Predicción de abandono en clientes de empresas de Telecomunicaciones mediante aprendizaje automático

David Alejandro Parra, Josué Rodríguez

I. INTRODUCCIÓN

En la actualidad se ha presentado un auge en el mercado de los diferentes servicios en línea que pueden prestar las compañías de telecomunicaciones y así mismo, ha incrementado la competencia entre estas a la hora de mantener consigo a sus clientes, ya que dentro de una misma categoría de producto tienen la libertad de elegir entre muchos proveedores y con una mala experiencia pueden desertar.
En este documento se abordará el desarrollo del proyecto de Inteligencia Artificial el cual tiene como finalidad recopilar y analizar datos para realizar un programa que permita hacer "churn prediction" o predicciones de abandono con los datos de una empresa de telecomunicaciones.

II. PREGUNTAS DE INTERÉS

A. *¿Es posible predecir quién es un cliente que desea cancelar los servicios?*

Esta es una de las principales finalidades del proyecto, predecir qué clientes podrían cancelar algún servicio o abandonar la compañía por completo, de tal forma que el programa pueda realizarle ofertas o incentivos antes de que suceda para asegurarse de que no se retire.

B. *¿Es posible predecir qué servicios necesita el cliente?*

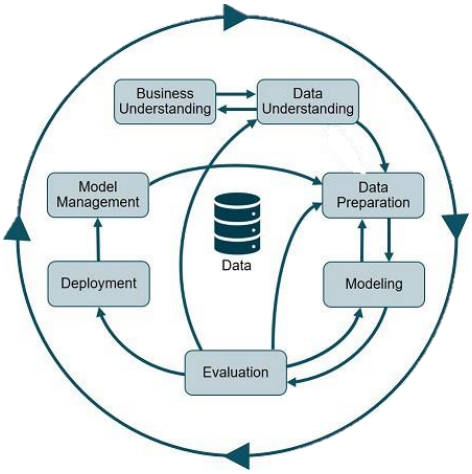
Otro objetivo que se podría esperar del programa es hacer uso de los patrones descubiertos en el comportamiento de los clientes con el fin de poder predecir qué servicios requieren los usuarios mayormente en ciertas temporadas para que así la compañía puede prepararse y ofrecer mejor cobertura y promociones que aseguren una mayor calidad de servicio.

III. TIPO DE PROBLEMA

Como se evidencio en las preguntas de interés, nos encontramos con un problema de **CLASIFICACIÓN**, dado que tenemos como entrada un conjunto de variables (predictores) que describen el comportamiento de un cliente y a través de los modelos obtendremos como salida determinar a qué categoría pertenece el cliente (abandona, no abandona).

IV. METODOLOGÍA

Para el desarrollo del proyecto se optó por usar la metodología CRISP-DM (Cross Industry Standar Process for Data Mining). Realizando los siguientes ajustes:



Evaluation → Modeling
Evaluation → Data Preparation
Evaluation → Data understanding
Deployment → Model management

Con el fin de obtener un despliegue eficaz y capaz de obtener los mejores resultados.

V. MÉTRICAS DE PROGRESO

Dentro de cualquier proyecto, es importante tener en cuenta las métricas a través de las cuales podemos ver reflejado el avance y la calidad obtenida en los procesos de las diferentes áreas de gestión. En nuestro caso específico, es de vital importancia establecer métricas que permitan evaluar y mejorar los modelos de aprendizaje automático y en general, el desarrollo del programa.

Para este fin, es importante tener en cuenta el tipo de problema que estamos abordando, como vimos en el punto III, se trata de un problema de clasificación, aunque también puede abordarse como un problema de regresión para estimar la relación entre la variable objetivo y otros valores de datos que influyen en la variable objetivo, expresados en valores continuos.

Para ello se puede hacer uso de diferentes técnicas de aprendizaje automático como KNN, regresión lineal, regresión polinomial, árboles de decisión, bosques aleatorios, entre algunos métodos más.

