UNIVERSIDAD DEL VALLE – FACULTAD DE INGENIERIA
Programa Académico de Estadística

Asignatura: Técnicas de Minería de Datos y Aprendizaje Automático
Profesor: Jaime Mosquera Restrepo
Fecha Entrega: Miércoles 26 de Abril – 6:00 p.m.

Laboratorio No. 1 - Preprocesamiento de Datos
(Estructura de Datos, Consistencia, Limpieza, Datos atípicos, Datos faltantes, Visualización)

El contexto del problema - el objetivo de análisis o pregunta de negocio.

El conjunto de datos *Calcium.xls* fue recogido por *Boyd, Delost y Holcomb* (1998) con el objetivo principal de analizar las diferencias en los niveles de Calcio (**CaMol**), Fosforo (**PhoMol**) y Fosfatasa Alcalina (**ALP**) para pacientes mayores de 65 años de edad en función del género (***Sex*** = Male or Female). El segundo objetivo fue determinar si la variación de las condiciones analíticas entre laboratorios o la edad de los pacientes, afecta a la distribución de las 3 variables de estudio.

Para cumplir con los objetivos, los investigadores realizaron una revisión retrospectiva de los procedimientos de laboratorio desarrollados en 6 diferentes instituciones prestadoras de servicios de médicos. Los datos contienen la información de 178 pacientes (92 Hombres y 86 Mujeres) mayores de 65 años, respecto a 3 variables cuantitativas: **Age** (Años), **ALP** (IU/L), **CaMol**(IU/L), **PhoMol**(IU/L) y 3 variables cualitativas: **Sex**, **Lab** y **AgeG**. Las variables cualitativas deben seguir la siguiente codificación:

Sex	1=Male; 2=Female
Lab	1=Metpath; 2=Deyor; 3=St. Elizabeth's; 4=CB Rouche; 5=YOH; 6=Horizon
AgeG	65-69; 70-74; 75-79; 80-84; 85-89 years

A usted se le solicita realizar un análisis exploratorio de datos. En ese sentido, se requiere que usted diseñe una visualización contundente datos, a través de tableros gráficos resumen, en la cual se evidencie la diferencia entre genero para todas las variables cuantitativas, además de posibles diferencias debidas a la edad y al laboratorio. Adicionalmente se requiere visualizar, de forma sintética, la estructura de correlación entre las variables cuantitativas.

Para más información acerca del contexto del estudio y de los rangos de referencia de las variables cuantitativas, es conveniente que usted revise el articulo original de la investigación (https://search.proquest.com/docview/204793424?pq-origsite=gscholar&fromopenview=true)

El preprocesamiento y limpieza de los datos

En una inspección rápida de la hoja de datos, usted podrá notar la presencia de algunos registros inconsistentes, datos faltantes y datos atípicos. Para evitar sesgos sobre los resultados y pérdida de registros, es necesario que usted, previo a realizar cualquier análisis, realice una actividad de limpieza de datos utilizando herramientas de software (R).

Desarrollo del Laboratorio en R.

Para realizar el ejercicio en R, usted debe seguir el siguiente libreto de limpieza y preprocesamiento:

1. Lea la hoja de datos y adecúe el formato de cada variable, verificando que dispone de una hoja de datos técnicamente correcta.
2. Construya el archivo: *consistencia.txt*, en el cual incluya las ecuaciones que usted considera necesarias para verificar la consistencia de los datos en el conjunto de variables. Aplique estas reglas sobre la hoja de datos y genere un pequeño reporte de sus resultados.
3. Visualice e identifique los registros que presentan datos faltantes.
4. Con los resultados de los puntos 2 y 3, usted dispone de un listado con los registros inconsistentes y con datos faltantes. Es necesario corregirlo.

