

En este trabajo final se pretende que el alumnado realice un informe final en relación a un problema de interés basado en datos recogidos para analizarlo. Con este objetivo se realizará un análisis exploratorio de los mismos y se tomarán decisiones en función de lo aprendido de los datos. Este análisis se realizará en dos fases:

1. **Análisis exploratorio univariante.** En esta fase se recomienda realizar un análisis exploratorio preliminar de los datos contenidos en **UNA sola de las bases de datos descritas al final de este documento**, o cualquier conjunto de datos que el alumnado considere de interés. En esta práctica el alumnado aplicará las distintas técnicas numéricas y gráficas aprendidas en clase. En un primer momento, se centrará en el **análisis de cada una de las variables de forma independiente** sin buscar, aún, posibles interacciones entre ellas (**análisis univariante**). Para ello se pide realizar los siguientes **análisis numéricos y gráficos** de cada variable, para detectar:

- a) **Recodificaciones o agrupaciones de datos** si lo considera oportuno mediante el visionado de la estructura del archivo de datos.
- b) **Valores perdidos** mediante la carga y visionado de datos. Para ello se deben realizar los siguientes pasos:
 - i.) De cada variable identificar el **% de valores perdidos**.
 - ii.) De las variables que tengan más del 5% de valores perdidos **analizar el patrón aleatorio** o no de los mismos. Para ello, estudiar la homogeneidad según grupos (NA y no NA) con otras variables. Si son continuas con un test de student, si son cualitativas o discretas con test de independencia Chi-cuadrado, etc.
(Investigar funciones para el contraste de medias como *t.test()*, etc. del lenguaje R)

En el caso de **homogeneidad el patrón es aleatorio** y, en este caso, se elige **sustituir** el NA por la media o la moda, según si es cuantitativa o cualitativa.
(Utilizar el código de las prácticas de clase.)

En el caso de que **no haya homogeneidad**, el **patrón no es aleatorio**. Esto habría que tratarlo con el investigador que plantea el problema bajo análisis porque **no se deberían ni eliminar ni sustituir**, pero como habitualmente no es factible esto, **se decide actuar como en el caso de patrón aleatorio, avisando de este hecho en el informe final**.

- c) Análisis descriptivo **numérico clásico** (medidas de tendencia central, dispersión, cuartiles, simetría, curtosis, etc.)
(Investigar funciones como *summary*, *skewness*, etc. del lenguaje R)
- d) **Valores extremos** (outliers) apoyándose en los resultados numéricos del apartado anterior así como en resultados gráficos (boxplots).
(Utilizar la función *boxplot()* del lenguaje R)

En el supuesto de que los haya **tomar la decisión de eliminarlos**, si el archivo de datos tiene suficientes registros, **o sustituirlos por la media o moda** según si la variable es cuantitativa o cualitativa.
(Utilizar el código de las prácticas de clase.)

- e) Muchas técnicas estadísticas no pueden evitar el **supuesto de normalidad**. En este sentido analizar este supuesto para las distintas variables continuas de la base de datos. Para ello se debe tratar de justificarlo o descartarlo de forma gráfica con gráficos de normalidad (qqplots, etc.).
(Investigar *qqplot()* del lenguaje R)

