

# RAPPORT D'ANALYSE AVANCÉE : PROJET MACHINE LEARNING RH

Optimisation des Ressources Humaines par l'Intelligence Artificielle :  
Attrition, Rémunération et Segmentation des Collaborateurs



# Introduction et Objectifs du Projet

Ce projet vise à analyser les données de ressources humaines d'une organisation type afin d'anticiper et de relever les défis stratégiques liés au capital humain. Pour ce faire, nous exploitons le Machine Learning autour de trois objectifs principaux :



## Prédire l'Attrition

Identifier les employés présentant un risque de départ volontaire afin de déployer des actions préventives ciblées.



## Expliquer le Salaire

Analyser les facteurs déterminants du revenu mensuel (MonthlyIncome) pour une meilleure transparence et équité.



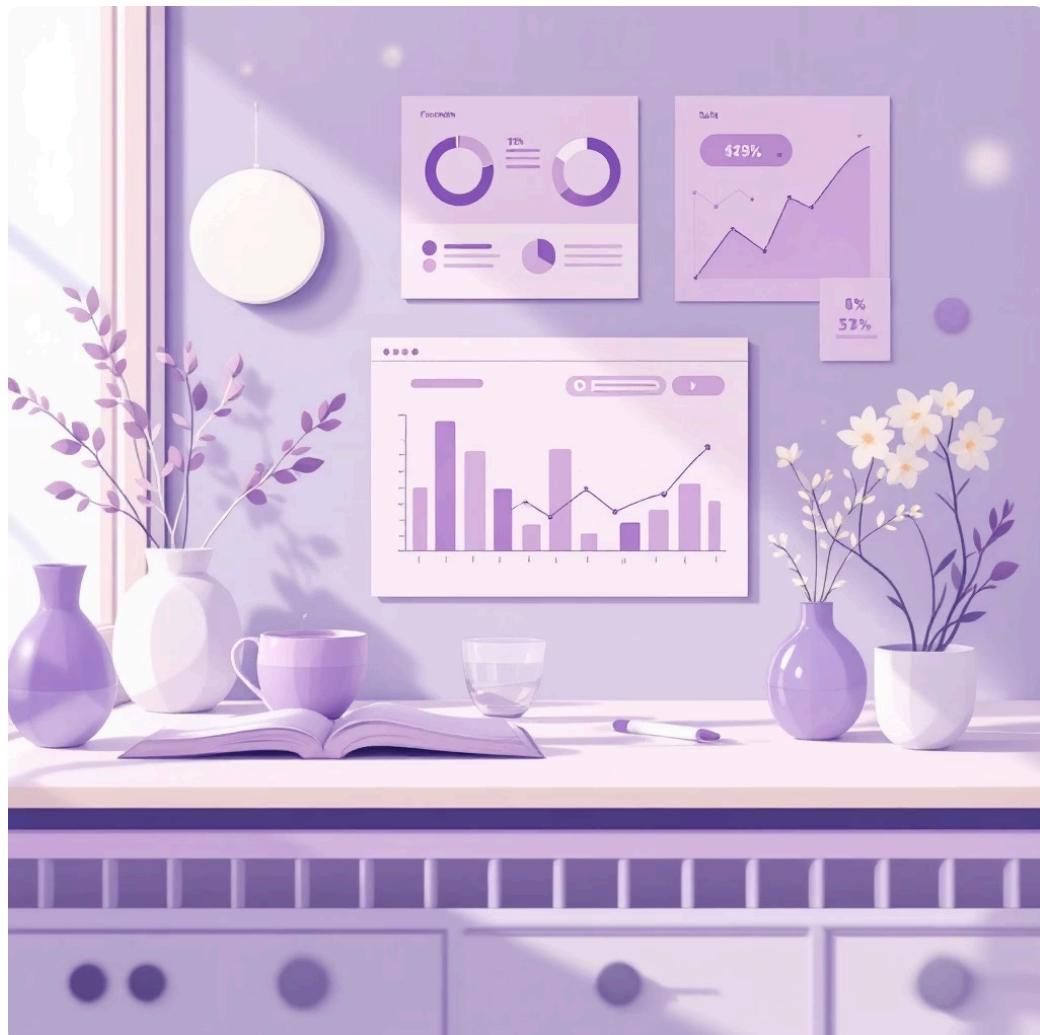
## Segmenter les Employés

Regrouper les collaborateurs en catégories homogènes afin d'élaborer des stratégies RH personnalisées.

