**Orange Summer Challenge 2024: Test Technique Machine Learning – IA**

Nom: VANA ZOKOM ELIE

Ville : Maroua

Email: [eliezokom@gmail.com](mailto:eliezokom@gmail.com)

**Introduction**

L'entreprise "HR Insights" souhaite prédire quels employés sont susceptibles de quitter l'entreprise afin de mettre en place des actions préventives pour réduire le taux d'attrition. Ce document détaille les étapes suivies pour préparer, analyser, entraîner des modèles et interpréter les résultats pour la prédiction de l'attrition des employés.

#### Étape 1 : Préparation et Exploration des Données

**1. Téléchargement et Chargement des Données**

* Les données ont été chargées dans un DataFrame Pandas.
* Les premières lignes du DataFrame et des statistiques descriptives ont été affichées pour comprendre la structure des données.
* Les valeurs manquantes ont été vérifiées, ainsi que les valeurs aberrantes.

**Étape 2 : Analyse Exploratoire des Données (EDA)**

**2. Analyse des Variables Cibles et Prédictives**

* La distribution de la variable cible (Attrition) a été visualisée.
* Les distributions des variables catégorielles et numériques ont été analysées.
* Les relations entre la variable cible et les autres variables ont été explorées à l'aide de boxplots et countplots.
* Une heatmap des corrélations entre les variables numériques a été générée.

**Étape 3 : Préparation des Données pour le Modèle**

**3. Encodage des Variables Catégorielles et Séparation des Données en Ensembles d'Entraînement et de Test**

* Les variables catégorielles ont été encodées en utilisant One-Hot Encoding.
* Les données ont été séparées en ensembles d'entraînement (train) et de test (test).
* Les variables numériques ont été standardisées.

**Étape 4 : Entraînement des Modèles**

**4. Choix des Modèles et Évaluation des Performances**

* Trois modèles de classification ont été entraînés : Logistic Regression, Random Forest, et SVM.
* Les performances des modèles ont été évaluées à l'aide des métriques telles que l'accuracy, la précision, le rappel, le F1-score et l'AUC-ROC.
* Une validation croisée a été effectuée pour évaluer la robustesse des modèles.

**Étape 5 : Interprétation et Recommandations**

**5. Interprétation des Résultats et Recommandations**

* Les résultats des modèles ont été interprétés pour identifier les variables les plus importantes.
* Les recommandations basées sur les insights obtenus à partir des données et des modèles ont été proposées.

**CONCLUSIONS**

* Les variables les plus importantes pour prédire l'attrition sont : MonthlyIncome, Age, TotalWorkingYears, MonthlyRate, EmployeeNumber, DistanceFromHome, YearsAtCompany, HourlyRate, DailyRate, et OverTime\_No.
* Les employés avec des revenus mensuels plus bas, vivant loin de l'entreprise, ou avec moins d'années de travail sont plus susceptibles de quitter l'entreprise.

**RECOMMANDATIONS**

#### 1. Augmenter les Salaires

**Observation :** Le revenu mensuel (MonthlyIncome) est la variable la plus importante pour prédire l'attrition. Les employés avec des salaires plus bas sont plus susceptibles de quitter l'entreprise.

**Recommandation :** Réévaluer les structures salariales pour s'assurer qu'elles sont compétitives par rapport au marché. Offrir des augmentations de salaire basées sur la performance et l'expérience pour retenir les talents clés.

#### 2. Améliorer les Opportunités de Carrière

**Observation :** Les années totales de travail (TotalWorkingYears) et les années passées dans l'entreprise (YearsAtCompany) sont des facteurs significatifs d'attrition.

**Recommandation :** Développer des plans de carrière clairs et offrir des opportunités de développement professionnel, telles que des formations et des programmes de mentorat. Encourager les promotions internes pour maintenir l'engagement des employés expérimentés.

#### 3. Faciliter le Télétravail ou le Travail Flexible

**Observation :** La distance du domicile au travail (DistanceFromHome) est également un facteur important. Les employés vivant loin de leur lieu de travail peuvent être plus enclins à quitter l'entreprise.

**Recommandation :** Mettre en place des politiques de télétravail ou offrir des options de travail flexible pour réduire le stress lié aux trajets quotidiens. Envisager des aides financières pour les frais de transport.

#### 4. Revoir les Politiques d'Heures Supplémentaires

**Observation :** Le fait de ne pas faire d'heures supplémentaires (OverTime\_No) est une variable importante. Les employés peuvent ressentir une surcharge de travail ou un manque de reconnaissance pour les heures supplémentaires effectuées.

**Recommandation :** Assurer une répartition équitable des heures supplémentaires et offrir des compensations appropriées pour les employés qui en font. Revoir les charges de travail pour éviter le surmenage.

**Lexique**

**1. Attrition**

**Définition :** L'attrition se réfère au phénomène par lequel des employés quittent une organisation volontairement ou involontairement. Cela peut inclure des départs à la retraite, des démissions ou des licenciements. La réduction de l'attrition est souvent un objectif clé pour les services des ressources humaines, car elle peut être coûteuse et perturbante pour l'entreprise.

**2. Random Forest**

**Définition :** Random Forest est un algorithme d'apprentissage automatique de type ensemble qui utilise une collection de modèles d'arbres de décision pour améliorer la précision des prédictions. Chaque arbre est construit en utilisant un échantillon aléatoire de données et un sous-ensemble aléatoire de caractéristiques.

**3. Feature Importance (Importance des Variables)**

**Définition :** L'importance des variables est une mesure utilisée pour déterminer quelles caractéristiques ont le plus d'influence sur les prédictions d'un modèle. Dans le contexte des Random Forest, l'importance est généralement calculée en mesurant la réduction de l'impureté (par exemple, la réduction de Gini) apportée par chaque caractéristique.

**4. One-Hot Encoding (Encodage One-Hot)**

**Définition :** L'encodage one-hot est une technique utilisée pour transformer des variables catégorielles en variables numériques. Chaque catégorie unique d'une variable est convertie en une nouvelle colonne binaire (0 ou 1), ce qui permet au modèle d'apprentissage automatique de traiter ces données.

**5. Cross-Validation (Validation Croisée)**

**Définition :** La validation croisée est une technique d'évaluation des modèles qui consiste à diviser les données en plusieurs sous-ensembles et à utiliser certains sous-ensembles pour entraîner le modèle et les autres pour le tester. Cette méthode permet de garantir que le modèle est robuste et qu'il généralise bien aux nouvelles données.

**6. AUC-ROC**

**Définition :** AUC-ROC (Area Under the Receiver Operating Characteristic Curve) est une mesure de performance pour les modèles de classification binaire. Elle représente la probabilité qu'un classificateur classe correctement une instance positive par rapport à une instance négative. Un AUC-ROC de 1 indique une classification parfaite, tandis qu'un AUC-ROC de 0.5 indique une classification aléatoire.