Introduction to Data Mining (3673-V), 8 cr – fall 2021

Course description	Data mining is the process of exploring large amounts of data to discover meaningful patterns or rules, a process becoming more and more important with the rapid growth of available information. The aim of the course is to provide an introduction to data mining techniques, focusing both on theory and practical applications. The course covers common methods for classification, prediction and clustering. We will look at methods such as Classification and Regression Trees, k-Nearest-Neighbors, Neural Networks, Hierarchical and Non-hierarchical Clustering. R and RStudio are used in the course. The course alternates and is given every other year
Learning goal	You are familiar with the field of data mining and have good knowledge of common data mining techniques and practices.
After completing the course, you will be able to	• explore, clean, and preprocess data; • make necessary data reductions, and partition data into training, validation and test datasets; • evaluate the predictive performance of data mining methods; • choose and apply appropriate data mining technique to well-defined study or research questions; • interpret and utilize results from data mining projects, and present the results in written reports; • analyze and evaluate data mining research reports.
Pre-requisites	The equivalent of a first course in statistical analysis (e.g. 7777 Introduction to Research methods, 7778 Statistical Analysis)
Total student workload	214 hours divided into Scheduled (contact) hours: 40 h Non-scheduled work: 174 h Contact hours: Lectures, exercises & computer labs (40 h). Self study: Preparing for the weekly lectures (54 h). Assignments (40 h). Article reviews (40 h). Final exam (40 h).
Literature and course material	Shmueli, G., Patel N.R. & Bruce P.C. (2016). Data Mining for Business Analytics. Third edition, Hoboken, NJ: Wiley James, G., Witten, D, Hastie, T. & Tibshirani, R. (2017) An introduction to statistical learning. First edition, New York: Springer. Selected chapters as specified by the instructor.
Assessment	Final exam (online, open book) 40%, assignments (30%) and article reviews (30%). Exam dates: 17.12.2021 and 10.1.2022
Teacher	Christian Johansson

Introduction to Data Mining (3673-V), 8 cr – fall 2021

Schedule	All lectures online (Teams) Monday, October 25th 8.30 – 10.00: Introduction (chapter 1 in Data Mining for Business Analytics) Wednesday, October 27th 8.30 – 10.00: Overview / data visualization / dimension reduction (chapters 2, 3, 4.1 – 4.7, 5.5) Wednesday, November 3rd 8.30 – 10.00: Linear regression / performance evaluation (chapters 6, 5.1 – 5.2, 4.9) Wednesday, November 10th 8.30 – 10.00: Logistic regression / performance evaluation (chapters 10, 5.3 – 5.4, 4.9) Wednesday, November 17th 8.30 – 10.00: Classification and regression trees (chapters 9.1 – 9.7, 9.9, 4.10) Wednesday, November 24th 8.30 – 10.00: Neural nets (chapter 11) Wednesday, December 1st 8.30 – 10.00: k-Nearest Neighbors and ensembles (chapters 7, 13.1, 9.8) Wednesday, December 8th 8.30 – 10.00: Cluster analysis (chapter 15)