MASTER 2 — GÉOSCIENCES

Parcours Géomarines • Géomatériaux • Pédologie

EXAMEN FINAL

Intelligence artificielle appliquée aux géosciences

Épreuve écrite — Session ……

Machine learning, données géoscientifiques et IA générative

Année universitaire	……	Date	……
Durée totale	1 h 30	Documents	Non autorisés
Ordinateur	Non autorisé	Calculatrice	Simple, autorisée
Enseignant	**Dr Kouadio**	Nombre de pages	17

Nom et prénoms :	
Matricule :	
Parcours :	☐ Géomarines ☐ Géomatériaux ☐ Pédologie

Cadre réservé au correcteur

Partie I — QCM	Partie II — Problème	Note finale
…… / 10	…… / 10	**…… / 20**

Organisation de l'épreuve

Partie	Contenu	Qui ?	Barème	Durée indicative
I — QCM	20 questions à choix multiple sur l'ensemble du cours	Tous les étudiants	10 points	30 min
II — Problème	Étude de cas d'ingénierie : 5 questions × 4 sous-questions	Sujet de **son** parcours uniquement	10 points	60 min
Total			**20 points**	**1 h 30**

Consignes générales

1. La **Partie I (QCM)** est commune et obligatoire pour tous les étudiants.
2. Pour la **Partie II**, chaque étudiant traite **uniquement** le problème de son parcours. Un autre sujet traité ne sera pas corrigé.
3. Aucun ordinateur n'est nécessaire : il n'y a ni programme à écrire ni code à exécuter. Les extraits Python sont à **lire et interpréter**.
4. Justifiez vos réponses, précisez toujours les **unités** (cmol(+)/kg, MPa, %…) et distinguez résultats de modèles, observations et décisions.
5. Rédigez les réponses de la Partie II sur la copie d'examen ; seule la grille du QCM se remplit sur le sujet.

PARTIE I — QCM commun (10 points)

Pour chaque question, une seule réponse est correcte. Reportez la lettre choisie (A, B, C ou D) dans la grille de réponses en fin de partie. Chaque bonne réponse vaut 0,5 point ; une réponse fausse, multiple ou absente vaut 0 (pas de point négatif).

1. Selon T. Mitchell, une machine « apprend » lorsque :

- **A)** elle exécute plus rapidement un programme écrit par l'humain
- **B)** elle dispose de davantage de mémoire et de puissance de calcul
- **C)** sa performance *P* sur une tâche *T* s'améliore avec l'expérience *E*
- **D)** elle peut répondre à toutes les questions qu'on lui pose

2. Quelle relation d'inclusion est correcte ?

- **A)** IA ⊃ Machine Learning ⊃ Deep Learning
- **B)** Deep Learning ⊃ Machine Learning ⊃ IA
- **C)** Le Machine Learning et l'IA sont deux domaines indépendants
- **D)** Le Deep Learning ne fait pas partie du Machine Learning

3. Laquelle de ces tâches relève de l'IA **générative** ?

- **A)** Prédire la CEC d'un sol à partir de son argile et de son carbone organique
- **B)** Classer des échantillons de roche en granite, basalte ou gneiss
- **C)** Estimer la teneur en COT d'un sédiment marin
- **D)** Rédiger un résumé de rapport géotechnique à partir d'une consigne écrite

4. Prédire la résistance en compression simple *Rc* (en MPa) d'une roche est un problème de :

- **A)** classification binaire
- **B)** régression
- **C)** clustering
- **D)** détection d'anomalies

5. On dispose de 500 analyses géochimiques **sans aucune étiquette** et l'on cherche des familles d'échantillons semblables. Il s'agit :

- **A)** d'apprentissage supervisé par régression
- **B)** d'apprentissage supervisé par classification
- **C)** d'apprentissage non supervisé (clustering, par exemple K-Means)
- **D)** d'apprentissage par renforcement

6. Une forêt aléatoire (*Random Forest*) :

- **A)** combine de nombreux arbres entraînés sur des tirages aléatoires et agrège leurs prédictions (vote ou moyenne)
- **B)** est un unique arbre de décision très profond
- **C)** ne peut être utilisée qu'en régression
- **D)** n'a pas besoin de données d'entraînement

7. Dans un réseau de neurones, la fonction d'activation sert principalement à :

- **A)** stocker les données d'entraînement
- **B)** séparer les données en entraînement et test
- **C)** encoder les variables catégorielles
- **D)** introduire une non-linéarité, sans laquelle le réseau resterait une fonction linéaire

8. Dans le code ci-dessous, y représente :

```
df = pd.read_csv("pedologie_regression_CEC_1200.csv")
X  = df[["argile_pct", "pH_eau", "carbone_organique_pct"]]
y  = df["CEC_cmol_charge_kg"]
```

A) la table complète des données
B) la variable cible que le modèle doit prédire
C) les variables explicatives
D) le jeu de test

