Machine Learning for Geomodelling

BearingPoint & Geoxilia

Télécom Paris - MS BIG DATA

Projet fil rouge

Babacar DIOUF
babacar.diouf@telecom-paris.fr

Mohammed Ouedh
mohammed@telecom-paris.fr

Parfait FANGUE
pfangue@telecom-paris.fr

Hiroto YAMAKAWA
hiroto.yamakawa@telecom-paris.fr

Octobre 2019 - Janvier 2020

Contents

1 Introduction

Représentation simplifiée de l'environnement géologique pour en montrer ses aspects importants, le geomodelling est une science appliquée, utilisée dans la gestion des ressources et risques naturelles (telles que les séismes) mais également dans la quantification des processus géologiques, dont les principales applications se trouvent dans le domaine de l'énergie. A l'heure actuelle, cette modélisation reste néanmoins une étape chronophage et coûteuse. Les progrès technologiques durant la dernière décennie a permis l'émergence du Machine Learning et de ses applications dans différents domaines. Naturellement, les industries plus traditionnelles du secteur de l'énergie s'y intéressent de plus en plus afin d'optimiser certains processus. Plus particulièrement, notre étude au cours de ce projet se concentrera sur l'exploration de diverses solutions de Machine Learning, dans le but d'optimiser cette phase de geomodelling, si importante pour la prospection et l'exploitation de gisements pétroliers. Sauf indications ultérieures, nous nous limiterons aux réservoirs de la mer du Nord, sur les champs suivants: "BEATRICE","CROMARTY","JACKY" et "ALWYN"

Après une courte description des termes techniques liées à ce domaine, nous développerons les motivations d'une approche orientée data, avant de présenter dans un second temps la démarche suivie durant cette première période. En ce sens, nous détaillerons nos premiers résultats, ainsi que les difficultés rencontrées au cours de ce travail. Nous terminerons sur une proposition des axes d'approfondissement que nous explorerons dans les prochains mois.

2 Présentation des entreprises

2.1 BearingPoint

Cabinet indépendant de conseil en management et en technologie. Avec l'acquisition de la compagnie HyperCube en 2012, BearingPoint propose depuis une offre de solutions d'analyse avancée, via une plateforme de data science accessible tournée vers l'industrialisation rapide d'applications spécifiques.

Figure 1: BearingPoint

2.2 Geoxilia

Startup spécialisée en géosciences constituée de géologues, géophysiciens et des ingénieurs réservoirs ayant plus de dix ans d'expérience dans le secteur des énergies. Spécialisée dans l'analyse et la prédiction des réservoirs d'hydrocarbures.

Figure 2: Geoxilia

3 Le Contexte et les enjeux

3.1 Le Geomodelling

Le géomodelling consiste à construire une représentation cohérente, bien qu'imparfaite, du sous-sol, basée sur des observations géophysiques et géologiques. En intégrant des données diverses et variées, elle permet aux experts de vérifier des hypothèses, mais également de visualiser des failles, des faciès, ou encore des propriétés pétrophysiques.

La représentation des structures des sous-sols est un aspect essentiel d'une large variété d'applications: étude des réservoirs, inspection des matériaux bruts et d'autres formes d'applications géologiques. Par exemple, dans l'industrie de l'énergie, des modèles réalistes générées à partir de données de puits (diagraphies/carottes) sont nécessaires en entrée des simulateurs de réservoirs, logiciels permettant de prédire le comportement des roches sous différents scénarios. Si de nombreuses solutions existent pour générer ces modèles géologiques, ces dernières sont en général confinées dans des logiciels commerciaux très coûteux.

3.2 les lithologies

La lithologie est la nature des roches formant un ensemble, une couche géologique. Plusieurs lithologies résident donc dans les sous-sols. Certaines plus particulièrement sont intéressantes puisqu'à l'intérieur s'y trouvent du pétrole à l'état brut. Ces hydrocarbures se sont formées dans des terrains propices (les roches mères) contenant des matières organiques sur des périodes de plusieurs millions d'années. Conjointement au pétrole se trouve du gaz et de l'eau. Ce dernier applique une pression hydrostatique sur les hydrocarbures (ayant des densités moins importantes) qui tentent alors de remonter vers la surface, sauf si une couche imperméable est atteinte et les empêchent de migrer.

Les hydrocarbures que nous recherchons se trouvent donc dans des roches perméables dites "réservoirs"(cela peut être du sable, des carbonates ou des dolomites), bordées par un toit imperméable "la roche piège", généralement composée d'argile (*clay* en anglais) ou de schiste. Identifier une épaisse couche de roche piège permet aux experts de localiser un hypothétique réservoir exploitable dans cette zone, méritant une étude plus approfondie.

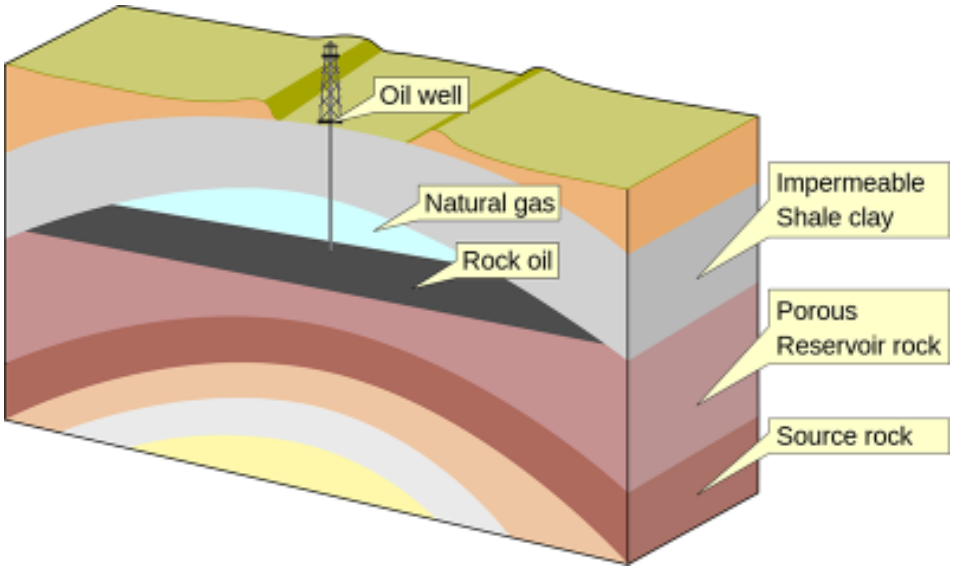


Figure 3: les différentes couches stratigraphiques

3.3 Le contexte et les enjeux

3.3.1 Problématiques

A l'heure d'aujourd'hui, et malgré les logiciels disponibles, le flux de travail typiquement utilisé en géoscience durant la phase de modélisation reste chronophage, avec de nombreuses tâches manuelles: parmi elles, des phases d'interprétations durant lesquelles une longue série de décisions successives importantes est prise par les experts, qui sont donc sujettes à des erreurs et aux biais humains.

