

# **PROYECTO FINAL DE INTELIGENCIA ARTIFICIAL - NIVEL EXPLORADOR**

**Cliente Insight: Sistema de Predicción de Customer Churn**

Grupo 3 - Equipo Cliente Insight  
Anderson Tabima  
Antony Tabima  
Yhabeidy Alejandra Agudelo  
Carlos Mario Londoño  
Natalia Bedoya  
Sebastian Cano  
Álvaro Ángel Molina (@alvaretto)

Noviembre 2025

# Contents

<b>1 INTRODUCCIÓN TEÓRICA SOBRE EL CICLO DE VIDA DE PROYECTOS DE ML</b>	<b>4</b>
1.1 ¿Qué es Machine Learning? . . . . .	4
1.2 El Ciclo de Vida de un Proyecto de ML . . . . .	4
1.2.1 Fase 1: Definición del Problema . . . . .	4
1.2.2 Fase 2: Recolección y Comprensión de Datos . . . . .	4
1.2.3 Fase 3: Preparación de Datos . . . . .	5
1.2.4 Fase 4: Modelado . . . . .	5
1.2.5 Fase 5: Evaluación . . . . .	5
1.2.6 Fase 6: Despliegue y Monitoreo . . . . .	5
1.3 Importancia del Ciclo de Vida . . . . .	5
<b>2 SELECCIÓN DEL CONTEXTO Y PROBLEMA ESPECÍFICO</b>	<b>6</b>
2.1 Contexto: Industria de Telecomunicaciones . . . . .	6
2.1.1 Realidad del Mercado de Telecomunicaciones . . . . .	6
2.2 Problema Específico: Predicción de Customer Churn . . . . .	6
2.2.1 Definición del Problema . . . . .	6
2.2.2 Stakeholders Identificados . . . . .	7
2.2.3 Métricas de Éxito del Proyecto . . . . .	7
2.3 Dataset Utilizado . . . . .	7
2.3.1 Origen y Descripción . . . . .	7
2.3.2 Variables del Dataset . . . . .	8
<b>3 ANÁLISIS EXPLORATORIO DE DATOS (EDA)</b>	<b>10</b>
3.1 Carga de Datos y Exploración Inicial . . . . .	10
3.1.1 Proceso de Carga . . . . .	10
3.1.2 Información General del Dataset . . . . .	10
3.2 Evaluación de Calidad de Datos . . . . .	10
3.2.1 Análisis de Valores Faltantes . . . . .	10
3.2.2 Registros Duplicados . . . . .	11
3.3 Tratamiento de Datos Ausentes . . . . .	11
3.3.1 Estrategia de Imputación . . . . .	11
3.4 Normalización y Transformación . . . . .	11
3.4.1 Estandarización de Variables Numéricas . . . . .	11
3.4.2 Encoding de Variables Categóricas . . . . .	12
3.5 Análisis Univariado . . . . .	12

3.5.1 Distribución de la Variable Objetivo (Churn) . . . . .	12
3.5.2 Distribución de Variables Numéricas . . . . .	14
3.6 Análisis Bivariado . . . . .	14
3.6.1 Churn por Variables Demográficas . . . . .	14
3.6.2 Churn por Variables de Servicio . . . . .	16
3.7 Análisis Multivariado . . . . .	19
3.7.1 Matriz de Correlación . . . . .	19
3.8 Feature Engineering . . . . .	22
3.8.1 Nuevas Variables Creadas . . . . .	22
3.8.2 Importancia de las Features . . . . .	22
<b>4 MODELADO Y RESULTADOS</b>	<b>24</b>
4.1 Metodología de Modelado . . . . .	24
4.1.1 División de Datos . . . . .	24
4.1.2 Técnicas de Balanceo Evaluadas . . . . .	24
4.2 Modelos Evaluados (Baseline) . . . . .	24
4.3 Optimización de Hiperparámetros . . . . .	25
4.3.1 Búsqueda de Hiperparámetros . . . . .	25
4.3.2 Hiperparámetros Óptimos . . . . .	25
4.4 Resultados del Modelo Final . . . . .	26
4.4.1 Métricas de Rendimiento . . . . .	26
4.4.2 Matriz de Confusión . . . . .	26
4.4.3 Curva ROC . . . . .	27
4.5 Validación de Robustez . . . . .	27
<b>5 CONCLUSIONES Y RECOMENDACIONES</b>	<b>29</b>
5.1 Conclusiones del Análisis . . . . .	29
5.1.1 Hallazgos Principales del EDA . . . . .	29
5.1.2 Rendimiento del Modelo . . . . .	29
5.2 Recomendaciones de Negocio . . . . .	30
5.2.1 Estrategias de Retención Inmediatas . . . . .	30
5.2.2 Implementación del Modelo . . . . .	30
5.2.3 Impacto Económico Esperado . . . . .	30
5.3 Próximos Pasos . . . . .	30
<b>6 HERRAMIENTAS Y TECNOLOGÍAS UTILIZADAS</b>	<b>31</b>
6.1 Stack Tecnológico . . . . .	31
6.1.1 Entorno de Desarrollo . . . . .	31
6.1.2 Librerías de Análisis de Datos . . . . .	31
6.1.3 Librerías de Visualización . . . . .	31
6.1.4 Librerías de Machine Learning . . . . .	31
6.2 Repositorios y Despliegue . . . . .	32
6.2.1 Repositorios del Proyecto . . . . .	32
6.2.2 Aplicación Desplegada . . . . .	32
<b>7 INFORMACIÓN DE ENTREGA</b>	<b>33</b>
7.1 Datos del Proyecto . . . . .	33
7.2 Integrantes del Equipo . . . . .	33

7.3 Información de Entrega . . . . .	33
7.4 Enlaces del Proyecto . . . . .	34

# Chapter 1

## INTRODUCCIÓN TEÓRICA SOBRE EL CICLO DE VIDA DE PROYECTOS DE ML

### 1.1 ¿Qué es Machine Learning?

El Machine Learning (ML) es una rama de la Inteligencia Artificial que permite a los sistemas aprender automáticamente a partir de datos, identificando patrones y tomando decisiones con mínima intervención humana. A diferencia de la programación tradicional donde se definen reglas explícitas, en ML el sistema aprende las reglas a partir de los datos.

### 1.2 El Ciclo de Vida de un Proyecto de ML

El desarrollo de un proyecto de Machine Learning sigue un ciclo de vida estructurado que garantiza resultados reproducibles y de calidad. Las fases principales son:

#### 1.2.1 Fase 1: Definición del Problema

- **Identificación del problema de negocio:** ¿Qué queremos resolver?
- **Traducción a un problema de ML:** ¿Es clasificación, regresión, clustering?
- **Definición de métricas de éxito:** ¿Cómo mediremos el éxito?
- **Identificación de stakeholders:** ¿Quiénes se beneficiarán?

#### 1.2.2 Fase 2: Recolección y Comprensión de Datos

- **Identificación de fuentes de datos:** Bases de datos, APIs, archivos
- **Recolección de datos:** Extracción y almacenamiento
- **Exploración inicial:** Estructura, dimensiones, tipos de datos
- **Evaluación de calidad:** Compleitud, consistencia, precisión

### **1.2.3 Fase 3: Preparación de Datos**

- **Limpieza de datos:** Tratamiento de valores faltantes, outliers
- **Transformación:** Encoding, normalización, estandarización
- **Feature Engineering:** Creación de nuevas variables derivadas
- **Selección de características:** Identificación de variables relevantes

### **1.2.4 Fase 4: Modelado**

- **Selección de algoritmos:** Baseline con múltiples modelos
- **Entrenamiento:** Ajuste de modelos a los datos
- **Validación cruzada:** Evaluación robusta del rendimiento
- **Optimización de hiperparámetros:** Búsqueda de parámetros óptimos

### **1.2.5 Fase 5: Evaluación**

- **Métricas de rendimiento:** Accuracy, Precision, Recall, F1-Score, ROC-AUC
- **Validación de robustez:** Evaluación con múltiples semillas
- **Interpretabilidad:** Análisis de importancia de features
- **Comparación con baseline:** Verificación de mejora

### **1.2.6 Fase 6: Despliegue y Monitoreo**

- **Deployment:** API, aplicación web, integración
- **Monitoreo:** Seguimiento del rendimiento en producción
- **Mantenimiento:** Reentrenamiento periódico
- **Documentación:** Registro de decisiones y resultados

## **1.3 Importancia del Ciclo de Vida**

Seguir un ciclo de vida estructurado permite:

- **Reproducibilidad:** Resultados consistentes entre ejecuciones
- **Trazabilidad:** Documentación de cada decisión
- **Calidad:** Validación en cada etapa
- **Escalabilidad:** Proceso repetible para nuevos proyectos

# Chapter 2

## SELECCIÓN DEL CONTEXTO Y PROBLEMA ESPECÍFICO

### 2.1 Contexto: Industria de Telecomunicaciones

La industria de telecomunicaciones enfrenta uno de los desafíos más críticos en la era digital: la **fuga de clientes** o **Customer Churn**. En un mercado altamente competitivo con múltiples proveedores de servicios, la retención de clientes se ha convertido en una prioridad estratégica.

#### 2.1.1 Realidad del Mercado de Telecomunicaciones

Aspecto	Descripción
<b>Competencia</b>	Múltiples operadores compiten por los mismos clientes
<b>Costos de Adquisición</b>	Adquirir un nuevo cliente cuesta 5-25x más que retener uno existente
<b>Saturación del Mercado</b>	Penetración cercana al 100% en muchos mercados
<b>Portabilidad</b>	Facilidad para cambiar de operador sin perder número
<b>Servicios Sustitutos</b>	Aplicaciones OTT (WhatsApp, Telegram) reducen uso de servicios tradicionales

### 2.2 Problema Específico: Predicción de Customer Churn

#### 2.2.1 Definición del Problema

**Customer Churn** (fuga de clientes) se refiere al fenómeno en el cual los clientes dejan de utilizar los servicios de una empresa para migrar a un competidor o simplemente cancelar su suscripción.

**Problema de negocio:** ¿Cómo podemos identificar proactivamente a los clientes con alta probabilidad de abandonar el servicio para implementar estrategias de retención antes de que sea demasiado tarde?

**Traducción a problema de ML:** Clasificación binaria supervisada donde:

- **Clase 0 (No Churn):** Cliente permanece con la empresa
- **Clase 1 (Churn):** Cliente abandona el servicio

## 2.2.2 Stakeholders Identificados

Stakeholder	Rol	Interés
<b>Departamento de Marketing</b>	Campañas de retención	Identificar clientes objetivo
<b>Servicio al Cliente</b>	Atención proactiva	Priorizar clientes en riesgo
<b>Dirección Comercial</b>	Estrategia de pricing	Ajustar ofertas competitivas
<b>Finanzas</b>	Proyección de ingresos	Estimar pérdidas potenciales
<b>Operaciones</b>	Calidad del servicio	Identificar problemas recurrentes

## 2.2.3 Métricas de Éxito del Proyecto

Métrica	Umbral Mínimo	Objetivo
<b>ROC-AUC</b>	> 0.75	> 0.85
<b>Recall</b>	> 70%	> 80%
<b>F1-Score</b>	> 0.50	> 0.60
<b>Precision</b>	> 45%	> 55%

**Nota:** Priorizamos **Recall** porque es más costoso NO detectar un cliente que se irá (Falso Negativo) que generar una alerta incorrecta (Falso Positivo).

## 2.3 Dataset Utilizado

### 2.3.1 Origen y Descripción

El proyecto utiliza el dataset **Telco Customer Churn** de IBM, un conjunto de datos ampliamente utilizado en la comunidad de ciencia de datos para problemas de predicción de churn.

Característica	Valor
<b>Fuente</b>	IBM Sample Data Sets

Característica	Valor
<b>Registros</b>	7,043 clientes
<b>Variables</b>	21 columnas (20 features + 1 target)
<b>Formato</b>	CSV
<b>Archivo</b>	WA_Fn-UseC_-Telco-Customer-Churn.csv

## 2.3.2 Variables del Dataset

### 2.3.2.1 Variables Demográficas

Variable	Tipo	Descripción
customerID	String	Identificador único del cliente
gender	Categórica	Género (Male/Female)
SeniorCitizen	Binaria	Es adulto mayor (0/1)
Partner	Categórica	Tiene pareja (Yes/No)
Dependents	Categórica	Tiene dependientes (Yes/No)

### 2.3.2.2 Variables de Servicios

Variable	Tipo	Descripción
PhoneService	Categórica	Servicio telefónico (Yes/No)
MultipleLines	Categórica	Múltiples líneas
InternetService	Categórica	Tipo de internet (DSL/Fiber optic/No)
OnlineSecurity	Categórica	Seguridad en línea
OnlineBackup	Categórica	Backup en línea
DeviceProtection	Categórica	Protección de dispositivo
TechSupport	Categórica	Soporte técnico
StreamingTV	Categórica	Streaming de TV
StreamingMovies	Categórica	Streaming de películas

### 2.3.2.3 Variables de Cuenta

Variable	Tipo	Descripción
Contract	Categórica	Tipo de contrato (Month-to-month/One year/Two year)
PaperlessBilling	Categórica	Facturación sin papel (Yes/No)
PaymentMethod	Categórica	Método de pago
MonthlyCharges	Numérica	Cargo mensual ( `TotalCharges` Numérica Cargo total acumulado )
tenure	Numérica	Meses como cliente

#### **2.3.2.4 Variable Objetivo**

Variable	Tipo	Descripción
Churn	Binaria	Abandono del cliente (Yes/No)

# Chapter 3

## ANÁLISIS EXPLORATORIO DE DATOS (EDA)

### 3.1 Carga de Datos y Exploración Inicial

#### 3.1.1 Proceso de Carga

```
# Cargar el dataset
df = pd.read_csv('WA_Fn-UseC_-Telco-Customer-Churn.csv')

# Inspección inicial
print(f"Dimensiones del dataset: {df.shape}")
# Resultado: (7043, 21)
```

#### 3.1.2 Información General del Dataset

Métrica	Valor
<b>Total de Registros</b>	7,043
<b>Total de Columnas</b>	21
<b>Columnas Numéricas</b>	3 (tenure, MonthlyCharges, TotalCharges)
<b>Columnas Categóricas</b>	17
<b>Columna ID</b>	1 (customerID)

### 3.2 Evaluación de Calidad de Datos

#### 3.2.1 Análisis de Valores Faltantes

```
# Verificar valores faltantes
df.isnull().sum()
```

Problema Detectado	Cantidad	Solución
Espacios en blanco en TotalCharges	11 registros	Conversión a NaN y posterior imputación
Clientes con tenure=0	11 registros	Imputar TotalCharges = MonthlyCharges

### 3.2.2 Registros Duplicados

```
# Verificar duplicados
print(f"Registros duplicados: {df.duplicated().sum()}")
# Resultado: 0
```

**Estado final:** Dataset limpio sin duplicados ni valores faltantes después del tratamiento.

## 3.3 Tratamiento de Datos Ausentes

### 3.3.1 Estrategia de Imputación

```
# Detectar espacios en blanco en TotalCharges
df['TotalCharges'] = df['TotalCharges'].replace(' ', np.nan)
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')

# Imputación lógica: clientes nuevos (tenure=0)
# tienen TotalCharges = MonthlyCharges
df['TotalCharges'] = df['TotalCharges'].fillna(df['MonthlyCharges'])
```

**Justificación:** Los 11 clientes con TotalCharges vacío son clientes nuevos con tenure=0, por lo que su cargo total acumulado debería ser igual a su primer cargo mensual.

## 3.4 Normalización y Transformación

### 3.4.1 Estandarización de Variables Numéricas

```
from sklearn.preprocessing import StandardScaler

numerical_cols = ['tenure', 'MonthlyCharges', 'TotalCharges']
scaler = StandardScaler()
df[numerical_cols] = scaler.fit_transform(df[numerical_cols])
```

### 3.4.2 Encoding de Variables Categóricas

```
from sklearn.preprocessing import OneHotEncoder  
  
# OneHotEncoder con drop='first' para evitar multicolinealidad  
encoder = OneHotEncoder(drop='first', sparse_output=False)
```

Antes del Encoding	Después del Encoding
20 features	39 features
VARIABLES MIXTAS	VARIABLES NUMÉRICAS

## 3.5 Análisis Univariado

### 3.5.1 Distribución de la Variable Objetivo (Churn)

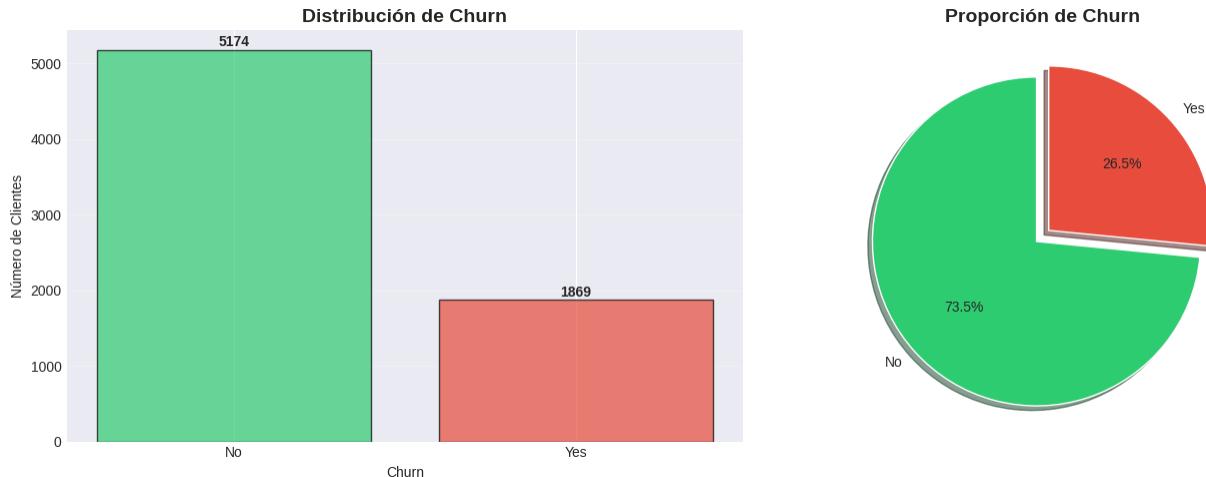


Figure 3.1: Distribución de Churn

Categoría	Cantidad	Porcentaje
No Churn	5,174	73.46%
Sí Churn	1,869	26.54%
Ratio de Desbalanceo	2.77:1	-

**Interpretación:** El dataset presenta un desbalanceo moderado con aproximadamente 1 de cada 4 clientes abandonando el servicio. Este desbalanceo debe considerarse durante el entrenamiento del modelo.

