Proyecto de Unidad:

Clasificacion de fidelizacion de clientes en la empresa Movistar

Gabriel Omar Quispe LLanque
2021204031@unam.edu.pe

Facultad de Ingeniería y arquitectura
Universidad Nacional de Moquegua

15 de diciembre de 2023

Resumen

Res.

Palabras clave: churn

Índice

1. Introducción **3**

2. Objetivos **3**

2.1 Objetivo General 3

2.2 Objetivos Específicos 3

3. Metodologia **3**

4. Procedimiento **3**

4.1 Comprensión de negocio 3

4.1.1 Datos generales 3

4.1.2 Servicios 3

4.2 Comprensión de datos 4

4.3 Preprocesamiento de datos 4

4.3.1 Imputacion de datos 4

4.3.2 Label encoder 5

4.3.3 Balanceo de Datos 5

4.4 Procesamiento de Datos 6

4.4.1 Standarizacion 6

4.4.2 División de Datos 6

4.5 Modelos de Entrenamiento 6

4.5.1 Modelo KNN 6

4.5.2 Modelo Naive Bayes 6

4.5.3 Modelo Arbol de desicion 6

4.5.4 Modelo Random Forest 6

5. Validacion de Modelos **6**

5.1 Matriz de Confusion 6

5.1.1 Verdadero Positivo (VP) 6

5.1.2 Falso Negativo (FN) 7

5.1.3 Verdadero Negativo (VN) 7

5.1.4 Falso Positivo (FP) 7

5.2 Exactitud 8

5.3 Presicion 8

5.4 Sensibilidad 8

6. Ejecucion de modelos **8**

6.1 Ingreso de usuario 8

6.2 Pruebas 8

7. Análisis y discusión de resultados **9**

7.1 Mejor Modelo 9

7.2 Matriz de Correlacion 9

8. Conclusion **9**

Bibliografía **9**

1. Introducción

La fidelización de clientes es un factor clave para el éxito de las empresas de telecomunicaciones. Con la alta competitividad en este mercado, resulta esencial que las compañías encuentren formas de retener a sus suscriptores y evitar la cancelación del servicio.

Este proyecto permitirá poner en práctica técnicas avanzadas de minería de datos e inteligencia artificial para resolver un problema real de negocios. Los resultados e insights obtenidos serán de utilidad para que Movistar u otras empresas de telecomunicaciones mejoren la relación con sus clientes y tomen mejores decisiones comerciales.[2]

2. Objetivos

2.1. Objetivo General

En este proyecto, se busca desarrollar y aplicar un modelo de entrenamiento para la clasificación de clientes en la empresa Movistar.

2.2. Objetivos Específicos

- Implementar diversos modelos de entrenamiento.
- Comparar el rendimiento de los diferentes modelos.
- Identificar el modelo de entrenamiento óptimo para la clasificación de clientes en la empresa Movistar.

3. Metodologia

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

4. Procedimiento

4.1. Comprensión de negocio

4.1.1. Datos generales

RUC: 20100017491
Razón Social: TELEFONICA DEL PERU S.A.A.
Tipo Empresa: Sociedad Anónima Abierta
Actividad Comercial: Telecomunicaciones
Fecha Inicio Actividades: 25 / Junio / 1920

4.1.2. Servicios

La empresa opera comercialmente en el mercado nacional bajo la marca Movistar, a través de la cual brinda principalmente los servicios mostrados en la figura posterior 1.

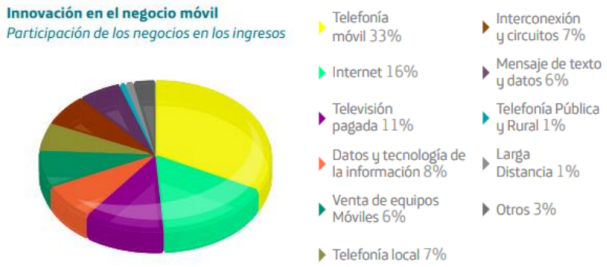


Figura 1. Servicios Ofrecidos

En enero de 2023, un total de 453 639 líneas móviles cambiaron de empresa operadora, 1.29 % menos respecto a las cifras de portabilidad reportadas en el mismo mes del año previo, informó el Organismo Supervisor de Inversión Privada en Telecomunicaciones (OSIPTEL) 2.