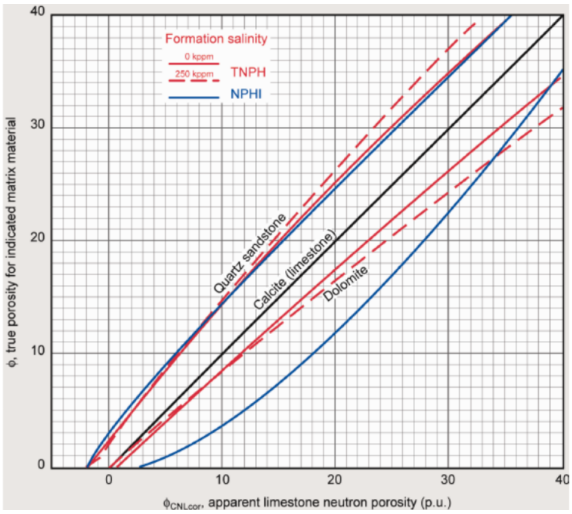


Figure 4: Exemple d'abaque utilisé par les experts - Neutron Porosity Information

3.3.2 Objectifs

Notre objectif au cours de ce projet est d'explorer diverses solutions orientées data, avec notamment l'application d'algorithmes de Machine Learning pour optimiser l'aide à la décision et anticiper les étapes critiques de façon efficace. Ccherchant à être aussi performante que les règles "métier", Cette approche orientée data pourrait également servir de proxy et offrir une alternative moins gourmande en calcul pour prédire le volume d'hydrocarbures piégés dans les réservoirs.

La phase initiale consiste à développer un premier modèle capable d'identifier correctement la lithologie à partir des variables d'entrées, et plus précisément si la roche identifiée est un argile ou non. C'est sur cet objectif que nous nous focalisons en cette première période.

LABEL/TARGET : la lithologie à partir des variables d'entrées.

3.4 Acquisition des données

Les diagraphies (LOGS en anglais) désignent les enregistrements continus des variations - en fonction de la profondeur - d'une caractéristique donnée des formations traversées par un sondage. Si plusieurs catégories existent (certaines se focalisant sur la boue, d'autres sur la vitesse de rotation des outils par exemple), nous nous intéressons particulièrement aux mesures de paramètres physiques obtenus à l'aide de sondes suspendues à un câble et descendues sur tout le long du puit - well en anglais - (de 1000 à 4000 m), pour faire des mesures directes tous les 0,5 mètres.

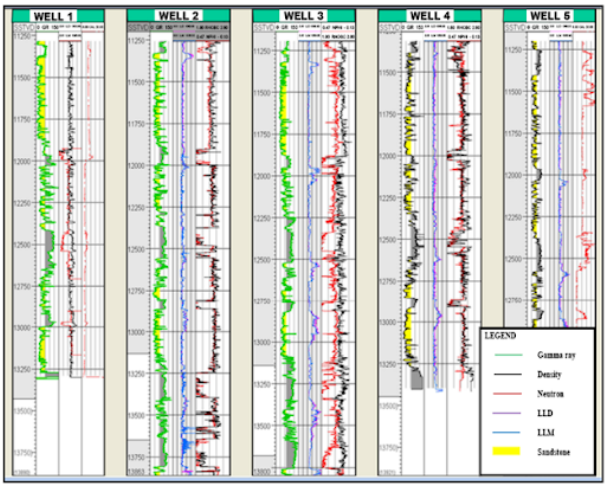


Figure 5: Exemple de diagraphies - Well Logs

Les logs principaux

- Caliper (CAL): le diamètre du trou du puit.
- Gamma Ray (GR): Mesure la radioactivité naturelle des formations géologiques.
- Neutron (PHIN): Mesure le contenu d'hydrogène de la formation. Des neutrons sont envoyées dans la formation à partir d'une source radioactive. Les neutrons ayant atteint le niveau thermique après interaction avec les atomes de la roche sont comptés en retour. On observe le ralentissement des neutrons, leur perte d'énergie et leur disparition.
- Density (RHOB): Des rayons gamma sont projetées dans la formation à partir d'une source radioactive puis les rayons gamma après interaction avec la roche sont comptés en retour. Mesure la densité des formations géologiques (RHOB).
- PHOTO ELECTRIC (PEF): Mesure l'absorption photo-électrique de la formation géologique.
- Sonic (DTC): On émet des ondes de compression P dans le puit et on mesure le temps que met l'onde pour se propager dans une formation d'un pied d'épaisseur. Mesure la vitesse de ralentissement des ondes de compression dans la formation géologique
- Resistivity (Micro, Shallow, Medium, Deep): Mesure la résistivité de la formation géologique à différentes profondeurs d'investigation (latérales).

4 Preprocessing

4.1 Description des fichiers

Dans le cadre de notre étude, nous avons reçu de la part de Geoxilia, des données portant sur les champs petroliers Alwyn, Jacky, Cromarty et Beatrice situés en mer du nord. Les données reçues se présentent sous les formats ".las" et ".xls".

• Les fichiers ≪xls≫ utiles pour notre étude sont au nombre de 3 et contiennent les lithologies associées à chaque profondeur pour un puit donné d'un champs, comme le montre l'extrait ci-dessous:

WELL	TOP	BASE	LITHO	DESCRIPTION	CHAMP
12/21C-6	121.92	153.01	UNDF	Undifferentiated	Jacky
12/21C-6	364.85	999.74	VALH	Valhall	Jacky
12/21C-6	999.74	1022.3	UHSM	Upper Hot Shale Member	Jacky
12/21C-6	1022.3	1077.77	SLTM	Siltstone Member	Jacky
12/21C-6	1077.77	1081.74	RZDM	Ryazanian Sand Member	Jacky
12/21C-6	1081.74	1594.41	SLTM	Siltstone Member	Jacky
12/21C-6	1594.41	1824.84	LHSM	Lower Hot Shale Member	Jacky
12/21C-6	1824.84	1985.16	OXSH	Oxfordian Shale	Jacky

Figure 6: Extrait d'un fichier welltops

- WELL représente le nom du puit
- Top et Base indiquent respectivement les profondeurs de départ et d'arrivée
- LITHO représente le code associé à la lithologie
- DESCRIPTION représente le nom de la lithologie

• Quant aux fichiers.las , on en dénombre 114 portant sur 68 puits distincts. Il s'agit de fichiers binaires permettant de stocker différentes mesures physiques (Gamma Ray, Résistivité...) communément appelées logs. La description des logs ou "curves" est en général renseignée dans l'en-tête du fichier. Dans nos fichiers, ces mesures sont collectées tous les 50 centimètres de profondeur comme le montre l'extrait ci-dessous :

