INSTITUTO POLITÉCNICO NACIONAL

Escuela Superior de Cómputo

Academia de Ingeniería de Software

Primer Avance de la propuesta de proyecto

Sistema de Satisfacción

Asignatura:

Introducción a la Ciencia de Datos

Equipo:

2

Integrantes:

- De Luna Ocampo Yanina *2020630226*
- Ramírez Méndez Kevin *2020630428*
- Salinas Velazquez Jacob *2020630397*

Fecha de entrega:

23 de Junio del 2021

2. Descripción general del proyecto

Los negocios están avanzando a pasos agigantados, gracias al avance que se ha tenido en las Ciencias de la Computación en los últimos años y al avance tecnológico que estas nos han brindado, creando un proceso de compra extremadamente personalizado para cada usuario, analizando cada click dado, cuánto tiempo vemos un producto, si regresamos a verlo, si pasamos a otro producto, etc. Los clientes desean vivir experiencias de compra únicas, trabajar en esa dirección ayuda a conseguir la fidelización de los clientes con la marca. Toda la información que un usuario deja a través de su navegación por el uso de un servicio debe ser procesada para facilitar que ese cliente pueda gozar de una buena experiencia al uso de este y crear un servicio personalizado expresamente para él, todo esto con la mentalidad de mantener a los clientes consumiendolo. Llega a ser atractivo ya que puede hacerse las 24 horas del día, accediendo a los datos rápidamente, dando un servicio no solo en un área pequeña o delimitada, sino a una nacional o internacionalmente.

Para este tipo de empresas representa una problemática la adquisición de nuevos clientes, que conservar clientes antiguos, ya que estos significan un gasto mayor para la empresa, esto debido a factores como son: la creación de campañas de marketing, la implementación de infraestructura y nuevas tecnologías para poder cubrir la demanda de estos, de igual manera, incluyendo la solución efectiva y rápida de problemas planteados por los clientes.

Por los problemas anteriormente mencionados las empresas buscan la mejor forma para poder aumentar sus ganancias y en el peor de los casos mantenerlas, tomando esto en cuenta, en este proyecto pretendemos analizar la información sobre los clientes de una empresa de telecomunicación con la finalidad de poder clasificar si un cliente esté satisfecho con el servicio y que esto lo lleve a quedarse dentro de la empresa.

3. Justificación.

Partimos bajo la premisa de que las empresas en el área de las telecomunicaciones prefieren que los clientes sean leales a su servicio a obtener nuevos clientes, ya que esto muchas veces, como ya hemos mencionado, genera más gastos. Esto es aún más notable en el mercado de las telecomunicaciones.

Para lograr maximizar el número de clientes es importante no solo tratar de atraer nuevos, sino también retener los existentes. Además, un nuevo cliente puede mostrarse renuente a los servicios que se ofrecen, mientras que los antiguos clientes se quedan por la calidad de los servicios o por el beneficio que pueden ofrecerle por permanecer.

En consecuencia, pretendemos analizar los registros de usuarios de un servicio de telecomunicaciones, para poder describir el comportamiento que tienen los usuarios, los factores que influyen al decidir si cancelar o no el servicio,así llegar a un modelo de clasificación que nos ayude a saber si cierto usuario se quedará afiliado o no mediante la satisfacción de este. Con esto poder tomar medidas para poder conservar clientes y observar la tendencia que se tiene entre estos.

La presente investigación se enfocará en el análisis de datos sobre los clientes pertenecientes a una empresa de telecomunicaciones, ya que debido al reciente aumento de gente adquiriendo estos tipos de servicios y el aumento de empresas dentro del mismo ámbito, crea la pérdida y descontento de clientes. Así, el presente trabajo permitirá mostrar las características similares que comparten los clientes satisfechos con su servicio y los clientes perdidos.

4. Reglas del negocio

Con el análisis de los datos que brinde esta investigación se pretende describir los factores que afectan a la fidelidad de un usuario, en este caso de telecomunicaciones, también se pretende poder clasificar usuarios que se quedan o no con el servicio para, a partir de esto, poder tomar decisiones como empresa y poder ampliar el rango de usuarios fieles a la misma.

Algunos puntos que se podrían mejorar a partir de los resultados del análisis son:

La experiencia de usuario personalizada, con base en los patrones que se generan en el uso del servicio. Para poder lograr esta atención personalizada, debemos tomar en cuenta que los datos deben ser de calidad, la empresa debe tener una buena capacidad de analizar los datos, de lo contrario, retendremos pocos clientes y el éxito de la empresa será mínimo.

Como segundo punto, consideramos la atención al cliente. La empresa podrá brindar respuestas a las problemáticas que se lleguen a presentar. Un ejemplo de esto sería el uso de tecnologías como son los chatbots que son capaces de simular conversaciones precisas y naturales con los humanos por medio de internet utilizando conocimientos de Inteligencia Artificial y Procesamiento de Lenguaje Natural, esta propuesta llega a ser una alternativa eficiente.

5. Descripción del modelo de almacenamiento

Con el fin de poder realizar un análisis a un grupo de clientes de una empresa de telecomunicaciones. Las categorías de datos que ocupamos, son las siguientes:

• Clientes que han cancelado el servicio el último mes.

• Servicios a los cuales cada cliente se ha suscrito.

• Información bancaria del cliente.

• Información demográfica del cliente.

Además, hemos decidido filtrar y organizar los datos de la siguiente manera:

- Datos
- Metadatos
  - Descripción
  - Rango
  - Tipo

Obteniendo así, la siguiente tabla:

Dato	Descripcion	Rango	Tipo
customerID	ID del cliente		Polinomial
gender	sexo del cliente	(hombre/mujer)	Binomial
SeniorCitizen	si el cliente es jubilado	(1, 0)	Binomial
Partner	Si el cliente es casado	(Yes, No)	Binomial
tenure	Cuantos meses la persona ha sido cliente de esta compañía		Entero
PhoneService	El servicio telefonico esta conectado	(Yes, No)	Binomial
MultipleLines	Hay multiples lineas de telefono conectadas simultaneamente	(Yes, No, No Phone Service)	Polinomial
InternetService	Tipo de proveedor de internet del cliente	(DSL, Fiber optic, No)	Polinomial
OnlineSecurity	El servicio de seguridad de internet esta conectado	(Yes, No, No internet Service)	Polinomial
OnlineBackup	El servicio de respaldo esta conectado	(Yes, No, No internet Service)	Polinomial
DeviceProtection	El cliente tiene el equipo asegurado	(Yes, No, No internet service)	Polinomial
TechSupport	El soporte servicio de soporte tecnico esta conectado	(Yes, No, No internet service)	Polinomial
StreamingTV	El servicio de stream esta conectado	(Yes, No, No internet Service)	Polinomial
StreamingMovies	El servicio de cine esta activado	(Yes, No, No internet Service)	Polinomial
Contract	Tipo de contrato	(Month-to-month, One year, Two year)	Polinomial
PaperlessBilling	si el cliente utiliza facturación electrónica	(Yes, No)	Binomial
PaymentMethod	método de pago	cheque electrónico, cheque enviado por correo, transferencia bancaria (automática), tarjeta de	Polinomial
MonthlyCharges	pago mensual actual		Entero
TotalCharges	el monto total que el cliente pagó por los servicios durante todo el tiempo		Entero
Churn	si hubo una deserción	(Yes, No)	Binomial

Estos datos se obtuvieron de la Base de Datos de una empresa de Telecomunicaciones y dado a la naturaleza estructurada de los mismos se trabajarán en formato CSV.

6. Análisis Exploratorio de Datos

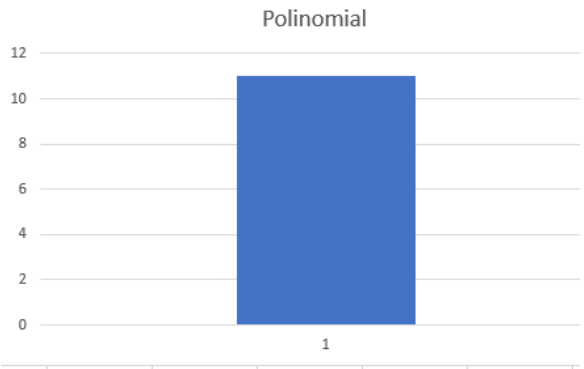
Con la finalidad de dar solución al problema que presenta la empresa de telecomunicación analizaremos un Dataset con los registros de una parte de sus clientes elegidos de forma al azar para evitar cualquier tipo de sesgo en nuestro espacio muestra. El tamaño de este dataset es de 5986 registros.

att1	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetServ...	OnlineSecur...	O
1869	7010-BRBUU	Male	0	Yes	Yes	72	Yes	Yes	No	No internet s...	N
4528	9688-YGXVR	Female	0	No	No	44	Yes	No	Fiber optic	No	Ye
6344	9286-DOJGF	Female	1	Yes	No	38	Yes	Yes	Fiber optic	No	N
6739	6994-KERXL	Male	0	No	No	4	Yes	No	DSL	No	N
432	2181-UAESM	Male	0	No	No	2	Yes	No	DSL	Yes	N
2215	4312-GVYNH	Female	0	Yes	No	70	No	No phone ser...	DSL	Yes	N
5260	2495-KZNFB	Female	0	No	No	33	Yes	Yes	Fiber optic	Yes	N
6001	4367-NHWMM	Female	0	No	No	1	No	No phone ser...	DSL	No	N
1480	8898-KASCD	Male	0	No	No	39	No	No phone ser...	DSL	No	N
5137	8016-NCFVO	Male	1	No	No	55	Yes	Yes	Fiber optic	Yes	Ye
3169	4578-PHJYZ	Male	0	Yes	Yes	52	Yes	No	DSL	No	Ye
4653	2091-MJTFX	Female	0	Yes	Yes	30	No	No phone ser...	DSL	No	N
2850	2277-DJJDL	Male	1	Yes	No	60	Yes	Yes	Fiber optic	No	N

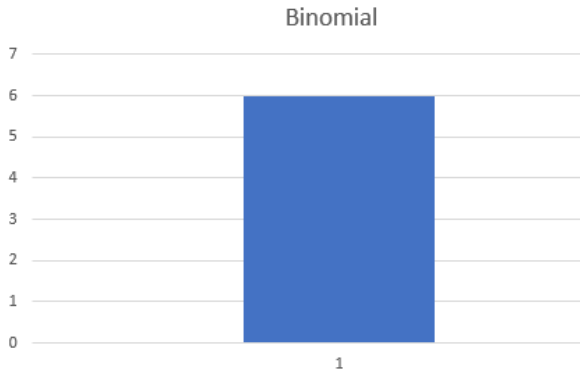
ExampleSet (5,986 examples, 0 special attributes, 22 regular attributes)

Frecuencia con la que se presentan los diferentes tipos de datos en nuestro análisis.

