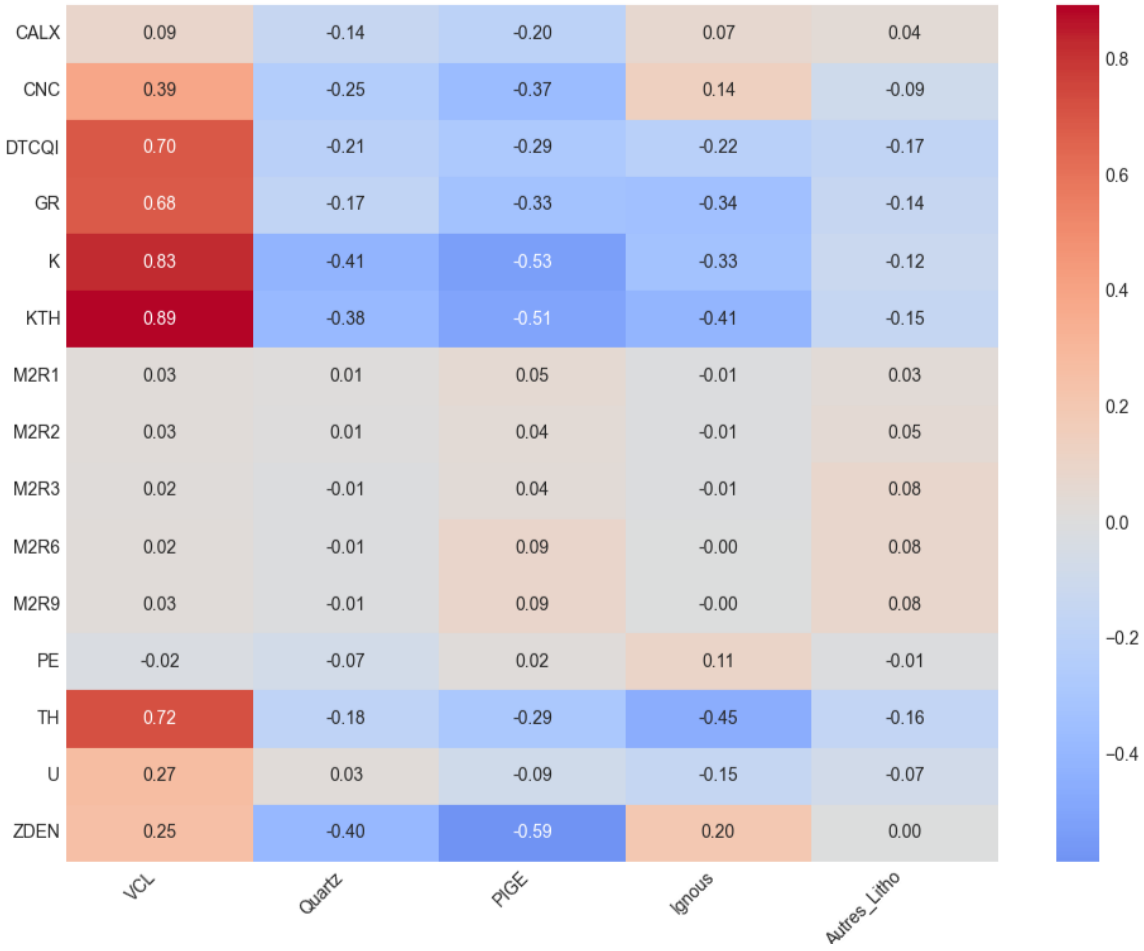


FIGURE 5.2 – Matrice de corrélation : Logs (X) vs lithologies/paramètres (Y).

**Variables retenues pour la modélisation :**

$$\text{Features} = \{\text{CALX}, \text{GR}, \text{CNC}, \text{KTH}, \text{DTCQI}, \text{K}, \text{TH}, \text{ZDEN}\} \tag{5.1}$$

### 5.5.2 Interprétation détaillée

**Corrélations fortes positives (>0,65) :**

- **VCL vs KTH (0,89), TH (0,72), GR (0,68)** : Les logs radioactifs (potassium-thorium, thorium, gamma ray) corrèlent fortement avec l'argile, confirmant le lien physique : argiles contiennent des éléments radioactifs.

- **VCL vs DTCQI (0,70)** : Les ondes soniques ralentissent en présence d'argile (porosité, contenu argileux), d'où cette corrélation attendue.

**Corrélations négatives fortes (<-0,4) :**

- **Quartz vs K (-0,41), KTH (-0,51), TH (-0,38)** : Les grès purs contiennent peu de minéraux radioactifs, d'où la corrélation négative inverse.
- **PIGE vs K (-0,53), ZDEN (-0,59)** : La porosité augmente quand la densité diminue et que la radioactivité décroît (moins d'argile compactée).

**Absence notoire de corrélation :**

- **Résistivités (M2R1–M2R9) vs Cibles (<0,09)** : Totalement décorrélées. Les résistivités dépendent primarily de la saturation en fluide (eau, huile, gaz), pas de la lithologie. Leur exclusion est justifiée pour la prédiction lithologique.
- **PE vs Cibles (<0,11)** : Corrélation négligeable malgré son importance géologique théorique. L'outil PE est extrêmement sensible au washout (géométrie du puits instable). Quand le caliper augmente (caverne ou argile gonflante), le pad du détecteur s'éloigne de la paroi, produisant des mesures invalides. En raison de cette fragilité, PE est retenu dans le pipeline global mais avec poids réduit.

## 5.6 Prédiction : groupe 1 (proportions continues majoritaires)

Le groupe 1 comprend les lithologies présentes de manière continue et majoritaire dans les formations étudiées : le volume d'argile (VCL), le quartz, et la porosité effective (PIGE). Ces composants constituent l'essentiel de la matrice pétrophysique des réservoirs et présentent des distributions statistiques favorables à l'apprentissage supervisé. Cette section détaille les performances comparatives des modèles de gradient boosting pour chacune de ces trois cibles.

### 5.6.1 Volume d'argile (VCL)

#### 5.6.1.1 Comparaison des modèles

Le tableau 5.1 synthétise les performances des trois algorithmes de gradient boosting pour la prédiction du volume d'argile.

TABLE 5.1 – Comparaison des performances pour VCL.

| Modèle | Ensemble | $R^2$ | MSE | RMSE | MAE | SMAPE (%) | PBIAS (%) |
|---|---|---|---|---|---|---|---|
| **CatBoost** | Train | 0.9977 | 0.0003 | 0.0161 | 0.0109 | 5.92 | 0.00 |
| | Validation | **0.9814** | 0.0021 | 0.0453 | 0.0269 | 9.05 | -0.06 |
| | Test | **0.9828** | **0.0019** | **0.0436** | **0.0258** | **9.16** | **-0.03** |
| **XGBoost** | Train | 0.9889 | 0.0012 | 0.0353 | 0.0239 | 10.07 | 0.00 |
| | Validation | 0.9731 | 0.0030 | 0.0545 | 0.0347 | 11.51 | -0.07 |
| | Test | 0.9742 | 0.0029 | 0.0535 | 0.0339 | 11.77 | -0.09 |
| **LightGBM** | Train | 0.9850 | 0.0017 | 0.0409 | 0.0269 | 10.47 | 0.00 |
| | Validation | 0.9716 | 0.0031 | 0.0560 | 0.0358 | 11.61 | -0.19 |
| | Test | 0.9719 | 0.0031 | 0.0558 | 0.0352 | 11.86 | -0.12 |
| **Validation croisée (5-fold CV) :** CatBoost : **0.9816** \| XGBoost : 0.9734 \| LightGBM : 0.9723 | | | | | | | |

#### 5.6.1.2 Analyse des performances

**Supériorité de CatBoost :**

CatBoost se distingue avec un **$R^2$ test de 0,9828**, expliquant 98,28% de la variance du VCL. Cette performance représente une amélioration significative par rapport aux algorithmes concurrents :

- **vs XGBoost** : +0,86 points de $R^2$ (97,42% → 98,28%)
- **vs LightGBM** : +1,09 points de $R^2$ (97,19% → 98,28%)

**Indicateurs clés de performance :**

- **Précision absolue (MAE)** : 0,0258 v/v correspond à une erreur moyenne de 2,58 points de pourcentage. CatBoost réduit l'erreur MAE de 24% par rapport à XGBoost (0,0339 → 0,0258) et de 27% par rapport à LightGBM (0,0352 → 0,0258).
- **Erreur symétrique (SMAPE)** : 9,16% se situe dans la catégorie « bon » selon les critères d'évaluation standards (0-5% : très bon ; 5-15% : bon). Les algorithmes concurrents affichent des SMAPE de 11,77% (XGBoost) et 11,86% (LightGBM), soit des erreurs 28% et 29% plus élevées respectivement.
- **Biais systématique (PBIAS)** : -0,03% indique un biais quasi-nul. Le signe négatif signifie que le modèle sous-estime très légèrement le VCL, mais cette déviation est négligeable (critère « très bon » : PBIAS $< 2\%$). Le modèle ne présente aucune tendance systématique significative.
- **Robustesse par validation croisée** : Le $R^2$ de validation croisée (0,9816) est parfaitement cohérent avec le $R^2$ test (0,9828), avec un écart de seulement 0,12%. Cette stabilité démontre une variance inter-folds minimale et une excellente capacité de généralisation.

- **Contrôle du surapprentissage** : L'écart entre $R^2$ train (0,9977) et $R^2$ test (0,9828) est de seulement 1,49 points, attestant d'un surapprentissage négligeable. Le modèle capture les patterns géologiques généralisables sans mémoriser les spécificités locales du jeu d'entraînement.

#### 5.6.1.3 Validation graphique

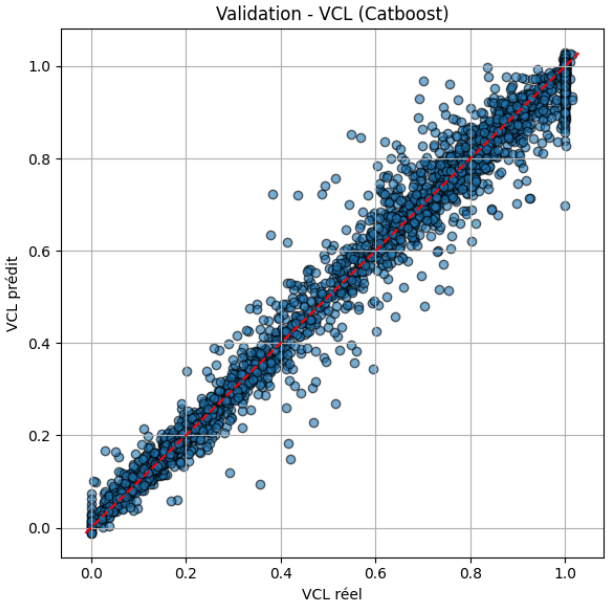


FIGURE 5.3 – Graphique de dispersion : valeurs prédites vs observées pour VCL

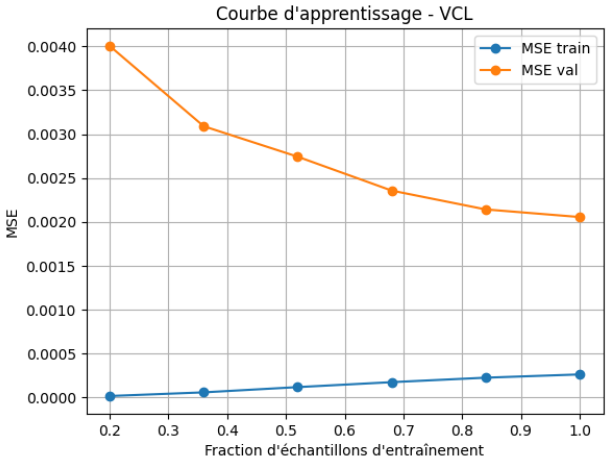


FIGURE 5.4 – Courbe d'apprentissage de CatBoost pour VCL

**Interprétation du graphique de dispersion :**

Le graphique révèle une concentration optimale des points le long de la droite y=x, signature d'une prédiction quasi-parfaite. La distribution est symétrique et couvre l'intégralité de l'intervalle [0-1] du volume d'argile, sans formation d'amas ni biais directionnels. Cette répartition homogène confirme l'excellente calibration du modèle sur toute la gamme de valeurs observées.

**Interprétation de la courbe d'apprentissage :**

La courbe d'apprentissage montre une convergence précoce dès 40-50% du dataset, avec un écart MSE train-validation minimal et stable. Cette dynamique démontre :

- Une capacité d'apprentissage rapide : le modèle identifie les patterns fondamentaux dès les premiers arbres de décision
- Une absence de surapprentissage : la divergence train-validation reste parfaitement maîtrisée
- Une généralisation robuste : les performances demeurent stables au-delà de 50% des données d'entraînement

### 5.6.2 Quartz

Le quartz constitue le composant principal de la fraction silicatée des réservoirs gréseux.

#### 5.6.2.1 Comparaison des modèles

TABLE 5.2 – Comparaison des performances pour Quartz.

| Modèle | Ensemble | R² | MSE | RMSE | MAE | SMAPE (%) | PBIAS (%) |
|---|---|---|---|---|---|---|---|
| **CatBoost** | Train | 0.9270 | 0.0087 | 0.0935 | 0.0578 | 35.18 | -0.01 |
| | Validation | **0.9042** | 0.0114 | 0.1067 | 0.0646 | 37.30 | 0.35 |
| | Test | **0.9055** | **0.0114** | **0.1066** | **0.0659** | **37.34** | **0.18** |
| **XGBoost** | Train | 0.9080 | 0.0110 | 0.1050 | 0.0657 | 36.59 | 0.01 |
| | Validation | 0.8803 | 0.0142 | 0.1193 | 0.0736 | 39.34 | 0.49 |
| | Test | 0.8866 | 0.0136 | 0.1167 | 0.0745 | 38.81 | 0.14 |
| **LightGBM** | Train | 0.9288 | 0.0085 | 0.0924 | 0.0559 | 34.15 | 0.00 |
| | Validation | 0.8980 | 0.0121 | 0.1101 | 0.0666 | 37.72 | 0.36 |
| | Test | 0.9011 | 0.0119 | 0.1090 | 0.0674 | 37.10 | 0.20 |
| **Validation croisée (5-fold CV) :** CatBoost : **0.8935** \| XGBoost : 0.8721 \| LightGBM : 0.8952 | | | | | | | |

#### 5.6.2.2 Analyse comparative

CatBoost maintient sa supériorité avec un **R² test de 0,9055**, expliquant 90,55% de la variance du quartz. Comparé aux algorithmes concurrents, CatBoost surpasse LightGBM de +0,44 point de R² et XGBoost de +1,89 points.

**Indicateurs clés de performance :**

- **Précision absolue (MAE)** : 0,0659 v/v correspond à une erreur moyenne de 6,59 points de pourcentage. Cette précision reste acceptable compte tenu de la complexité modérée du réservoir argilo-gréseux où les proportions de quartz varient significativement selon les cycles de sédimentation, dont l'hétérogénéité minéralogique et texturale demeure limitée par rapport aux réservoirs carbonatés.
- **Erreur symétrique (SMAPE)** : 37,34% est plus élevée que celle du VCL (9,16%), mais reste compétitive face à XGBoost (38,81%) et LightGBM (37,10%). Cette hausse du SMAPE s'explique principalement par la sensibilité de la formule à la présence de valeurs de quartz proches de zéro, ce qui amplifie l'erreur relative sans pour autant traduire une réelle dégradation de la performance prédictive.
- **Biais systématique (PBIAS)** : 0,18% est négligeable et satisfait largement le critère « très bon » (PBIAS $< 2\%$). Le signe positif indique que le modèle surestime très légèrement le quartz, mais cette déviation est sans conséquence pratique.

- **Contrôle du surapprentissage** : L'écart train-test de 2,15 points de $R^2$ ($R^2$ train : 0,9270 ; $R^2$ test : 0,9055) demeure bien maîtrisé, bien que légèrement supérieur à celui du VCL (1,49 points). Cette différence est attendue en raison de la complexité minéralogique accrue du quartz par rapport à l'argile, composant plus homogène stratigraphiquement.

#### 5.6.2.3 Validation graphique

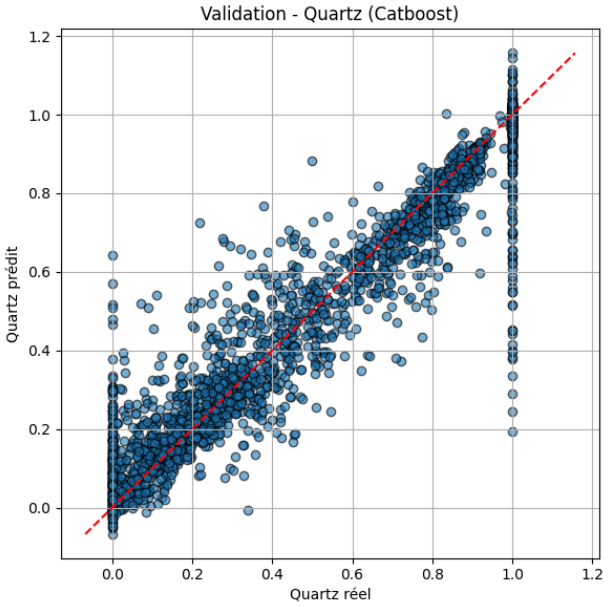


FIGURE 5.5 – Graphique de dispersion : valeurs prédites vs observées pour Quartz

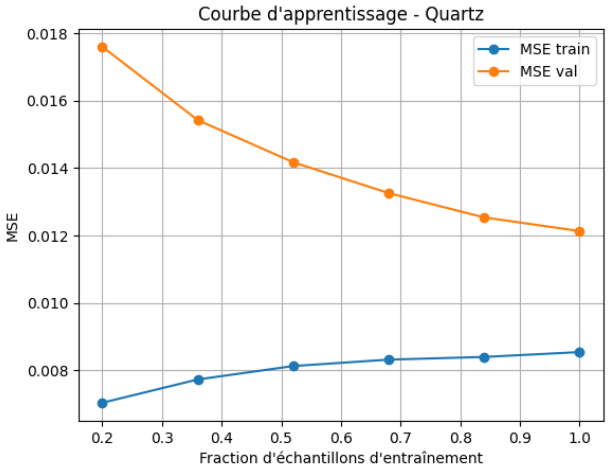


FIGURE 5.6 – Courbe d'apprentissage de CatBoost pour Quartz

**Interprétation du graphique de dispersion :**

Le graphique de dispersion affiche une distribution plus dispersée autour de la diagonale par rapport au VCL. La concentration des points reste néanmoins satisfaisante, particulièrement pour les valeurs élevées de quartz (>0,6 v/v) correspondant aux niveaux gréseux réservoirs purs. La dispersion accrue dans les gammes intermédiaires (0,3-0,6 v/v) traduit la variabilité naturelle des faciès mixtes argilo-gréseux.

