



## Proyecto: Telco Customer Churn MAT281

Estudiantes: Oscar Alcántara, Baitiare Corvalán, Carlos Flores,  
Arantza Ormeño, Aranxa Veloz, Diego Wielandt

Profesor: Francisco Alfaro

Departamento de Matemáticas  
Universidad Técnica Federico Santa María

Santiago, 30 de noviembre 2025

# Outline

- 1 Definición del problema
- 2 Análisis exploratorio
- 3 Visualización descriptiva
- 4 Preprocesamiento
- 5 Selección y comparación de modelos
- 6 Evaluación de modelos
- 7 Interpretación del modelo
- 8 Conclusiones y recomendaciones

# Contexto y motivación

- Industria de telecomunicaciones altamente competitiva y con mercados saturados.
- Retener clientes existentes suele ser entre 5 y 7 veces más barato que adquirir nuevos.
- El **churn** corresponde a los clientes que cancelan su contrato con la compañía.
- Anticipar qué clientes tienen alto riesgo de churn permite:
  - Focalizar campañas de retención.
  - Optimizar presupuesto comercial.
  - Mejorar la experiencia del cliente.

# Objetivo del proyecto

- **Objetivo de negocio:**

Diseñar un modelo que estime la **probabilidad de abandono** de cada cliente, para priorizar acciones de retención.

- **Variable objetivo:** Churn

- 1: el cliente abandona la compañía.
- 0: el cliente se mantiene.

- **Pregunta central:**

¿Qué características de los clientes se asocian a mayor riesgo de churn y cómo podemos utilizarlas para predecirlo?

# Outline

- 1 Definición del problema
- 2 Análisis exploratorio
- 3 Visualización descriptiva
- 4 Preprocesamiento
- 5 Selección y comparación de modelos
- 6 Evaluación de modelos
- 7 Interpretación del modelo
- 8 Conclusiones y recomendaciones

# Descripción general del dataset

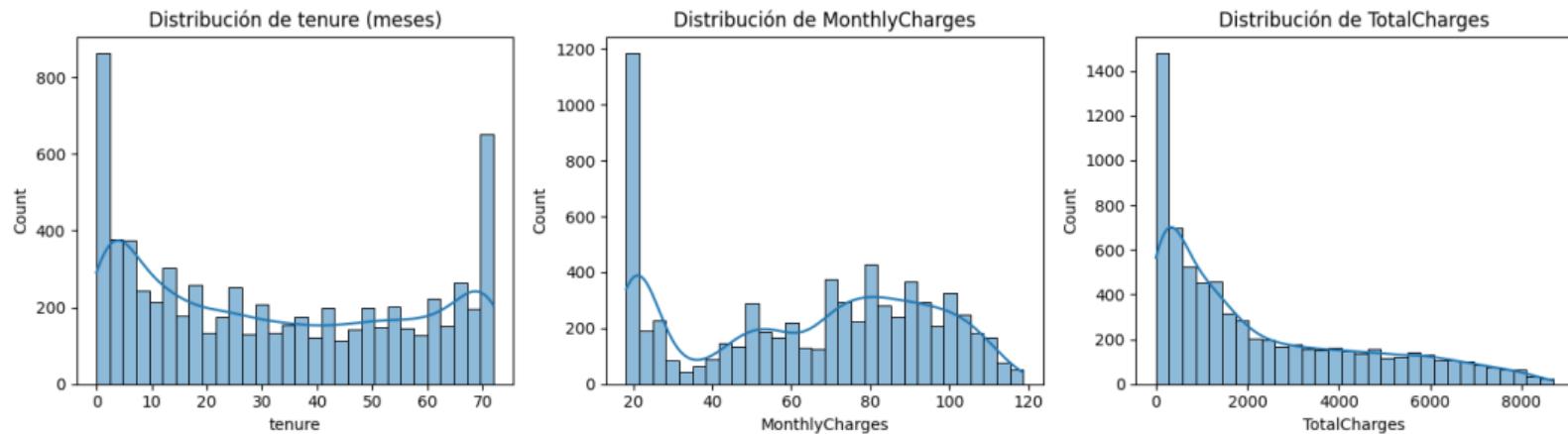
- Dataset **Telco Customer Churn** (Kaggle).
- Cada fila representa un cliente y su estado de churn.
- **Variables numéricas:** tenure, MonthlyCharges, TotalCharges, etc.
- **Variables categóricas:**  
tipo de contrato, servicios de internet, método de pago, servicios adicionales (streaming, seguridad online, etc.).
- **Tasa global de churn:** aproximadamente 26.5% de los clientes abandona el servicio.

