Laboratorio 3: Classification

# Información del dataset

El dataset consiste en un conjunto de datos publicitarios falsos, que indica si un usuario de Internet en particular hizo clic en un anuncio en el sitio web de una empresa. Este conjunto de datos contiene las siguientes características:

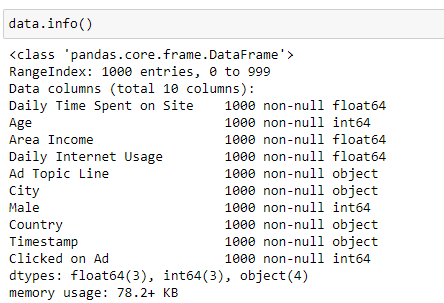
* **'Daily Time Spent on Site'**: tiempo del consumidor en el sitio en minutos
* **'Age'**: edad cliente en años
* **'Area Income'**: promedio renta del área geográfica del consumidor
* **'Daily Internet Usage'**: promedio minutos al día que el consumidor está en Internet
* **'Ad Topic Line'**: título del anuncio
* **'City'**: ciudad del consumidor
* **'Male'**: si el consumidor era hombre o no
* **'Country'**: país del consumidor
* **'Timestamp'**: hora a la que el consumidor hizo clic en el anuncio o en la ventana cerrada
* **'Clicked on Ad'**: 0 o 1 indicado si el cliente realiza clic en el anuncio

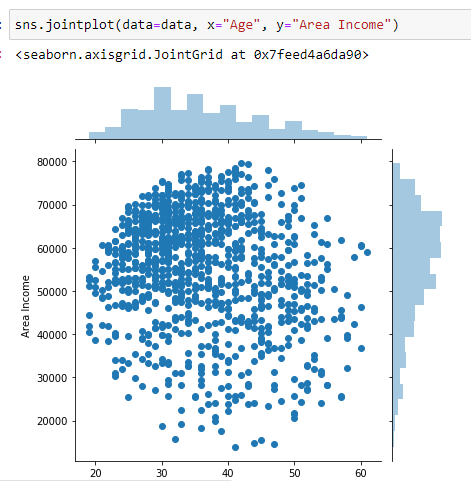
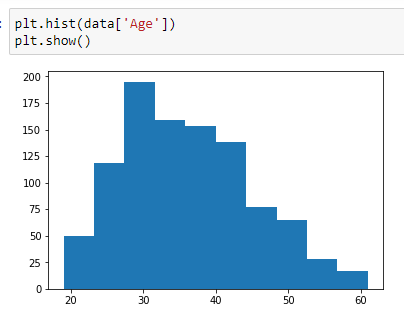
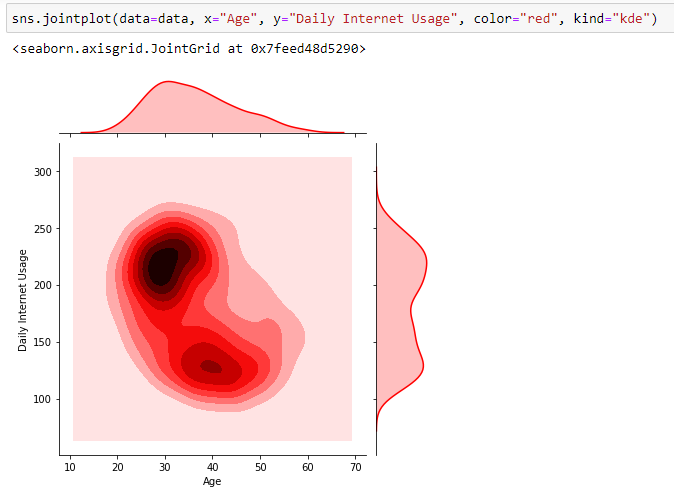
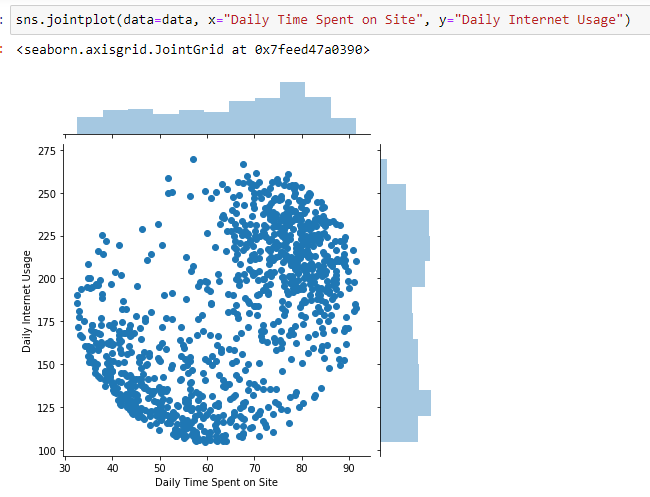
# Hipótesis u objetivo

