Consignes test Technique

Machine Learning / Régression / Classification

## 1. A propos d'Equancy

Equancy est un cabinet de conseil international, basé à Paris et Dubaï, spécialisé dans la transformation data des entreprises.

Nous planifions, concevons et mettons en œuvre des solutions Big Data, Data Science et Intelligence Artificielle pour nos clients. Nos projets vont de la mise en œuvre d'infrastructures spécialisées dans le traitement de la donnée de nos clients, de lacs de données jusqu'au développement de systèmes opérationnels intégrant des algorithmes de machine learning ou de deep learning. Nous sommes experts dans l'industrialisation de ces plates-formes, en appliquant les principes du devops à nos infrastructures data.

Nos clients sont de grands groupes français et internationaux (LVMH, Picard, Chanel VINCI, Volkswagen). Ils nous font confiance autant dans l'accompagnement au cadrage de leurs besoins que dans la réalisation des solutions data innovantes.

## 2. Contexte

Equancy 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.

## 3. 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 les groupes de profils techniques distincts
- Une prédiction des profils dont le métier n'est pas labellisé

## 4. Descriptif des données

Pour cet exercice on vous fournit le fichier suivant :

- 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

## 5. Conseils

Nos conseils de rédactions pour ce test :

- Vous devez utiliser Python, idéalement la version 3.
- Vous devez vous conformer à la directive de PEP 8 style guideline pour une meilleure lisibilité

- Veuillez justifier vos différents choix dans les commentaires, ils seront relus par un de nos data scientist

## 6. 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
```

## 7. Soumission

Le Projet sera livré sous forme de projet Gitlab.
Prenez soin de justifier consciencieusement vos choix et d'analyser vos résultats. Un code clair et commenté permet une meilleure compréhension de votre démarche.

Bonne chance !
Equancy Team