En consecuencia, usaremos las siguientes métricas de regresión para aprendizaje automático:

a) *MSE) – Error cuadrático medio*
b) *(R^2) – R al cuadrado*
c) *TP Rate,, FP Rate, Error Rate*
d) *Precisión*

VI. Análisis de los datos

Todos los datos se obtuvieron a través de un archivo de Excel en formato CSV (separado por comas), empezamos la exploración con la observación de los primeros datos y los atributos para comprender su estructura y tipo de datos, posteriormente verificamos si el conjunto de datos tenía datos nulos pero encontramos que no tenía ninguno, por lo determinamos que el dataset ya venía limpio. Finalmente cambiamos el tipo de dato de la variable 'TotalCharges' que era Object por defecto a float64.

I. Variables de interés:

El conjunto de datos tiene **7043 filas y 21 columnas.**

Hay **17 características categóricas:**

CustomerID: ID de cliente único para cada cliente
Gender: si el cliente es hombre o mujer
SeniorCitizen: si el cliente es una persona mayor o no (1, 0)
Partner: si el cliente tiene pareja o no (Sí, No)
Dependent: si el cliente tiene dependientes o no (Sí, No)
PhoneService: Si el cliente tiene servicio telefónico o no (Sí, No)
MultipeLines: Si el cliente tiene múltiples líneas o no (Sí, No, Sin servicio telefónico)
InternetService: Proveedor de servicios de Internet del cliente (DSL, Fibra óptica, No)
OnlineSecurity: si el cliente tiene seguridad en línea o no (Sí, No, Sin servicio de Internet)
OnlineBackup: si el cliente tiene una copia de seguridad en línea o no (Sí, No, Sin servicio de Internet)
DeviceProtection: si el cliente tiene protección de dispositivo o no (Sí, No, Sin servicio de Internet)
TechSupport: si el cliente tiene soporte técnico o no (Sí, No, Sin servicio de Internet)
StreamingTV: si el cliente tiene streaming de TV o no (Sí, No, Sin servicio de Internet)
StreamingMovies: si el cliente tiene películas en streaming o no (Sí, No, Sin servicio de Internet)
Contract: El plazo del contrato del cliente (Mes a mes, Un año, Dos años)
PaperlessBilling: El plazo del contrato del cliente (Mes a mes, Un año, Dos años)
PaymentMethod: el método de pago del cliente (Cheque electrónico, Cheque enviado por correo, Transferencia bancaria (automática), Tarjeta de crédito (automática))

3 características numéricas:

Tenure: Número de meses que el cliente ha permanecido en la empresa
MonthlyCharges: el monto cobrado al cliente mensualmente
TotalCharges: El monto total cobrado al cliente

Finalmente, hay una función de predicción:

Churn: Si el cliente abandonó o no (Sí o No)

II. Análisis exploratorio

Una vez limpio el dataset y comprendiendo cada una de las variables, procedimos a realizar diferentes gráficas que nos brindaran información sobre el perfil de una persona que potencialmente pueda abandonar.

Comenzamos analizando la relación que tienen las variables categóricas binarias *(Figura 1)*, es decir que solo contienen dos valores (0 : No – 1 : Si) con respecto a si el usuario abandona o no.

