1. Introducción

Objetivo del proyecto

- Detectar la probabilidad de "churn" (abandono) de clientes utilizando modelos supervisados.
- Segmentar a los clientes dentro de los grupos de **churn** y **no churn** mediante clustering (K Means).

Contexto del problema

La **retención de clientes** es un punto crucial en cualquier estrategia de negocio orientada a la rentabilidad. Mantener a los clientes que ya forman parte de nuestra base tiene un **costo considerablemente menor** que captar nuevos clientes, lo cual impacta directamente en la eficiencia y sostenibilidad de la compañía.

En este caso específico, los clientes en riesgo de fuga representan **2,862,926.9 dólares**, es decir, un **27% del total de nuestros usuarios**. Este dato evidencia la **prioridad estratégica** de identificar de manera proactiva a aquellos clientes con peligro de abandonar la compañía (**churn**) antes de que lo hagan.

Sin embargo, detectar **quiénes son los clientes con riesgo de fuga** no es suficiente. Es igualmente importante:

1. **Comprender su tipología**: Identificar los diferentes perfiles de clientes con mayor tendencia a abandonar.
2. **Entender las posibles causas de churn**: Factores como la calidad del servicio, el precio, la competencia u otros.

Con esta información, se podrán implementar acciones concretas para **reducir la tasa de abandono**, tales como:

- Mejorar nuestros servicios y la calidad de los productos.
- Desarrollar estrategias específicas de **fidelización** personalizadas según los segmentos identificados.

Metodología aplicada

- Uso de algoritmos de clasificación (Random Forest, Logistic Regression, XGBoost).
- Elección de un **Voting Classifier** para mejorar el equilibrio entre Recall y Precisión, además de crear un modelo más robusto para un target desbalanceado.
- Aplicación de **KMeans** para segmentar los clientes en grupos dentro de los resultados obtenidos.

2. Análisis Exploratorio de Datos

Descripción de los datos

- **Variables de identificación**: customerID.
- **Variables demográficas**: gender, SeniorCitizen, Partner, Dependents.

- **Variables relacionadas con el servicio:** PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.
- **Variables contractuales:** Contract, PaperlessBilling, PaymentMethod.
- **Variables de consumo:** tenure (meses de suscripción), MonthlyCharges, TotalCharges.
- **Variables objetivo:** Churn (Sí/No).

Preprocesamiento de datos

Eliminación de variables irrelevantes:

- Se eliminó la columna customerID por no aportar valor predictivo.

Codificación de variables categóricas:

- Se aplicó One-Hot Encoding a columnas con múltiples categorías: gender, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies y PaymentMethod.
- La columna Contract fue inicialmente codificada ordinalmente, pero posteriormente transformada con One-Hot Encoding para captar mejor la información de los contratos.

Conversión de variables binarias:

- Columnas con valores "Yes/No" (Partner, Dependents, PhoneService, PaperlessBilling, Churn) se convirtieron a True/False y luego a valores numéricos (1/0).

Transformación de columnas numéricas:

- TotalCharges: Convertida a float con manejo de errores.
- Valores nulos resultantes fueron rellenados con la mediana.

Escalado y limpieza:

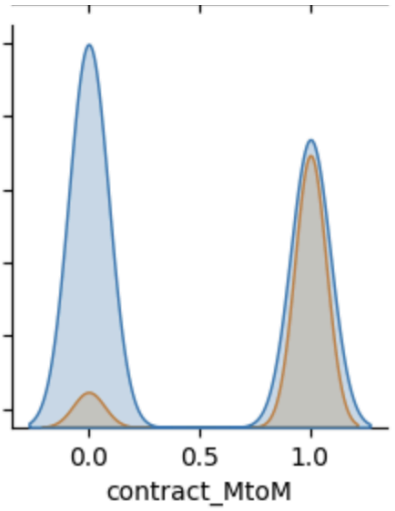
- Las columnas booleanas se transformaron completamente a valores 1 y 0 para asegurar compatibilidad con los modelos.

