

Apprentissage automatique

TP 1 - Importer et examiner les données

Halim Djerroud

Séance 1 — révision 3.0

Objectifs du TP

- Installer et prendre en main l'environnement de travail (Python, Jupyter, Pandas, Scikit-learn).
- Importer un jeu de données depuis Scikit-learn et depuis un fichier.
- Examiner la structure d'un dataset : dimensions, types de colonnes, statistiques descriptives.
- Repérer les premiers problèmes : valeurs manquantes, valeurs aberrantes, colonnes mal typées.

Commentaire :

Ce TP correspond à la séance 1. Les étapes suivantes de la préparation des données (visualisation, nettoyage, encodage, normalisation, séparation des jeux) seront traitées lors de la séance 2.

Outils et Environnement Le TP a été réalisé à l'aide des outils et environnements suivants :

- **Langage** : Python 3.x
- **Bibliothèques** : Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- **IDE** : Jupyter Notebook / VS Code / Google Colab

Une des solutions consiste à installer anaconda <https://www.anaconda.com/download/success>. Une fois anaconda installé :

```
source ~/anaconda3/bin/activate
```

Créer un nouvel environnement nommé ml :

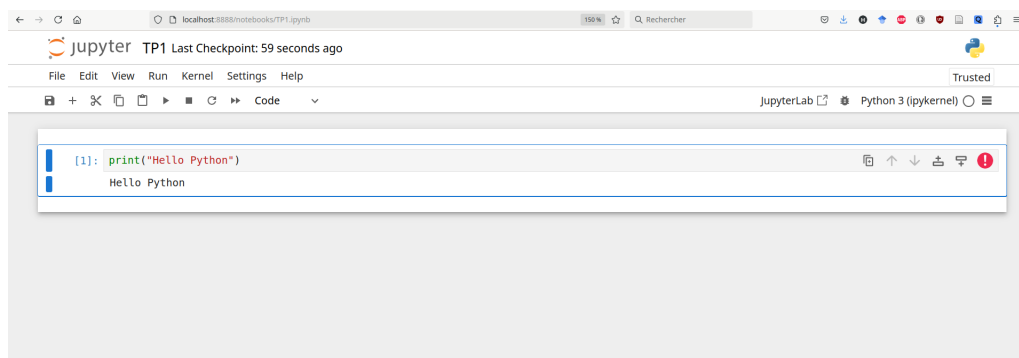
```
conda create --name ml
conda activate ml
(ml) conda install jupyter numpy scikit-learn matplotlib seaborn pandas
```

Lancer jupyter notebook :

```
cd /my/workspace
jupyter notebook
```

Dans un navigateur entrez l'adresse suivante :

<http://localhost:8888/tree>



1 Exercice 1 : Le dataset *California Housing*

Exercice inspiré du livre d'Aurélien Géron intitulé "Machine Learning avec Scikit-Learn", Chapitre 2 - 'Un projet de Machine Learning de bout en bout'.

1.1 Importer les Données

Utilisez la bibliothèque **Scikit-learn** pour charger le dataset California Housing et afficher les 5 premières lignes.

1.2 Examiner la Structure des Données

1. Affichez les informations générales sur les colonnes.
2. Expliquez le contenu de chaque colonne et indiquez dans quelle unité elle est exprimée.
3. Listez les types de données et les valeurs manquantes.

1.3 Utiliser le vrai dataset

Dans cette section, au lieu d'utiliser le dataset présent directement sur Scikit-Learn, nous allons utiliser le dataset original.

Téléchargez et extrayez le fichier *housing.tgz* sur le lien suivant :

<https://raw.githubusercontent.com/ageron/handson-ml/master/datasets/housing/housing.tgz>

1.3.1 Étudier la différence

Examinez le nouveau dataset et déterminez les différences entre les deux versions.

Question :

Quelles sont les principales différences entre le dataset sklearn et le dataset original ?

2 Pour aller plus loin : le dataset Iris

À traiter si vous avez terminé l'exercice 1. Ce dataset nous servira de fil rouge pour la classification, à partir de la séance 4.

2.1 Importer et examiner

Chargez le dataset Iris depuis Scikit-learn, construisez un **DataFrame** et répondez aux questions suivantes.

Question :

Combien y a-t-il d'observations et de caractéristiques ? Quels sont les noms des colonnes et que représente la colonne cible ?

Question :

Le dataset contient-il des valeurs manquantes ? Les trois espèces sont-elles représentées en proportions égales ?