

A photograph of two women in a collaborative workspace. In the foreground, a woman with short blonde hair is smiling and looking towards the right. In the background, a woman with dark curly hair is also smiling and looking towards the left. The wall behind them is covered with several framed photographs. The image is partially obscured by a purple and pink gradient overlay at the bottom.

20XX

DATA MINING

GABRIEL OMAR QUISPE LLANQUE

EMPRESA OPERADORA	LÍNEAS GANADAS	LÍNEAS PERDIDAS	RESULTADO NETO
CLARO	147 617	130 456	↑ 17 161
ENTEL	127 467	115 677	↑ 11 790
SUMA MÓVIL	15	1	↑ 14
GUINEA MOBILE	119	258	↓ -139
FLASH	218	785	↓ -567
BITEL	67 550	79 556	↓ -12 006
MOVISTAR	110 653	126 906	↓ -16 253

[*] El nuevo procedimiento de portabilidad móvil se inició el 16 de julio de 2014.
Fuente: PUNKU-OSIPTEL (fecha de corte: 01/02/2023)

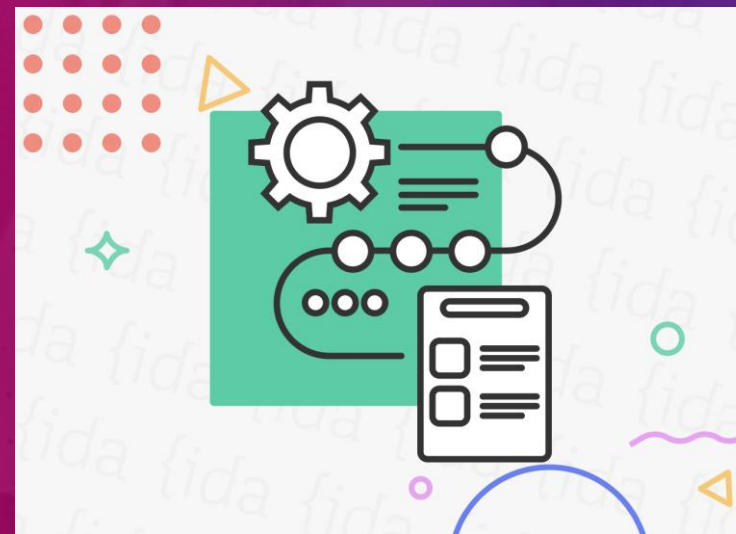
Predicción de retención de clientes en la empresa Movistar

En este proyecto usará técnicas de minería de datos e inteligencia artificial para predecir si los usuarios de Movistar se quedarán con este operador o cambiarán a otra compañía