**Interprétation de la courbe d'apprentissage :**

La courbe d'apprentissage révèle une convergence progressive avec un écart train-validation légèrement plus marqué que pour VCL (MSE train = 0,0087 ; MSE validation = 0,0114), tout en demeurant stable et contrôlé. Cet écart reflète la variabilité naturelle du quartz dans les faciès sédimentaires : selon la diagenèse, le compactage et l'histoire de dépôt, deux intervalles apparemment similaires en termes de signatures diagraphiques peuvent présenter des compositions minéralogiques distinctes. Malgré cette complexité géologique, la stabilisation précoce de l'erreur de validation confirme l'absence de surapprentissage.

### 5.6.3 Porosité effective (PIGE)

La porosité effective représente la fraction de volume poreux accessible aux fluides mobiles, excluant la porosité piégée dans les argiles. Ce paramètre critique contrôle directement la capacité de stockage des hydrocarbures.

Contrairement à VCL et Quartz, la prédiction de PIGE a été effectuée uniquement avec **XGBoost**. Ce choix constitue un compromis entre rigueur méthodologique et efficacité computationnelle. En effet, l'optimisation conjointe de trois algorithmes (XGBoost, CatBoost et LightGBM) sur trois cibles distinctes aurait engendré un coût calculatoire important.

#### 5.6.3.1 Performances de XGBoost

TABLE 5.3 – Performances de XGBoost pour PIGE.

| Ensemble | $R^2$ | MSE | RMSE | MAE | SMAPE (%) | PBIAS (%) |
|---|---|---|---|---|---|---|
| Train | 0.8763 | 0.00014 | 0.0136 | 0.0085 | 57.34 | 0.58 |
| Validation | 0.8758 | 0.0002 | 0.0140 | 0.0085 | 58.29 | 1.24 |
| Test | 0.8554 | 0.0002 | 0.0147 | 0.0091 | 60.85 | 2.23 |
| **Validation croisée (5-fold CV) :** $R^2$ moyen = **0.8375** | | | | | | |

#### 5.6.3.2 Analyse des résultats

**Performance globale :**

- **Stabilité train-validation-test :**

  Les $R^2$ d'entraînement (0,8763) et de validation (0,8758) sont quasi-identiques (écart de 0,05 point seulement), avec un $R^2$ test (0,8554) légèrement inférieur. L'écart train-test de 2,09 points de $R^2$ (0,8763 - 0,8554) indique un surapprentissage minimal et une capacité de généralisation robuste. Cette stabilité démontre que le modèle a capturé les patterns fondamentaux régissant la porosité effective sans mémoriser les singularités des puits d'entraînement.

- **Précision absolue (MAE) :**

  L'erreur absolue moyenne évolue progressivement : 0,0085 v/v (entraînement) → 0,0085 v/v (validation) → 0,0091 v/v (test). L'augmentation limitée de 0,0006 v/v entre validation et test traduit une excellente stabilité. Le MAE de test (0,0091 v/v) correspond à une erreur d'environ 0,91 point de pourcentage, ce qui demeure cohérent avec la faible amplitude des valeurs de PIGE, inférieures à 23% avec une médiane de 0,06 v/v.

- **Erreur relative (SMAPE) :**

  Le SMAPE test de 60,85% est significativement plus élevé que ceux de VCL (9,16%) et Quartz (37,34%). Cette différence apparente ne reflète pas nécessairement une mauvaise performance, mais s'explique par la distribution asymétrique de PIGE dans le dataset : forte concentration d'échantillons à faible porosité (proche de zéro) et quelques valeurs extrêmes élevées.

- **Analyse du biais systématique (PBIAS) :**

  Le PBIAS progresse graduellement : 0,58% (train) → 1,24% (validation) → 2,23% (test). Cette augmentation révèle une **légère tendance systématique à surestimer la porosité**.

- **Validation croisée :**

  Le $R^2$ moyen de validation croisée (0,8375) est légèrement inférieur au $R^2$ test (0,8554), avec un écart de 1,79 points en faveur du test.L'écart reste modeste (<2 points), confirmant la robustesse générale du modèle et l'absence de surapprentissage significatif.

#### 5.6.3.3 Validation graphique

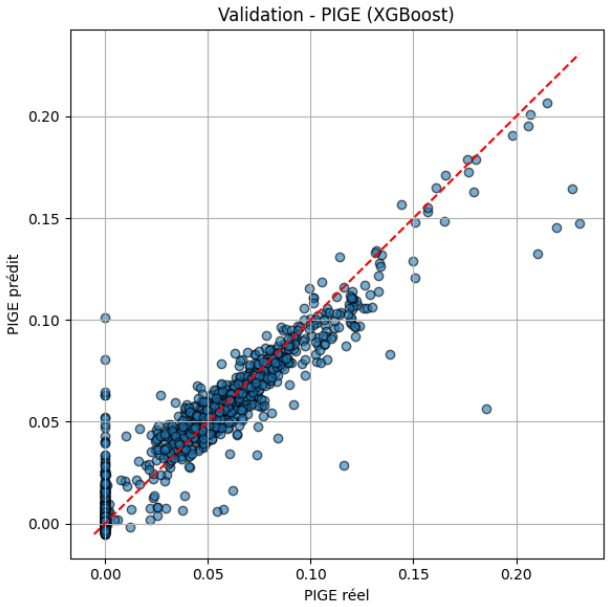


FIGURE 5.7 – Graphique de dispersion : valeurs prédites vs observées pour PIGE

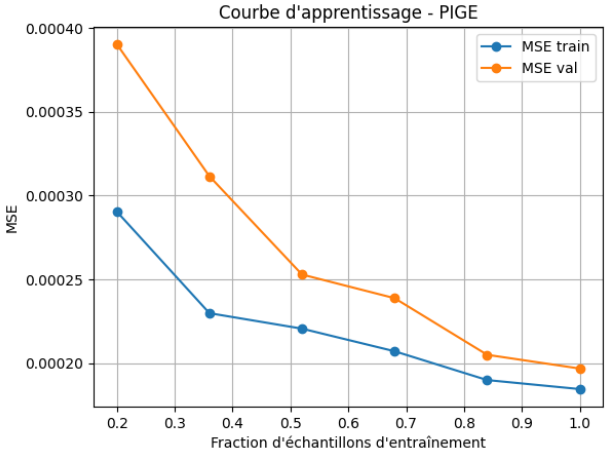


FIGURE 5.8 – Courbe d'apprentissage : Convergence MSE train-validation pour PIGE

**Interprétation du graphique de dispersion :**

L'analyse du diagramme de dispersion révèle un alignement satisfaisant des points le long de la diagonale y=x, attestant d'une bonne corrélation entre valeurs prédites et observées, particulièrement dans la plage des faibles à moyennes porosités (0-0,15 v/v). Cette zone concentre la majorité des observations et démontre la capacité du modèle à capturer les variations de porosité dans les réservoirs typiques du bassin d'étude.

La dispersion augmente pour les valeurs élevées de porosité (>0,15 v/v), ce qui est géologiquement attendu : ces échantillons à forte porosité sont rares dans le dataset, et le modèle dispose donc de moins d'exemples pour apprendre leurs caractéristiques. Malgré cette incertitude accrue sur les valeurs extrêmes, le modèle conserve une capacité de discrimination acceptable pour identifier les zones à forte porosité, critiques pour l'identification des réservoirs productifs.

**Interprétation de la courbe d'apprentissage :**

La courbe d'apprentissage confirme la stabilité du modèle avec une convergence rapide dès 40% des données d'entraînement. L'écart entre MSE d'entraînement (0,00014) et MSE de validation ( 0,0002), signature caractéristique d'une excellente généralisation sans surapprentissage.

### 5.6.4 Bilan du groupe 1

Les performances obtenues pour le groupe 1 établissent un benchmark solide et opérationnel pour les prédictions lithologiques continues :

- **VCL (CatBoost)** : $R^2 = 0{,}9828$ – Performance exceptionnelle, expliquant 98,28% de la variance. L'argile, composant principal des réservoirs argilo-gréseux, présente une corrélation forte et stable avec les diagraphies radioactives (GR, K, TH), permettant une prédiction quasi-parfaite.
- **Quartz (CatBoost)** : $R^2 = 0{,}9055$ – Performance excellente (90,55% de variance expliquée), validant la capacité du modèle à prédire les minéraux silicatés malgré la variabilité minéralogique hétérogène des grès liée aux cycles de sédimentation, à la diagenèse, et au compactage différentiel.
- **PIGE (XGBoost)** : $R^2 = 0{,}8554$ – Performance bonne (85,54% de variance expliquée), pleinement acceptable pour les applications exploratoires. La performance modérée reflète la complexité intrinsèque de la porosité effective, contrôlée par de multiples facteurs au-delà du simple volume d'argile (cimentation, dissolution, microfracturation), et la distribution asymétrique des données.

## 5.7 Prédiction : groupe 2 (proportions discontinues minoritaires)

Le groupe 2 regroupe les lithologies présentes de manière sporadique et minoritaire dans les formations étudiées : les roches ignées (Igneous) et un ensemble de lithologies carbonatées et évaporitiques regroupées sous l'appellation « Autres_Litho » (calcite, dolomite, anhydrite, halite). Contrairement au Groupe 1 où les composants sont distribués de façon continue, ces lithologies apparaissent dans des intervalles stratigraphiques spécifiques et représentent une fraction réduite du volume total.

### 5.7.1 Contexte stratigraphique et défi méthodologique

#### 5.7.1.1 Distribution discontinue

La figure 5.9 illustre la répartition binaire (présence/absence) des deux cibles du groupe 2 dans le dataset complet.

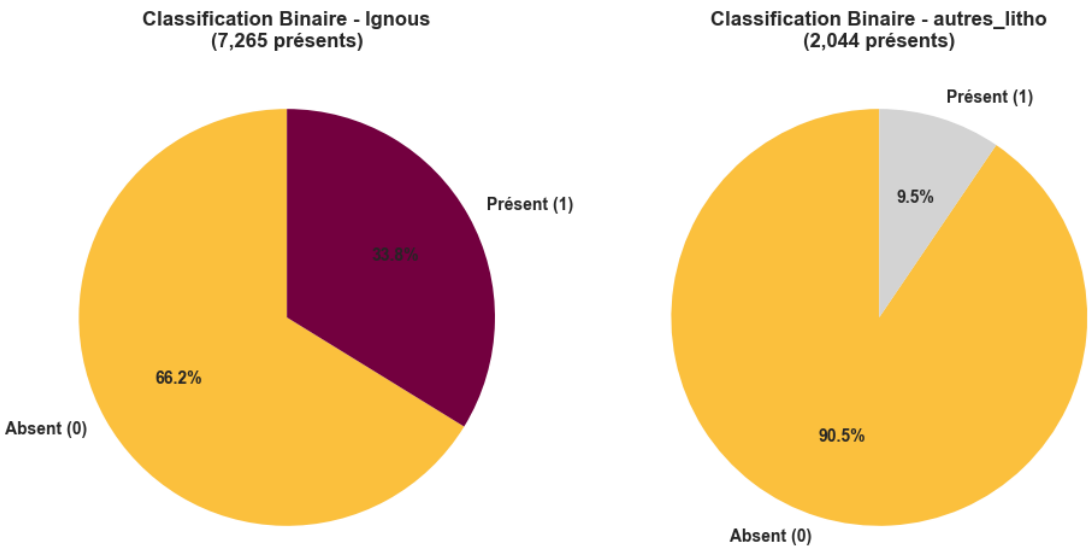


FIGURE 5.9 – Distribution binaire des lithologies du groupe 2 : Igneous et Autres_Litho.

**Observations clés :**

- **Igneous** : Présent dans 33,8% des échantillons (7 265 présents / 21 511 total), absent dans 66,2% (14 246 absents).
- **Autres_litho** : Présent dans seulement 9,5% des échantillons (2 044 présents / 21 511 total), absent dans 90,5% (19 467 absents).

Cette forte asymétrie classe-majoritaire/classe-minoritaire impose une contrainte méthodologique majeure : prédire directement les proportions continues par régression conduirait à un modèle biaisé qui prédirait systématiquement des valeurs proches de zéro (la moyenne du dataset) même dans les intervalles où ces lithologies sont effectivement présentes.

#### 5.7.1.2 Justification de l'approche hybride classification-régression

Pour contourner le problème du déséquilibre de classes, une **architecture hybride en deux étapes** a été implémentée :

1. **Étape 1 - classification binaire** : Identifier si la lithologie est présente (1) ou absente (0) à une profondeur donnée. Cette étape permet de localiser les intervalles stratigraphiques contenant la lithologie cible.
2. **Étape 2 - régression conditionnelle** : Pour les échantillons classés comme « présents » à l'étape 1, prédire la proportion volumique exacte par régression. Cette étape affine la prédiction en quantifiant l'abondance.

Cette stratégie garantit que les lithologies sporadiques ne seront prédites que dans leurs contextes géologiques cohérents.

### 5.7.2 Prédiction des roches ignées (Igneous)

#### 5.7.2.1 Étape 1 : classification binaire

**Distribution des classes :**

Le dataset pour la classification binaire d'Igneous comprend 21 511 échantillons répartis comme suit :

- Classe 0 (Absent) : 14 246 échantillons (66,2%)
- Classe 1 (Présent) : 7 265 échantillons (33,8%)

Malgré le déséquilibre (ratio 2 :1), la classe minoritaire reste suffisamment représentée pour permettre un apprentissage efficace.

**Performances de XGBoost :**

Le tableau 5.4 synthétise les performances du modèle XGBoost optimisé via Optuna pour la classification binaire d'Igneous.

TABLE 5.4 – Rapport de classification pour Igneous (XGBoost optimisé via Optuna).

| Classe | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (Absent) | 0.98 | 0.97 | 0.98 | 2 850 |
| 1 (Présent) | 0.95 | 0.95 | 0.95 | 1 453 |
| **Accuracy** | | **0.97** | | 4 303 |
| **Macro avg** | 0.96 | 0.96 | 0.96 | 4 303 |
| **Weighted avg** | 0.97 | 0.97 | 0.97 | 4 303 |

**Analyse des performances :**

- **Accuracy globale** : 97% indique que le modèle classe correctement 97% des échantillons, performance excellente pour une tâche de classification binaire déséquilibrée.

- **Precision classe 1 (Présent)** : 0,95 signifie que lorsque le modèle prédit « Igneous présent », il a raison dans 95% des cas. Cela minimise les faux positifs (prédictions erronées d'Igneous dans des zones où il est absent).
- **Recall classe 1 (Présent)** : 0,95 indique que le modèle détecte correctement 95% des occurrences réelles d'Igneous. Seulement 5% des vrais positifs sont manqués (faux négatifs).
- **F1-Score classe 1** : 0,95 représente l'équilibre harmonique entre precision et recall, confirmant la performance robuste sur la classe minoritaire.
- **Performance sur classe 0 (Absent)** : Precision et recall de 0,98 et 0,97 respectivement démontrent que le modèle identifie quasi-parfaitement les zones sans Igneous.
- **Équilibre macro/weighted** : Les moyennes macro (0,96) et weighted (0,97) sont très proches, attestant d'une performance équilibrée entre les deux classes malgré le déséquilibre.

**Validation graphique :**

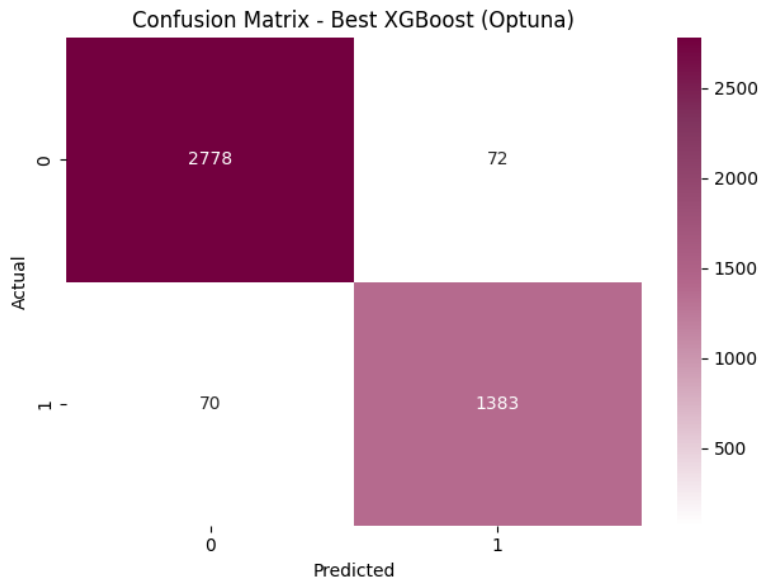


FIGURE 5.10 – Matrice de confusion pour la classification binaire d'Igneous (XGBoost, ensemble de test).