## Dataset Telco Customer Churn (Kaggle)

**Tamaño:** 7043 observaciones, 21 columnas.

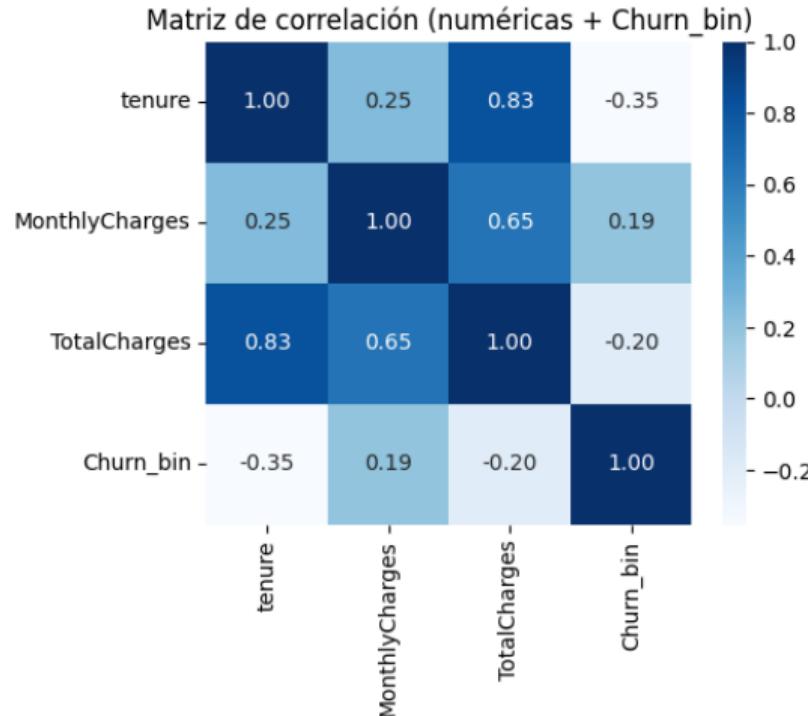
Cada fila representa un cliente y su estado de Churn.

# Distribución de variables numéricas



- La mayoría de los clientes son relativamente nuevos (tenure con valores bajos), aunque existe un grupo con permanencias altas.
- En MonthlyCharges, la distribución refleja distintos tipos de planes y niveles de pago, concentrados aprox. entre 20 y 100 USD.
- TotalCharges presenta una fuerte asimetría positiva: muchos clientes acumulan cargos bajos, consistente con clientes recientes.

# Relaciones entre variables



## Observaciones:

- Se analizó la correlación entre variables numéricas relevantes (tenure, MonthlyCharges, TotalCharges, etc.).
- TotalCharges presenta alta correlación con tenure, lo que es coherente con la acumulación de cargos en el tiempo.
- No se observa colinealidad extrema entre la mayoría de las variables, lo que permite incorporarlas conjuntamente en los modelos.

