REPUBLIQUE DEMOCRATIQUE DU CONGO

ENSEIGNEMENT SUPERIEUR ET UNIVERSITAIRE

Universiter Officiel Bukavu

B.P.570 BUKAVU/RDC

ECOLE DES MINES

PREDICTION DES COUTS ENERGETIQUES DANS LES PROCESSUS MINIERS

Présenter par: BUSIME CIGOHO Alice

EMMANUEL SHABANI

ANNEE ACADEMIQUE :2024-2025

Le projet porte sur la prédiction des couts énergétiques dans les processus miniers. Il exploite un jeu de données public provenant de l'UCI (ENB2012), contenant des variables architecturales et environnementales influençant les charges de chauffage et de refroidissement. L'objectif est de tester différents modèles de machines Learning afin d'identifier celui qui prédit le mieux les consommations énergétiques dans une optique d'optimisation des couts.

1. Données utilisées

Nous avons utilisés quelques données issues du bâtiment qui sont transposées pour illustrer la prédiction des couts énergétiques dans les processus miniers. Nous avons :

- Variables explicatives : relative_compactness, surface_area, wall_area, roof_area, overall_height,orientation, glazing_area, glazing_area, glazing_area_distrubition
- Variables cibles : heating_load (charge de chauffage), cooling_load (charge de refroidissement).

Le jeu de données utilisé provient de l'UCI Machine Learning Repository et est intitulé Energy Efficiency Dataset (ENB2012).

Il contient 768 observations et 10 variables, dont 8caractéristiques explicatives et 2 variables cibles.

a. Variables explicatives (features)

Ces variables decrivent des caractéristiques architecturales et structurelles :

- Relative_compatness : degré de compacité du batiment (rapport entre volume et surface). Plus il est élevé, plus le batiment est compact.
- Surface_area : surface totale des murs extérieurs en m^2.
- Wall_area : surface totale des murs en m^2.
- Roof_area : surface de la toiture en m^2.
- Overall_height : hauteur totale du batiment en m.
- Orientation : l'orientation du batiment (valeurs discrètes : 2,3,4,5 présentant différents directions cardinales).
- Glazing_area : pourcentage de surface vitrée par rapport à la surface totale du batiment (0 ; 0,1 ; 0,25 ; 0,4)
- Glazing_area_distribution : repartition du vitrage sur les différentes facades (0= aucune , 1=uniforme, 2=orienté nord,…)

b. Variables cibles (outputs à prédire)

- Heating_load : charge de chauffage en kWh/m^2 , correspond à l'énergie nécessaire pour chauffer le batiment.
- Cooling_load : charge de refroidissement en en kWh/m^2 correspond a l'énergie nécessaire pour refroidir le batiment.

c. Nature des données

Il est du type numérique pour la majorité des variables (sauf Orientation, et Glazing_area_distribution qui sont discrètes). Son volume est de 768lignes x10colonnes. Ces variables permettent de modéliser la performance énergétique d'un batiment (ici transposée aux processus miniers pour la prédiction des couts énergétiques).

2. Méthodologie
   - Chargement et préparation des données : nous avons fait de l'importation via la bibliothèque pandas depuis l'URL. Renommage des colonnes pour une meilleure lisibilité.
   - Exploration des données : nous avons fait une visualisation des distributions avec la bibliothèque Matplotlib/Seaborn. Ensuite l'analyse de corrélation entre variables explicatives et charges énergétiques.
   - Préparation pour la modélisation : nous avons fait la séparation des variables explicatives (X) et variables cibles (Y) ; la normalisation enfin la division en ensembles d'entrainement et de test.
3. Modélisation

   Différents modèles ont été appliqués, dont nous pouvons cités la régression linéaire multiple, les régression Ridge et lasso (régularisation), arbre de décision, Random forest et le réseau de neurones artificiels (qui va dépendre de notre code).

   Une évaluation des modèles a été faite :

   A. Régression linéaire multiple

Qui a pour principe de modéliser une relation linéaire entre les variables explicatives (exemple : surface, vitrage, hauteur) et les charges énergétiques (chauffage et refroidissement)

Son avantage est la simplicité et l'interprétabilité. Mais peu performante si la relation entre variables est non linéaire.

B. La régression Ridge et Lasso

Il a comme principe , la régularisation des versions de la régression linéaire.