9. L'instruction `pd.get_dummies(X, columns=["lithologie_observee"], dtype=int)` permet de :
A) supprimer les lignes contenant des valeurs manquantes
B) ramener toutes les variables numériques entre 0 et 1
C) transformer une variable catégorielle en colonnes indicatrices 0/1
D) séparer les données en entraînement et test

10. L'instruction `df.isnull().sum()` renvoie :
A) le nombre de valeurs manquantes dans chaque colonne
B) la somme des valeurs de chaque colonne
C) le nombre de lignes du fichier
D) le nombre de classes de la variable cible

11. Le jeu de **test** sert à :
A) ajuster les paramètres du modèle
B) choisir les hyperparamètres et comparer les modèles
C) augmenter la quantité de données d'entraînement
D) évaluer une seule fois la performance finale sur des données jamais utilisées pour apprendre ni pour choisir

12. Un modèle obtient $R^2 = 0{,}99$ sur l'entraînement mais $R^2 = 0{,}45$ sur le test. Le diagnostic le plus probable est :
A) un sous-apprentissage
B) un surapprentissage : le modèle a mémorisé le bruit de l'entraînement
C) un modèle parfait
D) une erreur de calcul du R^2

13. Pour prédire le COT, un collègue ajoute la colonne « COT mesuré » dans X. Cette erreur s'appelle :
A) un sous-apprentissage
B) un encodage One-Hot
C) une fuite de cible (*data leakage*) : les scores deviennent artificiellement excellents
D) une normalisation

14. Un modèle de CEC obtient MAE = 2,4 cmol(+)/kg sur le test. Cela signifie :
A) qu'en moyenne, l'écart absolu entre CEC prédite et CEC mesurée est d'environ 2,4 cmol(+)/kg
B) que le modèle se trompe de 2,4 % en moyenne
C) que chaque prédiction a une erreur d'exactement 2,4 cmol(+)/kg
D) que le modèle explique 2,4 % de la variance

15. Par rapport au MAE, le RMSE :

- **A)** s'exprime toujours en pourcentage
- **B)** est toujours inférieur au MAE
- **C)** ne peut être calculé qu'en classification
- **D)** pénalise davantage les grandes erreurs, car les écarts sont élevés au carré

16. Sur un nuage de points où la valeur **mesurée est en abscisse** et la valeur **prédite en ordonnée**, un point situé au-dessus de la diagonale $y = x$ correspond à :

- **A)** une prédiction parfaite
- **B)** une surestimation par le modèle
- **C)** une sous-estimation par le modèle
- **D)** une valeur manquante

17. Dans une matrice de confusion (lignes = classes réelles, colonnes = classes prédites), une case **hors diagonale** indique :

- **A)** des échantillons d'une classe réelle attribués à tort à une autre classe
- **B)** des échantillons correctement classés
- **C)** le nombre total d'échantillons du jeu de test
- **D)** l'accuracy du modèle

18. Une « hallucination » d'un assistant IA (LLM) désigne :

- **A)** un arrêt du logiciel par manque de mémoire
- **B)** une réponse trop courte
- **C)** un contenu plausible mais faux ou inventé, par exemple une référence bibliographique ou un DOI inexistant
- **D)** une réponse donnée dans une autre langue

19. Un prompt bien structuré contient, dans l'ordre recommandé en cours :

- **A)** uniquement la question, la plus courte possible
- **B)** le code complet et les données confidentielles du projet
- **C)** une liste de mots-clés sans phrase
- **D)** le rôle, le contexte, la tâche, les contraintes et le format attendu

20. Quel est le bon usage d'un assistant IA dans un rapport ou un mémoire ?

- **A)** Recopier le texte généré sans le relire, car il est toujours exact
- **B)** Vérifier chaque chiffre et chaque référence auprès des sources primaires, sans y copier de données confidentielles
- **C)** Citer les références proposées sans les lire
- **D)** Remplacer les essais de laboratoire par la réponse du LLM

Grille de réponses — Partie I

Inscrivez une seule lettre par case.

Question	1	2	3	4	5	6	7	8	9	10
Réponse										
Question	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Réponse										

PARTIE II — Problème d'ingénierie (10 points)

La Partie II propose **trois sujets indépendants**, un par parcours. Chaque sujet comporte cinq documents suivis chacun d'une question de 2 points (quatre sous-questions de 0,5 point). **Traitez uniquement le sujet de votre parcours.**

Parcours	Sujet à traiter	Thème de l'étude de cas
Pédologie	**Sujet 1**	Diagnostic pédologique et aménagement agricole
Géomatériaux	**Sujet 2**	Choisir des matériaux rocheux pour un projet routier à partir de résultats de machine learning
Géomarines	**Sujet 3**	Interpréter les sédiments marins et prioriser les contrôles environnementaux d'une zone côtière

Ce que l'on attend de vous

- **Données** : distinguer variables explicatives *X* et variable cible *Y*, classification et régression.
- **Code** : comprendre le rôle de chaque instruction Python, sans l'exécuter.
- **Classification** : comparer accuracy et F1-score macro, lire une matrice de confusion.
- **Régression** : interpréter MAE, RMSE et R^2 dans les unités de la cible, lire un nuage de points.
- **Décision** : formuler un avis d'ingénieur argumenté, prudent et vérifiable sur le terrain.

