FICHA IDENTIFICATIVA

Datos de la Asignatura

Código	44653
Nombre	Análisis exploratorio de datos
Ciclo	Máster
Créditos ECTS	4.5
Curso académico	2021 - 2022

Titulación(es)

Titulación	Centro	Curso	Periodo
2221 - Máster Universitario en Ciencia de Datos	Escuela Técnica Superior de Ingeniería	1	Primer cuatrimestre

Materias

Titulación	Materia	Caracter
2221 - Máster Universitario en Ciencia de Datos	5 - Análisis exploratorio de datos	Obligatoria

Coordinación

Nombre	Departamento
GOMEZ SANCHIS, JUAN	242 - Ingeniería Electrónica
IÑIGUEZ HERNANDEZ, MARIA DEL CARMEN	130 - Estadística e Investigación Operativa
MARTINEZ SOBER, MARCELINO	242 - Ingeniería Electrónica

RESUMEN

En esta asignatura se describen las primeras etapas de un problema de análisis de datos así como los modelos estadísticos lineales básicos asociados a los métodos de regresión y clasificación.

El 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).

En este módulo nos centraremos en las etapas de preparación de los datos y en los modelos lineales de regresión y clasificación que nos permitirán aprender de los *outputs* de interés a partir de un conjunto dado de *inputs* conocidos y un modelo estadístico que relaciona *inputs* y *outputs* de forma probabilística

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

Introducción a la Ciencia de Datos

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

2221 - Máster Universitario en Ciencia de Datos

- Que los/las estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo
- Ser capaces de valorar la necesidad de completar su formación técnica, científica, en lenguas, en informática, en literatura, en ética, social y humana en general, y de organizar su propio autoaprendizaje con un alto grado de autonomía
- Capacidad de análisis y síntesis, en la elaboración de informes, en la exposición, comunicación y defensa de ideas.
- 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.
- Capacidad de organización y planificación de actividades de investigación, desarrollo y consultoría en el área de ciencia de datos.
- Ser capaces de acceder a herramientas de información (bibliográficas y de empleo) y utilizarlas apropiadamente.
- Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.
- Extraer conocimiento de conjuntos de datos en diferentes formatos.
- Entender la utilidad de la ciencia de datos y sus elementos asociados, así como su aplicación en la resolución de problemas, eligiendo las técnicas más adecuadas a cada problema, aplicando de forma correcta las técnicas de evaluación y, finalmente,
  ?interpretando los modelos y resultados.

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.
Determinar las transformaciones más adecuadas para el problema a resolver.
Conocer los principios básicos del aprendizaje estadístico.
Conocer la metodología estadística básica y los modelos lineales de regresión y clasificación. Aplicar estos modelos en estudios reales.
Saber implementar un modelo lineal con las diferentes herramientas informáticas consideradas a lo largo del curso.

DESCRIPCIÓN DE CONTENIDOS

1. Introducción al análisis exploratorio de datos

En este bloque de introducción se presentarán los principales aspectos a tener en cuenta para realizar una correcta preparación de los datos y sus implicaciones en las siguientes etapas.

2. Adquisición y limpieza de datos

En este bloque se presentarán los diversos tipos de datos, (continuos, discreto) , importación de los formatos más habituales, conversión de datos, detección de datos anómalos.

3. Análisis estadístico descriptivo

En este bloque se muestran como caracterizar diferentes tipos de datos, mediante sus estadísitcos básico y diversos tipos de representaciones visuales básicas, como parte fundamental en la comprensión de los datos disponibles y la detección de errores en importación o en los valores originales de los mismos (análisis univariante, bivariante, multivariante, correlación, covarianza, etc)

4. Transformaciones en los datos

En este bloque se presentan métodos de transformación de datos. En esta etapa los datos son transformados para que el proceso de análisis sea más eficiente y se facilite la comprensión de la información que contienen.

5. Modelos lineales: modelos de regresión y modelos de análisis de la varianza.

Variables input y output. Regresión simple y regresión múltiple. Distribución normal multivariante. Regresión, correlación y causalidad. Estimación y contraste de hipótesis. Tabla ANOVA. Predicción.

6. Modelos de regresión estructurados

Selección de variables. Métodos de encogimiento: regresión ridge, lasso y elastic net. Regularización

7. Modelos lineales generalizados

Familia exponencial de distribuciones. Regressión lineal amb una matriu indicadora. Regressió logística

VOLUMEN DE TRABAJO

ACTIVIDAD	Horas	% Presencial
Clases teórico-prácticas	45,00	100
Elaboración de trabajos individuales	10,00	0
Estudio y trabajo autónomo	6,00	0
Lecturas de material complementario	1,50	0
Preparación de actividades de evaluación	6,00	0
Preparación de clases de teoría	10,00	0
Preparación de clases prácticas y de problemas	6,50	0
Resolución de casos prácticos	5,00	0
TOTAL	**90,00**	

METODOLOGÍA DOCENTE

Las clases combinarán el contenido teórico y el práctico, sin distinción entre sesiones de teoría y de laboratorio. Todas las sesiones se impartirán en aulas de informática.

Actividades teóricas. Desarrollo expositivo de la materia con la participación del estudiante en la resolución de cuestiones puntuales. Posible realización de cuestionarios individuales de evaluación.

Actividades prácticas. Aprendizaje mediante resolución de problemas, ejercicios y casos de estudio que permiten adquirir competencias sobre los diferentes aspectos de la materia.

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 aulas de ordenador.

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:

1. Evaluación de las actividades prácticas a partir de la elaboración de trabajos/memorias y/o exposiciones orales: 60% de la nota final.
2. Prueba objetiva, consistente en uno o varios exámenes que constarán tanto de cuestiones teórico-prácticas como de problemas: 40 % de la nota final

Las calificaciones obtenidas en el apartado 1 sólo se conservarán en las dos convocatorias del curso académico en que hayan sido realizadas, dado que su evaluación sólo es posible en el periodo de docencia.

REFERENCIAS

Básicas

- L. Han, M. Kamber, and J. Pei. (2012) Data Mining Concepts and Techniques (third Edition). Morgan Kaufman, Elsevier
- N. Zumel and J. Mount (2014). Practical Data Science with R. Manning Publications Co
- D. Pyle (1999). Data preparation for data mining. Academic Press
- G. J. Myatt and W. P. Johnson. (2014). Making Sense of Data I. Wiley.
- Y. Zao and J. Cen (2013) Data mining Applications with R. Academic Press
- R. D. Peng (2016) Exploratory Data Analysis with R. Lean Publishing (https://leanpub.com/exdata)
- G. James, E. Witten, T. Hastie, and R. Tibshirani. (2015). An Introduction to Statistical Learning with applications in R. Corrected 6th printing. Springer http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf
- https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R
- https://en.wikibooks.org/wiki/R_Programming
- T. Hastie, R. Tibshirani, and J. Friedman (2008). The Elements of Statistical Learning. Second Edition. Springer

Complementarias

- Cirillo (2016) RStudio for R Statistical Computing. Cookbook Paperback
- J .Albert and M. Rizzo. (2012) R by example. Springer

ADENDA COVID-19

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

En caso de que se produzca un modo híbrido de docencia (que combine presencialidad con no presencialidad) o un cierre de las instalaciones por causas sanitarias que afecten total o parcialmente a las clases de la asignatura, estas serán sustituidas preferentemente por sesiones no presenciales síncronas 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 permanecerán invariables, según aquello establecido por esta guía.