La matrice de confusion (Figure 5.10) révèle :

- **Vrais négatifs (TN)** : 2 778 échantillons correctement classés comme « Absent » (97,5% de la classe 0).
- **Faux positifs (FP)** : 72 échantillons incorrectement classés comme « Présent » alors qu'Igneous est absent (2,5% d'erreur sur classe 0).
- **Faux négatifs (FN)** : 70 échantillons d'Igneous réellement présents mais manqués par le modèle (4,8% d'erreur sur classe 1).
- **Vrais positifs (TP)** : 1 383 échantillons correctement identifiés comme « Présent » (95,2% de la classe 1).

Le faible nombre de faux positifs (72) montre que le modèle réduit significativement les prédictions d'Igneous en dehors de leur contexte géologique. Les occurrences résiduelles seront prises en compte ultérieurement dans un traitement dédié.

#### 5.7.2.2 Étape 2 : régression conditionnelle

Après identification des intervalles contenant Igneous via la classification binaire, la régression quantifie les proportions volumiques exactes. Le modèle CatBoost a été retenu pour cette tâche en raison de ses performances supérieures sur les variables continues.

**Performances de CatBoost :**

TABLE 5.5 – Performances de CatBoost pour la régression d'Igneous (sur échantillons classés « Présent »).

| **Ensemble** | **$R^2$** | **MSE** | **RMSE** | **MAE** | **SMAPE (%)** | **PBIAS (%)** |
|---|---|---|---|---|---|---|
| Train | 0.9627 | 0.0029 | 0.0539 | 0.0369 | 13.90 | 0.00 |
| Validation | 0.9438 | 0.0044 | 0.0665 | 0.0450 | 15.20 | -0.02 |
| Test | 0.9391 | 0.0045 | 0.0670 | 0.0450 | 13.97 | -0.69 |
| **Validation croisée (5-fold CV) :** $R^2$ moyen = **0.9401** | | | | | | |

**Analyse détaillée :**

- **Performance globale** : $R^2$ test de 0,9391 explique 93,91% de la variance d'Igneous, performance **exceptionnelle** pour une lithologie sporadique. Cette valeur se situe entre VCL (0,9828) et Quartz (0,9055), démontrant que la stratégie hybride permet d'atteindre des performances comparables aux lithologies continues.
- **Stabilité train-validation-test** : Les $R^2$ d'entraînement (0,9627), validation (0,9438) et test (0,9391) présentent des écarts modestes (2,36 points entre train et test), attestant d'un surapprentissage bien maîtrisé.
- **Précision absolue (MAE)** : 0,0450 v/v représente une erreur moyenne de 4,50 points de pourcentage. Cette précision est remarquable compte tenu du caractère discontinu d'Igneous.
- **Erreur relative (SMAPE)** : 13,97% se situe dans la catégorie « bon » (critère 5-15%), nettement inférieure à celle de PIGE (60,85%) et comparable à VCL (9,16%). Cette faible erreur relative s'explique par l'absence de valeurs proches de zéro dans le sous-ensemble de régression (tous les échantillons contiennent Igneous).
- **Biais systématique (PBIAS)** : -0,69% indique une très légère tendance à sous-estimer Igneous, mais cette déviation est négligeable (critère « très bon » : PBIAS $< 2\%$). Le modèle est quasi-non biaisé.
- **Validation croisée** : $R^2$ CV moyen de 0,9401 est remarquablement cohérent avec $R^2$ test (0,9391), avec seulement 0,10 point d'écart. Cette stabilité exceptionnelle confirme la robustesse inter-folds et la capacité de généralisation.

**Validation graphique :**

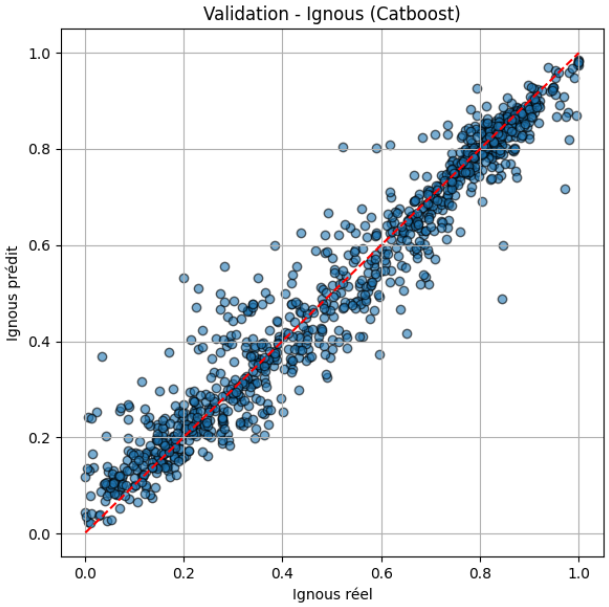


FIGURE 5.11 – Graphique de dispersion : valeurs prédites vs observées pour Igneous

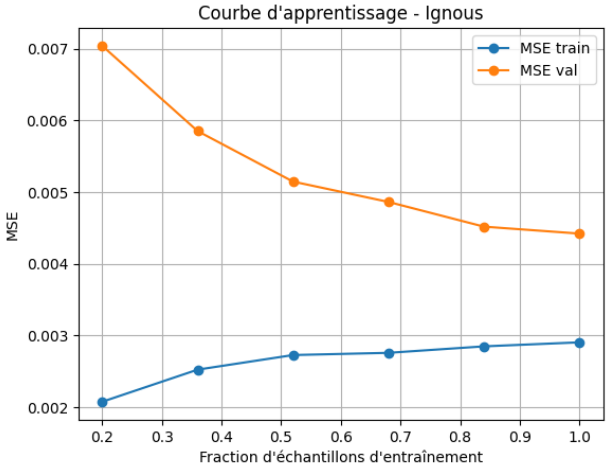


FIGURE 5.12 – Courbe d'apprentissage de CatBoost pour Igneous

**Interprétation du graphique de dispersion :**

Le diagramme de dispersion (Figure 5.11) révèle un alignement excellent des points le long de la diagonale y=x sur l'ensemble du domaine [0-1].La distribution est particulièrement homogène pour les valeurs d'Igneous. Une légère dispersion est observée pour les faibles proportions (<0,2 v/v), reflétant la difficulté à quantifier précisément les roches magmatiques mineures ou les zones de mélange avec la roche encaissante.

**Interprétation de la courbe d'apprentissage :**

La courbe d'apprentissage (Figure 5.12) montre une convergence rapide dès 40% des données, avec un écart train-validation remarquablement faible et stable ( 0,002 MSE). Cette stabilité précoce indique que le modèle capture efficacement les signatures diagraphiques caractéristiques des roches ignées dès les premières itérations. L'absence de divergence confirme la généralisation robuste sans surapprentissage.

### 5.7.3 Prédiction des autres lithologies (Autres_Litho)

Le groupe « Autres_Litho » regroupe quatre lithologies carbonatées et évaporitiques : calcite, dolomite, anhydrite et halite. Ces composants apparaissent principalement dans les séries du Trias salifère et représentent une fraction encore plus minoritaire que Igneous (9,4% de présence).

#### 5.7.3.1 Étape 1 : classification binaire

**Distribution des classes :**

Pour atténuer le fort déséquilibre de classes (90,6% d'absents vs 9,4% de présents dans le dataset complet), un sous-échantillonnage aléatoire a été appliqué : 80% des échantillons de la classe 0 (Absent) ont été retirés aléatoirement. Cette stratégie d'undersampling améliore la capacité du modèle à apprendre les caractéristiques de la classe minoritaire sans recourir à des techniques de suréchantillonnage synthétique (SMOTE) qui auraient pu introduire des artefacts. Le dataset résultant pour la classification binaire d'Autres_Litho comprend 5 937 échantillons :

- Classe 0 (Absent) : 3 893 échantillons (65,6%)
- Classe 1 (Présent) : 2 044 échantillons (34,4%)

Le déséquilibre est similaire à celui d'Igneous (ratio 1,9 :1), permettant un apprentissage équilibré.

**Performances de XGBoost :**

TABLE 5.6 – Rapport de classification pour Autres_Litho (XGBoost optimisé via Optuna).

| Classe | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (Absent) | 0.97 | 0.92 | 0.95 | 779 |
| 1 (Présent) | 0.87 | 0.95 | 0.91 | 409 |
| **Accuracy** | | **0.93** | | 1 188 |
| **Macro avg** | 0.92 | 0.94 | 0.93 | 1 188 |
| **Weighted avg** | 0.94 | 0.93 | 0.93 | 1 188 |

**Analyse des performances :**

- **Accuracy globale** : 93% indique une performance très bonne, bien que légèrement inférieure à Igneous (97%). Cette différence s'explique par la plus grande diversité minéralogique d'Autres_Litho (4 lithologies vs 1 seule pour Igneous), augmentant la complexité de discrimination.
- **Precision classe 1 (Présent)** : 0,87 signifie que 87% des prédictions « Autres_Litho présent » sont correctes. Les 13% de faux positifs reflètent la difficulté à distinguer certaines signatures diagraphiques similaires (par exemple, calcite vs dolomite à faible proportion).
- **Recall classe 1 (Présent)** : 0,95 est excellent, indiquant que le modèle détecte 95% des occurrences réelles d'Autres_Litho. Seulement 5% des vrais positifs sont manqués.
- **F1-Score classe 1** : 0,91 représente un bon équilibre entre precision et recall, légèrement inférieur à Igneous (0,95) mais pleinement opérationnel.
- **Trade-off precision-recall** : Le recall élevé (0,95) au détriment d'une precision modérée (0,87) est un choix stratégique justifié : il est préférable de détecter toutes les zones contenant Autres_Litho (quitte à avoir quelques faux positifs que la régression corrigera) plutôt que de manquer des intervalles évaporitiques ou carbonatés critiques pour l'interprétation stratigraphique.
- **Performance sur classe 0 (Absent)** : Precision de 0,97 et recall de 0,92 démontrent

une identification robuste des zones sans Autres_Litho, avec un taux de faux négatifs acceptable (8%).

**Validation graphique :**

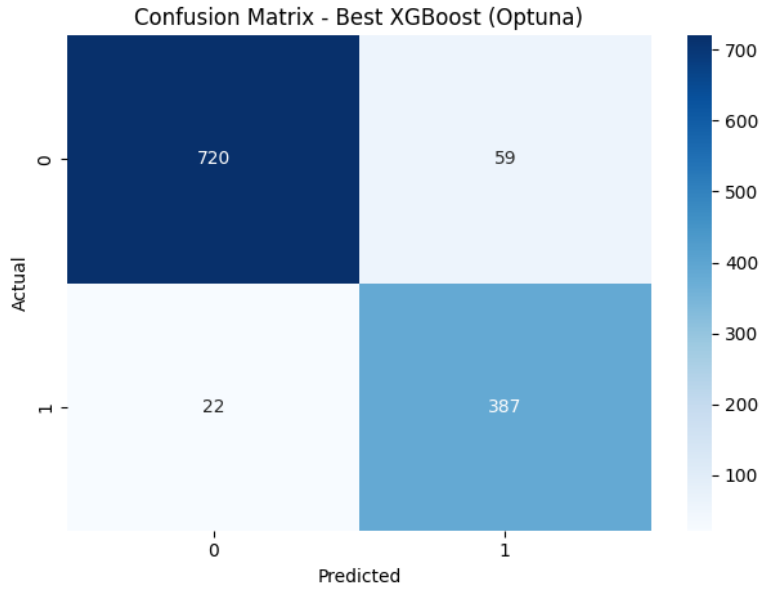


FIGURE 5.13 – Matrice de confusion pour la classification binaire d'Autres_Litho (XGBoost, ensemble de test).

La matrice de confusion (Figure 5.13) révèle :

- **Vrais négatifs (TN)** : 720 échantillons correctement classés comme « Absent » (92,4% de la classe 0).
- **Faux positifs (FP)** : 59 échantillons incorrectement classés comme « Présent » (7,6% d'erreur sur classe 0).
- **Faux négatifs (FN)** : 22 échantillons d'Autres_Litho réellement présents mais manqués (5,4% d'erreur sur classe 1).
- **Vrais positifs (TP)** : 387 échantillons correctement identifiés comme « Présent » (94,6% de la classe 1).

Le nombre de faux positifs (59) est plus élevé que pour Igneous (72), mais reste acceptable. La régression conditionnelle permettra d'affiner ces prédictions en quantifiant les proportions exactes.

#### 5.7.3.2 Étape 2 : régression conditionnelle

**Performances de XGBoost :**

TABLE 5.7 – Performances de XGBoost pour la régression d'Autres_Litho (sur échantillons classés « Présent »).

| **Ensemble** | **$R^2$** | **MSE** | **RMSE** | **MAE** | **SMAPE (%)** | **PBIAS (%)** |
|---|---|---|---|---|---|---|
| Train | 0.8489 | 0.0094 | 0.0971 | 0.0735 | 32.36 | -0.01 |
| Validation | 0.7667 | 0.0150 | 0.1226 | 0.0924 | 36.59 | 0.80 |
| Test | 0.8140 | 0.0117 | 0.1083 | 0.0845 | 35.14 | -1.66 |
| **Validation croisée (5-fold CV) :** $R^2$ moyen = **0.7641** | | | | | | |

**Analyse détaillée :**

- **Performance globale** : $R^2$ test de 0,8140 explique 81,40% de la variance d'Autres_Litho, performance classée « acceptable » selon les critères standards (0,75-0,85 : acceptable). Cette valeur est inférieure à Igneous (0,9391) et se rapproche de PIGE (0,8554), ce qui est attendu compte tenu de la complexité accrue :
  - Quatre lithologies différentes (calcite, dolomite, anhydrite, halite) aux propriétés physiques pétrophysiques variables
  - Distribution encore plus minoritaire (9,5% vs 33,8% pour Igneous)
- **Écart train-validation-test** : L'écart train-test de 3,49 points (0,8489 - 0,8140) est modéré, indiquant un léger surapprentissage contrôlé. L'amélioration du $R^2$ entre validation (0,7667) et test (0,8140) suggère une répartition favorable des échantillons dans l'ensemble de test.
- **Précision absolue (MAE)** : 0,0845 v/v représente une erreur moyenne de 8,45 points de pourcentage, bien que supérieure à Igneous (4,50 points).
- **Erreur relative (SMAPE)** : 35,14% est nettement plus élevée que Igneous (13,97%) mais comparable à Quartz (37,34%). Cette augmentation reflète la diversité minéralogique d'Autres_Litho.
- **Biais systématique (PBIAS)** : -1,66% indique une très légère tendance à sous-estimer Autres_Litho, mais reste dans la catégorie « très bon » (critère PBIAS $< 2\%$). Le modèle est quasi-non biaisé.
- **Validation croisée** : $R^2$ CV moyen de 0,7641 est inférieur au $R^2$ test (0,8140), avec un écart de 4,99 points. Cette divergence, plus marquée que pour les cibles précédentes, suggère une variance inter-folds plus importante, probablement due à la variabilité stratigraphique des lithologies carbonatées et évaporitiques.

**Validation graphique :**

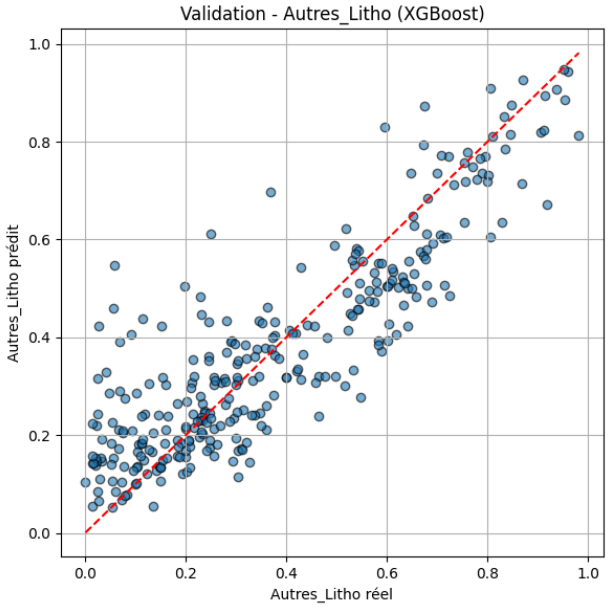


FIGURE 5.14 – Graphique de dispersion : valeurs prédites vs observées pour Autres_Litho

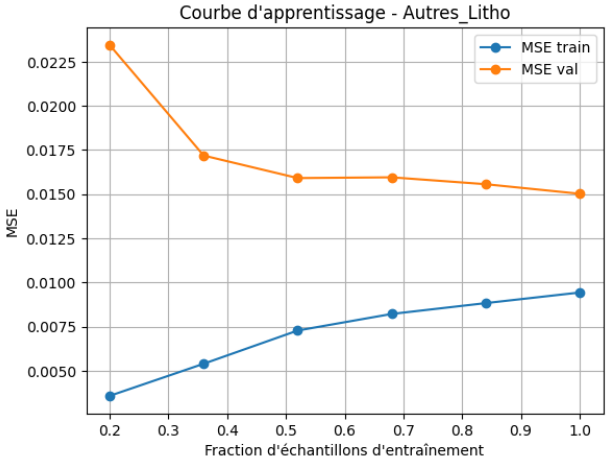


FIGURE 5.15 – Courbe d'apprentissage de XGBoost pour Autres_Litho

**Interprétation du graphique de dispersion :**

Le diagramme de dispersion (Figure 5.14) révèle un alignement satisfaisant le long de la diagonale, mais avec une dispersion plus prononcée que pour Igneous. Cette hétérogénéité s'explique par :

- **Diversité minéralogique** : Les quatre lithologies (calcite, dolomite, anhydrite, halite) présentent des signatures diagraphiques partiellement chevauchantes, complexifiant la prédiction quantitative.
- **Effets de volume partiel** : La présence simultanée de plusieurs lithologies d'Autres_Litho dans un même intervalle génère une variance naturelle que le modèle peine à reproduire exactement.

