

**Département Informatique**



## 

## TP2 – Machine Learning

## Prétraitement et visualisation

## de données

**Responsable du module :**

**Prof. OUANAN Mohamed**

**ZEKKOURI Hassan**

**Vendredi 06 Decembre 2019**



## REMERCIEMENT

Nous tenons à vous remercier monsieur Mohamed OUANAN pour votre formation et vos services.

Nous sommes également reconnaissant de nous avoir donner l’occasion de s’ouvrir sur un nouvel aspect de technologie qui est le Machine Learning et y mettre en œuvre nos compétences dont nous avons obtenu au cours du module.

Hassan ZEKKOURI

**PLAN**

Contents

[TP2 – Machine Learning 1](#_Toc26795115)

[Prétraitement et visualisation 1](#_Toc26795116)

[de données 1](#_Toc26795117)

[REMERCIEMENT 2](#_Toc26795118)

[INTRODUCTION 4](#_Toc26795119)

[I. Scikit-learn 5](#_Toc26795120)

[II. Jeux de données en Scikit-Learn 5](#_Toc26795121)

[III. TP1/TP2 5](#_Toc26795122)

[IV. Jupyter notebook | TP2 6](#_Toc26795123)

[Partie 1 : Imputation de données manquantes 6](#_Toc26795124)

[Partie 2 : Manipuler les variables catégoriques dans un jeu de données 6](#_Toc26795125)

[Partie 3 : Normalisation de données 6](#_Toc26795126)

[**FIN** 6](#_Toc26795127)

# INTRODUCTION

* Objectif :

Dans ce TP nous allons découvrir le module scikit-learn qui spécialisé en Machine Learning et on va essayer de découvrir et tester ses fonctionnalités !

# Scikit-learn

Scikit-learn est un logiciel écrit en Python, qui nécessite l'installation préalable du langage Python et des librairies NumPy1 et SciPy2 (pour le calcul scientifique), dans des versions qui doivent vérifier certaines contraintes de compatibilité. Le plus simple est d'installer une distribution de Python complète, comme Anaconda3, qui comprend la plupart des librairies courantes développées en Python, dont les trois citées plus haut. Le site officiel du logiciel Scikit-learn est :

<http://scikit-learn.org/stable/index.html>

La documentation en ligne est complète, et devra être consultée chaque fois que nécessaire :

<http://scikit-learn.org/stable/documentation.html>

Des tutoriaux sont disponibles à l'adresse suivante :

<http://scikit-learn.org/stable/tutorial/index.html>

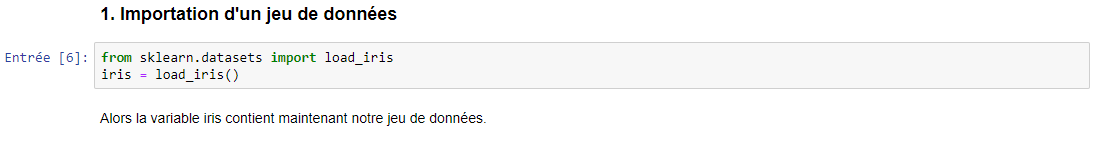
# Jeux de données en Scikit-Learn

Un certain nombre de jeux de données sont disponibles dans scikit-learn. Il est également possible de générer des données artificielles ou de récupérer des données externes.

Documentation relative au chargement de jeux de données :

<http://scikit-learn.org/stable/datasets/>

Les jeux de données disponibles dans scikit-learn sont : iris, boston, diabetes, digits, linnerud, sample images, 20newsgroups. Chacun de ces jeux de données se récupère à l'aide de la commande load\_nom-jeu qu'il faut dans un premier temps charger. Par exemple, pour récupérer le jeu iris :



# TP1/TP2

Pour bien documenter le travail au fur et à mesure nous avons choisie de travailler avec l’utile jupyter-notebook (IDE).

Tout le travail réalisé peut être consulté en ligne sur :

[https://notebooks.azure.com/DgrinderHZ/projects/machine-learning](https://notebooks.azure.com/DgrinderHZ/projects/machine-learning/tree/Master%20MST%20SIDI%20-%20FST%20Errachidia)

il suffit de l’ouvrir sur Azur Microsoft ou la télécharger pour Anaconda Jupyter.

# Jupyter notebook | TP2

### Partie 1 : Imputation de données manquantes

Lien de notebook :

<https://machinelearning-dgrinderhz.notebooks.azure.com/j/notebooks/Master%20MST%20SIDI%20-%20FST%20Errachidia/TP2/Machine%20Learning%20-%20TP2%20-%20Partie%201%20.ipynb>

### Partie 2 : Manipuler les variables catégoriques dans un jeu de données

Lien de notebook :

<https://machinelearning-dgrinderhz.notebooks.azure.com/j/notebooks/Master%20MST%20SIDI%20-%20FST%20Errachidia/TP2/Machine%20Learning%20-%20TP2%20-%20Partie%202.ipynb>

### Partie 3 : Normalisation de données

Lien de notebook :

<https://machinelearning-dgrinderhz.notebooks.azure.com/j/notebooks/Master%20MST%20SIDI%20-%20FST%20Errachidia/TP2/Machine%20Learning%20-%20TP2%20-%20Partie%203.ipynb>

# **FIN**

**Merci**

**données stocke toutes les données, si le fichier ne peut pas être ouvert lorsque le programme est en cours d'exécution, vous pouvez envisager de déplacer l'emplacement de ce dossier parameter.h définit le chemin du fichier, les paramètres de l'arborescence R, etc. Lors de la modification du chemin du fichier ou des paramètres de l'arborescence R, il vous suffit de modifier ce fichier imagedatabase.h Base de données d'images, qui doit appeler init et knnSearch lors de l'interface graphique. Un exemple est fourni dans knnQueryTest dans imagedatabasetest.cpp.**

**RTree.h L'implémentation de bas niveau de l'arborescence R. Vous pouvez ignorer ce fichier lorsque vous créez une interface graphique.**