Visualización inicial

	Feature	Importance
7	TotalCharges	0.182142
6	MonthlyCharges	0.171455
3	tenure	0.157688
21	contract_MtoM	0.064010
10	InternetService_Fiber optic	0.039744
19	PaymentMethod_Electronic check	0.037753
8	gender_Male	0.029074
5	PaperlessBilling	0.027799
12	OnlineSecurity_Yes	0.023841
1	Partner	0.023366
13	OnlineBackup_Yes	0.021972
15	TechSupport_Yes	0.021376
9	MultipleLines_Yes	0.020969
14	DeviceProtection_Yes	0.020382
23	contract_TwoYear	0.020308
0	SeniorCitizen	0.020195
2	Dependents	0.019890
17	StreamingMovies_Yes	0.018411
16	StreamingTV_Yes	0.017725
11	InternetService_No	0.016937
18	PaymentMethod_Credit card (automatic)	0.013818
20	PaymentMethod_Mailed check	0.012432
22	contract_OneYear	0.011480
4	PhoneService	0.007234

- TotalCharges, MonthlyCharges y tenure son las características más relevantes para predecir el churn.
- La variable contract_MtoM destaca como un factor clave.

El gráfico de densidad muestra la **distribución de clientes** con contratos "mes a mes" en relación al churn.



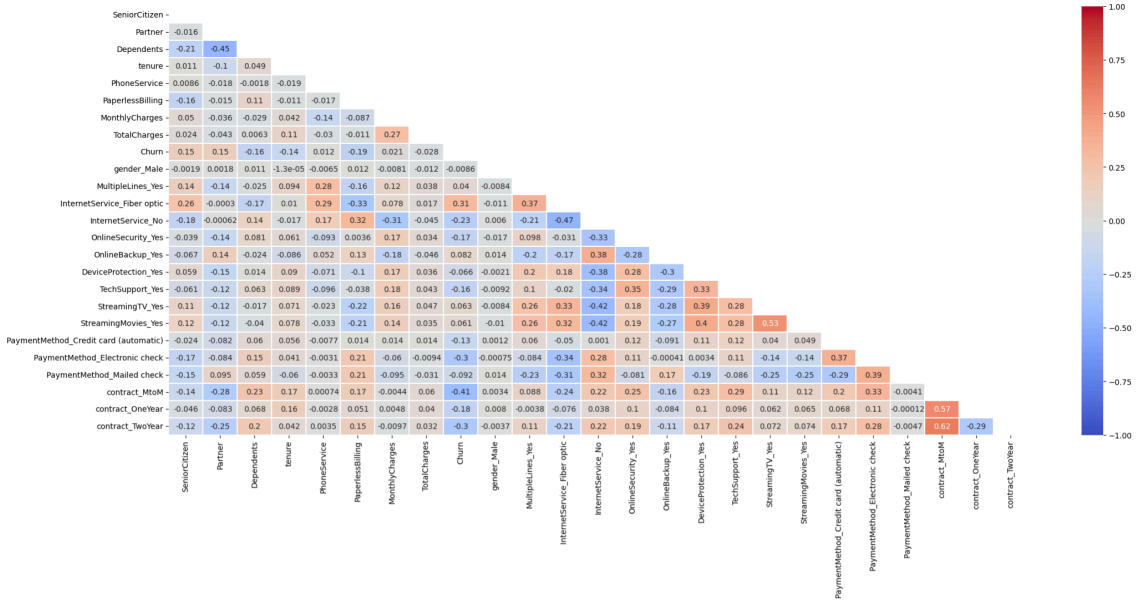
Observaciones clave:

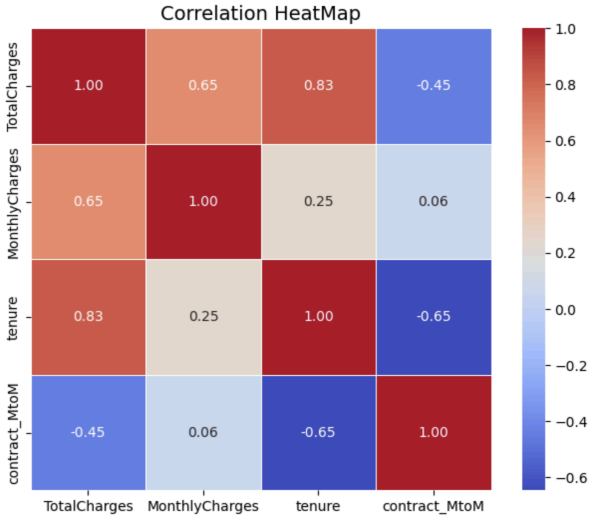
- Existe una concentración notable de **clientes con churn positivo** en contratos de tipo mensual.
- Esta tendencia refuerza la necesidad de analizar este grupo específico para estrategias de retención.

