UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL Y COMERCIAL

Técnica de Machine Learning para el cálculo de la probabilidad de fuga de los clientes de la empresa Bitel

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para:

Obtener el título profesional de Ingeniero Industrial y Comercial

AUTORES

Carla Benedicta Bernachea Collazos

Edward Chilet Paisig

Paola Guzmán Fernández

Victor Hugo Inche Contreras

Johana Mayra Leon Munive

ASESOR

Junior John Fabián Arteaga

ORCID Nº 0000-0001-9804-7795

Diciembre, 2021

RESUMEN

Según datos del Banco Mundial, la industria de las telecomunicaciones enfrenta cada año a una fuga de clientes que bordea el 30%. Estudios recientes han mostrado que tanto atributos cuantitativos: cantidad de minutos, mensajes, etc.; así como los cualitativos: edad, sexo, tipo de dispositivo tienen influencia en la fuga de clientes. En base a la literatura encontrada se definieron dos tipos de variables: demográficas y del comportamiento del consumidor las cuales son útiles para realizar la predicción de permanencia del cliente. Es por eso que haciendo uso de la técnica de regresión logística, se busca predecir la probabilidad de fuga de los clientes (Churn) de la empresa Bitel. Se realizó un exhaustivo trabajo de preprocesamiento y se llegó a entrenar un modelo de regresión logística con un accuracy score de 88%

Palabras clave: Fuga de clientes, regresión logística, machine learning, aprendizaje supervisado.

ABSTRACT

According to data from the World Bank, the telecommunications industry faces a customer churn of around 30% each year. Recent studies have shown that both quantitative attributes: amount of minutes, messages, etc; as well as the qualitative ones: age, sex, type of device have an influence on customer churn. Based on the literature found, two types of variables were defined: demographic and consumer behavior, which are useful for predicting customer permanence. That is why, using logistic regression technique, we seek to predict the probability of customer churn (Churn) of the Bitel company. An exhaustive preprocessing work was carried out and a logistic regression model was trained with an accuracy of 88%.

Keywords: Customer churn, logistic regression, machine learning, supervised learning.

INDICE DE CONTENIDOS

Índice de gráficos

Índice de tablas

Introducción

En la industria de las telecomunicaciones en el Perú, y en general en el mundo, debido a su carácter tan competitivo, y a la cantidad de operadores en el mercado; los clientes pueden elegir entre múltiples compañías para contratar el servicio y además, ejercer activamente sus derechos de cambiar de un operador a otro debido a la portabilidad numérica. Es decir que, los clientes tienen más ofertas y más información; lo que hace que exijan productos casi personalizados y mejores servicios a precios más bajos haciendo que las compañías telefónicas se enfoquen no solo en innovar, sino también en retener a los clientes; dado que cuesta de 5 a 10 veces más reclutar un nuevo cliente que retener uno existente. Es por eso que actualmente, el evitar la fuga de clientes (churn) es el principal problema comercial de los operadores telefónicos, incluso se ha vuelto aún más importante que la adquisición de los mismos; dado que es más rentable.

Muchas de estas empresas implementan estrategias que les permitan retener a los clientes y a la vez buscan sincronizar programas y procesos para mantenerlos por más tiempo; proporcionándoles productos y servicios que no solo se adapten a sus necesidades, sino que puedan exceder sus expectativas.

Con las estrategias de retención adecuadas, muchas compañías pueden reducir y gestionar la rotación y fuga de clientes al identificar su probabilidad de deserción. Esto, ahora gracias al *machine learning* y la técnica de regresión logística del aprendizaje supervisado ahora es posible, dado que son muy eficaces para este propósito y les permite a las empresas optimizar sus recursos financieros, de marketing, etc. para evitar que este problema.

En este estudio hemos abordado el problema de la fuga de clientes de la siguiente manera: en el capítulo 1, se realiza el planteamiento y la descripción detallada de la situación problemática analizando el contexto global, nacional y del sector. Asimismo, se define y delimita el alcance de la investigación desde la perspectiva espacial, temporal y conceptual. Luego en el capítulo 2, se exponen los conceptos y se evidencian los casos de éxito de la aplicación de la regresión logística y los modelos predictivos tanto en la industria de las telecomunicaciones como en otros sectores. En el tercer capítulo, se consigna la información de la empresa y su entorno para el posterior procesamiento. En el capítulo 4, se describe el marco de referencia y la metodología empleada de acuerdo a la información recolectada. Posteriormente, en el capítulo 5 se detallan los pasos ejecutados de manera secuencial para el

desarrollo de la solución, desde la recolección, procesamiento y limpieza de los datos disponibles, pasando por la selección minuciosa de las variables a incluir en el modelo, y su rol en la decisión que toma el cliente de dejar o no de compañía; hasta llegar a la medición del modelo y la elaboración de indicadores y reportes. Finalmente, en el capítulo 6 se presentan las conclusiones y hallazgos del estudio, así como las recomendaciones a tener en cuenta para optimizar el resultado del estudio o su posterior despliegue en escenarios futuros.

En base a la literatura encontrada, para el presente trabajo se definieron variables relacionadas con el comportamiento del consumidor, y de tipo demográficas, que junto con las variables cuantitativas relacionadas con el consumo de los clientes son útiles para realizar la predicción de permanencia de un cliente con el servicio contratado.

Capítulo I: Planteamiento del Problema

1.1. Descripción de la Realidad Problemática

Debido al avance tecnológico, todos los países han podido acelerar su desarrollo socioeconómico y mejorar los servicios, así como las oportunidades sociales que permitan conectar a los ciudadanos, transparentar sus acciones, eliminar barreras y construir un futuro mejor. Es así que se han podido transformar casi todos los sectores de la economía al introducir nuevos productos, servicios y modelos de negocio que permitan generar valor para las organizaciones y empleo para las personas.

Actualmente, la economía digital mundial representa según datos del Banco Mundial, casi el 22 % del producto bruto interno (PBI) mundial. Gracias al uso de las plataformas digitales los ciudadanos, incluso aquellos que viven en las regiones más remotas, han podido acceder a la información, obtener empleos, estudiar, e incluso recibir atención médica. Asimismo, el dinero móvil se ha convertido en la alternativa más fácil y segura del sistema bancario, sobre el modelo tradicional; lo cual ha propiciado la inclusión financiera, tal como se aprecia en el siguiente gráfico. (Ver figura 1)

Figura 1: Aporte de la economía digital al PBI mundial

Fuente: Banco Mundial (2021)

Y aunque no todos se han beneficiado de la misma manera puesto que todavía existen desigualdades en cuanto a la penetración, asequibilidad y desempeño de los servicios digitales; es un logro tangible que casi la mitad de la población mundial tenga acceso a internet desde 2016 ya sea a través de banda ancha móvil o fija.

Es por eso que en un mundo cada vez más digitalizado, el reducir la brecha digital resulta una cuestión crucial e incluso algunos autores ya hablan de una nueva clase de "pobres digitales".

Si nos trasladamos al contexto nacional; durante la última década la economía peruana ha presentado el mayor y más rápido crecimiento en la región, con una tasa de crecimiento promedio de 5.9% debido a las condiciones externas favorables que se venían presentando.

Sin embargo, debido a la súbita llegada del Covid19, la situación ha cambiado. Los reportes del Banco Mundial afirman que para el periodo comprendido entre octubre de 2020 y lo que va del 2021, en comparación con los niveles previos a la pandemia, al menos el 25% de las empresas peruanas redujeron sus ventas en 50%. Si bien a pesar de la crisis el 89% de las empresas mantuvieron a sus trabajadores, el 65% de ellas realizó un reajuste salarial, ya sea reduciendo los sueldos, las horas de trabajo del personal de su nómina u otorgando licencias.

Pero el Covid 19 también generó impactos positivos, puesto que el 34% de las empresas ha aumentado el uso del internet, las redes sociales y las plataformas digitales y obviamente la telefonía móvil.

Asimismo, el sector telecomunicaciones en el Perú es altamente competitivo; ello sumado a los avances regulatorios y tecnológicos que facilitan la portabilidad numérica; han otorgado a los clientes variadas opciones y empresas a las cuales cambiarse, en caso de que sus expectativas no hayan sido satisfechas por el proveedor actual.

De otra parte, en la última década han ingresado nuevas compañías telefónicas al mercado peruano, tanto gigantes conglomerados transnacionales como Viettel (en adelante Bitel) e incluso emprendimientos digitales como Cuy Mobile. Asimismo, la cantidad de líneas nuevas ha incrementado considerablemente en los últimos años, pasando de 26,61 millones de líneas móviles activas y una penetración de 94 líneas por cada 100 habitantes en 2009 a 40,83 millones de líneas móviles activas y una penetración de 124 líneas por cada 100 habitantes a junio de 2021. Es decir que la cantidad de líneas móviles activas ha crecido en un 35% en la última década. (Ver figura 2)

Figura 2: Líneas móviles en servicio ajustadas y penetración a nivel nacional

Fuente: Portal Punku – OSIPTEL (2021)

De acuerdo al portal PUNKU, la herramienta informática que permite obtener reportes estadísticos de los indicadores del mercado de telecomunicaciones basado en la información que las empresas operadoras reportan al OSIPTEL; la participación de mercado de la empresa Bitel ha crecido del 1% en 2014 al 17.8% para el tercer trimestre de 2020. Este es un crecimiento exponencial en un marco de solo 6 años. Sin embargo, si analizamos más detenidamente el crecimiento de Bitel; desde el año 2018 (año en el que alcanzó un 15.3% de participación del mercado), esta empresa solo ha crecido un 2.5% hasta 2020. Es más, entre el segundo y tercer trimestre del mismo año, Bitel ha perdido el 0.4% de participación del mercado debido a la fuga de clientes, para ello la empresa ha creado servicios de valor añadido como BITEL VIDEO, BITEL AVENTURAS e incluso una subdivisión dentro de su área VAS (Value Added Services), con enfoque en los e-sports y gaming; con el objetivo de satisfacer el nuevo mercado emergente de gamers. Sin embargo, aún no se ha podido evidenciar el impacto ni la relevancia de estas medidas en la retención de los clientes en general.

Asimismo, de acuerdo al último reporte del portal PUNKU emitido del 21 de octubre de este año (Ver figura 3), la evolución de las líneas móviles portadas mensualmente registra un promedio de 450 000 (cuatrocientas cincuenta mil) entre octubre de 2020 y octubre de 2021.

Figura 3: Evolución mensual de las líneas móviles portadas

Fuente: Portal Punku – OSIPTEL (2021)

Diversos factores pueden contribuir a la decisión de los clientes de terminar el servicio con su operador telefónico; por un lado se encuentran factores relacionados con el servicio o core de la compañía tales como: cobertura telefónica y de datos, calidad de la señal telefónica, los precios. Pero también existen factores relacionados con los servicios de valor agregado como: el servicio de atención al cliente, las promociones, descuentos, etc.

De igual forma, tal como se aprecia en la figura 4; como resultado de las cuarentenas y el aislamiento, durante los primeros nueve meses del 2020, los ingresos de las líneas de negocio de “Internet Fijo” y “Servicios Móviles” registraron incrementos de 4.5% y 0.6%, esto debido a que dichos servicios son imprescindibles y para poder realizar actividades como el teletrabajo, la teleducación y el ocio; los cuales impusieron su presencia debido a la pandemia.

Tabla 1: Crecimiento de ingresos operativos por línea de negocio

(en millones de S/)

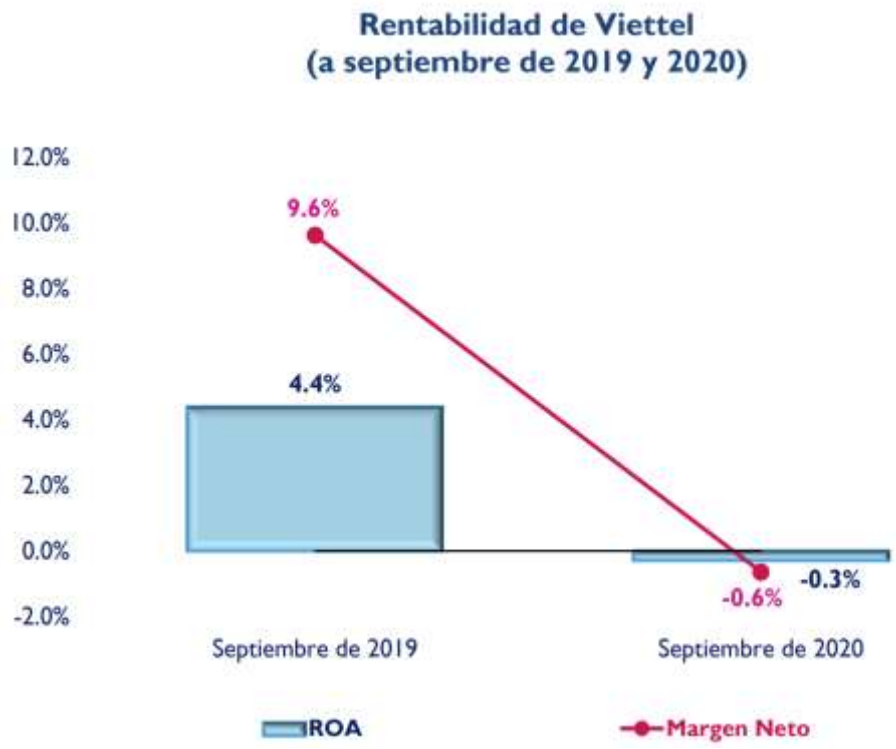
Líneas de negocio	Enero - Septiembre 2019		Enero - Septiembre 2020		Ene - Sept
	Ingresos	Participación	Ingresos	Participación	Δ% 2020
SERVICIOS MÓVILES	5,695	41.3%	5,726	48.2%	0.6%
VENTA DE EQUIPOS	3,333	24.2%	1,700	14.3%	-49.0%
INTERNET FIJO	1,416	10.3%	1,480	12.5%	4.5%
TELEVISIÓN DE PAGA	1,283	9.3%	1,262	10.6%	-1.6%
TRANSMISIÓN DE DATOS Y ALQUILER DE CIRCUITOS	741	5.4%	638	5.4%	-13.9%
TELEFONÍA FIJA DE ABONADOS	461	3.3%	313	2.6%	-32.1%
INTERCONEXIÓN	436	3.2%	356	3.0%	-18.3%
TELEFONÍA DE LARGA DISTANCIA	42	0.3%	30	0.3%	-27.4%
TELEFONÍA DE USO PÚBLICO	36	0.3%	16	0.1%	-55.9%
OTROS INGRESOS OPERATIVOS	330	2.4%	358	3.0%	8.4%
TOTAL DE INGRESOS	13,772		11,879		-13.7%
INGRESOS SIN VENTA DE EQUIPOS	10,439		10,179		-2.5%

Fuente: OSIPTEL (2021)

En el caso particular de Bitel, para el periodo comprendido entre enero y setiembre de 2020, sus ingresos operativos registraron un incremento de 6.9%, debido principalmente al buen desempeño de la unidad de negocio denominada “Servicios Móviles” que registró un crecimiento de 7.9%, y representó el 74.0% de los ingresos generados por la empresa. Esto también se evidenció en el incremento de su market share o participación de mercado en el mercado móvil el cual pasó de 16.8% (a diciembre del 2019) a 17.8% en septiembre de 2020, de acuerdo al reporte de del desempeño financiero del sector telecomunicaciones de Osiptel.

No obstante, pese a los buenos resultados operativos de la empresa, estos no fueron suficientes para cubrir sus gastos financieros; tal como lo muestra el reporte de rentabilidad de la compañía (Ver figura 5). Esto se tradujo en un resultado neto negativo. Es así que el Return on Assets (ROA) de la compañía ha caído al punto de convertirse en negativo (-0.3%) si se compara con el registrado durante los primeros nueve meses del 2019, donde el ROA fue de 4.4%.

Figura 4: Rentabilidad de Viettel



Fuente: OSIPTEL (2021)

Esta disminución del ROA debido a mayores gastos financieros y mayores obligaciones financieras a corto plazo, obliga a la compañía a buscar la manera de atraer más clientes y de retener a los actuales.

Debido a que la organización debe analizar y manejar una gran cantidad de datos, considerando que cada línea móvil contratada tiene diferentes características asociadas como: tipo de plan, cantidad de minutos, mensajes y datos asignados, si el usuario está satisfecho o ha presentado reclamos, etc. Sería difícil, por no decir prácticamente imposible, para una persona o un equipo de personas identificar y mapear cuales son las variables más relevantes y que inciden en la decisión de un cliente de rescindir los servicios de Bitel.

Ante esta situación, es necesario aplicar técnicas de machine learning que nos permitan trabajar con esta gran cantidad de datos y variables de manera oportuna y precisa con el objetivo de plantear estrategias de retención de los clientes en base a los resultados del análisis estadístico de la data. Es por ello que en este trabajo de investigación aplicaremos la técnica de Regresión Logística de Machine Learning para determinar ¿Cuál es la probabilidad de fuga de los clientes (Churn) de la empresa Bitel?

1.2. Justificación de la Investigación

Según el Informe Técnico Trimestral de Estadísticas de Uso de las Tecnologías de la Información y Comunicación de 2019 elaborado por el Instituto Nacional de Estadística e Informática (INEI) mostrado a continuación (Ver figura 6), al menos el 71,4% de los hogares peruanos se comunica utilizando exclusivamente telefonía móvil y el 20,6% de los peruanos emplean teléfono fijo y celular; mientras que solo el 1,3% de los hogares lo hace exclusivamente a través del teléfono fijo. De otra parte, existe un 6,7% de peruanos que no cuentan con ningún tipo de telefonía.

Tabla 2: Acceso al Servicio Telefónico

Acceso a teléfono	Jul-Ago-Sept 2018	Jul-Ago-Sept 2019 P/	Variación (Puntos porcentuales)
Total	**100,0**	**100,0**	
Solo teléfono fijo	1,6	1,3	-0,3
Solo teléfono móvil	68,2	71,4	3,2
Ambos	22,8	20,6	-2,2
Ninguno	7,4	6,7	-0,7

Fuente: Encuesta Nacional de Hogares – INEI (2020)

Estos datos evidencian la importancia que ha cobrado la telefonía móvil y como se ha posicionado como el principal mecanismo de comunicación de los peruanos, reportando un

crecimiento del 3,2% respecto al año anterior. Es pues, la telefonía móvil a nivel mundial, el medio de comunicación masivo por excelencia.

Sin embargo, el reporte del Ministerio de Transportes y Telecomunicaciones muestra que en el año 2020, en el Perú habían 37 072 040 líneas móviles activas, es decir, 12,1% menos que en 2019.

Tabla 3: Líneas Móviles en servicio

Servicios		Indicador	II T 2019	II T 2020	% Anual
Internet	Internet Fijo(*)	Suscriptores	2 427 857	2 586 007	6.5
	Internet Móvil[1]	Suscriptores	27 135 700	23 576 187	-13.1
Telefonía Móvil		Líneas en servicio	42 177 337	37 072 040	-12.1
Telefonía Fija	Telefonía Fija de Abonado(*)	Líneas en servicio	2 599 830	2 406 758	-7.4
	Telefonía Pública (*)	Teléfonos Públicos	119 589	104 693	-12.5
Radiodifusión por Cable(*)		Suscriptores	2 060 763	1 893 665	-8.1

(*) Información Preliminar

Fuente: Ministerio de Transportes y Comunicaciones (2020)

El mercado de la telefonía celular es altamente competitivo. Además de ello, como consecuencia de la portabilidad numérica, los consumidores cuentan con múltiples y variadas opciones en términos de planes y beneficios, así como compañías a su disposición para cambiarse a aquella de la cual reciba mayores prestaciones, en caso de que el servicio actual recibido no cumpla con sus expectativas. Es así que, la industria de las telecomunicaciones en el mundo se enfrenta anualmente a una fuga de clientes que bordea el 30 %, estimándose además que el costo de adquirir un nuevo cliente es 5 a 10 veces superior al costo de retener uno antiguo, tal como sostiene Junxiang, Lu (2020) en su artículo “Predicting customer churn in the telecommunications industry”.

Es de vital importancia por ello que las compañías analicen cuales son los factores que influyen en la decisión de los clientes para mantenerse o dejar el servicio contratado. Teniendo conocimiento de estos factores, se pueden desarrollar estrategias más específicas que permitan fidelizar y mantener a los clientes dentro de Bitel para continuar el crecimiento que anteriormente ha reportado y recuperar a aquellos clientes perdidos.

1.3. Delimitación de la Investigación

1.3.1. Delimitación Espacial

Este trabajo de investigación ha sido realizado a nivel Nacional. El Perú actualmente, cuenta 1.285 millones km² divididos en 24 departamentos y tiene de acuerdo a la estructura demográfica de la población peruana descrita por la Comisión Económica para América Latina y el Caribe (CEPAL), una población de 32.97 millones habitantes (Ver figura 8). Asimismo, existen 37 072 040 líneas móviles activas, de las cuales las de Bitel representan el 17,8% es decir 6’ 598,824 líneas móviles.

Tabla 4: Estructura Demográfica de la Población del Perú

Población total	32 971.9	(000)	2020
Tasa anual de crecimiento de la población	0.9	%	2020_2025
Rural	-0.6	%	2020_2025
Urbana	1.4	%	2020_2025
Tasa bruta de natalidad	16.9	%	2020_2025
Tasa bruta de mortalidad	5.9	%	2020_2025
Tasa de migración	-1.9	%	2020_2025
Esperanza de vida	77	años	2020_2025
Hombres	75	años	2020_2025
Mujeres	80	años	2020_2025

Fuente: CEPAL (2020)

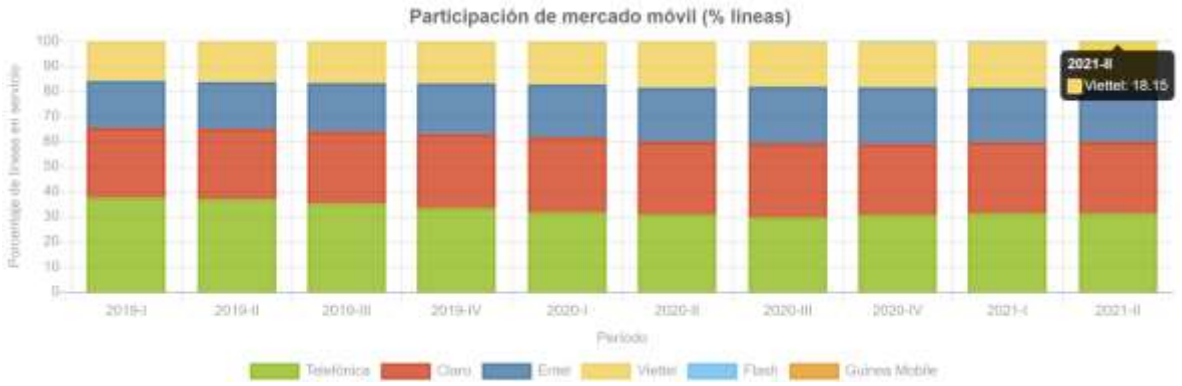
En términos específicos, la investigación se lleva a cabo en el área comercial de Bitel, que es la que se encarga del registro de las altas y fugas de usuarios. Este desarrollo tecnológico tendrá lugar en el área comercial que va a recolectar, procesar y analizar la data, a fin de construir un modelo preciso y fiable que proporcione resultados óptimos.

