

# Analyse des causes de démission chez TechNova Partners

Une analyse approfondie des données RH pour décrypter les facteurs clés du turnover et élaborer des stratégies de rétention efficaces.

Abdourahamane LY | lyabdourahamane66@gmail.com



# Sommaire



## Contexte et Problématique

Définir le cadre de l'étude et les questions clés sur le turnover.



## Sources de Données et Méthodologie

Présenter les données utilisées et l'approche analytique choisie.



## Analyse Exploratoire et Insights Clés

Explorer les corrélations et différences significatives dans les données.



## Prétraitement des Données

Détailler les étapes de préparation des données pour la modélisation.



## Comparaison et Sélection des Modèles

Évaluer les performances des modèles et choisir le plus adapté.



## Interprétation Approfondie du Modèle Final

Comprendre les facteurs déterminants de la démission via l'analyse du modèle.



## Recommandations RH Actionnables

Proposer des stratégies concrètes basées sur les résultats de l'étude.

# Contexte et problématique

## TechNova Partners

ESN spécialisée en transformation digitale, comptant plus de 1400 collaborateurs répartis dans divers départements.

## Problématique RH

Un taux de démission préoccupant (16 %) impacte la continuité des projets et génère des coûts de recrutement élevés.

## Objectif du projet

Identifier les causes racines des démissions et proposer des actions correctives ciblées, basées sur une approche data-driven.

Notre approche : analyser les facteurs de risque, développer un modèle prédictif explicite et formuler des recommandations concrètes pour la Direction des Ressources Humaines.

# Sources de données et méthodologie

## Sources intégrées

### SIRH

Informations complètes sur les 1470 employés (données démographiques, salariales, postes, départements, ancienneté).

### Évaluations

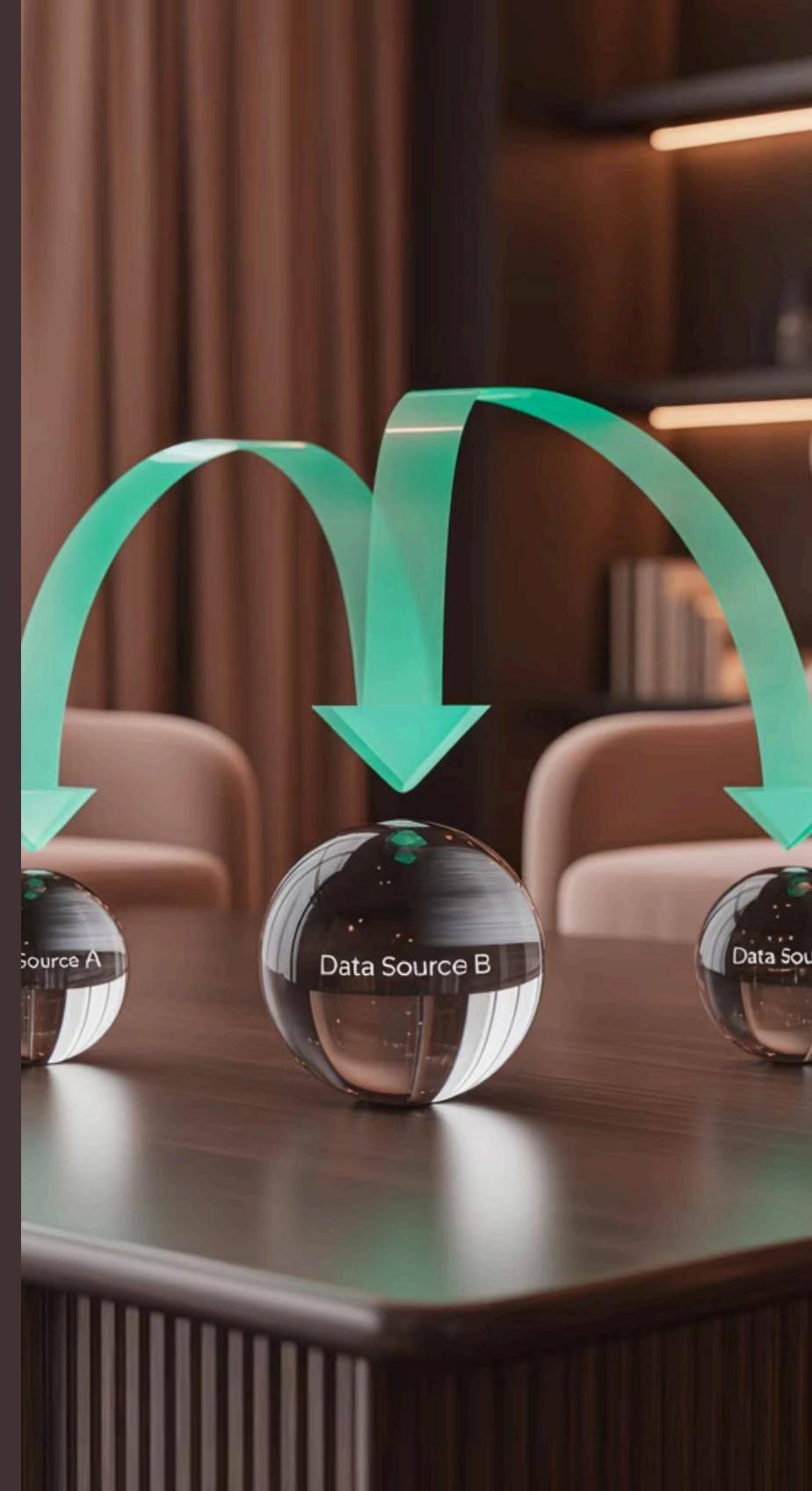
Données des évaluations (notes de performance, satisfaction, heures supplémentaires) liées à chaque employé.