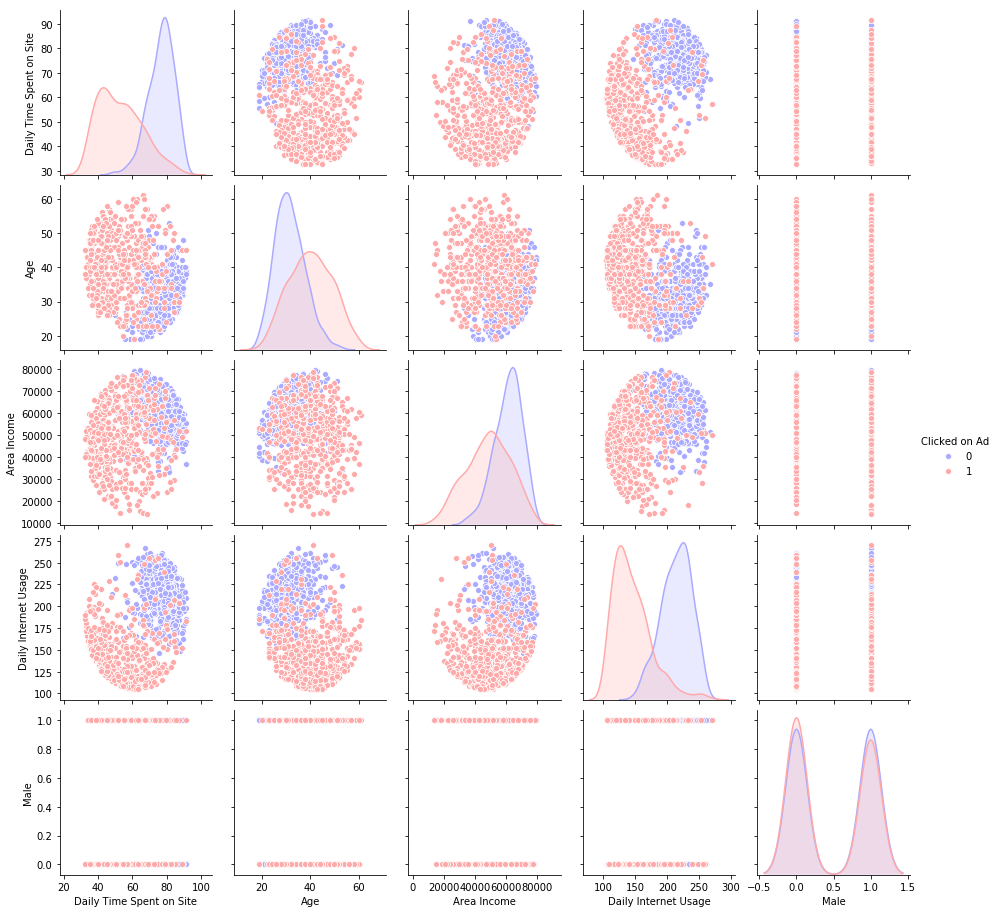
El objetivo de este laboratorio es construir un modelo que sea capaz de predecir si un usuario hará clic o no en un anuncio dependiendo las características de cada usuario.

# Solución y exploración

La solución se divide en dos partes: la exploración de la data, y el modelo para las predicciones.

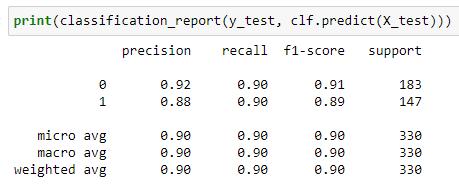
 En la parte de exploración, se corrieron algunos comandos para identificar columnas de la tabla y sus tipos de datos correspondientes:

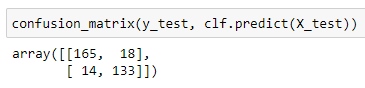
Posteriormente, se corrigieron inconsistencias en los tipos de datos y se continuó la exploración comenzando con la distribución de edades, ingresos mensuales y tiempo promedio en línea :

Finalmente, se separaron las poblaciones basándose en si el usuario hizo clic en el anuncio o no:

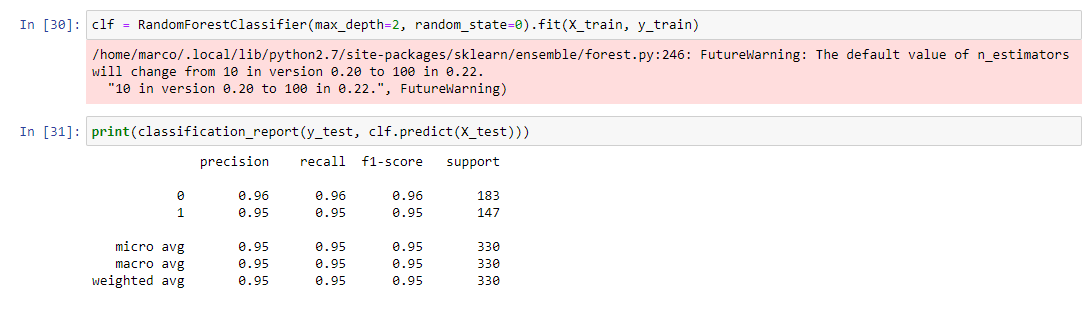
Finalmente, en la parte del modelo, se usó el paquete sklearn de Python para entrenar un modelo de regresión logística basándose en el dataset previamente descrito.

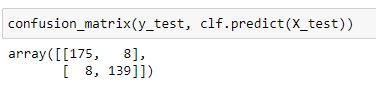
# Resultados: puntuales y respaldados por gráficos, tablas y visualizaciones en general

 El modelo de clasificación entrenado obtuvo una precisión mayor al 85% en sus predicciones tanto positivas como negativas:

 Con la siguiente confussion matrix:

Sin embargo, si en lugar de un modelo de regresión logística, se usa un Random Forest, los resultados mejoran a 95% de precisión:



 Con la siguiente confussion matrix: