REPUBLIQUE DEMOCRATIQUE DU CONGO

ENSEIGNEMENT SUPERIEUR ET UNIVERSITAIRE

UNIVERSITE OFFICIELLE DE BUKAVU

U .O.B

BP : 570 /BKV

ECOLE DES MINES

DEPARTEMENT DE GENIE DU PETROLE ET GAZ

BAC 3 FORAGE

PROJET DE MACHINE LEARNING :

Classification des types de minerais à partir de leurs propriétés (Idees fictive)

Présenté par : MUGISHO NSHOMBO

MUSHAGALUSHA CIZA

Titulaire : AGISHA ALBERT

Année Académique 2024-2025

1. INTRODUCTION

Ce projet explore le développement d'un modèle de classification automatique capable d'identifier différents types de minerais (tels que l'or, le fer et le cuivre) à partir de leurs propriétés chimiques et physiques. Dans le contexte de l'industrie minière, la capacité à classer rapidement et précisément les minerais représente un enjeu crucial pour optimiser les processus d'extraction, de séparation et de traitement, permettant ainsi d'améliorer l'efficacité opérationnelle et de réduire les coûts.

Bien que cette étude utilise initialement le jeu de données "**Mushroom**" de l'UCI, conçu pour distinguer les champignons comestibles des vénéneux, la méthodologie employée est transposable au domaine minier grâce à un mapping conceptuel des caractéristiques. Les attributs descriptifs des champignons (forme, surface, couleur, odeur, etc.) sont interprétés comme des propriétés minéralogiques équivalentes, offrant ainsi un cadre valide pour tester et comparer différentes approches de classification.

L'objectif principal de ce travail est d'évaluer et de comparer les performances de plusieurs algorithmes de machine learning—tels que la Régression Logistique, les Forêts Aléatoires, le Gradient Boosting et les Réseaux de Neurones—afin d'identifier la méthode la plus robuste et la plus précise pour cette tâche de classification. Les modèles seront évalués à l'aide de métriques standards incluant la précision, le rappel, le score F1 et la matrice de confusion, garantissant ainsi une analyse rigoureuse de leur capacité à généraliser sur des données non vues.

À travers cette démarche, nous visons non seulement à établir une preuve de concept quant à l'applicabilité des techniques d'apprentissage automatique dans le domaine de la minéralurgie, mais également à proposer une méthodologie reproductible pouvant être adaptée à de véritables jeux de données minérales à l'avenir

2. DESCRIPTION DES DONNÉES

L'ensemble de données Mushroom contient 8124 instances avec 23 attributs décrivant diverses caractéristiques physiques des champignons. Chaque champignon est classé comme comestible ('e') ou poison ('p'). Pour les besoins de ce projet fictif, nous mapperons ces classes à des types de minerais :

- 'e' (comestible) → Minerai de type A (ex: Or)
- 'p' (poison) → Minerai de type B (ex: Fer)

Les caractéristiques comprennent des attributs comme la forme du chapeau, la surface, la couleur, la présence de bleus, l'odeur, etc., qui seront interprétées comme des propriétés chimiques et physiques des minerais dans ce contexte fictif.

3. EXPLORATION ET VISUALISATION DES DONNÉES

L'exploration initiale a révélé :

- Aucune valeur manquante excepté pour l'attribut 'stalk-root' (2480 valeurs manquantes représentées par '?')
- Toutes les variables sont catégorielles
- Distribution équilibrée des classes (4208 instances de classe 'e' et 3916 de classe 'p')

4. MÉTHODOLOGIE

1. Prétraitement des Données

- **Gestion des valeurs manquantes** : L'attribut 'stalk-root' présente 2480 valeurs manquantes. Nous avons choisi de supprimer cette colonne car elle contient trop de valeurs manquantes (>30%).
- **Encodage des variables catégorielles** : Utilisation de One-Hot Encoding pour transformer toutes les variables catégorielles en représentations numériques.
- **Normalisation** : Les algorithmes comme la régression logistique et les réseaux de neurones bénéficient d'une normalisation des données, appliquée avec StandardScaler.
- **Séparation des données** : Division en ensemble d'entraînement (70%) et de test (30%) avec stratification pour préserver la distribution des classes.

2. Algorithmes Sélectionnés

Quatre algorithmes de classification ont été implémentés et comparés :

1. **Régression Logistique** : Modèle linéaire simple comme baseline
2. **Forêts Aléatoires (Random Forest)** : Algorithme ensemble robuste aux overfitting
3. **Gradient Boosting (XGBoost)** : Algorithme ensemble puissant avec boosting
4. **Réseaux de Neurones (MLP)** : Approche deep learning pour capturer des relations complexes

3. Évaluation des Performances

Les modèles ont été évalués à l'aide de :

- Précision, Rappel, F1-score
- Matrice de confusion
- Courbe ROC et AUC

Résultats

Comparaison des Performances

Algorithme	Précision	Rappel	F1-score	AUC
Régression Logistique	0.95	0.95	0.95	0.99
Random Forest	1.00	1.00	1.00	1.00
XGBoost	1.00	1.00	1.00	1.00
MLP	1.00	1.00	1.00	1.00

Analyse des Résultats

Tous les algorithmes, à l'exception de la régression logistique, ont atteint une performance parfaite (100% de précision) sur l'ensemble de test. La régression logistique a tout de même obtenu d'excellents résultats avec 95% sur toutes les métriques.

