FICHA IDENTIFICATIVA

Datos de la Asignatura

Código	36423
Nombre	Tratamiento de los datos
Ciclo	Grado
Créditos ECTS	6.0
Curso académico	2021 - 2022

Titulación(es)

Titulación	Centro	Curso	Periodo
1406 - Grado en Ciencia de Datos	Escuela Técnica Superior de Ingeniería	1	Segundo cuatrimestre

Materias

Titulación	Materia	Caracter
1406 - Grado en Ciencia de Datos	8 - Gestión de la información	Obligatoria

Coordinación

Nombre	Departamento
GOMEZ CHOVA, LUIS	242 - Ingeniería Electrónica
MARTINEZ SOBER, MARCELINO	242 - Ingeniería Electrónica

RESUMEN

El científico o la científica de datos se enfrenta a un conjunto de datos de muy diversa procedencia, formato, organización, codificación, etc. La correcta adquisición, organización, eliminación de posibles datos erróneos (*outliers*), imputación de datos faltantes (*missing data imputation*), transformación de los datos, selección de las características más relevante de un conjunto de alta dimensionalidad (*feature selection*), eliminación de datos redundantes, etc. es una de las etapas más costosas de un problema de análisis de datos. Esta etapa es crucial para el correcto tratamiento del problema y la fiabilidad y solidez de los resultados obtenidos en etapas posteriores de análisis (selección de modelos, clasificadores, agrupamiento, estimación, contrastes de hipótesis, etc). Todas estas tareas serán abordadas en la asignatura obligatoria, 36423 Tratamiento de Datos, que se imparte en el segundo cuatrimestre de primer curso.

Las clases de teoría se impartirán en castellano y las clases prácticas y de laboratorio según consta en la ficha de la asignatura disponible en la web del grado.

CONOCIMIENTOS PREVIOS

Relación con otras asignaturas de la misma titulación

No se han especificado restricciones de matrícula con otras asignaturas del plan de estudios.

Otros tipos de requisitos

Se recomienda haber superado la asignatura Datos, Ciencia y Sociedad que se imparte en el primer cuatrimestre del primer curso del grado.

COMPETENCIAS (RD 1393/2007) // RESULTADOS DEL APRENDIZAJE (RD 822/2021)

1406 - Grado en Ciencia de Datos

- (CG06) Capacidad de acceso y gestión de la información en diferentes formatos para su posterior análisis con el fin de obtener conocimiento a partir de datos.
- (CT03) Habilidad para defender su trabajo con rigor y argumentos, exponiéndolo de forma adecuada y precisa, apoyándose en los medios necesarios.
- (CT04) Ser responsables de su propio desarrollo profesional y de su especialización, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.
- (CE02) Conocer y aplicar de forma metodológica las técnicas de programación y la algoritmia necesarias para el procesado eficiente de información y la resolución informática de problemas que utilizan grandes volúmenes de datos.
- (CE06) Capacidad para representar y visualizar conjuntos de datos para la extracción de conocimiento.
- (CE11) Capacidad para diseñar e implementar la toma de datos, su integración, transformación, selección, comprobación de su calidad y veracidad a partir de distintas fuentes, teniendo en cuenta su carácter, heterogeneidad y variabilidad.
- (CE13) Saber diseñar, aplicar y evaluar algoritmos de Ciencia de Datos para la resolución de problemas complejos.
- (CB3) Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.

RESULTADOS DE APRENDIZAJE (RD 1393/2007) // SIN CONTENIDO (RD 822/2021)

- Conocer las técnicas y algoritmos para preprocesar y extraer las características más importantes de un conjunto de datos. (CB3,CG06,CT03,CT04,CE02,CE06)
- Determinar las transformaciones más adecuadas para el problema a resolver. (CB3,CT03,CT04,CE11)
- Saber caracterizar los datos atípicos o outliers. (CB3,CE13,CT03,CT04)
- Conocer qué problemas se tienen al tener conjuntos altamente desbalanceados. (CB3,CT03,CT04,CE13)

Como consecuencia de los resultados de aprendizaje adquiridos, el/la estudiante adquirirá las siguientes destrezas:

- Ser capaz de cargar cualquier fichero de datos de toda índole (texto, formatos tabulados, etc).
- Limpiar y completar el contenido del conjunto de datos una vez cargado (corrección de errores tipográficos, imputación de datos perdidos, etc..)
- Elegir el tipo de dato adecuado dependiendo de su naturaleza (entero, real, factor, texto, etc..)
- Filtrar muestras, seleccionar características, y crear nuevas a partir de formatos tabulados como el data frame.
- Realizar una caracterización de los datos dependiendo de su tipología.
- Crear visualizaciones de datos básicas con el fin de extraer conclusiones preliminares de los datos.

