ANX-PR/CL/001-01

LEARNING GUIDE

SUBJECT

103000391 - Knowledge Discovery In Data Bases

DEGREE PROGRAMME

10AK - Master Universitario En Software Y Sistemas

ACADEMIC YEAR & SEMESTER

2023/24 - Semester 1

Index

Learning guide

1. Description

1.1. Subject details

Name of the subject	103000391 - Knowledge Discovery In Data Bases
No of credits	4 ECTS
Type	Optional
Academic year ot the programme	First year
Semester of tuition	Semester 1
Tuition period	September-January
Tuition languages	English
Degree programme	10AK - Master Universitario en Software y Sistemas
Centre	10 - Escuela Tecnica Superior De Ingenieros Informaticos
Academic year	2023-24

2. Faculty

2.1. Faculty members with subject teaching role

Name and surname	Office/Room	Email	Tutoring hours *
Juan Pedro Caraca-Valente Hernandez (Subject coordinator)	D4301	juanpedro.caracavalente@upm.es	Tu - 09:00 - 12:00 Th - 10:00 - 13:00
Aurora Perez Perez	D4301	aurora.perez@upm.es	M - 10:30 - 13:30 Th - 10:30 - 13:30

* The tutoring schedule is indicative and subject to possible changes. Please check tutoring times with the faculty member in charge.

3. Skills and learning outcomes *

3.1. Skills to be learned

CEM2 - Analizar y sintetizar soluciones a problemas que requieran aproximaciones novedosas para la definición de la infraestructura computacional que permita el procesamiento y el análisis de datos de diversa naturaleza

CEM7 - Evaluar y aplicar las diversas teorías matemáticas y estadísticas, y los procesos, métodos y técnicas disponibles para la extracción y descubrimiento de conocimiento a partir de grandes volúmenes de datos

CEM8 - Aplicar los fundamentos teóricos y matemáticos adecuados al procesamiento y análisis de funciones y datos de diversa naturaleza, y evaluar y diseñar los métodos relacionados para su aplicación en dominios prácticos

CG1 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.

CG12 - Comprensión amplia de las técnicas y métodos aplicables en una especialización concreta, así como de sus límites

CG13 - Apreciación de los límites del conocimiento actual y de la aplicación práctica de la tecnología más reciente.

CG14 - Conocimiento y comprensión de la informática necesaria para la creación de modelos de información, y de los sistemas y procesos complejos

CG17 - Habilidades de gestión y capacidad de liderar un equipo que puede estar integrado por disciplinas y niveles distintos.

CG19 - Aproximación sistemática a la gestión de riesgos.

CG3 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.

CG4 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

CG7 - Especificación y realización de tareas informáticas complejas, poco definidas o no familiares

CG8 - Planteamiento y resolución de problemas también en áreas nuevas y emergentes de su disciplina

CG9 - Aplicación de los métodos de resolución de problemas más recientes o innovadores y que puedan implicar el uso de otras disciplinas

CGI20 - Adquirir conocimientos científicos avanzados del campo de la informática que le permitan generar nuevas ideas dentro de una línea de investigación.

CGI23 - Capacidad de leer y comprender publicaciones dentro de su ámbito de estudio/investigación, así como su catalogación y valor científico

3.2. Learning outcomes

RA68 - Ser capaz de analizar un dominio para determinar la relevancia de sus características temporales y las tareas de descubrimiento de conocimiento que se podrían plantear

RA70 - Ser capaz de realizar una evaluación completa del funcionamiento y utilidad de un proyecto de este tipo.

RA69 - Ser capaz de utilizar las técnicas de descubrimiento de conocimiento y su aplicabilidad en cada caso

* The Learning Guides should reflect the Skills and Learning Outcomes in the same way as indicated in the Degree Verification Memory. For this reason, they have not been translated into English and appear in Spanish.

4. Brief description of the subject and syllabus

4.1. Brief description of the subject

Knowledge Discovery techniques (or Data Mining) in large volumes of information are widely used today in different domains such as medicine, banking environments, industrial systems, etc. with a wide variety of applications such as data analysis, fraud detection, risk analysis, marketing campaigns, etc.

In this course all the stages of the Knowledge Discovery process will be reviewed and the most important techniques for each stage will be listed. Emphasis will be placed on techniques for data cleaning and preprocessing that, despite their importance, are often forgotten.

Next, the main techniques of Data Mining including Classification and Clustering techniques will be addressed. Some more recent methods for Data Analysis, including Deep Learning Techniques will also be covered.

