Job Title Classification

L'algorithme développer permet la classification d'intitulé de poste afin de déterminer s'il s'agit d'un Data Scientist ou pas. Il s'agit là d'obtenir un résultat binaire.

Dans un premier temps les données ont été chargé via pandas library afin de les manipuler comme dataFrame. Une fois les deux fichiers JSON chargé(**"list_job_tech_dataScience_NO" et "list_job_tech_dataScience_YES"**) on concatène nos deux dataframe.

Après la visualisation des données chargé, j'ai remarqué que les données sont déséquilibrées avec une majorité pour les classes de Data Scientist. Donc j'ai opté pour l'oversampling aléatoire afin d'équilibrer mes données avant d'entrainer les différents modèles de classification. Une fois cela effectué j'ai mis en place un pipeline. Tout d'abord il s'agit d'appliqué le CountVectoriser où les occurrences de chaque mots sont comptabilisé, cela permet de représenter les données textuelles sous forme de vecteurs numériques. Ensuite, on applique le TF-IDF, il s'agit là de calculer la fréquence de mots en divisant le nombre d'occurrence de chaque mot par le nombre total de mots.

Enfin, j'ai testé deux modèles de classification qui ont donné le même résultat en commençant par le RandomForestClassifier puis le Régression Linéaire.