Mapa de correlación

Observaciones clave:

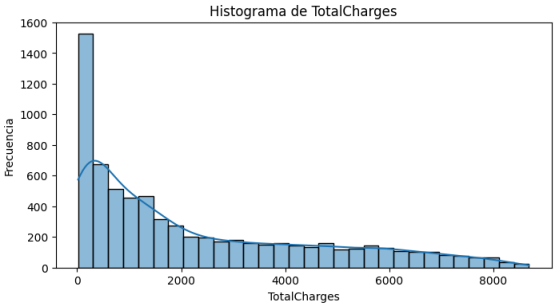
- TotalCharges, MonthlyCharges y tenure muestran alta correlación positiva entre ellas, lo que sugiere redundancia.
- La variable Churn muestra correlaciones negativas con tenure y positivas con contract_MtoM e InternetService_Fiber optic, lo que indica una posible relación con la **rotación de clientes** en contratos más cortos.

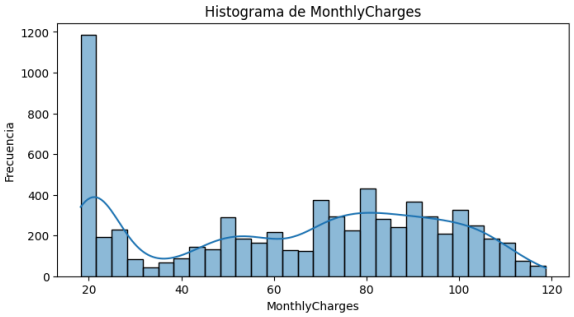


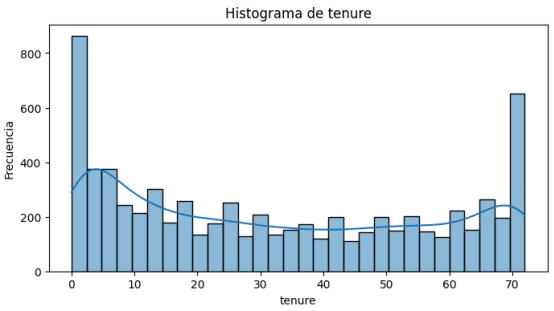


TotalCharges tiene correlación fuerte con tenure y **MonthlyCharges**, pero negativa con **contract_MtoM**, indicando posibles clientes de corta duración con contratos mensuales.

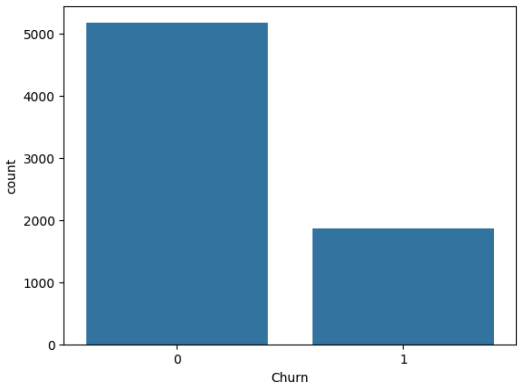
Aparte de esto, desechamos la idea de una posible redundancia entre estas variables.







Debido a la **asimetría** y fuerte sesgo hacia la izquierda, se ha aplicado una **transformación logarítmica** en TotalCharges y MonthlyCharges para normalizar los datos y mejorar su comportamiento en los modelos.



La mayoría de los clientes **no han hecho churn** (clase negativa).
Este desequilibrio deberá considerarse al entrenar los modelos para evitar sesgos predictivos.

3. Construcción del Modelo de Clasificación (Detección de Churn)

Modelos evaluados

- Random Forest
- Logistic Regression
- XG Boost

Selección del modelo final

El motivo principal ha sido encontrar un **equilibrio entre recall y precisión**. Con **Logistic Regression** se obtenía un mejor recall, logrando un balance favorable basado en la métrica F1-Score, mientras que con **XGBoost** se alcanzaba una **mayor precisión**. La combinación de estos modelos mediante un **Voting Classifier** permite aprovechar las fortalezas de cada uno y mejorar el rendimiento global.