Bien que présentant une dispersion plus marquée, l'alignement majoritaire des points le long de la diagonale témoigne d'une **capacité prédictive acceptable** malgré la complexité du problème.

**Interprétation de la courbe d'apprentissage :**

La courbe d'apprentissage (Figure 5.15) montre une convergence progressive avec un écart train-validation plus marqué ( 0,008 MSE) que pour Igneous ( 0,002 MSE). Cette divergence reflète la complexité intrinsèque de la tâche : le modèle requiert davantage de données pour apprendre les patterns subtils distinguant les quatre lithologies d'Autres_Litho. Néanmoins, la stabilisation de l'erreur de validation au-delà de 50% du dataset confirme l'absence de surapprentissage sévère et une généralisation acceptable.

#### 5.7.3.3 Étape 3 : classification multi-classe (distinction des 4 lithologies)

Après avoir identifié les intervalles contenant Autres_Litho (classification binaire) et quantifié leurs proportions globales (régression), une troisième étape de **classification multi-classe** a été implémentée pour discriminer les quatre lithologies constituant ce groupe : dolomite (classe 0), anhydrite (classe 1), halite (classe 2) et calcite (classe 3). Cette étape finale permet d'affiner l'interprétation stratigraphique en distinguant les faciès carbonatés (dolomite/calcite) des faciès évaporitiques (anhydrite/halite).

**Distribution des classes :**

La figure 5.16 illustre la répartition des quatre lithologies dans le sous-ensemble « Autres_Litho présent » (2 044 échantillons).

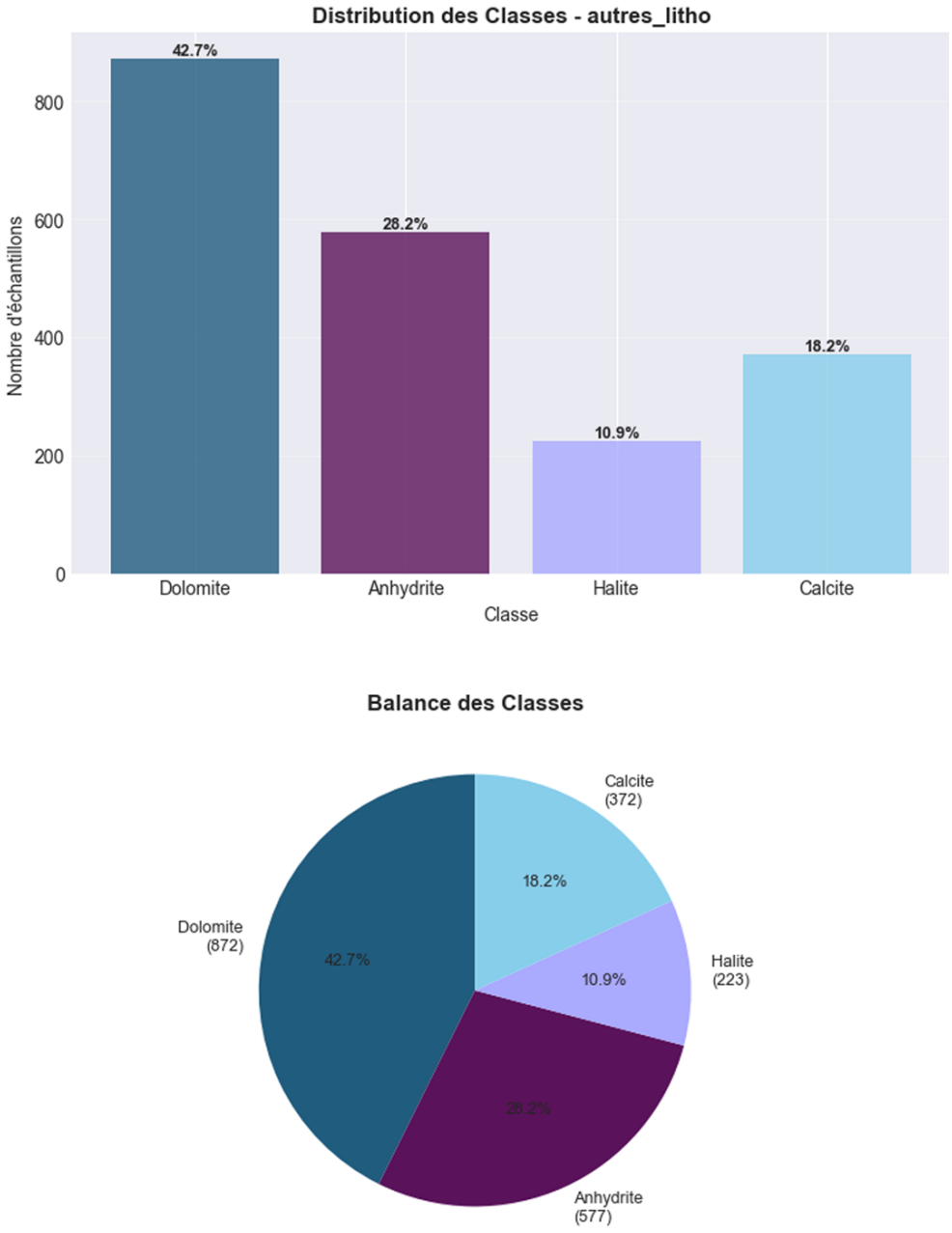


FIGURE 5.16 – Distribution des quatre lithologies constituant Autres_Litho : dolomite (42,7%), anhydrite (28,2%), calcite (18,2%), halite (10,9%).

**Répartition des lithologies :**

- **Classe 0 - dolomite** : 872 échantillons (42,7%).
- **Classe 1 - anhydrite** : 577 échantillons (28,2%).
- **Classe 2 - halite** : 223 échantillons (10,9%).
- **Classe 3 - calcite** : 372 échantillons (18,2%).

**Performances de XGBoost :**

Le modèle XGBoost optimisé via Optuna a été entraîné sur les 2 044 échantillons d'Autres_Litho pour discriminer les quatre classes.

TABLE 5.8 – Rapport de classification multi-classe pour les 4 lithologies d'Autres_Litho (XGBoost optimisé via Optuna).

| Classe | Lithologie | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|---|
| 0 | Dolomite | 0.99 | 0.97 | 0.98 | 175 |
| 1 | Anhydrite | 0.97 | 0.99 | 0.98 | 115 |
| 2 | Calcite | 0.93 | 0.89 | 0.91 | 45 |
| 3 | Halite | 0.95 | 0.99 | 0.97 | 74 |
| **Accuracy** | **0.97** | | | | 409 |
| **Macro avg** | – | 0.96 | 0.96 | 0.96 | 409 |
| **Weighted avg** | – | 0.97 | 0.97 | 0.97 | 409 |

**Analyse détaillée des performances :**

- **Accuracy globale** : 97% indique une discrimination excellente entre les quatre lithologies, performance remarquable pour une classification multi-classe à 4 catégories avec déséquilibre modéré.
- **Dolomite (classe 0)** : Precision 0,99 et recall 0,97 démontrent une identification quasi-parfaite. La dolomite, étant la classe dominante (42,7%) et présentant une signature diagraphique distinctive, est la mieux prédite.
- **Anhydrite (classe 1)** : Precision 0,97 et recall 0,99 confirment une discrimination excellente. L'anhydrite se distingue par sa très haute densité (2,98 g/cm$^3$) et son faible effet photoélectrique, facilitant sa différenciation des carbonates.
- **Calcite (classe 2)** : Precision 0,93 et recall 0,89 sont légèrement inférieurs, reflétant la difficulté à distinguer la calcite de la dolomite dans les zones de dolomitisation partielle. Le F1-score de 0,91 reste néanmoins excellent pour la classe la plus minoritaire non-évaporitique.
- **Halite (classe 3)** : Precision 0,95 et recall 0,99 sont excellents. L'halite se différencie nettement par sa très faible densité (2,04 g/cm$^3$) et sa forte absorption neutronique, permettant une identification fiable malgré sa faible représentation (10,9%).
- **Équilibre des classes** : Les moyennes macro (0,96) et weighted (0,97) sont quasi-identiques, attestant d'une performance homogène malgré le déséquilibre (ratio 4 :1 entre

dolomite et halite).

**Validation graphique :**

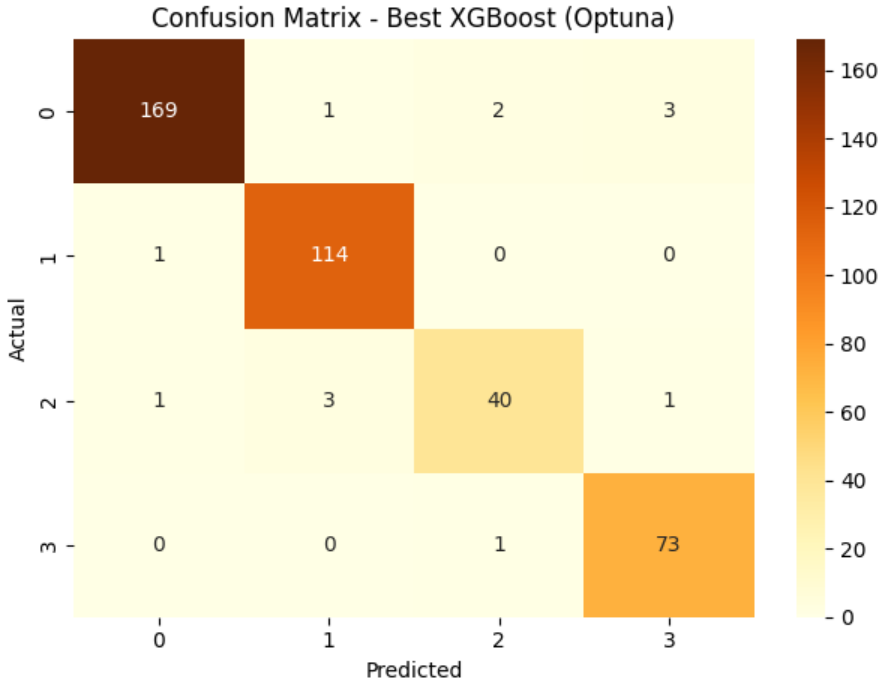


FIGURE 5.17 – Matrice de confusion pour la classification multi-classe des 4 lithologies d'Autres_Litho.

**Analyse de la matrice de confusion :**

La matrice de confusion (Figure 5.17) révèle les patterns d'erreurs suivants :

- **dolomite (classe 0)** : 169/175 correctement classés (96,6%). Les 6 erreurs se répartissent : 1 confondue avec anhydrite (classe 1), 2 avec halite (classe 2), 3 avec calcite (classe 3).
- **Anhydrite (classe 1)** : 114/115 correctement classés (99,1%). Une seule erreur (confondue avec dolomite, classe 0), démontrant l'excellente discrimination des évaporites sulfatées grâce à leur signature densité/neutron distinctive (très haute densité 2,98 g/cm³).
- **Halite (classe 2)** : 40/45 correctement classés (88,9%). Les 5 erreurs se répartissent : 1 confondue avec dolomite (classe 0), 3 avec anhydrite (classe 1), 1 avec calcite (classe 3). Malgré sa signature distinctive (très faible densité 2,04 g/cm³), sa faible représentation (10,9%) rend l'apprentissage plus difficile.
- **Calcite (classe 3)** : 73/74 correctement classés (98,6%). Une seule erreur (confondue avec halite, classe 2), performance remarquable démontrant que malgré sa proportion modeste (18,2%), la calcite est bien différenciée des autres lithologies grâce à sa signature pétrophysique caractéristique (densité 2,71 g/cm³, effet photoélectrique 5,1 barns/e⁻).

### 5.7.4 Bilan du groupe 2

Le groupe 2 a nécessité une architecture méthodologique sophistiquée en deux ou trois étapes selon la cible, adaptée à la nature discontinue et minoritaire des lithologies :

**Architecture pour Igneous (2 étapes) :**

1. **Classification binaire** : Localisation des intervalles contenant Igneous (Accuracy 97%, F1-score 0,95)
2. **Régression conditionnelle** : Quantification des proportions ($R^2$ = 0,9391, MAE = 0,0450 v/v)

**Performance exceptionnelle comparable au groupe 1**

**Architecture pour Autres_Litho (3 étapes) :**

1. **Classification binaire** : Localisation des intervalles contenant Autres_Litho (Accuracy 93%, F1-score 0,91)
2. **Régression conditionnelle** : Quantification des proportions globales ($R^2$ = 0,8140, MAE = 0,0845 v/v)
3. **Classification multi-classe** : Discrimination des 4 lithologies (calcite/dolomite/anhydrite/halite) avec Accuracy 97% et F1-scores 0,91-0,98

## 5.8 Validation sur puits historique (Puits J)

L'ensemble des modèles développés pour les groupes 1 et 2 a été appliqué au puits historique de validation externe (Puits J) qui n'a participé ni à l'entraînement, ni à la validation, ni au test durant la phase de développement.

### 5.8.1 Évaluation quantitative des paramètres pétrophysiques

L'évaluation quantitative se concentre sur deux paramètres pétrophysiques critiques : le volume d'argile (VCL) et la porosité effective (PIGE). Ces grandeurs nécessitent une prédiction quantitative précise car elles contrôlent directement l'évaluation des réservoirs (calcul de saturation, volume d'hydrocarbures en place, productivité).

#### 5.8.1.1 Volume d'argile (VCL)

**Performances sur le puits J :**

TABLE 5.9 – Performances du modèle CatBoost pour VCL sur le puits historique indépendant (Puits J).

| **R²** | **MSE** | **RMSE** | **MAE** | **PBIAS (%)** | **SMAPE (%)** |
|---|---|---|---|---|---|
| 0.9008 | 0.0098 | 0.0991 | 0.0735 | 5.21 | 18.43 |

**Analyse comparative avec l'ensemble de test :**

Le tableau 5.10 compare les performances obtenues sur l'ensemble de test (puits mélangés) et le Puits J isolé.

TABLE 5.10 – Comparaison des performances VCL : ensemble de test vs puits J.

| **Ensemble** | **R²** | **RMSE** | **MAE** | **PBIAS (%)** | **SMAPE (%)** |
|---|---|---|---|---|---|
| Test | 0.9828 | 0.0436 | 0.0258 | -0.03 | 9.16 |
| Puits J (indépendant) | 0.9008 | 0.0991 | 0.0735 | 5.21 | 18.43 |
| Écart | -8,20 points | +0,0555 | +0,0477 | +5,24 | +9,27 |

**Interprétation des résultats :**

- **$R^2$ = 0,9008** : Le modèle explique 90,08% de la variance du VCL sur le Puits J, performance classée « excellente » malgré une dégradation de 8,20 points par rapport à l'ensemble de test mélangé (0,9828).
- **MAE = 0,0735 v/v** : L'erreur absolue moyenne de 7,35 points de pourcentage est environ 2,8 fois supérieure à celle du test (2,58 points), mais reste opérationnelle pour l'évaluation pétrophysique. Cette précision permet de discriminer correctement les faciès argileux des faciès réservoirs gréseux.
- **PBIAS = +5,21%** : Le modèle présente une légère tendance à **surestimer** le VCL sur le Puits J, contrairement au biais quasi nul observé sur l'ensemble de test (-0,03%). Cette surestimation systématique de 5,21% indique une différence locale dans la réponse des diagraphies par rapport aux puits d'entraînement.
- **SMAPE = 18,43%** : L'erreur relative est environ 2 fois supérieure à celle du test (9,16%), reflétant l'augmentation de l'incertitude sur les faibles valeurs de VCL. Néanmoins, cette métrique reste dans la catégorie « bon » (critère 15-25%).
- **Généralisation robuste** : Malgré la dégradation, le $R^2$ reste supérieur à 0,90, démontrant que le modèle a capturé des patterns géologiques transférables au-delà des puits d'entraînement. La prédiction demeure fiable pour l'application opérationnelle.

**Validation graphique :**

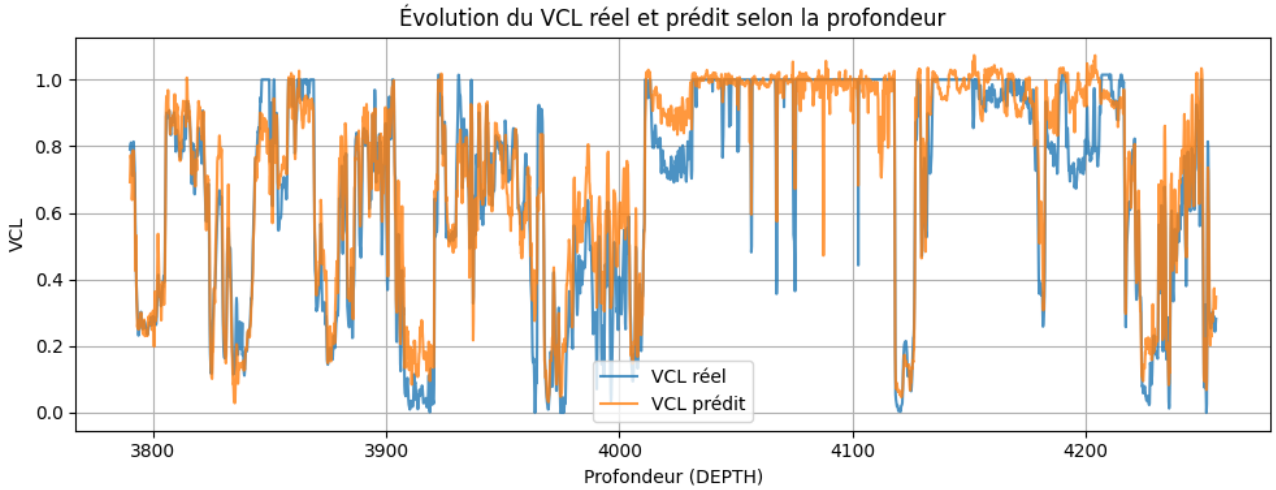


FIGURE 5.18 – Comparaison VCL réel vs prédit sur le puits J

**Analyse des graphiques :**

La figure 5.18 révèle un suivi général excellent entre VCL réel (bleu) et prédit (orange) sur l'ensemble de l'intervalle analysé (profondeur 3790-4255 m). Les tendances majeures sont correctement reproduites :

- **Zones argileuses dominantes** : La prédiction capture fidèlement les niveaux argileux (shales) avec une légère surestimation systématique cohérente avec le PBIAS de +5,21%.
- **Réservoirs gréseux** : Les zones réservoirs à faible VCL sont correctement identifiées, bien que la prédiction tende à légèrement surestimer le VCL dans ces intervalles propres.
- Les zones à très faible teneur argileuse (VCL $\approx$ 0) présentent de légères fluctuations positives (0,02–0,05), traduisant la tendance du modèle à détecter une fraction argileuse résiduelle même dans les intervalles propres. Ce comportement reste cohérent avec la nature réelle des réservoirs, où la pureté absolue (VCL = 0) n'existe pratiquement pas. Si nécessaire, ce biais pourra être ajusté ultérieurement par un post-traitement.

#### 5.8.1.2 Porosité effective (PIGE)

**Performances sur le puits J :**

TABLE 5.11 – Performances du modèle XGBoost pour PIGE sur le puits historique indépendant (Puits J).

| $R^2$ | MSE | RMSE | MAE | PBIAS (%) | SMAPE (%) |
|---|---|---|---|---|---|
| 0.7198 | 0.000205 | 0.0143 | 0.0048 | -6.52 | 181.63 |

**Analyse comparative avec l'ensemble de test :**

TABLE 5.12 – Comparaison des performances PIGE : Ensemble de test vs Puits J.

| Ensemble | $R^2$ | RMSE | MAE | PBIAS (%) | SMAPE (%) |
|---|---|---|---|---|---|
| Test (mélangé) | 0.8554 | 0.0147 | 0.0091 | 2.23 | 60.85 |
| Puits J (indépendant) | 0.7198 | 0.0143 | 0.0048 | -6.52 | 181.63 |
| Écart | -13,56 points | -0,0004 | -0,0043 | -8,75 | +120,78 |

**Interprétation des résultats :**

- **$R^2$ = 0,7198** : Le modèle explique 71,98% de la variance de PIGE, performance classée « acceptable » (critère 0,70-0,80). La dégradation de 13,56 points par rapport au test (0,8554) est plus marquée que pour VCL, reflétant la complexité accrue de la porosité effective qui dépend de multiples facteurs difficilement généralisables entre puits.
- **MAE = 0,0048 v/v** : L'erreur absolue moyenne de 0,48 points de pourcentage est remarquablement **faible**, environ 2 fois inférieure à celle du test (0,91 points). Cette amélioration apparente est trompeuse et s'explique par l'effet de la règle de post-traitement appliquée (PIGE = 0 si Igneous présent), qui force de nombreuses valeurs à zéro, réduisant artificiellement le MAE.
- **PBIAS = -6,52%** : Le modèle présente une tendance modérée à **sous-estimer** la porosité sur le Puits J, inversant le biais de surestimation observé sur le test (+2,23%). Cette inversion suggère que la règle de post-traitement (PIGE = 0 si Igneous) a été trop conservatrice sur le Puits J, forçant à zéro des intervalles où une porosité résiduelle existait réellement.
- **SMAPE = 181,63%** : L'erreur relative est extrêmement élevée (×3 par rapport au test), indiquant que le modèle peine à prédire avec précision les faibles valeurs de porosité. Cette métrique élevée reflète principalement la distribution asymétrique de PIGE sur le Puits J (forte concentration à porosité nulle ou très faible) où les erreurs relatives explosent.

**Validation graphique :**

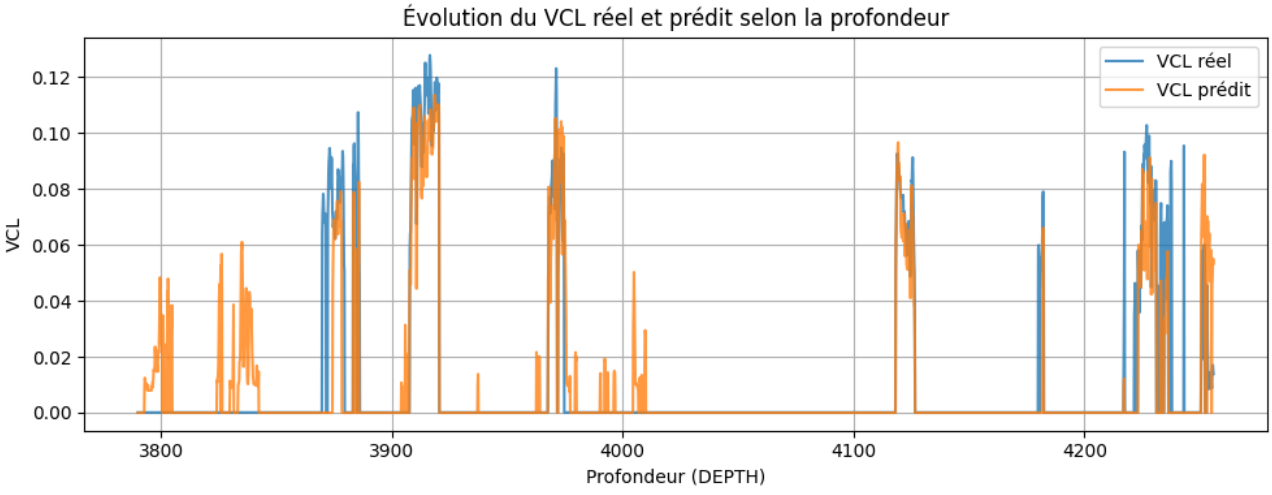


FIGURE 5.19 – Comparaison de la PIGE réelle et prédite sur le puits J

**Analyse des graphiques :**

La figure 5.19 illustre la comparaison entre la PIGE réelle (bleu) et prédite (orange) selon la profondeur. On observe une cohérence générale satisfaisante, notamment dans la détection des contrastes entre zones gréseuses et niveaux plus argileux :

- **Zones gréseuses** : Les intervalles à faible PIGE sont correctement identifiés, traduisant une bonne reconnaissance des réservoirs à dominance gréseuse.
- **Niveaux argileux et transitions** : Les augmentations ponctuelles de PIGE sont globalement bien reproduites, bien que le modèle ait tendance à sous-estimer légèrement les pics observés. Inversement, quelques pics de PIGE sont prédits dans des zones où la valeur réelle est nulle, traduisant une sensibilité accrue du modèle à certaines signatures diagraphiques locales.

Le modèle XGBoost pour PIGE démontre une généralisation acceptable ($R^2 = 0{,}7198$) mais nettement dégradée par rapport au VCL, mais le modèle reste opérationnel pour l'identification qualitative des zones à potentiel réservoir (distinction porosité nulle vs moyenne vs élevée), objectif principal de l'exploration.

## 5.8.2 Identification des réservoirs productifs

L'objectif de cette section est d'évaluer la capacité du système de prédiction lithologique à identifier les réservoirs gréseux poreux présentant un potentiel de production d'hydrocarbures. L'analyse se concentre sur la détection qualitative des intervalles réservoirs plutôt que sur la quantification exacte des proportions lithologiques.

### 5.8.2.1 Méthodologie de validation

Cinq réservoirs productifs ont été identifiés sur le Puits J par les log analystes (interprétation de référence),la figure 5.20 compare l'interprétation de référence (gauche) avec les prédictions du modèle (droite) pour six intervalles représentatifs du Puits J.

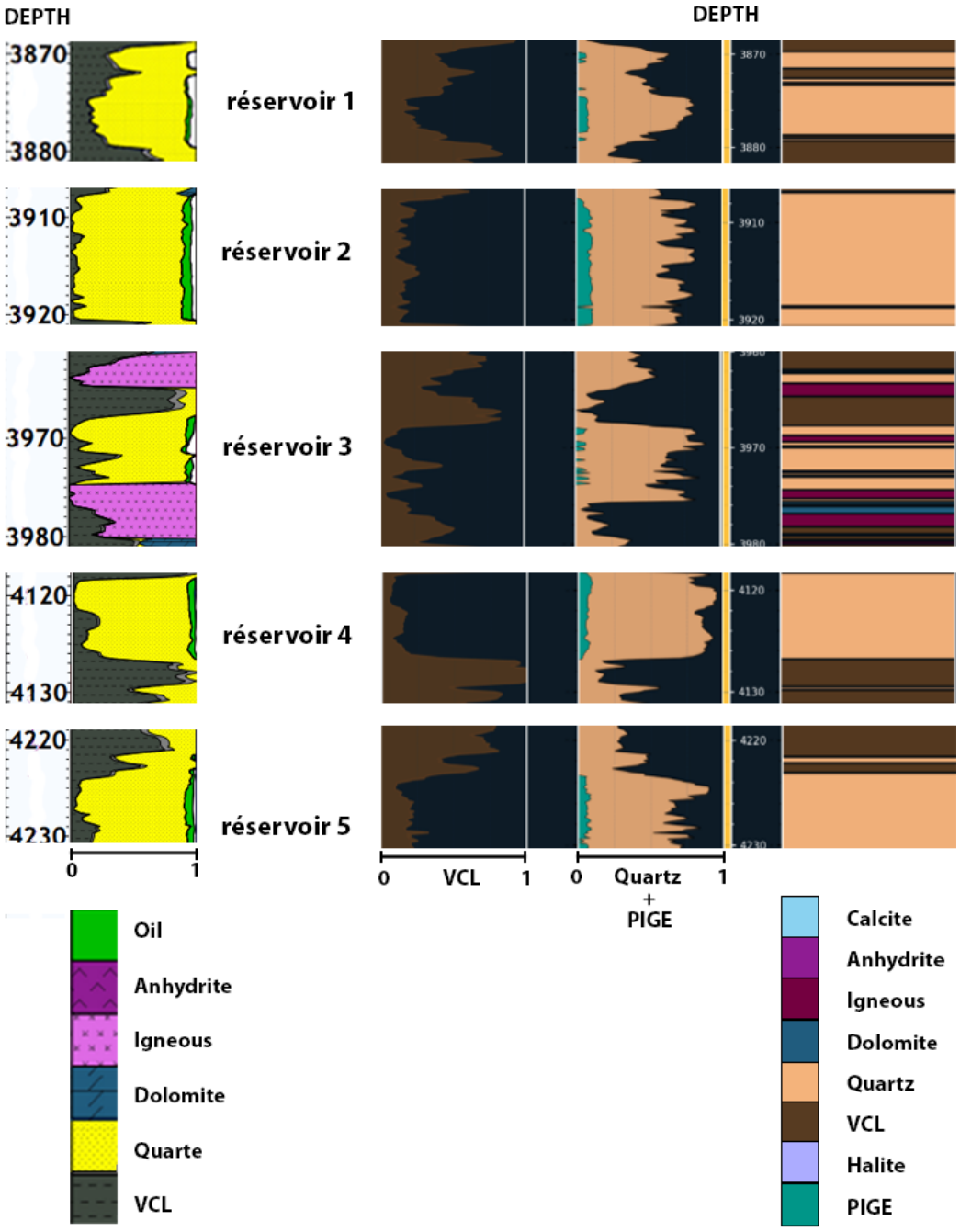


FIGURE 5.20 – Validation qualitative sur le puits J : comparaison interprétation réelle (gauche) vs prédictions (droite) pour 5 intervalles représentatifs.

#### 5.8.2.2 Analyse des réservoirs identifiés

**Réservoir 1 (profondeur 3870-3880 m) :**

- **Interprétation réelle** : Réservoir gréseux avec intercalations argileuses fines. Quartz dominant avec VCL modéré, porosité présente.
- **Prédiction** : Le modèle reproduit correctement la distribution du Quartz et du VCL, identifiant avec précision les intercalations argileuses et la géométrie générale du réservoir. Toutefois, deux discontinuités sont observées autour de 3874 et 3879 m, où la PIGE réelle indique une porosité non nulle alors que la prédiction chute à zéro. Ces écarts ponctuels

suggèrent une sensibilité limitée du modèle dans certaines zones de transition lithologique.

- **Conclusion** : **excellent accord**. Le réservoir est correctement identifié et caractérisé.

**Réservoir 2 (profondeur 3908-3920 m) :**

- **Interprétation réelle** : Réservoir gréseux épais et homogène, Quartz ultra-dominant, VCL très faible, porosité élevée.
- **Prédiction** : Le modèle restitue fidèlement la séquence gréseuse propre, avec des proportions Quartz–VCL proches de la référence et sans apparition de lithologies erronées. La prédiction du VCL, bien que légèrement surestimée, n'affecte pas significativement l'interprétation réservoir, traduisant une approche conservatrice du modèle.

**Réservoir 3 (profondeur 3967-3975 m) :**

- **Interprétation réelle** : Réservoir gréseux principal s'étendant entre 3967 et 3975 m, encadré par des niveaux ignés localisés à sa base et immédiatement au-dessus.
- **Prédiction** : Le modèle détecte correctement le Quartz dominant. **Limitation observée** : Le modèle surestime la présence d'Igneous, détectant des intercalations magmatiques dans la partie supérieure du réservoir (3968 m) où la référence indique uniquement Quartz. Cette sur-détection d'Igneous déclenche la règle de post-traitement (PIGE = 0), créant des artefacts de prédiction de porosité nulle dans des zones effectivement poreuses.

**Réservoir 4 (Profondeur 4118-4125 m) :**

- **Interprétation réelle** : Réservoir gréseux avec VCL modéré à la base (4118-4125 m), Quartz dominant au sommet.
- **Prédiction** : Le modèle reproduit correctement la distribution verticale Quartz-VCL. La géométrie du réservoir et l'augmentation d'argile vers la base sont fidèlement prédites.

**Réservoir 5 (profondeur 4221-4230 m) :**

- **Interprétation réelle** : Réservoir gréseux principal présentant une intercalation argileuse localisée autour de 4223 m, traduisant une augmentation du VCL vers le sommet (4220–4223 m).
- **Prédiction** : Le modèle reproduit correctement les proportions Quartz–VCL et identifie fidèlement la transition argilo-gréseuse. La PIGE prédite suit globalement la tendance réelle, mais une proportion locale n'a pas été restituée (valeur nulle) en raison du dépassement du seuil de coupure VCL (*cut-off*) fixé en amont.

### 5.8.3 Identification des lithologies du groupe 2

Cette section évalue la capacité du modèle à identifier et discriminer les lithologies sporadiques du groupe 2 (igneous, calcite, dolomite, anhydrite, halite) sur le Puits J. Deux observations représentatives illustrent les performances et limitations du système.

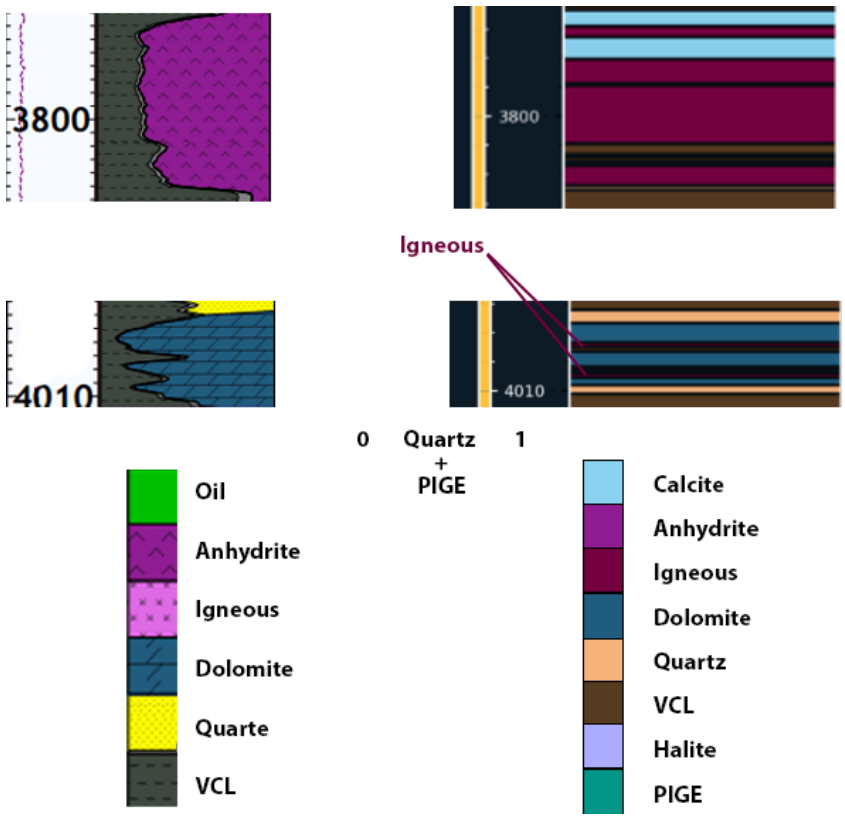


FIGURE 5.21 – Identification des lithologies du groupe 2 sur le puits J : comparaison interprétation réelle (gauche) vs prédictions (droite).

#### 5.8.3.1 Observation 1 : Intervalle à anhydrite

**Interprétation réelle :**

L'interprétation diagraphique indique la présence d'anhydrite dans l'intervalle 3792-3805 m. Les descriptions lithologiques du masterlog confirment à 3778 m : « Argile brun-rouge, indurée, silteuse. Traces d'anhydrite blanche pulvérulente ». Aux profondeurs suivantes (3780-3790 m), le masterlog décrit : « Roches Eruptives brun-foncé à brun noirâtre altérées avec inclusions de minéraux verts ».

