

Data Mining et Machine Learning

IRC

John Samuel

CPE Lyon

Année: 2024-2025

Courriel: john.samuel@cpe.fr



John SAMUEL



- **Enseignant-Chercheur**, Conception Logicielle et Big Data, CPE Lyon,
- **Intérêts et thèmes de recherche :**
Représentation de connaissances, le web sémantique, les services web, l'intégration de données, l'entrepôt de données, les systèmes distribués, système d'information géographique
- **Cours** : Programmation en C, Algorithmes en C, Data Mining et Machine Learning, Intelligence Artificielle et Deep Learning, Systèmes d'exploitation et Programmation Concurrente, Langages Web
- **Thèse** : Intégration des données issues de services web

Objectifs

- Maîtriser les techniques de **représentation, manipulation et prétraitement** des données pour en optimiser l'utilisation.
- Appliquer des méthodes avancées de **traitement des données** pour extraire des informations pertinentes et exploitables.
- **Construire des modèles de traitement** par apprentissage machine afin d'analyser et de prédire des tendances à partir de données.
- Intégrer les **données ouvertes liées** dans vos analyses pour enrichir vos résultats.

Composition du module:

- Cours: 12h
- Travaux pratiques et projet: 16h

Environnement de

- Système d'exploitation: Linux
 - Ubuntu
 - Machine virtuelle (VirtualBox) + Ubuntu
- Éditeur: Jupyter
- Logiciels: pandas, matplotlib, scikit-learn

Cours:

- Interactifs
- Les questions: chaque 20-30 mins
- Devoir surveillé: 60%

Travaux pratiques et Projet

- Projet: 40%
- 3 travaux pratiques et projet
- Programmation en binôme
- Soumission en ligne

Devoir surveillé (DS): 60%

- **Examen:** En-ligne sur E-campus
- **Durée:** 2 heures
- **Total:** 20 points
- **Documents:** autorisés
- **Types de documents autorisés:** tous les documents autorisés
- **Calculatrices :** non autorisées
- **Utilisation de l'internet :** non autorisée
- **Dépôt supplémentaire :** disponible pour les fichiers personnels

Vous recevrez un courrier détaillé avant l'examen

Travaux pratiques et projet

- Le projet sera **évalué**.
- La **dates limite** de soumission est précisée sur e-campus.
- Il est fortement conseillé de **travailler en binôme** pour favoriser la collaboration et l'efficacité.

Data Mining

Cours	Dates
Cours 1 (4h)	11 février
Cours 2 (4h)	13 février
Cours 3 (4h)	25 février

Data Mining

Travaux pratiques

Dates

TP 1

12 février

TP 2 et Projet

24 février

TP 3 et Projet

26 février

TP 4 et Projet

27 février

Travaux pratiques

- Exploration approfondie de **Jupyter**, une plateforme interactive prisée pour l'analyse de données. Création de notebooks interactifs, intégration de code et de visualisations pour une analyse interactive des données.
- Expérience pratique avec des jeux de **données ouvertes**, permettant une compréhension concrète des enjeux liés au traitement de données. Application de techniques avancées pour extraire des insights significatifs à partir de données hétérogènes.



Soumission: Travaux pratiques et Projet

TP	Points
TP 1	✗
TP 2	✗
TP 3	✗
Projet	✓ (20 points)

- Les TP, notés de 1 à 3, ne sont pas soumis à évaluation. Cependant, la participation active et l'assiduité sont fortement encouragées pour une meilleure compréhension des concepts enseignés.
- L'évaluation du projet compte pour 20 points.

Travaux pratiques

Chaque TP comporte plusieurs exercices. Chaque exercice est accompagné d'une indication de **niveau de difficulté** :

- ★: Facile
- ★★: Difficulté moyenne
- ★★★: Difficile

Liste de contrôle

Avant de déposer votre projet, vérifiez si vous respectez la liste de contrôle suivante:

- Les noms complets (prénom et noms) de la binôme sont correctement inclus dans le fichier CONTRIBUTORS.md.
- Le fichier README.md est rempli de manière exhaustive et conforme aux instructions fournies.
- Vous avez respecté les noms de fichiers tels que spécifiés dans chaque exercice des travaux pratiques.
- Votre code est accompagné de commentaires appropriés pour expliquer la logique et la fonctionnalité.
- Votre code peut être exécuté sans générer d'erreurs, et dans la mesure du possible, sans générer d'avertissements.
- Le rapport en format PDF (5 pages maximum, Arial 11pt)
- **⚠ Pas de fichiers Python (*.py)**

Modèle de code

Vous pouvez consulter <https://github.com/johnsamuelwrites/DataMining> en ligne ou le cloner sur votre machine à l'aide du terminal en utilisant les commandes suivantes.

