Test Technique :

# Machine Learning / Régression / Classification

# Contexte

Notre entreprise accompagne les entreprises dans la valorisation de leurs données. Pour cela, nous développons des outils qui permettent d’interpréter, de modéliser et d’ajouter de l’intelligence aux data sets propriétaires des entreprises. Cette intelligence vient par exemple de la création automatique de nouvelles features et/ou de l’intégration de données extérieures.

L'une des étapes la plus importante dans ce processus est de collecter, de compléter, de normaliser et de réaliser des analyses statistiques du dataset. Par conséquent, les bibliothèques de Python telles que pandas, numpy, matplotlib, sklearn, seaborn et toutes les autres bibliothèques de modèles d’expressions régulières sont essentielles pour obtenir les meilleurs résultats.

# Description de l’exercice

Le but de cet exercice est de déterminer votre capacité à optimiser ces bibliothèques écrites en Python.

Le contexte est celui de l’emploi. L’objectif de cet exercice est de déterminer le métier d’un candidat à partir des informations sur ses compétences.

A partir d'un dataset de compétences, vous réaliserez :

* Un clustering non supervisé afin d'identifier 2 groupes de profils techniques distincts
* Une prédiction des profils dont le métier n'est pas labellisé
* Un plan d’action pour répondre à un problème complexe. Il ne vous est pas demandé de coder, mais d’expliquer la démarche à adopter pour réaliser le projet, ainsi que des pistes que vous aurez trouvées sur le plan algorithmique.

# Descriptif des données

Pour cet exercice on vous fournit 2 fichiers :

* Test.ipynb : Un Jupyter notebook qui contient les questions, ainsi que des indications pour y répondre. C’est à la fois votre questionnaire et votre support de réponse pour ce test.
* Data.csv : Ce fichier contient un tableur de ~10.000 lignes décrivant le profil des candidats.

Ce tableau est composé de 6 colonnes :

* + **Entreprise** : correspond à une liste d'entreprises fictives
  + **Métier** : correspond au métier du candidat (Cette liste contient les valeurs : « data scientist », « lead data scientist », « data engineer » et « data architecte »
  + **Technologies** : correspond aux compétences maîtrisées par le profil
  + **Diplôme** : correspond à son niveau scolaire (Bac, Master, PhD...)
  + **Expérience** : correspond au nombre d'années d'expériences
  + **Ville** : correspond au lieu de travail
* Deux CV types donnés à titre d’exemples pour répondre à la dernière question

# Conseils

Nos conseils de rédactions pour ce test :

* Vous devez utiliser Python, idéalement la version 3.
* Vous n’avez pas le droit d’ajouter d’autres librairies que celles chargées dans le script
* Vous devez vous conformer à la directive de [PEP 8 style guideline](https://www.python.org/dev/peps/pep-0008/) pour une meilleure lisibilité
* Veuillez justifier vos différents choix dans les commentaires, ils seront relus par un de nos data scientist

# Compatibilité

Veuillez utiliser Jupyter Notebook pour vos réponses. Votre code doit être facilement déployable, nous devons pouvoir exécuter votre code nous-mêmes.

Vous trouverez ci-dessous la liste des bibliothèques que vous devez installer sur votre terminal :

pip3 install matplotlib

pip3 install pandas

pip3 install numpy

pip3 install seaborn

pip3 install -U scikit-learn

# Soumission

Envoyez-nous votre notebook rempli à xxxx..

Prenez soin de justifier consciencieusement vos choix et d’analyser vos résultats. Un code clair et commenté permet une meilleure compréhension de votre démarche.

# Pour conclure

Vous trouverez les questions dans le notebook Jupyter jointe au format .ipynb. Veuillez utiliser ce fichier de test pour vos réponses et le renvoyer une fois que vous avez terminé les tâches.

Bonne chance !

xxx