GUÍA DOCENTE
2021 - 2022

FICHA TÉCNICA DE LA ASIGNATURA

Datos de la asignatura	
Nombre completo	Fundamentos matemáticos del análisis de datos
Código	DMA-MBD-513
Impartido en	Máster en Big Data. Tec. y Analítica Avanzada/Master in Big Data Technologies and Advanced Analytics [Primer Curso]
Nivel	Master
Cuatrimestre	Semestral
Créditos	3,0 ECTS
Carácter	Obligatoria
Departamento / Área	Departamento de Matemática Aplicada
Responsable	Fernando San Segundo Barahona

Datos del profesorado	
Profesor	
Nombre	Fernando San Segundo Barahona
Departamento / Área	Departamento de Matemática Aplicada
Despacho	Alberto Aguilera 25 [D-201] 2377
Correo electrónico	fsansegundo@icai.comillas.edu
Profesores de laboratorio	
Profesor	
Nombre	Santiago Cano Casanova
Departamento / Área	Departamento de Matemática Aplicada
Despacho	Alberto Aguilera 25 [D-204] 2382
Correo electrónico	scano@icai.comillas.edu

DATOS ESPECÍFICOS DE LA ASIGNATURA

Contextualización de la asignatura

Aportación al perfil profesional de la titulación

This course is a primer in Statistics with a special emphasis on its mathematical foundations. The subject aims to introduce the student to the language of Statistics and to basic, but fundamental, concepts such as distributions, probability inference and statistical models. Our approach is practical and conceptual, using the computational tools as a mean to gain an insight into the central ideas of Statistics.

The subject also provides a first course of the R language by combining exposure of the main concepts in Statistics and tutorial R sessions. Our focus will be in a modern flavor of the R language provided by the *tidyverse* library and the use of the pipe operator to describe a data

analysis as a workflow.

We will also introduce some auxiliary tools that form an important part of the R data analysis ecosystem. Git and GitHub have become a natural setting for collaborative work and communication inside a team. Rmarkdown is used as the main format for our work and to increase the productivity and reproducibility of the analysis.

Prerequisitos

Basic knowledge of Calculus and Algebra is required (understand and manipulate equations, manipulate exponents and logarithms using their basic rules, full understanding of functions and inverse functions, understand limits, derivatives and integrals, know rules for product and summation, etc.) Basic knowledge of Statistics (descriptive statistics, discrete and continuous probability distribution models, sampling and basics of statistical inference) is highly recommended but not required.

Basic knowledge of Programming languages is required, ideally in R or Python.

Competencias - Objetivos

Competencias

Competences[1]	
General competences	
CG1.	Have acquired advanced knowledge and demonstrated, in a research and technological or highly specialized context, a detailed and well-founded understanding of the theoretical and practical aspects, as well as of the work methodology in one or more fields of study. *Haber adquirido conocimientos avanzados y demostrado, en un contexto de investigación científica y tecnológica o altamente especializado, una comprensión detallada y fundamentada de los aspectos teóricos y prácticos y de la metodología de trabajo en uno o más campos de estudio.*
CG2.	Know how to apply and integrate their knowledge, understanding, scientific rationale, and problem-solving skills to new and imprecisely defined environments, including highly specialized multidisciplinary research and professional contexts. *Saber aplicar e integrar sus conocimientos, la comprensión de estos, su fundamentación científica y sus capacidades de resolución de problemas en entornos nuevos y definidos de forma imprecisa, incluyendo contextos de carácter multidisciplinar tanto investigadores como profesionales altamente especializados.*
CG5.	Be able to transmit in a clear and unambiguous manner, to specialist and non-specialist audiences, results from scientific and technological research or state-of-the-art innovation, as well as the most relevant foundations that support them. *Saber transmitir de un modo claro y sin ambigüedades, a un público especializado o no, resultados procedentes de la investigación científica y tecnológica o del ámbito de la innovación más avanzada, así como los fundamentos más relevantes sobre los que se sustentan.*
CG6.	Have developed sufficient autonomy to participate in research projects and scientific or technological collaborations within their thematic area, in interdisciplinary contexts and, where appropriate, with a high knowledge transfer component.