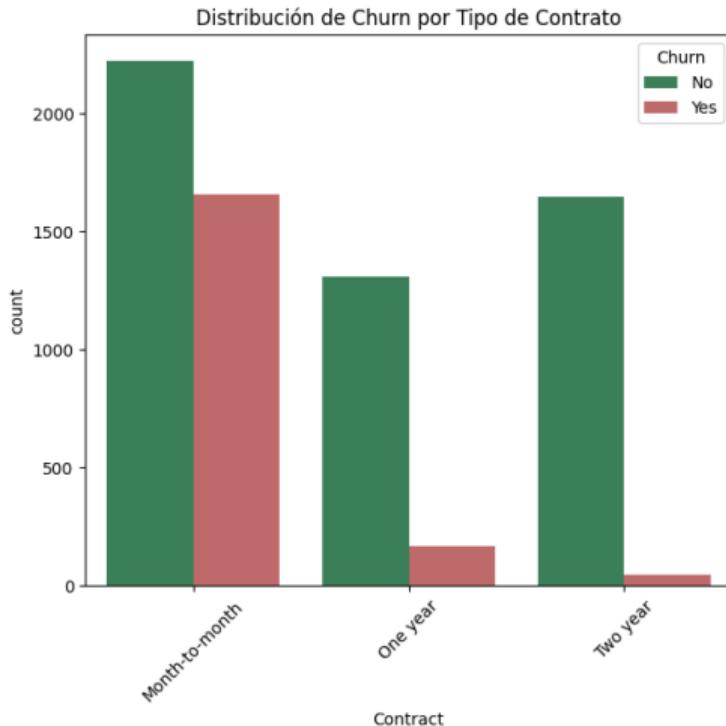
# Outline

- 1 Definición del problema
- 2 Análisis exploratorio
- 3 Visualización descriptiva
- 4 Preprocesamiento
- 5 Selección y comparación de modelos
- 6 Evaluación de modelos
- 7 Interpretación del modelo
- 8 Conclusiones y recomendaciones

# Patrones iniciales observados

- Los clientes con **contrato mensual** concentran gran parte del churn, mientras que los contratos de 1 y 2 años presentan tasas de abandono mucho menores.
- El churn es mayor en clientes con **baja antigüedad** (tenure pequeño), lo que sugiere que el riesgo es más alto en los primeros meses.
- Los **cargos mensuales altos** (MonthlyCharges) se asocian a una mayor probabilidad de abandono, especialmente en planes sin compromiso de largo plazo.
- No se observan diferencias tan marcadas por variables demográficas (por ejemplo, gender o SeniorCitizen), en comparación con variables contractuales y de uso del servicio.

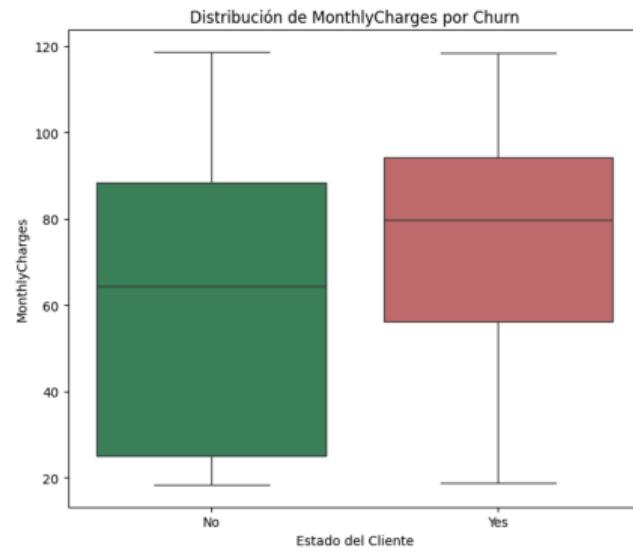
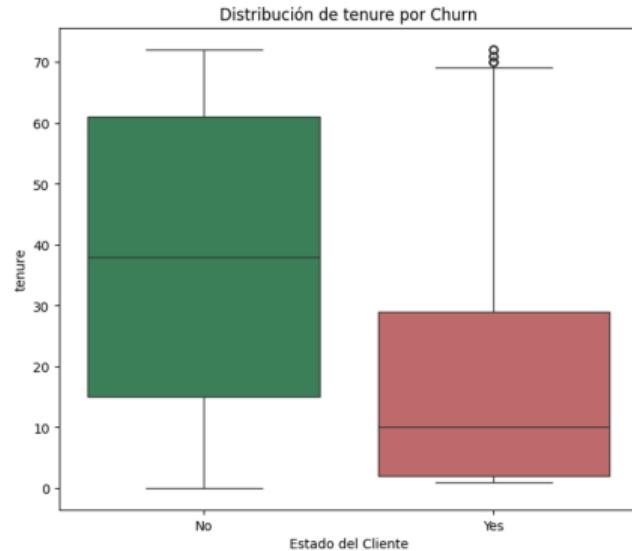
# Churn por tipo de contrato



Obs.:

- Los clientes con **contrato mensual** presentan la mayor proporción de churn.
- Los contratos de 1 y 2 años muestran porcentajes de abandono considerablemente menores.
- Esto sugiere que la **duración del contrato** es una de las variables más influyentes en el riesgo de abandono.

# Tenure y cargos mensuales según churn



- Los clientes que **abandonan** ( $\text{Churn} = \text{Yes}$ ) tienen, en promedio, **menor antigüedad** (tenure) que quienes se quedan.
- Además, pagan **cargos mensuales más altos** (MonthlyCharges).
- Esto sugiere que los **primeros meses** y los **planes más caros** son los segmentos con mayor riesgo de churn.

