



HR Analytics : Enquête sur les Facteurs de Départ des Employés

Phase 1 – Préparation des Preuves et Premières
Interrogations (Data Engineering & EDA)

Auteurs

Riad ZAID

Youssef SADIQUI

M Saad OUSSAMA

Encadrants

Pr. EL AZHARI

Pr. BENJBARA CHAIMAE

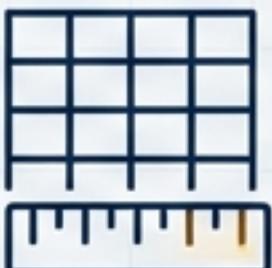
Le Dossier : Examen des Données Brutes



Source de la Donnée

Plateforme : Kaggle

Dataset : IBM HR Analytics Employee Attrition & Performance



Périmètre Initial

Taille : 1 470 observations (lignes)

Variables : 35 caractéristiques (colonnes)



Pistes d'Investigation Principales (Variables Clés)

Cible de l'enquête : Attrition (le départ de l'employé, Oui/Non)

Données quantitatives : Age, MonthlyIncome, TotalWorkingYears, etc.

Données qualitatives : Department, JobRole, Gender, etc.

Fiabiliser les Preuves : Nettoyage et Standardisation

1



Élimination des Doubles

Vérification et suppression des entrées dupliquées via `df.drop_duplicates()`.

2



Retrait des Informations Superflues

Suppression des colonnes constantes n'apportant aucune valeur à l'analyse ('EmployeeCount', 'Over18', 'StandardHours').

3



Vérification de Complétude

Confirmation de l'absence de valeurs manquantes dans le jeu de données.

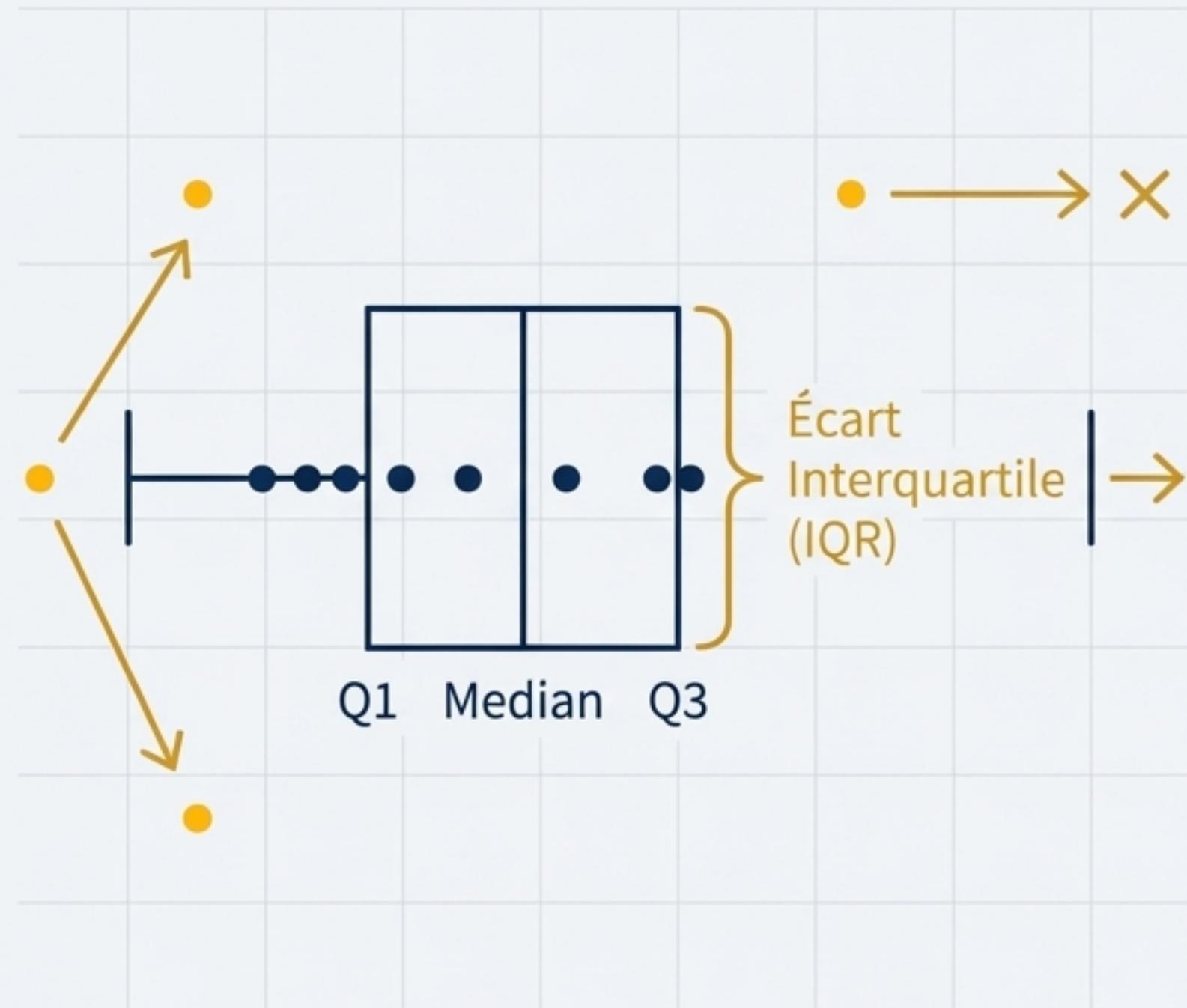
4



Standardisation du Texte

Harmonisation des entrées textuelles pour assurer la cohérence (Exemple : 'travel_rarely' unifié en 'Travel_Rarely').

Isoler les Anomalies : Le Traitement des Salaires Extrêmes



Méthodologie Appliquée

Technique : Détection par la méthode de l'écart interquartile (IQR), une approche statistique robuste pour identifier les valeurs aberrantes.

Constat

114 salaires ont été identifiés comme extrêmes par rapport à la distribution générale.

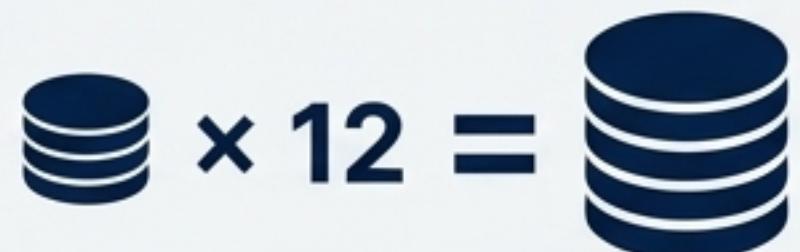
Action Corrective

Ces observations ont été retirées pour éviter de fausser les analyses de corrélation et les moyennes, garantissant ainsi une vision plus juste de la réalité de l'entreprise.

Enrichir le Dossier : Création de Nouvelles Variables d'Analyse (Feature Engineering)

Nouvelles Pistes Crées :

``AnnualIncome``



``MonthlyIncome``

``AnnualIncome``

Calcul du revenu annuel (``MonthlyIncome`` × 12) pour une perspective financière globale.

``Attrition_Numeric``

`Yes`

`No`



`1`

`0`

Conversion de la variable cible ``Attrition`` en format binaire (``Yes`` → 1, ``No`` → 0) pour permettre les calculs de corrélation et la future modélisation.

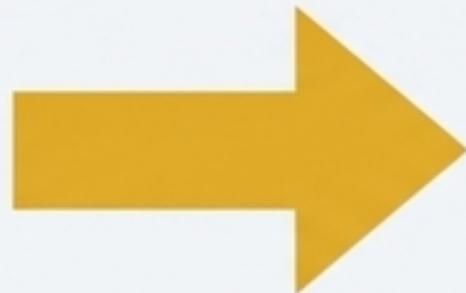
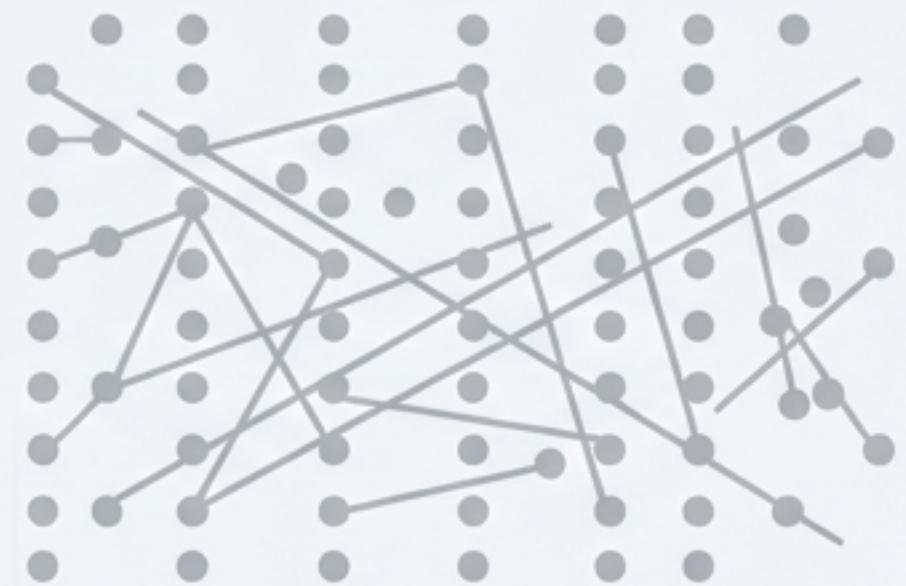
`Indexation Unique`



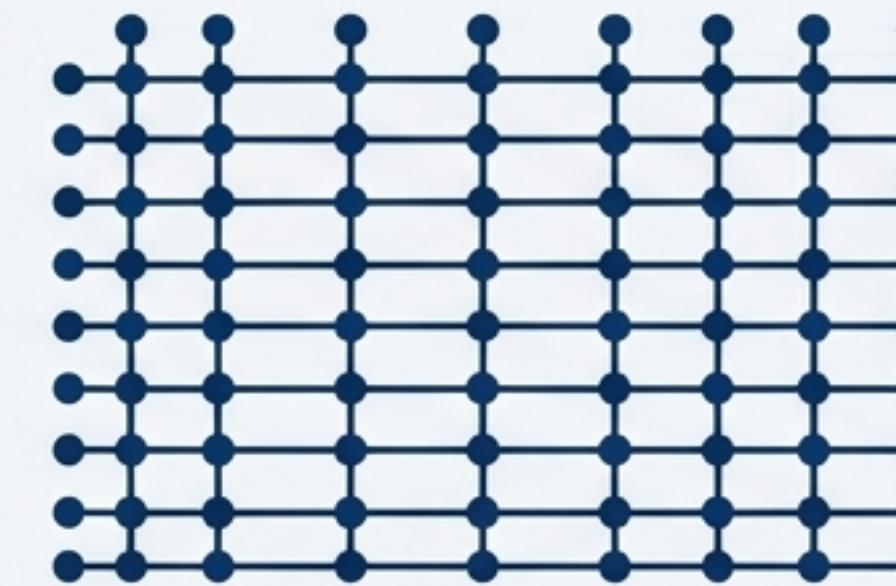
Utilisation de ``EmployeeNumber`` comme clé primaire pour un suivi précis de chaque individu.

