ESTADÍSTICA MULTIVARIANTE (Grado en Matemáticas)
Práctica final evaluable

En este trabajo final se pretende que el alumnado realice un informe profesional en relación a un problema de interés basado en un conjunto de datos de elección particular. Con este objtetivo, se realizará un análisis exploratorio y se tomarán decisiones en función de lo aprendido con los datos. Este análisis se realizará en dos fases:

1. **Analisis exploratorio univariante**

   En esta fase se recomienda realizar un análisis exploratorio preliminar de los datos contenidos en el conjunto de datos considerado. Para ello aplicará las distintas técnicas numéricas y gráficas utilizadas 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 recomienda realizar **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 mediana, según si es cuantitativa o cualitativa.
      (Utilizar el código fuente 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 en este caso no es factible, **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.)
   (*Utilizar el código fuente de las prácticas de clase.*)

   d) **Valores extremos** (outliers) apoyándose en los resultados numéricos del apartado anterior así como en resultados gráficos (boxplots).
   (*Utilizar el código fuente de las prácticas de clase.*)

   En el supuesto de que haya valores extremos se va a **tomar la decisión de eliminarlos**, si el archivo de datos tiene suficientes registros, **o sustituirlos por la media o mediana** según si la variable es cuantitativa o cualitativa.
   (Utilizar el código fuente de las prácticas de clase.)

   e) Muchas técnicas estadísticas no pueden evitar el **supuesto de normalidad**. En este sentido, se debe 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 de correlación entre variables con el test de Barlett.
   (Utilizar el código fuente 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 fuente de las prácticas de clase.)

   c) Si no se han tratado los valores pérdidos 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 fuente 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 fuente 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 fuente de las prácticas de clase.)

   f) Previo a la construcción de métodos de clasificación se debe analizar la normalidad multivariante de los datos con el test propuesto en clase de prácticas de análisis discriminante (tema 5).
   (Utilizar el código fuente 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 fuente 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, curva COR y distintas medidas de seguridad y validez interna (sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo).
   (Utilizar el código de las prácticas de clase.)

   i) Como añadido, se puede realizar un análisis clúster que confirme que la agrupación de la variable respuesta utilizada en los modelos de clasificación es adecuada. Si no se tuviera variable respuesta para el apartado g) anterior, habría que hacer primero este análisis clúster que la definiera para después utilizarla en la clasificación.

INDICACIONES

1. **Utilizar RMarkdown** para la realización del análisis exploratorio anterior para tener una visión general de las distintas salidas obtenidas con R. Este código puede tener 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 o html 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 intreprete 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 abrise 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.

5. **FECHA LÍMITE DE ENTREGA**

    Hasta el día previo al examen de la convocatoria ordinaria de la asignatura (**xx de xxxxx de 202x**).