**Rapport Projet : Prédiction de l'Absentéisme**

**1. Contexte et Objectif**

L'entreprise souhaite comprendre les facteurs qui influencent l'absentéisme des employés et prédire la probabilité qu'un salarié s'absente. Cette analyse vise à identifier les leviers d'action pour la direction RH afin de réduire les absences et améliorer le climat de travail.

**2. Données Utilisées**

Le jeu de données contient 1000 employés avec des informations comme : âge, ancienneté, salaire, satisfaction, heures supplémentaires, distance domicile-travail, type de contrat (CDD/CDI), service et nombre d'absences.

La variable cible est transformée en binaire : absent = 1 si l'employé a eu au moins une absence, 0 sinon. Il s'agit donc d'un problème de classification.

**3. Analyse Exploratoire**

* **Satisfaction** : les employés absents ont en moyenne une satisfaction plus faible.
* **Heures supplémentaires** : les employés avec plus d'heures sup ont plus de risque d'absence.
* **Distance domicile** : une longue distance est légèrement corrélée aux absences.
* **Type de contrat** : les CDD présentent un risque d'absence légèrement plus élevé.

**4. Modélisation**

**Prétraitement**

* **Normalisation** : Utilisation de StandardScaler pour standardiser les variables numériques.
* **Encodage** : LabelEncoder pour les variables catégorielles (type\_contrat, service).
* **PCA** : Réduction de dimensionnalité pour simplifier l’analyse.

**Algorithmes testés**

1. **Régression Linéaire** :
   * Modèle basique pour établir une baseline.
   * Métrique : Erreur absolue moyenne (MAE) non affichée, mais utilisée pour validation.
2. **Arbre de Décision (DecisionTreeRegressor)** :
   * Meilleure adaptabilité aux relations non linéaires.
   * Visualisation de l’arbre via plot\_tree (non affichée dans le snippet).

**5. Features Importances**

**Variables influençant les absences**

D’après les modèles et l’analyse :

1. **Ancienneté** : Corrélée positivement avec les absences (employés plus anciens plus absents ?).
2. **Âge** : Impact non linéaire (pic d’absences chez les 27–50 ans).
3. **Distance domicile-travail** : Distance élevée → légère hausse des absences.
4. **Type de contrat** : Les CDD ont moins d’absences que les CDI (hypothèse à vérifier).

**6. Recommandations RH**

**a) Gestion de l’Absentéisme :**

* **Analyse ciblée** : Identifier les services/contrats associés à des absences élevées (ex. CDD vs CDI).
* **Actions préventives** : Programmes de bien-être (ex. télétravail pour réduire la distance domicile-travail).

**b)** **Équité Salariale :**

* **Audit des salaires** : Vérifier si l’ancienneté et l’âge sont correctement récompensés, surtout dans le service "Support" où les salaires pourraient être moins attractifs.

**c) Fidélisation des Employés :**

* **Stratégies CDI** : Convertir les CDD en CDI pour les employés performants, en lien avec la moyenne d’ancienneté élevée.
* **Développement professionnel** : Offrir des formations aux employés avec moins de 5 ans d’ancienneté pour renforcer leur engagement.

d) **Amélioration de la Satisfaction**

* **Enquêtes approfondies** : Analyser les facteurs de satisfaction (ex. charge de travail, reconnaissance).
* **Bonus liés à la performance** : Introduire des incitations pour les équipes ayant une satisfaction > 4/5.

e) **Optimisation des Données :**

* **Modélisation prédictive** : Utiliser des algorithmes (ex. régression) pour prédire l’absentéisme ou le turnover à partir des variables existantes.
* **Encodage des catégories** : Transformer les variables textuelles (ex. "type\_contrat") en données exploitables pour l’IA.

**7. Conclusion**

Le modèle fournit des indications claires sur les leviers RH pour réduire l’absentéisme. La prochaine étape pourrait être de déployer ce modèle en continu, par exemple sous forme de tableau de bord RH, pour identifier les employés à risque en temps réel.