PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE | la clase ejecutiva

Instrucciones Miniproyecto 1

Procesamiento de datos en Python

Python se ha posicionado como el lenguaje de programación más utilizado para el desarrollo de proyectos de Aprendizaje de Máquina. Lo anterior, entre otros factores, es debido a que en este lenguaje se han desarrollado un conjunto de librerías altamente especializadas para las distintas etapas o tareas dentro de un proyecto en esta área.

Una de las etapas fundamentales de todo proyecto de Ciencia de Datos o Aprendizaje de Máquina es el preprocesamiento de las bases de datos. Los datos en la práctica vienen típicamente sucios, por lo tanto, es necesario trabajar (transformar, limpiar, visualizar, etc.) con ellos antes de aplicar los modelos de Aprendizaje de Máquinas.

A través de ejemplos simples, en esta actividad usted podrá:

- Explorar algunas de las funcionalidades de un conjunto de librerías muy utilizadas por los científicos de datos del ecosistema Python.
- Preprocesar y visualizar datos utilizando Python.
- Utilizar Jupyter Notebook para generar un reporte autocontenido de todos los puntos solicitados.

Trabajo a realizar

Esta sección del miniproyecto está dividida en 4 partes, las que se describen a continuación:

1. Lectura y análisis exploratorio de datos

a. Abrir Google Colab o Jupyter Notebook.
b. Importe (e instale en caso de ser necesario) librería pandas.
c. Cargar la base de datos de nombre *ejemplo_data.csv*. En esta parte recomendamos explorar las diferentes opciones de *read* que tiene disponible la librería Pandas, identificando los argumentos disponibles en cada una de ellas.
d. Identifique los tipos de varibles que hay disponibles en la base de datos (*df.types* o *df.info()*).
e. Utilizando la función *astype*, transforme el atributo ID a entero y el atributo activo a binario. Vuelva a consultar el estado de las variables.
f. Convierta el atributo unidades a entero y 2016 a flotante.

2. Estadísticas descriptivas

a. Cree un diccionario con 20 datos que contenga al menos dos atributos continuos y una variable categórica (por ejemplo: nombre, nota, edad).
b. Transforme dicho diccionario a un *data frame* de pandas.
c. Obtenga estadísticas descriptivas de tendencia central.
d. Obtenga estadísticas descriptivas de dispersión.

3. Transformación e imputación de datos

a. Importe (e instale en caso de ser necesario) librerías pandas y *sklearn*.
b. Cargar la base de datos de nombre *ejemplo_data2.csv*.
c. Para las variables numéricas, genere un diagnóstico de números perdidos. Luego, impute los valores de acuerdo a la media y de acuerdo a otro criterio seleccionado por usted. Explore las opciones de imputación del método *fillna()* de pandas.
d. Transforme las variables categóricas a numéricas, generando una variable *dummy* por cada categoría.

4. Visualización de datos

a. Importe librerías searborn, matplotlib y numpy.
b. En primer lugar, crearemos nuestra propia base de datos. Para ello, dentro de numpy utilizaremos *numpy.random.multivariate_normal*[1].
c. A modo de experimentación, le recomendamos explorar las opciones del método y los ejemplos disponibles en la documentación de ella.
d. Genere una base de datos que distribuya normal bivariada.
e. Para la base de datos creada por usted, genere un gráfico de dispersión utilizando seaborn. Explore los argumentos de los métodos utilizados.
f. Genere un *boxplot* para cada una de las dimensiones de la base de datos creada. Explore los argumentos de los métodos utilizados.

[1] https://docs.scipy.org/doc/numpy-1.15.1/reference/generated/numpy.random.multivariate_normal.html#numpy-random-multivariate-normal