

1 Your note was saved successfully. [View in Keep](#)



Close

## Informe Técnico Churn

### EDA

EDA.Churn.ipynb

Se realizó un preprocesamiento exhaustivo de los datos para eliminar valores perdidos, verificar valores nulos, imputaciones de valores faltantes, codificar variables categóricas y escalar las variables numéricas.

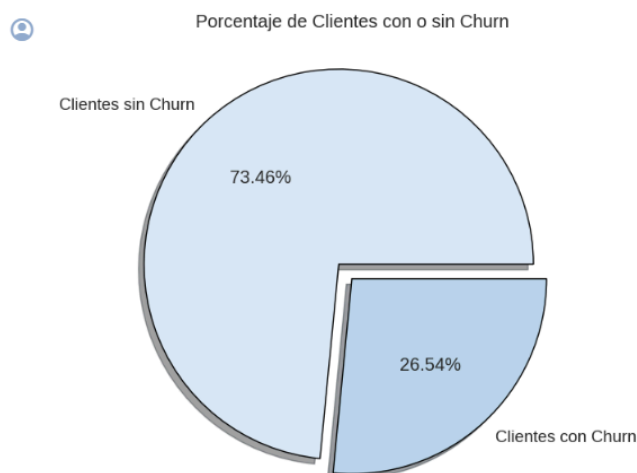
#	Column	Non-Null Count	Dtype
0	customerID	7043 non-null	object
1	gender	7043 non-null	object
2	SeniorCitizen	7043 non-null	int64
3	Partner	7043 non-null	object
4	Dependents	7043 non-null	object
5	tenure	7043 non-null	int64
6	PhoneService	7043 non-null	object
7	MultipleLines	7043 non-null	object
8	InternetService	7043 non-null	object
9	OnlineSecurity	7043 non-null	object
10	OnlineBackup	7043 non-null	object
11	DeviceProtection	7043 non-null	object
12	TechSupport	7043 non-null	object
13	StreamingTV	7043 non-null	object
14	StreamingMovies	7043 non-null	object
15	Contract	7043 non-null	object
16	PaperlessBilling	7043 non-null	object
17	PaymentMethod	7043 non-null	object
18	MonthlyCharges	7043 non-null	float64
19	TotalCharges	7043 non-null	float64
20	numAdminTickets	7043 non-null	int64
21	numTechTickets	7043 non-null	int64
22	Churn	7043 non-null	object

Observamos que el set de datos tiene 7043 variables no nulas

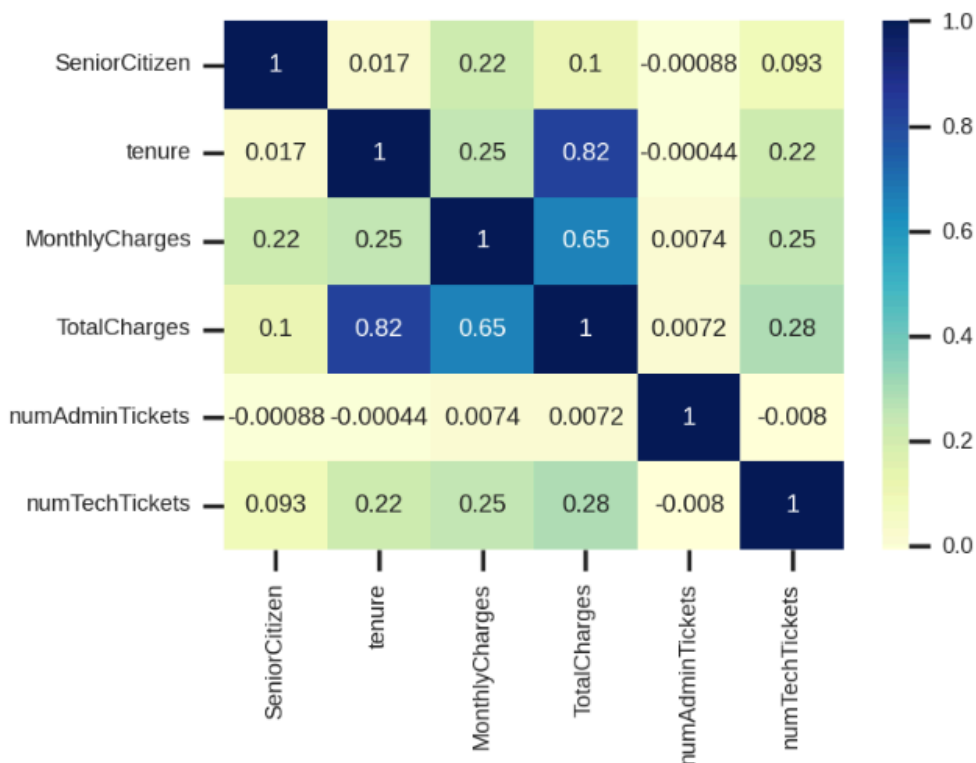
```
# Verificar valores nulos  
print(df.isnull().sum())
```

customerID	0
gender	0
SeniorCitizen	0
Partner	0
Dependents	0
tenure	0
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	0
numAdminTickets	0
numTechTickets	0
Churn	0
dtype: int64	

Nos interesa saber la tasa de abandono en Pandas, ya que el análisis de BI ya fue realizado en Power BI.



Vamos a ver las relaciones entre las variables



**tenure y TotalCharges:** Correlación positiva fuerte (0.82), indica que a mayor antigüedad del cliente, mayor es el monto total de los cargos a lo largo del tiempo.

**MonthlyCharges y TotalCharges:** Correlación positiva fuerte (0.65), indica que a mayor monto de cargos mensuales, mayor es el monto total de los cargos a lo largo del tiempo.

Se observa una correlación positiva entre la antigüedad del cliente y los cargos mensuales y totales, lo que sugiere que los clientes que permanecen más tiempo en la empresa tienden a gastar más.

