Modelo de fidelización para reducir la cancelación de los servicios que ofrece una empresa de Telecomunicaciones

Montenegro Fierro, Wilmer Juan

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Gestión en Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación, previo a la obtención del título de Magíster en Gestión en Sistemas de Información e Inteligencia de Negocios

Msc. Duque Cruz, Lorena Geselle

20 de noviembre del 2020

URKUND

Document Information

Analyzed document	2 TESIS_MONTENEGRO-FIERRO_WILMER-JUAN_Noviembre-2020 (urkund).docx (D85750822)
Submitted	11/18/2020 12:09:00 AM
Submitted by	Gualotuña Alvarez Tatiana Marisol
Submitter email	tmgualotunia@espe.edu.ec
Similarity	8%
Analysis address	tmgualotunia.espe@analysis.urkund.com

Sources included in the report

W	URL: https://repositorio.espe.edu.ec/bitstream/21000/18792/1/T-ESPE-038816.pdf Fetched: 7/1/2020 6:02:37 AM	8
SA	**Universidad de las Fuerzas Armadas ESPE / Solo Redacción Andrea Pinto.docx** Document Solo Redacción Andrea Pinto.docx (D78217430) Submitted by: mgalmache@espe.edu.ec Receiver: mgalmache.espe@analysis.urkund.com	5
SA	**Universidad de las Fuerzas Armadas ESPE / TESIS_versionfinal01082019.doc** Document TESIS_versionfinal01082019.doc (D54660967) Submitted by: lgduque@espe.edu.ec Receiver: lgduque.espe@analysis.urkund.com	11
SA	**tessis Iris.docx** Document tessis Iris.docx (D23242248)	4

Firma:

Firmado electrónicamente por:
LORENA GESELLE DUQUE CRUZ

Ing. Duque Cruz, Lorena Geselle Msc.

Director/a

C.C.: 1711592525

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"Modelo de fidelización para reducir la cancelación de los servicios que ofrece una empresa de Telecomunicaciones"** fue realizado por el señor **Montenegro Fierro, Wilmer Juan** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 20 de noviembre de 2020

Firma:

Firmado electrónicamente por:
LORENA GESELLE DUQUE CRUZ

.....

Ing. Duque Cruz, Lorena Geselle Msc.

Director/a

C.C.: 1711592525

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y

TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

RESPONSABILIDAD DE AUTORÍA

Yo **Montenegro Fierro, Wilmer Juan** con cédula de ciudadanía n° 1716313505, declaro que el contenido, ideas y criterios del trabajo de titulación: **"Modelo de fidelización para reducir la cancelación de los servicios que ofrece una empresa de Telecomunicaciones"** es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 20 de noviembre de 2020

Firma:

WILMER JUAN MONTENEGRO FIERRO

.....

Ing. Montenegro Fierro, Wilmer Juan

C.C.: 1716313505

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Montenegro Fierro, Wilmer Juan** autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: "**Modelo de fidelización para reducir la cancelación de los servicios que ofrece una empresa de Telecomunicaciones**" en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 20 de Junio de 2020

Firma:

WILMER JUAN MONTENEGRO FIERRO

.....

Ing. Montenegro Fierro, Wilmer Juan
C.C.: 1716313505

Dedicatoria

Dedico este gran esfuerzo a mi hijo Elian Nicolás Montenegro Paguay y a mi eterna compañera de vida Lucía Paguay Morejón, por su apoyo incondicional y paciencia. Ellos han sido, son y seguirán siendo mi inspiración, mi motivación para seguir adelante cumpliendo cada una de las metas que me proponga.

Agradecimiento

Agradezco a mis Padres Luis Eduardo Montenegro Carlosama y Martha Cecilia Fierro, quienes desde siempre me han apoyado e incentivado a seguir siempre adelante, a superarme, ellos durante toda mi vida me han dado el mejor ejemplo posible.

A mis Hermanos Guido; Edu y Kleber, por su apoyo y palabras de aliento, incentivando a no dejar que desmaye en esta meta que me propuse y que hoy es una realidad.

Índice de Contenidos

Índice de Tablas

Índice de Figuras

Resumen

En el Ecuador las empresas dedicadas a ofrecer servicios de telecomunicaciones se encuentran en innovación continua, para esto están utilizando las tendencias existentes en el mercado como mejora de servicio, satisfacción del cliente, fidelización de clientes otorgando bonos, ofertas, promociones masivas por medio de llamadas telefónicas y mensajes de texto. En la empresa de Telecomunicaciones objeto de estudio en los últimos meses se ha incrementado notablemente el porcentaje de clientes que optan por cancelar los servicios contratados y esto se traduce en una notable disminución de sus ingresos económicos.

El presente proyecto tiene como finalidad reducir el porcentaje de clientes que cancelan los servicios que ofrece la empresa de Telecomunicaciones objeto de estudio, por medio del diseño de un modelo analítico y a través del análisis de información histórica de los clientes.

Para el desarrollo del presente proyecto se utilizó la Metodología CRISP-DM (CRoss-Industry Standard Process for Data Mining) la misma que es un procedimiento Industrial Estándar para realizar Minería de Datos y permite encontrar un modelo útil y entendible que describe de forma clara y fácil los patrones encontrados.

Como resultado al concluir el presente trabajo de investigación se obtuvo un modelo predictivo que permite identificar a los clientes con alto nivel de deseo de cancelar los servicios contratados en el empresa de Telecomunicaciones objeto de estudio, a esta información se le otorga el tratamiento respectivo para reducir el porcentaje de clientes que optan por cancelar los servicios contratados, lo que consecuentemente provoca un incremento en sus ingresos.

PALABRAS CLAVE:

- **SATISFACCIÓN DEL CLIENTE**
- **CANCELACIÓN DE SERVICIOS**
- **RETENCIÓN Y FIDELIZACIÓN DE CLIENTES**
- **MODELOS PREDICTIVOS**

Abstract

In Ecuador, companies dedicated to offering telecommunications services are in continuous innovation, for this they are using existing trends in the market such as service improvement, customer satisfaction, customer loyalty by granting bonuses, offers, massive promotions through calls phone calls and text messages. In the Telecommunications company under study in recent months, the percentage of clients who choose to cancel the contracted services has increased notably and this translates into a notable decrease in their economic income.

The purpose of this project is to reduce the percentage of clients who cancel the services offered by the Telecommunications company under study, through the design of an analytical model and through the analysis of historical information from the clients.

For the development of this project, the CRISP-DM Methodology (CRoss-Industry Standard Process for Data Mining) was used, which is an Industrial Standard procedure to perform Data Mining and allows finding a useful and understandable model that describes clearly and easy found patterns.

As a result, at the end of this research work, a predictive model was obtained that allows identifying clients with a high level of desire to cancel the services contracted in the Telecommunications company under study, this information is given the respective treatment to reduce the percentage of clients who choose to cancel the contracted services, which consequently causes an increase in their income.

KEYWORDS:

- **CUSTOMER SATISFACTION**
- **CANCELLATION OF SERVICES**
- **RETENTION AND LOYALTY OF CUSTOMERS**
- **PREDICTIVE MODELS**

Capítulo I

Introducción

Antecedentes

Con la masificación del internet y los servicios de telecomunicaciones a nivel mundial, muchas empresas se han proliferado ofreciendo más y mejores servicios a los usuarios, este gran catálogo de servicios se ha convertido en un beneficio para los usuarios ya que pueden elegir entre una gran variedad de servicios, el que más se adapte a su necesidad y posibilidad económica.

En el Ecuador las empresas dedicadas a ofrecer servicios de telecomunicaciones se encuentran en innovación continua, para esto están utilizando las tendencias existentes en el mercado como mejora de servicio, satisfacción del cliente, retención y fidelización de clientes otorgando bonos, ofertas, promociones masivas por medio de llamadas telefónicas y mensajes de texto. Luego de verificar los indicadores de eficiencia de gestión actualmente realizados, se ha detectado que las llamadas y los mensajes de texto masivos no están teniendo los resultados esperados y más bien están creando apatía en los clientes hacia la empresa ya que se ven sobrecargados de información que no es de su interés.

Justificación, Importancia y Alcance

La fidelización de clientes es una de las principales fuentes de ingreso de las instituciones o empresas, es así que mientras más clientes se mantengan con sus servicios y estos sean cancelados en las fechas previstas, más rentable es la empresa y, mayor utilidad es la que percibe; por el contrario, si existe deserción de clientes o la cancelación de sus servicios la empresa deja de percibir ingresos y puede llegar a la quiebra o a su cierre.

El presente proyecto se justifica en la empresa de telecomunicaciones objeto de estudio, dicha empresa realiza campañas de retención y fidelización reactivas, para lo cual selecciona a los clientes que ingresaron una petición de cancelación de servicio, a dichos clientes se les realiza el proceso de retención mediante llamadas telefónicas desde el Contact Center, de los cuales pocos son los clientes que aceptan mantenerse por uno (1) a (3) meses a cambio de un beneficio como descuentos, amplitud de ancho de banda, megas gratis, decodificadores o canales adicionales, es así que se ve disminuida la cantidad de

clientes fidelizados mensualmente, lo que incide proporcionalmente en una disminución en el ingreso por este rubro.

El beneficio que se ofrece a la empresa con este proyecto es la realización de un análisis más detallado a través de un modelo predictivo que permite identificar los potenciales clientes con un alto nivel o intención de deserción o cancelación de uno o varios servicios contratados, este es un insumo muy valioso ya que previo a que el cliente ingrese su solicitud de cancelación de servicios, el área de marketing y Retencion & Fidelizacion puede realizar un proceso de retención proactiva por parte del *Contact Center* realizando la gestión necesaria para la solución de sus requerimientos, consultas, reclamos y el ofrecimiento de algún beneficio como compensación en caso de ser necesario, de esta forma se reducirá el número de clientes que opten por la cancelación o cierre de el o los servicios contratados, mejorando así las utilidades e ingresos mensuales para la empresa.

El alcance de este proyecto consiste en realizar un modelo analítico para reducir el porcentaje de clientes que cancelan los servicios contratados con la empresa de Telecomunicaciones objeto de estudio por medio del análisis de la información histórica correspondiente al comportamiento de sus clientes.

Planteamiento del Problema

Hoy en día la fidelidad de los clientes, es un tema frecuentemente analizado por los altos directivos/ejecutivos de las empresas, resultando difícil y costoso conseguir nuevos clientes. La fidelidad es un objetivo que se ha deteriorado por varios factores y se requiere establecer una estrategia que basado en el análisis de los datos históricos nos permita determinar los factores que causan este comportamiento.

Las áreas de Marketing y Retención & Fidelización de la empresa de telecomunicaciones objeto de estudio en los últimos meses no han obtenido los resultados esperados para evitar la deserción de los clientes que tienen contratados uno o varios servicios, esto porque las técnicas aplicadas para la retención de clientes llegan en un periodo tardío, en consecuencia las ofertas y/o beneficios ofrecidos a los clientes llegan en un punto en el que el cliente está cansado, no las desea y en su gran mayoría desisten o cancelan el servicio contratado.

Por estos motivos, las técnicas de retención de clientes que deciden o solicitan realizar la cancelación de uno o varios servicios no obtienen los resultados esperados llevando a la empresa objeto de estudio a buscar alternativas para implementar sistemas que les permitan generar recomendaciones e identificar los clientes que estén pensando en optar por la cancelación de uno o varios servicios y con esto realizar una retención proactiva y eficiente del cliente, evitando su deserción que se traduce en una considerable perdida o reducción de sus ingresos económicos.

Objetivo General

Diseñar una solución de Inteligencia de Negocios[1] que, mediante el análisis de información histórica de los clientes, permita realizar un modelo analítico para reducir el porcentaje de clientes que cancelan los servicios que ofrece la empresa de Telecomunicaciones objeto de estudio.

Objetivos Específicos

OE1: Realizar una revisión inicial básica de literatura, para verificar estudios o trabajos similares existentes mediante la comparación de técnicas existentes, aplicables y utilizadas en el análisis de riesgo de deserción de clientes con el propósito de seleccionar la más eficiente y aplicarla en el presente proyecto.

OE2: Construir un modelo predictivo que permita realizar un análisis de riesgo de deserción de clientes, como resultado de la revisión inicial básica de literatura, identificando el que mejor se adapte a las necesidades de la empresa de telecomunicaciones objeto de estudio, y que permita identificar las posibles causas que inciden en la cancelación de los servicios que ofrece.

OE3: Realizar la evaluación del modelo predictivo, a través del uso de técnicas de validación para determinar el porcentaje de confianza de sus resultados y de esta manera verificar si el modelo de predicción utilizado ayuda o no con la reducción del porcentaje

[1] Se denomina inteligencia empresarial, inteligencia de negocios o BI (del inglés business intelligence), al conjunto de estrategias, aplicaciones, datos, productos, tecnologías y arquitecturas técnicas, las cuales están enfocadas a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa.

de clientes que optan por realizar la cancelación de los servicios que ofrece la empresa de Telecomunicaciones objeto de estudio.

Preguntas de Investigación

Para la consecución de los objetivos específicos del proyecto de análisis, diseño e implementación de un modelo para mejorar la fidelización de clientes, se respondieron las siguientes preguntas para cada objetivo específico:

OE1 – RQ1.1: ¿Cuáles son los trabajos o estudios existentes que aportan en el desarrollo del presente proyecto?

OE1 – RQ1.2: ¿Cuáles son las metodologías, modelos y/o algoritmos utilizados por las empresas de telecomunicaciones para reducir la cancelación de servicios?

OE2 – RQ2.1: ¿Cuál o cuáles son las técnicas de minería de datos aplicables en el presente caso de estudio que permitan realizar la predicción de los clientes con alta intención de cancelar los servicios que ofrece la empresa de telecomunicaciones?

OE2 – RQ2.2: ¿Cuál o cuáles son las herramientas y algoritmos para minería de datos con mejores resultados en la comprobación de los modelos de predicción para el presente caso de estudio?

OE3 – RQ3.1: ¿Es posible obtener técnicas, reglas o un modelo que permita realizar un análisis de riesgo de cancelación de servicios según el comportamiento histórico de los clientes?

OE3 – RQ3.2: ¿Es factible validar el modelo propuesto para la reducción del porcentaje de cancelación de servicios mediante el método deductivo con técnicas de evaluación como, encuestas, entrevistas y/o comprobación estadística?

Hipótesis de Investigación

La implementación de un modelo de fidelización en la empresa de telecomunicaciones objeto de estudio basado en el análisis de la información histórica del comportamiento de sus clientes permitirá identificar los potenciales clientes con intención de cancelar sus servicios.

Señalamiento de las Variables de la Hipótesis

Variable Dependiente: Identificar los potenciales clientes con intención de deserción o cancelación de uno o varios servicios en la empresa objeto de estudio.

Variable Independiente: Análisis de la información histórica de los clientes que han cancelado sus servicios.

La demostración de la hipótesis planteada en el presente trabajo se la realizará mediante el método deductivo, usando técnicas de investigación como, encuestas, entrevistas y/o comprobación estadística, adicionalmente al poder evaluar los resultados de la predicción de los potenciales clientes con intención de deserción o cancelación de uno o varios servicios contratados utilizando métricas como exactitud, coeficiente Kappa, porcentaje de error, número de registros clasificados correctamente, con la finalidad de determinar la eficacia del algoritmo.

Capítulo II

Marco Teórico

Estado del Arte

Para el análisis del estado del arte se realizó una revisión sistemática de literatura haciendo uso de las fases de criterios de inclusión, exclusión y estrategia de búsqueda que son parte fundamental de un SMS[2], con el fin de encontrar trabajos de investigación relevantes sobre el presente trabajo de investigación y que nos ayuden a cumplir con el objetivo planteado. Como fuentes de búsqueda de la información para la investigación se usaron los siguientes repositorios académicos IEEExplore y Google Académico.

Criterios de inclusión y exclusión

Los resultados de las búsquedas en las bases digitales dependiendo del tema de interés o consultado retornaron una gran cantidad de artículos y tesis relacionados por lo que es importante definir ciertas características idóneas de los artículos a ser tomados en cuenta, para el presente análisis se tomaron en cuenta los siguientes criterios:

Criterios de inclusión.

- Se incluyen en la búsqueda trabajos a partir del año 2009 (10 años).
- El artículo hace referencia a trabajos de investigación realizados para la toma de decisiones.
- El artículo debe contener información referente al uso de metodologías de inteligencia de negocios para sistemas de recomendaciones.
- El artículo describe información sobre modelos analíticos implementados y enfocados a predicciones.
- El artículo hace referencia a herramientas implementadas en proyectos analíticos.

[2] Systematic Mapping Study (SMS): El SMS permite realizar un análisis de la literatura existente sobre un determinado tema a fin de identificar el estado del arte de un tema determinado.

Criterios de exclusión.

- Artículos que contengan temas de inteligencia de negocios no relacionados con el presente trabajo de investigación.
- Artículos que no estén relacionados a la toma de decisiones.
- No se tomaron en cuenta artículos científicos cuyo contenido en idiomas sean diferentes al inglés o español.

Definición de la estrategia de búsqueda

Para la definición de la estrategia de búsqueda se realizaron los pasos que se detallan a continuación:

Revisión Inicial: Se realizó una búsqueda o revisión inicial en los distintos repositorios académicos para saber si existen o no estudios relacionados con las preguntas de investigación.

Validación cruzada de estudios: Se procedió a verificar si los estudios encontrados en la revisión inicial cumplen o no cumplen con los criterios de inclusión y exclusión deseados, con lo cual finalmente se obtiene el listado de estudios a trabajar en las siguientes fases.

Integración del Grupo de Control: El grupo de control está conformado por aquellos estudios que cumplen con los criterios de inclusión y exclusión para lo cual se procede a realizar un análisis del título de los estudios, introducción, conclusiones y palabras clave. Los estudios del grupo de control seleccionados se muestran en la Tabla 1.

Tabla 1.

Tabla de Estudios Candidatos.

Grupo de control	Año de publicación	Título	Palabras Clave
EC1	2016	Sistema de predicción de clientes desertores de tarjetas de crédito para la banca peruana usando Support Vector Machine	Prediction system, Support Vector Machine, retention, defection, clients
EC2	2013	Diseño de un modelo de gestión al momento de interactuar con los clientes, basado en el Marketing relacional aplicado para el sector de las empresas de medicina prepagada en la ciudad de Quito	Customer, services, loyalty, fidelity, retain
EC3	2017	Fidelización de los clientes en la empresa Garzón S.A. período 2015-2016	Loyalty, Strategies, Service, Sales, Estrategías, Servicio, Post Venta, Clientes, Ventas, Retención
EC4	2012	Estrategias de mejoramiento de las relaciones con los clientes en empresas de servicios informáticos. Caso Sonda del Ecuador	customers, services, loyalty, satisfaction, sale, strategies
EC5	2015	Modelos Predictivos del Churn – Abandono de Clientes – Para Operadores De Telecomunicaciones	Predictive models, Churn, abandonment, telecommunications, predict, customers,, retention, loyalty
EC6	2017	Designing of Customer and Employee Churn Prediction Model Based on Data Mining Method and Neural Predictor	Churn, services, predict, customer, loyalty, predictive model
EC7	2013	Parallel Set Determination and K-means Clustering for Data Mining on Telecommunication Networks	Data mining, algorithms, telecomunications, services, reduce, churn, predictive models,
EC8	2017	The Research of Customer Loyalty Improvement in Telecom Industry Based on NPS Data Mining	Customer, loyalty, Data mining, telecomunications, customer, relationship

Grupo de control	Año de publicación	Título	Palabras Clave
EC9	2016	Customer Loyalty Prediction In Multimedia Service Provider Company With K-Means Segmentation And C4.5 Algorithm	Customer, loyalty, Algorithm, prediction, services, model
EC10	2013	Predicting customers' future demand using data mining analysis: A case study of wireless communication customer	Prediction, service, data mining, decision tree, customers, churn, predict, techniques

Nota. La tabla contiene los estudios candidatos usados en la elaboración del Estado del Arte.

