# **Preparación de datos con Python**

Actualmente existen múltiples herramientas que permiten una exploración profunda sobre los datos iniciales, las cuales son de gran utilidad para el analista que decide comenzar con este proceso. Desde una planilla de cálculo como Microsoft Excel, hasta módulos especializados en herramientas productivas, existen diversos mecanismos para llevar a cabo la preparación de los datos, teniendo como objetivo generar los insumos para un modelo de Machine Learning.

En esta ocasión, se pretende investigar los distintos mecanismos que dispone Python y las librerías desarrolladas por la comunidad, para el análisis y tratamiento de datos.

Todo el código Python será escrito en un Jupyter Notebook, aprovechando las facilidades que este tipo de documentos ofrece para la ejecución de este. Además, utilizaré la herramienta Google Colab, la cual resuelve las referencias a las librerías de Python a utilizar automáticamente.

## Importar librerías y lectura del Dataset

El primer paso consiste en importar todas las librerías a utilizar durante el proyecto.

* NumPy
* Pandas
* Seaborn
* MatPlotLib

Luego se debe leer el contenido del archivo “Titanic-Dataset.csv”. Al ejecutar el código se puede visualizar una vista previa de los registros. Debería verse de esta manera:

Icono

Descripción generada automáticamente

## Visualizar Dataset

Por medio de las siguientes instrucciones es posible identificar las columnas con valores numéricos. El resto de los atributos son de tipo categórico.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

La generación de gráficos para visualizar el Dataset puede realizarse mediante la librería Seaborn. Para ejemplificar, es posible generar un gráfico que ilustre la tasa de sobrevivientes, partiendo de la clase y el sexo del individuo.

Gráfico, Gráfico de barras

Descripción generada automáticamente

## Valores faltantes e imputación

Este Dataset presenta valores faltantes únicamente en los atributos Age, Cabin y Embarked. Para verificar esto, se utiliza el siguiente código:

Imagen que contiene Tabla

Descripción generada automáticamente

Como se puede ver en este conteo de valores faltantes, los atributos más impactados son Age y Cabin. En el caso de edad, será necesario aplicar una imputación de valores, siendo este atributo importante en la predicción del modelo. Probablemente, los atributos Cabin y Embarked no impacten significativamente en el desempeño del modelo, por lo que se podría considerar eliminarlos.   
  
La eliminación de atributos se realiza con el siguiente código. Además se eliminará el número de ticket, el cual no aporta ninguna información para la predicción.



Gráfico

Descripción generada automáticamente con confianza baja

Para decidir el valor que será imputado al atributo edad, es conveniente analizar la distribución de este. Esto se puede realizar mediante la librería matplotlib.

Gráfico, Gráfico de líneas, Histograma

Descripción generada automáticamente

La distribución observada es bastante centrada pero algo inclinada hacia un lado, por lo que una buena decisión sería imputar la edad utilizando la mediana de los datos. Esto se realiza mediante el siguiente código:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Al concluir este paso, tenemos asegurado que los datos no contienen datos faltantes. Por lo que serán de mayor utilidad y precisión a la hora de entrenar el modelo.