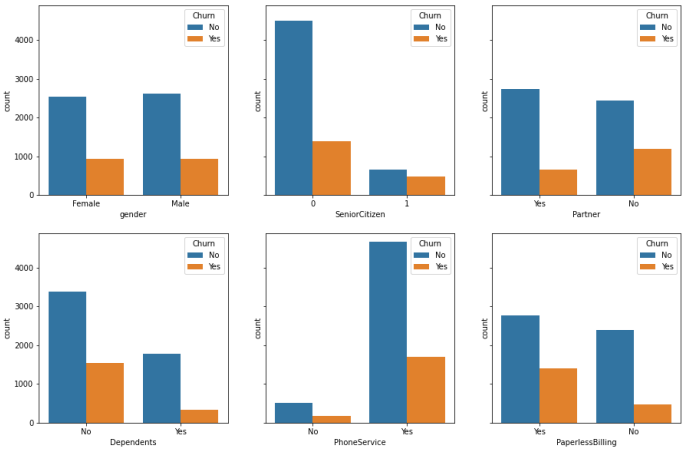


Figura 1

A primera vista, resalta un gran desequilibrio en las variables *SeniorCitizen* y *PhoneService*, por lo que se puede determinar que la mayoría de las personas en el conjunto de datos no son adultos mayores y tienen un servicio telefónico. Por el contrario, la tasa de abandono promedio para las variables *gender* y *Partner* es prácticamente la misma. Como conclusión de estas variables, que en su mayoría son información demográfica, podemos destacar a los adultos mayores, sin pareja y sin dependientes como un perfil de personas propensas a abandonar.

A continuación, en la *figura 2* encontraremos las gráficas relacionadas a los servicios que ofrece la compañía.

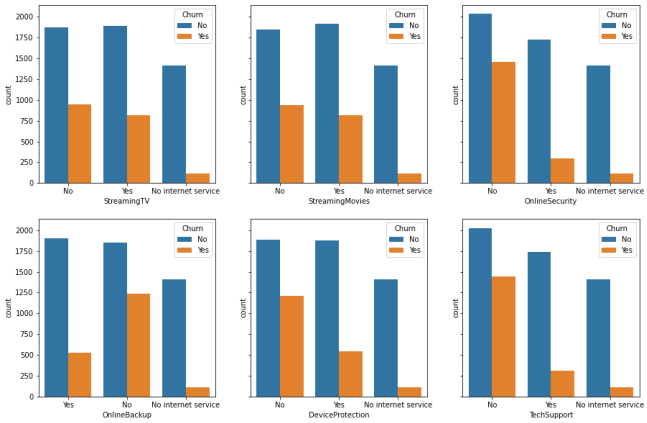


Figura 2

Observando los resultados, se evidencia que aquellos usuarios que hayan contratado los servicios de OnlineSecurity, OnlineBackup, DeviceProtection y TechSupport tienen una menor tasa de abandono, lo cuál es un factor que indica que la compañía está brindando un buen servicio en estas categorías, por otra parte los servicios de Streaming, ya sea de TV o de Movies se distribuyen igual entre si tienen o no el sercivio, por lo que no son relevantes en la tasa si el cliente abandona o no.

Finalmente procedimos a analizar la influencia que tenía el tipo de contrato *(figura 3.1),* el método de pago *(figura 3.2)* y la cantidad de meses que los clientes llevan en la compañía *(figura 3.3).*

Observamos que entre mayor es el tiempo de contrato menor es la tasa de abandono, por lo tanto, la empresa debe buscar una manera de asegurar que los clientes realicen contratos mínimamente de un año, para evitar la facilidad de abandonar al suscribirse a los servicios mes a mes.

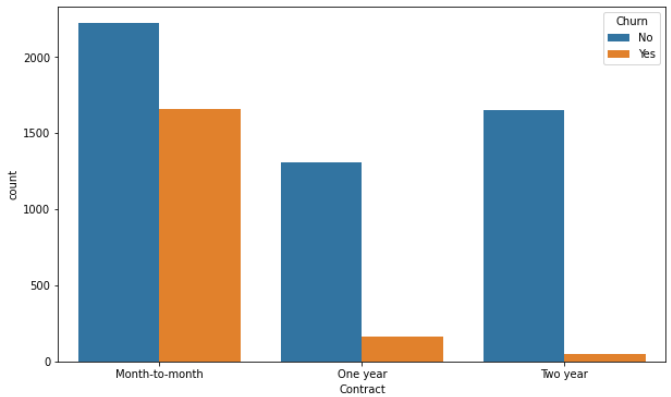


Figura 3.1

La empresa debe evaluar qué sucede con el método de pago Electronic check, ya que es el método más usado y, sin embargo, el que mayor tasa de abandono presenta, quizás evaluar si hay alguna falla o algún factor que pueda generar disgusto o malas experiencias entre los usuarios.

