République Islamique de Mauritanie
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Nouakchott

Analyse des Donnees d'Assurance Sante

Filière :
Statistique et Science de Donnees

Projet de Python

Sous le thème :

PREDICTION DU COUT DES SOINS MEDICAUX

Réalisé par :
Saade Bouh Aboubakar Hamar —C19871

Année Universitaire : 2024–2025

Table des matières

1 Introduction

Ce projet a pour but de prédire les coûts d'assurance médicale à l'aide d'un modèle de régression linéaire multiple.

2 Jeu de données

Le jeu de données utilisé contient 1338 lignes et 7 colonnes, sans valeurs manquantes :
— **age** : Âge de l'individu
— **sex** : Sexe (male/female)
— **bmi** : Indice de masse corporelle
— **children** : Nombre d'enfants à charge
— **smoker** : Statut de fumeur
— **region** : Région de résidence
— **charges** : Frais médicaux (variable cible)

3 Préparation des données

Les étapes suivantes ont été réalisées :
— Encodage des variables catégorielles avec OneHotEncoder
— Standardisation des variables numériques avec StandardScaler
— Division des données en ensemble d'entraînement (80%) et de test (20%)

4 Modélisation

Nous avons construit un pipeline contenant :
— un `ColumnTransformer` pour le prétraitement
— un modèle de régression linéaire (`LinearRegression`) de `scikit-learn`

5 Évaluation du modèle

Les métriques utilisées sont :
— R^2 : Coefficient de détermination
— MAE : Erreur absolue moyenne
— MSE : Erreur quadratique moyenne

6 Analyse des résultats

L'analyse met en évidence :
— Le statut de fumeur influence fortement les charges
— Un BMI élevé (>30) augmente les coûts
— L'âge a une relation positive avec les frais médicaux
— Le nombre d'enfants et la région ont un effet modéré

7 Interprétation des métriques

— $\mathbf{R^2 = 0.78}$: Le modèle explique 78% de la variance des frais médicaux.
— **MAE = 4193.29** : Erreur moyenne de 4193 unités monétaires.
— **MSE = 33 700 000** : Les grandes erreurs sont fortement pénalisées.
— **RMSE $\approx$ 5805.16** : Erreur quadratique moyenne en unités monétaires.

8 Conclusion

Le modèle de régression linéaire multiple fournit des résultats satisfaisants. Des pistes d'amélioration incluent l'utilisation de modèles plus complexes et l'ajout de données.

Fin