Bilan de la Préparation : De la Donnée Brute à la Preuve Fiable

AVANT



APRÈS



Taille : 1 470 lignes × 35 colonnes

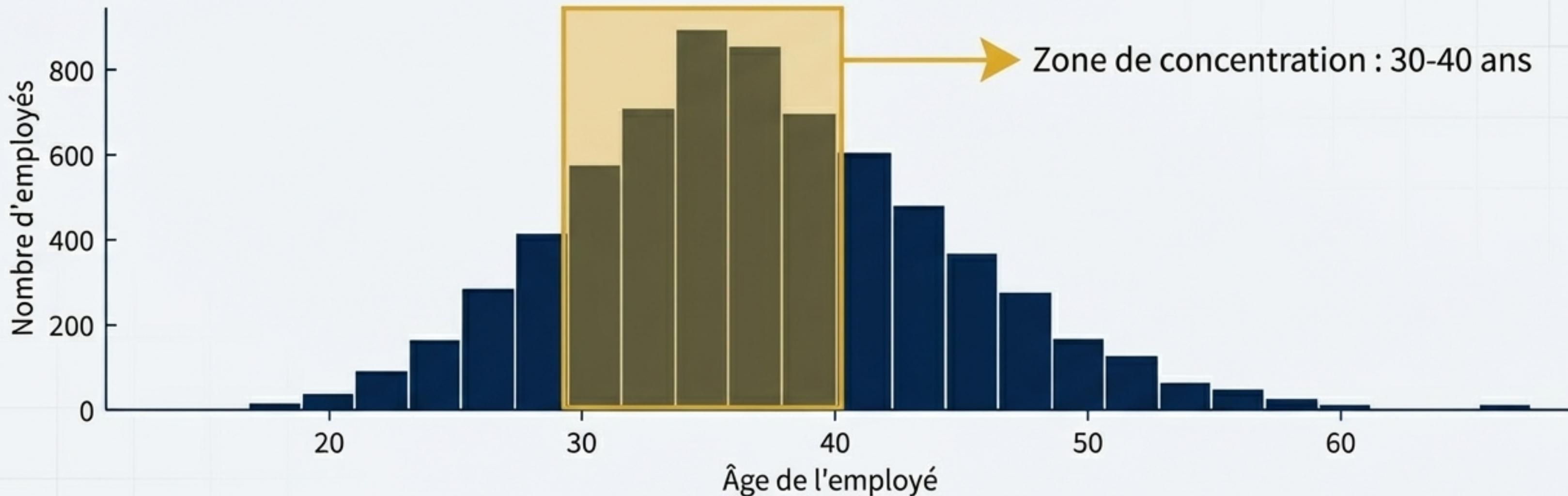
État : Données brutes, non standardisées,
avec anomalies.

Taille : 1 356 lignes × 32 colonnes

État : Jeu de données nettoyé,
structuré et enrichi.

Conclusion : Le dossier est maintenant prêt pour l'interrogatoire. Le jeu de données final est exporté (HR_Analytics_Structure_Complet.csv).

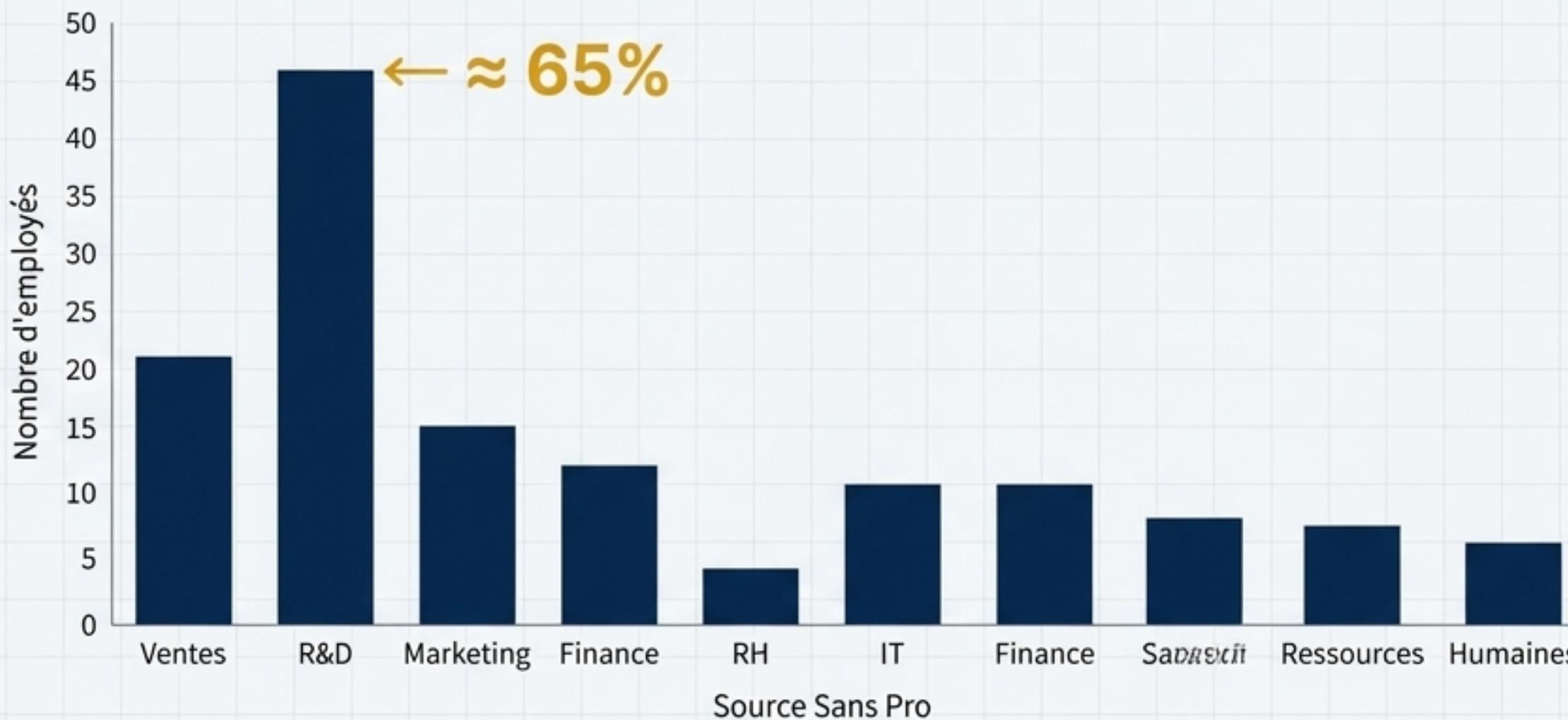
Premier Interrogatoire : Quelle est la structure d'âge de l'effectif ?



Premier Indice Révélé

La population de l'entreprise est majoritairement concentrée dans la tranche d'âge des 30-40 ans, suivant une distribution en cloche typique d'une organisation mature.

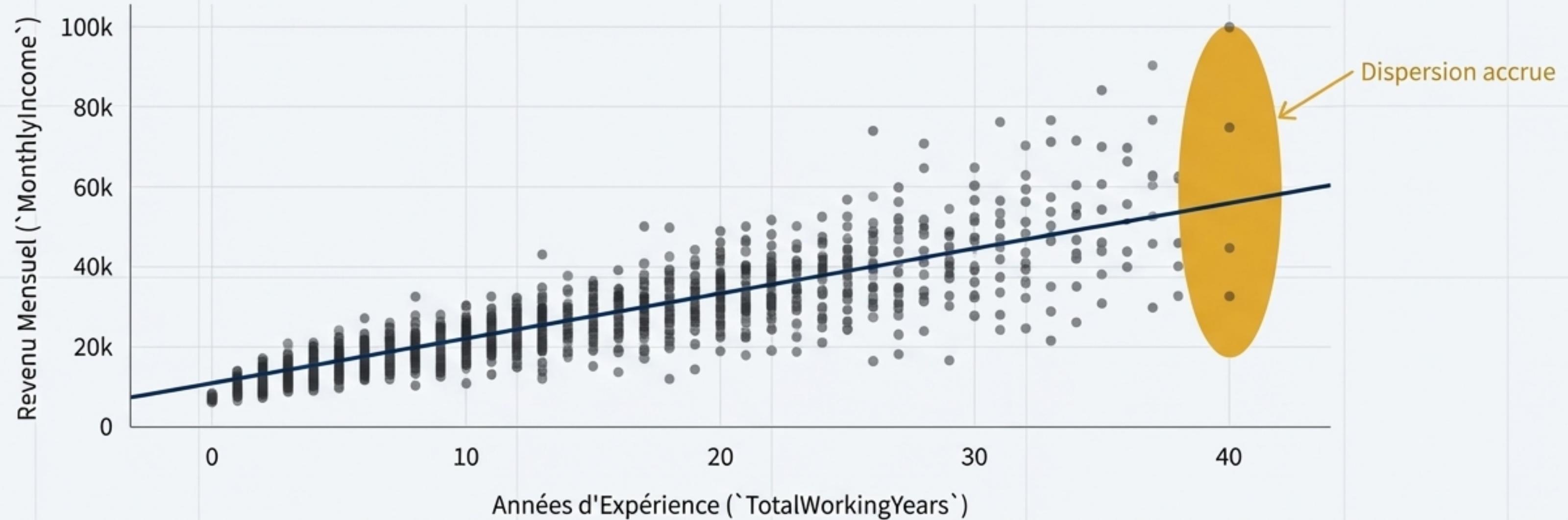
Répartition des Forces : Quel département domine l'organisation ?



Constat Structurel

Le département Recherche & Développement (R&D) constitue le cœur de l'effectif, représentant environ 65% du total. À l'inverse, les Ressources Humaines (RH) forment le pôle le plus restreint.

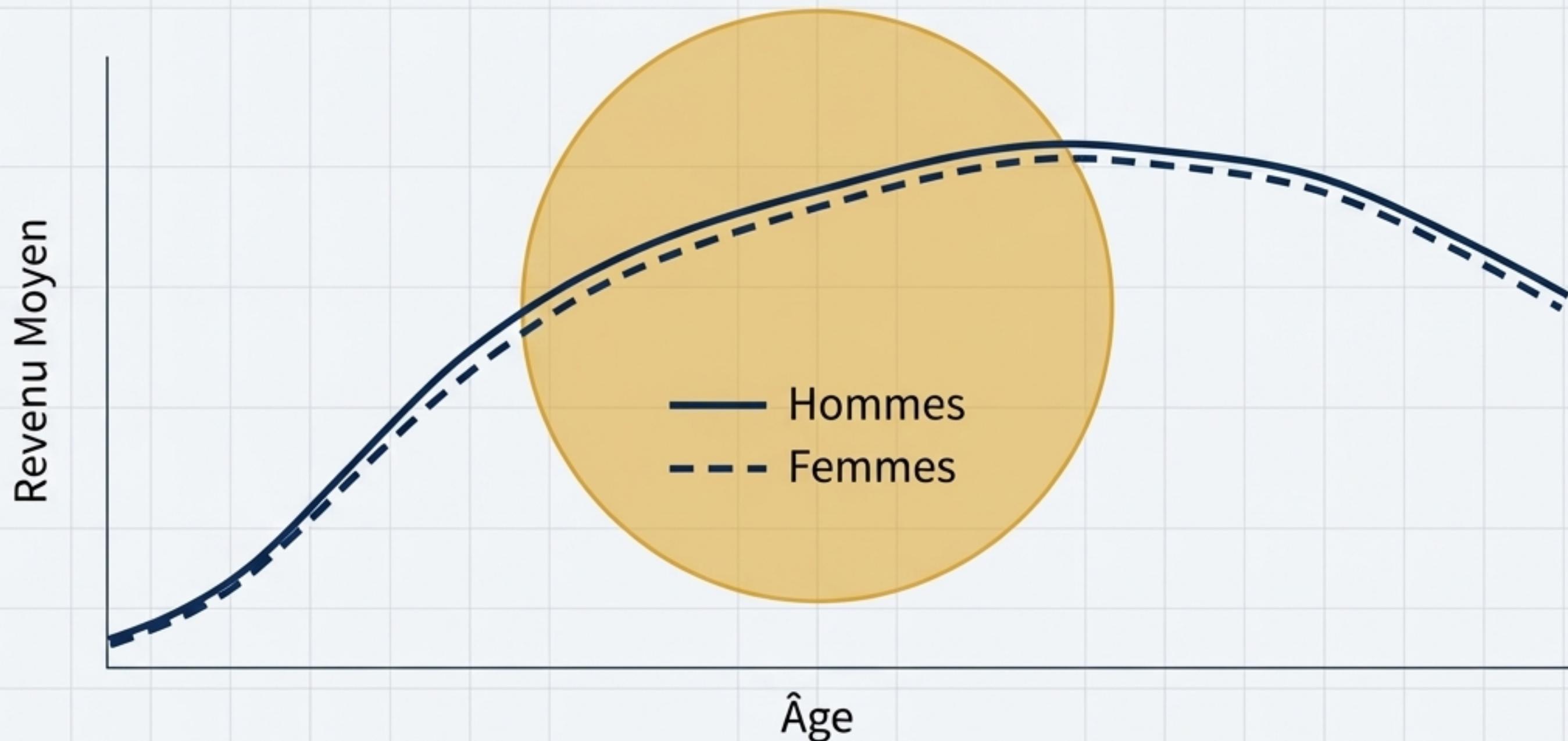
L'Expérience Paie-t-elle ? Corrélation entre Ancienneté et Revenu



Conclusion de l'Analyse

Une forte corrélation positive est confirmée. Le revenu augmente de manière significative avec l'expérience. On note également que la dispersion des salaires s'accentue avec l'ancienneté, suggérant une diversification des parcours de carrière au fil du temps.

