# INFORME FINAL - INTRODUCCIÓN A LA CIENCIA DE DATOS


## Predicción de Abandono de Clientes en Telco

Universidad Católica Argentina
Facultad de Química e Ingeniería
Licenciatura en Ciencia de Datos

**Estudiantes:**
**- Facundo Arimany**
**- Dante Scarpin**
**- Emiliano Palacios**
**- Matías Rivero**

**Docente: Pavón Claudio Gonzalo**
**Fecha de entrega: 27 de junio de 2025**

[Trabajo Final LCD - Github]

## 1. Introducción
En este proyecto aplicamos técnicas de Machine Learning supervisado para predecir si un cliente de una compañía de telecomunicaciones abandonará el servicio (churn), utilizando un dataset proporcionado por el docente basado en Telco Customer Churn. El flujo de trabajo se gestionó desde Google Colab y fue versionado en GitHub.

## 2. Exploración y Visualización Inicial
Se realizó una exploración del dataset, analizando la distribución de variables, tipos de datos y valores faltantes. Se incluyeron gráficos de caja y bigotes, así como diagramas de violín, para visualizar outliers y la distribución de datos en función de la variable objetivo.

## 3. Limpieza y Preprocesamiento
Se trataron valores nulos, se codificaron variables categóricas y se escalaron variables numéricas. Se empleó la técnica de escalado MinMaxScaler para normalizar las variables y se eliminaron columnas irrelevantes como 'customerID'.

## 4. Selección de Características
Se construyó un mapa de calor para visualizar correlaciones entre variables y detectar redundancias. Las variables más relevantes se seleccionaron para entrenar los modelos.

## 5. Modelos Entrenados
Se entrenaron los 3 modelos: Regresión Logística, Árbol de Decisión y Random Forest. Estos modelos fueron evaluados utilizando métricas de precisión, recall, accuracy y matriz de confusión, la clave en nuestro entrenamiento fue maximizar el Recall,  ya que es quien analiza nuestros falsos negativos, es decir, los clientes que el modelo predijo que no harían Churn (no se irían) pero realmente si hicieron Churn (si se fueron).

## 6. Resultados y comparación de modelos

```
=== Comparación de Modelos ===
```

| | Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| 0 | Regresión Logística | 0.74652 | 0.514035 | 0.809392 | 0.628755 |
| 1 | Árbol de Decisión | 0.76044 | 0.545220 | 0.582873 | 0.563418 |
| 2 | Random Forest | 0.80000 | 0.616188 | 0.651934 | 0.633557 |

Luego de comparar los datos pudimos ver que la Regresión Logística fue quien tiene un mayor porcentaje de Recall, lo cual es clave para nuestro objetivo y por ende, es el modelo por excelencia que debemos utilizar a pesar de sacrificar precisión. El Árbol de Decisión es el modelo que menos sirve para nuestro objetivo , si bien es un poco más equilibrado, su bajo Recall lo convierte en una opción poco viable. Y por último nos encontramos con los resultados del Random Forest, que resulta ser una opción interesante, siendo la opción más equilibrada y con mayores porcentajes de accuracy.

La comparación entre modelos indicó que la Regresión Logística será nuestra primera opción como mencionamos antes, debido a su alto Recall, mientras que RandomForest sería nuestra segunda opción, si bien es bastante inferior en cuanto a Recall respecto al primer modelo, que es lo que nos interesa, ofrece un mayor equilibrio en los otros datos, funcional en otros tipos de objetivos. Por último tenemos al Árbol de Decisión, que será descartado debido a que no tiene ningún beneficio respecto de los otros modelos.

## 7. Visualización con los modelos obtenidos

Finalmente realizamos varios gráficos con el objetivo de analizar, comparar e implementar los nuevos datos obtenidos por cada modelo, o por el modelo ejemplar (Regresión Logística)

## 8. Versionamiento y Herramientas

El desarrollo fue llevado a cabo en Google Colab y versionado en GitHub. Se realizaron commits descriptivos en cada etapa del análisis. Se utilizaron las librerías pandas, scikit-learn, seaborn y matplotlib.

