ANX-PR/CL/001-01

LEARNING GUIDE

SUBJECT

203000035 - Big Data Engineering

DEGREE PROGRAMME

20BC - Master Universitario En Biologia Computacional

ACADEMIC YEAR & SEMESTER

2021/22 - Semester 2

Index

Learning guide

1. Description

1.1. Subject details

Name of the subject	203000035 - Big Data Engineering
No of credits	3 ECTS
Type	Optional
Academic year ot the programme	First year
Semester of tuition	Semester 2
Tuition period	February-June
Tuition languages	English
Degree programme	20BC - Master Universitario en Biologia Computacional
Centre	20 - E.T.S. De Ingenieria Agronomica, Alimentaria Y De Biosistemas
Academic year	2021-22

2. Faculty

2.1. Faculty members with subject teaching role

Name and surname	Office/Room	Email	Tutoring hours *
Jose Manuel Moya Fernandez (Subject coordinator)		jm.moya@upm.es	- -

* 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

CE02 - Utilizar sistemas operativos, programas y herramientas de uso común en biología computacional, así como, manejar plataformas de cómputo de altas prestaciones, lenguajes de programación y análisis bioinformáticos

CE03 - Analizar e interpretar bioinformáticamente los datos que se derivan de las tecnologías ómicas, y proponer soluciones bioinformáticas en relación a dichos datos.

CE05 - Utilizar herramientas de biología computacional para el análisis genómico, incluida la genómica comparativa y biología evolutiva.

CE10 - Conocimiento de las técnicas de representación del conocimiento reutilizables y modelos de razonamiento en entornos centralizados y distribuidos a utilizar en la resolución de problemas que impliquen conducta inteligente.

CG01 - Poseer los conocimientos que constituyen la base científica y tecnológica de la Biología computacional, lo que permitirá el desarrollo de ideas originales en este campo, en un contexto de investigación o desarrollo.

CG03 - 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 el area de la Biología Computacional.

CT08 - Tener capacidad de análisis y síntesis para interpretar datos relevantes y abordar los problemas desde diferentes perspectivas.

3.2. Learning outcomes

RA48 - Entender qué es el análisis de grandes volúmenes de datos, sus características, los métodos y las plataformas más importantes

RA49 - Entender la plataforma Apache Spark, sus objetivos, sus componentes y su modelo de programación

RA50 - Escribir y depurar aplicaciones Spark para el análisis de datos masivos

RA51 - Utilizar Spark para desarrollar productos y servicios intensivos en datos, como sistemas de recomendación, predicción y diagnóstico, utilizando procesamiento paralelo, Spark Machine Learning Pipelines y Spark Streaming

RA52 - Utilizar SparkR para realizar análisis estadístico en grandes volúmenes de datos.

* 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

The analysis of large volumes of data is used to support decision making and to build data intensive products and services, such as recommendation, prediction and diagnostic systems. The collection of skills required to perform these functions have been grouped under the term Data Science. This course develops the necessary skills around the Apache Spark platform, which is rapidly becoming the reference platform for mass data analysis. The course is structured around practical projects that include prediction, using automatic learning algorithms, collaborative filtering and pattern search. These exercises demonstrate in a practical way how to manipulate data sets through parallel processing with PySpark, Spark SQL, Spark Machine Learning Pipelines and SparkR

In this course, you will:

- Understand what high-volume data analysis is, its features, methods, and major platforms.
- Understand the Apache Spark platform, its objectives, its components, and its programming model.
- Write and debug Spark applications for mass data analysis.
- Use Spark to develop data-intensive products and services, such as recommendation, prediction, and diagnostic systems, using parallel processing, Spark Machine Learning Pipelines, and Spark Streaming.
- Use PySpark or SparkR to perform statistical analysis on large volumes of data

4.2. Syllabus

1. Introduction to memory locality and latency and the Spark model

2. Spark Dataframes and Spark SQL basics

3. Exploratory data analysis with Spark

4. Introduction to machine learning in Spark

5. Linear regression in Spark

6. Logistic regression in Spark

7. Decision Trees, Random Forest and Boosting in Spark

8. K-Means clustering in Spark

9. Recommender systems in Spark

10. Natural language processing in Spark

11. Spark Streaming

12. Scalable neural networks with Spark

5. Schedule

5.1. Subject schedule*

Week	Face-to-face classroom activities	Face-to-face laboratory activities	Distant / On-line	Assessment activities
1	**1. Introduction to memory locality and latency and the Spark model** Duration: 02:00			
2	**1. Introduction to memory locality and latency and the Spark model** Duration: 02:00			
3	**2. Spark Dataframes and Spark SQL basics** Duration: 02:00			**Assignment 1 - Spark Dataframes exercise** Continuous assessment and final examination Not Presential Duration: 00:00
4	**3. Exploratory data analysis with Spark** Duration: 02:00			
5	**4. Introduction to machine learning in Spark** Duration: 02:00			
6	**5. Linear regression in Spark** Duration: 02:00			**Assignment 2 - Consulting project: Linear regression** Continuous assessment and final examination Not Presential Duration: 00:00
7	**6. Logistic regression in Spark** Duration: 02:00			**Assignment 3 - Consulting project: Logistic regression** Continuous assessment and final examination Not Presential Duration: 00:00
8	**7. Decision Trees, Random Forest and Boosting in Spark** Duration: 02:00			**Assignment 4 - Consulting projec: Tree methods** Continuous assessment and final examination Not Presential Duration: 00:00

9	**8. K-Means clustering in Spark** Duration: 02:00			**Assignment 5 - Consulting project: Clustering** Continuous assessment and final examination Not Presential Duration: 00:00
10	**9. Recommender systems in Spark** Duration: 02:00			
11	**10. Natural language processing in Spark** Duration: 02:00			
12	**11. Spark Streaming** Duration: 02:00			
13	**12. Scalable neural networks with Spark** Duration: 02:00			**Assignment 6 - Final project: Neural networks** Continuous assessment and final examination Not Presential Duration: 00:00
14	**12. Scalable neural networks with Spark** Duration: 02:00			
15	**Project review and discussion** Duration: 02:00			
16	**Project review and discussion** Duration: 02:00			
17				**Exam** Continuous assessment and final examination 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. Continuous assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
3	Assignment 1 - Spark Dataframes exercise		No Presential	00:00	15%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03
6	Assignment 2 - Consulting project: Linear regression		No Presential	00:00	15%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02
7	Assignment 3 - Consulting project: Logistic regression		No Presential	00:00	15%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02
8	Assignment 4 - Consulting projec: Tree methods		No Presential	00:00	15%	/ 10	CE05 CE03 CE10 CG03 CE02 CG01 CT08
9	Assignment 5 - Consulting project: Clustering		No Presential	00:00	15%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02

13	Assignment 6 - Final project: Neural networks		No Presential	00:00	20%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02
17	Exam		Face-to-face	02:00	5%	5 / 10	

6.1.2. Final examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
3	Assignment 1 - Spark Dataframes exercise		No Presential	00:00	15%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03
6	Assignment 2 - Consulting project: Linear regression		No Presential	00:00	15%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02
7	Assignment 3 - Consulting project: Logistic regression		No Presential	00:00	15%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02
8	Assignment 4 - Consulting projec: Tree methods		No Presential	00:00	15%	/ 10	CE05 CE03 CE10 CG03 CE02 CG01 CT08
9	Assignment 5 - Consulting project: Clustering		No Presential	00:00	15%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02

13	Assignment 6 - Final project: Neural networks		No Presential	00:00	20%	/ 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02
17	Exam		Face-to-face	02:00	5%	5 / 10	

6.1.3. Referred (re-sit) examination

Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
Práctica de programación en Spark y cuestionario asociado		Face-to-face	04:00	100%	5 / 10	CG01 CT08 CE05 CE03 CE10 CG03 CE02

6.2. Assessment criteria

This is the rubric used to evaluate all the programming assignments:

Specification (30%)

Readability (10%)

Reusability (10%)

Documentation (10%)

Delivery (20%)

Efficiency (20%)

7. Teaching resources

7.1. Teaching resources for the subject

Name	Type	Notes
Moodle de la asignatura	Web resource	
PySpark Documentation	Bibliography	http://spark.apache.org/docs/latest/api/python/index.html
Companion book: An Introduction to Statistical Learning with Applications in R, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani	Bibliography	http://faculty.marshall.usc.edu/gareth-james/ISL/

Companion book: Deep Learning, by Ian Goodfellow and Yoshua Bengio and Aaron Courville	Bibliography	https://www.deeplearningbook.org/