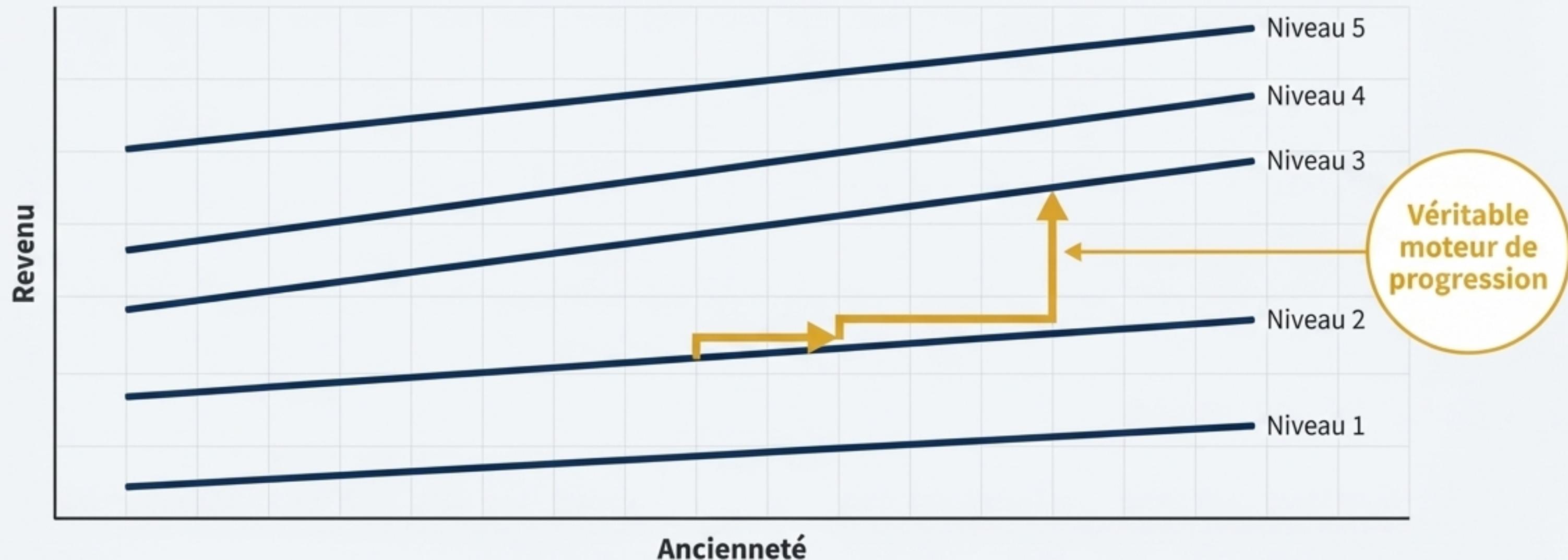
Enquête sur l'Équité : Le genre influence-t-il la rémunération ?



Verdict des Données

L'analyse ne révèle aucun écart significatif entre les courbes de revenu moyen des hommes et des femmes. À âge et expérience équivalents, la politique salariale semble équitable sur le critère du genre.

Mécanique d'Avancement : La progression est-elle liée au grade ou à l'ancienneté ?



Révélation Clé

La progression salariale est très nettement structurée par les paliers hiérarchiques. L'avancement au grade supérieur (`JobLevel`) est le véritable moteur de l'augmentation de revenu, bien plus que l'accumulation d'années d'ancienneté au sein d'un même niveau.

La Confrontation : Croiser les indices pour révéler le profil des départs

Approche : Analyse Multivariée 5D

Axe X : Âge

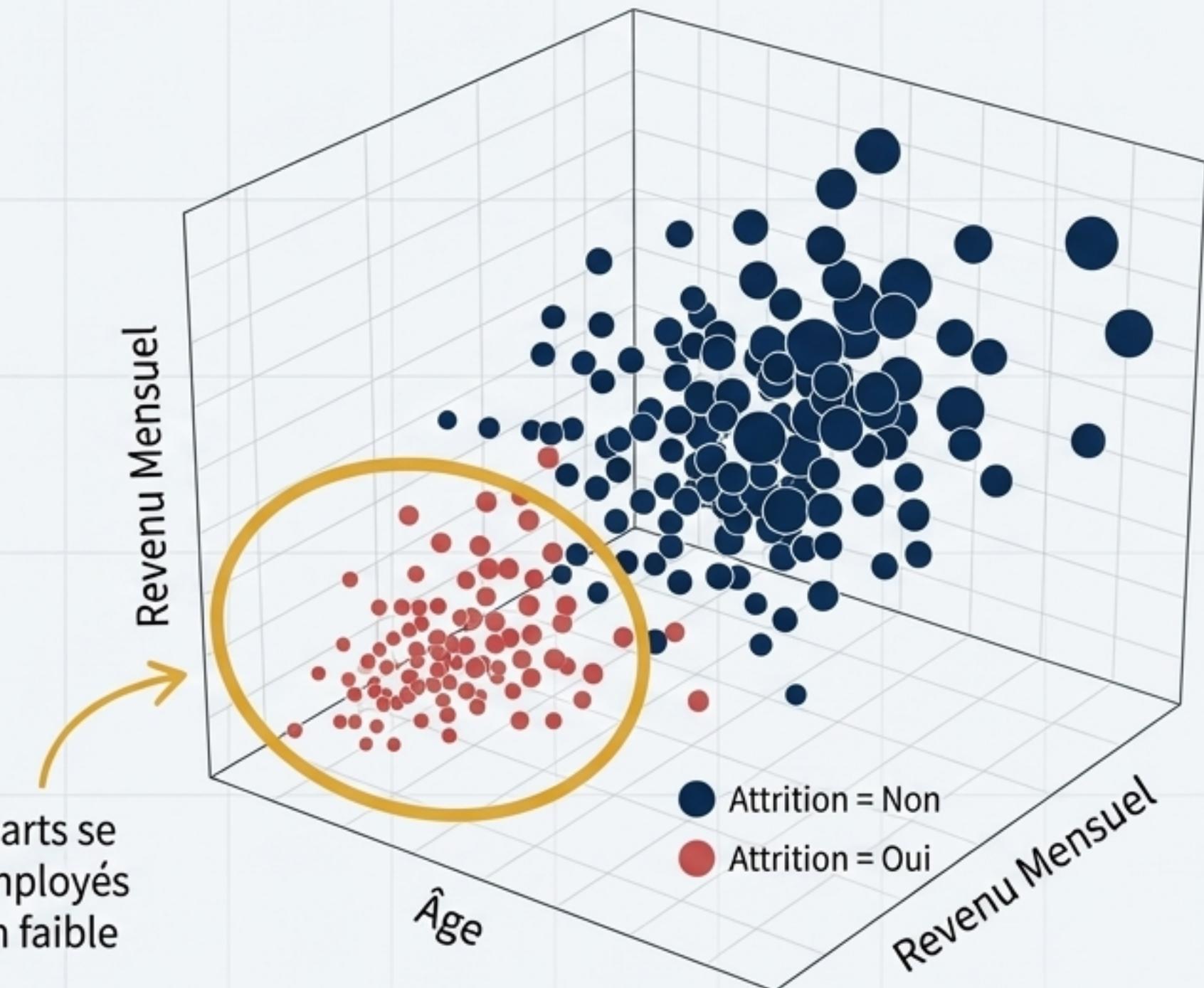
Axe Y : Revenu Mensuel

Axe Z (implicite) : Années d'Expérience

Couleur des points : Attrition (Oui/Non)

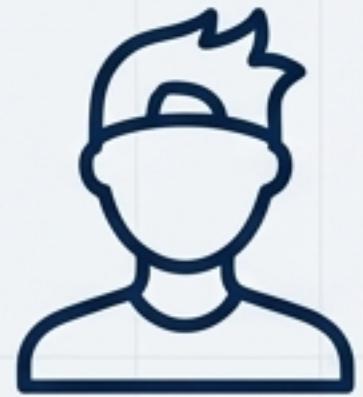
Taille des points : Niveau de Satisfaction

Le Schéma qui **Émerge** : Les départs se concentrent massivement ici : employés jeunes, à faible revenu et avec un faible niveau de satisfaction.



Portrait-Robot : Profil de l'Employé à Risque de Départ

Notre analyse exploratoire dresse le profil type d'un employé quittant l'entreprise :



Profil Démographique

Jeune et en début de carrière.



Situation Financière

Faible revenu mensuel.



Ressenti Professionnel

Faible niveau de satisfaction au travail ('JobSatisfaction').



Positionnement

Souvent situé à des niveaux hiérarchiques inférieurs.

Prochaines Étapes de l'Enquête : De l'Exploration à la Prédiction



Fondations Établies

Les schémas et profils identifiés dans cette phase constituent une base de connaissances solide et factuelle sur les dynamiques de départ.



Vers l'Action Proactive

La prochaine phase consistera à utiliser ces 'preuves' pour entraîner un modèle d'apprentissage automatique.