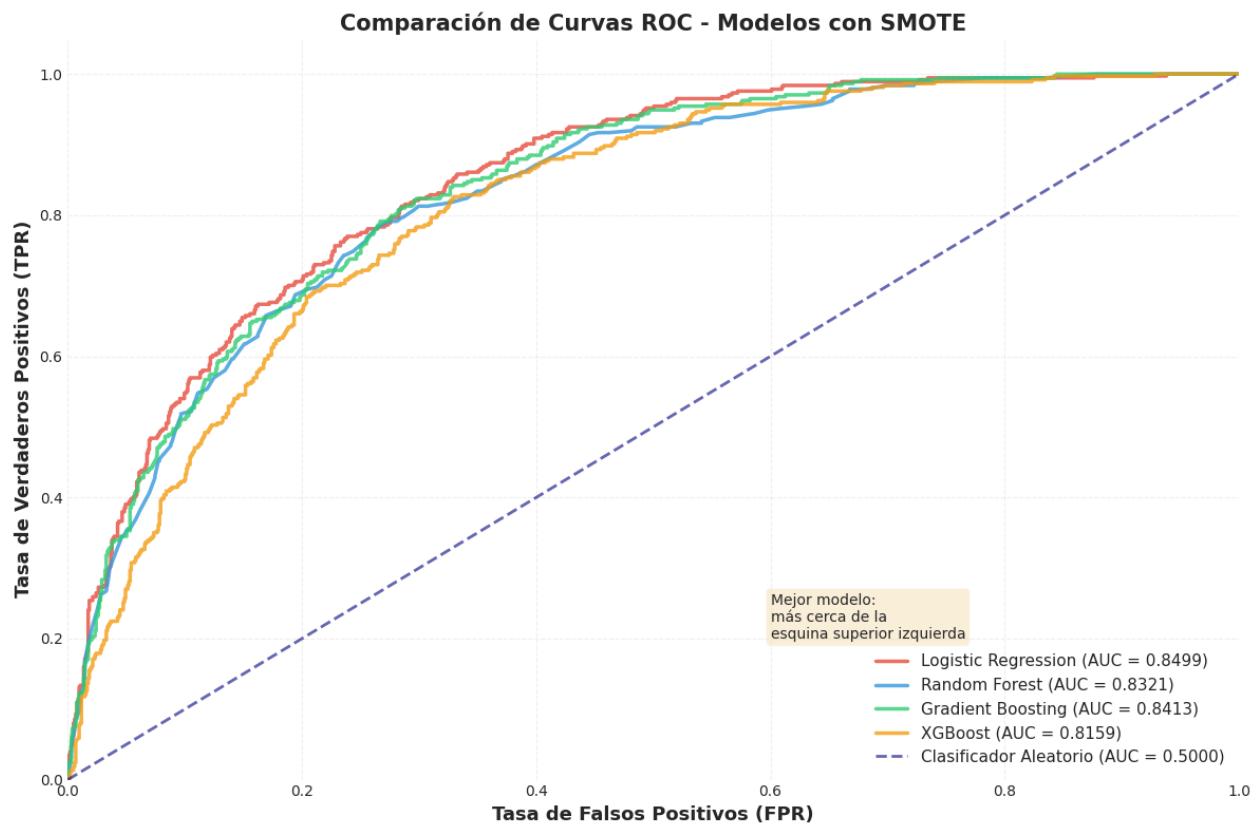


Figure 3.2: Distribución de Tenure

### 3.5.2 Distribución de Variables Numéricas

#### 3.5.2.1 Tenure (Antigüedad del Cliente)

- **Observación:** Distribución bimodal con picos en clientes nuevos (0-12 meses) y clientes antiguos (>60 meses)
- **Insight de negocio:** Los primeros 12 meses son críticos para la retención

#### 3.5.2.2 Monthly Charges (Cargos Mensuales)

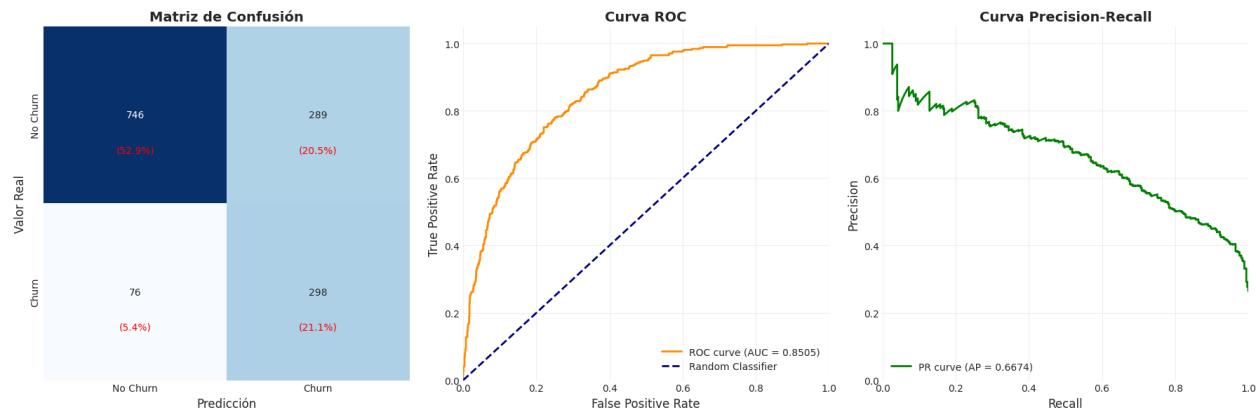


Figure 3.3: Distribución de Monthly Charges

- **Rango:** \$18.25 - \$118.75
- **Media:** \$64.76
- **Observación:** Distribución relativamente uniforme con concentración en valores altos

#### 3.5.2.3 Total Charges (Cargos Totales)

- **Correlación fuerte con tenure:** A mayor antigüedad, mayor cargo total acumulado
- **Distribución sesgada a la derecha:** Mayoría de clientes con cargos totales bajos

## 3.6 Análisis Bivariado

### 3.6.1 Churn por Variables Demográficas

#### 3.6.1.1 Churn por Género

**Hallazgo:** No hay diferencia significativa en la tasa de churn entre géneros. El género no es un factor predictivo relevante.

