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	RUIZ-GAZEN Anne
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 :

Aurore Archimbaud archimbaud@ese.eur.nl, Anne Ruiz-Gazen anne.ruiz-gazen@tse-fr.eu (T211), Silvia Gil Casals silvia.gil.casals@gmail.com

Several means of interaction are possible: after the classes, by email, on zoom or at the office of Anne Ruiz-Gazen with prior appointment.

Course's Objectives - Objectifs du cours :

The students are expected to understand the different notions by using some lectures notes, solving some exercises and implementing the methods with the softwares R and Python on some real data sets. More advanced methods in data mining will be covered at the end of the course through a project based learning approach. Students (in groups of 4 students) will choose a topic of interest among several modern topics of data mining. Some references will be provided and the students are expected to write a report and make an oral defense 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.

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.

In order to respect their teacher and class mates, 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:

- 4 points out of 20 on the computer lab session reports (0.5 per report, 4 reports with Aurore Archimbaud and 4 reports with Anne Ruiz-Gazen).

- 2 point out of 20 on exercises (2 series with Anne Ruiz-Gazen)

- 6 points out of 20 on the exam that will last 1.5 hour (end of the semester).

- 8 point out of 20 on the project (end of the semester).

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 Aurore Archimbaud on data visualization, outlier detection, and treatment of missing values
- 4 sessions with Anne Ruiz-Gazen on soft clustering and unsupervized and supervized dimension reduction
- 3 sessions with Silvia Gil Casals on neural networks, support vector machine and one-class support vector machine.

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