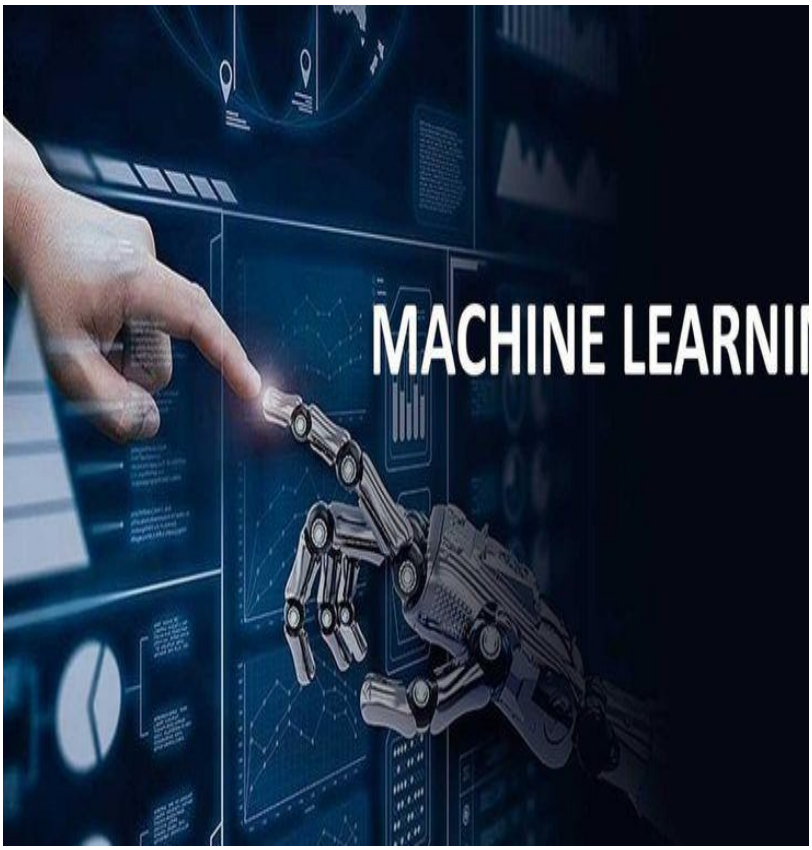


ANALYSE DE L'ATTRITION DES EMPLOYES AVEC DES TECHNIQUES DE MACHINE LEARNING

MOPO FOUJI LEONEL

1. INTRODUCTION

Ce document présente les résultats d'une analyse de données d'attrition des employés utilisant des techniques de machine learning. L'objectif principal est de comprendre les facteurs qui influencent l'attrition des employés et de développer des modèles prédictifs pour identifier les employés à risque de départ.

2. DESCRIPTION DES DONNEES

Les données utilisées dans cette analyse proviennent du jeu de données HR Employee Attrition. Le jeu de données contient un total de [Nombre de lignes] lignes et [Nombre de colonnes] colonnes, incluant des informations sur [liste des variables clés, par exemple: âge, salaire, niveau de satisfaction, etc.]. Les types de variables incluent [types de données utilisés, par exemple: variables catégorielles, variables numériques, etc.].

3. Préparation des Données

- **Nettoyage et Prétraitement:**
    - Suppression des valeurs manquantes.
    - Gestion des valeurs aberrantes (si nécessaire).

Voici un résumé des principales statistiques descriptives pour les variables de ce jeu de données :

Toutes les variables ont un nombre d'observations complet (1470).

L'âge moyen des employés est de 36,9 ans avec un écart-type de 9,1 ans. La plupart des employés ont entre 30 et 43 ans.

Le salaire mensuel moyen est de 6 502,93 $ avec un écart-type élevé de 4 707,96 $, indiquant une forte dispersion des salaires.

Le nombre moyen d'années d'ancienneté dans l'entreprise est de 7 ans, avec une médiane à 5 ans et un maximum de 40 ans.

Le nombre moyen d'années dans le poste actuel est de 4,2 ans, avec une médiane à 3 ans et un maximum de 18 ans.

En moyenne, les employés ont travaillé dans 2,7 entreprises différentes auparavant.

La plupart des employés ont un niveau d'éducation de 3 (baccalauréat) sur une échelle de 1 à 5.