Construcción de la cadena de búsqueda

La cadena de búsqueda se crea en base a los términos claves más relevantes, es decir las palabras que más se repiten en el grupo de control seleccionado previamente, los cuales se pueden agrupar en contextos para facilitar el trabajo, como por ejemplo: Mejora, Inteligencia de Negocios, Tipo de Minería de Datos, Entorno de Análisis, como se muestra en la Tabla2.

Tabla 2.

Palabras recurrentes en grupos de control.

Contexto	Palabra Clave	EC1	EC2	EC3	EC4	EC5	EC6	EC7	EC8	EC9	EC10	Núm. Rep.
Mejora	Retention	X	X	X	X	X	X		X		X	8
	Fidelity	X	X	X	X	X	X					6
	Customer	X	X	X	X	X	X	X	X	X	X	10
	Prediction	X	X			X	X		X	X	X	7
Inteligencia de Negocios	Support Vector Machine	X				X	X	X		X		5
	Predictive models					X	X	X				3
	Decision tree	X				X	X	X	X	X	X	7
Tipo de Minería de Datos	Data Mining	X	X	X	X	X	X	X	X	X	X	10
	Data Mining techniques	X		X		X	X				X	5

Contexto	Palabra Clave	EC1	EC2	EC3	EC4	EC5	EC6	EC7	EC8	EC9	EC10	Núm. Rep.
Entorno de Análisis	Algorithm	X			X	X	X	X	X	X	X	8
	Relationship	X	X	X	X	X	X	X	X		X	9
	Churn	X				X	X	X	X	X	X	7
	Telecommunications	X			X	X		X	X		X	6
	Services	X	X	X	X	X	X	X	X	X	X	10
	Products	X	X	X	X	X	X		X	X	X	9

Nota. La tabla muestra los resultados dados por el grupo de control con respecto a las palabras concurrentes.

La cadena de búsqueda se formó con la combinación de las palabras que más se repiten en cada contexto, utilizando conectores como OR para las palabras que están dentro del mismo contexto y el conector AND para las palabras que están en contextos distintos, la misma que se aplicó en IEEE Xplore[3] y en Google Académico[4] en búsquedas de título y contenido con el objetivo de encontrar un contexto con métodos de solución al problema planteado.

Se realizó la búsqueda únicamente para trabajos desarrollados a partir del año 2012 y conformando la cadena de búsqueda por la unión de las palabras claves que más se repiten en cada contexto, se preseleccionaron las siguientes cadenas y finalmente se escogió la que arrojo menos resultados.

Se realizaron varios intentos de cadena de búsqueda con diferente número de artículos resultantes, tal como se puede observar en la Tabla3.

[3] IEEE Xplore es una base de datos de investigación académica que proporciona acceso al texto completo de artículos y trabajos sobre Ciencias de la Computación, Ingeniería Eléctrica y Electrónica. Así mismo, IEEE Xplore Books contiene acceso parcial a más de 500 libros y libros electrónicos, en asociación con Wiley e-book y con MIT Press. Su sitio web es: https://ieeexplore.ieee.org/Xplore/home.jsp.

[4] Google Académico (en inglés, Google Scholar) es un buscador de Google enfocado y especializado en la búsqueda de contenido y literatura científico-académica.1 El sitio indexa editoriales, bibliotecas, repositorios, bases de datos bibliográficas, entre otros; y entre sus resultados se pueden encontrar citas, enlaces a libros, artículos de revistas científicas, comunicaciones y congresos, informes científico-técnicos, tesis, tesinas y archivos depositados en repositorios. Su sitio web es: https://scholar.google.es/

Tabla 3.

Determinación de las Cadenas de Búsqueda.

Cód.	IEEE			Google Académico			TOTAL
	Cadena de búsqueda propuesta	Artículos encontrados	Artículos Candidatos encontrados	Cadena de búsqueda propuesta	Artículos encontrados	Artículos Candidatos encontrados	Artículos Candidatos encontrados
CB1	((((RETENTION) OR (CUSTOMER) AND DECISION TREE) AND DATA MINING) OR SERVICES)	1913	5	((((RETENCION) O (CLIENTE) Y ARBOL DE DECISIONES) Y MINERIA DE DATOS) O SERVICIOS)	3780	1	1
CB2	(((((((RETENTION) OR CUSTOMER) AND DECISION TREE) AND DATA MINING) OR ALGORITHM) AND RELATHIONSHIP) OR PREDICTION)	1125	2	(((((((RETENCION) O CLIENTE Y ARBOL DE DECISIONES) Y MINERIA DE DATOS) O ALGORITMOS) Y RELACION) O PREDICCION)	474	2	2
CB3	((((((((RETENTION) OR FIDELITY)OR CUSTOMER) AND PREDICTIVE MODELS) OR DECISION TREE) AND DATA MINING) OR ALGORITHM) AND RELATIONSHIP)	213	1	((((((((RETENCION) O FIDELIDAD) O CLIENTE) Y MODELOS PREDICTIVOS) O ARBOLES DE DECISION) Y MINERIA DE DATOS) O ALGORITMO) Y RELACION)	89	2	2

Cód.	IEEE			Google Académico			TOTAL
	Cadena de búsqueda propuesta	Artículos encontrados	Artículos Candidatos encontrados	Cadena de búsqueda propuesta	Artículos encontrados	Artículos Candidatos encontrados	Artículos Candidatos encontrados
CB4	((((((((RETENTION) OR FIDELITY) OR CUSTOMER)AND PREDICTIVE MODELS) OR DECISION TREE) AND DATA MINING) AND CHURN) OR PRODUCTS)	479	3	((((((((RETENCION) O FIDELIDAD) O CLIENTE) Y MODELOS PREDICTIVOS) O ARBOLES DE DECISION) Y MINERIA DE DATOS) Y CANCELACION) O PRODUCTOS)	27	1	1
CB5	(CHURN) AND (CUSTOMER) AND (PREDICTION)AND (DECISION TREE)	42	2	(CANCELACION) Y (CLIENTE) Y (PREDICCION) Y (ARBOL DE DECISION)	20	2	2

Nota. La tabla contiene los resultados de las cadenas de búsqueda propuestas en las bibliotecas digitales.

Posterior a la verificación de las cadenas propuestas la que devolvió la menor cantidad de resultados y en la que también constaban varios de los estudios candidatos es:

(CHURN) AND (CUSTOMER) AND (PREDICTION) AND (DECISION TREE)

Una vez obtenidos los resultados se procedió a aplicar un filtro en donde se seleccionaron únicamente artículos publicados en el idioma inglés, los artículos que fueron

publicados a partir del año 2018 y aplicando los criterios de inclusión y exclusión quedando una selección de 5 artículos que conformarían el grupo de control, finalmente se realizó la revisión de los documentos encontrados, los cuales se listan en la Tabla 4.

Tabla 4.

Estudios del Grupo de Control.

Estudios	TITULO	AÑO	AUTORES
GC1	Designing of customer and employee churn prediction model based on data mining method and neural predictor	2017	Sepideh Hassankhani Dolatabadi ; Farshid Keynia
GC2	A comparative study of customer churn prediction in telecom industry using ensemble based classifiers	2017	Abinash Mishra ; U. Srinivasulu Reddy
GC3	Customer churn prediction using data mining approach	2018	Laila M. Qaisi ; Ali Rodan ; Kefaya Qaddoum ; Rizik Al-Sayyed
GC4	Prediction of Churning Behavior of Customers in Telecom Sector Using Supervised Learning Techniques	2018	Muhammad Ali ; Aziz Ur Rehman ; Shamaz Hafeez ; Dr. Muhammad Usman Ashraf
GC5	Churn prediction in telecommunication using machine learning	2017	Kriti Mishra ; Rinkle Rani

Nota. La tabla muestra información de los 5 artículos que conforman el grupo de control.

A continuación se presentan los resúmenes con la información más relevante de los documentos del grupo de control seleccionado.

Designing of customer and employee churn prediction model based on data mining method and neural predictor

El presente artículo se basa en el análisis de la competencia entre las empresas en el sector de servicios, predecir la pérdida de clientes para mejorar la retención clientes, el impacto de la lealtad a la marca, la pérdida de empleados en una organización, así como la dificultad de atraer a un nuevo cliente por cada cliente perdido, utiliza para su análisis técnicas de minería de datos como: Decision Tree, Naïve Bayes, Artificial Neural Network y Support Vector Machine (SVM). La principal contribución de este artículo es la demostración

de que las técnicas de aprendizaje automático (por ejemplo, SVM) se pueden usar para construir modelos predictivos, confiables y precisos para determinar o predecir la tasa de cancelación de clientes y/o empleados.(Dolatabadi & Keynia, 2017)

A comparative study of customer churn prediction in telecom industry using ensemble based classifiers

En este trabajo, se muestra el papel vital de la tasa de cancelación de clientes en los seguros de vida, banca y la industria de telecomunicaciones, habla también acerca del avance actual sobre el aprendizaje automático e inteligencia artificial, factores que han favorecido notablemente a que la predicción de la cancelación sea más realista y precisa. Resalta además que todos estos avances tecnológicos son esenciales para la detección temprana de clientes que corren un alto riesgo de abandonar la empresa o los servicios. En este documento, los clasificadores: namely Bagging, Boosting y Random Forest, se utilizaron para predecir la tasa de cancelación de clientes en la industria de telecomunicaciones. Compararon los clasificadores namely Bagging con los clasificadores well-known classifiers namely Decision Tree, Naïve Bayes Classifier y Support Vector Machine (SVM) y los resultados experimentales muestran que Random Forest tiene menos tasa de error, baja especificidad, alta sensibilidad y mayor precisión (91.66%) en comparación con otros métodos.(A. Mishra & Reddy, 2017)

Customer churn prediction using data mining approach

Este artículo habla sobre la predicción de la pérdida de clientes, indica que se está convirtiendo en una de las preocupaciones más importantes para las organizaciones en general y principalmente en el campo de las telecomunicaciones. Para este propósito, han realizado un estudio comparativo sobre tres modelos de clasificación predictiva de aprendizaje automático aplicados a un conjunto de datos para predecir la pérdida de clientes. Los modelos: árboles de decisión (DT), Naïve Bayes (NB) y reglas de rendimiento de inducción fueron evaluados para especificar el mejor rendimiento mediante varias medidas, como la precisión, la recuperación, la medida F y el área bajo cubierta (AUC).(Qaisi et al., 2018)

Prediction of Churning Behavior of Customers in Telecom Sector Using Supervised Learning Techniques

El presente artículo basado en que la minería de datos es un área extensa que co-relaciona diferentes ramas, es decir, estadísticas, base de datos, aprendizaje automático e inteligencia artificial y que dicha minería de datos se la puede usar en distintos sectores o industrias. La tasa de pérdida de clientes se refiere cuando el cliente ya no quiere mantener su relación con la empresa. La tasa de pérdida de clientes está jugando un papel importante en la gestión de clientes. Hoy en día, las compañías de telecomunicaciones se están enfocando en identificar clientes de alto valor y potenciales clientes para aumentar las ganancias y la participación de las empresas en el mercado. Además hacen hincapié en que conseguir un nuevo cliente es más difícil y costoso que el retener al cliente existente. En la investigación de este artículo realizan la predicción del comportamiento del cliente mediante el uso de varias técnicas de extracción de datos factor importante para ayudará a analizar el comportamiento del cliente y clasificará si se trata de un cliente con deseo o intención de cancelar sus servicios contratados. En esta investigación, utilizaron el conjunto de datos disponible en línea en Kaggle para predecir el comportamiento de la pérdida de clientes utilizando diferentes clasificadores como: SVM (Support Vector Machine), Bagging, Stacking, C50 / J48, PART, Naïve Bayes, Baysen Net, Adaboost y como resultado de este análisis muestra que el modelo dio un nivel de precisión de 99.8% utilizando algoritmos Bagging.

Churn prediction in telecommunication using machine learning

En el presente artículo, basados en que la industria de las telecomunicaciones tiene una fuerte competencia para retener clientes, se ha convertido en uno de los sectores de investigación del aprendizaje automático y minería de datos. Dado que el comportamiento de la tasa de cancelación de servicios se debe monitorear de cerca y de manera eficiente, se requiere de un modelo de predicción de la tasa de cancelación metódica para controlar o mitigar al máximo la cancelación de los servicios. Habla además de los principales contratiempos para lograr los rendimientos deseados en un clasificador son los enormes conjuntos de datos, el gran espacio de características y la distribución de clases desequilibrada. En este trabajo, exploran la implicación de Synthetic Minority Over-sampling TEchnique (SMOTE) para reducir el desequilibrio en los datos en colaboración con diferentes técnicas de reducción de características como la co-relación de características, la relación de ganancia, la ganancia de información y el método de evaluación de características OneR. Los clasificadores de árboles de clasificación y regresión (CART), CART empaquetado y árboles de decisión parcial (PART) están capacitados para analizar

el rendimiento en conjuntos de datos de espacio de funciones equilibrado y reducido. El desempeño de la predicción de los clasificadores se evalúa a través de medidas tales como el área bajo la curva (AUC), la sensibilidad y la especificidad. Finalmente, a través de simulaciones, concluye que el método propuesto basado en SMOTE, la co-relación y el conjunto funciona bien para pronosticar la cancelación o abandono. Por lo tanto, esta metodología puede ser útil para que la industria de las telecomunicaciones prediga la tasa de cancelación o abandono de sus servicios.(K. Mishra & Rani, 2017)

Resultados del Estado del Arte

Posterior a realizar un estudio de literatura de los artículos relacionados al tema de investigación, se identifica que todos los autores coinciden en la importancia de la creación de un modelo que permita reducir el porcentaje de cancelación de los servicios contratados en una empresa y principalmente que en base al análisis de la data histórica de los clientes se estudie su comportamiento con la finalidad de obtener los clientes con intención de cancelar sus servicios para poder aplicarles técnicas de retención y fidelización, en varios de los artículos revisados se realiza un análisis y comparación de los métodos de predicción utilizados para este fin común, llegando a concluir que las empresas aplican diferentes métodos para realizar el análisis de clientes entre ellos Data Mining, algoritmos de programación, redes neuronales, SVM (Support Vector Machine), Bagging, Stacking, C50 / J48, PART, Naïve Bayes, entre otras, siendo las más recomendadas Bagging y SVM (Support Vector Machine) ya que poseen mayor exactitud y precisión en sus resultados.

Los análisis realizados en los estudios y trabajos revisados permitirán obtener mejores resultados en el desarrollo del presente proyecto.

Contexto Organizacional

En este punto brevemente toparemos algunos aspectos para describir de manera general la razón de ser de la organización, tener una idea más clara del problema y la solución a desarrollar.

Descripción

Es una empresa de telecomunicaciones ecuatoriana cuyo giro de negocio es el ofrecer servicios convergentes de telecomunicaciones y TICs[5], a la comunidad para su desarrollo, comunicación e integración al mundo, es una empresa con once años en el mercado ecuatoriano y cuenta con más de 100 puntos de venta, asesoría, atención y soporte a nivel nacional. La empresa de telecomunicaciones se dedica a la venta de servicios de telefonía fija local, regional e internacional, acceso a internet estándar y de alta velocidad (DSL[6], GPON[7], Internet móvil 3G y 4G[8] LTE[9]), televisión satelital y telefonía móvil en el territorio nacional ecuatoriano.

Por motivos de confidencialidad el nombre de la empresa se mantendrá reservado.

Misión, Visión y valores

Tiene como misión brindar la mejor experiencia en todos los servicios que ofrece a la población con lo cual espera llegar a ser la empresa líder en el territorio ecuatoriano a través de una gestión de calidad y de excelencia.

Algunos de sus valores de manera general son:

- Compromiso con el cliente.
- Trabajar en equipo.
- Ser eficientes.
- Innovadores.

---

[5] Las TIC son el conjunto de tecnologías desarrolladas en la actualidad para una información y comunicación más eficiente, las cuales han modificado tanto la forma de acceder al conocimiento como las relaciones humanas. TIC es la abreviatura de Tecnologías de la Información y la Comunicación.

[6] La línea de abonado digital o línea de suscriptor digital, Digital Subscriber Line (DSL), es una familia de tecnologías que proporcionan el acceso a Internet mediante la transmisión de datos digitales a través del par trenzado de hilos de cobre convencionales de la red telefónica básica o conmutada.

[7] La Red Óptica Pasiva con Capacidad de Gigabit (GPON o Gigabit-capable Passive Optical Network en inglés) es una tecnología de acceso de telecomunicaciones que utiliza fibra óptica para llegar hasta el suscriptor.

[8] 3G y 4G son las siglas utilizadas para referirse a la tercera y cuarta generación de tecnologías de telefonía móvil.

[9] LTE es una tecnología de transmisión de datos de banda ancha inalámbrica que está principalmente diseñada para poder dar soporte al constante acceso de teléfonos móviles y de dispositivos portátiles a internet.

- Compromiso con la sociedad y el medio ambiente.

Objetivos de la Empresa de Telecomunicaciones

a) Mantener la satisfacción de sus clientes en los servicios ofertados (telefonía fija, móvil, internet y televisión satelital) a través de la atención personalizada de ejecutivos de ventas, post-venta y soporte técnico por medio de los distintos canales de acceso implementados para su comunicación.
b) Incrementar su participación en el mercado sobre los servicios ofertados (telefonía fija, móvil, internet y televisión satelital) a nivel nacional.
c) Evitar o reducir al máximo la cancelación de los servicios contratados ofreciendo un servicio de calidad y tratamiento anticipado hacia sus clientes.

Evaluación de la situación actual.

En la actualidad la empresa objeto de estudio cuenta con un reporte que le permite conocer la cantidad de clientes que cancelan sus servicios contratados de forma mensual. La retención anticipada que se realiza con los clientes se basa en un reporte empírico considerando ciertas variables que fueron el resultado de acuerdos entre la jefatura y supervisores en reuniones mantenidas, además se debe indicar que dichas variables no han sido analizadas ni modificadas desde hace más de un año, dicho reporte se lo distribuye a los asesores de Retencion&Fidelizacion para que realicen campañas de llamadas salientes en las cuales se indague sobre la satisfacción de los clientes y si existe alguna molestia o problema se gestiona la solución y se ofrece algún tipo de compensación con lo que se pretende retener a los clientes y evitar la cancelación de los servicios contratados, este proceso descrito no ha logrado reducir la cantidad de clientes que cancelan los servicios contratados puesto que el reporte no contiene una gran cantidad de clientes con una verdadera intención de cancelar sus servicios y se está dejando por fuera a los clientes que si tienen una gran intención de cancelación y al no ser gestionados mantienen su malestar y finalmente cancelan sus servicios contratados.

