FITXA IDENTIFICATIVA

Dades de l'Assignatura

Codi	44653
Nom	Anàlisi exploratòria de dades
Cicle	Màster
Crèdits ECTS	4.5
Curs acadèmic	2021 - 2022

Titulació/titulacions

Titulació	Centre	Curs	Període
2221 - M.U.Ciència de Dades	Escola Tècnica Superior d'Enginyeria	1	Primer quadrimestre

Matèries

Titulació	Matèria	Caràcter
2221 - M.U.Ciència de Dades	5 - Anàlisi exploratòria de dades	Obligatòria

Coordinació

Nom	Departament
GOMEZ SANCHIS, JUAN	242 - Enginyeria Electrònica
IÑIGUEZ HERNANDEZ, MARIA DEL CARMEN	130 - Estadística i Investigació Operativa
MARTINEZ SOBER, MARCELINO	242 - Enginyeria Electrònica

RESUM

En aquesta assignatura es descriuen les primeres etapes d'un problema d'anàlisi de dades així com els models estadístics lineals bàsics associats als mètodes de regressió i classificació.

El 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 faltantes (*missing data imputation*), transformació de les dades, selecció de les característiques més rellevant d'un conjunt d'alta dimensionalidad (*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 per a la fiabilitat i solidesa dels resultats obtinguts en etapes posteriors de l'anàlisi (selecció de models, classificadors, agrupament, estimació, contrastos d'hipòtesis, etc).

En aquest mòdul ens centrarem en les etapes de preparació de les dades i en els models lineals de regressió i classificació que ens permetran aprendre dels *outputs* d'interès a partir d'un conjunt donat d' *inputs* coneguts i un model estadístic que relaciona *inputs* i *outputs* de forma probabilística.

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

Introducció a la Ciència de Dades

2221 - M.U.Ciència de Dades

- Que els estudiants posseïsquen les habilitats d?aprenentatge que els permeten continuar estudiant d?una forma que haurà de ser en gran manera autodirigida o autònoma.
- Ser capaços de valorar la necessitat de completar la seua formació tècnica, científica, en llengües, en informàtica, en literatura, en ètica, social i humana en general, i d'organitzar el seu propi autoaprenentatge amb un alt grau d'autonomia
- Capacitat d'anàlisi i síntesi, en l'elaboració d'informes, en l'exposició, comunicació i defensa d'idees.
- Capacitat d'accés i gestió de la informació en diferents formats per al seu posterior anàlisi a fi d'obtindre coneixement a partir de dades.
- Capacitat d'organització i planificació d'activitats d'investigació, desenrotllament i consultoria en l'àrea de ciència de dades.
- Ser capaços d'accedir a ferramentes d'informació (bibliogràfiques i d'ocupació) i utilitzar-les apropiadament.
- Ser capaços d'assumir la responsabilitat del seu propi desenvolupament professional i de la seua especialització en un o més camps d'estudi, aplicant els coneixements adquirits en la identificació d'eixides professionals i jaciments d'ocupació.
- Extraure coneixement de conjunts de dades en diferents formats.
- 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.

Conèixer les tècniques i algorismes per preprocesar i extreure les característiques més importants d'un conjunt de dades.
Determinar les transformacions més adequades per al problema a resoldre.
Conèixer els principis bàsics de l'aprenentatge estadístic.
Conèixer la metodologia estadística bàsica i els models lineals de regressió i classificació. Aplicar aquests models en estudis reals.
Saber implementar un model lineal amb les diferents eines informàtiques considerades al llarg del curs.

DESCRIPCIÓ DE CONTINGUTS

1. Introducció a lanalisi exploratori dades

En aquest bloc d'introducció es presentaran els principals aspectes a tenir en compte per a realitzar una correcta visualització de les dades

2. Adquisició i neteja de dades

En aquest bloc es presentaran els diversos tipus de dades, (continus, discrets) , importació dels formats més habituals, conversió de dades, detecció de dades anòmales

3. Anàlisi estadístic descriptiu

En aquest bloc es mostren com caracteritzar diferents tipus de dades, mitjançant els seus estadísitcs bàsics i diversos tipus de representacions gràfiques com a part fonamental en la comprensió de les dades disponibles i la detecció d'errors en importació o en els valors originals dels mateixos (anàlisi univariant , bivariant , multivariant , correlació , covariància , etc.)

4. Transformacions en les dades

En aquest bloc es presenten mètodes de transformació de dades. En aquesta etapa les dades són transformats perquè el procés d'anàlisi sigui més eficient i es faciliti la comprensió de la informació que contenen

5. Models lineals: models de regressió i models danàlisi de la variança

Variables input i output. Regressió simple i regressió múltiple. Distribució normal multivariante. Regressió, correlació i causalitat. Estimació i contrast dhipótesis. Taula ANOVA. Predicció.

6. Models de regressió estructurats

Selecció de variables. Mètodes d'encongiment: regressió ridge, lasso i elastic net. Regularització

7. Models lineals generalitzats

Familia exponencial de distribuciones. Regresión lineal con una matriz indicadora. Regresión logística

VOLUM DE TREBALL

ACTIVITAT	Hores	% Presencial
Classes teoricopràctiques	45,00	100
Elaboració de treballs individuals	10,00	0
Estudi i treball autònom	6,00	0
Lectures de material complementari	1,50	0
Preparació d'activitats d'avaluació	6,00	0
Preparació de classes de teoria	10,00	0
Preparació de classes pràctiques i de problemes	6,50	0
Resolució de casos pràctics	5,00	0
TOTAL	**90,00**	

METODOLOGIA DOCENT

Les classes combinaran el contingut teòric i el pràctic, sense distinció entre sessions de teoria i de laboratori. Totes les sessions s'impartiran en aules d'informàtica.

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

Activitats pràctiques. Aprenentatge mitjançant resolució de problemes, exercicis i casos d'estudi que permeten adquirir competències sobre els diferents aspectes de la matèria.

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 aules d'ordinador.

AVALUACIÓ

L'avaluació de l'aprenentatge dels coneixements i competències obtinguts pels estudiants es realitzarà de forma continuada al llarg del curs, i constarà dels següents blocs d'avaluació:

1. Exercicis i treballs entregats durant el curs i/o exàmens parcials: 60% de la nota final.
2. Examen final: 40% de la nota final.

Les qualificacions obtingudes en l'apartat 1 sols es conservaran en les dues convocatòries del curs acadèmic en què s`hagen realitzat, donat que la seua avaluació sols és possible en el període de docència.

REFERÈNCIES

Bàsiques

- 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

Complementàries

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

ADDENDA COVID-19

Aquesta addenda només s'activarà si la situació sanitària ho requereix i previ acord del Consell de Govern

En cas que es produeixi una manera híbrid de docència (que combini presencialitat amb no presencialitat) o un tancament de les instal·lacions per causes sanitàries que afecten totalment o parcialment a les classes de l'assignatura, aquestes seran substituïdes preferentment per sessions no presencials síncrones seguint els horaris establerts.

Si el tancament afectés alguna prova d'avaluació presencial de l'assignatura, aquesta serà substituïda per una prova de naturalesa similar que es realitzarà en modalitat virtual a través de les eines informàtiques suportades per la Universitat de València.

Els percentatges de cada prova d'avaluació romandran invariables, segons allò establert per aquesta guia.