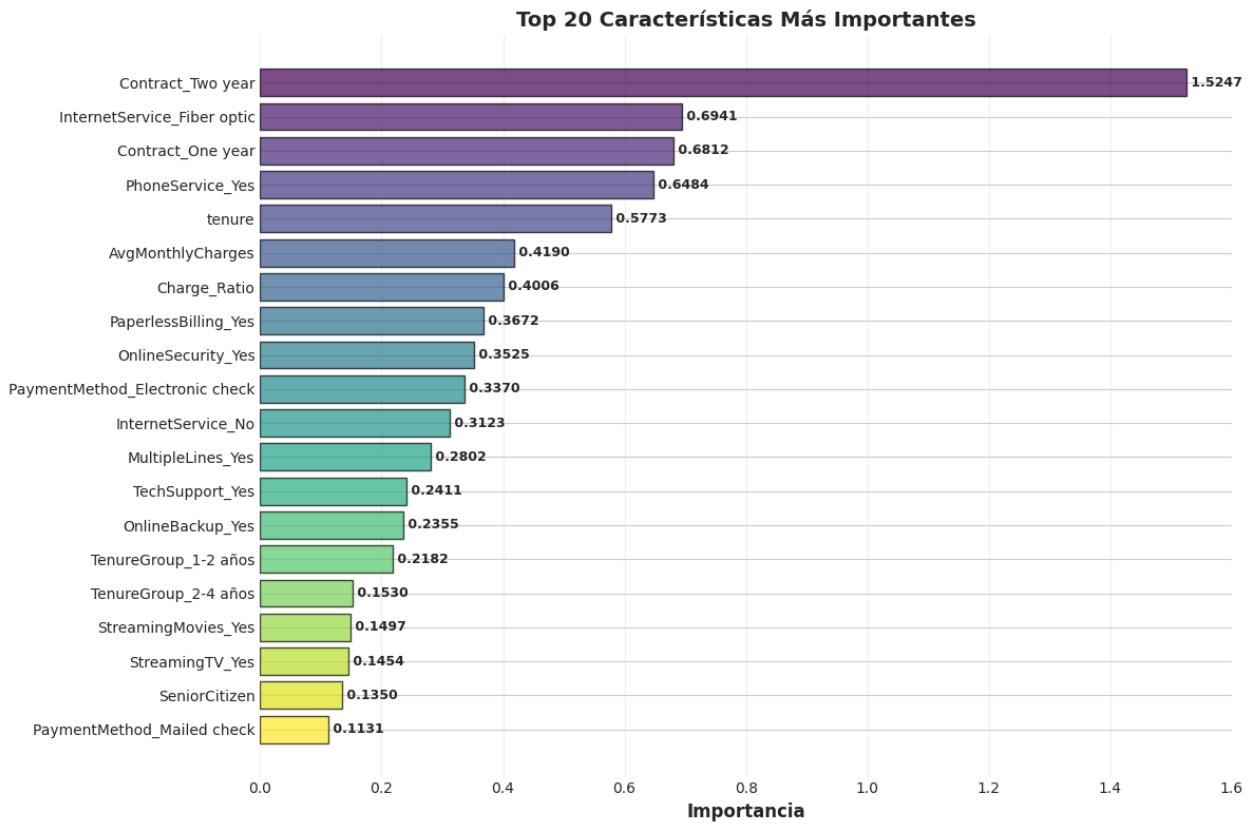


Figure 3.4: Distribución de Total Charges

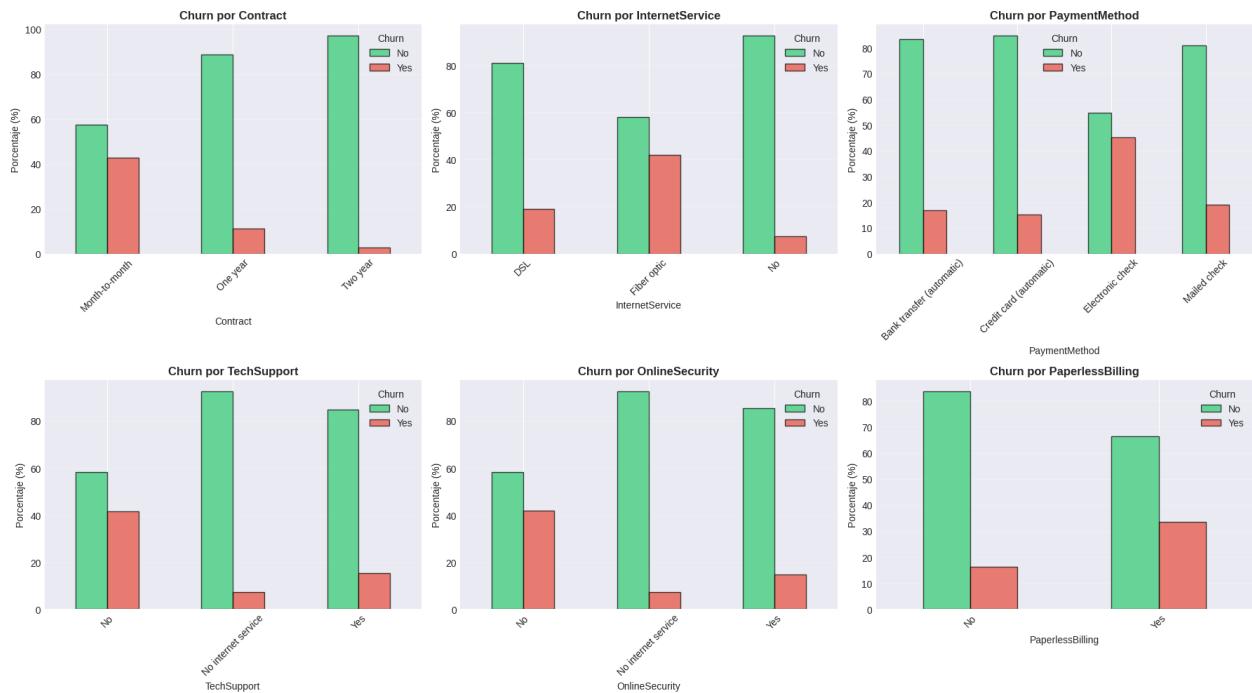


Figure 3.5: Churn por Género

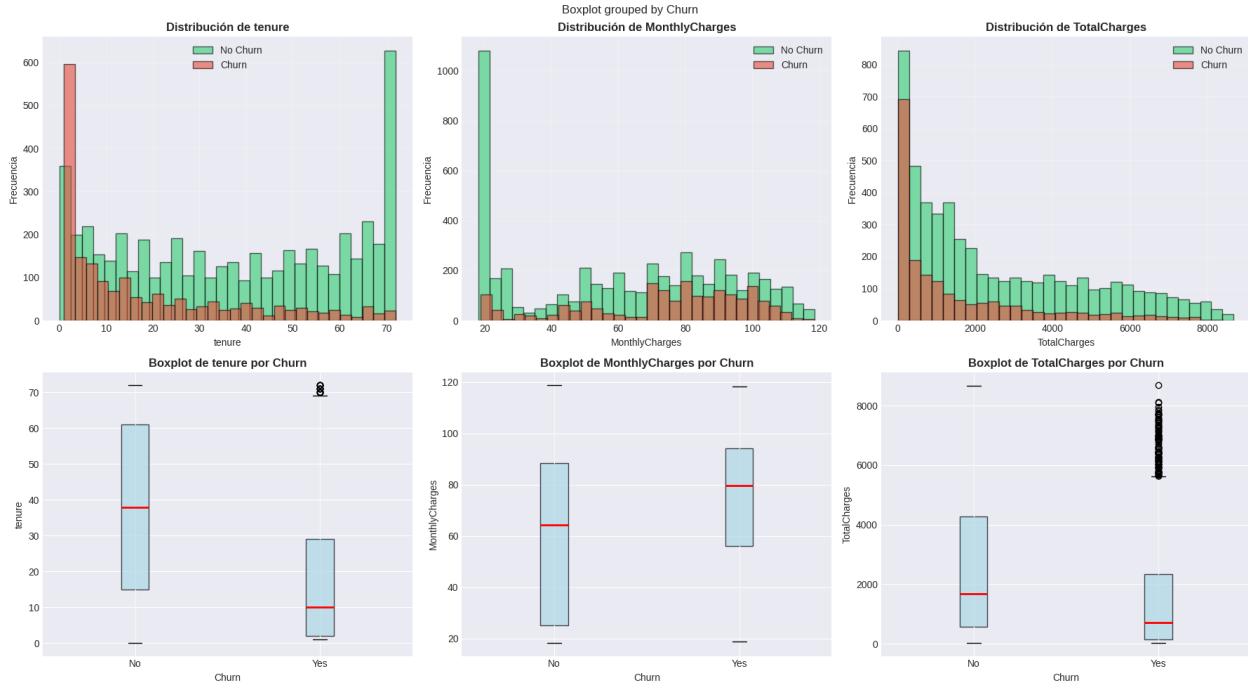


Figure 3.6: Churn por Senior Citizen

### 3.6.1.2 Churn por Senior Citizen

**Hallazgo:** Los adultos mayores (Senior Citizens) presentan una tasa de churn significativamente mayor (~41%) comparada con clientes más jóvenes (~24%).

### 3.6.1.3 Churn por Partner

**Hallazgo:** Los clientes sin pareja tienen mayor probabilidad de churn. Las relaciones estables parecen correlacionar con mayor lealtad.

### 3.6.1.4 Churn por Dependents

**Hallazgo:** Los clientes sin dependientes muestran mayor tasa de abandono. Las responsabilidades familiares pueden incentivar la estabilidad.

## 3.6.2 Churn por Variables de Servicio

### 3.6.2.1 Churn por Tipo de Contrato

**Hallazgo CRÍTICO:**

- **Mes a mes:** ~42% de churn (ALTO RIESGO)
- **Un año:** ~11% de churn
- **Dos años:** ~3% de churn