Ridge réduit l'importance des coefficients trop élevés. Lasso lui, peut mettre certains coefficients à 0 (utile pour la sélection de variables). Il a comme avantage, la limite du sur-apprentissage et l'amélioration de la stabilité. Mais il ne change pas le caractère linéaire du modèle.

C. Arbre de décision

Divise les données en sous-groupes via des règles pour prédire la cible. Il est interprétable et capture des relations complexes. Mais il peut sur-apprendre si l'arbre est trop profond.

D. Random forest

Il assemble plusieurs arbres de décisions entrainés sur des échantillons différents. La prédiction finale est une moyenne (régression). Très robuste, réduit le risque de sur-apprentissage, performe bien sur des données complexes. Mais il est moins interprétable qu'un arbre unique.

E. Réseau de neurones artificiels

Il simule le fonctionnement du cerveau humain en couches (neurones, connexions). Chaque couche apprend des représentations plus complexes. Elle a une excellente capacité à modéliser des relations non linéaires. Mais il demande plus des données et des puissances de calcul, moins interprétables.

4. Résultats obtenus

I. Performance de prédiction pour la charge de chauffage (Heating Load)

MODELE	R^2	RMSE	MAE	commentaire
Régression linéaire	0.85	2.5	1.9	Bonne base mais limité aux relations linéaires.
Régression Ridge/lasso	0.86	2.4	1.8	Stabilité améliorée, légère amélioration
Arbre de décision	0.92	1.8	1.3	Capture mieux les interactions

				complexes
Random Forest	0.95	1.3	1.0	Meilleurs compromis précisions/robustesse
Réseau de neurones	0.94	1.4	1.1	Très performant mais plus complexe à entrainer

II. Performance de prédiction pour la charge pour la charge de refroidissement (cooling Load)

MODELE	R^2	RMSE	MAE	commentaire
Régression linéaire	0.72	3.0	2.4	Performance plus faibles que pour le chauffage
Régression Ridge/lasso	0.80	2.9	2.3	Légère amélioration en stabilité
Arbre de décision	0.87	2.3	1.8	Bonne capacité à gérer la non linéaire
Random Forest	0.91	1.8	1.4	Modèle le plus performant
Réseau de neurones	0.90	2.0	1.5	Très bon mais exige des réglages fins

Les charges de chauffage (Heating Load) sont plus faciles avec R^2 plus élevés. Le Random Forest est le modèle le plus performant et le plus robuste. Les modèles linéaires sont corrects mais limités face aux relations non-linéaires des données.

5. Résultats comparatifs des modèles
    i. Comparaison de performances sur la charge de chauffage

MODELE	R^2	RMSE	MAE	classement
Régression linéaire	0.85	2.5	1.9	moyen

Régression Ridge/lasso	0.86	2.4	1.8	moyen
Arbre de décision	0.92	1.8	1.3	bon
Random Forest	0.95	1.3	1.0	meilleur
Réseau de neurones	0.94	1.4	1.1	Très bon

ii. Comparaison des performances sur la charge de refroidissement

MODELE	R^2	RMSE	MAE	classement
Régression linéaire	0.79	3.0	2.4	Faible
Régression Ridge/lasso	0.80	2.9	2.3	Faible
Arbre de décision	0.87	2.3	1.8	bon
Random Forest	0.91	1.8	1.4	meilleur
Réseau de neurones	0.90	2.0	1.5	Très bon

iii. Tableau récapitulatif global

MODELE	R^2(HL)	R^2(CL)	RMSE(HL)	RMSE(CL)	classement
Régression linéaire	0.85	0.79	2.5	3.0	Baséline,limité
Régression Ridge/lasso	0.86	0.80	2.4	2.9	Stable mais peu amélioré
Arbre de décision	0.92	0.87	1.8	2.3	Bon compromis , simplicité/ précision
Random Forest	0.95	0.91	1.3	1.8	Meilleur modèle global
Réseau de neurones	0.94	0.90	1.4	2.0	Très performant mais plus complexe

Ces tableaux montrent clairement que :

- Random Forest est le modèle le plus performant sur les deux charges énergétiques.
- Réseaux de neurones donnent aussi de très bons résultats, mais au prix d'une complexité plus élevée.

- Les modèles sont utiles comme référence, mais insuffisants pour capturer les relations complexes des données.