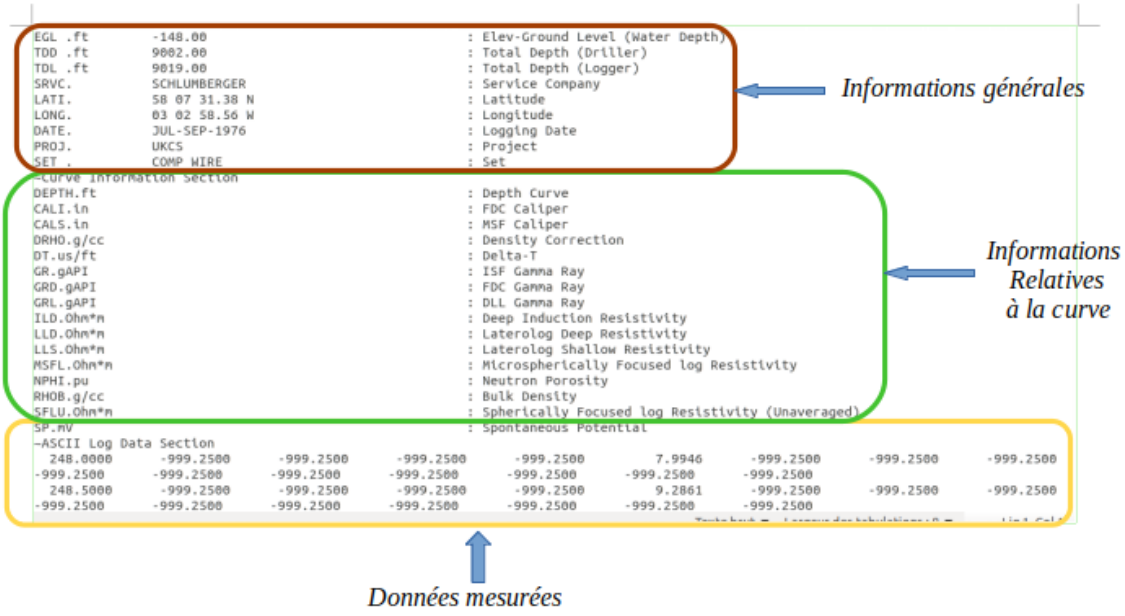


Figure 7: Extrait d'un fichier de data au format ".las"

Ces fichiers présentent la particularité d'être très hétérogènes d'un puit à l'autre (voir d'un champ à l'autre). Cette hétérogénéité se matérialise d'une part, par le fait que certaines variables identiques portent des noms différents d'un puits à l'autre et soient exprimées dans des unités différentes et d'autre part, par le fait que certaines variables soient absentes du glossaire international Schlumberger. L'extrait ci-dessous permet de mieux nous rendre compte de cette hétérogénéité:

Idx	UWI	Data	Passing	CAL*	DCOR*	DENS*	DTC*	DTS*	GR*	PE*	PHIN*
			%	66/114 wells	62/114 wells	81/114 wells	66/114 wells	5/114 wells	105/114 wells	28/114 wells	81/114 wells
0		9/95 curves	–	SCAL 12.41 in	SCOR 0.07 g/cc	SNB2 2.29 g/cc			GAB 58.45 api	SNP2 6.11 b/e	TNPL 13.02 pu
1	GB-11_30A-10	9/14 curves	–	CALI 12.23 IN	DRHO 0.01 0	RHOB 2.46 0	DTC2 85.98 0		GR 68.19 GAPI		NPHI 17.85 LPU
2		3/18 curves	–				DTMN 52.03 US/F		GR 46.30 GAPI		
3	GB-11_30A-A29	10/13 curves	–	CALI 15.67 IN	DRHO 0.03 G/C3	RHOB 2.44 G/C3	DT 85.92 US/F		GR 59.13 GAPI	PEF 6.59 B/E	NPHI 0.27 V/V
4		0/1 curves	–								
5	11920	9/14 curves	–	CALI 12.23 in	DRHO 0.01 g/cc	RHOB 2.46 g/cc	DTC2 85.98 us/ft		GRD 88.51 gAPI		NPHI 17.85 pu
6		9/10 curves	–	CAL_PP 8.05 IN	DENC_PP 0.01 G/C3	DENS_PP 2.44 G/C3	SONC_PP 105.85 US/F		GAMMA_PP 62.14 GAPI		NEUT_PP 0.27 V/V
7		2/18 curves	–				DT 54.35 US/F		RGR 36.64 GAPI		
8		3/18 curves	–				DTMN 52.16 US/F		GR 44.62 GAPI		
9		4/5 curves	–			RHOB 2.38 g/cm3	DT 83.61 us/ft		GR 62.38 gAPI		NPHI 0.22 m3/m3
10		7/9 curves	–	CAL_PP 0.67 INCH	DENC_PP -0.00 G/CC	DENS_PP 2.27 G/C3			GAMMA_PP 57.51 API		NEUT_PP 0.22 V/V
11	7462	8/9 curves	–	CALI 13.07 in	DRHO 0.04 g/cc	RHOB 2.61 g/cc	DT 87.20 us/ft		SGR 41.24 gAPI	PEF 3.00 b/e	NPHI 41.99 pu

Figure 8: Résumé des curves sur plusieurs fichiers ".las"

En observant cet extrait, on constate que le nom du champ Gamma Ray (GR) varie d'un puits à l'autre et se retrouve sous différents noms (GAB, GRD, GAMMA PP). On observe aussi que l'unité du Neutron (PHIN) varie d'un puit à l'autre (Volume/Volume pour certains et en pourcentage pour d'autres).

Ces hétérogénéités constatées nous ont contraint à effectuer des traitements en vue de l'intégration de nos données en base.

4.2 Traitement des fichiers

Comme évoqué dans la partie précédente, nous avons remarqué lors de l'analyse de nos fichiers, que certaines variables identiques pouvaient avoir des noms différents d'un puits à l'autre. Pour palier à ce problème, nous avons décidé de mettre en place un dictionnaire d'alias afin d'uniformiser les noms des différents champs.

Curves	Regroupements réalisés	Description
CAL	SCAL, APPC, CALI, CAL_PP, CALS, CAL, CALX, HCAL, SA, MCAL, CAL	CALIPER
DCOR	SCOR, DCOR, DENC_PP, DRHO, ZCOR, HDRA	DENSITY CORRECTION
DENS	DENS_PP, DEN, RHOB , ZDEN, HD , RHO8, SNB2, DENS, DENS_PP, DEN, RHOB, ZDEN, HD, RHO8	DENSITY
DTC	DT, SONC_PP, DTMN, DTMX, DTL, DTC2, DTC3, DTCO, SONC_PDTC, SONC_PP, DTMN, DTMX, DTL, DTC2, DTC3, DTCO, SONC_P, DT	COMPRESSIONAL SONIC TRAVEL TIME
DTS	DTS, TT, DTSM	SHEAR SONIC TRAVEL TIME
GR	GAMMA_PP, "GRC", "GRD", GAB, GRL, SGRC, SGRD, GR:1, GR:2, EHGR, RGR, SGR, CGR_PP, HGR, GR	GAMMA RAY
PE	SNP2, PEF, PEF8, PE_PP, PE	PHOTO ELECTRIC EFFECT
PHIN	NLIM, PHIN, NPHI, NEUT_PP, TNPL, SNPE, CNC, CN, TNPH, NEU, HNPO	NEUTRON POROSITY
RESD	R55P, SEDP, RESD, RDEP_PP, LLD, ILD, M0R6, M0R9, M0RX, M2R6, M2R9, M2RX, M4R6, M4R9, M4RX, AH090, AH060, RESD, AHO60, AHO90	DEEP RESISTIVITY
RESM	R25P, SEMP, RESM, RSHAL_PP, LLS, ILM, , MSFL, M0R3, M2R3, M4R3, RMED_PP, RLLS, RMLL, AH030, RESM, AHO30	MEDIUM RESISTIVITY
RESS	SESP, RESS, SFLU, MLL, SFL, M0R1, M0R2, M4R1, M4R2, M2R1, M2R2, RS, AH010, AH020, AHO10, AHO20	SHALLOW RESISTIVITY
RESMIC	RESMIC, RMIC_PP, SEXP	MICRO RESISTIVITY
SP	SP, SP_PP, SPD, SPL	SPONTANEOUS POTENTIAL

Figure 9: Alias des curves les plus importantes

Dans cette démarche, nous avons décidé conformément à notre dictionnaire d'alias, de rajouter les champs manquants dans certains fichiers afin d'avoir pour chaque fichier les 12 alias listés plus haut. Par contre, les champs rajoutés dans un fichier ne les contenant pas au départ se sont tous vus attribuer la valeur -9999, pour les distinguer des valeurs NaN (-999.25) contenues dans les fichiers.

