High-dimensional data analysis and machine learning

Course title - Intitulé du cours	High-dimensional data analysis and machine learning
Level / Semester - Niveau / semestre	M1 / S2
School - Composante	École d'Économie de Toulouse
Teacher - Enseignant responsable	MONDON Camille
Other teacher(s) - Autre(s) enseignant(s)	LAURENT Thibault
Lecture Hours - Volume Horaire CM	12
TA Hours - Volume horaire TD	0
TP Hours - Volume horaire TP	18
Course Language - Langue du cours	English - Anglais
TA and/or TP Language - Langue des TD et/ou TP	English - Anglais

Teaching staff contacts - Coordonnées de l'équipe pédagogique :

The course is divided in three parts. The first part (12 hours) and the second part (12 hours) are taught by Camille Mondon (PhD student), contact by email (camille.mondon@tse-fr.eu). The third part (6 hours) is taught by Thibault Laurent (research engineer), contact by email (thibault.laurent@tse-fr.eu).

Preferred means of interaction: email, prior appointment.

Course Objectives - Objectifs du cours :

This course is particularly relevant for students who are interested in pursuing their studies and career as data scientists. It does not contain advanced theory, but all methods and algorithms are described and implemented using *R* and results are analyzed in detail. The course is not difficult but requires much work throughout the semester and **attendance is required on March 3**.

The students are expected to develop skills in computational statistics and to be able to combine efficient programming with relevant statistical methods. The class will therefore include a large amount of practical applications. The course is divided into three parts.

The first part presents principal components analysis, clustering methods in the unsupervised context, and discriminant analysis and classification and regression tree procedure for supervised problems. All methods are implemented using R.

The second part of the class describes methods for supervised classification and regression problems. The course mainly covers the bootstrap and bagging approaches including random forests.

The third part focuses on an introduction to parallel computing to deal with big data.

Prerequisites - Pré-requis :

Proficient R programming, knowledge of descriptive statistics and principal components analysis.

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

For each of the first two parts, there are 4 weekly sessions of 3 hours. The slides are made available to students, but it is highly recommended not to miss any session in order to be able to implement the statistical methods and interpret the results. The third part consists in 2 sessions of 3 hours.

Personal laptops are accepted at the student's own risk (some sessions take place in a computer room). Students are expected to actively participate to the class. Late arrivals or missing students will be reported and can result in a grade penalty.

Grading system - Modalités d'évaluation :

The first two parts (eight lectures) are evaluated through some Multiple Choice Questionnaires (30%; during the lectures), some assignments (10%; between the lectures), and a project (60%; after these lectures).

Bibliography/references - Bibliographie/références :

Everitt, Brian, and Torsten Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with R*. New York, NY: Springer. DOI:10.1007/978-1-4419-9650-3.

Husson, François, Sebastien Lê, and Jérôme Pagès. 2017. *Exploratory Multivariate Analysis by Example Using R*. 2nd ed. New York: Chapman and Hall/CRC. DOI:10.1201/b21874.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. New York, NY: Springer US. DOI:10.1007/978-1-0716-1418-1.

Williams, Graham. 2011. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. New York, NY: Springer. DOI:10.1007/978-1-4419-9890-3.

Session planning - Planification des séances :

Mondays from January 6 to March 24 between 15:30 and 18:30.

Distance learning - Enseignement à distance :

Distance learning can be provided, when necessary, by implementing interactive virtual classrooms, MCQ tests and other online exercises / assignments, chatrooms and forums.