Q1 — Données	Q2 — Code Python	Q3 — Classification	Q4 — Régression	Q5 — Décision	Total
2 pts	2 pts	2 pts	2 pts	2 pts	**10 pts**

Sujet 1 — PÉDOLOGIE

Diagnostic pédologique et aménagement agricole

Réservé au parcours	Barème	Durée indicative	Ordinateur
Pédologie	10 points	60 minutes	Non requis

Mise en situation professionnelle

Vous êtes ingénieur pédologue dans un bureau d'études chargé d'orienter un projet d'aménagement agricole en Côte d'Ivoire. Les parcelles candidates présentent des propriétés variées : texture, acidité, drainage et matière organique.

Pour mieux cibler des investigations de terrain et des analyses de laboratoire coûteuses, votre équipe a construit deux outils d'aide au diagnostic : **(i)** une classification pédagogique des groupes de sols et **(ii)** une estimation de la capacité d'échange cationique (CEC). Les modèles ont déjà été entraînés. Votre mission consiste à examiner les données, interpréter les résultats et formuler une recommandation responsable.

Remarque : La classification pédagogique inspirée de la WRB ne remplace pas un diagnostic officiel de profil de sol.

Document 1 — Les données disponibles

Chaque jeu de données comporte 1 200 observations : entraînement (840), validation (180) et test (180). Principales colonnes :

Variable	Exemple / unité	Rôle possible
`texture`	Argileuse	Descripteur
`argile_pct`	48,4 %	Descripteur
`pH_eau`	4,61	Descripteur
`drainage`	Bien drainé	Descripteur
`carbone_organique_pct`	1,08 %	Descripteur
`type_sol_WRB_pedagogique`	Ferralsol	**Cible de classification**
`CEC_cmol_charge_kg`	9,25 cmol(+)/kg	**Cible de régression**

Question 1 — Données et apprentissage supervisé (2 points)

1.1 *(0,5 pt)* Pour prédire le type de sol, identifiez *Y* et trois variables *X* pertinentes. Pour prédire la CEC, quelle est la nouvelle variable *Y* ?

1.2 *(0,5 pt)* Classez `texture`, `drainage`, `argile_pct` et `pH_eau` en variables qualitatives ou quantitatives ; précisez la différence entre une classe et une valeur continue.

1.3 *(0,5 pt)* Expliquez en une phrase la distinction classification / régression appliquée à ce cas.

1.4 *(0,5 pt)* Expliquez les rôles respectifs des données d'entraînement, de validation et de test, et le risque d'utiliser les mêmes données pour apprendre et pour évaluer.

Document 2 — Quelques lignes de Python

Le data scientist de l'équipe vous transmet l'extrait suivant. *Vous n'avez pas à exécuter le code.*

```
import pandas as pd
df = pd.read_csv("pedologie_classification_1200.csv")
X  = df[["texture", "argile_pct", "pH_eau", "drainage"]]
y  = df["type_sol_WRB_pedagogique"]
X_encoded = pd.get_dummies(X, columns=["texture", "drainage"], dtype=int)
print(df.isnull().sum())
print(df["type_sol_WRB_pedagogique"].value_counts())
```

Question 2 — Prétraitement et lecture du code (2 points)

2.1 *(0,5 pt)* Quelle est la fonction de `read_csv()` ? Que représentent `df`, `X` et `y` ?

2.2 *(0,5 pt)* Pourquoi encoder `texture` et `drainage` ? Citez deux noms possibles de colonnes obtenues après l'encodage *One-Hot*.

2.3 *(0,5 pt)* Que vérifient `isnull().sum()` et `value_counts()` ? Proposez une action si des valeurs manquantes existent.

2.4 *(0,5 pt)* Pourquoi ne faut-il pas placer `type_sol_WRB_pedagogique` dans `X` pour prédire cette même colonne ? Nommez cette erreur.

Document 3 — Classification des groupes de sols

Résultats obtenus sur les 180 observations de test, avec 15 variables explicatives (encodage préparé à partir de l'entraînement).

Modèle	Accuracy (test)	F1-score macro (test)
Decision Tree	87,2 %	0,877
Random Forest	97,8 %	0,978

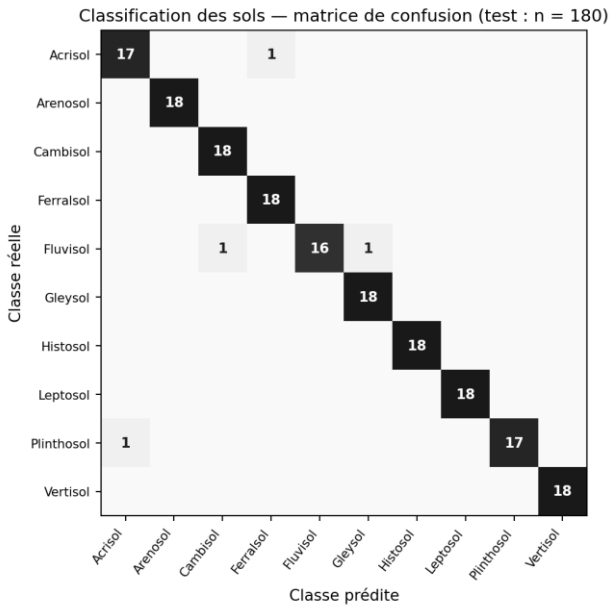


Figure 1 — Matrice de confusion du Random Forest (test, n = 180). Lignes : classes réelles ; colonnes : classes prédites. Les cases vides correspondent à zéro observation.

Question 3 — Lecture des résultats de classification (2 points)

3.1 *(0,5 pt)* Quel modèle choisiriez-vous sur la base des scores fournis ? Argumentez avec l'accuracy et le F1-score macro.

3.2 *(0,5 pt)* À partir de la matrice, indiquez le nombre total d'erreurs du Random Forest sur le jeu de test et le nombre de Fluvisols correctement reconnus.

3.3 *(0,5 pt)* Expliquez ce que signifie une confusion entre deux groupes de sols pour les activités de reconnaissance pédologique.

3.4 *(0,5 pt)* Pourquoi une performance élevée sur le jeu de test ne suffit-elle pas à supprimer les visites de terrain et les analyses de laboratoire ?

Document 4 — Estimation de la CEC

Les mêmes 180 observations de test servent à comparer les modèles de régression. Unité de la CEC : cmol(+)/kg.

Modèle	MAE	RMSE	R^2
Ridge	4,16	5,59	0,888
Random Forest Regressor	2,37	3,62	0,953

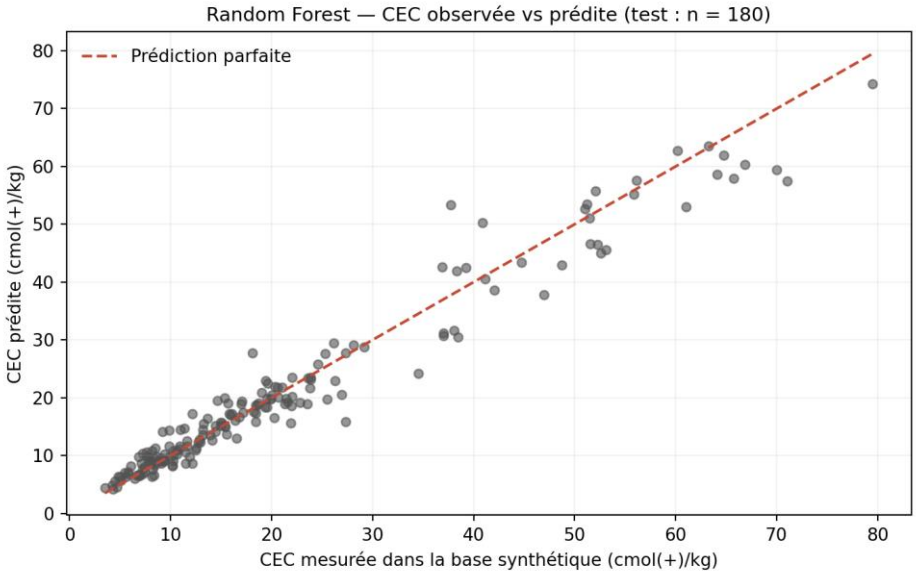


Figure 2 — CEC mesurée (abscisse) et CEC prédite par le Random Forest (ordonnée). Chaque point est un sol du jeu de test ; la diagonale pointillée correspond à une prédiction sans erreur.

Question 4 — Interprétation de la régression (2 points)

4.1 *(0,5 pt)* Quel modèle produit les meilleures estimations de CEC ? Utilisez au moins deux métriques pour justifier.

4.2 *(0,5 pt)* Interprétez concrètement un MAE de 2,37 cmol(+)/kg et expliquez la signification des points éloignés de la diagonale dans la figure 2.

4.3 *(0,5 pt)* Pourquoi le carbone organique et l'argile peuvent-ils être utiles pour estimer la CEC ? Une corrélation prouve-t-elle une causalité ?

4.4 *(0,5 pt)* Complétez `y = df["_____"]` pour une régression prédisant la CEC ; indiquez pourquoi cette cible est quantitative continue.

Document 5 — Votre recommandation de pédologue

Le responsable du projet souhaite examiner en priorité une parcelle destinée à des cultures exigeant un drainage satisfaisant et une disponibilité correcte en éléments nutritifs. *Les valeurs ci-dessous sont des observations et des estimations de CEC fournies pour le scénario, indépendantes des prédictions de la figure 2.*

Propriété	Parcelle A	Parcelle B	Parcelle C
Argile (%)	45	18	55
Sable (%)	35	72	20
pH	4,6	6,2	5,8
Carbone organique (%)	1,0	0,7	3,2
Drainage	Bon	Bon	Mauvais
CEC estimée (cmol(+)/kg)	11	6	28

Question 5 — Décision professionnelle (2 points)

5.1 *(0,5 pt)* Identifiez une contrainte agronomique plausible pour chacune des trois parcelles en vous appuyant sur les données.