	Haber desarrollado la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro de su ámbito temático, en contextos interdisciplinares y, en su caso, con una alta componente de transferencia del conocimiento.
CG7.	Being able to take responsibility for their own professional development and their specialization in one or more fields of study. *Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio.*
Specific competences	
CE6.	Being able to use the tools provided by the R ecosystem to perform a complete data analysis workflow. *Ser capaz de aplicar las herramientas del ecosistema de R para implementar un proceso de análisis de datos completo.*
CE7.	Have acquired advanced knowledge of the statistical concepts that lay the foundations of data Analysis. *Adquirir un nivel avanzado de comprensión de los conceptos estadísticos que sirven de fundamento al análisis de datos.*

[1] Competences in English are a free translation of the official Spanish version.

Resultados de Aprendizaje

Learning outcomes	
By the end of the course students should:	
RA1.	Be familiar with the R language and the tidyverse components.
RA2.	Be able to import data sets into R from common file formats, tidy the data and explore them.
RA3.	Have developed visualization and communication skills using ggplot and Rmarkdown.
RA4.	Be familiar with Git and services such as GitHub and the role they can play in the context of data analysis.
RA5.	Understand the concepts of Probability theory involved in the main results of classical Statistics.
RA6.	Be able to construct and understand statistical inferences, in particular confidence intervals and hypothesis test.
RA7.	Be familiar with the theory and R implementation of simple statistical models, such as univariate linear regression and logistic models.
RA8.	Understand the notions of goodness of fit and model diagnosis, applying them to the simple models in the previous

	item.
RA9.	Be able to implement a simulation approach to a simple random process.
RA10.	Be prepared to move on with R into the study of Machine Learning.

BLOQUES TEMÁTICOS Y CONTENIDOS

Contenidos – Bloques Temáticos

Theory

Unit 1. Types of Variables and Basic Data Structures

1.1 The basic structure of a data analysis.

1.2 Types of variables and tabular data.

1.3 Basic computational tools and data skills.

Unit 2. Graphics and Exploratory Data Analysis

2.1 The correspondence between basic variable type and graphs.

2.2 The Grammar of Graphics as implemented by ggplot.

2.3 Statistical summaries of data and dplyr.

Unit 3. Distributions

3.1 Distributions as theoretical models.

3.2 Measures of position and spread.

Unit 4. Probability

4.1 Discrete Probability. Laplace's Rule and Bayes Theorem.

4.2 Axioms of Probability.

4.3 Contingency Tables.

Unit 5. Random Variables

5.1 Discrete Random Variables.

5.2 Binomial Variables.

5.3 Continuous Random Variables.

5.4 Normal Random Variables.

Unit 6. Inference

6.1 Sampling Distributions and Central Limit Theorem.

6.2 Confidence Intervals.

6.3 Hypothesis Test.

Unit 7. Linear and Logistic Regression

7.1 Covariance.

7.2 Linear Regression Model.

7.3 Logistic Regression Model.

Unit 8. Bayesian Statistics

8.1 Introduction to Bayesian Statistics.

8.2 Monte Carlo Methods and Bayesian Software.

Laboratory

Lab 1. Exploratory Data Analysis

In the first lab we will check the student's software setup for the course and introduce the expected use of Git and Rmarkdown. The first assignment deals with data import and graphical exploration with ggplot.

Lab 2. Tidy Data

The main verb/functions of dplyr will be applied to illustrate the data wrangling part of the data analysis pipeline. Use of date/time data and strings will also be addressed in this lab session and second assignment.

Lab 3. Program

One of the goals of this lab is to introduce the student into simulation techniques with R. The third assignment will focus in the programming structures that support simulation while diving deeper into dplyr, tidyr and the main components of the tidyverse.

Lab 4. Models

Statistical modeling with R is presented in this lab session, using linear and logistic regression as examples. The fourth assignment applies all the tools in the course labs to a realistic data set.

Final project
The project offers the students the opportunity to showcase their skills by applying the methods in the course to complete a full data analysis workflow: import, tidy, transform, visualize, model and communicate.

