ANX-PR/CL/001-01
LEARNING GUIDE

SUBJECT
103000865 - Data Mining And Time Series

DEGREE PROGRAMME
10AZ - Master Universitario en Innovación Digital

ACADEMIC YEAR & SEMESTER
2020/21 - Semester 1
Index

Learning guide

1. Description...............................................................................................................................................................1
2. Faculty.....................................................................................................................................................................1
3. Skills and learning outcomes ...................................................................................................................................2
4. Brief description of the subject and syllabus .............................................................................................................3
5. Schedule..................................................................................................................................................................5
6. Activities and assessment criteria............................................................................................................................8
7. Teaching resources.................................................................................................................................................11
8. Other information....................................................................................................................................................11
1. Description

1.1. Subject details

<table>
<thead>
<tr>
<th>Name of the subject</th>
<th>103000865 - Data Mining And Time Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of credits</td>
<td>3 ECTS</td>
</tr>
<tr>
<td>Type</td>
<td>Optional</td>
</tr>
<tr>
<td>Academic year of the programme</td>
<td>Second year</td>
</tr>
<tr>
<td>Semester of tuition</td>
<td>Semester 3</td>
</tr>
<tr>
<td>Tuition period</td>
<td>September-January</td>
</tr>
<tr>
<td>Tuition languages</td>
<td>English</td>
</tr>
<tr>
<td>Degree programme</td>
<td>10AZ - Master Universitario en Innovación Digital</td>
</tr>
<tr>
<td>Centre</td>
<td>10 - Escuela Tecnica Superior de Ingenieros Informaticos</td>
</tr>
<tr>
<td>Academic year</td>
<td>2020-21</td>
</tr>
</tbody>
</table>

2. Faculty

2.1. Faculty members with subject teaching role

<table>
<thead>
<tr>
<th>Name and surname</th>
<th>Office/Room</th>
<th>Email</th>
<th>Tutoring hours *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juan Pedro Caraca-Valente Hernandez (Subject coordinator)</td>
<td>D4301</td>
<td><a href="mailto:juanpedro.caracavalente@upm.es">juanpedro.caracavalente@upm.es</a></td>
<td>Tu - 09:00 - 12:00  Th - 10:00 - 13:00</td>
</tr>
<tr>
<td>Aurora Perez Perez</td>
<td>D4301</td>
<td><a href="mailto:aurora.perez@upm.es">aurora.perez@upm.es</a></td>
<td>M - 10:30 - 13:30  Th - 10:30 - 13:30</td>
</tr>
</tbody>
</table>

* 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

CB07 - 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.

CE-CD04 - Capacidad para aplicar métodos avanzados para clasificación, modelado, segmentación y predicción a partir de un conjunto de datos.

CG02 - Que los estudiantes desarrollen la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro su ámbito temático explorando y generando nuevas ideas sistemáticamente, en contextos interdisciplinares y, en su caso, con una alta componente de transferencia del conocimiento.

CG03 - La capacidad de usar la lengua inglesa de manera competente, es decir, con capacitación para tareas complejas de trabajo y estudio.

3.2. Learning outcomes

RA76 - Know how to select and apply the best candidate techniques to Time Series Data Mining project.

RA75 - Understand the basics of the Knowledge Discovery Process, and its application to time series and complex data.

* 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, focusing on time series and complex data formats, 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.

In this subject we 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. On the other hand, in recent years social networks have become immensely popular and have generated a large amount of data, structured in the form of complex data networks, which are very interesting from the point of view of the discovery of hidden knowledge in them.

