**Ressources Python for Data Science**

Identique au plan d'étude d'une fonction en mathématiques, il est possible de définir un plan d'étude d'un jeu de données (dataset)

| Étapes | Outils, Rendu |
| --- | --- |
| Étude du domaine de définition, du domaine d'étude. | Fonctions usuelles, parité, symétries. ; dom de définition, d'étude. |
| Limites aux bornes | Limites fonctions usuelles, développements limités |
| Propriétés | Propriétés des fonctions usuelles. Étude de la continuité , dérivabilité, ... |
| Tableau de variation, étude du signe de la dérivée. | Fonctions dérivés première, seconde si dérivabilité |
| Graphes | Plusieurs possibilités Courbes / Surfaces |
| Propriétés géométriques | ex. coniques |
| Propriétés analytiques | Aire sous la courbe, courbe sur une surface,... |
|  |  |
|  |  |

**Étude de dataset (EDA)**

| Étapes | Outils, Points à voir (non exhaustifs) |
| --- | --- |
| Lecture du /des fichiers, jointures | read\_csv, delimiter, encodage |
| Valeurs manquantes, nettoyage de la donnée | Stratégie d'imputation : moyenne, mediane, regression, suppression de l'observation. |
| Type des variables | Pandas.dtype. Discrète, continue, Numérique, Catégorielle, clé primaire, clé secondaire |
| Distribution | Histogrammes |
| Feature Engineering | Scikit-Learn, pandas. Création de nouvelles colonnes (ex dates, adresses), bucketization / bins, encodage des variables catégorielles (ex : OHE : OneHotEncoding |
|  |  |
| Datavisualisation | Matplotlib, pandas-profiling, bokeh, plotly |
| Modélisation (ML, DL) | Scikit-learn, statsmodel, tensorflow/keras,pytorch/fastai,... ; Prediction, Classification si supervisé. Clustering , Reduction de dimension si non supervisé. |
| Re-DataViz | Aussi sur les courbes d'apprentissage du modele |
| API fication | Flask, Django, AzureML, ML Engine (GCP ou AWS) |
| Application complète | Django, Flask, Angular, ... |

Exemple de cycle de vie Data Science, il manque ici la mise en production et le monitoring du modèle et de l'application associée (ML Engineering vs Data Science)

**Ressources présentées lors des TP**:

Polycopié officiel, Notebooks Romain Jouin

Version ancienne de pandas, voir le livre aussi

https://github.com/TomAugspurger/effective-pandas

<https://datacarpentry.org/python-ecology-lesson/>

merge :

https://datacarpentry.org/python-ecology-lesson/05-merging-data/index.html#:~:text=The%20pandas%20merge%20function%20supports,(s)%20values%20are%20discarded.

Dataviz aussi,

<https://github.com/fxjollois/cours-2019-2020>

O'Reilly Python for DS

<https://github.com/jakevdp/PythonDataScienceHandbook/tree/master/notebooks>

Stratégie d'imputation

cf. Video slides Julie Josse

http://juliejosse.com/wp-content/uploads/2019/07/useRjosse\_2019.pdf

MOOC , video et concepts illustrés en interactif.

Notebooks sur Collab

<https://developers.google.com/machine-learning/crash-course/>

https://developers.google.com/machine-learning/crash-course/first-steps-with-tensorflow/programming-exercises

Kaggle Learn,

plusieurs cours

dont

<https://www.kaggle.com/learn/intro-to-deep-learning>

pandas-profiling

<https://github.com/pandas-profiling/pandas-profiling>

<https://www.datacorner.fr/pandas-profiling>/

<https://www.kaggle.com/learn/intro-to-deep-learning>

<https://www.kaggle.com/learn/computer-vision>

Pour les CNN

[https://cs231n.github.io/convolutional-networks/#overview](https://cs231n.github.io/convolutional-networks/%23overview)

https://cs231n.github.io/convolutional-networks/#overview

Le tuto sur le site tensorflow utilise le même dataset que le notebook TP6, avec les explications :

<https://www.tensorflow.org/tutorials/keras/classification>

<https://colab.research.google.com/github/tensorflow/tpu/blob/master/tools/colab/fashion_mnist.ipynb>

Deep Learning

Yann Le Cun à l'Académie des Sciences

<https://www.youtube.com/watch?v=cvV3lk8j6qc>

Collège de France

<https://www.college-de-france.fr/site/yann-lecun/_audiovideos.htm>

NLP Natural Language Processing.

<https://huggingface.co/>

contient une librairie python , un hub de modèles NLP (Bert, GPT, CamenBert, …) , un hub de datasets.

Spacy, algo classiques de manipulation de données textes,

Avec qq algos de ML (moins que sur huggingface)

<https://spacy.io/>

qq fcts sympathique comme les Vizualizers

https://spacy.io/usage/visualizers#jupyter

Stanford NLP

Ancienne librairie avec les tokenizer et ressources pour bcp de langues

https://nlp.stanford.edu/software/

Flask introduction

<https://openclassrooms.com/fr/courses/4425066-concevez-un-site-avec-flask>

<https://www.kaherecode.com/tutorial/demarrer-avec-flask-un-micro-framework-python>

Flask tuto

<https://www.kdnuggets.com/2019/01/build-api-machine-learning-model-using-flask.html>

<https://www.analyticsvidhya.com/blog/2020/04/how-to-deploy-machine-learning-model-flask/>

AutoML

<https://www.automl.org/book/>

<https://optuna.org/>

Meetup avec Edge-ML et mlBox (slides) , video

<https://nuit-blanche.blogspot.com/2017/06/paris-machine-learning-meetup-hors.html>

<https://github.com/AxeldeRomblay/MLBox>