# Reporte Telecom X

La base de datos que se utilizará en este reto es la de Telecom X,  contiene información sobre clientes, suscripciones y servicios de telecomunicaciones. El objetivo es realizar un análisis exploratorio de los datos para identificar patrones y tendencias en el comportamiento de los clientes, así como para detectar las principales razones por las cuales los clientes deciden abandonar el servicio.

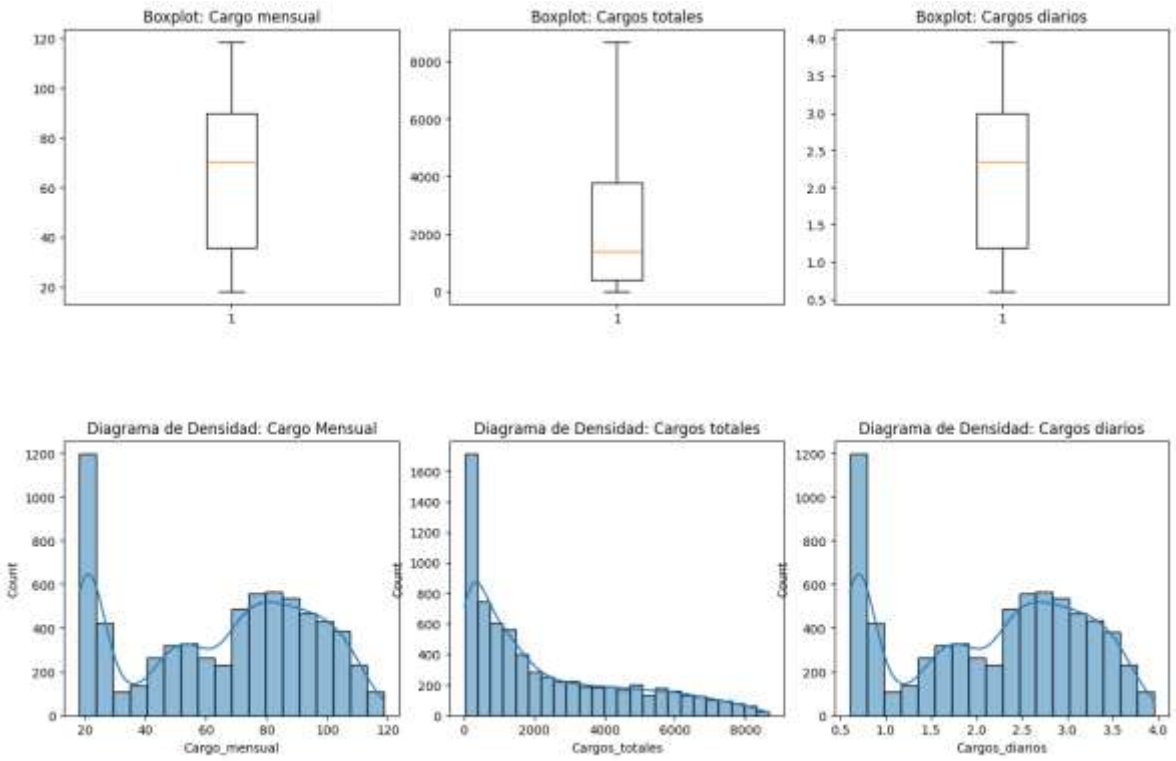
Para ello, se utilizarán técnicas de limpieza y transformación de datos, así como herramientas de visualización y análisis estadístico. El análisis se centrará en aspectos como la distribución de los clientes por tipo de suscripción, el uso de servicios, la satisfacción del cliente y la retención de clientes.

Al final tomando en consideración lo realizado en la parte de explocración de datos, se realizará un modelo de Machine Learning para predecir la probabilidad de que un cliente abandone el servicio.