Solución a desarrollar.

De acuerdo a lo descrito en el punto anterior la solución a desarrollar consiste en la construcción de un modelo analítico que en base al análisis del comportamiento de los clientes se identifique un posible patrón que siguen los clientes antes de optar

por la cancelación de sus servicios contratados, al conocer el patrón que siguen estos clientes mediante el modelo analítico extraer la base de clientes que están siguiendo dicho patrón y distribuirla a los asesores de Retencion&Fidelizacion para que realicen la gestión que corresponda de acuerdo a las directrices que indique su jefatura, con lo que se espera reducir la cantidad de clientes que cancelan los servicios contratados con la empresa de telecomunicaciones objeto de estudio.

Minería de Datos

La minería de datos y el descubrimiento de conocimiento en bases de datos (KDD – Knowledge Discovery in Database) atraen una gran cantidad de investigación y de atención de los medios durante los últimos tiempos. La extracción de datos y el descubrimiento de conocimiento en la base de datos están relacionados entre sí y con campos relacionados como aprendizaje automático, estadísticas y bases de datos. La minería de datos usa métodos robustos que ayudan a la reducción de los costos y riesgos para un negocio, de igual manera apoya el incremento de las rentas por cuanto extrae información crucial y estratégica a través de los datos disponibles.(Fayyad et al., 1996)

Técnicas De Minería De Datos

Las técnicas de Minería de Datos basan su clasificación en dos grandes categorías: Supervisadas (Predictivas) y No Supervisadas (Descriptivas). La técnica es un enfoque conceptual que permite extraer información de los datos y, por lo general es implementada por varios algoritmos. En la práctica, cada algoritmo representa una manera de desarrollar una técnica específica paso a paso; por lo que, es indispensable tener una comprensión de alto nivel de los algoritmos para conocer cuál de tantas técnicas son las más apropiadas para cada problema; adicionalmente, es necesario entender las características y los parámetros de los algoritmos para preparar los datos a analizar. (Molina López & García Herrero, 2006).

Técnicas Supervisadas

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados. La salida de la función puede ser un

valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente. Dentro de las técnicas Supervisadas se tiene:

a) La <u>Predicción</u>

- Regresión
- Arboles de Predicción
- Estimador de Núcleos

b) La <u>Clasificación</u>

- Tabla de Decisión
- Arboles de Decisión
- Inducción de Reglas
- Bayesiana
- Basado en Ejemplares
- Redes de Neuronas
- Lógica Borrosa
- Técnicas Genéticas

(Molina López & García Herrero, 2006)

Técnicas No Supervisadas

Son un método de Aprendizaje Automático donde un modelo es ajustado a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori. En el aprendizaje no supervisado, un conjunto de datos de objetos de entrada es tratado. Así, el aprendizaje no supervisado típicamente trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos.

El aprendizaje no supervisado también es útil para la compresión de datos: fundamentalmente, todos los algoritmos de compresión dependen tanto

explícita como implícitamente de una distribución de probabilidad sobre un conjunto de entrada.

Otra forma de aprendizaje no supervisado es la agrupación (en inglés, clustering), el cual a veces no es probabilístico.

Dentro de las técnicas Supervisadas se tiene:

a) El <u>Clustering</u>

- Numérico
- Conceptual
- Probabilístico

b) La <u>Asociación</u>

- A Priori (Molina López & García Herrero, 2006)

Metodologías de Minería De Datos

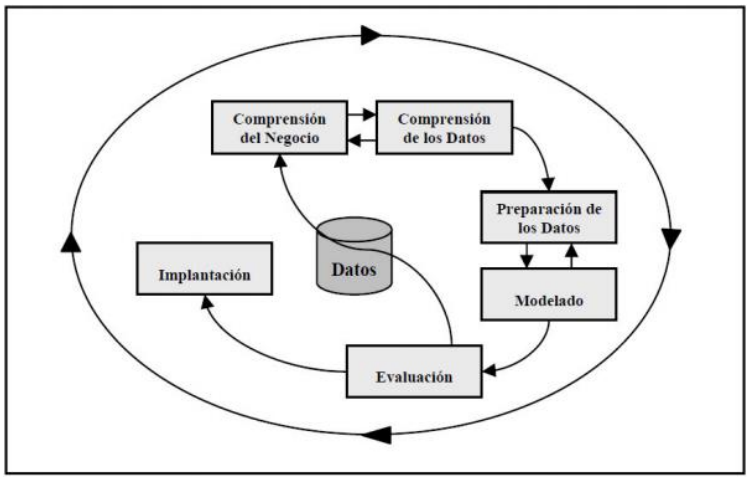
Existen tres metodologías que dominan el proceso de la minería de datos, estas son KDD, CRISP-DM y SEMMA, las mismas que se detalla brevemente a continuación:

CRISP-DM (CRoss-Industry Standard Process for Data Mining: Procedimiento Industrial Estándar para realizar Minería de Datos)

Es una metodología de Minería de datos para desarrollo de proyectos analíticos. CRISP-DM es como un proceso jerárquico, consiste en seis fases definidas de manera cíclica (Chapman et al., 2000) y son: análisis del problema, comprensión de datos, preparación de datos, modelado, evaluación y despliegue. Estas seis fases no son rígidas en procedimiento, de hecho muchas veces existe retroalimentación entre diferentes fases. En muchas ocasiones depende de la salida de una fase para saber que etapa o tarea de la etapa es la que va a seguir a continuación. Todo lo mencionado se puede verificar en la Figura 1.

Figura 1.

Fases de la metodología CRISP-DM



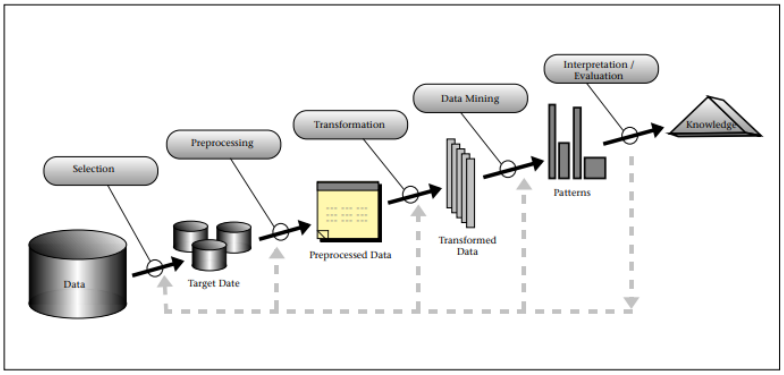
Nota. El grafico representa las fases de la metodología CRISP-DM. (Chapman et al., 2000).

KDD (Knowledge Discovery in Databases)

Es un proceso iterativo e interactivo que combina la experiencia en un problema con una variedad de técnicas de análisis de datos tradicionales y tecnologías avanzadas de aprendizaje automático por procedimientos computacionales. El objetivo es descubrir patrones y relaciones en los datos que puedan ser usados para hacer predicciones válidas. Propone o consta de 5 fases: Selección, pre procesamiento, transformación, minería de datos y evaluación e implantación, las mismas que se pueden observar en la Figura 2.

Figura 2.

Fases de la metodología KDD



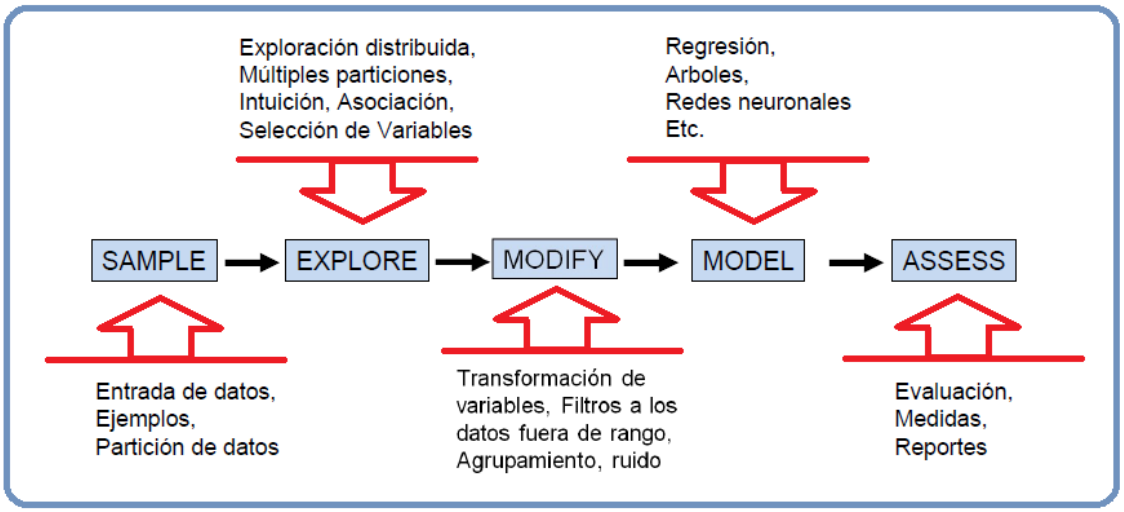
Nota. El grafico representa las fases de la metodología KDD. (Fayyad et al., 1996).

SEMMA (Sample, Explore, Modify, Model, Assess)

Esta metodología se define como el proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos. Su funcionamiento empieza tomando una muestra representativa de los datos, se aplican técnicas estadísticas de exploración y visualización, se seleccionan y transforman las variables, se modela con las variables para predecir y finalmente se evalúa la exactitud del modelo, tal y como se puede observar e la Figura 3.

Figura 3.

Fases y actividades de la metodología SEMMA



Nota. El grafico representa las fases de la metodología SEMMA. (Autoría propia).

Comparación entre las Metodologías de Minería de datos KDD, SEMMA y CRISP-DM según sus actividades específicas

Cada una de estas tres metodologías posee o define ciertas actividades a ejecutar, en este punto se realizó un análisis y comparación sobre las características que cumplen estas metodologías de acuerdo a cada fase del proceso de minería de datos. A continuación se detalla las características que cumple cada metodología con lo que se tendrá una mejor visión sobre la metodología que más o mejor se adapte al desarrollo de nuestro proyecto, esto se lo puede apreciar en la Tabla 5.

Tabla 5.

Tabla comparativa entre las metodologías KDD, SEMMA y CRISP-DM según sus actividades específicas.

FASE	CARACTERÍSTICAS	METODOLOGÍAS		
		CRISP-DM	KDD	SEMMA
Análisis del Problema	Evaluación de la organización	SI		
	Identificación de todas aquellas personas implicadas en el proyecto que tienen algún interés o intervienen en el	SI	SI	SI
	Definición del problema			SI
	Evaluación de fuentes de datos		SI	SI
	Análisis de soluciones potenciales	SI		
	Definición de objetivos del proyecto	SI		
	Determinación de criterio de éxito	SI	SI	
	Análisis de técnicas de DM	SI	SI	
	Especificación de documentos (entregables)			
Selección y Preparación de Datos	Análisis Exploratorio	SI	SI	
	Limpieza de datos	SI	SI	SI
	Transformación de variables y creación de atributos derivados	SI	SI	SI
	Análisis descriptivo de datos depurados			SI
	Revisión del conjunto de datos final con el usuario		SI	
Modelado	Selección de técnicas de Minería de Datos	SI	SI	
	Evaluación de resultados	SI	SI	S
	Evaluación de nuevos modelos	SI		SI
	Directivas para el descubrimiento de patrones		SI	
Evaluación	Interpretación de modelos	SI	SI	SI

FASE	CARACTERÍSTICAS	METODOLOGÍAS		
		CRISP-DM	KDD	SEMMA
Implementación	Comparación y ponderación de modelos	SI	SI	
	Revisión del proceso	SI	SI	SI
	Directivas para modelos no viables	SI	SI	
	Implementación de nuevos modelos	SI	SI	SI
	Programa de Mantenimiento	SI	SI	
	Resumen del proyecto	SI		
Total de Características que cumple:		19	17	10
% de Características que cumple:		76%	68%	40%

Nota. La tabla representa un cuadro comparativo entre las metodologías KDD, SEMMA y CRISP-DM según sus actividades específicas por cada fase.

Al concluir el análisis de la comparativa entre KDD, SEMMA y CRISP-DM según sus actividades específicas se concluye que la metodología que se adapta de mejor manera al desarrollo de nuestro proyecto es CRISP- DM, debido a que actualmente es la guía de referencia más utilizada en el desarrollo de proyectos de Minería de Datos y permite llevar el nivel de detalle de las actividades hasta 2, es decir que las tareas generales se proyectan o descomponen en tareas específicas.

Capítulo III

Propuesta de un Modelo Analítico aplicando la Metodología CRISP-DM

En este capítulo se utilizará la Metodología CRISP-DM para el desarrollo de la propuesta del modelo analítico, por lo que se aplicará las fases que comprenden dicha metodología. A continuación vamos a describir brevemente cada una de las fases.

Fase I. Fase de comprensión del negocio o problema.

Es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables.(Gallardo Arancibia, 2009)

Fase II. Fase de comprensión de los datos

Comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. (Gallardo Arancibia, 2009)

Fase III. Fase de preparación de los datos.

Una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.(Gallardo Arancibia, 2009)

Fase IV. Fase de modelado

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada al problema.

- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica. (Gallardo Arancibia, 2009)

Fase V. Fase de evaluación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. (Gallardo Arancibia, 2009)

Fase VI. Fase de implementación o despliegue.

Una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso. (Gallardo Arancibia, 2009)

La metodología CRISP-DM se eligió en base al análisis realizado en el capítulo 2, específicamente en el punto 2.4.

Comprensión Del Negocio

Es una empresa de telecomunicaciones ecuatoriana cuyo giro de negocio es el ofrecer servicios convergentes de telecomunicaciones y TICs, a la comunidad/sociedad/población para su desarrollo, comunicación e integración al mundo, es una empresa con más de once años en el mercado ecuatoriano y cuenta con alrededor de 100 puntos de venta, asesoría, atención y soporte a nivel nacional. La empresa de telecomunicaciones se dedica a la venta de servicios de telefonía fija local, regional e internacional, acceso a internet estándar y de alta velocidad (DSL, GPON, Internet móvil 3G y 4G LTE), televisión satelital y telefonía móvil en el territorio nacional ecuatoriano.

Área de retención y fidelización de clientes

El cliente es, ha sido y será siempre la razón de ser de cualquier empresa. De ellos depende el éxito o fracaso de cualquier negocio, por esta razón es primordial diseñar o construir estrategias enfocadas en la atención o asistencia hacia el cliente, principalmente dirigidas a satisfacer las necesidades de éste, con el fin de retener y fidelizar a cada uno de los clientes que ya tiene la empresa y por supuesto a captar nuevos usuarios que garanticen la continuidad y éxito del negocio.

Objetivos del Proyecto

El objetivo o resultado que se pretende obtener con este estudio es un modelo analítico que nos indique en base al patrón generado con el histórico (ya modelado y confirmado) quienes son los clientes con un alto porcentaje de intención para cancelar los servicios actualmente contratados con la empresa de telecomunicaciones objeto de estudio.

Definición del problema de minería de datos (evaluación de la situación actual)

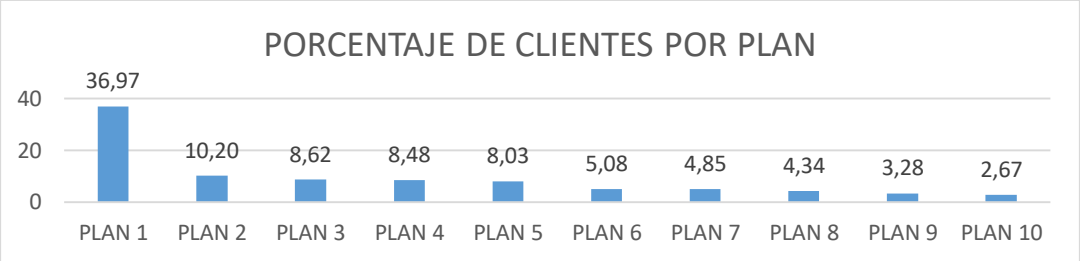
Al momento el Área de Retención y Fidelización obtiene la data para realizar campañas preventivas o anticipadas de una forma empírica, ellos realizan un barrido o selección de los clientes en base a unas 2 o 3 variables que las definieron en base a su experiencia, por ejemplo: si un cliente realizo reportes por servicio técnico (falla o mal funcionamiento del servicio) por lo menos 3 veces en el mes, ya es un posible candidato o tiene un 33,33% de probabilidades de salir seleccionado en la data para realizarle una campaña de retención preventiva o anticipada, si cumple con la condición de la segunda variable tendrá el 66,66% de probabilidades de ser seleccionado en la data y si cumple con la condición de la tercera variable, es ya un fijo candidato, esto les funciona pero no de la forma esperada“ ya que el análisis de las variables no está dado en base al estudio o análisis real del comportamiento (histórico) de los clientes previo a su decisión de cancelar el o los servicios contratados”.

Comprensión De Los Datos

En esta fase se realiza la recolección de datos inicial, con el objetivo o finalidad de tener una idea inicial del problema a resolver, además de familiarizarnos con los datos a estudiar y determinar la calidad de los mismos, adicional se identificará las relaciones existentes entre los campos y tablas que se van a usar del Data Warehouse existente, tal y como se muestra en la Figura 4.

Figura 4. Porcentaje de clientes por tipo de plan de datos celular (los 10 más representativos).

Porcentaje de clientes por tipo de plan de datos celular (los 10 más representativos).



Nota. El grafico representa el porcentaje de clientes por tipo de plan de datos celular. (Autoría propia).

Recolectar los Datos Iniciales

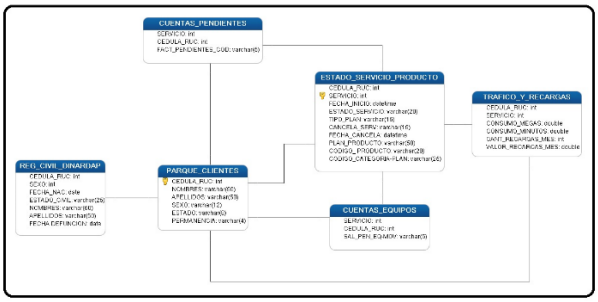
Los datos que se utilizan en el presente proyecto se refieren a la empresa de telecomunicaciones objeto de estudio y de sus clientes que incluyen información de tipo personal como nombres, apellidos, cédula de identidad, RUC, servicio(s) contratado(s), etc., por cuestiones legales y de confidencialidad de información se utilizaran datos enmascarados.

Descripción de los Datos

Los datos se encuentran almacenados en el Data Warehouse que posee actualmente la empresa de telecomunicaciones objeto de estudio, de dicho DWH usaremos únicamente ciertas tablas y campos con el objetivo de armar el Data Set a usar para el análisis y desarrollo del proyecto, Esto se lo puede apreciar en la Figura 5 y tabla 6.

Figura 5.

Tablas y campos usados del DWH existente.



Nota. La figura representa, del DWH existente, las tablas y campos usados para el presente caso de estudio. (Autoría propia).

Tabla 6.

Descripción de tablas a usar del DWH existente.

