Ejercicio 2. Ciencias de Datos 22.enero.2021

El objetivo de este ejercicio es iniciar al estudiante en el preprocesamiento de datos. Este preprocesamiento involucra: a) conocer el tipo de distribución de los datos, b) resumir los datos, c) muestrear los datos y d) graficar datos.

Emplear el dataset proporcionado (weather de la ciudad de Szeged, Hungría), el cual contiene casi 96500 instancias con 12 variables. https://www.kaggle.com/budincsevity/szeged-weather/data No se debe emplear los datos "crudos" tal como se obtengan, sino que se debe obtener una muestra sobre la que se trabajará. Para ello es necesario conocer los tipos de datos, valores, distribución de valores, etc.

1. Revisar Capítulos 3 y 4 del libro *Python Data Science Handbook*
2. Revisar el libro *Statistics and Machine Learning in Python* ftp://ftp.cea.fr/pub/unati/people/educhesnay/pystatml/StatisticsMachineLearningPythonDraft.pdf
3. Determinar el tipo de distribución que siguen los datos.
    a. https://pythonhealthcare.org/2018/05/03/81-distribution-fitting-to-data
    b. https://hackdeploy.com/fitting-probability-distributions-with-python
4. Con base en lo anterior realizar CUATRO tipos de muestreo (a discreción -dos aleatorios y dos no aleatorios. Leer el documento "samplingtechniques.pdf".
    a. https://machinelearningmastery.com/statistical-sampling-and-resampling
    b. https://people.duke.edu/~ccc14/sta-663/ResamplingAndMonteCarloSimulations.html
    c. https://towardsdatascience.com/probability-sampling-with-python-8c977ad78664
5. Resumir y visualizar (a discreción) el dataset, leer "repaso_estadistica.pdf"
    a. https://docs.python.org/3/library/statistics.html

Entrega

- Lunes 31 de enero de 2021, 23.59 hrs.
- Compartir notebooks generados por correo: ilopez4a@gmail.com