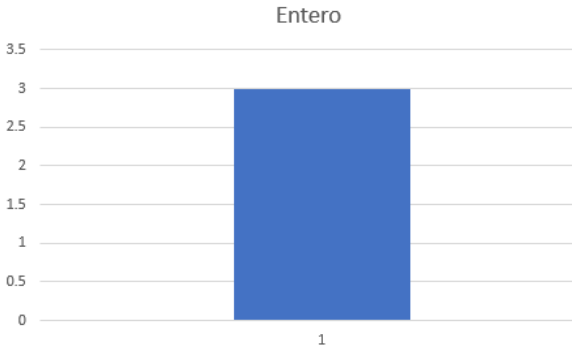
- Continuos (Binomial)
- Enteros
- Categóricos Nominales (Polinomiales)



Gráfica individual del tipo Polinomial



Gráfica individual del tipo Binomial

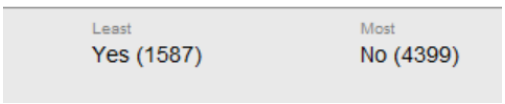


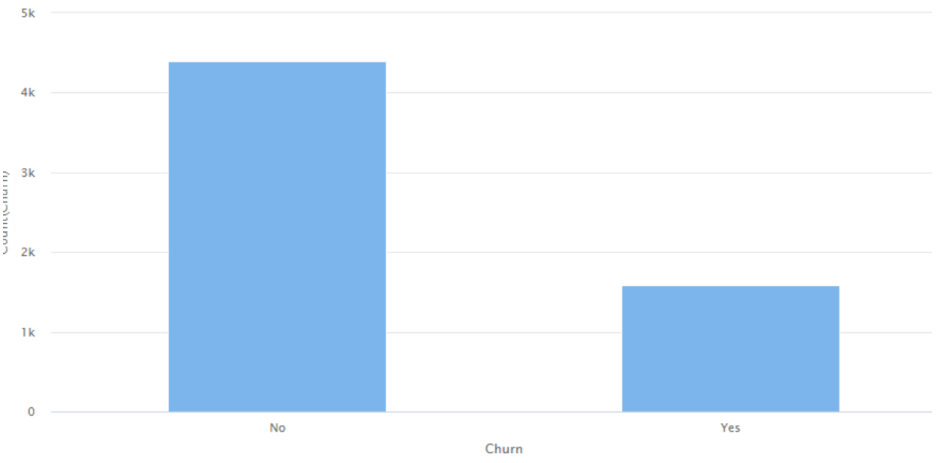
Gráfica individual del tipo Entero

Podemos observar claramente que la tendencia de nuestro análisis es a datos categóricos nominales siendo incluso ese el tipo de dato del valor que queremos modelar(**Churn**). Definimos **Churn** como el porcentaje del total de clientes de una empresa que deja de hacer negocios durante un periodo de tiempo específico, ayudando a evaluar el grado de satisfacción de los clientes.

Análisis del Atributo de Interés

Churn

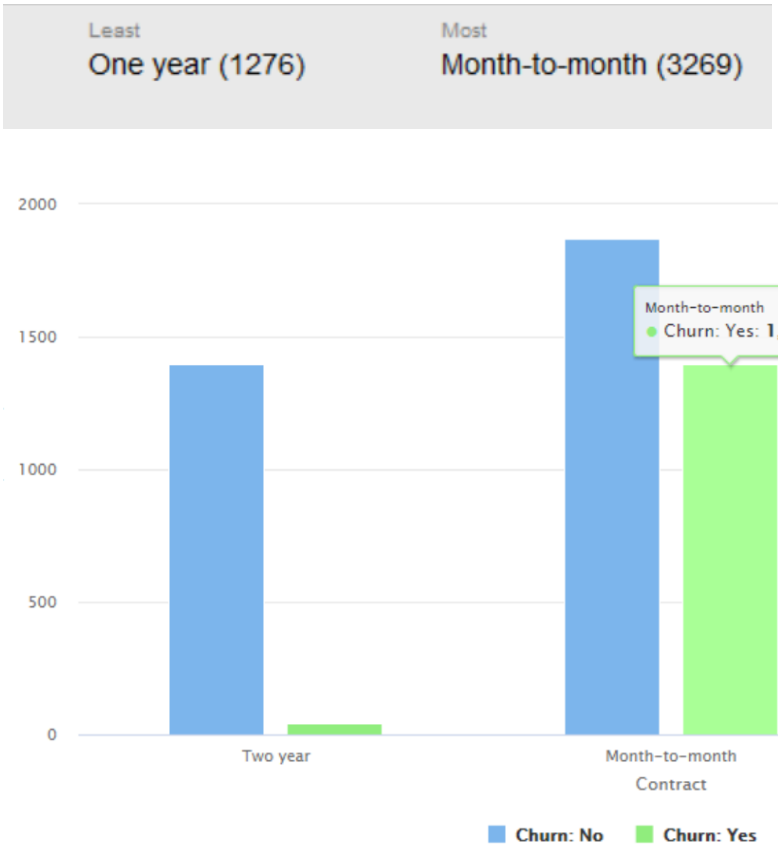




Se trata del atributo a modelar, podemos observar que en nuestro espacio muestra la gran mayoría de los clientes permanecen con el servicio.

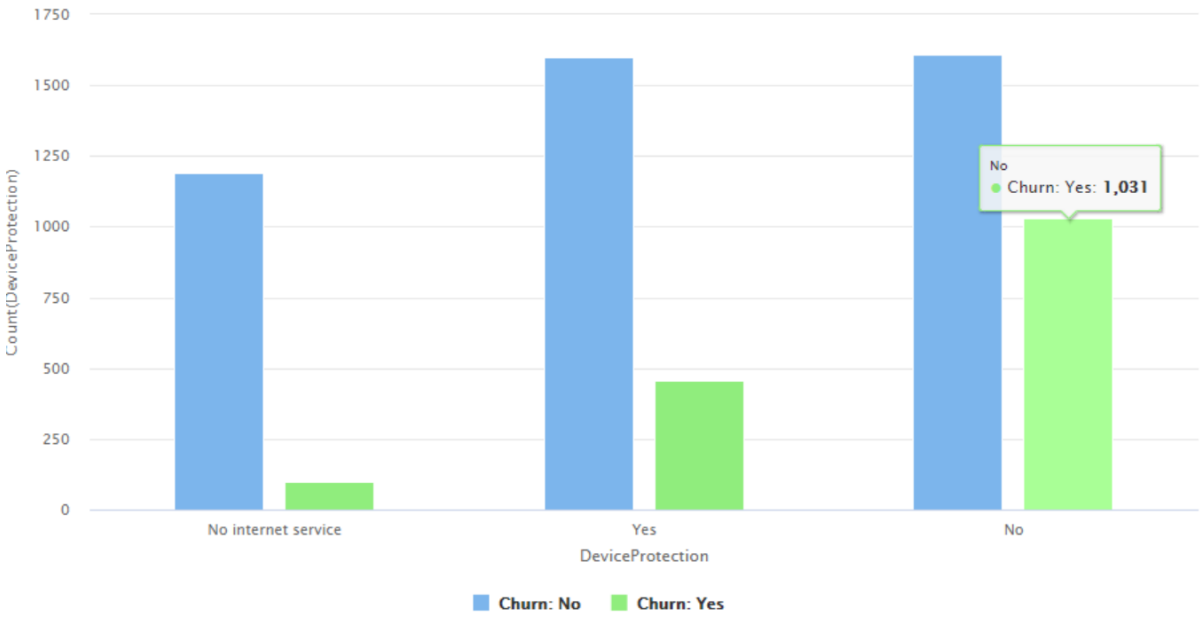
Análisis por Correlación con el atributo a modelar

Contract



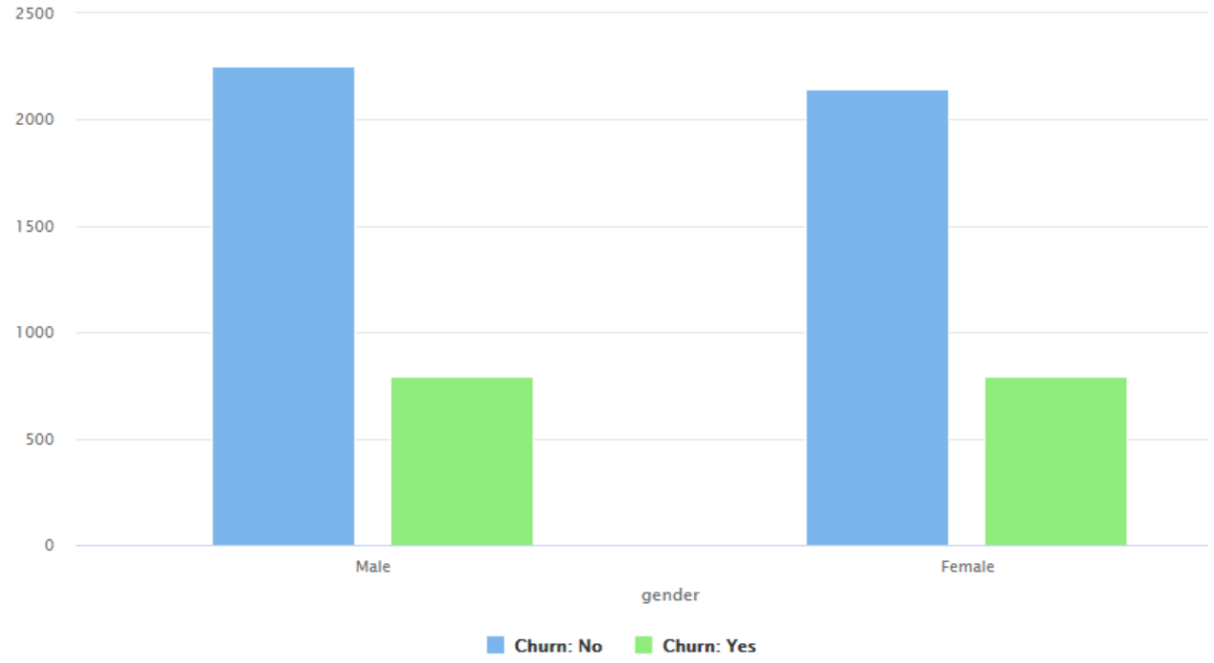
Casi todos los clientes tienen un contrato por mes y la mayoría de estos dejan el servicio. Esto cobra sentido si analizamos que es más difícil cancelar un servicio cuando lo pagas por año.

Device Protection



Hay una frecuencia semejante entre los usuarios que tienen sus dispositivos asegurados y los que no, de entre los últimos existe una tendencia a finiquitar su contrato con la empresa.

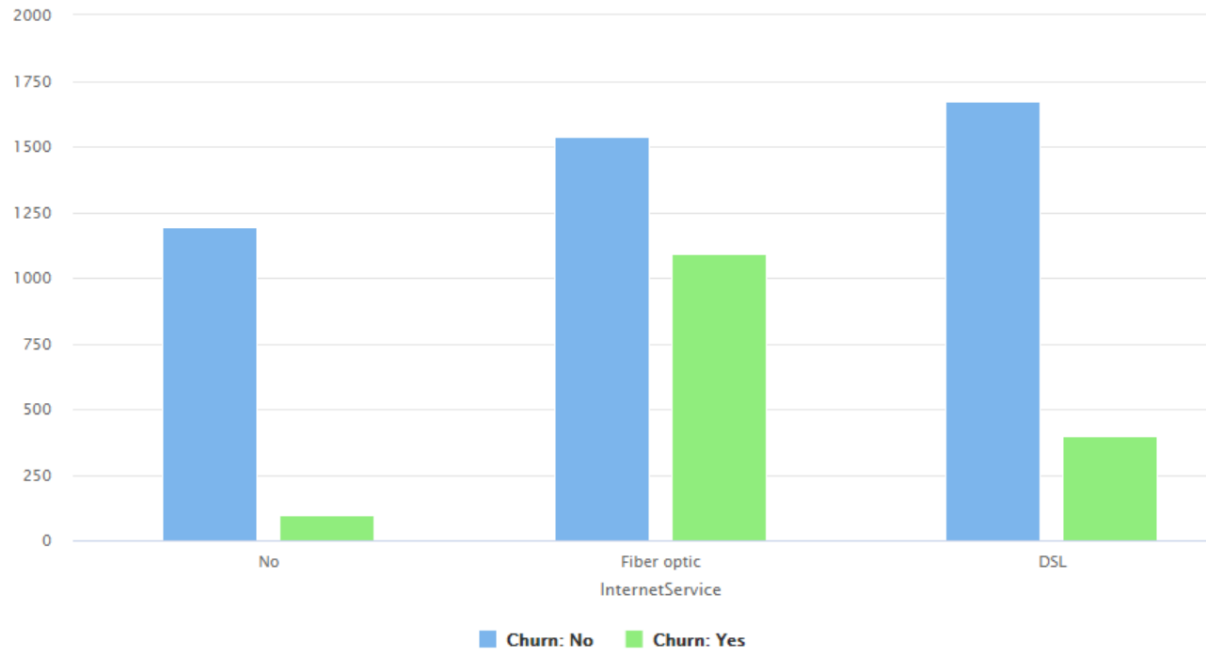
Gender



Se podría decir que no hay una fluctuacion relevante en el sexo de los clientes del espacio muestra. De la misma forma es evidente que el sexo no influye en la decision de seguir o no en la empresa.

Internet Service

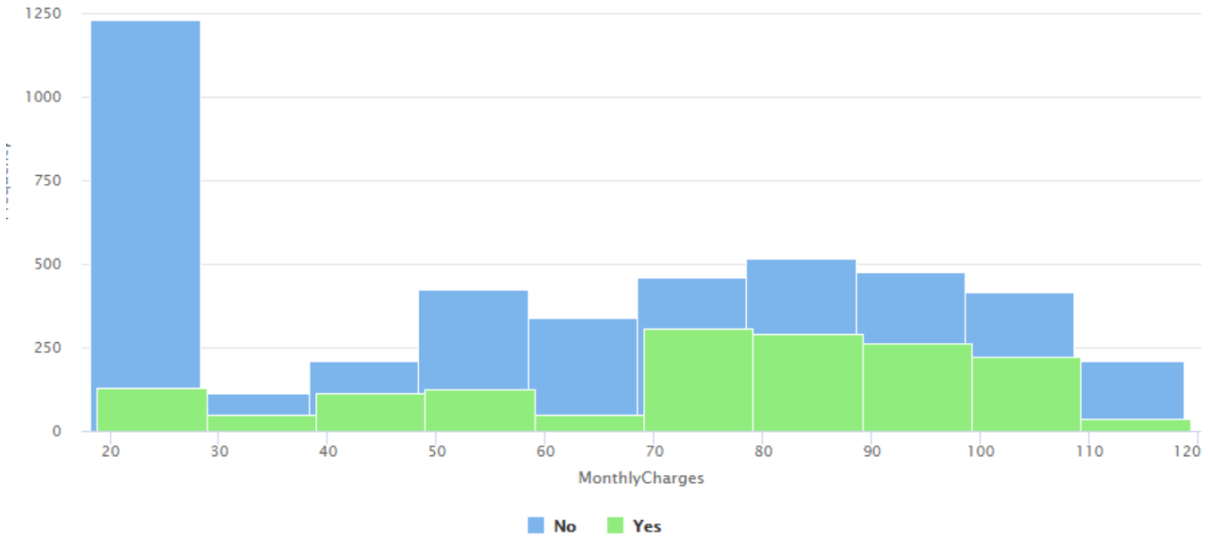
Least	Most
No (1291)	Fiber optic (2627)



Hay una cantidad muy parecida de usuarios que usan Fibra óptica y los que tienen DSL, sin embargo se nota una insatisfacción mayor con el servicio con los usuarios que usan fibra óptica.

Monthly Charges

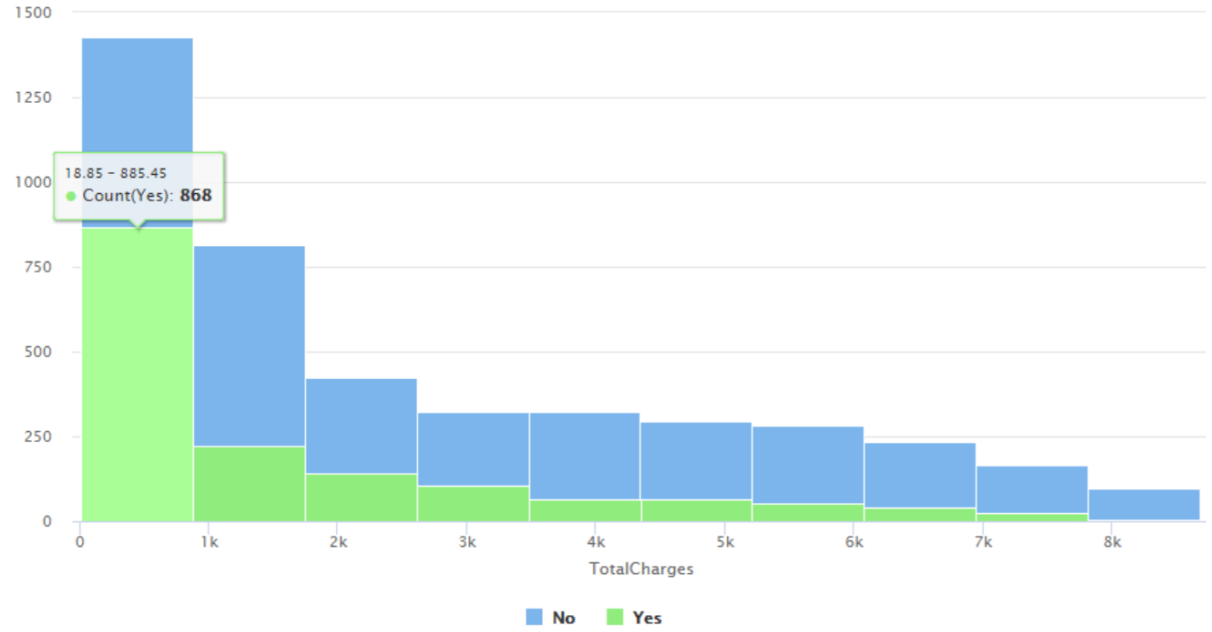
Min	Max	Average
18.250	118.750	64.802



Existe una distribución marcada entre los precios que pagan los clientes mensualmente. A simple vista no se nota una correlación directa entre el pago y la insatisfacción del cliente.