5.2 *(0,5 pt)* Quelle parcelle privilégieriez-vous pour une investigation approfondie avant aménagement ? Justifiez le compromis entre pH, CEC et drainage ; *plusieurs choix argumentés sont acceptables.*

5.3 *(0,5 pt)* Pourquoi la CEC élevée de la parcelle C ne prouve-t-elle pas, à elle seule, que cette parcelle est la meilleure ?

5.4 *(0,5 pt)* Proposez deux analyses ou vérifications complémentaires nécessaires avant la décision finale, en précisant leur utilité.

Sujet 2 — GÉOMATÉRIAUX

Choisir des matériaux rocheux pour un projet routier à partir de résultats de machine learning

Réservé au parcours	Barème	Durée indicative	Ordinateur
Géomatériaux	10 points	60 minutes	Non requis

Mise en situation professionnelle

Vous êtes ingénieur géomatériaux dans un bureau d'études chargé d'appuyer le choix de matériaux provenant de plusieurs carrières pour un projet de construction routière en Côte d'Ivoire. Les essais mécaniques complets sur tous les échantillons sont coûteux.

Votre équipe a constitué deux bases de données de 1 200 échantillons : l'une sert à identifier la **lithologie** (classification), l'autre à estimer la **résistance à la compression simple** *Rc*, exprimée en MPa (régression). Les modèles sont déjà entraînés. Votre mission : comprendre les données et quelques instructions Python, évaluer les prédictions et produire un avis d'ingénieur.

Remarque : Aucune prédiction ne vaut qualification réglementaire d'un granulat.

Document 1 — Base de données et variables

Variable du fichier	Exemple	Signification
`densite_seche_g_cm3`	2,70	Masse volumique sèche (g/cm³)
`porosite_pct`	6,5	Porosité (%)
`vitesse_vp_m_s`	4 800	Vitesse des ondes P (m/s)
`alteration_visuelle_pct`	12	Indice visuel d'altération (%)
`lithologie_observee`	Granite	Nature pétrographique observée
`y_lithologie`	Basalte	**Cible de classification**
`y_resistance_compression_simple_MPa`	126,5	**Cible de régression** *Rc* (MPa)
`split`	test	Partition du jeu de données

Répartition de chaque fichier : entraînement = 840 ; validation = 180 ; test = 180. Les deux fichiers ne contiennent aucune valeur manquante.

Question 1 — Comprendre la mission et les données (2 points)

1.1 *(0,5 pt)* Pour identifier la lithologie, indiquez la cible *Y* et citez trois variables *X* quantitatives pertinentes.

1.2 *(0,5 pt)* Précisez si `lithologie_observee` et `porosite_pct` sont catégorielles ou numériques. Quelle différence existe-t-il ici entre classification et régression ?

1.3 *(0,5 pt)* Expliquez les rôles des ensembles d'entraînement, de validation et de test.

1.4 *(0,5 pt)* Citez une raison pour laquelle la résistance *Rc* seule ne suffit pas à déclarer un granulat apte à la construction routière.

Document 2 — Lecture d'instructions Python

Vous n'avez pas à exécuter le code.

```
import pandas as pd
df = pd.read_csv("M2_geomateriaux_regression_Rc_1200.csv")
X  = df[["porosite_pct", "vitesse_vp_m_s",
         "densite_seche_g_cm3", "lithologie_observee"]]
y  = df["y_resistance_compression_simple_MPa"]

X_encoded = pd.get_dummies(X, columns=["lithologie_observee"], dtype=int)
print(df.isnull().sum())
```

Question 2 — Interpréter un code, sans programmer (2 points)

2.1 *(0,5 pt)* À quoi servent `pd.read_csv()` et `df` ? Identifiez `X` et `y` dans l'extrait.

2.2 *(0,5 pt)* Pourquoi encoder `lithologie_observee` ? Donnez un exemple de nouvelle colonne obtenue.

2.3 *(0,5 pt)* Qu'indique `df.isnull().sum()` ? Quelle situation pratique ferait envisager un traitement des données ?

2.4 *(0,5 pt)* Un collègue ajoute `y_resistance_compression_simple_MPa` à `X` pour prédire cette même résistance. Identifiez l'erreur et sa conséquence.

Document 3 — Identification des lithologies

Modèle	Accuracy (test)	F1-score macro (test)
Decision Tree	72,8 %	0,722
Random Forest	76,1 %	0,758

La matrice ci-dessous croise les classes réelles (lignes) et les classes prédites (colonnes) par le Random Forest sur 180 échantillons de test. Les teintes foncées indiquent des effectifs plus élevés.