El Voting Classifier combina las predicciones de varios modelos y selecciona el resultado final utilizando en este caso las probabilidades de clase **(soft voting)**.

Comparativa mejores resultados

Modelo (best)	**Precisión**	**Recall**	**ROC-AUC**
Random Forest(M5)	(N)0.91 (P)0.50	(N)0.71 (P)0.82	0.7623
Logistic Regression(M5)	(N)0.87 (P)0.51	(N)0.77 (P)0.68	0.7620
XGBoost(M2)	(N)0.84 (P)0.67	(N)0.91 (P)0.52	0.7147
Voting Classifier	**(N)0.88 (P)0.59**	**(N)0.82 (P)0.70**	**0.8507**

Resultados del modelo final con todos los datos

```
Matriz de Confusión:
 [[4349  825]
 [ 554 1315]]

Informe de Clasificación:
              precision    recall  f1-score   support

           0       0.89      0.84      0.86      5174
           1       0.61      0.70      0.66      1869

    accuracy                           0.80      7043
   macro avg       0.75      0.77      0.76      7043
weighted avg       0.81      0.80      0.81      7043


ROC-AUC: 0.8655741149671476
```

4. Segmentación de Clientes con K Means

Objetivo

- Detectar subgrupos dentro de los clientes clasificados como **Churn** y **No Churn**.

Metodología

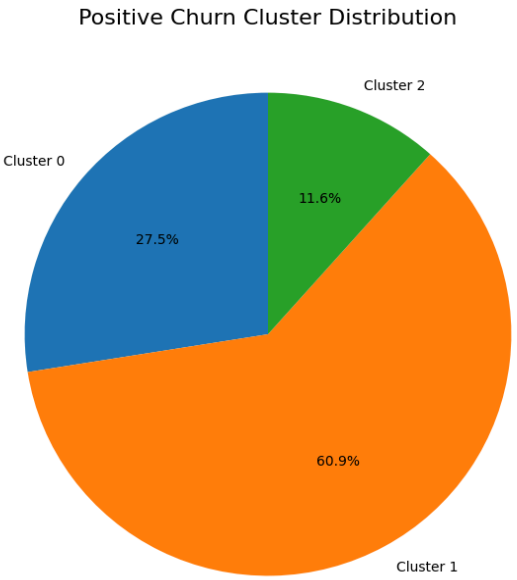
- Aplicación de **K Means** por separado en los grupos de churn y no churn.
- Selección del número óptimo de clusters utilizando el **método Elbow** y el **Silhouette Score**.

Resultados del clustering

Características de los clientes con *churn POSITIVO* (que abandonan)?

Principales características:

- Clientes con **baja antigüedad** y **pocos servicios contratados**.
- Alta proporción de **contratos mensuales** (fácil cancelación).
- Predominio de facturación electrónica y **altos cargos mensuales**.
- Baja adopción de servicios adicionales como TechSupport u OnlineSecurity.
- Mayor uso de métodos de pago **Electronic Check**.



Se identificaron **3 clusters** en los clientes con probabilidad de Churn (positivo):

Cluster 0 (27.5%): Clientes **nuevos,** con bajos cargos mensuales, pocos servicios contratados y **contratos mensuales.**

- **Causa:** Bajo valor percibido.
- **Estrategia:** Ofertas de servicios básicos con promociones limitadas.

Cluster 1 (60.9%): Clientes con **cargos altos,** contratos mensuales y algunos servicios adicionales.

- **Causa:** Precio elevado y falta de personalización.
- **Estrategia:** Planes modulares y descuentos personalizados.

Cluster 2 (11.6%): Clientes de **larga antigüedad** y **altos gastos totales,** con múltiples servicios contratados.

