

**Optimisation de la Rétention des
Talents par le Machine Learning**

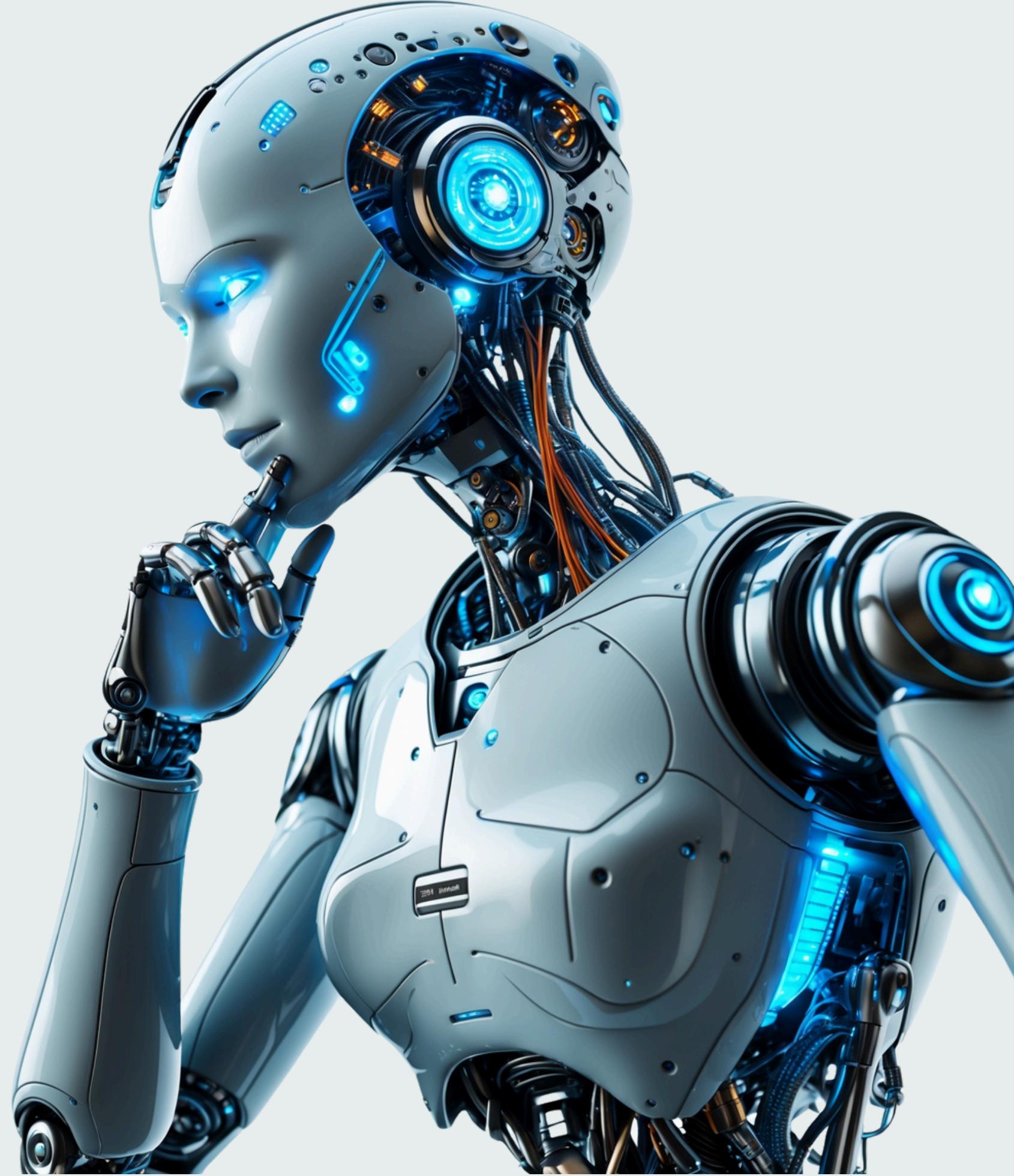
Analyse prédictive du Churn Employé - Approche Data-Driven

PRÉSENTÉ PAR:

MICKAEL ANDRIEU
CHRISTSYLVIA HODONOU

PROFESSEURS:

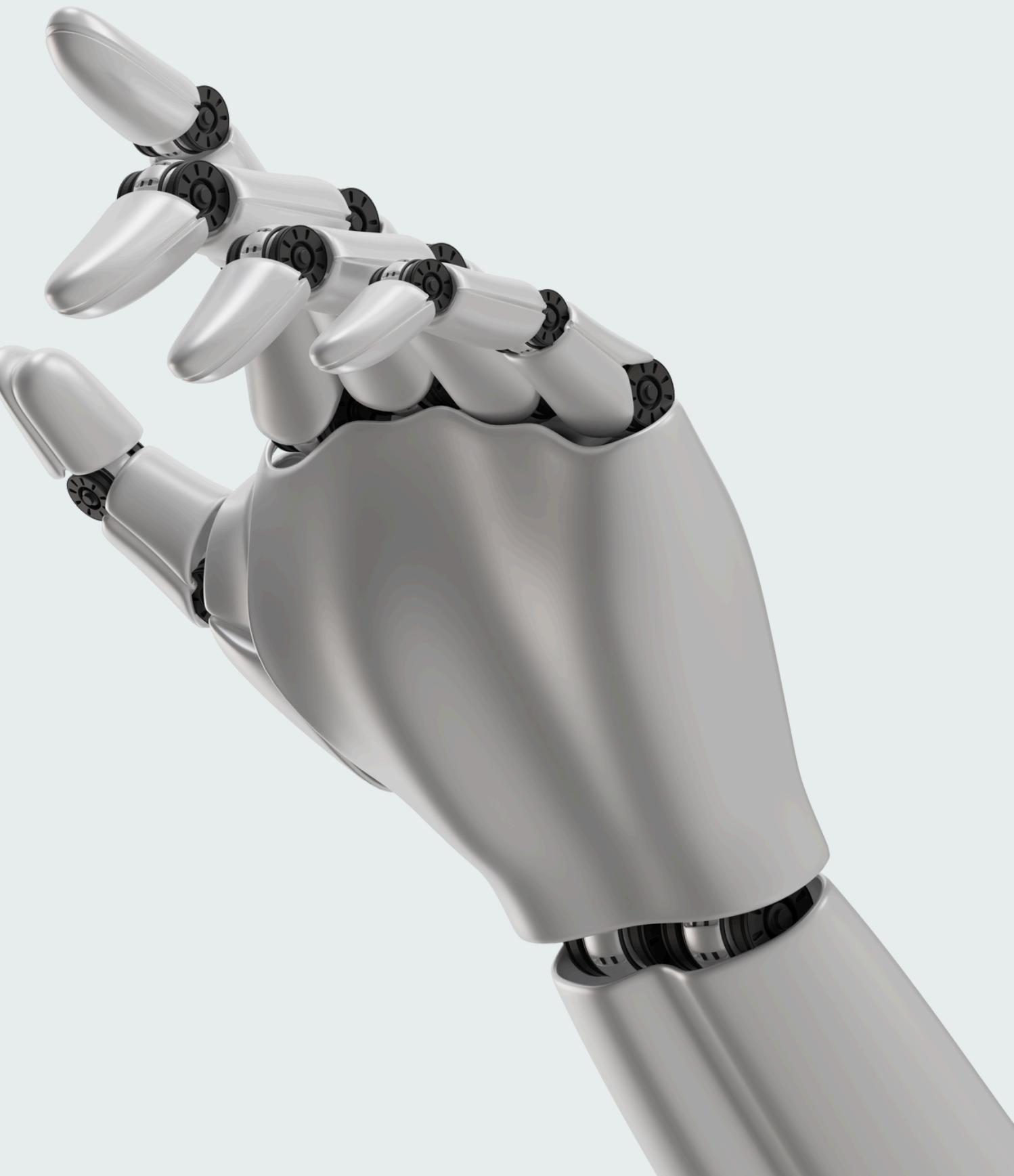
MANUEL LESAICHERRE
PIERRE MOUSSALLY



Prediction du churn employé

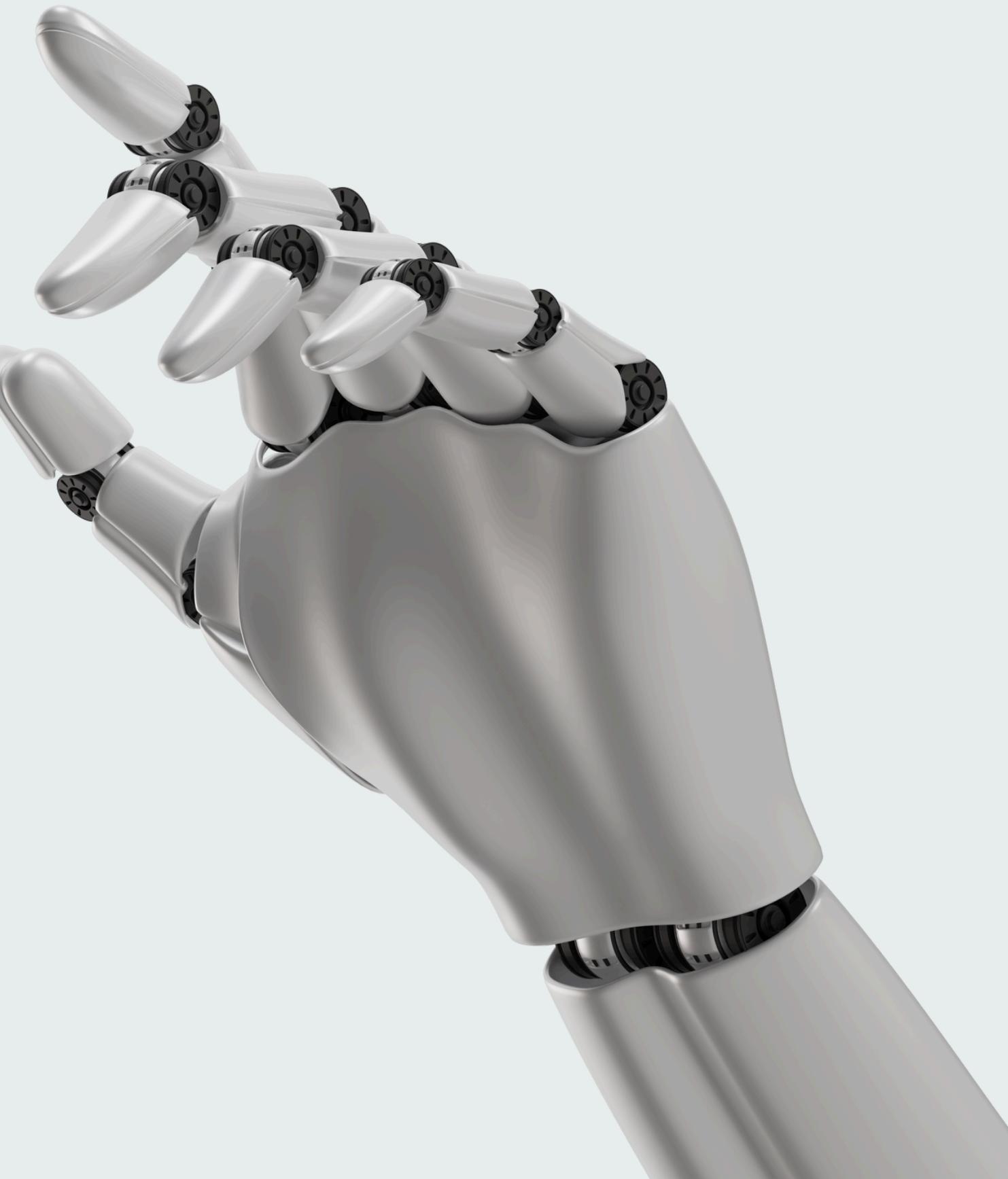
Score de risque + watchlist actionnable pour les RH

- Dataset: Employee Churn Data
- Environnement: Mac + VS Code, venv Python, notebooks Jupyter
- Suivi: GridSearchCV (recall) + MLflow



Sommaire

- 01 Analyse des Besoins Métiers (Le "Pourquoi")
- 02 Légal & Ethique (éventuellement si données réelles)
- 03 Présentation du Dataset
- 04 Analyse Exploratoire (EDA) - Insights
- 05 Concepts Algorithmiques retenus
- 06 Modélisation et Optimisation
- 07 Comparaison des Modèles
- 08 Analyse de la Feature Importance
- 09 Tableau du classement des employés les plus à risques
- 10 Deploiement & chiffrage (approche)



Analyse des Besoins Métiers (Le "Pourquoi")

Problématique:

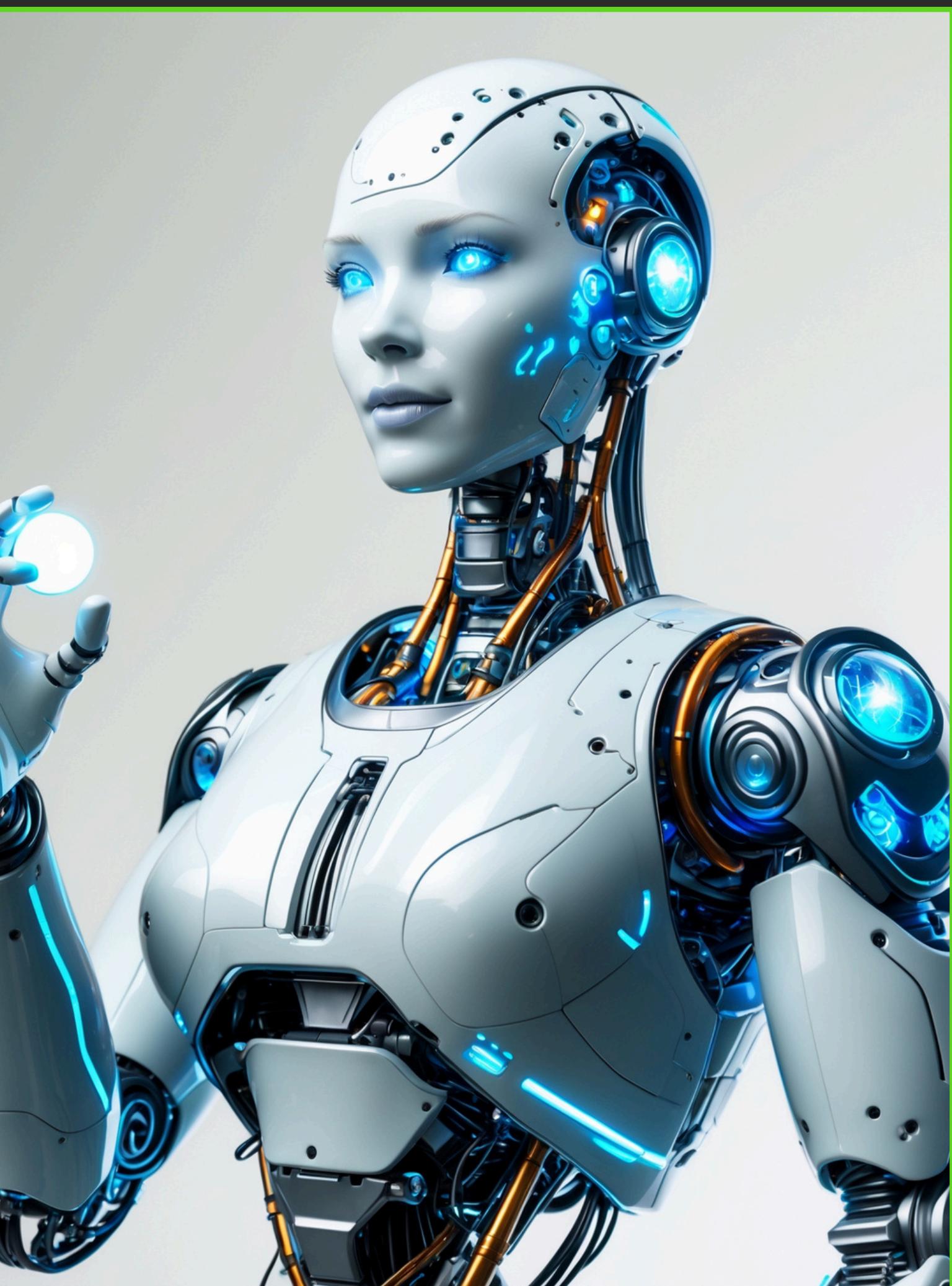
Le départ volontaire des collaborateurs coûte cher
(recrutement, formation, perte de savoir-faire).

Objectif :

Identifier les profils à risque avant leur départ
pour permettre des actions RH préventives.

Engagement :

Une solution fondée sur l'honnêteté scientifique,
la traçabilité et la reproductibilité.



Legal & ethique (si donnees réelles)

Bonnes pratiques pour un usage RH responsable

Points de vigilance

- Base legale & transparence (information des employés, finalité claire)
- Minimisation: supprimer les identifiants (Employee ID) des features
- Variables sensibles (ex: Genre): risque de biais -> audit équité + option d'exclusion
- Usage: aide à la décision (priorisation), pas une décision automatique
- Traçabilité: journalisation des versions, données, métriques (MLflow)

Note: ce projet utilise un dataset public/synthétique. En entreprise, un DPO/Legal doit valider la collecte et l'usage avant mise en production.



Présentation du Dataset

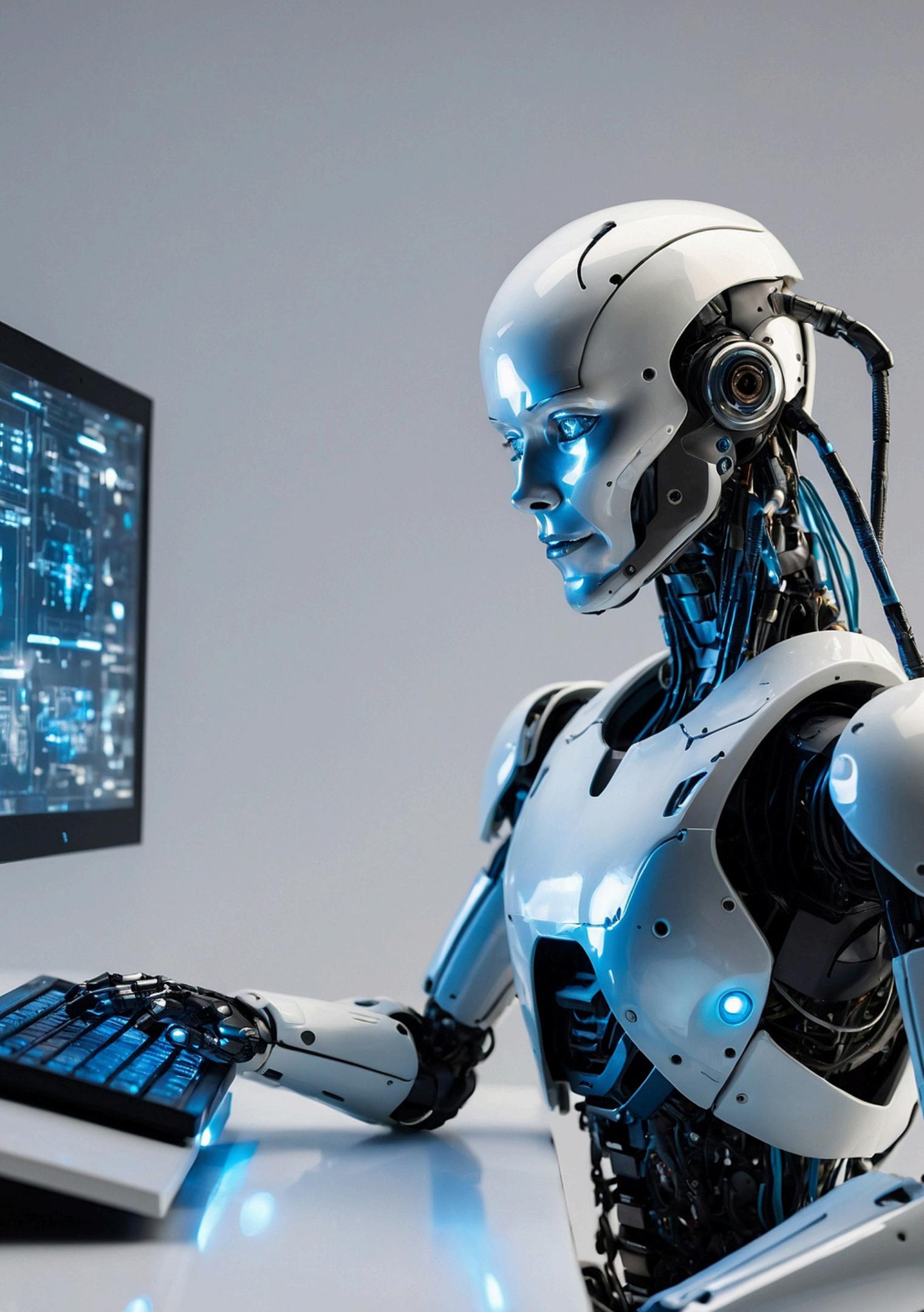
Source : Kaggle

Volume : 10 000 employés (lignes), 22 colonnes (variables)