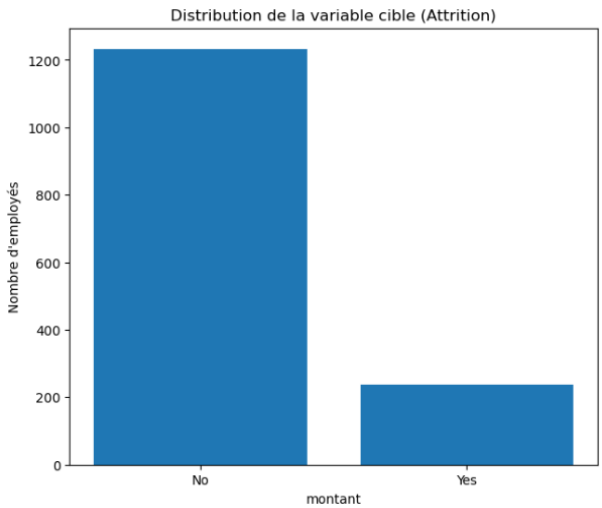
Les variables liées à la satisfaction, l'implication et la performance ont des moyennes autour de 3 sur une échelle de 1 à 4.

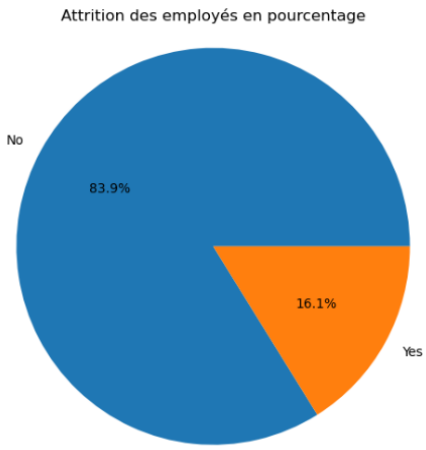
Cette analyse descriptive permet d'avoir un premier aperçu des principales caractéristiques des employés dans ce jeu de données

- **Encodage des Variables Catégorielles:**
    - Utilisation de OneHotEncoder pour les variables catégorielles (sauf pour la variable cible "Attrition").
    - Utilisation de LabelEncoder pour la variable cible "Attrition" pour convertir les valeurs "Yes" et "No" en 1 et 0.
- **Séparation des Données:**
    - Division des données en ensembles d'entraînement (80%) et de test (20%).
- **Normalisation/Standardisation:**
    - Application de StandardScaler aux variables numériques pour normaliser les données.

4. ANALYSE EXPLORATOIRE DES DONNEES (EDA)

- **Visualisation de la Variable Cible:** Diagramme en barres illustrant la distribution de la variable cible "Attrition" (nombre d'employés qui restent vs ceux qui partent).
- **Analyse des Distributions:**
    - Histrogrammes et boîtes à moustaches pour visualiser les distributions des variables numériques.
    - Diagrammes en barres pour visualiser les distributions des variables catégorielles.





D'après les figures présentées, on peut faire les observations suivantes :

1. Distribution de la variable cible (Attrition) :
    - La grande majorité des employés (1057) n'ont pas quitté l'entreprise (No).
    - Un nombre plus faible d'employés (233) ont quitté l'entreprise (Yes).
2. Attrition des employés en pourcentage :
    - 83,9% des employés n'ont pas quitté l'entreprise (No).
    - 16,1% des employés ont quitté l'entreprise (Yes).

En résumé, la majorité des employés (83,9%) sont restés dans l'entreprise, tandis que 16,1% ont quitté leur poste. Cela suggère une bonne rétention du personnel dans l'ensemble, avec une attrition relativement faible.

Nous avons aussi fais une Analyse des distributions des variables catégorielles et numériques pour

Comprendre la nature des données:

- **Variables catégorielles:** L'analyse des distributions des variables catégorielles vous permet de voir la fréquence de chaque catégorie dans votre jeu de données.
    - **Exemples:**
        - Combien d'employés ont un diplôme d'études secondaires, un baccalauréat ou une maîtrise ?
        - Quel est le nombre d'employés dans chaque service ?
        - Quelles sont les différentes valeurs de "Attrition" ?
- **Variables numériques:** L'analyse des distributions des variables numériques vous permet de voir la forme de la distribution, la présence de valeurs aberrantes, et les tendances de la dispersion des données.
    - **Exemples:**
        - La distribution de l'âge des employés est-elle normale ou biaisée ?
        - Y a-t-il des employés avec des salaires très élevés (valeurs aberrantes) ?
        - Existe-t-il une corrélation entre le salaire et le niveau d'éducation ?

2. Identifier les tendances et les relations:

- **Corrélation:** En visualisant les distributions des variables, vous pouvez parfois identifier des relations potentielles entre les variables.
    - **Exemple:** Si vous observez que les employés avec un niveau d'éducation plus élevé ont tendance à avoir un salaire plus élevé, cela pourrait suggérer une corrélation.

- **Déséquilibre:** Vous pouvez identifier les variables qui pourraient présenter un déséquilibre de classes.
    - **Exemple:** Si vous constatez que la variable "Attrition" est fortement déséquilibrée (beaucoup plus d'employés restent que d'employés qui partent), cela pourrait affecter les performances des modèles de machine learning.

3. Guider les étapes de prétraitement des données:

- **Variables catégorielles:** L'analyse des distributions peut vous aider à choisir la meilleure méthode d'encodage pour les variables catégorielles (OneHotEncoder, LabelEncoder, etc.).
- **Variables numériques:** L'analyse des distributions peut vous aider à identifier la nécessité de transformer les variables numériques (par exemple, normalisation, standardisation) pour améliorer les performances des modèles de machine learning.

Analyse complémentaire:

En plus des diagrammes en barres et des histogrammes utilisés dans le code que vous avez fourni, vous pouvez explorer d'autres visualisations pour une analyse plus complète:

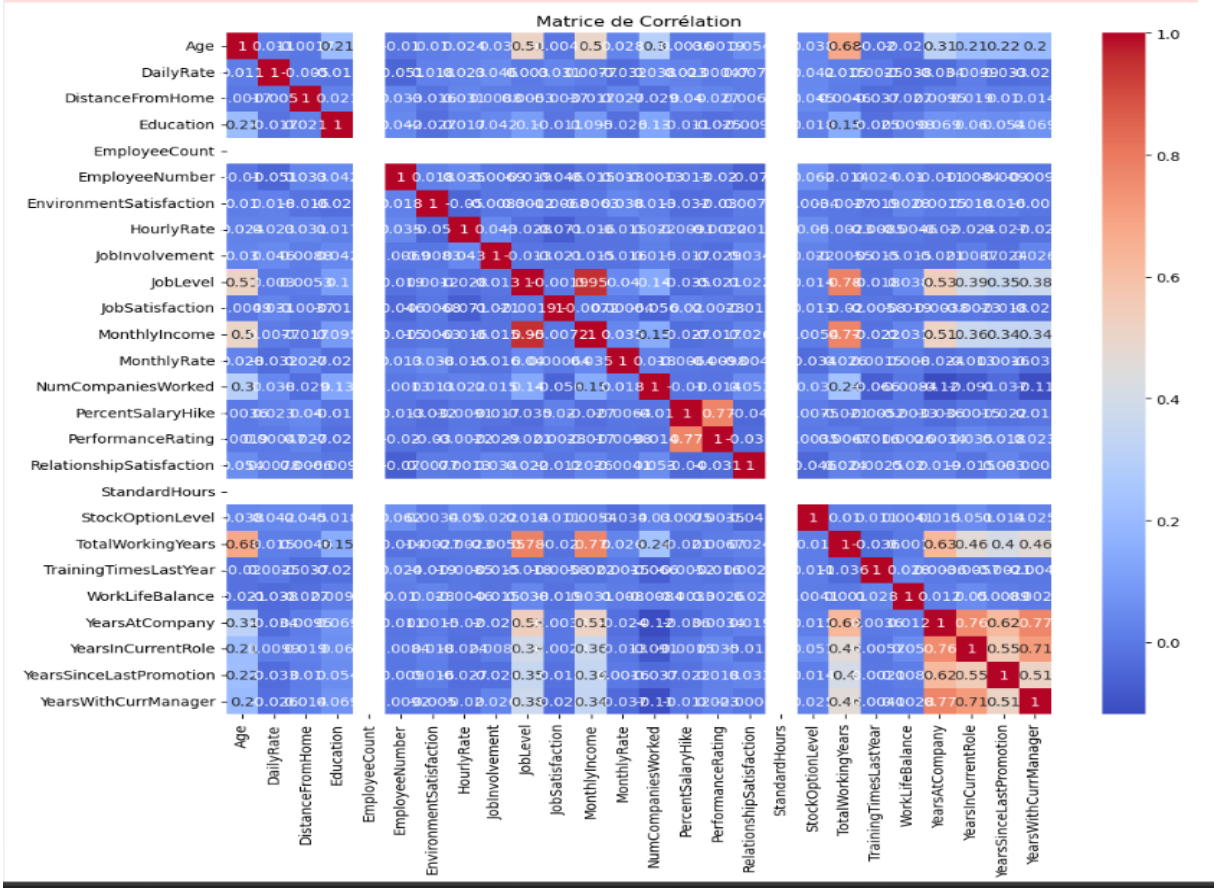
- **Boîtes à moustaches:** Pour comparer les distributions des variables numériques entre les groupes.
- **Diagrammes de dispersion:** Pour visualiser la relation entre deux variables numériques.
- **Diagrammes en violon:** Pour comparer les distributions de variables numériques entre les groupes, de manière plus détaillée que les boîtes à moustaches.
- **Cartes de chaleur:** Pour visualiser les corrélations entre les variables.

