**Recap segunda sesión: Limpieza de datos.**

**Eliminación de columnas, establecimiento del tipo de los datos, normalización de los datos.**

**En esta sesión de trabajo hemos seguido explorando el conjunto de datos eliminando las columnas irrelevantes. Comprobamos que lo s datos tenían el tipo que les correspondía y normalizamos los datos para que éstos siguieran el mismo formato.**

Lo primero que hicimos fue comprobar si teníamos alfun registro duplicado, eliminamos 14 registros.

A lo largo de la sesión, añadimos una hoja par apoder hacer la normalización de los datos con la función BuscarV en lugar de encadenar funciones SI.

**DESCRIPCION DE LAS COLUMNAS**

CustomerID: Identificador único de cada cliente.

#Utilizamos un formato condicional para resaltar si había algún identificador duplicado. Encontramos que sí seguía habiendo duplicados

Churn Label: Etiqueta que indica si el cliente se fue (Yes ) o no (No).

Churned: Variable binaria (0 o 1) que indica si el cietne abandonó el servicio (1) o se mantuvo (0).

#Comprobamos que ambas columnas tenían exactamente los mismos datos. Todos los ‘yes’ eran 1 y lo s’no’ eran 0, eliminamos la columna Churned ya que en este caso nos interesan mas los valores de texto (para las visualizaciones) que los numéricos

Account Length (in months): Tiempo que el cliente ha estado suscrito al servicio, en meses.

Customer Service Calls: Número de llamadas realizadas al servicio de atención al cliente.

Avg. Monthly GB Download: Promedio mensual de datos descargados por el cliente en GB

Unlimited Data Plan: Indica si el cliente tiene un plan de datos ilimitado (1: Sì , 0:No)

#Cambiamos el formato de los datos de 0/1 a ‘no’/’yes’

Extra Data Charges: Cargos adicionales por uso de datos más allá del límite del plan.

State: Estado de residencia del cliente.

Phone Number Número de teléfono del cliente.

#Eliminamos la columna ya que no nos aporta valor en el análisis

Gender: Género del cliente.

#Cambiamos los datos para que todas las categorías estén en ingles

Age: Edad del cliente.

Under 30: Indica si el cliente es menor de 30 años (1: Sì , 0:No).

#Comprobamos que los valores afirmativos de esta columna fueran coherentes con los valors de la columna Age.

Senior: Indica si el cliente es mayor de 65 años (1: Sì , 0:No).

#Realizamos la misma comprobación que en la columna anterior. Aquí vimos que había clientes con 65 años que tenían asignado ‘yes’ y otros que tenían asignado ‘No’, por lo tanto. Se quedo en contactar con el cliente para ver que deberíamos poner a estos clientes

COMPROBACION SENIOR:

#Columna creada para ver si los valores de la columna Senior coincidían con Age

Group: Indica si el cliente tiene contratado un plan familiar.

Number of Customers in Group: Número de clientes que pertenecen al mismo plan familiar.

Device Protection & Online Backup: Indica si el cliente tiene protección de dispositivo y respaldo en línea.

Contract Type: Tipo de contrato(Mensual, anual, etc.).

Payment Method: Método de pago utilizado por el cliente.

#Esta columna tenia registros en minúsculas y mayúsculas, se estandarizaron poniendolo en formato Nombre Propio

Monthly Charges: Cargo mensual que el cliente paga por el servicio.

#Esta columna tenia valores numéricos y otros con la unidad monetaria. Se elimino la unidad monetaria y se añadió al nombre de columna, se convirtieron los datos a numero.

Total Charges: Cargos totales acumulados durante la suscripción.

Churn Category: Categoría que indica el motivo de la deserción (churn).

Churn Reason: Razón específica por la que el cliente abandonó el servicio.

Contact Date: Fecha del último contacto con el cliente.

#Se formateo la fecha para que Excel interpretara los datos como datos de tipo fecha

Last Transaction Date: Fecha de la última transacción realizada por el cliente.

Tenure(in months): Tiempo total de suscripción del cliente.

#Esta columna se elimino, ya que aportaba información similar a otra pero con datos irrelevantes según el cliente

Customer Segment: Clasificación del cliente según su nivel de uso o importancia (Alto, Medio, Bajo).

#Se formatean los datos para que todos estén en ingles

Average Monthly Expenses: Gasto promedio mensual del cliente

#Se redujo el numero de decimales de la columna

Number of Complaints of support: Número de quejas o interacciones con el servicio de soporte

Preferec Contact Method: Método preferido por el cliente para ser contactado (email teléfono,etc.).

#Se formatean los datos para que todos estén en ingles

Applied Discount: Indica si se ha aplicado algún descuento al cliente.

#Se formatean los datos para que todos estén en ingles

Hobby: Pasatiempos del cliente.

Favorite TV Show: Programa de televisión favorito del cliente.

#Eliminamos esta columna ya que no nos aporta valor en el analisis

Internal Notes: Notas internas relacionadas con el cliente.

#Se formatean los datos para que todos estén en ingles

COLUMNAS RELEVANTES PARA NUESTRO ANÁLISIS:

-Churn Label: Ojo que esta es importante.

-Duración de la cuenta.

-Frecuencia de llamadas a atención al cliente.

-Número de quejas o soporte.

-Método de pago utilizado.

-Tipo de contrato.

-Segmento de cliente y su relación con el abandono.