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)

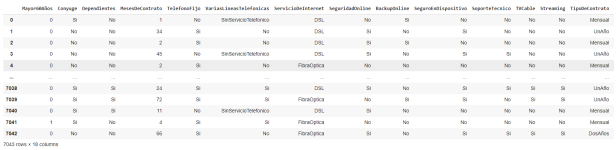
Figura 2. Cambio de Operadoras-2023

Lo que indica que cuando un usuario cambia de operadora movil, cambia todos su servicios que tiene con ella, es por ello la importancia de mantener a los clientes.

Por lo tanto en base del mapa de procesos de la empresa movistar, se logra identificar que el principal problema para la retencion de clientes corresponde a la oficina de gestion de relacion con el cliente, que esta encargado del aseguramiento del servicio y el aseguramiento de la calidad de los servicios prestados.

4.2. Comprensión de datos

La DataSet es la siguiente doinde se observa 7043 filas y 18 columnas, donde tenemos datos de tipo continua y categorica, las cuales esta es de 2 y mas categorias.



	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

7043 rows × 18 columns

Figura 3. Data set

Columna	Descripción
Mayor 60 Años	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.
Meses De Contrato	Meses que lleva el cliente con su contrato actual.
Telefono Fijo	Indica si el cliente tiene servicio de teléfono fijo. 0 = Sin servicio, 1 = Sí tiene.
Varias Lineas Telefonicas	Indica si el cliente tiene varias líneas de teléfono fijo. 0 = No, 1 = Sí.
Servicio De Internet	Tipo de servicio de internet contratado. Valores: DSL, Fibra óptica, Sin servicio.
Seguridad Online	Indica si el cliente tiene servicio de seguridad online. 0 = No, 1 = Sí.
Backup Online	Indica si el cliente tiene servicio de backup online. 0 = No, 1 = Sí.
Seguro En Dispositivo	Indica si el cliente tiene seguro para dispositivos. 0 = No, 1 = Sí.
Soporte Tecnico	Indica si el cliente tiene servicio de soporte técnico. 0 = No, 1 = Sí.

Tabla 1. Dataset Parte 1

Columna	Descripción
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.
Tipo De Contrato	Tipo de contrato del cliente. Valores: Mensual, Anual, etc.
Pago Online	Indica si el cliente paga online. 0 = No, 1 = Sí.
Forma De Pago	Forma de pago del cliente. Valores: Cheque, Débito en cuenta, etc.
Cuenta Mensual	Monto en dólares de la cuenta mensual del cliente.
Churn	Indica si el cliente dejó la compañía. 0 = No, 1 = Sí.

Tabla 2. Dataset Parte 2

4.3. Preprocesamiento de datos

4.3.1. Imputacion de datos

Esta data es obtenida de un trabajador de Movistar encargado de llamar a los clientes que son clientes de movistar y se estan poniendo en contacto con otras operadoras para contratar sus servicios.

Ellos realizan informes semanales de las llamadas que hicieron a los clientes intentando que se mantengan con movistar, ofreciendole nuevas promociones, descuento o solucionar sus problemas, Al ser la dataset producto de un informe, se obtiene que no se tiene problemas de imputacion de datos, es decir que no tenemos valores null.

```
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
```

Figura 4. Imputacion

4.3.2. Label encoder

En nuestra data set tenemos valores continuas y categoricas, para que se pueda entrenar esta dataset es 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

	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

Figura 5. Label Encoder

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

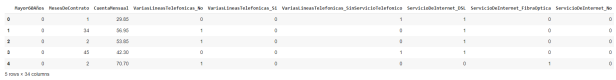


Figura 6. Label Encoder 2

Se junta toda la data trabajada, se observa que se tiene ahora mas columnas a diferencia de la dataset original. Como se observa en la data set original se presenta 18 columnas ya hora gracias a las tecnicas de Label Encoder se tiene 39 Columnas con datos categoricos de 0 y 1 que puedan ser comprendidas por en ordenador.

4.3.3. Balanceo de Datos

Para el balanceo de datos se necesita identificar nuestro target, el target es (Churn) debido, que este nos indica si un cliente se ira del operador o no y el cual se debe pronosticar, Por lo cual se tomara el resto de columnas como variables que daran el resultado de churn.

$f(x) = y$, donde y sera el churn , $f(x)$ seran las columnas restantes.

Como se observa, la data set se encuentra desbalanceada, por lo que si entrenamos en estas condiciones los resultados tiende a que y = 0.Es decir que las respuestas se inclinaran mas a que los clientes se quedaran con la compañia. Para este tipo de problemas existen metodos de balance de datos como lo son el over sampling y el under sampling.