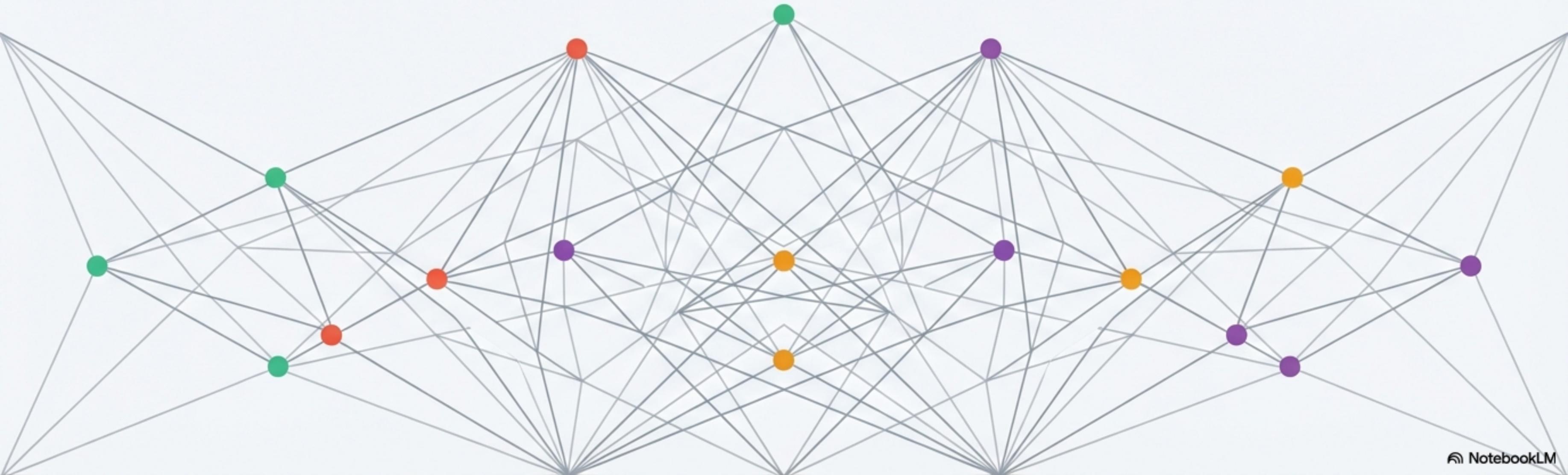


Objectif Final

Développer un outil prédictif capable d'identifier en amont les employés à risque d'attrition, afin de permettre des actions RH ciblées et préventives.

Analytics RH : Transformer les Données en Décisions Stratégiques

Application de la Modélisation Prédictive aux Défis Clés du Capital Humain (Phases 3 à 5)



L'Évolution Stratégique : De la Gestion Réactive à l'Intelligence Prédictive

La fonction RH passe d'un rôle administratif à celui de partenaire stratégique. En exploitant la modélisation prédictive, nous pouvons désormais anticiper les besoins, optimiser les politiques et agir sur les défis humains avant qu'ils n'impactent l'entreprise.



Défi 1 : La Rémunération

Établir une Politique Salariale Juste, Compétitive et Basée sur les Données



Le Défi

- * **'Question Métier'**: Comment estimer le `Monthly Income` d'un employé ou d'un candidat de manière objective et cohérente ?
- * ***Enjeux Stratégiques***:
 - **Équité Interne**: Garantir une structure salariale non-discriminatoire.
 - **Compétitivité Externe**: Attirer et retenir les meilleurs talents face au marché.
 - **Maîtrise Budgétaire**: Planifier la masse salariale avec précision.

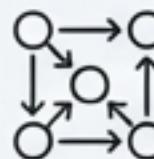
La Régression pour Modéliser et Prédire les Niveaux de Salaire

La Boîte à Outils Algorithmique



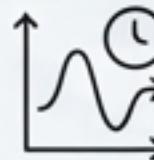
Régression Linéaire Simple

Pour établir une ligne de base (ex: salaire vs. années d'expérience).



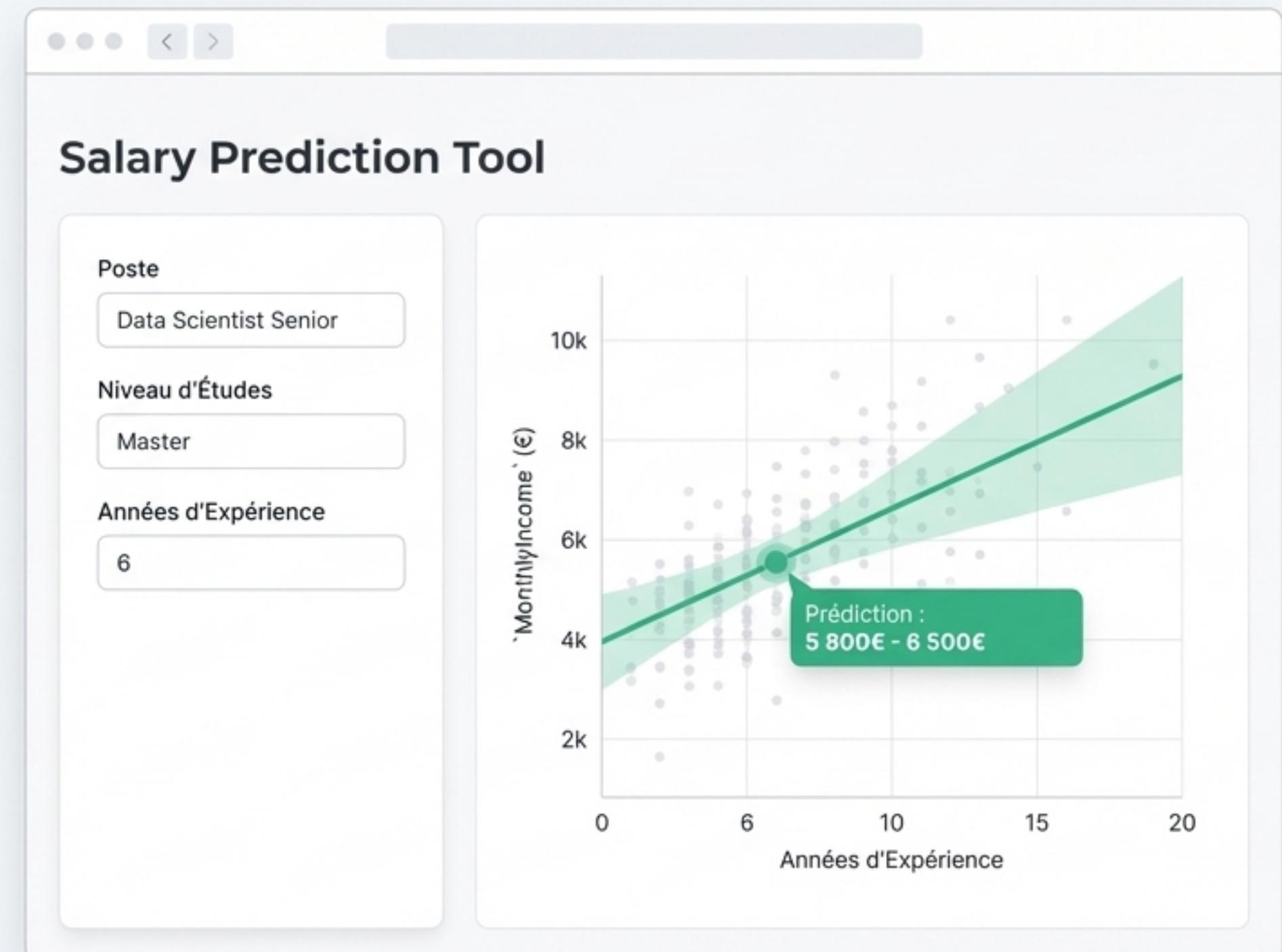
Régression Linéaire Multiple

Pour un modèle robuste intégrant de multiples facteurs (poste, diplôme, performance, etc.).



ARIMA

Pour analyser et prédire les tendances temporelles des salaires au sein de l'entreprise ou du marché.



L'Application et l'Impact

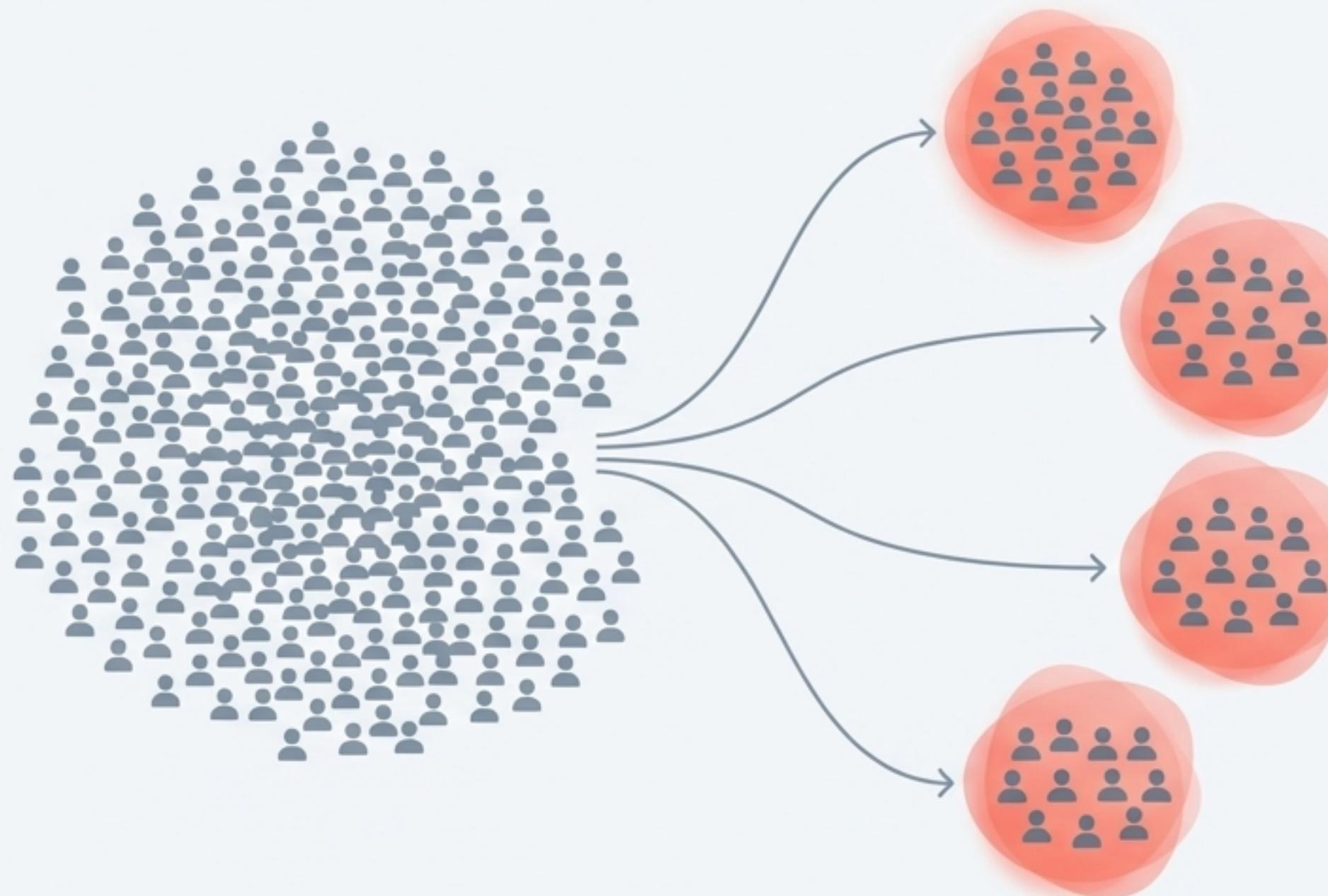
- **Output**

Un outil d'aide à la décision qui prédit une fourchette de salaire en fonction du profil.

- **Impact**

Offres de recrutement calibrées, revues salariales objectives, et anticipation des dérives budgétaires.

Identifier des Segments d'Employés pour Personnaliser les Politiques RH



Le Défi

- **Question Métier**

Nos employés sont-ils un groupe monolithique, ou existe-t-il des 'profils' distincts avec des besoins et des motivations différents ?

- **Enjeux Stratégiques**

- **Engagement:** Augmenter l'engagement en répondant aux attentes spécifiques de chaque groupe.
- **Développement:** Créer des parcours de carrière et de formation sur-mesure.
- **Efficacité des Politiques RH:** Allouer les ressources là où elles auront le plus d'impact.

Le Clustering pour Révéler les Groupes Homogènes au Sein des Effectifs



La Boîte à Outils Algorithmique

- **K-Means**

Pour créer un nombre prédéfini de segments distincts et non-superposés.