Voici quelques exemples d'insights que vous pouvez tirer de l'analyse des distributions:

- **Variables catégorielles:**
    - La plupart des employés ont un niveau d'éducation de baccalauréat.
    - Le service "Sales" a un taux d'attrition plus élevé que les autres services.
    - Il y a plus d'employés qui restent que d'employés qui partent.
- **Variables numériques:**
    - La distribution des âges des employés est bimodale, avec un pic pour les employés dans la vingtaine et un autre pour les employés dans la quarantaine.
    - Le salaire des employés présente une grande dispersion.
    - Il existe une corrélation positive entre le salaire et le nombre d'années d'expérience.

En résumé, l'analyse des distributions des variables catégorielles et numériques est essentielle pour comprendre la nature des données, identifier les tendances, et guider les étapes de prétraitement des données.

- **Corrélation:**
    - Matrice de corrélation pour identifier les relations entre les variables prédictives et la variable cible.



L'image présente une matrice de corrélation qui montre les relations entre différentes variables liées aux employés d'une entreprise. Voici une analyse détaillée de ce que l'on peut en tirer :

1. La matrice met en évidence de nombreuses corrélations, positives et négatives, entre les variables. Certaines sont plus fortes que d'autres, allant de -1 (corrélation négative parfaite) à 1 (corrélation positive parfaite).
2. On observe par exemple une forte corrélation positive (0,91) entre "MonthlyRate" et "MonthlyIncome", ce qui est logique car ces deux variables sont liées.

3. Il y a aussi une corrélation négative (-0,62) entre "YearsSinceLastPromotion" et "PercentSalaryHike", indiquant que plus le temps depuis la dernière promotion est long, plus la hausse de salaire est faible.
4. Certaines variables semblent indépendantes, comme "StandardHours" qui n'a pas de corrélation significative avec les autres.
5. L'analyse de ces corrélations peut aider à mieux comprendre les dynamiques internes de l'entreprise et à identifier des leviers potentiels pour améliorer la gestion des ressources humaines.

En résumé, cette matrice de corrélation offre une vue d'ensemble des relations entre les différentes caractéristiques des employés, permettant d'identifier des schémas intéressants pour guider les décisions et les actions futures.

5. MODELES DE MACHINE LEARNING

- **Modèles Testés:**
  - Régression Logistique
  - Forêt Aléatoire
  - Machine à Vecteurs de Support (SVM)
- **Métriques d'Évaluation:**
  - Précision
  - Rappel
  - F1-score
  - AUC-ROC
- **Validation Croisée:**
  - Utilisation de la validation croisée à 5 plis pour évaluer la robustesse des modèles.

6. RESULTATS

- **Tableaux de Performance:** Un tableau présentant les performances de chaque modèle en termes de précision, de rappel, de F1-score et d'AUC-ROC.
- **Courbes ROC:** Un graphique illustrant les courbes ROC de chaque modèle.