### Sondage interne

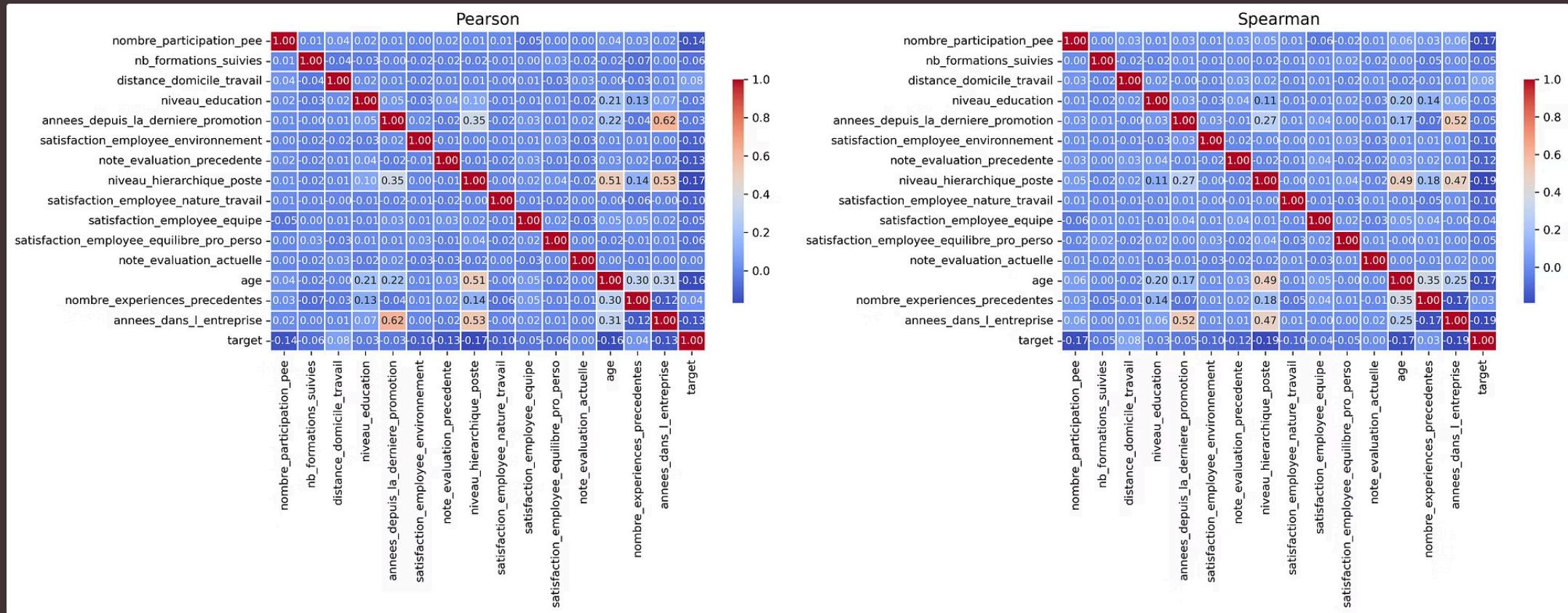
Résultats du sondage sur le bien-être, la formation, les déplacements et l'équilibre vie professionnelle/personnelle.

## Préparation des données

- Jointure des trois sources via les identifiants uniques (id\_employee, eval\_num, code\_sondage).
- La jointure interne a généré un dataset final de 1470 observations et 23 variables.
- La variable cible est "a\_quitte\_l\_entreprise", avec 16 % de démissions.

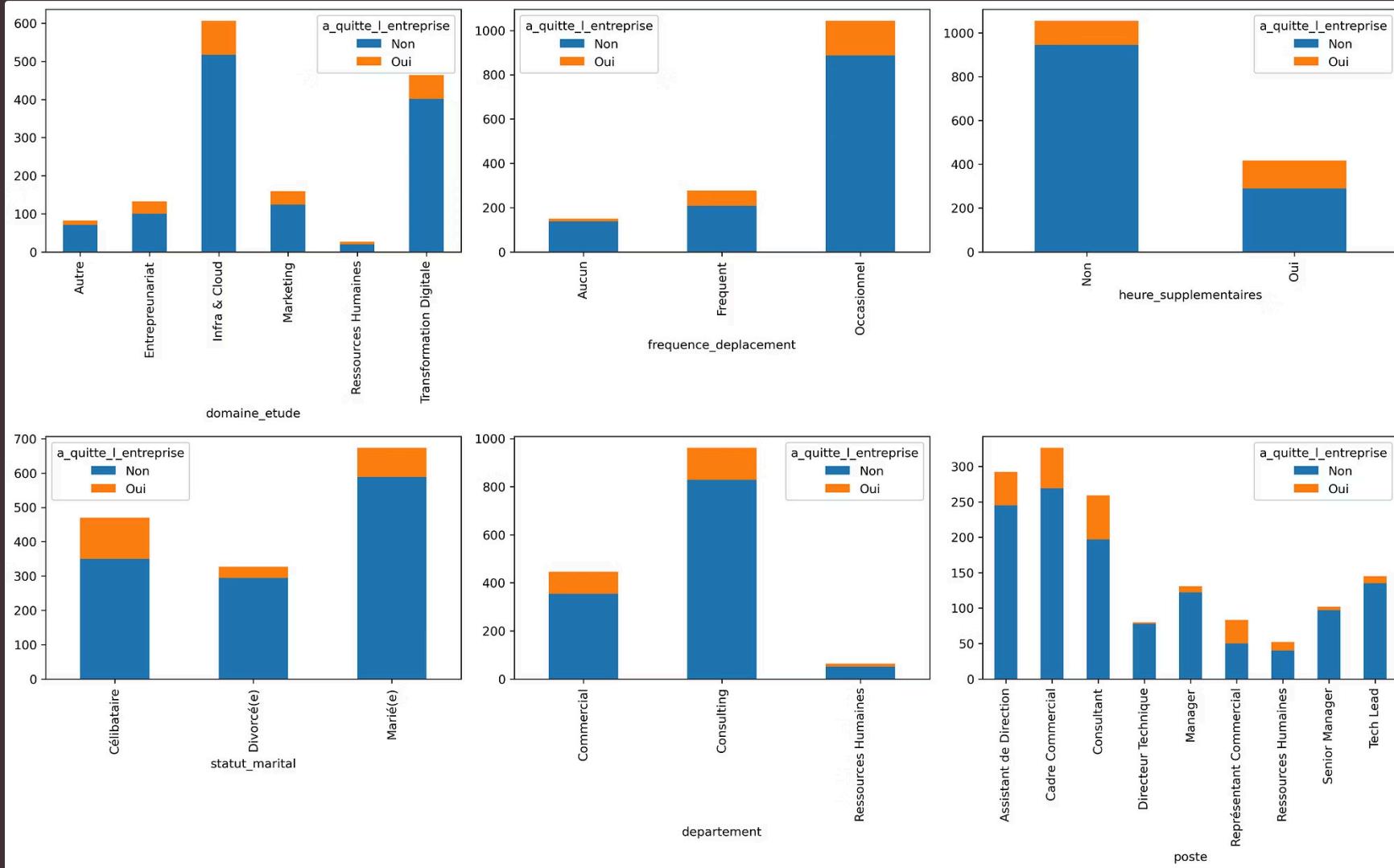


# Matrice de Corrélation des Variables Numériques



L'analyse de la matrice de corrélation révèle que nos variables numériques présentent une faible corrélation avec la variable cible (départ de l'entreprise). Cela indique que leur impact prédictif sur les démissions est susceptible d'être limité.

# Comparaison des Variables Catégorielles avec le Départ de l'Entreprise



L'analyse révèle que les variables catégorielles joueront un rôle prépondérant dans la compréhension et la prédiction des démissions. Des différences significatives sont observées entre les profils des employés qui ont quitté l'entreprise et ceux qui sont restés, à l'exception de certaines catégories comme les Directeurs Techniques, les Senior Managers et les Ressources Humaines.

# Analyse des associations et différences significatives

## Test du Chi-carré

Ce tableau présente les résultats du test du Chi-carré appliqué aux variables catégorielles afin d'évaluer leur association avec la variable cible « a\_quitté\_l'entreprise ».

VARIABLES	$\chi^2$ ( Chi2 )	ddl	p-value
domaine_etude	16.02	5	0.007
frequence_deplacement	24.18	2	0.000
heure_supplementaires	87.56	1	0.000
status_marital	46.16	2	0.000
departement	10.80	2	0.000
poste	86.19	8	0.000

## ANOVA

Ce tableau présente les résultats de l'analyse de variance évaluant les différences entre les moyennes des variables, regroupées selon la variable cible « a\_quitté\_l'entreprise ».

