**Astuces & informations pratiques trouvées en Machine Learning à la CEMP**

Certains algos ont + tendance à prédire une erreur sur le 1 ou le 0 ça dépend. Donc on peut parfois inverser le pb.

**Imbalanced Dataset**

PU learning ou semi-supervisé. SVM de scikit-learn peut faire du semi-supervisé.

SMOTE: for Imbalanced Classification with Python -> data augmentation for the minority class

[SMOTE for Imbalanced Classification with Python - MachineLearningMastery.com](https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/)

Installation de SMOTE :

*En créant à la main un fichier* ***pip.ini (dans le dossier de mon environnement C:\Users\A3180813\AppData\Local\miniforge3\envs\test\_env)*** *et en mettant les commandes que Franck mettait dans****pip.conf*** *j’arrive à installer imblearn (pour avoir SMOTE). Je l’installe par pip du coup mais je pense qu’en ajoutant les artifactory à condarc ça devrait permettre de l’installer dans conda si besoin.Sinon sous le artifactory de « base » que j’utilise dans conda on à pas lime mais on a shap donc ça me va.*

**Kmeans**

Graphe kemans pour clustering. Epaisseur de la coque de bateau : nb de points. Calcule la valeur de la silouhette et la contribution à la de chaque point. On cherche à avoir des rectangles. a droite de la barre rouge: contribue négativement à la valeur de la silhouette moyenne. mais ce graphe est très long à sortir.

[Silhouette Analysis in K-means Clustering | by Mukesh Chaudhary | Medium](https://medium.com/@cmukesh8688/silhouette-analysis-in-k-means-clustering-cefa9a7ad111)

Pour des données mixtes (catégorielles, continues) : k-prototypes, pour des données catégorielles : k-modes

**Palentir**

Palentir foundry -> à voir. Tourne sur AWS mais on n'utile pas les modules AWS en direct. C'est des modules qui gèrent des bases de données. Palentir: utilise AWS mais pas les modules en direct, crée des interfaces + simples et louent des serveurs. Mais ils ont restreint plein de trucs.

GCP: notebook dans le cloud. La CE a accès à ça ?

**GridSearch**

Au lieu d’utiliser le gridSearch, il y a d’autres méthodes plus efficaces (chercher lesquelles) comme RandomSearch par ex. (Il y en a encore une autre mieux retrouver le nom)

**Programmation**

Programmer en PySpark : remplacer les fonctions python par des fonction PySpark dès que possible car cela permet de paralléliser les calculs. Cependant, PySpark ne possède pas toutes les fonctions ni toutes les options disponibles en python.

**Librairies qui existent**

PyPDF : analyse un doc pdf tout seul

Missingno : graphe pour visualiser les NaN

**Image processing** : super blog sur le traitement d’image en python et les OCR : [Image Processing | TheAILearner](https://theailearner.com/image-processing/). Rappel fonctionnement d’un noyau de convolution : [references - Any use of non-rectangular-shaped kernels in convolutional neural networks? Especially when analyzing game boards - Cross Validated (stackexchange.com)](https://stats.stackexchange.com/questions/235032/any-use-of-non-rectangular-shaped-kernels-in-convolutional-neural-networks-espe)

**Boîtes cool à Toulouse d’après Mathieu** :

Syngenta

Dataiku

Scale

Geotrend

Le métier de data scientist va évoluer vers des notebooks dans le cloud, des pipelines de données dans le cloud.

**Alteryx** :

Environnements pythons sur Alteryx :

Jupytertoolenv (pas de XGboost mais marche en serveur)

Machinelearningenv (XGbosst mais marche pas en serveur)

Après tu as les env locals, sauf que tu peux pas installer le sdk AYX