## 9. Referencias


• Dataset Telco Customer Churn – proporcionado por el docente
• Scikit-learn documentation
• Kaggle y Stack Overflow para consultas puntuales

### Informacion del Data Set

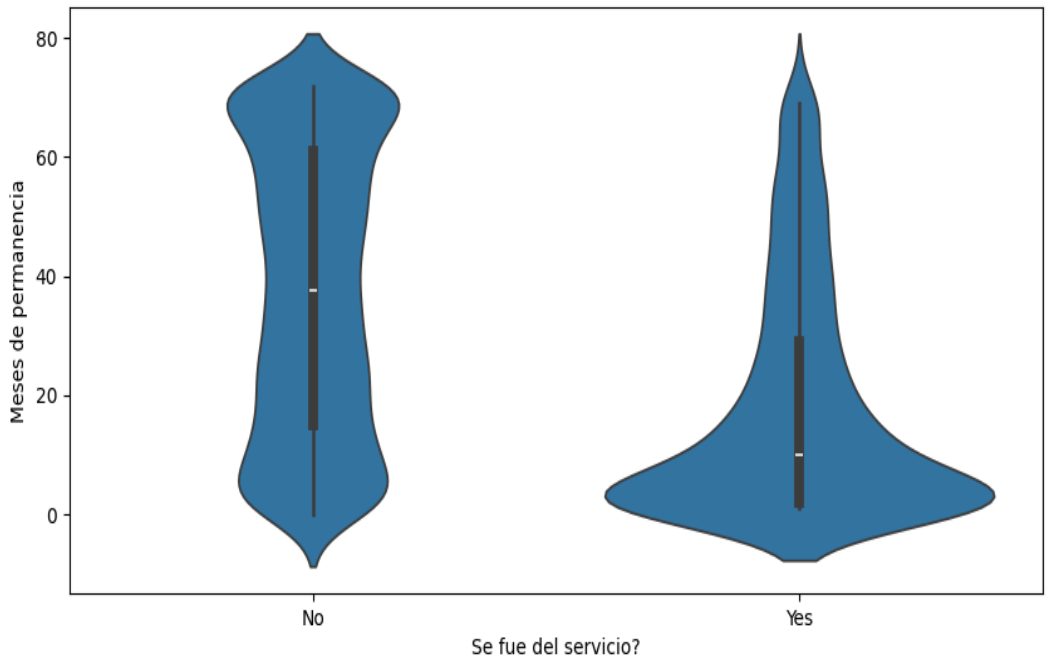
A continuación, se presentan algunas de las variables más relevantes del dataset de Telco Customer Churn:

- **customerID**: Identificador único del cliente.
- **gender**: Género del cliente: Male o Female.
- **SeniorCitizen**: Indica si el cliente es mayor a 65 años (1) o no (0).
- **Partner**: Si el cliente tiene pareja (Yes/No).
- **Dependents**: Si el cliente tiene personas a cargo (Yes/No).
- **tenure**: Cantidad de meses que el cliente ha permanecido con la empresa.
- **PhoneService**: Si el cliente tiene servicio telefónico (Yes/No).
- **MultipleLines**: Si tiene múltiples líneas telefónicas (Yes/No/No phone service).
- **InternetService**: Tipo de servicio de internet: (DSL/Fiber optic/No internet service).
- **OnlineSecurity**: Si tiene servicio de seguridad en línea.
- **OnlineBackup**: Si tiene servicio de backup online.
- **DeviceProtection**: Si cuenta con protección de dispositivo.
- **TechSupport**: Si tiene soporte técnico.
- **StreamingTV**: Si tiene servicio de streaming de TV.
- **StreamingMovies**: Si tiene servicio de streaming de películas.
- **Contract**: Tipo de contrato: Month-to-month, One year, Two year.
- **PaperlessBilling**: Si el cliente utiliza facturación sin papel (Yes/No).
- **PaymentMethod**: Método de pago: Electronic check, Mailed check, etc.
- **MonthlyCharges**: Cargo mensual del cliente.
- **TotalCharges**: Total de cargos acumulados por el cliente.
- **Churn**: Variable objetivo: indica si el cliente abandonó el servicio.

## Gráficos y análisis de Dataset

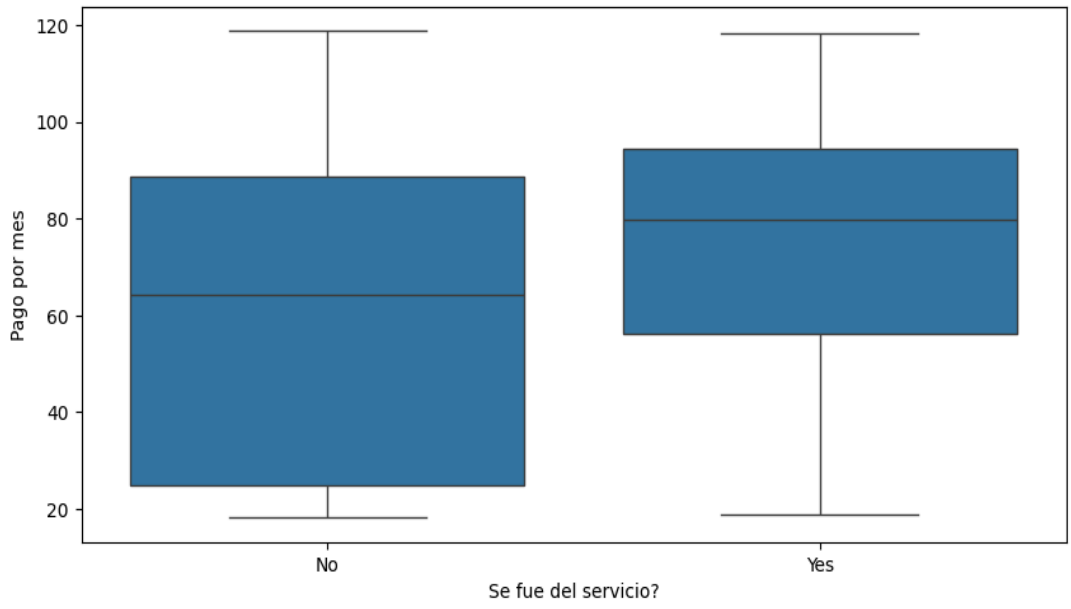
Este gráfico es un **gráfico de violín**. Muestra la distribución de la variable "Meses de permanencia" (tenure) para dos grupos: clientes que "No" se fueron del servicio (no churn) y clientes que "Sí" se fueron del servicio (churn).

En resumen, el gráfico te permite visualizar cómo se relaciona el tiempo que un cliente permanece en el servicio con la probabilidad de que abandone o no dicho servicio.

