Data Mining

Course title - Intitulé du cours	Data Mining
Level / Semester - Niveau /semestre	M2 / S1
School - Composante	Ecole d'Economie de Toulouse
Teacher - Enseignant responsable	GIL CASALS Silvia et HALFORD Max
Lecture Hours - Volume Horaire CM	30h
TA Hours - Volume horaire TD	0
TP Hours - Volume horaire TP	0
Course Language - Langue du cours	Anglais
TA and/or TP Language - Langue des TD et/ou TP	Anglais

Teaching staff contacts - Coordonnées de l'équipe pédagogique :

Silvia Gil Casals silvia.gil.casals@gmail.com, Max Halford maxhalford25@gmail.com

Several means of interaction are possible: after the classes, by email, on Zoom.

Course's Objectives - Objectifs du cours :

The students will be presented with several concepts on the of basis lectures notes. They will then have to solve lab exercises using Python, on some real datasets. They will learn about supervised learning, anomaly detection, text processing, unsupervised learning, structure decomposition, feature engineering, data visualisation, and data manipulation in general.

Students (in groups of 4 students) will choose a topic of interest among several modern topics of data mining. This choice will be made during the first lecture with Max Halford. Some starting points will be provided. Students are expected to write a report and make an oral defence in order to present the project to the other students.

Prerequisites - Prérequis :

Probability and Statistics as taught in the first year of Master in Econometrics and Statistics. The more knowledge of Python, the better.

Practical information about the sessions - Modalités pratiques de gestion du cours :

During the computer lab sessions, the students can bring their own laptop or tablet or use the computers in the room. We will be leveraging GitHub Codespaces for getting access to a decent Python environment.

In order to respect their teacher and classmates, the students are expected not to be more than 5 minutes late.

Grading system - Modalités d'évaluation :

Please find below the grading details for the data mining course:

- 6 points will come from short quizzes at the start of each session
- 14 points will come from the project

Bibliography/references - Bibliographie/références :

- Bilodeau, M. and Brenner, D. (2008), Theory of multivariate statistics. Springer Science & Business Media.
- Izenman, A. (2009). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer Texts in Statistics. Springer New York.
- Jolliffe, I. (2013). Principal Component Analysis. Springer Series in Statistics. Springer New York.
- Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning series. MIT Press.
- Murphy, K. P. (2022). Probabilistic Machine Learning: An introduction. MIT Press.

Some printed lecture notes will be provided with more references available at the library at the Manufacture.

Session planning - Planification des séances :

- 3 sessions with Silvia Gil Casals on neural networks, support vector machine and one-class support vector machine.
- 7 sessions with Max Halford on finding structure in a dataset, clustering, feature engineering, supervised learning, text processing, and data visualisation.

Distance learning – Enseignement à distance :

Distance learning can be provided when necessary by implementing, for example: / En cas de nécessité, un enseignement à distance sera assuré en mobilisant, par exemple :

- Interactive virtual classrooms / Classe en ligne interactive
- Recorded lectures (videos) / Vidéo enregistrée de la présentation du matériel pédagogique
- MCQ tests and other online exercises and assignments / QCM et exercices en ligne
- Remote (online) tutorials (classes) / TP/TD à distance
- Chatrooms / Forums