# Outline

- 1 Definición del problema
- 2 Análisis exploratorio
- 3 Visualización descriptiva
- 4 Preprocesamiento
- 5 Selección y comparación de modelos
- 6 Evaluación de modelos
- 7 Interpretación del modelo
- 8 Conclusiones y recomendaciones

# Limpieza y transformación de datos

- Remoción de columnas no predictivas (`customerID`).
- Conversión de `TotalCharges` a numérico y manejo de nulos.
- Normalización de categorías y binarización de variables (Yes/No).
- Filtrado de **outliers** en `tenure` para reducir distorsiones.

# Codificación y partición del dataset

- Separación en variables explicativas ( $\mathbf{X}$ ) y objetivo ( $y = \text{Churn}$ ).
- Codificación de categóricas con **one-hot encoding**.
- Escalamiento de variables numéricas para modelos sensibles a la escala.
- División del dataset: 75% entrenamiento y 25% prueba (estratificado).

# Desbalance de clases y SMOTE

- Churn = 1 es minoritario  $\Rightarrow$  desbalance de clases.
- Accuracy puede ser engañosa en este escenario.
- Se aplicó **SMOTE** en el set de entrenamiento para:
  - Crear ejemplos sintéticos de la clase minoritaria.
  - Mejorar el aprendizaje del modelo.
- Resultado: clases más balanceadas y mejor Recall/F1 en Churn = 1.

# Outline

- 1 Definición del problema
- 2 Análisis exploratorio
- 3 Visualización descriptiva
- 4 Preprocesamiento
- 5 Selección y comparación de modelos
- 6 Evaluación de modelos
- 7 Interpretación del modelo
- 8 Conclusiones y recomendaciones

# Modelos supervisados considerados

- **Regresión Logística (SMOTE)**: modelo lineal e interpretable; baseline probabilístico.
- **Random Forest (class\_weight=balanced)**: ensamble que captura no linealidades y variables mixtas.
- **SVM RBF (SMOTE)**: modelo de margen máximo con frontera no lineal.
- **KNN (SMOTE)**: clasificador basado en vecinos; baseline simple de comparación.

# Pipelines y ajuste de hiperparámetros

- Preprocesamiento unificado:
  - Numéricas: imputación (mediana) + StandardScaler.
  - Categóricas: imputación (frecuente) + OneHotEncoder.
- **LR, RF y SVM** optimizados con **GridSearchCV** (3-fold, métrica F1).
- Hiperparámetros ajustados:
  - LR:  $C$ .
  - RF: `n_estimators`, `max_depth`, `min_samples_split`.
  - SVM:  $C$  y  $\gamma$ .
- **KNN**: baseline con  $k = 5$ , sin tuning.

# Comparación de modelos en el conjunto de prueba

Modelo	Accuracy	Precisión	Recall	F1-score	ROC-AUC
SVM	0.774	0.556	0.739	0.634	0.826
Random Forest	0.772	0.553	0.737	0.632	0.844
Logistic Regression	0.750	0.519	0.801	0.630	0.844
KNN	0.698	0.457	0.741	0.565	0.774

- El **SVM** obtiene el mejor **F1-score** y la mayor Accuracy, seguido muy de cerca por Random Forest y Regresión Logística.
- Random Forest y Regresión Logística logran el mejor **ROC-AUC** ( 0.84), mostrando buena capacidad discriminatoria.
- **KNN** presenta el peor desempeño y se mantiene solo como baseline.

