ANÁLISIS DE DATOS – 2º de BIOLOGÍA
PRÁCTICA 4

Guión para realizar la práctica: Estudio de la variable **Especie** como función de las variables **Anchura** y **Longitud** tanto del pétalo como del sépalo. Se utilizarán los datos que están en el archivo: `iris.dat`.[1] Es un archivo de texto con un caso por fila, con las variables separadas por ";". Los nombres de las variables no están en la primera fila, por lo que tras importar los datos a SPSS habrá que darles el nombre apropiado. El orden en el que vienen las variables es el que aparece más abajo.

Se trata buscar un criterio para clasificar un ejemplar en su especie utilizando las cuatro medidas dadas. Los datos corresponden a tres especies del género *Iris* nativos del Canadá: *I. setosa*, *I. versicolor, I. virginica*, aunque la discriminación la haremos solamente para las especies *versicolor* y *virginica.*

Previamente al estudio haremos un análisis descriptivo de las variables a estudiar, que vamos a codificar como:
LS: Longitud del sépalo
AS: Anchura del sépalo
LP: Longitud del pétalo
AP: Anchura del pétalo
ES: Especie

Para ello, utilizamos **Analizar → Estadísticos descriptivos → Explorar...**
En el cuadro de diálogo pasamos a la "Lista de dependientes" las variables **LS**, **AS**, **LP**, y **AP** y en la "Lista de factores" solamente **ES**. Los botones "Estadísticos…" y "Opciones…" no los tocamos; en el botón "Gráficos…" solamente marcamos en "Diagramas de cajas" la opción "Niveles de los factores juntos".

Cuestión 1: Extrae a mano en un cuadro las medias y desviaciones típicas de cada una de las variables para cada una de las especies
Responde en la hoja de respuestas

Cuestión 2: A la vista de los diagramas de cajas, ¿qué especie o especies resulta simple clasificar por medio de una sola de las cuatro variables?
Responde en la hoja de respuestas

A continuación vamos a buscar un método de clasificación para las especies *I. versicolor* e *I. virginica* por medio de regresión logística, utilizando las cuatro variables; para ello recodificamos primeramente las especies una de ellas (p. ej., *I. virginica*) con "1" y las otras con "0" (podemos llamar a la nueva variable **Virginica** de forma que el valor "1" indica que la especie es *virginica* y el valor "0" que no lo es). A continuación, seleccionamos en la base de datos solamente estas dos especies (recuérdese: Datos Seleccionar casos… Si se satisface la condición… etc.).

Una vez echa la selección, utilizamos el procedimiento **Analizar → Regresión → Logística binaria...** Seleccionamos la variable **Virginica** en "Dependientes" y las cuatro variables **LS**, **AS**, **LP**, y **AP** como "Covariables". Solamente modificamos las opciones del botón "Opciones..." donde seleccionamos "En el último paso" en "Visualización".

Observa la tabla "Variables en la ecuación".

Cuestión 3: Escribe la ecuación logística estimada utilizando los resultados que da esta tabla.
Responde en la hoja de respuestas

1 Fisher, R. A. *The use of multiple measurements in taxonomic problems.* Annual Eugenics, **7**, Part II, 179–188 (1936).

Cuestión 4: Escribe el criterio de discriminación que da esta ecuación. Utiliza el valor Y=0.5 para dividir los casos.
Responde en la hoja de respuestas

Cuestión 5: Escribe los p-valores resultantes para cada uno de los coeficientes y observa de qué coeficientes no se puede afirmar que sean significativamente distintos de cero.
Responde en la hoja de respuestas

Entre las tablas obtenidas se encuentra la "Tabla de clasificación" que muestra (con valor de corte 0.5) cuantos de los 100 datos se clasifican bien (y cuantos mal) según el criterio de clasificación anterior.

Cuestión 6: ¿Cuántos hay en cada situación?
Responde en la hoja de respuestas

Parece que los p-valores correspondientes a los coeficientes de las variables **LS** y **AS** son algo altos (mayor el primero. Vamos a ver de qué tamaño son las correlaciones lineales entre las variables explicativas.

Vamos a **Analizar** → **Correlaciones** → **Bivariadas...** y elegimos las cuatro variables.

Cuestión 7: ¿Qué dos variables presentan un coeficiente de correlación más alto y cuál es este valor?
Responde en la hoja de respuestas

Dada la correlación lineal estimada entre **LS** y **LP**, repitamos la regresión logística binaria suprimiendo la variable **LS**.

Cuestión 8: ¿Qué coeficientes y p-valores se obtienen ahora?
Responde en la hoja de respuestas

Cuestión 9: ¿Cuál es la nueva ley de clasificación?
Responde en la hoja de respuestas

Cuestión 10: ¿Cuántos datos de cada especie se clasifican correctamente utilizando este criterio?
Responde en la hoja de respuestas