

**FITXA IDENTIFICATIVA****Dades de l'Assignatura**

Codi	36423
Nom	Tractament de les dades
Cicle	Grau
Crèdits ECTS	6.0
Curs acadèmic	2022 - 2023

Titulació/titulacions

Titulació	Centre	Curs	Període
1406 - Grau en Ciència de Dades	Escola Tècnica Superior d'Enginyeria	1	Segon quadrimestre

Matèries

Titulació	Matèria	Caràcter
1406 - Grau en Ciència de Dades	8 - Gestió de la informació	Obligatòria

Coordinació

Nom	Departament
GOMEZ CHOVA, LUIS	242 - Enginyeria Electrònica
MARTINEZ SOBER, MARCELINO	242 - Enginyeria Electrònica

RESUM

El científic o la científica de dades s'enfronta a un conjunt de dades de molt diversa procedència, format, organització, codificació, etc. La correcta adquisició, organització, eliminació de possibles dades errònies (outliers), imputació de dades mancants (missing data imputation), transformació de les dades, selecció de les característiques més rellevants d'un conjunt d'alta dimensionalitat (feature selection), eliminació de dades redundants, etc. és una de les etapes més costoses d'un problema d'anàlisi de dades. Aquesta etapa és crucial per al correcte tractament del problema i la fiabilitat i solidesa dels resultats obtinguts en etapes posteriors d'anàlisi (selecció de models, classificadors, agrupament, estimació, contrastos d'hipòtesis, etc.). Totes aquestes tasques seran abordades en l'assignatura obligatòria, 36423 Tractament de Dades, que s'imparteix en el segon quadrimestre de primer curs.

Les classes de teoria s'impartiran en castellà i les classes pràctiques i de laboratori segons consta en la fitxa de l'assignatura disponible en la web del grau.



CONEIXEMENTS PREVIS

Relació amb altres assignatures de la mateixa titulació

No heu especificat les restriccions de matrícula amb altres assignatures del pla d'estudis.

Altres tipus de requisits

Es recomana haver superat l'assignatura Dades, Societat i Universitat que s'imparteix en el primer quadrimestre del primer curs del grau.

COMPETÈNCIES

1406 - Grau en Ciència de Dades

- (CG06) Capacitat d'accés i gestió de la informació en diferents formats per a la seva posterior anàlisi amb la finalitat d'obtenir coneixement a partir de dades.
- (CT03) Habilitat per defensar el seu treball amb rigor i arguments, exposant-ho de forma adequada i precisa, recolzant-se en els mitjans necessaris.
- (CT04) Ser responsables del seu propi desenvolupament professional i de la seva especialització, aplicant els coneixements adquirits en la identificació de sortides professionals i jaciments d'ocupació.
- (CE02) Conèixer i aplicar de forma metodològica les tècniques de programació i l'algorísmia necessàries per al processament eficient d'informació i la resolució informàtica de problemes que utilitzen grans volums de dades.
- (CE06) Capacitat per representar i visualitzar conjunts de dades per a l'extracció de coneixement.
- (CE11) Capacitat per dissenyar i implementar la presa de dades, la seva integració, transformació, selecció, comprovació de la seva qualitat i veracitat a partir de diferents fonts, tenint en compte el seu caràcter, heterogeneïtat i variabilitat.
- (CE13) Saber dissenyar, aplicar i avaluar algorismes de Ciència de Dades per a la resolució de problemes complexos.
- (CB3) Que els estudiants tinguin la capacitat d'arreglar i interpretar dades rellevants (normalment dins de la seua àrea d'estudi) per emetre judicis que incloguen una reflexió sobre temes rellevants d'índole social, científica o ètica.

RESULTATS DE L'APRENTATGE

- Conèixer les tècniques i algoritmes per preprocessar i extraure les característiques més importants d'un conjunt de dades. (CB3, CG06, CT03, CT04, CE02, CE06)



- Determinar les transformacions més adequades per al problema a resoldre. (CB3, CT03, CT04, CE11)
- Saber caracteritzar les dades atípics o outliers. (CB3, CE13, CT03, CT04)
- Conèixer quins problemes es tenen quan tractem amb conjunts altament desbalancejat. (CB3, CT03, CT04, CE13)

Com a conseqüència dels resultats d'aprenentatge adquirits, el / la estudiant adquirirà les següents destreses:

- Ser capaç de carregar qualsevol fitxer de dades de tota índole (text, formats tabulats, etc.).
- Netejar i completar el contingut del conjunt de dades un cop carregat (correcció d'errors tipogràfics, imputació de dades perdudes, etc ..)
- Escollir el tipus de dada adequat depenent de la seva naturalesa (sencer, real, factor, text, etc ..)
- Filtrar mostres, seleccionar característiques, i crear noves a partir de formats tabulats com el data frame.
- Realitzar una caracterització de les dades depenent de la seva tipologia.
- Crear visualitzacions de dades bàsiques per tal d'extreure conclusions preliminars de les dades.