VARIABLES	F ( Statistic )	p-value
domaine_etude	3.23	0.007
frequence_deplacement	12.27	0.000
heure_supplementaires	94.66	0.000
status_marital	23.78	0.000
departement	5.43	0.004
poste	11.37	0.000

Toutes les variables étudiées montrent une liaison statistique significative avec la variable cible, ce qui indique qu'elles peuvent contribuer à la prédiction du départ de l'entreprise. Cependant, nous avons également observé une forte dépendance entre certaines variables (par exemple, « poste » et « département », « poste » et « domaine\_etude », « département » et « domaine\_etude », « statut\_marital » et « poste »). Par conséquent, les variables « poste » et « domaine\_etude » seront exclues de notre modélisation afin d'éviter la multicolinéarité et de ne retenir que les informations non redondantes.

# Prétraitement des Données

## Séparation X/y

Nous commençons par séparer les variables explicatives (X) de la variable cible (y).



## Encodage Catégories

Les variables catégorielles sont ensuite encodées numériquement avec Label Encoding.

## Mise à l'Échelle

Les variables numériques sont mises à l'échelle avec MinMaxScaler.

## Assemblage ColumnTransformer

Les variables transformées sont assemblées avec ColumnTransformer.

## Split Train/Test

Enfin, les données sont divisées en ensembles d'entraînement et de test avec une stratification sur la cible.

# Performances Comparées des Modèles

Models	Recall		F1-Score		Precision		AUC-ROC	Balanced Accuracy
	classe 0	classe 1	classe 0	classe 1	classe 0	classe 1		
Dummy	0.000	<b>1.000</b>	0.000	0.276	0.000	0.160	0.500	0.500
Logistic Regression	<b>0.923</b>	0.447	<b>0.910</b>	0.483	0.898	<b>0.525</b>	0.794	0.685
Gradient Boosting	0.802	0.766	0.868	0.545	<b>0.947</b>	0.424	<b>0.810</b>	<b>0.784</b>
Balanced Random Forest	0.887	0.638	0.907	<b>0.571</b>	0.928	0.517	0.808	0.763

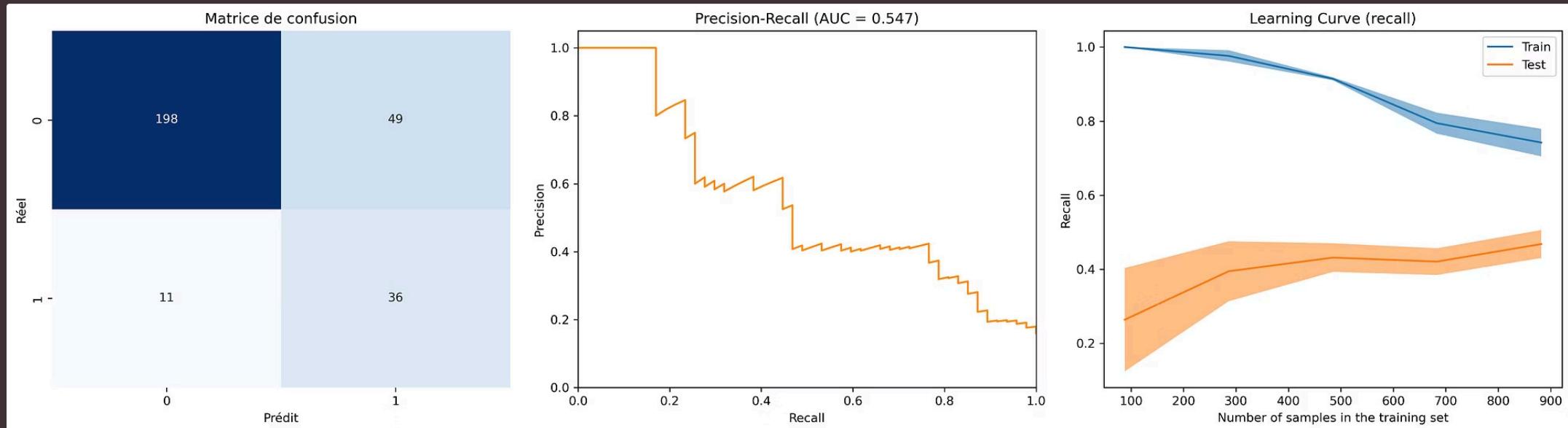
Les performances des modèles sont évaluées en tenant compte de l'importance des métriques, en particulier pour la classe 1 (employés démissionnaires). Étant donné le déséquilibre des classes (247 observations pour la classe 0 "Reste" et 47 pour la classe 1 "Démissionne"), un oversampling via SMOTETomek (Imblearn) a été utilisé pour rééquilibrer les données lors de l'entraînement.

