

PROJET DATA SCIENCE & HR ANALYTICS

# Rapport de Benchmark

Comparaison des Algorithmes de Prédiction de Salaire

**Modèles analysés :**

Régression Linéaire Simple

Régression Linéaire Multiple

Séries Temporelles (ARIMA)

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Définition des Indicateurs (KPIs)</b>	<b>2</b>
<b>3</b>	<b>Résultats du Benchmark</b>	<b>2</b>
<b>4</b>	<b>Analyse et Interprétation</b>	<b>2</b>
4.1	Échec de la Régression Simple . . . . .	2
4.2	Illusion du modèle ARIMA . . . . .	2
4.3	Supériorité de la Régression Multiple . . . . .	3
<b>5</b>	<b>Conclusion Générale</b>	<b>3</b>

# 1 Introduction

L'objectif de ce projet est de développer un modèle prédictif capable d'estimer le salaire mensuel (*MonthlyIncome*) des employés en fonction de diverses caractéristiques RH (expérience, poste, département, etc.).

Nous avons testé et comparé trois approches distinctes pour identifier la plus performante pour notre application finale.

## 2 Définition des Indicateurs (KPIs)

Pour évaluer nos modèles, nous utilisons les métriques suivantes :

- **MAE (Mean Absolute Error)** : L'erreur moyenne en valeur absolue. Elle indique de combien de Dirhams le modèle se trompe en moyenne.
- **RMSE (Root Mean Squared Error)** : Pénalise les fortes erreurs. Utile pour vérifier la stabilité du modèle.
- **MAPE (Mean Absolute Percentage Error)** : L'erreur moyenne exprimée en pourcentage. C'est l'indicateur clé pour la décision métier.
- $R^2$  (**Coefficient de détermination**) : Indique la proportion de la variance du salaire expliquée par le modèle (de 0 à 1).

## 3 Résultats du Benchmark

Le tableau ci-dessous synthétise les performances obtenues sur notre jeu de données de test (20% du dataset).

Algorithme	$R^2$	MAPE	MAE	Verdict
1. Régression Simple	0.49	39.5 %	1 825 \$	✗ Rejeté
2. Régression Multiple	0.87	20.1 %	910 \$	✓ Retenu
3. ARIMA (Séries Temp.)	N/A	8.63 %	1 092 \$	~ Analytique

TABLE 1 – Comparatif de performance des modèles prédictifs

## 4 Analyse et Interprétation

### 4.1 Échec de la Régression Simple

Le modèle basé uniquement sur les années d'expérience (*TotalWorkingYears*) affiche une erreur de près de **40%**. Cela démontre que l'ancienneté seule ne suffit pas à justifier le salaire. Ce modèle est trop imprécis pour une utilisation individuelle.

### 4.2 Illusion du modèle ARIMA

Le modèle ARIMA affiche un excellent MAPE de **8.63%**. Cependant, ce résultat est trompeur pour notre besoin. ARIMA prédit la courbe du salaire *moyen* de l'entreprise. Il

attribue le même salaire à tous les employés ayant la même ancienneté, sans distinction de poste (Directeur vs Assistant). Il est utile pour les tendances macro-RH, mais pas pour la prédiction individuelle.

### 4.3 Supériorité de la Régression Multiple

La Régression Linéaire Multiple (incluant *JobRole* et *Department*) est le modèle le plus équilibré. Avec un  $R^2$  **de 0.87**, il explique près de 90% des variations de salaires. Son erreur moyenne de **20%** est acceptable pour un outil d'estimation.

## 5 Conclusion Générale

Suite à ce benchmark, nous avons décidé d'intégrer la **\*\*Régression Linéaire Multiple\*\*** dans notre application finale. C'est le seul algorithme capable de distinguer les spécificités métiers (Poste, Département) tout en conservant une bonne précision globale.