-
- f) Cualquier otra cuestión que se considere de interés para un buen entendimiento de los datos.
2. **Análisis exploratorio multivariante.** En segundo lugar, se comprobarán los supuestos subyacentes a la aplicación de las distintas técnicas multivariantes de reducción de la dimensión, como pueden ser el ACP o el AF, antes de ser aplicadas. Para ello se pide:
- a) Comprobar los supuestos de correlación entre variables con el test de Barlett.
(Utilizar el código de las prácticas de clase.)
 - b) Se asume que en el análisis univariante anterior se han identificado y tratado los outliers, si no es así hay que hacer el análisis de los mismos antes de aplicar técnicas de reducción de la dimensión.
(Utilizar el código de las prácticas de clase.)
 - c) Si no se han tratado los valores perdidos NA porque no superaran el 5% según lo indicado en el análisis univariante anterior, llegado a este momento hay que tomar decisiones sobre ellos para poder utilizar técnicas de reducción de la dimensión.
(Utilizar el código de las prácticas de clase.)
 - d) En este punto se pide realizar un estudio de la posibilidad de reducción de la dimensión mediante variables observables. Es conveniente elegir el número óptimo de componentes principales por las distintas técnicas gráficas introducidas en clase.
(Utilizar el código de las prácticas de clase.)
 - e) Del mismo modo, se pide realizar una reducción de la dimensión mediante variables latentes, eligiendo previamente el número óptimo de factores a considerar.
(Utilizar el código de las prácticas de clase.)
 - f) Previo a la construcción de métodos de clasificación analizar la normalidad multivariante de los datos con el test propuesto en clase de prácticas.
(Utilizar el código de las prácticas de clase.)
 - g) A continuación se procederá a construir un clasificador mediante un análisis discriminante lineal y otro cuadrático.
(Utilizar el código de las prácticas de clase.)
 - h) Finalmente realizaremos una validación muy básica de los clasificadores obtenidos mediante la representación gráfica de su respectiva matriz de confusión.
(Utilizar el código de las prácticas de clase.)

INDICACIONES

1. **Utilizar RMarkdown** para la realización del análisis exploratorio anterior para tener una visión general de las distintas salidas obtenidas con R así como el texto que describa las opiniones del alumnado en función de las mismas.
2. Realizar el **informe final con un procesador de textos científicos**, preferiblemente LaTeX. También se aceptarán trabajos escritos en Word, Writer, etc. También se puede incrustar LaTeX en el documento RMarkdown y presentar el pdf que genera.
3. El informe final podría incluir las siguientes secciones.
 - **Resumen o abstract** de no más de 200 palabras poniendo en contexto el problema elegido, indicando que técnicas se han aplicado y con qué objetivo para terminar con una línea o dos que describa una conclusión final.
 - **Introducción** de no más de 400 palabras que extienda un poco el resumen anterior. Esta sección debe terminar con un párrafo de dos o tres líneas definiendo el objetivo del trabajo a realizar.
 - **Materiales y Métodos.** Esta sección podría incluir una subsección, 'Materiales' que describa brevemente la base de datos, informando de lo que almacenan las distintas variables y aportando una tabla con los estadísticos descriptivos básicos (media y desviación típica para variables cuantitativas; % y totales para variables categóricas). La segunda subsección, 'Métodos estadísticos', de no más de 400 palabras indicará las distintas técnicas estadísticas utilizadas. Se insiste en que en esta subsección **se indican las técnicas, no se explican** ni se dan clases magistrales de las mismas.
 - **Resultados.** Esta sección debería mostrar un resumen de los resultados más destacados obtenidos en el desarrollo de esta práctica. Debe ser una exposición objetiva de resultados sin interpretación en el contexto del problema.
 - **Discusión** de no más de 600 palabras que interprete los resultados obtenidos. Esta sección debería comenzar recordando cuál era el objetivo u objetivos que se anunciaban en el párrafo final de la introducción, para a continuación discutir los que se han conseguido y cómo en función de los resultados.
 - **Conclusión** de no más de 250 palabras que haga una síntesis de lo conseguido, hable de las fortalezas del trabajo realizado, informe de las limitaciones del mismo y que haga propuestas de mejora o indique otros caminos que podrían seguirse o abrirse en el contexto analizado.
4. **FORMA DE ENTREGA.** El alumnado subirá a la tarea creada para esta práctica en la plataforma PRADO cuatro ficheros: el **código fuente RMarkdown**, la **salida html obtenida con RMarkdown**, el **archivo fuente de LaTeX** con el informe final (si se opta por otro procesador de textos, se pedirá el archivo editable creado por el alumno) y un **archivo pdf con el informe final** compilado en el caso de LaTeX, o guardado como pdf desde cualquier otro editor de texto utilizado.

A continuación se encuentra una **descripción de cada base de datos** y sus variables. Se recomienda que el **informe de RMarkdown contenga esta descripción** en la sección de materiales y métodos.

