



TECHNICAL SHEET OF THE SUBJECT

Data of the subject	
Subject name	Data Acquisition and Transformation
Subject code	DTC-MBD-517
Mainprogram	Official Master's Degree in Big Data
Involved programs	Máster Universitario en Big Data [First year]
Credits	3,0 ECTS
Type	Obligatoria
Department	Department of Telematics and Computer Sciences

Teacher Information	
Teacher	
Name	Sofia Sánchez Urbano
Department	Department of Telematics and Computer Sciences
EMail	ssurbano@icai.comillas.edu
Teacher	
Name	Alberto José López Espinosa
Department	Department of Telematics and Computer Sciences
EMail	ajlespinosa@icai.comillas.edu

SPECIFIC DATA OF THE SUBJECT

Contextualization of the subject
Contribution to the professional profile of the degree
Ability to collect, clean and structure large amounts of data through information published on the Internet.
Creation of databases from the scratch.
Prerequisites
A basic knowledge of Python is strictly required.

Competencies - Objectives	
Competences	
Conocimientos o contenidos	
CO1	Entender los fundamentos de la analítica de datos y su aplicación en diversas áreas de la inteligencia artificial, destacando la integración en soluciones complejas y multidisciplinares para el análisis avanzado de datos masivos atendiendo a la diversidad de problemas específicos de cada área.



Syllabus 2025 - 2026

CO2	Comprender las técnicas de procesamiento de datos, las arquitecturas y herramientas más habituales y apropiadas para condiciones y requisitos de casos específicos.
Competencias	
CP1	Integrar las arquitecturas, técnicas de inteligencia artificial, análisis avanzado de datos y de visualización y de cumplimiento legal para ofrecer la solución global óptima.
CP4	Implementar las técnicas de procesamiento de datos y usar las herramientas más habituales y apropiadas a las condiciones y requisitos de casos específicos.
CP7	Aplicar conocimientos avanzados en Big Data y analítica de datos para desarrollar soluciones innovadoras en proyectos y en investigación, aportando y evaluando soluciones óptimas para el procesamiento y análisis de datos a gran escala.
Habilidades o destrezas	
HA1	Comunicar de manera oral y escrita con rigor técnico, claridad expositiva y coherencia argumentativa a todo tipo de interlocutores, técnicos y no técnicos.
HA2	Trabajar en equipos de carácter pluridisciplinar y/o internacional y organizar y liderar adecuadamente las dinámicas de grupo.
HA3	Desarrollar las habilidades interpersonales que requieren los entornos profesionales actuales (empatía, tolerancia, respeto, capacidad para aunar intereses contrapuestos).
HA4	Gestionar, organizar y planificar adecuadamente el trabajo y el tiempo, cumpliendo objetivos y estándares de calidad.
HA5	Mantener una formación y aprendizaje continuo y adaptación a los cambios tecnológicos y científicos.

THEMATIC BLOCKS AND CONTENTS

Contents - Thematic Blocks

Unit 01. Data extraction and transformation. ETL process. Grammar and regular expressions. CSV + JSON.

Unit 02. API REST. HTTP requests. Parsing responses with Postman and Python. Intercept XHR requests.

Unit 03. The web as a source of information: HTML and CSS. Webscraping project. Legal aspects webscraping. HTML parsing with BeautifulSoup4 and Selenium. Interacting with JS.

Unit 04. Cleaning and quality of the extracted data. Integrity and normalization. Encoding problems. PDF data extraction. OCR.

Theory:

Lesson 01. Presentation.

Lesson 02. Python & Requests.

Lesson 03. Http requests.



Syllabus 2025 - 2026

Lesson 04. Response formats.

Lesson 05. Web scraping with Bs4.

Lesson 06. Project phases.

Lesson 07. Web scraping with basic Selenium.

Lesson 08. Project structure.

Lesson 09. Logging.

Lesson 10. Data cleaning.

Lesson 11. Recap.

Lesson 12. Career path.