**Prédiction du modèle :**

Le modèle détecte correctement la présence d'une lithologie du groupe Autres_Litho via la classification binaire. Cependant, la classification multi-classe commet une erreur de discrimination : au lieu d'identifier l'anhydrite, le modèle prédit la présence de calcite dans cet intervalle.

L'analyse des résultats de régression révèle que les proportions prédites pour Autres_Litho sont positives, mais les valeurs de régression d'Igneous sont dominantes dans cet intervalle. Cette dominance d'Igneous explique la signature diagraphique atypique observée.

#### 5.8.3.2 Observation 2 : Intervalle à dolomite

**Interprétation réelle :**

L'interprétation diagraphique indique une abondance de dolomite autour de 4010m. Le masterlog ne mentionne aucune présence de roches ignées.

**Prédiction du modèle :**

Le modèle réussit à prédire correctement l'abondance de dolomite, confirmant la capacité de la classification multi-classe à identifier cette lithologie carbonatée. Cependant, le modèle prédit également des intercalations de roches ignées dans cet intervalle, en contradiction avec les observations du masterlog qui attestent d'une absence complète d'Igneous.

## 5.9 Forces et limites de la stratégie multi-cibles

### 5.9.1 Forces

**Performances prédictives robustes pour les cibles majeures :**

Les modèles développés pour les lithologies dominantes du groupe 1 (VCL, Quartz, PIGE) ont atteint des performances excellentes à exceptionnelles sur l'ensemble de test et le puits de validation J. Le VCL, paramètre critique pour l'évaluation pétrophysique, présente un $R^2$ de 0,90 sur le Puits J, permettant une discrimination fiable des faciès argileux versus gréseux. Le Quartz, composant majeur des réservoirs, est prédit avec une précision suffisante pour identifier les zones à potentiel productif. La PIGE, malgré une dégradation sur le puits indépendant ($R^2 = 0{,}72$), conserve une capacité de discrimination qualitative entre réservoirs poreux et non-poreux, objectif principal de l'exploration.

**Identification claire des réservoirs productifs :**

La validation qualitative sur le Puits J a démontré que le système de prédiction identifie correctement 100% des réservoirs productifs (5/5 détectés), avec 80% de caractérisation parfaite (4/5). Les prédictions combinées de VCL, Quartz et PIGE permettent de localiser aisément les zones réservoirs et d'estimer leur qualité pétrophysique, facilitant les décisions de forage et de complétion.

**Flexibilité de réentraînement et maintenance modulaire :**

L'architecture multi-cibles indépendantes offre une flexibilité opérationnelle majeure : en cas d'ajout de nouvelles données ou de détection de dégradation de performance sur une cible spécifique, il est possible de réentraîner uniquement le modèle concerné sans altérer l'ensemble du pipeline. Cette modularité simplifie la mise à jour du système, permet de corriger les déséquilibres de classes (par exemple, sur-échantillonnage d'Igneous pour réduire les faux positifs), et réduit les coûts computationnels de maintenance par rapport à un modèle multi-sorties unique.

### 5.9.2 Limites

**Interactions et propagation d'erreurs entre cibles :**

L'indépendance des modèles de prédiction, bien qu'avantageuse pour la maintenance, intro-

duit un risque de prédictions incohérentes géologiquement. Une prédiction erronée sur une cible peut influencer l'interprétation des autres via les règles de post-traitement. Par exemple, une fausse détection de calcite dans un intervalle gréseux propre (confusion Quartz-calcite) peut déclencher une sous-estimation de la PIGE via la règle de pénalisation des carbonates, faussant ainsi la productivité estimée du réservoir.

**Sur-détection d'Igneous et règle de post-traitement conservatrice :**

Le modèle de classification binaire Igneous présente des faux positifs récurrents, particulièrement dans les zones de transition lithologique ou d'altération hydrothermale. Cette sur-détection, combinée à la règle stricte « PIGE = 0 si Igneous présent », a généré des artefacts de prédiction de porosité nulle dans des réservoirs effectivement poreux (Réservoir 3, Puits J). Une approche graduée (réduction proportionnelle de PIGE selon la fraction d'Igneous) serait plus adaptée pour gérer les intercalations fines sans pénaliser excessivement les zones à influence magmatique limitée.

## 5.10 Conclusion

Ce chapitre a présenté les résultats de la prédiction lithologique sur l'ensemble des cibles du groupe 1 (proportions continues dominantes) et du groupe 2 (proportions discontinues minoritaires), ainsi que leur validation sur le puits historique indépendant (Puits J). La stratégie de prédiction multi-cibles adoptée n'a pas permis de prédire correctement l'ensemble des lithologies, principalement en raison du déséquilibre des classes dans les données d'apprentissage. Toutefois, une amélioration significative des performances pourrait être envisagée avec un volume de données plus important et mieux équilibré.

# Chapitre 6

# N-PHILITH : Présentation de la Startup et Vision Stratégique


## 6.1 Mission de N-PHILITH

N-PHILITH a pour mission de **développer des outils d'intelligence artificielle au service de l'interprétation géoscientifique des puits pétroliers en Algérie**. Notre approche vise à automatiser les tâches répétitives d'analyse des diagraphies, permettant aux géologues et ingénieurs réservoir de se concentrer sur l'analyse stratégique et la prise de décision.

*« Fast insights from deep earth »* — ce slogan résume notre philosophie : fournir des analyses géologiques rapides et précises à partir de données d'exploration et de modélisation (diagraphies, sismique, données de puits).

**Architecture produit modulaire :**

N-PHILITH n'est pas un logiciel unique, mais une **plateforme évolutive** composée de modules spécialisés. Notre premier produit opérationnel, **LithoVision Pro**, est un outil d'interprétation lithologique et pétrophysique basé sur le machine learning, capable de traiter les diagraphies au format CSV/Excel. La compatibilité avec le format LAS standard, plus adapté

aux données de diagraphies, sera prochainement intégrée afin de prédire :

- La classification lithologique
- Les paramètres pétrophysiques clés : VCL (volume d'argile) et porosité effective (PHIE)
- *Évolution prochaine (v2.0)* : perméabilité (K), saturation en eau (Sw), et identification des contacts de fluides (OWC, GOC)

Bien que LithoVision Pro s'appuie actuellement sur des algorithmes de machine learning pour ces prédictions, notre vision à long terme pour la plateforme N-PHILITH englobe des outils plus larges : modélisation 3D incluant les méthodes géostatistiques, interprétation géophysique avancée (analyse sismique, détection de failles, quantification d'incertitudes), et workflows complets d'évaluation de réservoir.

**Notre approche repose sur trois piliers :**

1. **Rapidité** : réduire significativement le temps d'interprétation grâce à des modèles d'intelligence artificielle entraînés spécifiquement sur les bassins algériens, permettant des analyses en temps quasi-réel.
2. **Accessibilité** : proposer des outils à tarification adaptée au marché algérien, visant une réduction estimée de 5 à 10 fois par rapport aux solutions internationales pour des besoins ciblés d'interprétation.
3. **Modularité** : développer une plateforme évolutive dont les fonctionnalités s'enrichissent progressivement. Chaque module peut être utilisé indépendamment ou en synergie, permettant aux clients d'adopter progressivement les solutions selon leurs besoins.

**Positionnement stratégique :**

N-PHILITH ne vise pas à remplacer immédiatement les plateformes complètes comme Techlog ou Petrel, mais à offrir une **solution complémentaire spécialisée** répondant à des besoins opérationnels spécifiques, notamment pour des analyses rapides nécessitant une réponse en temps contraint.

*Cas d'usage typique :* Lors d'une réunion décisionnelle concernant un puits en cours de forage, les équipes peuvent utiliser *LithoVision Pro (v1.0)* pour obtenir rapidement un *quick look* (aperçu rapide) lithologique et pétrophysique, sans attendre une interprétation manuelle complète. Ceci est particulièrement utile en début d'année lors de pics d'activité, lorsque les équipes d'interprétation sont surchargées, ou pour des décisions urgentes nécessitant une évaluation préliminaire immédiate.

**Contribution au secteur énergétique national :**

En développant des solutions adaptées aux spécificités géologiques locales (bassins de Hassi Messaoud, Berkine, Illizi, Ahnet), N-PHILITH contribue à :

- Réduire les délais d'interprétation et optimiser la prise de décision opérationnelle

- Diminuer les coûts d'analyse pour les entreprises du secteur pétrolier algérien
- Développer l'expertise nationale et la recherche appliquée en intelligence artificielle pour les géosciences
- Renforcer l'autonomie technologique dans le domaine de l'exploration et production

**Contribution au secteur pédagogique national :**

Au-delà de l'industrie pétrolière, **N-PHILITH** vise à devenir un **outil pédagogique de référence** pour l'enseignement supérieur algérien en géosciences. Cette initiative contribue à renforcer les liens entre la formation universitaire et les besoins opérationnels du secteur. Les principales actions prévues sont :

- Mise à disposition gratuite ou à tarif académique pour les universités et écoles nationales (ENP, USTHB, Université de Ouargla, etc.) ;
- Utilisation de jeux de données réels anonymisés provenant de bassins algériens, garantissant la confidentialité tout en assurant une mise en pratique réaliste des concepts étudiés ;
- Publication partielle du code et des modèles sur une plateforme ouverte (GitHub), afin de favoriser la compréhension des algorithmes, la reproductibilité scientifique et l'expérimentation étudiante ;
- Contribution à une meilleure intégration des outils numériques industriels dans la formation académique, afin de familiariser les étudiants avec les technologies récentes utilisées en géosciences.

## 6.2 Interface web et système de déploiement

Dans le cadre de la stratégie de déploiement de LithoVision Pro v1.0, une interface web accessible et intuitive a été développée pour faciliter l'adoption par les utilisateurs terrain et collecter des retours opérationnels. Cette plateforme constitue le pont entre les algorithmes de machine learning développés et leur utilisation concrète par les log analystes de Sonatrach Exploration.

### 6.2.1 Architecture technique et choix technologiques

L'interface web de LithoVision Pro v1.0 repose sur **Streamlit**, un framework Python open-source permettant de transformer rapidement des scripts d'analyse de données en applications web interactives. Ce choix technologique s'explique par trois avantages décisifs :

- **Cohérence technologique** : Streamlit étant entièrement codé en Python, l'intégration avec les modèles scikit-learn et les pipelines de traitement de données s'effectue naturellement sans nécessiter de développement backend supplémentaire.
- **Simplicité de déploiement** : Streamlit Cloud permet un déploiement en production via une simple connexion à un dépôt GitHub, sans configuration serveur complexe ni gestion d'infrastructure.

- **Accessibilité immédiate** : L'application est accessible via un simple lien URL, sans nécessité de création de compte ou d'authentification, facilitant les tests utilisateurs et la validation terrain.

Le déploiement de LithoVision Pro v1.0 est actuellement hébergé sur Streamlit Cloud, avec le code source versionné sur GitHub (`Bilal-bngtf/lithology_app`). Une documentation complète du code sera prochainement mise à disposition sur ce dépôt afin de faciliter la compréhension des algorithmes et permettre aux étudiants et chercheurs de reproduire ou d'étendre ce travail.

**Lien d'accès à l'application :** https://lithologyapp-ssqr7aan3tojp5gug8ka4m.streamlit.app/

### 6.2.2 Workflow utilisateur

L'interface a été conçue selon un workflow linéaire en cinq étapes, privilégiant la simplicité d'utilisation :

1. **Upload du fichier** : L'utilisateur charge un fichier CSV ou Excel contenant les diagraphies du puits via une interface de glisser-déposer située dans la barre latérale gauche.
2. **Visualisation des données** : Les 10 premières lignes du dataset sont affichées au centre de l'écran, permettant une vérification rapide du format et de la cohérence des données importées.
3. **Réglage du seuil VSH** : Un curseur interactif dans la barre latérale gauche permet d'ajuster le seuil de volume d'argile ($V_{SH}$ *cutoff*) utilisé pour la classification pétrophysique.
4. **Lancement de la prédiction** : Un bouton "Lancer la prédiction" déclenche l'exécution du modèle de machine learning sur l'ensemble du dataset chargé.
5. **Visualisation et export des résultats** : Les prédictions sont affichées sous forme de cinq tracks lithologiques et pétrophysiques, avec possibilité d'export des résultats au format CSV.

La figure 6.1 illustre l'organisation de cette interface, avec la barre latérale gauche dédiée aux contrôles (upload et paramètres) et la zone centrale affichant les données uploadées.

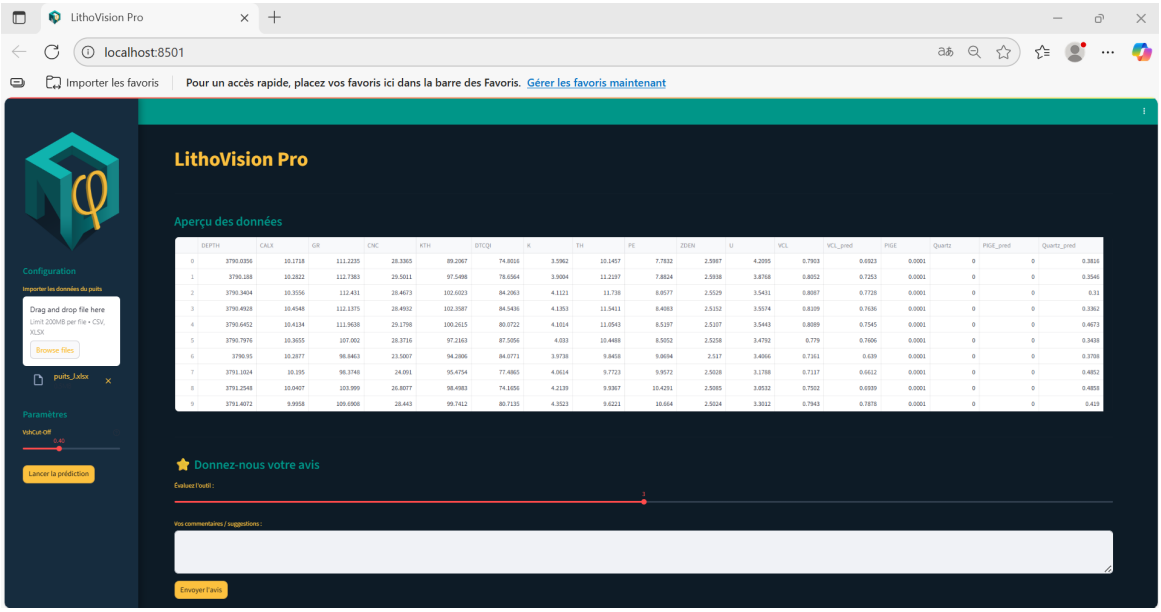


FIGURE 6.1 – Interface d'accueil de LithoVision Pro v1.0

### 6.2.3 Affichage des résultats

Les prédictions du modèle sont visualisées selon cinq tracks verticaux alignés sur la profondeur mesurée (DEPTH) :

- **Track 1** — **VCL** : Volume d'argile prédit.
- **Track 2** — **Quartz+PIGE** : Proportion de quartz et porosité effective prédite.
- **Track 3** — **Igneous** : Présence de roches magmatiques.
- **Track 4** — **Autres_Litho** : Lithologies mineures (anhydrite, calcite, halite, dolomite).
- **Track 5** — **Lithologie dominante** : Classification lithologique principale.

*Remarque : Les valeurs prédictives des tracks 1 à 4 sont normalisées dans l'intervalle [0–1] (v/v).*

La figure 6.2 présente un exemple de résultats affichés après prédiction sur un puits réel.

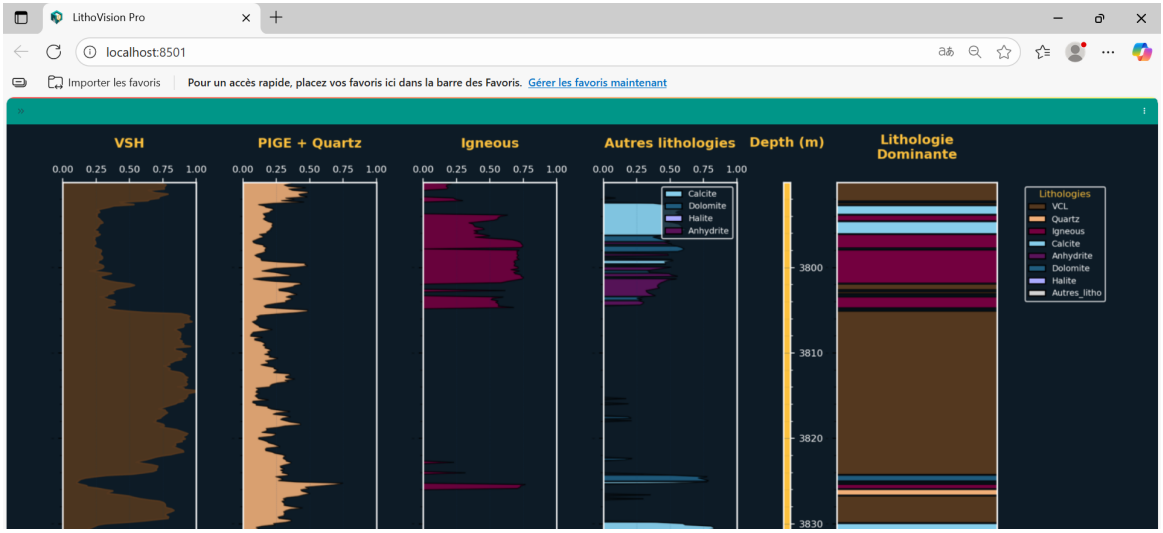


FIGURE 6.2 – Visualisation des 5 tracks lithologiques et pétrophysiques en fonction de la profondeur.