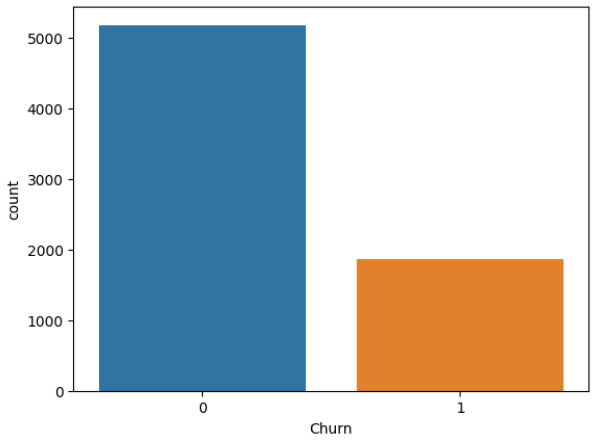


Figura 7. Balanceo de Datos

La tecnica que mas se presta para nuestro problema es el over sampling, debido a que crea datos ficticios, aumentando la data lo que beneficia al entrenamiento.Ser aplica el metodo descrito y se obtiene el siguiente resultado.

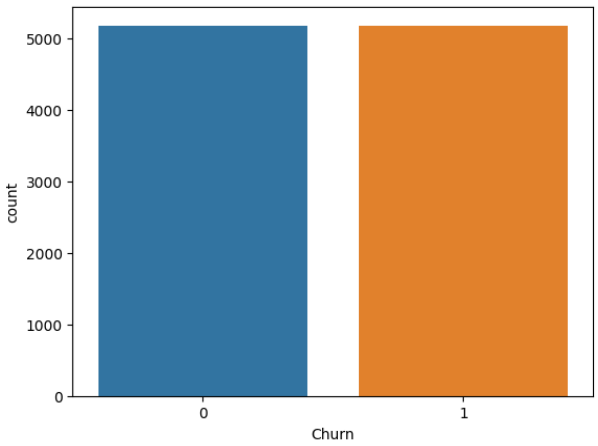


Figura 8. Over Sampling

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 lista para ser entrenada.

	Conyuge	Dependientes	TelefonoFijo	PagoOnline	Churn	Mayor60Años	MesesDeContrato	CuentaMensual	VariasLineasTelefonicas_No	VariasLineasTelefonicas_Si
0	1	0	0	1	0	0	1	29.85	0	0
1	0	0	1	0	0	0	34	56.95	1	0
2	0	0	1	1	1	0	2	53.85	1	0
3	0	0	0	0	0	0	45	42.30	0	0
4	0	0	1	1	1	0	2	70.70	1	0
...	...	...	...	...	...	...	...	...	...	...
7038	1	1	1	1	0	0	24	84.80	0	1
7039	1	1	1	1	0	0	72	103.20	0	1
7040	1	1	0	1	0	0	11	29.60	0	0
7041	1	0	1	1	1	1	4	74.40	0	1
7042	0	0	1	1	0	0	66	105.65	1	0

7043 rows × 39 columns

(a) Sin Over Sampling (a)

	Conyuge	Dependientes	TelefonoFijo	PagoOnline	Mayor60Años	MesesDeContrato	CuentaMensual	VariasLineasTelefonicas_No	VariasLineasTelefonicas_Si
0	1	0	0	1	0	1	29.650000	0	0
1	0	0	1	0	0	34	56.950000	1	0
2	0	0	1	1	0	2	53.850000	1	0
3	0	0	0	0	0	45	42.300000	0	0
4	0	0	1	1	0	2	70.700000	1	0
...	...	...	...	...	...	...	...	...	...
10043	1	0	1	1	0	5	80.921075	0	1
10044	0	0	1	1	0	3	74.347278	0	1
10045	0	0	1	1	0	18	70.164647	1	0
10046	0	0	1	0	0	30	93.024726	0	1
10047	0	0	1	0	0	4	50.880441	0	0

10348 rows × 39 columns

(b) Con Over Sampling (b)

Figura 9. Balance de datos

4.4. Procesamiento de Datos

4.4.1. Standarizacion

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.

4.4.2. División de Datos

Se configuró una división de datos donde el 70 % se utiliza como conjunto de entrenamiento y el 30 % restante como conjunto de prueba. Esta división permite verificar la eficacia de nuestro proceso de entrenamiento.