Avant de réaliser l'analyse exploratoire, une première lecture du jeu de données ,basée sur une intuition métier RH, nous amène à penser que le Churn pourrait être influencé par :

- Satisfaction
- Heures supplémentaires
- Salaire
- Équilibre vie pro vs vie perso
- Ancienneté

Ces variables nous semblent être un bon point de départ pour analyser le churn, avant de confronter ces intuitions aux résultats de l'EDA.



Analyse Exploratoire (EDA) - Insights



Constat 1

Un niveau de satisfaction bas (< 0.4) est fortement corrélé au churn.



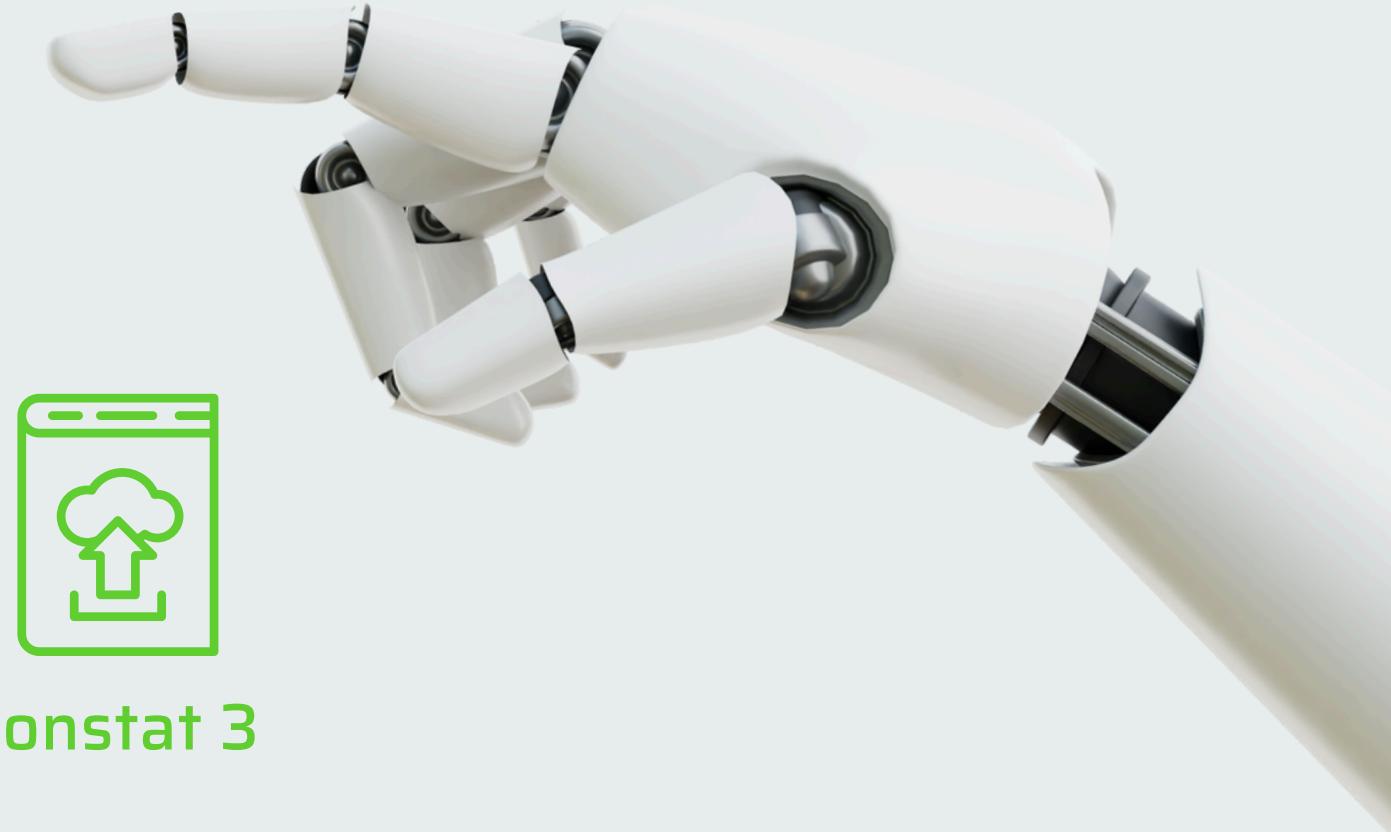
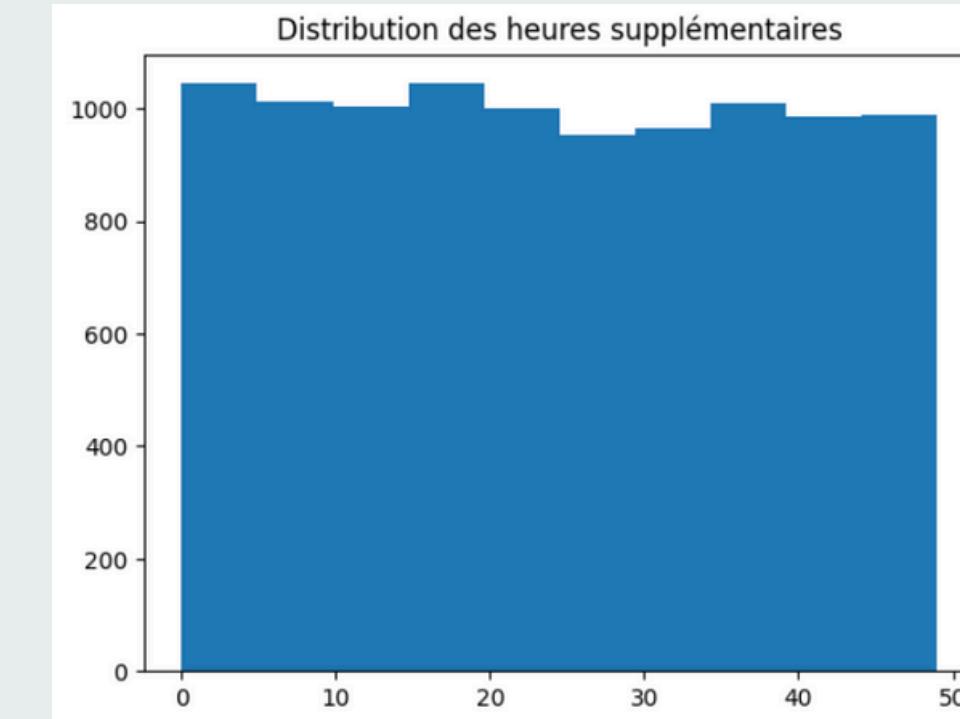
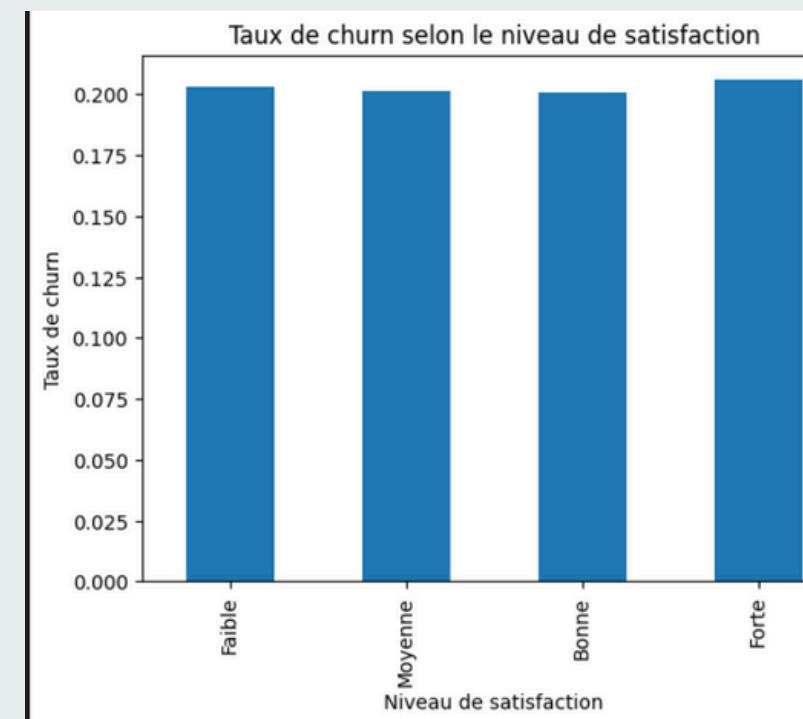
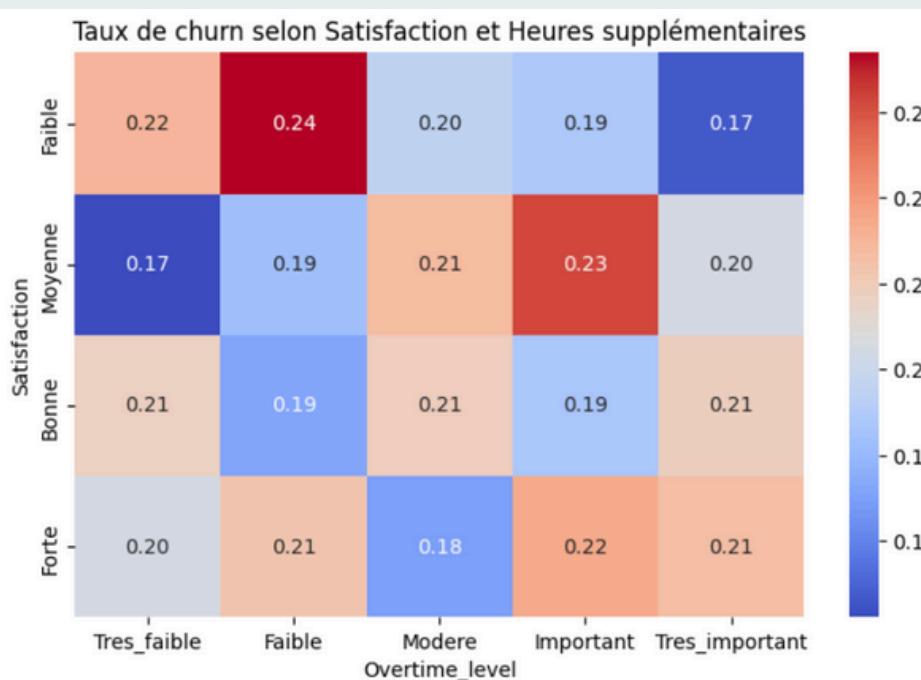
Constat 2

Le surmenage (heures mensuelles élevées) et le manque de promotion sur 5 ans sont des facteurs aggravants.



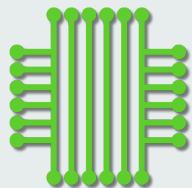
Constat 3

Déséquilibre : Les départs sont minoritaires (classe minoritaire), nécessitant une attention particulière lors de la modélisation.



Concepts Algorithmiques retenus

Modèles testés :



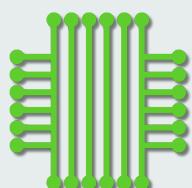
Dummy Classifier

Pour la simplicité et l'interprétabilité.



Random Forest

Pour gérer les relations non-linéaires et les interactions de variables.



XGBoost

Pour la performance pure via le gradient boosting.

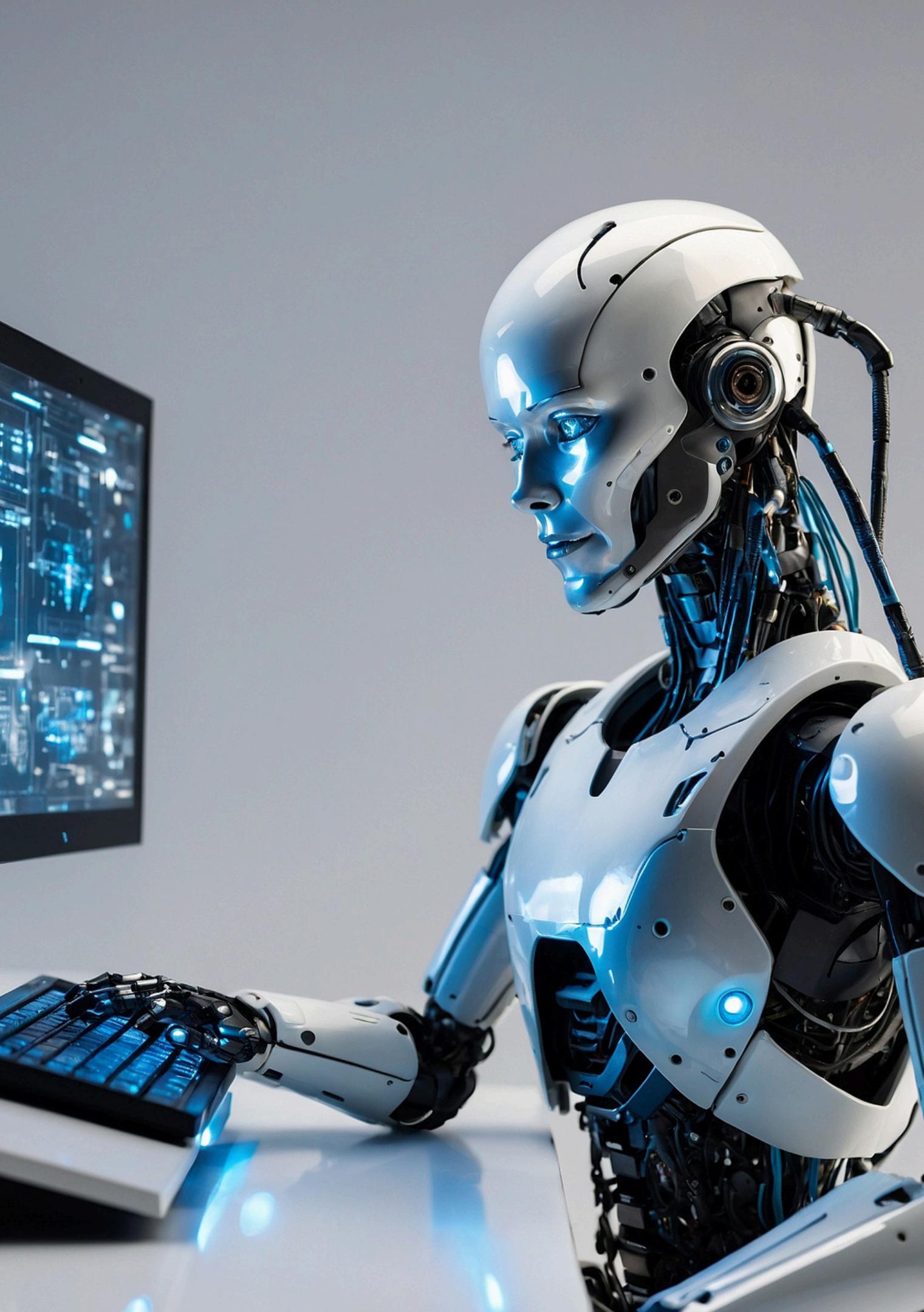