Total Charges

Min	Max	Average
18.800	8684.800	2298.061

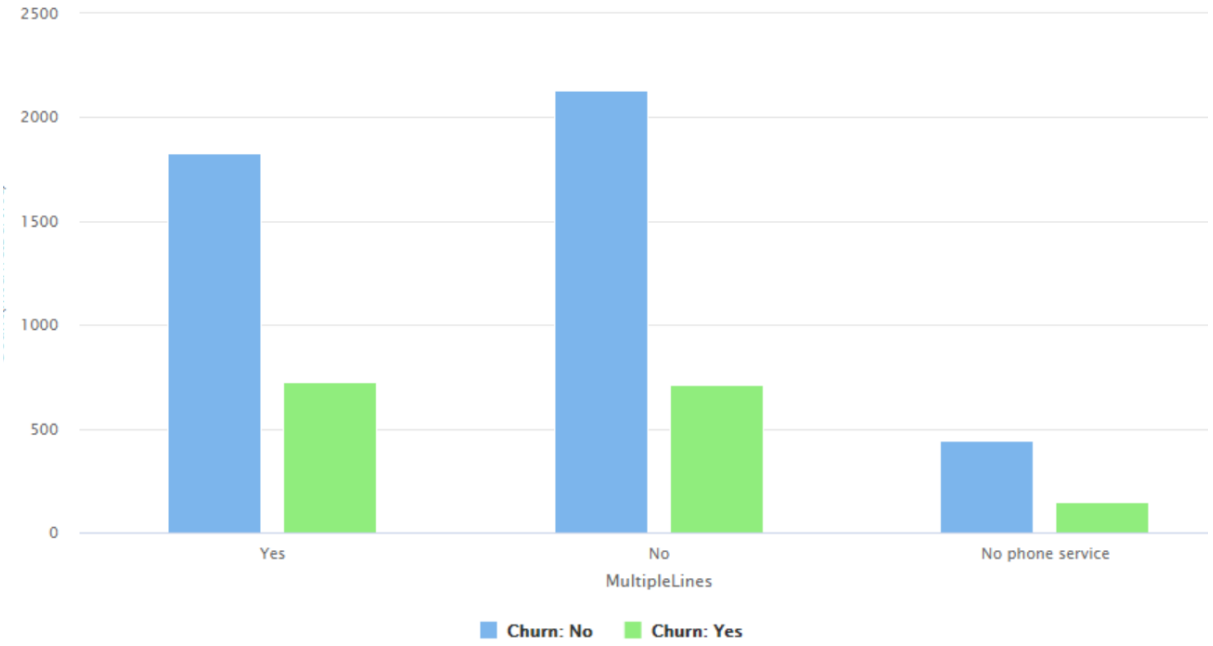


Los clientes que menos han pagado denotan la mayoría, esto se puede deber a que son clientes nuevos y a su vez estos son los que más tienden a dejar el servicio. Comparando con el número de registros notamos que hacen falta 10 datos(Como se muestra en la siguiente imagen).

Name	Type	Missing
TotalCharges	Real	10

Multiple Lines

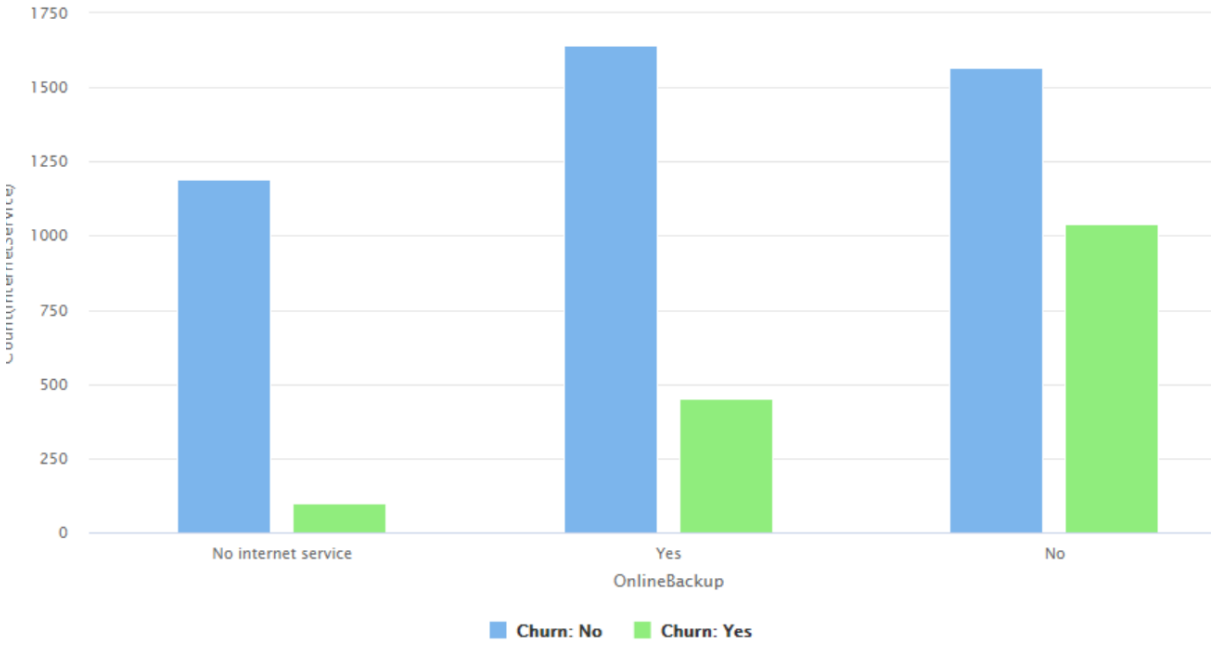
Least	Most
No phone service (590)	No (2848)



No se nota una relación directa entre la insatisfacción de los clientes y el hecho de que tengan 1 o varias líneas.

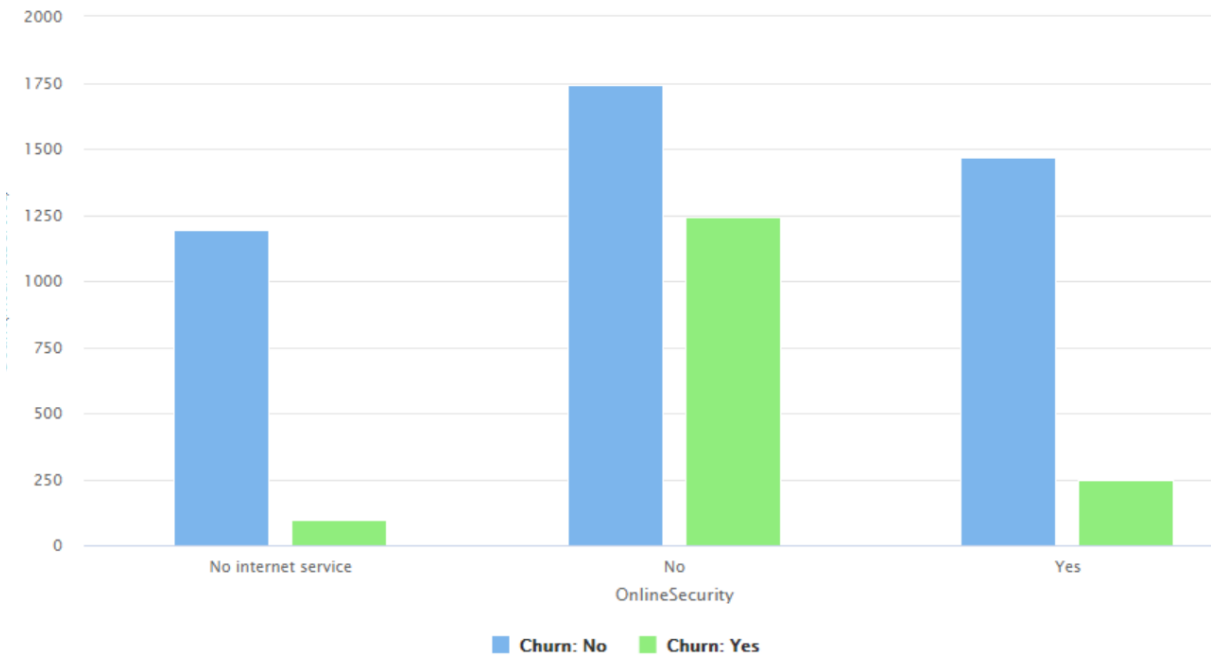
Online Backup

Least	Most
No inter [...] ce (1291)	No (2605)



Hay una cantidad parecida de clientes que tienen respaldo online y los que no, se puede determinar que la mayoría de clientes insatisfechos no tienen un respaldo.

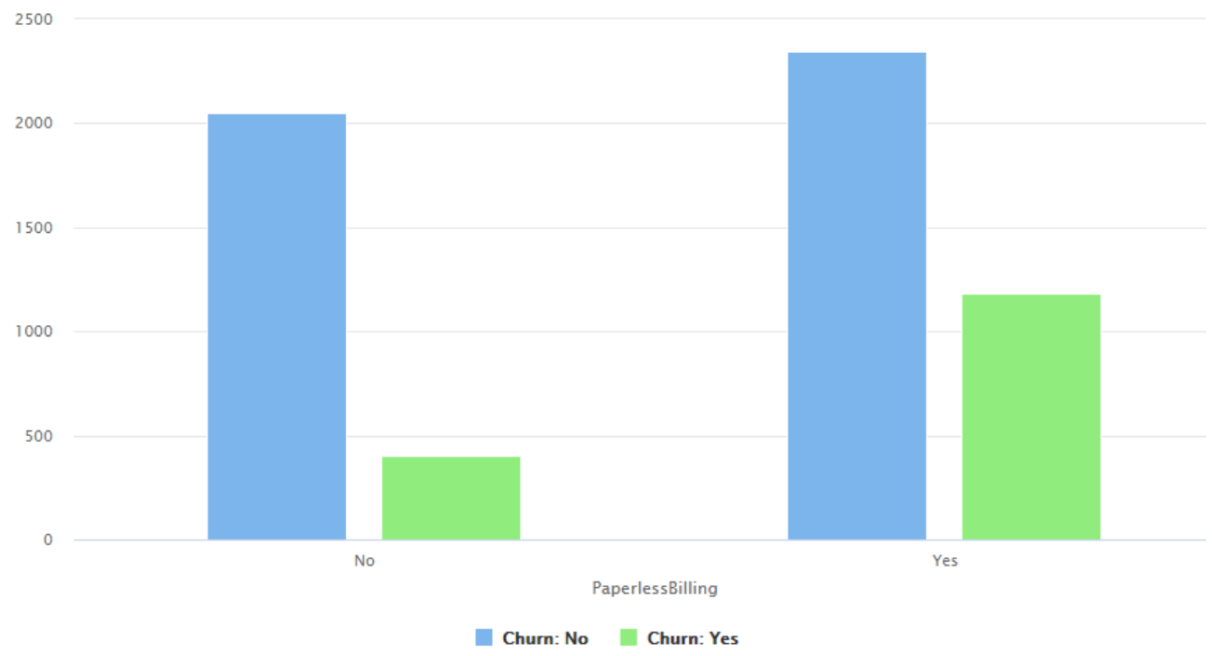
Online Security



En el espacio muestra que existen más clientes que no cuentan con un servicio de seguridad al navegar en linea, es este sector el que deserta más.

Paperless Billing

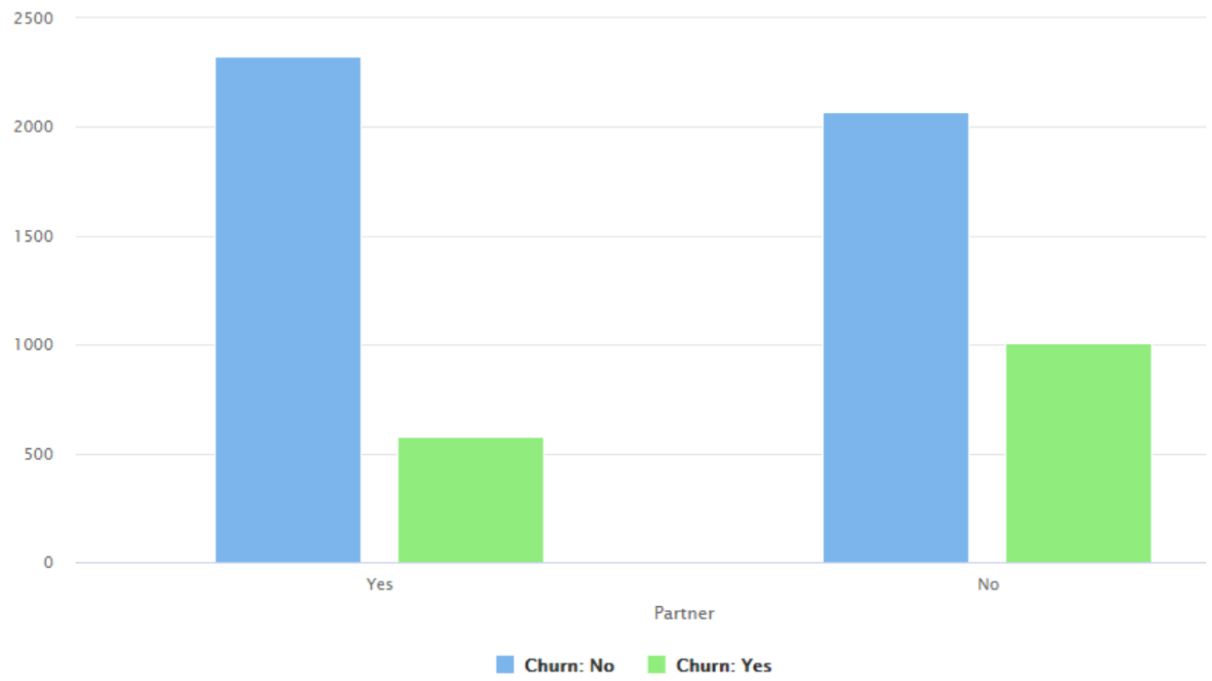
Least	Most
No (2458)	Yes (3528)



Hay un mayor sector de la población analizada que usa facturación electrónica. Este mismo sector es el que tiene una tendencia mayor a cancelar su contrato.

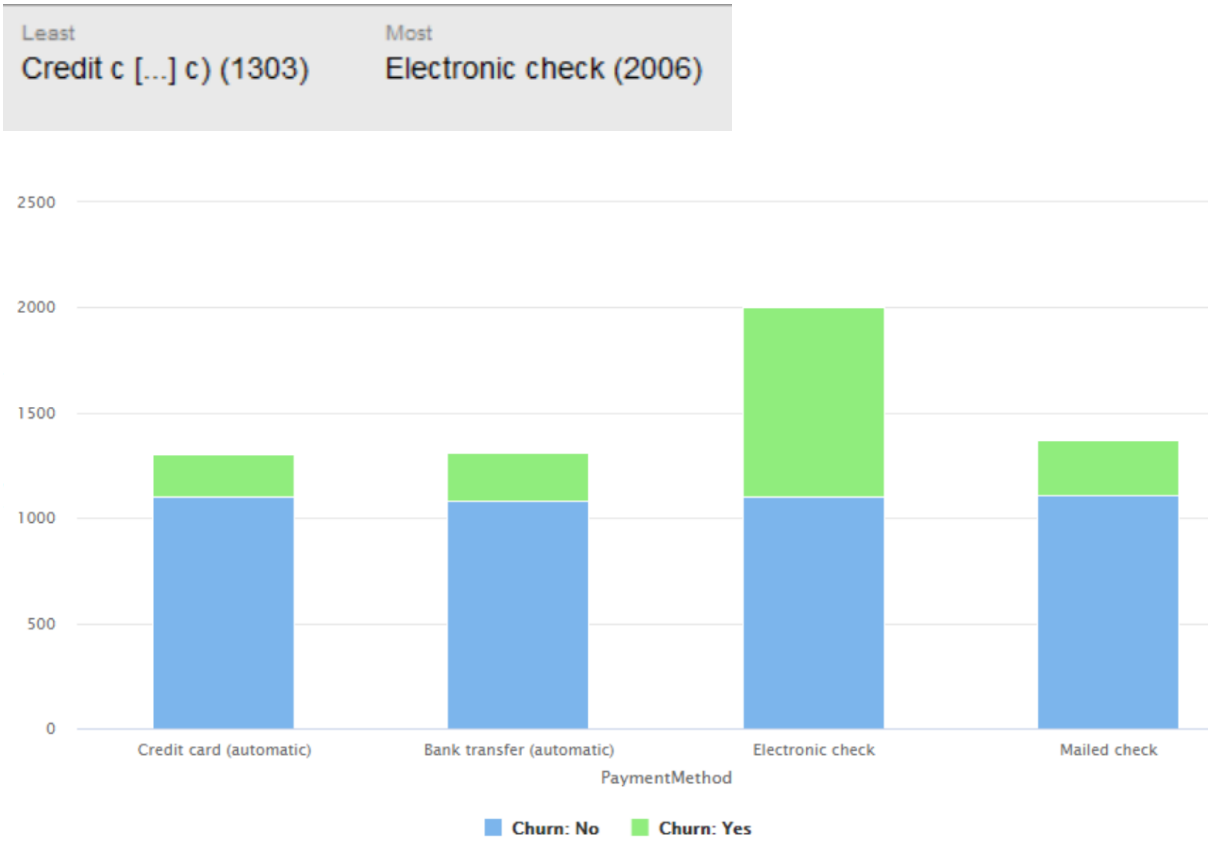
Partner

Least	Most
Yes (2904)	No (3082)



De la población analizada se podría decir que casi la mitad están casados y la otra mitad no, resulta interesante observar que las personas que no están casadas tienden más a cancelar su contrato, aunque esto podría ser un sesgo en los datos.

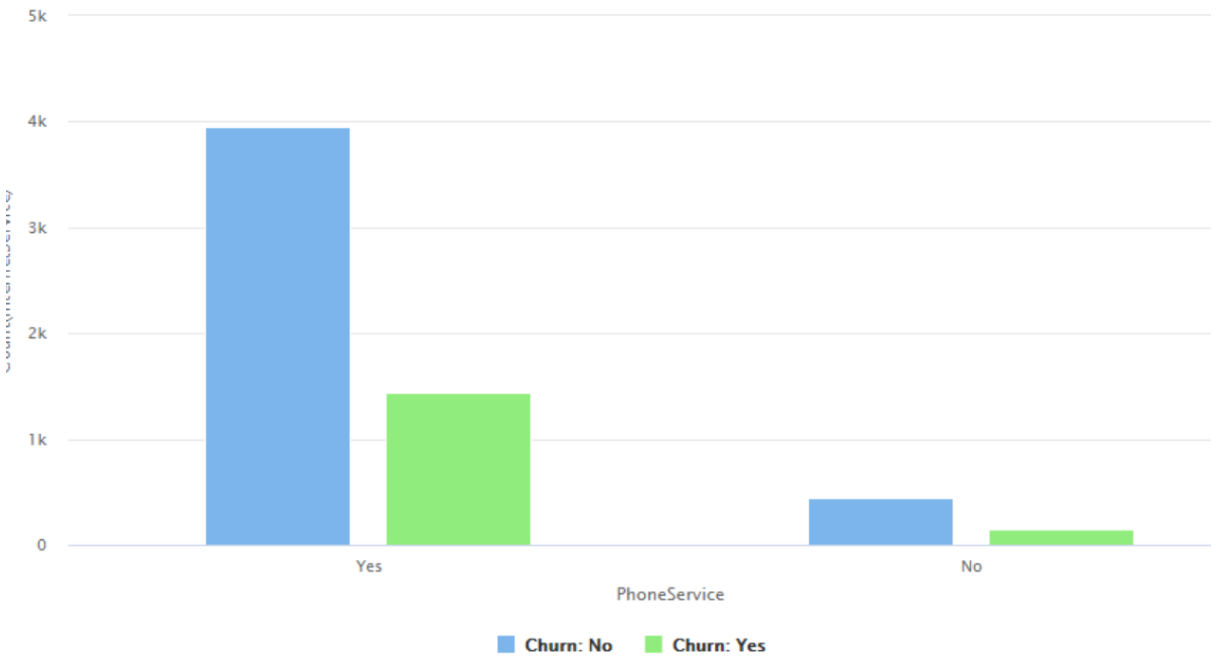
Payment Method



De las cuatro opciones de pago que ofrece el servicio tres son equiparables y la opción de cheques electrónicos es la que más se utiliza, siendo también la opción con mayor número de deserción, esto puede ser causa de la facilidad y popularidad del mismo método.