Se procede a verificar cómo actúa AutoML con datos desbalanceados

churn\_pycaret.ipynb

AUC falla, queda en 0

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
knn	K Neighbors Classifier	0.7717	0.0000	0.7717	0.7791	0.7747	0.4309	0.4321	0.6310
lda	Linear Discriminant Analysis	0.7620	0.6676	0.7620	0.6658	0.6987	0.2266	0.2382	0.5870
lr	Logistic Regression	0.7370	0.9210	0.7370	0.7798	0.6277	0.0127	0.0741	1.0430

Se procede a balancear los datos

## Desbalanceo de datos en modelos de clasificación: Un problema y sus soluciones

En el ámbito del aprendizaje automático, el **desbalanceo de clases** es un problema común que surge cuando la distribución de las clases en un conjunto de datos no es uniforme. Esto significa que una clase (clase mayoritaria) tiene un número significativamente mayor de ejemplos que la otra clase (clase minoritaria). Este desequilibrio puede afectar negativamente el rendimiento de los modelos de clasificación, especialmente cuando se trata de clasificar la clase minoritaria.

### Impacto del desbalanceo de datos:

El desbalanceo de datos puede conducir a diversos problemas en la clasificación:

- **Sesgo del modelo:** Los modelos de clasificación tienden a favorecer la clase mayoritaria durante el entrenamiento, lo que puede llevar a una baja precisión en la clasificación de la clase minoritaria.
- **Baja sensibilidad:** La sensibilidad, que mide la capacidad del modelo para identificar correctamente los ejemplos de la clase minoritaria, se ve afectada negativamente por el desbalanceo.
- **Baja especificidad:** La especificidad, que mide la capacidad del modelo para identificar correctamente los ejemplos que no pertenecen a la clase minoritaria, también puede verse afectada.

### Estrategias para abordar el desbalanceo de datos:

Existen diversas estrategias para abordar el problema del desbalanceo de datos en la clasificación:

#### 1. Técnicas de muestreo:

- **Sobremuestreo:** Esta técnica consiste en aumentar el número de ejemplos de la clase minoritaria. Se pueden utilizar técnicas como **SMOTE** (Synthetic Minority Oversampling Technique) o **ADASYN** (Adaptive Synthetic Sampling Approach) para generar nuevos ejemplos de la clase minoritaria de manera sintética.
- **Submuestreo:** Esta técnica consiste en reducir el número de ejemplos de la clase mayoritaria. Se puede realizar un muestreo aleatorio o un muestreo estratificado para mantener la distribución de las características dentro de la clase mayoritaria.

## 2. Penalización del error:

Se pueden utilizar algoritmos de clasificación que penalizan los errores en la clase minoritaria con mayor severidad. Esto se puede lograr ajustando los pesos de las clases o utilizando funciones de costo personalizadas.

## 3. Selección de algoritmos:

Algunos algoritmos de clasificación son más sensibles al desbalanceo de datos que otros. Por ejemplo, los algoritmos basados en árboles de decisión pueden ser más propensos a sesgarse hacia la clase mayoritaria, mientras que los algoritmos de K-Nearest Neighbors (KNN) pueden ser más robustos al desbalanceo.

## 4. Enfoque en la clase minoritaria:

En algunos casos, puede ser más importante enfocarse en la clasificación precisa de la clase minoritaria, incluso si esto significa sacrificar un poco la precisión en la clasificación de la clase mayoritaria. En este caso, se pueden utilizar métricas de evaluación como la **precisión por clase** o el **área bajo la curva ROC (AUC)** para evaluar el rendimiento del modelo.

El desbalanceo de datos es un problema importante en la clasificación que puede afectar negativamente el rendimiento de los modelos. Existen diversas estrategias para abordar este problema, como el sobremuestreo, el submuestreo, la penalización del error, la selección de algoritmos y el enfoque en la clase minoritaria. La elección de la estrategia adecuada depende del conjunto de datos específico y la tarea de clasificación.

**Es importante tener en cuenta que el desbalanceo de datos es solo uno de los factores que pueden afectar el rendimiento de un modelo de clasificación. Es importante considerar otros aspectos como la calidad de los datos, la elección del algoritmo de clasificación y la optimización de los hiperparámetros para obtener un modelo efectivo.**

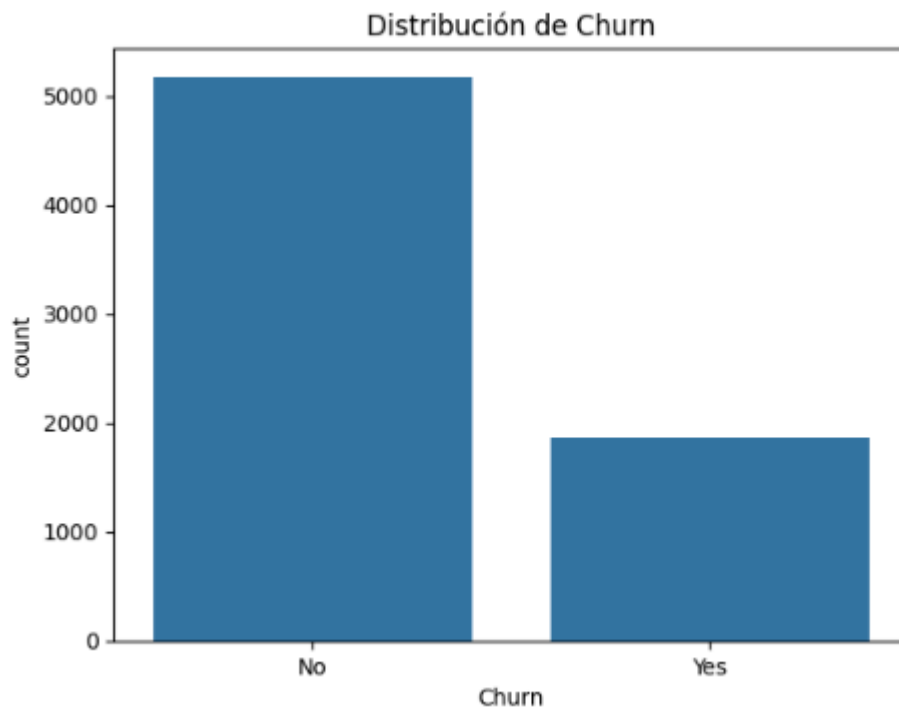
## Balanceo de Datos

 [ChurnV3\\_desbalanceo.ipynb](#)