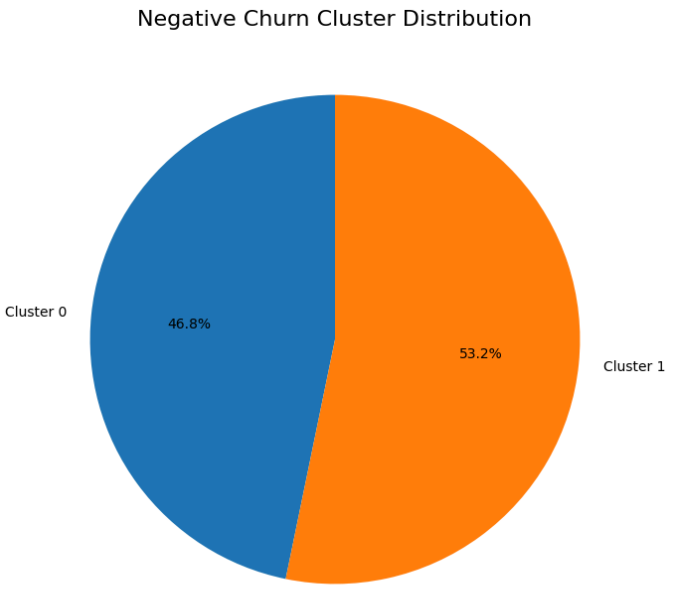
- **Causa:** Falta de engagement activo.
- **Estrategia:** Programas de fidelización y contacto proactivo.

Características de los clientes con *churn NEGATIVO* (leales)

Principales características:

- Mayor proporción de clientes con **contratos a largo plazo** (anuales o bianuales).
- Uso elevado de servicios adicionales como OnlineBackup, TechSupport y StreamingTV.
- **Cargos mensuales más bajos** y mayor satisfacción.
- Predominio del método de pago **Credit Card Automatic.**

- Alta proporción de clientes con **familia o dependientes.**



Se identificaron **2 clusters** en los clientes con probabilidad de Churn (positivo):

- **Cluster 0 (46.8%):** Clientes **leales y de larga antigüedad**, con múltiples servicios contratados, incluidos Internet y fibra.
    - Son clientes con contratos variados (mensuales, anuales, bianuales).

- **Cluster 1 (53.2%):** Clientes más **nuevos** y con menores cargos mensuales, que tienden a no contratar fibra ni servicios adicionales.
    - Segmento más joven/adulto, menos comprometido con servicios adicionales.

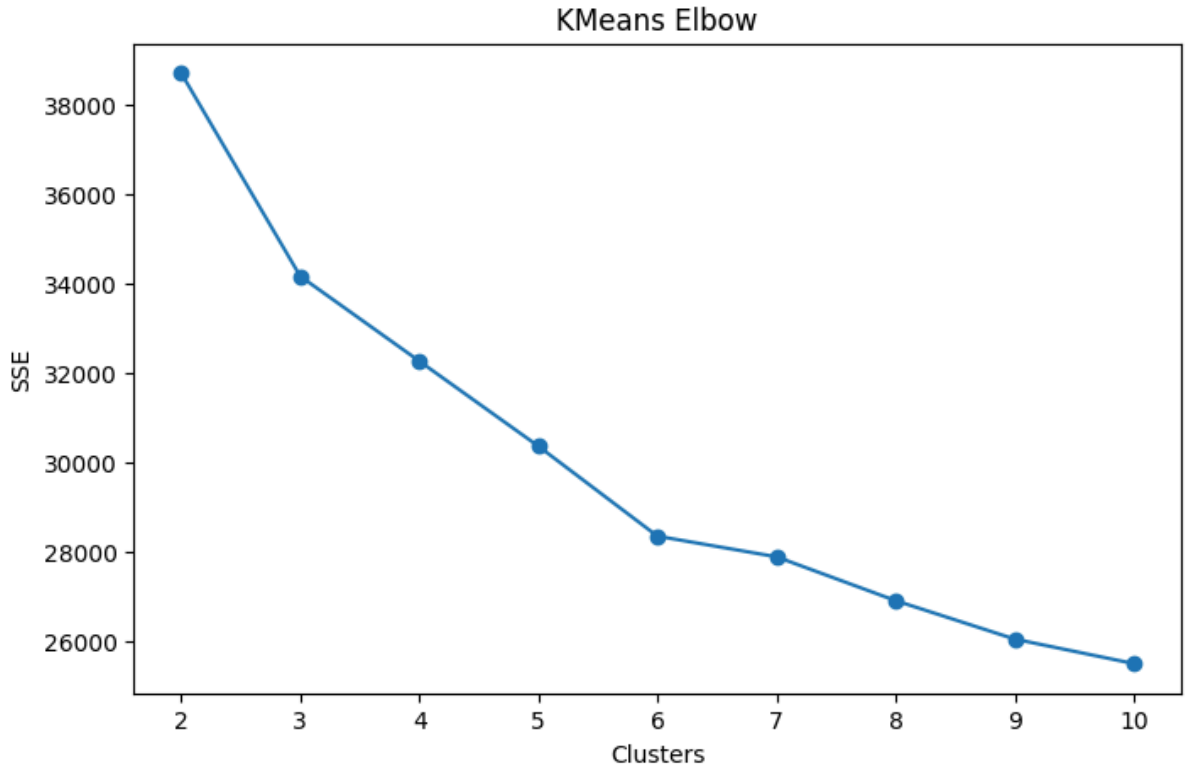
5. Evaluación y Conclusiones

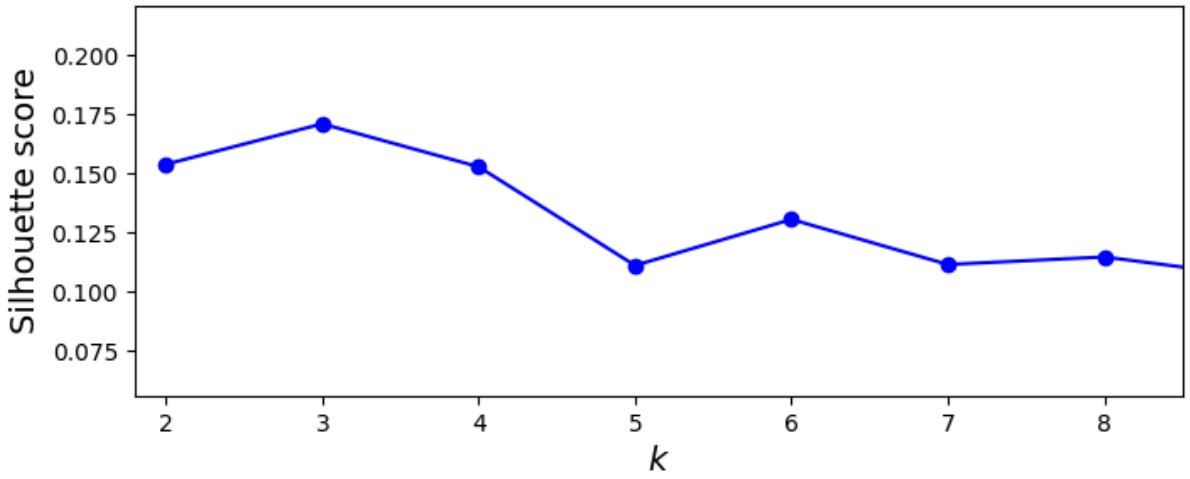
Evaluación del modelo

- En las pruebas realizadas, el **Voting Classifier** demuestra ser un **predictor robusto**, logrando un buen equilibrio para generalizar en datos nuevos. Con un **Train Score de 0.80** y un **Test Score de 0.78**, el modelo ofrece confianza incluso frente al desbalance de clases en el target, asegurando un rendimiento consistente al incorporar datos nuevos.

Evaluación de la segmentación

- Se seleccionaron **3 clusters** para el análisis de **churn POSITIVO**, ya que proporcionan una segmentación clara y generalizable; no obstante, la posibilidad de un cuarto cluster será revisada en el futuro para capturar características adicionales.





- Para el **churn NEGATIVO**, se seleccionaron **2 clusters**, ya que la segmentación ha generalizado correctamente y ofrece interpretaciones claras y coherentes. La elección proporciona confianza en los resultados y permite una identificación precisa de los grupos leales.

Resultados finales

- Muestra la utilidad del proyecto:
    - Capacidad de predecir churn de manera efectiva.

    - Identificación de subgrupos que permiten personalizar estrategias de retención.

Conclusiones clave

Principales hallazgos:

- Clientes con **contratos mensuales,** altos cargos y pocos servicios adicionales tienen mayor riesgo de churn.
- Clientes **leales** suelen tener contratos largos y múltiples servicios.

Beneficios del Voting Classifier:

- Mejor equilibrio entre **recall y precisión,** logrando un rendimiento robusto.

Beneficios del KMeans:

- Segmentación clara que permite estrategias adaptadas a las **necesidades específicas** de cada grupo.