# Choix du Modèle Final (En fonction de la classe 1)

MODELS	Recall × 0.40	F1-Score × 0.25	Precision × 0.15	AUC-ROC × 0.10	Balanced Accuracy × 0.10	Score Composite
<b>Logistic Regression</b>	0.447	0.483	0.525	0.794	0.685	0.526
<b>Gradient Boosting</b>	0.766	0.545	0.424	0.810	0.784	<b>0.666</b>
<b>Balanced Random Forest</b>	0.638	0.571	0.517	0.808	0.763	0.633

Le modèle de Gradient Boosting est sélectionné comme choix optimal pour prédire les départs des employés, en raison de son meilleur compromis entre un rappel élevé et un bon score F1.

# Résultats du Modèle Gradient Boosting

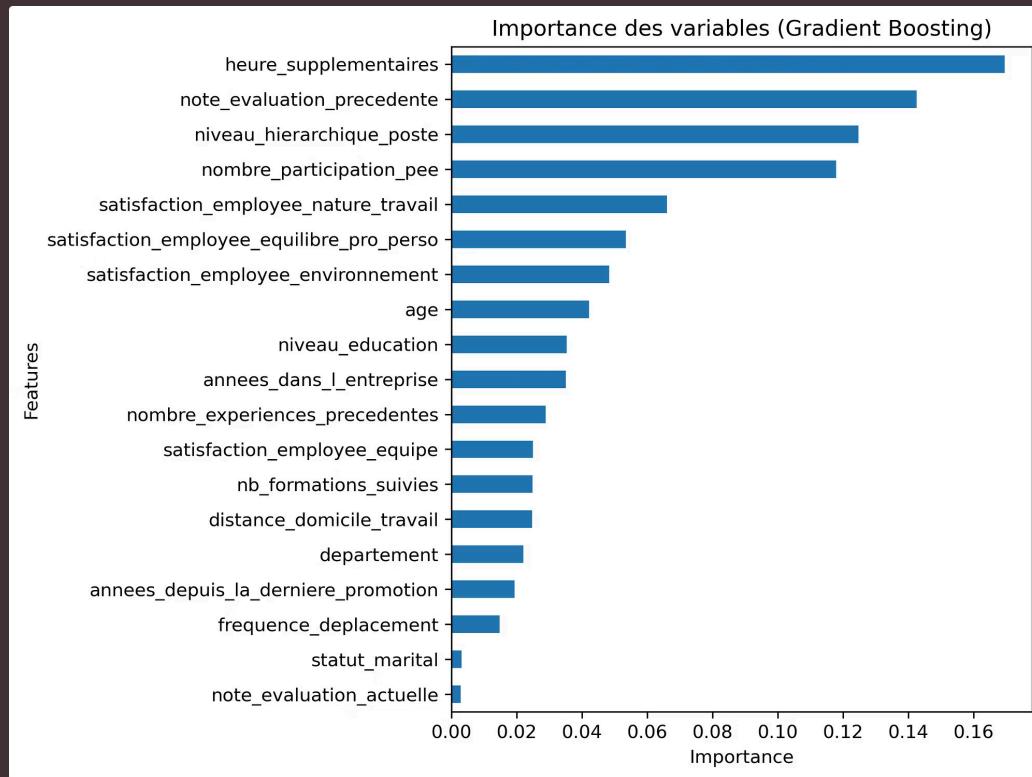


Le modèle détecte 76 % des démissionnaires. La précision est plus faible (42 %), donc quelques fausses alertes, mais elles restent acceptables car elles permettent d'anticiper les départs. Globalement, le modèle atteint un **bon équilibre entre rappel et précision** (AUC ROC = 0.81).