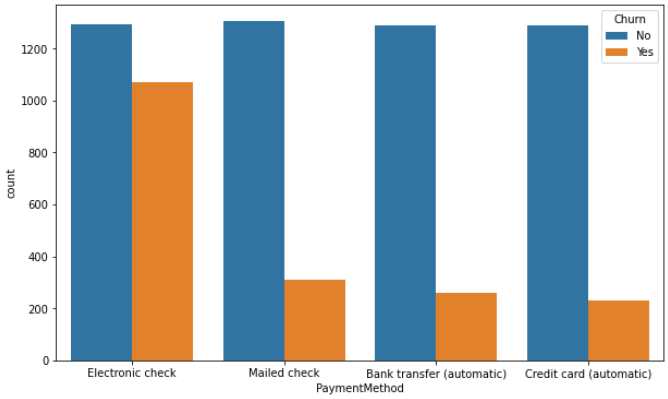


Figura 3.2

Con respecto a la permanencia de los clientes en la empresa, se puede apreciar que la mayoría de los clientes son nuevos (llevan pocos meses afiliados) o llevan mucho tiempo afiliados a la empresa, por lo que se debe hallar una manera de que los clientes superen el año afiliados y así asegurar una mayor tasa de fidelización.

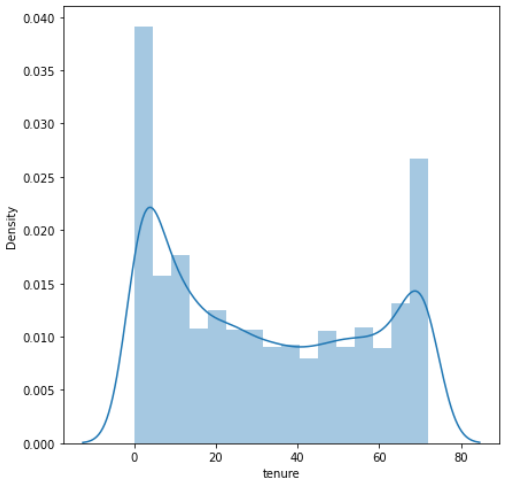


Figura 3.3

VII. ENTRENAMIENTO DE LOS MODELOS

Teniendo en cuenta la investigación que realizamos sobre los posibles modelos de machine learning que se pueden aplicar sobre problemas de este tipo (clasificación binaria), optamos por elegir e implementar cuatro modelos de los que hablaremos a continuación, dando un breve resumen de como funciona el algoritmo y finalmente, los resultados obtenidos al aplicarlos para este problema.

Para realizar el entrenamiento de los modelos primero debemos de borrar las columnas que no vamos a utilizar, en este caso es el customerID, pues no representa nada de importancia para nuestro modelo. Posteriormente se realizó un cambio en todas las variables de tipo “object”, buscando la mejor manera para representarlos de forma numérica.

Para ello revisamos nuevamente cada atributo con el fin de encontrar formas para poder realizar separaciones *(figura 4).*

```
gender: ['Female' 'Male']
Partner: ['Yes' 'No']
Dependents: ['No' 'Yes']
PhoneService: ['No' 'Yes']
MultipleLines: ['No' 'Yes']
InternetService: ['DSL' 'Fiber optic' 'No']
OnlineSecurity: ['No' 'Yes']
OnlineBackup: ['Yes' 'No']
DeviceProtection: ['No' 'Yes']
TechSupport: ['No' 'Yes']
StreamingTV: ['No' 'Yes']
StreamingMovies: ['No' 'Yes']
Contract: ['Month-to-month' 'One year' 'Two year']
PaperlessBilling: ['Yes' 'No']
PaymentMethod: ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
Churn: ['No' 'Yes']
```

Figura 4. Objetos encontrados en cada atributo.

