RAPPORT FINAL – PROJET MACHINE LEARNING

Analyse RH : Attrition, Salaire, Segmentation des employés

---

1. Introduction

L'objectif de ce projet est d'analyser les données RH d'une entreprise fictive afin de prédire :

1. **L'attrition des employés** (Classification)
2. **Le salaire MonthlyIncome** (Régression)
3. **Les groupes RH homogènes** (Clustering)

Ces trois axes permettent :

- d'anticiper les départs volontaires,
- de comprendre les facteurs influençant le salaire,
- de segmenter les employés pour adapter les stratégies RH.

Le projet suit l'ensemble du cycle Machine Learning :

Préparation → Modélisation → Évaluation → Interprétation → Recommandations RH.

2. Description des données

Les données comportent **1470 employés** et **35 variables**, regroupées en :

- **Variables numériques**

Exemple : Age, DistanceFromHome, MonthlyRate, PercentSalaryHike, TotalWorkingYears…

- **Variables catégorielles nominales**

Department, JobRole, Gender, MaritalStatus, BusinessTravel…

- **Variables catégorielles ordinales**

JobSatisfaction (1–4), EnvironmentSatisfaction (1–4), WorkLifeBalance (1–4)...

- **Variable cible**

- **Attrition** (Oui/Non) pour la classification
- **MonthlyIncome** pour la régression

3. Prétraitement des données

3.1 Encodage

Type	Méthode	Justification
Numériques	StandardScaler	Normalisation pour SVM et modèles linéaires

Type	Méthode	Justification
Nominales	OneHotEncoder	Évite les relations ordinales fictives
Ordinales	OrdinalEncoder	Respecte l'ordre des niveaux de satisfaction

3.2 Pipeline

Un **ColumnTransformer** puis un **Pipeline** garantissent :

- pas de fuite de données,
- reproductibilité,
- automatisation du preprocessing.

3.3 Split

- **80% train / 20% test**
- **Stratification sur Attrition**

4. Modélisation – Classification (Attrition)

4.1 Modèles testés

- KNN
- Logistic Regression
- SVM (RBF)
- Decision Tree
- Random Forest
- XGBoost

Évalués avec **Stratified 5-Fold** et **F1-macro**.

Résultats CV :

Modèle	F1-macro CV
KNN	0.586
Logistic Regression	0.650
SVM (RBF)	**0.667**
Decision Tree	0.573
Random Forest	0.592
XGBoost	0.661

Le meilleur modèle est SVM (RBF).

4.2 Optimisation – GridSearchCV (SVM)

Hyperparamètre	Valeur optimale
C	1
gamma	0.01
PCA	0.95

Score CV optimal : 0.6666

4.3 Amélioration du Recall (RH = éviter les faux négatifs)

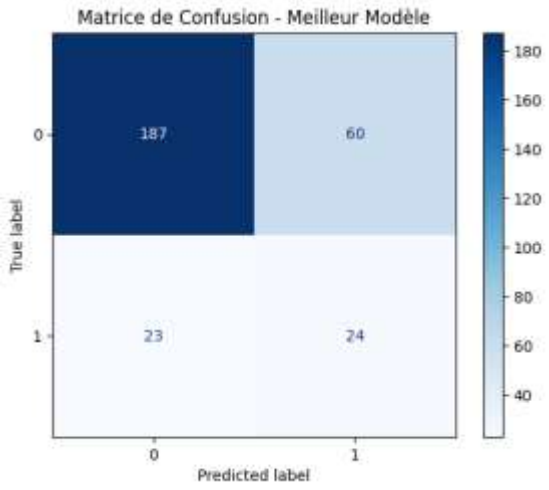
Un tuning du seuil de décision a permis :

Meilleur compromis :

- **Threshold = 0.35**
- Accuracy = 0.718
- Precision = 0.286
- Recall = **0.511**
- F1 = 0.366

Orientation métier : **favoriser le rappel pour détecter les employés à risque**.

Matrice de confusion :



Interprétation :

- **TP (24)** : employés réellement à risque correctement détectés
- **FN (23)** : employés à risque manqués par le modèle
- **FP (60)** : employés faussement signalés comme « risque »
- **TN (187)** : employés correctement identifiés comme stables

- Le modèle **capte plus de vrais départs**, au prix de plus de faux positifs.
- Ce compromis est logique lorsqu'on cherche à maximiser le recall.

5. Modélisation – Régression (Prédiction du salaire)

5.1 Variables

La variable **JobLevel** est retirée car parfaitement corrélée au salaire.