4.5. Modelos de Entrenamiento

4.5.1. Modelo KNN

El modelo KNN (K-Nearest Neighbors) es un algoritmo de aprendizaje automático utilizado para clasificación y regresión. La idea principal detrás de KNN es bastante simple. Supongamos que tienes un conjunto de datos con puntos en un espacio. Cuando quieres predecir el resultado para un nuevo punto, el modelo KNN busca los "vecinos"más cercanos a ese punto en el espacio de características.

4.5.2. Modelo Naive Bayes

El modelo Naive Bayes es un algoritmo de aprendizaje automático que se utiliza principalmente para la clasificación y el filtrado. Se basa en el teorema de Bayes y asume una ïngenuidad.º independencia condicional entre las características.[1]

4.5.3. Modelo Arbol de desicion

El modelo de árbol de decisión es un algoritmo de aprendizaje automático que se utiliza para resolver problemas de clasificación y regresión. En cada nodo, elige la característica que mejor divide los datos en subgrupos más puros (homogéneos). La "pureza"se mide mediante métricas como la ganancia de información o la impureza de Gini.

4.5.4. Modelo Random Forest

Random Forest es un algoritmo de aprendizaje automático que se basa en la construcción de múltiples árboles de decisión y combina sus resultados para mejorar la precisión y la robustez del modelo. En cada nodo de cada árbol, se selecciona aleatoriamente un subconjunto de características. Esto ayuda a decorrelacionar los árboles y a hacer que cada uno se enfoque en diferentes aspectos del conjunto de datos.[3]

5. Validacion de Modelos

5.1. Matriz de Confusion

La matriz de confusión 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). 10.

		Predicción	
		Sí	**No**
Real	Sí	VP	FN
	No	FP	VN

Figura 10. Matriz de Confusion

5.1.1. Verdadero Positivo (VP)

Ocurre cuando en nuestros datos de entrenamiento la etiqueta es 1 y el modelo predice correctamente un 1 en los datos de prueba.

5.1.2. Falso Negativo (FN)

Sucede cuando en nuestros datos de entrenamiento la etiqueta es 1, pero el modelo predice incorrectamente un 0 en los datos de prueba.

5.1.3. Verdadero Negativo (VN)

Sucede cuando en nuestros datos de entrenamiento la etiqueta es 0 y el modelo predice correctamente un 0 en los datos de prueba.

5.1.4. Falso Positivo (FP)

Ocurre cuando en nuestros datos de entrenamiento la etiqueta es 0, pero el modelo predice incorrectamente un 1 en los datos de prueba.
Se observa los VP, FN, FP, VN de los modelos correspondientes.