In this subject we also want to explore areas of Knowledge Discovery less known, but equally important. There are domains where information is presented mostly in the form of Time Series which require a very specialized

treatment. Examples of these are medical domains such as Electrocardiography or Audiometry, financial domains, etc. Time series are a challenge to the traditional techniques of Data Mining and often require the use of novel solutions. Special emphasis will be made on Temporal Abstraction tecnhiques.

4.2. Syllabus

1. Introduction
    1.1. Data Types, Time Series
    1.2. Basic Concepts
2. Knowledge Discovery Process
    2.1. Knowledge Discovery Process Stages
    2.2. Data Preprocessing for basic data types and time series
3. KDD Tools
    3.1. Background
    3.2. A KDD Tool: WEKA
4. Data Mining Techniques
    4.1. Classification
    4.2. Advanced Methods for Data Analysis
    4.3. Clustering
    4.4. Time Series Techniques
5. Evaluation
    5.1. Objectives
    5.2. Evaluation Techniques

5. Schedule

5.1. Subject schedule*

Week	Classroom activities	Laboratory activities	Distant / On-line	Assessment activities
1	**1. Introduction** Duration: 02:00 Lecture			
2	**2.1 Knowledge Discovery Process** Duration: 00:45 Lecture **2.2 Data Preprocessing for basic data types and time series** Duration: 01:00 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
3	**3 KDD Tools** Duration: 00:20 Lecture **3.1 Background and 3.2 WEKA** Duration: 01:25 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
4	**3.2 Case Study: WEKA** Duration: 01:00 Problem-solving class **Domain Analysis and KDD Process** Duration: 00:45 Cooperative activities			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
5	**4.1 Classification Techniques** Duration: 01:45 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
6	**4.1 Classification Techniques** Duration: 01:45 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
7	**4.2 Advanced Methods for Data Analysis** Duration: 01:45 Lecture **Case Study: Advanced Methods** Duration: 00:15 Cooperative activities			**Project Stage 1: Domain Analysys, Data study, Objective definition** Group work Continuous assessment Not Presential Duration: 00:20

8	**4.2 ClusteringTechniques** Duration: 01:45 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
9	**4.2 Clustering Techniques** Duration: 01:45 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
10	**4.3 Time Series Data Mining** Duration: 01:45 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
11	**4.3 Time Series Data Mining** Duration: 01:45 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
12	**4.3 Time Series Data Mining** Duration: 01:45 Lecture **Case Study: Time Series Data Mining** Duration: 00:15 Cooperative activities			**Project Stage 2: Application of Data Mining Techniques** Group work Continuous assessment Not Presential Duration: 00:20
13	**4.3 Time Series Data Mining** Duration: 01:45 Lecture			**Progressive Evaluation Activities** Other assessment Continuous assessment Presential Duration: 00:15
14	**5 Evaluation** Duration: 01:45 Lecture **Group Discussion** Duration: 00:15 Additional activities			
15				**Project Stage 3: Evaluation** Group work Continuous assessment Not Presential Duration: 00:20 **Project Presentation** Group presentation Continuous assessment Presential Duration: 02:00
16				
17				**Project complete** Group work Final examination Not Presential Duration: 01:00

Depending on the programme study plan, total values will be calculated according to the ECTS credit unit as 26/27 hours of student face-to-face contact and independent study time.

* The schedule is based on an a priori planning of the subject; it might be modified during the academic year, especially considering the COVID19 evolution.

6. Activities and assessment criteria

6.1. Assessment activities

6.1.1. Assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
2	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM2 CG1 CG4 CG8 CEM8
3	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM2 CG1 CG4 CEM8
4	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM2 CG1 CG4 CG7 CG8 CG9 CG12
5	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM2 CG1 CG4 CG8 CG9 CG13 CG14
6	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM7 CEM2 CG1 CG4 CG8 CG9 CG19 CGI20 CGI23 CEM8

7	Project Stage 1: Domain Analysys, Data study, Objective definition	Group work	No Presential	00:20	10%	/ 10	CG1 CG7 CG8 CG12 CG17
8	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM7 CG3 CG7 CG12 CGI20 CGI23
9	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM2 CG3 CG4 CG8 CG9 CG19 CEM8
10	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM2 CG1 CG3 CG4 CG7 CG8 CG9 CG12 CG13 CG14 CG17 CG19 CGI20
11	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM2 CG1 CG3 CG4 CG7 CG8 CG9 CG12 CG13 CG19 CEM8
12	Project Stage 2: Application of Data Mining Techniques	Group work	No Presential	00:20	10%	/ 10	CEM7 CEM2 CG1 CG3 CG4 CG7 CG8 CG9 CG12 CG13 CG14

							CG17 CG19 CGI20 CGI23 CEM8
13	Progressive Evaluation Activities	Other assessment	Face-to-face	00:15	3%	/ 10	CEM7 CEM2 CG1 CG3 CG4 CG7 CG8 CG9 CG12 CG13 CG14 CG17 CG19 CGI20 CGI23 CEM8
15	Project Stage 3: Evaluation	Group work	No Presential	00:20	10%	/ 10	CEM7 CEM2 CG1 CG3 CG4 CG7 CG8 CG9 CG12 CG13 CG14 CG17 CG19 CGI20 CGI23 CEM8
15	Project Presentation	Group presentation	Face-to-face	02:00	40%	/ 10	CEM7 CEM2 CG1 CG3 CG4 CG7 CG8 CG9 CG12 CG13 CG14 CG17 CG19 CGI20 CGI23 CEM8

6.1.2. Global examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
17	Project complete	Group work	No Presential	01:00	100%	/ 10	CEM7 CEM2 CG1 CG3 CG4 CG7 CG8 CG9 CG12 CG13 CG14 CG17 CG19 CGI20 CGI23 CEM8

6.1.3. Referred (re-sit) examination

Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
Project complete	Group work	Face-to-face	00:00	100%	5 / 10	CEM7 CEM2 CG1 CG3 CG4 CG7 CG8 CG9 CG12 CG13 CG14 CG17 CG19

						CGI20 CGI23

6.2. Assessment criteria

For the evaluation of this subject we will take into account, on the one hand, Progressive Evaluation Activities (PEA) that will be carried out daily in class and, on the other hand, the Data Mining Project.

PEA

The PEA can only be evaluated through out the course, therefore they can not be re-evaluated for the Final Evaluation mode or in the extra call. In these activities we will account for the attendance to class, active participation of the student and the evaluable exercises that will be performed almost every class.

Data Mining Project

The Data Mining Project will be evaluated according to the three phases described below and the corresponding weights.

This project will be done individually or in groups of 2 people. The work will be done incrementally and will be presented in the following phases:

· Phase 1: students will choose a domain to which data they have access, analyze their characteristics and establish the objectives to be achieved through the Data Mining Project. They will write a report indicating the different tasks that would be carried out in each stage of the Knowledge Discovery process according to the specific needs of the domain and the objectives.

· Phase 2: through the use of a Knowledge Discovery software tool, Data Mining algorithms will be applied to the data of each domain. In addition, the student will analyze the limitations of the algorithms available in

the tool and possible improvements.

· Phase 3: an evaluation plan will be made to assess the results obtained and the plan will be executed.

The 3 deliveries of the Data Mining Project are mandatory and will be evaluated according to the weights assigned in the table in the previous section (summative evaluation).

The Data Mining Project will be presented in class. Each group will have 15 minutes for the oral presentation plus 5 minutes of questions.

Qualification standards

The subject will be evaluated on 10 points, divided into 3 points for continuous assessment (this part can only be done during the course) and 7 for the Data Mining Project. To pass the subject it will be necessary to attend at least 70% of the classes and obtain a final grade of no less than 5 points.

The dates for the delivery of each part of the Data Mining Project will be published at the beggining of the course.

In the extra call, those parts of the Data Mining Project that are pending may be delivered . Continuous assessment will not be repeated, so the grade of the subject will be obtained exclusively from the Data Mining Project.

7. Teaching resources

7.1. Teaching resources for the subject

Name	Type	Notes
WEKA	Web resource	Official webpage of the Data Mining Tool WEKA, with tutorials and free download http://www.cs.waikato.ac.nz/ml/weka/
Data Mining: Concepts and Techniques	Bibliography	Book about Data Mining Techniques. J.Han y M. Kamber. Ed. Morgan Kauffman, 2006.
Data Mining: Concepts, Models, Methods, and Algorithms	Bibliography	Book about Data Mining Techniques. M. Kantardzic (eds.), John Wiley & Sons, 2003
From Data Mining to Knowledge Discovery in Databases	Bibliography	Paper: fundational works on nowadays Data Mining. U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, 1996
Subject webpage	Web resource	https://muss.fi.upm.es/asigDCBD.php
Moodle	Others	https://moodle.upm.es/titulaciones/oficiales/course/view.php?id=406

8. Other information

8.1. Other information about the subject

During the course, we will try to use as many data files related to Sustainable Development Goals of UN as possible, specially number 13 Climate Action