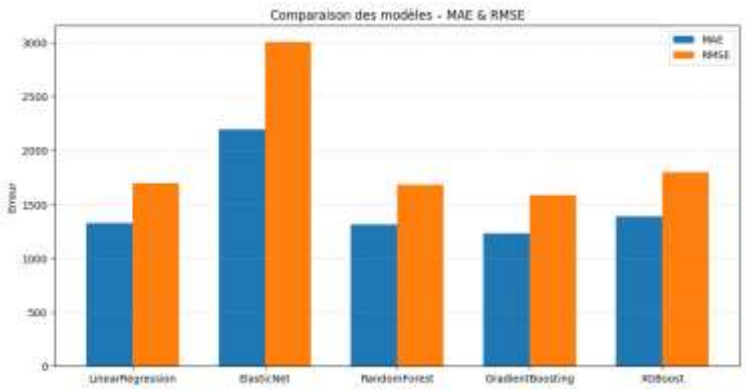
5.2 Modèles testés

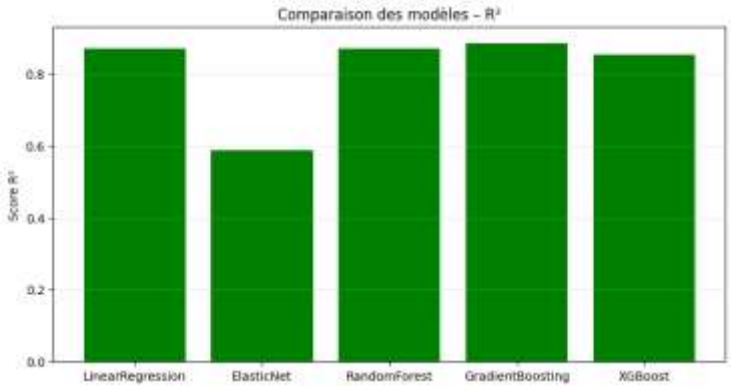
- Linear Regression
- ElasticNet
- RandomForestRegressor
- GradientBoostingRegressor

Résultats :

Modèle	MAE	RMSE	R^2
Linear Regression	1326	1692	0.869
ElasticNet	2192	3004	0.587
Random Forest	1315	1684	0.870
Gradient Boosting	**1232**	**1583**	**0.885**

Comparaison Graphique des modeles





Meilleur modèle : GradientBoostingRegressor

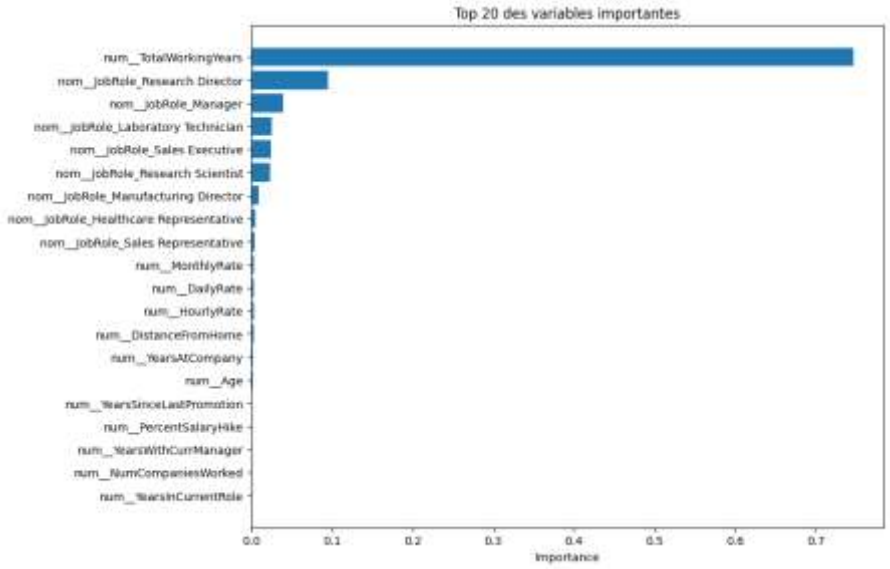
5.3 Optimisation GridSearch

Résultat final :

- MAE = **1244**
- R^2 = **0.882**

Le modèle de base (non optimisé) restait meilleur, ce qui arrive quand les hyperparamètres par défaut sont déjà optimaux.

Visualisation des variables importantes



6. Clustering – Segmentation des employés

6.1 Méthodes testées

- KMeans
- Agglomerative
- DBSCAN
- HDBSCAN
- Spectral Clustering

6.2 Scores obtenus

Méthode	Silhouette ↑	CH ↑	DB ↓	Clusters
KMeans	0.087	**110.5**	2.971	3
Agglomerative	0.092	110.2	2.435	3
DBSCAN	1 cluster	N/A	N/A	1
HDBSCAN	**0.161**	45.6	3.357	2
Spectral	**0.102**	76.1	**1.886**	3

Spectral Clustering donne les **clusters les plus distincts** (meilleur Davies-Bouldin).
HDBSCAN a meilleur silhouette mais seulement 2 clusters (moins utile métier).

7. Interprétation des clusters (Spectral)

Cluster 0 — Seniors performants (R&D / Sales)

- 37 ans, 12 ans d’expérience
- 7 ans dans l’entreprise
- JobSatisfaction = 4

- **Faible risque d’attrition**, employés stables et expérimentés.

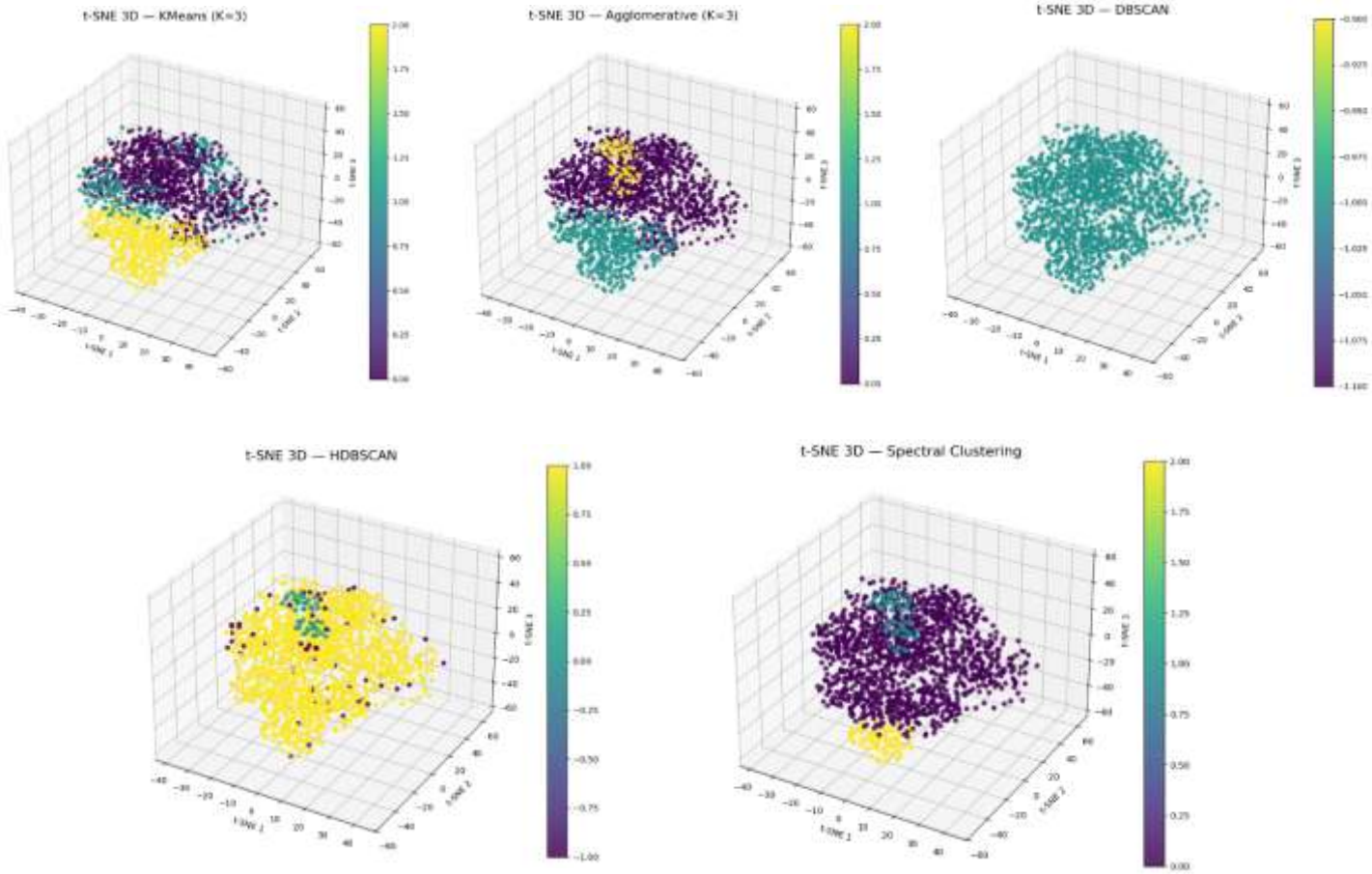
Cluster 1 — Employés administratifs stabilisés (HR)

- 36 ans, 10 ans d’expérience
- 6 ans dans l’entreprise
- Satisfaction moyenne
    - **Stables**, mais nécessité de formations continues.

Cluster 2 — Jeunes commerciaux en début de carrière

- 30 ans
- 5 ans d’expérience
- 3 ans dans l’entreprise
    - **Risque d’attrition élevé**, besoin d’accompagnement et de progression rapide.

8. Visualisations principales (TSNE 3D)



9. Recommandations RH

✓ Pour réduire l'attrition :

- Suivre de près les employés du **Cluster 2**
- Augmenter les opportunités d'évolution rapide
- Améliorer la satisfaction des jeunes employés

✓ Pour fidéliser les seniors :

- Programmes de leadership
- Reconnaissance et développement de carrière

✓ Pour optimiser le recrutement :

- Identifier le cluster cible selon le poste
- Réduire le turnover dans les équipes commerciales

10. Conclusion générale

Ce projet Machine Learning complet a permis de :

1. Prédire l'attrition

Le modèle **SVM optimisé + Threshold = 0.35** est capable d'identifier la majorité des employés à risque.

2. Expliquer le salaire MonthlyIncome

Le modèle **Gradient Boosting** atteint **$R^2 = 0.885$**, montrant une forte capacité prédictive.

3. Segmenter les employés en groupes cohérents

Le **Spectral Clustering** révèle trois segments RH exploitables pour une stratégie de rétention.