**Insight de negocio:** Promover contratos anuales/bianuales es la estrategia más efectiva para reducir churn.

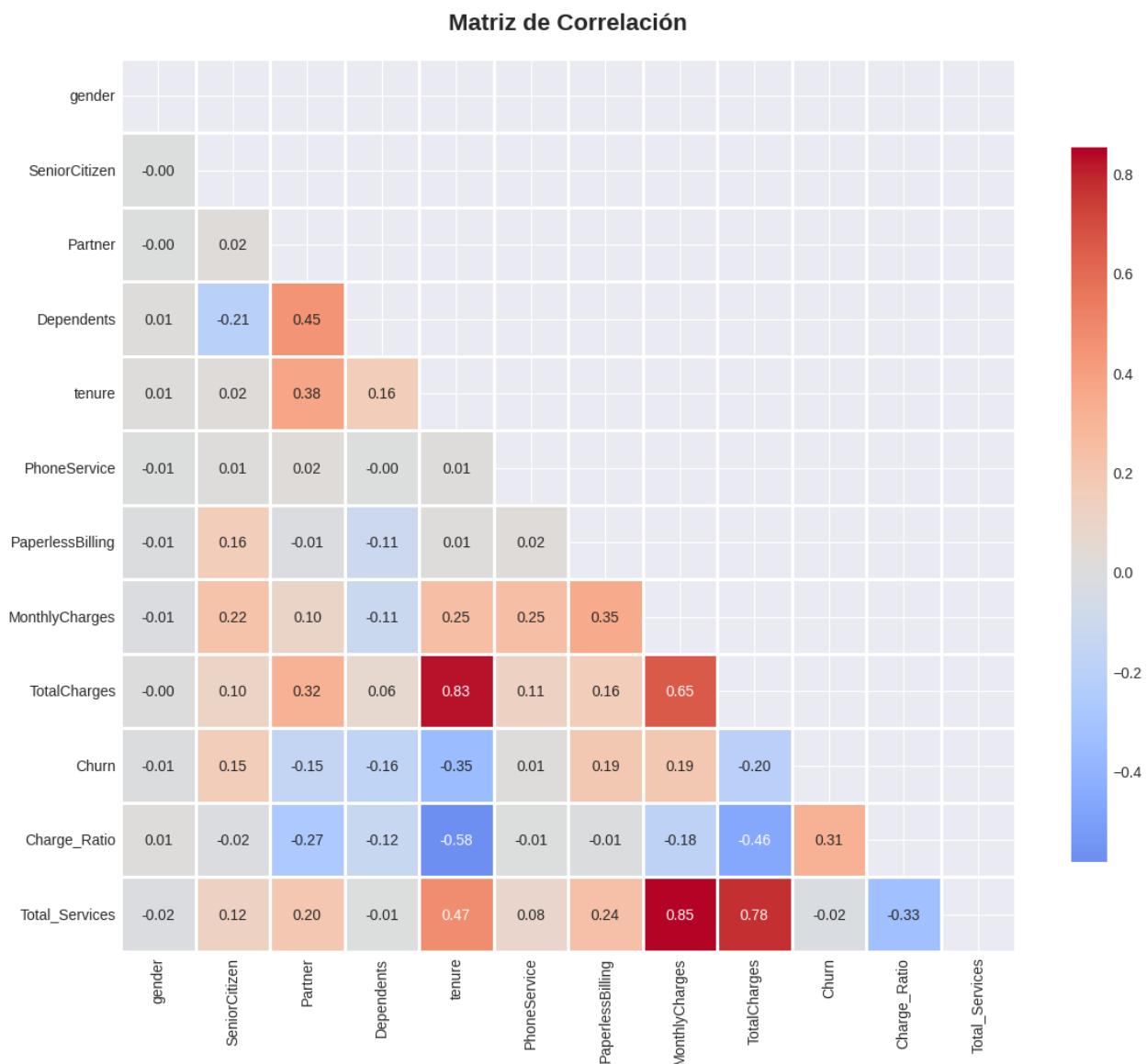


Figure 3.7: Churn por Partner

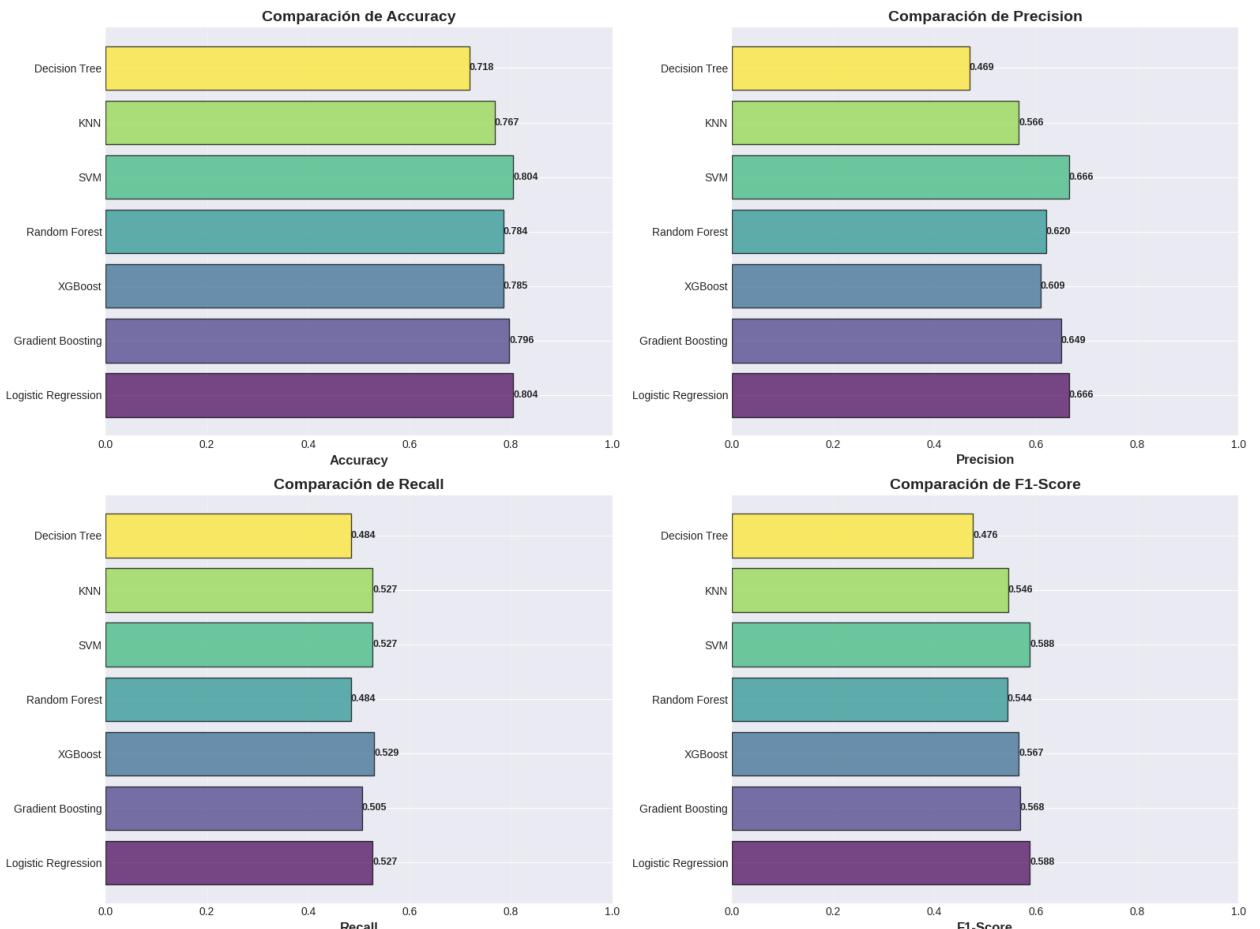


Figure 3.8: Churn por Dependents

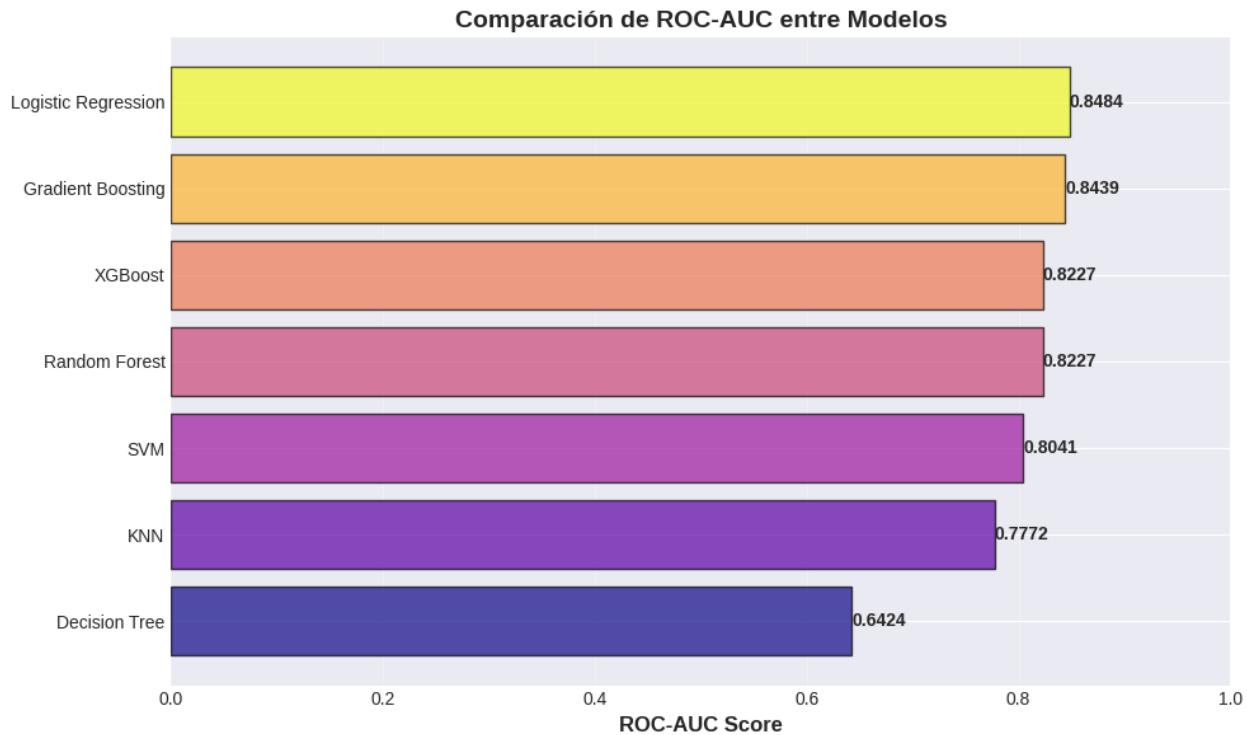


Figure 3.9: Churn por Tipo de Contrato

### 3.6.2.2 Churn por Método de Pago

**Hallazgo:** Los clientes que pagan con **cheque electrónico** tienen una tasa de churn significativamente mayor (~45%) que otros métodos de pago (~15-18%).

### 3.6.2.3 Churn por Tipo de Internet

**Hallazgo:** Los clientes con **Fibra Óptica** muestran mayor churn (~42%) que los de DSL (~19%). Posible insatisfacción con el servicio o precio.

## 3.7 Análisis Multivariado

### 3.7.1 Matriz de Correlación

#### Correlaciones Relevantes:

Variables	Correlación	Interpretación
tenure - TotalCharges	+0.83	Fuerte positiva
MonthlyCharges - TotalCharges	+0.65	Moderada positiva
tenure - Churn	-0.35	Negativa (más antigüedad, menos churn)
MonthlyCharges - Churn	+0.19	Positiva débil