METODOLOGÍA DOCENTE

Aspectos metodológicos generales de la asignatura

This course has a practical "hands on, head first" approach to Statistics. We put conceptual understanding in the first place, and use the computational tools in R to make the concepts come alive in the classroom and in the students practice.

Metodología Presencial: Actividades

In-class activities	Competences
• **Lectures:** The lecturer will introduce the fundamental concepts of each unit, along with some practical recommendations, and will go through worked examples to support the explanation. Active participation will be encouraged by raising open questions to foster discussion and by proposing online quizzes and short application exercises to be solved in class either on paper or using a software package.	CG1, CG7, CE6, CE7
• **Lab sessions:** Under the instructors supervision, students, divided in small groups, will apply the concepts and techniques covered in the lectures to walk through the steps of the data analysis workflow.	CG1, CG2, CG5, CG6, CG7, CE6, CE7
• **Tutoring** for groups or individual students will be organized upon request.	–

Metodología No presencial: Actividades

Out-of-class activities	Competences
• Personal study of the course material and resolution of the proposed exercises.	CG1, CG7, CE6, CE7
• Lab session preparation to make the most of in-class time.	CG1
• Lab results analysis and report writing.	CG2, CG5, CE6, CE7
• Development of a final project in small groups.	CG1, CG2, CG5, CG6, CG7, CE6, CE7

RESUMEN HORAS DE TRABAJO DEL ALUMNO

STUDENT WORK-TIME SUMMARY			
IN-CLASS HOURS			
Lectures	**Lab sessions**		**Assessment**
20	8		2
OUT-OF-CLASS HOURS			
Self-study	**Lab preparation**	**Lab report writing**	**Final project**
20	8	8	24
		ECTS credits:	**3 (90 hours)**

EVALUACIÓN Y CRITERIOS DE CALIFICACIÓN

Assessment activities	Grading criteria	Weight
Quizzes	▪ Understanding of the theoretical concepts.	10%
Final exam	▪ Understanding of the theoretical concepts. ▪ Application of these concepts to problem-solving. ▪ Ability to use R and Git to implement a data analysis workflow ▪ Critical analysis of numerical exercises' results.	30%
Lab assignments	▪ Application of theoretical concepts to real problem-solving. ▪ Ability to use the R ecosystem and Git. ▪ Written communication, modeling and visualization skills.	35%
Final project	▪ Problem analysis. ▪ Quality of the proposed solution. ▪ Teamwork. ▪ Written communication, modeling and visualization skills. ▪ Use of collaborative software (Git) ▪ There will be an intra-group evaluation method to differentiate among team members.	25%

Calificaciones

Grading

Regular assessment

- **Theory** will account for 40%, of which:

- Quizzes: 10%
- Final exam: 30%

- **Lab** will account for the remaining 60%, of which:

- Lab assignments: 35%
- Final project: 25%

To pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 4 out of 10 points, and the laboratory mark (the weighted average of the assignments and the final project) must be at least 5 out of 10 points. Otherwise, the final grade will be the lowest of the three marks.

Retake

Lab marks will be preserved as long as the weighted average of the assignments and the final project results in a passing grade. Otherwise, a new project will have to be developed and handed in. In addition, all students will take a final exam. The resulting grade will be computed as follows:

- **Theory** will account for 40%, of which:

- Quizzes: 10%
- Final exam: 30%

- **Lab** will account for the remaining 60%, of which:

- If the student passed the lab during regular assessment
  - Lab assignments: 35%
  - Final project: 25%
- Otherwise
  - Final project: 60%

As in the regular assessment period, to pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 4 out of 10 points, and the mark of the laboratory must be at least 5 out of 10 points. Otherwise, the final grade will be the lowest of the three marks.