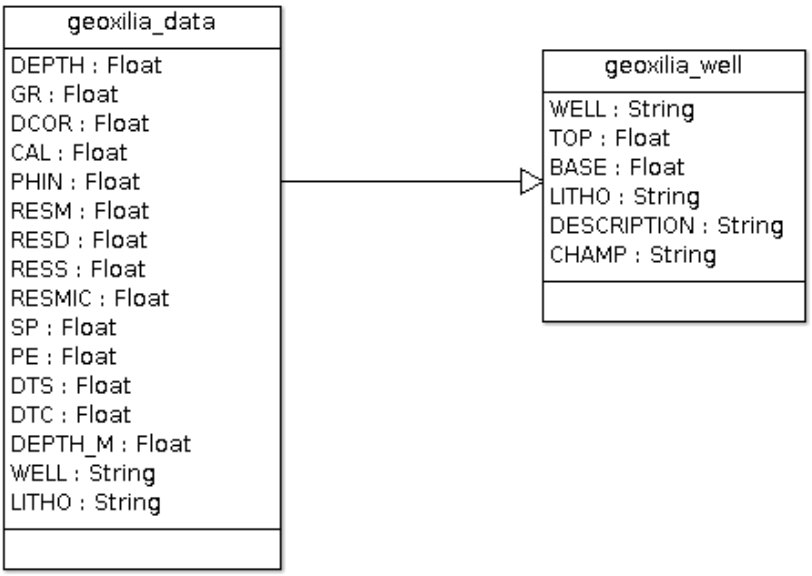
Par la suite, nous avons réalisé **l'uniformisation des unités** parmi les mesures identiques comme :

- Les profondeurs, qui ont été toutes converties des inches en mètres.
- Les mesures de porosité (PHIN) qui ont toutes été ramenées sous forme de ratio car certaines étaient exprimées en pourcentage

Une fois toutes les corrections effectuées, nous avons sauvegardé toutes nos données au format ".csv" afin de faciliter l'insertion de ces dernières en base.

4.3 Création de la base de données

Pour la création de notre base de données, nous avons retenu un schéma contenant 2 tables: - une table "geoxilia_data" contenant les données des différents puits (fichiers ≪.las≫) - une table "geoxilia_well" contenant les données sur les lithologies associées à chaque puit :



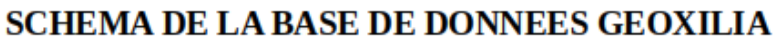


Figure 10: Schéma de la base de données créée

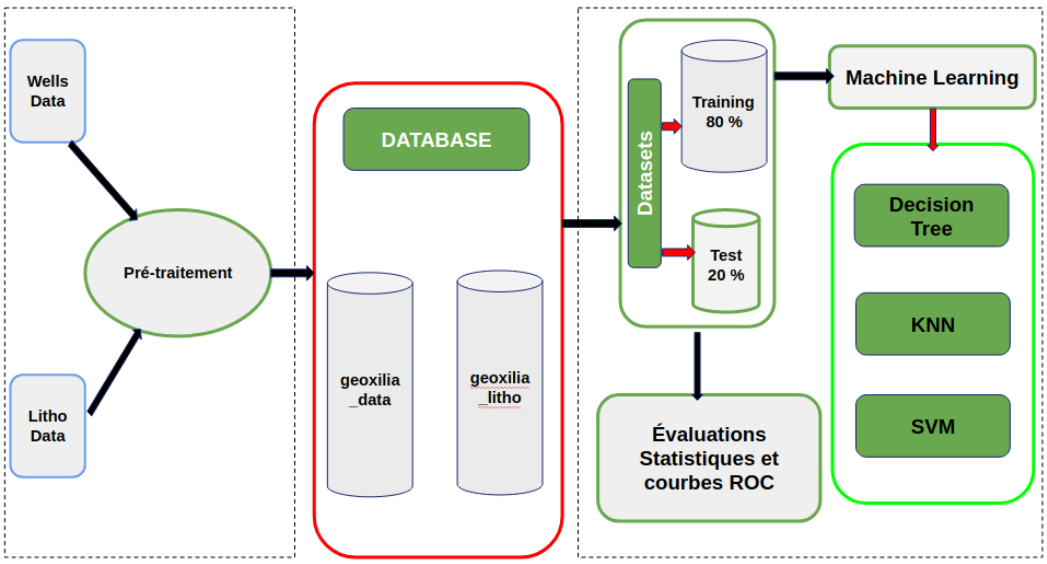


Figure 11: ETL pipeline pour la mise en base des données brutes avant l'analyse

5 Analyse Exploratoire

Dans cette partie, nous explorons dans un premier temps la base de données mise en place avoir une meilleure connaissance de nos données. Par la suite, nous procéderons à une analyse descriptive pour mieux comprendre nos variables.

5.1 Exploration des données

Ci-dessous quelques lignes de la base de données:

DEPTH	GR	CAL	DCOR	DENS	DTC	DTS	PE	PHIN	RESD	RESM	RESS	RESMIC	SP	DEPTH_M	WELL	LITHO
2848.0	30.1875				116.0	-9999.0	-9999.0	0.0	3.6055	3.5996	-9999.0	-9999.0	-44.6875	868.0704000000001	11/30A-A17	LCSD
2848.5	21.8125				117.375	-9999.0	-9999.0	0.0	3.6602	3.2617	-9999.0	-9999.0	-42.9063	868.2228	11/30A-A17	LCSD
2849.0	21.875				118.1875	-9999.0	-9999.0	0.0	3.6719	3.2266	-9999.0	-9999.0	-41.6875	868.3752000000001	11/30A-A17	LCSD
2849.5	17.1875				114.5625	-9999.0	-9999.0	0.0	3.7285	3.1621	-9999.0	-9999.0	-40.9688	868.5276	11/30A-A17	LCSD
2850.0	17.9375				112.5625	-9999.0	-9999.0	0.0	3.5488	3.2422	-9999.0	-9999.0	-39.7188	868.6800000000001	11/30A-A17	LCSD

Figure 12: Extrait de la base de données

Conformément aux décisions prises durant la mise en base, nous observons dans cet extrait qu'il y a bien des valeurs '-9999' et 0.0, correspondant aux des variables initialement manquantes dans certains puits (identifiéa grâce à la variable « WELL »). Par la suite, nous les identifierons comme anomalie de la base de données.