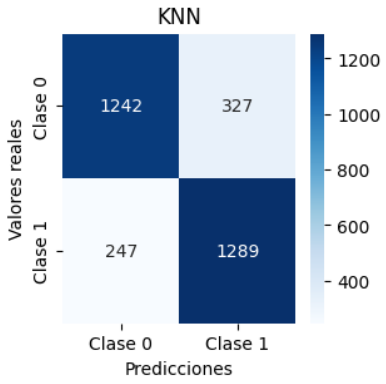


Figura 11. Matriz de Confusión KNN

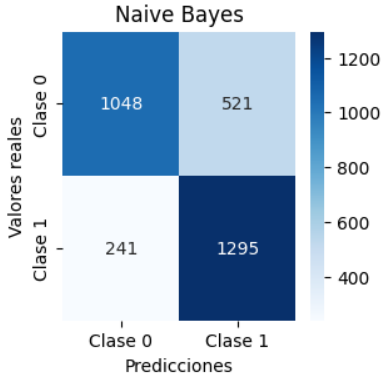


Figura 12. Matriz de Confusión Naive Bayes

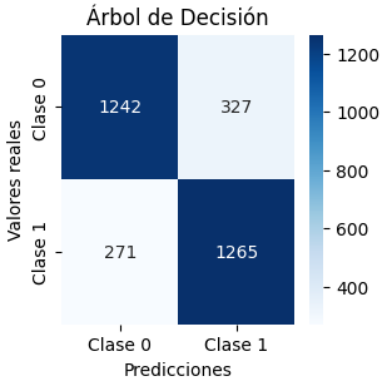


Figura 13. Matriz de Confusión Árbol de Decisiones

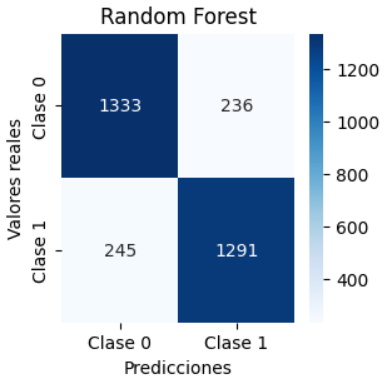


Figura 14. Matriz de Confusión Random Forest

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

Tabla 3. Resultados matriz de Confusion

5.2. Exactitud

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

Tabla 4. Resultados de los modelos por Exactitud

5.3. Presicion

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

Tabla 5. Resultados de los modelos por Presicion

5.4. Sensibilidad

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

Tabla 6. Resultados de los modelos Por Sensibilidad

6. Ejecucion de modelos

6.1. Ingreso de usuario

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

Las caracterizticas del usuario son las siguientes.

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

Tabla 7. Datos de Características

Para realizar la prueba con este usuario, necesitamos presentar sus características en el formato que utilizamos en la etapa de codificación de etiquetas (label Encoder) es decir con las 39 columnas que quedo, en lugar de las 18 que era inicialmente. El formato sería el siguiente:

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

$$0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1]$$

6.2. Pruebas

Se realizo la prediccion del ususario en los diferentes modelos implemntados, se obtuvo que todos los modelos indicaban que el usuario Xpablito clasifica como un usuario NO CHURN.

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

Figura 15. KNN

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

Figura 16. Naive Bayes

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

Figura 17. Arbol de Decisiones

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

Figura 18. Random Forest

7. Análisis y discusión de resultados

7.1. Mejor Modelo

El mejor modelo para este problema, teniendo en cuenta las métricas anteriores, es el modelo **Random Forest**, que se comporta de la mejor manera.

- Mejor exactitud con un 84 %.
- Mejor precisión con un 84 %.
- Sensibilidad del 84 %.
- Nayor Verdapositivos con 1333.
- Mayor Verdaderos Negativos con 1291.
- Menor Falsos Negativos 236.
- Segundo Menor Falso Poitivo con 245.

7.2. Matriz de Correlacion

Las columnas con Mayor importancia que definen si un usuario Clasifica como CHURN o no son la columna de mayor a 60 años, Mese de Contrato Y la cuota mensual.

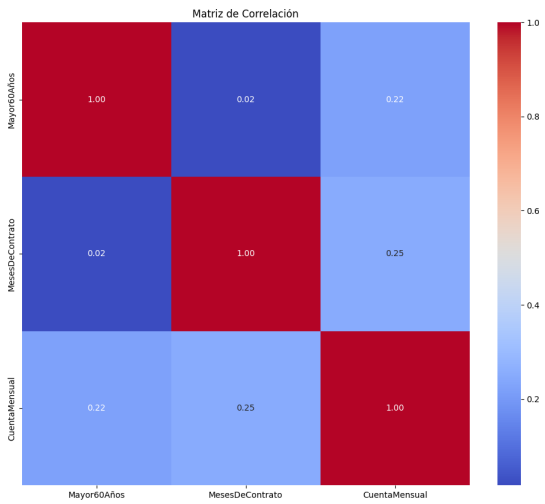


Figura 19. Matriz de Correlacion

8. Conclusion

Aplicando los diferentes métodos de aprendizaje automático, se concluye que el modelo más eficaz para clasificar la fidelización de un usuario en la empresa Movistar es el Random Forest. Este modelo ha demostrado una precisión y exactitud destacadas, así como un rendimiento superior en términos de Verdaderos Positivos (VP), Verdaderos Negativos (VN) y Falsos Positivos (FP).

La identificación precisa de usuarios propensos a cancelar el servicio es crucial, y el Random Forest ha demostrado ser altamente efectivo en esta tarea. Adicionalmente, 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 influyentes en la clasificación de usuarios CHURN.

Por lo tanto, se recomienda tomar precauciones y desarrollar estrategias específicas para usuarios que cumplen con estas características, ya que son indicadores clave de posibles cancelaciones de servicio.

Bibliografía

[1] Panagiotis Antoniadis. «Decision Tree vs. Naive Bayes Classifier». En: *Baeldung* (2023).

[2] Márcio Guia. «Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis». En: *Orcid* (2019).

[3] Li-Min Wang. «Combining decision tree and Naive Bayes for classification». En: *Orcid* (2006).