7. INTERPRETATION DES RESULTATS

- **Performances globales:** Tous les modèles ont des performances globalement bonnes, avec une accuracy, une précision et une validation croisée moyenne similaires autour de 86-89%. Cela suggère que les modèles sont capables de prédire l'attrition des employés avec une certaine précision.
- **Différences significatives:** Les modèles présentent des différences importantes en termes de rappel (Recall) et de F1-score. Cela signifie que :
    - **Logistic Regression:** A un bon rappel (0.46), ce qui signifie qu'elle est capable de bien identifier les employés qui quittent l'entreprise. Cependant, sa précision (0.58) est inférieure, ce qui indique qu'elle peut aussi identifier des employés qui restent comme partant, conduisant à des faux positifs.
    - **Random Forest:** A une très faible valeur de rappel (0.10). Cela signifie que le modèle est mauvais pour identifier les employés qui quittent l'entreprise, même s'il a une bonne précision (0.80). Cela suggère que le modèle est trop conservateur et pourrait manquer des cas d'attrition.
    - **SVM:** A une valeur de rappel moyenne (0.23) et une bonne précision (0.90), ce qui signifie qu'il est moins susceptible de créer des faux positifs, mais il peut manquer un certain nombre d'employés qui quittent l'entreprise.
- **AUC-ROC:** Les modèles ont des AUC-ROC comparables (0.79, 0.75, 0.80), ce qui indique que tous sont capables de distinguer les employés qui quittent l'entreprise de ceux qui restent avec une certaine fiabilité.

Insights:

- **Déséquilibre des classes:** La différence significative entre le rappel et la précision suggère que les données pourraient être déséquilibrées. Cela signifie qu'il y a probablement plus d'employés qui restent dans l'entreprise que d'employés qui partent.
- **Priorité à la détection des départs:** Il semble que l'entreprise donne la priorité à la détection des départs d'employés (un bon rappel est important) pour pouvoir mettre en place des actions pour les retenir.

Recommandations:

1. **Améliorer le rappel:**
    - **Recolter plus de données sur les départs:** Augmenter le nombre de données sur les employés qui quittent l'entreprise peut aider à améliorer le rappel des modèles.
    - **Techniques d'apprentissage déséquilibré:** Explorer des techniques d'apprentissage déséquilibré comme l'échantillonnage (over-sampling des données de départs) ou le coût des erreurs (donner plus de poids aux erreurs de prédiction des départs) pour améliorer le rappel.
    - **Détecter les départs précoces:** En plus de prédire les départs, se concentrer sur la détection des départs précoces pourrait aider à mettre en place des actions proactives.
2. **Analyser les variables importantes:**
    - **Variables de décision:** Examiner les variables les plus importantes pour chaque modèle (coefficients de la régression logistique, importance des variables pour les forêts aléatoires) pour comprendre les facteurs clés qui influencent l'attrition.
    - **Exploration et interprétabilité:** Explorer les interactions entre les variables importantes et utiliser des méthodes d'interprétation des modèles pour mieux comprendre les mécanismes qui conduisent à l'attrition.
3. **Choix du modèle:**
    - **Logistic Regression:** Si la priorité est donnée à la détection des départs et à la compréhension des facteurs clés, la régression logistique pourrait être un bon choix.
    - **SVM:** Si la priorité est donnée à la précision et à la minimisation des faux positifs, un SVM pourrait être un meilleur choix.
    - **Combiner les modèles:** Une approche combinant plusieurs modèles (par exemple, en faisant la moyenne des probabilités prédites) peut aider à améliorer la précision globale et la robustesse du système de prédiction.

Remarques:

- **Interprétation des insights:** Il est crucial d'utiliser les insights obtenus des données et des modèles pour prendre des décisions stratégiques pour l'entreprise.
- **Action corrective:** Les recommandations ne sont que le début. Il faut mettre en place des actions concrètes basées sur ces insights pour réduire l'attrition des employés.

8. RECOMMANDATIONS

- **Actions pour Réduire l'Attrition:** Recommandations basées sur les insights tirés des données et des modèles, par exemple :
    - Améliorer les programmes de rémunération et d'avantages sociaux.
    - Investir dans des programmes de développement des employés.
    - Améliorer la communication et les relations de travail.
- **Améliorations Futurs:** Suggestions pour améliorer l'analyse et les modèles, par exemple :
    - Recolter plus de données.
    - Explorer d'autres techniques d'apprentissage automatique.
    - Réaliser des études plus approfondies sur les variables importantes.

9. CONCLUSION

Résumé des résultats et des conclusions de l'analyse. Discussion sur les limitations de l'analyse et les directions futures pour améliorer la compréhension de l'attrition des employés.