TABLA	DESCRIPCION TABLA	CAMPOS USADOS
CUENTAS_EQUIPOS	Contiene información de los equipos adquiridos por cliente a la empresa de telecomunicaciones objeto de estudio.	CEDULA_RUC SERVICIO SAL_PEN_EQ-MOV
CUENTAS_PENDIENTES	Contiene información sobre las cuentas pendientes por cliente a la empresa de telecomunicaciones objeto de estudio.	CEDULA_RUC SERVICIO FACT_PENDIENTES_COD
ESTADO_SERVICIO_PRODUCTO	Contiene información relevante sobre las cuentas pendientes por cliente a la empresa de telecomunicaciones objeto de estudio.	CEDULA_RUC SERVICIO FECHA_INICIO ESTADO_SERVICIO TIPO_PLAN CANCELA_SERV FECHA_CANCELA PLAN_PRODUCTO CODIGO_ PRODUCTO CODIGO_CATEGORIA-PLAN

TABLA	DESCRIPCION TABLA	CAMPOS USADOS
PARQUE_CLIENTES	Contiene información sobre los clientes de la empresa de telecomunicaciones objeto de estudio.	CEDULA_RUC NOMBRES APELLIDOS SEXO ESTADO PERMANENCIA
TRAFICO_Y_RECARGAS	Contiene información del consumo de megas, minutos, etc. del plan contratado.	CEDULA_RUC SERVICIO CONSUMO_MEGAS CONSUMO_MINUTOS CANT_RECARGAS_MES VALOR_RECARGAS_MES
REG_CIVIL_DINARDAP	Contiene información del cliente / ciudadano.	CEDULA_RUC SEXO FECHA_NAC ESTADO_CIVIL NOMBRES APELLIDOS FECHA_DEFUNCION

Nota. La tabla muestra una breve descripción de las tablas a usar del DWH existente.

Exploración de los Datos

Una vez concluida y comprendida la descripción de los datos, se continúa con la exploración de los datos, esto nos ayudará a determinar la consistencia de los mismos.

Los planes de datos celulares[10], donde se concentran la mayor cantidad de clientes son Plan 1, seguido por Plan 2, Plan3, Plan 4, Plan 5, etc. El total de clientes que tienen contratados planes de datos celulares con las empresa de Telecomunicaciones objeto de estudio y sus servicios que se encuentran en estado activo son **323446**. Por motivos de confidencialidad hemos enmascarado los nombres de los planes de datos celulares DENOMINANDOLOS COMO PALN1 ETC, todo lo mencionado se puede observar en la Tabla 7

[10] Plan de datos celular es aquel que en estos días se ofrece a los clientes que poseen celulares inteligentes, aquellos que están preparados para hacer uso del Internet y así puedan navegar por la Web, enviar y recibir e-mails, llamadas telefónicas, usar redes sociales como Facebook, whatsapp, twitter, etc. Muchos de estos teléfonos inteligentes pueden acceder a redes de alta velocidad ofrecidas por las compañías de telecomunicaciones y telefonía celular.

Tabla 7.

Número de clientes por tipo de plan de datos celular.

NOMBRE DEL PLAN DE TELEFONIA MOVIL	NUMERO DE CLIENTES POR PLAN
PLAN 1	119571
PLAN 2	32997
PLAN 3	27897
PLAN 4	27415
PLAN 5	25971
PLAN 6	16447
PLAN 7	15684
PLAN 8	14039
PLAN 9	10595
PLAN 10	8638
PLAN 11	7437
PLAN 12	5495
PLAN 13	4370
PLAN 14	1141
PLAN 15	866
PLAN 16	747
PLAN 17	610
PLAN 18	371
PLAN 19	293
PLAN 20	291
PLAN 21	272
PLAN 22	261
PLAN 23	251
PLAN 24	232
PLAN 25	220
PLAN 26	194
PLAN 27	185
PLAN 28	140
PLAN 29	92
PLAN 30	87
PLAN 31	87
PLAN 32	86
PLAN 33	64
PLAN 34	58
PLAN 35	53
PLAN 36	36

NOMBRE DEL PLAN DE TELEFONIA MOVIL	NUMERO DE CLIENTES POR PLAN
PLAN 37	33
PLAN 38	23
PLAN 39	22
PLAN 40	21
PLAN 41	20
PLAN 42	18
PLAN 43	16
PLAN 44	10
PLAN 45	10
PLAN 46	8
PLAN 47	8
PLAN 48	7
PLAN 49	6
PLAN 50	6
PLAN 51	5
PLAN 52	5
PLAN 53	4
PLAN 54	4
PLAN 55	4
PLAN 56	4
PLAN 57	3
PLAN 58	2
PLAN 59	2
PLAN 60	2
PLAN 61	2
PLAN 62	1
PLAN 63	1
PLAN 64	1
PLAN 65	1
PLAN 66	1
PLAN 67	1
PLAN 68	1
PLAN 69	1
TOTAL	**323446**

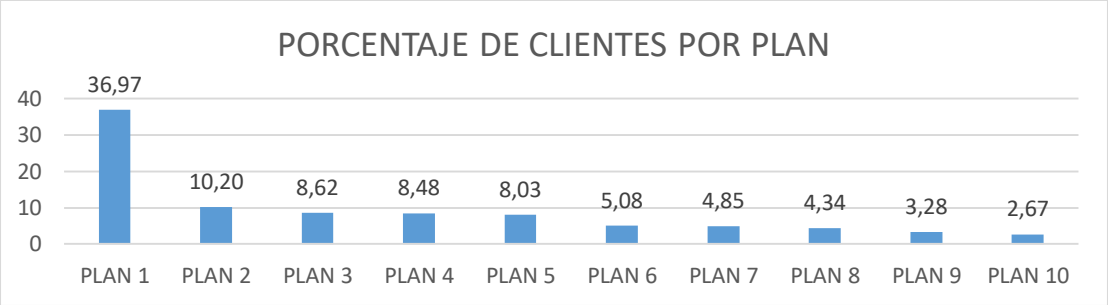
Nota. La tabla muestra el listado existente al momento en cuanto a número de clientes por plan de telefonía móvil.

En la Figura 6 se muestran los planes de datos celulares donde se concentran la mayor cantidad de clientes y son desde el Plan 1 hasta el Plan 10, para efectos de

análisis trabajaremos con el Plan 1[11] que es el más popular y cuenta con la mayor cantidad de clientes en estado activos.

Figura 6.

Porcentaje de clientes por tipo de plan de datos celular (los 10 más representativos).



Nota. La figura representa (en porcentajes de clientes) los 10 planes de datos celulares más representativos. Porcentaje de clientes por tipo de plan de datos celular (los 10 más representativos). (Autoría propia).

Verificación de la calidad de los Datos

Luego de realizada la exploración de los datos hemos podido confirmar que los datos son consistentes, y nos permitirán trabajar para obtener los resultados planteados en los objetivos del presente proyecto. Los datos son de calidad por lo siguiente:

- Los datos no contienen errores, ya que se sometieron a un proceso ETL, es decir se realizó una limpieza de datos, estandarización de datos, modelamiento de datos, etc.

- Los datos a ser ingresados en el data warehouse existente cuentan con validaciones que evitan el ingreso de datos erróneos, eliminando el ruido en el conjunto de datos.

[11] Plan1, es el nombre enmascarado que se le ha colocado al plan de datos celular más usado o con la mayor cantidad de clientes que tiene la empresa de telecomunicaciones objeto de estudio.

- Los valores nulos fueron procesados cuando se realizó la construcción del data warehouse.

- El modelo de base de datos origen cumple con las reglas de normalización de datos, es decir, evita la redundancia, todas las tablas se encuentran relacionadas, no existe valores duplicados, los valores en las diferentes columnas son congruentes en relación al tipo de dato, logrando la integridad de los datos.

Preparación De Los Datos

La fase de preparación de los datos nos ayuda a dar el formato y adecuar los datos que serán usados en las técnicas de minería de datos seleccionadas para trabajar en el presente proyecto.

Una vez realizada la recolección inicial de los datos, se procede a su respectiva preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, que para este caso de estudio son las técnicas de predicción.

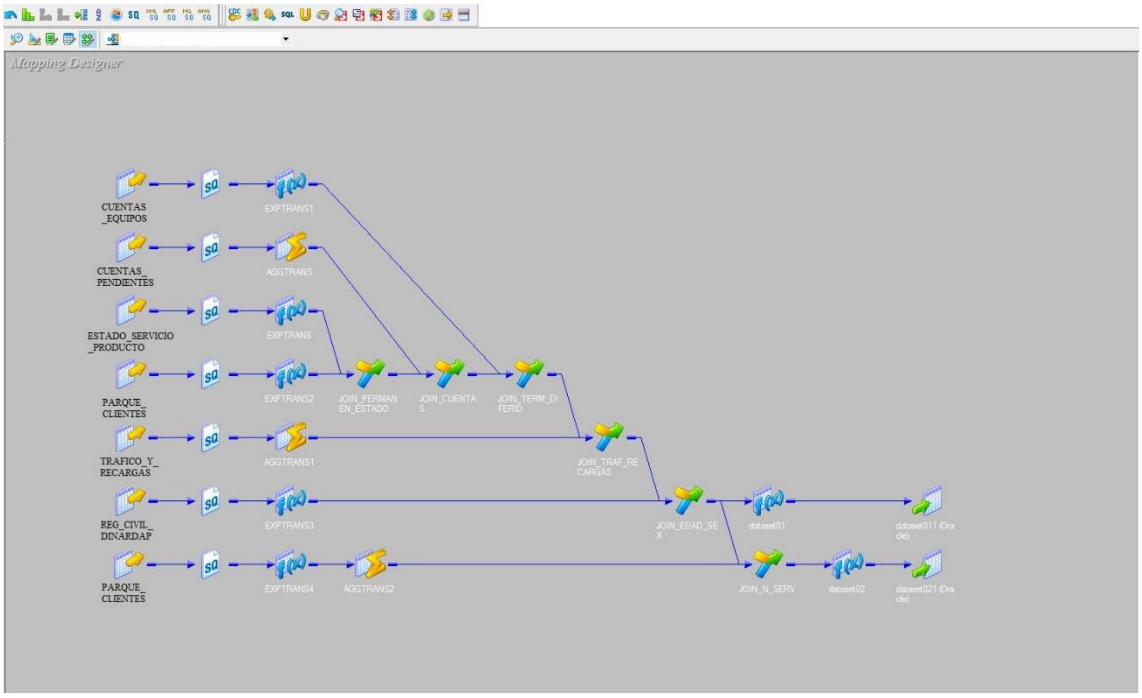
La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza y construcción de los datos, integración de diferentes o similares orígenes de datos y cambios de formato o formateo de los datos. (Gallardo Arancibia, 2009).

Para cumplir con esta fase del proyecto, en lo concerniente a las etapas de Selección de Datos, Limpieza de los Datos y construcción de los datos se procedió a realizar un ETL en la herramienta Informática Power Center, esto se lo puede apreciar en la Figura 7, el hecho o la razón de usar esta herramienta de ETL radica fundamentalmente en que la empresa de telecomunicaciones objeto del presente estudio, cuenta con la mencionada herramienta y existe personal capacitado para el uso de la misma, como resultado de este ETL se obtiene el DataFrame que será quien contiene la data a ser analizada y usada en la herramienta Knime en las etapas restantes

de esta fase que son Integración y Formateo de los Datos, también se usa Knime para la siguiente fase de modelado para los tres (3) modelos a ser evaluados.

Figura 7.

ETL realizado en la herramienta Informática Power Center para la obtención del DataFrame.



Nota. La figura representa el ETL realizado en la herramienta Informática Power Center para obtener el DataFrame. (Autoría propia).

Figura 8.

Cuadrante de Gartner para Herramientas de Minería de Datos.

Nota. La figura representa las herramientas de minería de datos que se encuentran en el cuadrante de Gartner del año 2019. (Piatetsky, 2020).

Informática Power Center es una de las herramientas ETLs (extracción, transformación y carga) más populares en la actualidad, se ha afianzado como una de las herramientas de integración de datos más potentes del mercado y se la utiliza para la construcción de data warehouse empresariales. Por otra parte en la Figura 8 se muestra que Knime es una herramienta que se encuentra bien posicionada dentro del cuadrante de Gartner, KNIME continúa innovando al respaldar las últimas tendencias en aprendizaje automático, incluidas Python, Spark, H2O y otras plataformas de aprendizaje profundo, y fomentando una comunidad de usuarios vibrante y atractiva.

Selección de los Datos

En esta etapa se seleccionan los campos necesarios del Data Warehouse que nos permiten cumplir con los objetivos del negocio planteados en el proyecto. Para esto se analizaron los campos con los que se cuenta en la data de clientes que cancelaron y los que no cancelaron sus servicios contratados con la empresa de telecomunicaciones objeto de estudio. Los mismos se detallan o describen en la Tabla 8.

Tabla 8.

Detalle de los campos seleccionados (tabla origen, descripción y tabla destino).

TABLA ORIGEN	CAMPOS USADOS	DESCRIPCION CAMPOS	TABLA DESTINO
CUENTAS_EQUIPOS	CEDULA_RUC	Cédula o RUC del cliente	DATAFRAME
	SERVICIO	Número que identifica el servicio contratado por el cliente	
	SAL_PEN_EQ-MOV	Saldo pendiente a la fecha por equipo celular adquirido	
CUENTAS_PENDIENTES	CEDULA_RUC	Cédula o RUC del cliente	
	SERVICIO	Número que identifica el servicio contratado por el cliente	
	FACT_PENDIENTES_COD	Información de las facturas pendientes de pago	
ESTADO_SERVICIO_PRODUCTO	CEDULA_RUC	Cédula o RUC del cliente	
	SERVICIO	Número que identifica el servicio contratado por el cliente	
	FECHA_INICIO	Fecha de inicio del contrato del servicio	
	ESTADO_SERVICIO	Estado a la fecha del servicio (activo/inactivo)	

TABLA ORIGEN	CAMPOS USADOS	DESCRIPCION CAMPOS	TABLA DESTINO
	TIPO_PLAN	Nombre del plan adquirido y atado al servicio contratado (Plan1, Plan2, Pan3, etc.)	
	CANCELA_SERV	Identifica o señala si el servicio fue cancelado o no (SI/NO)	
	FECHA_CANCELA	Fecha en la que se produjo o ejecuto la cancelación del servicio	
	CODIGO_ PRODUCTO	Identificador del producto (1= fijo, 2=móvil, 3=internet)	
PARQUE_CLIENTES	CEDULA_RUC	Cédula o RUC del cliente	
	NOMBRES	Primer y segundo nombre del cliente	
	APELLIDOS	Primer y segundo apellido del cliente	
	SEXO	Sexo del cliente	
	ESTADO	Estado del cliente (activo/inactivo)	
	PERMANENCIA	Numero de meses a la fecha que el cliente tiene contratado un servicio con la empresa	
TRAFICO_Y_RECARGAS	CEDULA_RUC	Cédula o RUC del cliente	
	SERVICIO	Número que identifica el servicio contratado por el cliente	
	CONSUMO_MEGAS	Numero de megas consumidos mes a mes	
	CONSUMO_MINUTOS	Numero de minutos consumidos mes a mes	
	CANT_RECARGAS_MES	Numero de recargas	

TABLA ORIGEN	CAMPOS USADOS	DESCRIPCION CAMPOS	TABLA DESTINO
		realizadas mes a mes	
	VALOR_RECARGAS_MES	Valor total de las recargas realizadas mes a mes	
REG_CIVIL_DINARDAP	CEDULA_RUC	Cédula o RUC del ciudadano	
	SEXO	Sexo del ciudadano	
	FECHA_NAC	Fecha de nacimiento del ciudadano	
	ESTADO_CIVIL	Estado civil del ciudadano	
	NOMBRES	Primer y segundo nombre del ciudadano	
	APELLIDOS	Primer y segundo apellido del ciudadano	
	FECHA_DEFUNCION	Fecha de defunción del ciudadano	

Nota. La tabla muestra el listado de los campos usados (extraídos del DWH existente), así como la información de la tabla origen y la tabla destino, así como una breve descripción de cada uno de los campos.

Los campos escogidos / seleccionados son los que tienen relación con los objetivos de la minería de datos planteados en el proyecto.

Limpieza de los Datos

Para la limpieza de los datos se decidió trabajar con la data referente a los clientes que tienen o tuvieron contratado un plan llamado “PLAN 1” ya que corresponde al plan con mayor cantidad de clientes, de un total de 323446 el PLAN 1 corresponde al 36.97% tal y como se puede observar en la Tabla 7 y Figura 6.

En el proceso de limpieza de datos se procedió con la eliminación de registros con valores blancos y nulos, se eliminó registros cuya cédula o RUC no contienen la longitud adecuada.

Construcción de los Datos

En esta etapa se utiliza la herramienta Informática Power Center, la cual nos permite realizar la transformación de los campos necesarios para el desarrollo de la analítica mismos que son almacenados en un DataFrame (archivo Excel).

Atributos Derivados

Los atributos derivados se muestran en la Tabla 9, donde se especifican los campos, descripción del campo y las transformaciones realizadas en el ETL de la herramienta Informática PowerCenter.

Tabla 9.

Transformación de Campos.

CAMPO	DESCRIPCION	TRANSFORMACION
PERMANENCIA	Información mensual de la permanencia o no de un cliente	Numero de meses que un cliente lleva en la empresa
ESTADO	Estado a la fecha de un cliente	Segmentación del estado de un servicio (1=activo / 0=inactivo)
FACTURAS PENDIENTES	Información de las facturas pendientes de pago	Conteo del número de facturas sin pagar o pendientes de pago por cliente y numero_servicio
SAL_PEND_EQUIPO	Saldo pendiente a la fecha por equipo celular adquirido	Segmentación saldo pendiente por equipo celular adquirido (1=si / 0=no)
CONSUMO_MEGAS	Consumo mensual de megas	Promedio mensual del consumo de megas por cliente y numero de servicio
CONSUMO_MINUTOS	Consumo mensual de minutos	Promedio mensual del consumo de minutos por cliente y numero de servicio
CANTIDAD DE RECARGAS	Valor (en dólares) de las recargas realizadas mes a mes	Promedio mensual (en dólares) de las recargas realizadas por cliente y numero de servicio
ESTADO_CAL	Estado a la fecha de un cliente	Segmentación del estado de un servicio (1=activo / 0=inactivo)
NUMERO DE CLIENTES POR PLAN	Número de servicio / número de documento	Conteo del número de clientes por cada plan existente

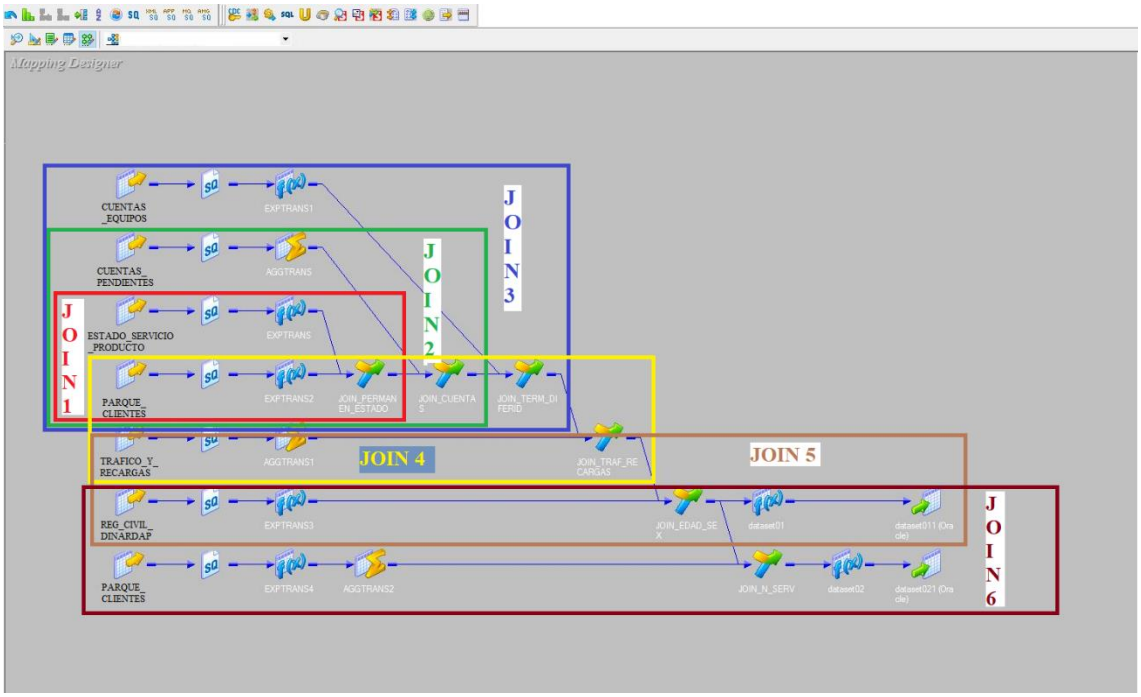
Nota. La tabla muestra el listado de los campos resultantes de transformaciones realizadas, así como la descripción del campo y una explicación de la transformación.