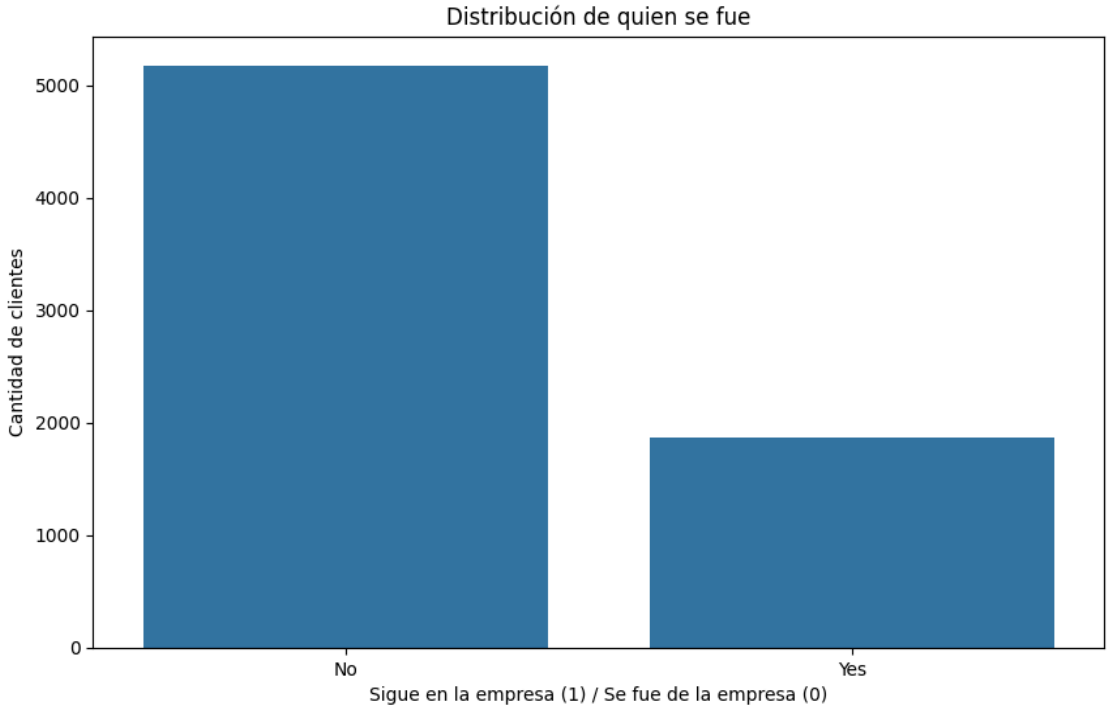



El segundo gráfico es un **gráfico de caja y bigotes (boxplot)**. Muestra la distribución de los "Pagos por mes" (MonthlyCharges) para dos grupos de clientes: aquellos que "No" se fueron del servicio y aquellos que "Sí" se fueron del servicio.

El gráfico te ayuda a comparar cómo varían los pagos mensuales entre los clientes que permanecen y los que abandonan el servicio, mostrando la mediana, los cuartiles y los posibles valores atípicos.



El tercer gráfico  muestra una distribución simple sobre cuántos clientes abandonaron la empresa, y cuántos se quedaron. Sirve para ver qué porcentaje de Churn tiene el DataSet

