



ÉCOLE
D'INGÉNIEURS
PARIS-LA DÉFENSE

Projet - Python for data analysis

Tenants et aboutissant du problème

Xavier ALEXIADE - DIA 2



Les données

D'après le fichier page-blocks.names, ce dataset est un ensemble d'analyses de blocs de différents documents. Ces analyses peuvent être utilisées pour retrouver si un bloc est en fait un texte, une ligne verticale/horizontale ou une image.

Ainsi, le dataset a pour informations sur le document :

- height : la hauteur du bloc
- lenght : la longueur du bloc
- area : la surface du bloc
- eccen : le rapport longueur/hauteur du bloc
- p_black : le % de pixels noirs dans la surface
- p_and : le % de pixels noirs après l'application du RLSA
- mean_tr : la moyenne de transition pixel blanc/noir
- blackpix : le nombre total de pixels noirs dans le bloc
- blackand : le nombre total de pixels noirs dans le bloc après le RLSA
- wb_trans : le nombre total de transition pixel blanc/noir



Les données

L'analyse de ces données peut donner comme résultat une de ces catégories :

- texte
- ligne horizontale
- image
- ligne verticale
- graphique

Les classes étant peu nombreuses on peut garder celles fournies.



Les données

Toutes les données du dataset sont sous forme numériques, il va donc être beaucoup plus simple d'appliquer des méthodes de machine learning sans avoir à numériser des données sous forme de string/catégories.

Problématique

Prédire au mieux un ensemble de données tests et donner la classe auxquelles ces données appartiennent.





Questions posées

- Dans un premier temps, je me pose la question de savoir si toutes les valeurs sont présentes ou s'il y a du nettoyage à faire ?
 - > Dans notre cas, toutes les données sont présentes et il n'y a pas de "NaN values".
- Puis, comme dit précédemment, je cherche si les classes (ou les outputs) ne sont pas trop nombreux ce qui rendrait l'analyse et la prédiction trop hasardeuse.
 - > Il se trouve qu'il n'y a que 5 classes possibles ce qui est très bien.
- Enfin, je me demande comment les colonnes influent sur les classes. Est-ce qu'une donnée élevée quelque part donne quasi obligatoirement une classe ?
 - > Avec l'analyse du dataset avec Pandas, je remarque qu'il y a quelques colonnes qui sont très liées à certaines catégories.



Etude du jeu de données

L'étude via plusieurs méthode de machine learning s'est faite de manière assez simple.

J'ai créer un modèle dans chacune des méthodes utilisées puis j'ai comparé leur accuracy avec le F1-score.