## Técnicas de sobremuestreo:

Dado que el conjunto de datos presentaba un desequilibrio de clases (más clientes que no abandonan que clientes que abandonan), se aplicaron técnicas de

sobremuestreo como SMOTE, ADASYN y SMOTEENN para balancear las clases antes del entrenamiento de los modelos.



```
# Verificar el desbalanceo de clases
class_counts = df['Churn'].value_counts()
print("Recuento de cada clase:")
print(class_counts)
```

```
Recuento de cada clase:
Churn
No      5174
Yes     1869
Name: count, dtype: int64
```

## Aplicación de Smote

```
Después de aplicar SMOTE:
Churn
No      5174
Yes     5174
```

Se obtienen las siguientes métricas

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
<b>et</b>	Extra Trees Classifier	0.9008	0.9589	0.9008	0.9008	0.9008	0.8016	0.8017	1.4520
<b>lightgbm</b>	Light Gradient Boosting Machine	0.9008	0.9710	0.9008	0.9009	0.9008	0.8016	0.8018	0.9780

Se observa una gran mejora en precisión y AUC

## Aplicación de Adasyn

```
Después de aplicar ADASYN:
Churn
Yes    5228
No     5174
```

El resultado es el siguiente

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
<b>xgboost</b>	Extreme Gradient Boosting	0.8885	0.9598	0.8885	0.8891	0.8884	0.7769	0.7776	0.3910

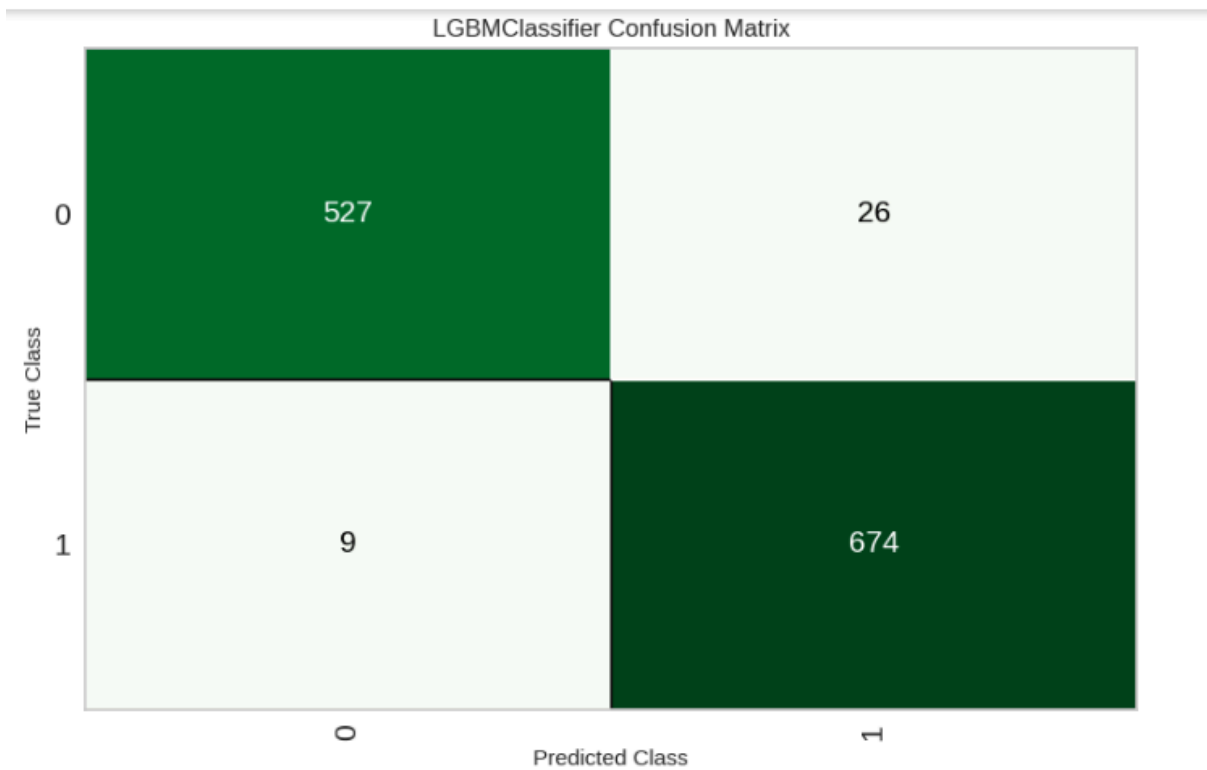
## Aplicación de Smoteen

```
Después de aplicar SMOTEENN:
Churn
Yes    3412
No     2765
```

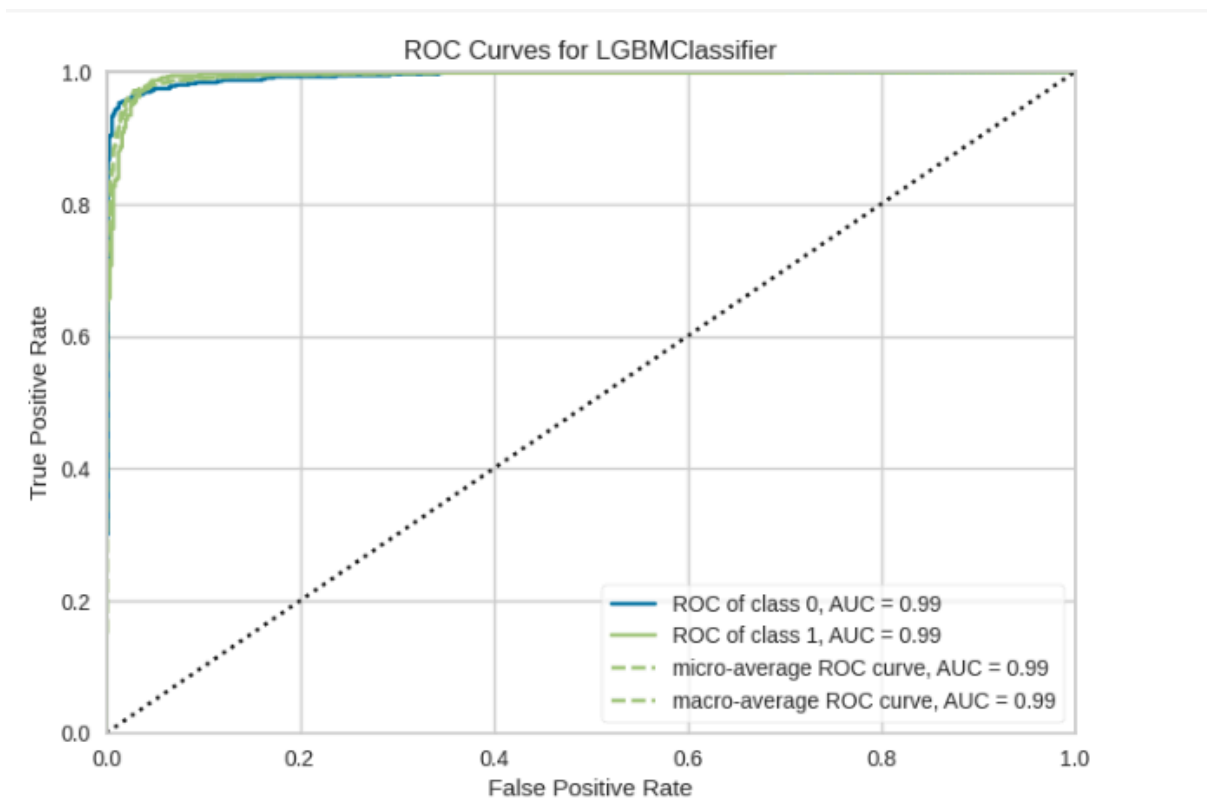
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
<b>lightgbm</b>	Light Gradient Boosting Machine	0.9690	0.9943	0.9690	0.9691	0.9690	0.9373	0.9374	1.0220
<b>xgboost</b>	Extreme Gradient Boosting	0.9688	0.9945	0.9688	0.9689	0.9688	0.9369	0.9370	0.3400

Se verifica un gran rendimiento en Accuracy y AUC con el ensamble LightGBM

Se obtiene la siguiente matriz de confusión



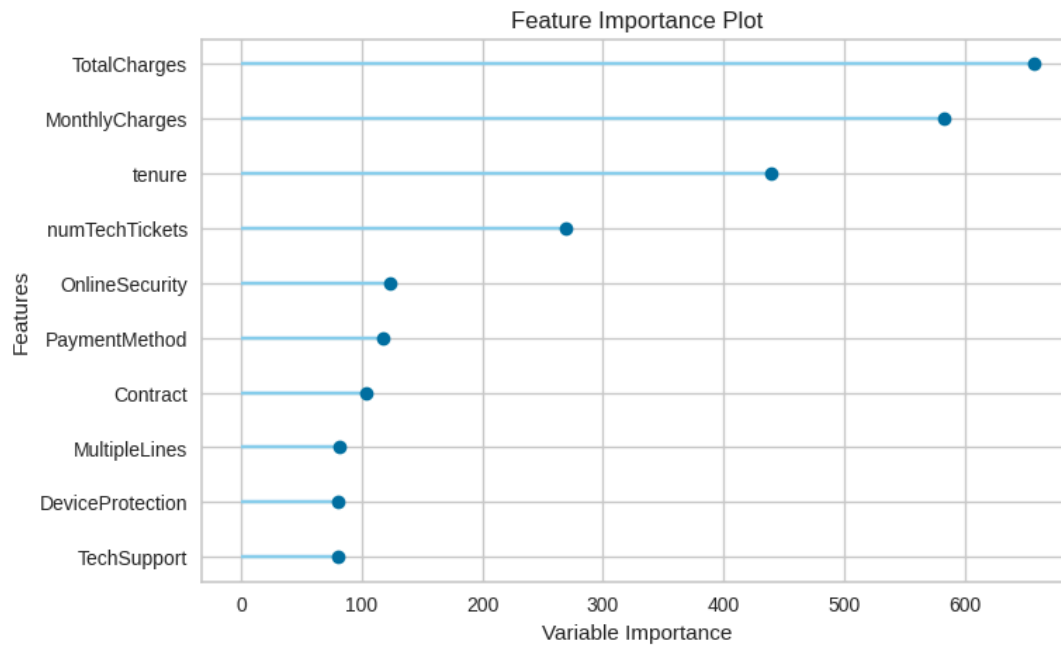
Curva ROC



No se va a buscar mejorar el desempeño por las métricas alcanzadas.

Importancia de características





## Predicción del Churn con Modelos de Aprendizaje Automático

### Introducción:

El churn, o abandono de clientes, es una métrica esencial para las empresas que buscan mejorar su retención y aumentar sus ingresos. Los modelos de aprendizaje automático (ML) ofrecen una herramienta poderosa para predecir el churn y tomar medidas preventivas para evitar la pérdida de clientes.

### Objetivo:

Este informe técnico tiene como objetivo evaluar el rendimiento de diferentes modelos de ML para la predicción del churn en un conjunto de datos específico. Se compararán los modelos en base a métricas como precisión, AUC, recall, F1-score, kappa y MCC.

### Metodología:

#### 1. Conjunto de datos:

Se utilizó un conjunto de datos con las siguientes variables:

- customerID
- gender
- SeniorCitizen
- Partner
- Dependents
- tenure

- PhoneService
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies
- Contract
- PaperlessBilling
- PaymentMethod
- MonthlyCharges
- TotalCharges
- numAdminTickets
- numTechTickets
- Churn

Contiene información sobre clientes y su historial de interacción con la empresa. El conjunto de datos se dividió en conjuntos de entrenamiento y prueba.

## **2. Modelos de aprendizaje automático:**

Se entrenaron y evaluaron cuatro modelos de ML:

- Extra Trees Classifier
- Light Gradient Boosting Machine (LGBM)
- XGBoost
- Random Forest Classifier

## **3. Evaluación del rendimiento:**

Se evaluó el rendimiento de los modelos en base a las siguientes métricas:

- Precisión: Proporción de predicciones correctas.
- AUC: Área bajo la curva ROC.
- Recall: Proporción de casos positivos correctamente identificados.
- F1-score: Media armónica entre precisión y recall.
- Kappa: Medida de acuerdo entre las predicciones del modelo y la clasificación real.
- MCC: Coeficiente de correlación de Matthews.

## **SMOTE**

```

) from pycaret.classification import *

# Configurar el entorno de PyCaret para clasificación VC=5, Split0.8, Balanceo
clf = setup(data=ros_dataset, target='Churn', session_id=123, log_experiment=True,
            experiment_name='Churn', fix_imbalance=True, normalize=True,
            normalize_method='zscore', train_size=0.8, fold=5)

```

Se indica balanceo de datos , que normalice los datos y que los normalice con zscore, con un split de 0.8 y una validación cruzada de 5

```

# Comparación modelos
best = compare_models()

```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
<b>et</b>	Extra Trees Classifier	0.9008	0.9589	0.9008	0.9008	0.9008	0.8016	0.8017	1.4520
<b>lightgbm</b>	Light Gradient Boosting Machine	0.9008	0.9710	0.9008	0.9009	0.9008	0.8016	0.8018	0.9780

## ADASYN

```

#Comparación de modelos
best = compare_models()

```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
<b>xgboost</b>	Extreme Gradient Boosting	0.8885	0.9598	0.8885	0.8891	0.8884	0.7769	0.7776	0.3910

## SMOTEENN

```

#Comparación de modelos
best = compare_models()

```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
<b>lightgbm</b>	Light Gradient Boosting Machine	0.9690	0.9943	0.9690	0.9691	0.9690	0.9373	0.9374	1.0220
<b>xgboost</b>	Extreme Gradient Boosting	0.9688	0.9945	0.9688	0.9689	0.9688	0.9369	0.9370	0.3400

## Resultados:

- La mayoría de los modelos alcanzaron un buen rendimiento, con una precisión cercana al 90% y un AUC superior al 95% después de aplicar técnicas de muestreo.
- XGBoost y LGBM, en conjunto con SMOTEENN, obtienen los mejores resultados, con una precisión superior al 96% y un AUC superior al 99%.
- Las variables más importantes para predecir el churn son los cargos totales, los cargos mensuales y la antigüedad.


## Conclusiones:

- Los modelos de ML son herramientas efectivas para predecir el churn.
- **XGBoost y LGBM, combinados con SMOTEENN, son los modelos más adecuados para esta tarea.**
- Las variables relacionadas con los costos y la antigüedad del cliente son importantes para la predicción del churn.

## Recomendaciones:

- Se sugiere implementar el modelo de mejor rendimiento (XGBoost o LGBM con SMOTEENN) en un sistema de producción para predecir el churn y tomar medidas preventivas.
- Se recomienda realizar un análisis más profundo de las variables que influyen en el churn para identificar oportunidades de mejora en la atención al cliente y la retención de clientes.

## Análisis de clusters

 Pycaret cluster churn.ipynb

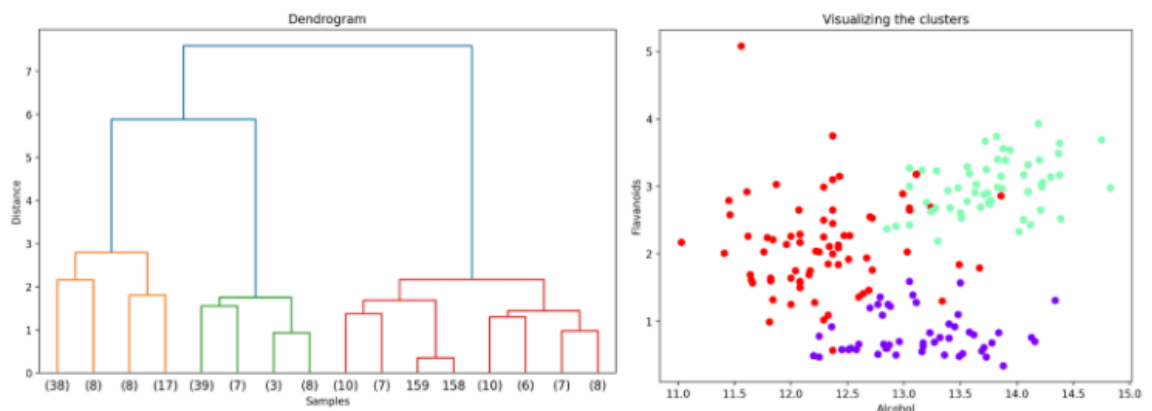
## Clustering en Aprendizaje Automático

El clustering, también conocido como agrupamiento, es una técnica de aprendizaje automático **no supervisado** que se utiliza para agrupar datos en función de su similitud. El objetivo del clustering es identificar grupos naturales o clusters de datos sin conocer previamente las etiquetas o clases a las que pertenecen los datos.

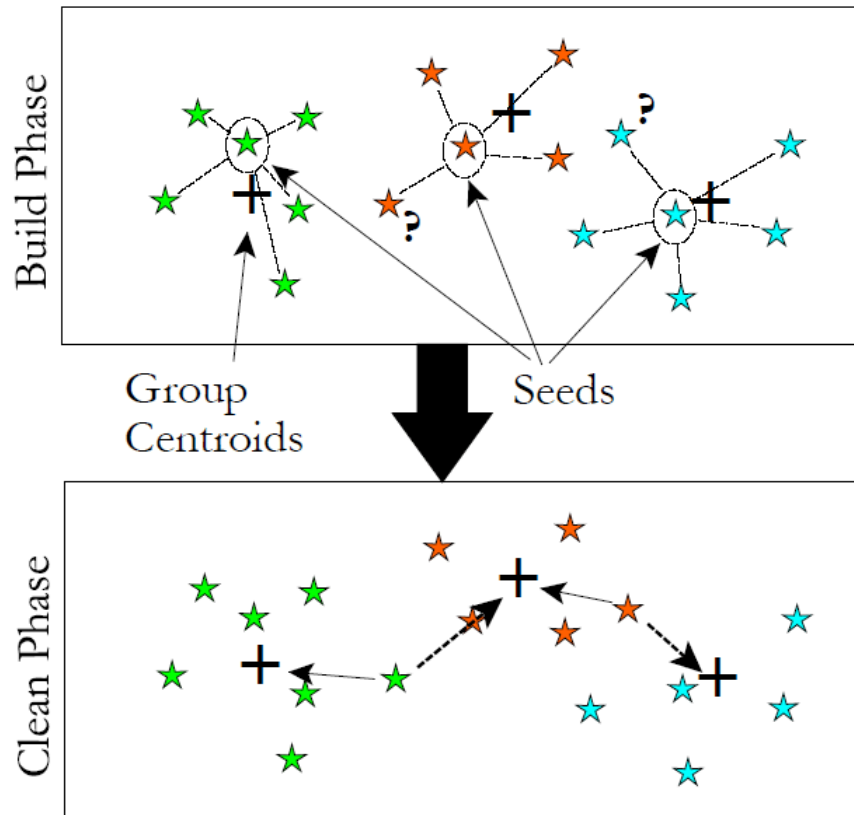
## Tipos de Clustering

Existen dos tipos principales de clustering:

- **Clustering jerárquico:** Construye una jerarquía de clusters, donde cada cluster en un nivel es un subconjunto de los clusters del nivel anterior.



- **Clustering basado en particiones:** Divide los datos en un conjunto fijo de clusters sin jerarquía.



## Algoritmos de Clustering

Existen diversos algoritmos de clustering, cada uno con sus propias fortalezas y debilidades. Algunos de los algoritmos más comunes incluyen:

- **K-means:** Un algoritmo de clustering basado en particiones que asigna cada punto de datos al cluster más cercano en términos de **distancia euclidiana**.
- **Clustering de densidad local (DBSCAN):** Un algoritmo de clustering basado en la densidad que identifica clusters como regiones de alta densidad de puntos de datos, separados por regiones de baja densidad.
- **Clustering de mezcla gaussiana:** Un algoritmo de clustering probabilístico que asume que los datos provienen de una mezcla de distribuciones gaussianas, y luego identifica los clusters como las componentes de la mezcla.

## Evaluación del Clustering

Evaluar la calidad de los clusters generados es un desafío en el clustering no supervisado. No existe una métrica única que sea universalmente aplicable, pero algunas medidas comunes incluyen:

- **Silhouette score:** Mide la cohesión dentro de los clusters y la separación entre clusters.
- **Calinski-Harabasz index:** Evalúa la compacidad de los clusters y la separación entre ellos.
- **Davies-Bouldin index:** Mide la dispersión dentro de los clusters y la distancia entre ellos.

## **Selección del Algoritmo de Clustering**

La elección del algoritmo de clustering adecuado depende de las características específicas de los datos y la tarea en cuestión. Se deben considerar factores como la distribución de los datos, la forma de los clusters y la presencia de ruido.

## **Aplicaciones del Clustering**

El clustering tiene una amplia gama de aplicaciones en diversos campos, incluyendo:

- **Análisis de marketing:** Segmentar clientes en función de sus características de compra o comportamiento.
- **Análisis de redes sociales:** Identificar comunidades o grupos de usuarios en redes sociales.
- **Bioinformática:** Agrupar genes o proteínas en función de su expresión o estructura.
- **Procesamiento de imágenes:** Segmentar imágenes en regiones con características similares.

## **Consideraciones adicionales**

- **Preprocesamiento de datos:** Es importante preprocesar los datos antes de aplicar el clustering, como normalización o eliminación de outliers.
- **Evaluación del clustering:** No existe una medida única para evaluar la calidad del clustering. Se pueden utilizar diferentes métricas, como la distancia entre clusters o la cohesión dentro de los clusters.
- **Interpretación de los clusters:** Una vez que se han identificado los clusters, es importante interpretarlos para comprender su significado y utilidad.

El clustering es una técnica fundamental en el aprendizaje automático que permite descubrir patrones y relaciones ocultas en conjuntos de datos sin etiquetas. La elección del algoritmo de clustering adecuado, la evaluación de la calidad de los

clusters y la aplicación de técnicas de preprocesamiento y selección de características son aspectos cruciales para obtener resultados exitosos. El clustering tiene una amplia gama de aplicaciones en diversos campos, desde el análisis de mercado hasta la bioinformática, y sigue siendo una herramienta valiosa para la exploración y el análisis de datos.

## Aspectos Técnicos del Código

Decidimos explorar los efectos del cluster en la segmentación de clientes.

Cómo partimos de un aprendizaje supervisado para ver la posibilidad de abandono, debemos eliminar la variable Churn

```
# Eliminar la variable 'Churn' del dataset
df.drop(columns=['Churn'], inplace=True)
```

Para que no haya problemas con la implementación del algoritmo, borramos Customer ID e imputamos los cargos con 0, ya que pueden no existir, por ser primera factura.

```
# Borrar columna Customer ID
df.drop(columns=['customerID'], inplace=True)
```

```
# Imputar valores faltantes en 'TotalCharges' con 0
df['TotalCharges'].fillna(0, inplace=True)
```

Se hace un preprocesamiento con Label Encoder y Standar Scaler

```
# Seleccionar solo las características numéricas para estandarizar
numeric_features = df.select_dtypes(include=['int64', 'float64']).columns

# Estandarizar características numéricas
scaler = StandardScaler()
df[numeric_features] = scaler.fit_transform(df[numeric_features])
```

De la observación del set de datos, consideramos adecuado empezar con 3 clusters

```
from sklearn.cluster import KMeans
# Definir el número de clústeres
n_clusters = 3
# Crear una instancia del modelo KMeans
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
```

Obtuvimos las siguientes conclusiones

## Interpretaciones

**Clúster 0:** Este grupo parece estar compuesto por clientes que tienen algunas características distintivas:

Son más propensos a ser ciudadanos mayores (SeniorCitizen). Tienen una menor probabilidad de tener pareja (Partner) o dependientes (Dependents). Tienen una mayor antigüedad (tenure) en la compañía. Es más probable que tengan servicio telefónico (PhoneService), pero menos probable que tengan servicio de internet (InternetService). Tienen menos probabilidades de utilizar servicios adicionales como seguridad en línea, respaldo en línea, etc. Son más propensos a tener contratos a largo plazo (One year o Two year). Prefieren la facturación sin papel (PaperlessBilling). Prefieren pagar con cheque electrónico (PaymentMethod). Tienen cargos mensuales y totales promedio más bajos. Tienen menos tickets de administración y técnico.