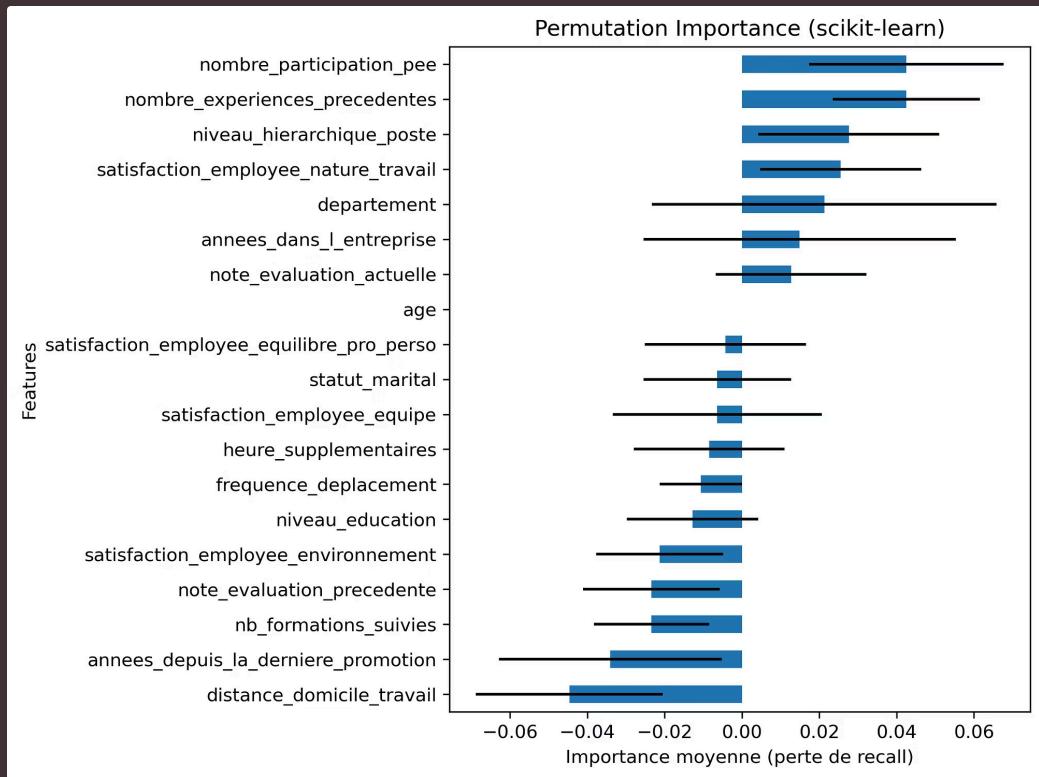
La courbe d'apprentissage montrent un **écart entre entraînement et test**, signe de sur-apprentissage. Mais cet écart diminue nettement quand on augmente les données. Cela indique que **plus de données permettraient d'améliorer la généralisation** du modèle.

# Interprétation du modèle – Importances globales

Les variables **heures supplémentaires, notes d' évaluation précédentes et niveau hiérarchique** apparaissent comme les plus influentes. Cela reflète surtout la structure interne du modèle, qui peut amplifier des corrélations entre variables.

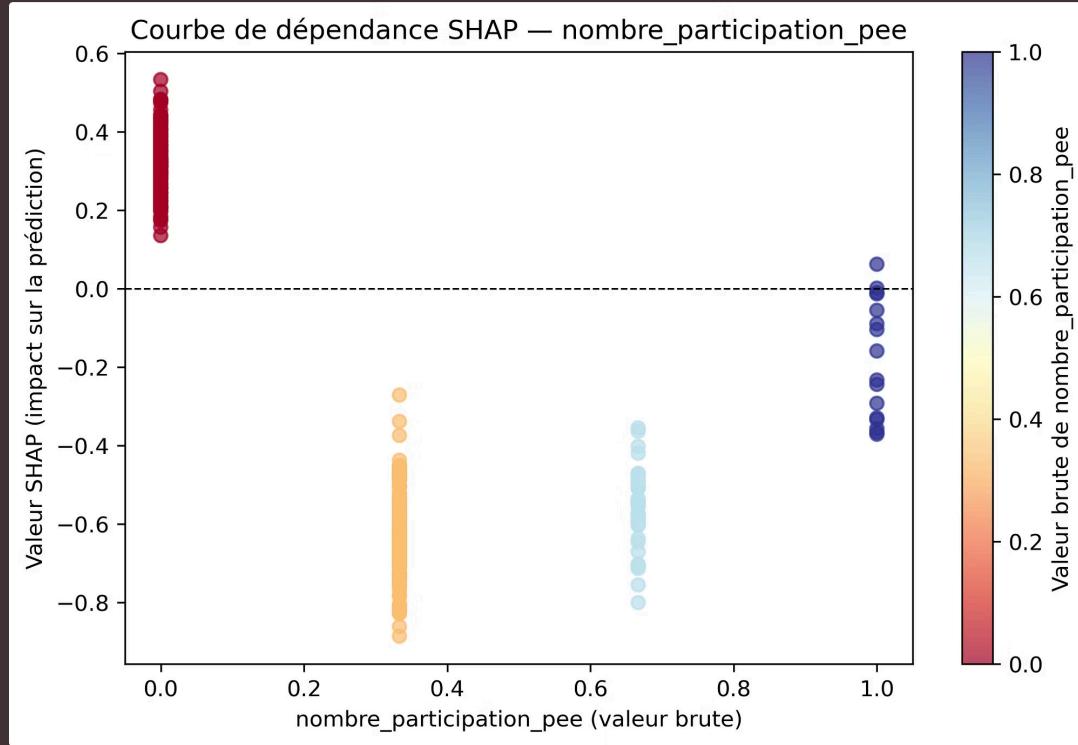


Ici, c' est la **participation au PEE, l' expérience professionnelle antérieure et le niveau hiérarchique** qui ressortent comme plus déterminants. Cette approche montre quelles variables affectent réellement le **recall** quand elles sont perturbées.