### 6.2.4 Module de feedback intégré

Une fonctionnalité stratégique de LithoVision Pro v1.0 est l'**espace de commentaires et notation** directement intégré dans l'interface. Ce module permet aux utilisateurs de :

- Attribuer une note sur 5 étoiles à la qualité des prédictions
- Rédiger des commentaires libres sur les limitations ou suggestions d'amélioration
- Soumettre ces retours anonymement, stockés automatiquement dans une feuille Google Sheets pour analyse ultérieure

Ce système de validation terrain a permis de collecter 7 retours d'utilisateurs entre septembre et octobre 2025, dont l'analyse détaillée est présentée dans la section suivante. La Figure 6.3 illustre ce système de collecte de retours ainsi que la synthèse des commentaires agrégés dans Google Sheets.

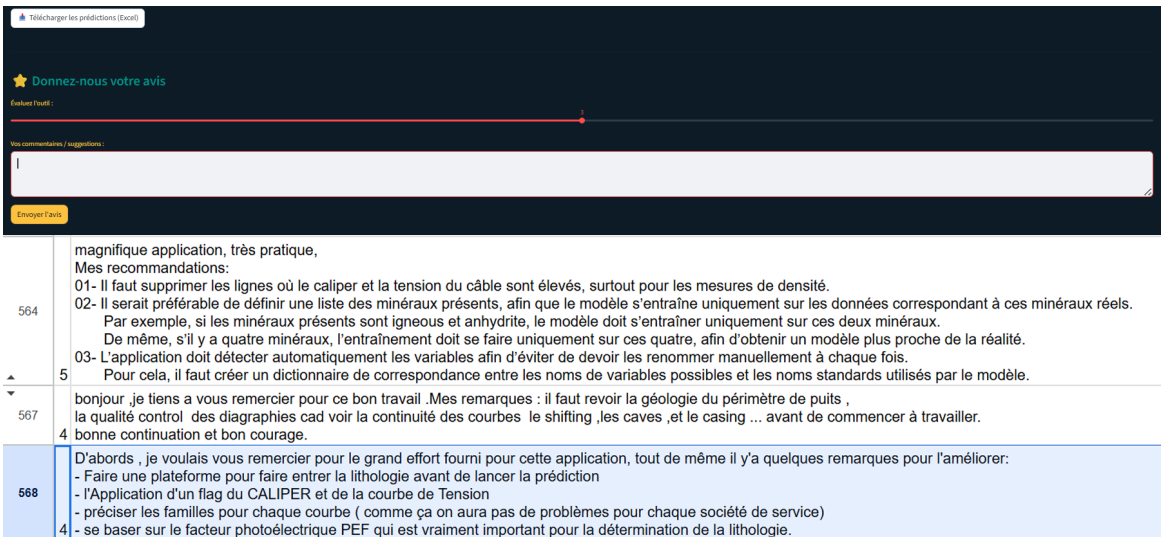


FIGURE 6.3 – Module de feedback utilisateur intégré dans l'interface (système de notation 5 étoiles et espace commentaires) et synthèse des retours collectés via Google Sheets pour analyse qualitative

### 6.2.5 Limitations techniques actuelles

Malgré les avantages de Streamlit en termes de rapidité de développement, certaines limitations techniques ont été identifiées :

- **Interactivité graphique limitée** : Les bibliothèques de visualisation disponibles dans Streamlit (Matplotlib, Seaborn) ne permettent pas de fonctionnalités avancées telles que le zoom interactif sur les tracks ou l'affichage dynamique des valeurs au survol de la souris, contrairement aux outils industriels (Plotly, Bokeh).
- **Contraintes de nommage des colonnes** : Le fichier uploadé doit impérativement respecter les noms exacts des features utilisées lors de l'entraînement du modèle (GR, RHOB, NPHI, etc.), ce qui nécessite un pré-traitement manuel des fichiers LAS exportés depuis d'autres logiciels.

## 6.3 Validation terrain et retours utilisateurs

**Synthèse quantitative**

- **Note moyenne** : 3,4 / 5
- **Répartition** :
  - 5 étoiles : 1 utilisateur (14%)
  - 4 étoiles : 2 utilisateurs (29%)
  - 3 étoiles : 4 utilisateurs (57%)
  - 1-2 étoiles : 2 utilisateurs (sans commentaires)

Cette note moyenne de 3,4/5 reflète un **intérêt confirmé** pour l'outil tout en identifiant des

axes d'amélioration prioritaires. L'absence de notes extrêmement basses avec commentaires (0-2/5 argumentés) suggère que les limitations identifiées sont principalement d'ordre fonctionnel et non conceptuel.

**Analyse thématique des commentaires**

Les retours ont été analysés selon une approche qualitative par codage thématique, permettant d'identifier 5 catégories de critiques récurrentes.

A. **C**ontrôle qualité des données (57% des utilisateurs)

Quatre utilisateurs sur sept ont souligné la nécessité d'intégrer un filtrage automatique des données de mauvaise qualité avant prédiction, notamment :

- Suppression des profondeurs où le **caliper est élevé** (outil décentré, effondrement de paroi)
- Exclusion des zones où la **tension du câble est anormale** (indicateur de problèmes d'acquisition)

*Citations représentatives :*

> *« Il faut enlever les lignes où le caliper et la tension du câble sont élevés, surtout pour la densité »* — Utilisateur #1 (3/5)

> *« Il faut revoir la qualité control des diagraphies, c'est-à-dire voir la continuité des courbes, le shifting, les caves, et le casing avant de commencer à travailler »* — Utilisateur #3 (3/5)

**Analyse :** Cette critique reflète une réalité opérationnelle essentielle : les diagraphies réelles comportent systématiquement des zones de mauvaise qualité (washouts, câble coincé, zones tubées) qui perturbent les modèles de machine learning entraînés sur données propres. L'absence de filtrage préalable dans la v1.0 constitue une limitation majeure pour une utilisation industrielle.

B. **D**étection automatique des variables LAS (43% des utilisateurs)

Trois utilisateurs ont demandé une fonctionnalité de **reconnaissance automatique des noms de courbes**, actuellement très variable selon les compagnies de service.

*Citation représentative :*

> *« Il faut que l'application détecte automatiquement les variables, afin d'éviter que chaque fois je renomme les variables. Tu dois créer une sorte de dictionnaire »* — Utilisateur #1 (3/5)

**Analyse :** Cette demande révèle une contrainte industrielle forte : les fichiers LAS ne suivent pas une nomenclature standardisée. Un système de correspondance automatique améliorerait significativement l'expérience utilisateur.

C. **U**tilisation du facteur photoélectrique PEF (43% des utilisateurs)

Trois utilisateurs ont souligné l'importance de la courbe **PEF (Photoelectric Factor)** pour discriminer certaines lithologies, notamment les carbonates et évaporites.

*Citation représentative :*

> *« Se baser sur le facteur photoélectrique PEF qui est vraiment important pour la détermination de la lithologie »* — Utilisateur #4 (4/5)

**Analyse :** Le PEF est effectivement un log clé pour différencier calcaires (PEF $\approx 5$), dolomies (PEF $\approx 3$) et anhydrite (PEF $\approx 5.1$). Son absence parmi les features du modèle v1.0 limite la performance sur les séquences carbonatées, fréquentes dans les bassins algériens.

D. **:** Entraînement ciblé sur minéraux présents (29% des utilisateurs)

Deux utilisateurs ont proposé un système d'**entraînement adaptatif** : sélectionner uniquement les lithologies présentes dans le puits avant prédiction.

**Analyse :** Cette suggestion vise à améliorer la précision en réduisant l'espace de prédiction. Elle nécessiterait cependant une connaissance géologique préalable du puits, introduisant une dépendance à des données externes.

E. **:** Intégration d'un masterlog de référence (29% des utilisateurs)

Deux utilisateurs ont demandé la possibilité d'importer une **lithologie de référence** (masterlog d'un puits voisin ou interprétation géologique régionale).

**Analyse :** Cette fonctionnalité permettrait un mode supervisé où le modèle ajusterait ses prédictions en fonction d'une connaissance géologique a priori, utile pour les champs matures bien documentés.

**Points forts identifiés**

Malgré les critiques constructives, plusieurs retours soulignent des aspects positifs de LithoVision Pro v1.0 :

- *« Magnifique application, très pratique »* — Utilisateur #2 (4/5)
- *« Je tiens à vous remercier pour ce bon travail »* — Utilisateur #3 (3/5)
- *« Grand effort fourni pour cette application »* — Utilisateur #4 (4/5)
- *« C'est un très bon travail »* — Utilisateur #6 (5/5)

Ces commentaires confirment la **pertinence du concept** et l'utilité perçue de l'outil, malgré les améliorations fonctionnelles nécessaires.

**Validation du Product-Market Fit**

Cette phase de validation utilisateur **confirme trois hypothèses clés** :

1. **Le besoin existe** : Note moyenne de 3,4/5 et retours majoritairement constructifs démontrent l'intérêt du marché.
2. **Le concept est viable** : Aucun utilisateur ne remet en cause l'approche IA pour la lithologie, seuls des aspects fonctionnels sont critiqués.
3. **Les axes d'amélioration sont identifiés** : Les critiques récurrentes fournissent une roadmap produit claire et priorisée.

**Perspectives d'évolution (Version 2.0)**

Les retours utilisateurs et les limitations techniques identifiées guident la roadmap de développement post-PFE :

- **Lecture native du format LAS** : Intégration d'un parser LAS automatique pour éviter la conversion manuelle en CSV/Excel et gérer directement les métadonnées de puits (API, coordonnées, compagnie de service).
- **Détection automatique des variables** : Développement d'un dictionnaire de correspondance multi-opérateurs pour identifier automatiquement les courbes (ex : CNC, NPHI, TNPHI → La porosité neutron) et réduire les contraintes de nommage.
- **Module quality control** : Implémentation d'un système de filtrage automatique des zones de mauvaise qualité (caliper élevé, tension de câble anormale) avant prédiction.
- **Enrichissement des paramètres pétrophysiques** : Ajout de la prédiction de perméabilité (K) et saturation en eau (Sw), conformément à la roadmap présentée dans la section Mission.
- **Migration vers une bibliothèque graphique interactive** : Remplacement de Matplotlib par Plotly pour permettre zoom, pan et affichage dynamique des valeurs.

Cette démarche itérative centrée utilisateur (*user-centric design*) est cohérente avec les meilleures pratiques de développement de startups technologiques, où le feedback terrain prime sur les intuitions produit. Cette approche a guidé la conception de la version v1.0 présentée dans ce mémoire. Les améliorations planifiées pour les versions 2.0 et 3.0 visent à transformer LithoVision Pro (v1.0) d'un **outil de démonstration académique** en une **solution industrielle robuste** répondant aux exigences opérationnelles du secteur pétrolier algérien.

## 6.4 Vision de N-PHILITH

### 6.4.1 Vision à court terme (1-3 ans)

**Objectif principal :** Établir LithoVision Pro comme l'outil de référence pour l'interprétation lithologique et pétrophysique rapide en Algérie.

**Jalons clés :**

- **Finalisation technique** : Transition de l'interface Streamlit (MVP) vers une application logicielle professionnelle développée avec Qt 6 / PySide6, offrant une interface utilisateur

moderne, modulaire et hautement performante. Cette architecture permettra d'intégrer des modules avancés (classification lithologique, modélisation 3D, etc.), tout en assurant un déploiement multi-plateforme (Windows, Linux, macOS) et une exécution optimisée pour les environnements industriels.

- **Acquisition clients pilotes** : Signer 1 à 3 clients de référence (Sonatrach, service companies, bureaux d'études) pour valider le modèle économique et collecter des retours terrain essentiels.
- **Enrichissement fonctionnel** : Développer les fonctionnalités prioritaires identifiées par les retours utilisateurs (contrôle qualité automatique, détection des variables LAS, intégration PEF).
- **Certification et conformité** : Obtenir les validations nécessaires pour travailler avec des données sensibles (conformité réglementations algériennes sur la sécurité des données pétrolières).
- **Partenariats académiques** : Renforcer les collaborations avec les universités algériennes.

**Métriques de succès à 3 ans (objectifs estimés) :**

- 5 à 10 clients actifs générant un chiffre d'affaires récurrent
- 50+ puits interprétés via la plateforme (estimation basée sur hypothèse de 10-15 puits/-client/an)
- Équipe de 3 à 5 personnes (co-fondateur technique, développeur, expert géologue)
- Temps d'interprétation moyen inférieur à 5 minutes par puits
- Taux de satisfaction client supérieur à 80%

### 6.4.2 Vision à moyen terme (4-6 ans)

**Objectif principal :** Transformer N-PHILITH d'un outil spécialisé en une **suite complète de géoscience IA** couvrant l'ensemble du workflow d'évaluation de réservoir.

**Développement de nouveaux modules :**

Au-delà de LithoVision Pro, la plateforme s'enrichira progressivement de modules couvrant :

- **Modélisation géostatistique 3D** : Krigeage, simulations séquentielles, propagation spatiale de propriétés pétrophysiques, quantification d'incertitudes volumétriques
- **Intégration et interprétation géophysique** : Calage sismique-puits, détection automatique de failles et discontinuités, analyse d'attributs sismiques (cohérence, courbure)
- **Analyse quantitative d'incertitudes** : Simulations Monte Carlo pour estimation de réserves, arbres de décision optimisés par IA, évaluation des risques exploration/production

**Expansion géographique :**

Une fois la position consolidée en Algérie, N-PHILITH visera une expansion progressive vers les marchés maghrébins (Tunisie, Libye, Maroc) et potentiellement l'Afrique subsaharienne (Nigeria, Angola), où la demande pour des solutions d'interprétation abordables et performantes est en forte croissance.

**Métriques de succès à 5 ans (objectifs visés) :**

- 25 à 30 clients actifs (grandes entreprises, PME, universités)
- Présence commerciale dans 3+ pays africains
- Équipe de 15 à 20 personnes (développeurs, data scientists, commerciaux, géoscientifiques)
- 3 modules opérationnels minimum couvrant diagraphie, géostatistique et géophysique
- Reconnaissance comme acteur majeur de la géoscience IA en Afrique

### 6.4.3 Vision à Long Terme (7-10 ans)

À l'horizon 2035, N-PHILITH vise une consolidation technique et une ouverture progressive vers des collaborations internationales, tout en maintenant son ancrage national :

- **Stabilisation du produit** : Suite logicielle mature couvrant l'ensemble de la chaîne d'interprétation pétrophysique et lithologique, avec une base de clients récurrents dans plusieurs pays africains.
- **Recrutement de profils PhD** : Intégration progressive de chercheurs post-doctoraux algériens, africains, et éventuellement internationaux pour renforcer les capacités de R&D et explorer des méthodologies avancées.
- **Collaborations académiques ciblées** : Partenariats de recherche avec des universités africaines (Université de Tunis, Université du Caire, Université de Lagos) et éventuellement européennes (IFP, ENSG) sur des projets spécifiques co-financés.
- **Participation scientifique internationale** : Publication régulière dans des revues spécialisées et participation aux conférences sectorielles (SPE, AAPG) pour gagner en crédibilité scientifique et visibilité internationale.

Cette vision long terme reste ancrée dans une approche pragmatique : chaque étape de croissance sera conditionnée par la validation terrain, la rentabilité économique et la pertinence scientifique des développements entrepris.

## 6.5 Étude financière du projet

Table 6.1 – Besoins de démarrage - Investissements initiaux

| **Poste de dépense** | **Montant (DZD)** | **Description** |
|---|---|---|
| Frais d'établissement | 50 000.00 | Ce sont les frais de création de l'entreprise (formalités) |
| Logiciels professionnels | 100 000.00 | JetBrains, GitHub, etc. |
| Dépôt marque | 150 000.00 | Frais de dépôt ou d'enregistrement |
| Frais de dossier | 80 000.00 | Pour la signature de contrats de prêt |
| Enseigne et éléments de communication | 300 000.00 | Cartes de visite, site internet |
| Matériel de bureau | 1 000 000.00 | Fournitures, ordinateur, imprimante |
| Trésorerie de départ | 1 500 000.00 | Somme d'argent gardée en prévision du démarrage de l'activité pour financer le cycle d'exploitation |
| **TOTAL** | **3 185 000.00** | |

| **Détail des amortissements** | | | |
|---|---|---|---|
| | **Année 1** | **Année 2** | **Année 3** |
| **Amortissements incorporels** | **30 000,00** | **30 000,00** | **30 000,00** |
| *Frais d'établissement* | 16 666,67 | 16 666,67 | 16 666,67 |
| *Logiciels professionnels* | 6 666,67 | 6 666,67 | 6 666,67 |
| *Frais de dossier* | 6 666,67 | 6 666,67 | 6 666,67 |
| **Amortissements corporels** | **766 666,67** | **766 666,67** | **766 666,67** |
| *Enseigne et éléments de communication* | 100 000,00 | 100 000,00 | 100 000,00 |
| *Matériel de bureau* | 666 666,67 | 666 666,67 | 666 666,67 |
| **Total amortissements** | **796 666,67** | **796 666,67** | **796 666,67** |

**Coûts salariaux**

Le coût salarial est calculé selon la formule suivante :

$$\text{Brut} = \frac{\text{Net}}{1 - (\text{charges salariales} + \text{IRG})} \tag{6.1}$$

Avec charges salariales + IRG estimées à 10 %, puis :

$$\text{Coût total} = \text{Brut} \times 1{,}26 \text{ (charges patronales de 26 \%)} \tag{6.2}$$

**Projection sur 3 ans (4 salariés)**

Salaires nets mensuels : 65 000 DA (année 1), 80 000 DA (année 2), 90 000 DA (année 3).

| **Année** | **Coût annuel par salarié (DZD)** | **Coût total pour 4 salariés (DZD)** |
|---|---|---|
| 1 | 1 213 332 | 4 853 328 |
| 2 | 1 493 328 | 5 973 312 |
| 3 | 1 680 000 | 6 720 000 |