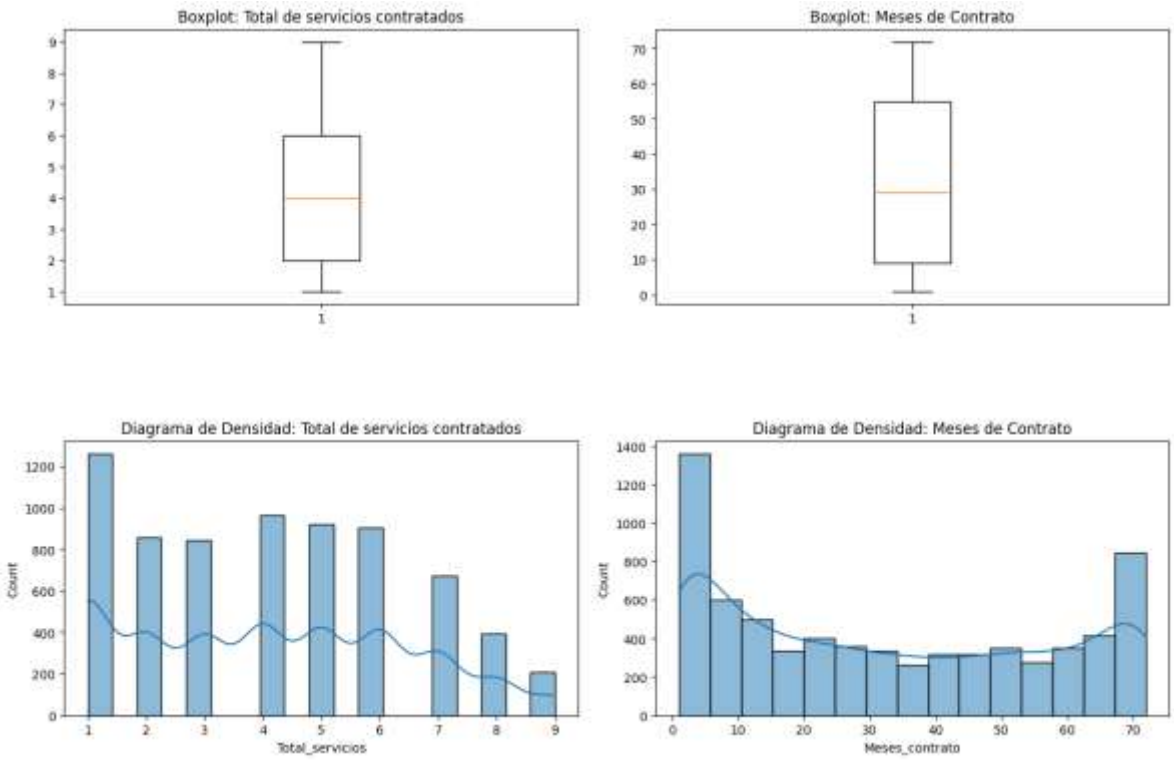
El propósito general de este reto es desarrollar un modelo de Machine Learning que permita predecir si un cliente abandonará el servicio de telecomunicaciones, basándose en las características de los clientes y sus patrones de uso.

## Principales hallazgos del análisis exploratorio

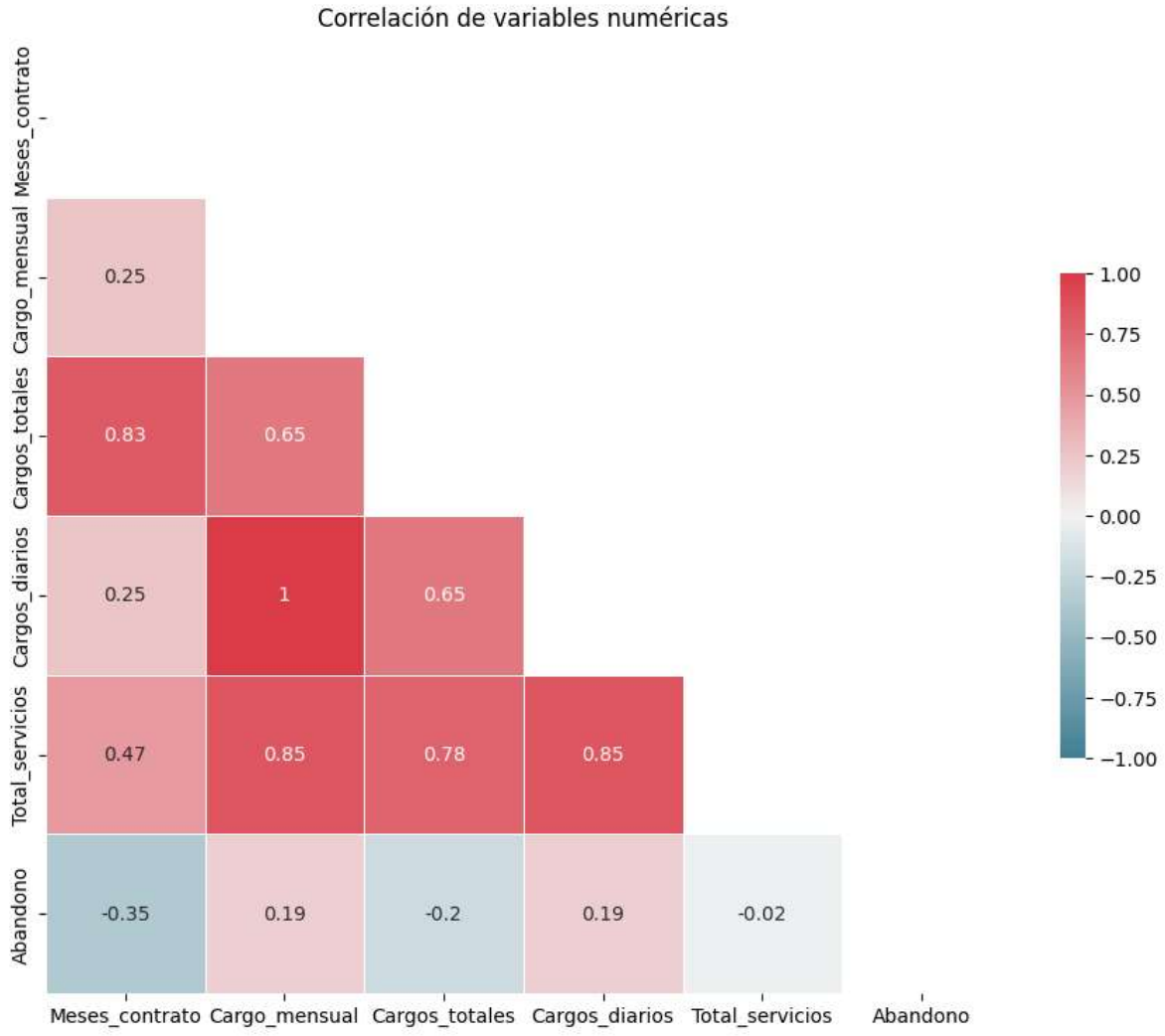
Al analizar la distribución de las variables numéricas encontramos que las que se refieren a los cargos del clientes encontramos en que su mayoría están cargados hacia la izquierda por lo tanto la gran mayoría de los clientes tienen bajos cargos, aunque podemos observar que se empieza a generar una especie de forma de campana en los cargos mensuales a partir de las 70 unidades.




Por otro lado podemos observar que el total de servicios varía mucho entre clientes, aunque en su mayoría las personas solamente contrataron un servicio. De igual forma podemos observar que los meses de contrato siguen una distribución bastante interesante ya que la mayoría de los clientes se encuentran en extremos (clientes con pocos meses contrato y clientes con muchos meses de contrato).



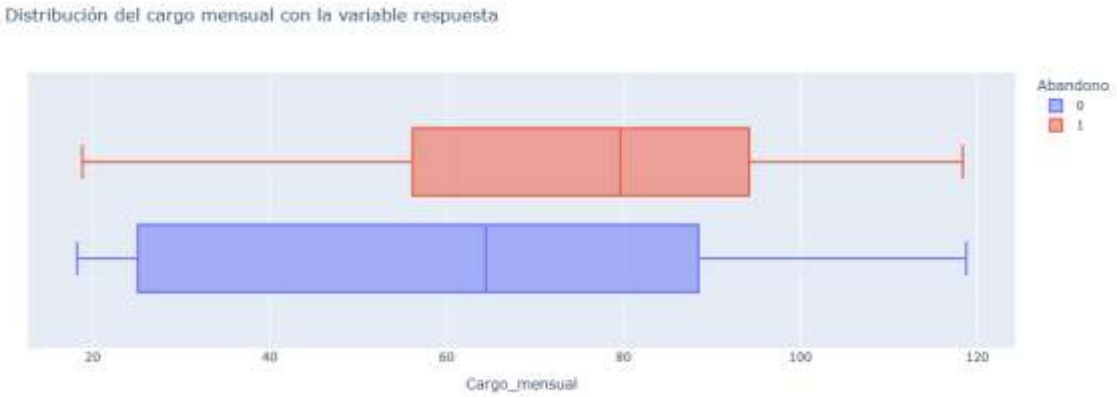

Al analizar la correlación de las variables numéricas con la variable respuesta (Abandono), encontramos una relación bastante débil entre ellas exceptuando por los meses de contrato que tiene una relación negativa con nuestra variable objetivo. Lo que quiere decir que entre mayor sean los meses de contrato menor será la probabilidad de abandono.



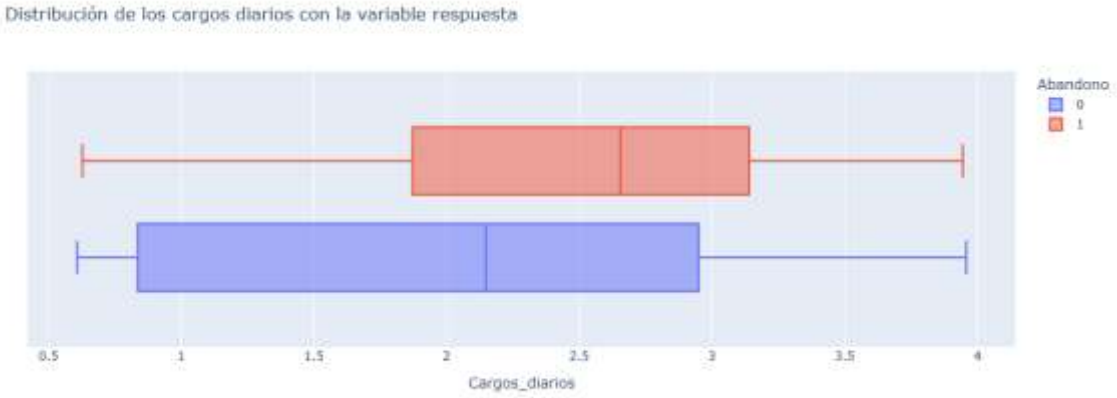

Al analizar las distribuciones específicas de nuestras variables numéricas con la de respuesta encontramos diferentes hallazgos interesantes.

Como se mencionó anteriormente encontramos que las personas que han abandonado la empresa tienen una menor cantidad de meses contratados, lo que nos lleva a suponer que las personas que tienen un mayor nivel de lealtad tienden a no abandonar la empresa.


Al revisar los cargos mensuales notamos que las personas que abandonaron la empresa tienden a tener mayores cargos que las personas que no lo han hecho.



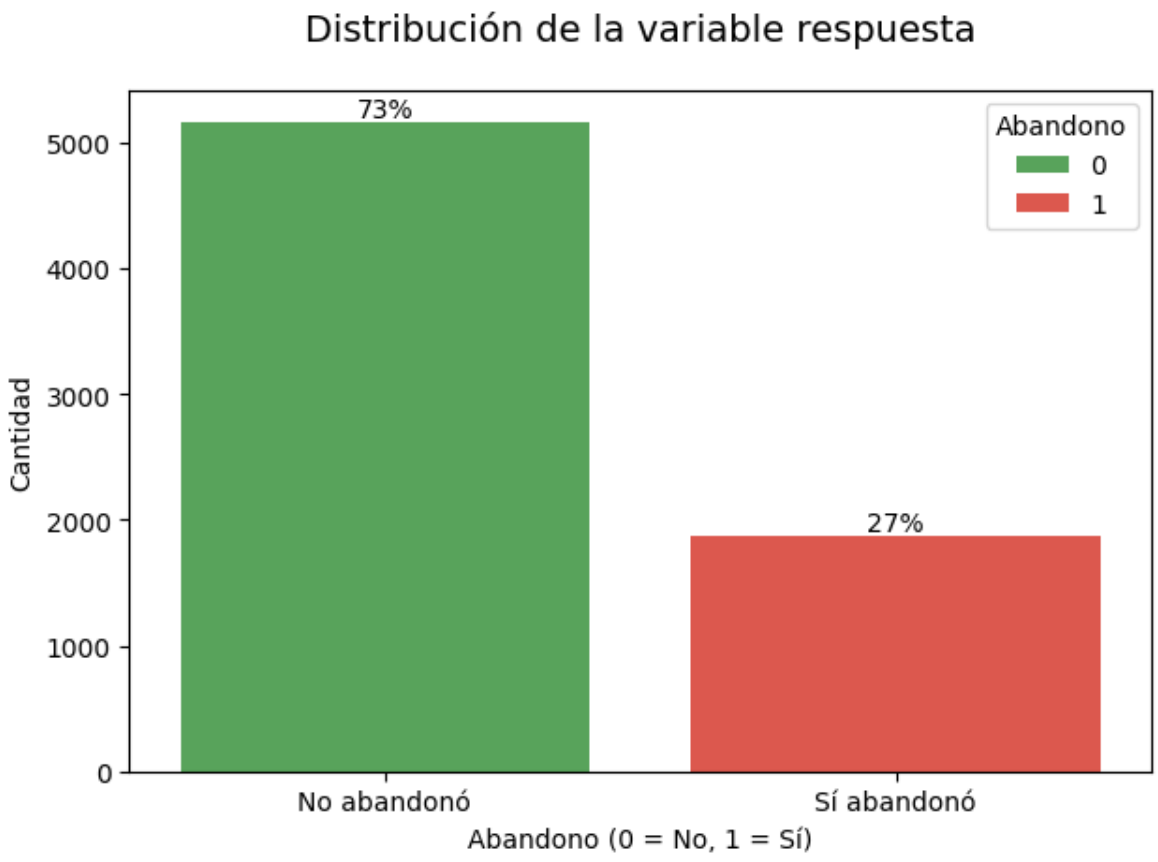

La distribución de los cargos diarios con la variable respuesta es igual que la distribución de los cargos mensuales, esto se debe a que esta variable fue creada por nosotros.




La distribución de los cargos totales con la variable respuesta demuestran que las personas que abandonaron la empresa tienden a tener un menor cargo total, lo cual parece un poco contradictorio con lo anterior que revisamos del cargo mensual.


Por último al revisar la distribución del total de servicios contratados con la variable respuesta encontramos que la gran mayoría de los clientes que abandonaron la empresa contratan entre 3 a 5 servicios, mientras que las personas que no abandonaron la empresa tienen una distribución más amplia (de 2 a 6 servicios contratados).


La distribución de nuestra variable respuesta demuestra que nuestra base de datos cuenta con más clientes que no han abandonado la empresa que los clientes que si lo han hecho. Debido a este suceso se tendrá que aplicar una técnica de balanceo para entrenar de mejor al modelo y este pueda identificar mejor a los clientes que han abandonado la empresa.



## Modelo de predicción y principales hallazgos del modelo de predicción

Después de probar tres modelos KNN, Árbol de decisiones y Random Forest, además de optimizar sus hiperparámetros a optimizar la predicción de personas que abandonaron la empresa. Encontramos que el mejor modelo es el árbol de decisiones.