**Clúster 1:** Este grupo parece estar compuesto por clientes que tienen un conjunto diferente de características:

Son más propensos a tener pareja (Partner) y dependientes (Dependents). Tienen más probabilidades de tener múltiples líneas (MultipleLines), así como otros servicios adicionales como seguridad en línea, respaldo en línea, etc. Tienen más probabilidades de tener contratos a largo plazo (One year o Two year). Prefieren la facturación en papel (PaperlessBilling). Tienen cargos mensuales y totales promedio más altos. Tienen más tickets de administración y técnico.

**Clúster 2:** Este grupo representa clientes con características diferentes nuevamente:

Son menos propensos a ser ciudadanos mayores (SeniorCitizen). Tienen menos probabilidad de tener pareja (Partner). Tienen una menor antigüedad (tenure) en la compañía. Son menos propensos a tener servicio telefónico (PhoneService) e internet (InternetService). Tienen menos probabilidades de utilizar servicios adicionales. Tienen menos probabilidades de tener contratos a largo plazo (One year o Two year). Prefieren pagar con cheque electrónico (PaymentMethod). Tienen cargos mensuales y totales promedio más altos. Tienen menos tickets de administración y técnico en comparación con el Clúster 1.

Vamos a probar Pycaret

```
# Configuración del entorno de PyCaret
exp_clu = setup(data=df, normalize=True, normalize_method='zscore', ignore_features=['Churn'])
```

Normalizamos con zscore

Creamos 3 clusters

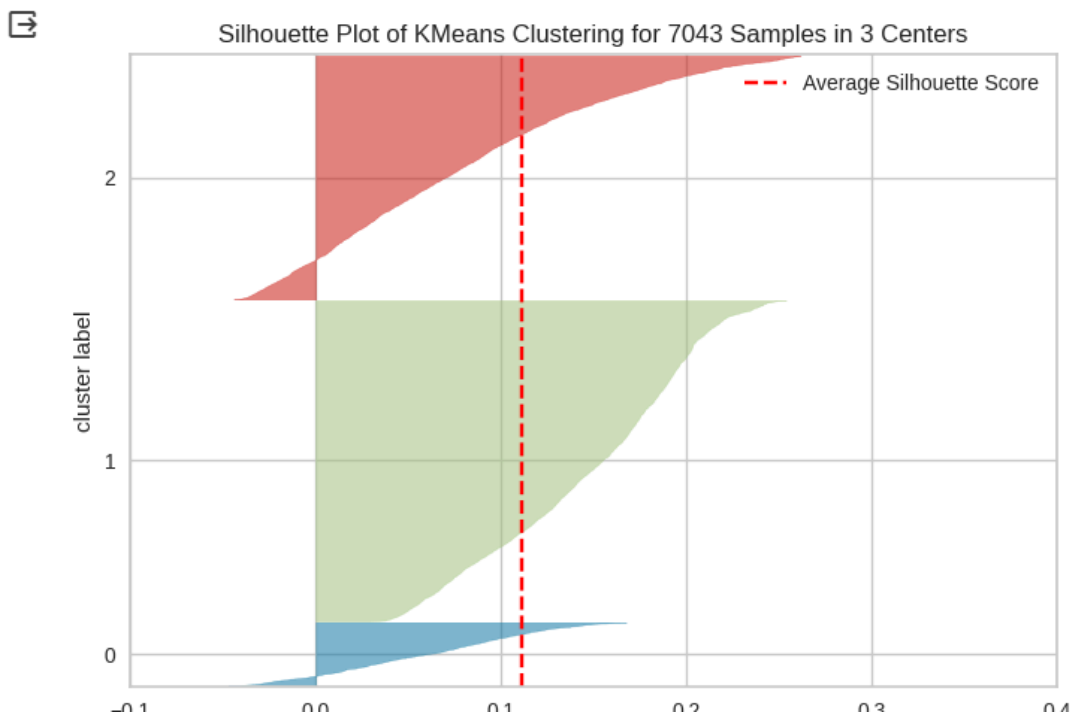


```
[ ] kmeans= create_model ('kmeans', num_clusters=3)
```

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.1113	624.8126	2.6188	0	0	0

Estos resultados sugieren que los clústeres no son altamente homogéneos en cuanto a la separación entre las clases conocidas (Homogeneidad y Completeness son 0) y no hay una fuerte similitud entre las agrupaciones obtenidas y una agrupación de referencia (índice de Rand es 0). Sin embargo, el valor de Silhouette sugiere que los objetos están en general bien emparejados con sus propios clústeres. La Calinski-Harabasz indica una buena estructura de los clústeres, pero el valor de Davies-Bouldin sugiere que la separación entre los clústeres podría mejorarse.

A pesar de obtener que Elbow nos indica 6 clusters decidimos seguir adelante con 3, reafirma nuestra estrategia Silhouette. En caso que no funcionen las estrategias a implementar, se cambiará a 6.



## Recomendaciones

El análisis ha identificado 3 clústeres de clientes con características distintivas. A continuación, se presenta una interpretación detallada de cada clúster y algunas recomendaciones basadas en sus hallazgos:

### Clúster 0: Clientes de bajo perfil

#### Características:

Mayor probabilidad de ser SeniorCitizen (ciudadanos mayores). Menor probabilidad de tener Partner (pareja) o Dependents (dependientes). Mayor duración de la antigüedad (tenure). Mayor probabilidad de tener PhoneService (servicio telefónico). Menor probabilidad de tener InternetService (servicio de internet). Menor probabilidad de tener servicios de seguridad en línea, respaldo en línea, protección de dispositivos, soporte técnico, televisión en streaming y películas en streaming. Mayor probabilidad de tener contratos a largo plazo (One year o Two year). Menor probabilidad de usar papel para facturación. Mayor probabilidad de pagar con cheque electrónico. Menor valor promedio de cargos mensuales y totales. Menor cantidad de tickets de administración y técnico. Interpretación: Este clúster parece estar compuesto por clientes de bajo perfil que utilizan principalmente el servicio telefónico y tienen una menor participación en servicios adicionales. Son clientes leales con contratos a largo plazo y tienden a pagar con cheque electrónico.

