

**1. General information****Course:** DATA MINING**Type:** CORE COURSE**Degree:** 347 - DEGREE PROGRAMME IN COMPUTER SCIENCE ENGINEERING (CR)**Center:** 108 - SCHOOL OF COMPUTER SCIENCE OF C. REAL**Year:** 4**Main language:** Spanish**Use of additional languages:****Web site:** Available in Campus Virtual <https://campusvirtual.uclm.es>**Code:** 42348**ECTS credits:** 6**Academic year:** 2023-24**Group(s):** 20**Duration:** First semester**Second language:****English Friendly:** Y**Bilingual:** N**Lecturer:** FERNANDO GUALO CEJUDO - Group(s): 20

| Building/Office | Department | Phone number | Email | Office hours |
|-----------------|---------------------------------------|--------------|------------------------|--------------|
| | TECNOLOGÍAS Y SISTEMAS DE INFORMACIÓN | | Fernando.Gualo@uclm.es | |

Lecturer: ARTURO PERALTA MARTIN-PALOMINO - Group(s): 20

| Building/Office | Department | Phone number | Email | Office hours |
|------------------|---------------------------------------|--------------|------------------------|--------------|
| FERMIN CABALLERO | TECNOLOGÍAS Y SISTEMAS DE INFORMACIÓN | 926295300 | Arturo.Peralta@uclm.es | |

2. Pre-Requisites

We recommend students are familiar with Computer Science concepts, like those in the previous courses of the Degree Program. In addition, this subject is based on the skills and knowledge acquired in the following ones:

- Logic
- Statistics
- Algorithm Design
- Intelligent systems
- Knowledge Based Systems

3. Justification in the curriculum, relation to other subjects and to the profession

Data mining and machine learning are linked to the field of statistics and computer algorithms. They are based on techniques for the extraction of knowledge from data sets. In recent years, these disciplines are gaining importance due to the increase in data production -propitiated by phenomena such as the rise of the Internet or social networks- or the development of new techniques for obtaining genetic information. From a professional point of view, there is a rising demand for data scientists in fields as diverse as marketing, market analysis, security, or biology.

4. Degree competences achieved in this course**Course competences**

| Code | Description |
|--------|---|
| CM05 | Ability to acquire, formalise, and represent human knowledge in a computable form for the solution of problems throughout a digital system in any application context, especially the one linked to computational aspects, perception, and behaviour in intelligent frames. |
| CM7 | Ability to know and develop computational learning techniques, and design and implement applications and systems which could use them, including the ones for the automatic extraction of information and knowledge from great batches of information. |
| INS01 | Analysis, synthesis, and assessment skills. |
| INS04 | Problem solving skills by the application of engineering techniques. |
| INS05 | Argumentative skills to logically justify and explain decisions and opinions. |
| PER02 | Ability to work in multidisciplinary teams. |
| PER04 | Interpersonal relationship skills. |
| PER05 | Acknowledgement of human diversity, equal rights, and cultural variety. |
| SIS01 | Critical thinking. |
| SIS03 | Autonomous learning. |
| SIS09 | Care for quality. |
| UCLM03 | Accurate speaking and writing skills. |

5. Objectives or Learning Outcomes**Course learning outcomes**

Description

Development and implementation of a small to medium-sized information retrieval system.

Description and application of different phases of the discovery process of knowledge extraction from large volumes of data.

Knowledge and development of computational learning techniques, both supervised and unsupervised, and design and implement applications and systems that use them.

6. Units / Contents

Unit 1: Introduction

Unit 1.1 Artificial Intelligence, KDD and Data Mining

Unit 1.2 Intelligent Data Analysis in Big Data Environments

Unit 2: Exploratory Data Analysis

Unit 3: Data Mining Tasks

Unit 3.1 Clustering

Unit 3.2 Dimensionality Reduction

Unit 3.3 Association Rules Extraction

Unit 3.4 Anomaly Detection

Unit 3.5 Classification

Unit 3.6 Regression

Unit 4: Data Mining Applications

Unit 5: Study Cases

ADDITIONAL COMMENTS, REMARKS

LABORATORY

a complete KDD process will be developed throughout the course. The student will propose the domain linked to their interests of work and/or research or some topic proposed by the professors.

1. Problem Selection
2. Data Selection.
3. Pre-processing.
4. Transformation.
5. Data Mining.
6. Use of patterns discovered in an application.

