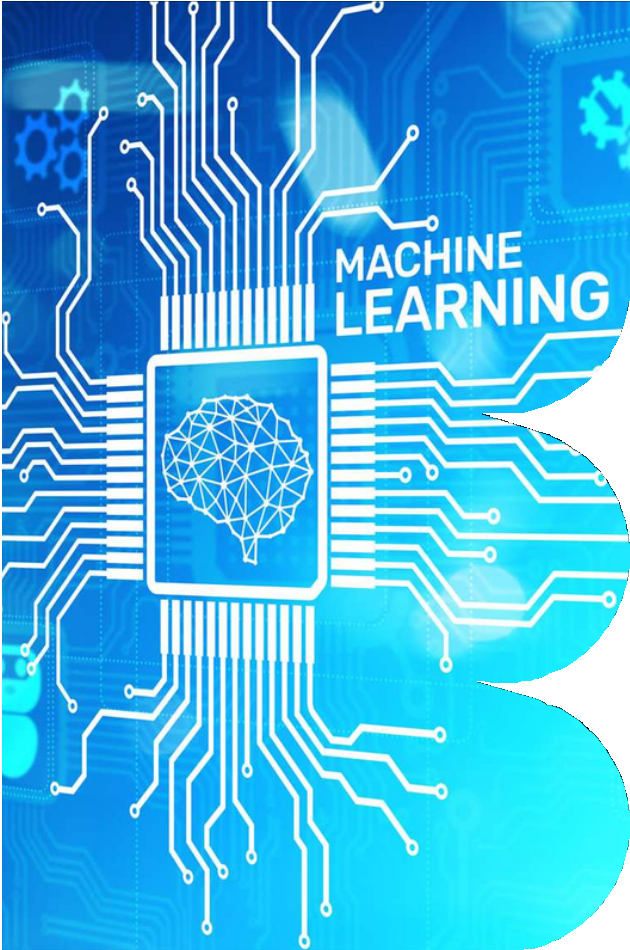


Rapport de projet de Machine Learning & Text Mining



PRÉSENTÉ À

Mr. ZAKRANI

PRÉSENTÉ PAR

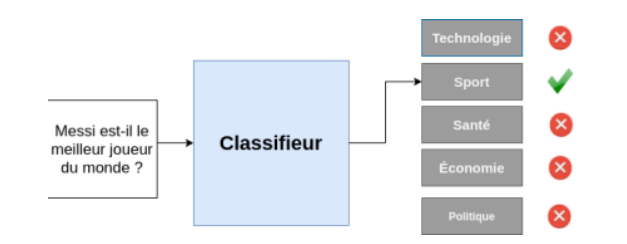
CHAFIKI Fatima elzahra

**1 - Définition du problème à résoudre :**

Dans ce projet , On a un problème de classification des données textes càd la Classification

Multi-Class  qui consiste à associer du texte à une classe parmi plusieurs (plus de deux) , dans notre cas les textes sont des articles de différentes catégorie : sport, économie, culture, politique et société.

Donc le but de ce projet : concevoir et développer un modèle intelligent pour classer les articles selon leur catégorie



**2 - Acquisition des données d’apprentissages et de tests :**

**La collecte** de **données** est une phase primordiale dans le machine learning , donc on a lit les

Fichier textes qui contient les article et les découper en phrases pour avoir un dataset qui contient 2 colonnes les phrases et leurs catégorie .

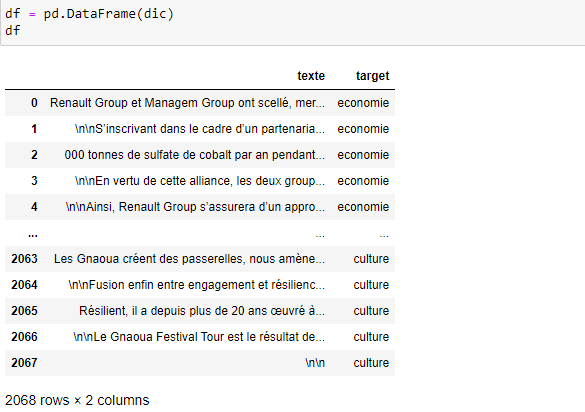


Dans le code ci-dessus , On a collecté les articles et les affecter une classe qui est la catégorie .