DESCRIPCIÓ DE CONTINGUTS

1. Introducció al tractament de dades

- 1.1. Per què analitzar dades.
- 1.2. Visió global d'un problema de tractament de dades.

2. Obtenció de dades

- 2.1. Introducció.
- 2.2. Introducció de dades.
- 2.3. Repositoris de dades.
- 2.4. Formats de fitxers de dades.
- 2.5. Fusió de dades procedents de diferents fonts.



2.6. Accés a bases de dades.

3. Visualització de dades.

- 3.1. Gràfics explicatius i exploratoris.
- 3.2. Sistemes gràfics en R: base, grid, lattice, ggplot2.
- 3.3. Llibreria ggplot2: representacions bàsiques.

4. Preparació de les dades

- 4.1. Estructura d'un conjunt de dades per a la seva anàlisi: operacions bàsiques.
- 4.2. Manipulació de dades. Llibreria tidyr.
- 4.3. Maneig de dades. Llibreria dplyr.

5. Anàlisi exploratòria de dades I. Definicions

- 5.1. Explorant un nou dataset.
- 5.2. Caracterització de variables.
- 5.3. Visualització de relacions entre variables.

6. Anàlisi Exploratori de dades II. Anomalies.

- 6.1. Anomalies en variables numèriques: Outliers. Caracterització d'outliers. Mètodes de detecció.
- 6.2. Anomalies en variables numèriques: dades perdudes i absents

7. Treballant amb dades de tipus text

- 7.1. Introducció
- 7.2. Bases de l'anàlisi de dades de text.
- 7.3. Funcions bàsiques per al maneig de caràcters en R.
- 7.4. Expressions regulars.
- 7.5 Codificació de caràcters: ascii vs Unicode.

8. Pràctiques de Tractament de les dades

En aquest bloc es presentaran una sèrie de supòsits pràctics a manera de pràctiques de laboratori duts a terme en aula d'informàtica.

- Pràctica 1. Importació de dades.
- Pràctica 2. Visualització de dades.
- Pràctica 3. Preparació de dades amb tidyr.
- Pràctica 4. Maneig de dades amb dplyr.
- Pràctica 5. Dades anòmals.
- Pràctica 6. Anàlisi exploratòria de dades.
- Pràctica 7. Anàlisi completa d'un conjunt de dades.

**VOLUM DE TREBALL**

ACTIVITAT	Hores	% Presencial
Classes de teoria	34,00	100
Pràctiques en laboratori	20,00	100
Pràctiques en aula	6,00	100
Elaboració de treballs en grup	5,00	0
Elaboració de treballs individuals	10,00	0
Estudi i treball autònom	15,00	0
Lectures de material complementari	5,00	0
Preparació d'activitats d'avaluació	15,00	0
Preparació de classes de teoria	15,00	0
Preparació de classes pràctiques i de problemes	15,00	0
Resolució de qüestionaris on-line	10,00	0
TOTAL	150,00	

METODOLOGIA DOCENT

Les classes combinaran el contingut teòric i pràctic.

MD1 - Activitats teòriques. Desenvolupament expositiu de la matèria amb la participació de l'estudiant en la resolució de qüestions puntuals. Realització de qüestionaris individuals d'avaluació.

En les activitats teòriques de caràcter presencial es desenvoluparan els temes de l'assignatura proporcionant una visió global i integradora, analitzant amb major detall els aspectes clau i de major complexitat, fomentant, en tot moment, la participació de l'alumnat (CB03, CT03).

MD2 - Activitats pràctiques. Aprenentatge mitjançant resolució de problemes, exercicis i casos d'estudi a través dels quals s'adquireixen competències sobre els diferents aspectes de la matèria. (CB03, CG06, CE02, CE06, CE11, CE13)

Les activitats teòriques es complementen amb activitats pràctiques amb l'objectiu d'aplicar els conceptes bàsics i ampliar-los amb el coneixement i l'experiència que es vagin adquirint durant la realització dels treballs proposats.