Integración de los Datos

La integración de los datos, consiste en la creación de nuevas estructuras, a partir de los datos seleccionados o datos de origen, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, unión o juntura de tablas, campos o nuevas tablas donde se resumen ciertas características específicas de múltiples registros o de otros campos en nuevas tablas de resumen. En la Figura 9 se muestra que en esta etapa se describen los JOINs realizados con las tablas del DWH existente, necesarios para la construcción del DataSet, insumo de la fase de Modelado.

Figura 9.

ETL Informática Power Center con sus JOINs.



Nota. La figura representa los JOINs realizados en el ETL de la herramienta Informática Power Center para obtener el DataFrame. (Autoría propia).

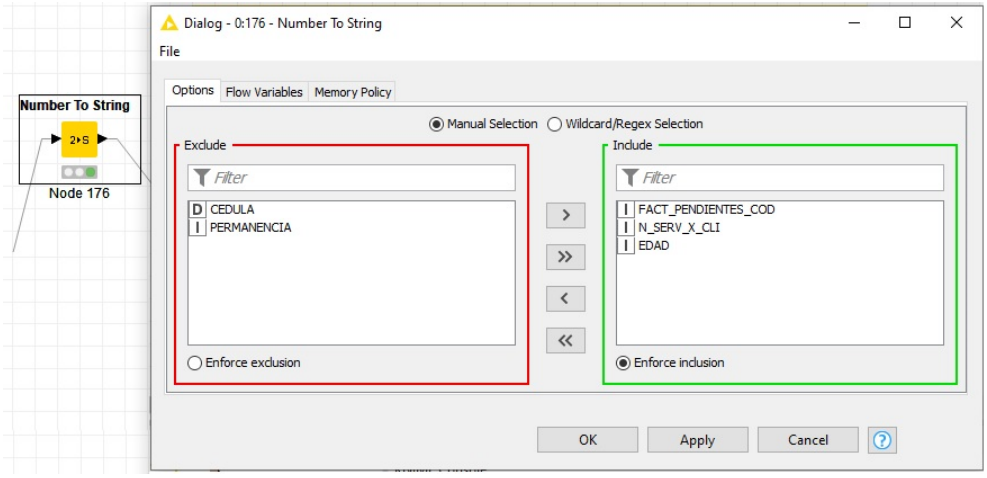
En la figura 9 se describen los JOINs entre las tablas del DWH existente, JOINs que permitieron llegar a formar el DataSet que será insumo de la fase de Modelado. El JOIN 1 muestra la juntura realizada entre la tabla PARQUE_CLIENTES y la tabla ESTADO_SERVICIO_PRODUCTO, este resultado se junta con la tabla CUENTAS_PENDIENTES por medio del JOIN 2, este resultado se junta con la tabla CUENTAS_EQUIPOS por medio del JOIN 3, este resultado se junta con la tabla TRAFICO_Y_RECARGAS por medio del JOIN 4, este resultado se junta con la tabla REG_CIVIL_DINARDAP por medio del JOIN 5, y finalmente parta obtener el DataSet final con todos los campos necesarios para elaborar nuestros modelos se junta este resultado con la tabla PARQUE_CLIENTES.

Formateo de los Datos

Para el proceso de formateo de datos se usó el nodo “Number to String” para convertir a formato tipo cadena, debido a que estos campos se encontraban en formato numérico en el DataFrame extraído esto se puede visualizar en la Figura 10.

Figura 10.

Nodo Number to String



Nota. La figura representa el nodo Number to String de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Modelado

En la fase de modelado se elige las técnicas de minería de datos adecuadas para cumplir con los objetivos de la minería de datos, luego se genera el plan de prueba, a continuación, se aplica las técnicas de minería de datos escogidas sobre los datos para construir el modelo y finalmente se evalúa el modelo para determinar si cumple con los criterios de éxito.

Selección Técnica de Modelado

Se construyeron tres modelos, basados en las técnicas de minería de datos descritas, y son:

- Árboles de Decisión
- Naive Bayes
- Redes Neuronales

Árboles de decisión: Es considerada una de las técnicas más eficaces para la clasificación supervisada ya que combina técnicas matemáticas y computacionales para ayudar a la descripción, la categorización y la generalización de un conjunto de datos y es muy fácil de entender e interpretar debido a la su forma gráfica de presentar los resultados utilizando lógica booleana. Esta técnica es capaz de manejar datos numéricos al igual que datos categorizados. Pueden ser analizados grandes cantidades de datos utilizando recursos informáticos básicos o estándares en un plazo razonable de tiempo.

Naive Bayes: Este algoritmo que se basa en probabilidades condicionadas con datos conocidos, es rápido y muy fácil de usar, motivo por el cual es uno de los modelos de clasificación más usados. Se basa en calcular probabilidades de datos conocidos y de acuerdo a los resultados y una fórmula, se puede calcular la probabilidad de que la entrada sea de una u otra clase con lo cual proporciona sus resultados.

Redes Neuronales: Es o son una adaptación artificial (semejante) de lo que hace el cerebro humano, las funciones son muy parecidas o semejantes a las conexiones neuronales reales: capaces de aprender en base a la experiencia, generalizar casos anteriores a nuevos casos, identifica características esenciales a partir de entradas que representan información irrelevante etc.

Los algoritmos que se acaba de describir fueron escogidos tomando en consideración su popularidad, rapidez, fácil uso y robustez. Además, estos algoritmos permiten manejar datos nominales y no nominales en sus implementaciones y para el presente caso de estudio planteado el atributo a predecir es nominal puesto que es: 1 => “SI_CANCELA” o 0 => “NO_CANCELA” dependiendo si el cliente cancela o no su o sus servicios contratados con la empresa de telecomunicaciones.

Creación del modelo de Minería de Datos

Se construyó 3 diferentes modelos para cumplir con los objetivos planteados en el presente proyecto, los cuales se basan en la generación del modelo predictivo, orientado a predecir si el cliente cancelará alguno de los servicios contratados con la empresa de Telecomunicaciones. De acuerdo a la herramienta KNIME, escogida para la aplicación de los modelos, se realizan las diferentes configuraciones de los parámetros establecidas para cada uno de ellos.

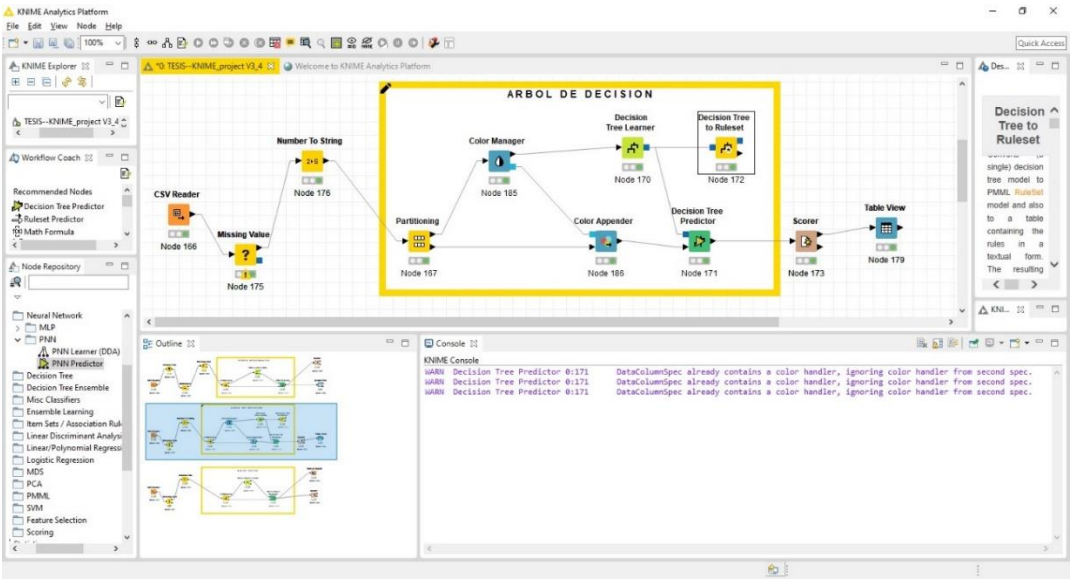
La construcción de los modelos describe cada una de las técnicas seleccionadas: Árboles de Decisión, Naive Bayes y Redes Neuronales.

Árboles de Decisión

La aplicación de la técnica de árboles de decisión, nos ayuda a identificar las variables más relevantes o con mayor grado de correlación, y que son las que intervienen en la decisión de si una persona o cliente cancelara o no algún servicio que actualmente tenga contratado con la empresa de telecomunicaciones objeto de estudio, se define como el atributo clase al campo cancela_serv (variable a predecir). La Figura 11 muestra el modelo de árbol de decisión realizado en la herramienta Knime.

Figura 11.

Modelo Árbol de Decisión



Nota. La figura representa el modelo de Árbol de Decisión construido o desarrollado en la herramienta Knime para el presente trabajo de estudio.

La técnica de árboles de decisión nos permite visualizar de mejor manera la clasificación de los datos con la ayuda de un esquema gráfico, el cual permite conocer el modelo predictivo.

A continuación se describen los nodos que se utilizó para la generación del modelo predictivo del árbol de decisión en la herramienta Knime:

- **Nodo CSV Reader**

Figura 12.

Nodo CSV Reader

Nota. La figura representa el nodo CSV Reader de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para abrir y/o leer archivos CSV. Este nodo nos permite obtener todos los campos, variables involucradas o que forman parte del proceso de minería de datos y que son el resultado del proceso ETL en el DWH existente y que nos dio como resultado el DataFrame con el cual hemos trabajado.

- **Missing Value**

Figura 13.

Nodo Missing Value

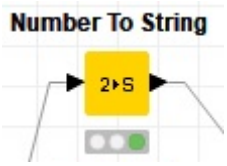
Nota. La figura representa el nodo Missing Value de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para manejar o tratar los valores perdidos o inexistentes de la tabla o archivo de entrada. Este nodo nos permite dar un tratamiento específico a los registros del DataFrame en base a alguna característica de un campo dentro de nuestro archivo de entrada.

- **Number to String**

Figura 14.

Nodo Number to String

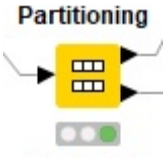


Nota. La figura representa el nodo Number to String de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para convertir números en cadenas. Este nodo nos permite cambiar el tipo de dato de un campo específico del DataFrame de número a cadena de texto.

- **Partitioning**

Figura 15.
Nodo Partitioning

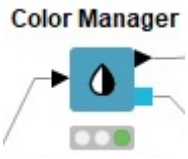


Nota. La figura representa el nodo Partitioning de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para partir o dividir la data de entrada en entrenamiento y prueba. Este nodo nos permite configurar los porcentajes de prueba y entrenamiento del archivo de entrada, en nuestro caso el archivo de entrada DataFrame posee 11395 registros y en el nodo partitioning hemos configurado un 80% para data de entrenamiento del modelo que corresponde a 95516 registros y un 20% para probar o validar el modelo propuesto, y que corresponde a 23879.

- **Color Manager**

Figura 16.
Nodo Color Manager

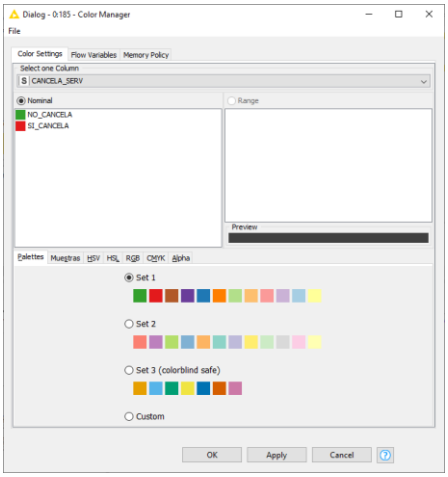


Nota. La figura representa el nodo Color Manager de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para asignar un color especifico a un determinado campo tipo cadena, esto con la finalidad de mejorar la interpretación de la respuesta de uno o varios campos del archivo de entrada. Para efectos de nuestro análisis y construcción del modelo se configuro dos colores para las dos posibles respuestas del campo cancela_serv, tal y como se muestra en la Figura 17.

Figura 17.

Configuración nodo Color Manager

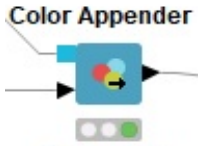


Nota. La figura representa la configuración del nodo Color Manager de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Color Appender**

Figura 18.

Nodo Color Appender

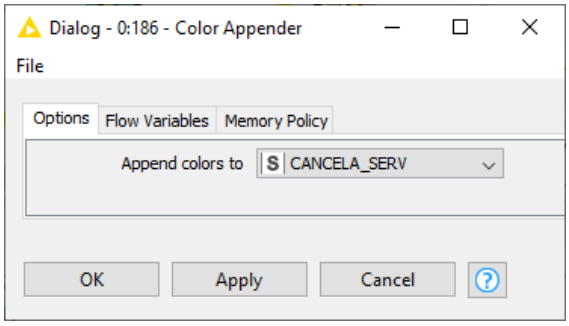


Nota. La figura representa el nodo Color Appender de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para agregar los colores configurados a las columnas de la data de entrada en el nodo color manager hacia el nodo color appender. Para efectos de nuestro análisis y tal como ya fueron configurados los colores en el nodo color manager, en este caso se realizó la configuración de colores para la columna CANCELA_SERV, tal y como se muestra en la figura 19.

Figura 19.

Configuración nodo Color Appender



Nota. La figura representa la configuración del nodo Color Appender de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Decision Tree Learner**

Figura 20.

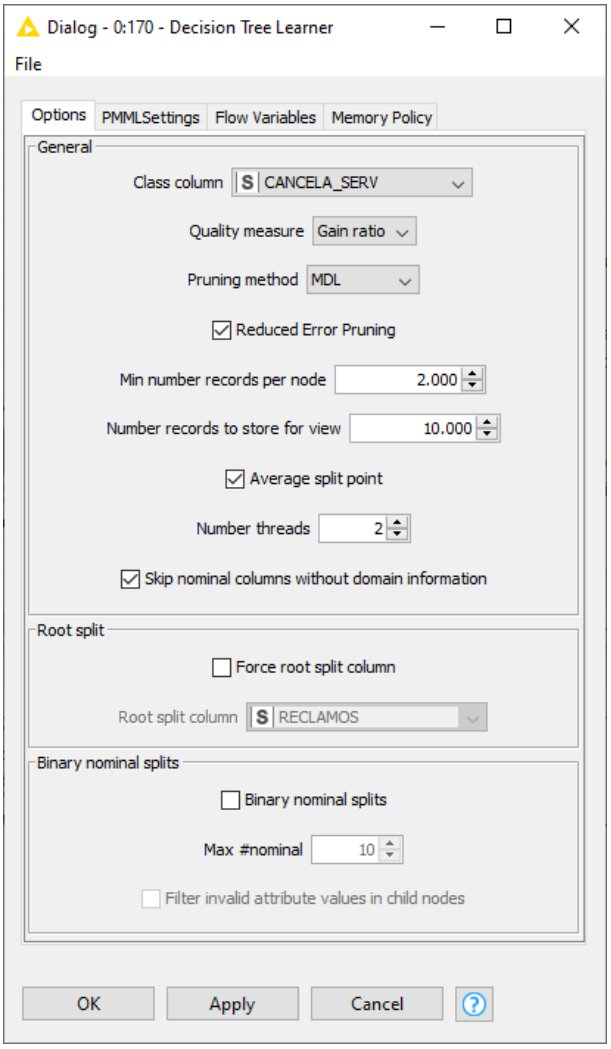
Nodo Decision Tree Learner

Nota. La figura representa el nodo Decision Tree Learner de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para el aprendizaje del árbol de decisión. Para nuestro caso de estudio se configuro en la sección clase a la columna o variable CANCELA_SERV, en la medida de calidad se configuro Gain ratio, en el método de poda se seleccionó MDL además de chequear la opción de Reduce Error Pruning, en el número mínimo de registros por nodo se configuro 2.000, en el número de registros a almacenar para la vista se configuró 10.000, entre otras configuraciones, a continuación en la Figura 21 se mostrara todas las configuraciones realizadas en el presente nodo.

Figura 21.

Configuración nodo Decision Tree Learner

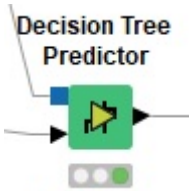


Nota. La figura representa la configuración del nodo Decision Tree Learner de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Decision Tree Predictor**

Figura 22.

Nodo Decision Tree Predictor

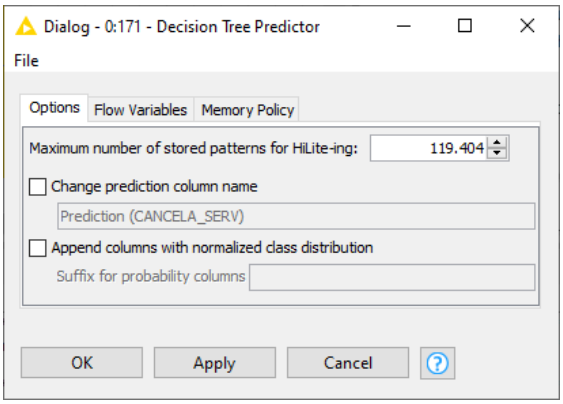


Nota. La figura representa el nodo Decision Tree Predictor de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para predecir la data de prueba en base al entrenamiento dado en el nodo Decision Tree Learner. Para nuestro caso de estudio específicamente este nodo nos permite predecir el valor de clase CANCELA_SERV para nuevos patrones, y este resultado se mostrara o visualizara en la columna denominada Prediction (CANCELA_SERV). En la Figura 23 se puede observar la pantalla de configuración del Nodo Decision Tree Predictor.