1.3.2. Delimitación Temporal

La investigación usará información formal de la empresa Bitel, una empresa de origen vietnamita que inició sus operaciones en el Perú desde el 2014, que ingresó al mercado con la estrategia de ofrecer un servicio de telefonía móvil de buena calidad a un precio más accesible,

apuntando a clientes de nivel socioeconómico C y D en principio, para luego buscar escalar a los primeros niveles. Actualmente, el mercado peruano se encuentra distribuido de la siguiente manera: Movistar (31.38%) mantiene el liderazgo, seguido de Claro (28.45%) y Entel (21.59%) y luego está Bitel (18.15%). Dado que Guinea Mobile no logró obtener una participación mayor al 1%, no figura en la estadística; tal como se aprecia en el reporte de participación porcentual de mercado emitido por OSIPTEL (Ver figura 9).

Figura 5: Participación porcentual de mercado móvil



Fuente: Portal Punku – OSIPTEL (2021)

En términos específicos, el presente estudio toma del universo de los 6’ 598,824 de clientes que tiene Bitel actualmente, una cantidad aleatoria de registros que representa un porcentaje de la base total de líneas prepago activas del operador tanto prepago como postpago, al momento del inicio del estudio. Por políticas internas de la empresa, no se puede trabajar con toda la base de clientes. Sin embargo, si se denota que se han excluido de este porcentaje las líneas móviles asignadas a los empleados, las cuales ascienden a 500 líneas móviles postpago.

Debido a que el modelo cuenta con variables que se modifican con el tiempo, se hizo necesario trabajar con las líneas que estuvieron activas desde Julio hasta Octubre de 2021, a fin de poder contar con historia para agregar y calcular las variables transaccionales en los casos de estudio. Finalmente, la base con la que se trabajara tiene un total 28092 líneas prepago y postpago que permanecieron activas y luego cancelaron el servicio durante estos 4 meses.

1.3.3. Delimitación Conceptual

Según lo definido en por Junxiang Lu en su artículo “Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS”,

el fenómeno churn es “la acción de cancelar el servicio prestado por la compañía”, ya sea por decisión del cliente haciendo uso de su voluntad o por parte de la empresa.

Esta investigación se centra en estudiar el problema que representa la fuga de clientes o churn por su nombre en inglés, para la empresa de telefonía móvil Bitel. Esta fuga puede estar relacionada con diversos factores, los mismos que pueden tener relación o no con la calidad del servicio.

Para poder realizar este estudio, haremos uso de la técnica de regresión logística del machine learning, que forma parte del aprendizaje supervisado, dado que lo que buscamos es construir una función capaz de predecir el valor de la probabilidad de fuga de los clientes; tomando como punto de partida el conocimiento existente de la variable a modelar, que son los factores que influyen en la decisión de los clientes de dar por concluido el servicio.

Cabe mencionar también que las variables de la base de datos de clientes no se modifican a lo largo del tiempo y se establecen en el momento en que el cliente activa la línea. Sin embargo, las variables de consumo que se encuentran en la base de datos de telecomunicaciones si varían a lo largo del tiempo. Es necesario tener en cuenta este detalle a la hora de agregarlas para la implementación del modelo.

Capítulo II: Marco Teórico

2.1. Antecedentes de la Investigación

- **Jain H., Khunteta A. y Srivastava S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. Procedia Computer Science, 167, 101-112. Recuperado de: https://www.sciencedirect.com/science/article/pii/S1877050920306529**

Objetivo:

La presente investigación tiene como objetivo utilizar dos técnicas de aprendizaje automático para predecir el abandono de clientes, la regresión logística y *Logit Boost*. Nos menciona que en la actualidad todas las industrias tienen problemas con la rotación de los clientes, estos constantemente están cambiando sus servicios de un proveedor a otro, y en el caso de las telecomunicaciones es mucho mayor debido a la gran competitividad del mercado de este sector. Por lo tanto, estas empresas están constantemente analizando la rotación de sus clientes para así tener una mayor retención.

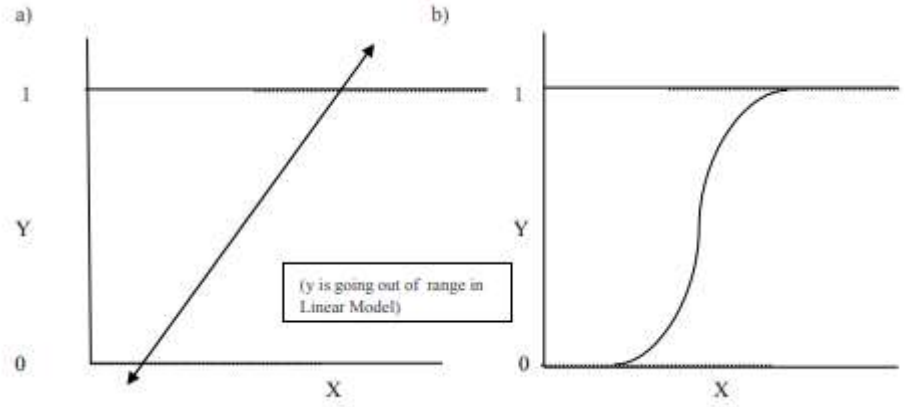
Metodología:

En este estudio se utilizaron dos técnicas de Machine Learning para predecir el abandono de clientes: regresión logística y *Logitboost*.

- ✓ La regresión logística toma las entradas de valor real y hace la predicción como una clase de entrada que pertenece a la clase 0.Si la predicción es> 0.5, entonces toma la salida como clase 0; de lo contrario, toma la salida como clase 1.

  No se eligió el modelo lineal, ya que el valor de predicción puede resultar fuera del rango.

Figura 6: Regresión lineal vs. Regresión logística



Fuentes: Jain, Khunteta y Srivastava (2020)

- ✓ Logit Boost es un modelo de regresión logística aditiva. Logit Boost toma diferentes ejemplos de entrenamiento repetidamente debido a que el algoritmo de aprendizaje base genera una nueva regla de predicción débil, que causa tantas rondas y el algoritmo de impulso posterior debe convertir estos reglas débiles en una regla de predicción fuerte que, normalmente, se vuelve mucho más precisa que una regla débil.

$$\log \frac{(y = 1|x)}{(y = -1|x)} = \sum_{m=1}^{M} f_m(x).$$

Para el experimento, la base de datos utilizada fue real, perteneciente a una empresa estadounidense de telecomunicaciones llamada Orange, esta base tenía 3333 instancias, y se dividió en dos grupos, uno de texto y otro de entrenamiento. Por último, se llevó acabo Regresión Logística y Logit Boost con la ayuda de la herramienta de aprendizaje automático Weka y se evaluó el rendimiento con diferentes criterios de rendimiento.

Resultados:

Los resultados muestran que tanto las técnicas de Regresión logística como Logit Boost obtuvieron buenos resultados. La precisión alcanzó el 85,2385% para el caso de Regresión logística y el 85,1785 para Logit Boost.

- **J. Lu. Predicting customer churn in the telecommunications industry — an application of survival analysis modeling using sas. SAS User Group International Online Proceedings, 27,114–27. Recuperado de: https://support.sas.com/resources/papers/proceedings/proceedings/sugi27/p114-27.pdf**

A través de este artículo, el autor desea proporcionar una herramienta que ayude a las empresas de telecomunicaciones a gestionar la reducción de la deserción o fuga de sus clientes. Se propone una estrategia para determinar, no sólo qué clientes tienen un alto riesgo; sino también saber qué tan pronto lo harán, es decir después de cuánto tiempo posterior a la firma del contrato de servicios.

Objetivo:

Comprender el riesgo que significa una potencial fuga de clientes (churn) en términos cuantitativos dentro de una empresa de telecomunicaciones.

Emplear técnicas estadísticas para predecir de manera oportuna la probabilidad de fuga. Asimismo, hacer uso del análisis de supervivencia para determinar qué cliente abandonará la compañía y cuándo.

Metodología:

Comparar la eficacia del uso de técnicas de análisis de supervivencia versus las técnicas machine learning basadas en métodos estadísticos como la regresión logística y los árboles de decisión.

Para poder realizar la investigación, se tomaron una muestra de 41 374 clientes activos de alto valor de toda la base de clientes de una empresa de telecomunicaciones de Estados Unidos. Esta información fue obtenida de cuatro bases de datos principales: financiera y de marketing a nivel de bloque, datos demográficos a nivel del cliente proporcionados a través de un proveedor externo, datos internos del cliente y registros de contacto del cliente.

Se analizaron atributos demográficos de los clientes tales como: sexo, edad, estado civil, etc. También datos internos de la compañía como; tipo de plan, código de segmentación del cliente, si tenían contratados productos y/o servicios adicionales, si pagaban puntualmente sus recibos, descuento y promociones, si tenían líneas adicionales, etc. Todo esto se analizó durante los quince meses posteriores al inicio del estudio.

Resultado:

De todos los datos recolectados, para el estudio solo se emplearon 29 variables exploratorias, como conjunto de datos final para llevar a cabo el análisis de supervivencia.

Los hallazgos de este estudio son útiles para que las empresas de telecomunicaciones puedan elaborar estrategias que optimicen la retención de clientes y/o los recursos con los que gestionan este problema.

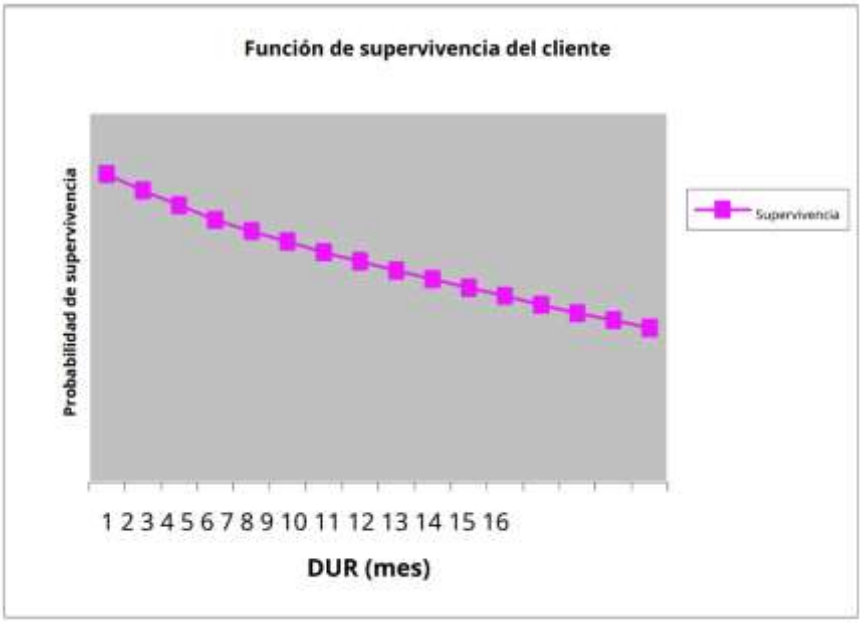
El autor sostiene además, que implementar la regresión logística y los árboles de decisión no toma mucho tiempo y que puede ser relativamente fácil si se compara con otras técnicas.

Sin embargo, la principal desventaja que poseen estos métodos es que no permiten predecir en qué momento los clientes abandonaran el servicio o al menos estimar su periodo de permanencia.

Como alternativa, el autor propone el uso del análisis de supervivencia, otra herramienta estadística que permitió mediante la clasificación de las probabilidades de supervivencia pronosticadas por los clientes en orden ascendente; identificar que los dos deciles superiores capturan entre el 55 y el 60% de las fugas de los clientes y los cinco deciles superiores capturan casi el 90% de las fugas. Con ello se pueden plantear estrategias comerciales para fidelizar a los clientes, optimizar la retención y así poder reducir el fenómeno de churn.

Como parte del análisis de supervivencia realizado en este estudio, el cual parte de un análisis univariante inicial de todas las variables categóricas cruzadas, quedando así solo 21 variables. Luego de procesar la información, se obtuvieron los siguientes gráficos, que denotan la función de supervivencia hallada así

Figura 7: Función de supervivencia del cliente

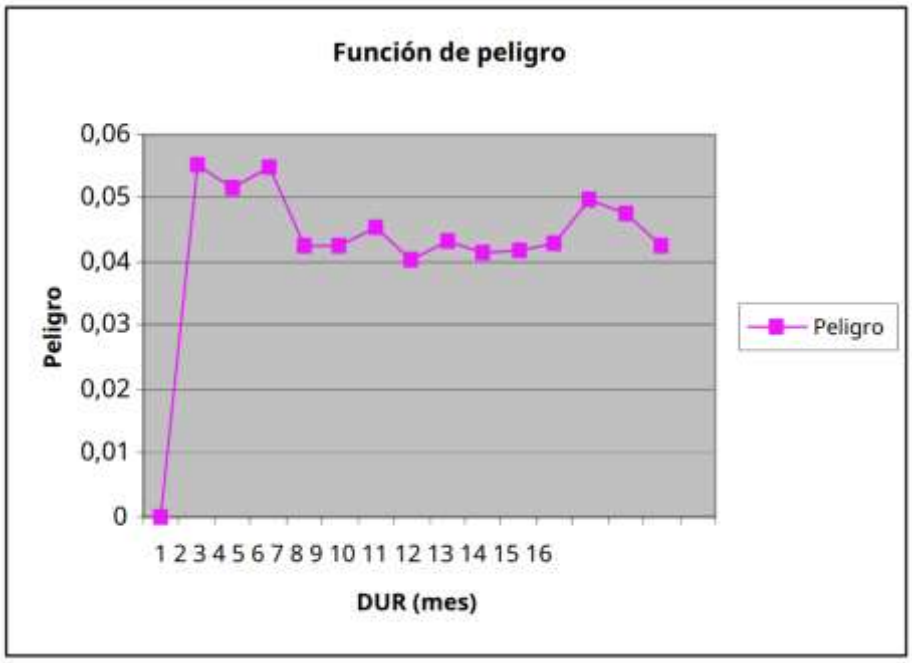


Fuente: Información del artículo estudiado (2014)

La efectividad del modelo y el éxito del mismo se validan de dos maneras. La primera fue a través de la clasificación de las probabilidades de supervivencia pronosticadas durante un tiempo específico en orden ascendente en deciles y luego se el resultado obtenido con el número de clientes que abandonaron la compañía (fugados) durante este período de tiempo especificado en cada decil.

La segunda manera de validar el éxito del modelo consiste en colocar las probabilidades de supervivencia pronosticadas en el mismo orden y luego comparar el número de clientes fugados hasta este tiempo especificado en cada decil.

Figura 8: Función de riesgo del cliente



Fuente: Información del artículo estudiado (2014)

- **Sandhya, K, Thaslima S, Vindhya, R y Srilakshmi P.(2021) Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression. International Journal of Innovative Research in Computer Science & Technology. 9(4).27-29. https://doi.org/10.21276/ijircst.2021.9.4.6**

Objetivos:

En el presente artículo se plantea que existen diferentes estrategias para generar mayores ingresos a una empresa de telecomunicación, primero está la búsqueda de nuevos clientes, el segundo es que los clientes existentes compren o actualicen su servicio por uno mejor y el último es que se mantenga la base de clientes por el mayor tiempo. De estas se calcula su retorno de inversión obteniendo como resultado que la última estrategia viene a ser la más rentable, dejando como conclusión que el costo de retener un cliente es mucho menor que adquirir uno nuevo y el hecho de venderle un producto o servicio adicional a uno que ya se tiene.

Para centrarse en la estrategia de retención de clientes, la empresa debe conocer las variables que implican el movimiento de clientes de un proveedor a otro. Para ello se propone utilizar técnicas de aprendizaje automático para predecir los eventos de fuga de clientes, aprendiendo de los datos almacenados de la empresa.

La cual serán usadas para aplicarlo con los métodos y algoritmos de aprendizaje automático basados en árboles y regresión para la creación de un modelo enfocado en la rotación de clientes

Metodología:

Se usó la programación en R, la cual es un software estadístico y de análisis de datos, para construir el modelo de predicción de abandono de clientes. El sistema contará con tres opciones, ver el análisis de rendimiento la cual muestra resultados obtenidos al aplicar la regresión logística y el árbol de decisiones; las pruebas, en la cual se construye una lista de clientes el cual tiene una alta probabilidad de abandono, y por último el de entrenamiento y prueba; el cual construye el modelo predictivo de la rotación de clientes de una empresa. De esta forma se podrá tener el factor riesgo respecto a los clientes de la empresa y su probabilidad de abandono.

Respecto a la validación de datos a usar, se plantea como un requisito funcional obtener la base de datos de una fuente confiable o una base de datos real de una empresa de telecomunicación, esta información servirá como referencia para el modelo predictivo a desarrollar.

Resultado:

Llegar a predecir de forma eficiente el comportamiento de un cliente en la industria de telecomunicaciones sirviéndose del análisis de sus datas históricas, servirá de base para poder entender mejor la situación de la empresa en el presente, de la misma forma los factores de abandono y las listas de personas con mayor porcentaje de probabilidad de abandono serán en quienes se centrará las estrategias para la retención de clientes. La cual creará un plan preventivo que aportará a un uso más eficiente de su base de datos, así como reaccionar anticipadamente a situaciones de fuga de clientes masiva.

- **Hashmi, N., Butt, N. A., & Iqbal, M. (2013). Customer churn prediction in telecommunication a decade review and classification. International Journal of Computer Science Issues (IJCSI), 10(5), 271.**

Objetivo:

Identificar la frecuencia de uso que han tenido los algoritmos que se usan para predecir el abandono de clientes (churn), a lo largo del periodo 2002 - 2013, mediante el conteo de investigaciones provenientes de revistas de buen prestigio. Así también, realizar un ranking de cuál es la técnica más utilizada para resolver la fuga de clientes, según los autores consultados.

Metodología:

El autor explora en bases de datos de revistas prestigiosas publicadas entre 2002 y 2013 orientadas al sector de telecomunicaciones para obtener una literatura académica. Los consultados fueron:

- ✓ Elsevier
- ✓ IEEE Xplore
- ✓ SpringerLink
- ✓ ScienceDirect
- ✓ Biblioteca digital ACM
- ✓ Búsqueda académica de Microsoft

No se han tomado en cuenta: artículos de conferencias, boletines, notas de conferencias, libros, tesis doctorales, trabajos no publicados y actas de conferencias.

Los artículos fueron examinados en base a las siguientes dimensiones:

- ✓ Distribución de artículos por tipo de conjunto de datos
- ✓ Distribución de artículos por técnicas
- ✓ Distribución de artículos por Revistas
- ✓ Distribución de artículos por año de publicación

Resultados:

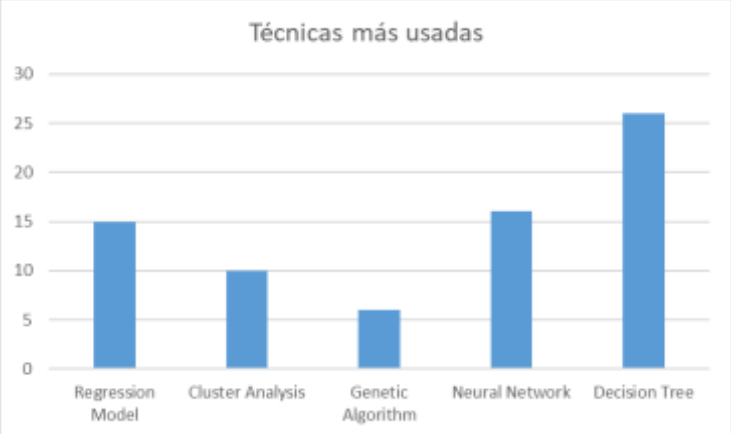
Se identificó a Decision Tree (DT) como el algoritmo más usado en la predicción de clientes, seguido de Neural Network (NN), Logistic Regression, Cluster Analysis y Genetic Algorithm.

Tabla 5: Frecuencia en uso de técnicas de predicción

Technique	Frequency
Decision Tree	26
Neural Network	16
Logistic Regression	15
Cluster Analysis	10
Genetic Algorithm	6
Markov Model	4
Naïve Bayes	4
k-nearest-neighbor	3
Bayesian Belief Network	3
Association Rule	2
Support Vector Machine	2
Bagging	2
CART	2
CHAID	2
K-Means	1
Fuzzy C means	1
influence diffusion model	1
Chr-PmRF	1
partial least squares (PLS)	1
C5.0	1
Structural Equation Model	1
Total	**104**

Fuente: Hashmi, Butt & Iqbal (2013)

Figura 9: Técnicas más usadas entre 2002-2013



Fuente: Elaboración propia basada en información del artículo estudiado

Por último, se ha identificado que el año donde más artículos de prestigio se han emitido fue en el 2012, seguido del 2010 y el 2011.

Tabla 6: Clasificación de artículos según el año de su publicación

Year of Publication	Frequency of Papers
2002	1
2003	3
2004	2
2005	3
2006	7
2007	6
2008	4
2009	3
2010	10
2011	7
2012	11
2013 (June)	4
Total	**61**

Fuente: Hashmi, Butt & Iqbal (2013)

- **Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications. 38(12), 15273-15285.**

Objetivo:

El objetivo de este artículo de investigación es la creación de un modelo de predicción de churn del área de tarjetas de crédito, mediante la selección de variables relevantes utilizando datos de tarjetas de crédito de un banco chino y realizando el análisis de estos a través los algoritmos de machine learning: Regresión logística y árboles de decisión.

Metodología:

La data fue recolectada de un banco chino anónimo. Se extrajo una muestra de 5456 clientes de manera aleatoria desde una base total de 60 millones de clientes. Esta base consideraba datos demográficos, transaccionales, información de la tarjeta de crédito etc. De

esta muestra se obtuvieron 135 variables las cuales fueron categorizadas en información personal del cliente, información básica de la tarjeta, información de riesgo del cliente e información transaccional. Mediante un análisis de multicolinealidad se redujeron a 95 variables: 12 de información personal, 33 de información de tarjeta, 4 de riesgo y 46 transaccionales. Debido a que la muestra contenía 91% de churners, se realizó pruebas con diferentes proporciones del dataset de test hasta llegar a una proporción de 1:1 entre usuarios churn y no churn en el dataset de test. Al cual se aplicó las técnicas de regresión logística y de árboles de decisión Se crearon modelos basados en las distintas categorías antes mencionadas. Creando modelos donde sólo contenían una categoría y otros donde era mixto.

Resultado:

El resultado fue que en el caso de la regresión logística el modelo 6 el cual contenía un mix de todas las categorías teniendo un error promedio del 12%. En el caso de los árboles de decisión, el mejor modelo fue el 7 que de igual manera contiene un mix de variables y que tuvo un error promedio de 16%

2.2. Bases Teóricas

2.2.1. Inteligencia Artificial

Este concepto aparece en la sociedad con el que ahora se considera el padre de la inteligencia artificial, Alan Turing, aunque él no fue quien acuño este término, si fue el primero en cuestionar si una máquina podía pensar y cómo esto sería posible, dentro de su artículo *Computing machinery and intelligence*, basándose en la psicología cognitiva, la cual sostiene que nuestro cerebro tiene un sistema de procesamiento de datos que se asemeja al de una computadora digital.

Turing fue el creador del conocido Test de Turing, el cual plantea que si un entrevistador al entablar una conversación con un humano y una computadora, al mismo tiempo, no logra diferenciar cuál es el humano y cuál la computadora, se concluirá que dicha computadora ha adquirido inteligencia o simplemente, es inteligente.

Años después, McCarthy (1956) califica a la inteligencia artificial como una ciencia, con la cual se logran hacer máquinas inteligentes, especialmente programas de cómputo inteligentes. Actualmente, un concepto más moderno es el que ofrece Rouhiainen (2018) quien

lo resume como una habilidad que necesitan los ordenadores para realizar actividades que solo pueden ser realizadas con la inteligencia humana.

En el desarrollo de esta investigación resumimos los conceptos de los autores consultados y nos referiremos a la inteligencia artificial como un conjunto de tecnologías que permiten replicar el intelecto humano en su comportamiento, mediante un aparato tecnológico.

A continuación, un resumen de las áreas en las que se utiliza Inteligencia Artificial y cuáles han sido sus principales aportes dentro de estas:

Tabla 7: Principales aportes de la Inteligencia Artificial

Área	**Principales aportes**
Gestión y Control	• Análisis Inteligente • Fijación de Objetivos • Planificación • Comunicación
Fabricación	• Diseño • Planificación • Programación • Monitorización • Control • Gestión de Proyectos • Robótica simplificada/visión computarizada • Diagnóstico en plantas • Integración de funciones de fabricación.
Educación	• Adiestramiento práctico • Exámenes • Diagnóstico
Equipamiento	• Diseño • Diagnóstico • Adiestramiento • Mantenimiento • Configuración • Monitorización • Ventas
Ingeniería	• Diseño • Control • Análisis
Cartografía	• Interpretación de fotografías • Diseño • Resolución de problemas cartográficos

Profesiones	• Abogacía • Medicina • Contabilidad • Geología • Química
Software	• Enseñanza • Especificación • Diseño • Verificación • Mantenimiento
Sistemas de armamento	• Guerra electrónica • Identificación de objetivos • Control adaptativo • Proceso de imágenes • Proceso de señales
Proceso de datos	• Educación • Interfaces en lenguaje natural • Acceso inteligente a datos/gestores de base de datos • Análisis inteligente de datos
Finanzas	• Planificación • Análisis • Consultoría

Fuente: Elaboración propia en base a Rauch-Hindin

La inteligencia artificial engloba distintas herramientas analíticas, este estilo de trabajo grupal entre las herramientas han permitido que la inteligencia artificial sea más eficiente al resolver problemas complicados. Así también, la inteligencia artificial puede ser clasificada en dos: IA débil e IA fuerte:

2.2.1.1. **Inteligencia Artificial Débil**

Se refiere a la simulación de inteligencia basada en imitar el razonamiento que manejamos los humanos, pero en un nivel básico y racional. Este tipo de IA ya ha sido creado y lo disfrutamos a diario mediante tecnologías cotidianas.

En la revisión de la literatura propia de la IA; Mendez G., Ramirez R., Mora G. (2020) indican que la IA débil es la solución oportuna para problemas relacionados a:

- Aprendizaje Automático, o Machine Learning: se alimenta al algoritmo con una gran cantidad de datos que permita identificar las mejores opciones y a su vez, las

excepciones. El concepto es usado frecuentemente en su versión en inglés incluso en el habla hispana, por lo que será utilizado en adelante.

- Métodos Probabilísticos
- Computación Evolutiva
- Sistemas Difusos

2.2.1.2. Inteligencia Artificial Fuerte

Este tipo de inteligencia aún no ha sido desarrollada ya que es muy compleja, busca enseñarle a una máquina a que aprenda y razone por sí mismo. Esto solo se ha visto en ficción como una visión de lo que se pretende conseguir: una máquina con inteligencia a un humano, no en conocimientos, si no en sus capacidades cognitivas.

Este tipo de inteligencia en la actualidad es un debate entre grupos que apoyan su estudio y ven con buenos ojos que la IA alcance e incluso supere la capacidad intelectual del humano para beneficio de nuestra sociedad, enfrentados a los opositores que piden frenar este avance por miedo a que se cree una IA fuerte tan completa que pueda ser orientada a causas destructivas, o incluso las diseñadas con buenos propósitos puedan desarrollar un método perjudicial para lograrlo.

Sobre este enfrentamiento, Tegmark (2020) resalta la importancia de la educación en esta sociedad encaminada al desarrollo tecnológico. Sabiendo que el principal miedo que tienen los opositores de la IA es que los humanos seremos reemplazados por máquinas inteligentes en nuestra ocupación laboral y muchas profesiones van a desaparecer, el autor defiende que esto ya sucede hoy en día, y se considera normal como y como indicador de progreso. Los trabajadores son reemplazados por robots a diario y muchas profesiones se ven destinadas al olvido, pero en el camino se van creando nuevas ocupaciones, justamente ante la necesidad de atender la IA.

Tabla 8: Comparación de IA débil e IA fuerte

IA DÉBIL	IA FUERTE
Pocas Redes Neuronales	Muchas Redes Neuronales

No puede razonar por sí solo	Imitan el razonamiento humano
Aprende de ejemplos similares	Aprende como un humano
Orientado a problemas muy concretos	Resuelve problemas abiertos
Reactivo	Proactivo
Sin Flexibilidad	Flexible
Programado por un humano	Auto programable

Fuente: Elaboración propia

2.2.2. Machine Learning

Conocido también como aprendizaje automático, es una rama de la Inteligencia Artificial que tiene como objetivo dotar a las máquinas con capacidad de aprender.

Según Alba y González (2020) el Machine Learning es una rama que tiene como objetivo descubrir patrones automáticamente a partir de datos, así luego puede imitarlos cuando se le pida decidir sobre una situación similar.

Por otro lado, Mendez, Ramirez y Mora (2020) la clasifican como una sub área de la inteligencia artificial, que engloba una serie de algoritmos y herramientas de modelado que se utilizan para procesar datos, que está presente en la mayoría de disciplinas científicas.

2.2.2.1. **Procesos de Machine Learning**

El proceso de inducción de conocimiento debe estar basado en respuesta al problema establecido. El Machine Learning busca presentar las situaciones en donde una máquina inteligente razona mejor que un ser humano ya que ha podido analizar toda la data.

Según Alba y Gonzales (2020) el proceso de Machine se define en las siguientes etapas:

2.2.2.1.1. Recopilación de Datos.

Para esta primera etapa, el objetivo es conseguir la mayor cantidad de información sobre el proceso analizado, esta labor suele ser la más difícil del proceso ya que obtener información confidencial de una empresa puede ser muy difícil. Se conoce la importancia de garantizar una considerable cantidad de información, pero de igual manera se debe procurar lograr un equilibrio con la calidad de data que se recopila, esto para garantizar el rendimiento del modelo

y evitar que al momento de realizar una limpieza, quede una cantidad menor a la necesaria para construir el modelo. Esta identificación de datos relevantes y su disponibilidad solo pueden ser realizadas por expertos, es un proceso que no se puede automatizar.

Identificar data sets: conjunto de datos mayormente tabulados, representado por filas y columnas. Puede encontrarse en página web, redes sociales, registro de llamadas en call center, foros, blogs, etc.

Recuperar Datos: se recoge la información encontrada en los data sets sin discriminar ningún tipo de dato, procurando recuperar datos de diferentes fuentes, la variedad también es fundamental para un mejor modelado. Adicionalmente, se debe tomar en cuenta la vigencia de los datos en el tiempo.

2.2.2.1.2. Pre procesado

Esta etapa consta de una serie de operaciones cuyo objetivo es hacer que los datos recolectados estén listos para el modelado. Algunos defectos más comunes que se encuentran en esta etapa es:

Valores Ausentes: Es común que por error del sistema, falla en la recolección de datos, o algún otro motivo encontremos valores nulos dentro de nuestros datos. El mayor problema de los valores nulos es que disminuyen el rendimiento del modelo de aprendizaje ya que un valor nulo no se puede trabajar numéricamente. Para solucionar este tema se puede aplicar diferentes técnicas como: interpolación, eliminación de valores nulos, etc.

Inconsistencia de datos: Son los errores en el formato de los datos, se da cuando los datos provienen de bases no estructuradas, data lakes, etc.

Valores duplicados: Puede suceder que en alguna ocasión encontramos registros duplicados en nuestro conjunto de datos. Es importante detectar estos registros y convertirlos en su correcto formato y valor, o en caso contrario, eliminarlos.

Duplicidad de los datos: Se puede dar por un error en la base de datos primaria o un error en la recolección y unificación de los datos. Al tener una mayor cantidad de datos similares, el modelo puede verse sesgado de manera inadvertida. Este tipo de error es el más sencillo de arreglar, ya que no es muy difícil darse cuenta de registros duplicados.

Outliers: También llamados valores extremos, son valores que se encuentran sumamente alejados de la media de los demás valores, generalmente se toman 3 desviaciones estándar alejados de la media. Es recomendable eliminarlos para evitar distorsionar la distribución de los parámetros cuantitativos.

Más allá de los errores que se pueden detectar en esta etapa, esta etapa también concentra los esfuerzos preparas la data para el modelado, entre las técnicas para adaptar la data tenemos:

Estandarización: Ya que muchas veces los algoritmos de machine learning se basan utilizan la distancia euclidiana para obtener la distancia de dos medidas, las escalas de valores de dos variables que difieran mucho. Ejemplo una escala de 0 a 10, con una escala de -1000 a 2000 puede afectar el rendimiento y la precisión de los modelos ya que el algoritmo considera ambas escalas como iguales, a pesar que expresan diferentes dimensiones. En base a esto, se utilizan distintas técnicas con el objetivo de estandarizar estas variables, una de ellas es minmax, mediante la cual utilizando la media, y los valores extremos se obtienen una escala de valor entre 0 y 1, permitiendo al modelo tomar en cuenta las escalas de cada dimensión sin problema.

Dummy Variables: Una variable ficticia es una variable que toma valores de 0 y 1, donde los valores indican la presencia o ausencia de algo (por ejemplo, un 0 puede indicar un placebo y 1 puede indicar un fármaco). Cuando una variable categórica tiene más de dos categorías, se puede representar mediante un conjunto de variables ficticias, con una variable para cada categoría. Las variables numéricas también se pueden codificar de forma ficticia para explorar efectos no lineales. Las variables ficticias también se conocen como variables indicadoras, variables de diseño, contrastes, codificación one-hot y variables de base binaria.

2.2.2.1.3. Aprendizaje Supervisado

Según Ruiz, G. (2019). “Los algoritmos de aprendizaje supervisado están basados en un modelo predictivo, el cual está compuesto por dos grupos de datos uno para realizar el entrenamiento, otro de prueba y un mecanismo que permita evaluar si el algoritmo está haciendo las cosas bien”

El primer grupo sirve para que era el modelo aprenda a clasificar las muestras, estos son etiquetados u organizados para dar un orden a la nueva información que estará siendo ingresado al sistema

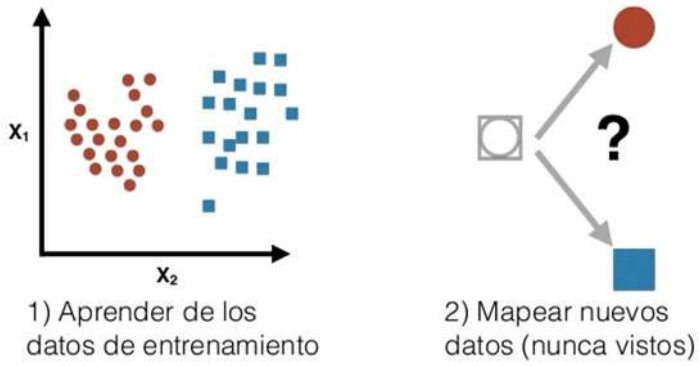
En este tipo de aprendizaje existen dos algoritmos el de clasificación y de regresión.

- **Algoritmo de clasificación**

Se usan cuando el resultado deseado es una etiqueta discreta, es decir cuando las respuestas están limitadas a finitos resultados posibles, al ser parte del aprendizaje supervisado a priori se conoce a que grupo pertenece el elemento en estudio.

En el caso que el modelo entrenado este realizado para predecir cualquiera de los dos grupos posibles, mediante la búsqueda de patrones en los datos tomados de referencia; luego tiende a comparar los nuevos grupos y predecir como por ejemplo en el caso que sea una clasificación binaria tendrá que elegir entre 0 ó 1, o si se busca predecir si un alumno aprobara el curso, si un cliente comprara un producto o si el cliente abandonara su actual empresa de servicio.

Figura 10: Diagrama de clasificación

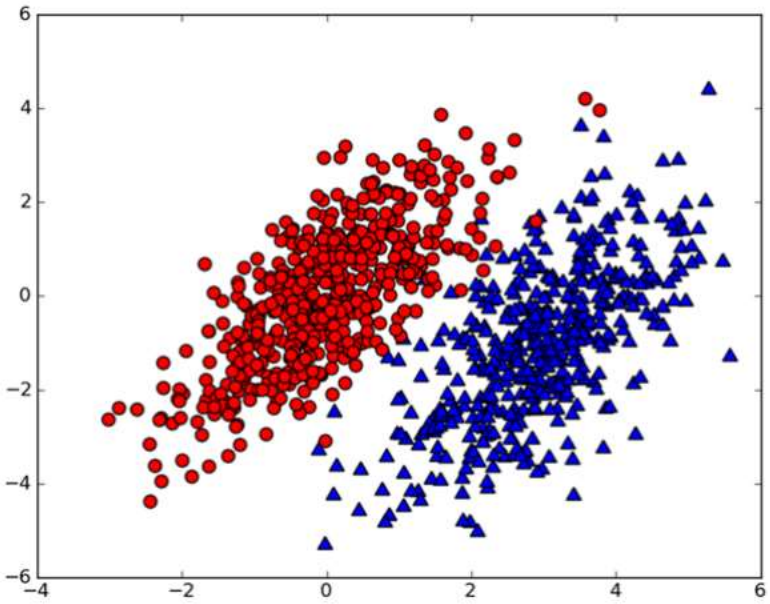


Fuente: Salcedo (2018)

- **Clasificador Naive Bayes**

Es un método probabilístico que busca aprender en base a la aproximación, considerando la mayor cantidad de evidencia posible. El clasificador ingenuo de Bayes considera que cada una de estas características contribuyen de manera independiente a la probabilidad de que este elemento sea validado, dentro de un mismo grupo de su propia especie, con una pequeña cantidad de datos de entrenamiento se puede estimar las variables para la clasificación.

Figura 11: Gráfico de clasificador Naiye Bayes



Fuente: Harrington P. (2012)

En la figura 11 se muestra dos distribuciones con parámetros conocidos que describen su distribución en el espacio

- **Regresión Logística**

Chittaroni (2002) clasificaba a la regresión logística como un instrumento estadístico para realizar análisis multivariado, que podía ser de uso explicativo como predictivo. Además, este instrumento es útil en el caso de contar con una variable dependiente dicotómica (puede tomar valor 1: presente o 0: no presente) y un conjunto de variables predictoras o independientes. Estas variables pueden ser cualitativas o cuantitativas. En el caso que la variable sea de tipo cualitativa, no podría ser procesada directamente, por lo que se debe realizar una conversión, pasar de ser una variable cualitativa a ser una variable simulada o como es más comúnmente conocida variable dummy.

El modelo de regresión logística viene dado de la siguiente formula:

$$\text{Logit}(P) = \log (P)/(1-p) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Siendo:

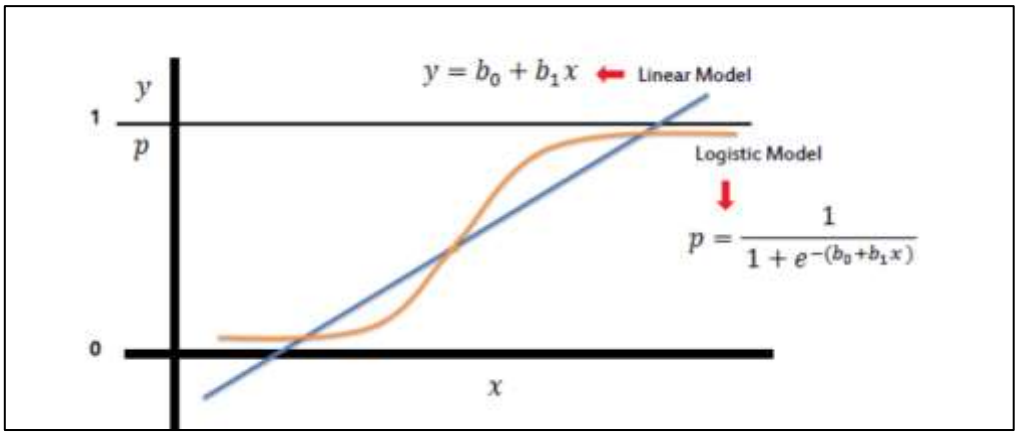
P: Probabilidad de evento de interés

b_0, b_1, b_2, …b_n: Los parámetros

X_0, X_1, X_2, …X_n: Variables independientes

Despejando el valor de P se tendría una ecuación de 2 variables (regresión logística binaria) y es representado gráficamente del plano xy, la cual será representada en la figura 12.

Figura 12: Regresión Logística para 2 variables



Fuente: Quiza J. (2018)

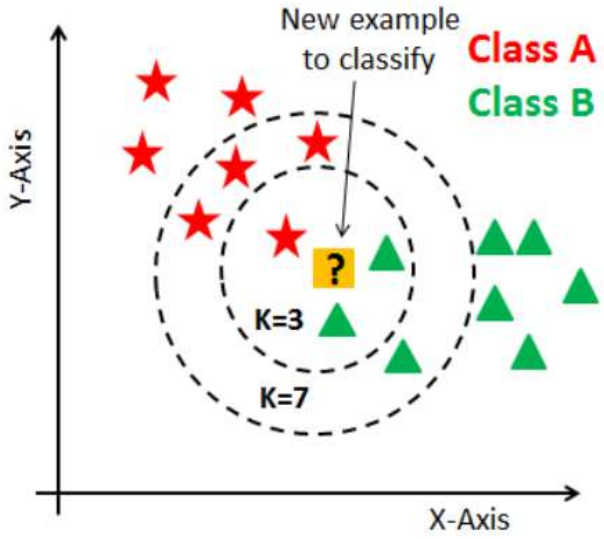
La regresión logística de naranja viene diferenciada de la regresión lineal en azul, para diferenciar el llamado límite de decisión, la cual indica que el valor menor a 0.5 será igual 0, mientras que todo valor por encima de 0.5 será igual a 1.

- **Algoritmo KNN**

Es una técnica que busca agrupar la muestra dada con los valores más cercanos del que se está tratando de predecir, se busca una clasificación sujeta al tipo de datos que lo rodean.

Existe una etiqueta previa de los datos anteriores, lo que hace el modelo es memorizar la fase de entrenamiento que sirven luego de base para la predicción, según se muestra en la figura 13 se realiza el cálculo de acuerdo a los elementos más próximos al elemento a analizar.

Figura 13: Gráfico de representación del algoritmo KNN



Fuente: Ichi Pro (2019)

- **Algoritmo de regresión**

El objetivo de estos algoritmos es predecir la relación entre ciertas características y una variable objetiva continua. El análisis de regresión se centra en fijar como dependiendo una variable y ver su comportamiento entre otras variables independientes.

Con este modelo planteado se puede construir un proceso de aprendizaje automático que ayude a entender la tendencia de los objetos de estudio y realizar pronósticos en el tiempo de acuerdo a este.

Los principales algoritmos de regresión son la regresión lineal y el árbol de decisión; son usados para predecir valores continuos mediante rectas que buscan aproximar el comportamiento de la variable dependiente en el tiempo.

2.2.2.1.4. Aprendizaje no supervisado

Para el aprendizaje no supervisado no tenemos una variable objetivo como lo teníamos en clasificación y regresión, quiere decir que no tenemos datos clasificados o etiquetados, por lo tanto, se busca predecir un valor desde nuestros datos de entrada es decir este tipo de aprendizaje obtiene información de las características propias de la base de ingreso sin conocer previamente la referencia de salida.

En este tipo de aprendizaje existen dos categorías específicas que se conocen como clustering y reducción dimensional.