#### Recomendaciones:

Ofertas personalizadas: Se podrían ofrecer paquetes de servicios básicos a precios atractivos para este segmento, enfocándose en el servicio telefónico y la atención al cliente para clientes de la tercera edad. Canales de comunicación: Es importante considerar canales de comunicación tradicionales como el teléfono y el correo postal para llegar a este segmento. Estrategias de retención: Implementar estrategias de retención para clientes de larga data, como descuentos o programas de fidelización.

### **Clúster 1: Clientes activos**

#### Características:

Mayor probabilidad de tener Partner (pareja), Dependents (dependientes), MultipleLines (múltiples líneas), servicios de seguridad en línea, respaldo en línea, protección de dispositivos, soporte técnico, televisión en streaming y películas en streaming. Mayor probabilidad de tener contratos a largo plazo (One year o Two year). Mayor probabilidad de usar papel para facturación. Mayor valor promedio de cargos mensuales y totales. Mayor cantidad de tickets de administración y técnico. Interpretación: Este clúster parece estar compuesto por clientes activos que utilizan múltiples servicios, incluyendo internet, servicios de seguridad y entretenimiento en streaming. Son clientes con mayor participación en la empresa y generan más tickets de atención al cliente.

#### Recomendaciones:

Upselling y cross-selling: Aprovechar la predisposición de este segmento a utilizar múltiples servicios para ofrecerles upgrades, paquetes combinados o nuevos productos complementarios. Atención al cliente personalizada: Brindar una atención al cliente de alta calidad y rápida respuesta a sus solicitudes, considerando la mayor

cantidad de tickets que generan. Canales digitales: Enfocarse en canales de comunicación digitales como el sitio web, la aplicación móvil y el correo electrónico para facilitar la interacción con este segmento.

## **Clúster 2: Clientes potenciales**

Características:

Menor probabilidad de ser SeniorCitizen (ciudadanos mayores). Menor probabilidad de tener Partner (pareja). Menor duración de la antigüedad (tenure). Menor probabilidad de tener PhoneService (servicio telefónico). Menor probabilidad de tener InternetService (servicio de internet). Menor probabilidad de tener servicios de seguridad en línea, respaldo en línea, protección de dispositivos, soporte técnico, televisión en streaming y películas en streaming. Menor probabilidad de tener contratos a largo plazo (One year o Two year). Mayor probabilidad de pagar con cheque electrónico. Mayor valor promedio de cargos mensuales y totales. Menor cantidad de tickets de administración y técnico. Interpretación: Este clúster parece estar compuesto por clientes potenciales que aún no utilizan al máximo los servicios disponibles. Son clientes con menor antigüedad y menor participación en servicios adicionales.

Recomendaciones:

Campañas de marketing personalizadas: Dirigir campañas de marketing personalizadas a este segmento, destacando los beneficios de los diferentes servicios disponibles y ofreciendo promociones atractivas. Programas de fidelización: Implementar programas de fidelización que recompensen a los clientes por aumentar su uso de los servicios. Análisis de churn: Monitorear de cerca la tasa de churn en este segmento para identificar y abordar posibles factores de riesgo.

## **Consideraciones sobre el número de clústeres:**

Si bien el método Elbow sugiere 6 clústeres, es importante considerar que la elección del número óptimo de clústeres depende del contexto específico del negocio y de los objetivos del análisis. En este caso, la agrupación en 3 clústeres proporciona una segmentación clara y significativa de los clientes con características y comportamientos distintivos, lo que permite formular recomendaciones estratégicas específicas para cada grupo.

En este caso, la interpretación de los 3 clústeres obtenidos proporciona información valiosa y segmenta a los clientes de manera significativa para acciones comerciales.

Recomendaciones adicionales:

Análisis de sensibilidad: Se recomienda realizar un análisis de sensibilidad para evaluar la robustez de los resultados y verificar que la interpretación de los clústeres no cambia significativamente al variar el número de clústeres en un rango razonable.

## Conclusiones

Se han analizado las características promedio de cada cluster y se ha llegado a la conclusión de que el clustering puede ser una herramienta valiosa para las empresas que buscan:

**Segmentar el mercado de manera efectiva:** Los resultados sugieren que los clusters identificados se diferencian en características como género, edad, estado civil, dependientes, antigüedad en la empresa, servicios utilizados, métodos de pago y cargos mensuales. Dicha información puede ser utilizada para crear segmentos de mercado más homogéneos y desarrollar estrategias de marketing personalizadas para cada grupo.

**Optimizar las campañas de marketing:** Por ejemplo, el cluster 1, con un mayor uso de servicios en línea y mayor antigüedad en la empresa, podría ser un objetivo ideal para campañas de upselling o cross-selling. En cambio, el cluster 0, con un menor uso de servicios y mayor proporción de clientes nuevos, podría requerir estrategias de marketing enfocadas en la adquisición y la fidelización.

**Mejorar la retención de clientes:** El cluster 1, con mayor incidencia de tickets de soporte, podría ser un grupo de clientes con mayor riesgo de abandono. Se podrían implementar estrategias de retención específicas para este segmento, como programas de fidelización o atención al cliente personalizada.

**Desarrollar nuevos productos y servicios:** El análisis del comportamiento y las preferencias de cada cluster puede guiar el desarrollo de nuevos productos o servicios que satisfagan mejor sus necesidades específicas. Por ejemplo, el cluster 1, con mayor uso de servicios en línea, podría ser un objetivo para el desarrollo de nuevas aplicaciones o servicios digitales.

### Beneficios potenciales del clustering en este caso:

- Implementar el clustering como parte de un proceso continuo de análisis y mejora de las estrategias de marketing.
- Utilizar los resultados del clustering para crear perfiles de clientes detallados y desarrollar estrategias de marketing personalizadas para cada segmento.
- Monitorear el rendimiento de las campañas de marketing segmentadas y realizar ajustes según sea necesario.
- Integrar el clustering con otras herramientas de análisis de datos para obtener una visión completa del comportamiento del cliente.

El clustering es una herramienta poderosa que puede ayudar a las empresas a mejorar sus estrategias de marketing, segmentar su mercado de manera efectiva, optimizar sus campañas, aumentar la retención de clientes, desarrollar nuevos productos y servicios y crear experiencias personalizadas para cada cliente.