# Outline

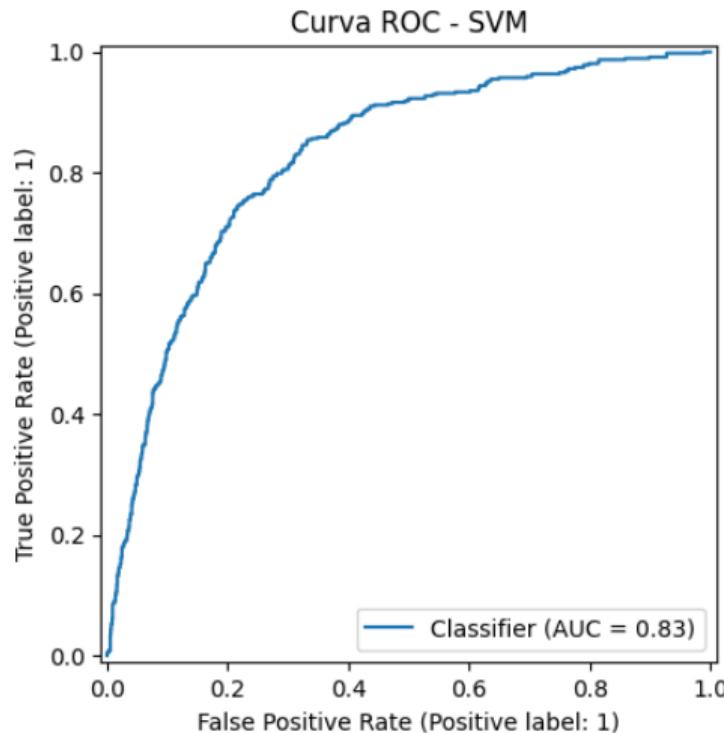
- 1 Definición del problema
- 2 Análisis exploratorio
- 3 Visualización descriptiva
- 4 Preprocesamiento
- 5 Selección y comparación de modelos
- 6 Evaluación de modelos
- 7 Interpretación del modelo
- 8 Conclusiones y recomendaciones

# Métricas en el conjunto de prueba

Modelo	Accuracy	Precisión	Recall	F1-score	ROC-AUC
SVM	0.77	0.56	0.74	0.63	0.83
Random Forest	0.77	0.55	0.74	0.63	0.84
Logistic Regression	0.75	0.52	0.80	0.63	0.84
KNN	0.70	0.46	0.74	0.57	0.77

- Los tres modelos principales (SVM, Random Forest y Regresión Logística) alcanzan **F1-score** y **ROC-AUC** similares.
- **SVM** obtiene el mejor balance entre Accuracy y F1-score, mientras que Random Forest y Regresión Logística logran la mejor ROC-AUC.
- **KNN** queda claramente por debajo y se mantiene sólo como baseline.

# Curva ROC del modelo seleccionado



## Observaciones:

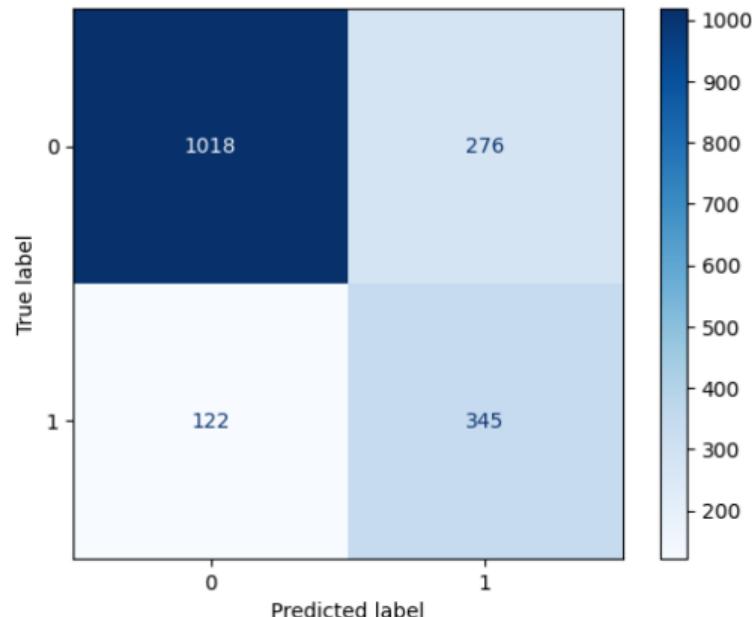
- La curva ROC del modelo final se mantiene claramente por encima de la diagonal (clasificador aleatorio).
- El área bajo la curva (**ROC-AUC**  $\approx$  0.83–0.84) indica una buena capacidad para distinguir entre clientes que harán churn y los que se mantienen.
- Esto permite fijar distintos **umbrales de decisión** dependiendo de cuán agresiva se quiera hacer la campaña de retención.

# Matriz de confusión del modelo final

## Observaciones:

- Los **verdaderos positivos** corresponden a clientes con churn que el modelo identifica correctamente como de alto riesgo.
- Los **falsos negativos** son clientes que hacen churn pero el modelo clasifica como “se queda”: son los casos más costosos para el negocio.
- Los **falsos positivos** representan clientes que no se irían, pero se marcarían como de riesgo (sobrecosto en campañas de retención).
- El modelo logra un compromiso razonable entre capturar churn (**Recall**) y no sobrerreaccionar con demasiados falsos positivos.