   Si requiere corregirlo, en el siguiente enlace: https://docs.google.com/spreadsheets/d/1-9uLmMwl_95cEWrShuqEOdcrJH_Kanm3/edit?usp=sharing&ouid=111645287266110588487&rtpof=true&sd=true encontrará una herramienta de consulta que le permitirá acceder a todos los datos correctos (*seleccione variable y número de registro y la herramienta le generará el dato correcto*). Es posible que la inconsistencia detectada o el dato faltante sea producto de una omisión en la fase de digitación. Ahora que conoce el verdadero dato, usted puede corregirlo.

5. Sobre el conjunto de variables cuantitativas, realice un diagnóstico de datos atípicos. Utilice los dos enfoques, univariado y multivariado. Para cada dato atípico identificado, decida si debe ser retenido o aislado del análisis de datos.
6. Ahora usted tiene una hoja de datos con unos pocos datos faltantes. Algunos de ellos son originalmente ausentes y otros se convirtieron en ausentes por ser datos atípicos aislables. Sugiera el método adecuado para realizar la imputación de estos datos y ejecútela.
7. Genere un resumen de los cambios realizados en la hoja de datos. *ReporteCambios.txt*

Perfecto, ahora usted tiene una hoja de datos limpia, guárdela en el archivo *clean_calcium.csv*. Ha llegado el momento de visualizar los datos.

Visualización de datos.

8. Utilice su pericia de estadístico para resumir los datos en uno o pocos tableros gráficos, en los cuales se pueda evidenciar:

   i. La distribución de los pacientes por edad, laboratorio y género.
   ii. Las diferencias de las variables clínicas: **ALP**, **CaMol, PhoMol** entre los grupos de edad (**AgeG**) , Sexo (**Sex**) y Laboratorio (**Lab**).
   iii. La estructura de correlación entre las variables cuantitativas: **Age**, **ALP**, **CaMol**, **PhoMol**.

   Nota 1: Por favor sea muy cuidadoso con la gestión de los gráficos. Ubique nombres adecuados para los ejes, leyendas y títulos. Sea consistente con el manejo de los colores e intente que su representación sea lo más contundente posible, que hable por si sola.

   Nota 2: El resultado se puede grabar en un archivo independiente (*Tablero_Grafico.tiff*) por cada tablero grafico generado o generar un solo

archivo (*Tablero_Grafico.html*) que contenga todos los tableros. Para generar el archivo html requerirá construir un Script de Rmarkdown.

Productos Entregables del Laboratorio.

Como entregable del presente laboratorio, usted debe ubicar en el campus virtual dos archivos comprimidos (o enlaces web que dirijan a los archivos comprimidos, cuando el archivo supere el tamaño máximo de carga el campus virtual) que contenga los siguientes elementos:

1. ***Solución_R.zip:*** *contiene los soportes de la solución del laboratorio en R*

a. El archivo *consistencia.txt*
b. El archivo *ReporteCambios.txt*
c. La nueva hoja de datos *clean_calcium.csv*
d. El script R, *Script_R.txt*, editado adecuadamente con una división desplegable asociada a cada uno de los puntos desarrollados (Puntos 1 a 6)
e. Una imagen de cada tablero gráfico generado, almacenada en formato tiff o un archivo html con la publicación del conjunto de tableros de R_markdown. En el caso de entregar múltiples imágenes, cada imagen debe llevar el nombre Tablero_Grafico1.tiff. El número en el nombre de archivo variará en función de la cantidad de tableros generados.

Condiciones de entrega.

1. Trabajo en equipo - El laboratorio debe ser desarrollado en grupos de 2 personas.
2. Forma y tiempo de Entrega – Entrega en el campus virtual. La asignación estará disponible para la carga de los entregables hasta el miércoles 26 de abril – 6:00 pm.

Success in your first data mining experience

Referencia Bibliográfica

Boyd, J., Delost, M., and Holcomb, J., (1998). Calcium, phosphorus, and alkaline phosphatase laboratory values of elderly subjects, Clinical Laboratory Science, 11(4), 223-227.