Phone Service

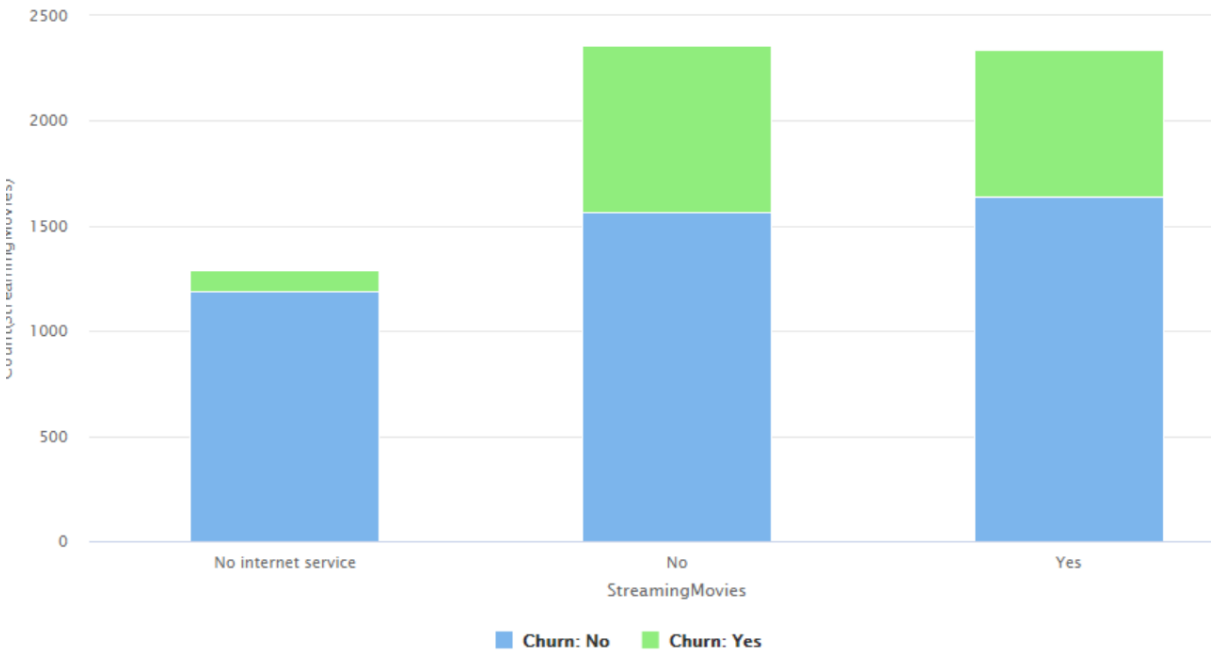
Least	Most
No (590)	Yes (5396)



La gran mayoría de los clientes cuenta con servicio telefónico, de estos una pequeña parte tiende a cancelar su contrato.

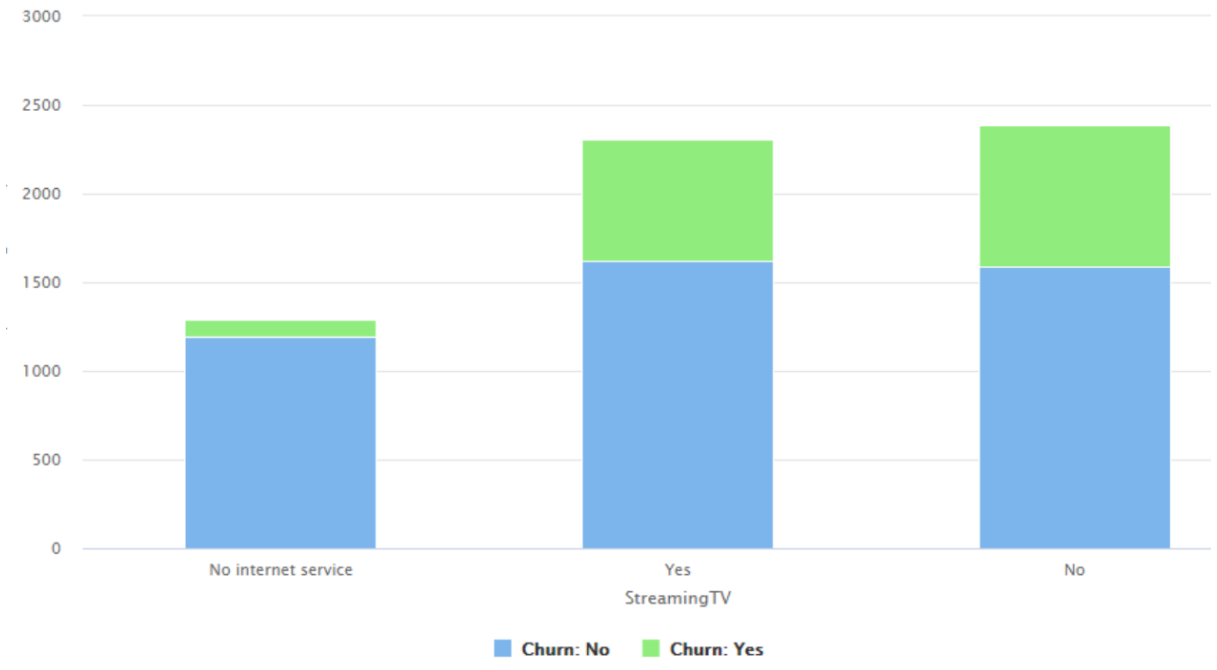
Streaming Movies

Least	Most
No inter [...] ce (1291)	No (2356)



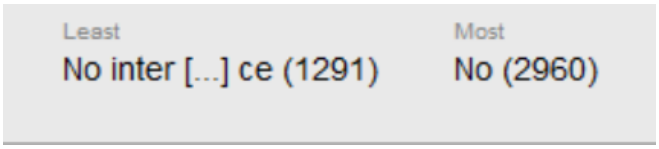
Por la gráfica se podría interpretar una que existen dos modas en la población, tener o no tener servicios de Peliculas en linea, en ambos hay cierta población que tiende a cancelar su contrato.

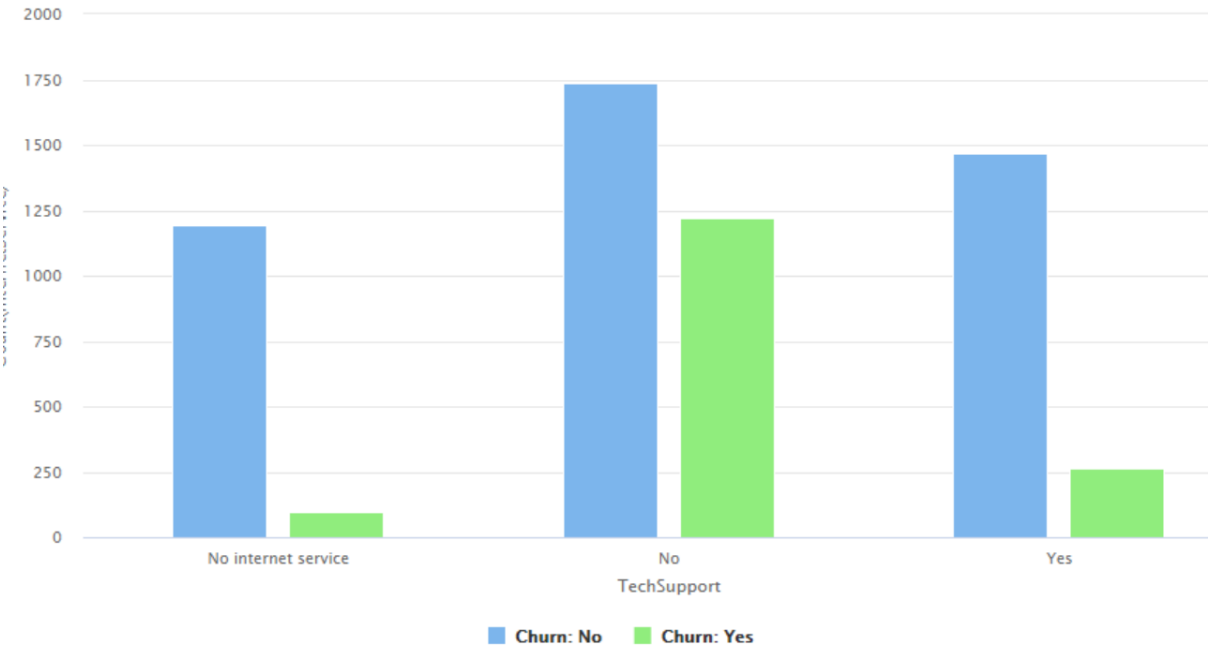
Streaming TV



Al igual que el atributo anterior hay una similaridad entre las poblaciones que tienen servicio de TV y los que no.

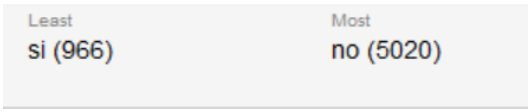
Tech Support

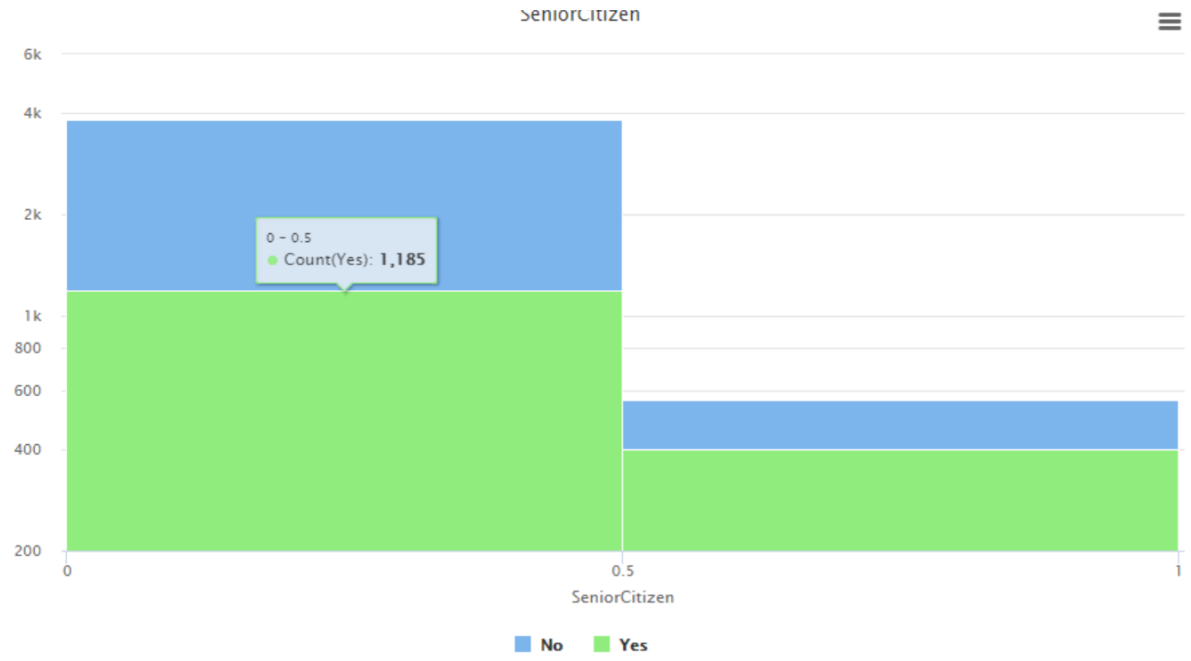




La mayor parte de la población no ha recibido soporte técnico y esto está correlacionado de gran manera con la decisión de los clientes de quedarse o no.

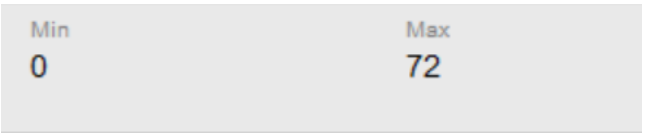
Senior Citizen

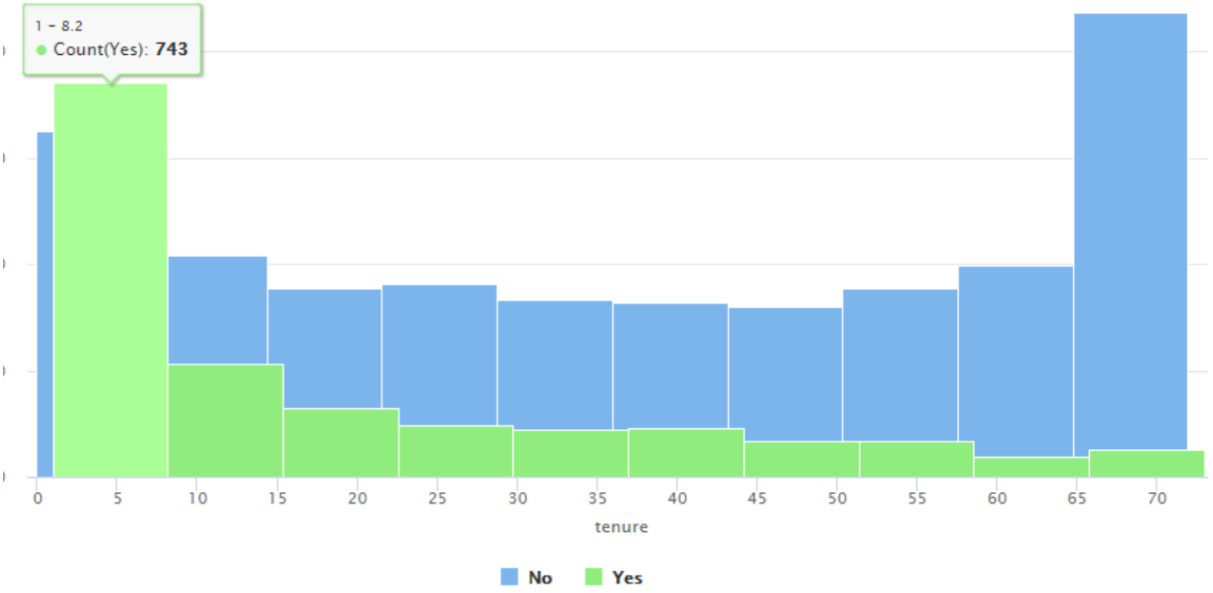




La gran mayoría de la población no es jubilada.

Tenure





La mayor parte de la población muestra es nueva en la empresa, entre menos tiempo lleve la persona con el servicio, más probable es que lo deje. Esto implica una relación inversamente proporcional.

7. Solucion analitica de datos

A partir del análisis exploratorio y de la necesidad propia del problema de llegar a clasificar si las personas se quedarán o no con el servicio, el modelo a utilizar debe de:

- Trabajar bien con datos Categóricos Nominales.
- Robusto con outliers.
- Se enfoque en la clasificación de datos de forma binaria(si-no)

Modelos Candidatos

Árboles de Decisión

- Pros
    - Robusto
    - No se necesita una preparación exigente de los datos, soporta valores nulos.
    - Datos Categóricos.
    - Clasifica.
    - Simple y entendible
    - Visualizable
    - No necesita ser un problema linealmente separable.
- Contras
    - Propenso a overfitting

Naive Bayes

- Pros
    - Robusto
    - Clasifica
    - Trabaja con datos categóricos
- Contras
    - Si la variable categórica tiene una categoría que no se observó al entrenar el modelo, este asigna una probabilidad de 0. Para solucionar esto se utilizará la Corrección de Laplace.
    - Asunción de predictores independientes, no puede mostrar dependencia entre clases.

Regresión Logística

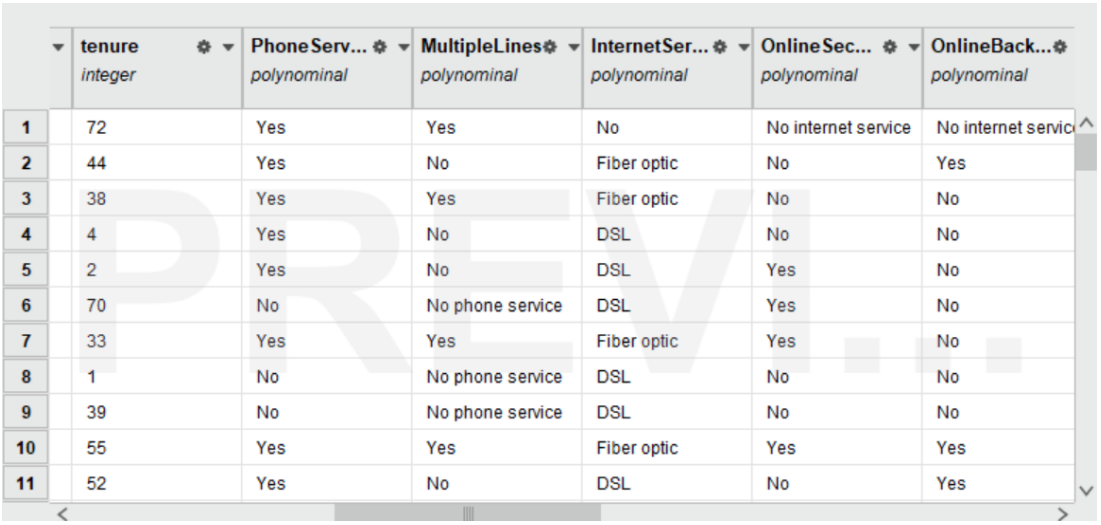
- Pros
    - Robusto
    - Determina la asociación entre los atributos
    - Es fácil de interpretar
    - Se actualiza fácilmente con nuevos datos
- Contras
    - El problema tiene que ser linealmente separable
    - No acepta labels categóricos

A partir de analizar las propuestas de modelos de clasificación hemos decidido utilizar el algoritmo de **Árboles de Decisión**.

Preparación y tratamiento de la Data

1. Darle formato al CSV

    Elegiremos el tipo de dato que más representa a cada atributo.



	tenure *integer*	PhoneServ... *polynominal*	MultipleLines *polynominal*	InternetSer... *polynominal*	OnlineSec... *polynominal*	OnlineBack... *polynominal*
1	72	Yes	Yes	No	No internet service	No internet servic
2	44	Yes	No	Fiber optic	No	Yes
3	38	Yes	Yes	Fiber optic	No	No
4	4	Yes	No	DSL	No	No
5	2	Yes	No	DSL	Yes	No
6	70	No	No phone service	DSL	Yes	No
7	33	Yes	Yes	Fiber optic	Yes	No
8	1	No	No phone service	DSL	No	No
9	39	No	No phone service	DSL	No	No
10	55	Yes	Yes	Fiber optic	Yes	Yes
11	52	Yes	No	DSL	No	Yes

2. Seleccionar los Atributos

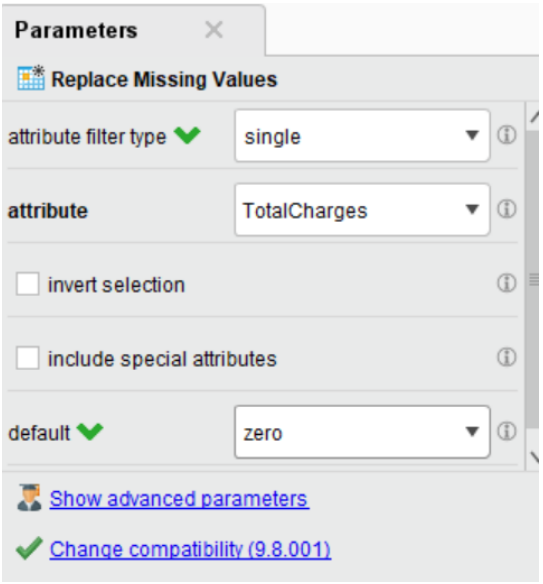
    Basándonos en el Análisis Exploratorio de Datos elegiremos los atributos que presenten un impacto en la decisión de cancelar o no el contrato.

    Los Atributos que serán excluidos del modelo serán:

    - att1
    - Customer ID
    - gender
    - Monthly Charges
    - Partner

3. Tratar valores faltantes

El atributo Total Charges cuenta con datos faltantes, haciendo un análisis rápido podemos ver que estos hacen referencia a un Tenure = 0. Esto quiere decir que estos 10 usuarios no han estado ni un mes con el servicio, lo que significa que no han pagado y podemos tomar esos datos como 0.



4. Discretizar Datos Continuos

Para tener una mejor clasificación de nuestros atributos y como una buena práctica, hemos decidido discretizar los datos continuos. Estos datos son:

Senior Citizen

Dividido en sus dos posibles valores (Si-No)

Senior Citizen		
5020	84%	No
966	16%	Si
5986	100%	

Total Charge

Se dividirá en tres bajo, medio y alto tomando el 33% de la distribución de los datos para cada categoría.

TotalCharge		
5986	100%	Cargo
2296	38%	
1039	17%	
568	9%	2618.6
427	7%	
386	6%	
359	6%	
333	6%	6085
274	5%	
191	3%	
103	2%	8684.8
5976	100%	
missing 10		

Tenure

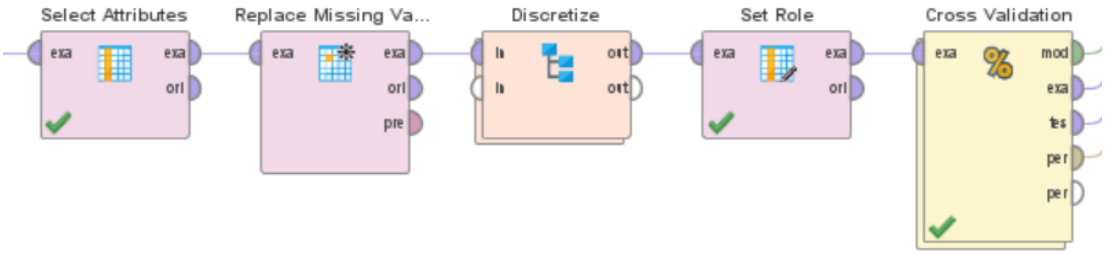
Se dividirá en tres: bajo, medio y alto tomando el 33% de la distribución de los datos para cada categoría.

Tenure		
5986	100%	Meses
1357	23%	
640	11%	14.4
490	8%	
471	8%	
428	7%	
424	7%	43.2
384	6%	
419	7%	
437	7%	
936	16%	72
5986	100%	

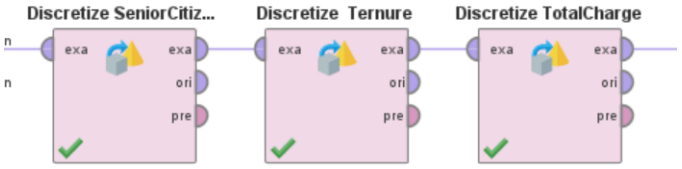
5. Aplicar el Modelo de Árboles de Decisión

Con los datos ya preparados solo falta utilizar el modelo elegido para poder clasificar los registros.

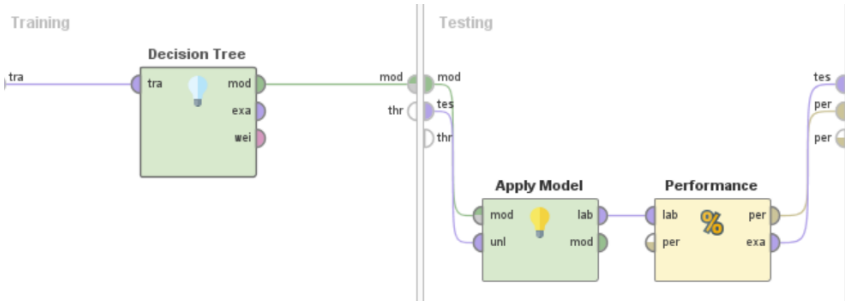
Modelo en Rapidminer



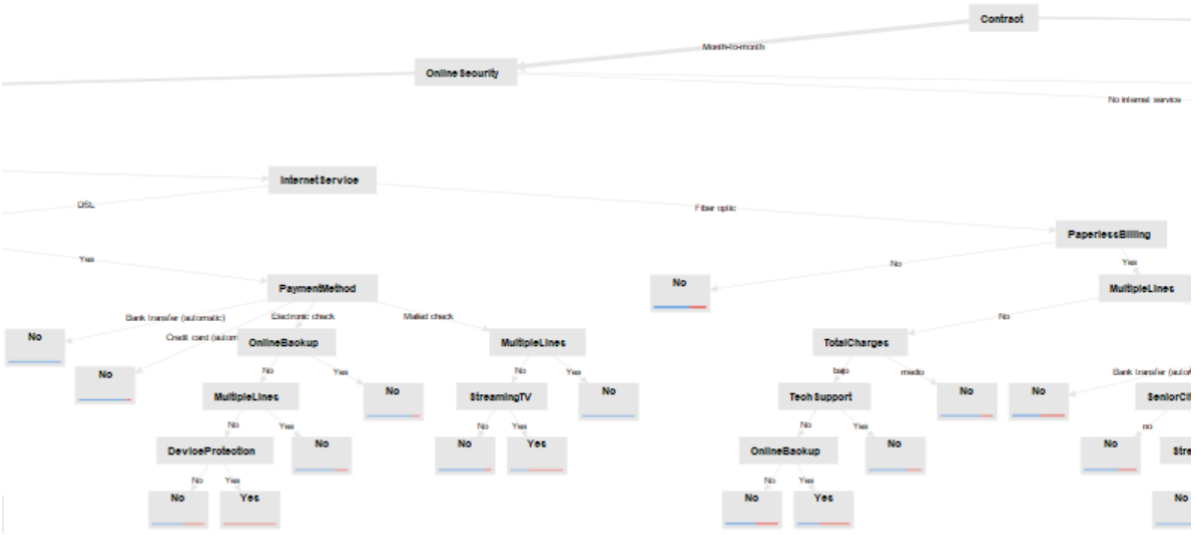
Proceso Principal



Subproceso Discretize



Subproceso Cross Validation



Fragmento del Árbol de Decisión

6. Rendimiento del Modelo

Para medir la efectividad de modelos como este donde la respuesta es categórica y nominal se utiliza una matriz de confusión.

accuracy: 77.50% +/- 1.47% (micro average: 77.50%)

	true No	true Yes	class precision
pred. No	3854	802	82.77%
pred. Yes	545	785	59.02%
class recall	87.61%	49.46%	

Nuestro modelo tiene una precisión del 77.50%

Predicción de clasificación del modelo

Churn	prediction(Churn)	confidence(No)	confidence(... ↓
Yes	Yes	0.272	0.728
No	Yes	0.272	0.728
Yes	Yes	0.272	0.728
Yes	Yes	0.272	0.728
Yes	Yes	0.272	0.728
No	Yes	0.272	0.728
Yes	Yes	0.273	0.727
Yes	Yes	0.273	0.727
No	Yes	0.273	0.727
Yes	Yes	0.282	0.718
Yes	Yes	0.282	0.718
Yes	Yes	0.282	0.718
Yes	Yes	0.282	0.718
Yes	Yes	0.282	0.718

8. Conclusiones

La atención al cliente es el factor clave para que la satisfacción sea positiva, ya que a partir del análisis realizado hemos recabado que casi la totalidad de los usuarios los cuales dejaron el servicio tuvieron una baja o nula atención al cliente y soporte técnico, lo cual indica que la satisfacción y la atención al cliente están relacionadas directamente.

El sector de usuario que no cuenta con un servicio de seguridad de internet también es propenso a dejar el servicio, lo cual indica insatisfacción al tener contratado el servicio. Esto se puede solucionar ofreciendo promociones para Antivirus a los clientes que cuenten con un servicio de Internet.

Existen dos cualidades las cuales también indican una cantidad considerable de cancelación de servicio, las cuales son la falta de respaldo y los usuarios los cuales no cuentan con pareja, indicando que regularmente este tipo de usuario no se sienten satisfechos con el servicio.

Por otra parte, se ubica que los clientes presentan problemas con la línea de teléfono y la fibra óptica, al presentarse estos problemas al inicio de su contrato tienden a irse. Al ser más probable que el cliente se vaya en los primeros meses de uso del servicio, es recomendable atraer su atención con precios atractivos y promociones que incluyan paquetes con antivirus y servicios de streaming, también se hace énfasis en mejorar la calidad del servicio y atención al cliente.

Utilizar el modelo de clasificación previamente mencionado, como observamos, es de suma importancia para visualizar de forma factible los atributos que tienen relevancia en la decisión de cancelar o no el servicio, a su vez nos facilita el análisis de las características que no estén a nuestro favor y que posiblemente queden fuera de nuestra visión, así, ayudándonos a llegar a más personas que no tengan estas similitudes o que quizá no tengan más de una de estás, garantizando que la pérdida será menor y contaremos con la fidelidad de más clientes.