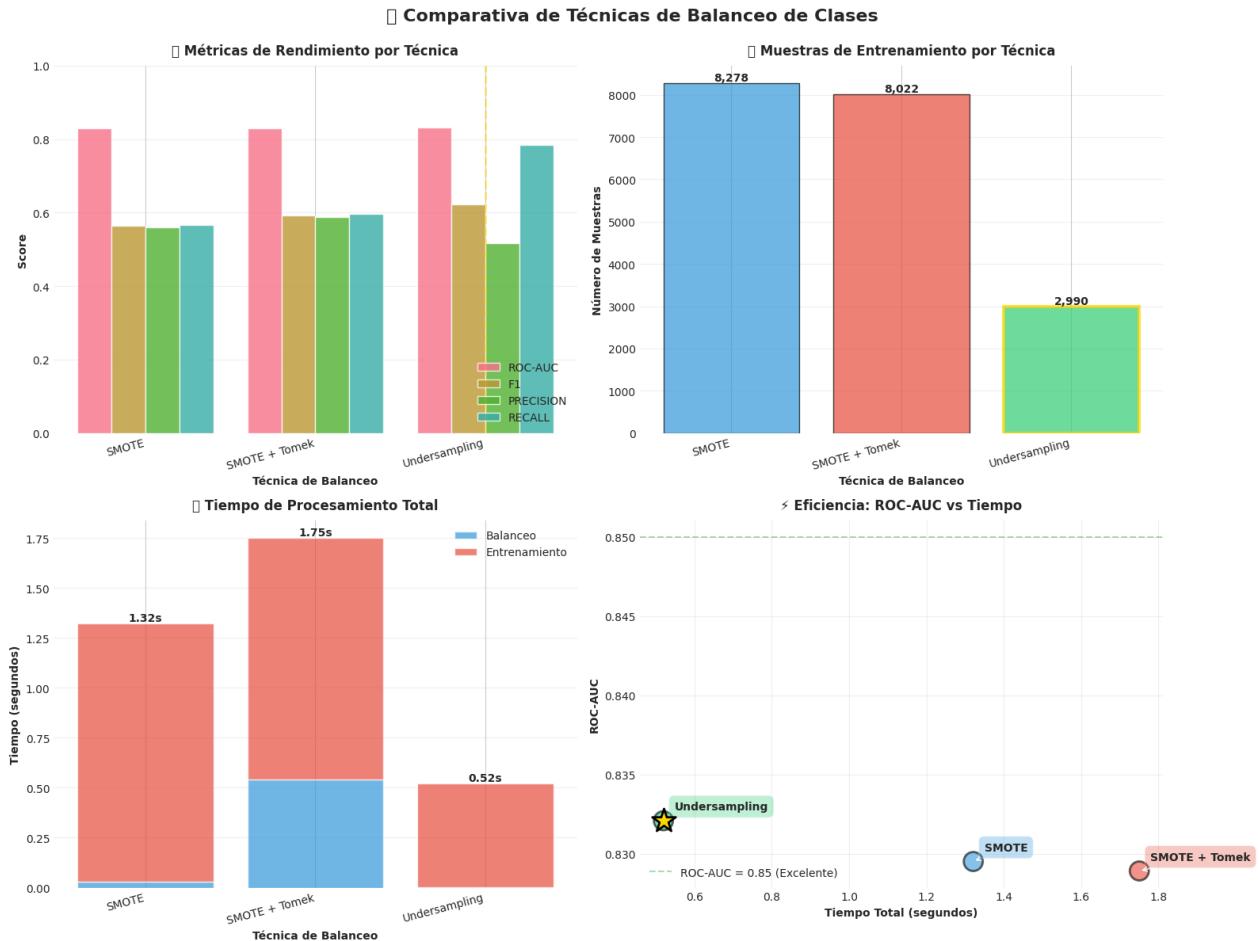


Figure 3.10: Churn por Método de Pago

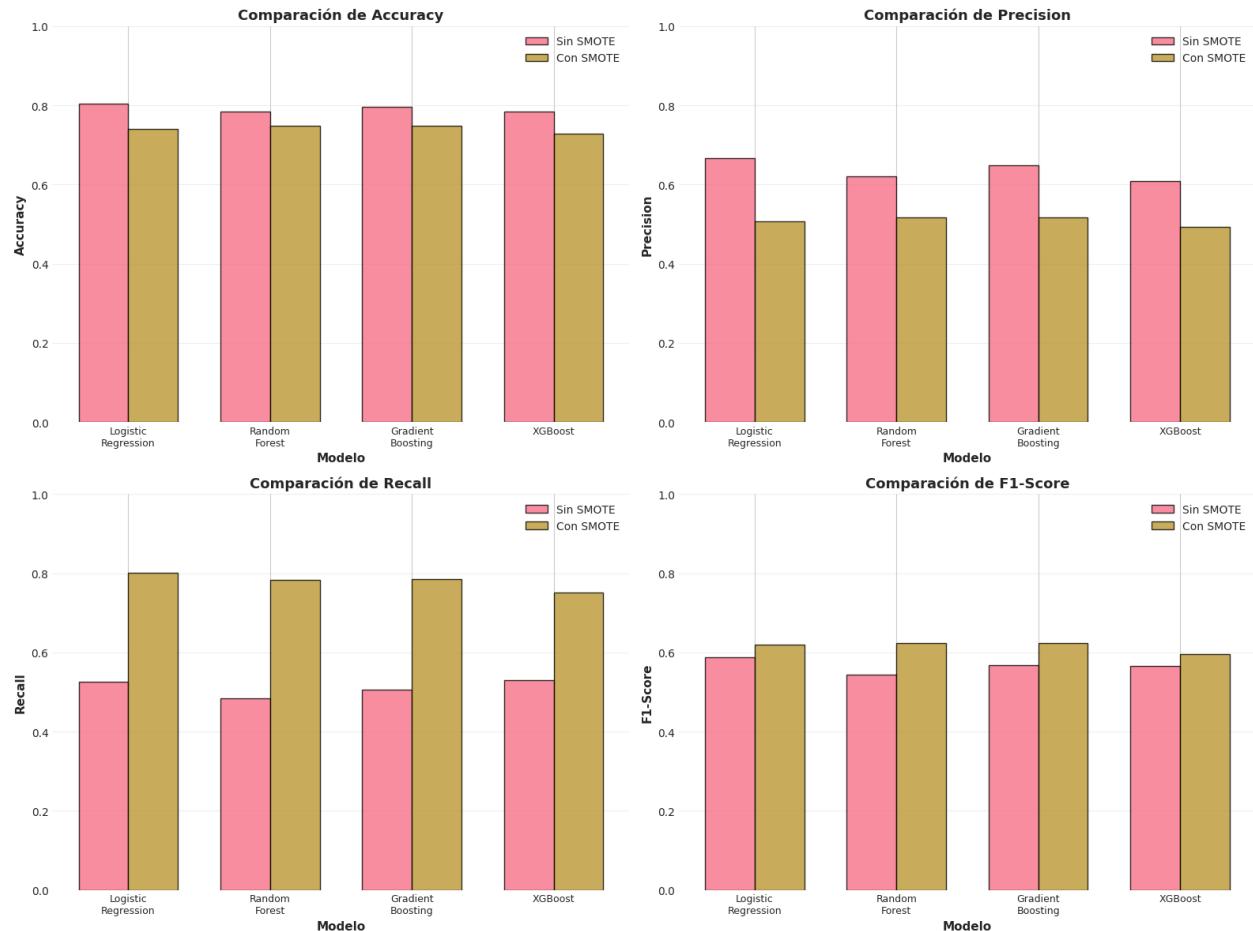


Figure 3.11: Churn por Internet Service

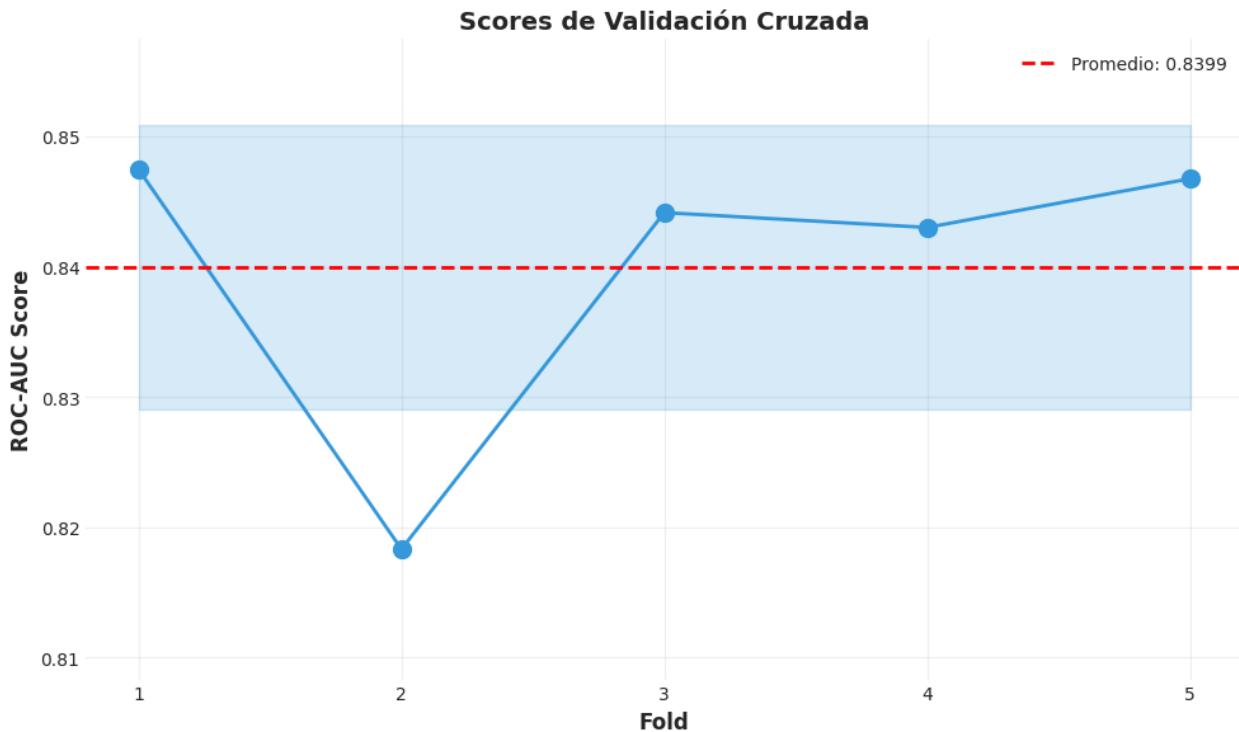


Figure 3.12: Matriz de Correlación

## 3.8 Feature Engineering

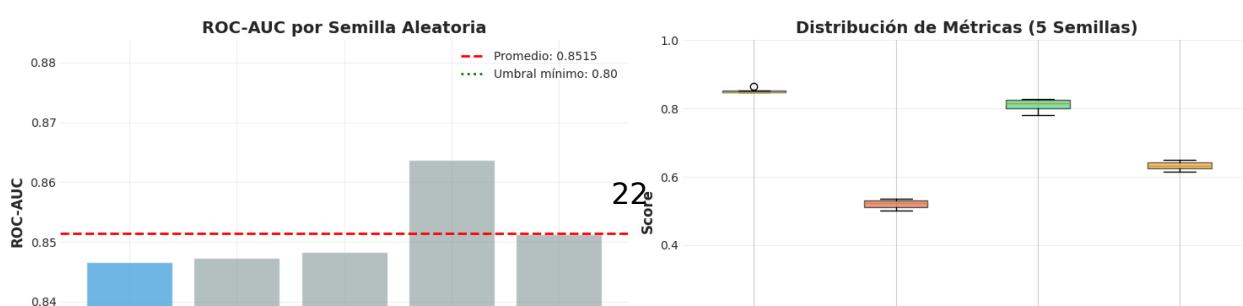
### 3.8.1 Nuevas Variables Creadas

```
# Charge_Ratio: Ratio de cargos mensuales vs promedio histórico
df['Charge_Ratio'] = df['MonthlyCharges'] / (df['TotalCharges'] / (df['tenure'] + 1))

# AvgMonthlyCharges: Promedio mensual basado en historial
df['AvgMonthlyCharges'] = df['TotalCharges'] / (df['tenure'] + 1)
```