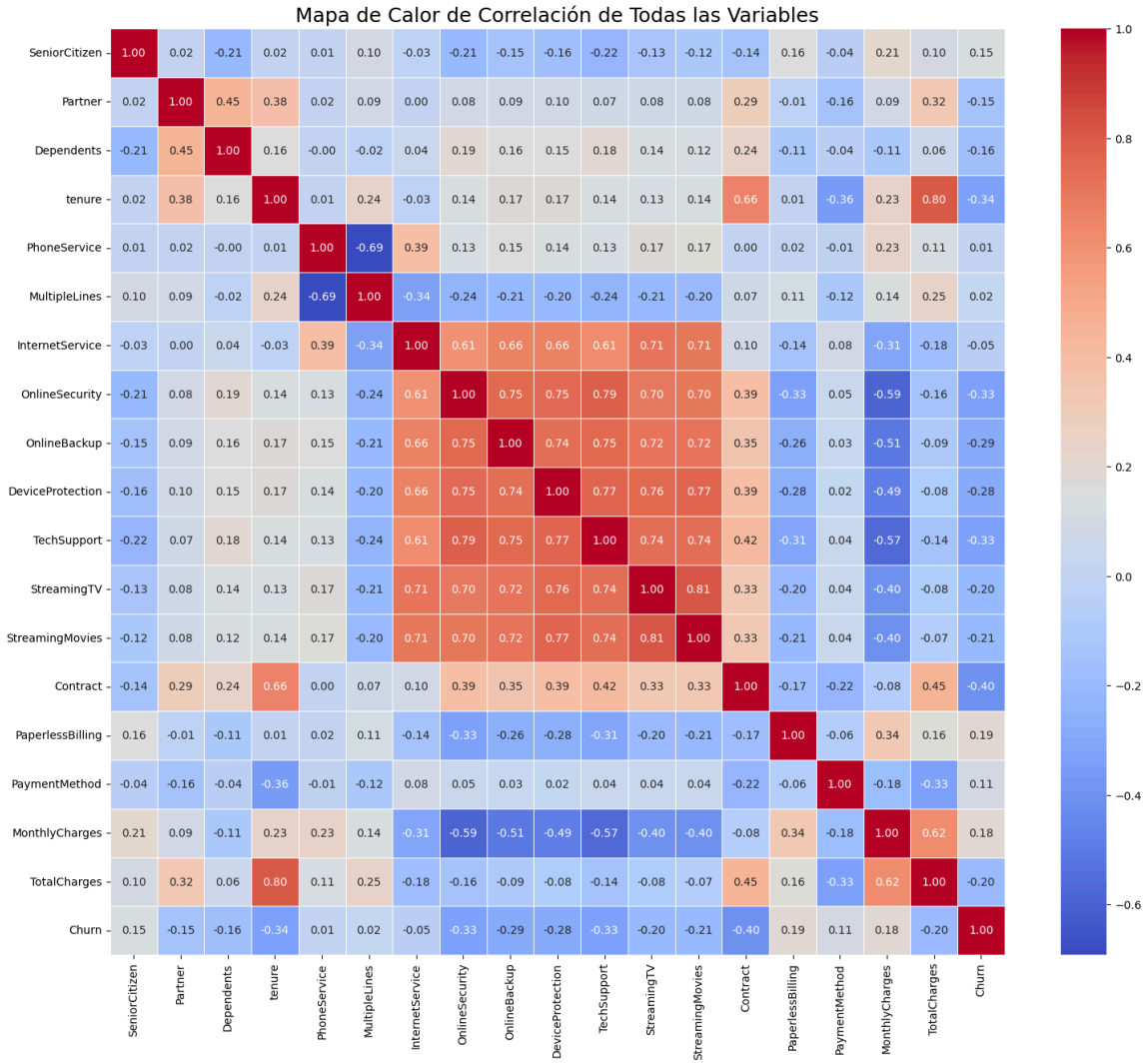



## Selección de características relevantes (Mapa calor)

Este gráfico es un **Mapa de Calor de Correlación**. Muestra la fuerza y dirección de la relación lineal entre cada par de variables en tu conjunto de datos.

El mapa de calor te permite identificar rápidamente qué variables están fuertemente relacionadas entre sí (valores cercanos a 1 o -1, colores intensos) y cuáles tienen poca o ninguna relación (valores cercanos a 0, colores tenues), es decir que mientras las correlaciones más se acerquen al valor 1, son más probables de irse, mientras que las correlaciones que se acerquen al valor -1, son más probables a quedarse en la empresa.

