**INFORME FINAL**

Como nuestra nueva tarea en el descubrimiento de la razón por la que los clientes dejan la empresa, se nos asignó la creación de un modelo de machine learning para poder determinar si un cliente nuevo va a quedarse con la empresa o no.

**PARTE I: Carga de los datos.**

La primera parte del trabajo consistió en cargar el archivo usado en la primera parte del trabajo, este archivo contenía la información necesaria para la creación de los modelos de machine learning. Con ayuda de pandas se cargo y analizo el archivo con la información pertinente.

El data frame cargado constaba de una serie de 24 columnas diferentes mostradas a continuación:

A screenshot of a computer

AI-generated content may be incorrect.

Podemos determinar que existen variables que no son necesarias en este análisis tales como la columna 0 y la columna 1 por lo tanto procedemos a eliminarlas.

Después de un rápido análisis de los valores del data frame podemos observar que existen variables del tipo booleano (si y no) las cuales podemos convertir en tipo numérico, para un análisis mas practico, sin embargo debemos tener cuidado ya que no podemos simplemente ponerlos como 1 o 0, ya que en automático el modelo que se quiere construir le dará más peso, siendo así un error muy grave, por esta razón se procederá a hacer un encoding de las columnas.

**PARTE II: One-hot Encoding**

Debido a que se quiere hacer el calculo de conservación de clientes, la variable dependiente va a ser la columna “Churn” es importante tomar en cuenta este dato.

Existen dos maneras de hacer el encoding, una es mediante el método dummies de pandas y otro es el one-hot encoding, el cual en este case será usado debido a que este último permite el poder hacer encoding en nuevos datos, pudiendo hacer más versátil el resultado final (modelo de machine learning).

Los datos al ser transformados podemos verlos a continuación:

A screenshot of a computer

AI-generated content may be incorrect.

**PARTE III: Balanceo de clases**

Una buena practica es el balancear las clases debido a que si tenemos una cantidad mayor en los datos (ya sea positivos o negativos) y se puee hacer de dos manera, para este trabajo se decidió hacer un “Oversampling” usando el método SMOTE, una ves realizado el oversamplin en la variable “Churn” podemos observar que las variables ahora son 50-50.

A computer screen shot of a computer code

AI-generated content may be incorrect.

Debido a que los datos originales presentaban valores nulos fue necesario hacer un chequeo en estos para poder eliminar los valore nulos del archivo, este fue un paso extra que no debió ser necesario si se hubiera validado que la información del principio estaba correcta.

**PARTE IV: Análisis de correlación**

Esta es un aparte fundamental e importante, debido de que a partir de este punto podemos determinar cuales son las variables que se relacionan con nuestra variable “Churn” y así poder crear un modelo que pueda predecir de manera correcta los valores futuros.

A screenshot of a computer

AI-generated content may be incorrect.

Al ejecutar el análisis de correlación podemos ver que los índices van de -1 a +1, siendo los que podemos determinar existe una correlación los que esta mas cerca del valor +1.

Al hacer un heat map usando solo la variable “Churn” podemos apreciar lo siguiente:

A graph with a number of data

AI-generated content may be incorrect.

**PARTE V : Selección de variables**

Después del análisis de correlacion se realizo un box plot usando las variables “Churn” y “Costumer ternure”, para poder darnos una idea del tiempo que dua un cliente con la empresa mostrando lo siguiente:

A diagram of a variety of colored squares

AI-generated content may be incorrect.

Con este box plot podemos darnos cuenta que la cantidad de clientes que se quedan con la empresa (un 75%) lo hacer por un termino de máximo 30 meses. Lo que nos da a entender porque los contratos de 2 años eran los que tiene mas éxito.

**PARTE VI: Division de datos y entrenamiento de los modelos**.

Una ves determinado que la variable “Churn” es la variable dependiente, se procede a hacer la división de los datos en un porcentaje de:

* 30% para el entrenamiento
* 70% para la validación

Usando la librería sklearn se procede a la división de valores.

A computer code with black text

AI-generated content may be incorrect.

Una vez realizada la división de los valores procedemos a la creación de los modelos de machine learning, siendo el primero:

* **Árbol de decisión:** Una vez entrenado el modelo se realiza el reporte de clasificación.

A screenshot of a computer program

AI-generated content may be incorrect.

Se nos solicita hacer un modelo de regresión lógica por lo tanto primero debemos estandarizar los datos, utilizando el siguiente código realizamos la estandarización:

A white background with black text

AI-generated content may be incorrect.

Una ves conseguido la estandarización entrenamos nuestro modelos y vemos el reporte de clasificación.

* **Regresion lógica:** Se entrena el modelo de regresion lógica dando como resultado.

A screenshot of a computer program

AI-generated content may be incorrect.

**CONCLUSION:**

Como podemos observar los modelos nos dan el mismo resultado, por lo que podemos concluir que si usamos cualquiera de estos modelos la predicción será la misma para ambos resultados, podemos considerar lo siguiente para mejorar los modelos:

* Reducir algunas variables del data frame para evitar que ambos modelos nos den el mismo resultado
* Quiza un análisis de overfittting sería necesario para determinar porque los valores son los mismos
* Revisar porque el archivo original tenia valores nulos desde el principio