fil rouge

MACHINE LEARNING FOR GEOMODELLING

BABACAR DIOUF

MOHAMMED OUEDRHIRI

PARFAIT FANGUE

HIROTO YAMAKAWA

PRÉSENTATION DE l'ÉQUIPE

Geoxilia
Startup spécialisée en géosciences constituée de géologues, géophysiciens et des ingénieurs réservoirs. Spécialisée dans l'analyse et la prédiction des réservoirs d'hydrocarbures.

BearingPoint
Cabinet indépendant de conseil en management et en technologie, proposant une offre de solutions en data science depuis 2012, avec l'acquisition de HyperCube.

Jamyl BRAHAMI
Managing Director

Adeline AUVINET
Senior Geologist

Pauline MAURY
Lead Data Scientist

Eya KALBOUSSI
Data Scientist

Groupe Fil Rouge
Babacar DIOUF
Mohammed OUEDRHIRI
Parfait FANGUE
Hiroto YAMAKAWA

Tuteur académique:
Thomas BONALD

PLAN

1. Le contexte et les enjeux
2. La démarche
3. Preprocessing
4. Analyse Exploratoire
5. Machine Learning
6. Prochaines étapes

LE CONTEXTE ET LES ENJEUX

Le GEOMODELLING ?

Problématiques :

Peut-on trouver des **alternatives plus efficaces** aux **méthodes traditionnelles** utilisées par les experts durant la réalisation d'un modèle géologique ?

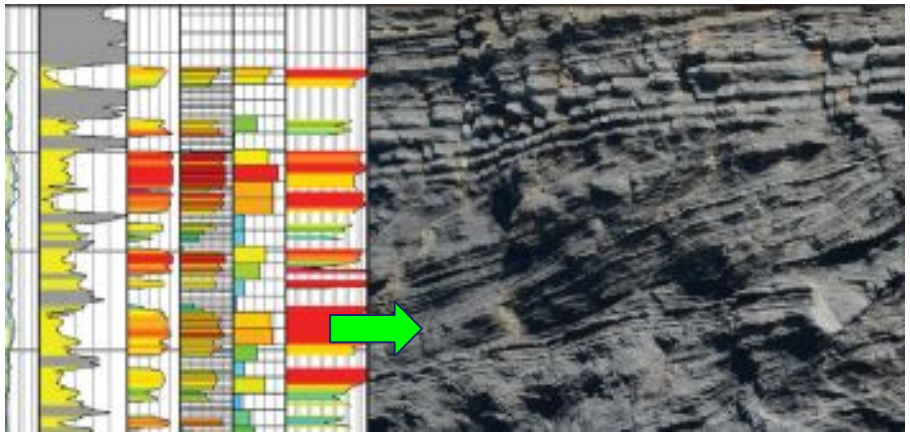
Objectifs :

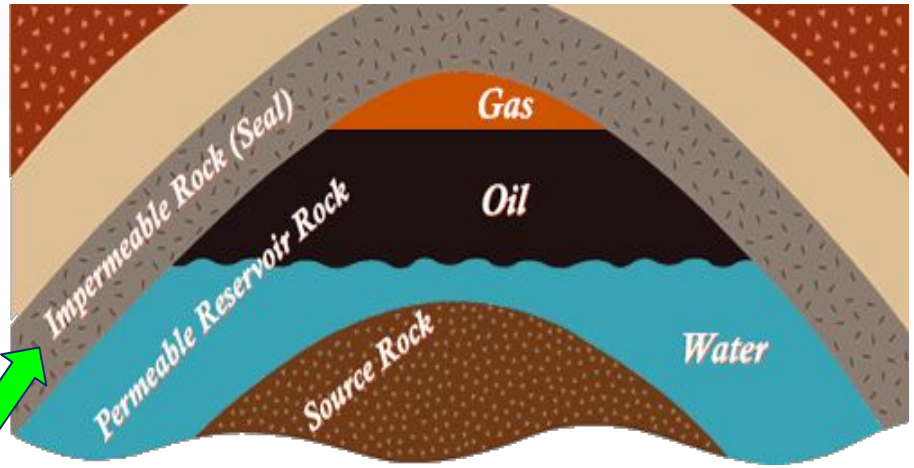
Explorer et entraîner des algorithmes de Machine Learning pour **optimiser la phase de modélisation** et anticiper les étapes critiques

LABEL / TARGET : Lithologie = couche de roche Ex :{**ARGILE**}

Études portées sur les réservoirs de la mer du Nord

Quatre champs : BEATRICE / CROMARTY / JACKY / ALWYN





LE CONTEXTE ET LES ENJEUX

Logs importants selon Geoxilia - **logs (= Variables explicatives)**

1	2	3	4	5	6	7
CALIPER (CAL) Diamètre du puit	**GAMMA RAY (GR)** radioactivité naturelle	**DENSITE (DENS & DCOR)** Densité et densité corrigé	**PHOTO ELECTRIC (PEF)** absorption photo-électrique	**NEUTRON (PHIN)** contenu d'hydrogène	**SONIC (DTC)** ralentissement de la vitesse des ondes de compression	**RESISTIVITE (RESD, RESM)** résistivité à différentes profondeurs latérales

Remarque : à l'échelle des réservoirs d'hydrocarbures de la mer du Nord, **la profondeur** des champs étudiés n'est pas assez élevée. Nous faisons l'hypothèse que ce paramètre n'a pas d'influence sur l'identification des lithologies.

Note : Mesures prises tous les 0,5 m sur le long du puit (de 1000 à 4000 m) pendant ou après le forage.

PLAN

1. Le contexte et les enjeux
2. **La démarche**
3. Preprocessing
4. Analyse Exploratoire
5. Machine Learning
6. Prochaines étapes

LA DÉMARCHE

Compréhension du métier
Réunion avec Geoxilia et Bearingpoint pour apprendre sur le métier et le travail des géophysiciens

Mid OCT 2019

Early NOV 2019

Cleaning des données
Uniformisation des données transmises par Geoxilia sous un format exploitable.

Late DEC 2019

Mise en base
Créer un dataset d'apprentissage avec les données de champs matures sur plusieurs puits.

Analyse exploratoire
Identification des variables explicatives importantes pour sélectionner le modèle approprié

Mid JAN 2020

Late JAN 2020

TODAY

Premiers algorithmes
Comparaison des modèles de Machine Learning pour identifier les lithologies: Algorithme supervisé

PLAN

1. Le contexte et les enjeux
2. La démarche
3. **Preprocessing**
4. Analyse Exploratoire
5. Machine Learning
6. Prochaines étapes

PREPROCESSING - Quelques chiffres

Données

- 114 fichiers (.las) contenant les logs sur différents puits (**68 puits distincts**)
- 3 fichiers (.xls) contenant les lithologies associées à chaque profondeur (**78 lithologies**)

