# Módulo Análisis Exploratorio de Datos y Estadística

# Caso Practico Final: Santander

1. Para esta práctica se va a hacer uso de un conjunto de datos de una antigua competición sobre recomendación de productos del Santander en Kaggle, si bien ha sido reducido y modificado para la presente práctica. Respecto a este conjunto de datos, la información que se nos da en la competición es:

En esta competición se dispone de información de año y medio de datos sobre el comportamiento de clientes en la compra de nuevos productos. Los datos empiezan el día 28/01/2015 y tiene información mensual sobre los productos que tiene un cliente como "tarjeta de crédito", "cuenta de ahorros", ... El objetivo de esta competición era predecir qué productos adquirirá en el mes de la fecha 28/06/2016. Los productos están en las columnas llamadas "ind\_(xyz)\_ult1" que son las columnas que están desde la posición 25 a la 48 en el conjunto de datos de entrenamiento.

En los datos de la competición, **data\_santander\_red,** se puede comprobar que existe información de distintos tipos. Exactamente:

* Instante temporal: identifica el momento en el que se guarda el registro
* Identificador de usuario: nos hacer un seguimiento de la persona mes a mes
* Datos personales: información que identifican a una persona (edad, sexo, lugar de residencia, ...)
* Relación persona/empresa: información que identifica la relación de la persona con el banco (antiguedad, indice de actividad, ...)
* Productos de la persona: productos que disponía la persona el día del registro

La relación completa de variables está en el anexo.

## **Ejercicio 1** -Lectura y pasos previos-: Con los datos de data\_santander\_red (variable “renta”):

1. Leerlo como un dataframe (con la librería pandas) y llamarlo df\_santander
2. Observar los tipos de datos y cambiar “ind\_viv\_fin\_ult1” a tipo float64 (astype(float)), ya que pretendemos usarlo en fases posteriores y se requiere que sea numérico.
3. Tratar de responder a las siguientes preguntas de entendimiento inicial de los datos:
   * ¿Disponemos de un conjunto de datos suficientemente grande para nuestro objetivo?
   * ¿Nos faltan variables que podrían ser interesantes?
   * ¿Existe alguna variable que no sepamos lo que significa?
   * ¿Alguna de las variables que tenemos tiene un tipo raro (ej: es de tipo String en vez de numérico)?
4. Revisar el número de provincias que existen en los datos y compararlo con las provincias de España, ¿falta alguna? (df\_santander['nomprov'].nunique())
5. Ordena las provincias alfabéticamente y luego calcula los valores únicos (unique())
6. Revisa si los datos de los clientes que tenemos están homogéneamente distribuidos en España o se centran más en algunas provincias que en otras.
7. Identificar el formato del dataset.
8. Dividir la muestra de train (train\_santander\_red) y test (test\_santander\_red) en una proporción 70%-30%.

Uno de los puntos más importantes en cualquier trabajo de ciencia de datos es entender bien el problema. Para esta labor se suele hacer uso de expertos en la materia (trabajadores en la empresa, usuarios del programa, ...) y de la preguntas o hipótesis que tiene que ir haciéndose el científico de datos. Estas dudas pueden ser funcionales (ej: ¿Cuando un cliente pide un préstamo por primera vez necesita detallar su patrimonio?) o más situacionales (ej: ¿La empresa está más presente en algunos territorios que otros?, ¿La edad influye en los productos que un cliente compra?). La utilidad de estas hipótesis radica en explicar la realidad y desde discernir qué variables pueden ser interesantes hasta la optimización debido a ineficiencias que se producen.

## **Ejercicio 2** -Limpieza y Analisis descriptivo- sobre el fichero de data\_santander\_red:

1. Datos missing:
   * Observar si hay tipos de datos missing y graficarlos
   * Calcular el % que supone para cada variable y proponer varias soluciones para la variable “ind\_nomina\_ult1”.
   * Reemplazar para “age” y “antigüedad” el valor “NA” por np.NaN.
   * Si hay alguna variable que esté completamente vacía eliminarla.
   * La variable “antigüedad” mayor de 160 años debe ser un error por lo que se debe transformar ese caso a dato no disponible (np.NaN)
2. Calcular la media, mediana, desviación estándar, rango, rango intercuartílico, coeficiente de asimetría y la curtosis para la “edad” y “renta” y explicar los resultados. En particular, explicar lo que observe sobre los niños de 4 años. Aplicar las operaciones a nivel de grupo y columna de la media para esas dos variables y saque una conclusión sobre el grupo de las mujeres.
3. Revise si existe algún outlier. Para ello, calcular los percentiles 0.05 (Q1) y 0.95(Q3) utilizando la función quantile() de Pandas. Luego, calcular el (pseudo)rango intercuartílico (IQR) como la diferencia entre Q3 y Q1. Luego definir los límites inferior y superior para identificar outliers como 1.5 veces el IQR por debajo de Q1 y por encima de Q3. Finalmente, encontrar los outliers al seleccionar los valores que están por debajo del límite inferior o por encima del límite superior.

A continuación, se pide realizar una Visualización básica de datos con Python.