DESCRIPCIÓN DE CONTENIDOS

1. Introducción al tratamiento de datos

1.1. Por qué analizar datos.
1.2. Visión global de un problema de tratamiento de datos.

2. Obtención de datos

2.1. Introducción.
2.2 Introducción de datos.
2.3 Repositorios de datos.
2.4 Formatos de ficheros de datos.
2.5. Fusión de datos procedentes de diferentes fuentes.
2.6. Acceso a bases de datos

3. Visualización de datos

3.1. Gráficos explicativos y exploratorios.
3.2. Sistemas gráficos en R: base, grid, lattice, ggplot2.
3.3. Librería ggplot2: representaciones básicas.

4. Preparación de los datos

4.1. Estructura de un conjunto de datos para su análisis: operaciones básicas.
4.2. Manipulación de datos. Librería tidyr.
4.3. Manejo de datos. Librería dplyr.

5. Análisis exploratorio de datos I. Definiciones

5.1. Explorando un nuevo dataset.
5.2. Caracterización de variables.
5.3. Visualización de relaciones entre variables.

6. Análisis Exploratorio de datos II. Anomalías.

6.1.Anomalías en variables numéricas: Outliers. Caracterización de outliers. Métodos de detección.
6.2. Anomalías en variables numéricas: datos perdidos y ausentes

7. Trabajando con datos de tipo texto

7.1. Introducción
7.2. Bases del análisis de datos de texto.
7.3. Funciones básicas para el manejo de caracteres en R.
7.4. Expresiones regulares.
7.5 Codificación de caracteres: ascii vs Unicode.

8. Prácticas de Tratamientos de los datos

En este bloque se presentarán una serie de supuestos prácticos a modo de prácticas de laboratorio llevados a cabo en el aula de informática.
Práctica 1. Importación de datos.
Práctica 2. Visualización de datos.
Práctica 3. Preparación de datos con tidyr.
Práctica 4. Manejo de datos con dplyr.
Práctica 5. Datos anómalos.
Práctica 6. Análisis exploratorio de datos.
Práctica 7. Análisis completo de un conjunto de datos.

VOLUMEN DE TRABAJO

ACTIVIDAD	Horas	% Presencial
Clases de teoría	34,00	100
Prácticas en laboratorio	20,00	100
Prácticas en aula	6,00	100
Elaboración de trabajos en grupo	5,00	0
Elaboración de trabajos individuales	10,00	0
Estudio y trabajo autónomo	15,00	0
Lecturas de material complementario	5,00	0
Preparación de actividades de evaluación	15,00	0
Preparación de clases de teoría	15,00	0
Preparación de clases prácticas y de problemas	15,00	0
Resolución de cuestionarios on-line	10,00	0
TOTAL	**150,00**	

METODOLOGÍA DOCENTE

Las clases combinarán el contenido teórico y práctico.

MD1 - Actividades teóricas. Desarrollo expositivo de la materia con la participación del estudiante en la resolución de cuestiones puntuales. Realización de cuestionarios individuales de evaluación.

En las actividades teóricas de carácter presencial se desarrollarán los temas de la asignatura proporcionando una visión global e integradora, analizando con mayor detalle los aspectos clave y de mayor complejidad, fomentando, en todo momento, la participación del alumnado (CB03, CT03).

MD2 - Actividades prácticas. Aprendizaje mediante resolución de problemas, ejercicios y casos de estudio a través de los cuales se adquieren competencias sobre los diferentes aspectos de la materia. (CB03, CG06,CE02,CE06,CE11,CE13)

Las actividades teóricas se complementan con actividades prácticas con el objetivo de aplicar los conceptos básicos y ampliarlos con el conocimiento y la experiencia que se vayan adquiriendo durante la realización de los trabajos propuestos.

MD4 -Trabajos en laboratorio y/o aula ordenador. Aprendizaje mediante la realización de actividades desarrolladas de forma individual o en grupos reducidos y llevadas a cabo en laboratorios y/o aulas de ordenador. (CB03, CG06,CE02,CE06,CE11,CE13)

Además de las actividades presenciales, los estudiantes deberán realizar tareas personales (fuera del aula) sobre: cuestiones y problemas, así como la preparación de clases y exámenes (estudio). Estas tareas se realizarán principalmente de manera individual, con el fin de potenciar el trabajo autónomo, pero adicionalmente se incluirán trabajos, especialmente la preparación y resolución de prácticas laboratorio, que requieran la participación de pequeños grupos de estudiantes (2-3) para fomentar la capacidad de integración en grupos de trabajo.

Se utilizarán lan plataforman de e-learning (Aula Virtual) Universitat de València, Microsoft Teams y Blackboard Collaborate, como soporte de comunicación con el alumnado. A través de ellas se tendrá acceso al material didáctico utilizado en clase, así como los problemas y ejercicios a resolver.

EVALUACIÓN

La evaluación del aprendizaje de los conocimientos y competencias conseguidas por los estudiantes se hará de forma continuada a lo largo del curso, y constará de los siguientes bloques de evaluación:

Primera y segunda convocatorias

SE1 - Prueba objetiva, consistente en uno o varios exámenes que constan tanto de cuestiones teórico-prácticas como de problemas (evaluación de competencias CB03, CG06, CT03, CE02, CE06, CE11, CE13) (48%) (**Nota: Todos los porcentajes están referidos a la nota final**)

SE1-1 (30%) Examen de teoría

SE1-2 (18%) Examen de laboratorio)

SE2 - Evaluación de las actividades prácticas a partir de la elaboración de trabajos/memorias y/o exposiciones orales (evaluación de competencias CB03, CG06, CT03, CT04, CE02, CE06, CE11, CE13) (32%)

SE2-1 (20%) Realización de un miniproyecto consistente en el análisis completo de un conjunto de datos.

SE2-2 (12%) Asistencia y evaluación de las sesiones de laboratorio (Actividad NO RECUPERABLE)

SE3 - Evaluación continua de cada alumno, basada en la participación y grado de implicación del alumno en el proceso de enseñanza-aprendizaje, teniendo en cuenta la asistencia regular a las actividades presenciales previstas y la resolución de cuestiones y problemas propuestos periódicamente. (20%)

SE3-1 (5%) Asistencia regular a las actividades presenciales previstas (evaluación de competencias CB04, CG01). (Actividad NO RECUPERABLE)

SE3-2 (15%) Resolución de cuestiones y problemas propuestos (evaluación de competencias CB02, CB04, CG01, CT03). (Actividad NO RECUPERABLE)

La nota final de la asignatura se calculará como la media ponderada de cada uno de los apartados anteriores, de acuerdo al siguiente criterio: SE-1 (48%), SE-2 (32%), SE-3 (20%).

Consideraciones particulares sobre la evaluación:

- Es necesario obtener una calificación mínima de 4 (sobre 10) en los apartados de evaluación SE1-1 (examen Teoría), SE1-2 (examen de laboratorio) y SE2-1 (mini proyecto), para promediar.
- Las actividades SE2-2, SE3-1 y SE3-2 no son recuperables.
- La actividad SE1-2 se realizará al acabar el examen de teoría el día de la convocatoria oficial.

En cualquier caso, el sistema de evaluación se regirá por lo establecido en el Reglamento de Evaluación y Calificación de la Universidad de Valencia para Grados y Másteres:

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

REFERENCIAS

Básicas

- R.K.Pearson (2018) Exploratory Data Analysis Using R. CRC.
- H. Wickham, G. Grolemund. (2016) R for data Science. OReilly 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. Buttres 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)

Complementarias

- 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)

ADENDA COVID-19

Esta adenda solo se activará si la situación sanitaria lo requiere y previo acuerdo del Consejo de Gobierno

La metodología docente de la asignatura seguirá el Modelo Docente aprobado por la Comisión Académica de Título de Ciencia de Dades (https://go.uv.es/cienciadatos/ModelDocentGCD). En caso de que se produzca un cierre de las instalaciones por causas sanitarias que afecte total o parcialmente a las clases de la asignatura, estas serán sustituidas por sesiones no presenciales siguiendo los horarios establecidos. Si el cierre afectara a alguna prueba de evaluación presencial de la asignatura, esta será sustituida por una prueba de naturaleza similar que se realizará en modalidad virtual a través de las herramientas informáticas soportadas por la Universitat de València. Los porcentajes de cada prueba de evaluación se mantendrán, según lo establecido en esta guía