MD4 - Treballs en laboratori i / o aula ordinador. Aprenentatge mitjançant la realització d'activitats desenvolupades de forma individual o en grups reduïts i dutes a terme en laboratoris i / o aules d'ordinador. (CB03, CG06, CE02, CE06, CE11, CE13)

A més de les activitats presencials, els estudiants hauran de realitzar tasques personals (fora de l'aula) sobre: qüestions i problemes, així com la preparació de classes i exàmens (estudi). Aquestes tasques es realitzaran principalment de manera individual, per tal de potenciar el treball autònom, però addicionalment s'inclouran treballs, especialment la preparació i resolució de pràctiques laboratori, que requereixin la participació de petits grups d'estudiants (2-3) per fomentar la capacitat d'integració en grups de treball.



S'utilitzarà les plataformes d'e-learning (Aula Virtual) de la Universitat de València, Microsoft Teams i Blackboard Collaborate, com a suport de comunicació amb l'alumnat. A través d'elles es tindrà accés al material didàctic utilitzat en classe, així com els problemes i exercicis a resoldre.

AVALUACIÓ

L'avaluació de l'aprenentatge dels coneixements i competències obtinguts pels estudiants es farà de manera contínua al llarg del curs, i comptarà dels següents blocs d'avaluació:

Primera i segona convocatòries:

SE1 - Prova objectiva, consistent en un o diversos exàmens que consten tant de qüestions teòriques-pràctiques com de problemes (avaluació de competències CB03, CG06, CT03, CE02, CE06, CE11, CE13) (48%) (Nota: Tots els percentatges estan referits a la nota final)

SE1-1 (30%) Examen de teoria

SE1-2 (18%) Examen de laboratori)

SE2 - Avaluació de les pràctiques a partir de l'elaboració de treballs / memòries i / o exposicions orals (avaluació de competències CB03, CG06, CT03, CT04, CE02, CE06, CE11, CE13) (32%).

SE2-1 (20%) Realització d'un miniproyecto consistent en l'anàlisi completa d'un conjunt de dades.

SE2-2 (12%) Sessions Laboratori (Activitat NO RECUPERABLE)

SE3 - Avaluació contínua de cada alumne, basada en la participació i grau d'implicació de l'alumne en el procés d'ensenyament-aprenentatge, tenint en compte la regular assistència a les activitats presencials previstes i la resolució de qüestions i problemes proposats periòdicament. (20%)

SE3-1 (5%) Assistència regular a les activitats presencials previstes (avaluació de competències CB04, CG01). (Activitat NO RECUPERABLE)

SE3-2 (15%) Resolució de qüestions i problemes proposats (avaluació de competències CB02, CB04, CG01, CT03). (Activitat NO RECUPERABLE)

La nota final de l'assignatura es calcularà com la mitjana ponderada de cada un dels apartats anteriors, d'acord amb el següent criteri: SE-1 (48%), SE-2 (32%), SE-3 (20%).

Consideracions particulars sobre l'avaluació:

- És necessari obtenir una qualificació mínima de 4 (sobre 10) en els apartats d'avaluació SE1-1 (examen Teoria), SE1-2 (examen de laboratori) i SE2-1 (mini projecte), per fer la mitjana.
- Les activitats SE2-2, SE3-1 i SE3-2 no són recuperables.



- L'activitat SE1-2 es realitzarà a l'acabar l'examen de teoria el dia de la convocatòria oficial.

En qualsevol cas, el sistema d'avaluació es regirà pel que estableix el Reglament d'Avaluació i Qualificació de la Universitat de València per a Grans i Mestres:

<https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdictoSeleccionado=5639>

REFERÈNCIES

Bàsiques

- R.K.Pearson (2018) Exploratory Data Analysis Using R. CRC.
- H. Wickham, G. Grolemund. (2016) R for data Science. O'Reilly Media Inc.
<http://r4ds.had.co.nz/>
- B. S. Baumer, D. T. Kaplan, N. J. Horton (2017) Modern Data Science with R. Boca Raton : Taylor & Francis CRC Press.
- R. Buttrech y, L.R. Whitaker (2018). A data scientist's guide to acquiring, cleaning and managing data in R . Wiley. (disponible e-libro)
- W. Graham, (2017). The Essentials of Data Science: Knowledge Discovery Using R. Chapman and Hall/CRC. (disponible e-libro)

Complementàries

- L. Han, M. Kamber, and J. Pei. (2012) Data Mining Concepts and Techniques (third Edition). (disponible e-libro)
- N. Zumel and J. Mount (2014). Practical Data Science with R. Manning Publications Co.
- A.Cirillo (2017) R Data Mining. Pack Publishing (disponible e-libro)
- C.Aggarwal (2015) Data mining: the textbook. Springer (disponible e-libro)