7. Activities, Units/Modules and Methodology

| Training Activity | Methodology | Related Competences (only degrees before RD 822/2021) | ECTS | Hours | As | Com | Description |
|--|----------------------------------|--|----------|---|----|-----|--|
| Class Attendance (theory) [ON-SITE] | Lectures | CM05 CM7 | 0.6 | 15 | N | - | Teaching of the subject matter by lecturer (MAG) |
| Individual tutoring sessions [ON-SITE] | | CM05 CM7 INS05 SIS01 SIS09 UCLM03 | 0.18 | 4.5 | N | - | Individual or small group tutoring in lecturer's office, classroom or laboratory (TUT) |
| Study and Exam Preparation [OFF-SITE] | Self-study | CM05 CM7 INS01 SIS01 SIS03 SIS09 | 1.8 | 45 | N | - | Self-study (EST) |
| Other off-site activity [OFF-SITE] | Practical or hands-on activities | CM05 CM7 INS01 INS04 PER02 PER04 PER05 SIS03 SIS09 UCLM03 | 0.9 | 22.5 | N | - | Lab practical preparation (PLAB) |
| Problem solving and/or case studies [ON-SITE] | Problem solving and exercises | CM05 CM7 INS01 INS04 PER02 PER04 PER05 SIS01 SIS09 | 0.6 | 15 | Y | N | Worked example problems and cases resolution by the lecturer and the students (PRO) |
| Writing of reports or projects [OFF-SITE] | Self-study | CM05 CM7 INS01 INS04 INS05 PER02 PER04 PER05 SIS01 SIS03 SIS09 UCLM03 | 0.9 | 22.5 | Y | N | Preparation of essays on topics proposed by lecturer (RES) |
| Laboratory practice or sessions [ON-SITE] | Practical or hands-on activities | CM05 CM7 INS04 PER02 PER04 PER05 SIS03 SIS09 | 0.72 | 18 | Y | Y | Realization of practicals in laboratory /computing room (LAB) |
| Progress test [ON-SITE] | Assessment tests | CM05 CM7 INS01 INS04 INS05 PER02 SIS01 SIS09 UCLM03 | 0.1 | 2.5 | Y | N | Progress test 1 of the first third of the syllabus of the subject (EVA) |
| Progress test [ON-SITE] | Assessment tests | CM05 CM7 INS01 INS04 INS05 PER02 SIS01 SIS09 UCLM03 | 0.1 | 2.5 | Y | N | Progress test 2 of the two first thirds of the syllabus of the subject (EVA) |
| Progress test [ON-SITE] | Assessment tests | CM05 CM7 INS01 INS04 INS05 PER02 SIS01 SIS09 UCLM03 | 0.1 | 2.5 | Y | N | Progress test 3 of the complete syllabus of the subject (EVA) |
| Total: | | | 6 | 150 | | | |
| Total credits of in-class work: 2.4 | | | | Total class time hours: 60 | | | |
| Total credits of out of class work: 3.6 | | | | Total hours of out of class work: 90 | | | |

As: Assessable training activity

Com: Training activity of compulsory overcoming (It will be essential to overcome both continuous and non-continuous assessment).

| 8. Evaluation criteria and Grading System | | | |
|---|-----------------------|----------------------------|--|
| Evaluation System | Continuous assessment | Non-continuous evaluation* | Description |
| Progress Tests | 15.00% | 0.00% | Progress test 2 Non-compulsory activity that can be retaken. To be carried out at the end of the second third of the teaching period. |
| Progress Tests | 27.50% | 0.00% | Progress test 3. Non-compulsory activity that can be retaken. To be carried out during the non-teaching period |
| Theoretical papers assessment | 15.00% | 15.00% | Non-compulsory activity that can be retaken. To be carried out before end of teaching period |
| Progress Tests | 7.50% | 0.00% | Progress test 1. Non-compulsory activity that can be retaken (rescheduling). To be carried out at the end of the first third of the teaching period. |
| Laboratory sessions | 25.00% | 25.00% | Compulsory activity that can be retaken. To be carried out during lab sessions |
| Oral presentations assessment | 10.00% | 10.00% | Non-compulsory activity and can be retaken. The students in the continuous mode will be evaluated in theory/laboratory sessions The students of non-continuous mode will be evaluated from this activity through of an alternative system. |
| Final test | 0.00% | 50.00% | Compulsory and recoverable activity to be carried out on the date scheduled for the final ordinary examination . |
| Total: | 100.00% | 100.00% | |

According to art. 4 of the UCLM Student Evaluation Regulations, it must be provided to students who cannot regularly attend face-to-face training activities the passing of the subject, having the right (art. 12.2) to be globally graded, in 2 annual calls per subject, an ordinary and an extraordinary one (evaluating 100% of the competences).

Evaluation criteria for the final exam:

Continuous assessment:

In compulsory activities, a minimum mark of 40% is required in order to pass that activity and have the possibility to therefore pass the entire subject. A compulsory activity cannot be divided into eliminatory parts, nor can minimum marks be established for each of its parts. In the case of the activities that may be retaken (i.e., rescheduling), an alternative activity or test will be offered in the resit/retake exam call (convocatoria extraordinaria).