Se realizó un cambio en el género, tomando ‘Female’ como 1 y ‘Male’ como 0. Cambiamos todos los ‘Yes’ y ‘No’ por 1 y 0 respectivamente y por último separamos en nuevas columnas las opciones de las columnas "internetService", "contract" y "PaymerMethod", indicando con 1 si cuenta con el servicio y con 0 de lo contrario, quedando así todo en términos numéricos. Por último, escalamos los variables continuas 'tenure', 'MonthlyCharges' y 'TotalCharges' para poder realizar comparaciones apropiadas.

A. *Regresión Logística*

El primer modelo que se implementó fue el de regresión logística, el cual se tomará como baseline, ya que es uno de los modelos de machine learning más simples y utilizados para clasificación binaria. Este modelo busca estimar la relación entre la variable binaria dependiente y las variables independientes.

	precision	recall	f1-score	support
0	0.83	0.89	0.86	1023
1	0.65	0.53	0.58	386
accuracy			0.79	1409
macro avg	0.74	0.71	0.72	1409
weighted avg	0.78	0.79	0.79	1409

Figura 5. Métricas obtenidas en el modelo de Regresión Logística.

B. *Support vector machine*

Para eficiencia del modelo vamos a tomar los cambios realizados en el modelo de redes neuronales, teniendo eso en cuenta, podemos iniciar con la explicación.
Lo primero que realizamos es la configuración del clasificador, en el cual le vamos a colocar que tenga un escalado automático y que la margen que se toma también sea automática, para que se adapte a los cambios del modelo.
Ya teniendo eso, podemos entrenar el modelo y al momento de evaluarlo tenemos que nos da una precisión bastante buena, tanto para los clientes que van a renunciar como los que no.

	precision	recall	f1-score	support
0	0.86	0.92	0.89	1023
1	0.74	0.59	0.66	386
accuracy			0.83	1409
macro avg	0.80	0.76	0.77	1409
weighted avg	0.82	0.83	0.82	1409

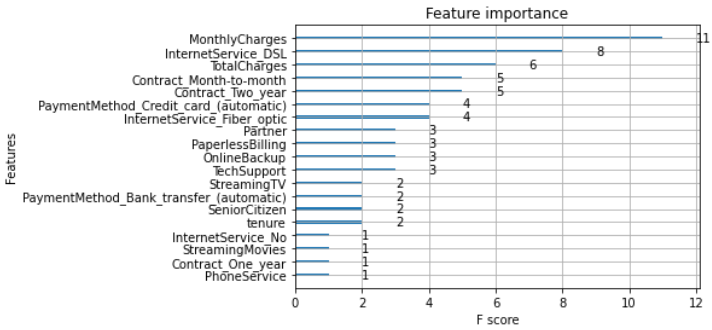
Figura 6. Métricas obtenidas en el modelo SVM

C. *XGBoost*

Este es un modelo basado en arboles de decisión, consiste en una secuencia de arboles denominada CART (“Classification and Regression Trees”). Los árboles se agregan secuencialmente a fin de aprender del resultado de los árboles previos y corregir el error producido por los mismos, hasta que ya no se pueda corregir más dicho error (esto se conoce como “gradiente descendente”)

```
Train accuracy XGB:  0.8068867589634363
Test accuracy XGB:  0.7927608232789212
Recall Score:  0.6787564766839378
```

Además, este modelo arroja información importante sobre cuáles son las características más importantes para realizar una predicción.



D. *Redes neuronales*

Este modelo consta de 3 capas, en la primera capa van a haber 26 nodos, en la segunda 15 y en la de salida 1 nodo, esto con el fin de que el aprendizaje automático sea más rápido y eficiente, además de eso se va a utilizar el optimizador ‘adam’ el cual se va a ajustar automáticamente para determinar cuánto va mermar o subir cada predicción, este proceso se va a repetir por 500 veces, cada vuelta las redes se ajustan para generar un valor más parecido con la retroalimentación, la cual lo determina la precisión.