Resumen de los modelos

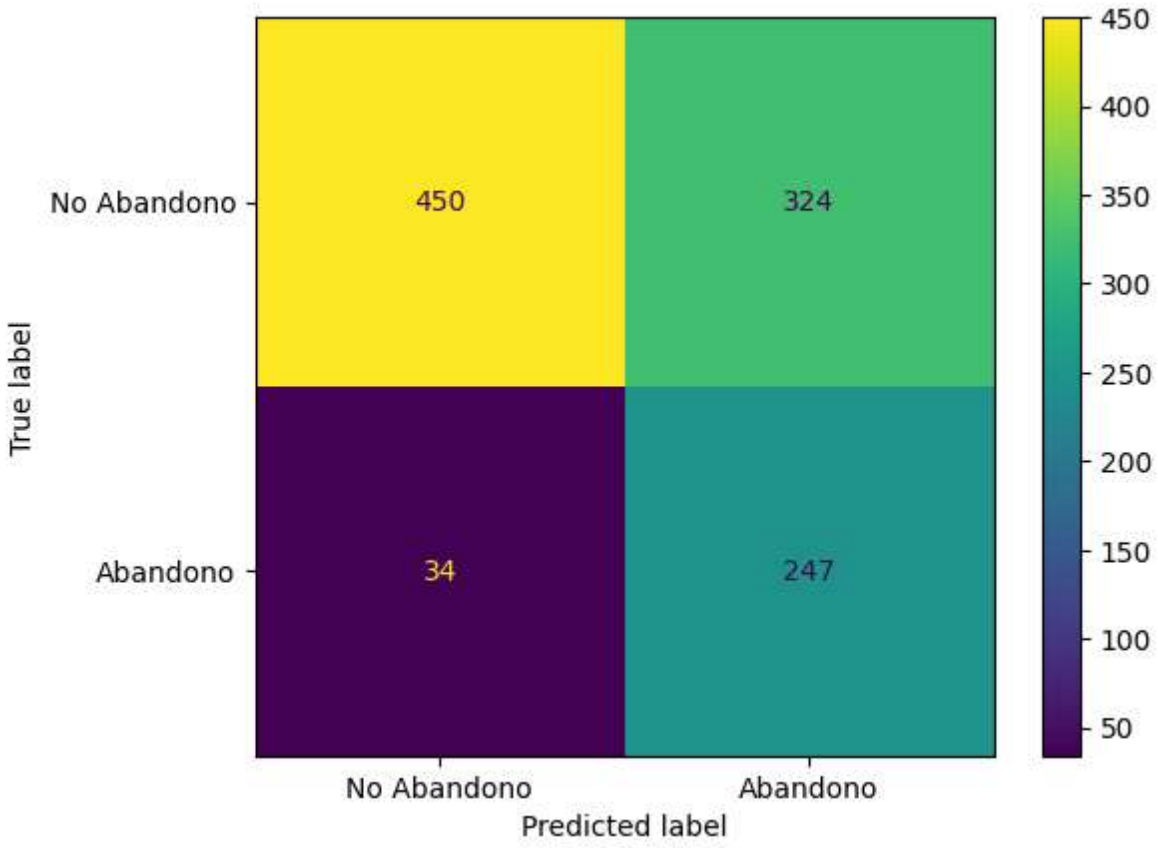
| **Modelo** | **Exactitud** | **Precisión** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Árbol de decisiones con hiperparámetros | 0.661 | 0.433 | 0.879 | 0.580 |
| KNN con hiperparámetros | 0.734 | 0.500 | 0.769 | 0.606 |
| Random Forest con hiperparámetros | 0.778 | 0.563 | 0.747 | 0.642 |

**El recall hace referencia a la proporción de datos positivos (personas que abandonaron) correctamente predichas*.

Me gustaría resaltar que ciertas variables se eliminaron a la hora de entrenar los modelos debido a que presentaron multicolinealidad, las variables que se eliminaron fueron:

- Cargo mensual
- Total_servicios (debido a una baja correlación con la variable respuesta)
- Tipo_servicio_internet_Sin servicio
- Cargo_mensual
- Tipo_contrato_Mes a mes
- Metodo_pago_Cheque por enviado por correo
- Cargos_totales