Notre approche couvre l'intégralité du cycle de vie du Machine Learning : de la Préparation des données à la Modélisation, l'Évaluation, l'Interprétation des résultats et enfin l'élaboration de Recommandations RH concrètes.

# Description et Préparation des Données

Notre analyse repose sur un jeu de données complet et détaillé, comprenant 1470 dossiers d'employés et 35 variables distinctes.



## Classification des Variables :

- **Numériques** : L'âge, la distance au domicile, le revenu mensuel, le pourcentage d'augmentation salariale, l'ancienneté, etc.
- **Catégorielles Nominales** : Le département, le rôle, le genre, le statut marital, le type de voyage d'affaires, etc.
- **Catégorielles Ordinales** : La satisfaction au travail (1–4), la satisfaction de l'environnement (1–4), et l'équilibre vie professionnelle/personnelle (1–4).
- **Variables Cibles** : L'attrition (Oui/Non) pour nos modèles de classification, et le MonthlyIncome pour nos analyses de régression.

Le prétraitement des données a été rigoureusement mené, intégrant des méthodes d'encodage spécifiques et un pipeline robuste. Cette démarche assure la qualité optimale des données et la reproductibilité de nos modèles.

### StandardScaler

Normalisation des variables numériques, essentielle pour les modèles linéaires.

### OneHotEncoder

Encodage des variables nominales pour éviter toute relation ordinaire artificielle.

### OrdinalEncoder

Préservation de l'ordre inhérent aux niveaux des variables ordinaires.

# Modélisation Prédictive de l'Attrition : Stratégies de Classification

Afin de prédire l'attrition des employés, divers algorithmes de classification ont été évalués. Leur performance a été mesurée par le score F1-macro, validé au moyen d'une approche croisée stratifiée à 5 plis.



## SVM (RBF)

F1-macro : 0.667 (Meilleure performance préliminaire)



## Régression Logistique

F1-macro : 0.650



## XGBoost

F1-macro : 0.661

Parmi les modèles testés, le **SVM (RBF)** a démontré la meilleure performance F1-macro avant toute optimisation.

Modèle	F1-macro (CV)
KNN	0.586
Régression Logistique	0.650
SVM (RBF)	0.667
Arbre de Décision	0.573
Random Forest	0.592
XGBoost	0.661

# Optimisation du Modèle d'Attrition et Stratégie RH

L'optimisation du modèle SVM via GridSearchCV a affiné ses performances, en privilégiant le **Rappel**, une métrique cruciale pour les Ressources Humaines.

## Optimisation avec GridSearchCV (SVM)

C

Valeur optimale : 1

Gamma

Valeur optimale : 0.01

PCA

Valeur optimale : 0.95

Un score CV optimal de **0.6666** a été atteint suite à cette optimisation.

- Perspective métier :** Privilégier un rappel élevé assure l'identification d'un maximum d'employés à risque, quitte à accepter une légère baisse de précision.

## Amélioration du Rappel

Afin de maximiser la détection des départs potentiels, le seuil de décision (Threshold) a été ajusté à **0.35**.

- Précision = 0.718
- Exactitude = 0.286
- Rappel = 0.511
- Score F1 = 0.366



# Modélisation de Régression pour la Prédiction de Salaire

Afin de prédire le salaire mensuel (MonthlyIncome), diverses approches de modélisation par régression ont été explorées. La variable JobLevel a été exclue de cette analyse en raison de sa forte corrélation avec l'objectif.



Modèle	MAE	RMSE	R <sup>2</sup>
Linear Regression	1326	1692	0.869
ElasticNet	2192	3004	0.587
Random Forest	1315	1684	0.870
<b>Gradient Boosting</b>	<b>1232</b>	<b>1583</b>	<b>0.885</b>

Parmi les modèles évalués, le **GradientBoostingRegressor** se distingue par ses performances supérieures, affichant un R<sup>2</sup> de 0.885. Il est notable que l'optimisation par GridSearch n'a pas permis de surpasser les performances initiales du modèle, indiquant que ses hyperparamètres par défaut étaient déjà optimalement configurés.



