Guia docente				
Datos Identificativos				2021/22
Asignatura (*)	Modelización Estadística de Datos de Alta Dimensión		Código	614G02013
Titulación	Grao en Ciencia e Enxeñaría de Datos			
Descriptores				
Ciclo	Periodo	Curso	Tipo	Créditos
Grado	1º cuatrimestre	Segundo	Obligatoria	6
Idioma	CastellanoGallegoInglés			
Modalidad docente	Presencial			
Prerrequisitos				
Departamento	Matemáticas			
Coordinador/a	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	http://dm.udc.es/staff/ricardo_cao/			
Descripción general	Esta asignatura proporciona un primer contacto del alumnado con la modelización estadística de grandes conjuntos de datos: técnicas de análisis multivariante, herramientas estadísticas y programas informáticos avanzados para el análisis de datos de alta dimensión, identificación de las ventajas y limitaciones de los diferentes métodos, y procedementos de crítica, diagnosis e interpretación de los resultados en relación con el problema propuesto.			

Plan de contingencia	1. Modificaciones en los contenidos: No se realizarán cambios. 2. Metodologías: ? Metodologías docentes que se mantienen: Pruebas respuesta breve: se realizará una prueba de respuesta breve, aproximadamente en la mitad del cuatrimestre, mediante la plataforma Moodle.udc.es. El día del examen oficial, en enero de 2022, se realizará la segunda prueba de respuesta breve, también a través de la plataforma Moodle.udc.es. Cada una de estas pruebas computa un 20% en la evaluación. Trabajo tutelado: trabajo práctico, en parejas, que computa en la evaluación (20%). La presentación oral (que computa el 10%) se realizará a través de TEAMS. Pruebas sobre prácticas con R: se realizará una prueba, aproximadamente en la mitad del cuatrimestre, y se entregará mediante la plataforma Moodle.udc.es. El día del examen oficial, en enero de 2022, se realizará el segundo ejercicio evaluable en R, y también se hará la entrega a través de la plataforma Moodle.udc.es. Cada una de estas pruebas computa un 20% en la evaluación. ? Metodologías docentes que se modifican: Las sesiones magistrales: no computan en la evaluación. Se impartirán usando TEAMS en la franja horaria que tiene asignada la materia en el calendario de aulas de la facultad. Además, estas sesiones por TEAMS se podrán complementar con vídeos explicativos. Las prácticas TIC: no computan en la evaluación. En la modalidad presencial consistían en análisis de datos usando software estadístico (R). Se sustituyen por vídeos donde se explica con detalle el desarrollo de la práctica. Esos vídeos podrían ser realizados y grabados en TEAMS en la misma hora de la clase, o subidos a la plataforma con anterioridad. Además, se realizarán tutorías grupales semanales por TEAMS (o más, según demande el alumnado) para seguimiento y apoyo de los alumnos. 3. Mecanismos de atención personalizada al alumnado: Herramienta: Correo Electrónico, Vídeo conferencia (Teams), Moodle. Temporalización: Correo Electrónico: Diariamente. De uso para hacer consultas, solicitar encuentros virtuales para resolver dudas y hacer el seguimiento de los trabajos tutelados. Vídeo conferencia (TEAMS): Dos sesiones semanales, para el avance de los contenidos, en la franja horaria que tiene asignada la materia en el calendario de aulas de la facultad. También se realizarán tutorías individuales y grupales fijadas previamente mediante correo electrónico. Moodle: Diariamente, según la necesidad del alumnado. Se presentarán ?foros temáticos? asociados a los módulos de la materia, para formular las consultas necesarias. 4. Modificaciones en la evaluación: Se mantendrá el peso de la cualificación en cada una de las pruebas. La diferencia está en que la presentación del trabajo tutelado se realizará a través de TEAMS, y las pruebas (parciales y/o finales) de conceptos se realizarán por Moodle y las de prácticas en R se entregarán por esa misma plataforma. Observaciones de evaluación: A lo largo del curso se realizan dos parciales, uno para la parte de los Bloques 0-2 y otro de la parte de los Bloques 3-4, que permiten liberar la parte correspondiente de la materia. 5. Modificaciones de la bibliografía o webgrafía: No se realizarán cambios. Ya disponen de todos los materiales de trabajo de manera digitalizada en Moodle.

Competencias / Resultados del título	
Código	Competencias / Resultados del título

A17	CE17 - Capacidad para la construcción, validación y aplicación de un modelo estocástico de un sistema real a partir de los datos observados y el análisis crítico de los resultados obtenidos.
A20	CE20 - Conocimiento de las herramientas informáticas en el campo del análisis de los datos y modelización estadística, y capacidad para seleccionar las más adecuadas para la resolución de problemas.
B2	CB2 - Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio
B3	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
B7	CG2 - Elaborar adecuadamente y con cierta originalidad composiciones escritas o argumentos motivados, redactar planes, proyectos de trabajo, artículos científicos y formular hipótesis razonables.
B8	CG3 - Ser capaz de mantener y extender planteamientos teóricos fundados para permitir la introducción y explotación de tecnologías nuevas y avanzadas en el campo.
B9	CG4 - Capacidad para abordar con éxito todas las etapas de un proyecto de análisis de datos: exploración previa de los datos, preprocesado, análisis, visualización y comunicación de resultados.
B10	CG5 - Ser capaz de trabajar en equipo, especialmente de carácter multidisciplinar, y ser hábiles en la gestión del tiempo, personas y toma de decisiones.
C1	CT1 - Utilizar las herramientas básicas de las tecnologías de la información y las comunicaciones (TIC) necesarias para el ejercicio de su profesión y para el aprendizaje a lo largo de su vida.

Resultados de aprendizaje			
Resultados de aprendizaje	Competencias / Resultados del título		
Conocer las principales técnicas del análisis estadístico multivariante.	A17	B2 B8 B9 B10	C1
Conocer los principales problemas que pueden surgir al trabajar con datos de alta dimensión.	A17 A20	B2 B3 B9 B10	C1
Saber seleccionar las principales variables y modelos en problemas reales.	A17 A20	B2 B3 B8 B9	C1
Ser capaz de aplicar las principales técnicas de análisis multivariante a conjuntos de datos reales o simulados.	A17 A20	B2 B3 B7 B8 B9 B10	C1
Ser capaz de interpretar los resultados y conocer las limitaciones de los métodos de análisis estadístico multivariante.	A17 A20	B2 B3 B7 B8 B9 B10	C1
Saber manejar con soltura programas informáticos avanzados de análisis estadístico.	A20	B2 B10	C1

Contenidos

Tema	Subtema
0. Distribuciones multidimensionales	0.1 Concepto de distribución multidimensional 0.2 Matriz de varianzas-covarianzas. Transformaciones lineales 0.3 Normal multidimensional: definición y propiedades
1. Métodos de reducción de la dimensión	1.1 Objetivos del Análisis de Componentes Principales (ACP) 1.2 Transformaciones para conseguir incorrelación 1.3 Obtención de las componentes principales 1.4 Componentes principales y cambios de escala 1.5 Interpretación de las componentes principales 1.6 Análisis factorial 1.7 Escalamiento multidimensional
2. Clasificación no supervisada	2.1 Objetivos de la clasificación no supervisada: métodos jerárquicos y no jerárquicos 2.2 Análisis clúster: planteamiento y objetivos 2.3 Árbol jerárquico o dendograma 2.4 Similitudes y discrepancias entre observaciones 2.5 Criterios para la formación de grupos: encadenamiento simple, completo, promedio del grupo, método del centroide, método de Ward 2.6 Métodos no jerárquicos basados en distancias: vecinos más cercanos, k medias, métodos basados en estimación de la densidad
3. Clasificación supervisada	3.1 Objetivos de la clasificación supervisada: reglas de clasificación y criterios de error 3.2 Análisis factorial discriminante: planteamiento, objetivos y cálculo de los factores discriminantes 3.3 Análisis discriminante lineal de Fisher y análisis discriminante cuadrático 3.4 Regla discriminante de máxima verosimilitud, regla Bayes, reglas discriminantes no paramétricas 3.5 Relación con los modelos de regresión con respuesta binaria 3.6 Estimación de la probabilidad de clasificación incorrecta: validación cruzada y bootstrap
4. Modelos para datos de alta dimensión	4.1 Selección de variables en regresión: contrastes de significación. 4.2 El problema de los contrastes múltiples: false discovery rate (FDR) y familywise error rate (FWER) 4.3 Modelos de regresión de coeficientes dispersos: regresión riscal (ridge regression), lasso y sus variantes 4.4 Selección de variables y modelos con coeficientes dispersos en el caso de clasificación

Planificación				
Metodologías / pruebas	Competencias / Resultados	Horas lectivas (presenciales y virtuales)	Horas trabajo autónomo	Horas totales
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
Prueba de respuesta 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
(*)Los datos que aparecen en la tabla de planificación són de carácter orientativo, considerando la heterogeneidad de los alumnos				

Metodologías	
Metodologías	**Descripción**
Presentación oral	Presentación con ordenador.
Prácticas a través de TIC	Análisis estadístico de conjuntos de datos usando R.
Prueba de respuesta múltiple	Prueba de repuesta múltiple sobre conceptos.
Solución de problemas	Elección de las herramientas estadísticas y estrategias para resolver problemas. Formulación de modelos para datos multivariantes. Formulación de algoritmos para el análisis de datos de alta dimensión.

Atención personalizada	
Metodologías	**Descripción**
Prácticas a través de TIC Solución de problemas	Asistencia y participación en las clases teóricas. Examen escrito de múltiple opción. Trabajo de análisis de datos multivariantes. Supuesto práctico a realizar por el alumno.

Evaluación			
Metodologías	**Competencias / Resultados**	**Descripción**	**Calificación**
Presentación oral	A1 B2 B3 B4 C4	Presentación oral del trabajo en parejas.	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 el software estatístico libre R.	40
Solución de problemas	A17 A33 A2 B2 B5 B6 B7 B8 B10	Contenido del trabajo en parejas relacionado con los temas 0-3.	10
Prueba de respuesta múltiple	A19 A24 A25 A1 B3 B8	Prueba(s) de comprensión de los conceptos impartidos.	40

Observaciones evaluación

La evaluación se realizará por medio de dos pruebas sobre prácticas con R, un trabajo en parejas, así como dos pruebas escritas de conceptos. La primera de las pruebas prácticas y la primera de conceptos se realizarán aproximadamente en la mitad del cuatrimestre, y corresponderán a los temas 0-2. Las segundas de cada una de esas pruebas se realizarán el día fijado para el examen final, en el mes de enero de 2022. Estas segundas pruebas corresponderán a toda la materia del curso, pero los/las alumnos/as que hayan superado cada una de las pruebas de mitad del cuatrimestre, podrán liberarse de la materia de los temas 0-2, tratando solamente sus pruebas sobre los temas 3-4. La calificación tanto de la(s) prueba(s) de conceptos, como de la(s) prueba(s) sobre prácticas con R representarán el 40% de la calificación global, cada una. El 20% restante corresponderá al trabajo por parejas, que tiene que ser presentado en público por los alumnos, durante la segunda mitad del cuatrimestre. La mitad de la puntuación de este trabajo (10% de la calificación global) corresponde a la presentación oral del mismo.
En resumen, las ponderaciones de la evaluación quedarán de la siguiente forma:
Trabajo práctico en parejas: 20% del total (10% resolución del ejercicio práctico en R y 10% presentación oral).Exámenes de conceptos: se realizarán dos exámenes de conceptos (cada uno con ponderación del 20% sobre el total). El primer examen, relacionado con los Bloques 0-2, tendrá lugar a mitad del cuatrimestre. El segundo examen, relacionado con los Bloques 3-4, se realizará el día del examen oficial. Se permite liberar materia, de forma que los estudiantes que aprueben el primer examen parcial, ya no se examinarán de los Bloques 0-2 en el examen oficial, al menos que quieran subir nota. Sin embargo, los estudiantes que suspendan o no se presenten al parcial, irán al examen oficial con esas dos partes, y la suma de ambas valdría un 40% sobre el total.Exámenes prácticos: siguen la misma idea que los exámenes de conceptos. Se realizarán dos exámenes prácticos (cada uno con ponderación del 20% sobre el total), utilizando el software estadístico R. El primer examen, relacionado con las prácticas en R de los Bloques 0-2, tendrá lugar a mitad del cuatrimestre. El segundo examen, relacionado con las prácticas en R de los Bloques 3-4, se realizará el día del examen oficial. Se permite liberar materia, de forma que los estudiantes que aprueben el primer examen parcial, ya no se examinarán de las prácticas de los Bloques 0-2 en el examen oficial, al menos que quieran subir nota. Sin embargo, los estudiantes que suspendan o no se presenten al parcial, irán al examen oficial con esas dos partes, y la suma de ambas valdría un 40% sobre el total.Para superar la materia será necesario obtener una calificación de por lo menos 5 sobre 10 en el conjunto de la materia.
En la oportunidad de julio, los alumnos podrán liberarse de hacer las pruebas correspondientes en las que su calificación en la oportunidad de enero fuera de por lo menos 4 sobre 10.
En la primera oportunidad (enero-febrero), solo los alumnos que no se hayan presentado a ninguna de las pruebas evaluables que figuran arriba obtendrán la calificación de NO PRESENTADO. En julio obtendrán la calificación de NO PRESENTADO los alumnos que no se hubieran presentado al examen final de esa fecha.
Si algún estudiante quiere hacer alguna de las pruebas en un idioma oficial específico (gallego o español), debe avisar al profesorado por lo menos 1 semana antes de la correspondiente prueba.

Fuentes de información	
Básica	- Anderson, T.W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley - Chatfield, C., Collins, A. J. (1980). Introduction to multivariate analysis. Chapman & Hall - Giraud, C. (2014). Introduction to High-Dimensional Statistics. Chapman & Hall/CRC - Goldstein, M., Dillon,W. R. (1984). Multivariate Analysis: Methods and Applications. Wiley - Jambu, M. (1991). Exploratory and Multivariate Data Analysis. Boston, Academic Press - Jobson, J.D. (1994). Applied Multivariate Data Analysis. Springer-Verlag - Johnson, R. A., Wichern, D. W. (2007). Applied multivariate statistical analysis. Prentice Hall - Koch, I. (2014). Analysis of Multivariate and High-Dimensional Data. Cambridge University Press - Mardia, K.V., Kent, J.T., Bibby, J.M. (1994). Multivariate Analysis. Academic Press. Academic Press - Muirhead, R.J. (1982). Aspects of multivariate statistical theory. John Wiley & Sons - Rencher, A.C. (1998). Multivariate Statistical Inference and Applications. Wiley - Wainwright, M.J. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press
Complementária	

Recomendaciones
Asignaturas que se recomienda haber cursado previamente

Introducción a las Bases de Datos/614G02008
Álgebra Lineal/614G02001
Cálculo Multivariable/614G02006
Matemática Discreta/614G02002
Fundamentos de Programación II/614G02009
Fundamentos de Programación I/614G02004
Inferencia Estadística/614G02007
Probabilidad y Estadística Básica/614G02003
Asignaturas que se recomienda cursar simultáneamente
Modelos de Regresión/614G02012
Asignaturas que continúan el temario
Técnicas de Simulación y Remuestreo/614G02036
Análisis Estadístico de Datos Complejos/614G02031
Aprendizaje Automático III/614G02026
Recuperación de Información/614G02027
Aprendizaje Automático I/614G02019
Aprendizaje Automático II/614G02021
Análisis Estadístico de Datos con Dependencia/614G02022
Otros comentarios

(*) La Guía Docente es el documento donde se visualiza la propuesta académica de la UDC. Este documento es público y no se puede modificar, salvo cosas excepcionales bajo la revisión del órgano competente de acuerdo a la normativa vigente que establece el proceso de elaboración de guías