Ya realizado el entrenamiento nos damos cuenta de que los valores de salida son decimales entre el 1 y 0, así que creamos un sesgo en el cual si está por encima de 0.5 lo tome como 1 y si no, entonces 0. Al realizar el cambio, ejecutar los elementos de prueba y evaluar el modelo se obtiene una precisión del 83% para calcular los clientes que se quedan, lamentablemente no es tan bueno para predecir los usuarios desean salir.

	precision	recall	f1-score	support
0	0.83	0.85	0.84	1023
1	0.58	0.55	0.57	386
accuracy			0.77	1409
macro avg	0.71	0.70	0.70	1409
weighted avg	0.76	0.77	0.77	1409

Figura 6. Métricas obtenidas en el modelo de red neuronal

VIII. Mejora de datos

Al revisar la distribución del conjunto de datos nos dimos cuenta de que se encontraba desbalanceado, pues mientras que datos sobre clientes que se quedaban había 5000 en el caso contrario solo habían 1800, significando que nuestros modelos no iban a funcionar como lo esperábamos, resultado que vimos plasmado en el entrenamiento de cada modelo.

Ilustración 1. Diferencia en el conjunto de datos

Para corregir este error teníamos dos opciones, la primera era encontrar un nuevo dataset, cosa que no podíamos hacer debido al tipo de problema al que nos enfrentábamos, y la segunda, que consistía en realizar un resampling para poder tener una mejor distribución y entrenar mejor el modelo.

Para realizar el resampling de forma adecuada y no caer en el sobre entrenamiento de los modelos con datos repetidos implementamos una forma en la que se generaran aleatoriamente datos en base a los que ya teníamos, permitiendo así realizar un nuevo entrenamiento a los modelos y mejorar la distribución de los datos.

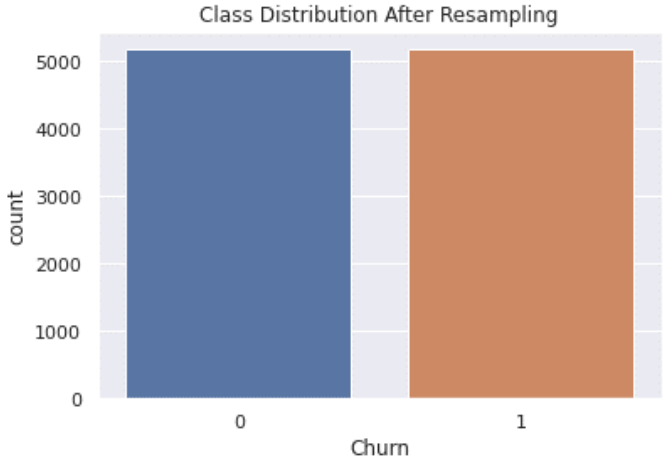


Ilustración 2. Datos despues resampling

IX. Nuevo entrenamiento de los modelos

Al realizar el cambio anteriormente mencionado revisamos todos los modelos para ver la mejoría que se había conseguido.

A. *Regresión Logística*

En la regresión logística tuvimos una mejoría al pasar de un 65% a un 72% con el nuevo entrenamiento, a primera vista se ve una mejoría de un 7% en la precisión que deseamos.

	precision	recall	f1-score	support
0	0.78	0.69	0.73	1040
1	0.72	0.80	0.76	1030
accuracy			0.75	2070
macro avg	0.75	0.75	0.75	2070
weighted avg	0.75	0.75	0.75	2070

Ilustración 3. Metricas Regresión Logística

Cambio considerable para nuestro problema, pero no lo suficiente como lo esperábamos.