TEACHING METHODOLOGY

General methodological aspects of the subject

In-class Methodology: Activities

Theory: Presentation of the basic concepts by the teacher and practical recommendations to carry them out. Student participation will be encouraged to generate a debate environments in class. Practice: Introduction to the practice associated with each of the theoretical classes and monitoring it in class. Cooperation between students to solve problems will be encouraged.	CO1, CO2, CP1, CP4, CP7, HA1, HA2, HA3, HA4, HA5
---	---

Non-Presential Methodology: Activities

Practice: Resolution of the practices presented in class.	CO1, CO2, CP1, CP4, CP7, HA1, HA2, HA3, HA4, HA5
---	---

SUMMARY STUDENT WORKING HOURS

CLASSROOM HOURS

Clases magistrales expositivas y participativas: Exposición de contenidos fundamentales por parte del profesor impulsando la reflexión y participación del alumno.	Ejercicios prácticos y resolución de problemas: Sesiones prácticas con uso de software: Actividad formativa con ordenador que, bajo la guía del profesor-tutor, fomenta el aprendizaje autónomo y/o cooperativo del alumno mediante la ejecución de programas para la consecución de los objetivos marcados
---	--



Syllabus 2025 - 2026

14.00	16.00	
NON-PRESENTIAL HOURS		
Estudio personal: Reflexión y análisis individual de los contenidos teóricos y prácticos de las materias y/o asignaturas del Master	Trabajos: Los alumnos tendrán que hacer trabajos breves (individuales y/o en grupo), por indicación del profesor	Proyectos: Los alumnos tendrán que hacer trabajos de tamaño medio o grande (individuales y/o en grupo), por indicación del profesor
10.00	25.00	25.00
ECTS CREDITS: 3,0 (90,00 hours)		

EVALUATION AND CRITERIA

The use of AI to produce full assignments or substantial parts thereof, without proper citation of the source or tool used, or without explicit permission in the assignment instructions, will be considered plagiarism and therefore subject to the University's General Regulations.

Evaluation activities	Evaluation criteria	Weight
Final exam.	Selection of correct answers among given options. The aim is to understand theoretical concepts seen in class.	30
Script practices.	Creation of scripts and individual requests. The cleanliness of the code, its effectiveness and the level of documentation will be assessed.	20 %
Project practices.	Creation of projects and execution flows. The cleanliness of the code, its effectiveness and the level of documentation will be assessed.	50 %

Ratings

To pass the course the exam score must be greater than 5 out of 10.

To pass the course, the average score for the practices must be greater than 5 out of 10.

Attendance at 85% of the lessons is mandatory to be able to take the exam according to Article 93 of the General Regulations of the University of Comillas.

In this subject, the use of AI is allowed for planning and testing, but never as a final solution. AI may be used for activities prior to the task, such as brainstorming, outlining, and initial research.

This level (Level 2*) focuses on the effective use of AI for planning, synthesis, and idea generation, but assessments must emphasize the student's ability to develop and refine these ideas independently.



Syllabus 2025 - 2026

You may use AI to plan, develop ideas, and research. Your final submission must demonstrate how you developed and refined these ideas.

The use of AI to produce complete assignments or significant parts of them, without citing the source or tool used, or without explicit permission in the assignment description, will be considered plagiarism and will be regulated according to the General Regulations of the University.

*Perkins, Furze, Roe & MacVaugh (2024). The AI Assessment Scale

WORK PLAN AND SCHEDULE

Activities	Date of realization	Delivery date
Lesson 01. Presentation. Lesson 02. Python & Requests. Practice 01. Postman. Practice 02. Aemet.	week 01	week 03
Lesson 03. Http Requests. Lesson 04. Response formats. Practice 03. NBA. Practice 04. Wikipedia Continents.	week 02	week 04
Lesson 05. Web scraping with Bs4. Practice 05. Wikipedia Comillas. Day November 03th no lesson.	week 03	week 05
Lesson 06. Project phases. Practice 06. Wikipedia Universities. Day November 10th no lesson.	week 04	week 06
Lesson 07. Web scraping with basic Selenium. Lesson 08. Project structure. Practice 07. CNMC. Practice 08. Borme I.	week 05	week 07
Lesson 09. Logging.		



Syllabus
2025 - 2026

Lesson 10. Data cleaning. Practice 09. Borme II. Practice 10. Borme III.	week 06	week 08
Lesson 11. Review. Lesson 12. Professional path.	week 07	week 09

BIBLIOGRAPHY AND RESOURCES

Basic Bibliography
Theory lessons. Official documentation of open source libraries: requests, bs4, selenium & chromedriver.
Complementary Bibliography
Web scraping with Python, Ryan Mitchel, O’reilly. ISBN: 1491985577 Python Crash Course, Eric Matthes. ISBN: 1593279280

