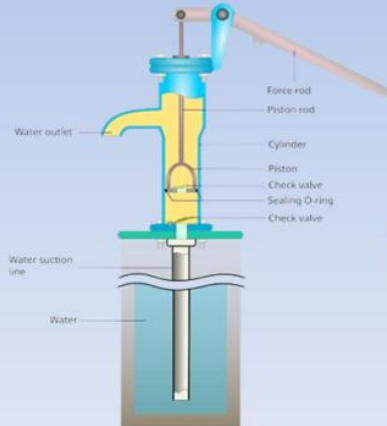


ANALYSE PRÉDICTIVE DE L'ÉTAT DES POINTS D'EAU EN TANZANIE

(PROJET DE MACHINE LEARNING)



OVERVIEW & GOAL

O

OVERVIEW

- Jeu de données : sociodémographique s, transactions, comportements
- Méthodes testées : Régression Logistique & Arbre de Décision

C

CHALLENGE

- Grande dimensionnalité (nombreuses variables)
- Données déséquilibrées
- Importance d'interopérabilité des résultats

A

APPROACH


- Prétraitement : encodage, normalisation, sélection de variables
- Construction de modèles supervisés
- Évaluation via métriques (accuracy, recall, f1-score, matrice de confusion)

G

GOAL

- Identifier le modèle le plus performant et interprétable
- Produire des recommandations opérationnelles basées sur les résultats

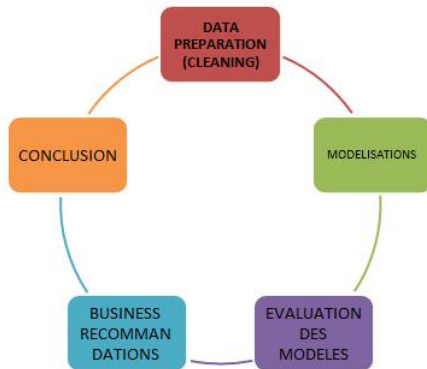
DATA UNDERSTANDING



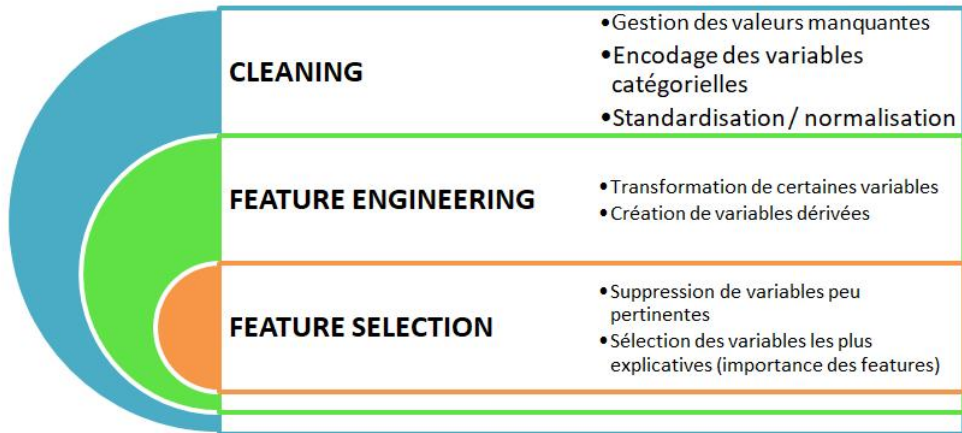
STRUCTURE DE DONNEES	<ul style="list-style-type: none">• Nombre d'observations : ~59k• Variables : numériques & catégorielles
DISTRIBUTION & TENDANCES	<ul style="list-style-type: none">• Variables déséquilibrées• Variabilité importante sur certaines features
INSIGHTS PRELIMINAIRES	<ul style="list-style-type: none">• Quelques variables semblent discriminantes• Indications utiles pour la sélection des modèles

EXPLORATORY DATA ANALYSIS

STEPS



DATA PREPARATION



MODELISATION

LOGISTIQUE REGRESSION

Accuracy (train, balanced): 0.814983164983165

Classification Report (balanced):

	precision	recall	f1-score	support
functional	0.90	0.80	0.85	32259
functional needs repair	0.43	0.94	0.59	4317
non functional	0.87	0.81	0.84	22824
accuracy			0.81	59400
macro avg	0.73	0.85	0.76	59400
weighted avg	0.85	0.81	0.83	59400

Matrice de confusion (balanced):

```
[[25916  3752  2591]
 [  148 4064   105]
 [ 2812 1582 18430]]
```

EVALUTATION DU MODELE

LIMITES

- Quelques confusions entre functional et needs repair.
- Performance légèrement déséquilibrée selon les classes.

