University of Reading

MSc Data Science and Advanced Computing

CSMDM21 - Data Analytics and Mining

Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

Lecture notes and videos powered by
Prof. Giuseppe Di Fatta

CSMDM21 - Data Analytics and Mining

❑ CSMDM21 is a 20-credit module (for a total of 200 hours)

- https://www.reading.ac.uk/module/document.aspx?modP=CSMDM21&modYR=2122
- ~20 hours of lectures
- ~10 hours of interactive practical sessions (online and in a PC lab):
  - please see the timetable for time and location (online vs PC lab)
- ~170 hours of guided independent study

❑ Course Assessment:

- **100%** based on coursework (1 major assignment)
  - Assignment release date: see Blackboard
  - Coursework submission deadline: see Blackboard for exact date/time
    - electronic submission in Bb: one zip archive containing your report (pdf) and your solution (data workflow and code)

CSMDM21 - Data Analytics and Mining

❑ Recommended textbook:

- See Blackboard (Bb) for the recommended textbook and further books, some including links for online access to them.
- "Introduction to Data Mining", Tan, Steinbach, Kumar, Addison-Wesley
  - https://www-users.cs.umn.edu/~kumar001/dmbook/index.php
  - Some lecture slides of this course are based on the recommended textbook and its instructor resources. These are provided solely for the use of teaching this courses.

❑ **Adopted tools** (free and open source)

- **KNIME**, a data science and machine learning platform (https://www.knime.com)
- the **R** programming language (https://www.r-project.org) and RStudio (https://rstudio.com)

These are all preinstalled in the lab PCs: https://www.reading.ac.uk/internal/its/AppsAnywhere.aspx

➢ You should also download and install their latest version in your personal computer or laptop.

Resources

- Blackboard is the content management system adopted at the University of Reading:
  - Information on the module (incl. roadmap and schedule)
  - Information on the lecturer
  - Video lectures, handouts and other learning resources
  - Link to external resources (free and open source software)
  - Weekly practical assignments
  - Major coursework assignment and submission point
  - Your grade and feedback

- Please refer to the schedule to learn what is happening every week
  - schedule of topics
  - schedule of practical work
  - some interactive sessions are run online and some in a PC lab

CSMDM21: Aims, Objectives and Outcomes

- **Aims**:

  Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories. Automated **Data Analytics and Mining** techniques are becoming essential components to any information system. Typically data have to be cleaned, pre-processed, selected, merged, etc., and finally processed for the extraction of knowledge in terms of data models.

- **Objectives**:

  This module introduces concepts, methodologies, algorithms and tools for the design of data science workflows. In particular, the language R and the data science platform KNIME are introduced and adopted for the hands-on activities. Student will learn general Data Science principles and Data Mining algorithms and will apply them in different application domains.

- **Assessable learning outcomes:**

  Students are expected to understand the general **Knowledge Discovery in Data** process, various Data Mining algorithms and the two specific tools (KNIME and R) adopted in this module. Students will be able to use these tools to apply the methods to generate data models. Students will be able to evaluate the quality of the data models and compare them. During practical activities as well as for the assessment students will adopt state-of-the-art tools for implementing data analytics and mining solutions in different applicative domains.

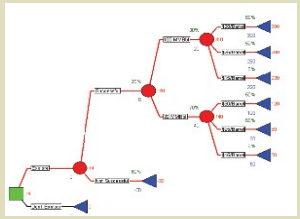
Schedule (download the PDF file from Blackboard)

Type	Mode
L	Pre-recorded Lecture Videos - Watch the videos any time before the practical session
P	Interactive Practical Sessions (f2f in Polly Vacher G45) - Attend the session: follow the tutorial PDF file and carry out the exercises

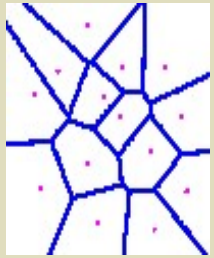
CSMDM16 / CSMDM21 Data Analytics and Mining - Autumn Term 2021

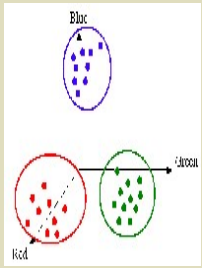
Week	Type	Topic	Date	Time	Assessment
1	P	Introduction to this module	Sept 28 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	
1	L	Introduction to Data Mining and Data Science Platforms Introduction to Data and Data Preprocessing			
2	L	Introduction to KNIME			
2	P	P01: practical on KNIME Basics	Oct 5 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	Assignment release: 2021 Oct 5
3	L	Proximity Measures			
3	P	P02: practical on Data Transformation in KNIME	Oct 12 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	Suggest to work on Problem #1 task #1
4	L	Clustering			
4	P	P03: practical on Clustering in KNIME	Oct 19 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	Suggest to work on Problem #1 task #2
5	L	Classification and Model Evaluation			
5	P	P04: practical on Classification in KNIME and model evaluation	Oct 26 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	Suggest to work on Problem #2 task #4, 5 (KNIME)
6		No teaching week (catch up on lectures and exercises)			
7	L	Advanced KNIME, parameter optimisation, ensamble methods			
7	P	P05: practical on advanced KNIME	Nov 9 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	
8	L	Introduction to R			
8	P	P06: practical on R Basics	Nov 16 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	
9	L	Clustering and Classification in R			
9	P	P07: practical on model evaluation in R - part 1	Nov 23 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	Suggest to work on Problem #1 task #3
10	L	Association Rule Mining			
10	P	P08: practical on model evaluation in R - part 2	Nov 30 (Tue)	Group A: 9:00 - 10:00 Group B: 11:30 - 12:30	Suggest to work on Problem #2 task #4, 5 (R)
11		No teaching (catch up on lectures and exercises, work on Assignment) Suggest to work on Problem #2 task #6 and finalize the report			**Assignment deadline:** **2021 Dec 10 (Fri) 12:00 noon**

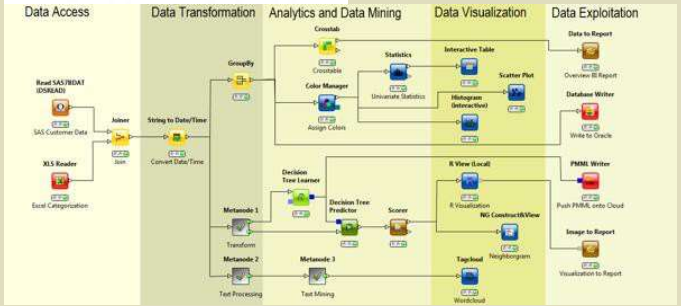
Outline

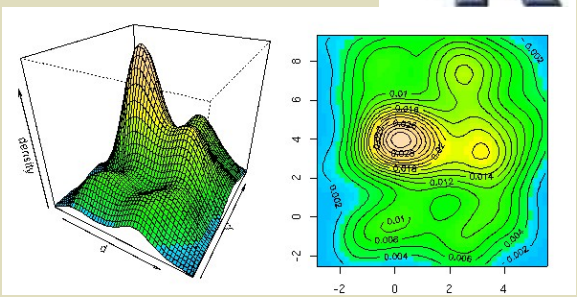


- Introduction to Data Science
  - Data and data preprocessing
- Data Mining tasks: Classification, Clustering, Association Rules
- KNIME: a Data Science and Machine Learning platform
- R: a programming language for computational statistics, Data Mining, visualisation









Next video lecture:

➢Introduction to Data Analytics and Mining