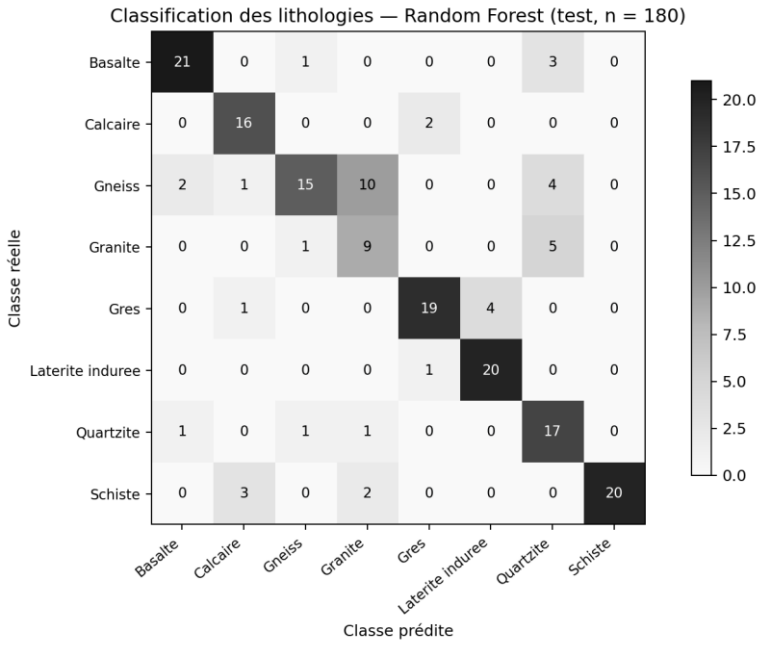


Figure 1 — Matrice de confusion du Random Forest pour la classification des lithologies (test, n = 180).

Question 3 — Lecture des résultats de classification (2 points)

3.1 *(0,5 pt)* Quel modèle choisiriez-vous au regard de l'accuracy et du F1-score macro ? Justifiez.

3.2 *(0,5 pt)* À l'aide de la matrice, expliquez la signification des cases de la diagonale et d'une case située hors diagonale.

3.3 *(0,5 pt)* Identifiez une confusion entre deux lithologies et expliquez pourquoi elle peut constituer un risque lors d'une première sélection en carrière.

3.4 *(0,5 pt)* Même avec une accuracy élevée, pourquoi une vérification pétrographique reste-t-elle nécessaire ?

Document 4 — Estimation de la résistance en compression

Modèle	MAE (MPa)	RMSE (MPa)	R^2
Ridge	17,64	22,42	0,862
Random Forest Regressor	13,75	18,73	0,904



Figure 2 — Rc mesurée (abscisse) et Rc prédite par le Random Forest (ordonnée) sur le jeu de test. La droite pointillée représente une prédiction parfaite (y = x).

Question 4 — Interprétation de la régression (2 points)

4.1 *(0,5 pt)* Quel modèle est le plus performant selon les trois métriques ?

4.2 *(0,5 pt)* Interprétez concrètement, en MPa, le MAE du meilleur modèle.

4.3 *(0,5 pt)* Que représente un point situé au-dessus de la droite $y = x$? Pourquoi un point très éloigné de cette droite appelle-t-il à la prudence ?

4.4 *(0,5 pt)* Pourquoi la porosité, la vitesse *Vp* et le degré d'altération peuvent-ils être utiles pour estimer *Rc* ? Donnez un mécanisme physique pour deux de ces variables.

Document 5 — Votre avis d'ingénieur

Trois sites candidats à l'approvisionnement présentent les caractéristiques indicatives ci-dessous. *Les valeurs de Rc sont des prédictions du modèle ; ces trois cas sont indépendants du jeu de test.*

Propriété	Carrière A	Carrière B	Carrière C
Porosité (%)	4	17	8
Vp (m/s)	5 400	2 700	4 100
Altération visuelle (%)	5	45	20
Rc prédite (MPa)	165	82	128

Le chantier nécessite des granulats résistants, durables et adaptés à la couche routière envisagée. Les résultats d'abrasion, de fragmentation, de durabilité et de granulométrie ne sont pas encore disponibles.

Question 5 — Décision d'ingénieur (2 points)

5.1 *(0,5 pt)* Quelle carrière étudieriez-vous en premier ? Justifiez avec au moins deux propriétés.

5.2 *(0,5 pt)* La carrière A peut-elle être autorisée immédiatement sur la seule base de la *Rc* prédite ? Pourquoi ?

5.3 *(0,5 pt)* Proposez deux essais ou contrôles supplémentaires adaptés aux granulats routiers et expliquez leur rôle.

5.4 *(0,5 pt)* Rédigez en deux ou trois phrases un avis au chef de projet : choix provisoire, vérifications à mener et incertitude du modèle.

Sujet 3 — GÉOMARINES

Interpréter les sédiments marins et prioriser les contrôles environnementaux d'une zone côtière

Réservé au parcours	Barème	Durée indicative	Ordinateur
Géomarines	10 points	60 minutes	Non requis

Mise en situation professionnelle

Vous êtes ingénieur géomarin dans un bureau d'études missionné pour une étude préalable à un aménagement offshore sur le plateau continental du golfe de Guinée. Votre équipe cherche à caractériser les faciès sédimentaires et à repérer les stations susceptibles de présenter une accumulation importante de **carbone organique total (COT)**.

