Modelos probabilísticos y análisis estadístico / Métodos estadísticos para data analytics

Segundo taller

- **Descripción de los datos:** Pima Indians Diabetes Database.
- **Donador:** Vincent Sigillito.
- **Publicación en la que fue usado el dataset:** Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Annual Symposium on Computer Application in Medical Care (p. 261). American Medical Informatics Association.
- **Información de la muestra:** Todos los pacientes son mujeres mayores de 21 años pertenecientes a la etnia Pima. Otros detalles concernientes con el algoritmo ADAP pueden ser consultados en el artículo. Número of casos (pacientes): 768. Existen valores perdidos: si. Número de variables: 8 más la clase (novena columna)
- [**preg:**] Número de veces embarazadas
- [**plas:**] Concentración de glucosa plasmática después de 2 horas de aplicada una prueba oral de tolerancia a la glucosa (GTT)
- [**pres:**]Presión arterial diastólica (mm Hg)
- [**skin:**] Espesor del pliegue cutáneo ubicado sobre el músculo tríceps (mm)
- [**test:**] Niveles de insulina en suero a las 2 horas (U/ml)
- [**mass:**]Índice de masa corporal (IMC) (peso en kg/(estatura en m)^2)
- [**pedl:**]Función pedigree de la diabetes (tipo de diabetes)
- [**age:**] Edad (años)
- [**class:**] Atributo de clase (0: negativo para diabetes; 1: positivo para diabetes)
- **Motivación:** El correcto análisis de los datos es el principal insumo para la toma de decisiones acertadas. Los médicos y los investigadores están tomando decisiones críticas todos los días. Por lo tanto, es necesario que estas personas tengan algún conocimiento básico de análisis de datos. Esta actividad tiene como objetivo reforzar las habilidades de los estudiantes de maestría de la Facultad de Ciencias Naturales e Ingeniería de la Universidad de Bogotá Jorge Tadeo Lozano para procesar y describir datos.

- **Instrucciones:** Desarrolle los siguientes ejercicios y dé respuesta a las preguntas en un archivo de texto cuyo nombre corresponda con el suyo. Envíe este archivo al correo: **rodrigo.gil@utadeo.edu.co** antes de la media noche del miércoles 23 de septiembre. Adjunte también el código en R que le permitió solucionar cada uno de los puntos.

Puntos a desarrollar

Con base en los datos descriptos anteriormente (DatosDiabetis.csv) desarrolle cada uno de los siguientes ejercicios.

1. **Ejercicio 1: Resumen de información en tablas**.

   - Construya una tabla (sólo una tabla) con estadísticas descriptivas básicas: media, mediana y coeficiente de variación; para cada una de las variable pero separando la muestra en dos grupos dependiendo del valor de la novena columna (0: Negativo para diabetes; 1: positivo para diabetes).
   - Asigne un nombre conveniente a la tabla y redacte un párrafo (evite que sea una sola oración) describiendo el conjunto de datos con base en los resultados mostrados en la tabla. Por tratarse de varias variables puede ser más fácil se enfoque sólo en algunas (mínimo tres). Esta descripción deberá contener comparaciones entre las dos clases definidas por la novena columna (0:negativo para diabetes; 1:positivo para diabetes). Es común que el párrafo anteceda a la tabla y que dentro del texto se haga referencia a ella, por ejemplo: En la tabla 1 se puede observar que ....

2. **Ejercicio 2: Resumen de información en gráficas**.

   - Construya una gráfica (solo una) que deje ver la relación entre las variables mass y skin; pero en la cual también se pueda distinguir de manera clara el efecto de la variable clase (class) sobre la relación entre las variables. Es decir, asigne colores diferentes a los puntos dependiendo si son 0 o 1 (variable class).
   - Asigne un título apropiado a la gráfica que construyó de manera que se pueda entender claramente la información contenida en ella. Redacte un pequeño párrafo interpretando la información contenida en la gráfica. Al igual que el caso anterior, el texto precede a la gráfica.

3. **Ejercicio 3: Explorar correlaciones**.
   Explore el nivel de asociación entre las variables mass y skin para las dos clases por separado (variable class). Sea cuidadoso al momento de seleccionar el método de correlación y una vez lo haya definido justifique brevemente su decisión. Al momento de realizar la correlación reflexione acerca de los siguiente aspectos:

   - ¿Se deben excluir datos como aquellos que indican valores de cero para mass y skin?
   - ¿Se deberá comprobar que el valor de la correlación es significativamente diferente de cero?

   Tenga en cuenta estos elementos al momento de redactar la interpretación.