Mini-Projet
Apprentissage Supervisé Linéaire

(M1 Intelligence Artificielle)

Étudiant :
C34645 – Zahra Yeselk Boubacar

Encadré par :
Dr. EL BENANY Mohamed Mahmoud

Année universitaire :
2025–2026

1 Introduction

L'apprentissage supervisé est une branche essentielle du machine learning qui consiste à apprendre un modèle à partir de données étiquetées. L'objectif de ce mini-projet est de consolider les bases de cette approche à travers l'étude de deux modèles fondamentaux : la régression linéaire et la régression logistique.
La régression linéaire est utilisée pour prédire une variable continue, tandis que la régression logistique permet de résoudre des problèmes de classification. Ce rapport présente une synthèse des choix méthodologiques adoptés ainsi qu'une interprétation des résultats obtenus, illustrée par des visualisations issues des expérimentations.

2 Partie 1 : Régression Linéaire

2.1 Présentation des données

Pour la régression linéaire, nous avons utilisé le dataset *Medical Insurance Cost*, qui contient des informations sur des assurés telles que l'âge, l'indice de masse corporelle (BMI), le statut de fumeur, le nombre d'enfants et la région de résidence. La variable cible est `charges`, qui représente le coût médical annuel. Cette variable est numérique, ce qui satisfait la condition requise pour l'application d'un modèle de régression linéaire.

2.2 Analyse des corrélations

Une analyse exploratoire des données a été réalisée à l'aide d'une matrice de corrélation (heatmap) afin d'identifier les relations entre les variables numériques.
Cette analyse permet de repérer les variables les plus corrélées à la variable cible `charges`, ce qui constitue un choix méthodologique important avant la modélisation.

2.3 Modélisation

Le modèle de régression linéaire est formalisé par l'équation suivante :

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon$$

où y représente la variable cible, x_i les variables explicatives, β_i les coefficients du modèle et ε le terme d'erreur. Les variables catégorielles ont été transformées en variables numériques à l'aide de la méthode One-Hot Encoding afin de pouvoir être intégrées au modèle.

2.4 Évaluation des performances

Le modèle a été évalué sur un jeu de test à l'aide de l'erreur quadratique moyenne (MSE) et du coefficient de détermination R^2.
La proximité des points avec la diagonale indique que le modèle fournit des prédictions globalement proches des valeurs réelles, ce qui traduit de bonnes performances.

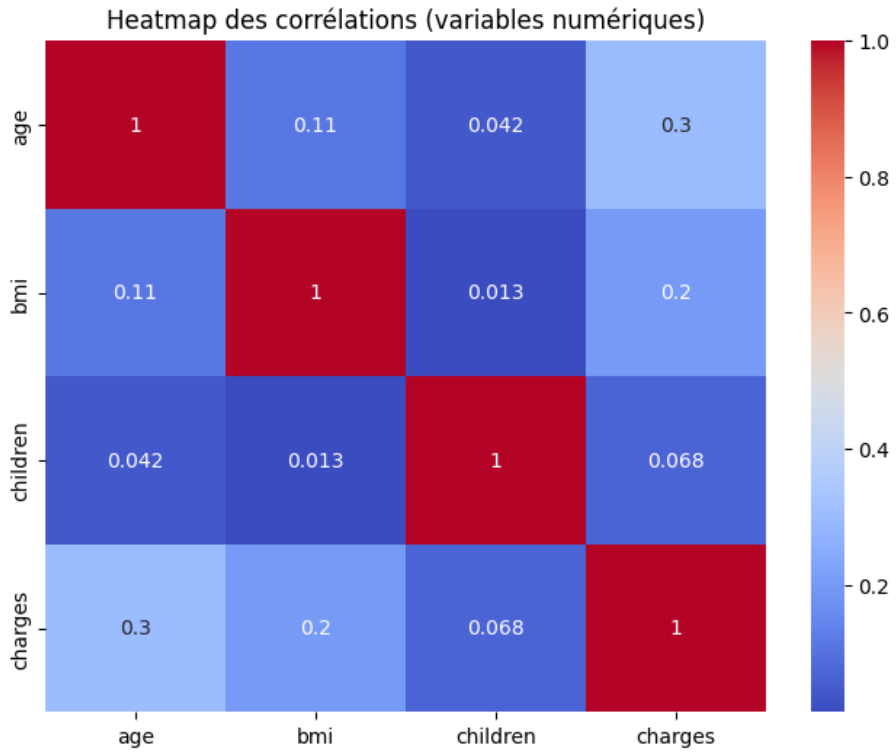

FIGURE 1 – Matrice de corrélation des variables numériques

2.5 Interprétation des coefficients

L'interprétation des coefficients β_i permet d'identifier l'importance des variables dans la prédiction.
L'analyse montre que le statut de fumeur est la variable la plus influente, suivi par l'âge et le BMI. Certaines régions présentent un effet négatif par rapport à la région de référence.

3 Partie 2 : Régression Logistique

3.1 Présentation des données

La régression logistique a été appliquée au dataset Iris, fourni par la bibliothèque `scikit-learn`. Le problème a été transformé en une classification binaire en distinguant la classe *setosa* des autres classes.