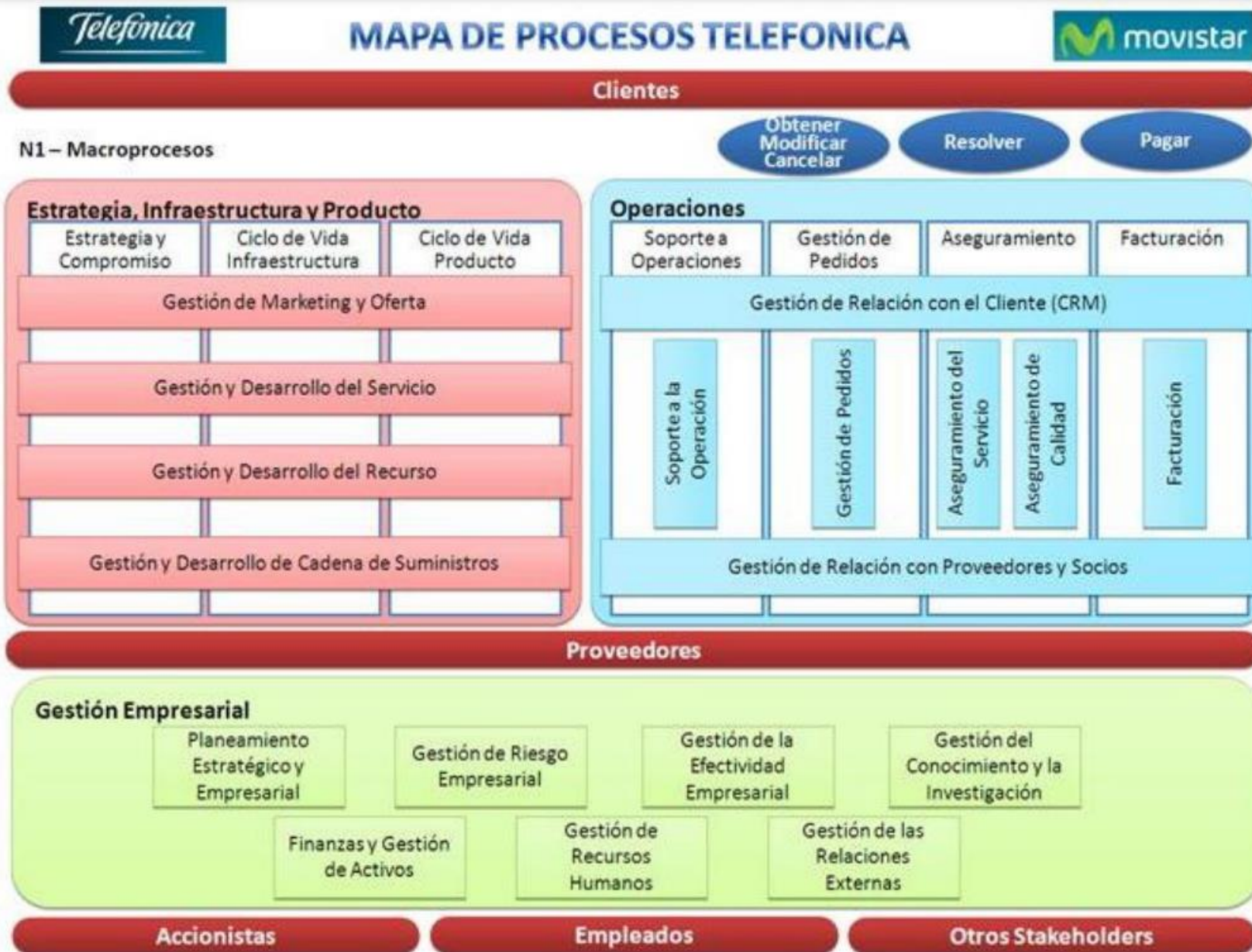
METODOLOGIA

La metodología que se usara para este proyecto sera CRISP-DM (Cross-Industry Standard Process for Data Mining). proporciona una estructura para guiara los profesionales a través de las fases de un proyecto de minería de datos.



Se logra identificar que el principal problema para la retención de clientes corresponde a la **oficina de gestión de relación con el cliente**

Que esta encargado del aseguramiento del servicio y el aseguramiento de la calidad de los servicios prestados.



COMPRESION DE NEGOCIOS

COMPRESION DE DATOS

- **Mayor60Años:** Indica si el cliente tiene más de 60 años. 0 = No, 1 = Sí.
- **Conyuge:** Indica si el cliente tiene cónyuge o pareja. 0 = No, 1 = Sí.
- **Dependientes:** Indica si el cliente tiene dependientes. 0 = No, 1 = Sí.
- **MesesDeContrato:** Meses que lleva el cliente con su contrato actual.
- **TelefonoFijo:** Indica si el cliente tiene servicio de teléfono fijo. 0 = Sin servicio, 1 = Sí tiene.
- **VariasLineasTelefonicas:** Indica si el cliente tiene varias líneas de teléfono fijo. 0 = No, 1 = Sí.
- **ServicioDeInternet:** Tipo de servicio de internet contratado. Valores: DSL, Fibra óptica, Sin servicio.
- **SeguridadOnline:** Indica si el cliente tiene servicio de seguridad online. 0 = No, 1 = Sí.
- **BackupOnline:** Indica si el cliente tiene servicio de backup online. 0 = No, 1 = Sí.
- **SeguroEnDispositivo:** Indica si el cliente tiene seguro para dispositivos. 0 = No, 1 = Sí.
- **SoporteTecnico:** Indica si el cliente tiene servicio de soporte técnico. 0 = No, 1 = Sí.
- **TVCable:** Indica si el cliente tiene servicio de TV por cable. 0 = No, 1 = Sí.
- **Streaming:** Indica si el cliente tiene servicio de streaming de video/música. 0 = No, 1 = Sí.
- **TipoDeContrato:** Tipo de contrato del cliente. Valores: Mensual, Anual, etc.
- **PagoOnline:** Indica si el cliente paga online. 0 = No, 1 = Sí.
- **FormaDePago:** Forma de pago del cliente. Valores: Cheque, Débito en cuenta, etc.
- **CuentaMensual:** Monto en dólares de la cuenta mensual del cliente.
- **Churn:** Indica si el cliente dejó la compañía. 0 = No, 1 = Sí.

	Mayor60Años	Conyuge	Dependientes	MesesDeContrato	TelefonoFijo	VariasLineasTelefonicas	ServicioDeInternet	SeguridadOnline	BackupOnline	SeguroEnDispositivo	SoporteTecnico	TVCable	Streaming	TipoDeContrato
0	0	Si	No	1	No	SinServicioTelefonico	DSL	No	Si	No	No	No	No	Mensual
1	0	No	No	34	Si	No	DSL	Si	No	Si	No	No	No	UnAño
2	0	No	No	2	Si	No	DSL	Si	Si	No	No	No	No	Mensual
3	0	No	No	45	No	SinServicioTelefonico	DSL	Si	No	Si	Si	No	No	UnAño
4	0	No	No	2	Si	No	FibraOptica	No	No	No	No	No	No	Mensual
...
7038	0	Si	Si	24	Si	Si	DSL	Si	No	Si	Si	Si	Si	UnAño
7039	0	Si	Si	72	Si	Si	FibraOptica	No	Si	Si	No	Si	Si	UnAño
7040	0	Si	Si	11	No	SinServicioTelefonico	DSL	Si	No	No	No	No	No	Mensual
7041	1	Si	No	4	Si	Si	FibraOptica	No	No	No	No	No	No	Mensual
7042	0	No	No	66	Si	No	FibraOptica	Si	No	Si	Si	Si	Si	DosAños

PREPROCESAMIENTO

20XX

Mayor60Años	0
Conyuge	0
Dependientes	0
MesesDeContrato	0
TelefonoFijo	0
VariasLineasTelefonicas	0
ServicioDeInternet	0
SeguridadOnline	0
BackupOnline	0
SeguroEnDispositivo	0
SoporteTecnico	0
TVCable	0
Streaming	0
TipoDeContrato	0
PagoOnline	0
FormaDePago	0
CuentaMensual	0
Churn	0
dtype: int64	

LABEL ENCODER

En nuestra data set tenemos valores continuas y categoricas, para que se pueda entrenar esta datasetes necesario tratar la data para que puede ser comprendida por la computadora

Para la transformacion de la data de tipo categorica, cree un diccionario, de tal manera que solo quede 0 y 1. Diccionario: si = 1, no = 0

Para el caso de la data de tipo categorica que posea mas de 2 estados, se usara el metodo get.dummies que hara sera construir mas columnas segun las variables de la columna original, lo que posteriormente sera rellenado con 1 y 0 segun corresponda

	Conyuge	Dependientes	TelefonoFijo	PagoOnline	Churn
0	1	0	0	1	0
1	0	0	1	0	0
2	0	0	1	1	1
3	0	0	0	0	0
4	0	0	1	1	1

INPUTACION DE DATOS

	Mayor60Años	MesesDeContrato	CuentaMensual	VariasLineasTelefonicas_No	VariasLineasTelefonicas_Si	VariasLineasTelefonicas_SinServicioTelefonico	ServicioDeInternet_DSL	ServicioDeInternet_FibraOptica	ServicioDeInternet_No
0	0	1	29.85	0	0	1	1	0	0
1	0	34	56.95	1	0	0	1	0	0
2	0	2	53.85	1	0	0	1	0	0
3	0	45	42.30	0	0	1	1	0	0
4	0	2	70.70	1	0	0	0	1	0

5 rows x 34 columns

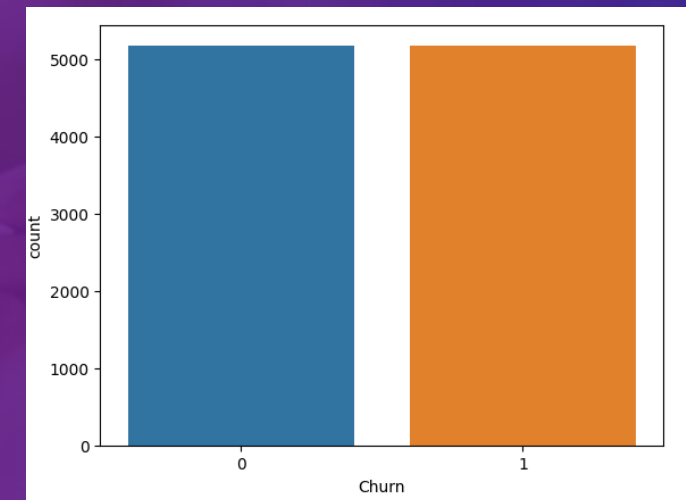
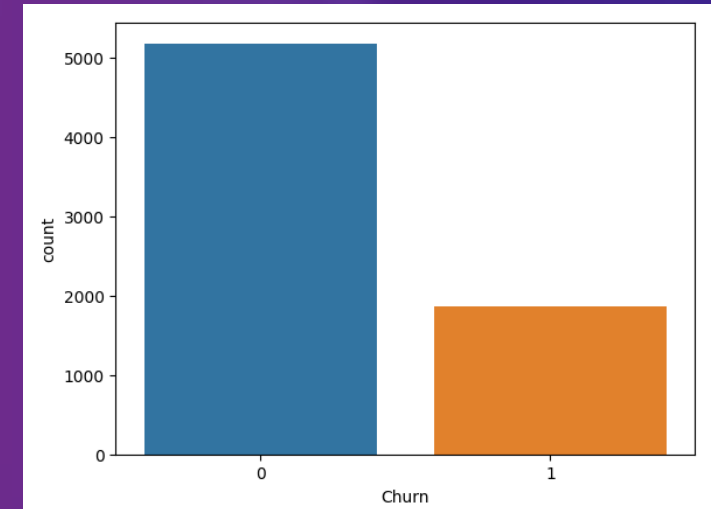
BALANCEO DE DATOS

20XX

OVER SAMPLING

La tecnica que mas se presta para nuestro pro-blema es el over sampling, debido a que crea datos ficticios, aumentando la data lo que beneficia al en-trenamiento.

Como se observa que nuestra data ya se encuentra balanceada y la data que se genero de manera artificial se añadio a nuestra tabla, aumentando en mas de 3000 registros. Ahora nuestro Data Set esta listo para ser entrenada



PROCESAMIENTO DATA

Se configuro el 30% para datos de prueba

Se configuro el 70 % para datos de entrenamiento

Ademas de dividir la data set en variable dependiente e independiente

```
Mayor60Años      0
Conyuge           0
Dependientes      0
MesesDeContrato   0
TelefonoFijo      0
VariasLineasTelefonicas  0
ServicioDeInternet  0
SeguridadOnline   0
BackupOnline      0
SeguroEnDispositivo  0
SoporteTecnico    0
TVCable           0
Streaming         0
TipoDeContrato    0
PagoOnline        0
FormaDePago       0
CuentaMensual     0
Churn             0
dtype: int64
```

20XX

Variable
independiente

ESTANDARIZAR

Se utiliza el StandardScaler de la biblioteca scikit-learn para estandarizar (normalizar) las características. Esto implica ajustar los datos para que tengan una media de cero y una desviación estándar de uno



Variable
dependiente

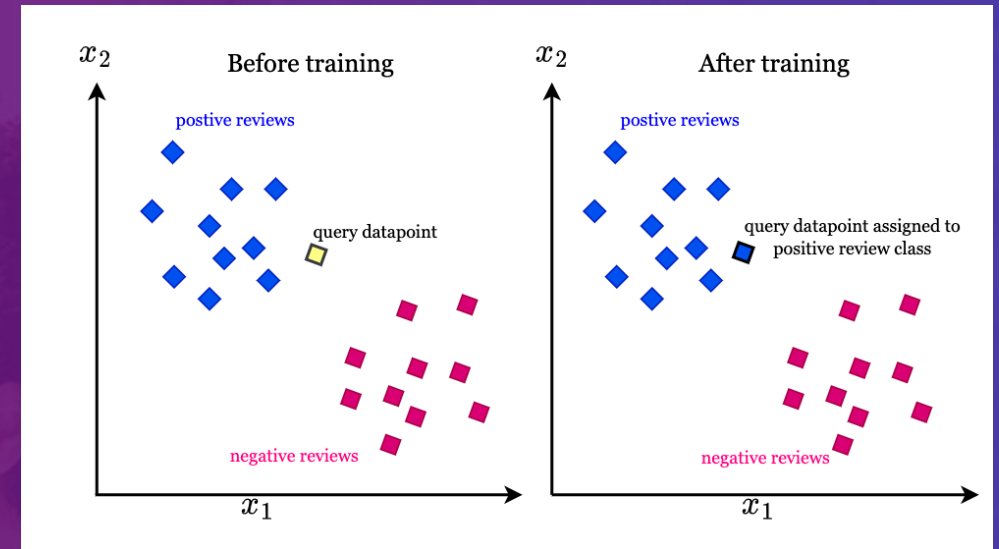
MODELOS DE ENTRENAMIENTO

20XX

KNN

Dime con quien andas y te dire quien eres

KNN clasifica un punto nuevo mirando los puntos cercanos en lugar de memorizar patrones. El valor de "k" representa la cantidad de vecinos más cercanos que se toman en cuenta al realizar una predicción. Por ejemplo, si $k=3$, se observarían los tres puntos más cercanos, y el nuevo punto se clasificaría según la mayoría de esas tres clasificaciones.



```
KNeighborsClassifier  
KNeighborsClassifier(metric='euclidean')
```

MODELOS DE ENTRENAMIENTO

20XX

NAIVE BAYES

Lo simple no falla

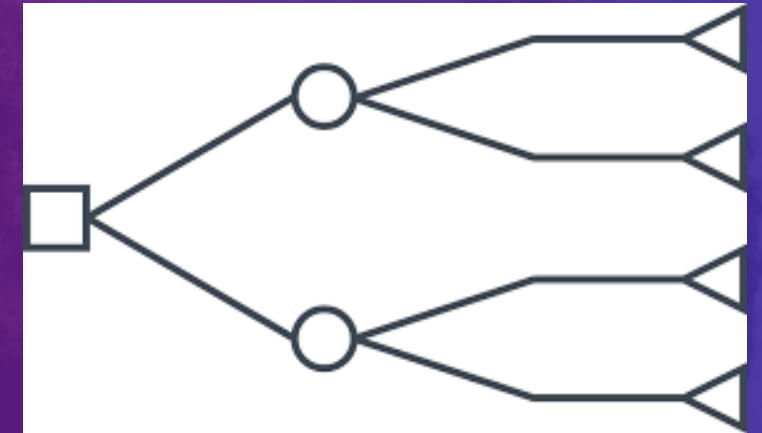
Naive Bayes es un modelo de machine learning que hace predicciones usando la probabilidad de que ciertos eventos estén relacionados, asumiendo que las características son independientes entre sí. Es rápido y eficiente, especialmente para clasificación de texto y problemas simples.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

```
▾ BernoulliNB  
BernoulliNB()
```

ARBOL DE DECISIONES

Un árbol de decisiones es un modelo de machine learning que toma decisiones paso a paso, dividiendo un conjunto de datos en grupos más pequeños basándose en características específicas. Se parece a un juego de "adivina quién", donde cada pregunta reduce las opciones hasta llegar a una decisión. Es utilizado para clasificación y regresión, siendo fácil de entender y visualizar.



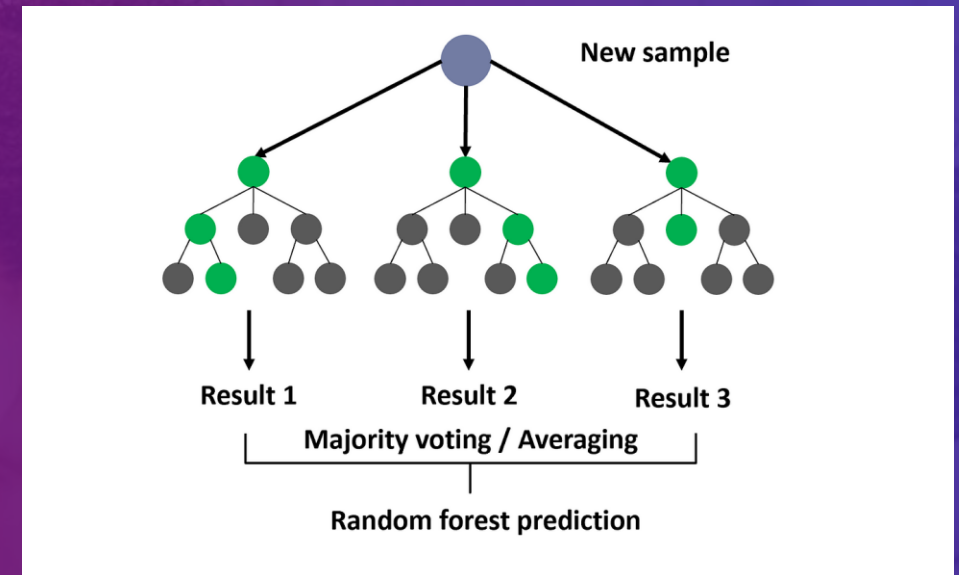
```
DecisionTreeClassifier  
DecisionTreeClassifier(criterion='entropy', random_state=42)
```


MODELOS DE ENTRENAMIENTO

20XX

RANDOM FOREST

combina múltiples árboles de decisión para tomar decisiones más robustas y precisas. En lugar de depender de un solo árbol, crea un "bosque" de árboles y toma decisiones basadas en la mayoría de votos de los árboles individuales. Esto mejora la generalización y rendimiento del modelo, haciendo que Random Forest sea eficaz para una variedad de tareas, incluyendo clasificación y regresión.



```
RandomForestClassifier  
RandomForestClassifier(criterion='entropy', random_state=42)
```

VALIDACION DE MODELOS

MATRIZ DE CONFUSION

20XX

es una herramienta que se utiliza en el campo de la clasificación en aprendizaje automático. Proporciona un resumen de rendimiento del modelo, mostrando la cantidad de verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos FN

