TELEDUC
Educación a Distancia

Herramientas avanzadas de programación en *Python* para procesamiento de datos

Glosario

- **Archivo CSV:** Archivo de texto plano que replica una matriz de datos. Cada fila representa a medidas o valores de cada instancia. Cada columna representa distintos tipos de datos para cada variable específico. En general, los datos dentro de las columnas se separan por una "," o un ";". Además, se usa que la primera fila de este archivo lleve el nombre de las columnas.
- **Ciencia de datos:** "Es la disciplina que permite encontrar patrones predecibles en sets de datos estructurados, y no estructurados" (Dhar, 2013). También, "es un concepto que busca unificar estadísticas, análisis de datos, inteligencia artificial y cualquier otro método similar para poder entender y analizar fenómenos con los datos." (Hayashi, 1998). Un ejemplo de su uso es la astronomía, donde diariamente se genera un exabyte de datos (la cantidad de información que se genera en internet en un solo día).
- **Data *Frame*:** Los data *frames* son la estructura más usada en la librería *Pandas*. Se puede imaginar como una matriz de datos. Cada columna, representa datos para una variable específica. Y cada fila corresponde a medidas o valores de cada instancia. Cada columna será de un tipo específico, y cada fila podrá tener valores de distintos tipos. En esta estructura de datos, se agregan filas y columnas a preferencia. Las columnas de un data *frame* son series. Es importante destacar que cada fila tendrá un índice, que nos permitirá acceder a esa fila en particular y su información específica.

- **Encoding:** Es la codificación del archivo. *Python* y *Pandas* ocupan uno por *default*, que a veces no carga bien caracteres en español (como tildes, ¡ y ñ). Para evitarlo se emplea el *encoding* "latin-1" al cargar un data *frame*.

- **Estadísticos descriptivos:** Describen las estadísticas básicas de un conjunto de datos de forma cuantitativa. Específicamente, en *Pandas* se calcula:

  count: Cantidad de ocurrencias en una columna.
  mean: promedio simple entre los valores de una columna.
  std: Desviación estándar. Distancia promedio de todos los valores respecto al promedio.
  min: valor mínimo entre los valores de una columna.
  max: valor máximo entre los valores de una columna.

25%,50%,75%: valores al separar en cuartiles los datos.

- **Librería:** Otra forma de llamar a los paquetes, que son códigos encapsulados en Python que permiten ejecutar operaciones específicas por medio de sus funciones. Se necesita cargarlos mediante el comando import.

- ***Pandas:*** Es una librería de código abierto, que provee estructuras de datos de alto rendimiento y fáciles de usar. Además, provee herramientas para el análisis de datos en *Python*. Tiene dos estructuras de datos básicas: series y data *frames*.

- **Paquete random:** permite generar un número aleatorio. Esto se puede hacer de la siguiente manera:

```
import random

num_aleatorio = random.randint(a,b)
```

Donde [a,b] definen el rango en el que podemos crear el número aleatorio. En este caso el número aleatorio generado se almacena en la variable num_aleatorio.

- **Serie:** Es una lista de datos con un largo fijo, aquí no podemos agregar o quitar elementos, sólo modificarlos. Además, los datos de una serie son del mismo tipo.