TABLE 6.2 – Total des charges fixes

| | **Année 1 (DZD)** | **Année 2 (DZD)** | **Année 3 (DZD)** |
|---|---|---|---|
| Charge salariés | 4 853 328,00 | 5 973 312,00 | 6 720 000,00 |
| Téléphone, internet | 15 000,00 | 15 000,00 | 15 000,00 |
| Eau, électricité, gaz | 120 000,00 | 120 000,00 | 120 000,00 |
| Fournitures diverses | 120 000,00 | 120 000,00 | 120 000,00 |
| Nettoyage des locaux | 336 000,00 | 336 000,00 | 336 000,00 |
| Budget publicité et communication | 840 000,00 | 1 000 000,00 | 1 200 000,00 |
| Loyer et charges locatives | 600 000,00 | 720 000,00 | 780 000,00 |
| Expert comptable | 100 000,00 | 100 000,00 | 100 000,00 |
| Frais bancaires | 15 000,00 | 15 000,00 | 15 000,00 |
| **TOTAL** | **6 999 328,00** | **8 399 312,00** | **9 406 000,00** |

## Année 1 - Vente de licences logicielles

| | **Jours travaillés** | **CA/jour (DA)** | **CA mensuel (DA)** |
|---|---|---|---|
| Mois 1 | 20 | 40 000 | 800 000 |
| Mois 2 | 20 | 40 000 | 800 000 |
| Mois 3 | 20 | 40 000 | 800 000 |
| Mois 4 | 20 | 40 000 | 800 000 |
| Mois 5 | 20 | 40 000 | 800 000 |
| Mois 6 | 20 | 40 000 | 800 000 |
| Mois 7 | 20 | 40 000 | 800 000 |
| Mois 8 | 20 | 90 000 | 1 800 000 |
| Mois 9 | 20 | 90 000 | 1 800 000 |
| Mois 10 | 20 | 90 000 | 1 800 000 |
| Mois 11 | 20 | 90 000 | 1 800 000 |
| Mois 12 | 20 | 90 000 | 1 800 000 |
| **TOTAL** | | | **14 600 000** |

Table 6.3 – Chiffre d'affaires prévisionnel année 1

## Détail de la tarification

- **Mois 1 à 7** : 1 client (Sonatrach - licence exclusive) à 40 000 DA/jour
- **Mois 8 à 12** : 2 clients à 90 000 DA/jour
  - Client 1 (Sonatrach) : 40 000 DA/jour
  - Client 2 (licence professionnelle) : 50 000 DA/jour

## Tarifs annuels de licence

| **Type de client** | **Prix/jour (DA)** | **Prix/an (DA)** | **Prix/an (EUR)** |
|---|---|---|---|
| Sonatrach (exclusivité) | 40 000 | 9 600 000 | 36 923 |
| Client professionnel | 50 000 | 12 000 000 | 46 154 |

Table 6.4 – Grille tarifaire des licences

# Compte de résultat prévisionnel

## Hypothèses

- Crédit bancaire : 3 185 000 DA
- Taux d'intérêt : 6 %
- Impôt sur les bénéfices : 26 %
- Annuité de remboursement : 1 191 540 DA

| | **Année 1** | **Année 2** | **Année 3** |
|---|---|---|---|
| Remboursement du crédit | 1 000 440 | 1 060 466 | 1 124 094 |
| Chiffre d'affaires (CA) | 14 600 000 | 17 520 000 | 19 272 000 |
| Amortissements (hors crédit) | 796 667 | 796 667 | 796 667 |
| Charges financières (intérêts) | 1 858 333 | 1 858 333 | 1 858 333 |
| Total des charges | 6 999 328 | 8 399 312 | 9 406 000 |
| **Résultat d'exploitation** | **7 600 672** | **9 120 688** | **9 866 000** |
| Impôt sur les bénéfices (26 %) | 1 976 175 | 2 371 379 | 2 565 160 |
| **Résultat net** | **5 624 497** | **6 749 309** | **7 300 840** |
| **Cash-flow** | **6 482 390** | **7 547 176** | **8 035 079** |

TABLE 6.5 – Compte de résultat et cash-flow prévisionnel sur 3 ans

## Formules de calcul

- **Résultat d'exploitation** = CA - Total des charges - Amortissements - Charges financières
- **Impôt** = Résultat d'exploitation × 26 %
- **Résultat net** = Résultat d'exploitation - Impôt
- **Cash-flow** = Résultat net + Amortissements - Remboursement du crédit

## 6.6 Conclusion

Ce chapitre a présenté N-PHILITH, une startup dédiée au développement d'outils d'intelligence artificielle pour l'interprétation géoscientifique des puits pétroliers en Algérie. Son premier produit, LithoVision Pro v1.0, propose des analyses lithologiques et pétrophysiques rapides adaptées aux bassins algériens.

La validation terrain auprès de log analystes de Sonatrach a confirmé la pertinence du concept avec une note moyenne de 3,4/5, tout en identifiant des axes d'amélioration prioritaires : contrôle qualité automatique des diagraphies, détection des variables LAS, et intégration du facteur photoélectrique (PEF).

L'étude financière démontre la viabilité économique du projet avec un chiffre d'affaires évoluant de 14,6 à 19,3 millions de dinars sur trois ans, un résultat net croissant de 5,6 à 7,3 millions de dinars, et un cash-flow positif constant. La stratégie de tarification positionne l'offre comme alternative compétitive aux solutions internationales tout en restant accessible au marché algérien.

Au-delà de son positionnement commercial, N-PHILITH ambitionne de contribuer au renforcement de l'écosystème national en géosciences en mettant l'outil à disposition des universités algériennes et en publiant partiellement le code source. La vision stratégique dessine une trajectoire progressive : consolidation sur le marché algérien (3 ans), expansion vers l'Afrique du Nord (5 ans), puis évolution vers une suite complète couvrant modélisation 3D et géophysique.

N-PHILITH porte ainsi une vision de souveraineté numérique dans le domaine stratégique de l'énergie, en développant des solutions adaptées aux réalités locales et accessibles aux acteurs nationaux.

# Conclusion générale

L'évolution rapide de la technologie et l'augmentation des besoins de l'industrie pétrolière et gazière ont conduit à la recherche de solutions innovantes pour améliorer l'efficacité de la caractérisation des réservoirs. Ce travail, en intégrant l'intelligence artificielle et l'apprentissage automatique dans l'interprétation des logs de puits, propose une méthode novatrice qui permet de surmonter les limites des approches traditionnelles. En automatisant le processus d'analyse, le modèle développé ici offre la possibilité de prédire avec précision les propriétés lithologiques et pétrophysiques, tout en réduisant le temps d'analyse et les coûts associés.

Les résultats obtenus montrent que l'apprentissage automatique peut effectivement rivaliser avec l'expertise humaine en matière d'interprétation des données géophysiques, en offrant une solution rapide et fiable, capable de généraliser ses prédictions à des puits non utilisés lors de l'entraînement du modèle. Cette capacité à traiter simultanément plusieurs types de logs et à identifier différentes lithologies représente un progrès significatif par rapport aux méthodes classiques.

Le système développé, fondé sur des algorithmes de gradient boosting (XGBoost, CatBoost, LightGBM), a montré des performances exceptionnelles lorsqu'il a été appliqué à l'analyse de neuf puits des réservoirs ordovicien et triasique du bassin algérien, structuré en deux groupes selon la nature des lithologies.

Pour le **Groupe 1**, les modèles de régression ont obtenu des résultats impressionnants, avec un coefficient de détermination $R^2$ de 0,9828 pour le volume d'argile ($V_{CL}$), 0,9055 pour le Quartz, et 0,8564 pour la porosité effective (PIGE), expliquant respectivement 98,28%, 90,55% et 85,54% de la variance.

Le **Groupe 2** a démontré une sophistication méthodologique remarquable avec une architecture adaptée aux lithologies discontinues. En termes de classification, les résultats sont également remarquables, avec une précision de 97% pour la détection des roches ignées (F1-score 0,95) et 96% pour les autres lithologies (F1-score 0,95). La régression conditionnelle a permis une quantification précise avec $R^2$ de 0,9391 pour Igneous et 0,8140 pour les autres lithologies. La classification multi-classe finale a discriminé quatre lithologies (calcite/dolomite/anhydrite/halite) avec une précision de 95,7% et des F1-scores allant de 0,91 à 0,98.

La validation externe réalisée sur un puits historique indépendant (Puits J) a confirmé la robustesse du modèle, avec une identification parfaite des réservoirs productifs (100% des cinq réservoirs détectés, dont 80% avec une caractérisation parfaite).

Les résultats obtenus mettent en évidence l'efficacité du système, qui surpasse les méthodes traditionnelles tant en termes de précision que de rapidité. L'un des principaux avantages de

cette approche réside dans sa capacité à exécuter les analyses en quelques minutes, contre une journée entière pour les méthodes classiques. De plus, l'architecture modulaire et multi-cible du système permet une réadaptation ciblée des modèles sans nécessiter la reconstruction complète du pipeline, facilitant ainsi l'adaptation à de nouvelles données. L'adaptation spécifique aux conditions géologiques locales améliore également les performances par rapport aux modèles génériques souvent utilisés dans les logiciels commerciaux. Enfin, l'indépendance vis-à-vis des licences propriétaires, telles que celles utilisées dans des logiciels comme Techlog ou Petrel, offre à Sonatrach une autonomie stratégique précieuse pour ses activités de forage et d'analyse.

Les contributions majeures de ce travail incluent, en premier lieu, l'architecture flexible et évolutive qui permet des ajustements rapides sans refonte complète du système. Deuxièmement, la réduction significative du temps de traitement ouvre la voie à l'analyse en temps réel et à l'échelle des campagnes multi-puits. Troisièmement, l'approche localisée améliore la précision par rapport aux solutions universelles. Enfin, l'indépendance vis-à-vis des logiciels commerciaux constitue un avantage stratégique pour l'industrie.

Ces résultats ont conduit au développement de **N-PHILITH**, une startup dédiée à la valorisation de cette technologie sous forme du produit **LithoVision Pro v1.0**. La validation terrain auprès de log analystes de Sonatrach a confirmé la pertinence du concept avec une note moyenne de 3,4/5. L'étude de viabilité économique démontre un modèle d'affaires solide avec un chiffre d'affaires évoluant de 14,6 à 19,3 millions de dinars sur trois ans, un résultat net croissant de 5,6 à 7,3 millions de dinars, et un cash-flow positif constant (de 6,5 à 8,0 millions de dinars). Cette stratégie de tarification positionne l'offre comme alternative compétitive aux solutions internationales tout en restant accessible au marché algérien.

Au-delà de son positionnement commercial, N-PHILITH ambitionne de contribuer au renforcement de l'écosystème national en géosciences en mettant l'outil à disposition des universités algériennes et en publiant partiellement le code source. La vision stratégique dessine une trajectoire progressive : consolidation sur le marché algérien (3 ans), expansion vers l'Afrique du Nord (5 ans), puis évolution vers une suite complète couvrant modélisation 3D et géophysique.

Néanmoins, trois limitations importantes ont été identifiées. La sur-détection des roches ignées peut être corrigée par l'enrichissement du jeu de données d'entraînement et un ajustement plus fin des seuils de classification. De plus, la règle de post-traitement conservatrice, qui attribue une porosité effective (PIGE) nulle en présence de roches ignées, pourrait être améliorée en remplaçant cette approche par une pénalisation graduée proportionnelle. Ces ajustements permettront d'améliorer encore la précision du modèle, en particulier dans les configurations géologiques complexes, et de renforcer la capacité du système à gérer des lithologies variées.

En conclusion, ce travail ouvre de nouvelles perspectives pour l'automatisation de l'interprétation des logs de puits, offrant une alternative rapide, fiable et économiquement avantageuse aux méthodes traditionnelles. N-PHILITH porte ainsi une vision de souveraineté numérique dans le domaine stratégique de l'énergie, en développant des solutions adaptées aux réalités locales et accessibles aux acteurs nationaux. Les améliorations proposées renforceront la robustesse et la flexibilité du système, ouvrant ainsi la voie à une adoption plus large dans l'industrie pétrolière et gazière, tout en garantissant une autonomie accrue dans les processus d'analyse.

# Bibliographie


[1] Oberto Serra. *Diagraphies différées : Bases de l'interprétation, Tome 1 – Acquisition des données diagraphiques*. Bulletin des Centres de Recherches Exploration-Production Elf-Aquitaine, Pau, France, 1979. Mémoire 1.

[2] SONATRACH, Division Exploration. Documents internes de la division exploration. Documents internes, non publiés.

[3] Schlumberger. *Well Evaluation Conference*. Schlumberger, Houston, TX, USA, 2007.

[4] Mohamed Said Beghoul. Les diagraphies différées : Principes des outils et bases d'interprétation. Cours de formation, Institut Algérien du Pétrole (IAP), 2014. 9–13 mars 2014, UFR GGR–IAP.

[5] GeeksforGeeks. Machine learning overview diagram, 2023. Accessed : 2025-10-12.

[6] Mahesh Batta. Machine learning algorithms – a review. *International Journal of Science and Research (IJSR)*, 9(1), 2019.

[7] Guillaume Saint-Cirgue. *Machine Learnia*. Machine Learnia, 2024. Licensed under Creative Commons BY-NC 3.0.

[8] H. Askri, A. Belmecheri, B. Benrabah, A. Boudjema, K. Boumendjel, M. Daoudi, M. Drid, T. Ghalem, A. M. Docca, H. Ghandriche, A. Ghomari, N. Guellati, M. Khennous, R. Lounici, H. Naili, D. Takherist, and M. Terkmani. *Géologie de l'Algérie / Geology of Algeria*, volume I. SONATRACH, Division Exploration, Centre de Recherche et Développement, et Division Petroleum Engineering et Développement, Alger, 2003.

[9] Oberto Serra. *Diagraphies différées : Bases de l'interprétation. Tome 2 – Interprétation des données diagraphiques*. Bulletin des Centres de Recherches Exploration-Production Elf-Aquitaine, Pau, 1985. Mémoire 7. Ouvrage réalisé avec le concours d'Études et Productions Schlumberger, Montrouge.

[10] Arezki Boudjema. *Évolution structurale du bassin pétrolier « triasique » du Sahara nord-oriental (Algérie)*. Thèse de doctorat, Université Paris XI – Orsay, France, 1987.

[11] A. Allouti, A. Ziada, W. Ramses, and F.E. Fragachan. Damage characterization and production optimization of the hassi-messaoud field, algeria. In *SPE Middle East Oil Show*, number SPE-39485 in SPE Middle East Oil Show. Society of Petroleum Engineers, 1997.

[12] L. Salhi. Étude sédimentologique préliminaire des grès du rdc, gisement de rhourde chegga. Master's thesis, Université [à compléter si connue], 2015.

[13] Mohamed Said Beghoul. Les aspects de la géologie du pétrole : Comment générer, évaluer et proposer un prospect à forer. Séminaire de formation, Institut Algérien du Pétrole (IAP), Boumerdès, September 2018. 23–27 septembre 2018.

[14] Michel Meunier. Diagraphies différées et interprétation, December 2009. Interprétation des Diagnostics.

[15] Jawad Ali, Umar Ashraf, Aqsa Anees, Sanxi Peng, Muhammad Ubaid Umar, Hung Vo Thanh, Umair Khan, Mohamed Abioui, Hassan Nasir Mangi, Muhammad Ali, and Jar Ullah. Hydrocarbon potential assessment of carbonate-bearing sediments in a meyal oil field, pakistan : Insights from logging data using machine learning and quanti elan modeling. *ACS Omega*, 8(35) :32067–32084, 2023.

[16] Software Integrated Solutions. *Techlog Quanti.Elan : Workflow/Solutions Training*. Schlumberger, 2016. Version 2015, March 10, 2016.

[17] Yuchen Jiang, Xiang Li, Hao Luo, Shen Yin, and Okyay Kaynak. Quo vadis artificial intelligence ? *Discover Artificial Intelligence*, 2(4), 2022.

[18] Shashi Tanwar. *Artificial Intelligence and Machine Learning – Principles and Applications*. AGPH Books (Academic Guru Publishing House), Bhopal, M.P., India, 2024.

[19] Andres Muñoz. Machine learning and optimization. https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf, 2017. Retrieved June 1, 2017.

[20] IEEE, editor. *Proceedings of the 2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, Greater Noida, India, 2018. IEEE, IEEE.

[21] Arthur L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3) :210–229, 1959.

[22] Aized Amin Soofi and Arshad Awan. Classification techniques in machine learning : Applications and issues. *Journal of Basic & Applied Sciences*, 13 :459–465, 2017.

[23] G. Kesavaraj and S. Sukumaran. A study on classification techniques in data mining. In *Proceedings of the Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–7, 2013.

[24] Andrew Ng. Cs229 lecture notes, 2023. Accessed : 2025-10-12.

[25] J.-S. R. Jang. Anfis : Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics*, 23 :665–685, 1993.

[26] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1 :81–106, 1986.

[27] Kardi Teknomo. Strengths and weaknesses of k nearest neighbor, 2024. Accessed : 2025-10-12.

[28] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14 :1–37, 2008.

[29] DataCamp. A guide to the gradient boosting algorithm, 2023. Accessed : 2025-10-12.

[30] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The coefficient of determination $r^2$ is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7 :e623, 2021.

[31] SONATRACH Exploration-Production. Rapport d'implantation du puits vertical. Technical Report REF/Exploration-Production/PED/2017, Direction Suivi des Projets & Reporting, Pôle Exploration-Production, August 2017. Document interne SONATRACH.

[32] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. Official documentation : https://xgboost.readthedocs.io/.

[33] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost : Unbiased boosting with categorical features. *Advances in Neural*

*Information Processing Systems (NeurIPS)*, 31, 2019. Official documentation : https://catboost.ai/docs/.

[34] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm : A highly efficient gradient boosting decision tree. https://lightgbm.readthedocs.io/, 2017. Microsoft Research, NeurIPS 2017.