Innovation

Utilisation de modèles d'ensemble pour capturer la complexité des comportements RH.



Modélisation et Optimisation

- 01** **Pipeline Technique** : StandardScaler (numérique) +OneHotEncoder (catégoriel).
- 02** **Optimisation** : Recherche d'hyperparamètres via GridSearchCV (ex: n_estimators, max_depth).
- 03** **Tracking** : Utilisation de **MLFlow** pour consigner chaque run (paramètres, métriques) et garantir la traçabilité demandée.

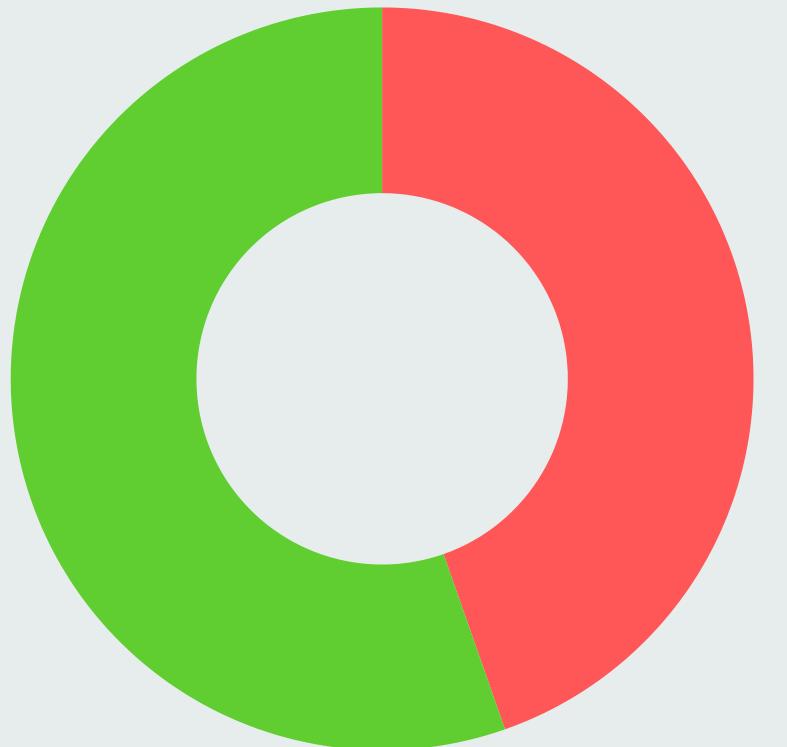


Comparaison des Modèles

Synthèse des Résultats

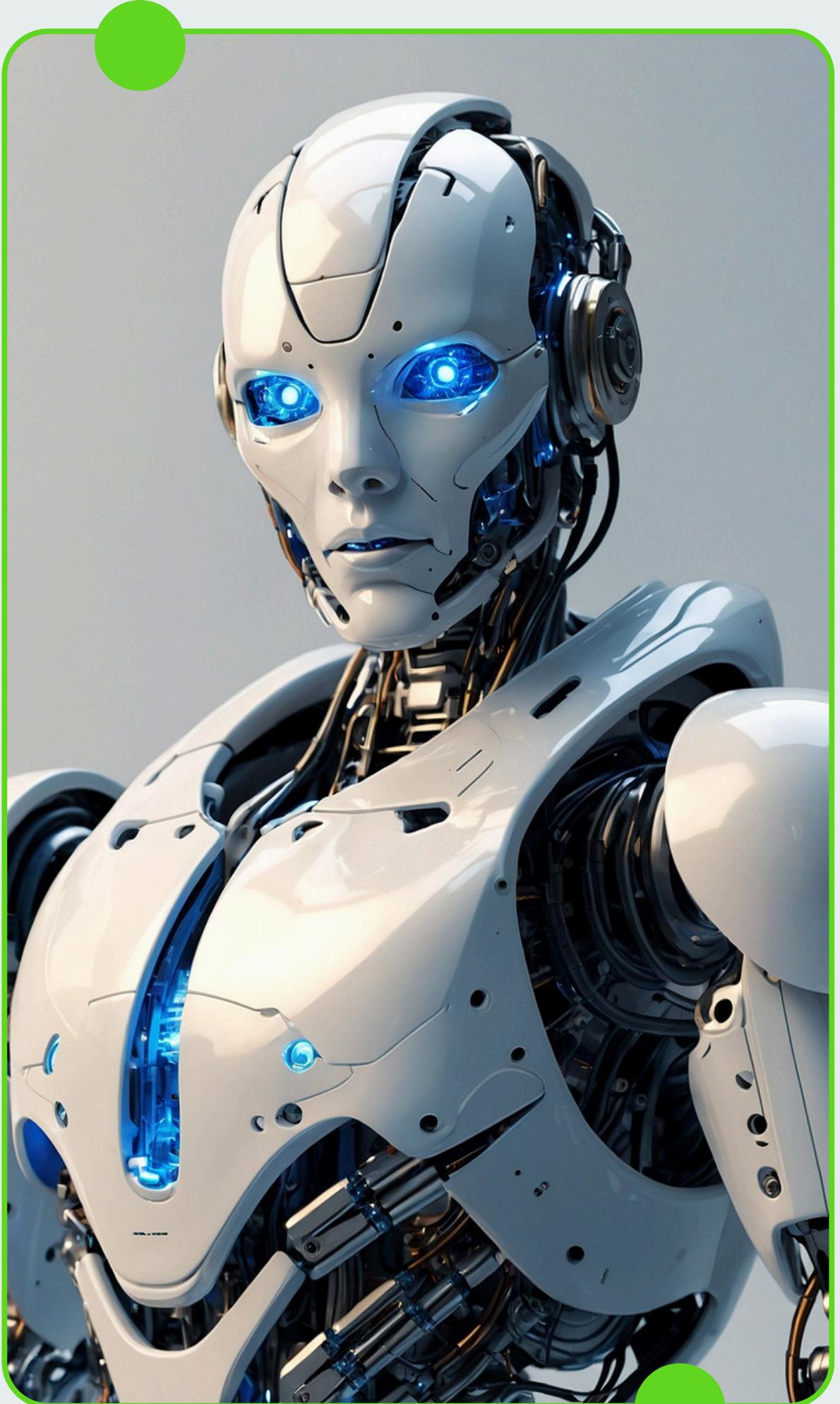
DummyClassifier : Accuracy
~**79%** (Trop simple).