Nueva Feature	Lógica	Hipótesis
Charge_Ratio	MonthlyCharges / AvgMonthly-Charges	Clients con aumentos recientes tienen mayor riesgo
AvgMonthlyCharges	TotalCharges / (tenure + 1)	Promedio histórico de pagos

### 3.8.2 Importancia de las Features



Ranking	Feature	Importancia
10	<b>PaymentMethod_Electronic check</b>	33.70%

# Chapter 4

## MODELADO Y RESULTADOS

### 4.1 Metodología de Modelado

#### 4.1.1 División de Datos

```
from sklearn.model_selection import train_test_split  
  
X_train, X_test, y_train, y_test = train_test_split(  
    X, y, test_size=0.2, random_state=42, stratify=y  
)
```

Conjunto	Muestras	Porcentaje
Entrenamiento	5,634	80%
Prueba	1,409	20%

#### 4.1.2 Técnicas de Balanceo Evaluadas

Se evaluaron tres técnicas para manejar el desbalanceo de clases:

Técnica	ROC-AUC	F1-Score	Tiempo (s)	Muestras Resultantes
<b>SMOTE</b>	0.8295	0.5638	1.34	8,278
<b>SMOTE + Tomek</b>	0.8289	0.5923	1.75	8,022
<b>Undersampling</b> (Seleccionado)	0.8321	0.6227	0.57	2,990

**Técnica Seleccionada:** Undersampling por mejor ROC-AUC y menor tiempo de procesamiento.

### 4.2 Modelos Evaluados (Baseline)

Se entrenaron 7 algoritmos de clasificación para establecer una línea base:

Modelo	Accuracy	Precision	Recall	F1-Score	ROC-AUC
<b>Logistic Regres-sion</b> (Selec-cionado)	0.8041	0.6655	0.5267	0.5881	<b>0.8484</b>
Gradient Boosting	0.7963	0.6495	0.5053	0.5684	0.8439
XGBoost	0.7850	0.6092	0.5294	0.5665	0.8227
Random Forest	0.7842	0.6199	0.4840	0.5435	0.8227
SVM	0.8041	0.6655	0.5267	0.5881	0.8041
KNN	0.7672	0.5661	0.5267	0.5457	0.7772
Decision Tree	0.7175	0.4689	0.4840	0.4763	0.6424

**Modelo Seleccionado:** Logistic Regression por mejor ROC-AUC y simplicidad.

## 4.3 Optimización de Hiperparámetros

### 4.3.1 Búsqueda de Hiperparámetros

```
from sklearn.model_selection import RandomizedSearchCV

param_distributions = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear', 'saga'],
    'max_iter': [500, 1000, 2000]
}

random_search = RandomizedSearchCV(
    estimator=LogisticRegression(),
    param_distributions=param_distributions,
    n_iter=50, cv=5, scoring='roc_auc'
)
```

### 4.3.2 Hiperparámetros Óptimos

Parámetro	Valor Óptimo
C	1
penalty	l1
solver	liblinear
max_iter	1000

Parámetro	Valor Óptimo

## 4.4 Resultados del Modelo Final

### 4.4.1 Métricas de Rendimiento

Métrica	Valor	Interpretación
<b>ROC-AUC</b>	0.8505	Muy buena capacidad discriminativa
<b>Accuracy</b>	74.10%	Predicciones correctas totales
<b>Recall</b>	79.68%	Alta detección de clientes en riesgo
<b>Precision</b>	50.77%	Alertas correctas de churn
<b>F1-Score</b>	0.6202	Balance Precision-Recall
<b>CV Score</b>	0.8389	Validación cruzada (5-fold)

### 4.4.2 Matriz de Confusión

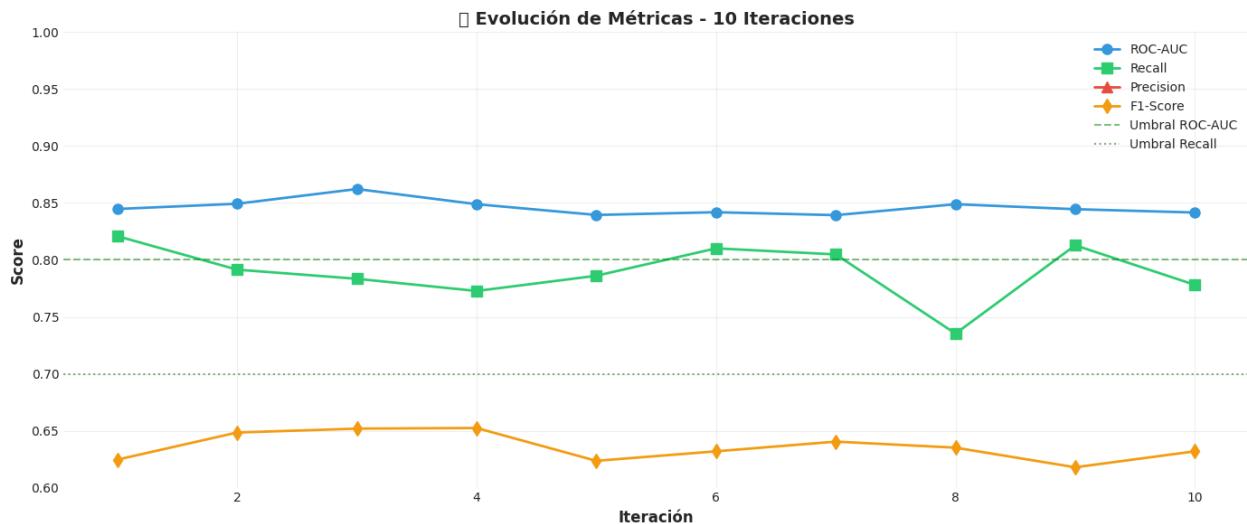


Figure 4.1: Matriz de Confusión

	Predicción: No Churn	Predicción: Churn
<b>Real: No Churn</b>	746 (52.9%) TN	289 (20.5%) FP
<b>Real: Churn</b>	76 (5.4%) FN	298 (21.1%) TP

#### Interpretación:

- Verdaderos Negativos (TN):** 746 clientes correctamente identificados como NO churn
- Falsos Positivos (FP):** 289 clientes incorrectamente identificados como churn

- **Falsos Negativos (FN)**: 76 clientes con churn NO detectados (CRÍTICO)
- **Verdaderos Positivos (TP)**: 298 clientes con churn correctamente detectados