Des prélèvements systématiques seraient coûteux : les modèles de machine learning servent ici à **orienter** les investigations, et non à certifier la qualité de l'environnement. Deux bases de 1 200 stations ont été constituées : classification des faciès et prédiction du COT (en %). Votre travail consiste à comprendre les variables, lire de courts extraits Python, interpréter les figures et formuler un avis scientifique argumenté.

Document 1 — Les données disponibles

Variable	Exemple	Interprétation
`profondeur_eau_m`	75	Profondeur d'eau (m)
`sable_pct`	35	Fraction sableuse (%)
`silt_pct`	40	Fraction silteuse (%)
`argile_pct`	25	Fraction argileuse (%)
`energie_hydrodynamique_indice`	0,21	Indice relatif de l'énergie
`oxygene_dissous_mg_L`	3,2	Oxygène dissous (mg/L)
`potentiel_redox_mV`	155	Potentiel redox (mV)
`y_facies_sedimentaire`	Sable_vaseux	**Cible de classification**
`y_carbone_organique_total_pct`	1,45	**Cible de régression** (%)
`split`	test	Partition de la base

Chaque fichier compte 840 stations pour l'entraînement, 180 pour la validation et 180 pour le test. La somme graviers + sable + silt + argile doit être proche de 100 % (aux arrondis près).

Question 1 — Comprendre les données et les cibles (2 points)

1.1 *(0,5 pt)* Pour prédire un faciès, identifiez *Y* et trois variables explicatives *X* pertinentes.

1.2 *(0,5 pt)* Distinguez une variable catégorielle et deux variables numériques du tableau ; expliquez ce que sont la classification et la régression dans ce projet.

1.3 *(0,5 pt)* Que représentent les ensembles d'entraînement, de validation et de test ? Pourquoi ne faut-il pas conclure uniquement sur l'entraînement ?

1.4 *(0,5 pt)* Une station présente 10 % de graviers, 35 % de sable, 40 % de silt et 20 % d'argile. Quel contrôle de cohérence effectuez-vous et quelle action proposez-vous ?

Document 2 — Comprendre le code Python

Vous n'avez pas à exécuter le code.

```
import pandas as pd
df = pd.read_csv("M2_geomarine_regression_COT_1200.csv")
X  = df[["sable_pct", "argile_pct",
         "oxygene_dissous_mg_L", "facies_observe"]]
y  = df["y_carbone_organique_total_pct"]

X_encoded = pd.get_dummies(X, columns=["facies_observe"], dtype=int)
print(df.isnull().sum())
```

Question 2 — Lecture du code et prétraitement (2 points)

2.1 *(0,5 pt)* Expliquez l'utilité de `read_csv()`, `df`, `X` et `y`.

2.2 *(0,5 pt)* À quoi sert `get_dummies()` ? Donnez le nom plausible d'une colonne issue de l'encodage de `facies_observe`.

2.3 *(0,5 pt)* Quelle vérification réalise `isnull().sum()` ? Comment réagiriez-vous si plusieurs valeurs d'oxygène dissous étaient absentes ?

2.4 *(0,5 pt)* Un collègue inclut le COT mesuré dans `X` alors qu'il cherche à prédire ce même COT. Expliquez l'erreur et son impact potentiel.

Document 3 — Classification des faciès sédimentaires

Modèle	Accuracy (test)	F1-score macro (test)
Decision Tree	95,0 %	0,949
Random Forest	96,1 %	0,959

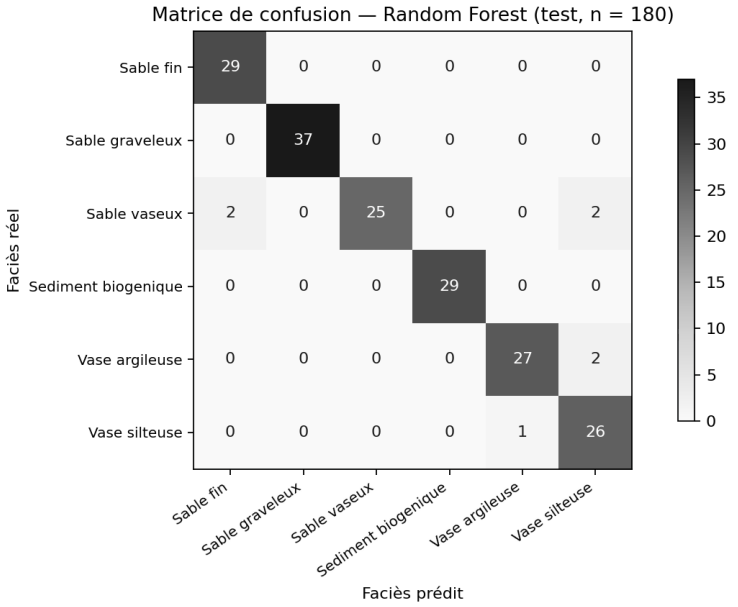


Figure 1 — Matrice de confusion du Random Forest (180 stations de test). Lignes : faciès réels ; colonnes : faciès prédits.

