Test Technique

Machine Learning / Régression / Classification

1. Contexte

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.

2. 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 compilé par Aquila, 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

3. Descriptif des données

Pour cet exercice Aquila 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

4. 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 pour une meilleure lisibilité
- Veuillez justifier vos différents choix dans les commentaires, ils seront relus par un de nos data scientist

5. 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
```

6. Soumission

Envoyez-nous votre code ou l'URL du référentiel à l'adresse cmeunier@aquiladata.fr.
Une explication de la façon dont vous avez abordé les problèmes, de la logique qui sous-tend les choix principaux de conception et du fonctionnement de votre solution (par exemple, une vue d'ensemble de l'architecture simple) est également la bienvenue.

7. Pour conclure

Vous avez environ 1 heure et 30 minutes pour répondre à toutes les questions.
Vous trouverez les questions dans la pièce jointe au format .ipynb. Veuillez utiliser ce fichier de test pour vos réponses et le renvoyer une fois que vous avez terminé les tâches.

Nous vous rappelons juste que nous sommes davantage intéressés par votre potentiel, plutôt que par vos compétences techniques actuelles : ainsi les questions de ce test ne sont pas éliminatoires, si vous ne savez pas répondre à une question il n'y a pas de soucis, mais apprendre rapidement est essentiel.

Bonne chance !

Aquila Data Team