La comparaison révèle que certaines variables surévaluées par l' importance native (ex. : heures supplémentaires) perdent leur poids en permutation. À l' inverse, d' autres (ex. : participation au PEE) émergent comme critiques pour la détection des démissions. Les **variables comportementales et d' engagement** comptent davantage que les données démographiques.

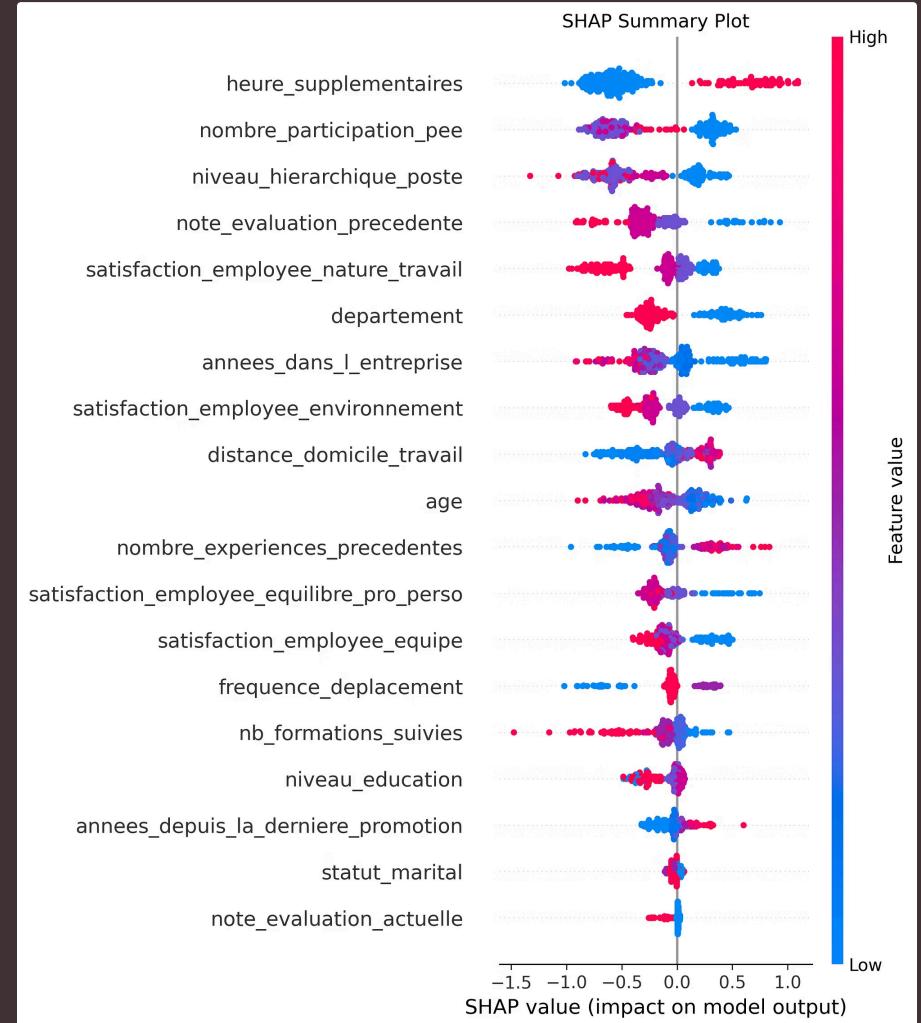
# Interprétation du modèle – Shap



Interprétation (figure gauche) :