B. *Support Vector Machine*

En cuanto al modelo SVM no vimos un cambio tan grande, pues pasamos de un 74% a un 79%, notando una mejoría de un 5%.

	precision	recall	f1-score	support
0	0.86	0.76	0.81	1040
1	0.79	0.87	0.83	1030
accuracy			0.82	2070
macro avg	0.82	0.82	0.82	2070
weighted avg	0.82	0.82	0.82	2070

Ilustración 4 metrias SVM

C. *XGBoost*

Al realizar el nuevo ajuste de hiper parámetros tenemos basado en el análisis de importancia.

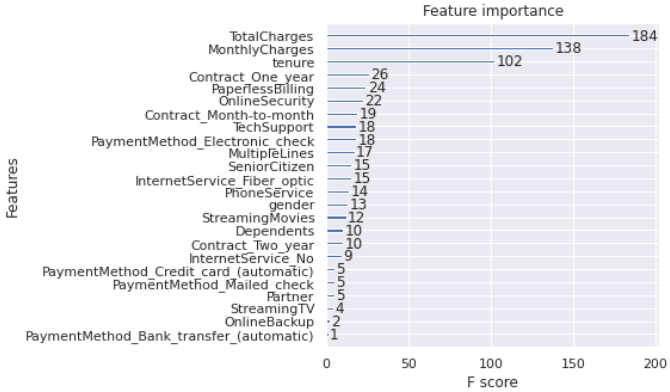


Ilustración 5. rasgos de importancia XGBoost

Tenemos que la mejoría del modelo es la más notable, pues ahora da una precisión de 84% en los clientes que se quieren salir del servicio, permitiendo así trabajar con más tranquilidad en cuanto a realizar predicciones.

	precision	recall	f1-score	support
0	0.96	0.82	0.88	1040
1	0.84	0.96	0.90	1030
accuracy			0.89	2070
macro avg	0.90	0.89	0.89	2070
weighted avg	0.90	0.89	0.89	2070

Ilustración 6. Métricas XGBoost

D. Redes Neuronales.

Al volver a entrenar el modelo de redes neuronales tenemos una mejoría de pasar de un 58% a un 75%.

	precision	recall	f1-score	support
0	0.96	0.82	0.88	1040
1	0.84	0.96	0.90	1030
accuracy			0.89	2070
macro avg	0.90	0.89	0.89	2070
weighted avg	0.90	0.89	0.89	2070

Ilustración 7. Métricas Redes neuronales

En conclusión, podemos ver una notable mejoría en todos los modelos al realizar el balanceo del conjunto de datos, dando como resultado una mejora en la precisión de todos los modelos.

X. DESPLIEGUE

Para realizar el despliegue utilizamos el modelo XGBoost por las métricas que pudimos evidenciar anteriormente. Optamos por realizarlo con Plotly y Dash, herramientas que facilitan el despliegue.

Ilustración 8. Imagen de inicio

En la ilustración 8 podemos ver como se ve el despliegue de la aplicación, dividimos el campo en dos partes, la primera es cuando se quiere agregar un cliente en especifico y saber si éste quiere terminar contrato, para ello debe de agregar todos los datos del cliente.

A. Predicción de un cliente

Primero se le va a preguntar sobre su información demográfica, datos como su género, si se encuentra jubilado, si está asociado y si es dependiente (ilustración 9).

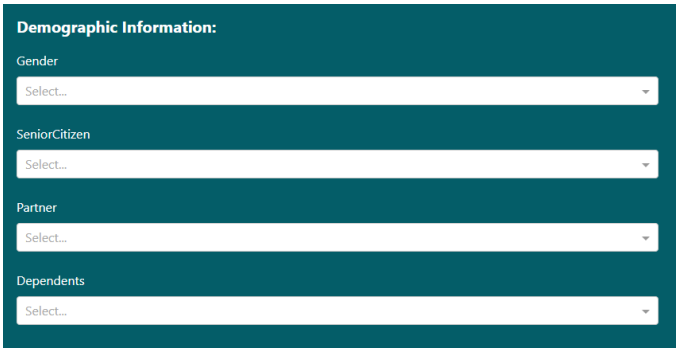


Ilustración 9. Información demográfica

Luego se pide los servicios asignados al cliente, los cuales pueden ser, servicio de teléfono, múltiple líneas, servicio de internet, seguridad online, backup online, protección de dispositivos, soporte tecnológico, streamingTV y StreamingMovies (Ilustración 10).