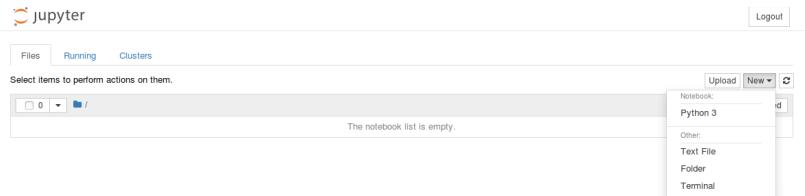
```
$ git clone https://github.com/johnsamuelwrites/MachineLearning
```

```
$ cd MachineLearning ; ls
```

Et pour les dernières modifications:

```
$ git pull
```

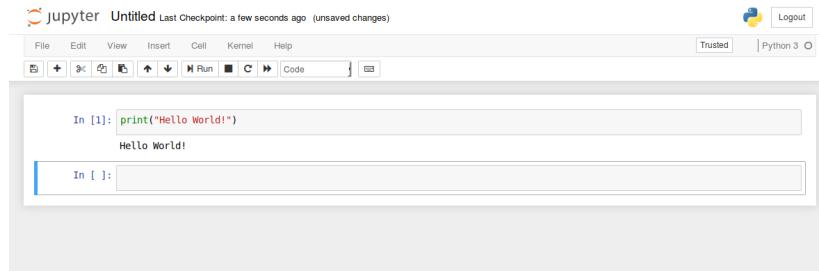
Travaux pratiques: Notebooks Jupyter



- Lorsque vous créez un nouveau notebook (onglet "New"), vous aurez la possibilité de choisir parmi plusieurs kernels.
- Assurez-vous de sélectionner le kernel "Python 3" ou "Python 3+" pour ce TP.

Data Mining: Notebooks Jupyter

Travaux pratiques: Notebooks Jupyter



- **Découpage structuré** : Divisez le code en cellules, facilitant la compréhension, la modification, et le débogage par étapes.
- **Interactivité** : Exécutez du code Python directement dans le notebook, permettant une exploration interactive des données.
- **Avantages - Communication complète** : Combine explications, visualisations, et code dans un seul document, facilitant la compréhension et le partage avec d'autres.

Data Mining: Notebooks Jupyter

Travaux pratiques: notebook Jupyter

The screenshot shows a Jupyter Notebook interface with the title "jupyter MyNotebook Last Checkpoint: an hour ago (unsaved changes)". The top menu includes File, Edit, View, Insert, Cell, Kernel, Help, Logout, Trusted, and Python 3. Below the menu are standard notebook controls: New, Open, Save, Run, Cell, Code, and Help.

Cell In [2] contains the code `print("Hello World!")`, which has been run and produced the output "Hello World!".

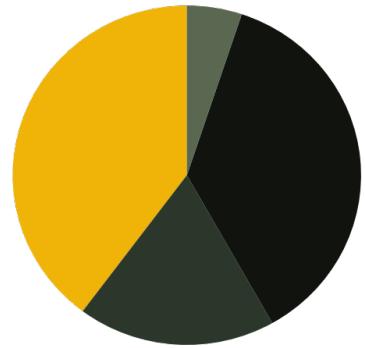
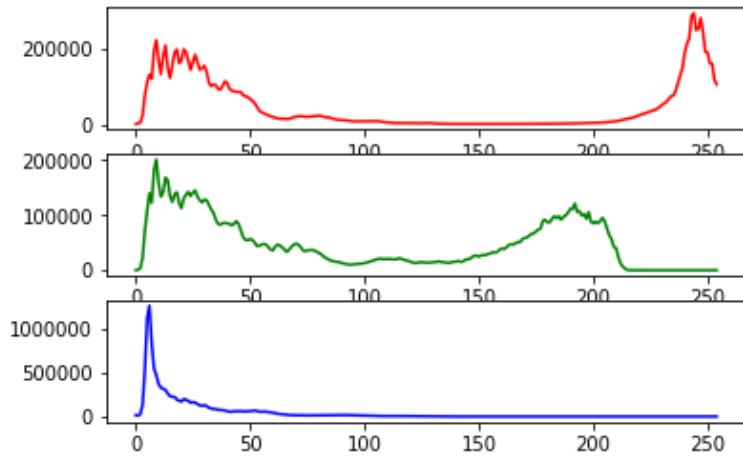
Cell In [14] contains the code `from sklearn import datasets
print(datasets.__all__)`. This has run and displayed a long list of available datasets from the scikit-learn library, including "load_boston", "load_diabetes", "load_digits", "load_files", "load_iris", "load_breast_cancer", "load_linnerud", "load_milcomp", "load_sample_image", "load_svmlight_file", "make_blobs", "make_circles", "make_classification", "make_gaussian_quantiles", "make_low_rank_matrix", "make_moons", "make_multilabel_classification", "make_regression", "make_s_curve", "make_sparse_coded_signal", "make_sparse_spd_matrix", "make_sparse_uncorrelated", "make_spd_matrix", "make_swiss_roll", and "mlrdata_filename".