Figura 23.

Configuración Nodo Decision Tree Predictor

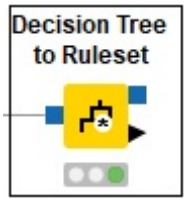


Nota. La figura representa la configuración del nodo Decision Tree Predictor de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Decision Tree to Ruleset**

Figura 24. Nodo Decisión Tree to Ruleset

Nodo Decisión Tree to Ruleset

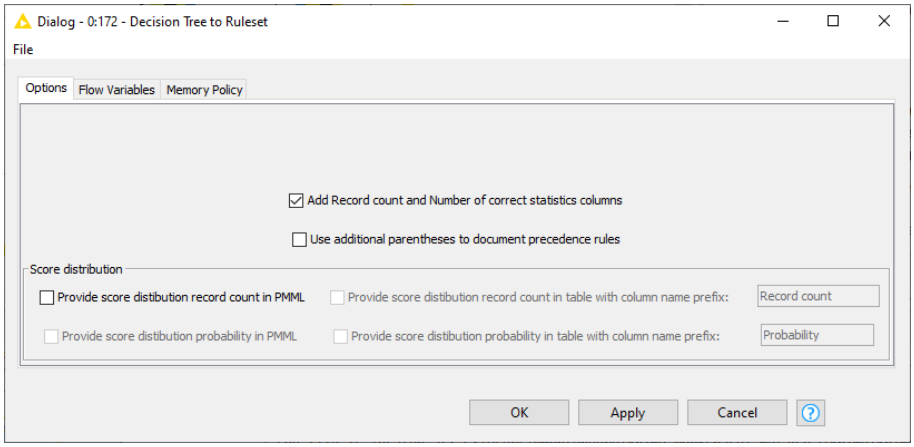


Nota. La figura representa el nodo Decisión Tree to Ruleset de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para visualizar de forma textual las reglas resultantes que se aplican en la ejecución del modelo construido. A continuación en la Figura 25 se muestra la configuración realizada en este nodo para nuestro caso de estudio y en la Figura 26 se observa las reglas resultantes de nuestro modelo de árbol de decisión.

Figura 25.

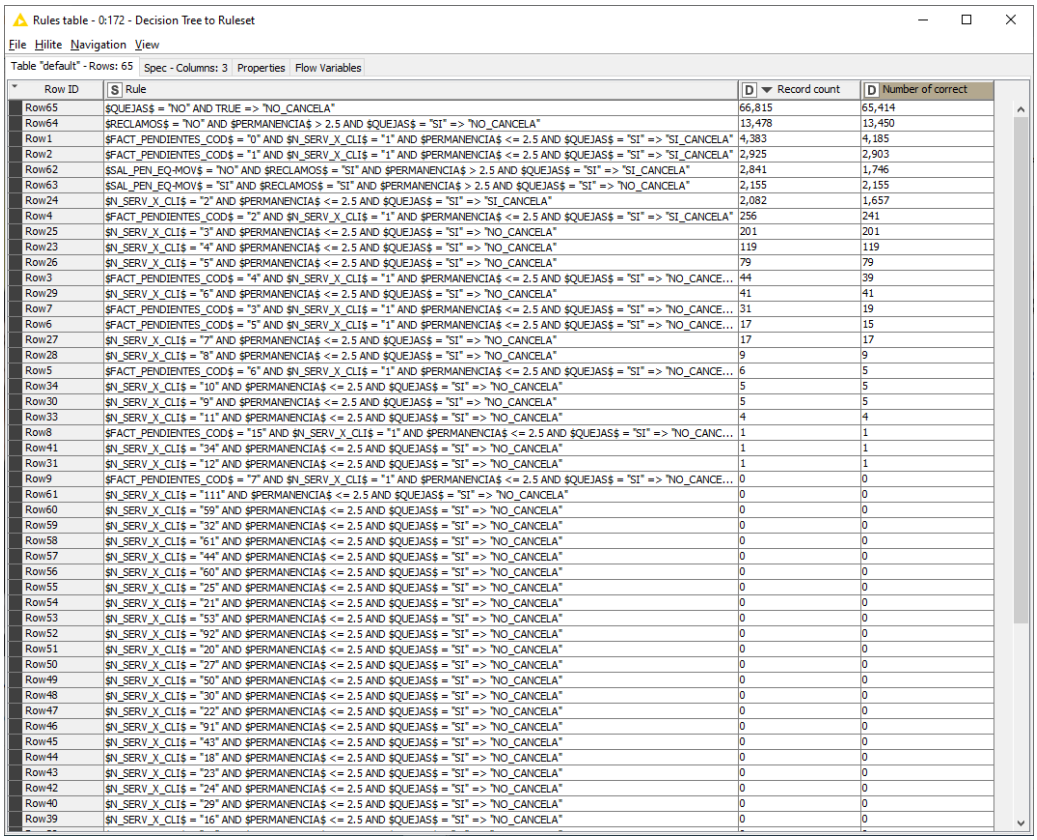
Configuración Nodo Decisión Tree to Ruleset



Nota. La figura representa la configuración del nodo Tree to Ruleset de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Figura 26.

Tabla de reglas resultantes del nodo Decisión Tree to Ruleset (orden descendente en base al campoRecord count).



Rules table - 0:172 - Decision Tree to Ruleset

File Hilite Navigation View

Table "default" - Rows: 65 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	Rule	Record count	Number of correct
Row65	$QUEJAS$ = "NO" AND TRUE => "NO_CANCELA"	66,815	65,414
Row64	$RECLAMOS$ = "NO" AND $PERMANENCIA$ > 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	13,478	13,450
Row1	$FACT_PENDIENTES_COD$ = "0" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "SI_CANCELA"	4,383	4,185
Row2	$FACT_PENDIENTES_COD$ = "1" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "SI_CANCELA"	2,925	2,903
Row62	SAL_PEN_EQ-MOV = "NO" AND $RECLAMOS$ = "SI" AND $PERMANENCIA$ > 2.5 AND $QUEJAS$ = "SI" => "SI_CANCELA"	2,841	1,746
Row63	SAL_PEN_EQ-MOV = "SI" AND $RECLAMOS$ = "SI" AND $PERMANENCIA$ > 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	2,155	2,155
Row24	$N_SERV_X_CLI$ = "2" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "SI_CANCELA"	2,082	1,657
Row4	$FACT_PENDIENTES_COD$ = "2" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "SI_CANCELA"	256	241
Row25	$N_SERV_X_CLI$ = "3" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	201	201
Row23	$N_SERV_X_CLI$ = "4" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	119	119
Row26	$N_SERV_X_CLI$ = "5" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	79	79
Row3	$FACT_PENDIENTES_COD$ = "4" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCE...	44	39
Row29	$N_SERV_X_CLI$ = "6" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	41	41
Row7	$FACT_PENDIENTES_COD$ = "3" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCE...	31	19
Row6	$FACT_PENDIENTES_COD$ = "5" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCE...	17	15
Row27	$N_SERV_X_CLI$ = "7" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	17	17
Row28	$N_SERV_X_CLI$ = "8" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	9	9
Row5	$FACT_PENDIENTES_COD$ = "6" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCE...	6	5
Row34	$N_SERV_X_CLI$ = "10" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	5	5
Row30	$N_SERV_X_CLI$ = "9" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	5	5
Row33	$N_SERV_X_CLI$ = "11" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	4	4
Row8	$FACT_PENDIENTES_COD$ = "15" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANC...	1	1
Row41	$N_SERV_X_CLI$ = "34" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	1	1
Row31	$N_SERV_X_CLI$ = "12" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	1	1
Row9	$FACT_PENDIENTES_COD$ = "7" AND $N_SERV_X_CLI$ = "1" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCE...	0	0
Row61	$N_SERV_X_CLI$ = "111" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row60	$N_SERV_X_CLI$ = "59" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row59	$N_SERV_X_CLI$ = "32" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row58	$N_SERV_X_CLI$ = "61" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row57	$N_SERV_X_CLI$ = "44" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row56	$N_SERV_X_CLI$ = "60" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row55	$N_SERV_X_CLI$ = "25" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row54	$N_SERV_X_CLI$ = "21" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row53	$N_SERV_X_CLI$ = "53" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row52	$N_SERV_X_CLI$ = "92" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row51	$N_SERV_X_CLI$ = "20" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row50	$N_SERV_X_CLI$ = "27" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row49	$N_SERV_X_CLI$ = "50" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row48	$N_SERV_X_CLI$ = "30" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row47	$N_SERV_X_CLI$ = "22" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row46	$N_SERV_X_CLI$ = "91" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row45	$N_SERV_X_CLI$ = "43" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row44	$N_SERV_X_CLI$ = "18" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row43	$N_SERV_X_CLI$ = "23" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row42	$N_SERV_X_CLI$ = "24" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row40	$N_SERV_X_CLI$ = "29" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0
Row39	$N_SERV_X_CLI$ = "16" AND $PERMANENCIA$ <= 2.5 AND $QUEJAS$ = "SI" => "NO_CANCELA"	0	0

Nota. La figura representa la tabla de reglas resultantes de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Scorer**

Figura 27.

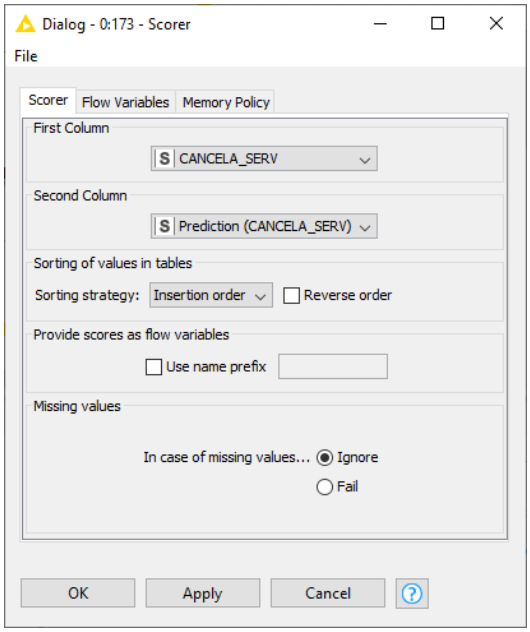
Nodo Scorer

Nota. La figura representa el nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para visualizar a la salida la matriz de confusión con el número de coincidencias en cada celda, esto como resultado al comparar el campo de entrenamiento con el campo predicho, este resultado permite evaluar que tan bueno o no resulta ser el modelo. Para nuestro caso específico mostrará el resultado al evaluar los campos CANCEL_SERV con el campo Prediction (CANCELA_SERV) y se podrá interpretar que tan confiable o preciso es nuestro modelo de árbol de decisión. En la figura 28 se muestra la configuración de nodo y en la Figura 29 se observa la matriz de confusión resultante.

Figura 28

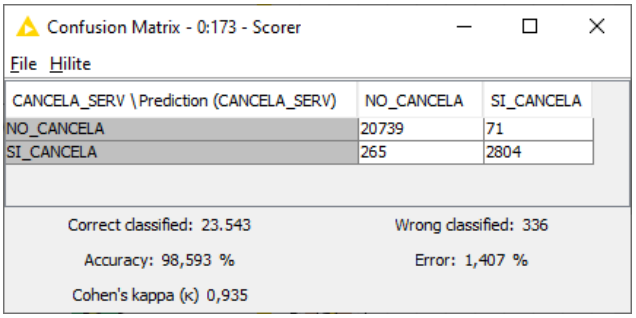
Configuración nodo Scorer



Nota. La figura representa la configuración del nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Figura 29.

Resultado Matrix de confusión (nodo Scorer)



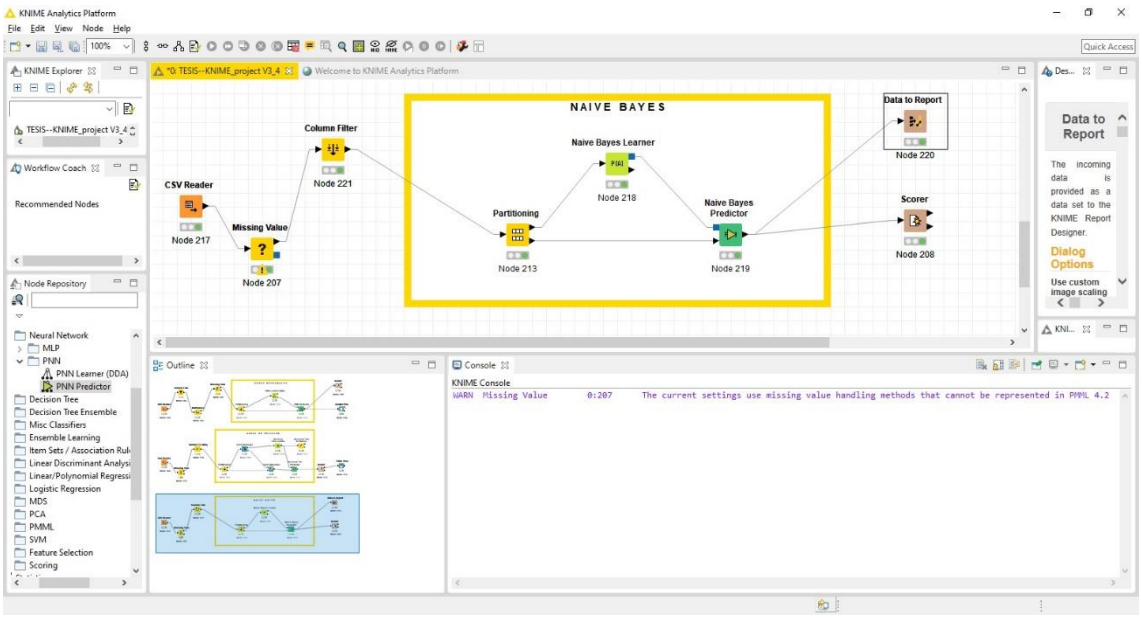
Nota. La figura representa el resultado de la matriz de confusión del nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Naive Bayes

La aplicación de la técnica Naive Bayes, nos ayuda a identificar las variables más relevantes o con mayor grado de correlación, y que son las que intervienen en la decisión de si una persona o cliente cancelará o no algún servicio que actualmente tenga contratado con la empresa de telecomunicaciones objeto de estudio, se define como el atributo clase al campo cancela_serv (variable a predecir). La Figura 30 muestra el modelo Naive Bayes realizado en la herramienta Knime.

Figura 30.

Modelo Naive Bayes



Nota. La figura representa el modelo Naive Bayes construido o desarrollado en la herramienta Knime para el presente trabajo de estudio.

A continuación se describen los nodos que se utilizó para la generación del modelo predictivo de Naive Bayes:

- **Nodo CSV Reader**

Figura 31.

Nodo CSV Reader

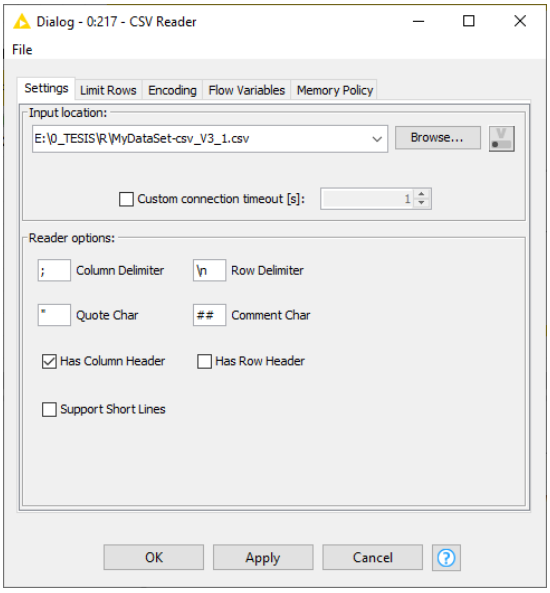
Nota. La figura representa la configuración del nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para abrir y/o leer archivos CSV. Este nodo nos permite obtener todos los campos, variables involucradas o que forman parte del proceso de minería de datos y que son el resultado del proceso ETL en el DWH existente y que nos dio como resultado el DataFrame con el cual hemos trabajado. En la Figura 32 se

muestran las configuraciones realizadas en el Nodo CSV Reader para nuestro caso de estudio específico.

Figura 32.

Configuración nodo CSV Reader



Nota. La figura representa la configuración del nodo CSV Reader de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Missing Value**

Figura 33.
Nodo Missing Value

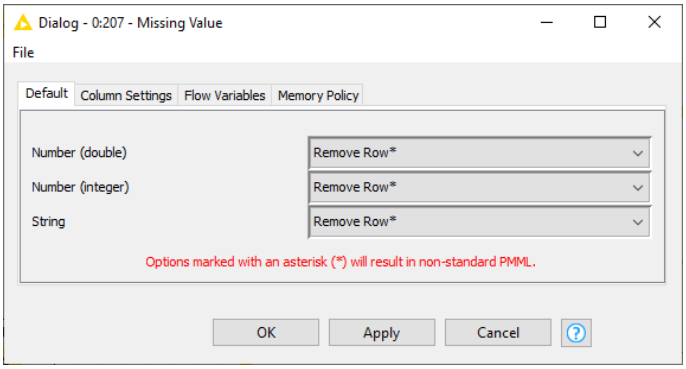
Nota. La figura representa la configuración del nodo CSV Reader de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para manejar o tratar los valores perdidos o inexistentes de la tabla o archivo de entrada. Este nodo nos permite dar un tratamiento específico a los registros del DataFrame en base a alguna característica de un campo dentro de

nuestro archivo de entrada. En la figura 34 se muestran las configuraciones realizadas en el Nodo Missing Value para nuestro caso de estudio específico.

Figura 34.

Configuración Nodo Missing Value



Nota. La figura representa la configuración del nodo Missing Value de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Column Filter**

Figura 35.

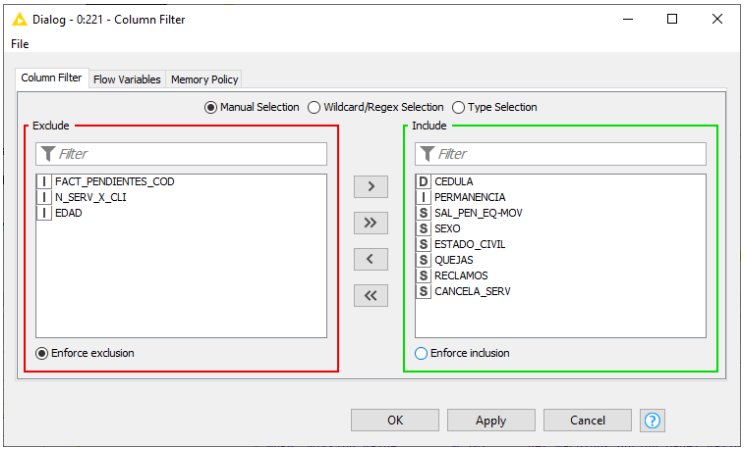
Nodo Column Filter

Nota. La figura representa el nodo Column Filter de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para convertir números en cadenas. Este nodo nos permite cambiar el tipo de dato de un campo específico del DataFrame de número a cadena de texto. En la figura 36 se muestran las configuraciones realizadas en el Nodo Column Filter para nuestro caso de estudio específico.