Ilustración 10. Servicios asignados

Por ultimo se requiere la información de la cuenta, los cuales son, el numero de meses que ha estado en la empresa, tipo de contrato (anual, mensual o cada dos años), facturación (anual, mensual o cada dos años), métodos de pago (tarjeta de crédito, pago electrónico, pago a través de correo, transferencia), el monto cobrado al cliente mensualmente y el importe total cobrado al cliente (ilustración 11).

Ilustración 11. Información de la cuenta.

Al cargar todos los datos del cliente se nos mostrara un mensaje con la información si el cliente va a abandonar o no.

B. Predicción de un conjunto de clientes

Primero se debe cargar el csv con todos los datos que deseamos explorar.

Ilustración 12. Carga de datos

La aplicación nos va a mostrar una visualización del csv con la predicción realizada.

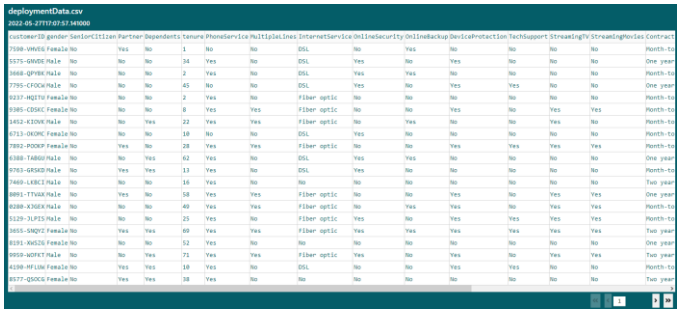


Ilustración 13. visualización del csv

Luego nos muestra los identificadores de los clientes y su respectiva predicción (ilustración 14).

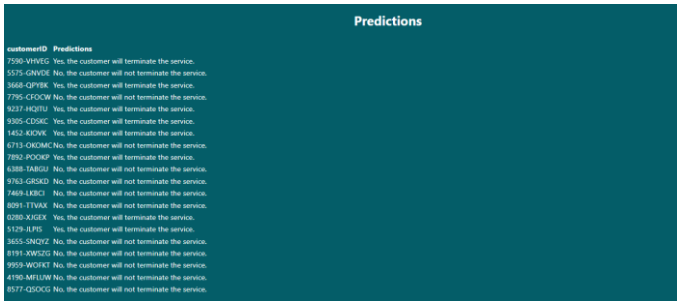


Ilustración 14. clientes y su predicción

Luego nos muestra graficas en cuanto a la distribución demográfica e información de la cuenta.

Ilustración 15. balance demográfico de la predicción

Ilustración 16. graficas balance información cuenta

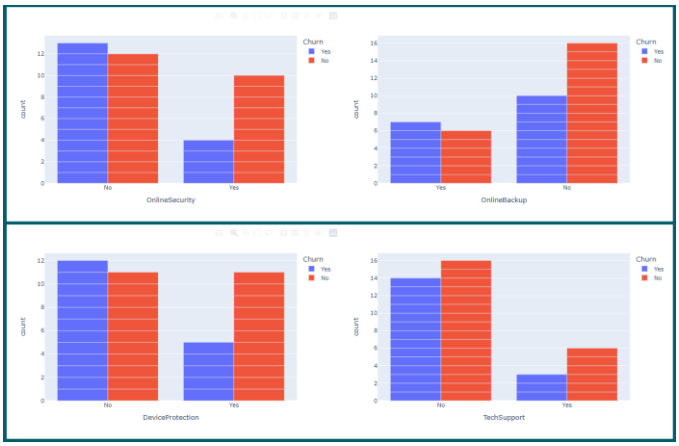


Ilustración 17. Graficas balance información cuenta