- **Segmentación basada en agrupamiento (clustering)**

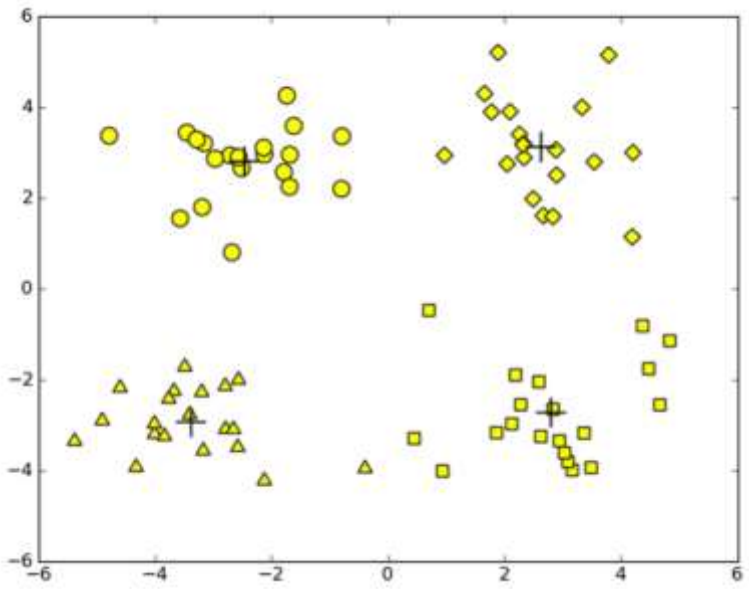
Es una técnica exploratoria que sirve para analizar datos en las que se organiza la información por grupos, sin conocer previamente la etiqueta o estructura propia que compone. Lo que realiza este método es agrupar datos con características idénticas a otros datos que tengan las características muy similares.

- **Algoritmo de K-means**

El principal objetivo es optimizar la partición de la imagen en áreas conforme a unas características dadas. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo. Se suele usar la distancia cuadrática.

Según Ruiz, G (2019) Se deben realizar los siguientes pasos: Inicialización: con el número de grupos, k, se establecen k centroides en el espacio de los datos. Asignación objetos a los centroides: se asigna el objeto a su centroide más cercano. Actualización centroides: la posición del centroide se ajusta tomando la posición del promedio de los objetos de cada grupo.

Figura 14: Diagrama K-means



Nota: Los centros de los conglomerados están marcados con una cruz.

Fuente: Harrington (2012)

- **Reducción de dimensionalidad**

Esta forma de agrupamiento busca reducir la cantidad de variables que controla un elemento, la idea de esta técnica es facilitar el uso de los conjuntos de datos puesto que es imperioso que estos se puedan agrupar más fácilmente, así como reducir el procesado de algoritmos muy complejos y por último facilitar la compresión de los resultados.

Análisis de componentes principales o PCA

El primer método para la reducción de la dimensionalidad se llama análisis de componentes principales. (PCA). En PCA, el conjunto de datos se transforma de su sistema de coordenadas original a un nuevo sistema coordinado. El nuevo sistema de coordenadas es elegido por los propios datos, en la cual se busca la reducción de dimensiones logrando una eficiencia en la cantidad de almacenamiento, el tiempo de procesamiento y la validación de los datos relevantes dentro del modelado.

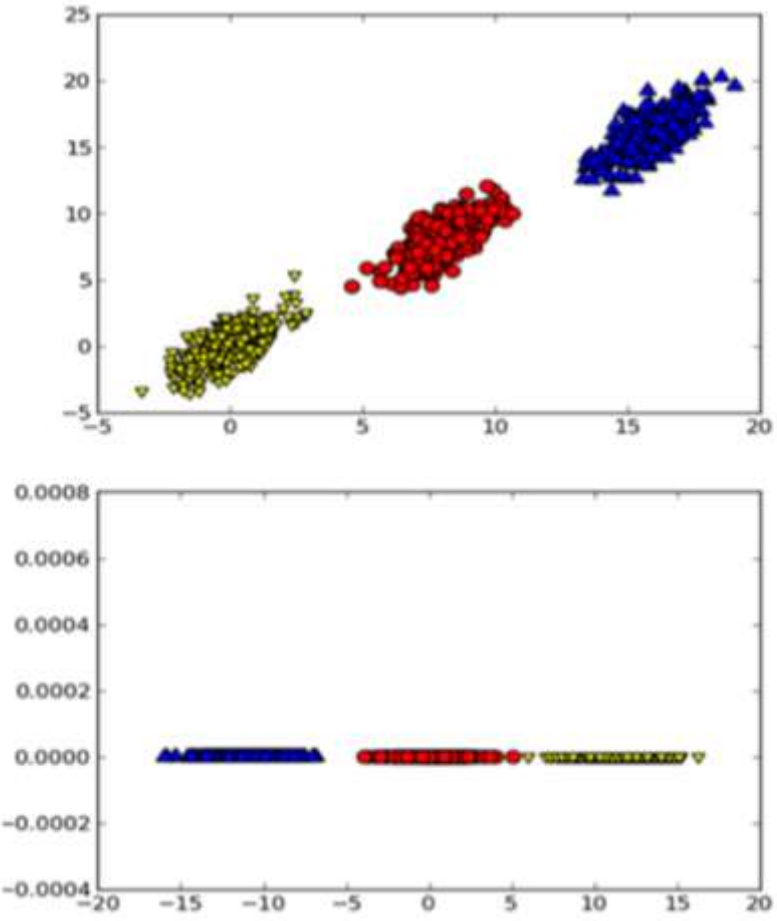
Según Ruiz G. (2019). Para la implementación del PCA, se usa la biblioteca de scikit.learn de Python, se debe seguir los siguientes pasos.

Paso 1 se realiza la división de los datos en dos conjuntos de entrenamiento y pruebas.

Paso 2 se realiza una normalización escalar estándar de los datos seleccionados.

Paso 3 Aplicamos PCA a los datos seleccionados.

Figura 15: Gráfico PCA reducido a solo 2 dimensiones



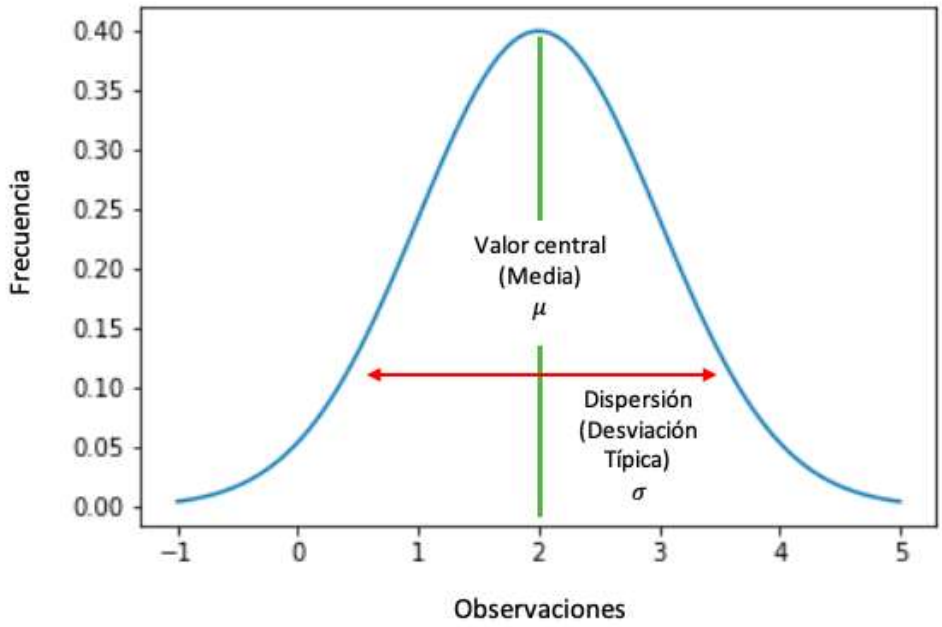
Fuente: Harrington P. (2012)

- **Distribución normal**

Es una de las distribuciones de probabilidades de variable continua que con más frecuencia aparece en estadística y en teoría de probabilidades.

La grafica de su función de densidad es acampanada y simétrica respecto a ciertos parámetros estadísticos. Este modelo teórico es capaz de aproximar de manera muy cercana el valor de una variable aleatorio a una función que dependa solo de la media y desviación estándar. El uso del modelo normal puede justificarse asumiendo que cada observación se obtiene como la suma de unas pocas causas independientes.

Figura 16: Gráfico de distribución normal



Nota: Gráfico de distribución normal.

Fuente: Rodo (2020)

2.2.2.1.5. Evaluación de modelo (Error Analysis)

En esta última etapa se procede a validar la precisión que tiene el modelo mediante una comparación entre la clasificación existente previamente y los resultados obtenidos del entrenamiento. Es decir, recordar cuál fue el problema que identificamos y cuestionar si el modelado que se ha construido resuelve debidamente dicho problema.

Para validar la veracidad del modelo, se necesita diseñar una métrica que indique la efectividad de la predicción y la coincidencia de valores. Si la exactitud es menor o igual a 50% la veracidad no será válida.

La Evaluación de modelo puede desarrollarse mediante:

Evaluación mediante un conjunto de test: se debe dividir el conjunto de datos en dos subconjuntos independientes: el de entrenamiento y el de test. El de entrenamiento es usado solo en la fase de aprendizaje y el de test para estimar el error.

Evaluación mediante validación cruzada (cross-validation): se deben dividir los datos en “k” k subconjuntos disjuntos con un tamaño similar (k-fold cross validation), esta

manera, se puede construir y evaluar un modelo, usando la unión de k-1 subconjuntos para el aprendizaje y el restante como test. Finalmente, cuando se obtiene la media aritmética de los ratios de error obtenidos se consigue el ratio de error para la muestra final.

Precisión del modelo de clasificación (Accuracy Score)

Tabla 9: Matriz de confusión

		Clasificación Real	
		Positivo	Negativo
Clasificación según Predicción	Positivo	VERDADEROS POSITIVOS (VP)	FALSOS POSITIVOS (FP)
	Negativo	FALSOS NEGATIVOS (FN)	VERDADEROS NEGATIVOS (VN)

Fuente: Elaboración propia

- Verdaderos positivos (VP): cantidad de positivos que fueron clasificados correctamente como positivos por el modelo.
- Verdaderos negativos (VN): cantidad de negativos que fueron clasificados correctamente como negativos por el modelo.
- Falsos positivos (FP): cantidad de negativos que fueron clasificados incorrectamente como positivos. Error tipo 1 (Falsos positivos).
- Falsos negativos (FN): cantidad de positivos que fueron clasificados incorrectamente como negativos. Error tipo 2 (Falsos Negativos).

Precisión = $\frac{VP+VN}{VP+VN+FP+FN}$

Ratio de Verdadero Positivo = $\frac{VP}{VP+FN}$

Ratio de Falso Positivo = $\frac{VN}{VN+FP}$

Precisión = $\frac{VP}{VP+FP}$

Rellamado = $\frac{VP}{VP+FN}$

2.2.3. Churn

Kamalraj, Malathi (2013) analiza el término directamente aplicado en el sector de telecomunicaciones, indicando que se trata del movimiento de clientes que cambian de los servicios de un proveedor por otro basado en distintas razones. Las principales razones para incitar este movimiento son la insatisfacción con la calidad del servicio, elevados costos, planes poco atractivos, mal soporte, etc.

También existen razones basadas estrictamente en las condiciones del consumidor, como puede ser la imposibilidad de pago o un cambio de residencia, donde la empresa no está disponible o el servicio no está disponible.

Sin embargo, la empresa también puede ser la decisora de cortar relaciones con el cliente, optando por cancelar el contrato por motivos de política de la empresa.

Se recopila una serie de variables propuesta por Van den Poel et al. (2004), donde clasifica a cuatro de ella como las más importantes:

- El comportamiento del cliente: se identifica que servicios y la frecuencia con que es utilizado por el cliente. Para nuestra investigación, que estudia el sector de telecomunicaciones, se evalúa la cantidad y duración de las llamadas, el período entre llamadas, el uso de la red para el intercambio de datos, etc.

- Las percepciones del cliente: identifica la opinión que el cliente va creando del servicio en uso. Normalmente se conocen estos datos mediante encuestas a los clientes orientando las preguntas a conocer el estado de su satisfacción. Se debe tomar en cuenta

la conexión de datos, la calidad del servicio, la satisfacción con el manejo de problemas reportados, satisfacción respecto a la ubicación, imagen y reputación de la empresa.

- La demografía: incluyen edad, sexo, educación, estatus social, los datos geográficos también se utilizan para calcular la rotación.

- El macro entorno del cliente: describe cambios importantes que se desarrollan en el mundo y experiencias del cliente que afectan la forma en que utiliza el servicio. Se puede considerar a las personas que han sufrido directamente la llegada de un desastre natural y que pueden confiar en usar sus celulares durante el mismo, aumenta las probabilidades de seguir utilizando el servicio.

2.2.4. Variables en la fuga de clientes

2.2.4.1. Variables demográficas

Según Celik (2019) en estudios anteriores, se muestran incluidas en el modelo variables demográficas como la edad y el sexo, así como otras variables relacionadas con el sector, que son sugeridas por los expertos y que se cree que tienen un efecto en la deserción de los clientes.

Las variables demográficas brindan información acerca de las características del cliente, tanto en edad, sexo, profesión o estatus socioeconómico; de esta forma describen el estilo de vida que tiene el cliente y como esta afecta en su comportamiento. Por ello termina siendo relevante conocer esta información puesto que, dependiendo de cada una de estos factores el comportamiento del cliente suele cambiar, ya que las necesidades de las personas cambian dependiendo la edad que tienen, estilo de vida que lleven, necesidad socioeconómico o profesión que tengan.

2.2.4.2. Variables de comportamiento del consumidor

El comportamiento del cliente se refiere a los hábitos o características de consumo de un individuo, incluidas las tendencias sociales, los patrones de frecuencia y los factores de fondo que influyen en su decisión de consumir algo. Las empresas estudian el comportamiento de los clientes para comprender a su público objetivo y crear productos y ofertas de servicios más atractivos. En el ecosistema digital de hoy, se puede recopilar y almacenar datos, que van más allá de las variables demográficas como edad, género, edad, ingreso, etc. Como menciona

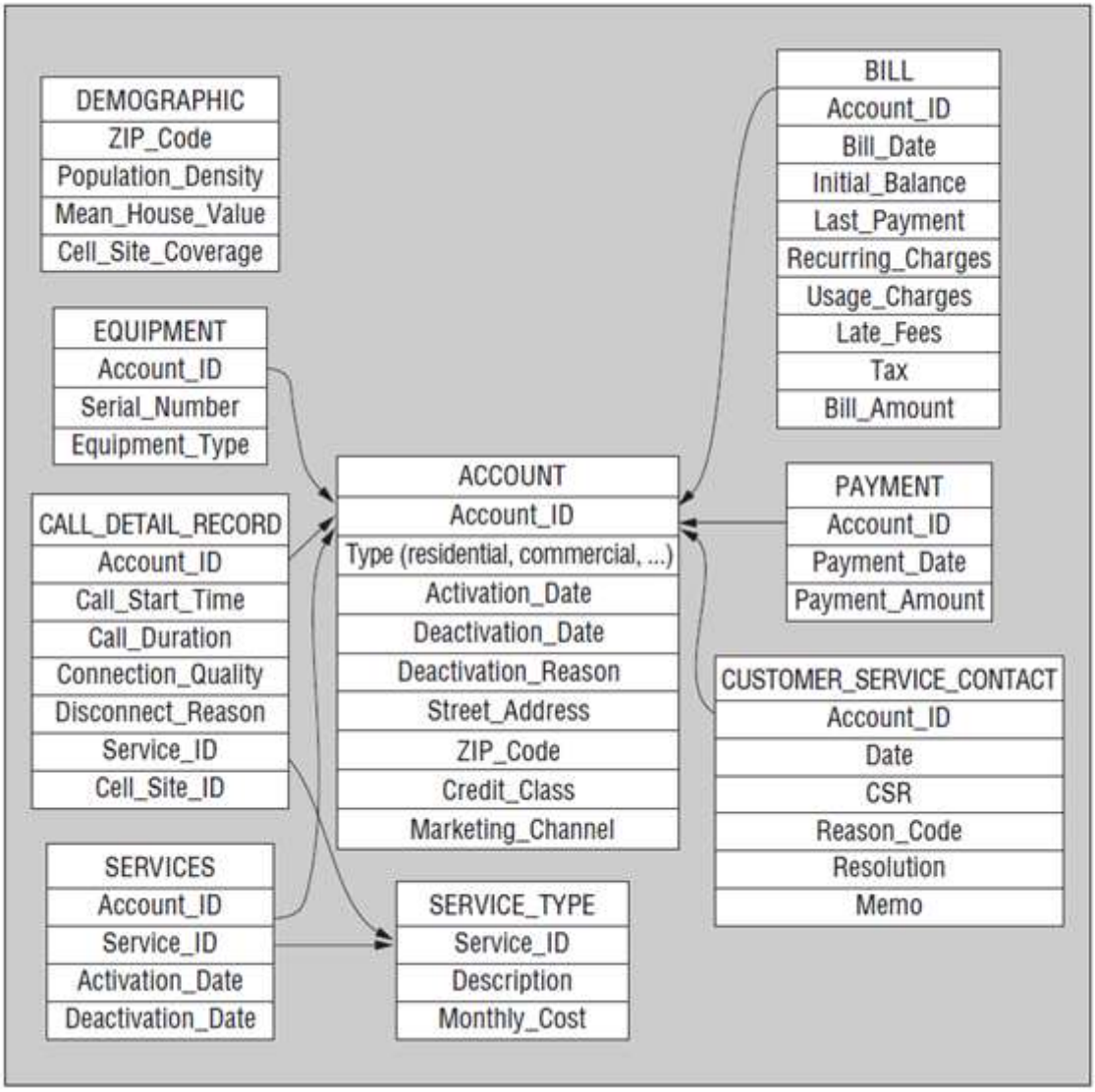
Fiegerman (2013), compañías como Netflix están basando sus sistemas de recomendaciones variables basadas en el comportamiento del consumidor, como los hábitos de consumo de películas, que tipo de películas ven, que recomendaciones ignoran, etc. Más que en la demografía a la que uno pertenece.

Figura 17: Modelo predictivo de Netflix y sus variables

Fuente: Varsini, Priya (2020)

En el campo de las telecomunicaciones, las variables de comportamiento de consumidor son diferentes a otros sectores ya que la empresa de telecomunicaciones brinda distintos tipos de servicios, ya sea llamada telefónica, consumo de datos, etc. Yan, Wolniewicz y Dodier (2003), mencionan un set de variables en su estudio de predicción de churn tal como se muestra en la figura 18. En su estudio tomaron en cuenta las variables de comportamiento de consumidor como la cantidad de minutos de llamadas, el tipo de servico (prepago o postpago), la fecha de activación del servicio, fecha de desactivación del servicio, etc. A partir de las conclusiones de su estudio, se concluyó que si bien no se podía realizar un modelo sin variables demográficas ya que aún presentaban relevancia dentro de su modelo, las variables del comportamiento del consumidor tenían una gran relevancia dentro del modelo.

Figura 18: Variables utilizadas en estudio



Fuente: Yan, L., Wolniewicz, R. H., & Dodier, R. (2004)

Capítulo III: Entorno Empresarial

3.1. Descripción de la empresa

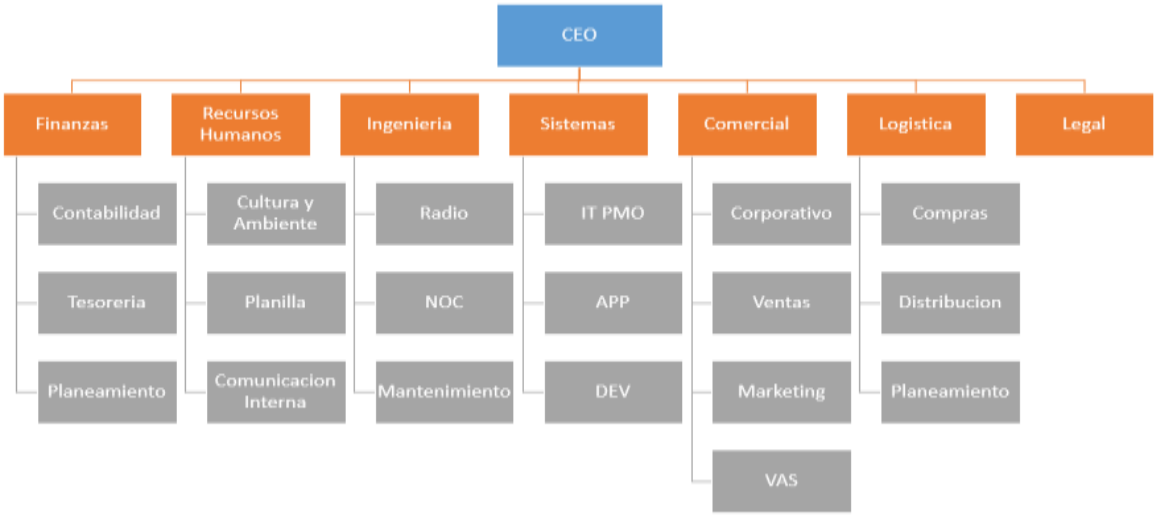
Bitel, nombre comercial de Viettel Peru S.A.C es una empresa de telecomunicaciones que opera en Perú desde el año 2014. Es parte del grupo VIETTEL, el cual es una compañía transnacional de telecomunicaciones vietnamita con sede en Hanoi, Vietnam. Bitel cuenta con una concesión del Ministerio de Transportes y Telecomunicaciones que le permite operar en la C de 1900 MHz por 20 años. Bitel fue la 4 empresa de telecomunicaciones extranjera en ingresar al Perú, detrás de Telefónica, Entel y Claro. Gestión (2019) señala que la estrategia de Bitel ha sido en ofrecer un servicio asequible al consumidor y llegar a todos los lugares del Perú, esta estrategia le valió obtener el 16.3% del mercado peruano de telefonía móvil para el año 2019. Bitel presta servicios de telefonía móvil prepago y postpago, servicio de internet fijo y móvil, y soluciones de telecomunicaciones a empresas y organizaciones del estado.