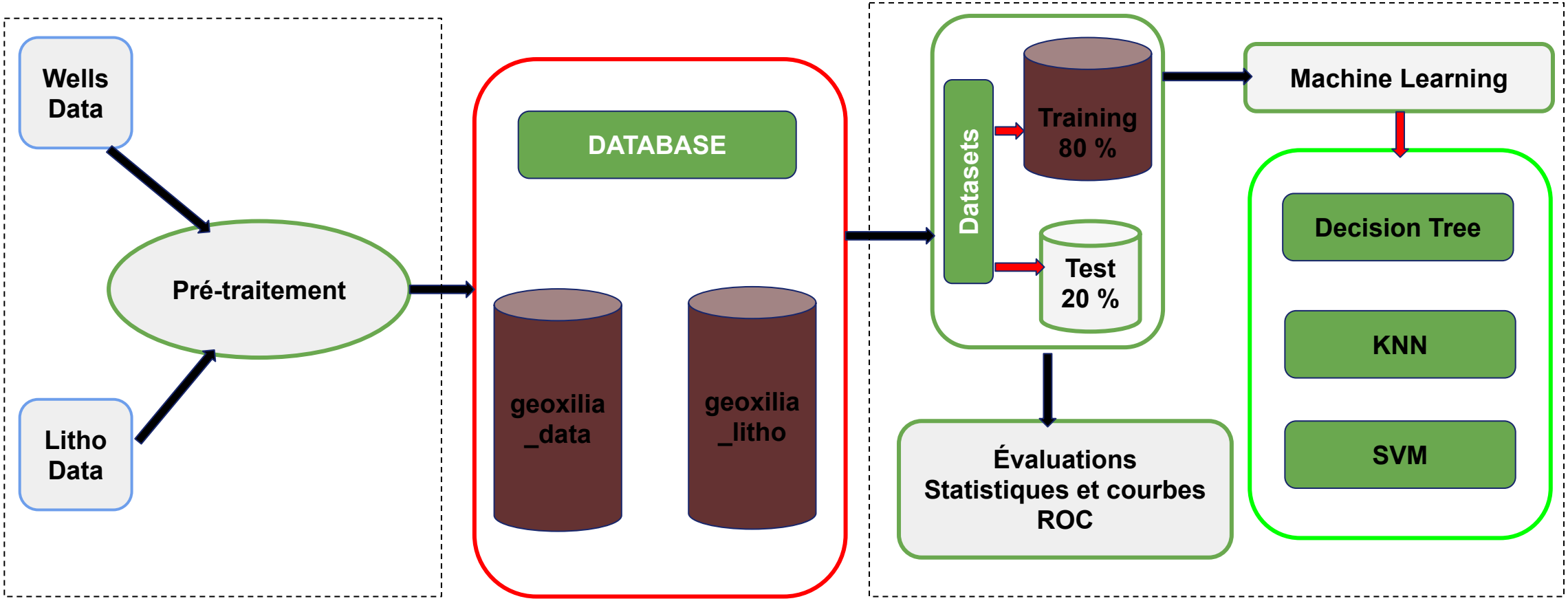
Problèmes

- **Données inexploitables** en l'état pour de la prédiction
- Données **très hétérogènes**
- **Diversité** des unités pour une même mesure physique

Solutions

- **Pré-traitement des données** et mise en place d'une base de données sur SQLITE (Faible volume de données < 1 GB)

PREPROCESSING - Quelques chiffres



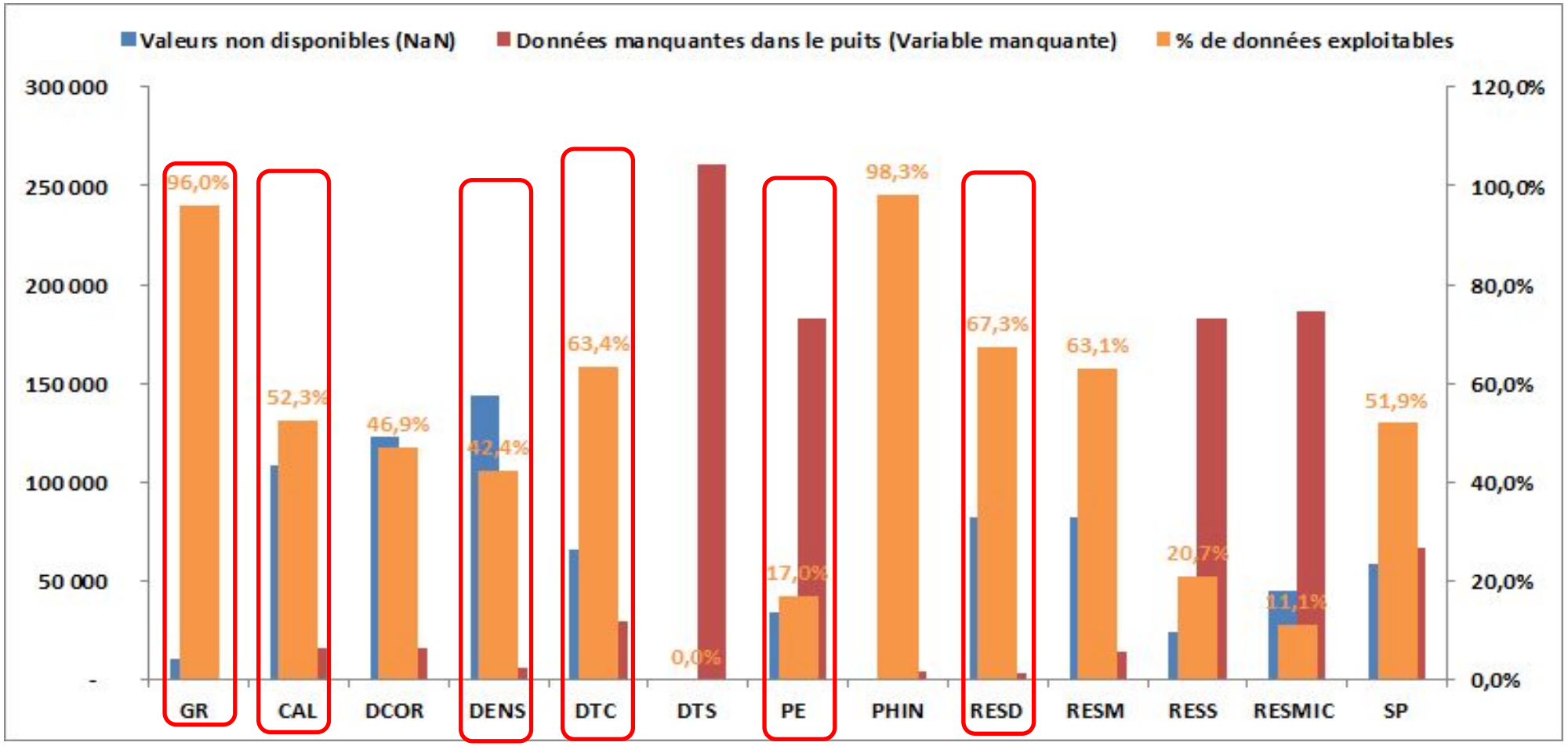
PREPROCESSING - Quelques chiffres

Observations	Pourcentage	Description
698355	**100%**	Nombre total d'observations dans la base de données
260584	**37%**	Observations labellisées

PLAN

1. Le contexte et les enjeux
2. La démarche
3. Preprocessing
4. **Analyse Exploratoire**
5. Machine Learning
6. Prochaines étapes

