Guía docente de Asignatura– Grado en Estadística Aplicada

Datos generales de la asignatura

Asignatura:	**Estudio y Depuración de Datos - 801598**
Curso académico:	2019-20
Carácter	Obligatoria
Curso:	Segundo
Semestre:	4
Créditos ECTS	
Presenciales:	2,4
No presenciales:	3,6
Total	6,0
Actividades docentes	
Clases teóricas:	35%
Seminarios:	12,5%
Clases prácticas:	52,5%
Total	100%
Departamentos responsables:	Departamento de Estadística y Ciencia de los Datos
Profesores:	Daniel Gómez González / Francisco Javier López Ipiña Mattern / Magdalena Ruth Ferrán Aranaz

Datos específicos de la asignatura

Breve descriptor:	Depuración, codificación, transformaciones y tratamiento previo al análisis estadístico de una base de datos.
Requisitos:	Conocimientos básicos de descripción y exploración de datos, azar y probabilidad, estimación. Partimos de los conocimientos de SAS que se imparten en la asignatura Software estadístico I.
Competencias	
Generales:	CG 9-AD1- Reducir la información de interés para su tratamiento y análisis. CG 10-AD1- Realizar trabajos con otros estudiantes y debatir sobre el análisis de datos adecuado.
Específicas:	CE 2-AD1- Depurar un conjunto cualquiera de datos para su posterior análisis estadístico. CE 5-AD1- Buscar y encontrar patrones de comportamiento en los datos. CE 21-AD1- Utilizar correctamente el software estadístico programable.

Contenidos

TEMA 1. FICHEROS DE DATOS: TRANSFORMACIONES Y CODIFICACIÓN

1.1 Tipos de variables: Nominales, Ordinales y Continuas.

1.2 Transformaciones de variables

1.3 Recuento de valores en los casos

1.4 Recodificación de variables.

1.5 Categorización de variables.

1.6 Asignación de rangos

1.7 Chequeo y recodificación de variables.

1.8 Categorización.

1.9 Manipulación de fechas.

TEMA 2. CONTROL DE INTEGRIDAD DE LOS DATOS.

2.1.- El problema de los datos atípicos.

2.2 Detección de casos atípicos

2.2.1 Gráficos de control.

2.2.2 Control del valor de los datos estandarizados.

2.3 Detección de casos atípicos en distribuciones multidimensionales. Distancia de Mahalanobis.

2.3.1. Gráficos de Caja y Bigotes.

2.3.2.- Diagramas de Dispersión.

2.4.- Detección y tratamiento de duplicados.

TEMA 3. VALORES PERDIDOS.

3.1 El problema de los datos perdidos.

3.2. Análisis de valores perdidos.

3.2.1. Diagnóstico de valores perdidos.

3.2.2. Métodos de cálculo estadísticos muestrales.

3.2.3. Imputación de valores perdidos: Algoritmo EM y método de regresión.

3.3. Análisis y estimación de valores perdidos.

TEMA 4. DATOS MISSING II: IMPUTACIÓN MÚLTIPLE.

4.1. Imputación de valor perdidos en variables unidimensionales.

4.2. Introducción a los métodos de imputación múltiple de valores perdidos.

4.2.1. Análisis de patrones de datos missing.

4.2.2. Imputación múltiple de variables numéricas y categóricas.

4.2.3. Utilización de los ficheros generados mediante imputación Múltiple.

TEMA 5. NORMALIDAD MULTIVARIANTE Y TRANSFORMACIONES BOX-COX PARA CONSEGUIR NORMALIDAD.

5.1 La distribución Normal Multivariante.

5.1.1. Propiedades

5.1.2. Análisis de la hipótesis de normalidad univariante.

5.1.2.1. Métodos Gráficos

5.1.2.2. Contrastes de hipótesis.

5.2. Transformaciones Box-Cox.

5.3. Contraste de hipótesis de Normalidad Multivariante.

Evaluación

Se utilizará el procedimiento de evaluación continua exclusivamente para aquellos alumnos que asistan a las clases de prácticas.
Estos alumnos podrán realizar trabajos mediante software de aplicación específico y ser evaluados por ellos.
La nota final tendrá en cuenta tanto la evaluación continua como la prueba final. Se calculará como el máximo entre:

a) La calificación de la prueba final.
b) La media ponderada de la evaluación contínua y la prueba final, siendo el peso de la evaluación contínua de al menos el 35%.

En todo caso, sí se podrá superar la asignatura mediante el procedimiento de evaluación continua.

Bibliografía

- Cody, Ron. "Cody's Data Cleaning Techniques Using SAS Software". Ed SAS Publishin; 1999.
- Boehmke, Bradley. "Data Wrangling with R". Springer. 2016.