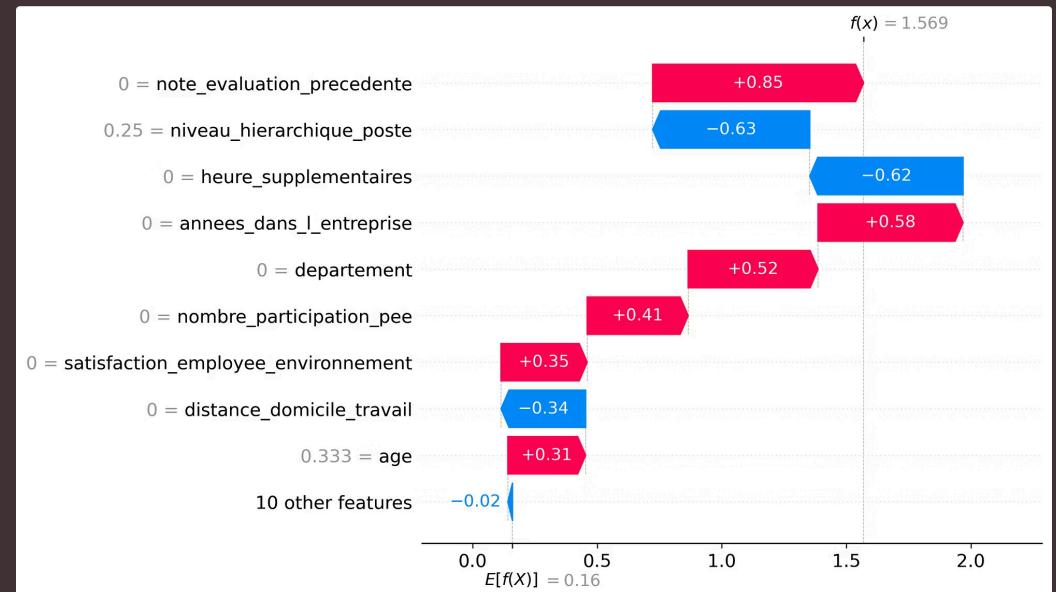
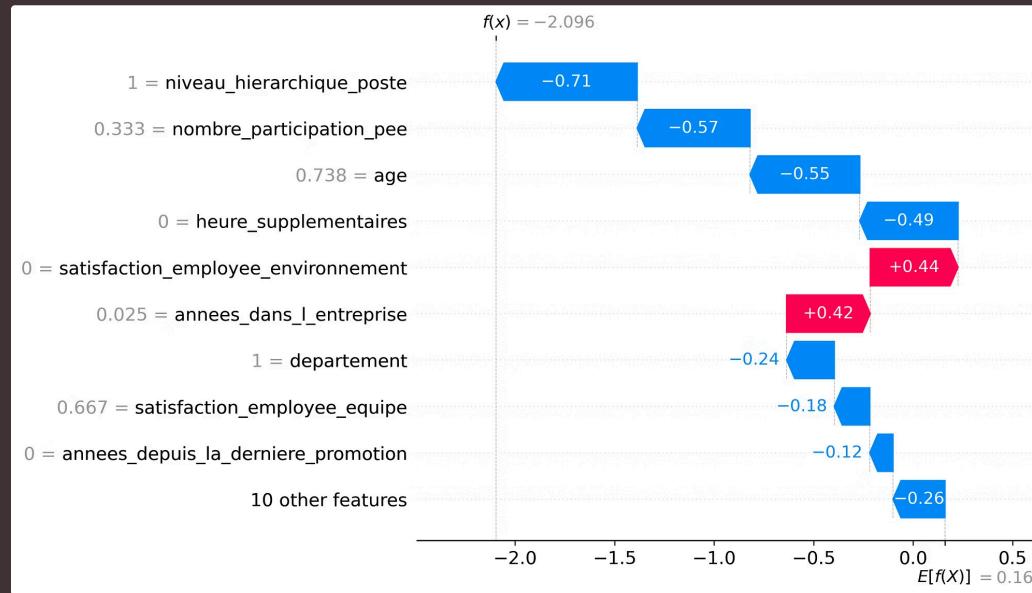
Les employés qui ne participent pas au PEE affichent un **risque accru de démission**.  
Inversement, une participation élevée au PEE favorise la **fidélisation**.

Interprétation (figure droite) : Les variables clés sont principalement liées au **travail** et à l'**engagement** (heures supplémentaires, participation au PEE, satisfaction). Les données démographiques ont un impact moindre.



# Interprétation du modèle – Comparaison des importances locales

Pour cet employé, les facteurs de stabilité (hiérarchie, participation, âge, heures sup) l'emportent clairement sur les signaux de risque, ce qui explique la prédiction “reste” .



Pour cet employé, les facteurs favorisant la mobilité (évaluation, ancienneté, département, participation, satisfaction, âge) l’ emportent nettement sur les signaux de rétention, ce qui explique la prédiction “démissionnaire”

# Recommandations RH actionnables



## Réduire les facteurs de risque comportementaux

- **Participation PEE:** Une communication ciblée et des incitations peuvent encourager l'adhésion au PEE, dont l'absence est un facteur de risque élevé de départ.
- **Satisfaction au travail:** Lancer des enquêtes de satisfaction rapides et trimestrielles pour ajuster les conditions de travail et le contenu des missions.



## Optimiser la gestion des parcours professionnels

- **Niveau hiérarchique & expériences passées:** Mettre en place un suivi personnalisé des talents à haut risque, en s'appuyant sur la détection précoce du modèle.
- Proposer des passerelles internes et des missions transversales pour maintenir l'engagement.



## Gérer la charge et l'organisation du travail

- Ajuster la politique des heures supplémentaires pour prévenir les charges excessives grâce à un suivi managérial mensuel.
- Former les managers à identifier et traiter les signaux d'alerte comportementaux.



## Pilotage continu grâce au modèle prédictif

- Utiliser le modèle de Gradient Boosting comme outil de veille RH pour un scoring mensuel des risques de départ.
- Mettre à jour le modèle avec les nouvelles données pour affiner les prédictions.

**Notre modèle devient un outil opérationnel de rétention des talents**

## Merci de votre attention !

N'hésitez pas si vous avez des questions.

Abdourahamane LY | lyabdourahamane66@gmail.com