Figura 36.

Configuración nodo Column Filter

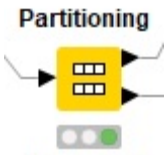


Nota. La figura representa la configuración del nodo Column Filter de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Partitioning**

Figura 37.

Nodo Partitioning

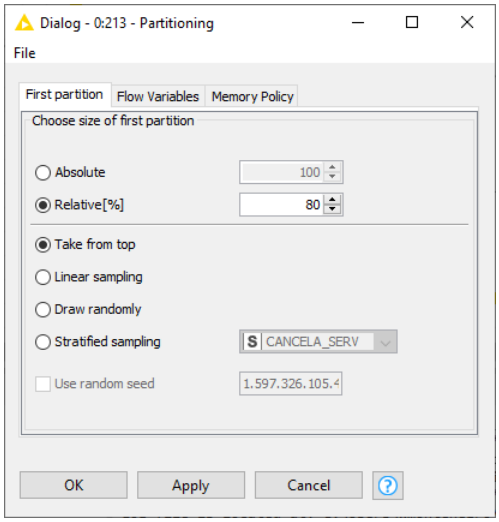


Nota. La figura representa el nodo Partitioning de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para partir o dividir la data de entrada en entrenamiento y prueba. Este nodo nos permite configurar los porcentajes de prueba y entrenamiento del archivo de entrada, en nuestro caso el archivo de entrada DataFrame posee 11395 registros y en el nodo partitioning hemos configurado un 80% para data de entrenamiento del modelo que corresponde a 95516 registros y un 20% para probar o validar el modelo propuesto, y que corresponde a 23879. En la Figura 38 se muestran las configuraciones realizadas en el Nodo Partitioning para nuestro caso de estudio específico.

Figura 38.

Configuración nodo Partitioning

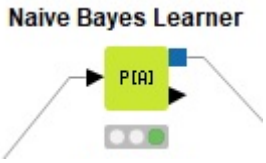


Nota. La figura representa la configuración del nodo Partitioning de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Naive Bayes Learner**

Figura 39.

Nodo Naibe Bayes Learner

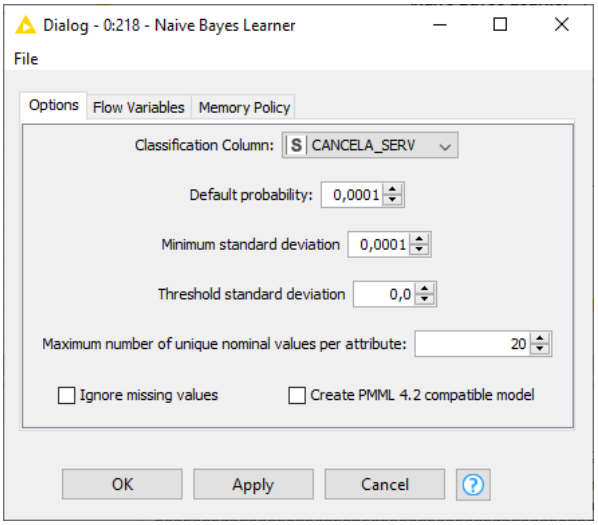


Nota. La figura representa el nodo Naibe Bayes Learner de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para el aprendizaje de Naive Bayes. Para nuestro caso de estudio se configuro en la sección classification column o variable CANCELA_SERV, en Default probability se configuro el valor de 0.0001, entre otras configuraciones. En la Figura 40 se muestran las configuraciones realizadas en el Nodo Naive Bayes Learner para nuestro caso de estudio específico.

Figura 40.

Configuración nodo Naibe Bayes Learner

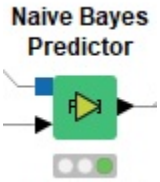


Nota. La figura representa la configuración del nodo Naibe Bayes Learner de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Naive Bayes Predictor**

Figura 41.

Nodo Naive Bayes Predictor



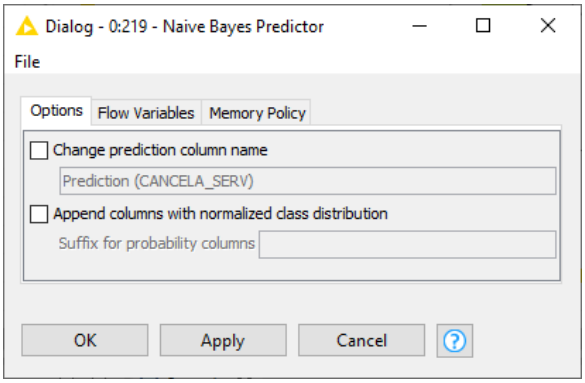
Nota. La figura representa el nodo Naive Bayes Predictor de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para predecir la data de prueba en base al entrenamiento dado en el nodo Naive Bayes Learner. Para nuestro caso de estudio específicamente este nodo nos permite predecir el valor de la columna CANCELA_SERV para nuevos patrones, Y este resultado se mostrara o visualizara en la columna denominada

Prediction (CANCELA_SERV). En la figura 42 se muestran las configuraciones realizadas en el Nodo Naive Bayes Predictor para nuestro caso de estudio específico.

Figura 42.

Configuración nodo Naive Bayes Predictor



Nota. La figura representa la configuración del nodo Naive Bayes Predictor de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Scorer**

Figura 43.

Nodo Scorer

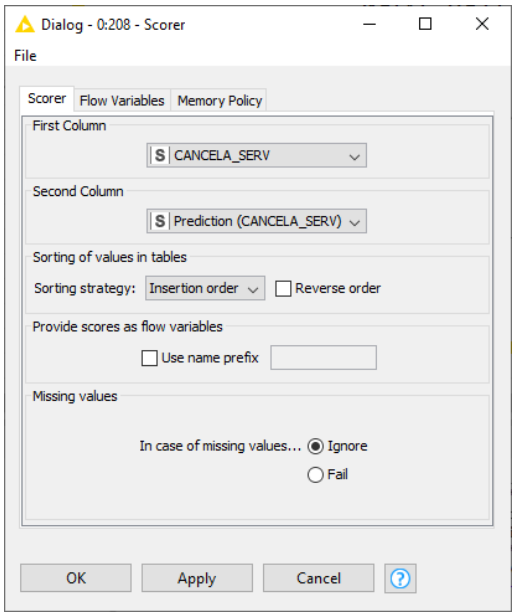
Nota. La figura representa el nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para visualizar a la salida la matriz de confusión con el número de coincidencias en cada celda, esto como resultado al comparar el campo de entrenamiento con el campo predicho, este resultado permite evaluar que tan bueno

o no resulta ser el modelo. Para nuestro caso específico mostrará el resultado al evaluar los registros de la columna CANCEL_SERV con el campo Prediction (CANCELA_SERV) y se podrá interpretar que tan confiable o preciso es nuestro modelo Naive Bayes. En la Figura 44 se muestra la configuración de nodo y en la Figura 45 se observa la matriz de confusión resultante.

Figura 44.

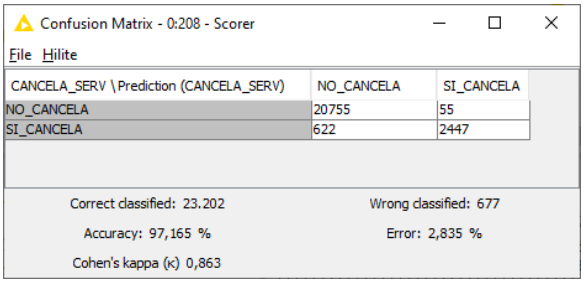
Configuración nodo Scorer



Nota. La figura representa la configuración del nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Figura 45.

Matriz de confusión nodo Scorer



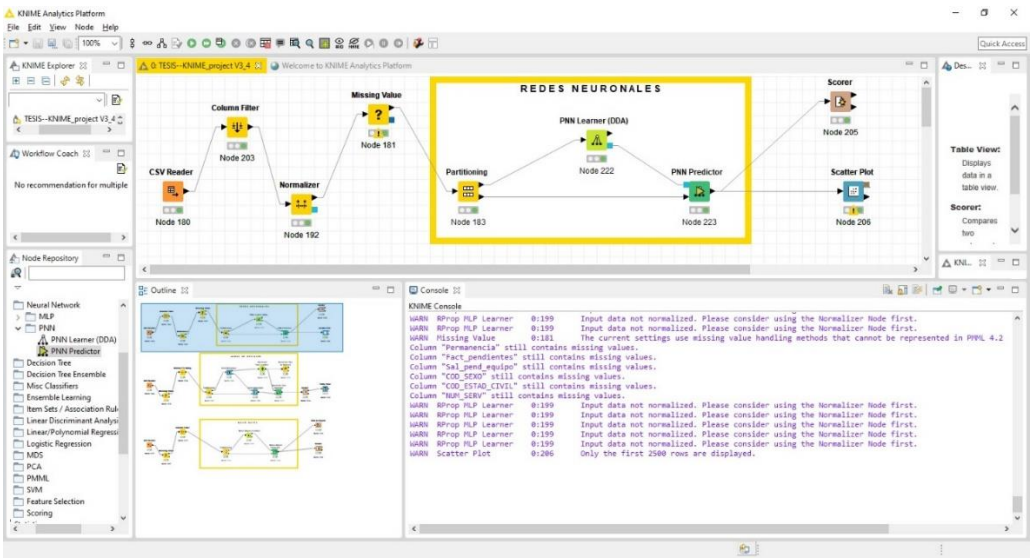
Nota. La figura representa el resultado de la matriz de confusión del nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Redes Neuronales

La aplicación de la técnica de Redes Neuronales, nos ayuda a identificar las variables más relevantes o con mayor grado de correlación, y que son las que intervienen en la decisión de si una persona o cliente cancelara o no algún servicio que actualmente tenga contratado con la empresa de telecomunicaciones objeto de estudio, se define como el atributo clase al campo cancela_serv (variable a predecir). La Figura 46 muestra el modelo de Redes Neuronales realizado en la herramienta Knime.

Figura 46.

Modelo Redes Neuronales



Nota. La figura representa el modelo de Redes Neuronales construido o desarrollado en la herramienta Knime para el presente trabajo de estudio.

A continuación se describen los nodos que se utilizó para la generación del modelo predictivo de redes neuronales:

- **Nodo CSV Reader**

Figura 47.

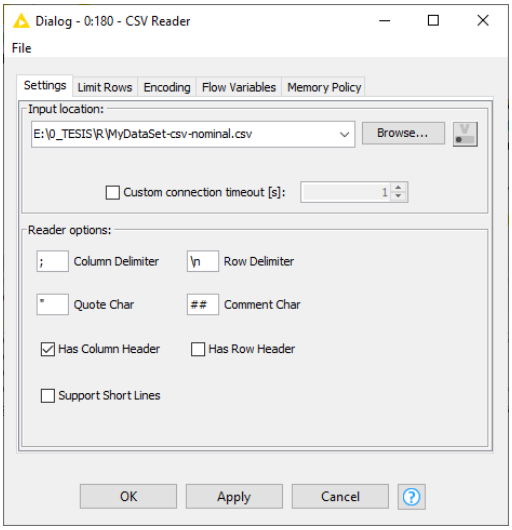
Nodo CSV Reader

Nota. La figura representa el nodo *CSV Reader* de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para abrir y/o leer archivos CSV. Este nodo nos permite obtener todos los campos, variables involucradas o que forman parte del proceso de minería de datos y que son el resultado del proceso ETL en el DWH existente y que nos dio como resultado el DataFrame con el cual hemos trabajado. En la Figura 48 se muestran las configuraciones realizadas en el Nodo CSV Reader para nuestro caso de estudio específico.

Figura 48.

Configuración nodo CSV Reader



Nota. La figura representa la configuración del nodo CSV Reader de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Missing Value**

Figura 49.

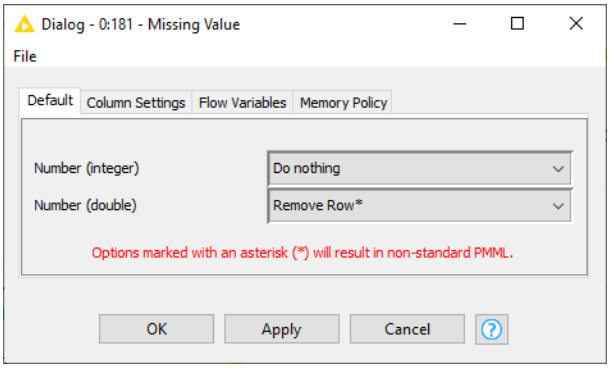
Nodo Missing Value

Nota. La figura representa el nodo Missing Value de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para manejar o tratar los valores perdidos o inexistentes de la tabla o archivo de entrada. Este nodo nos permite dar un tratamiento específico a los registros del DataFrame en base a alguna característica de un campo dentro de nuestro archivo de entrada. En la Figura 50 se muestran las configuraciones realizadas en el Nodo Missing Value para nuestro caso de estudio específico.

Figura 50.

Configuración Nodo Missing Value

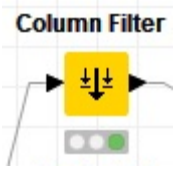


Nota. La figura representa la configuración del nodo Missing Value de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Column Filter**

Figura 51.

Nodo Column Filter

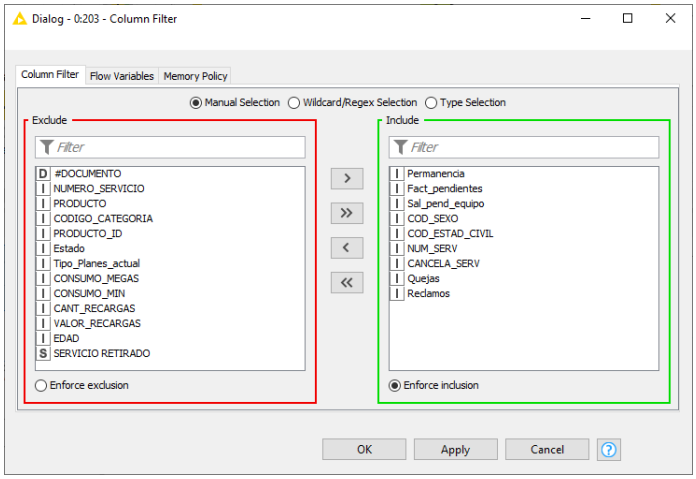


Nota. La figura representa el nodo Column Filter de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para convertir números en cadenas. Este nodo nos permite cambiar el tipo de dato de un campo específico del DataFrame de número a cadena de texto. En la figura 52 se muestran las configuraciones realizadas en el Nodo Column Filter para nuestro caso de estudio específico.

Figura 52.

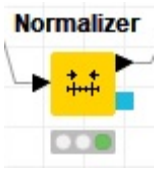
Configuración Nodo Column Filter



- **Normalizer**

Figura 53.

Nodo Normalizer

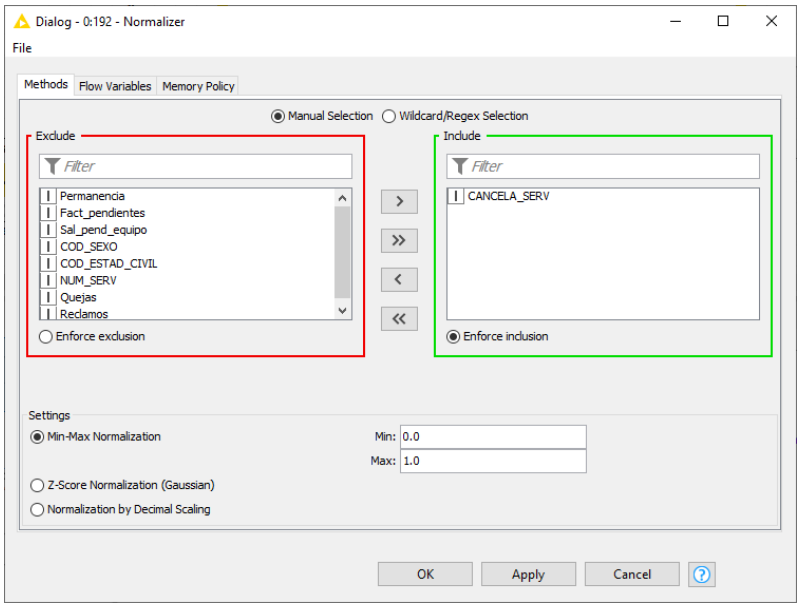


Nota. La figura representa el nodo Normalizer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para normalizar los valores de todas las columnas numéricas (se debe o puede seleccionar las columnas o campos que se desea normalizar sus valores). Este nodo nos permite normalizar los valores de un campo específico del DataFrame. En la Figura 54 se muestran las configuraciones realizadas en el Nodo Normalizer para nuestro caso de estudio específico.

Figura 54.

Configuración nodo Normalizer

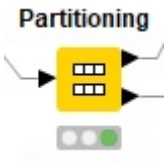


Nota. La figura representa la configuración del nodo Normalizer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Partitioning**

Figura 55.

Nodo Partitioning

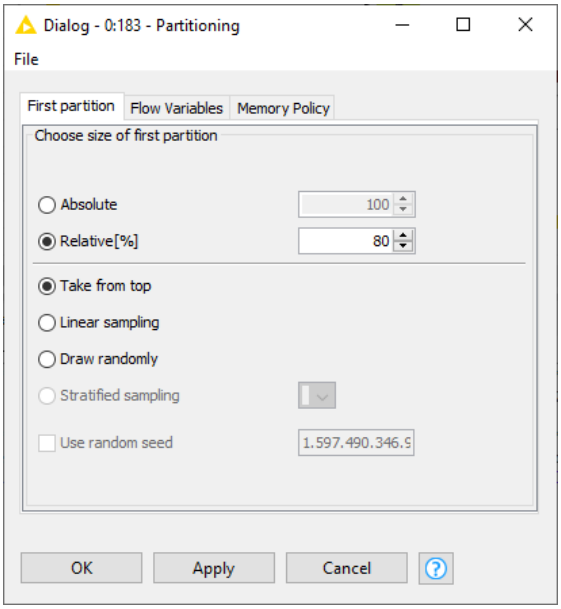


Nota. La figura representa el nodo Partitioning de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para partir o dividir la data de entrada en entrenamiento y prueba. Este nodo nos permite configurar los porcentajes de prueba y entrenamiento del archivo de entrada, en nuestro caso el archivo de entrada DataFrame posee 11395 registros y en el nodo partitioning hemos configurado un 80% para data de entrenamiento del modelo que corresponde a 95516 registros y un 20% para probar o validar el modelo propuesto, y que corresponde a 23879. En la Figura 56 se muestran las configuraciones realizadas en el Nodo Partitioning para nuestro caso de estudio específico.

Figura 56.

Configuración nodo Partitioning

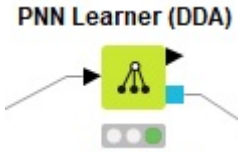


Nota. La figura representa la configuración del nodo Partitioning de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **PNN Learner (DDA)**

Figura 57.