Un zoom a été effectué pour identifier les variables impactées par ces anomalies et les quantifier :

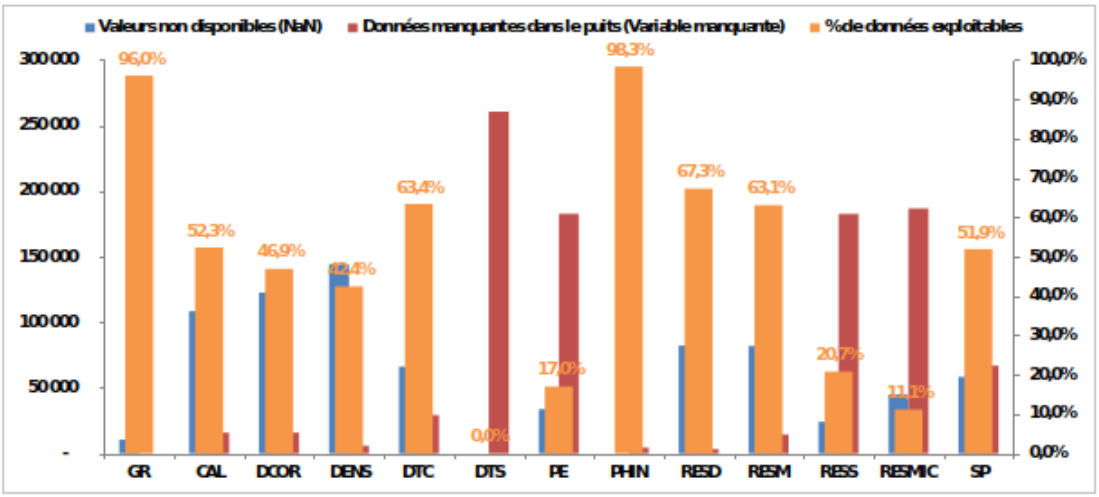


Figure 13: Quantification des anomalies de la base de données

Les résultats montrent que selon les variables considérées, le taux d'exploitabilité des données varie de 0 à 98,3. Autrement dit, le nombre d'observations évolue en fonction des variables sélectionnées. Afin d'avoir un nombre conséquent de données, nous avons fait le choix de mettre en place plusieurs datasets. Ces datasets dépendent de l'importance de chaque variable, selon les recommandations des experts de Geoxilia dans la classification de la lithologie.

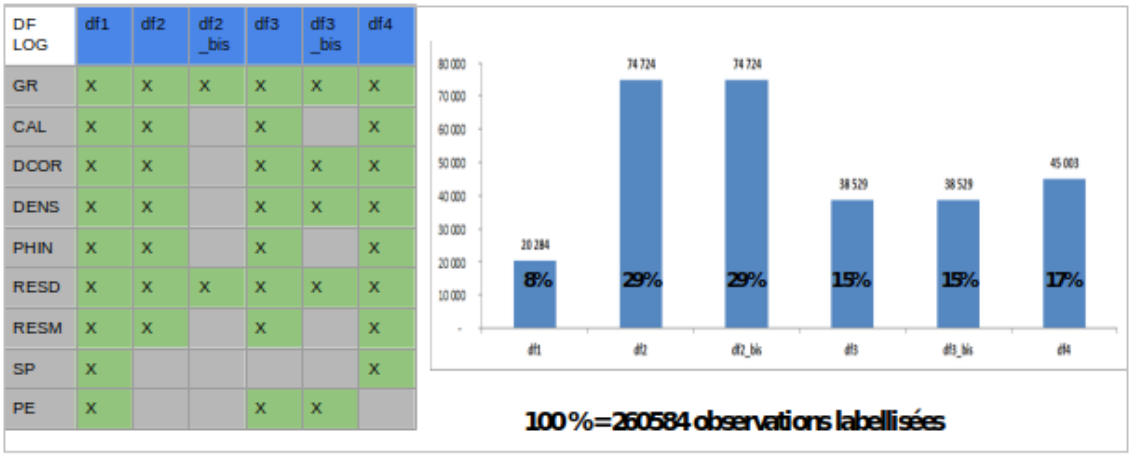


Figure 14: Différents datasets utilisés

Le graphique ci-dessus le nombre d'observations de chaque table, ainsi que le pourcentage par rapport à la taille de la base initiale (qui est de 260584). La variable ≪LITHO≫ contient également des lithologies dynamiques, c'est-à-dire qui évoluent avec le temps: à un instant t donné, la lithologie peut être classifiée ≪sable≫ puis à un instant t+1 classifiée ≪argile≫. Ces lithologies, identifiées par 'DEVH', 'PIPE', 'LI' dans notre base et correspondent à 1652 observations, ont été supprimées car elles ne sont pas pertinentes. De plus, leur faible nombre ne devrait impacter le modèle.

5.2 Statistiques descriptives

Pour cette partie, la méthodologie est la même pour tous les datasets construits ci-dessus. Afin d'éviter les répétitions, nous présenterons simplement quelques résultats obtenus avec le dataset df3_bis.

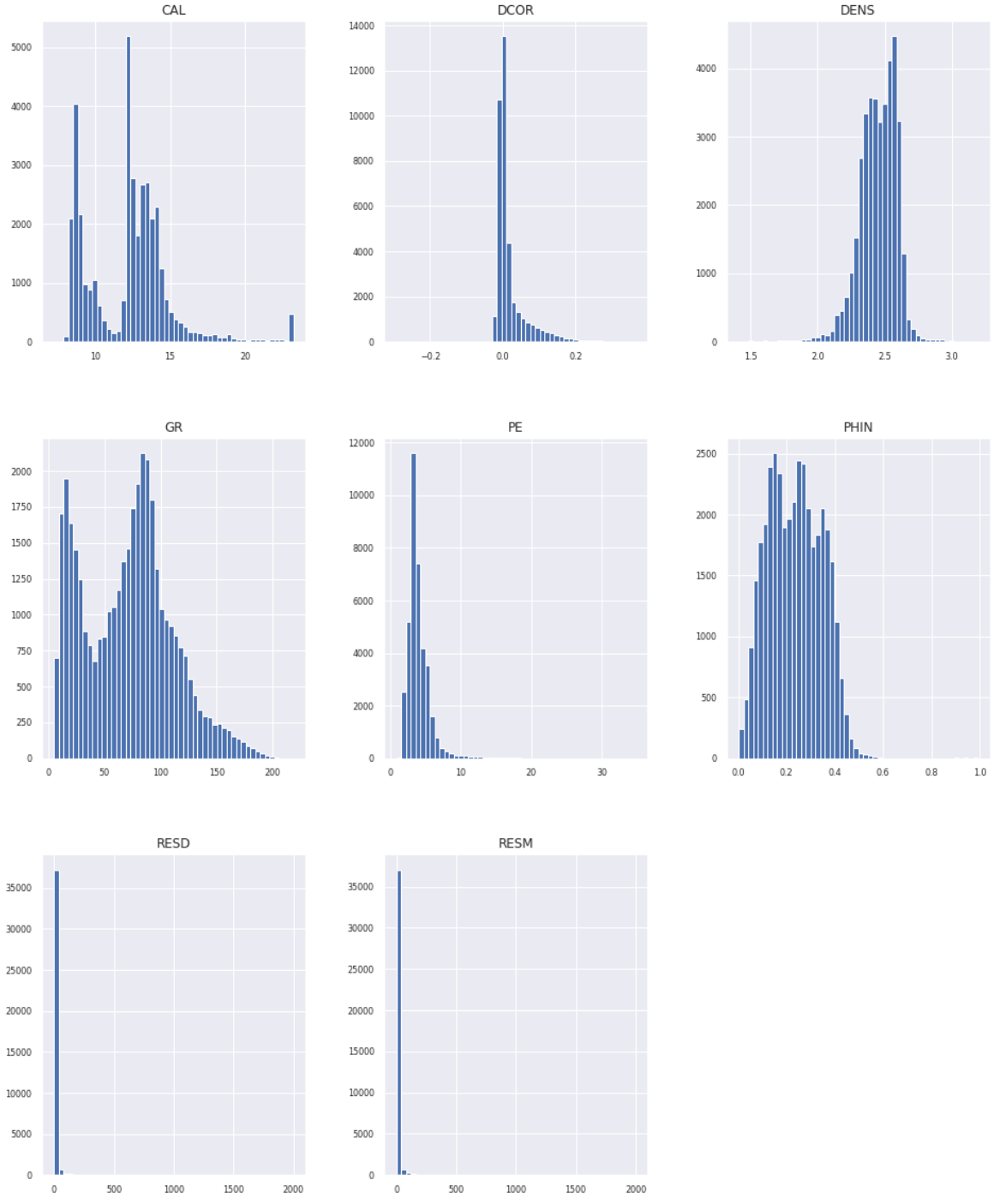


Figure 15: Diagrammes de distribution des variables explicatives

5.3 Matrice de corrélation

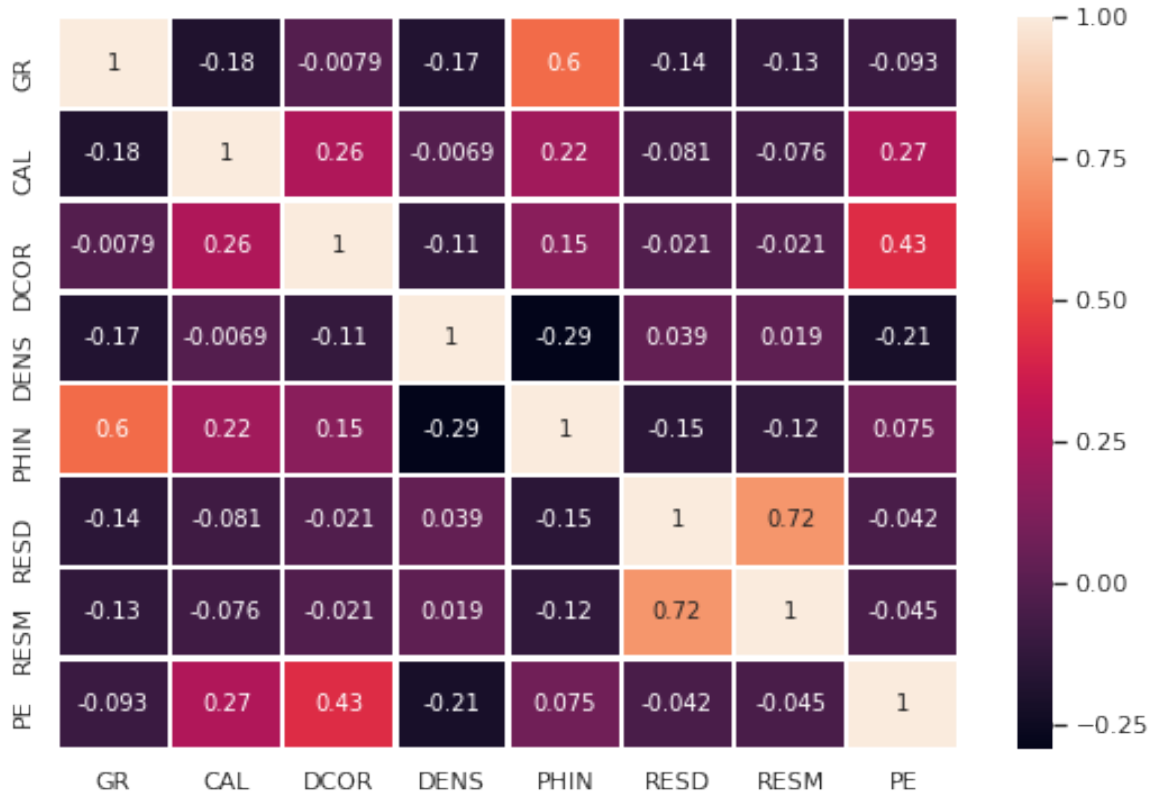


Figure 16: Matrice de corrélation

En observant cettre matrice de corrélation, on note non seulement une forte corrélation entre 'GR' et 'PHIN' mais également entre les variables RESD et RESM, ce qui n'était pas visible dans les autres datasets.