Matriz de Confusión - SVM



# Outline

- 1 Definición del problema
- 2 Análisis exploratorio
- 3 Visualización descriptiva
- 4 Preprocesamiento
- 5 Selección y comparación de modelos
- 6 Evaluación de modelos
- 7 Interpretación del modelo
- 8 Conclusiones y recomendaciones

# Factores más influyentes en el churn

- Los modelos entrenados indican que las variables más relevantes son:
  - **Tipo de contrato** (Contract): mensual vs 1–2 años.
  - **Antigüedad** del cliente (tenure).
  - **Cargos mensuales** (MonthlyCharges).
  - **Cargos totales** acumulados (TotalCharges).
  - Tipo de **servicio de internet** y algunos servicios adicionales.
- Estas variables coinciden con los patrones vistos en el análisis exploratorio y en la visualización descriptiva.
- En conjunto, el modelo captura bien la idea de que el churn se concentra en clientes con **contrato flexible, poca antigüedad y planes de mayor costo**.

# Interpretación del modelo y perfiles de clientes

- La tasa de churn es de aproximadamente **26.5%**: cerca de 1 de cada 4 clientes abandona el servicio.
- Las características típicas de los clientes que **churnan** son:
  - **Menor tiempo de permanencia** (tenure bajo).
  - **Cargos mensuales altos** (MonthlyCharges elevados).
  - **Cargos totales bajos**, coherentes con contratos recientes.
  - Con mayor frecuencia, **contrato mensual**.
- En contraste, los clientes de **bajo riesgo** suelen:
  - Tener **antigüedad alta**.
  - Poseer **contratos de 1 o 2 años**.
  - Presentar **cargos mensuales moderados**.
- El modelo seleccionado (SVM con SMOTE) alcanza un **F1-score** cercano a 0.63 y un **ROC-AUC** cercano a 0.83, lo que permite segmentar de forma útil a los clientes según su riesgo de churn.

# Outline

- 1 Definición del problema
- 2 Análisis exploratorio
- 3 Visualización descriptiva
- 4 Preprocesamiento
- 5 Selección y comparación de modelos
- 6 Evaluación de modelos
- 7 Interpretación del modelo
- 8 Conclusiones y recomendaciones

# Conclusiones principales

- El dataset Telco Customer Churn presenta una tasa de abandono de aproximadamente **26.5%**, por lo que el churn es un problema relevante para la compañía.
- El análisis exploratorio y las visualizaciones muestran que el churn se concentra en clientes con:
  - **Contrato mensual**.
  - **Baja antigüedad** (tenure reducido).
  - **Cargos mensuales altos** (MonthlyCharges elevados).
- Los modelos entrenados (LR, Random Forest, SVM, KNN) logran un buen desempeño, destacando el **SVM con SMOTE** como modelo final.
- El modelo seleccionado alcanza un **F1-score** cercano a 0.63 y un **ROC-AUC** cercano a 0.83, lo que permite distinguir razonablemente bien entre clientes de alto y bajo riesgo de churn.

# Recomendaciones y trabajo futuro

- **Campañas focalizadas de retención:**

- Priorizar clientes con contrato mensual, baja antigüedad y cargos mensuales altos.
- Ofrecer descuentos o beneficios para migrar a contratos de mayor plazo.

- **Uso operativo del modelo:**

- Integrar el **score** de churn en los sistemas comerciales para segmentar clientes según riesgo.
- Definir umbrales de probabilidad de churn según el presupuesto disponible para retención.

- **Trabajo futuro:**

- Incorporar nuevas fuentes de información (contacto con soporte, reclamos, encuestas de satisfacción, etc.).
- Re-entrenar y recalibrar periódicamente el modelo a medida que cambian el mercado y el comportamiento de los clientes.



## Proyecto: Telco Customer Churn MAT281

Estudiantes: Oscar Alcántara, Baitiare Corvalán, Carlos Flores,  
Arantza Ormeño, Aranxa Veloz, Diego Wielandt

Profesor: Francisco Alfaro

Departamento de Matemáticas  
Universidad Técnica Federico Santa María

Santiago, 30 de noviembre 2025