FORCES

- Bonne accuracy globale (0.81).
- Classes majoritaires bien identifiées.
- Classe critique (needs repair) bien détectée (Recall élevé).
- Modèle robuste et utile pour orienter les interventions.

RANDOM FOREST

Accuracy (train, RF): 0.9996

Classification Report (Random Forest):

	precision	recall	f1-score	support
functional	1.00	1.00	1.00	32259
functional needs repair	1.00	1.00	1.00	4317
non functional	1.00	1.00	1.00	22824
accuracy			1.00	59400
macro avg	1.00	1.00	1.00	59400
weighted avg	1.00	1.00	1.00	59400

Matrice de confusion (RF):

```
[[32238   17    4]
 [    0 4317    0]
 [    1    2 22821]]
```

EVALUTATION DU MODELE

LIMITES

- Classe minoritaire encore mal prédite : malgré le `class_weight=balanced`, la classe functional needs repair reste difficile à identifier. Le modèle favorise surtout functional et non functional.
- F1-score global relativement faible (0.60) → équilibre limité entre classes. Biais vers les classes majoritaires : la performance globale (accuracy élevée) masque les erreurs sur la classe rare → indicateur trompeur si on regarde uniquement l'accuracy.

FORCES

- Bonne capacité à distinguer functional et non functional. Identification claire des variables les plus importantes (précision et rappel équilibrés).
- Gestion du déséquilibre améliorée grâce à `class_weight=balanced`.
- Bonne performance sur la classe non functional (recall = 0.58, précision = 0.90).

DECISION TREE

Accuracy (train): 0.7619023569023569

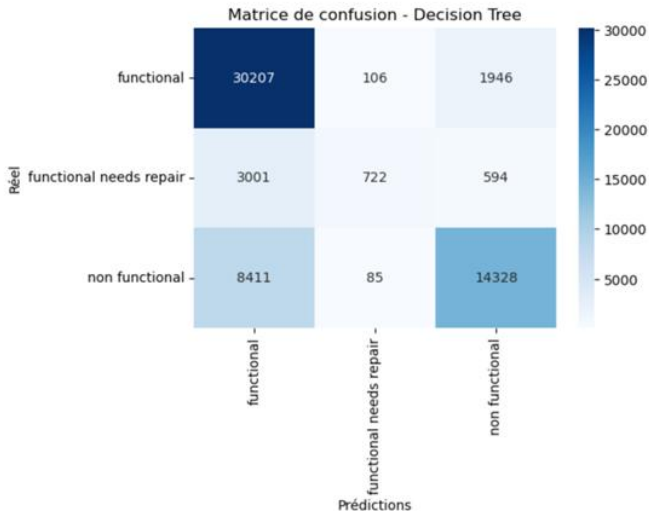
Classification report:

	precision	recall	f1-score	support
functional	0.73	0.94	0.82	32259
functional needs repair	0.79	0.17	0.28	4317
non functional	0.85	0.63	0.72	22824
accuracy			0.76	59400
macro avg	0.79	0.58	0.61	59400
weighted avg	0.78	0.76	0.74	59400

Matrice de confusion :

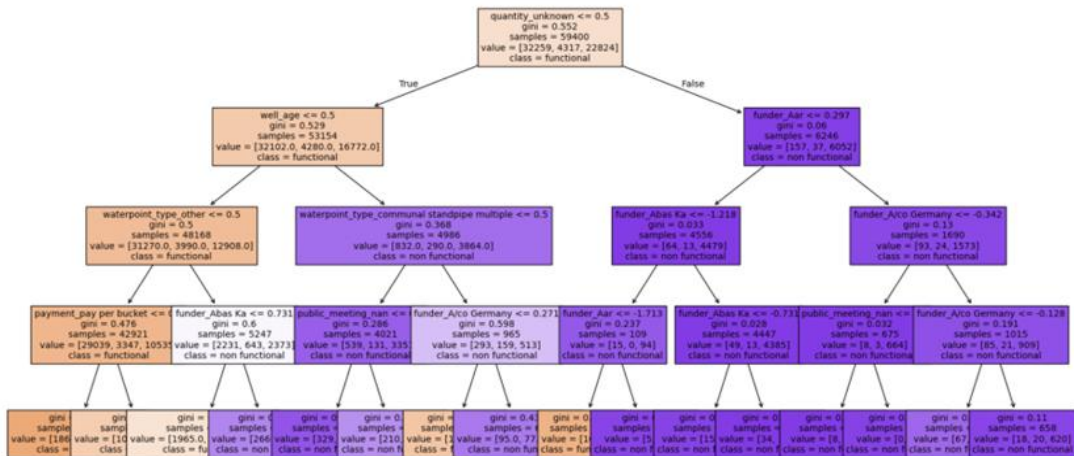
```
[[30207  106  1946]
 [ 3001   722    594]
 [ 8411    85 14328]]
```

MATRICE DE CONFUSION



GRAPHIC DECISION TREE

Arbre de décision simplifié - top 20 features



EVALUATION DU MODELE

LIMITES

- Classe minoritaire encore mal prédite : malgré le `class_weight=balanced`, la classe functional needs repair reste difficile à identifier le modèle favorise surtout functional et non functional.
- F1-score global relativement faible → équilibre limité entre classes. Biais vers les classes majoritaires : la performance globale (accuracy élevée) masque les erreurs sur la classe rare → indicateur trompeur si on regarde uniquement l'accuracy.

FORCES

- Bonne capacité à distinguer functional et non functional Identification claire des variables les plus importantes (précision et rappel équilibrés).
- Gestion du déséquilibre améliorée grâce à `class_weight=balanced`.
- Bonne performance sur la classe non functional.

COMPARAISON DES TROIS MODELES



LOGISTIC REGRESSION

- Bon compromis global avec une accuracy élevée (~ 0.85).
- Distingue bien fonctionnel et non fonctionnel.
- Faiblesses : la classe intermédiaire fonctionnel needs repair reste mal captée, car la frontière linéaire ne suffit pas pour modéliser cette catégorie plus ambiguë.

DECISION TREE

- Plus interprétable, on peut visualiser facilement les règles utilisées (par ex. `gps_height`, `age` puits, `funder`).
- Faiblesses : performance plus faible (~ 0.76 accuracy) et tendance au sur-apprentissage si la profondeur n'est pas bien contrôlée.

RANDOM FOREST

- Combine plusieurs arbres \rightarrow meilleure robustesse et meilleure capacité à modéliser la complexité.
- Gère bien le déséquilibre entre classes grâce à `class_weight="balanced"`.
- Offre souvent le meilleur équilibre précision/rappel, notamment pour la classe intermédiaire, même si l'interprétabilité diminue par rapport à un seul arbre.

BUSINESS RECOMMENDATIONS

BUSINESS RECOMMENDATION 1

Prioriser la maintenance préventive des puits "functional needs repair"

Nos modèles montrent que cette classe est la plus difficile à prédire correctement.

Cela traduit une zone grise opérationnelle : ces puits fonctionnent encore, mais avec un risque élevé de panne.

Recommandation : mettre en place un programme de suivi régulier (ex. inspections trimestrielles) pour ces puits afin de réduire leur transition vers la catégorie non fonctionnel.

BUSINESS RECOMMENDATION 2

Utiliser les variables géographiques et techniques pour cibler les interventions

Les variables comme `gps_height`, `construction_year`, et la localisation (`region`, `basin`) influencent fortement le statut des puits.

Certains contextes géographiques présentent plus de non-functional wells.

Recommandation : orienter les investissements en maintenance et réhabilitation vers les zones à risque identifiées par les modèles (ex. altitude basse, puits anciens).

BUSINESS RECOMMANDATION 3

Mettre en place un tableau de bord de suivi basé sur la prédiction

Le modèle Random Forest peut servir comme un outil d'aide à la décision.

Recommandation : créer un dashboard opérationnel qui affiche :

Les puits classés par risque de panne,

Une alerte précoce pour les puits à surveiller,

Les priorités d'allocation des ressources (techniciens, financements, pièces de rechange).

CONCLUSION

L'analyse a montré que la régression logistique offre les meilleures performances globales pour distinguer les puits fonctionnels et non fonctionnels, tout en restant robuste et généralisable. L'arbre de décision apporte une bonne interprétabilité mais souffre de surapprentissage, tandis que la random forest améliore légèrement la précision grâce à son approche ensembliste, au prix d'une complexité plus élevée.

Dans le cadre de ce projet, nous retenons la régression logistique avec pondération des classes comme méthode principale, car elle combine une bonne performance globale, une relative simplicité de mise en œuvre et une interprétation claire des résultats.