DEGREE CURRICULUM

BIG DATA PROJECT

Coordination: COMAS RODRIGUEZ, CARLOS

Academic year 2023-24

BIG DATA PROJECT 2023-24

Subject's general information

<table>
<tr><td>Subject name</td><td>BIG DATA PROJECT</td></tr>
<tr><td>Code</td><td>103090</td></tr>
<tr><td>Semester</td><td>1st Q(SEMESTER) CONTINUED EVALUATION</td></tr>
<tr><td>Typology</td><td>
<table>
<tr><th>Degree</th><th>Course</th><th>Character</th><th>Modality</th></tr>
<tr><td>Master's Degree in Informatics Engineering</td><td>2</td><td>OPTIONAL</td><td>Attendance-based</td></tr>
</table>
</td></tr>
<tr><td>Course number of credits (ECTS)</td><td>6</td></tr>
<tr><td>Type of activity, credits, and groups</td><td>
<table>
<tr><th>Activity type</th><th>PRALAB</th><th>TEORIA</th></tr>
<tr><th>Number of credits</th><td>3</td><td>3</td></tr>
<tr><th>Number of groups</th><td>1</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>Coordination</td><td>COMAS RODRIGUEZ, CARLOS</td></tr>
<tr><td>Department</td><td>COMPUTER ENGINEERING AND DIGITAL DESIGN</td></tr>
<tr><td>Teaching load distribution between lectures and independent student work</td><td>30% of the time are lectures (3 hours/week) and 70% is based on autonomous work.</td></tr>
<tr><td>Important information on data processing</td><td>Consult this link for more information.</td></tr>
<tr><td>Language</td><td>English</td></tr>
</table>

BIG DATA PROJECT 2023-24

Teaching staff	E-mail addresses	Credits taught by teacher	Office and hour of attention
COMAS RODRIGUEZ, CARLOS	carles.comas@udl.cat	3	
VIRGILI GOMA, JORDI	jordi.virgili@udl.cat	3	

Subject's extra information

To follow this course, the student should have knowledge of Python and it recommended to have completed the subject Massive Data Processing.

Learning objectives

1. Understand and apply statistical techniques of data mining.
2. Use appropriately statistical packages for data analysis.
3. Understand and apply statistical techniques related to the operation of Big Data.
4. Use appropriately statistical packages for such exploitation.
5. Proposes adequate visualization of the analysed information that facilitates understanding.
6. Know the common applications in the area of big data and be able to develop solutions for these problems.
7. Systematically deploy the set of common techniques for solving big data problems.
8. Communicate effectively the results of the project to technical partners and customers.

Competences

Specific skills

The specific competencies for the exercise of the profession of Computer Engineer, and to which the Master's in Computer Engineering of the UdL gives access are:

*Ability to integrate technologies, applications, services and systems specific to Computer Engineering, in a general way, and in broader and multidisciplinary contexts.

*Capacity for strategic planning, elaboration, direction, coordination, and technical and economic management in the fields of computer engineering related, among others, to: systems, applications, services, networks, infrastructures or computer facilities and centers or software development factories, respecting the adequate fulfillment of the quality and environmental criteria and in multidisciplinary work environments.

*Ability to manage research, development and innovation projects in companies and technology centers, with a guarantee of safety for people and goods, the final quality of products and their approval.

*Ability to model, design, define architecture, deploy, manage, operate, administer and maintain applications, networks, systems, services and computer content.

*Ability to understand and know how to apply the operation and organization of the Internet, new generation network technologies and protocols, component models, intermediate software and services.

*Ability to ensure, manage, audit and certify the quality of developments, processes, systems, services, applications and computer products.

*Ability to design, develop, manage and evaluate certification mechanisms and guarantee security in the processing and access to information in a local or distributed processing system.

*Ability to analyze the information needs that arise in an environment and carry out in all its stages the process of building an information system.

*Ability to design and evaluate operating systems and servers, and applications and systems based on distributed computing.

*Ability to understand and be able to apply advanced knowledge of high-performance computing and numerical or computational methods to engineering problems.

*Ability to design and develop computer systems, applications and services in embedded and ubiquitous systems.

*Ability to apply mathematical, statistical, and artificial intelligence methods to model, design, and develop applications, services, intelligent systems, and knowledge-based systems.

*Ability to use and develop methodologies, methods, techniques, specific use programs, rules and standards of computer computing.

*Ability to conceptualize, design, develop and evaluate the human-computer interaction of computer products, systems, applications and services.

*Ability to create and exploit virtual environments, and to create, manage and distribute multimedia content.

Transversal skills

On the other hand, the University of Lleida itself and the Escola Politècnica Superior establish a series of transversal competencies in the design of the curriculum of all the degrees of the School that include:

*Correction in written oral expression.

*Proficiency in a foreign language.

*Mastery of ICT.

*Respect for the fundamental rights of equality between men and women, the promotion of human rights and the values of a culture of peace and democratic values.

*Ability to plan and organize personal work.

*Ability to consider the socio-economic context as well as sustainability criteria in engineering solutions.

*Ability to convey information, ideas, problems and solutions to both specialized and non-specialized audiences.