3.1.1. Reseña histórica y actividad económica

3.1.2. Descripción de la organización

3.1.2.1. Organigrama.

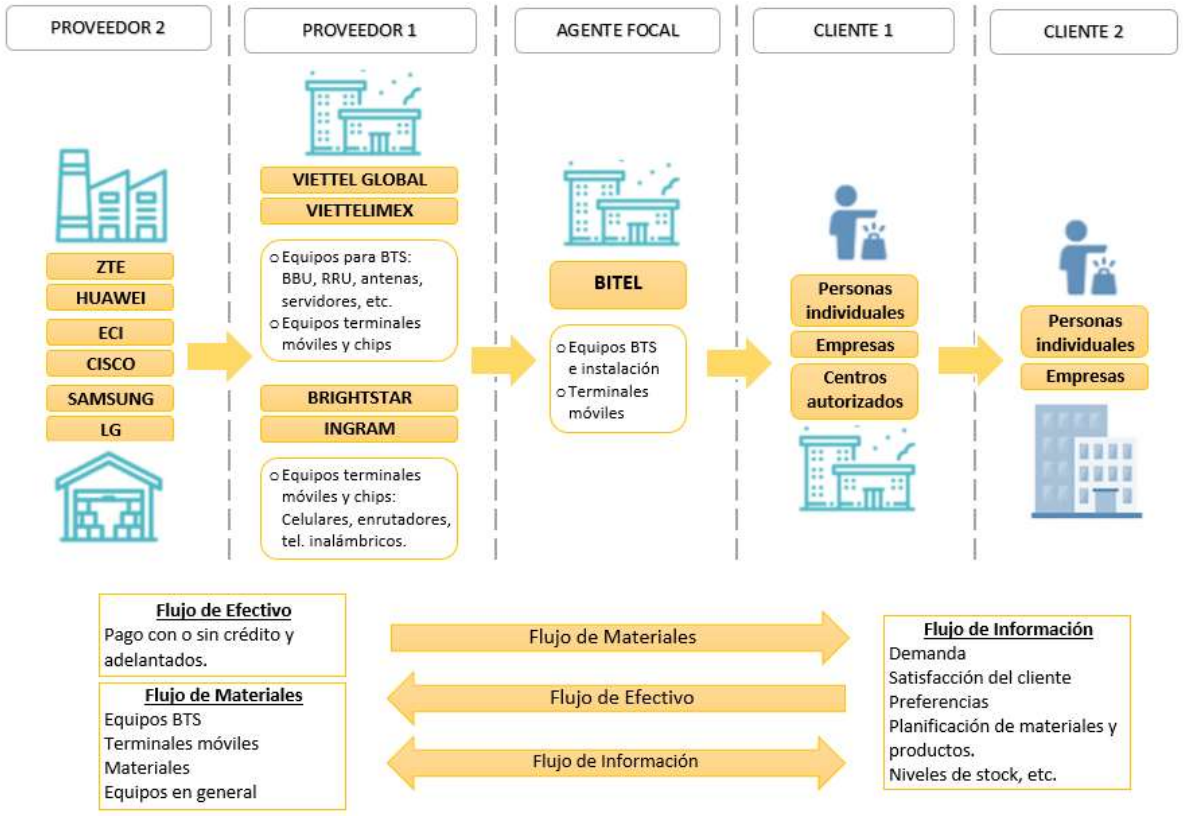
Figura 19: Organigrama de la empresa Bitel



Fuente: Bitel Perú (2021)

3.1.2.2. Cadena de suministros.

Figura 20: Cadena de suministro de Bitel



Fuente: Elaboración propia

Bitel adquiere sus suministros con el objetivo de vender móviles e implementar infraestructura de red. Para el caso de los móviles ofrecen marcas como Samsung, LG y Huawei y para la red ZTE, ECI, Huawei y Cisco. Para adquirir los productos de estas marcas Bitel usa intermediarios. Entre estos proveedores de proveedores de Bitel, ZTE es la marca principal para el funcionamiento de una BTS (Base Transceiver Station) y su red inalámbrica, y para el caso de móviles el más importante es la empresa Fortune Ship International en China, quien es el fabricante de todos los celulares marca Bitel.

Los intermediarios o proveedores directos más importantes de Bitel para adquirir las marcas antes mencionadas son Viettel Global y Viettelimex que son empresas del mismo grupo Viettel Global y las nacionales Ingram y Brightstar.

- Viettel Global y Viettelimex: compras para equipos de infraestructura de red y móviles de marca Bitel.

- Ingram y Brightstar: compras de móviles de marca Huawei, Lenovo, LG y Nokia

3.1.3. Datos generales estratégicos de la empresa

3.1.3.1. Visión, misión y valores o principios.

- Misión

Ser la empresa líder de telecomunicaciones en el Perú.

- Visión

Posicionarnos para 2023, como la empresa líder en telefonía móvil, acercando opciones de comunicación a todas las personas y empresas del Perú.

- Valores

  - ✓ Respeto a la ley
  - ✓ Cultura de Integridad
  - ✓ Entereza y Respeto

3.1.3.2. Objetivos estratégicos.

- ✓ Incrementar la rentabilidad económica y financiera.
- ✓ Ampliar la infraestructura de red.
- ✓ Incrementar nuestro segmento de mercado.
- ✓ Mejorar la satisfacción de los clientes.
- ✓ Mejorar el abastecimiento y distribución de productos y servicios.
- ✓ Ser reconocida como la empresa líder en el sector C y D.

3.1.3.3. Evaluación interna y externa. FODA cuantitativo.

La técnica FODA cuantitativa tiene como objetivo evaluar condiciones actuales de la empresa para identificar mejoras, estandarización de procesos que marchan bien, y eliminar posibles desventajas. Sin embargo, esta técnica indica que no toda debilidad debe ser mejorada, esto si dicha debilidad no está afectando el aprovechamiento de las oportunidades o si esta

debilidad pudiera ser un conductor para activar una amenaza. Esto debido a que la realidad dista de las teorías, y debemos procurar invertir en mejorar solo las debilidades que sean relevantes.

Lo mismo sucede para las fortalezas, muchas veces son declaraciones teóricas que no aportan a aprovechar las oportunidades que tiene la empresa, y tampoco ayudan a enfrentar las amenazas.

Iniciamos la evaluación de la matriz FODA, identificando las fortalezas y debilidades que tiene Bitel hoy en día, las oportunidades y amenazas que el entorno otorga. Describimos las siguientes variables:

Tabla 10: Matriz FODA de Bitel

FORTALEZAS	**DEBILIDADES**
F1: Aceptable satisfacción de sus clientes por cumplir con el servicio prometido.	D1: Baja presencia en el mercado A y B que predominan en compras.
F2: Alto nivel tecnológico.	D2: Falta de fidelidad de clientes.
F3: Ofrece servicios de alta calidad a menor costo que la competencia.	D3: Baja velocidad de respuesta en servicios de Internet. (descargas)
F4: Segunda operadora con mejor atención al cliente: capacidad de respuesta, empatía y venta transparente.	D4: Bitel aún es considerada una marca débil.
OPORTUNIDADES	**AMENAZAS**

O1: Incremento en la aceptación en clientes del sector A y B.	A1: Competidores muy arraigados al mercado.
O2: Desarrollo de nuevos productos en respuesta del emergente mercado de e-sports y gaming.	A2: Entidades reguladoras más estrictas.
O3: Nuevos mercados: mayor cantidad de zonas rurales que solicitan servicios de telefonía.(población rural poco atendida)	A3: Inestabilidad del panorama político - económico del país.
O4: Trabajo y clases no presenciales, generan demanda de servicio de telefonía y equipos celulares.	A4: Incremento de portabilidades a corto plazo.

Fuente: Elaboración propia

A continuación, debemos enfrentar los factores endógenos (fortalezas y debilidades) versus los factores exógenos (oportunidades y amenazas), puntuando del 1-7 según el análisis. Para los factores endógenos, se debe cuestionar si una fortaleza aporta al aprovechamiento de una oportunidad, siendo 1 poco aporte y 7 un buen aprovechamiento (1er cuadrante) y como las mismas fortalezas permiten afrontar las amenazas identificadas.

Para los factores exógenos, cuestionar si las debilidades no permiten aprovechar las oportunidades y como la debilidad permite activar las amenazas identificadas. En este caso, la puntuación es contraria, si calificamos con 7 se deduce que la debilidad permite que se active la amenaza, y si puntuamos con 1, todo lo contrario o en poca medida. Para las oportunidades, puntuar con el máximo (7) significa que la debilidad identificada no permite aprovechar la oportunidad con la que se está contrastando y si valoramos con una nota baja (1 la más baja) afirmamos que la debilidad no afecta en el aprovechamiento de la oportunidad.

Para fines de la presente investigación, los cinco integrantes del equipo analizaremos individualmente y colocaremos las puntuaciones que consideramos correspondan, por ello, mostraremos las cinco matrices de valoración en la sección de Anexos y a continuación una sexta matriz con el promedio de todas las anteriores. Es a partir de esta última matriz que

podremos producir conclusiones y comparar con las expectativas que teníamos acerca de la relevancia de cada variable y definir con cuáles se debe trabajar.

Tabla 11: Matriz FODA cuantitativo

MATRIZ FODA CUANTITATIVO - PROMEDIO GRUPO 3											
		OPORTUNIDADES					AMENAZAS				
		O1	O2	O3	O4	PROMEDIO	A1	A2	A3	A4	PROMEDIO
FORTALEZAS	F1	5.2	2.4	5.6	5.8	4.75	5.4	1.8	2.4	5.2	3.7
	F2	4.8	6.4	5.4	5	5.4	4.6	3	2.8	3.8	3.55
	F3	4.2	3.4	6	4.2	4.45	5.4	2.4	4.6	5	4.35
	F4	5.8	2.6	4.4	3.2	4	6	4.6	1.8	6.6	4.75
	PROMEDIO	5	3.7	5.35	4.55		5.35	2.95	2.9	5.15	
DEBILIDADES	D1	7	4.6	2	4	4.4	7	1.6	2.6	4.2	3.85
	D2	5.2	1.8	3.6	3.6	3.55	6.2	1.6	1.8	6.8	4.1
	D3	6.6	5	3.8	4.4	4.95	5.2	2.6	1.6	5	3.6
	D4	6.4	5.2	5.8	5	5.6	5.8	2.2	2.4	4.4	3.7
	PROMEDIO	6.3	4.15	3.8	4.25		6.04	2	2.1	5.1	

Fuente: Elaboración propia

Conclusiones del análisis de FODA cuantitativo:

Se deduce que la Fortaleza 2: Alto nivel tecnológico, es la más importante para lograr aprovechar las oportunidades descritas, y por el contrario, la Fortaleza 4: Segunda operadora con mejor atención al clientes: capacidad de respuesta, empatía y venta transparente, es la fortaleza que menos impacta en el aprovechamiento de las oportunidades.

Se debe ejecutar la Oportunidad 3: Nuevos mercados: mayor cantidad de zonas rurales que solicitan servicios de telefonía (población rural poco atendida), ya que sobresale por encima del resto de oportunidades como la que mejor prospecto de éxito tiene. Posteriormente, se deberían considerar la Oportunidad 1: Incremento en la aceptación en clientes del sector A y B, la Oportunidad 4: Trabajo y clases no presenciales, generan demanda de servicio de telefonía y equipos celulares. En el caso de la Oportunidad 2: Desarrollo de nuevos productos en respuesta del emergente mercado de e-sports y gaming, se entiende que haya quedado en último puesto

ya que es una oportunidad de integración vertical y su éxito no se vería reflejado en los servicios actualmente ofrecidos.

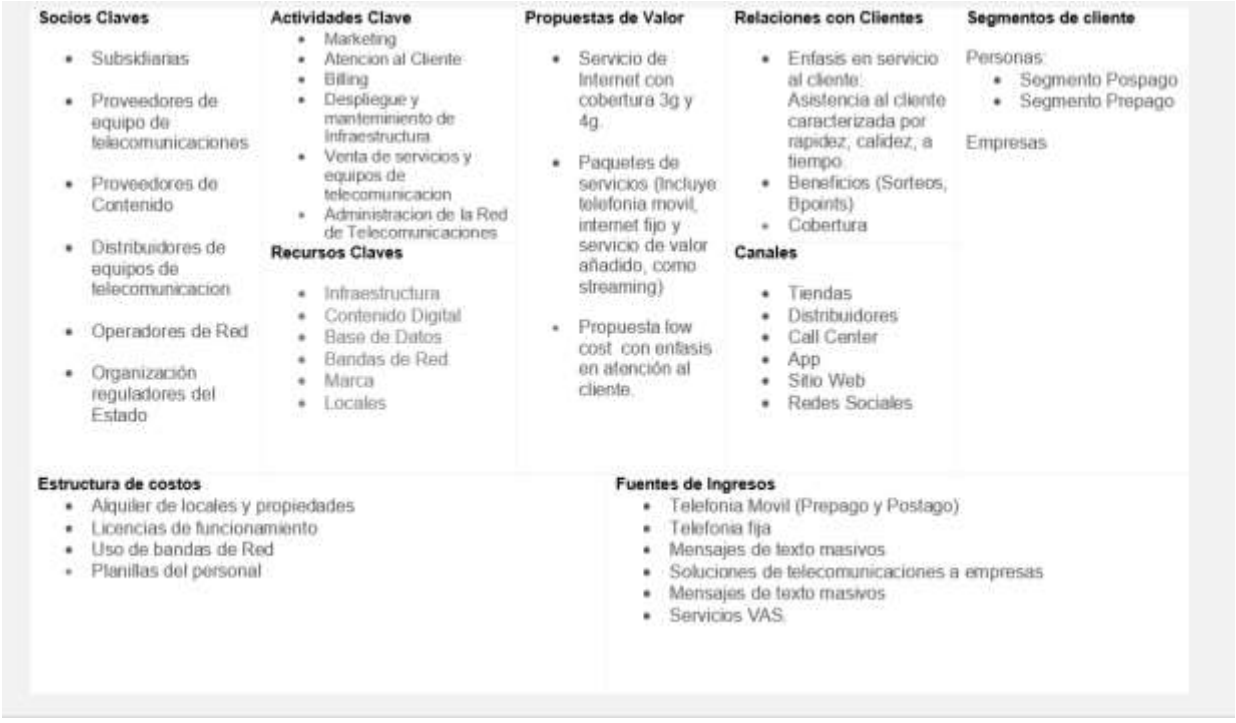
La Fortaleza 4: Segunda operadora con mejor atención a los clientes: capacidad de respuesta, empatía y venta transparente, es la que mejor puede combatir las amenazas a las que se enfrenta la empresa actualmente, también debería considerarse la Fortaleza 3: Ofrece servicios de alta calidad a menor costo que la competencia, ya que ha obtenido un promedio muy similar a la fortaleza 4.

Entre las debilidades identificadas, la Debilidad 4: Bitel aún es considerada una marca débil, se debe resolver con prioridad ya que nos frena al aprovechamiento de oportunidades y puede permitir la activación de algunas amenazas.

3.2. Modelo de negocio actual (CANVAS)

Mediante el modelo canvas se plantea una visión estratégica de la empresa que permita conocer los aspectos claves del negocio y de cómo se relacionan entre sí. Se plantea desde la información de los socios claves, la propuesta valor que se ofrece a través de los distintos canales y de que forma esta genera ingresos monetarios a la empresa.

Figura 21: Modelo de Negocios de Bitel

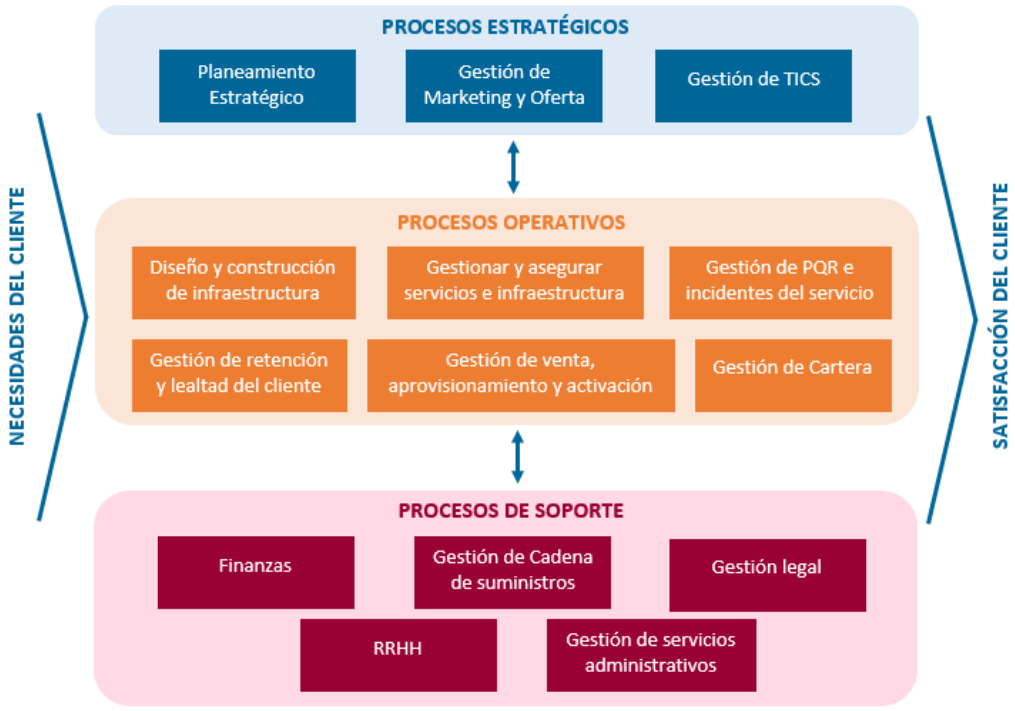


Fuente: Elaboración propia

3.3. Mapa de procesos actual

Una vez identificados, clasificados y jerarquizados los procesos es conveniente representarlos gráficamente, de manera que se puede tener un gráfico representativo de las diferentes áreas y sus funciones dentro de la empresa, así como la interrelación entre las necesidades y la satisfacción del cliente.

Figura 22: Mapa de procesos de Bitel



Fuente: Elaboración propia (2021)

Capítulo IV: Metodología De La Investigación

4.1. Diseño de la Investigación

El presente trabajo, constituye una investigación experimental, según lo planteado por Hernández, Fernández y Baptista (2014), puesto que es un estudio donde se manipulan intencionalmente la variable o las variables independientes para evaluar el efecto que tienen en la variable dependiente. En este caso específico, la variable independiente Regresión Logística se manipulará y veremos cómo afecta a la variable dependiente que es la probabilidad de fuga de los clientes.

Figura 23: Esquema de experimento y variables

Fuente: Hernández, Fernández y Baptista (2014)

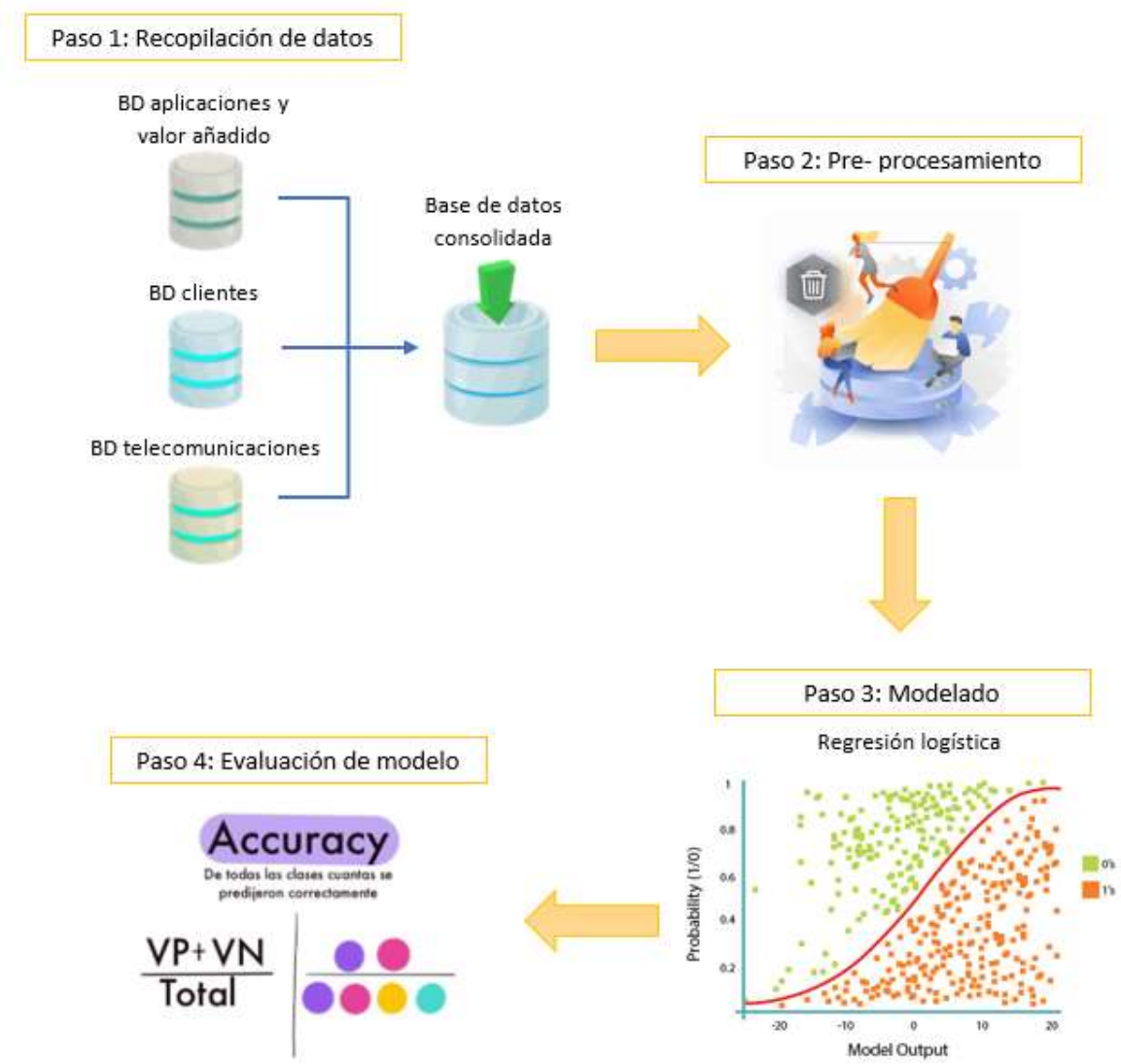
Asimismo, se trata de una investigación transversal debido a que la base de datos utilizada se dio durante un periodo de tiempo determinado. Para esta investigación en particular, se toma como periodo comprendido entre los meses de julio a octubre del 2021.

Por otro lado, la investigación tiene un enfoque cuantitativo ya que se hará una recolección y análisis de datos para predecir escenarios con el uso de una técnica de Machine Learning. Además de que los pasos se harán de manera secuencial sin eludir otros. Finalmente, se medirán y analizarán las variables para extraer una serie de conclusiones.