The progress tests will be common for all the theory/laboratory groups of the subject and will be evaluated by the lecturers of the subject in a serial way, i.e., each part of the progress tests will be evaluated by the same lecturer for all the students. A student is considered to pass the subject if she/he obtains a minimum of 50 points out of 100, taking into account the points obtained in all the evaluable activities, and also has passed all the compulsory activities.

For students who do not pass the subject in the final exam call (convocatoria ordinaria), the marks of activities already passed will be conserved for the resit/retake exam call (convocatoria extraordinaria). If an activity is not recoverable, its assessment will be preserved for the resit/retake exam call (convocatoria extraordinaria) even if it has not been passed. In the case of the passed recoverable activities, the student will have the opportunity to receive an alternative evaluation of those activities in the resit/retake exam call and, in that case, the final grade of the activity will correspond to the latter grade obtained.

The mark of the passed activities in any call, except for the progress tests, will be conserved for the subsequent academic year at the request of the student, provided that mark is equal or greater than 50% and that the activities and evaluation criteria of the subject remain unchanged prior to the beginning of that academic year.

The failure of a student to attend the progress test 3 will automatically result in her/him receiving a "Failure to attend" (no presentado). If the student has not passed any compulsory evaluation activity, the maximum final grade will be 40%.

Non-continuous evaluation:

Students may apply at the beginning of the semester for the non-continuous assessment mode.

In the same way, the student may change to the non-continuous evaluation mode as long as she/he has not participated during the teaching period in evaluable activities that together account for at least 50% of the total mark of the subject. If a student has reached this 50% of the total obtainable mark or the teaching period is over, she/he will be considered in continuous assessment without the possibility of changing to non-continuous evaluation mode.

Students who take the non-continuous evaluation mode will be globally graded, in 2 annual calls per subject, an ordinary and an extraordinary one (evaluating 100% of the competences), through the assessment systems indicated in the column "Non-continuous evaluation".

In the "non-continuous evaluation" mode, it is not compulsory to keep the mark obtained by the student in the activities or tests (progress test or partial test) taken in the continuous assessment mode.

Specifications for the resit/retake exam:

Evaluation tests will be conducted for all recoverable activities. Due to the nature of the progress tests, in the resit/retake exam (convocatoria extraordinaria) there will be a single progress test that includes the three progress tests. The failure of a student to attend this progress test will automatically result in her/him receiving a "Failure to attend" (no presentado), except in the case that the student conserves the mark for progress tests from the final exam call (convocatoria ordinaria). In the latter case, the student's carrying out of any other evaluable activity in the resit/retake exam call (convocatoria extraordinaria) will result in a numerical mark.

Specifications for the second resit / retake exam:

Same characteristics as the resit/retake exam call.

| 9. Assignments, course calendar and important dates | |
|---|-------|
| Not related to the syllabus/contents | |
| Hours | hours |
| General comments about the planning: The course is taught in three weekly sessions of 1.5 hours.T | |

| 10. Bibliography and Sources | | | | | | |
|------------------------------|------------|------------------|------|------|------|-------------|
| Author(s) | Title/Link | Publishing house | City | ISBN | Year | Description |

| | | | | | |
|---|---|----------------------------|---------------|------|---|
| Jan Van der Plass | Python Data Science Handbook | O'Reilly | 9781491912058 | 2016 | This accessible and classroom-tested textbook/reference presents an introduction to the fundamentals of the emerging and interdisciplinary field of data science. The coverage spans key concepts adopted from statistics and machine learning and the practical application of data science. |
| | https://learning.oreilly.com/library/view/python-data-science/9781491912126/ | | | | |
| Igual, Laura, Seguí, Santi | Introduction to Data Science | Springer | 9783319500171 | 2017 | |
| | https://link.springer.com/book/10.1007%2F978-3-319-50017-1 | | | | |
| Adriaans, P. W.; Zantinge, D. | Data Mining. | Addison-Wesley | | 1996 | |
| Berry, M. J. A.; Linoff, G. | Data Mining Techniques. | Wiley Computer Publishing. | New York | 1996 | |
| Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. | The KDD Process for Extracting Useful Knowledge from Volumes of Data. | | | 1996 | |
| Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthrusamy, R. (Eds) | Advances in Knowledge Discovery and Data Mining. | AAAI/MIT Press | Cambridge MA | 1996 | |
| Jefrey Leek | The Elements of Data Analytic Style | LeanPub | | 2014 | |
| | http://worldpece.org/sites/default/files/datastyle.pdf | | | | |