# PREDICCIÓN DE SUSCRIPCIONES PARA LA CAMPAÑA DE MARKETING EN EL SECTOR BANCARIO

## Introducción y objetivos

En el presente informe se desarrollará la aplicación de conceptos de Machine Learning a un caso real de negocio. Utilizando un modelo de aprendizaje supervisado, se busca evaluar las características y el comportamiento de los clientes de un banco para predecir si se suscribirán a la próxima campaña de marketing. El objetivo de la modelización es que el banco pueda focalizar su campaña hacia los clientes más predispuestos a aceptar la suscripción, eficientizando así sus recursos.

## Descripción del dataset

El banco provee un dataset de 45.211 clientes y 17 variables que describen a cada uno de ellos. Se cuenta con una variable "Subscription" que indica si el cliente se suscribe o no a la campaña, por lo cual es el resultado a predecir por el modelo (1: rechaza, 2: acepta).

Las variables que se utilizarán para realizar la predicción son:

Variables numéricas:

- Age: edad del cliente.
- Balance (euros): promedio de saldo en la cuenta a lo largo de un año.
- Last Contact Day: último día de contacto con el cliente en el mes.
- Last Contact Duration: duración del último contacto con el cliente, medido en segundos.
- Campaign: cantidad de contactos al cliente durante la campaña, incluyendo el último.
- Pdays: cantidad de días que pasaron desde el último contacto con el cliente de una campaña anterior
- Previous: cantidad de contactos previos a esta campaña para cada cliente.

Variables categóricas:

- Job: tipo de empleo del cliente.
- Marital Status: estado civil del cliente.
- Education: educación máxima alcanzada por el cliente.
- Credit: indica si el cliente tiene deuda de crédito o no.
- Housing Loan: indica si tiene seguro de hogar o no.
- Personal Loan: indica si tiene préstamos tomados o no.
- Contact: tipo de contacto con el cliente.

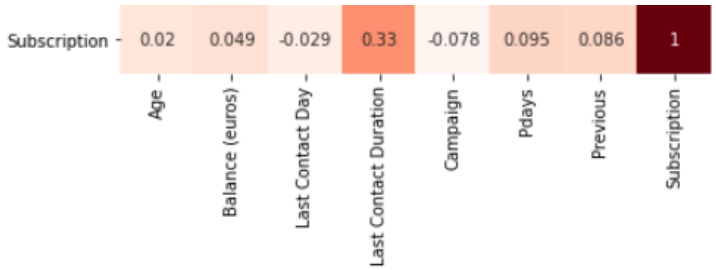
- Poutcome: resultado de la campaña anterior respecto al cliente.

Se detectó además que el 76% de las filas del dataset contienen al menos un valor nulo. Luego de realizar un preprocesamiento sobre los valores faltantes, se redujo el porcentaje a un 37% y se eliminaron estas filas.

## Análisis exploratorio de datos

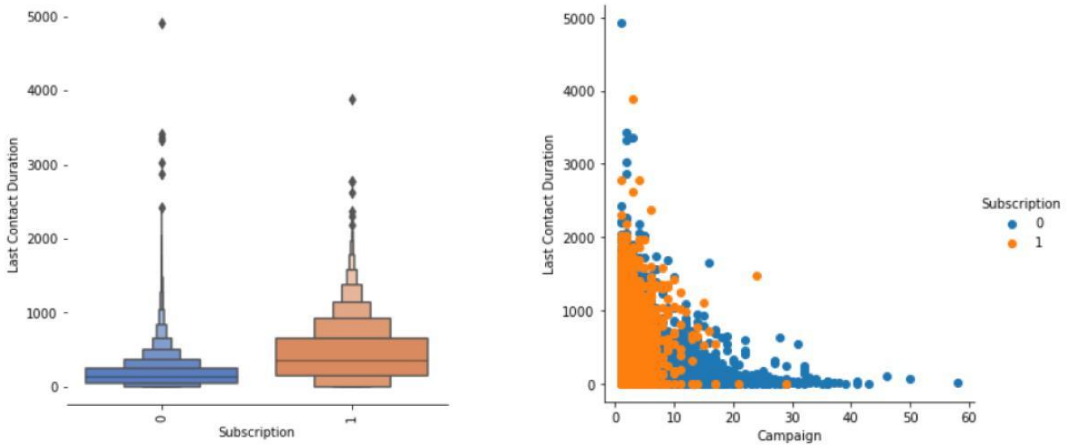
Con el fin de lograr un mayor entendimiento de los datos e identificar las relaciones entre las diferentes variables, se realizó un análisis exploratorio de datos (EDA).

En primer lugar se identificó la correlación lineal entre las variables numéricas, considerando principalmente a la variable "Subscription", ya que indicará cuáles son las variables que mejor predicen, en principio, el resultado de la campaña.




Se puede observar que la variable "Last Contact Duration" es la que mejor correlaciona. Es decir, luego de lograr un contacto prolongado con un cliente, es más probable que el resultado de la campaña sea satisfactorio. Se observa esta distribución en el siguiente boxplot.