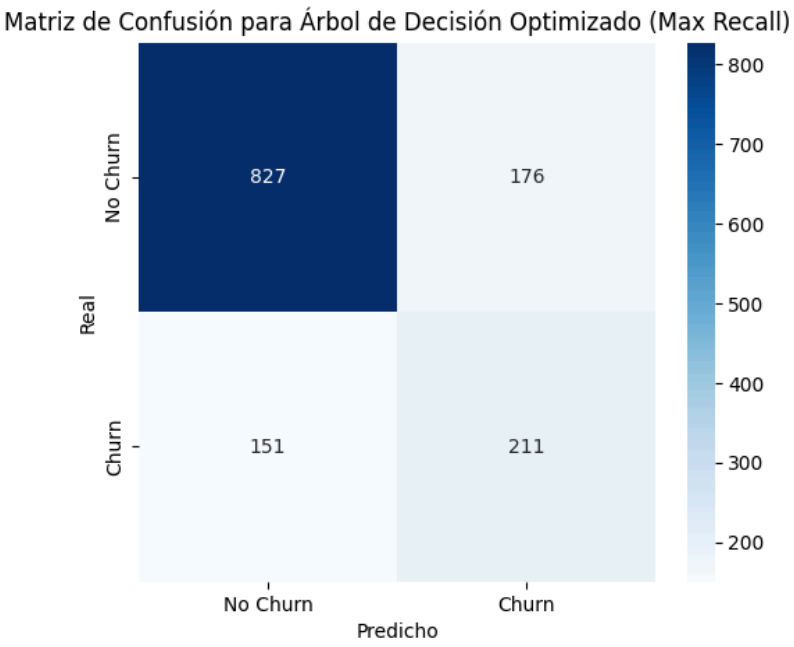
Esto es útil para entender las interdependencias en tus datos y la relación de cada variable con la variable objetivo 'Churn' .



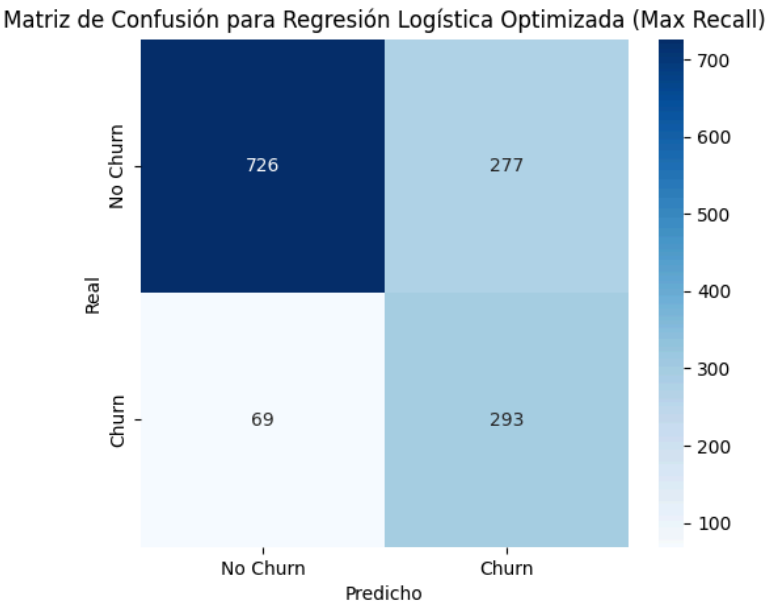

# Evaluación del rendimiento del modelo

## Matrices de confusión de cada modelo luego de entrenarlo

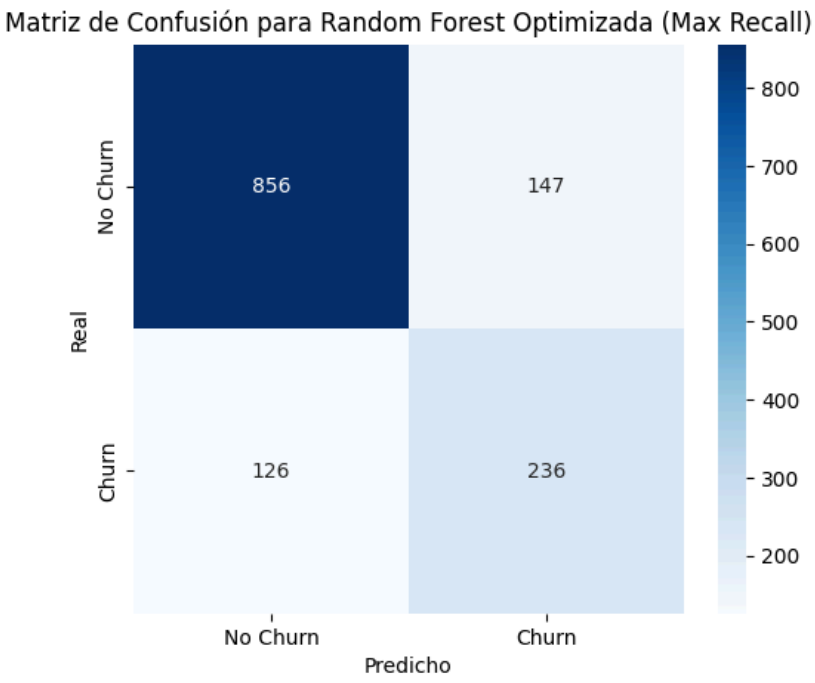
## Árbol de Decisión :