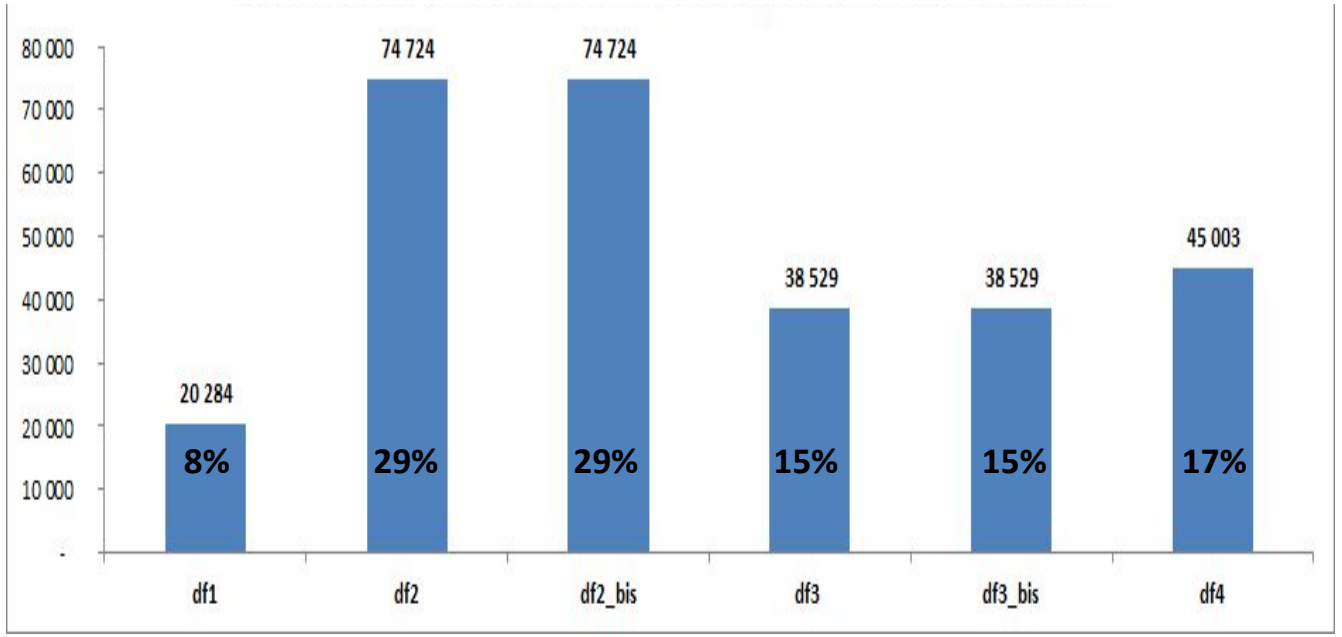
ANALYSE EXPLORATOIRE



ANALYSE EXPLORATOIRE

Différentes bases utilisées selon l'exploitabilité des variables

DF LOG	df1	df2	df2 _bis	df3	df3 _bis	df4
GR	X	X	X	X	X	X
CAL	X	X		X		X
DCOR	X	X		X	X	X
DENS	X	X		X	X	X
PHIN	X	X		X		X
RESD	X	X	X	X	X	X
RESM	X	X		X		X
SP	X					X
PE	X			X	X	



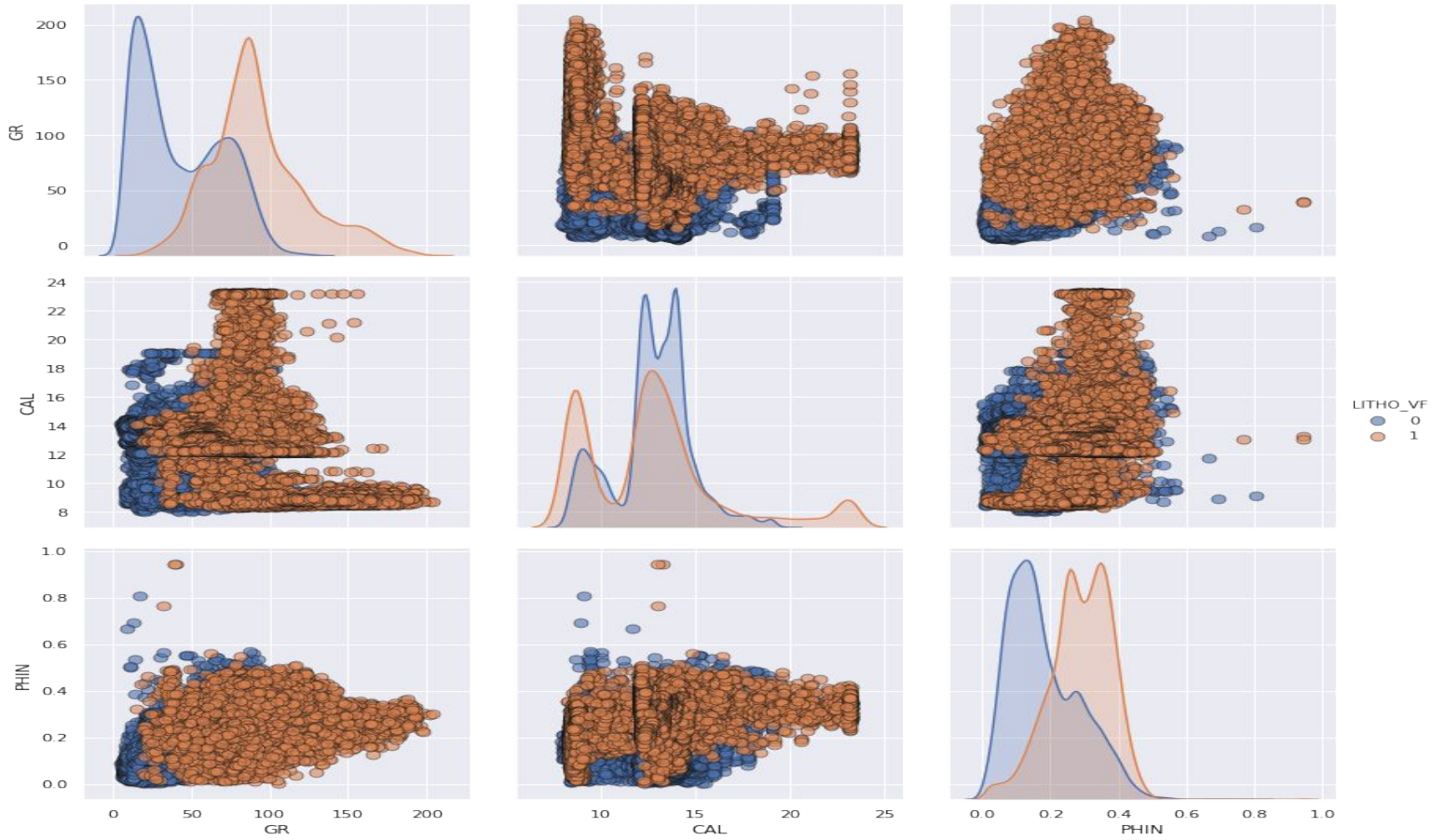
100 % = 260584 observations labellisées

ANALYSE EXPLORATOIRE

Matrix Correlation

ANALYSE EXPLORATOIRE

Classifieurs non linéaires mieux adaptés à nos données



PLAN

1. Le contexte et les enjeux
2. La démarche
3. Preprocessing
4. Analyse Exploratoire
5. **Machine Learning**
6. Prochaines étapes

MACHINE LEARNING – Premiers résultats

Classifieurs utilisés:

- CART
- SVM
- KNN

Encodage du target / Label

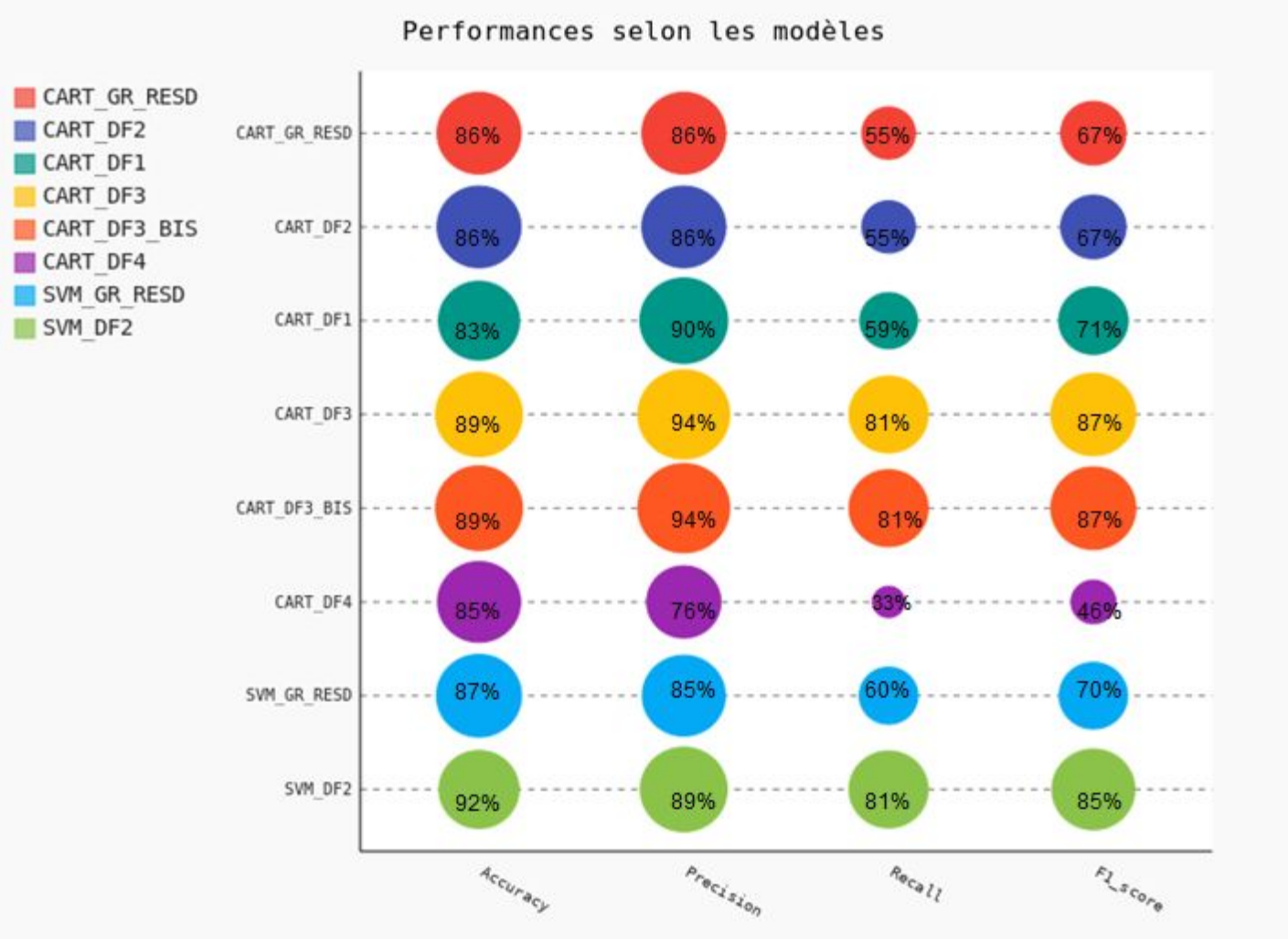
- 1 pour **argile**
- 0 pour **non argile**

Sélection de modèle:

- Cross Validation : **10 folds**

Indicateurs de Performance:

- F1-score
- Accuracy
- Précision



MACHINE LEARNING – Zoom sur le modèle CART retenu

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

Base utilisée: *df3_bis*

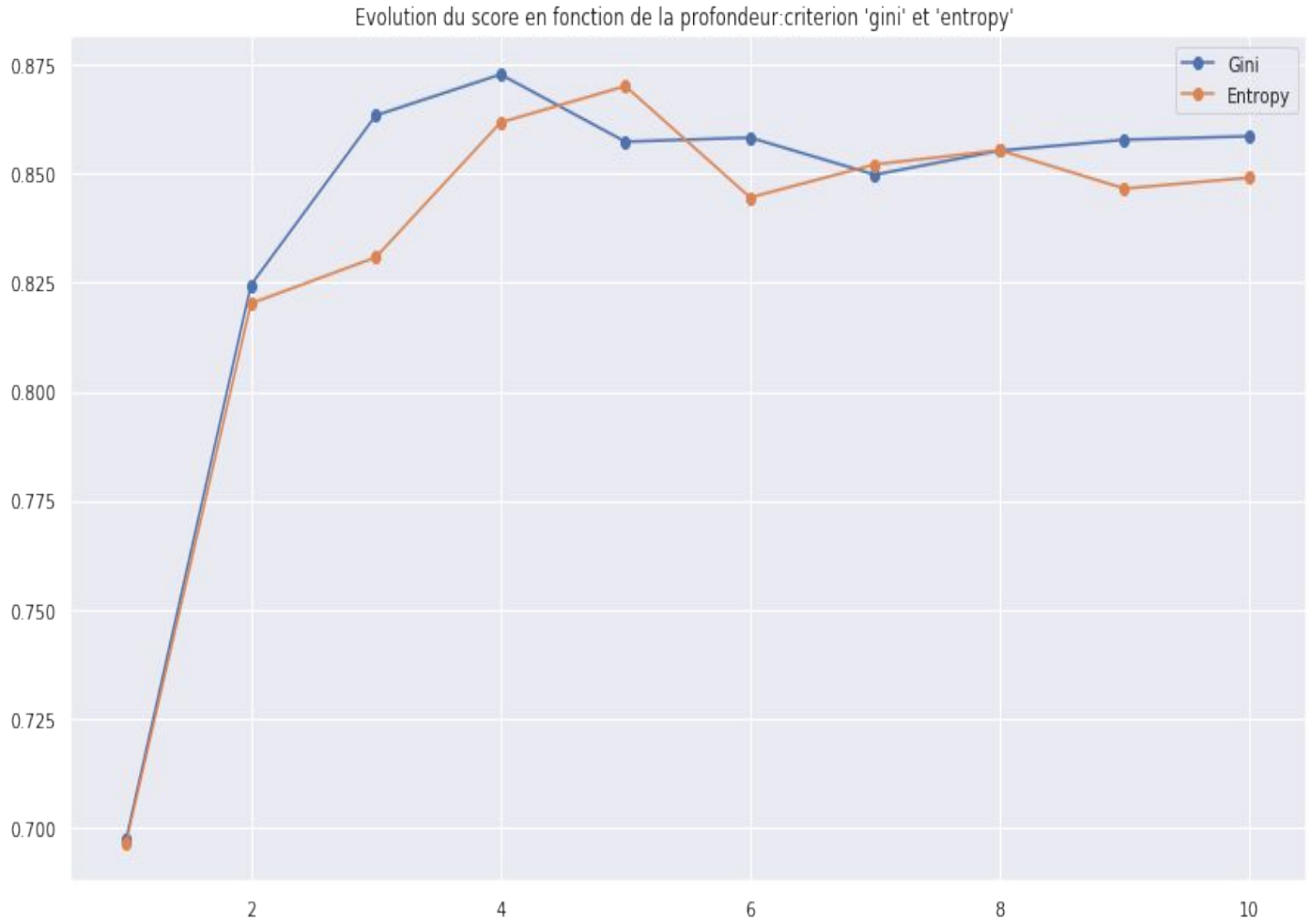
CART :
hyperparamètre : profondeur de 4

indicateurs de performances:

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

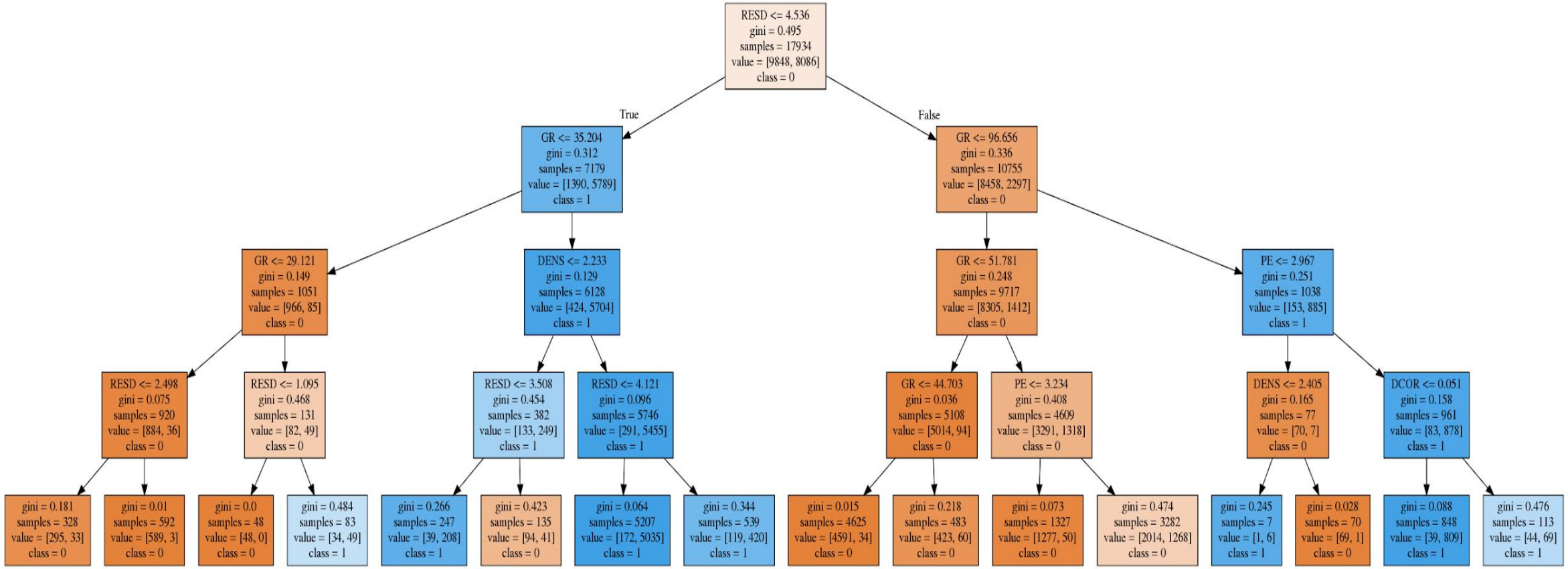
Modèle de sélection:
Cross Validation : 10

Avantage *:*
Performant et facile à interpréter



MACHINE LEARNING – Modèle sélectionné VS règles métiers

Résultats du modèle avec les mêmes variables utilisées par les règles “métier”: arbre de classification



PLAN

1. Le contexte et les enjeux
2. La démarche
3. Preprocessing
4. Analyse Exploratoire
5. Machine Learning
6. **Prochaines étapes**

PROCHAINES ÉTAPES - perspectives

Comparaison du modèle avec les règles métier → Modèles non supervisés → Modèles capable de gérer les NaN → Prédiction des volumes

Comparer les performances de notre modèle aux règles métier utilisées pour identifier les lithologies.

Exploitation de **modèles non supervisés** sur le jeu de données non labellisées.

Mettre en place un algorithme permettant de remplacer les valeurs nulles.

S'intéresser à la **prédiction des volumes** d'hydrocarbures dans les réservoirs.

RESSOURCES

- Pour calculs imposants : **AWS - ElasticMapReduce**
- *Utilisation des crédits restants sur les comptes Educate*

MERCI

ANNEXE : Modèle CART avec df1

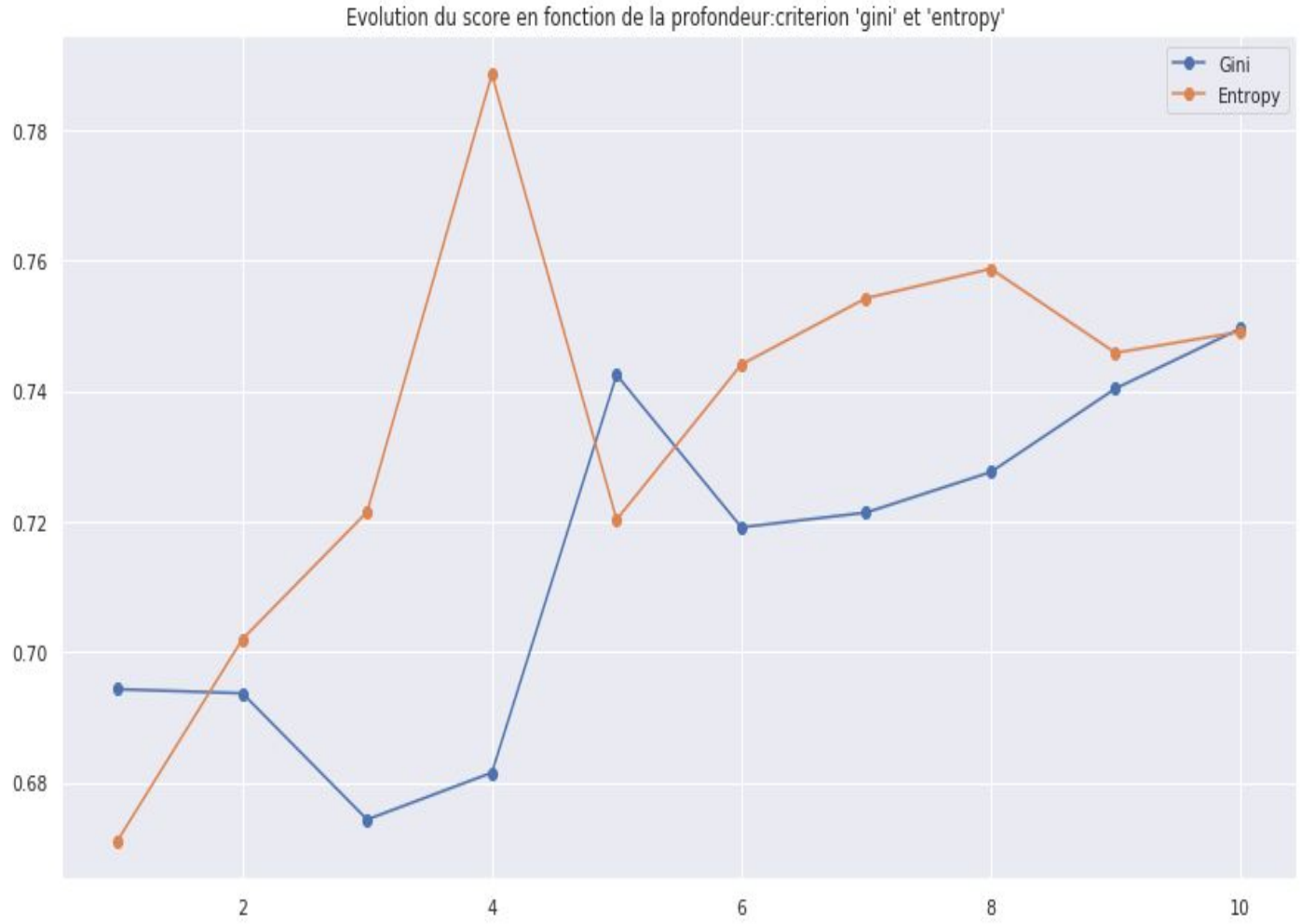
Variables sélectionnées:
GR, CAL, DCOR, DENS, PHIN, RESD, RESM, SP, PE
Base utilisée: df1

CART :
hyperparamètre :
profondeur de 5

indicateurs de performances:

1. **Accuracy Score** : 0.83
2. **F1-score**: 71%
3. **Précision**: 90%

Modèle de sélection:
Cross Validation : 10



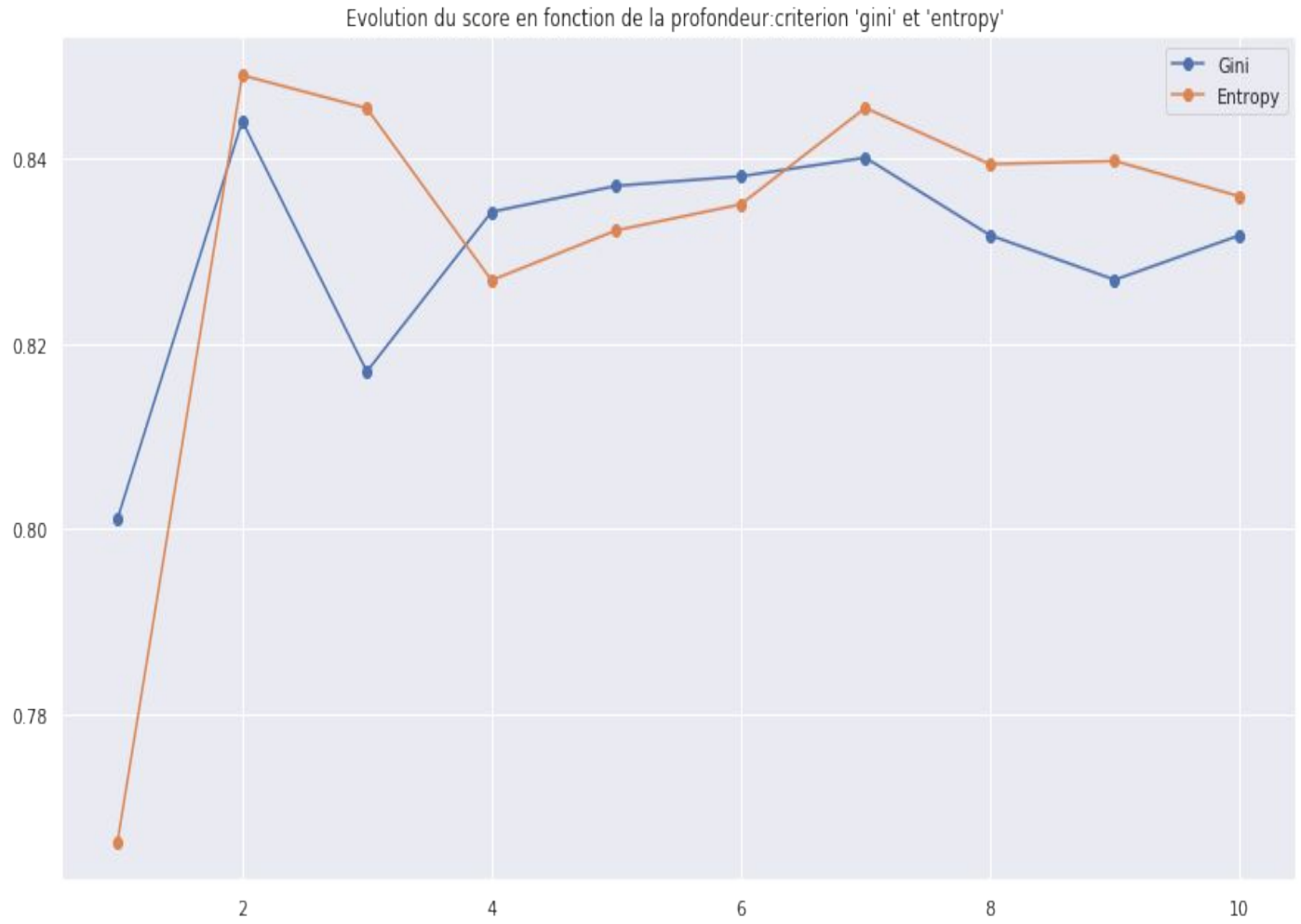
ANNEXE : Modèle CART avec df2

Variables sélectionnées:
GR, CAL, DCOR, DENS, PHIN, RESD, RESM
Base utilisée: df2

CART :
hyperparamètre :
profondeur de 2

indicateurs de performances:
1. **Accuracy Score** : 0.86
2. **F1-score**: 67%
3. **Précision**: 86%

Modèle de sélection:
Cross Validation : 10



ANNEXE : Modèle CART avec df2_GR_RESD

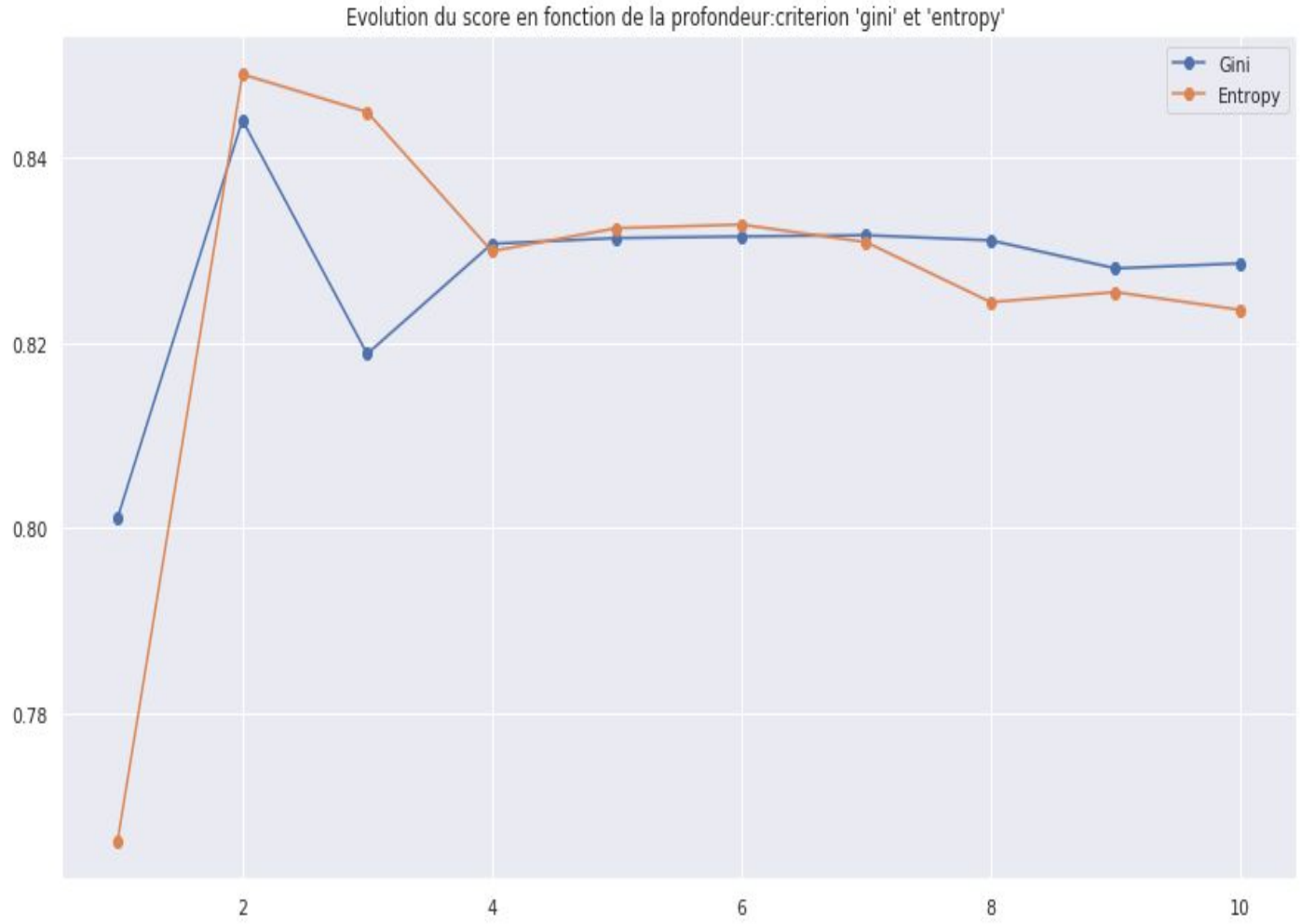
Variables sélectionnées:
GR, RESD

Base utilisée: df3_bis

CART :
hyperparamètre :
profondeur de 4

indicateurs de performances:
1. **Accuracy Score** : 0.86
2. **F1-score**: 67%
3. **Précision**: 86%

