# Data Mining

Nom : **EUDES** Prénom : **Richard** Année : **M2** Semestre : **10**

Nature : **CM** Volume horaire : **24 H** ECTS / Coef : **2**

| | |
|---|---|
| Prérequis | Une connaissance de niveau M1 des statistiques et de l'optimisation (recherche opérationnelle). Une expérience antérieure du logiciel SAS est utile (la programmation SAS base est un plus). |

| | |
|---|---|
| Résumé | Ce cours couvre les compétences attendues d'un data miner / data scientist pour structurer un protocole d'analyse de données à l'aide de workflows (SAS Enterprise Miner) dans la création de modèles non supervisés (segmentations, analyses d'association et de séquence) et supervisés (arbre de décision, régression et modèles de réseaux neuronaux). Il résume l'ensemble des étapes statistiques nécessaires à la préparation de données et à la construction de modèles pertinents et interprétables. En outre, ce cours permet d'introduire des éléments de gestion de projets propres aux démarches analytiques modernes (CRISP, SEMMA). |

| | |
|---|---|
| Objectifs | - définir un projet et explorer les données graphiquement (dataviz)<br>- modifier les données pour de meilleurs résultats d'analyse (transformer les variables, les données manquantes,…)<br>- construire et comprendre des modèles prédictifs tels que des arbres de décision, des modèles de régression, des réseaux de neurones (et d'autres algorithmes comme SVM, PLS Regression, Knn, ..)<br>- comparer et expliquer des modèles complexes<br>- générer et utiliser le code de score (scoring).<br>- appliquer l'analyse d'association et de séquence aux données de transaction (techniques de clustering)<br>- utiliser d'autres outils de modélisation tels que l'induction de règles, le Gradient Boosting et les Support Vector Machines<br><br>Ce cours peut vous aider à vous préparer aux examens de certification suivants: Modélisation prédictive avec SAS Enterprise Miner. |

| | |
|---|---|
| Bibliographie | Beck, A. 1997. "Herb Edelstein discusses the usefulness of data mining." *DS Star.* Vol. 1, N0. 2. Available www.tgc.com/dsstar/.<br>Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. New York: Oxford University Press.<br>Breiman, L. et al. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.<br>Hand, D. J. 1997. *Construction and Assessment of Classification Rules*. New York: John Wiley & Sons, Inc.<br>Hand, D. J. 2005. "What you get is what you want? – Some dangers of black box data mining."<br>*M2005 Conference Proceedings*, Cary, NC: SAS Institute Inc.<br>Hand, D. J. 2006. "Classifier technology and the illusion of progress." *Statistical Science* 21:1-14.<br>Hand, D. J. and W. E. Henley. 1997. "Statistical classification methods in consumer credit scoring:<br>a review." *Journal of the Royal Statistical Society A* 160:523-541.<br>Hand, David, Heikki Mannila, and Padraic Smyth. 2001. *Principles of Data Mining*. Cambridge, Massachusetts: The MIT Press.<br>Harrell, F. E. 2006. *Regression Modeling Strategies*. New York: Springer-Verlag New York, Inc. |

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag New York, Inc.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley & Sons, Inc.

Kass, G. V. 1980. "An exploratory technique for investigating large quantities of categorical data." *Applied Statistics* 29:119-127.

Mosteller, F. and J. W. Tukey. 1977. *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

Piatesky-Shapiro, G. 1998. "What Wal-Mart might do with Barbie association rules." *Knowledge Discovery Nuggets*, 98:1. Available http://www.kdnuggets.com/.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. New York: Cambridge University Press.

Rubinkam, M. 2006. "Internet Merchants Fighting Costs of Credit Card Fraud," *AP Worldstream*. The Associated Press.

Rud, Olivia Parr. 2001. *Data Mining Cookbook: Modeling Data, Risk, and Customer Relationship Management*. New York: John Wiley & Sons, Inc.

Sarle, W. S. 1983. *Cubic Clustering Criterion*. SAS Technical Report A-108. Cary, NC: SAS Institute Inc.

Sarle, W. S. 1994a. "Neural Networks and Statistical Models," *Proceedings of the Nineteenth Annual SAS® Users Group International Conference*. Cary: NC, SAS Institute Inc., 1538-1550.

Sarle, W. S. 1994b. "Neural Network Implementation in SAS® Software," *Proceedings of the Nineteenth Annual SAS® Users Group International Conference.* Cary: NC, SAS Institute Inc., 1550-1573.

Sarle, W. S. 1995. "Stopped Training and Other Remedies for Overfitting." *Proceedings of the 27th Symposium on the Interface*.

SAS Institute Inc. 2002. *SAS® 9 Language: Reference, Volumes 1 and 2*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2002. *SAS® 9 Procedures Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2002. *SAS/STAT® 9 User's Guide, Volumes 1,2, and 3*. Cary, NC: SAS Institute Inc.

Weiss, S. M. and C. A. Kulikowski. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Mateo, CA: Morgan Kaufmann.

# PLAN

Le support de cours étant en anglais, il se structure de la manière suivante :