## Regresión Logística:




## Random Forest:

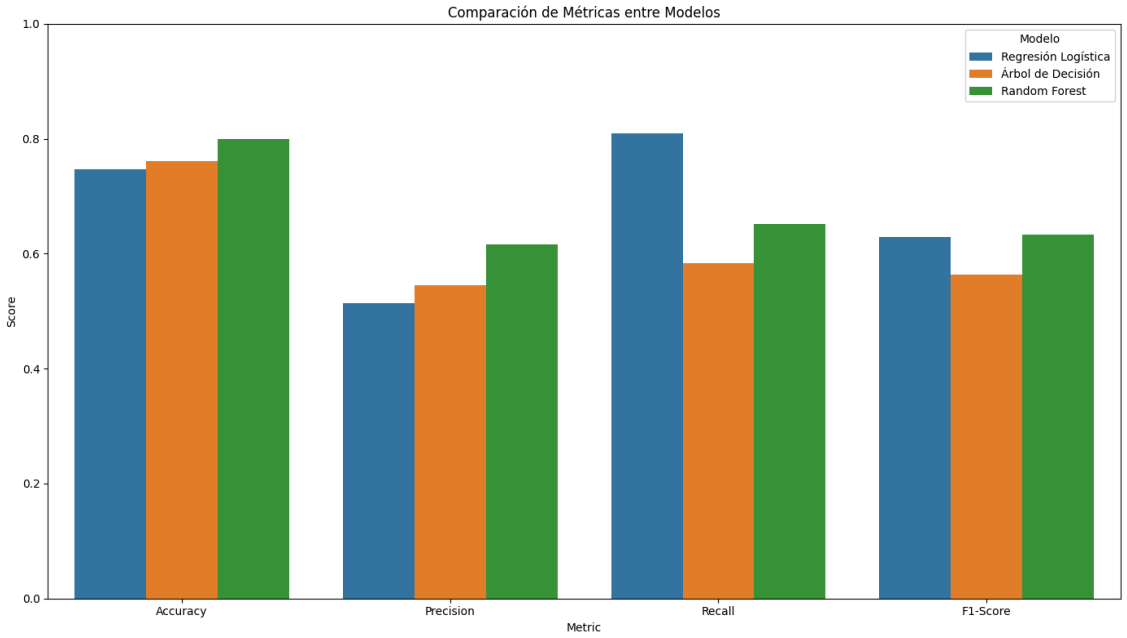



El modelo de **Regresión Logística** es el mejor para cumplir el objetivo de capturar a la mayor cantidad posible de clientes que se van (el mayor **Recall de un 81%**), pero genera muchas **falsas alarmas**. Por otro lado, el **Random Forest** es el modelo más equilibrado y eficiente, ofreciendo el mejor balance general entre detectar a los clientes en riesgo y la precisión de sus predicciones.

