

Test Technique Quantmetry

Objectif

Cet exercice a pour objectif d'évaluer vos compétences en data science, à travers votre connaissance des statistiques, de l'apprentissage automatique, et vos capacités en programmation.

Votre rendu comportera **au choix** :

- ▶ Un fichier texte de réponse aux questions de 6 pages au maximum (format PDF ou Word) et un fichier contenant votre code en R ou en Python (précisez la version en en-tête du fichier). Votre script devra pouvoir être rejoué facilement. N'hésitez pas à fournir un fichier README ou à spécifier dans le corps du script les actions nécessaires à la relance de votre code.
- ▶ Ou un notebook Python correctement mis en page contenant vos réponses aux questions et le code utilisé.

Enoncé

Le jeu de données contenu dans `data.csv` décrit des candidatures au poste de chercheur d'or chez OrFée. Votre objectif consiste à prédire le succès ou l'échec d'une candidature. Le jeu de données comporte 11 colonnes :

- ◇ *date* – date de la candidature
- ◇ *age* – âge du candidat
- ◇ *diplome* – plus haut diplôme obtenu (bac, licence, master, doctorat)
- ◇ *specialite* – spécialité du diplôme (géologie, forage, détective, archéologie, ...)
- ◇ *salaire* – salaire demandé
- ◇ *dispo* – oui : disponibilité immédiate, non : pas disponible immédiatement
- ◇ *sexe* – féminin (F) ou masculin (M)
- ◇ *exp* – nombre d'années d'expérience
- ◇ *cheveux* – couleur des cheveux (châtain, brun, blond, roux)
- ◇ *note* – note (sur 100) obtenue à l'exercice de recherche d'or
- ◇ *embauche* – le candidat a-t-il été embauché ? (0 : non, 1 : oui)

1. Statistiques descriptives

1. Décrivez le jeu de données. Présentez seulement les analyses et éventuels retraitements qui vous paraissent les plus pertinents et faites une première conclusion sur les variables à sélectionner en vue de la prédiction du succès ou de l'échec d'une candidature.
2. Y a-t-il une dépendance statistiquement significative entre :
 - (a) La spécialité et le sexe ?
 - (b) La couleur de cheveux et le salaire demandé ?
 - (c) Le nombre d'années d'expérience et la note à l'exercice ?
3. Quelle procédure statistique vous permettrait de dire si les distributions en âge des deux populations (femmes et hommes) sont identiques ou non ? Décrivez les étapes de votre raisonnement.

2. Machine Learning

1. Concevez un modèle permettant de prédire la variable *embauche* et expliquez votre choix d'algorithme. Si votre modèle comporte des spécificités de paramétrage, justifiez également vos choix de paramètres.
2. Décrivez et justifiez le critère de performance utilisé.
3. Proposez deux à trois pistes d'amélioration de votre modèle.