- **Clustering Hiérarchique (CAH)**

Pour explorer les relations et la proximité entre les groupes à différents niveaux de granularité.



L'Application et l'Impact

- **Output**

Identification de personas d'employés basés sur leurs données (performance, engagement, ancienneté, compétences, etc.).

- **Impact**

Politiques de télétravail adaptées, programmes de formation ciblés, communication interne personnalisée.



Les Piliers Stables

Ancienneté élevée, performance constante



Les Hauts Potentiels

Jeunes, performance > 90%, fort potentiel d'évolution



Les Experts Spécialisés

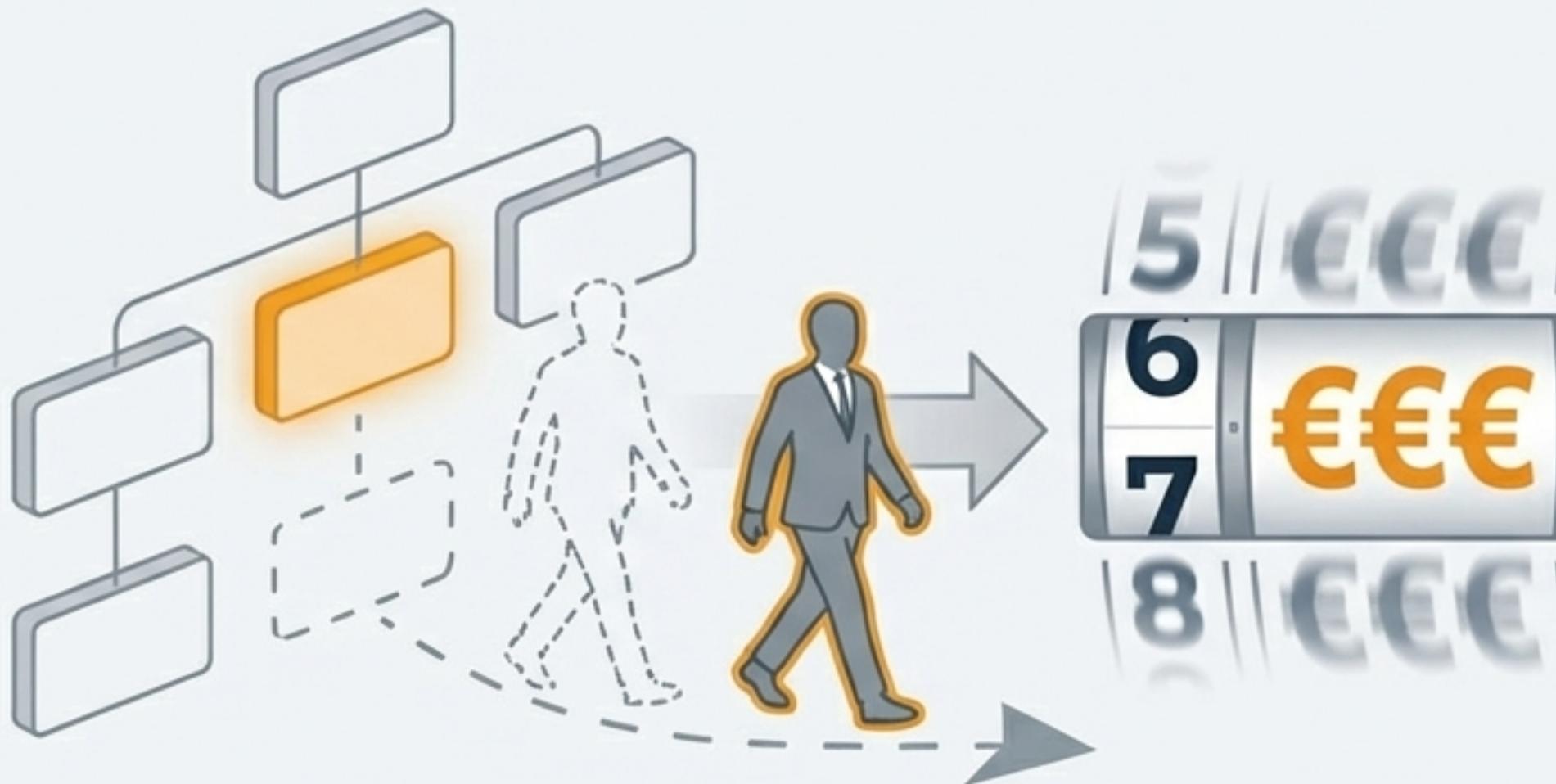
Rôle de niche, compétences rares



Les Talents à Risque

Bonne performance, faible satisfaction

Anticiper et Prévenir l'Attrition pour Conserver les Talents



Le Défi

- **Question Métier**

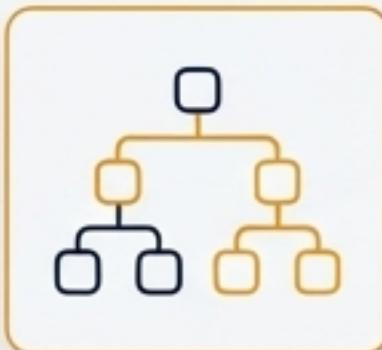
Quels sont les employés les plus susceptibles de quitter l'entreprise ('Attrition = Yes') dans les 6 prochains mois ?

- **Enjeux Stratégiques**

- **Coût:** Réduire les coûts directs et indirects liés au recrutement et à la formation de remplaçants.
- **Continuité:** Maintenir la connaissance et la productivité au sein des équipes.
- **Marque Employeur:** Éviter la perception d'un environnement de travail instable.

La Classification : Un Arsenal d'Algorithmes pour Prédire le Risque de Départ

Modèles pour l'Interprétabilité (Comprendre le "Pourquoi")

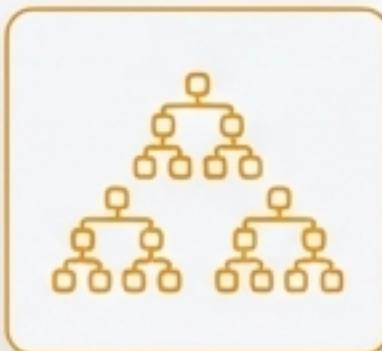


Arbre de Décision: Visualise les chemins de décision menant à l'attrition.

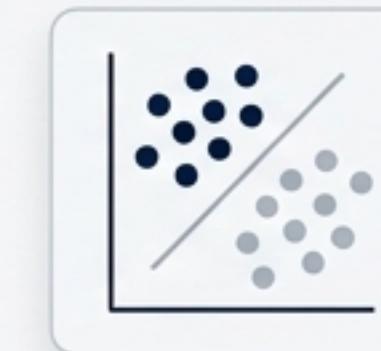


Régression Logistique: Identifie le poids de chaque facteur de risque (ex: salaire, distance domicile-travail).

Modèles pour la Performance Prédictive (Maximiser la Précision)



Random Forest: Agrège de multiples arbres de décision pour une prédiction robuste.

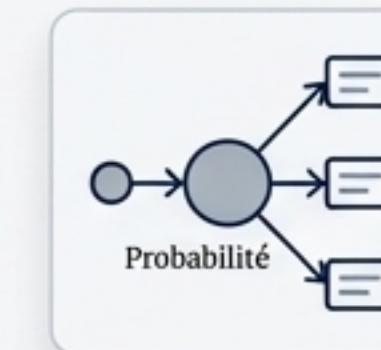


SVM (Support Vector Machine): Efficace pour trouver la frontière optimale entre les 'partants' et les 'restants'.

Autres Approches Efficaces



K-NN (K-Nearest Neighbors): Classe un employé en fonction de ses 'voisins' les plus proches.



Naïve Bayes: Modèle probabiliste rapide et simple, excellent comme baseline.

De la Prédiction à l'Action : Cibler les Interventions de Rétention

L'Application et l'Impact

- Output:** Un tableau de bord de risque d'attrition, mis à jour en continu, qui identifie les employés à risque et les principaux facteurs contributifs.

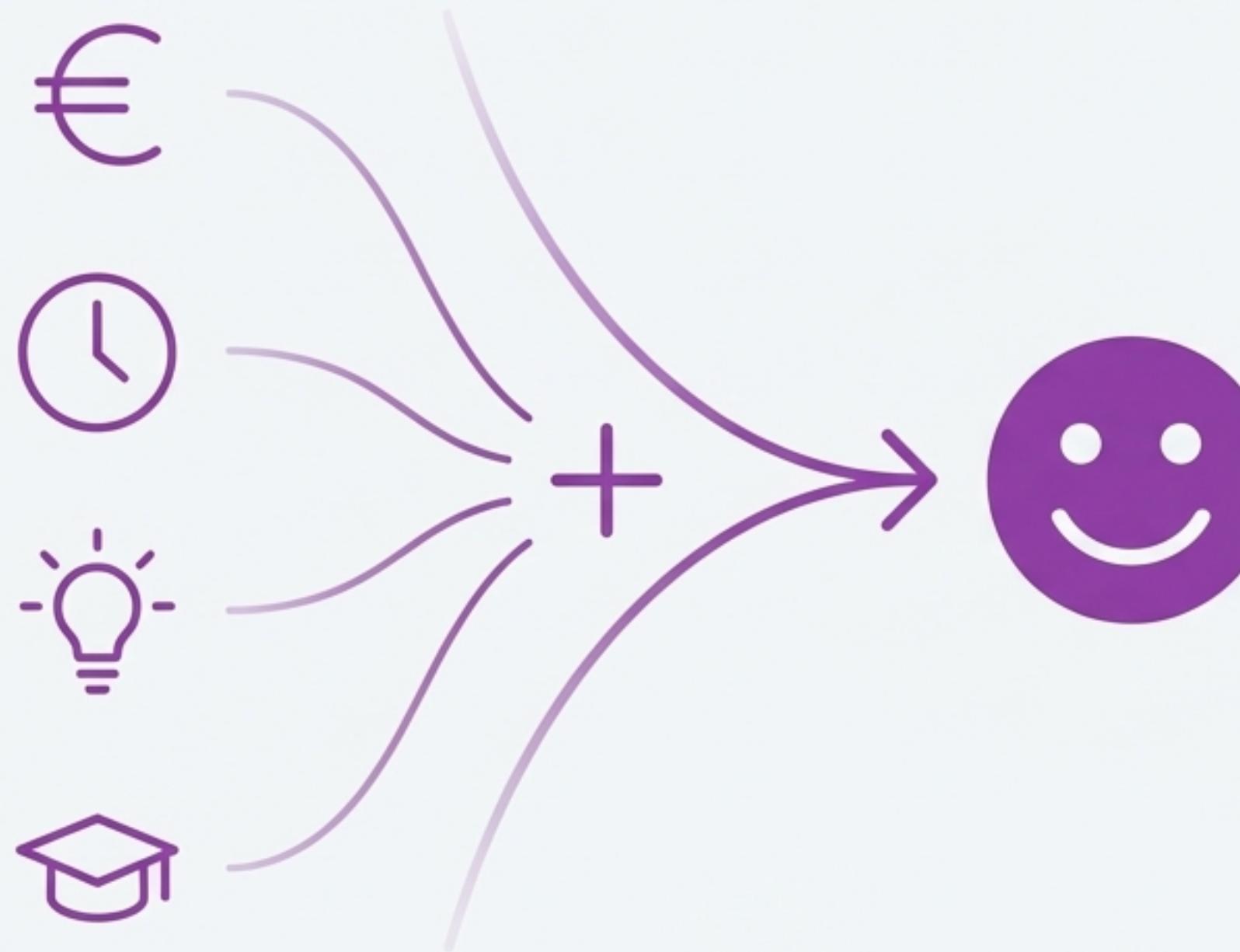
Impact

- Pour les Managers:** Permet d'initier des discussions de rétention ciblées et préventives.
- Pour les RH:** Aide à concevoir des plans d'action globaux (ex: révision de la politique de déplacement, programmes de mentorat).

