UNIVERSITÀ DI PAVIA

Anno Accademico 2021/2022

DATA SCIENCE AND BIG DATA ANALYTICS

Enrollment year	2021/2022
Academic year	2021/2022
Regulations	DM270
Academic discipline	ING-INF/05 (DATA PROCESSING SYSTEMS)
Department	DEPARTMENT OF ELECTRICAL,COMPUTER AND BIOMEDICAL ENGINEERING
Course	COMPUTER ENGINEERING
Curriculum	Data Science
Year of study	1°
Period	2nd semester (07/03/2022 - 17/06/2022)
ECTS	6
Lesson hours	60 lesson hours
Language	English
Activity type	ORAL TEST
Teacher	NOCERA ANTONINO (titolare) - 6 ECTS
Prerequisites	Basic Knowledge on database management systems and on data manipulation.
Learning outcomes	This course aims at identifying and provide an in-depth of the main tools and systems for mining big data and, more in general, for working in the current scenario of data science. At the end of this course the student must be able to: understand data, formulate hypothesis on data, transform and model data, define a methodology for the analysis, test and confirm results. The student will learn basic concepts of cloud computing along with the best practice to being able to operate in this context. He will learn the main architectures for processing big-datasets, such as MapReduce. He will understand how to work with NoSQL databases to face the main issues in the big data realm.

Course contents	The life cycle of a data science project. Python libraries for data manipulation and data preparation. Using the main data mining techniques with Python. A brief overview of other popular tools for data science and data mining. The Big Data paradigm and the main issues in this context. NoSQL Databases and MongoDB. Main Cloud architectures for Big Data. Open Stack as an example of open source solution to Cloud Computing. Hadoop, HDFS, Map Reduce and Hive. Overview of Apache Spark. Text mining and Social Network Analysis as case studies.
Teaching methods	This course is organized in lectures, laboratory and cooperative learning. Lectures are used to present theoretical concepts and all the notions about this course. During the lectures, the student will also understand how to apply these notions. Laboratory is used as a mean to allow the student to apply the concepts and techniques shown during the lectures to real-world case studies. Finally, this course leverages cooperative learning, group working and brainstorming. This allows for the development of many transversal skills, such as: team working capabilities, conflict management, and the capability to acquire and exploit different ideas from a team.
Reccomended or required readings	1. Data Science and Big Data Analytics - Discovering, Analyzing, Visualizing and Presenting Data. Wiley. 2. Big Data Fundamentals – Concepts, Drivers & Techniques. Prentice Hall, 2015. 3. Data Mining - Practical Machine Learning Tools and Techniques. Elsevier. 4. Data Mining - Concept and Techniques. Elsevier. 5. Notes provided by Professor
Assessment methods	The assessment consists of an oral discussion about a group project work each student is involved in. The student is in charge of preparing a report about his project work. During the oral discussion the report presented by the student will bel used as a mean to go in-depth in the theoretical concepts used therein. To prepare the report, the student will have to use the tools, introduced during lectures, to extract knowledge form real-life datasets. During the assessment the student must prove a good knowledge of the

	main concepts introduced in this course, to be able to handle the lifecycle of a data science project and to know the main architectures, tools and NoSQL solutions to work in the data analytics and Big Data contexts. The assessment will carefully consider the level of expertise in the use of the tools, the ability of the student to build projects adopting these tools, the level of understanding of the notions taught in this course, the methodological rigor and appropriateness of the technical vocabulary.
Further information	
Sustainable development goals - Agenda 2030	$lbl_legenda_sviluppo_sostenibile