Cell In [] is currently empty.

- Explorer l'utilisation de la bibliothèque scikit-learn pour la mise en œuvre de tâches d'apprentissage machine à l'aide de jeux de données intégrés.

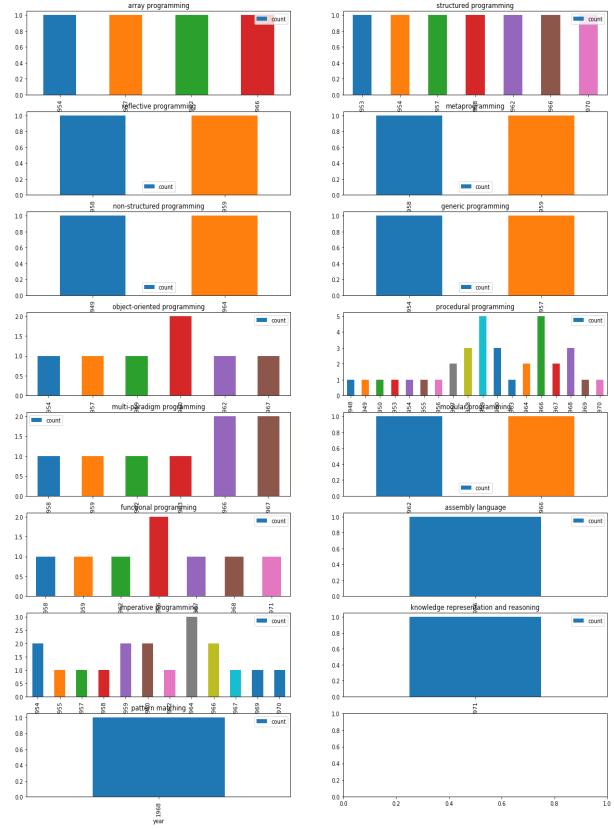
Data Mining: Notebooks Jupyter

Travaux pratiques: Visualisation et notebook Jupyter



Data Mining: Notebooks Jupyter

Travaux pratiques: Visualisation et notebook Jupyter



Data Mining

Travaux pratiques: Wikidata (Open Data)

The screenshot shows the Wikidata Query Service interface. On the left, there's a sidebar with various icons for navigation. The main area has tabs for "Wikidata Query", "Examples", "Help", and "Tools". The "Wikidata Query" tab is active. A "Query Helper" panel on the left shows a filter set to "instance of programming language". Below it, there's a "Show" button and a "Limit 100" option. The main query editor window contains the following SPARQL code:

```
1 SELECT ?languageLabel (YEAR(?inception) as ?year)
2 WHERE
3 {
4   #instances of programming language
5   ?language wdt:P31 wd:Q9143;
6     wdt:P571 ?inception;
7     rdfs:label ?languageLabel.
8   FILTER(lang(?languageLabel) = "en")
9 }
10 ORDER BY ?year
11 LIMIT 100
```

Below the query, the results are listed under the heading "languageLabel". The results are:

- ENIAC coding system
- ENIAC Short Code
- Von Neumann and Goldstine graphing system

At the bottom right, there are download options: "Code", "Download", and "Link". The "Download" dropdown menu is open, showing options for "JSON file", "JSON file (verbose)", "TSV file", "TSV file (verbose)", and "CSV file".

Sites web

- <https://jupyter.org/>
- <https://www.wikidata.org/>

Couleurs

- Color Tool - Material Design

Images

- Wikimedia Commons