## Visualización de los datos obtenidos

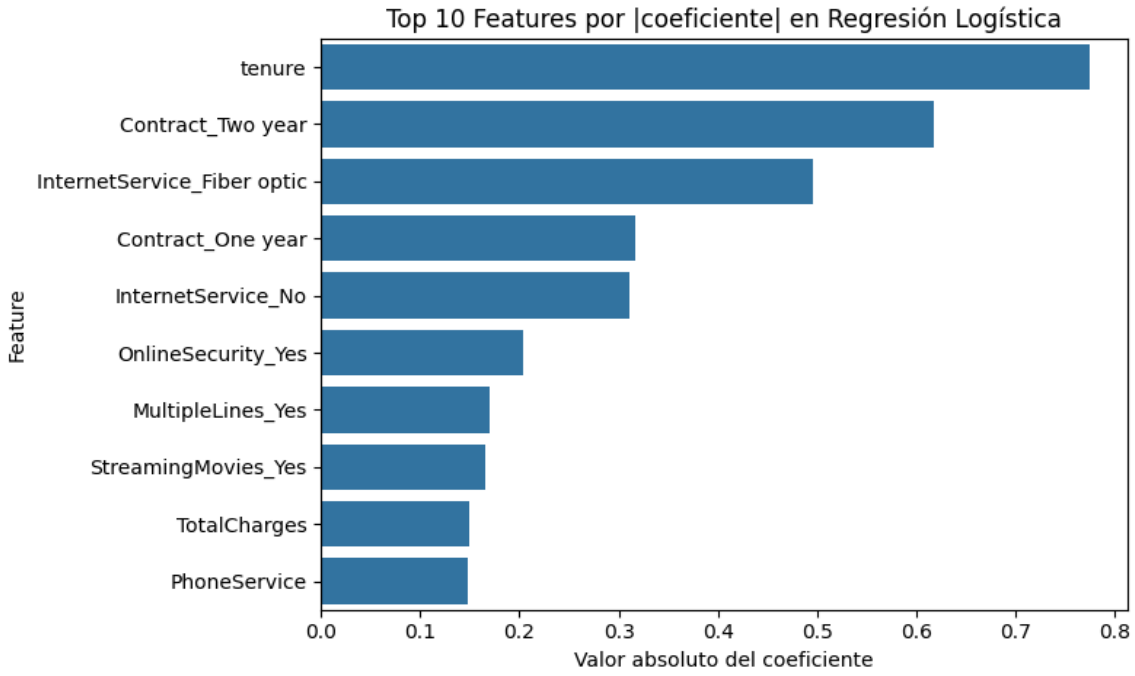
Este gráfico es un **gráfico de barras comparativo**. Muestra el rendimiento de tres modelos de clasificación (Regresión Logística, Árbol de Decisión y Random Forest) a través de cuatro métricas clave: Accuracy, Precision, Recall y F1-Score.

El gráfico te permite comparar visualmente qué modelo se desempeña mejor en cada una de estas métricas, facilitando la elección del modelo más adecuado según el objetivo específico (por ejemplo, maximizar el Recall para la detección de churn).




Este gráfico es un **gráfico de barras horizontales** que muestra las "Top 10 Features por |coeficiente|" para tu modelo de Regresión Logística.

Visualiza las diez características más influyentes que el modelo de Regresión Logística utiliza para predecir el abandono de clientes, basándose en la magnitud absoluta de sus coeficientes. Un coeficiente más grande (en valor absoluto) indica una mayor importancia de esa característica en la predicción del modelo.



**Curva ROC del modelo de Regresión Logística**:

Este gráfico es una **Curva ROC** para la Regresión Logística. Muestra qué tan bien el modelo puede distinguir entre clientes que se irán y los que se quedarán a diferentes umbrales de predicción  (Falsos y Verdaderos Positivos). La curva naranja representa el rendimiento del modelo, y el **AUC (Área Bajo la Curva) de 0.85** indica una buena capacidad de clasificación, ya que está lejos de la línea azul de predicción aleatoria y cerca de la esquina superior izquierda (Verdaderos Positivos).