Après On va découper les articles en des phrases .



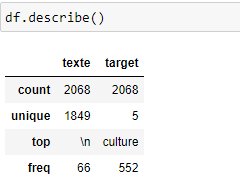
Le résultat final  de dataset :



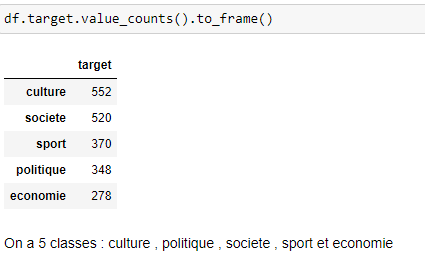
**3 - Analyser, explorer les données :**

Dans cette étape , on va analyser et explorer les données pour extraire des information autour de données .

a- les informations statistiques de dataset :



b- Analyse des classes :



On constate qu’on a un déséquilibrage des classes car la quantité des classes varie .

**4 - Préparer et nettoyer les données :**

a- nettoyage de texte :

Pour avoir un modèle performant , On a besoin de faire un traitement des données ,dans notre cas on va nettoyer les données **enlever les caractères** qui ne sont pas des lettres comme “#”, “-“, “!”, les **chiffres** ou bien encore **les mots qui n’ont pas de sens** ou qui ne font pas partie de la langue analysée.

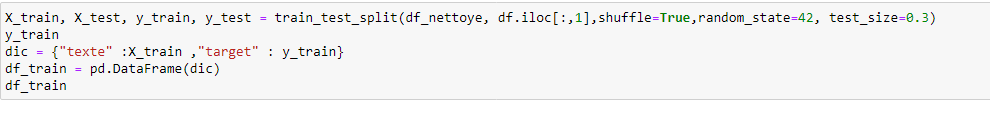


Le résultat final de dataset :



b-découpage des données :

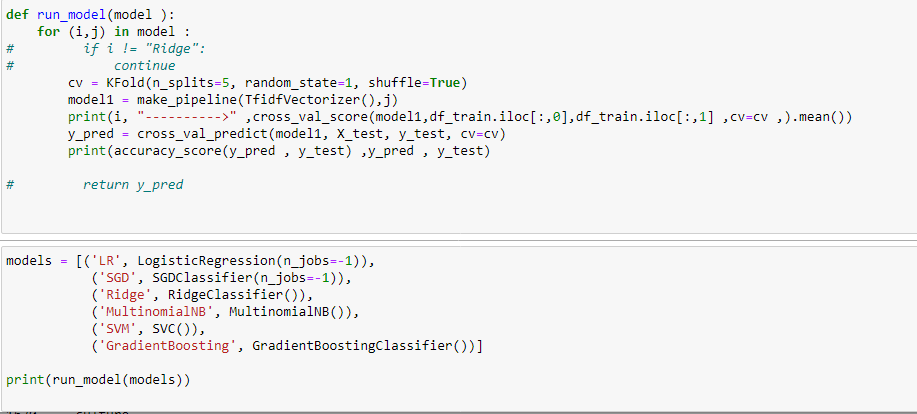
Pour valider notre modèle , on doit pas donner 100% de dataset pour le training , dans note cas on a donné 80% .



**5 - Choisir un modèle d’apprentissage :**

Dans cette étape on va choisir un modèle performant qui donne une accuracy elevée , c est pou

On va entrainer plusieurs modèle sur le dataset et ensuite choisir le plus approprié .



Donc on peut constater que l'accuracy de modèle SGD est la plus élevée .