4.2. Syllabus

1. Introduction
   1.1. Data Types
   1.2. Basic Concepts

2. Knowledge Discovery Process
   2.1. Knowledge Discovery Process Stages
   2.2. Data Preprocessing

3. KDD Tools
   3.1. Background
   3.2. A KDD Tool: WEKA

4. Data Mining Techniques for Time Series and Complex data
4.1. Classification

4.2. Clustering

4.3. Genetic Algorithms

4.4. Time Series Techniques

5. Evaluation

5.1. Objectives

5.2. Evaluation Techniques
### 5. Schedule

#### 5.1. Subject schedule*

<table>
<thead>
<tr>
<th>Week</th>
<th>Face-to-face classroom activities</th>
<th>Face-to-face laboratory activities</th>
<th>Distant / On-line</th>
<th>Assessment activities</th>
</tr>
</thead>
</table>
| 1    | 1. Introduction  
Duration: 02:00 | 1. Introduction  
Duration: 02:00 | 1. Introduction  
Duration: 02:00 | Continuous assessment  
Not Presential  
Duration: 00:20 |
| 2    | 2.1 Knowledge Discovery Process  
Duration: 01:45 | 2.1 Knowledge Discovery Process  
Duration: 01:45 | 2.1 Knowledge Discovery Process  
Duration: 01:45 | |
|      | Case Study: Knowledge Discovery Process  
Duration: 00:15 | Case Study: Knowledge Discovery Process  
Duration: 00:15 | Case Study: Knowledge Discovery Process  
Duration: 00:15 | |
| 3    | 2.2 Data Preprocessing  
Duration: 01:45 | 2.2 Data Preprocessing  
Duration: 01:45 | 2.2 Data Preprocessing  
Duration: 01:45 | |
|      | Case Study: Data Preprocessing  
Duration: 00:15 | Case Study: Data Preprocessing  
Duration: 00:15 | Case Study: Data Preprocessing  
Duration: 00:15 | |
| 4    | 3 KDD Tools  
Duration: 00:20 | 3 KDD Tools  
Duration: 00:20 | 3 KDD Tools  
Duration: 00:20 | |
|      | 3.1 Background and 3.2 WEKA  
Duration: 01:40 | 3.1 Background and 3.2 WEKA  
Duration: 01:40 | 3.1 Background and 3.2 WEKA  
Duration: 01:40 | |
| 5    | 3.2 Case Study: WEKA  
Duration: 01:20 | 3.2 Case Study: WEKA  
Duration: 01:20 | 3.2 Case Study: WEKA  
Duration: 01:20 | Project Stage 1: Domain Analysys, Data study, Objective definition |
|      | Domain Analysis and KDD Process  
Duration: 00:40 | Domain Analysis and KDD Process  
Duration: 00:40 | Domain Analysis and KDD Process  
Duration: 00:40 | Continuous assessment  
Not Presential  
Duration: 00:20 |
| 6    | 4.1 Classification Techniques  
Duration: 01:45 | 4.1 Classification Techniques  
Duration: 01:45 | 4.1 Classification Techniques  
Duration: 01:45 | |
|      | Case Study: Classification Techniques  
Duration: 00:15 | Case Study: Classification Techniques  
Duration: 00:15 | Case Study: Classification Techniques  
Duration: 00:15 | |
| 7    | 4.1 Classification Techniques  
Duration: 01:45 | 4.1 Classification Techniques  
Duration: 01:45 | 4.1 Classification Techniques  
Duration: 01:45 | |
|      | Case Study: Classification Techniques  
Duration: 00:15 | Case Study: Classification Techniques  
Duration: 00:15 | Case Study: Classification Techniques  
Duration: 00:15 | |
<table>
<thead>
<tr>
<th>Page</th>
<th>Activity</th>
<th>Duration</th>
<th>Activity</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4.2 Clustering Techniques</td>
<td>01:45</td>
<td>4.2 Clustering Techniques</td>
<td>01:45</td>
</tr>
<tr>
<td></td>
<td>Case Study: Clustering Techniques</td>
<td>00:15</td>
<td>Case Study: Clustering Techniques</td>
<td>00:15</td>
</tr>
<tr>
<td>9</td>
<td>4.2 Clustering Techniques</td>
<td>01:45</td>
<td>4.2 Clustering Techniques</td>
<td>01:45</td>
</tr>
<tr>
<td></td>
<td>Case Study: Clustering Techniques</td>
<td>00:15</td>
<td>Case Study: Clustering Techniques</td>
<td>00:15</td>
</tr>
<tr>
<td>10</td>
<td>4.3 Genetic Algorithms</td>
<td>01:45</td>
<td>4.3 Genetic Algorithms</td>
<td>01:45</td>
</tr>
<tr>
<td></td>
<td>Case Study: Genetic Algorithms</td>
<td>00:15</td>
<td>Case Study: Genetic Algorithms</td>
<td>00:15</td>
</tr>
<tr>
<td>11</td>
<td>4.3 Genetic Algorithms</td>
<td>01:45</td>
<td>4.3 Genetic Algorithms</td>
<td>01:45</td>
</tr>
<tr>
<td></td>
<td>Case Study: Genetic Algorithms</td>
<td>00:15</td>
<td>Case Study: Genetic Algorithms</td>
<td>00:15</td>
</tr>
<tr>
<td>12</td>
<td>4.4 Time Series Data Mining</td>
<td>01:45</td>
<td>4.4 Time Series Data Mining</td>
<td>01:45</td>
</tr>
<tr>
<td></td>
<td>Case Study: Time Series Data Mining</td>
<td>00:15</td>
<td>Case Study: Time Series Data Mining</td>
<td>00:15</td>
</tr>
<tr>
<td>13</td>
<td>4.4 Time Series Data Mining</td>
<td>01:45</td>
<td>4.4 Time Series Data Mining</td>
<td>01:45</td>
</tr>
<tr>
<td></td>
<td>Case Study: Time Series Data Mining</td>
<td>00:15</td>
<td>Case Study: Time Series Data Mining</td>
<td>00:15</td>
</tr>
<tr>
<td>14</td>
<td>4.4 Time Series Data Mining</td>
<td>01:45</td>
<td>4.4 Time Series Data Mining</td>
<td>01:45</td>
</tr>
<tr>
<td></td>
<td>Case Study: Time Series Data Mining</td>
<td>00:15</td>
<td>Case Study: Time Series Data Mining</td>
<td>00:15</td>
</tr>
<tr>
<td>15</td>
<td>5 Evaluation</td>
<td>01:45</td>
<td>5 Evaluation</td>
<td>01:45</td>
</tr>
<tr>
<td></td>
<td>Case Study: Data Mining Project Evaluation</td>
<td>00:15</td>
<td>Case Study: Data Mining Project Evaluation</td>
<td>00:15</td>
</tr>
</tbody>
</table>
| 16 | Project Stage 3: Evaluation  
Continuous assessment  
Not Present  
Duration: 00:20 |
| 17 | Project complete  
Final examination  
Not Present  
Duration: 01:00 |
|     | Project Presentation  
Continuous assessment and final examination  
Not Present  
Duration: 02:00 |
|     | Attendance to class, participation and evaluable exercises  
Continuous assessment  
Not Present  
Duration: 00:00 |