1. **DB_1:** El fichero de datos DB_1.sav contiene, entre otras, las variables znac_def, zmortinf, zfertil, zinc_pob, ztasa_na, zurbana, zalfabet, zcaloria, zlog_pib, zpib_cap, zpoblac, zdensida, que son las variables estandarizadas de las originales de igual denominación sin la z inicial, y que respectivamente son los valores para cada país del mundo de:
 - Tasa Nacimientos/Defunciones (nac_def).
 - Mortalidad infantil: muertes por 1000 nacimientos vivos (mortinf).
 - Fertilidad: numero promedio de hijos (fertil).
 - Aumento de la poblacion en % anual (inc_pob).
 - Tasa de natalidad por 1.000 habitantes (tasa_na).
 - Habitantes en ciudades en % (urbana).
 - Personas Alfabetizadas en % (alfabet).
 - Ingesta diaria de calorías (calorias).
 - Log(10) de PIB_CAP (log_pib).
 - Producto interior bruto per-capita (pib_cap).
 - Poblacion en miles (poblac).
 - Habitantes por Km2 (densidad).
2. **DB_2:** Un grupo constituido por 13 empresas se ha clasificado según las puntuaciones obtenidas en 8 indicadores económicos en el archivo DB_2.sav:
 - X1: Indicador de volumen de facturación de la empresa.
 - X2: Indicador del nivel de nueva contratación.
 - X3: Indicador del total de clientes
 - X4: Indicador de beneficios de la empresa.
 - X5: Indicador de nivel de retribución salarial de los empleados.
 - X6: Indicador de nivel de organización empresarial dentro de la empresa.
 - X7: Indicador de nivel de relaciones con otras empresas.
 - X8: Indicador de nivel de equipamiento (ordenadores, maquinaria, etc.).
3. **DB_3:** En el conjunto constituido por 34 estados del mundo se han observado 11 variables cuyos resultados se recogen en el archivo DB_3.sav. Estas variables se han estandarizado, pues están tomadas con unidades de medida muy diferentes. Estas variables son:
 - Ztlibrop: Número de libros publicados.
 - Ztejercci: Cociente entre el número de individuos en ejército de tierra y población total del estado.
 - Ztpobact: Cociente entre población activa y total.
 - Ztenergi: Tasa de consumo energético.
 - Zpservi: Población del sector servicios.
 - Zpagricu: Población del sector agrícola.
 - Ztmedico: Tasa de médicos por habitante.
 - Zespvida: Esperanza de vida.
 - Ztminfan: Tasa de mortalidad infantil.
 - Zpobdens: Densidad de población

-
- Zpoburb: Porcentaje de población urbana.

4. **DB_4:** Se ha realizado una encuesta a un grupo de trabajadores de una gran empresa para conocer lo que piensan sobre la organización y funcionamiento de la misma. Estos datos se encuentran recogidos en el archivo DB_4.sav. Los encuestados responden en una escala de 0 a 10 según el grado de acuerdo con cada uno de los ítems que aparecen a continuación:

- CODIGO: Código de trabajador.
- ITEM1: No hay que hacer cambios constantemente en el modo de proceder.
- ITEM2: Funcionará mejor si cada uno introduce su modo y estilo.
- ITEM3: Los problemas provienen de no hacer reformas importantes.
- ITEM4: Hay que generar más interacción entre los trabajadores.
- ITEM5: Cada uno debe obrar según su estilo propio.
- ITEM6: Los problemas se evitan con más disciplina.
- ITEM7: Aquí hace falta más libertad al trabajador y menos órdenes.
- ITEM8: No basta trabajar. El progreso se basa en buena gestión.
- ITEM9: Guiarse por las opiniones de unos y otros puede ser peligroso.
- ITEM10: Las cosas no funcionan cuando no se tiene en cuenta a todos.
- CATEGORI: Grupo de edad del trabajador:
 - 1,00 MJ (Muy joven)
 - 2,00 J (joven)
 - 3,00 Ma (mayor)
 - 4,00 Mma (muy mayor)