El alcance será correlacional pues se tiene como finalidad conocer la relación de las variables en un contexto específico.

4.2. Metodología de implementación de la solución

Figura 24: Metodología de implementación de la solución



Fuente: Elaboración propia

Paso 1: Recolección de datos

Se utilizarán tres bases de datos brindadas por la empresa Bitel para elaborar los registros de fuga de clientes durante el periodo comprendido entre los meses de julio a octubre de 2021. La base de datos obtenida al integrar las bases de datos clientes, telecomunicaciones y valor añadido contiene 28092 registros de usuarios de la empresa con sus respectivas características o atributos; los cuales son de tipo cualitativo y cuantitativo respectivamente. Estos datos fueron recolectados de las siguientes bases de datos:

- Base de datos de Clientes: plan de telefonía (producto_type), marca de celular (device_type), nombres, edad, sexo, cliente migrado de otra empresa (portabilidad), etc.

- Base de datos de Telecomunicaciones: cantidad de minutos, mensajes (sms), datos (megabytes) etc.

- Base de datos de Aplicaciones y VAS: uso de app mibitel, suscripción a servicios de valor añadido (vas), publicidad, etc.

Finalmente, para facilitar el análisis se unirán los datos en un solo archivo, que es el que servirá para la siguiente etapa del modelado.

Paso 2: Pre-procesamiento

En esta etapa se hará el análisis de la base de datos consolidada y se procederá a realizar la limpieza, es decir, la eliminación de datos ausente o missing values y de los datos otuliers o espurios es decir aquellos valores atípicos o inusuales que pudieran haberse originado por errores de inserción de data.

Paso 3: Modelado

Para este siguiente paso se realizará el modelado de los datos usando la técnica de Regresión Logística del Machine Learning. Se hará la programación con Python a fin de poder identificar el modelo predictivo del churn para datos o entradas futuras.

Paso 4: Evaluación de Modelo

En esta última etapa, se valida la precisión del modelo al comparar los resultados obtenidos con la clasificación existente.

4.3. Metodología para la medición de resultados de la implementación

4.3.1. Instrumentos de medida

Tabla 12: Instrumentos de medida

Variables	Validación	Datos

Churn	Automático por decisión del cliente	Registro del Churn actual por cada usuario y sus características.
Modelo de Regresión Logística	Experto en Machine Learning	Registro de veces que la regresión clasifica correctamente el churn. Registro de veces que la regresión clasifica incorrectamente el churn.

Fuente: Elaboración propia

4.3.2. Operacionalización de Variables

Tabla 13: Operacionalización de Variables

Variable Independiente - X	Regresión Logística: técnica para lograr clasificar el churn a partir de las características de cada usuario.
Variable Dependiente - Y	Clasificación real del churn
Indicadores	Nivel de *Accuracy:* $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$ *Confusion Matrix* True Class / Positive / Negative Predicted Class – Positive: True Positive Count (TP) | False Positive Count (FP) Predicted Class – Negative: False Negative Count (FN) | True Negative Count (TN)

Fuente: Elaboración propia

4.4. Cronograma de actividades y presupuesto

Esta investigación se realizó en un determinado tiempo. A continuación, se detalla las actividades que se realizaron para llevarlo a cabo.

Tabla 14: Cronograma de actividades

Actividades	Setiembre				Noviembre				Noviembre				Diciembre			
	SEM1	SEM2	SEM3	SEM4	SEM1	SEM2	SEM3	SEM4	SEM1	SEM2	SEM3	SEM4	SEM1	SEM2	SEM3	SEM4
CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA																
Descripción de la Realidad Problemática																
Justificación de la Investigación																
Delimitación de la Investigación																
CAPÍTULO II: MARCO TEÓRICO																
Antecedentes de la Investigación																
Bases Teóricas																
CAPÍTULO III: ENTORNO EMPRESARIAL																
Descripción de la empresa																
Modelo de negocio actual																
Mapa de procesos actual																
CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN																
Diseño de la Investigación																
Metodología de implementación de la solución																
Metodología para la medición de resultados de la implementación																
Cronograma de actividades y presupuesto																
CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN																
Propuesta solución																
Medición de la solución																
CIERRE																
Corrección del informe																
Sustentación																

Fuente: Elaboración propia

Para el desarrollo de esta investigación se está estimando un presupuesto de los gastos realizados para llevarlo a cabo por todos los participantes de este estudio. A continuación, se muestra la tabla del presupuesto realizado, considerar que los montos fueron aproximados.

Tabla 15: Presupuesto de la investigación

Recurso	Cantidad	Costo
Equipo: Laptop	5	S/. 2500.00
Software: Jupyter Notebook	5	S/. 0.00
Electricidad	5	S/. 200.00

Internet	5	S/. 500.00
Total		S/. 3200.00

Fuente: Elaboración propia

Capítulo V: Desarrollo de la Solución

5.1. Propuesta solución

5.1.1. Planeamiento y descripción de Actividades

El modelo integrará información de las bases de datos “Telecomunicaciones”, “Aplicaciones y VAS” y “Clientes”. En segundo lugar, se procederá a depurar eliminando los registros que contengan valores nulos en los principales parámetros, con el fin de obtener datos listos para su aplicación. Luego, se modelará la información a través del algoritmo de aprendizaje supervisado regresión logística. Finalmente, se evaluará la precisión del modelo con el fin de obtener un modelo estadísticamente válido, mediante el indicador de precisión denominado accuracy score.

5.1.2. Desarrollo de actividades. Aplicación de herramientas de solución.

5.1.2.1. Recolección de datos

Se extrajeron los datos con los parámetros que el equipo definió como principales bases de datos ORACLE: telecomunicaciones, clientes y aplicaciones, mediante el siguiente código (los dblinks, nombres de tablas y columnas se describen utilizando alias con el fin de mantener el anonimato de estructura de la base de datos de Viettel Perú), mencionado que la tabla primaria master.gen_report de la base de telecomunicaciones ya estaba creada con una muestra aleatoria de clientes (personas naturales) tanto activos como clientes que habían optado por terminar la relación de servicio con VIETTEL PERU entre los meses de Julio a Octubre de 2021:

Figura 25: Sentencia SQL para la extracción de datos

```
109
110     select
111             tel.sms SMS,
112             tel.llamadas VOICE,
113             sum(tel.sum_up_down) Datos,
114             cl.product_type,
115             cl.edad,
116             cl.sexo,
117             cl.dias_Activo TiempDePermanencia,
118             isnull(c.reclamos,0) reclamos,
119             cl.shop_name ciudad,
120             ap.vas vas,
121             cl.apr_ads ads,
122             tel.device_type,
123             case when cl.portabilidad is not null then 1
124             else 0 end portabilidad,
125             cl.mibitel
126     from master.gen_report tel
127     inner join master.clientes@dblink_clientes cl on cl.cliente_id = tel.id
128     left join master.app@dblink_app ap on ap.cliente_id = tel.id
129     where fecha_salidad >= '01-07-2021' and fecha_salidad < '01-11-2021'
130
131
```

Fuente: Elaboración propia

A través de este query o consulta, se obtuvo un total de 28092 registros. Luego se pasó a exportar esta información a un archivo csv, denominado base_1.csv con el fin de poder explorar la información de manera rápida a través de tablas dinámicas.

5.1.2.2. **Pre procesamiento**

5.1.2.2.1. Análisis de Variables

Con la base de datos consolidada se procedió a analizar la información y se obtuvo las siguientes variables:

Tabla 16:Variables de estudio

Nombre	Grupo de Variable	Descripción	Unidad de medida
Cliente ID	Identificador	Código del cliente	-
Product type	Comportamiento del consumidor	Tipo de plan	Prepago/ Postpago
Device type	Comportamiento del consumidor	Marca de equipo	-
ADS	Comportamiento del consumidor	Acepta publicidad	Si/ No

VAS	Comportamiento del consumidor	Servicio de Valor añadido	Si/ No
SMS	Comportamiento del consumidor	Mensajes	Unidades
Voice	Comportamiento del consumidor	Llamadas	Minutos
Datos	Comportamiento del consumidor	Internet	Megabytes
Portabilidad	Comportamiento del consumidor	Clientes que migraron anteriormente	Si/ No
Mi bitel	Comportamiento del consumidor	Usan la app mi bitel	Si/ No
Tiempo de permanencia	Comportamiento del consumidor	Antigüedad como cliente	Meses
Quejas y/o Reclamos	Comportamiento del consumidor	Quejas del servicio	Si/ No
Sexo	Demográfica	Género	Hombre/ Mujer
Cuidad	Demográfica	Lugar de contratación	-
Edad	Demográfica	Edad del cliente	Años

Fuente: Elaboración propia

A continuación, se realiza el análisis para depurar las variables que no son significativas para la predicción de fuga de clientes. Para esta toma de decisiones, hemos tomado como referencia los estudios previos y las herramientas estadísticas que sustenten el impacto que tienen dichas variables en el churn.

- **Análisis teórico**

Según el estudio realizado por Castro, J. y Pérez, E. (2020) que también evalúa el abandono de clientes en una compañía de telecomunicaciones en México, se puede apreciar que las siguientes variables resultan indispensables o de gran influencia para predecir la fuga del cliente:

- La edad del cliente: existe un gran número de clientes jóvenes que tienden a migrar de servicio.
- Género: la fuga del cliente no se relaciona con el género del cliente.
- Tiempo de permanencia: la tasa de fuga del cliente se ve disminuida conforme el cliente permanece mayor tiempo con el servicio.

Por otro lado, para Vargas, M. y Pineda, W. (2019) en su trabajo de grado de un Modelo Logístico con Datos Funcionales Aplicado al Churn en un Operador Móvil en Colombia, nos menciona que las variables más significativas para la predicción del churn son:

- Antigüedad del usuario en la compañía
- Cantidad de megas usadas
- Cantidad de llamadas realizadas

De igual manera, Rodríguez M. (2020) en su investigación y análisis de la predicción de Churn en una empresa de Telecomunicaciones de España, para la obtención del grado de Máster indica que las variables más relevantes para poder realizar la predicción del churn de un cliente son:

- Antigüedad del usuario en la compañía
- Servicios de valor añadido: múltiples líneas en la compañía, servicios de streaming.

De los diferentes estudios observamos que la ciudad (city_name) en la que se realizó el registro de las líneas telefónicas no tiene gran relevancia en el problema de la fuga de clientes, puesto que los clientes de todas las ciudades dan por concluido el servicio independientemente de la cobertura.

Asimismo, observamos que las variables como edad, antiguedad en la empresa y las variables de consumo: cantidad de minutos, mensajes y megabytes son incluidas en más de un estudio; siendo muy relevantes para el análisis y la predicción de fuga de los clientes (churn)

- **Análisis estadístico**

Dentro de las variables numéricas se encuentran SMS_1, DATA_1, VOICE_1, Tiempo de permanencia y Edad.

Se plantea el siguiente cuadro estadístico de los factores más representativos para el respectivo análisis de las variables cuantitativas.

Tabla 17: Descripción Estadística

Variable	Cantidad	Promedio	Mediana	Moda	Desviación estándar	Mín.	Máx.
SMS_1	28092	7.57	4.00	0.00	32.03	0	1351

DATA_1	28092	6298.36	4886.50	0.00	7325.62	0	86812
VOICE_1	28092	133.80	98.00	0.00	223.68	0	3806.62
Tiempo De Permanencia	28092	37.57	38.00	12.00	22.04	0	75
Edad	28092	32.09	32.00	40.00	7.81	19	45

Fuente: Elaboración Propia

La cantidad de data valida obtenida para el presente análisis es de 28092 valores.

- ✓ **Cantidad de mensajes de texto enviados por el cliente**

La variable representada como SMS_1 define la cantidad de mensajes de texto es una variable representativa dentro del análisis, puesto que representa la actividad del cliente con los servicios que brinda la empresa, si bien los mensajes de texto han quedado bastante relegados hoy en día, la mayor cantidad de los clientes que usan el servicio se encuentra entre los 30 y 40 años siendo esto los que más usaban los mensajes de texto como método cotidiano de comunicación.

Para esta variable se encontró que el valor promedio de mensajes de texto enviado por persona es de 7.57 por mes, mientras que el valor que más se repite es de 0, ya que con los servicios de internet que brinda el servicio cuenta con vías alternas de comunicación.

- ✓ **Cantidad de megabits usadas por el cliente durante un mes**

Desde los últimos años el uso de los datos o megabits se ha incrementado, por ello la variable DATA_1 es de uso frecuente en este servicio, la cual representa la actividad y el uso diario del cliente con su telefonía móvil.

Se obtuvo como promedio el uso de 6298 megabits por mes con un valor reiterativo de 0 megabytes usados por mes; mientras que la desviación estándar indica un alto grado de dispersión de sus valores.

- ✓ **Cantidad de minutos en llamada usado por el cliente en un mes**

La variable VOICE_1 llega a ser una de las variables relevantes para calcular la interactividad del cliente con su proveedor de telefonía móvil, puesto que define la cantidad de minutos usados en su plan contratado o por medio de recarga prepago.

La cantidad de minutos usados por el cliente durante un mes es en promedio de 133 minutos con una moda de 0 minutos en llamada por mes y con un alto índice de dispersión de sus valores.

- ✓ **Tiempo de permanencia.**

Según Beltran (2019) Después de realizar diversos métodos para evaluar sus diferentes variables, se obtiene que la tenencia o tiempo de permanencia como una de las variables más relevantes para poder obtener la probabilidad de fuga de un cliente.

Esta variable indica el tiempo de estadía de un cliente en la empresa Bitel, el cual llega a ser un factor altamente representativo con el que se puede evaluar el tiempo total de meses que un cliente permanece afiliado a una compañía de telefónica.

El valor promedio obtenido para esta variable es de 37.57 meses, siendo el valor más repetitivo encontrado de 12 meses como el tiempo que la mayoría de clientes toma en abandonar el servicio.

- ✓ **Edad**

Con respecto a la variable demográfica edad, tenemos una edad promedio de 32 años, mientras que respecto al valor nominal que más agrupa valores es la edad de 40 años con 1126 registros y el valor con la menor frecuencia es la edad de 29 años con 947 registros.

Tabla 18: Frecuencia de la edad de los clientes

Rango edad	fi	Fi	hi	Hi
[19-21>	3110	3110	0.111	0.111
[21-25>	3195	6305	0.114	0.224
[25-28>	2989	9294	0.106	0.331
[28-31>	2986	12280	0.106	0.437
[31-34>	3116	15396	0.111	0.548
[34-37>	3084	18480	0.110	0.658
[37-40>	3222	21702	0.115	0.773
[40-43>	3243	24945	0.115	0.888
[43-45]	3147	28092	0.112	1

Fuente: Elaboración propia

Existe una distribución de los clientes bastante equitativa teniendo como los más representativas los clientes mayores a 31 años.

- ✓ **Device_Type**

Esta variable representa la marca de celular que usa el cliente es una variable de comportamiento del consumidor. A continuación, se muestra la frecuencia de los valores.

Tabla 19: Frecuencia de Valores de la Variable Device_Type

Device Type	Cantidad de Mobiles	% del Total
Samsung	8474	30.17%
HUAWEI	4148	14.77%
Redmi	1931	6.87%
Motorola	1685	6.00%
LG	1259	4.48%
ZTE	941	3.35%
SKY DEVICES	855	3.04%
Fortuneship	826	2.94%
HYUNDAI	819	2.92%
Lenovo	633	2.25%
B mobile	508	1.81%
...		
Cinterion	2	0.01%
Richpad	2	0.01%
Own FUNVALUE	2	0.01%
NUU	2	0.01%
KINGCOMM	2	0.01%
Vozz	2	0.01%
lephone	2	0.01%
Total general	**28092**	**100.00%**

Fuente: Elaboración Propia

Como se puede apreciar en la tabla anterior, el 80% de todos los valores de la variable device_type se encuentran agrupados en solo once categorías o marcas de celular, En base a esto y con el fin de poder trabajar con la data de manera más sencilla, se procedió a agrupar los otros 150 valores con la nomenclatura otros, quedando como se aprecia en la siguiente tabla.

Tabla 20: Frecuencia de Valores de la Variable Device_Type Homologada

Device Type	Cantidad de Mobiles	% del Total
Samsung	8474	30.17%
HUAWEI	4148	14.77%
Redmi	1931	6.87%
Motorola	1685	6.00%
LG	1259	4.48%
ZTE	941	3.35%
SKY DEVICES	855	3.04%
Fortuneship	826	2.94%
HYUNDAI	819	2.92%
Lenovo	633	2.25%
B mobile	508	1.81%
Otros	6013	21.40%
Total general	**28092**	**100.00%**

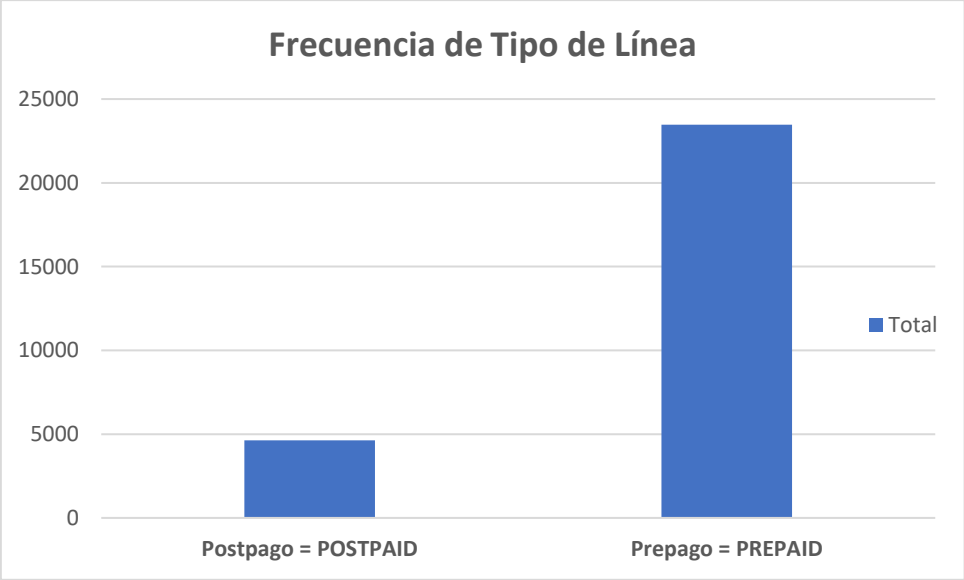
- ✓ **Product_Type**

Esta variable representa el tipo de plan o servicio del cliente. Es decir que puede ser prepago o post pago. A continuación, se muestra la tabla de frecuencia de los valores que puede tomar esta variable:

Tabla 21: Valores de la Variable Product_Type

Product_Type	Cantidad de Líneas
PostPago	4632
Prepago	23460
TOTAL	**28092**

Figura 26: Frecuencia de Product_Type



Fuente: Elaboración Propia

Como se aprecia en la tabla anterior, la gran mayoría de los clientes de Bitel son de tipo prepago, esto se debe a que el público objetivo de la compañía se encuentra en los sectores C y D.

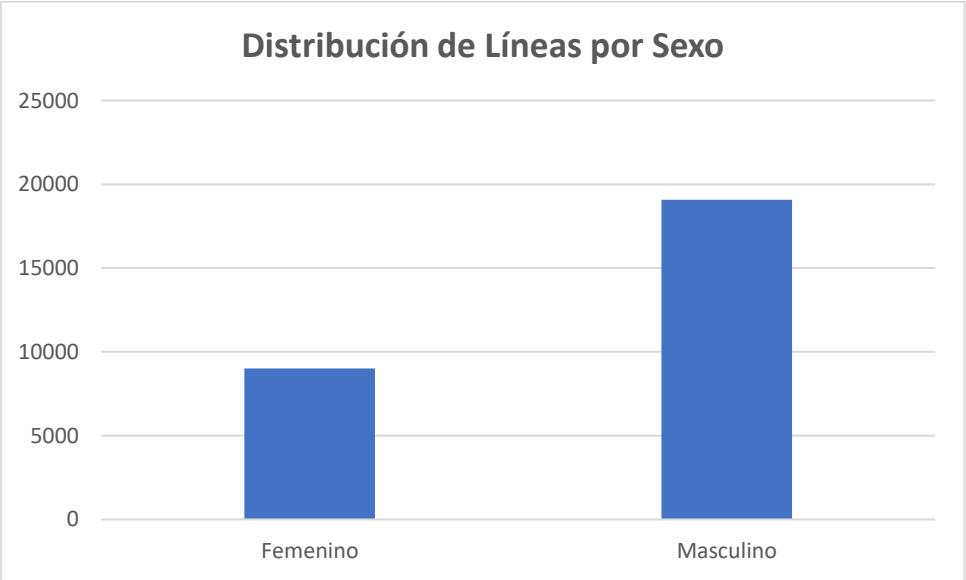
- ✓ **Sexo**

Esta variable, es de tipo demográfica y representa el sexo del cliente. Es decir que puede ser femenino o masculino. A continuación, se muestra la tabla de frecuencia de los valores que puede tomar esta variable:

Tabla 22: Valores de la Variable Sexo

Sexo	Cantidad de Líneas
Femenino	9016
Masculino	19076
TOTAL	**28092**

Figura 27: Frecuencia de la Variable Sexo



Fuente: Elaboración Propia

Como se puede observar en el gráfico, 19076 de los registros corresponde a clientes de género masculino, es decir que representan el 67,91% del total.

