# Conceptos y Aplicaciones en Big Data

## 2do semestre 2021

## Práctica 7 – Spark SQL - MLlib

1) Rehaga los ejercicios 4, 5, 6 y 7 de la práctica 5, utilizando la API de los DataFrame de Spark SQL.

2) El banco desea saber para cada nacionalidad:
   - El top 3 de cantidad de clientes
   - El top 3 de cantidad de cajas de ahorro en total
   - El top 3 de cantidad de dinero prestado en préstamos
   - El top 3 de clientes más deudores (acumulado de las cajas de ahorro con saldo negativo)

   Resuélvalo utilizando las funciones window de Spark SQL.

3) El banco desea saber para cada clase (año de nacimiento de sus clientes)
   - El top 3 de cantidad de clientes
   - El top 3 de cantidad de cajas de ahorro en total
   - El top 3 de cantidad de dinero prestado en préstamos
   - El top 3 de clientes más deudores (acumulado de las cajas de ahorro con saldo negativo)

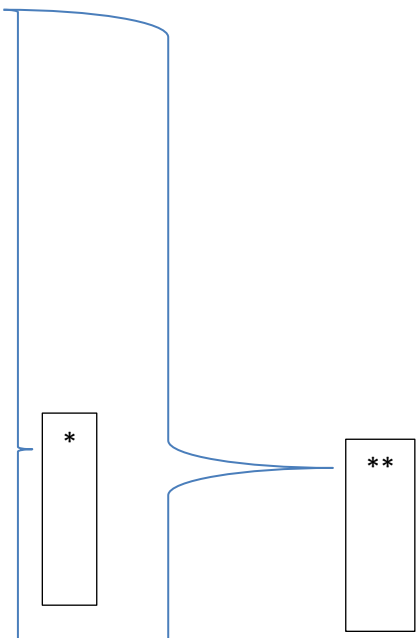
   Resuélvalo utilizando las funciones window de Spark SQL.

4) El banco desea conocer, para cada nacionalidad, el mínimo, máximo, promedio y mediana de cada grupo de clientes: “ricos”, “medios” y “pobres” (use ntile de 3 para subcategorizar a los clientes de cada nacionalidad).

5) Plantee un algoritmo iterativo que permita conocer, para cada nacionalidad, el rango de edad tal que:
   a) El rango de edad debería ser el más chico posible
   b) El rango de edad debería ser como mínimo 2
   c) La cantidad de clientes del rango debe ser mayor o igual a 10
   d) La suma del saldo de todas las cajas de ahorro de cada cliente dentro del rango debe ser mayor a 1.000.000
   e) Todos los criterios mencionados deben cumplirse en al menos el 75% de los clientes.

Ejemplo:

| Cliente | Nacionalidad | Saldo | Edad |
|---|---|---|---|
| 2 | ARG | 36 | 100 |
| 9 | ARG | 52 | 100 |
| 2 | ARG | 57 | 100 |
| 7 | ARG | 35 | 100 |
| 2 | ARG | 37 | 100 |
| 10 | ARG | 51 | 99 |
| 4 | ARG | 54 | 99 |
| 4 | ARG | 36 | 98 |
| 6 | ARG | 43 | 98 |
| 6 | ARG | 55 | 98 |
| 3 | ARG | 39 | 97 |
| 8 | ARG | 54 | 97 |
| 10 | ARG | 41 | 97 |
| 8 | ARG | 35 | 97 |
| 9 | ARG | 37 | 97 |
| 4 | ARG | 59 | 97 |
| 8 | ARG | 36 | 97 |
| 7 | ARG | 58 | 96 |
| 7 | ARG | 54 | 96 |
| 2 | ARG | 44 | 94 |
| 10 | ARG | 33 | 94 |
| 3 | ARG | 44 | 94 |
| 6 | ARG | 56 | 93 |
| 9 | ARG | 52 | 92 |
| 10 | ARG | 31 | 92 |
| 8 | ARG | 34 | 92 |
| 5 | ARG | 44 | 91 |
| 6 | ARG | 55 | 91 |
| 7 | ARG | 54 | 91 |




* Con un rango de edad de 6 (100-94) los saldos suman 990. No cumple

** Con un rango de edad de 7 (100-93) los saldos suman 1046. Si cumple

Si bien este ejemplo muestra solo la condición para el primer cliente de la tabla, esta misma condición debería cumplirse en al menos el 75% de los clientes.

6) Utilice el dataset del banco para hacer las siguientes predicciones:
   a) Cantidad de cajas de ahorro que tendrá un cliente según su año de nacimiento.
      - Considere tres categorías:
        - ❖ Clientes que solo tienen una caja de ahorro
        - ❖ Clientes que tienen 2 o 3 cajas de ahorro
        - ❖ Clientes que tienen 4 o más cajas de ahorro

b) Monto total entre todas las cajas de ahorro que tendrá un cliente según su mes de nacimiento y la cantidad de cajas de ahorro.
   - Considere dos categorías:
     - ❖ Saldo total negativo
     - ❖ Saldo total positivo

c) Cantidad de préstamos que sacará un cliente según su edad, el monto total entre todas sus cajas de ahorro y la cantidad total de cajas de ahorro.
   - Considere tres categorías:
     - ❖ Clientes que no han sacado préstamos
     - ❖ Clientes que tienen 1 o 2 préstamos
     - ❖ Clientes que tienen 3 o más préstamos

Para hacer cada una de las predicciones realice lo siguiente:

i. Prepare el dataset según lo solicita cada consigna (variables dependientes – variable objetivo)
ii. Separe el dataset en dos subconjuntos al azar (80% de los ejemplos para el que llamaremos *train* y el 20% restante para el que llamaremos *test*)
iii. Entrene un modelo DecisionTreeClassifier (u otro si desea investigar la librería) con el subconjunto *train* para obtener un modelo clasificador.
iv. Realice la predicción con el modelo conseguido usando el subconjunto *test*.
v. Realice una comparación entre la respuesta del modelo y el dato real del dataset *test*, para determinar la tasa de acierto en cada una de las categorías

7) Estimators y transformers

a) Implemente un estimador que realice la regresión lineal entre dos variables de un dataframe.

b) Implemente un transformador que dado un dateframe con la variable X devuelva otro dataframe con el agregado de la variable Y estimada.

Dadas las variables X y Y se debe calcular:

$$Y = \propto X + \beta$$

donde:

$$\propto = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

y:

$$\beta = \frac{\sum y - \alpha \sum x}{n}$$

Dado un dataframe con la variable *X* y los parámetros $\alpha$ y $\beta$ la variable *Y* se estima como:

$$Y = \propto X + \beta$$