# Segmentation des Employés par Clustering

Afin d'optimiser la gestion des ressources humaines, nous avons appliqué diverses méthodes de clustering pour identifier des groupes d'employés homogènes.

## KMeans

Silhouette: 0.087, 3 Clusters

## Agglomerative

Silhouette: 0.092, 3 Clusters

## DBSCAN

1 Cluster (pertinence limitée)

## HDBSCAN

Silhouette: 0.161, 2 Clusters

## Spectral Clustering

Silhouette: 0.102, 3 Clusters (Davies-Bouldin optimisé)

Malgré un score de silhouette supérieur pour HDBSCAN, le **Spectral Clustering** a été sélectionné en raison de son indice Davies-Bouldin inférieur, révélant des clusters intrinsèquement plus distincts et mieux définis.

# Interprétation des Clusters (Spectral Clustering)

L'analyse des trois clusters, identifiés par le Spectral Clustering, met en lumière des profils d'employés distincts. Chaque groupe présente des caractéristiques uniques qui appellent des stratégies RH ciblées et adaptées.

## Cluster 0 : Seniors performants

Ce groupe se compose de professionnels expérimentés, principalement au sein des départements R&D et Ventes (âge moyen : 37 ans, 12 ans d'ancienneté). Ils affichent un niveau de satisfaction élevé et présentent un **faible risque d'attrition**.

## Cluster 1 : Employés administratifs

Majoritairement rattachés aux fonctions administratives et RH, ces employés possèdent une expérience solide (environ 36 ans, 10 ans d'ancienneté). Leur satisfaction est modérée, indiquant qu'ils sont **stables mais nécessitent un soutien par la formation continue**.

## Cluster 2 : Jeunes commerciaux

Ce cluster regroupe des profils plus jeunes (environ 30 ans, 5 ans d'expérience) et est fortement représenté dans les équipes commerciales. En début de carrière, ils présentent un **risque d'attrition élevé**, ce qui souligne l'importance d'un accompagnement renforcé et d'opportunités d'évolution rapides.

La compréhension approfondie de ces profils distincts permet d'orienter les interventions RH de manière plus efficace, contribuant ainsi à l'amélioration de la rétention et au développement stratégique des talents au sein de l'organisation.



# Recommandations RH Stratégiques

Suite à une analyse approfondie de l'attrition, des rémunérations et de la segmentation des employés, nous présentons ici nos recommandations stratégiques clés pour optimiser la gestion des Ressources Humaines.

## Réduire l'Attrition



- Assurer un suivi proactif du Cluster 2 (jeunes commerciaux).
- Accélérer l'accès à des opportunités d'évolution de carrière significatives.
- Renforcer la satisfaction et l'engagement des jeunes collaborateurs.

## Fidéliser les Talents Seniors



- Mettre en œuvre des programmes de leadership et de mentorat ciblés.
- Proposer des parcours de reconnaissance et de développement de carrière sur mesure.



## Optimiser le Recrutement

- Identifier précisément les profils correspondant aux clusters cibles pour chaque poste.
- Élaborer des stratégies efficaces pour réduire le turnover au sein des équipes commerciales.

# Synthèse Stratégique du Projet

Ce projet de Machine Learning a permis de développer des outils prédictifs et descriptifs cruciaux, optimisant ainsi une gestion RH proactive et stratégique.

## Prédiction Précise de l'Attrition

Le modèle SVM optimisé, doté d'un seuil de décision de 0.35, s'avère performant pour identifier les collaborateurs à risque, facilitant ainsi des interventions RH ciblées et préventives.

## Analyse Approfondie du Salaire (MonthlyIncome)

Le modèle Gradient Boosting, avec un coefficient de détermination  $R^2$  de 0.885, atteste d'une capacité prédictive élevée, mettant en lumière les principaux facteurs déterminants du salaire mensuel.

## Segmentation Stratégique des Employés

Le Spectral Clustering a révélé trois segments RH distincts et significatifs, offrant une base solide pour l'élaboration de stratégies de rétention et de développement personnalisées et efficaces.