- ✓ **City_Name**

Esta variable, es de tipo demográfica y representa la ciudad en la que se registró el cliente. Esta variable puede tomar los siguientes valores:

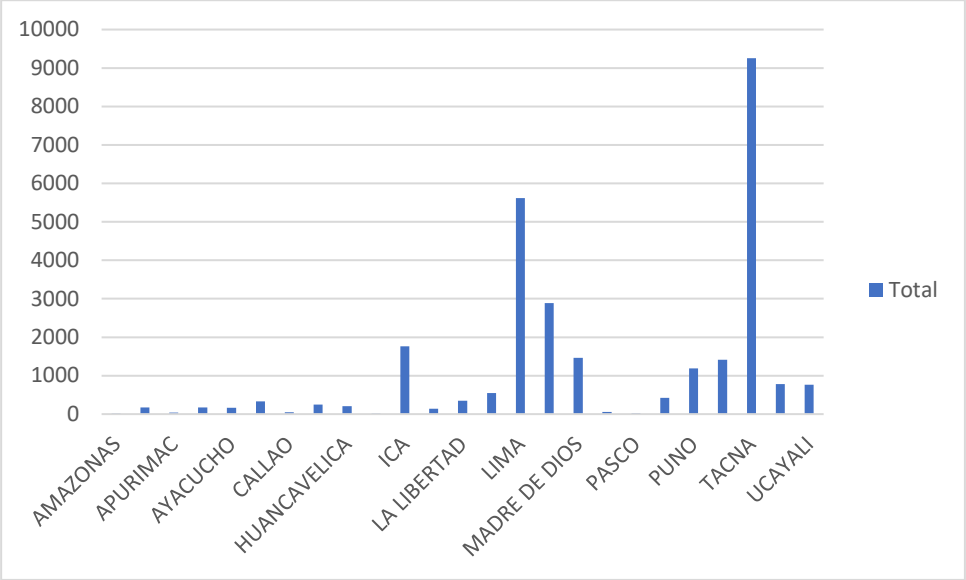
Tabla 23: Valores de la Variable City_Name

Ciudad	Cantidad de Líneas
AMAZONAS	16
ANCASH	171
APURIMAC	39
AREQUIPA	171
AYACUCHO	162
CAJAMARCA	333
CALLAO	48
CUSCO	246
HUANCAVELICA	207
HUANUCO	18
ICA	1764
JUNIN	138
LA LIBERTAD	348
LAMBAYEQUE	546
LIMA	5618
LORETO	2883
MADRE DE DIOS	1467
MOQUEGUA	60
PASCO	27
PIURA	423
PUNO	1185
SAN MARTIN	1413
TACNA	9258
TUMBES	783
UCAYALI	768
Total	**28092**

Fuente: Elaboración Propia

Como podemos apreciar, la ciudad con mayor cantidad de líneas telefónicas asignadas es Tacna que tiene 9258 registros, seguida de Lima con 5618. A diferencia de otros operadores de telefonía móvil, el grueso de los clientes de Bitel se encuentra fuera de la capital.

Figura 28: Distribución Total de Líneas Telefonicas por Ciudad

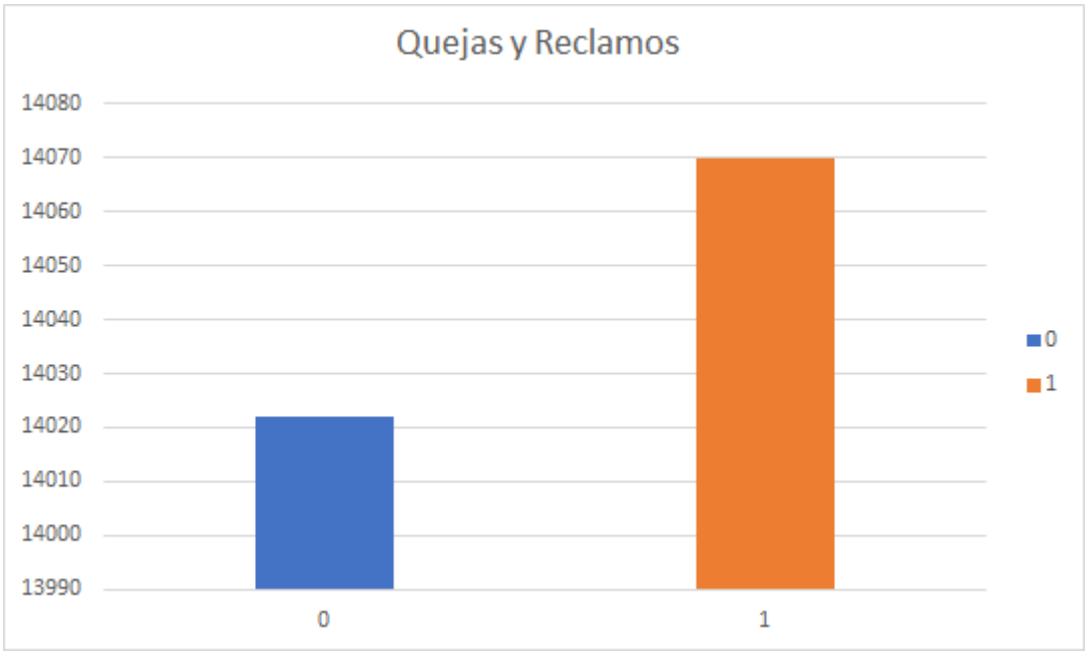


Fuente: Elaboración Propia

- ✓ **Quejas y Reclamos**

Con respecto a la variable dicotómica cualitativa "quejas y reclamos", el valor 0 representa que el cliente no tuvo quejas y 1 que si realiza una o más quejas. Como se aprecia en la figura siguiente, hay prácticamente una proporción de 1:1 entre las personas que tuvieron quejas y las que no.

Figura 29: Frecuencia de Quejas y Reclamos

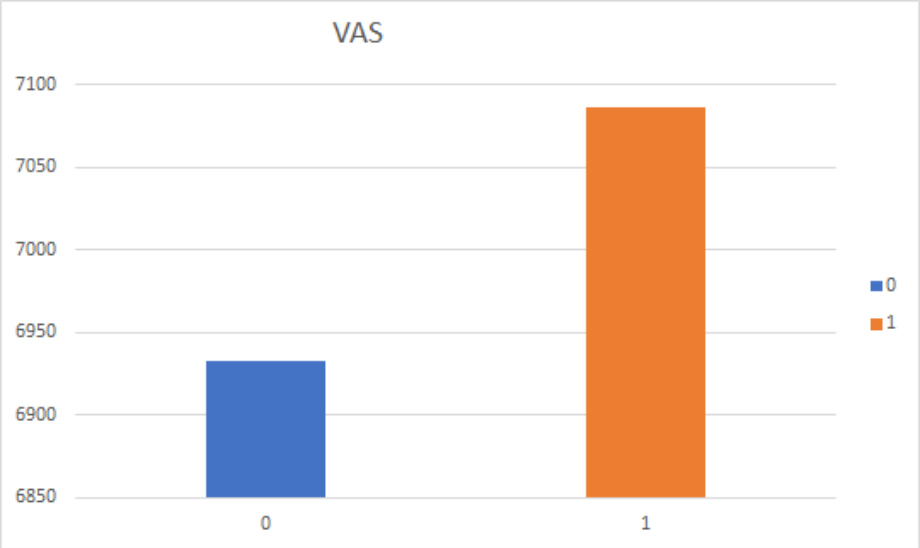


Fuente: Elaboración Propia

- ✓ **Vas (Servicio de Valor Añadido)**

Con respecto a la variable dicotómica cualitativa "Vas", vemos que por un pequeño margen, hay más gente que ha consumido servicios de valor añadido que gente que no.

Figura 30: Frecuencia de la variable Vas



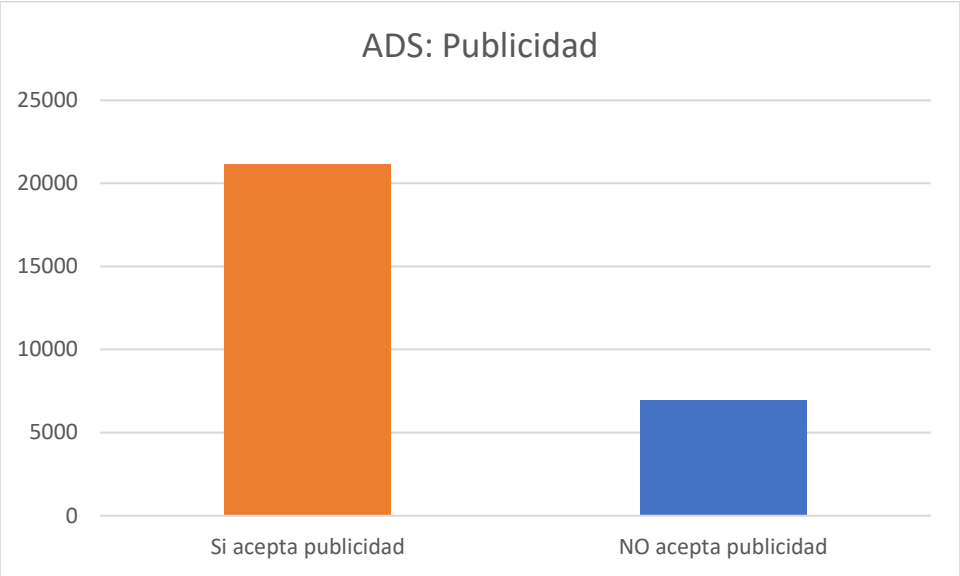
Fuente: Elaboración Propia

- ✓ **ADS (Acepta publicidad)**

En la investigación de Rivero, Henrique y Serra (2019), analizan la actitud hacia la publicidad e implicaciones en la intención de compra, y se va determinar variables que influyen en la portabilidad numérica de las líneas móviles de las principales empresas de telefonía móvil en el Perú. Entre estas variables se propone la publicidad como crítica ya que el acceso y atención a esta, incrementa la probabilidad de que el cliente conozca lo beneficios que ofrecen los proveedores competidores y piense en realizar la portabilidad numérica.

Esta será considerada una dummy que identifica si el usuario acepta publicidad durante su navegación. 1: si acepta publicidad o 0: no acepta publicidad.

Figura 31:Frecuencia de la variable ADS



Fuente: Elaboración propia

- ✓ **Portabilidad (Clientes que migraron anteriormente)**

Según el Relanzamiento de la Portabilidad Numérica Móvil, emitido por el Organismo Supervisor de Inversión Privada en Telecomunicaciones - OSIPTEL (2018), detalla evidencia teórica y empírica para explicar la importancia de la portabilidad numérica móvil, ya que implica reducir costos a los consumidores, aprovechando los beneficios del nuevo proveedor, lo que contribuiría a incrementar la competencia entre empresas operadoras en el mercado ya que se ven obligadas a mejorar las condiciones que les brinda a cada cliente, para retenerlo o atraerlo desde la competencia.

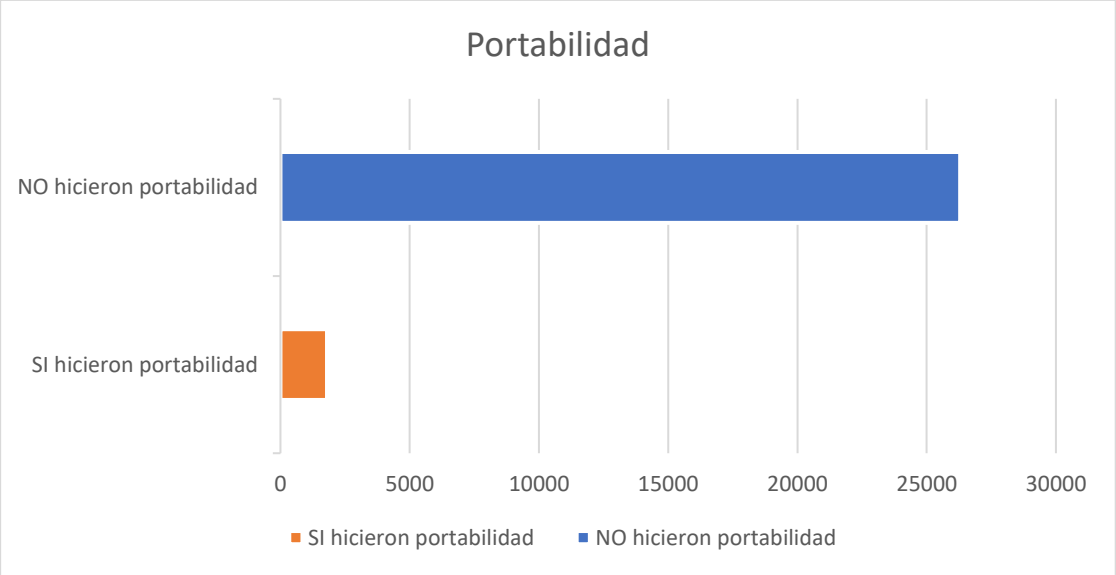
La Encuesta Residencial de Servicios de Telecomunicaciones ERESTEL (2016) por encargo de OSIPTEL, detalla puntos a tomar en cuenta para hallar clientes con interés de realizar portabilidad numérica.

- Conocimiento sobre el beneficio de portabilidad numérica
- Dificultad del proceso de portabilidad
- Ha realizado anteriormente, un proceso de portabilidad

La portabilidad será considerada una dummy que identifica si el interés del consumidor califica como interés presente o interés ausente por realizar portabilidad numérica en corto plazo mediante su historial de portabilidades. 1: cliente migró anteriormente y 0: cliente no migró anteriormente.

Solo un 7% de los clientes del estudio fueron provenientes por portabilidad de otras empresas, competidores: Claro, Movistar, Entel. Este indicador refleja la poca captación de nuevos clientes que prefieren contratar Bitel, a costo de oportunidad dejar su proveedor actual.

Figura 32:Frecuencia de la variable Portabilidad



Fuente: Elaboración propia

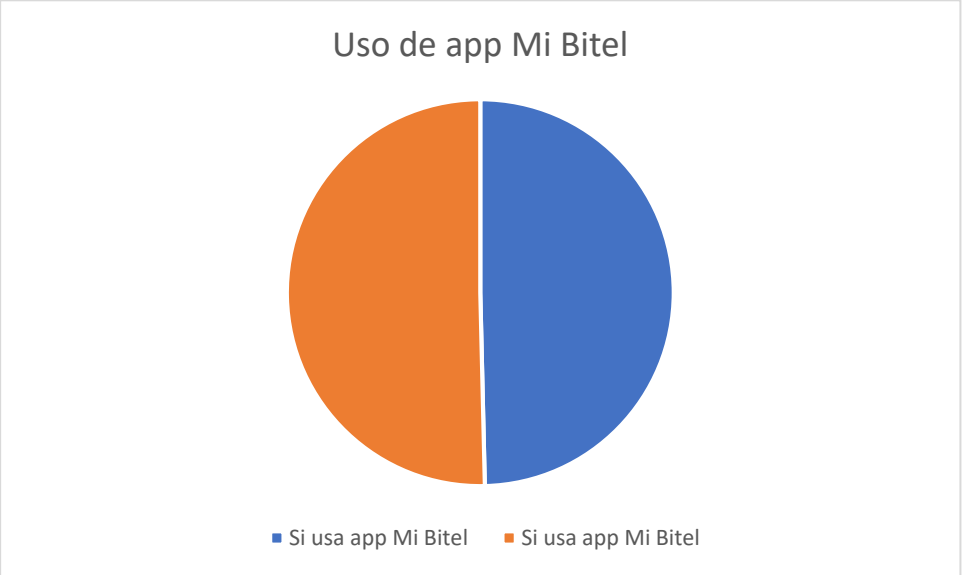
- ✓ **Mi Bitel (Uso de la app)**

Bitel Perú ofrece a sus clientes, el acceso a su app Mi Bitel, a través de la cual se puede gestionar tus servicios Bitel, la descarga y uso de la App Mi Bitel se puede tomar de manera gratuita. Mi Bitel ofrece funcionalidades como: pago de servicios, visualización de recibos, recargas Bitel, consulta de saldo, descuentos, etc.

La disposición de este beneficio es considerada en la dimensión de la calidad de servicio. Según Vavra (2016) de acuerdo a la ISO 9001:2000 la calidad de servicio debe contener: fiabilidad, capacidad de respuesta, empatía, seguridad y tangibilidad. Estas características serán las que determinen el uso de la app.

El uso de la app Mi Bitel será considerada una dummy que identifica si el usuario hace uso de la app o no. Siendo 1: usa la app Mi Bitel y 0: no usa la app Mi Bitel. Un 50% de los clientes de Bitel, hace uso de la app Mi Bitel. Esto refleja la baja fidelización que tienen actualmente los clientes, las causas pueden ser diversas: la app no implica un valor añadido, no existen incentivos, no es una app amigable, etc.

Figura 33: Frecuencia de la variable Mi Bitel



Fuente: Elaboración propia

5.1.2.2.2. Análisis de multicolinealidad

Con el fin de hallar si las variables tienen una relación directa y fuerte entre ellas se realizará el análisis de la multicolinealidad usando el factor de inflación de la varianza con el fin de evaluar esta relación. Se subió la base de datos a una carpeta con el objetivo de realizar un pre procesamiento más profundo. Al estar ya subida la base pre procesada, se ejecutó el software Anaconda y mediante la herramienta júpiter notebooks. Se importaron las principales librerías y módulos necesarios para aplicar el procesamiento de los datos y el modelado de la técnica de machine learning escogida (Figura 31). A partir de estas importaciones, se modeló el archivo csv a un dataframe de la librería pandas (Figura 32).

Figura 34: Importación de librerías

```
import pandas as pd
import numpu as np
import statsmodels
import sklearn
```

Fuente: Elaboración propia

Figura 35: Importación de archivo csv a dataframe

```
df = pd.read_csv('base (1).csv',sep=';')
df.head(5)
```

	PRODUCT_TYPE	ADS	SEX	CITY_NAME	DEVICE_TYPE	FLAG_PORTABILIDAD	VAS	MIBITEL	SMS_1	DATA_1	VOICE_1	Tiempo De Permanencia	Quejas o Reclamos	Edad	CHURN
0	PREPAID	0	M	AYACUCHO	B mobile	1	1	1	6	4371	133,0	8	0	23	0
1	PREPAID	0	F	ICA	B mobile	0	0	0	8	4430	116,0	11	0	28	1
2	PREPAID	0	F	ICA	B mobile	0	0	1	6	7543	146,0	57	0	45	1
3	PREPAID	1	F	ICA	B mobile	0	1	0	8	1553	55,0	9	0	21	1
4	PREPAID	1	F	ICA	B mobile	0	0	0	8	8207	122,0	70	0	27	1

Fuente: Elaboración propia

Después de esto, se necesita convertir las variables cualitativas nominales a cuantitativas ya que para hallar el factor inflatorio de varianza es necesario que todas la variables sean numericas. Para lograr esto se utilizó el modulo “LabelEncoder” de la libreria sklearn y corremos su funcion en las variables cualitativas “Product_Type”,”Sex”,”City_Name” y “Device Type”. Tal como se muestra en la imagen a continuacion.

Figura 36: Conversión de variables cualitativas a numéricas

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
df['PRODUCT_TYPE'] = label_encoder.fit_transform(df['PRODUCT_TYPE'])
df['SEX'] = label_encoder.fit_transform(df['SEX'])
df['CITY_NAME'] = label_encoder.fit_transform(df['CITY_NAME'])
df['DEVICE_TYPE'] = label_encoder.fit_transform(df['DEVICE_TYPE'])
```

Fuente: Elaboración propia

Figura 37: Dataframe con valores numéricos

	PRODUCT_TYPE	ADS	SEX	CITY_NAME	DEVICE_TYPE	FLAG_PORTABILIDAD	VAS	MIBITEL	SMS_1	DATA_1	VOICE_1	Tiempo De Permanencia	Quejas o Reclamos	Edad	CHURN
0	1	0	1	4	0	1	1	1	6	4371	133.0	8	0	23	0
1	1	0	0	10	0	0	0	0	8	4430	116.0	11	0	28	1
2	1	0	0	10	0	0	0	1	6	7543	146.0	57	0	45	1
3	1	1	0	10	0	0	1	0	8	1553	55.0	9	0	21	1
4	1	1	0	10	0	0	0	0	8	8207	122.0	70	0	27	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
28087	0	1	1	12	7	0	0	0	0	0	0.1	74	0	22	1
28088	1	1	1	23	7	0	1	1	0	0	0.0	10	0	43	1
28089	1	1	1	23	7	0	0	0	0	0	0.0	37	0	43	1
28090	1	1	1	23	7	0	1	1	0	0	0.0	53	0	24	1
28091	1	1	1	23	7	0	0	1	0	0	0.0	28	1	24	1

Fuente: Elaboración propia

Por último teniendo ya solo valores numéricos, se corre la función variance_inflation_factor de la librería “statsmodels”.

Figura 38: Código para encontrar el FIV de las variables

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif_scores = pd.DataFrame()
vif_scores["Attribute"] = X.columns
vif_scores["VIF Scores"] = [variance_inflation_factor(X.values, i) for i in range(len(X.columns))]

display(vif_scores)
```

Fuente: Elaboración propia

Figura 39: FIV de los atributos

	Attribute	VIF Scores
0	PRODUCT_TYPE	8.678627
1	ADS	4.226866
2	SEX	2.994329
3	CITY_NAME	11.433260
4	DEVICE_TYPE	5.060691
5	FLAG_PORTABILIDAD	1.243385
6	VAS	1.966286
7	MIBITEL	1.963575
8	SMS_1	1.123402
9	DATA_1	2.539104
10	VOICE_1	1.791161
11	Tiempo De Permanencia	3.731792
12	Quejas o Reclamos	1.964865
13	Edad	12.881480

Fuente: Elaboración propia

De acuerdo a Ringle, C. ,Anderson, R. , Tatham, R & Black, W. (2015), el límite máximo del FIV (Factor de inflación de la varianza) es de 5, en base a esto se procede a eliminar las variables “Edad”, “City_Name” y “Product_Type” ya que presentan una relación fuerte con las demás variables independientes. Asimismo, anteriormente se mencionó que la variable edad es importante para predecir la fuga de cliente, porque el área de marketing necesita conocer a su público objetivo para sus campañas de retención.

Figura 40: Eliminación de variables

```
3 df = df.drop(columns=['Edad','CITY_NAME','PRODUCT_TYPE'])
4 df
5
```

Fuente: Elaboración Propia

5.1.2.2.3. Normalización

Con el objetivo de obtener mejores resultados, se normalizan los parámetros cuantitativos, es decir, las columnas “SMS_1”, “DATA_1”,”VOICE_1” y “Tiempo De Permanencia” de la figura 38. Para lo cual se creó un sub-dataframe df_norm que contuviera sólo estos parámetros mostrando en la Figura 39, se normalizaron (Figura 40) y a este mismo dataframe se le combinó los parámetros cualitativos del primer dataframe, como se aprecia en la Figura 42.