Question 3 — Interprétation de la classification (2 points)

3.1 *(0,5 pt)* Comparez les deux modèles à partir de l'accuracy et du F1-score macro.

3.2 *(0,5 pt)* Interprétez une cellule diagonale et une cellule hors diagonale de la matrice.

3.3 *(0,5 pt)* Repérez une confusion entre faciès ; décrivez une conséquence possible pour le plan d'échantillonnage.

3.4 *(0,5 pt)* Pourquoi une accuracy élevée sur le jeu de test ne dispense-t-elle pas d'une analyse granulométrique en laboratoire ?

Document 4 — Prédiction du carbone organique total (COT)

Modèle	MAE (% COT)	RMSE (% COT)	R^2
Ridge	0,291	0,376	0,880
Random Forest Regressor	0,278	0,378	0,878

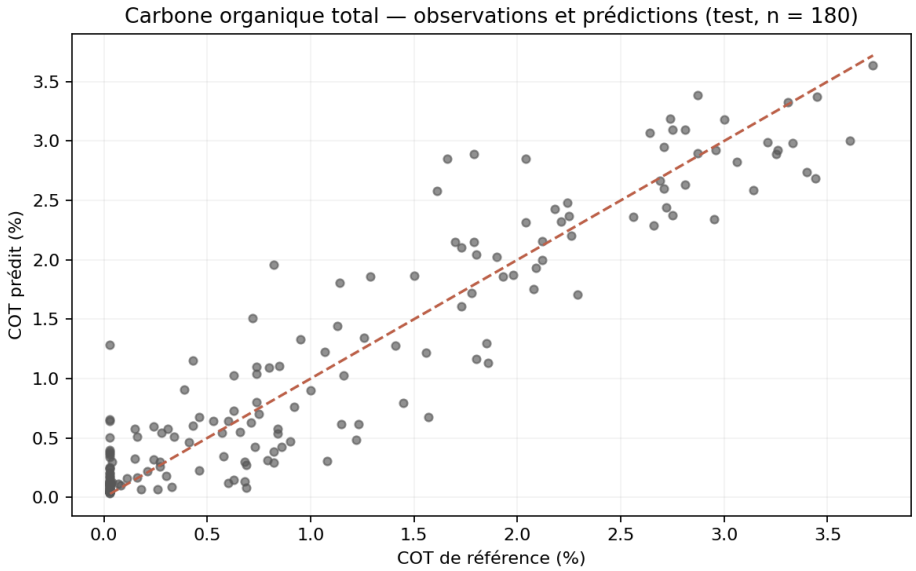


Figure 2 — COT de référence (abscisse) et COT prédit par le Random Forest (ordonnée) sur 180 stations de test. La diagonale en pointillés correspond à une prédiction parfaite.

Question 4 — Analyse des performances de régression (2 points)

4.1 *(0,5 pt)* Les trois métriques désignent-elles le même meilleur modèle ? Comparez MAE, RMSE et R^2, puis expliquez le compromis.

4.2 *(0,5 pt)* Donnez le sens concret du MAE du meilleur modèle, dans l'unité du COT.

4.3 *(0,5 pt)* Que signifie un point situé au-dessus de la diagonale ? Expliquez pourquoi des points éloignés de celle-ci sont préoccupants.

4.4 *(0,5 pt)* Expliquez deux mécanismes susceptibles de relier la fraction fine, l'énergie hydrodynamique ou l'oxygénation à l'accumulation ou à la préservation du carbone organique.

Document 5 — Décision de l'ingénieur géomarin

Trois stations candidates ont été reconnues. *Ces valeurs sont indépendantes des 180 stations du jeu de test.*

Propriété	Station A	Station B	Station C
Profondeur d'eau (m)	25	75	140
Graviers (%)	0	0	0
Sable (%)	85	35	10
Silt (%)	10	40	45
Argile (%)	5	25	45
Oxygène dissous (mg/L)	6,5	3,2	1,8
COT prédit (%)	0,6	1,8	3,4

L'aménagement envisagé nécessite de repérer les zones à caractériser en priorité. *Une teneur élevée en COT ne démontre pas, à elle seule, une contamination chimique ou une pollution.*

Question 5 — Votre avis d'ingénieur (2 points)

5.1 *(0,5 pt)* Associez à chaque station un faciès plausible à partir de sa granulométrie et justifiez brièvement.

5.2 *(0,5 pt)* Quelle station recommanderiez-vous d'investiguer en premier pour vérifier une accumulation de COT ? Justifiez par deux observations.

5.3 *(0,5 pt)* Pourquoi une concentration élevée en COT ne prouve-t-elle pas une contamination ? Indiquez deux mesures complémentaires utiles à l'évaluation environnementale.

5.4 *(0,5 pt)* En deux ou trois phrases, adressez votre recommandation au chef de mission : priorité, incertitudes et vérifications de terrain.

— FIN DE L'ÉPREUVE —