



Guía Docente				
Datos Identificativos			2021/22	
Asignatura (*)	Modelización Estadística de Datos de Alta Dimensión	Código	614G02013	
Titulación				
Descriptorios				
Ciclo	Período	Curso	Tipo	Créditos
Grao	1º cuatrimestre	Segundo	Obrigatoria	6
Idioma	CastelánGalegoInglés			
Modalidade docente	Presencial			
Prerrequisitos				
Departamento	Matemáticas			
Coordinación	Cao Abad, Ricardo	Correo electrónico	ricardo.cao@udc.es	
Profesorado	Cao Abad, Ricardo López Cheda, Ana	Correo electrónico	ricardo.cao@udc.es ana.lopez.cheda@udc.es	
Web	<a href="http://dm.udc.es/staff/ricardo_cao/">http://dm.udc.es/staff/ricardo_cao/</a>			
Descrición xeral	Esta materia proporciona un primeiro contacto do alumnado coa modelización estatística de grandes conxuntos de datos: técnicas de análise multivariante, ferramentas estatísticas e programas informáticos avanzados para a análise de datos de alta dimensión, identificación das vantaxes e limitacións dos diferentes métodos, e procedementos de crítica, diagnose e interpretación dos resultados en termos do problema proposto.			



<b>Plan de continxencia</b>	<p>1. Modificacións nos contidos: Non se realizarán cambios.</p> <p>2. Metodoloxías: ? Metodoloxías docentes que se manteñen: Probas resposta breve: realizarase unha proba de resposta breve, aproximadamente na metade do cuadrimestre, mediante a plataforma Moodle.udc.es. No día do exame oficial, en xaneiro de 2022, realizarase a segunda proba de resposta breve, tamén a través da plataforma Moodle.udc.es. Cada unha destas probas computa un 20% na avaliación. Traballo tutelado: traballo práctico, en parellas, que computa na avaliación (20%). A presentación oral (que computa o 10%) realizarase a través de TEAMS. Probas sobre prácticas con R: realizarase unha proba, aproximadamente na metade do cuadrimestre, e entregarase mediante a plataforma Moodle.udc.es. No día do exame oficial, en xaneiro de 2022, realizarase o segundo exercicio avaliable en R, e tamén se fará a entrega a través da plataforma Moodle.udc.es. Cada unha destas probas computa un 20% na avaliación.</p> <p>? Metodoloxías docentes que se modifican: As sesións maxistrals: non computan na avaliación. Impartiranse usando TEAMS na franxa horaria que ten asignada a materia no calendario de aulas da facultade. Ademais, estas sesións por TEAMS poderanse complementar con vídeos explicativos. As prácticas TIC: non computan na avaliación. Na modalidade presencial consistían en análises de datos usando software estatístico (R). Substitúense por vídeos onde se explica con detalle o desenvolvemento da práctica. Eses vídeos poderían ser realizados e gravados en TEAMS na mesma hora da clase, ou subidos á plataforma con anterioridade. Ademais, realizaranse titorías grupais semanais por TEAMS (ou mais segundo o demande o alumnado) para seguimento e apoio dos alumnos.</p> <p>3. Mecanismos de atención personalizada ao alumnado: Ferramenta: Correo Electrónico, Vídeo conferencia e Moodle.</p> <p>Temporalización: Correo Electrónico: Diariamente. De uso para facer consultas, solicitar encontros virtuais para resolver dúbidas e facer o seguimento dos traballos tutelados. Vídeo conferencia (Teams): Dúas sesións semanais, para o avance dos contidos, na franxa horaria que ten asignada a materia no calendario de aulas da facultade. Tamén se realizarán titorías individuais e grupais fixadas previamente mediante correo electrónico. Moodle: Diariamente, segundo a necesidade do alumnado. Presentaranse ?foros temáticos? asociados aos módulos da materia, para formular as consultas necesarias.</p> <p>4. Modificacións na avaliación: Manterase o peso da cualificación en cada unha das probas. A diferenza está en que a presentación do traballo tutelado se realizará a través de TEAMS, e as probas (parciais e/ou finais) de conceptos realizaranse por Moodle e as de prácticas en R entregaranse por esa mesma plataforma.</p> <p>Observacións de avaliación: Ao longo do curso realízanse dous parciais, un para a parte dos Bloques 0-2 e outro da parte dos Bloques 3-4, que permiten liberar a parte correspondente da materia.</p> <p>5. Modificacións da bibliografía ou webgrafía: Non se realizarán cambios. Xa dispoñen de todos os materiais de traballo da maneira dixitalizada en Moodle.</p>
-----------------------------	---

## Competencias / Resultados do título

Código	Competencias / Resultados do título
--------	-------------------------------------



Resultados da aprendizaxe			
Resultados de aprendizaxe	Competencias / Resultados do título		
Coñecer as principais técnicas da análise estatística multivariante.	A17	B2 B8 B9 B10	C1
Coñecer os principais problemas que poden xurdir ao traballar con datos de alta dimensión.	A17 A20	B2 B3 B9 B10	C1
Saber seleccionar as principais variables e modelos en problemas reais.	A17 A20	B2 B3 B8 B9	C1
Ser quen de aplicar as principais técnicas de análise multivariante a conxuntos de datos reais ou simulados.	A17 A20	B2 B3 B7 B8 B9 B10	C1
Ser quen de interpretar os resultados e coñecer as limitacións dos métodos de análise estatístico multivariante.	A17 A20	B2 B3 B7 B8 B9 B10	C1
Saber manexar con soltura programas informáticos avanzados de análise estatística.	A20	B2 B10	C1

Contidos	
Temas	Subtemas
0. Distribucións multidimensionais	0.1 Concepto de distribución multidimensional 0.2. Matriz de varianzas-covarianzas. Transformacións lineais 0.3. Normal multidimensional: definición e propiedades
1. Métodos de redución da dimensión	1.1 Obxectivos da Análise de Compoñentes Principais (ACP) 1.2 Transformacións para conseguir incorrelación 1.3 Obtención das compoñentes principais 1.4 Compoñentes principais e cambios de escala 1.5 Interpretación das compoñentes principais 1.6 Análise factorial 1.7 Escalamiento multidimensional



2. Clasificación non supervisada	<p>2.1 Obxectivos da clasificación non supervisada: métodos xerárquicos e non xerárquicos</p> <p>2.2 Análise clúster: deseño e obxectivos</p> <p>2.3 Árbore xerárquica ou dendograma</p> <p>2.4 Similitudes e discrepancias entre observacións</p> <p>2.5 Criterios para a formación de grupos: encadeamento simple, completo, promedio do grupo, método do centroide, método de Ward</p> <p>2.6 Métodos non xerárquicos baseados en distancias: veciños máis próximos, k medias, métodos baseados na estimación da densidade</p>
3. Clasificación supervisada	<p>3.1 Obxectivos da clasificación supervisada: regras de clasificación e criterios de erro</p> <p>3.2 Análise factorial discriminante: deseño, obxectivos e cálculo dos factores discriminantes</p> <p>3.3 Análise discriminante lineal de Fisher e análise discriminante cadrático</p> <p>3.4 Regra discriminante de máxima verosimilitude, regra Bayes, regras discriminantes non paramétricas</p> <p>3.5 Relación cos modelos de regresión con resposta binaria</p> <p>3.6 Estimación da probabilidade de clasificación incorrecta: validación cruzada e bootstrap</p>
4. Modelos para datos de alta dimensión	<p>4.1 Selección de variables en regresión: contrastes de significación</p> <p>4.2 O problema dos contrastes múltiples: false discovery rate (FDR) e familywise error rate (FWER)</p> <p>4.3 Modelos de regresión de coeficientes dispersos: regresión riscal (ridge regression), lasso e as súas variantes</p> <p>4.4 Selección de variables e modelos con coeficientes dispersos no caso de clasificación</p>

Planificación				
Metodoloxías / probas	Competencias / Resultados	Horas lectivas (presenciais e virtuais)	Horas traballo autónomo	Horas totais
Presentación oral	A1 B2 B3 B4 C4	30	36	66
Prácticas a través de TIC	A9 A12 A17 A18 A19 A20 A26 A33 A3 A4 A5 A6 A8 B7 B9 B10 C1 C2 C3	14	21	35
Proba de resposta múltiple	A19 A24 A25 A1 B3 B8	2	6	8
Solución de problemas	A17 A33 A2 B2 B5 B6 B7 B8 B10	14	21	35
Atención personalizada		6	0	6

\*Os datos que aparecen na táboa de planificación son de carácter orientativo, considerando a heteroxeneidade do alumnado

Metodoloxías	
Metodoloxías	Descrición
Presentación oral	Presentación con ordenador
Prácticas a través de TIC	Análise estatística de conxuntos de datos usando R.
Proba de resposta múltiple	Proba de resposta múltiple sobre conceptos.



Solución de problemas	Elección das ferramentas estatísticas e estratexias para resolver problemas. Formulación de modelos para datos multivariantes. Formulación de algoritmos para a análise de datos de alta dimensión.
-----------------------	---

### Atención personalizada

Metodoloxías	Descrición
Prácticas a través de TIC	Asistencia e participación nas clases teóricas. Exame escrito de múltiple opción.
Solución de problemas	Traballo de análise de datos multivariantes. Suposto práctico a realizar polo alumno.

### Avaliación

Metodoloxías	Competencias / Resultados	Descrición	Cualificación
Presentación oral	A1 B2 B3 B4 C4	Presentación oral do traballo por parellas.	10
Prácticas a través de TIC	A9 A12 A17 A18 A19 A20 A26 A33 A3 A4 A5 A6 A8 B7 B9 B10 C1 C2 C3	Práctica(s) de ordenador usando o software estatístico libre R.	40
Solución de problemas	A17 A33 A2 B2 B5 B6 B7 B8 B10	Contido do traballo en parella relacionado cos temas 0-3.	10
Proba de resposta múltiple	A19 A24 A25 A1 B3 B8	Proba(s) de comprensión dos conceptos impartidos.	40

### Observacións avaliación



A avaliación realizarase por medio de dúas probas sobre prácticas con R, un traballo por parellas, así como dúas probas escritas de conceptos. A primeira das probas prácticas e a primeira de conceptos realizaranse aproximadamente na metade do cuadrimestre, e corresponderán aos temas 0-2. As segundas de cada unha desas probas realizaranse o día fixado para o exame final, no mes de xaneiro de 2022. Estas segundas probas corresponderán a toda a materia do curso, mais os/as alumnos/as que teñan superado cada unha das probas de metade do cuadrimestre, poderán liberarse da materia dos temas 0-2, tratando só as súas probas sobre os temas 3-4. A cualificación tanto da(s) proba(s) de conceptos, como da(s) proba(s) sobre prácticas con R representarán o 40% da cualificación global, cada unha. O 20% restante corresponderá ao traballo por parellas, que ten que ser presentado en público polos alumnos, durante a segunda metade do cuadrimestre. A metade da puntuación deste traballo (10% da cualificación global) corresponde á presentación oral do mesmo.

En resumo, as ponderacións da avaliación quedarán da seguinte forma:

**Traballo práctico en parellas:** 20% do total (10% resolución do exercicio práctico en R e 10% presentación oral). **Exames de conceptos:** realizaranse dous exames de conceptos (cada un con ponderación do 20% sobre o total). O primeiro exame, relacionado cos Bloques 0-2, terá lugar a metade do cuadrimestre. O segundo exame, relacionado cos Bloques 3-4, realizarase o día do exame oficial. Permítese liberar materia, de forma que os estudantes que aproben o primeiro exame parcial, xa non se examinarán dos Bloques 0-2 no exame oficial, agás que queiran subir nota. Sen embargo, os estudantes que suspendan ou non se presenten ao parcial, irán ao exame oficial con esas dúas partes, e a suma de ambas valerá un 40% sobre o total. **Exames prácticos:** seguen a mesma idea que os exames de conceptos. Realizaranse dous exames prácticos (cada un con ponderación do 20% sobre o total), utilizando o software estatístico R. O primeiro exame, relacionado coas prácticas en R dos Bloques 0-2, terá lugar na metade do cuadrimestre. O segundo exame, relacionado coas prácticas en R dos Bloques 3-4, realizarase o día do exame oficial. Permítese liberar materia, de forma que os estudantes que aproben o primeiro exame parcial, xa non se examinarán das prácticas dos Bloques 0-2 no exame oficial, agás que queiran subir nota. Sen embargo, os estudantes que suspendan ou non se presenten ao parcial, irán ao exame oficial con esas dúas partes, e a suma de ambas valerá un 40% sobre o total. Para superar a materia será necesario obter unha calificación de alomenos 5 sobre 10 no conxunto da materia.

Na

oportunidade de xullo os alumnos poderán liberarse de facer as probas correspondentes nas que a súa calificación na oportunidade de xaneiro fora de alomenos 4 sobre 10.

Na primeira oportunidade (xaneiro), só os alumnos que non se teñan presentado a ningunha das probas avaliáveis que figuran

arriba obterán a calificación de NON PRESENTADO. En xullo obterán a calificación de NON PRESENTADO os alumnos que non se tiveran presentado ao exame final desa data.

Se algún estudante quere facer algunha das probas nun idioma oficial específico (galego ou español), debe avisar ó profesorado alomenos 1 semana antes da correspondente proba.

## Fontes de información

Fontes de información	
<b>Bibliografía básica</b>	<ul style="list-style-type: none"> <li>- Anderson, T.W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley</li> <li>- Chatfield, C., Collins, A. J. (1980). Introduction to multivariate analysis. Chapman &amp; Hall</li> <li>- Giraud, C. (2014). Introduction to High-Dimensional Statistics. Chapman &amp; Hall/CRC</li> <li>- Goldstein, M., Dillon, W. R. (1984). Multivariate Analysis: Methods and Applications. Wiley</li> <li>- Jambu, M. (1991). Exploratory and Multivariate Data Analysis. Boston, Academic Press</li> <li>- Jobson, J.D. (1994). Applied Multivariate Data Analysis. Springer-Verlag</li> <li>- Johnson, R. A., Wichern, D. W. (2007). Applied multivariate statistical analysis. Prentice Hall</li> <li>- Koch, I. (2014). Analysis of Multivariate and High-Dimensional Data. Cambridge University Press</li> <li>- Mardia, K.V., Kent, J.T., Bibby, J.M. (1994). Multivariate Analysis. Academic Press. Academic Press</li> <li>- Muirhead, R.J. (1982). Aspects of multivariate statistical theory. John Wiley &amp; Sons</li> <li>- Rencher, A.C. (1998). Multivariate Statistical Inference and Applications. Wiley</li> <li>- Wainwright, M.J. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press</li> </ul>
<b>Bibliografía complementaria</b>	



## Recomendacións

### Materias que se recomenda ter cursado previamente

Introdución ás Bases de Datos/614G02008  
Álxebra Lineal/614G02001  
Cálculo Multivariable/614G02006  
Matemática Discreta/614G02002  
Fundamentos de Programación II/614G02009  
Fundamentos de Programación I/614G02004  
Inferencia Estatística/614G02007  
Probabilidade e Estatística Básica/614G02003

### Materias que se recomenda cursar simultaneamente

Modelos de Regresión/614G02012

### Materias que continúan o temario

Técnicas de Simulación e Remostraxe/614G02036  
Análise Estatística de Datos Complexos/614G02031  
Aprendizaxe Automática III/614G02026  
Recuperación de Información/614G02027  
Aprendizaxe Automática I/614G02019  
Aprendizaxe Automática II/614G02021  
Análise Estatística de Datos con Dependencia/614G02022

### Observacións

(\*A Guía docente é o documento onde se visualiza a proposta académica da UDC. Este documento é público e non se pode modificar, salvo casos excepcionais baixo a revisión do órgano competente dacordo coa normativa vixente que establece o proceso de elaboración de guías