# **Instalación y primero pasos para usar be-GReaT**

1. **Instalación**

Para realizar la instalación, usaremos una distribución de GNU/Linux, en este caso Linux Mint.

Primero, deberemos poseer la versión 3.9 de Python. En el caso de no poseerla, realizar las siguientes instrucciones en la consola:

1. sudo apt install software-properties-common
2. sudo add-apt-repository ppa:deadsnakes/ppa
3. sudo apt-get update
4. sudo apt install python3.9
5. (opcional) python3.9 –version (para comprobar que se instaló correctamente)
6. pip install be\_great

Con esto habríamos instalado correctamente be-GReaT para nuestra distribución.

Segundo, vamos a hacer uso de algunos datasets (conjunto de datos tabulados) que nos proporciona la librería SKlearn. Para instalarla debemos realizar las siguientes instrucciones en la consola:

1. sudo apt-get update
2. pip3 install -U scikit-learn

Con todo esto configurado, solo haría falta crear un directorio y los archivos de ejemplo, para aprender la sintaxis y poder observar los resultados proporcionados por be-GReaT.

1. **Primer ejemplo: Diabetes**

Para este primer ejemplo crearemos un archivo vacío de extensión .py (EjemploDiabetes.py).

El objetivo es el siguiente: A partir de los datos reales brindados por el dataset de SKlearn, crear unos datos sintéticos usando be-GReaT que nos permitan anonimizar los anteriores y realizar una serie de comprobaciones que nos permitan ver si estos son prácticamente iguales para el estudio.

Aquí dejo una imagen con el código fuente:



Figura 1: Código fuente

1. **Ejemplos de regresión proporcionados por be-GReaT**

El siguiente ejemplo está disponible en el GitHub que contiene la librería be-GReaT. En este, se nos sitúa en un escenario donde poseemos un dataset con información acerca de casas de california. A continuación, se muestra dos muestras (Una con los datos originales y otra con los datos sintéticos):

Tabla

Descripción generada automáticamente con confianza media

Figura 2: Muestra de datos originales

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 3: Muestra de datos sintéticos

Al tratar de realizar un estudio con los datos, por ejemplo, realizar un estudio sobre la regresión (entre la longitud y latitud) de cada dataset, podemos observar como la regresión es prácticamente idéntica en ambos casos; mostrándose así, la gran funcionalidad que nos proporciona la generación de datos sintéticos. La capacidad de crear datos nuevos que no pertenecen a nadie (por lo que no tienen la necesidad de ser protegidos), pero a su vez, sirven de la misma manera que los originales para estudiarlos:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura 4: Regresión de los datos originales

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura 5: Regresión de los datos sintéticos

1. **Conclusión:**

La generación de datos sintéticos es un gran paso en una buena dirección hacia la anonimización de los datos. Sin embargo, la calidad de estos depende fundamentalmente de que “tan entrenado” está el modelo que estemos usando por lo que es importante entrenar al modelo constantemente con distintos datasets, que le permitan ampliar su conocimiento para así exprimir su potencial al máximo.

Para consultar mas información acerca del proyecto, véase:

* <https://github.com/kathrinse/be_great/>
* <https://kathrinse.github.io/be_great/>
* <https://arxiv.org/abs/2210.06280>