3.2 Prétraitement et modélisation

Les variables d'entrée ont été normalisées afin d'améliorer la convergence du modèle. La régression logistique modélise la probabilité d'appartenance à une classe à l'aide de la fonction sigmoïde :

$$P(y = 1|x) = \frac{1}{1 + e^{-z}}$$

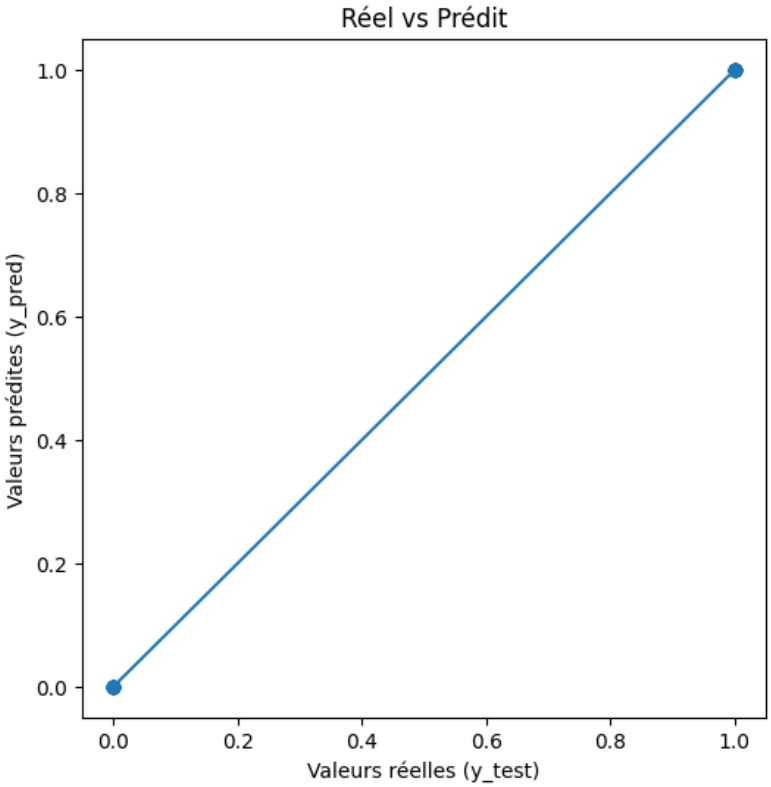

FIGURE 2 – Comparaison entre les valeurs réelles et les valeurs prédites

3.3 Évaluation des performances

Les performances du modèle ont été évaluées à l'aide d'une matrice de confusion ainsi que des métriques Accuracy, Précision et Recall.
Les résultats obtenus montrent des performances parfaites, avec des valeurs égales à 1.0 pour l'Accuracy, la Précision et le Recall. Ce résultat s'explique par le fait que la classe *setosa* est naturellement bien séparée des autres classes dans le dataset Iris.

4 Conclusion

Ce mini-projet a permis de mettre en pratique les concepts fondamentaux de l'apprentissage supervisé. La régression linéaire s'est montrée efficace pour la prédiction de variables continues, tandis que la régression logistique a donné d'excellents résultats pour la classification binaire. L'interprétation des coefficients et des métriques d'évaluation a permis de mieux comprendre l'impact des variables et la qualité des modèles construits.

Lien vers le Notebook

Le notebook Python contenant l'ensemble du code, des visualisations et des résultats est disponible à l'adresse suivante :
`https://colab.research.google.com/drive/1prM125opUxJXhulocr70JATG5AJR0nqW?usp=sharing`

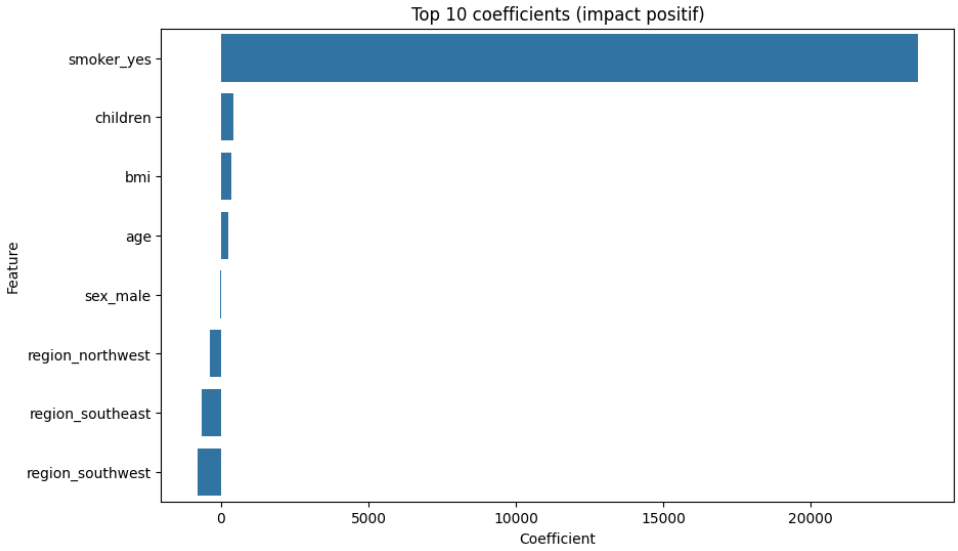

FIGURE 3 – Importance des variables selon les coefficients du modèle

```
Matrice de confusion :
 [[20  0]
 [ 0 10]]
```

FIGURE 4 – Matrice de confusion de la régression logistique