Figura 41: Selección de parámetros cuantitativos

```
df_num = df[['SMS_1','DATA_1','VOICE_1','Tiempo De Permanencia']]
df_num
```

Fuente: Elaboración propia

Figura 42: Sub-Dataframe con parámetros cuantitativos

	SMS_1	DATA_1	VOICE_1	Tiempo De Permanencia
0	6	4371	133.0	8
1	8	4430	116.0	11
2	6	7543	146.0	57
3	8	1553	55.0	9
4	8	8207	122.0	70
...	...	...	...	...
28087	0	0	0.1	74
28088	0	0	0.0	10
28089	0	0	0.0	37
28090	0	0	0.0	53
28091	0	0	0.0	28

Fuente: Elaboración propia

Figura 43: Proceso de normalización

```
df_norm = (df_num - df_num.min()) / (df_num.max() - df_num.min())
df_norm
```

Fuente: Elaboración propia

Figura 44: Unión de parámetros cualitativos

```
df_norm['ADS'] = df['ADS']
df_norm['SEX'] = df['SEX']
df_norm['DEVICE_TYPE'] = df['DEVICE_TYPE']
df_norm['FLAG_PORTABILIDAD'] = df['FLAG_PORTABILIDAD']
df_norm['VAS'] = df['VAS']
df_norm['MIBITEL'] = df['MIBITEL']
df_norm['Quejas o Reclamos'] = df['Quejas o Reclamos']
df_norm['CHURN'] = df['CHURN']
df_norm

```

Fuente: Elaboración propia

Figura 45: Dataframe con parámetros cuantitativos normalizados y parámetros cualitativos

	SMS_1	DATA_1	VOICE_1	Tiempo De Permanencia	ADS	SEX	DEVICE_TYPE	FLAG_PORTABILIDAD	VAS	MIBITEL	Quejas o Reclamos	CHURN
0	0.004441	0.050350	0.034939	0.106667	0	M	B mobile	1	1	1	0	0
1	0.005922	0.051030	0.030473	0.146667	0	F	B mobile	0	0	0	0	1
2	0.004441	0.086889	0.038354	0.760000	0	F	B mobile	0	0	1	0	1
3	0.005922	0.017889	0.014449	0.120000	1	F	B mobile	0	1	0	0	1
4	0.005922	0.094538	0.032049	0.933333	1	F	B mobile	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
28087	0.000000	0.000000	0.000026	0.986667	1	M	Otros	0	0	0	0	1
28088	0.000000	0.000000	0.000000	0.133333	1	M	Otros	0	1	1	0	1
28089	0.000000	0.000000	0.000000	0.493333	1	M	Otros	0	0	0	0	1
28090	0.000000	0.000000	0.000000	0.706667	1	M	Otros	0	1	1	0	1
28091	0.000000	0.000000	0.000000	0.373333	1	M	Otros	0	0	1	1	1

Fuente: Elaboración propia

Habiendo realizado el pre procesamiento de los datos cuantitativos, procedimos a la realización de las variables dummy de los parámetros cualitativos, ya que, la técnica de regresión logística no permite trabajar con valores no numéricos. Creamos una lista con los nombres de las variables cualitativas (Figura 43) y utilizamos la función de pandas "get_dummies" para obtener las variables dummies, pasando nuestro data frame de 12 columnas a 46 (Figura 44 y Figura 45).

Figura 46: Lista con los parámetros cualitativos

```
var_cat = ['Quejas o Reclamos','ADS','SEX','DEVICE_TYPE','FLAG_PORTABILIDAD','VAS','MIBITEL']
```

Fuente: Elaboración propia

Figura 47: Creación de variables Dummies

```
df_f =  pd.get_dummies(df_norm, columns=var_cat)
df_f
```

Fuente: Elaboración propia

Figura 48: Dataframe con variables cuantitativas normalizadas y variables cualitativas en forma dummy

	SMS_1	DATA_1	VOICE_1	Tiempo De Permanencia	CHURN	Quejas o Reclamos_0	Quejas o Reclamos_1	ADS_0	ADS_1	SEX_F	...	DEVICE_TYPE_Redmi	DEVICE_TYPE_SKY DEVICES	DEVICE_T
0	0.004441	0.050350	0.034939	0.106667	0	1	0	1	0	0	...	0	0	
1	0.005922	0.051030	0.030473	0.146667	1	1	0	1	0	1	...	0	0	
2	0.004441	0.086889	0.038354	0.760000	1	1	0	1	0	1	...	0	0	
3	0.005922	0.017889	0.014449	0.120000	1	1	0	0	1	1	...	0	0	
4	0.005922	0.094538	0.032049	0.933333	1	1	0	0	1	1	...	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
87	0.000000	0.000000	0.000026	0.986667	1	1	0	0	1	0	...	0	0	
88	0.000000	0.000000	0.000000	0.133333	1	1	0	0	1	0	...	0	0	
89	0.000000	0.000000	0.000000	0.493333	1	1	0	0	1	0	...	0	0	
90	0.000000	0.000000	0.000000	0.706667	1	1	0	0	1	0	...	0	0	
91	0.000000	0.000000	0.000000	0.373333	1	0	1	0	1	0	...	0	0	

Fuente: Elaboración propia

4.1.1.1. **Modelado**

A partir del dataframe ya pre procesado, se realizó la separación en dos dataframe “X”, que contenga las variables independientes e “y”, que contiene la variable dependiente Churn, tal como se muestra en la figura 46. Luego ambas fueron divididas en una proporción de 1:5 entre los datasets de entrenamiento y de test.

Figura 49: Separación de dataframe en train y test

```
X = df_f.drop(['CHURN'], axis=1)
y = df_f['CHURN']

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2)
```

Fuente: Elaboración propia

Como se muestra en la figura 47, a partir del dataset de train se realizó el aplicó la técnica de machine learning de regresión logística, con la semilla 0 con el objetivo de asegurar la replicabilidad del código.

Figura 50: Aplicación de la técnica de regresión logística

```
:  1 modelo1 = LogisticRegression(random_state=0)
   2 modelo1.fit(X_train,y_train)
   3 res = modelo1.predict(X_test)
```

Fuente: Elaboración propia.

4.2. Medición de la solución

4.2.1. Análisis de Indicadores cuantitativo y/o cualitativo.

Al realizar la comparación entre el modelo realizado a partir del dataset de entrenamiento y el dataset de test validamos que la métrica de accuracy score es de 0.8799, es decir que el valido correctamente un 88% de los registros del dataset de test, tal como se muestra en la figura 48.

Figura 51: Accuracy score del modelo

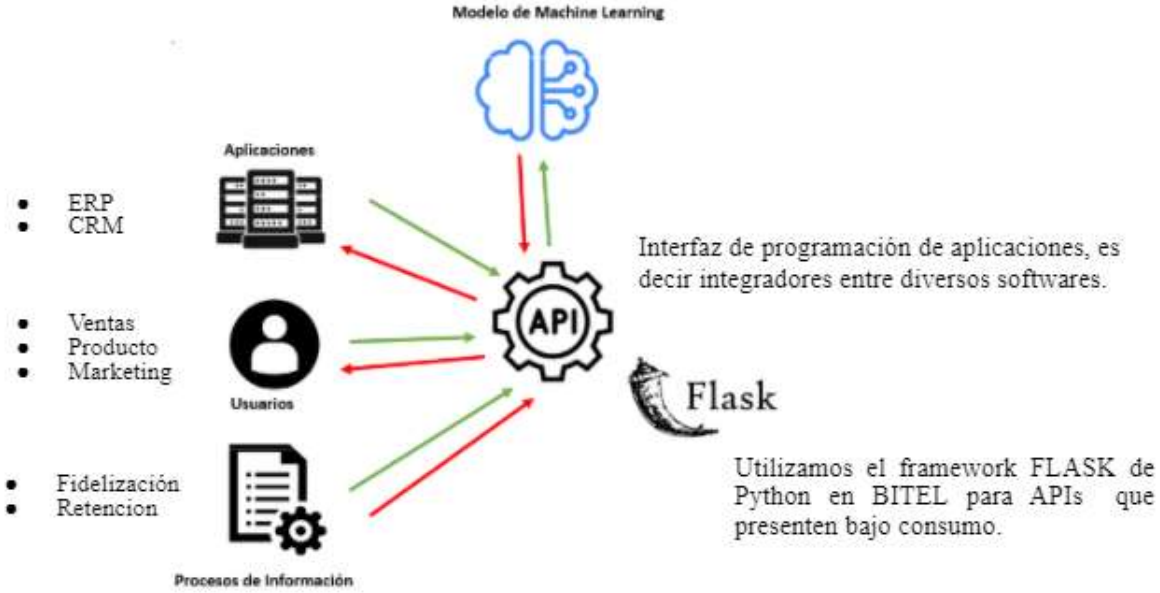
```
1 accuracy_score(y_test,res)
0.8798718633208756
```

Fuente: Elaboración propia

5.1.3. Simulación de solución. Aplicación de Software

Con el objetivo que el modelo pueda ser utilizado por múltiples plataformas y ya que se desea que el modelo funcione tanto para procesos de información internos como por usuarios específicos, se recomienda la implementación del modelo a través de una rest api utilizando el framework FLASK de python. Tal como se muestra en la siguiente figura, el procedimiento que seguiría el proceso sería que las aplicaciones, procesos o usuarios que quieran predecir el churn de una línea o un grupo de líneas hagan una llamada a la api y la api recoja su archivo en formato csv, después que la api prediga el churn de las líneas en base al modelo trabajado y por último, la api devuelve la información con las predicciones del modelo.

Figura 52: Simulación de solución



Fuente: Elaboración propia

Capítulo VI: Conclusiones y Recomendaciones

6.1. Conclusiones

Las empresas de telecomunicaciones, las cuales tienen modelos de negocio basados en suscripción tienen como principal reto mantener su base de clientes además de captar nuevos, lo cual representa un gran desafío, teniendo en cuenta que los clientes cada vez son más exigentes y tienen mayor acceso a la información sobre sus derechos como usuarios, los procesos de atención de quejas y/o reclamos, las entidades a las cuales pueden acudir en segunda instancia, los costos en el mercado de productos sustitutos, etc.

En el presente trabajo se ha desarrollado de manera teórica y práctica la implementación de un modelo de regresión logística con el objetivo de predecir la fuga de clientes de la empresa Bitel Perú. Se obtuvo de manera exitosa un modelo que integraba variables cualitativas como son las de consumo o transaccionales, así como aquellas de tipo demográficas y las relacionadas con el comportamiento del cliente.

Se integraron distintas bases de datos, una base de datos que contenía información del consumo de una muestra de los clientes de Bitel como consumo de datos en megabytes, minutos y mensajes (sms), otra base de datos con sexo, si usan la aplicación de mibitel, etc. A partir de esta base de datos se obtuvo una muestra de 16512 registros. Luego realizando la exploración de la data encontramos que dos parámetros (Sexo y Device Type) presentaban valores nulos, un 47% de todos los registros, luego de la eliminación de estos registros obtuvimos 8572 registros. Después de estos, se subió el archivo csv en el software jupiter notebook a través de la librería pandas de python, donde se eliminó las variables que presentaban una fuerte relación con otras variables independientes, luego se se normalizo los parámetros cuantitativos, además, se crearon dummies variables para los variables cuantitativas con el objetivo de poder entrenar el modelo con todas las variables disponibles. Luego de realizar el preprocesamiento de la data, se dividió entre datasets de entrenamiento y de test en una proporción de 4:1. Luego de entrenar la data con la semilla 0 y validar el modelo con el dataset de test, obtuvimos un accuracy score de 88%, lo que significa que el modelo predijo correctamente un 88% de los registros del dataset de test.

Se espera que, a partir de este trabajo y modelo entrenado, la empresa Bitel pueda predecir los clientes con potencial de fuga y tomar acciones y estrategias para disminuir este potencialmente.

6.2. Recomendaciones

Existen variables y datos nuevos que sería importante considerar e incorporar en un trabajo futuro para mayor nivel de detalle; como es la información relacionada con la razón por la que el cliente cancela el contrato: calidad del servicio, costo del servicio, etc. que, si bien no han sido incluidas en estudios previos, podrían integrarse en el modelo.

Existen otros métodos estadísticos complementarios empleados para dar respuesta al problema de la fuga de clientes o churn como el análisis de supervivencia, que permiten estudiar la ocurrencia y el momento de los eventos. Este método permite hacer un seguimiento de la pérdida de clientes de manera efectiva durante largos periodos de tiempo y optimiza los parámetros que son difíciles de manejar con los métodos estadísticos convencionales; además de ofrecer gráficos y resultados que van más allá de la estimación de la probabilidad de fuga. El análisis de supervivencia permite segmentar poblaciones en base a la duración hasta la ocurrencia de un evento.

En este estudio no distingue a los clientes individuales (personas con DNI) de los corporativos (empresas o personas con RUC). Sin embargo, es deseable investigar la pérdida de clientes corporativos por separado de la pérdida de clientes individuales en el futuro. Sin embargo si se ha eliminado de la muestra a las líneas otorgadas a los empleados, puesto que la cancelación de estas líneas no constituye una fuga de clientes como tal.

Referencia Bibliográfica

- Ana G. y Fernando A. (2017) Machine Learning en la Industria: El caso de la Siderurgia, 55-63. Recuperado de: https://dialnet.unirioja.es/servlet/articulo?codigo=6207513
- Arce Chíncaro, E., & Mejía Puente, M. (2011). Application of a credit evaluation model in order to reduce the risk of the clients' portfolio of insurance companies.
- B. Huang, T. Kechadi, B. Buckley, G. Kiernan, E. Keogh, T.Rashid (2010), "A new feature set with new windows techniques for customer churn prediction in land-line telecommunications", Expert Systems with Applications, 37, pp. 3657-3665.
- Barrientos, F; Ríos, S. (2013). Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones. Revista Ingeniería de Sistemas. Volumen XXVII.
- Beltran, B. (2019). Predicción de fuga de clientes en empresas de telefonía móvil: El caso de estudio de virgin mobile [Tesis de magíster, Universidad del desarrollo]. Repositorio los libertadores. Recuperado de: https://repository.libertadores.edu.co/bitstream/handle/11371/2077/Baena_Camilo_2018.pdf?sequence=1&isAllowed=y
- Chitarroni, H. (2002, diciembre). La regresión logística. Recuperado de: https://racimo.usal.edu.ar/83/1/Chitarroni17.pdf
- Ditterrich, T. G. (1997). Machine learning research: four current direction. Artificial intelligence magzine, 4, 97-136.
- F. Barrientos (2011). Diseño e implementación de una metodología de predicción de fuga de clientes en una compañía de telecomunicaciones.
- Harrington P. (2012). Machine Learning in action. Manning Publications Co.www.wowebook.com
- Hashmi, N., Butt, N. A., & Iqbal, M. (2013). Customer churn prediction in telecommunication a decade review and classification. International Journal of Computer Science Issues (IJCSI), 10(5), 271.

- Informe Técnico Trimestral de Estadísticas de Uso de las Tecnologías de la Información y Comunicación de 2019 elaborado por el Instituto Nacional de Estadística e Informática (INEI).

- Jain H., Khunteta A. y Srivastava S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. Procedia Computer Science, 167, 101-112. Recuperado de: https://www.sciencedirect.com/science/article/pii/S1877050920306529

- Junxiang, Lu (2002). Predicting customer churn in the telecommunications industry — an application of survival analysis modeling using sas. SAS User Group International (SUGI27) Online Proceedings, pages 114–27. Recuperado de: https://support.sas.com/resources/papers/proceedings/proceedings/sugi27/p114-27.pdf

- Méndez-Gurrola, I. I. (2020). Aprendizaje automático aplicado en física: Una revisión de la literatura científica. Instituto de Arquitectura Diseño y Arte.

- Pérez, V. P. 2014. Modelo de predicción de Fuga de clientes de telefonía Móvil Post pago. Memoria para Optar al Título de Ingeniero Civil Industrial. Departamento de Ingeniería Industrial. Universidad de Chile.

- Pocohuanca. (2018). Determinantes del crecimiento de la portabilidad numérica de líneas Móviles de Entel Perú y Movistar periodo: julio 2014 – diciembre 2017.

- Quiza, J. (2018). Regresión Logística con Pytorch. Recuperado de: https://medium.com/datos-y-ciencia/regresi%C3%B3n-log%C3%ADstica-con-pytorch-d2212b912b6b

- Revista Mediterránea de Comunicación. 2019, 10(1): 203-213.

- Ringle, C. M., Wende, S., & Becker, J. M. (2015). SmartPLS 3. Bönningstedt: SmartPLS.

- Rodríguez, M. (2020) Análisis de predicción de Churn en una empresa de Telecomunicaciones de España. Recuperado de: https://eprints.ucm.es/id/eprint/61981/1/TFM-Miner%C3%ADa2020%20MICHEL%20RODRIGUEZ.pdf

- Rosas Díaz, K. M. (2021). Sistema e-commerce con el uso de drones para la distribución de productos de la empresa Victoria, Los Olivos.

- Ruiz, G. (2019). Modelo de análisis de datos utilizando técnicas de aprendizaje supervisado y no supervisado, para identificar patrones en la información generada por los pacientes, sometidos a juegos diseñados como un instrumento de apoyo terapéutico [Tesis Magister, Universidad Jorge Tadeo Lozano]. Repositorio institucional de Universidad Jorge Tadeo Lozano. Recuperado de: https://expeditiorepositorio.utadeo.edu.co/bitstream/handle/20.500.12010/8502/Trabajo%20de%20grado.pdf?sequence=1&isAllowed=y

- Vavra, F (2016). Como medir la satisfacción del cliente según la ISO 9001:2000. España, Madrid: Fundación Confemetal.

Anexos

Anexo 1: Matriz FODA cuantitativo 1

MATRIZ FODA CUANTITATIVO - CARLA BERNACHEA											
		OPORTUNIDADES					AMENAZAS				
		O1	O2	O3	O4	PROMEDIO	A1	A2	A3	A4	PROMEDIO
FORTALEZAS	F1	7	1	5	6	4.75	5	1	2	6	3.5
	F2	5	7	6	3	5.25	6	1	5	5	4.25
	F3	7	3	7	5	5.5	7	1	7	7	5.5
	F4	5	1	6	4	4	6	1	1	7	3.75
	PROMEDIO	6	3	6	4.5		6	1	3.75	6.25	
DEBILIDADES	D1	7	5	1	5	4.5	7	1	3	6	4.25
	D2	4	2	5	2	3.25	7	1	1	7	4
	D3	7	7	2	6	5.5	4	1	1	6	3
	D4	7	4	4	6	5.25	7	3	2	6	4.5
	PROMEDIO	6.25	4.5	3	4.75		6.25	1.5	1.75	6.25	

Anexo 2: Matriz FODA cuantitativo 2

MATRIZ FODA CUANTITATIVO - EDWARD CHILET											
		OPORTUNIDADES					AMENAZAS				
		O1	O2	O3	O4	PROMEDIO	A1	A2	A3	A4	PROMEDIO
FORTALEZAS	F1	7	4	6	5	5.5	6	1	2	3	3
	F2	7	6	7	7	7	5	6	2	1	4.5
	F3	7	6	4	3	6.75	7	3	6	5	5.25
	F4	7	4	3	2	4	6	7	1	6	5
	PROMEDIO	7	5	5	4.5						
DEBILIDADES	D1	7	4	2	3	4	7	3	2	4	4
	D2	4	2	5	2	3.25	6	1	1	7	3.75
	D3	7	6	1	4	4.5	6	2	2	3	3.25
	D4	7	3	5	6	5.25	7	5	2	4	4.5
	PROMEDIO	6.25	3.75	3.25	4.75		6.5	2.75	1.75	4.5	

Anexo 3: Matriz FODA cuantitativo 3

MATRIZ FODA CUANTITATIVO - PAOLA GUZMAN											
		OPORTUNIDADES					AMENAZAS				
		O1	O2	O3	O4	PROMEDIO	A1	A2	A3	A4	PROMEDIO
FORTALEZAS	F1	1	3	5	7	4	4	1	5	6	4
	F2	2	7	3	6	4.5	3	2	4	5	3.5
	F3	3	4	6	2	4.25	2	3	6	1	3
	F4	5	2	4	1	3	5	4	3	7	4.75
	PROMEDIO	2.75	4	4.5	4						
DEBILIDADES	D1	7	3	5	6	5.25	7	2	5	6	5
	D2	6	1	4	7	4.5	5	3	4	7	4.75
	D3	5	2	3	4	3.5	6	4	3	5	4.25
	D4	4	6	7	3	4.5	4	1	6	3	3.5
	PROMEDIO	5.5	3	4.75	5		5.5	2.5	4.5	5.25	

Anexo 4: Matriz FODA cuantitativo 4

MATRIZ FODA CUANTITATIVO - VICTOR INCHE											
		OPORTUNIDADES					AMENAZAS				
		O1	O2	O3	O4	PROMEDIO	A1	A2	A3	A4	PROMEDIO
FORTALEZAS	F1	5	2	7	6	5	5	5	1	5	4
	F2	5	5	5	6	5.25	4	2	1	2	2.25
	F3	2	1	7	7	4.25	5	4	2	7	4.5
	F4	5	4	5	5	4.75	6	5	2	6	4.75
	PROMEDIO	4.25	4	5	5		5	4	1.5	5	
DEBILIDADES	D1	7	7	1	1	4	7	1	1	1	2.5
	D2	6	2	1	1	2.5	7	1	1	6	3.75
	D3	7	7	6	1	5.25	5	1	1	5	3
	D4	7	7	5	4	5.75	4	1	1	3	2.25
	PROMEDIO	6.75	5.75	3.25	1.75		5.75	1	1	3.75	

Anexo 5: Matriz FODA cuantitativo 5

MATRIZ FODA CUANTITATIVO - MAYRA LEON											
		OPORTUNIDADES					AMENAZAS				
		O1	O2	O3	O4	PROMEDIO	A1	A2	A3	A4	PROMEDIO
FORTALEZAS	F1	6	2	5	5	4.5	7	1	2	6	4
	F2	5	7	6	3	5.25	5	4	2	6	4.25
	F3	2	3	6	4	3.75	6	1	2	5	3.5
	F4	7	3	4	4	4.5	7	6	2	7	5.5
	PROMEDIO	5	3.75	5.25	4		6.25	3	2	6	
DEBILIDADES	D1	7	4	1	5	4.25	7	1	2	4	3.5
	D2	6	2	3	6	4.25	6	2	2	7	4.25
	D3	7	3	7	7	6	5	5	1	6	4.25
	D4	7	6	7	6	6.5	7	1	1	6	3.75
	PROMEDIO	6.75	3.75	4.5	6		6.25	2.25	1.5	5.75	