## **Ejercicio 3** -Visualización y Otros- sobre el fichero de data\_santander\_red:

1. Mediando análisis de correlaciones, identificar si la edad es un factor que permite identificar si una persona va a comprar un producto o no.
2. A continuación, hacer una matriz de correlaciones para entender cómo se correlaciona cada producto con el resto.
3. Realizar un análisis cluster sobre las variables de “age” y “renta”. Verifique que para “age” primero debe de convertir la variable a float. Dado que hay missing impute el valor medio para cada variable. Después graficarlo y obtener conclusiones.
4. Cree la variable “numero\_productos” como la suma de todos los productos qu consume la persona:

cols = ["ind\_tjcr\_fin\_ult1","ind\_reca\_fin\_ult1","ind\_pres\_fin\_ult1","ind\_plan\_fin\_ult1","ind\_hip\_fin\_ult1","ind\_fond\_fin\_ult1","ind\_ecue\_fin\_ult1","ind\_dela\_fin\_ult1","ind\_deme\_fin\_ult1","ind\_deco\_fin\_ult1","ind\_ctpp\_fin\_ult1","ind\_ctop\_fin\_ult1","ind\_ctma\_fin\_ult1","ind\_ctju\_fin\_ult1","ind\_cno\_fin\_ult1","ind\_cder\_fin\_ult1","ind\_cco\_fin\_ult1"] df\_santander['numero\_productos']= df\_santander[cols].sum(axis=1)

A continuación, usando el paquete seaborn y las variables “renta” y “numero\_productos” realizar un gráfico de puntos con regresión lineal. ¿Qué se puede apreciar?

1. Usando seaborn y las variables “renta” y “antiguedad” realizar un gráfico de dispersión apoyándose en alguna variable categórica (por ejemplo hue=”nomprov”). Después haga lo mismo tomando logaritmos para la renta e interprete los resultados.
2. Calcular el intervalo de confianza para el promedio de renta al 95%. Decida si usa una Normal o una T-Student.
3. Usando gráficas, ¿qué tipo de distribución siguen las dos variables “renta” y “numero\_productos”? ¿Considera que es necesario una transformación de las variables?
4. ¿Qué variables piensa que serán más relevantes para los análisis predictivos posteriores?
5. Con los análisis realizados en estos ejercicios y con cualquier otro análisis exploratorio de los datos a través de los gráficos (histograma, correlaciones, covarianzas, gráfico de pastel, boxplot, scatterplot,… ) y variables que considere relevantes, extraiga sus conclusiones sobre esta fase de análisis exploratorio, considerando el objetivo de la competición de predecir qué productos adquirirá después del set de datos.

## Anexo: Relación de variables

| **Nombre de columna** | **Descripción** |
| --- | --- |
| fecha\_dato | The table is partitioned for this column |
| ncodpers | Customer code |
| ind\_empleado | Employee index: A active, B ex employed, F filial, N not employee, P pasive |
| pais\_residencia | Customer's Country residence |
| sexo | Customer's sex |
| age | Age |
| fecha\_alta | The date in which the customer became as the first holder of a contract in the bank |
| ind\_nuevo | New customer Index. 1 if the customer registered in the last 6 months. |
| antiguedad | Customer seniority (in months) |
| indrel | 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month) |
| ult\_fec\_cli\_1t | Last date as primary customer (if he isn't at the end of the month) |
| indrel\_1mes | Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner ),P (Potential),3 (former primary), 4(former co-owner) |
| tiprel\_1mes | Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential) |
| indresi | Residence index (S (Yes) or N (No) if the residence country is the same than the bank country) |
| indext | Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country) |
| conyuemp | Spouse index. 1 if the customer is spouse of an employee |
| canal\_entrada | channel used by the customer to join |
| indfall | Deceased index. N/S |
| tipodom | Addres type. 1, primary address |
| cod\_prov | Province code (customer's address) |
| nomprov | Province name |
| ind\_actividad\_cliente | Activity index (1, active customer; 0, inactive customer) |
| renta | Gross income of the household |
| segmento | segmentation: 01 - VIP, 02 - Individuals 03 - college graduated |
| ind\_ahor\_fin\_ult1 | Saving Account |
| ind\_aval\_fin\_ult1 | Guarantees |
| ind\_cco\_fin\_ult1 | Current Accounts |
| ind\_cder\_fin\_ult1 | Derivada Account |
| ind\_cno\_fin\_ult1 | Payroll Account |
| ind\_ctju\_fin\_ult1 | Junior Account |
| ind\_ctma\_fin\_ult1 | Más particular Account |
| ind\_ctop\_fin\_ult1 | particular Account |
| ind\_ctpp\_fin\_ult1 | particular Plus Account |
| ind\_deco\_fin\_ult1 | Short-term deposits |
| ind\_deme\_fin\_ult1 | Medium-term deposits |
| ind\_dela\_fin\_ult1 | Long-term deposits |
| ind\_ecue\_fin\_ult1 | e-account |
| ind\_fond\_fin\_ult1 | Funds |
| ind\_hip\_fin\_ult1 | Mortgage |
| ind\_plan\_fin\_ult1c | Pensions |
| ind\_pres\_fin\_ult1 | Loans |
| ind\_reca\_fin\_ult1 | Taxes |
| ind\_tjcr\_fin\_ult1 | Credit Card |
| ind\_valo\_fin\_ult1 | Securities |
| ind\_viv\_fin\_ult1 | Home Account |
| ind\_nomina\_ult1 | Payroll |
| ind\_nom\_pens\_ult1 | Pensions |
| ind\_recibo\_ult1 | Direct Debit |