*Ability to conceive, design and implement projects and / or provide new solutions, using engineering tools. Have motivation for quality and continuous improvement.

Gender perspective in teaching.

Basic actions

• In the guide and teaching material and in the classroom, make sure that the language is inclusive and not sexist.

• In teaching materials, make sure that the images do not perpetuate gender stereotypes.

• In the teaching material, make sure that the examples and exercises counter gender stereotypes.

• In the teaching material, make sure that the context of the examples and the exercises cover various topics.

• As far as possible, include statements with social and / or gender relevance.

More advanced actions

• In projects, promote the study of some aspect of social and / or gender relevance.

• Explicitly emphasize the social and / or gender relevance in the activities (projects, cases, practices).

• Contextualize the statements of the exams in order to highlight the social and / or gender relevance of the subject.

• Incorporate the variables ‘gender’ and ‘sex’ in the analysis (statistical analysis, solution design, etc.).

• Incorporate in the Teaching Guide objectives related to social and / or gender relevance.

Subject contents

1. Introduction to PCA and the EM algorithm
2. Principal component analysis (PCA).
    1. Data matrices and associated spaces
    2. Principal component analysis
    3. Interpretation and quality of the results of a PCA
3. Expectation-Maximization Algorithm
    1. Maximum Likehood Estimation (MLE)
    2. The Expectation-Maximization (EM) Algorithm
    3. EM for Missing Data
    4. Gaussian Mixture Model (EM clustering)
4. Business Intelligence
    1. Supporting enterprise decisions on big data insights
    2. Smart Data
    3. Visualisation with matplotlib
    4. Databricks, Google Cloud,...
5. Interactive Exploration of Big Data
    1. Apache Spark for querying and interactively exploring great volumes of heterogeneous data
    2. SparkSQL
    3. Spark R and Exploratory Data Analysis

Methodology

All the courses of the Big Data Analytics block (including this one), will be graded by an unique, common, project involving all the courses topics (data gathering, processing, learning, statistics, visualization, etc.).

Students will work on that project from the beginning to the final courses. During regular courses, different topics will be introduced, showing their relation to the common project and how all topics fit together to create a real-world complex task or project.

The three courses forming Big Data Analytics will use the same technological base setup:

- Python as the base programming language.
- Hadoop/Spark (with Java if required)
- Although during the courses other technological suites will be introduced:
    - Scala
    - NodeJS
    - Etc.

Development plan

BIG DATA PROJECT 2023-24

Week	Description	Face-to-Face Activities	Autonomous Student Activities
1	Introduction to CPA and The EM algorithm	Exhibition events and methodology Lecture and participatory classes	Study and Exercises resolution
2	Data matrices and associated spaces and PCA	Lecture and participatory classes	Study and Exercises resolution
3	Bank Holiday (Sant Miquel)		
4	Interpertation and quality of the result of a CPA	Lecture and participatory classes	Study and Exercises resolution
5	Maximum Likehood Estimation (MLE)	Lecture and participatory classes	Study and Exercises resolution
6	The Expectation-Maximization (EM) Algorithm	Lecture and participatory classes	Study and Exercises resolution
7	Oral Presentations	CPA and EM Presentations	Study and Exercises resolution Project Development
8	Apache Spark Intro and Demos Big Data Project Review	Lecture and participatory classes Project Presentations	Cases Study Project Development
9	Big Data Exploration	Lecture and participatory classes	Project Development
10	Big Data Exploration	Lecture and participatory classes	Project Development
11	Big Data Exploration	Lecture and participatory classes	Project Development
12	Big Data Exploration	Lecture and participatory classes	Project Development
13	Business Intelligence	Lecture and participatory classes	Project Development
14	Business Intelligence	Lecture and participatory classes	Project Development
15	Business Intelligence	Project Presentations	Project Development

Evaluation

The assessment for this course is based on continuous evaluation. Depending on the sanitary situation some of these activities could be done as a classroom activity or virtually using tools of the Virtual Campus.

ID	Evaluation activities	%	Dates	Mandatory	I/G (1)
Block 1: OP1	Oral Presentation and project report Resolution of a practical problem (CPA)	25%	Week 7	Y	Group
Block 2: OP2	Oral Presentation and project report Resolution of a practical problem (EM)	25%	Week 7	Y	Group
Block 3: PP	Writing Work Project Planning	15%	Week 10	Y	Group
Block 4: PD	Writing Work Project Deliverable	15%	Week 15	Y	Group

ID	Evaluation activities	%	Dates	Mandatory	I/G (1)
Block 5: OP3	Oral Presentation Project	20%	Week 15	Y	Group

(1) Individual / Group

The final mark will be calculated using the following formula:

Final grade = 0,25*OP1 + 0,25*OP2 + 0,15*PP + 0,15*PD + 0,2*OP3

Bibliography

[Kar15] Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, "Learning Spark: Lightning-Fast Big Data Analysis", O'Reilly, 2015

[Ryz15] Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills, "Advanced Analytics with Spark: Patterns for Learning from Data at Scale", O'Reilly, 2015

[Bae14] Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications"