Brief Projet Fin Machine Learning

Contexte :

Je suis CEO d'une boite qui s'occupe de faire des statistiques sur l'emploi dans le secteur du développement informatique et de la data en France.
Je m'intéresse tout particulièrement aux différences de salaires entre ces métiers + villes.
Je vous mandate afin de me fournir une étude sur ce marché à présenter sous forme de Dashboard.

NB : Vous travaillerez en mode agile.

Rendu : Votre lien Github + Présentation client avec PowerPoint (au plus tard le jeudi : **28/05/2020 à 18h00**)

Présentation :

1/ Une première présentation avec moi sur les aspects techniques (idéalement prévoir un powerpoint) - Durée 1h max
2/ Présentation devant la promo en mode client - 15mn

Votre mission :

1/ Faire un script de scraping sur indeed (https://www.indeed.fr/) qui permette à l'user de spécifier le type d'annonces qu'il souhaite récupérer :

- **Métier (développeur, data scientist…)**
- **Type de contrat recherché (CDI, CDD, freelance…)**
- **Lieu de recherche (Paris, Toulouse, …)**

Les infos à scraper :

- **Titre**
- **Nom de la boite**
- **Adresse**
- **Salaire**
- **Descriptif du poste**
- **Date de publication de l'annonce**

Vous pouvez vous concentrer sur les annonces :

- **Métiers : développeur, data scientist, data analyst, business intelligence.**
- **Localisation : Paris, Lyon, Toulouse, Nantes et Bordeaux.**
- **Type de contrat : tous**

2/ Prévoir un script qui permette de stocker automatiquement les infos scrapées dans une bdd Mongo (le script devra prendre en compte le fait de remplacer ou de ne pas tenir compte

d’une annonce si cette dernière est déjà dans la bdd).

3/ Récupérer les annonces pour lesquelles on a un salaire. (Il faudra un peu cleaner…)
Sur ces annonces l’objectif est de prédire le salaire en fonction des features à votre disposition (à vous de tester ce qui est pertinent)
Exemple de modèles à tester : **Random Forest, Logistic Regression, Kernel RBF, Gradient Boosting Classifier, XGBClassifier…**

4/ Sur les annonces pour lesquels il n’y a pas de salaire, déduire et compléter ce champ en fonction des résultats de 3/

5/ Créer un dashboard avec Flask/Dash. A vous de voir ce qui est pertinent de montrer au client.

6/ BONUS 1 : Faire un script qui permette d’actualiser automatiquement vos résultats chaque semaine.

7/ BONUS 2 : Dockeriser

8/ BONUS 3 : Pour chacune des entreprises scrapées, créer un script qui permette de récupérer sur LinkedIn toutes les infos disponibles sur ces dernières (taille de l’entreprise, spécialité, adresse mail, site, description, nombre d’employés etc.).
Ajouter ces informations dans la bdd.
ATTENTION DE NE PAS VOUS FAIRE BAN VOTRE COMPTE (Tips : Avec Selenium, chercher l’entreprise dans les suggestions et non sur la page de résultat)

9/ BONUS 4 : Avec le nom de l’entreprise (ou l’url du site récupéré via profil LinkedIn), rechercher sur le site web de l’entreprise le mail générique ainsi que leurs potentiels recherche de poste (onglet « On recrute », « Nous rejoindre » etc..).
Compléter la bdd mongo avec ces informations.

10/ BONUS 5 : Ne restez pas bloquer !!! Consultez-nous si besoin d’éclairage

Du courage !!!