5.4 Visualisation du label en fonction de 3 couples de variables

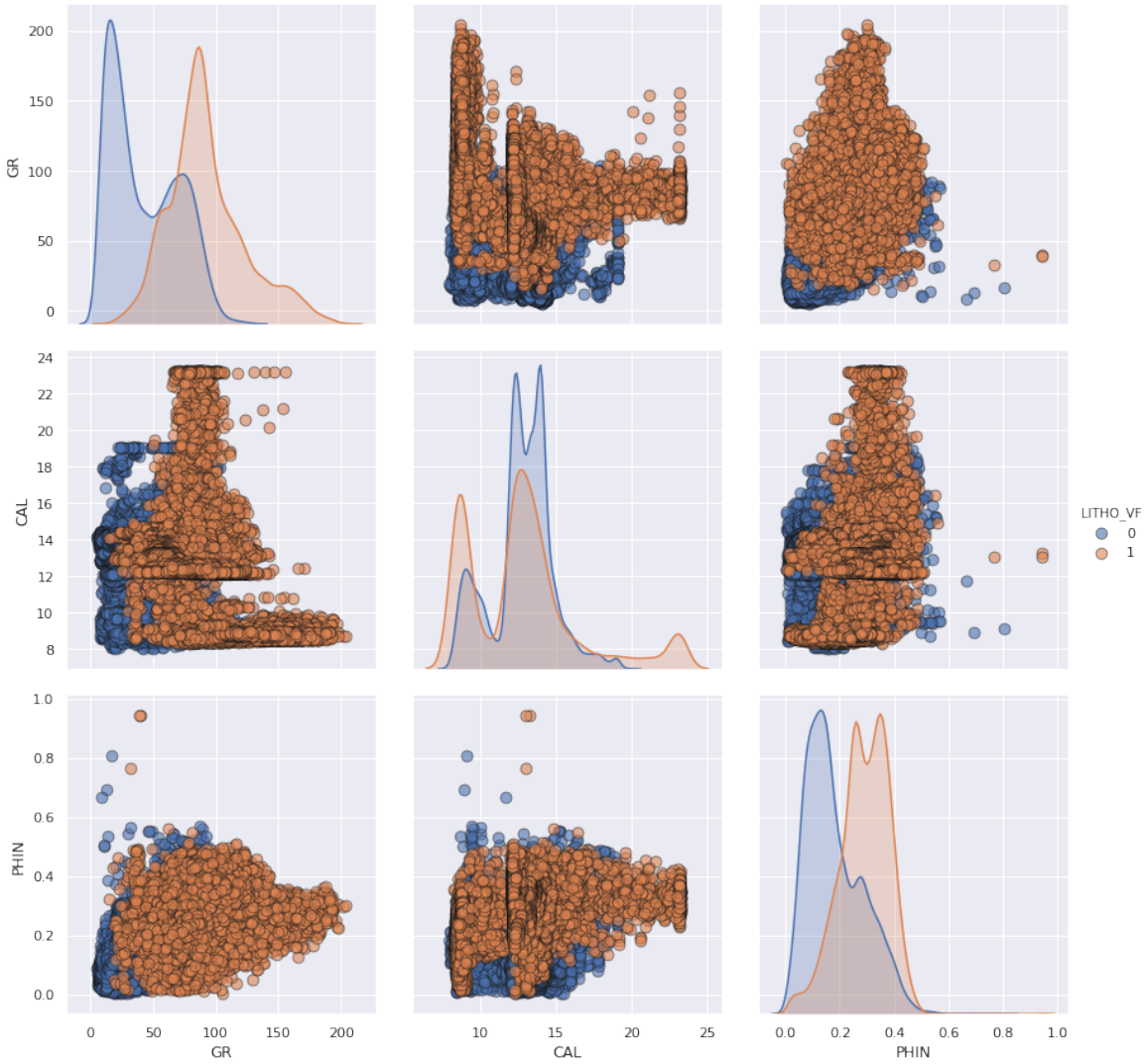


Figure 17: Séparabilité de nos labels

La figure ci-dessus montre que nos données à classifier ne sont pas linéairement séparables. Ainsi, dans la partie Machine Learning, nous utiliserons des modèles adaptés à ce type de données.

6 Premier Algorithmes de Machine Learning

Nous avons utilisé les modéles CART et SVM pour répondre à cette problématique de classification de données non linéairement separables. En effet, l'algorithme CART a été testé sur toutes les bases. Les hyper-paramétres de ces algorithmes CART ont été fixés en procédant à un "tuning" de la profondeur, de 1 à 10 . Nos bases étant de petite taille, pour chaque profondeur testée, une "cross validation" a été réalisée pour récupérer "l'accuracy". La profondeur optimisant l'accuracy a été retenue pour chaque modèle CART. Les indicateurs tels que:

- la "précision" : rapport des observations correctement prédites positives et le nombre total des observations prévues positives
- le "recall" : rapport des observations correctement prédites positives et le nombre total des observations dans la classe actuelle
- le "f1-score" : moyenne pondérée de la précision et du recall
- l'"accuracy" : rapport des observations correctement prédites et le nombre total des observations

Pour comparer ces algorithmes CART, nous avons jugé pertinent d'utiliser le modèle SVM qui est très adapté à ce genre de problématique:

Ci-après la synthèse des résultats obtenus avec ces différents modèles:

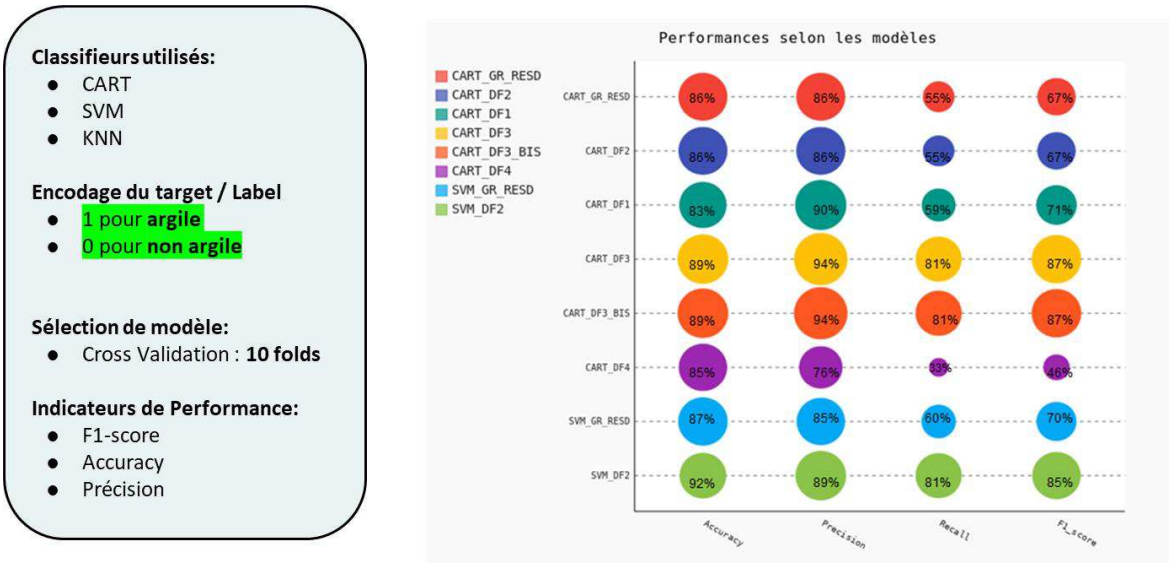


Figure 18: Comparaison des performances des classifieurs

Les modèles CART_DF3_bis (modèle CART réalisé à partir de la base df3_bis) et SVM_DF2 (SVM réalisé à partir de la base df3) présentent les indicateurs de performance les plus intéressants. Le choix du modèle **CART_DF3_bis** se justifie du fait que:

- Sur ce projet, nous avons décidé de donner plus d'impact au risque de prédire argile alors que c'est du non argile (Investissement à perte par l'entrepreneur) que l'inverse (manque à gagner de la part de l'entrepreneur).Ce premier risque est mesuré par la précision :

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

avec TP: True Positive et FN: False Negative. C'est le modèle CART_DF3_bis qui **minimise le mieux** ce risque d'investissement à perte via sa précision qui est la plus élevée.