VS



Régression logistique: Recall
~98-99% (Excellente capture
des motifs).

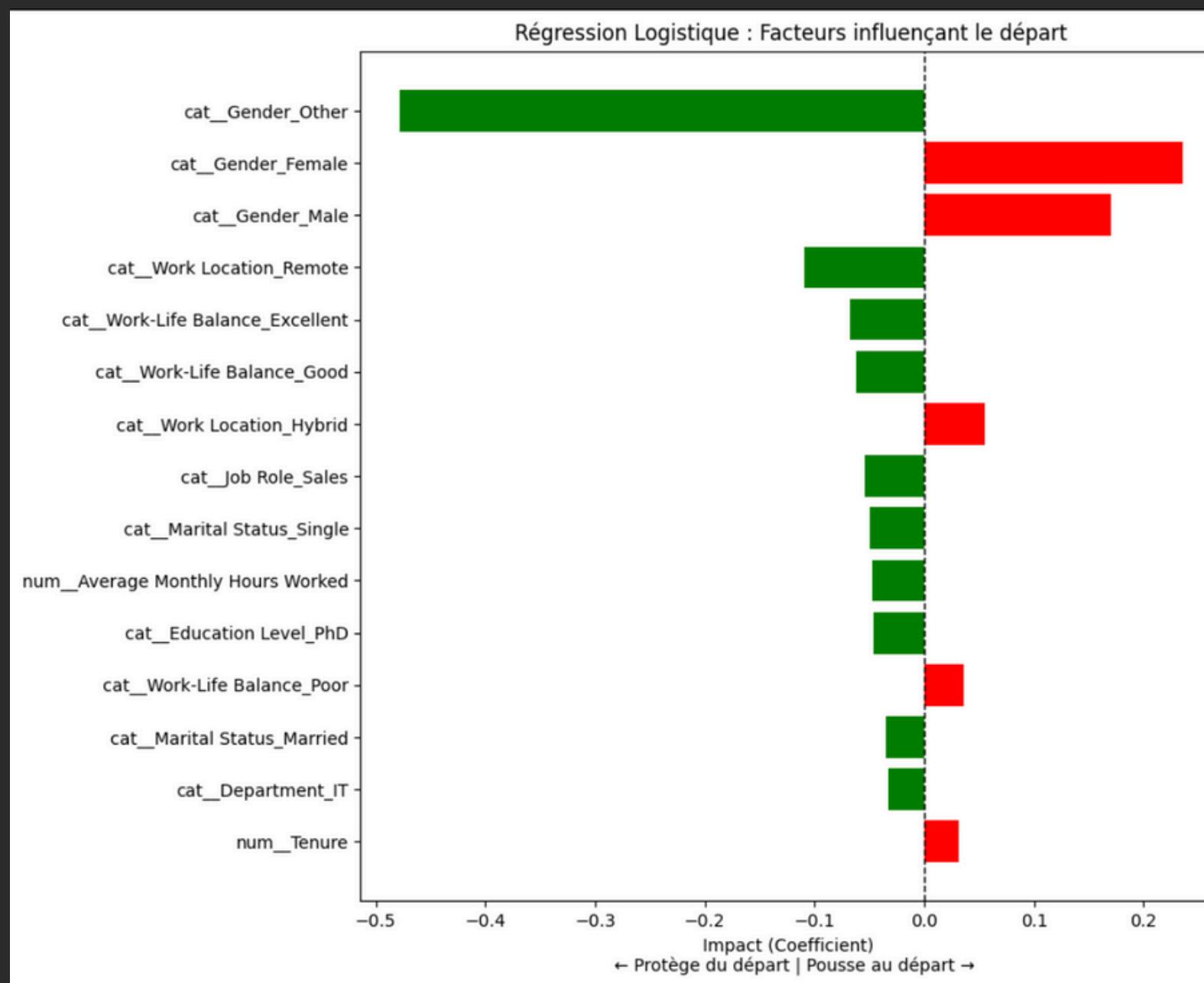
Métriques clés : Focus sur le **Recall** et le **F1-Score** pour minimiser les faux négatifs (employés qui partent sans être détectés) car l'accuracy seule est trompeuse.



Analyse de la Feature Importance

Globale

- Les variables les plus influentes sont le genre, le type de travail (à distance ou non) et l'équilibre vie pro vie perso
- Interprétation société :** Une hausse de charge/horaires peut impacter différemment selon la réalité sociale (double journée, flexibilité, pression). Le **genre** peut refléter un biais de données ou une inégalité organisationnelle (pas une cause biologique).
- Garde-fou :** À vérifier : performance/alertes par sexe + test du modèle sans la variable 'genre'.



Locale (Explicabilité)

- Capacité du modèle à expliquer pourquoi un individu spécifique a un score de risque élevé (ex: forte charge de travail + évaluation décevante).
- Utilisation métier :** "Justifier une alerte → comprendre quoi traiter (charge, WLB, perf, distance...).
- Action :** On propose des mesures sur les facteurs modifiables (organisation, management, planning), pas sur les attributs sensibles.

