


# Exploración de los DATOS



- ❖ PANDAS para levantar y extraer información relevante de nuestro *Dataset*.
- ❖ A partir de la exploración de los datos se manipulan las variables y se hace una limpieza inicial.
- ❖ Se obtiene una primera versión del Dataset procesado y se define la variable TARGET para nuestro modelo de *clasificación*.

# Visualización de las variables



- ❖ Seaborn para graficar distribuciones y relación entre variables de nuestro conjunto de datos.
  - ❖ Se intenta tener una aproximación sobre la influencia de los datos con respecto a nuestra variable TARGET.
  - ❖ Se obtiene información sobre la **correlación** entre las variables de nuestro Dataset.
- 

# Data cleaning and preprocessing

- ❖ Se trabaja sobre el Dataset para depurarlo y poder utilizarlo como entrada de un modelo de Machine Learning.
- ❖ Se tienen variables numéricas y variables categóricas, a las cuales hay que “encodear” para que puedan ser utilizadas.
- ❖ Hay datos faltantes y otros que con anotaciones ambiguas que se deben descartar de nuestro Dataset.
- ❖ Se considera que hay datos que no estarían disponibles o serían inexistentes a la hora de predecir si un cliente va a tomar un préstamo personal por primera vez, entonces se descartan estas variables.

# ENTRENAMIENTO del modelo

- ❖ Con el Dataset obtenido del paso anterior se entrenan un modelo de Regresión Logística y un Random Forest.
- ❖ Se obtiene una alta precisión para ambos y se visualizan distintas métricas alcanzadas.
- ❖ Se elige una métrica a maximizar enfocándonos en nuestro problema.

# Búsqueda de variables significativas y conclusiones

- ❖ A partir del análisis inicial de la correlación entre variables, se proponen subconjuntos reducidos del Dataset para explorar la performance de nuestros modelos a partir de dichos Datasets.
- ❖ Se mide la métrica elegida anteriormente para cada subconjunto de variables.
- ❖ Se obtienen 5 variables suficientemente significativas para alcanzar un excelente resultado en relación a la métrica elegida para evaluar nuestra clasificación.