Modèle de sélection:
Cross Validation : 10



ANNEXE : Modèle CART avec df3

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

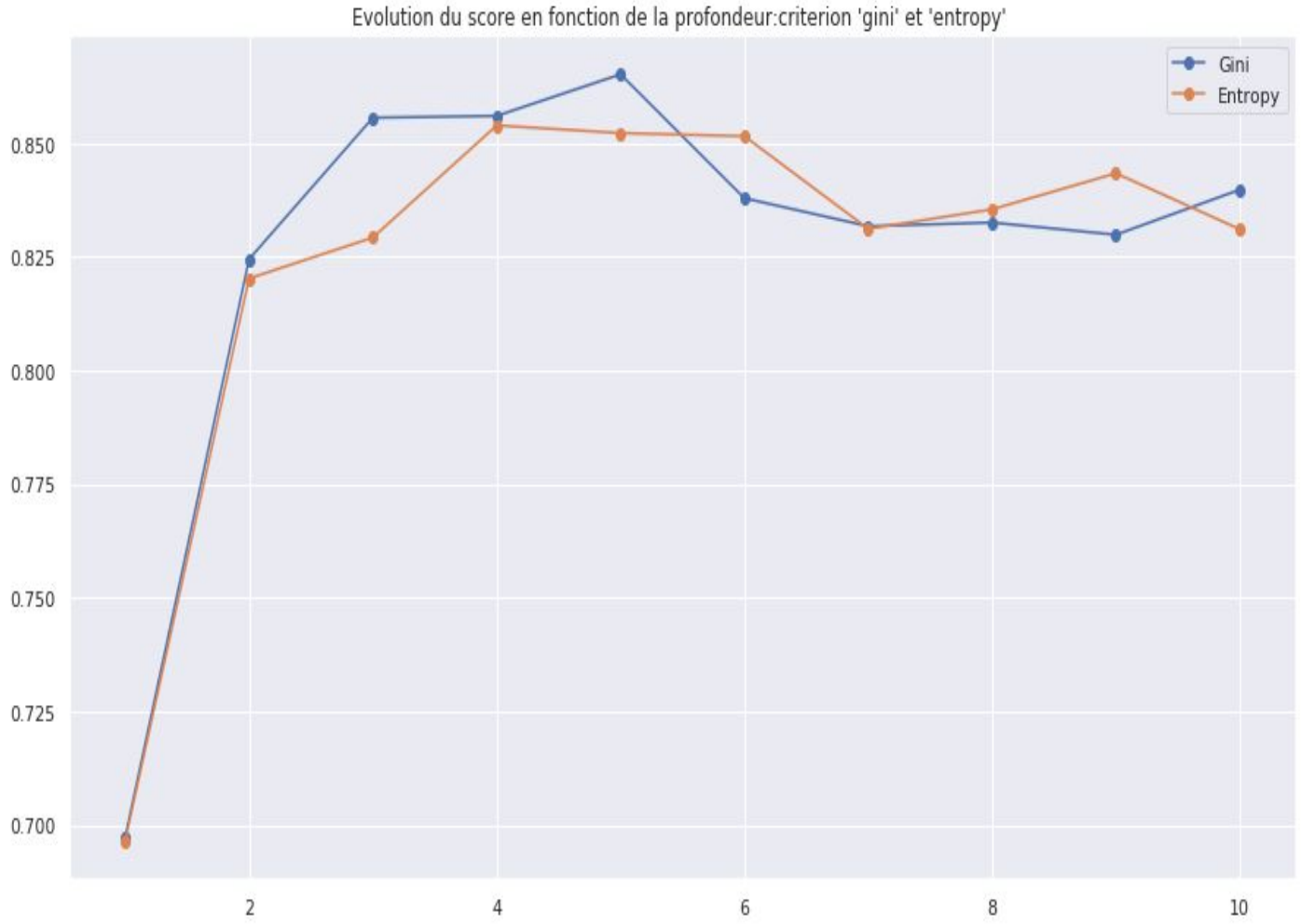
Base utilisée: df3

CART :
hyperparamètre :
profondeur de 5

indicateurs de performances:

1. **Accuracy Score** : 0.89
2. **F1-score**: 87%
3. **Précision**: 94%

Modèle de sélection:
Cross Validation : 10



ANNEXE : Modèle CART avec df4

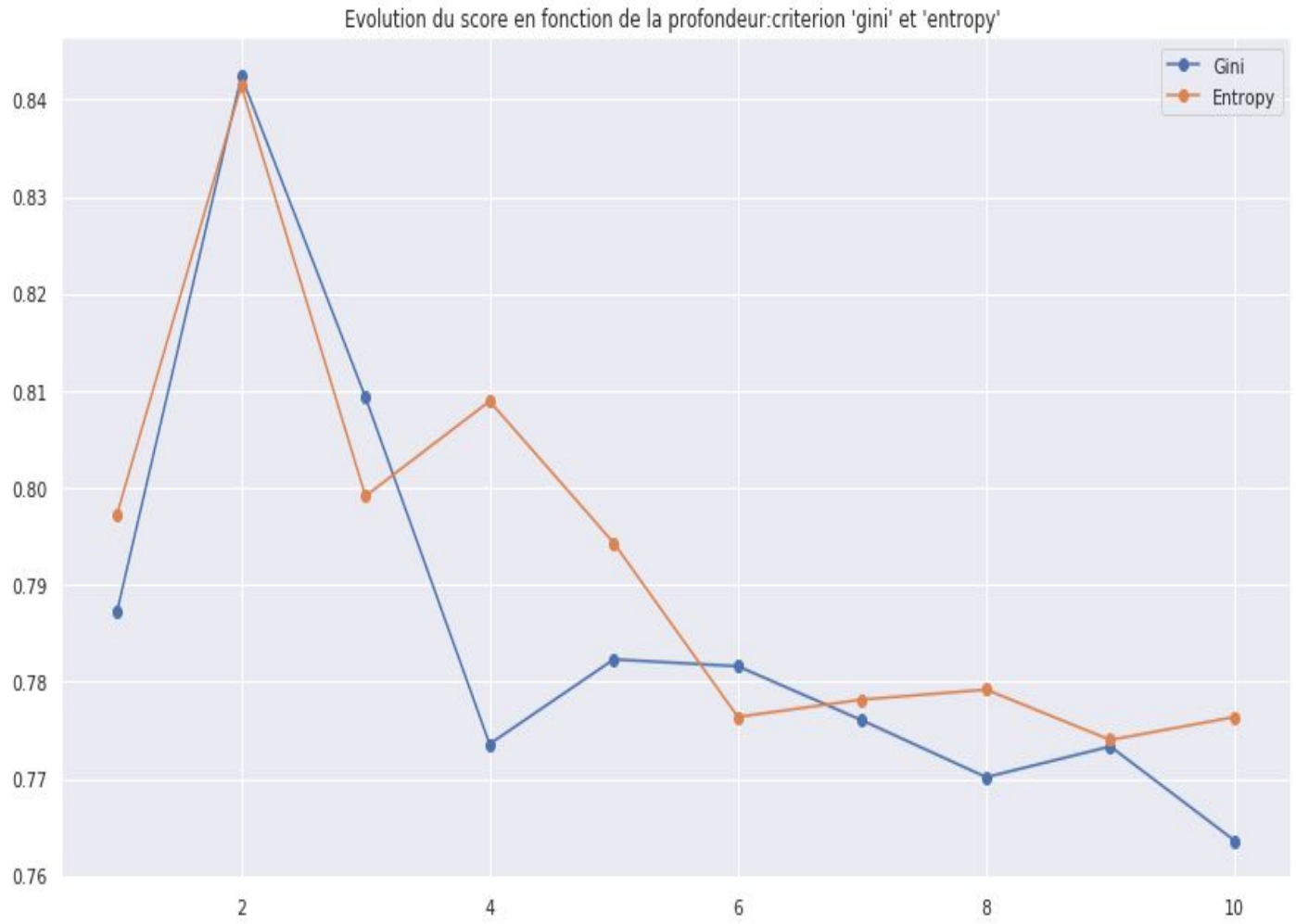
Variables sélectionnées:
GR, DCOR, DENS, RESD, PE

Base utilisée: df4

CART :
hyperparamètre :
profondeur de 5

indicateurs de performances:
1. **Accuracy Score** : 0.85
2. **F1-score**: 46%
3. **Précision**: 76%

Modèle de sélection:
Cross Validation : 10



ANNEXE : Modèle SVM avec df2_bis

Variables sélectionnées:
GR, RESD

Base utilisée: df2_bis

SVM :
hyperparamètres:
gamma = 0.01
C = 100

		Matrix confusion	
		Predict classes	
		0	1
Actual classes	0	7 297	290
	1	1 095	1 639

	precision	recall	f1-score	Accuracy
0	0.87	0.96	0.91	0.87
1	0.85	0.60	0.70	

ANNEXE : Modèle SVM avec df3_bis

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

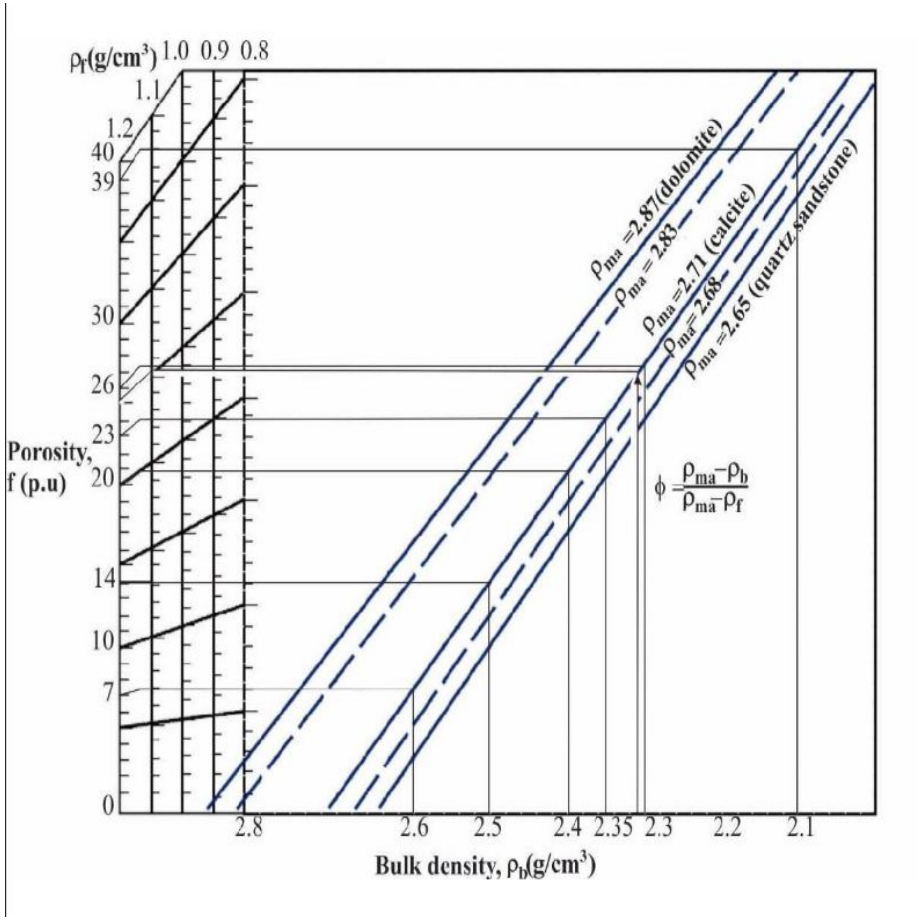
Base utilisée: df3_bis

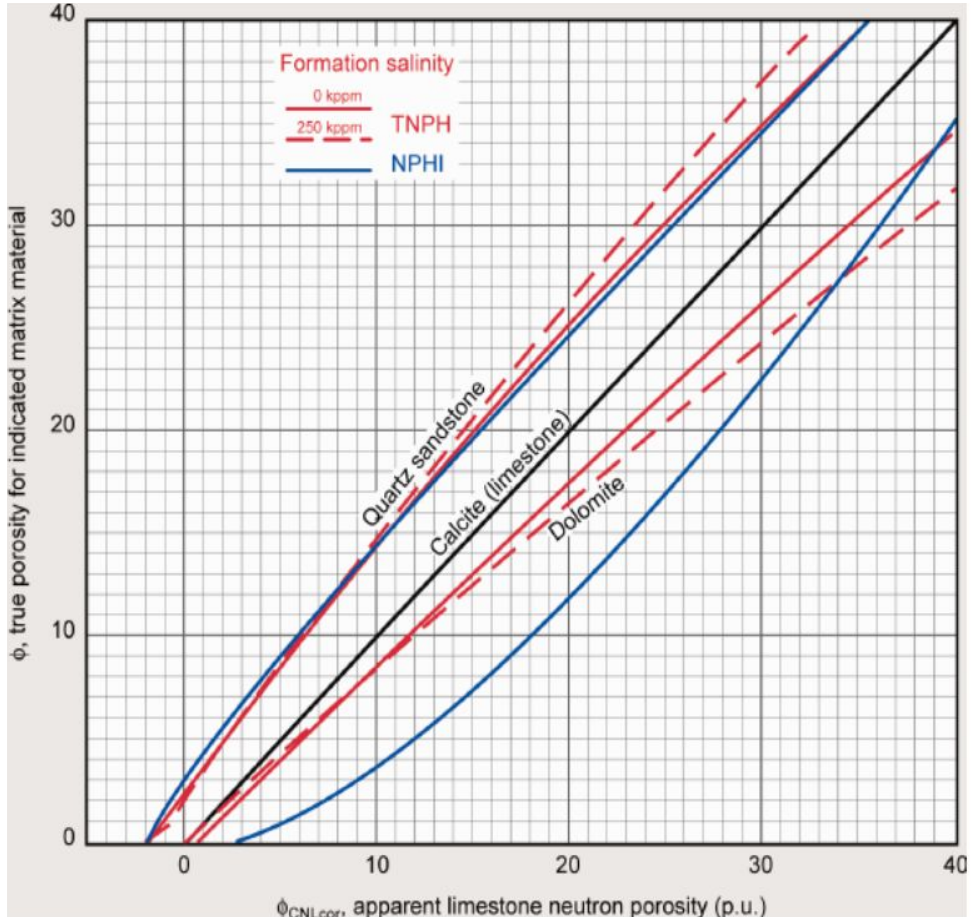
SVM :
hyperparamètres:
gamma = 0.01
C = 100

		Matrix confusion	
		Predict classes	
		0	1
Actual classes	0	7 325	262
	1	531	2 203

	precision	recall	f1-score	Accuracy
0	0.93	0.97	0.95	0.92
1	0.89	0.81	0.85	

ANNEXE : les abaques experts





ANNEXE : Répartition du Travail