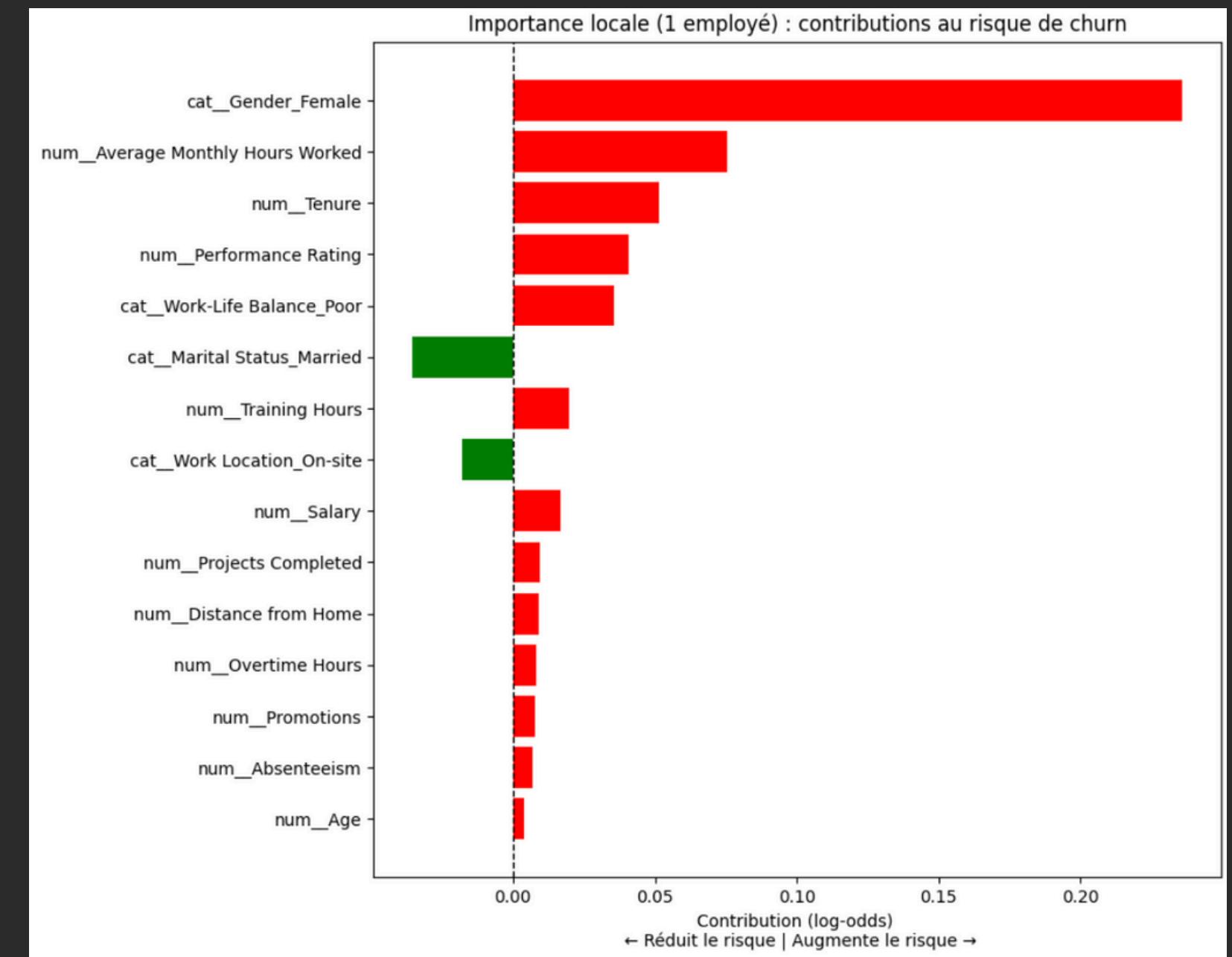


TABLEAU DU CLASSEMENT DES EMPLOYÉS LES PLUS À RISQUES

Score de risque + watchlist actionable pour les RH

	Employee ID	churn_risk_pct_str	risk_band
0	E07846	59.98%	Élevé
1	E00295	58.68%	Élevé
2	E07450	58.5%	Élevé
3	E04387	57.78%	Élevé
4	E08355	57.56%	Élevé
5	E05964	57.44%	Élevé
6	E09265	57.22%	Élevé
7	E06458	57.03%	Élevé
8	E03497	56.96%	Élevé
9	E03290	56.84%	Élevé
10	E08120	56.79%	Élevé
11	E02646	56.75%	Élevé
12	E08723	56.74%	Élevé
13	E07000	56.73%	Élevé
14	E07123	56.72%	Élevé
15	E03591	56.52%	Élevé
16	E06245	56.43%	Élevé
17	E05376	56.37%	Élevé
18	E00073	56.33%	Élevé
19	E01271	56.3%	Élevé

	Employee ID	churn_risk_pct_str	risk_band
667	E03876	51.21%	Moyen
668	E00163	51.21%	Moyen
669	E00371	51.21%	Moyen
670	E01145	51.21%	Moyen
671	E06236	51.2%	Moyen
672	E09438	51.2%	Moyen
673	E01454	51.19%	Moyen
674	E07722	51.19%	Moyen
675	E06617	51.19%	Moyen
676	E06917	51.19%	Moyen
677	E07534	51.18%	Moyen
678	E03769	51.17%	Moyen
679	E02648	51.17%	Moyen
680	E01064	51.17%	Moyen
681	E08076	51.16%	Moyen
682	E08670	51.16%	Moyen
683	E02156	51.16%	Moyen
684	E08639	51.16%	Moyen
685	E05257	51.15%	Moyen
686	E08230	51.15%	Moyen

	Employee ID	churn_risk_pct_str	risk_band
1333	E06418	48.63%	Faible
1334	E09454	48.63%	Faible
1335	E09605	48.63%	Faible
1336	E09902	48.62%	Faible
1337	E01453	48.62%	Faible
1338	E05512	48.61%	Faible
1339	E09889	48.61%	Faible
1340	E01237	48.61%	Faible
1341	E03666	48.6%	Faible
1342	E05862	48.6%	Faible
1343	E01848	48.59%	Faible
1344	E09915	48.59%	Faible
1345	E09562	48.58%	Faible
1346	E01052	48.58%	Faible
1347	E07531	48.57%	Faible
1348	E00574	48.57%	Faible
1349	E09969	48.57%	Faible
1350	E06346	48.56%	Faible
1351	E08665	48.56%	Faible
1352	E05212	48.56%	Faible

Deploiement & chiffrage (approche)

Architecture minimale

- Batch (quotidien/hebdo): calcul scores + export watchlist
- Stockage: CSV/DB + historisation des scores
- Suivi: MLflow + monitoring (drift, metriques)
- Dashboard: Tableau/BI pour KPIs & tendances



Merci pour votre
attention

