Tarea 2 Módulo 1 - Análisis de Datos 2022-1

D. Sierra-Porta

April 13, 2022

1 Asignación - Trabajo de campo

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-beta.ics.uci.edu/ de la UC Irvine Machine Learning Repository.

Para tener una mejor idea y dirigirse más rápido a cada uno de los dataset vamos a la página web: http://odds.cs.stonybrook.edu/

Lo que queremos hacer es lo siguiente:

1. Descarge y carge 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 info para esta tarea está en el primer notebook.

7. Adicionalmente use los algoritmos que encontrará en el segundo notebook 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.