- **L'interprétation des résultats du CART**: Les variables qui discriminent le mieux les classes "Argile" et "Non argiles" sont montrées par le modèle CART ainsi que leur pouvoir à classifier

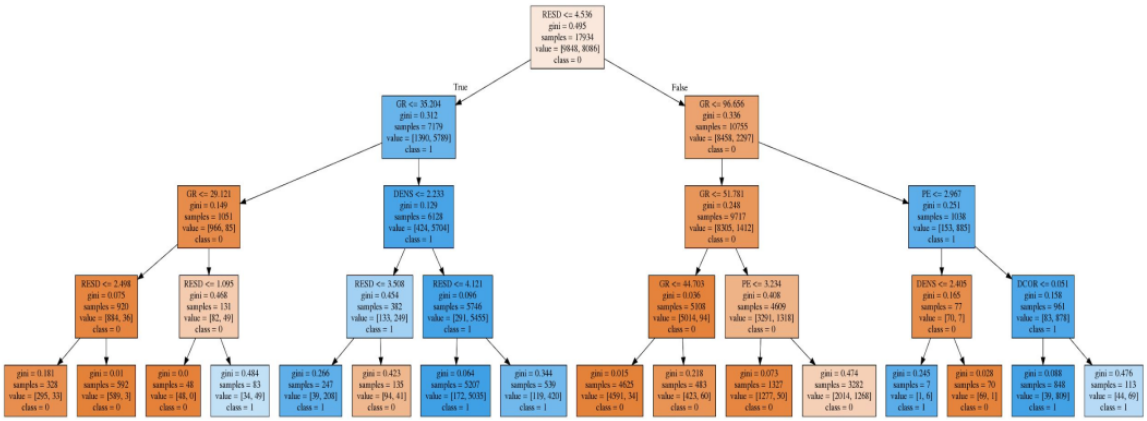


Figure 19: Arbre de décision

- **Sa cohérence par rapport aux règles métier utilisées actuellement**: en effet, les variables retenues par ce modèle sont celles utilisées par le métier, à l'exception de la variable "DCOR" (qui a un pouvoir de discrimination pas très élevé dans notre modèle).

Variables sélectionnées:
GR, DCOR, DENS, RESD, PE

Base utilisée: *df3_bis*

CART :
hyperparamètre : profondeur de 4

indicateurs de performances:
1. **Accuracy Score** : 0.85
2. **F1-score**: 87%
3. **Précision**: 94%

Modèle de sélection:
Cross Validation : 10

Avantage *:*
Performant et facile à interpréter

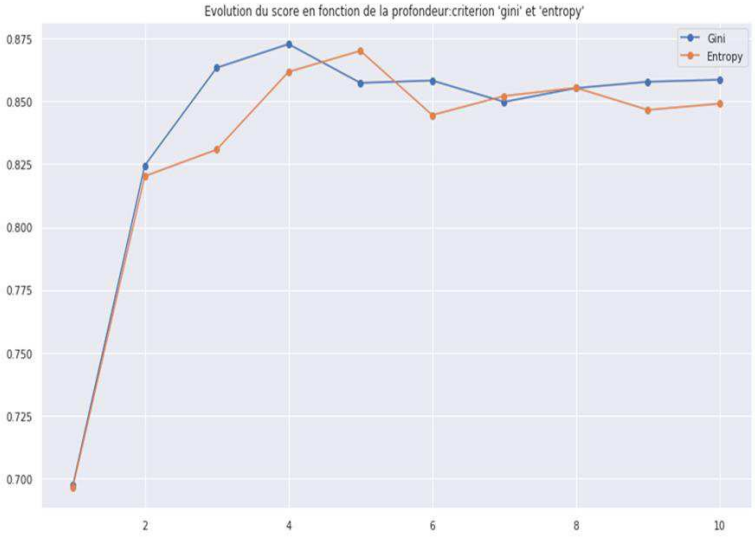


Figure 20: Algorithme CART_DF3_bis et choix de la profondeur

7 Conclusion et perspectives

Durant cette première partie, si la majorité du temps a été passée sur la phase de cleaning, cet investissement a néanmoins porté ses fruits. Une base de données propre étant le socle de tout travail de data scientist, celle-ci nous a en effet permis d'effectuer une analyse exploratoire cohérente et d'obtenir des premiers résultats prometteurs. Durant les cinq prochains mois, nous explorerons en profondeur les axes suivants:

7.1 Apprendre les règles métiers

Le flux de travail traditionnel repose sur un paramétrage manuel du modèle géologique, à l'aide d'équations physiques et d'une succession de décisions relevant de l'expérience métier.

- Nous pouvons essayer d'encoder ces différentes équations et règles métiers, afin de comparer leurs résultats avec nos modèles de Machine Learning.
- En l'occurrence, ces modèles devraient avoir le potentiel d'être plus précis, car ces derniers peuvent être entraînés en utilisant des jeux de données bien plus grands que ce que les êtres humains peuvent manipuler. La qualité de la prédiction pourrait également augmenter, puisque dans une approche orientée data, nous diminuons le biais humain.
- Un tel modèle permet un diagnostic rapide des incertitudes et sensibilités de modélisation pour mettre en lumière les éléments critiques pouvant affecter la fiabilité du modèle.
- Enfin, les différentes règles utilisées peuvent être apprises directement depuis les données. Avec un accès à des données de qualité, nous pourrions rapidement inférer un large panel de propriétés.

7.2 Gérer les valeurs manquantes et explorer les algorithmes non supervisés

Bien que le potentiel du Machine Learning en geoscience soit prometteur, il existe quelques problèmes pratiques. Si les jeux de données utilisés peuvent être volumineux, les labels (lithologies) nécessaires pour entraîner les modèles sont peu nombreux et difficiles à produire, comme nous avons pu le constater. L'amélioration des performances de nos modèles supervisés est ainsi limitée par la rareté des labels, et le coût pour les obtenir.

- Les méthodes non supervisées pourront certainement être une piste pour contourner ce problème.
- Des algorithmes capables de gérer les valeurs manquantes seront également étudiées.
- Une autre approche intéressante serait d'utiliser les labels comme variables explicatives afin de prédire les valeurs manquantes dans les logs.

7.3 Prédire le volume d'hydrocarbures

Enfin, nous espérons terminer ce projet avec des premiers travaux sur la prédiction de volumes d'hydrocarbures grâce à des approches Machine Learning.

- La complexité géologique des réservoirs rendant souvent impossible la modélisation exacte avec seulement des mesures de logs, cette étude nécessitera des données supplémentaires.

7.4 Ressources en calcul

- Si toutes les opérations ont pu fonctionner en local jusqu'à présent, certains des points cités plus haut nécessiteront probablement une puissance de calcul importante. si ce cas venait à se présenter, nous utiliserons les services d'Amazon Web Services. Nous pourrons mettre à profit les crédits offerts par la plateforme aux étudiants pour réaliser l'ensemble de ces calculs.

References

[1] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.

[2] Albert Einstein. *Zur Elektrodynamik bewegter Körper*. (German) [*On the electrodynamics of moving bodies*]. Annalen der Physik, 322(10):891–921, 1905.

[3] Diagraphie - well logs
`https://fr.wikipedia.org/wiki/Diagraphie`

[4] Diagraphie, géophysique
`https://www.universalis.fr/encyclopedie/diagraphies-geophysique`

[5] Earth Science Australia
`http://earthsci.org/mineral/energy/fuels/fuels.html`

[6] E.R. (ROSS) CRAIN, P.Eng. [*Crain's Petrophysical Handbook*. (English)]
`https://www.spec2000.net/`

[7] Eirik Larsen, S.J. Purves, D. Economou and B. Alaei *Is Machine Learning taking productivity in petroleum geoscience on a Moore's Law trajectory?*. (English)
`https://www.researchgate.net/publication/329697568`

[8] B. Michel *Modélisation de la production d'hydrocarbures dans un bassin pétrolier*. Mathématiques [math]. Université Paris Sud - Paris XI, 2008. Français. tel-00345753f