Se presenta a su derecha un scatterplot para mostrar la relación entre la cantidad de contactos y su duración. En este gráfico se puede identificar la correlación negativa entre la cantidad de contactos y el resultado de la campaña. Además, conforme avanza la cantidad de contactos es más difícil lograr uno prolongado, por lo cual se concluye que no es conveniente insistir con clientes que no tuvieron respuesta satisfactoria en sus primeros contactos.




En cuanto a las variables categóricas, se analizó su relación con el resultado en forma nominal y relativa. Segmentando por tipo de trabajo, se encontró que los clientes que son estudiantes o retirados son los más predispuestos a aceptar la campaña, aunque la mayor cantidad de suscripciones provienen de personas que trabajan en management o como technician. Por otra parte, los clientes que ya habían tenido resultados exitosos en la campaña anterior son ampliamente más propensos a suscribirse nuevamente.

## Materiales y métodos (Algoritmos utilizados)

Para llevar a cabo la experimentación, se aplica aprendizaje supervisado con regresión logística a partir del dataset preprocesado. Este tipo de aprendizaje consiste en inferir propiedades y estructuras en la distribución de los datos con el fin de predecir un resultado, en el caso planteado la suscripción a una campaña de marketing.

Para poder evaluar las variables categóricas dentro del modelo, se generaron variables *dummies*, que consisten en una apertura de cada clase en diferentes columnas que tomarán valores booleanos. Además se realizó una estandarización de las variables numéricas con el fin de normalizar la magnitud de los valores de cada variable, haciendo que adopten una distribución normal y mejorando la performance del modelo.

En la etapa de entrenamiento del modelo, se utiliza el dataset etiquetado con resultados reales para cada cliente. Se buscará encontrar los parámetros $w$ que minimicen el error entre el resultado de la predicción ($\hat{y}$) respecto al resultado real de la campaña ($y$). Con este objetivo se plantea la regresión logística, que devuelve valores entre 0 y 1 según la probabilidad de que el cliente acepte la suscripción ($y = 1$) o la rechace ($y = 0$).

Se computa la regresión logística con las siguientes fórmulas:

$$z = w^T x \qquad (1)$$

$$g(z) = \frac{1}{1+e^{-z}} \qquad (2)$$

(1) Se realiza una regresión lineal mediante una matriz *w* que contiene la contribución de cada variable al resultado final. A mayor *w* para una variable, más fuerte su correlación con el resultado.
(2) Se aplica la función sigmoide a los resultados de la regresión lineal, para expresarlo como probabilidad de clase y poder realizar la clasificación binaria.

El entrenamiento del modelo se realiza minimizando la función costo de entropía cruzada, hasta obtener la matriz *w* más conveniente. Se incluyen los métodos de *Grid Search* y *Cross Validation* para encontrar la mejor opción de hiperparámetros entre diferentes alternativas planteadas. Por otra parte, fue necesario ajustar la importancia relativa de las clases para

Como índice de evaluación de los resultados se adoptará el *Accuracy*, que indica la proporción de muestras clasificadas correctamente sobre el total.

En un segundo experimento se aplica *Principal Component Analysis*. Se trata de un método de reducción de dimensionalidad, particularmente útil cuando los datos son de alta dimensionalidad y correlación. El objetivo es identificar una cantidad reducida de variables que representen a los datos originales en un subespacio de menor dimensión combinando linealmente a las variables originales. El objetivo de este método es reducir el ruido en los datos y disminuir el tiempo y memoria utilizados en el procesamiento.

# Experimentos y resultados

Como primer experimento, se entrenó el modelo para el total de las variables numéricas y las *dummies* generadas. De esta forma se obtuvo un Accuracy = 0.8887

En segundo lugar, se entrenó el modelo luego de realizar el PCA, con lo cual las variables numéricas se redujeron de 7 a 2 componentes. Se obtuvo un Accuracy = 0.8951

# Discusión y conclusiones

Como conclusión del análisis, respecto a la distribución de los datos se puede observar que la variable más relacionada a resultado de la campaña es la duración del último contacto con el cliente, por lo cual se recomienda enfocar los recursos hacia aquellos clientes que estuvieron dispuestos a mantener un contacto prolongado.

Respecto al modelo de Machine Learning, se observa que si bien el rendimiento es aceptable, se ve influido por el desbalance de clases con el que cuenta el dataset y la

ponderación relativa de sus pesos no logra remediarlo del todo. Se concluye además que la implementación del PCA no tuvo influencia significativa en el resultado, ya que la precisión del modelo aumentó muy levemente. Este comportamiento puede deberse a que las variables no están tan correlacionadas como se esperaba al momento de elegir la nueva dimensionalidad, o bien porque el dataset cuenta con muchas variables categóricas que no pueden ser reducidas.

# Referencias


* Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The Elements of Statistical Learning
* Stoltzfus, J. (2011). Logistic Regression: A Brief Primer
* Dunteman, G. (1989). Principal Components Analysis