ANX-PR/CL/001-01

LEARNING GUIDE

SUBJECT

103000866 - Massively Parallel Machine Learning

DEGREE PROGRAMME

10AZ - Master Universitario En Innovación Digital

ACADEMIC YEAR & SEMESTER

2022/23 - Semester 1

Index

Learning guide

1. Description

1.1. Subject details

Name of the subject	103000866 - Massively Parallel Machine Learning
No of credits	4.5 ECTS
Type	Optional
Academic year ot the programme	Second year
Semester of tuition	Semester 3
Tuition period	September-January
Tuition languages	English
Degree programme	10AZ - Master Universitario en Innovación Digital
Centre	10 - Escuela Tecnica Superior De Ingenieros Informaticos
Academic year	2022-23

2. Faculty

2.1. Faculty members with subject teaching role

Name and surname	Office/Room	Email	Tutoring hours *
Bonifacio Alberto Mozo Velasco (Subject coordinator)	4307	a.mozo@upm.es	W - 16:00 - 17:30 Th - 16:00 - 17:30

* 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

CB06 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

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

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

CG07 - Capacidad de trabajar y comunicarse también en contextos internacionales.

3.2. Learning outcomes

RA46 - Apply parallelization strategies to machine learning algorithms

RA47 - Design massively parallel versions of supervised machine learning algorithms

RA44 - Understand the basic mechanisms for designing parallel applications in Big Data regimes

* 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

This course introduces to the parallelization of Machine Learning algorithms (supervised and unsupervised) on Big Data distributed computing platforms. In particular, MapReduce and RDD paradigms are explained jointly with the corresponding distributed platforms supporting them (Hadoop and Apache Spark). Using several well-known machine learning algorithms, the student will learn how to parallelize machine learning algorithms using the corresponding Hadoop and Apache Spark APIs.

4.2. Syllabus

1. Introduction to Big Data
2. Machine Learning (supervised and unsupervised techniques)
3. MapReduce and Hadoop
4. RDDs and Spark
5. Parallelization of machine learning algorithms

5. Schedule

5.1. Subject schedule*

Week	Classroom activities	Laboratory activities	Distant / On-line	Assessment activities
1	**Lecture on Unit 1** Duration: 02:30			
2	**Lecture on Unit 2** Duration: 02:30			
3	**Practical classes** Duration: 02:00			
4	**Practical classes** Duration: 02:00			
5	**Lecture on Unit 3** Duration: 02:00			
6	**Lecture on Unit 3** Duration: 02:00			
7	**Lecture on Unit 4** Duration: 02:00			
8	**Lecture on Unit 4** Duration: 02:00			
9				**Individual Exam** Continuous assessment Presential Duration: 02:00
10	**Project development and tutorial in laboratory** Duration: 02:00		**Project Development** Duration: 03:00	
11	**Project development and tutorial in laboratory** Duration: 02:00		**Project Development** Duration: 03:00	
12	**Project development and tutorial in laboratory** Duration: 02:00		**Project Development** Duration: 03:00	

13	**Project development and tutorial in laboratory** Duration: 02:00		**Project Development** Duration: 03:00	
14				**Student Project Presentations** Continuous assessment and final examination Presential Duration: 02:00
15				
16				
17				**Final Exam** Final examination Not Presential Duration: 02:00

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
9	Individual Exam		Face-to-face	02:00	35%	5 / 10	CG03 CB07 CB06 CG07
14	Student Project Presentations		Face-to-face	02:00	65%	5 / 10	CG03 CG07 CB06

6.1.2. Global examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
14	Student Project Presentations		Face-to-face	02:00	65%	5 / 10	CG03 CG07 CB06
17	Final Exam		No Presential	02:00	35%	5 / 10	CB07 CB06 CG03 CG07

6.1.3. Referred (re-sit) examination

Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
The student will take an individual written exam covering Units 1-4		Face-to-face	02:00	35%	5 / 10	CG03 CG07 CB07 CB06
The student will hand the practical assignments up to the day of the written exam.		Face-to-face	00:00	65%	5 / 10	CB06 CG03 CG07

6.2. Assessment criteria

Assessment Criteria

Continous assessment

The following partial grades will be obtained (grades from 0 to 10):

N1: Grade for the individual exam.

N2: Grade for the student project.

The final grade is obtained as follows: N = 0.3*N1 + 0.7*N2

As a general criterion, to pass the course (in the first or second call) it is required:

- **N1 >= 5 and N2 >= 5.**

Final exam

The final grade is the grade obtained in the final exam

7. Teaching resources

7.1. Teaching resources for the subject

Name	Type	Notes
Resilient Distributed Datasets	Bibliography	Zaharia, Matei, et al. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012.
MapReduce	Bibliography	Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.
UPM Moodle Web Site	Web resource	The web site UPM-Moodle (http://moodle.upm.es) will be used to provide course material to students. Students will use this web site to submit for evaluation the results of their assignments and individual works.
Machine Learning. A probabilistic approach.	Bibliography	Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

8. Other information

8.1. Other information about the subject

Sustainable Development Goals:
This subject is aligned with Goal 4 Quality Education that aims to "Guarantee inclusive, equitable and quality education and promote lifelong learning opportunities for all" that defined within the Sustainable Development Goals (SDGs) established by the United Nations Organization (UN).

Other considerations:
In anticipation of possible recurrences of the COVID epidemic and in the event of a possible suspension of all face-to-face educational activities, the exceptional procedure that will be carried out will be that the evaluation activities initially planned as face-to-face activities could be developed through online platforms. The online platforms to be used will be those provided by the University, both those accessible through Moodle and those available as a telematic resource.