Rapport de Mini-Projet : Apprentissage Supervisé Linéaire

Amanetoullah Cheikh Mohamed Brahim (C22643)
Master 1 Intelligence Artificielle

Janvier 2026

1 Introduction

Ce rapport présente les travaux réalisés dans le cadre du module Machine Learning. L'objectif est de consolider les bases de l'apprentissage supervisé à travers l'implémentation et l'analyse de deux modèles fondamentaux : la Régression Linéaire pour la prédiction de variables continues et la Régression Logistique pour la classification binaire. L'intégralité du projet a été développée sous l'environnement **Google Colab**.

2 Méthodologie et Étapes de Réalisation

Le projet a été conduit selon un pipeline de science des données rigoureux, structuré en cinq étapes clés :

Configuration & Accès :
Environnement **Google Colab** avec montage du Drive et extraction automatisée des fichiers via *zipfile*.

Ingénierie des données :
Nettoyage, encodage des variables qualitatives (*Label Encoding*) et normalisation (*StandardScaler*) pour optimiser la convergence des modèles.

Entraînement :
Division des données en ensembles d'apprentissage (**80%**) et de test (**20%**) pour prévenir le sur-apprentissage.

Évaluation :
Validation des performances par les métriques R^2 et MSE pour la régression, ainsi que l'Accuracy et la matrice de confusion pour la classification.

3 Partie 1 : Régression Linéaire (Medical Insurance)

3.1 Choix et Préparation des Données

Nous avons utilisé le dataset *Medical Insurance Cost*. La variable cible est **charges**. Les variables catégorielles (sex, smoker, region) ont été encodées pour être intégrées dans l'équation mathématique du modèle.

3.2 Analyse des Corrélations

La matrice de chaleur (**Figure 1**) indique que la variable *smoker* est le facteur le plus corrélé aux coûts, suivie de l'âge et de l'IMC.

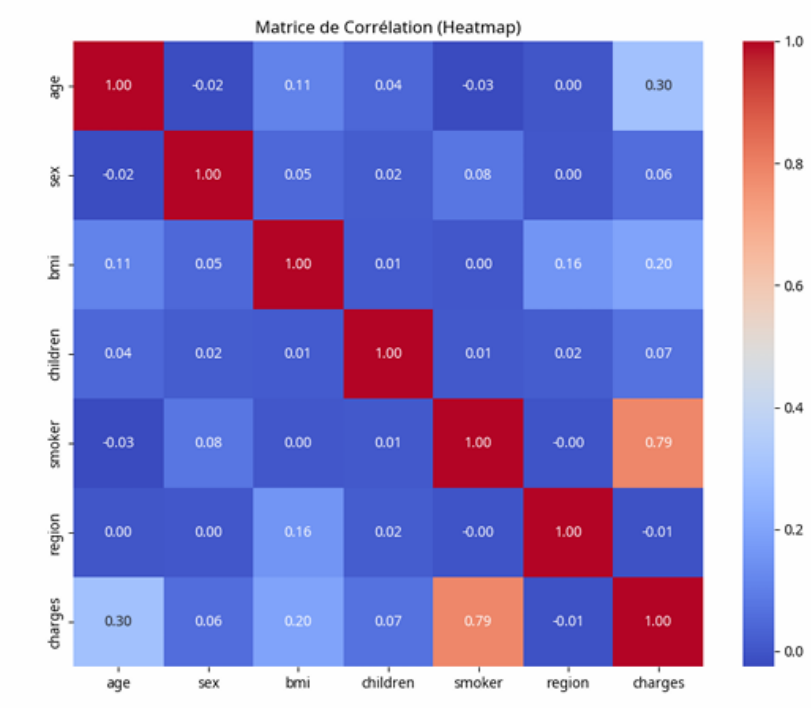

FIGURE 1 – Matrice de corrélation (Heatmap) montrant les facteurs influençant les charges.

3.3 Modélisation et Résultats

Le modèle est formalisé par : $y = \beta_0 + \sum \beta_i x_i + \epsilon$. Les performances obtenues après évaluation sont : $R^2 = 0.7833$ et $MSE = 33\,596\,915.8$.

3.4 Interprétation des Coefficients β_i

L'impact majeur est porté par la variable **smoker** ($\beta = 23593.98$), confirmant que le tabagisme augmente drastiquement les frais médicaux, suivi par l'IMC et l'âge.

4 Partie 2 : Régression Logistique (Iris Dataset)

4.1 Choix et Préparation des Données

Pour cette partie, nous avons utilisé le dataset **Iris** (natif dans Scikit-Learn). Le problème a été transformé en une classification binaire : distinguer la classe *Iris-Setosa* (Classe 0) des autres espèces. Les données ont été normalisées pour garantir une convergence optimale du modèle.

4.2 Modélisation et Performance

La probabilité est calculée via la fonction sigmoïde : $P(y = 1|x) = \frac{1}{1+e^{-z}}$. Le modèle obtient une **Accuracy de 100%**. Ce résultat s'explique par la séparabilité linéaire claire de la classe Setosa par rapport aux autres.

4.3 Matrice de Confusion

La matrice confirme l'absence totale d'erreurs (Vrais Négatifs : 20, Vrais Positifs : 10).

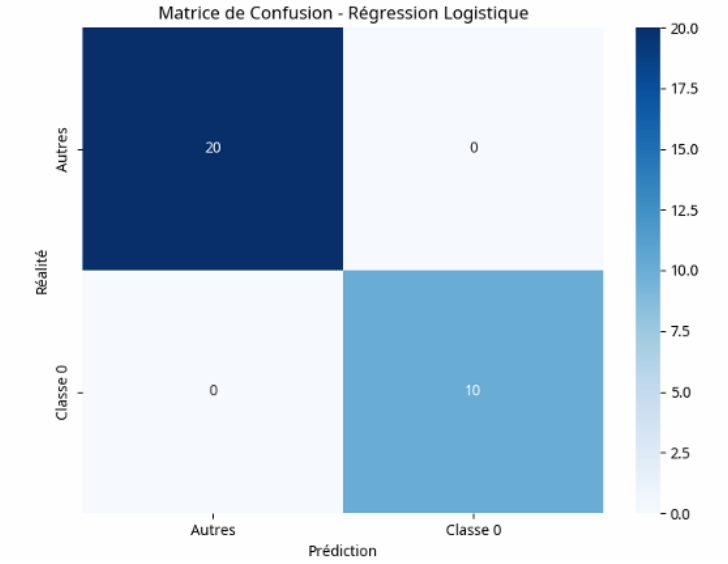

FIGURE 2 – Matrice de confusion - Classification binaire Iris.

5 Conclusion

Ce projet démontre l'efficacité des modèles linéaires. La Régression Linéaire a permis de quantifier l'impact du mode de vie sur les coûts d'assurance, tandis que la Régression Logistique a prouvé sa robustesse pour la classification de données linéairement séparables.