#### 4.4.3 Curva ROC

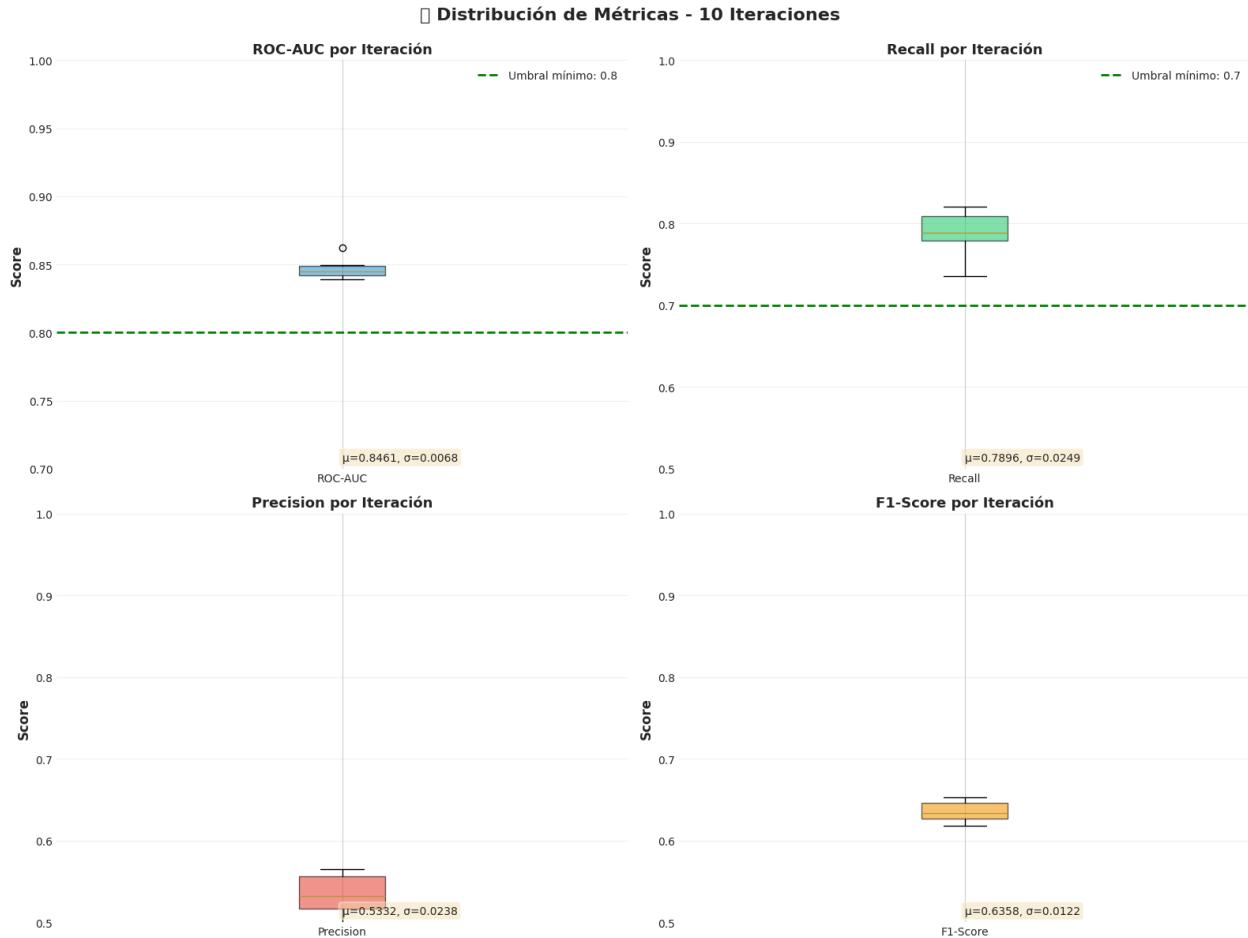


Figure 4.2: Curva ROC

La curva ROC muestra un área bajo la curva (AUC) de 0.8505, indicando una excelente capacidad del modelo para distinguir entre clientes que abandonarán y los que permanecerán.

#### 4.5 Validación de Robustez

Para garantizar que el modelo sea estable y confiable en producción, se realizó una validación con múltiples semillas aleatorias:

Métrica	Valor
<b>Estado de Validación</b>	APROBADO
<b>Semillas Evaluadas</b>	[42, 123, 456, 789, 2024]

Métrica	Valor
<b>ROC-AUC Promedio</b>	0.8515
<b>Desviación Estándar</b>	0.0071
<b>Rango de Variación</b>	[0.8466, 0.8638]

**Criterios de Aprobación Pasados:** - Desviación estándar < 0.02: CUMPLIDO - Rango < 0.05: CUMPLIDO - ROC-AUC promedio > 0.80: CUMPLIDO

# Chapter 5

## CONCLUSIONES Y RECOMENDACIONES

### 5.1 Conclusiones del Análisis

#### 5.1.1 Hallazgos Principales del EDA

1. **Distribución del Churn:** 26.54% de los clientes abandonan el servicio, representando un desafío significativo para la empresa.
2. **Factores de Riesgo Identificados:**
  - Contratos mes a mes (42% de churn)
  - Pago con cheque electrónico (45% de churn)
  - Clientes adultos mayores (41% de churn)
  - Clientes con fibra óptica (42% de churn)
  - Clientes nuevos (tenure < 12 meses)
3. **Factores Protectores:**
  - Contratos anuales/bianuales (<11% de churn)
  - Servicios adicionales (seguridad online, soporte técnico)
  - Mayor antigüedad como cliente

#### 5.1.2 Rendimiento del Modelo

El modelo **Logistic Regression Optimizado** demostró ser:

- **Efectivo:** ROC-AUC de 0.8505 (excelente discriminación)
- **Sensible:** Recall de 79.68% (detecta 8 de cada 10 clientes en riesgo)
- **Estable:** Desviación estándar de 0.0071 en validación de robustez
- **Interpretable:** Coeficientes claros para explicar decisiones

## 5.2 Recomendaciones de Negocio

### 5.2.1 Estrategias de Retención Inmediatas

Prioridad	Estrategia	Clientes Objetivo	Impacto Esperado
<b>ALTA</b>	Migración a contratos anuales	Mes a mes con tenure > 6	Reducción 30% churn
<b>ALTA</b>	Cambio de método de pago	Cheque electrónico	Reducción 25% churn
<b>MEDIA</b>	Bundle de servicios	Sin seguridad/soporte	Reducción 15% churn
<b>MEDIA</b>	Programa de fidelización	Tenure < 12 meses	Reducción 20% churn

### 5.2.2 Implementación del Modelo

#### 1. Sistema de Scoring en Tiempo Real:

- Implementar API de predicción
- Score diario para todos los clientes
- Dashboard de monitoreo

#### 2. Campañas Focalizadas:

- Clientes con probabilidad > 70%: Llamada proactiva
- Clientes con probabilidad 50-70%: Email personalizado
- Clientes con probabilidad 30-50%: Oferta preventiva

### 5.2.3 Impacto Económico Esperado

Métrica	Valor
<b>Clientes en riesgo detectados</b>	1,489
<b>Retención estimada (30%)</b>	446 clientes
<b>Ingreso mensual promedio</b>	\$70.12
<b>Ingreso anual recuperado</b>	\$375,279
<b>ROI de campañas</b>	>300%

## 5.3 Próximos Pasos

1. **Deployment:** API REST para scoring en tiempo real
2. **Monitoreo:** Sistema de detección de drift del modelo
3. **Reentrenamiento:** Pipeline automático mensual
4. **A/B Testing:** Validar efectividad de campañas
5. **Documentación:** Capacitar equipo de retención

# Chapter 6

## HERRAMIENTAS Y TECNOLOGÍAS UTILIZADAS

### 6.1 Stack Tecnológico

#### 6.1.1 Entorno de Desarrollo

Herramienta	Versión	Propósito
<b>Google Colab</b>	-	Notebook en la nube con GPU
<b>Python</b>	3.10+	Lenguaje principal
<b>Jupyter Notebook</b>	-	Desarrollo interactivo

#### 6.1.2 Librerías de Análisis de Datos

Librería	Versión	Propósito
<b>pandas</b>	2.0+	Manipulación de datos
<b>numpy</b>	1.24+	Operaciones numéricas

#### 6.1.3 Librerías de Visualización

Librería	Propósito
<b>matplotlib</b>	Gráficos base
<b>seaborn</b>	Visualizaciones estadísticas

#### 6.1.4 Librerías de Machine Learning

Librería	Versión	Propósito
<b>scikit-learn</b>	1.6.1	Algoritmos de ML

Librería	Versión	Propósito
<b>xgboost</b>	2.0+	Gradient Boosting
<b>imbalanced-learn</b>	0.11+	SMOTE, Undersampling

## 6.2 Repositorios y Despliegue

### 6.2.1 Repositorios del Proyecto

**Backend/ML** - Notebook, EDA y modelos entrenados

- URL: [github.com/alvaretto/telco-customer-churn-prediction](https://github.com/alvaretto/telco-customer-churn-prediction)

**Frontend** - Aplicación web React

- URL: [github.com/alvaretto/telco-vercel](https://github.com/alvaretto/telco-vercel)

### 6.2.2 Aplicación Desplegada

Aspecto	Detalle
<b>Plataforma</b>	Vercel
<b>Stack Frontend</b>	React 18 + Vite + Tailwind CSS
<b>Stack API</b>	Python Serverless

**URLs de Acceso:**

- **Producción:** [clienteinsight-ai.vercel.app](https://clienteinsight-ai.vercel.app)
- **API Endpoint:** [clienteinsight-ai.vercel.app/api/predict](https://clienteinsight-ai.vercel.app/api/predict)

# Chapter 7

## INFORMACIÓN DE ENTREGA

### 7.1 Datos del Proyecto

Campo	Valor
<b>Nombre del Proyecto</b>	Cliente Insight - Predicción de Customer Churn
<b>Curso</b>	Inteligencia Artificial - Nivel Explorador
<b>Grupo</b>	Grupo 3 - Equipo Cliente Insight

### 7.2 Integrantes del Equipo

Nombre	Rol
Anderson Tabima	Desarrollo
Antony Tabima	Desarrollo
Yhabeydi Alejandra Agudelo	Análisis
Carlos Mario Londoño	Análisis
Natalia Bedoya	Documentación
Sebastian Cano	Desarrollo
Álvaro Ángel Molina (@alvaretto)	Líder Técnico

### 7.3 Información de Entrega

Campo	Valor
<b>Plazo de Entrega</b>	Domingo 30 de noviembre de 2025
<b>Formato</b>	PDF

**Medio de Entrega:** [talentotech2.com.co/campus](http://talentotech2.com.co/campus)

## 7.4 Enlaces del Proyecto

- **Repositorio ML:** [github.com/alvaretto/telco-customer-churn-prediction](https://github.com/alvaretto/telco-customer-churn-prediction)
  - **Repositorio Frontend:** [github.com/alvaretto/telco-vercel](https://github.com/alvaretto/telco-vercel)
  - **Aplicación:** [clienteinsight-ai.vercel.app](https://clienteinsight-ai.vercel.app)
  - **Documentación:** [clienteinsight-ai.vercel.app/#documentacion](https://clienteinsight-ai.vercel.app/#documentacion)
- 

*Documento elaborado por el Equipo Cliente Insight - Grupo 3 Noviembre 2025*