Tarea 2 – Data Mining

Sesión 2 y 3

Actividad.

Tomando en cuenta lo que vimos en clase y los dos notebooks que trabajamos acerca de limpieza de datos: datos faltantes y datos extremos, y también regresión lineal, vamos a hacer un ejercicio completo. Para esto vamos a ir a la página siguiente en donde vamos a descargar datos, cada uno según su predilección y gusto. Todos los dataset los vamos a encontrar en la página web https://archive.ics.uci.edu/ de la UC Irvine Machine Learning Repository. Por favor seleccione un dataset pequeño, no más de 1000-2000 observaciones, para evitar el colapso de su equipo y ayudarle en la interpretación de los resultados.

Lo que queremos hacer es lo siguiente:

1. Descargue y cargue los datos en su notebook, usando Pandas preferiblemente. Imprima para tener una idea de los datos.
2. Calcule la cantidad de datos faltantes y luego si no tiene cree artificialmente y aleatoriamente datos faltantes.
3. Use las metodologías vistas en clase para rellenar, o poner, datos en los lugares de los datos faltantes.
4. Haga gráficos para evaluar la calidad de este llenado y las formas de las distribuciones.
5. Escriba conclusiones acerca del proceso, ¿Qué metodología de imputación le sirvió mejor?
6. Determine si tiene datos outliers o datos extremos y diga cuales son y elimínelos. La información para esta tarea está en el primer notebook.
7. Adicionalmente use los algoritmos que encontrará en el segundo notebook (linear_regresion) para inferir regresiones, determine parámetros, determine el grado de regresión o coeficiente de regresión (bondad de ajuste) y escriba conclusiones.
8. Debe hacer todo en un sólo notebook y presentar conclusiones sobre el tema.