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. Continuous assessment

<table>
<thead>
<tr>
<th>Week</th>
<th>Description</th>
<th>Modality</th>
<th>Type</th>
<th>Duration</th>
<th>Weight</th>
<th>Minimum grade</th>
<th>Evaluated skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Project Stage 1: Domain Analysis, Data study, Objective definition</td>
<td>No Presential</td>
<td>00:20</td>
<td>10%</td>
<td>/ 10</td>
<td>CG02</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Project Stage 2: Application of Data Mining Techniques</td>
<td>No Presential</td>
<td>00:20</td>
<td>10%</td>
<td>/ 10</td>
<td>CB07, CG02, CG03</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Project Stage 3: Evaluation</td>
<td>No Presential</td>
<td>00:20</td>
<td>10%</td>
<td>/ 10</td>
<td>CB07, CG02, CG03</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Project Presentation</td>
<td>No Presential</td>
<td>02:00</td>
<td>40%</td>
<td>/ 10</td>
<td>CB07, CG02, CG03, CE-CD04</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Attendance to class, participation and evaluable exercises</td>
<td>No Presential</td>
<td>00:00</td>
<td>30%</td>
<td>/ 10</td>
<td>CB07, CG02, CE-CD04</td>
<td></td>
</tr>
</tbody>
</table>

6.1.2. Final examination

<table>
<thead>
<tr>
<th>Week</th>
<th>Description</th>
<th>Modality</th>
<th>Type</th>
<th>Duration</th>
<th>Weight</th>
<th>Minimum grade</th>
<th>Evaluated skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>Project complete</td>
<td>No Presential</td>
<td>01:00</td>
<td>60%</td>
<td>/ 10</td>
<td>CB07, CG02, CG03, CE-CD04</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Project Presentation</td>
<td>No Presential</td>
<td>02:00</td>
<td>40%</td>
<td>/ 10</td>
<td>CB07, CG02, CG03, CE-CD04</td>
<td></td>
</tr>
</tbody>
</table>

6.1.3. Referred (re-sit) examination
### 6.2. Assessment criteria

For the evaluation of this subject we will take into account, on the one hand, the attendance and participation in class and, on the other hand, the Data Mining Project.

For the Continuous Evaluation the attendance to class, the active participation of the student and the evaluable exercises that are raised in class will be considered.

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

### Data Mining Project

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 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 beginning 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

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEKA</td>
<td>Web resource</td>
<td>Official webpage of the Data Mining Tool WEKA, with tutorials and free download <a href="http://www.cs.waikato.ac.nz/ml/weka/">http://www.cs.waikato.ac.nz/ml/weka/</a></td>
</tr>
<tr>
<td>From Data Mining to Knowledge Discovery in Databases</td>
<td>Bibliography</td>
<td>Book about some foundational works on nowadays Data Mining Techniques U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, 1996</td>
</tr>
<tr>
<td>Subject webpage</td>
<td>Web resource</td>
<td><a href="http://www.dlsiis.fi.upm.es/master_muss/asig">http://www.dlsiis.fi.upm.es/master_muss/asig</a> DCBD.html</td>
</tr>
</tbody>
</table>

8. Other information

8.1. Other information about the subject

Classes will be taught in English (Spanish only if every student speaks spanish correctly). Documentation (including transparencies used in class) will be provided in English

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