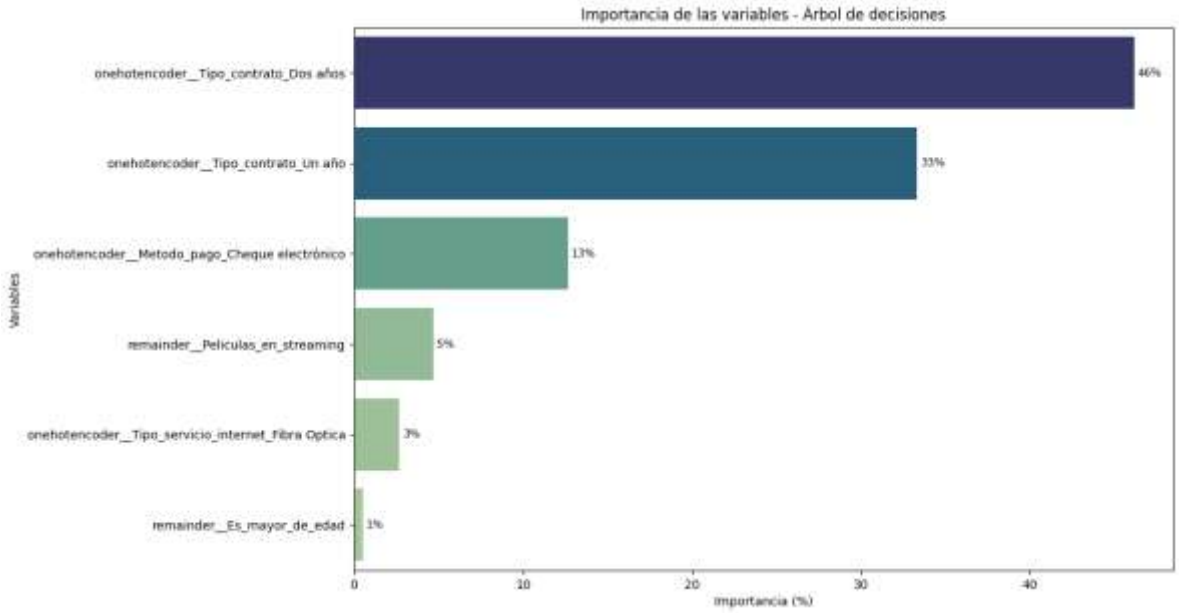
Matriz de confusión con datos de prueba no vista por el modelo.




Podemos darnos cuenta que el modelo tiene una alta capacidad para predecir a las personas que van a abandonar la empresa y es buena para evitar una clasificación incorrecta. Lamentablemente al buscar este objetivo el modelo se vio comprometido y tiende a clasificar mal a las personas que no van a abandonar la empresa, de igual forma no es el mejor modelo para predecir a las personas que no abandonaron la empresa.

Este compromiso se hizo pensando en el objetivo del proyecto el cual es priorizar a las personas que van a abandonar la empresa.

Al revisar la importancia de que el modelo les dio a las variables, notamos que las principales son las referentes a un tipo de contrato, de igual forma otra se enfoca en un método de pago, al servicio de streaming, la fibra óptica y la edad del cliente.




## Estrategias

La principal estrategia que le recomendamos a la empresa que siga es una que busque mejorar la lealtad con los clientes. Esto principalmente porque sabemos que este factor es altamente determinante en nuestra variable respuesta y esto se comprobó con la importancia que le da el modelo a las variables referentes con el tipo de contrato que tiene el cliente con la empresa.

Con esta campaña se espera mejorar el servicio para hacer que los clientes duren más tiempo en el servicio y con ello puedan quedarse en la empresa, esto debido a que existe una pequeña correlación de los clientes con mayor tiempo de contrato y su permanencia en la empresa.

La campaña se puede enfocar en dar descuentos en el servicio, entrar en un programa de atención personalizada o acceder a ciertos servicios de manera gratuita.

## Conclusiones

El modelo presento una buena capacidad para predecir a las personas que van a abandonar la empresa y consideramos que este compromiso es óptimo considerando el objetivo de la empresa.

La estrategia propuesta busca lograr reducir el número de las personas que abandonan la empresa basándonos en la correlación de los meses de contrato con la permanencia y la importancia que les dio el modelo a las variables referentes con la duración del contrato.

Por último consideramos importante realizar una investigación más a fondo de las personas que abandonaron la empresa y encontrar las principales razones por las cuales la abandonaron y así buscar otras variables que puedan mejorar la eficacia del modelo.