Course rules

- Class attendance is mandatory according to Article 93 of the General Regulations (Reglamento General) of Comillas Pontifical University and Article 6 of the Academic Rules (Normas Académicas) of the ICAI School of Engineering. Not complying with this requirement may have the following consequences:
  - Students who fail to attend more than 15% of the lectures may be denied the right to take the final exam during the regular assessment period.
  - Regarding laboratory, absence to more than 15% of the sessions can result in losing the right to take the final exam of the regular assessment period and the retake. Missed sessions must be made up for credit.
- Students who commit an irregularity in any graded activity will receive a mark of zero in the activity and disciplinary procedure will follow (cf. Article 168 of the General Regulations (Reglamento General) of Comillas Pontifical University).

PLAN DE TRABAJO Y CRONOGRAMA

Actividades

In and out-of-class activities	Date/Periodicity	Deadline
Final exam	After the lecture period	–
Lab sessions	Weeks 1 to 4	–
Review and self-study of the concepts covered in the lectures	After each lesson	–
Lab preparation	Before every lab session	–
Lab report writing	–	One week after the end of each session
Final project	From week 3	Last week

	In-class activities			Out-of-class activities				Learning outcomes
Week	Time [h]	Lecture	Laboratory	Time [h]	Self-study	Lab preparation and report writing	Other activities	Code
1	2	Course overview (0.5h) Software setup and check (0.5h) Types of Variables and Basic Data Structures (1h)		2	Review and self-study (2h)			RA1, RA2. RA4
	2	Graphics and Exploratory Data Analysis (2h)		2	Review and self-study (2h)			RA3
						Lab		

	2		Lab 1 (2h)	4		preparation (1.5h) Report writing (2h)	Quiz 1 (0.5 h)	RA1, RA2. RA3, RA4
2	2	Distributions (2h)		2	Review and self-study (2h)			RA3, RA5, RA9
	2	Probability (2h)		2	Review and self-study (2h)			RA3, RA5, RA9
	2		Lab 2 (2h)	4		Lab preparation (1.5h) Report writing (2h)	Quiz 2 (0.5 h)	RA3, RA5, RA9
3	2	Random Variables (2h)		6	Review and self-study (2h)		Final project (4h)	RA3, RA6, RA9
	2	Inference (2h)		7	Review and self-study (3h)		Final project (4h)	RA3, RA6, RA9
	2		Lab 3 (2h)	4		Lab preparation (1.5h) Report writing (2h)	Quiz 2 (0.5 h)	RA3, RA6, RA9
4	2	Linear and Logistic Regression (2h)		6	Review and self-study (2h)		Final project (4h)	RA3, RA7, RA8, RA9, RA10
	2	Bayesian Statistics (2h)		7	Review and self-study (3h)		Final project (4h)	RA3, RA7, RA8, RA9, RA10
	2		Lab 4 (2h)	4		Lab preparation (1,5h) Report writing (2h)	Quiz 2 (0.5 h)	RA3, RA7, RA8, RA9, RA10

5	2	Final exam (2h)		10	Review and self-study (2h)		Final project (8h)	RA1 to RA10

BIBLIOGRAFÍA Y RECURSOS

Bibliografía Básica

Basic bibliography

- Slides prepared by the lecturer (available in Moodlerooms).
- Grolemund, G. and Wickham, H. (2017). R for Data Science. O'Reilly. Freely available online at r4ds.had.co.nz. ISBN-13: 978-1491910399
- Peng, R. (2016) R Programming for Data Science. Freely available online at bookdown.org/rdpeng/rprogdatascience/.
- R A. Irizarry (2021) Introduction to Data Science. Freely available online at rafalab.github.io/dsbook ISBN-13: 978-0367357986
- Kabacoff, R., (2021). R in Action. 3rd Edition. Manning Publications. ISBN-13: 978-1617296055
- Kurt W. (2019) Bayesian Statistics the Fun Way. No Starch Press. ISBN-13:

Bibliografía Complementaria

Complementary bibliography

- Xie Y. (2019) R Markdown: The Definitive Guide. Freely available online at bookdown.org/yihui/rmarkdown . ISBN-13: 978 -1138359338
- Zumel I. and Mount J. (2019) Practical Data Science with R. Manning Publications. ISBN-13: 978-1617295874
- Kruschke, J., (2014), Doing Bayesian Data Analysis, 2nd Edition. Academic Press. ISBN-13: 978-0124058880