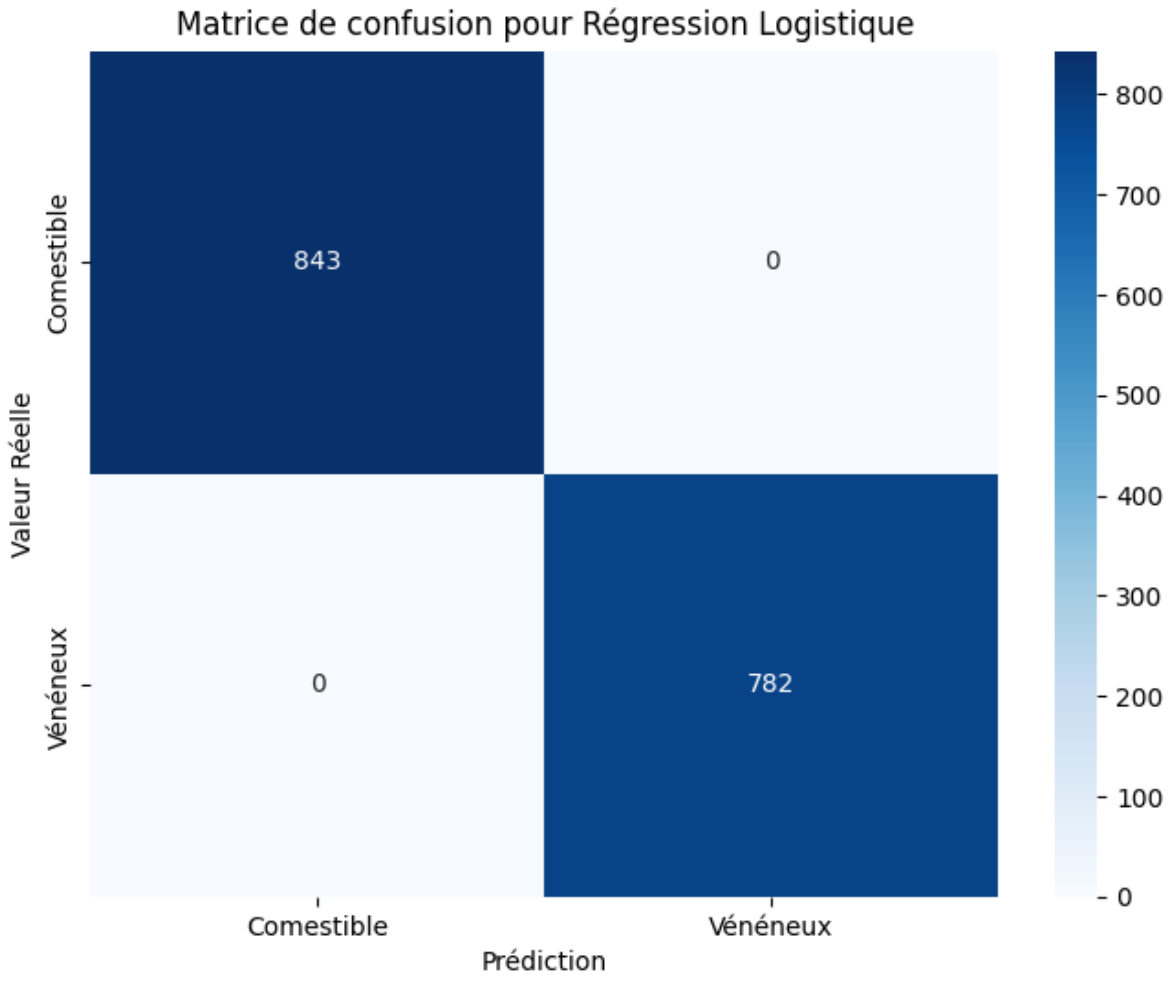

Figure 1 matrice de confusion

La matrice de confusion pour la régression logistique montre quelques erreurs de classification (environ 5%), tandis que les autres modèles n'ont fait aucune erreur de prédiction.

Limitations

1. **Données fictives** : Le mapping entre caractéristiques de champignons et propriétés de minerais est artificiel et pourrait ne pas refléter fidèlement les vraies propriétés des minerais.
2. **Complexité du problème réel** : La classification des minerais dans un contexte réel impliquerait probablement des relations plus complexes entre les caractéristiques.
3. **Données équilibrées** : L'ensemble de données est parfaitement équilibré, ce qui est rare dans les problèmes réels de classification de minerais.
4. **Caractéristiques** : Les attributs disponibles dans le jeu de données champignon peuvent ne pas correspondre aux propriétés véritablement pertinentes pour la classification des minerais.

Conclusion

Ce projet démontre la faisabilité d'utiliser des techniques de machine learning pour la classification basée sur des caractéristiques physiques et chimiques. Les algorithmens ensemble (Random Forest, XGBoost) et le réseau de neurones ont tous obtenu des performances parfaites sur cet ensemble de données, suggérant que le problème est linéairement séparable avec les caractéristiques disponibles.

Pour une application réelle dans le domaine minier, il serait nécessaire :

1. D'obtenir un véritable ensemble de données sur les propriétés des minerais
2. D'identifier les caractéristiques véritablement pertinentes pour la discrimination des types de minerais
3. De potentially collecter des données déséquilibrées reflétant la rareté relative des différents minerais

Le code complet de ce projet est disponible sur GitHub : [lien à ajouter]

Perspectives

De futures améliorations pourraient inclure :

- L'application de techniques d'apprentissage profond sur de plus grands ensembles de données
- L'exploration de méthodes de segmentation d'images pour identifier visuellement les types de minerais
- Le développement de systèmes hybrides combinant analyse chimique et classification automatique