# Rapport du projet

### Objectif du projet

L'objectif est de prédire le salaire en USD (*salary_in_usd*) à partir de variables liées au poste, au niveau d'expérience, à la localisation et aux caractéristiques de l'entreprise. Deux modèles de régression ont été comparés :

- **Régression linéaire**
- **Random Forest Regressor**

La performance des modèles a été évaluée principalement avec :

- le **$R^2$** (coefficient de détermination)
- le **RMSE** (Root Mean Squared Error)

### Données et Prétraitement

Le jeu de données inclut des variables telles que : work_year, experience_level, employment_type, job_title, salary_currency, employee_residence, remote_ratio, company_location, company_size.

Les étapes de preprocessing ont consisté à :

- sélectionner les variables pertinentes
- gérer les valeurs manquantes
- encoder les variables catégorielles
- normaliser les données
- diviser le dataset (75% train / 25% test)

### Méthodologie

**1. Régression Linéaire**

- Test de **toutes les combinaisons possibles de 4 variables explicatives**
- Évaluation avec $R^2$ et RMSE

**2. Random Forest**

- Optimisation des hyperparamètres via **GridSearchCV** (n_estimators, max_depth, min_samples_split)
- Évaluation sur les mêmes critères ($R^2$, RMSE)

**Résultats**

❖ **Régression Linéaire**

- Performances très faibles
- $R^2$ souvent **négatif** en validation croisée
- RMSE élevé
  => Le modèle n'explique pas correctement la variance du salaire
  => Probable non-linéarité + multicolinéarité

❖ **Random Forest**

- Performances nettement supérieures
- $R^2$ moyen ≈ **0,395** (≈ 39,5% de variance expliquée)
- RMSE raisonnable
- Meilleure combinaison de variables :
  **experience_level, job_title, employee_residence, remote_ratio**

=> Random Forest capture mieux les relations complexes du dataset.

**Conclusion**

La régression linéaire n'est pas adaptée aux données étudiées, probablement à cause de relations non linéaires et des contraintes imposées.
Le modèle **Random Forest** est le plus performant et serait retenu pour des prédictions futures.

Des améliorations potentielles incluent :

- feature engineering avancé
- optimisation d'hyperparamètres avec recherche bayésienne
- validation croisée renforcée
- tests de modèles plus complexes (Boosting, Neural Networks)

**Annexes**

Sans optimisation d'hyperparamètres et avec un split 80/20 :

- Random Forest atteint environ **40% de variance expliquée**
- Meilleure combinaison :
  **experience_level, job_title, employee_residence, company_location**

Scripts fournis :

- pipeline.py : preprocessing
- main.py : entraînement + évaluation
  Résultats détaillés : Results.csv
  Tests additionnels : Annexe_Results.csv