Tableau de Bord du Risque d'Attrition - T3

Employé	Score de Risque	Principaux Facteurs de Risque	Statut
Employé A	85%	⌚ Heures supplémentaires élevées ↗ Absence d'augmentation récente 👤 Manager peu noté	● Action Requise
Employé B	72%	🚗 Long trajet domicile-travail 👉 Peu d'opportunités de formation 👤 Équipe en sureffectif	● Action Requise
Employé C	65%	👤 Changement de manager récent 😊 Satisfaction faible (sondage) ⌚ Heures supplémentaires élevées	○ À surveiller

Découvrir les Combinaisons de Facteurs qui Mènent à une Forte Satisfaction



Le Défi

Question Métier Au-delà des facteurs individuels (salaire, manager), quelles *combinaisons* de conditions créent un environnement où les employés sont le plus satisfaits ?

Enjeux Stratégiques

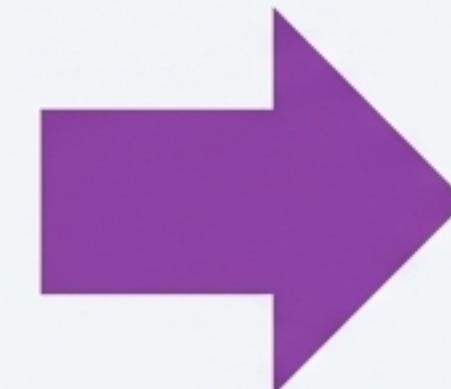
- **Culture d'Entreprise:** Construire une culture basée sur ce qui compte vraiment pour les employés.
- **Optimisation des Investissements RH:** Concentrer les efforts sur les initiatives ayant un effet multiplicateur.
- **Innovation:** Créer une ‘recette’ unique pour devenir un employeur de choix.

Les Règles d'Association pour Révéler les 'Recettes' de la Satisfaction

La Boîte à Outils Algorithmique

- **Algorithme Apriori:** Détourné de son usage initial en 'market basket analysis' (analyse du panier de la ménagère), cet algorithme identifie les ensembles de facteurs qui apparaissent fréquemment ensemble chez les employés les plus satisfaits.

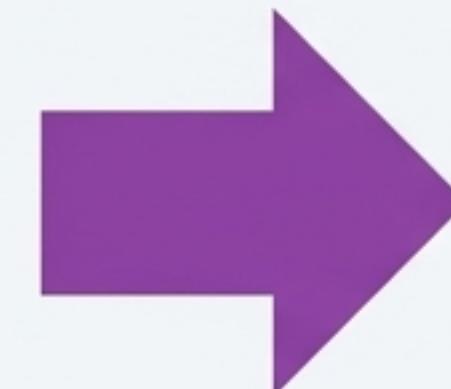
{Bon équilibre vie pro/perso,
Autonomie dans le travail}



{Satisfaction
Élevée}

(Confiance: 85%)

{Opportunités de formation,
Feedback régulier du manager}



{Satisfaction
Élevée}

(Confiance: 82%)

L'Application et l'Impact

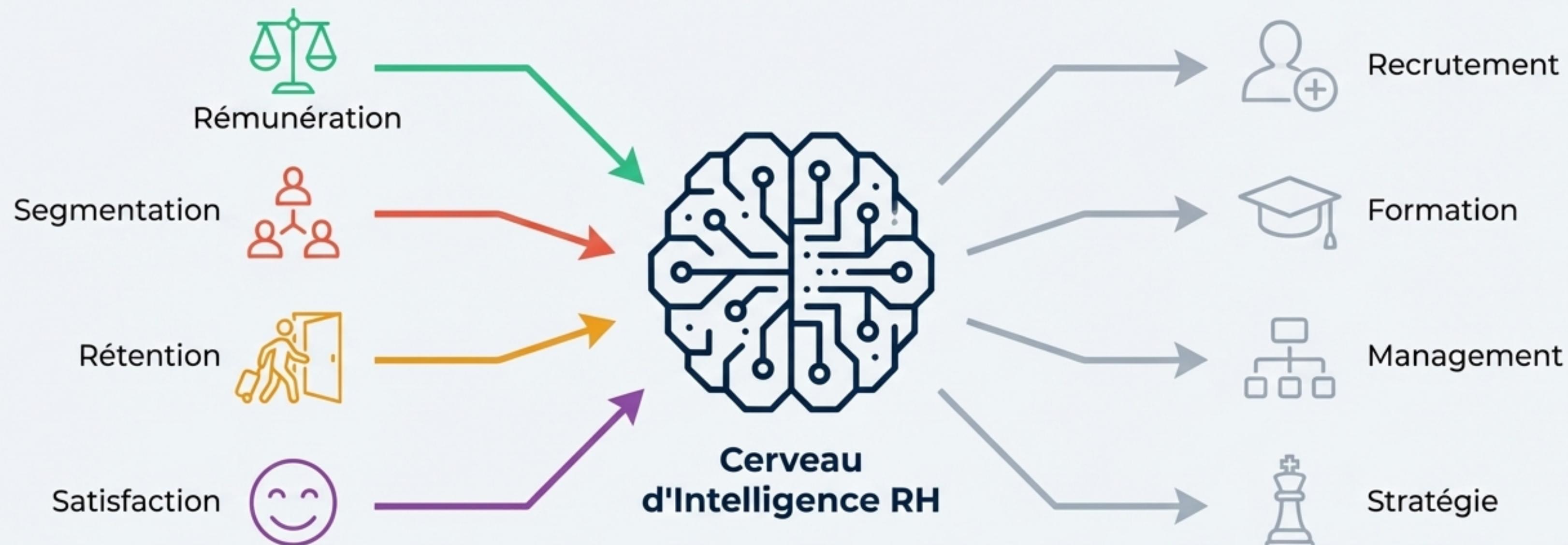
- **Output:** Des règles d'association claires et actionnables.
- **Impact:** Permet de concevoir des politiques RH qui ciblent les combinaisons gagnantes, plutôt que des facteurs isolés.

Une Boîte à Outils Algorithmique pour Chaque Défi RH Stratégique

Défi RH	Approche Analytique	Question Clé Adressée	Impact Business
 Rémunération Juste	Régression	Combien faut-il payer ?	Équité & Maîtrise des Coûts
 Personnalisation de l'Expérience	Clustering	Qui sont nos employés ?	Engagement & Efficacité
 Rétention des Talents	Classification	Qui risque de partir ?	Continuité & Stabilité
 Optimisation de la Satisfaction	Règles d'Association	Qu'est-ce qui génère la satisfaction ?	Culture & Marque Employeur

Vers un Écosystème RH Intégré et Augmenté par la Donnée

Vision: Ces modèles ne sont pas des solutions isolées, mais les briques fondamentales d'un système de gestion des talents intelligent. En les connectant, nous créons une vue à 360° du cycle de vie de l'employé, permettant des décisions plus rapides, plus justes et plus stratégiques à chaque étape.



NOTRE BOÎTE À OUTILS D'ANALYSE RH : LES MEILLEURS MODÈLES POUR CHAQUE DÉFI



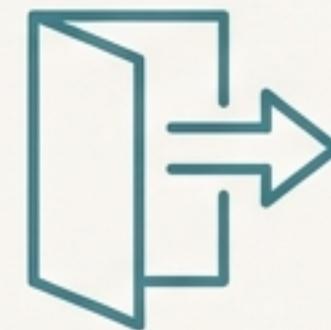
Défi : Prédire le Salaire

Solution Choisie : Régression Linéaire Multiple



Défi : Segmenter les Talents

Solution Choisie : K-Means



Défi : Anticiper l'Attrition

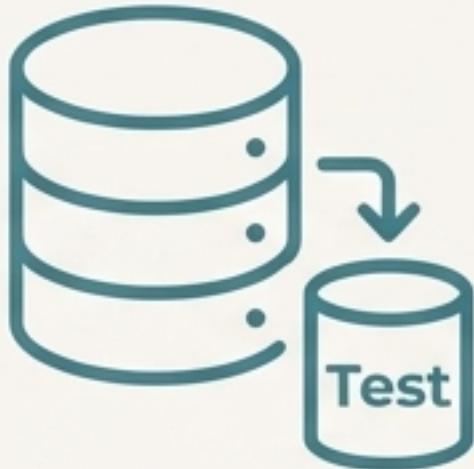
Solution Choisie : Random Forest



Défi : Comprendre la Satisfaction

Approche Validée : Règles d'Association

Une Méthodologie Rigoureuse pour des Recommandations Fiables



Jeu de Données Unifié

Tous les algorithmes ont été évalués sur un jeu de données unique, avec une répartition standard de 80% pour l'entraînement et 20% pour le test, garantissant une comparaison équitable.



Validation Croisée (5-Folds)

Nous avons utilisé une validation croisée à 5 plis ("5-folds") pour nous assurer que la performance mesurée des modèles est stable et non due au hasard de la répartition des données.



Métriques Standardisées

Chaque type de problème a été évalué avec des métriques standards et appropriées (R^2 , AUC, Score de Silhouette, etc.), permettant une comparaison objective des performances.

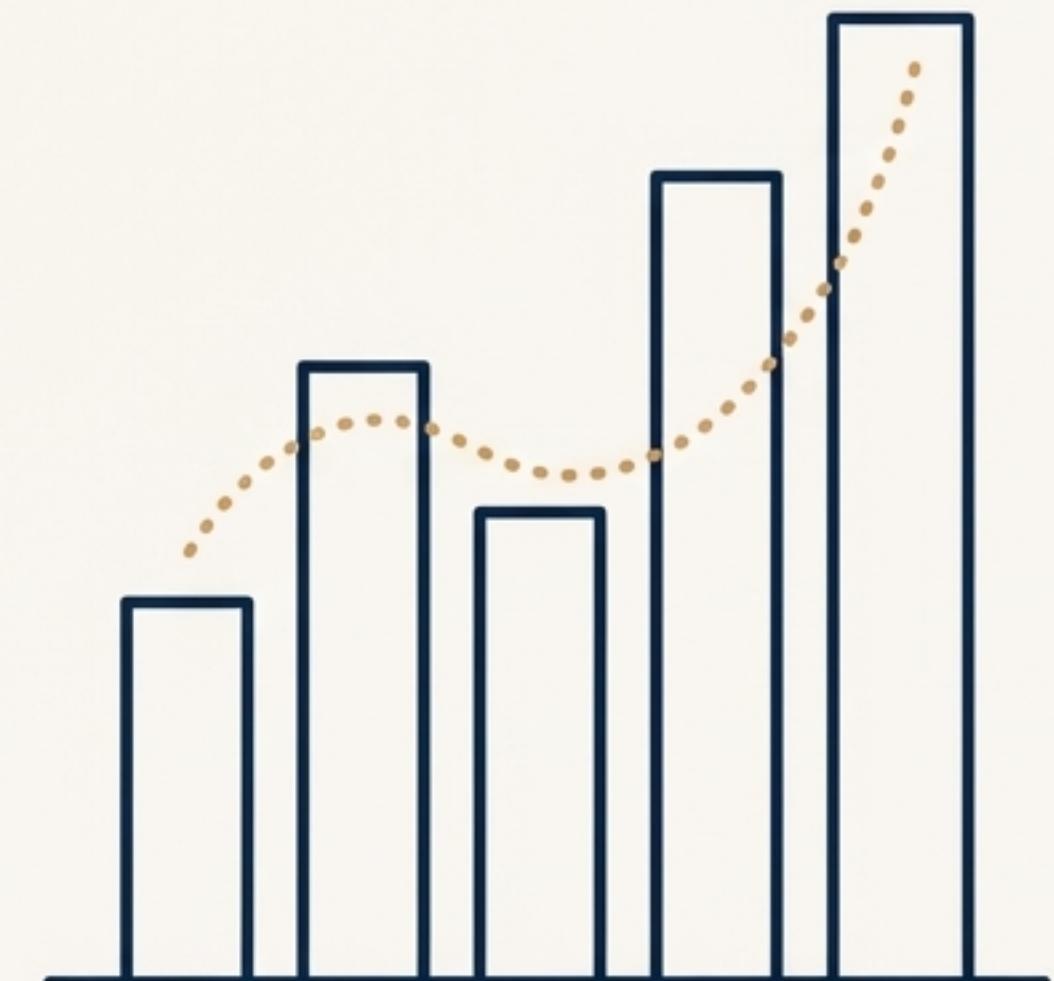
Définir des Niveaux de Rémunération Équitables et Compétitifs

Le Défi Stratégique

Comment s'assurer que notre politique de rémunération est à la fois juste en interne et attractive sur le marché ? Il est essentiel de modéliser les salaires sur des facteurs objectifs pour éviter les biais et justifier les décisions.