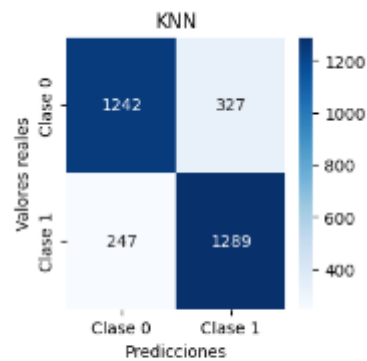


Figura 11. Matriz de Confusión KNN

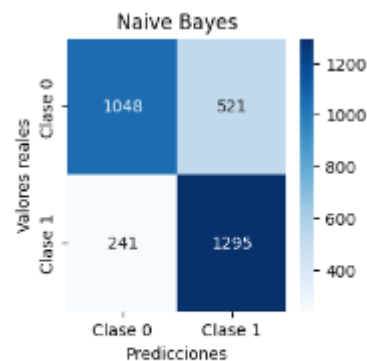


Figura 12. Matriz de Confusión Naive Bayes

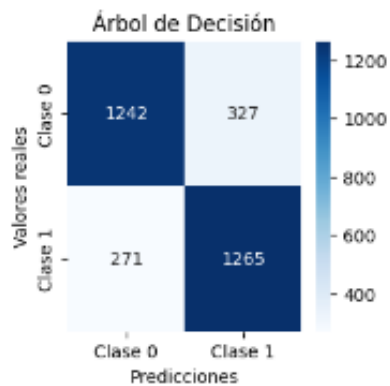


Figura 13. Matriz de Confusión Árbol de Decisiones

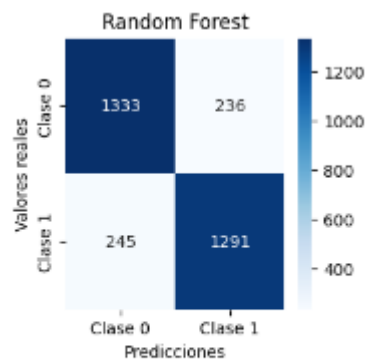


Figura 14. Matriz de Confusión Random Forest

		Predicción	
		Sí	No
Real	Sí	VP	FN
	No	FP	VN

Modelo	VP	FN	FP	VN
KNN	1242	327	247	1289
Naive Bayes	1048	521	241	1295
Árbol de Decisión	1242	327	271	1265
Random Forest	1333	236	245	1291

VALIDACION DE MODELOS

20XX

Modelo	Exactitud
KNN	0.8151368760064412
Naive Bayes	0.7545893719806763
Árbol de Decisión	0.8074074074074075
Random Forest	0.8450885668276973



Modelo	Precisión
KNN	0.7976485148514851
Naive Bayes	0.7131057268722467
Árbol de Decisión	0.7945979899497487
Random Forest	0.8454485920104781

Modelo	Sensibilidad
KNN	0.8391927083333334
Naive Bayes	0.8430989583333334
Árbol de Decisión	0.8235677083333334
Random Forest	0.8404947916666666


- **Exactitud (Accuracy):** Es la medida de cuántas predicciones correctas hizo tu modelo en comparación con todas las predicciones. Un alto porcentaje de exactitud significa que tu modelo está acertando la mayoría de las veces.
- **Precisión (Precision):** Mide cuántas de las instancias que el modelo predijo como positivas son realmente positivas. En otras palabras, si el modelo dice que algo va a suceder, la precisión te dice cuántas veces tenía razón.
- **Sensibilidad (Recall):** Mide cuántas de las instancias positivas reales el modelo logró capturar con sus predicciones positivas. En términos simples, si algo realmente sucedió, la sensibilidad te dice cuántas veces el modelo lo identificó correctamente.

IMPLEMENTACION

20XX

Para realizar la prueba de prediccion de nuestros modelos de entrenamiento, se realizo la **creación de un usuario**, este usuario **no esta en nuestra dataset** por lo que estamos **probando la prediccion** de los diferentes modelos en un caso real:

CON LAS SIGUIENTES CARACTERIZTICAS



Característica	Valor
Mayor60Años	No
Conyuge	No
Dependientes	No
MesesDeContrato	0
TelefonoFijo	Sí
VariasLineasTelefonicas	No
ServicioDeInternet	FibraOptica
SeguridadOnline	No
BackupOnline	Sí
SeguroEnDispositivo	No
SoporteTecnico	Sí
TVCable	No
Streaming	Sí
TipoDeContrato	UnAño
PagoOnline	Sí
FormaDePago	DebitoEnCuenta
CuentaMensual	39.90

$$X_{\text{pablito}} = [0, 0, 1, 1, 0, 0, 39.90, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1]$$

KNN - El ususario clasifica como : NO CHURN (Se mantiene en la compañía)

NB - El ususario clasifica como : NO CHURN (Se mantiene en la compañía)

AD - El ususario clasifica como : NO CHURN (Se mantiene en la compañía)

RF - El ususario clasifica como : NO CHURN (Se mantiene en la compañía)

CONCLUSION

EL MEJOR MODELO ES **RANDOM FOREST**
POR LO SIGUIENTE:

- Mejor exactitud con un 84 %
- Mejor precisión con un 84 %.
- Sensibilidad del 84 %
- Mayor Verdadero positivos con 1333
- Mayor Verdaderos Negativos con 1291
- Menor Falsos Negativos 236
- Segundo Menor Falso Positivo con 245

Se destaca la importancia de prestar atención a tres columnas específicas en la matriz de correlación: la **edad del usuario** (mayor a 60 años), la **duración del contrato** y la **cuota mensual**. Estas tres características han demostrado ser las más influyen-tes en la clasificación de usuarios CHURN