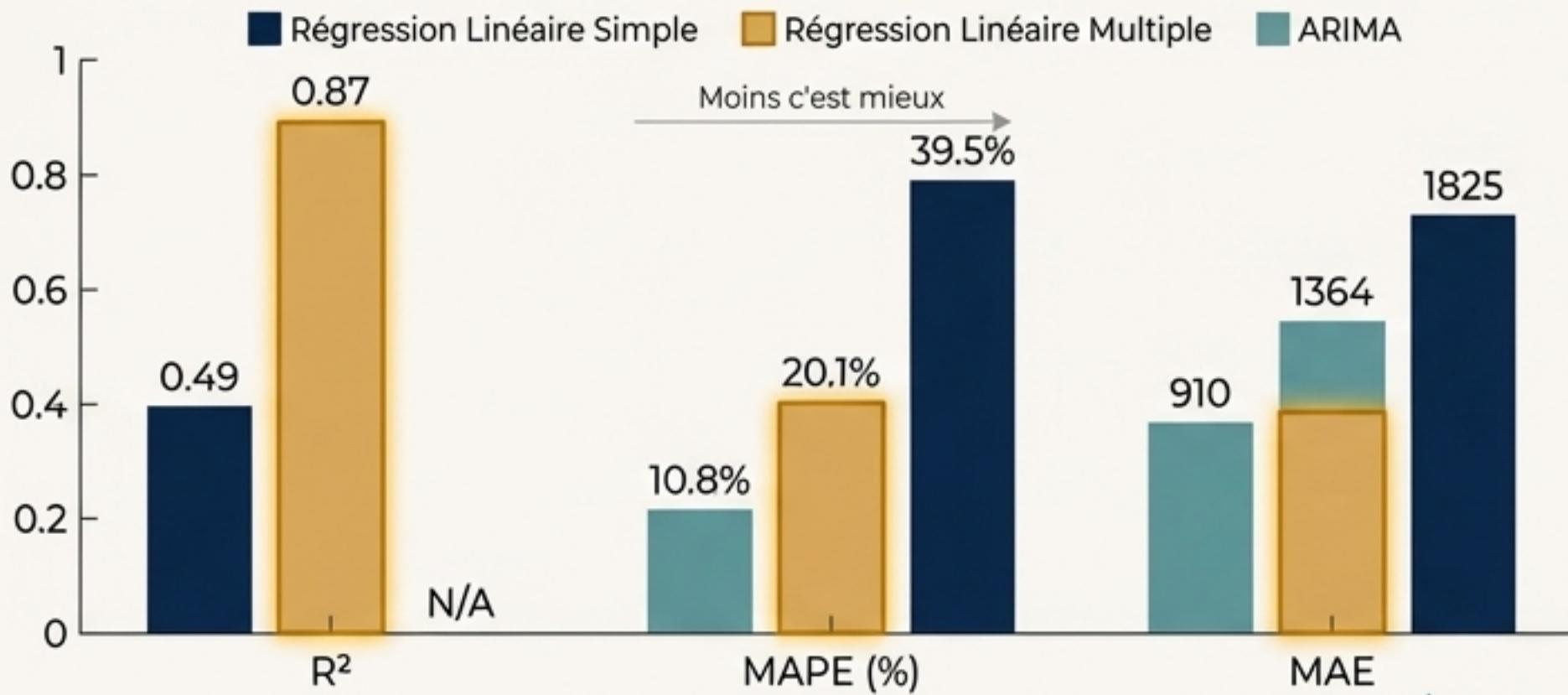
L'Approche Analytique

Évaluation de trois modèles de régression pour identifier celui offrant le meilleur équilibre entre précision (MAPE, MAE) et capacité à expliquer les variations de salaire (R^2), un critère clé pour la transparence RH.



La Régression Linéaire Multiple s'impose par sa Performance Globale

Performance des Modèles et Analyse



Analyse Clé

Si l'ARIMA offre le plus faible taux d'erreur (MAPE), son R^2 non applicable et sa nature 'boîte noire' limitent son utilité pour expliquer pourquoi un salaire est ce qu'il est. La Régression Linéaire Multiple offre le meilleur compromis avec une excellente capacité explicative ($R^2 = 0.87$) et une erreur prédictive deux fois plus faible que le modèle simple.

Tableau Comparatif des Résultats

Algorithme	R^2	MAPE	MAE	Temps (s)	Verdict
Régression Linéaire Simple	0.49	39.5%	1825	0.02	Rejeté
Régression Linéaire Multiple	0.87	20.1%	910	0.15	CHOISI
ARIMA	N/A	10.8%	1364	0.25	Analytique

Comprendre les Profils des Collaborateurs au-delà des Structures Formelles

Le Défi Stratégique

Les organigrammes ne disent pas tout. Pour personnaliser la gestion des talents (formation, carrière, engagement), nous devons identifier des groupes de collaborateurs qui se ressemblent réellement en termes de comportements, de compétences et d'attributs.

L'Approche Analytique

Comparaison d'algorithmes de clustering sur leur capacité à former des groupes (1) bien séparés les uns des autres (Score de Silhouette), (2) internement cohérents (Inertie) et (3) facilement interprétables par les managers RH.



K-Means pour la meilleure séparation, CAH pour une interprétabilité maximale

Algorithme	Silhouette	Inertie	Stabilité	Interprétabilité	Verdict
K-Means	0.5842	4200	Élevée	Moyenne	<input checked="" type="checkbox"/> CHOISI
CAH	0.5410	4350	Élevée	Excellente	<input checked="" type="checkbox"/> Validation
DBSCAN	0.5100	N/A	Faible	Faible	<input type="checkbox"/> Rejeté

CHOISI : K-Means



Force Principale

Score de Silhouette le plus élevé (0.5842), indiquant des clusters très distincts et bien définis.



Cas d'Usage Idéal

Parfait pour une segmentation opérationnelle rapide, la personnalisation de masse et l'identification de profils types à grande échelle.

VALIDATION : Classification Ascendante Hiérarchique (CAH)



Force Principale

Interprétabilité excellente. Le dendrogramme visuel permet de comprendre et d'expliquer la logique de la segmentation aux équipes métier.



Cas d'Usage Idéal

Essentiel pour l'analyse exploratoire et pour valider la pertinence business des segments identifiés.

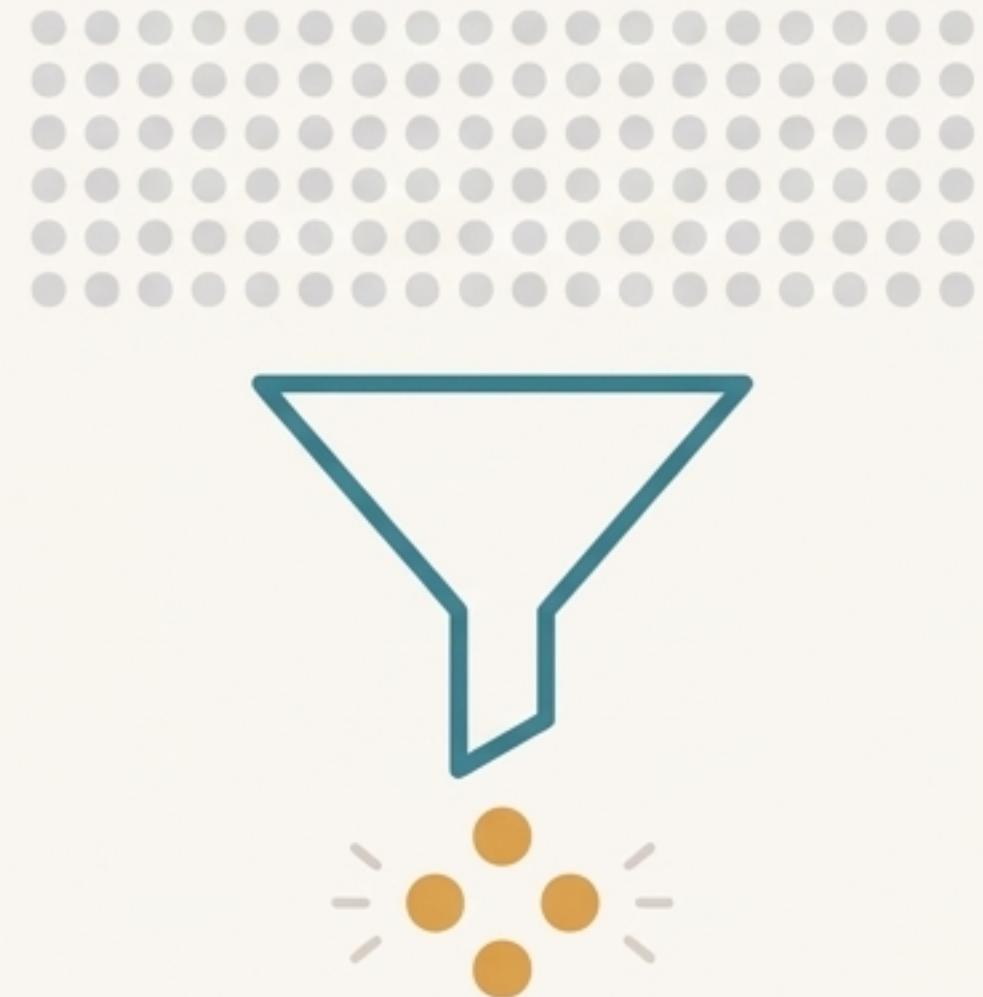
Anticiper les Départs pour Mieux Retenir les Talents

Le Défi Stratégique

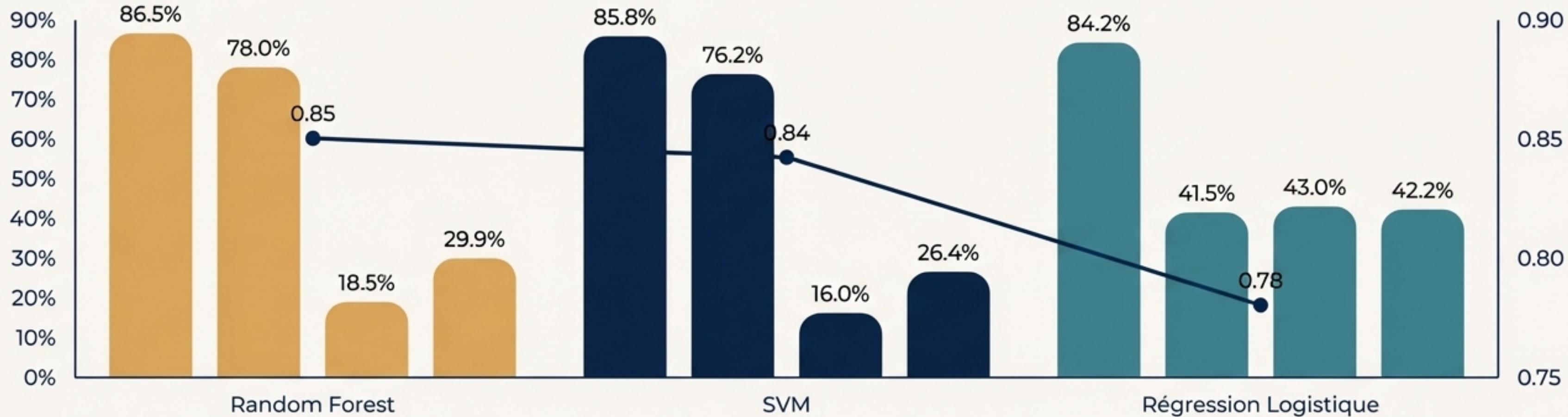
Le départ d'un talent coûte cher et perturbe les équipes. L'enjeu n'est pas seulement de savoir qui pourrait partir, mais de le faire avec suffisamment de confiance pour engager des actions de rétention (primes, entretiens) sans les gaspiller sur des collaborateurs fidèles.

L'Approche Analytique

Évaluation de modèles de classification. Le critère clé est le compromis entre une bonne Précision (ne pas cibler à tort) et un bon Rappel (ne pas manquer les vrais départs), synthétisé par des métriques globales comme l'AUC.



Random Forest offre le meilleur pouvoir prédictif global



Analyse Clé

La Régression Logistique identifie plus de départs potentiels (Rappel de 43%), mais se trompe souvent (Précision de 41.5%). Le Random Forest est plus conservateur (Rappel de 18.5%) mais bien plus fiable : quand il signale un risque, il a raison dans 78% des cas. Son AUC de 0.85 confirme sa supériorité globale pour distinguer les deux populations.

Le Verdict : un choix stratégique pour optimiser l'impact des actions de rétention

CHOISI : Random Forest

La Raison Stratégique

Dans un contexte de ressources limitées, il est plus coûteux d'engager des actions de rétention sur 100 personnes pour en sauver 43 (logique du Rappel élevé) que de cibler 20 personnes avec une forte certitude d'en sauver 18 (logique de la Précision élevée).

Avantage Clé

La Précision de 78% garantit que les efforts des managers et les investissements financiers sont dirigés vers les collaborateurs les plus véritablement à risque, maximisant le retour sur investissement des programmes de rétention.

AUTRES MODÈLES

SVM :

Une alternative très solide avec des performances quasi-identiques.

Régression Logistique :
Pertinent uniquement si la stratégie est de 'n'oublier personne', quitte à sur-solliciter les équipes.

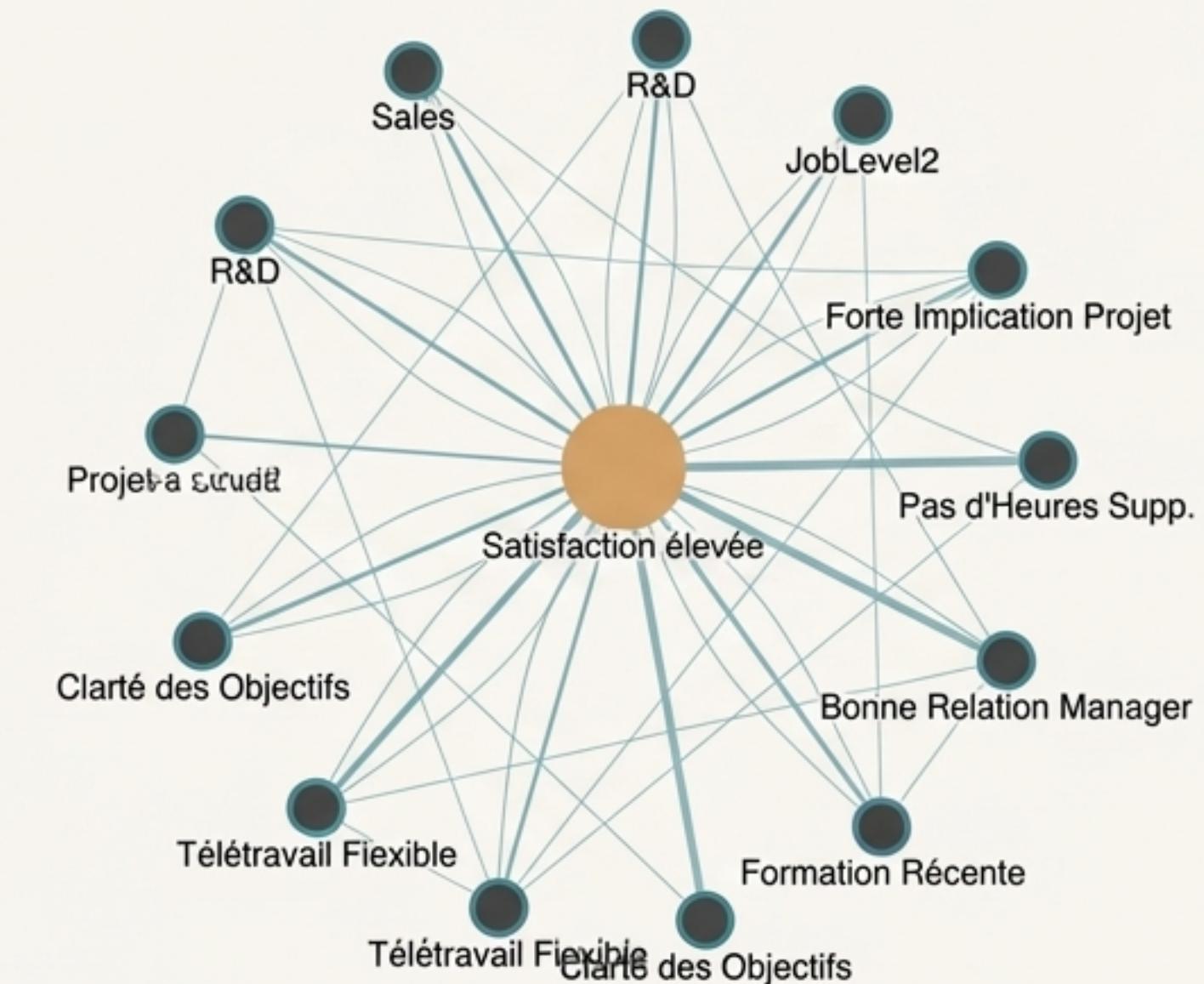
Découvrir les Combinations de Facteurs qui Mènent à la Satisfaction

Le Défi Stratégique

Les enquêtes de satisfaction nous disent 'quoi' mais rarement 'pourquoi'. Comment identifier les combinaisons de facteurs qui, ensemble, créent un environnement où les collaborateurs s'épanouissent ?

L'Approche Analytique

Utilisation des règles d'association pour trouver des 'patterns' cachés dans les données. Nous ne cherchons pas un modèle prédictif, mais des règles du type 'Si un employé a les caractéristiques A et B, alors il est souvent satisfait'. La métrique clé est le 'Lift', qui mesure à quel point la co-occurrence est plus fréquente que le hasard.



Trois Levier Clés de la Satisfaction

Révélés par les Données

Le "Sweet Spot" Commercial

$\{Sales\} + \{JobLevel2\} \rightarrow \{Satisfaction élevée\}$

Confidence: 72%, **Lift:** 1.55

Interprétation Business

Les collaborateurs du département Sales au niveau de seniorité 2 sont 1.55 fois plus susceptibles d'être satisfaits que la moyenne. Ce groupe est un pilier de performance et d'engagement qu'il faut absolument choyer et développer.

L'Engagement par l'Innovation

$\{R&D\} + \{Forte Implication Projet\} \rightarrow \{Satisfaction élevée\}$

Confidence: 68%, **Lift:** 1.39

Interprétation Business

Pour la population R&D, l'implication forte dans leurs projets est un moteur de satisfaction majeur. L'autonomie et la valorisation de leur impact sont des leviers essentiels.

Les Fondamentaux du Bien-être

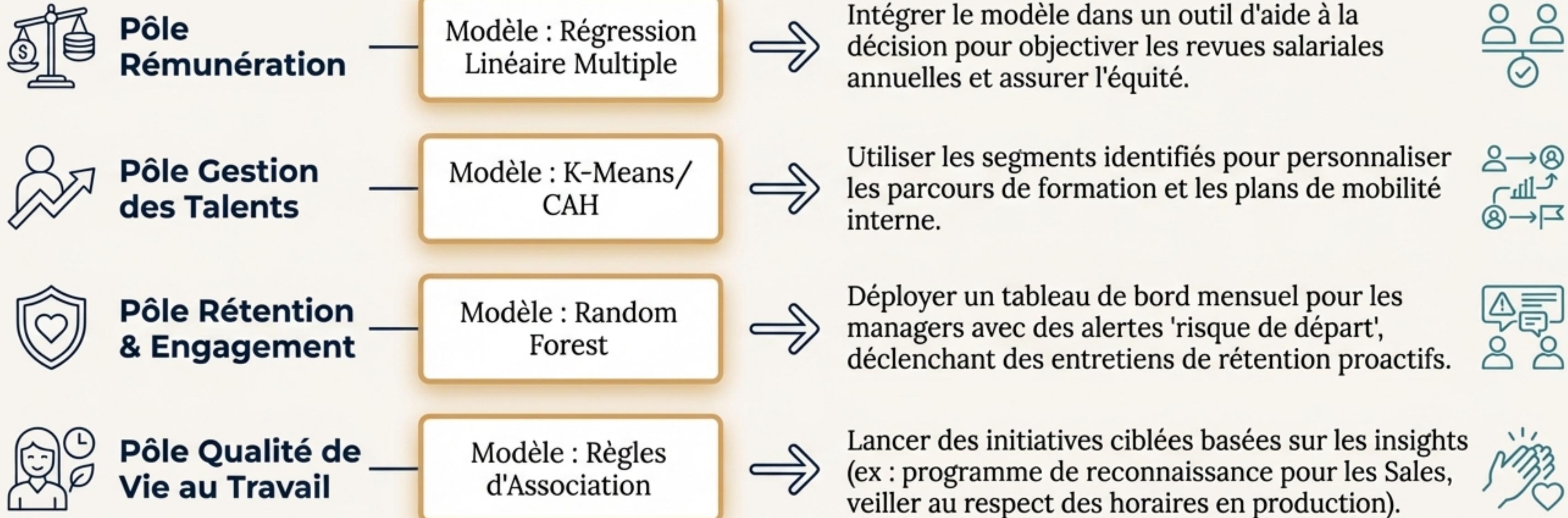
$\{Pas d'Heures Supp.\} + \{Bon Env. de Travail\} \rightarrow$

Support: 15.3%, **Confidence:** 75%,
Lift: 1.25

Interprétation Business

C'est la règle la plus répandue. Un équilibre vie pro/vie perso respecté et un environnement de travail de qualité forment le socle de la satisfaction pour une large partie des collaborateurs.

De l'Analyse à l'Action : Notre Feuille de Route pour une RH "Data-Driven"



Ce benchmark nous équipe d'outils analytiques robustes pour rendre nos décisions RH plus intelligentes, plus justes et plus proactives.