Nodo PNN Learner (DDA)



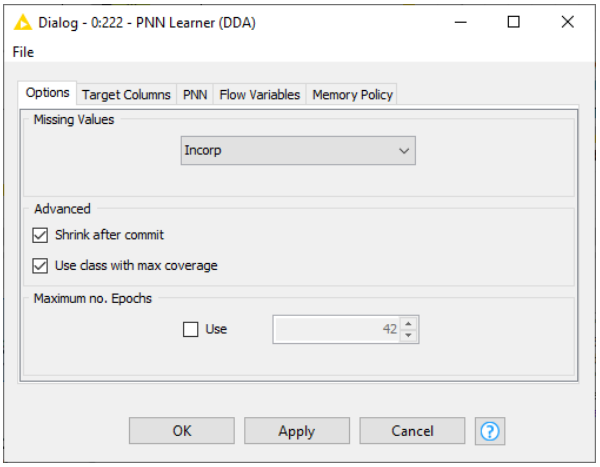
Nota. La figura representa el nodo PNN Learner (DDA) de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para el aprendizaje de Redes Neuronales. Para nuestro caso de estudio se configuro en la sección Target Columns la columna o variable

CANCELA_SERV, en Missing Values se configuro Incorp, entre otras configuraciones. En la Figura 58 se muestran las configuraciones realizadas en el Nodo PNN Learner (DDA) para nuestro caso de estudio específico.

Figura 58.

Configuración Nodo PNN Learner (DDA)

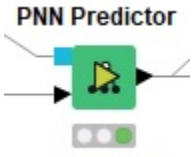


Nota. La figura representa la configuración del nodo PNN Learner (DDA) de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **PNN Predictor**

Figura 59.

Nodo PNN Predictor

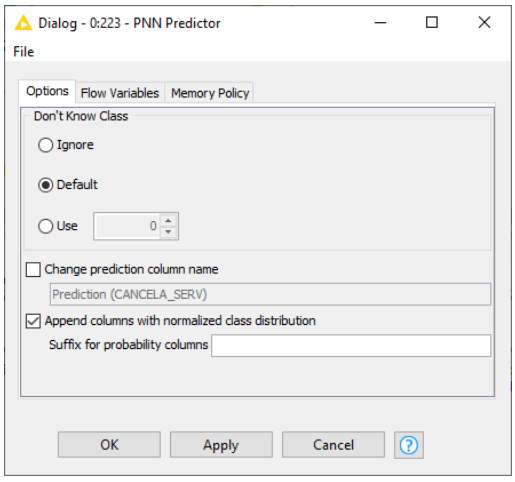


Nota. La figura representa el nodo PNN Predictor de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para predecir la data de prueba en base al entrenamiento dado en el nodo PNN Learner (DDA). Para nuestro caso de estudio específicamente este nodo nos permite predecir el valor de la columna CANCELA_SERV para nuevos patrones, Y este resultado se mostrara o visualizara en la columna denominada Prediction (CANCELA_SERV). En la Figura 60 se muestran las configuraciones realizadas en el Nodo PNNPredictor para nuestro caso de estudio específico.

Figura 60.

Configuración nodo PNN Predictor



Nota. La figura representa la configuración del nodo PNN Predictor de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

- **Scorer**

Figura 61.

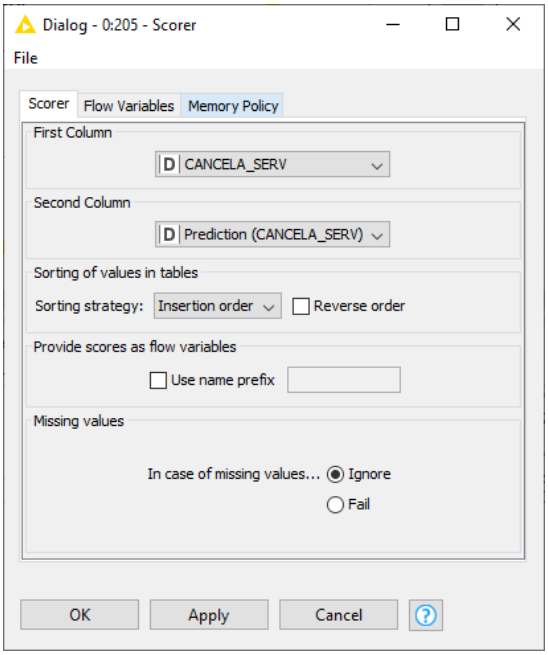
Nodo Scorer

Nota. La figura representa el nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Nodo usado para visualizar a la salida la matriz de confusión con el número de coincidencias en cada celda, esto como resultado al comparar el campo de entrenamiento con el campo predicho, este resultado permite evaluar que tan bueno o no resulta ser el modelo. Para nuestro caso específico mostrará el resultado al evaluar los registros de la columna CANCEL_SERV con el campo Prediction (CANCELA_SERV) y se podrá interpretar que tan confiable o preciso es nuestro modelo de Redes Neuronales. En la Figura 62 se observa la configuración de nodo y en la Figura 63 se muestra la matriz de confusión resultante.

Figura 62.

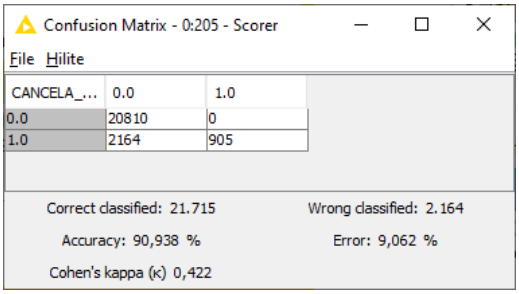
Configuración nodo Scorer



Nota. La figura representa la configuración del nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Figura 63.

Matriz de confusión nodo Scorer



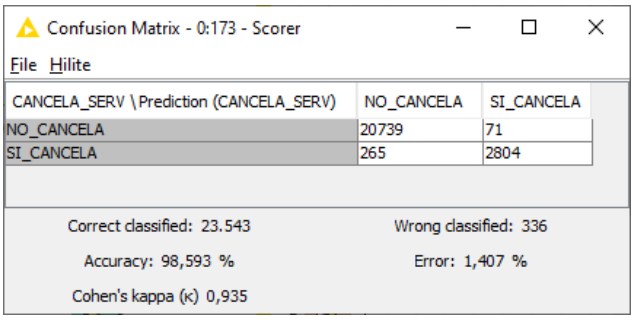
Nota. La figura representa el resultado de la matriz de confusión del nodo Scorer de la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Evaluación o Validación Del Modelo

En cada uno de los flujos de trabajo con los modelos desarrollados o propuestos se incorporó un nodo llamado Scorer con la finalidad de obtener la matriz de confusión de cada modelo y así poder evaluar cada uno de ellos para poder determinar cuál de los modelos planteados es el que mejor se adapta a nuestro caso de estudio y por ende el que más aciertos de predicción tenga, así como el menor porcentaje de error. En las figuras 64, 65 y 66 se puede observar las matrices de confusión resultantes de cada modelo.

Figura 64.

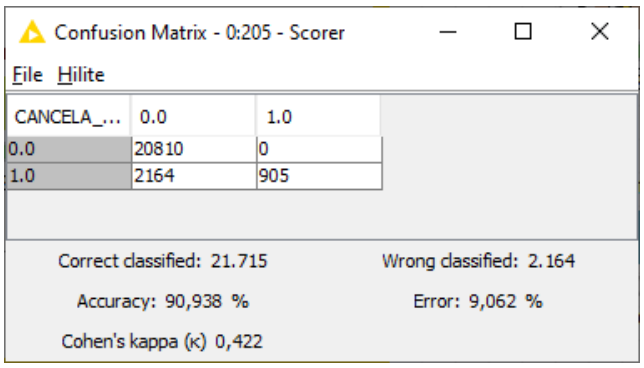
Matriz de confusión modelo Decisión Tree



Nota. La figura representa el resultado de la matriz de confusión del nodo Scorer en el modelo Decision Tree realizado en la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Figura 65.

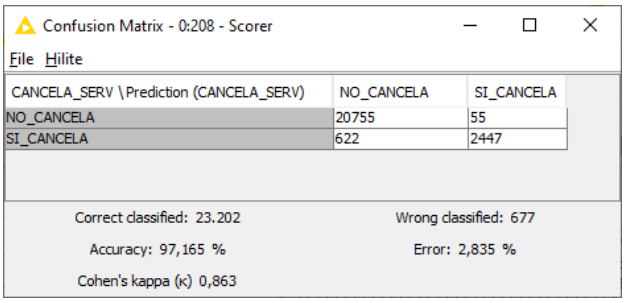
Matriz de confusión modelo Neural Network



Nota. La figura representa el resultado de la matriz de confusión del nodo Scorer en el modelo Neural Network realizado en la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

Figura 66.

Matriz de confusión modelo Naive Bayes



Nota. La figura representa el resultado de la matriz de confusión del nodo Scorer en el modelo Naive Bayes realizado en la herramienta Knime, este nodo se usó en la construcción de los modelos del presente trabajo de estudio.

En base a las figuras anteriores, se realiza una comparación entre las matrices de confusión, tal y como se muestra en la Tabla 10.

Tabla 10.

Resumen y comparativa de resultados de las matrices de Confusión generadas en cada uno de los modelos.

Técnica	Clasificados correctamente	Exactitud	Coeficiente Kappa	Clasificados Incorrectamente	Error
Árbol de decisión	23543	98,593	0,935	336	1,407
Red Neuronal	21715	90,938	0,422	2164	9,062
Naive Bayes	23202	97,165	0,863	677	2,835

Nota. La tabla representa a modo de resumen una comparación de los resultados de las matrices de confusión que se generaron para los 3 modelos elaborados.

De acuerdo a la Tabla 8, el número de datos clasificados correctamente para las tres técnicas es bastante alto consiguiendo porcentajes de exactitudes superiores al 90%, de la misma manera porcentajes de error bajos. También se muestran los valores del coeficiente Kappa, que indica que el porcentaje de concordancia de la variable predicha y la variable real son altas. No obstante, se puede concluir que la técnica de Árbol de Decisión en el modelo construido es la más exacta y la que mejor se adapta y predice la data ingresada para nuestro caso de estudio específico.

Despliegue de la Información

Finalmente en la fase de despliegue de la información, que es la última y en la que se explota, el conocimiento adquirido a través del modelo de árbol de decisión generado, se explica el proceso que permite a los altos mandos de la empresa de Telecomunicaciones tomar decisiones rápidas y acertadas a través del modelo de minería de datos de árbol de decisión encontrado, además se muestran los resultados obtenidos para que las gerencias puedan comprender de una forma rápida y fácil. Adicionalmente se indica el método que se usara para el respectivo mantenimiento al modelo generado.

Despliegue

Para poder realizar el despliegue del modelo generado es necesario lo siguiente:

- Acceso al Data warehouse existente
- Acceso a la herramienta ETL Informática Power Center.
- Solicitar un servidor virtual para la herramienta KNIME con accesos a los servidores del DWH y ETL descritos o mencionados.

- Instalación de la herramienta KNIME Analytics Platform v4.1.2 o versión superior.

Conseguidos los pasos anteriores se puede realizar lo siguiente:

- Configuración de periodicidad en el ETL para generación del Data Frame (input del modelo de árbol de decisión generado) según las necesidades de los dueños del proceso o según la demanda que el negocio lo requiera.
- Ejecutar el ETL para actualización de la data del DataFrame.
- Ejecutar el modelo de árbol de decisión de la herramienta Knime generado.
- Verificar el comportamiento del modelo en base a la data actualizada del DataFrame.

Monitoreo y mantenimiento

El monitoreo y mantenimiento del modelo de árbol de decisión generado debe considerar la frecuencia y/o periodicidad con la que se actualiza la información en los sistemas transaccionales y en el DWH existente, esta información se actualiza de dos formas, la primera se actualiza a cada momento (en línea), mientras que otro tipo de información (sin tanta relevancia) se actualiza a día caído. Los dueños del proceso deben definir la periodicidad de ejecución del ETL que permite crear o generar la data del DataFrame en base a sus necesidades o lo que el negocio requiera.

El proceso de minería de datos es recomendable realizarlo quincenal y mensualmente ya que de esta forma se podrá tener una visión temprana del comportamiento de los clientes y así poder realizar campañas de retención anticipada o proactiva a aquellos clientes que dentro de cada periodo (15 días) ya formen parte del segmento que desea cancelar sus servicios, si esperamos a fin de mes es posible que las campañas de retención anticipada no generen el mismo impacto con lo que tendríamos más clientes que si cancelan sus servicios contratados con la empresa de Telecomunicaciones objeto de estudio, sin embargo, esto podría variar de acuerdo a las necesidades del negocio o altos mandos.

En la figura 66 se puede observar un segmento del resultado que arroja el modelo para la data de la empresa de telecomunicaciones objeto de estudio, aquí se observa que clientes tienen un alto grado de intención para cancelar sus servicios contratados; y es sobre ellos a los que se debe realizar la gestión de retención anticipada ofreciéndoles beneficios que sean de su agrado y así que ellos desistan la idea de cancelar sus servicios.

Documentación de resultados

La documentación de resultados obtenidos presenta el informe en forma resumida de estos resultados y lo más relevante del proyecto de minería de datos realizado, en el Capítulo 4., se realiza la descripción y resultados obtenidos una vez ejecutado el modelo de árbol de decisión generado o resultante pata el presente caso de estudio.

Capítulo IV

Evaluación de resultados del Modelo predictivo ganador

En este capítulo se realizará la evaluación de los resultados obtenidos en el modelo predictivo ganador, mediante el análisis e interpretación de los mismos. Dicho análisis e interpretación se explica y detalla nivel a nivel (cinco primeros) del modelo Árbol de Decisión, en los cuales con valores y porcentajes tenemos de mayor a menor las principales causas o motivos que tienen los clientes para optar por la cancelación de los servicios contratados.

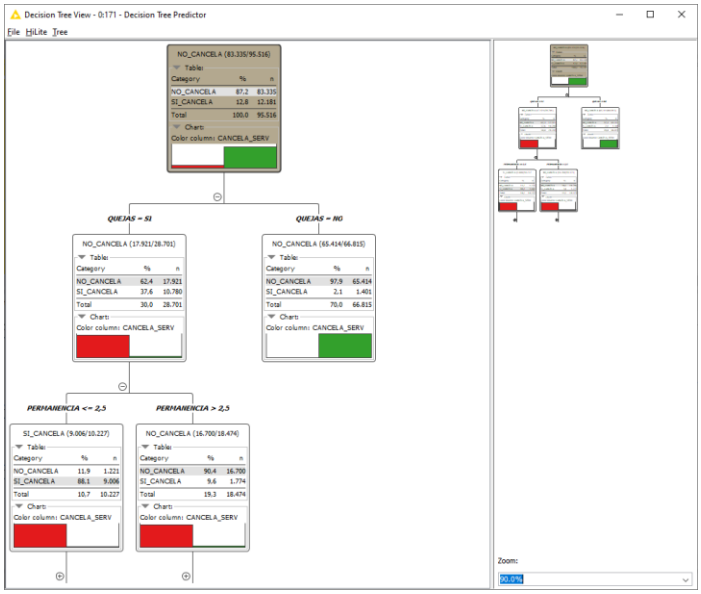
Análisis e Interpretación

A continuación en las Figuras 67 y 68 se puede observar los cinco (5) primeros niveles generados o resultantes en el aprendizaje del modelo Árbol de Decisión. A continuación se describen o explican de la siguiente forma:

En el nivel uno (1) se puede verificar que del total de la data ingresada (119404) para el entrenamiento del modelo de Árbol de Decisión se utilizó el (80%) y correspondiente a 95516, el (12.8% de la data de entrenamiento) que son 12181 corresponde a los clientes que si cancelan por lo menos un servicio contratado con la empresa de Telecomunicaciones y por la otra parte el (87.2%) que son 83335 corresponden a los clientes que no cancelan los servicios contratados. En el siguiente nivel (2) se verifica que la data se divide en 70% para los clientes que NO tienen ingresado quejas por algún problema técnico con el servicio contratado y por otra parte un 30% de clientes que si tienen ingresado por lo menos una queja por falla técnica en el servicio contratado, analizando este nivel y porcentajes ya descritos del 70% 1401 clientes si cancelan sus servicios contratados habiendo ingresado por lo menos un reclamo técnico por su servicio contratado, y los 65414 clientes no cancelaran su servicio contratado; en este mismo nivel del 30% de clientes que si tienen ingresadas por lo menos una queja del servicio técnico contratado, 10780 clientes si cancelan sus servicios contratados, mientras que los 17921 clientes que si tienen ingresados por lo menos un reclamo por falla en el servicio técnico contratado NO cancela sus servicios contratados.

Figura 67.

Ramas del mejor Modelo de predicción Árbol de Decisión (Niveles 1,2 y 3).

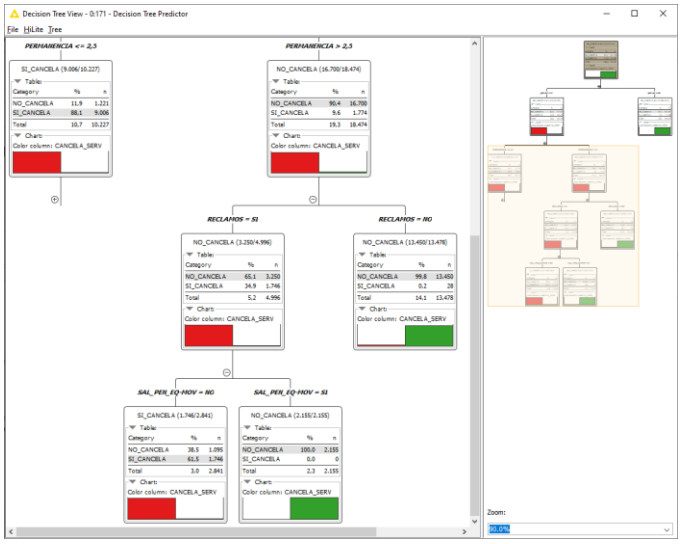


Nota. La figura representa las Ramas del mejor Modelo de predicción Árbol de Decisión (Niveles 1,2 y 3) desarrollado en la herramienta Knime.

En el siguiente nivel (3) se verifica que el 30% de clientes que tienen ingresado quejas por algún problema técnico con el servicio contratado se divide en un tiempo de permanencia menor o igual a 2.5 meses con un 10.7% y por otra parte un 19.3% de clientes que tienen un tiempo de permanencia mayor a 2.5 meses. Del 10.7%, 9006 clientes que tienen una permanencia menor o igual a 2.5 meses si cancelan por lo menos un servicio contratado y el restante 1221 clientes no cancelan sus servicios contratados; por la otra rama del árbol que es el 19.3% de clientes que tienen un tiempo de permanencia mayor a 2.5 meses, 16700 clientes no cancelan sus servicios contratados, mientras que 1774 clientes si cancelan por lo menos un servicio contratado a pesar de tener un tiempo de permanencia mayor a 2.5 meses.

Figura 68.

Ramas del mejor Modelo de predicción Árbol de Decisión (Niveles 3,4 y 5).



Nota. La figura representa las Ramas del mejor Modelo de predicción Árbol de Decisión (Niveles 3, 4 y 5) desarrollado en la herramienta Knime.

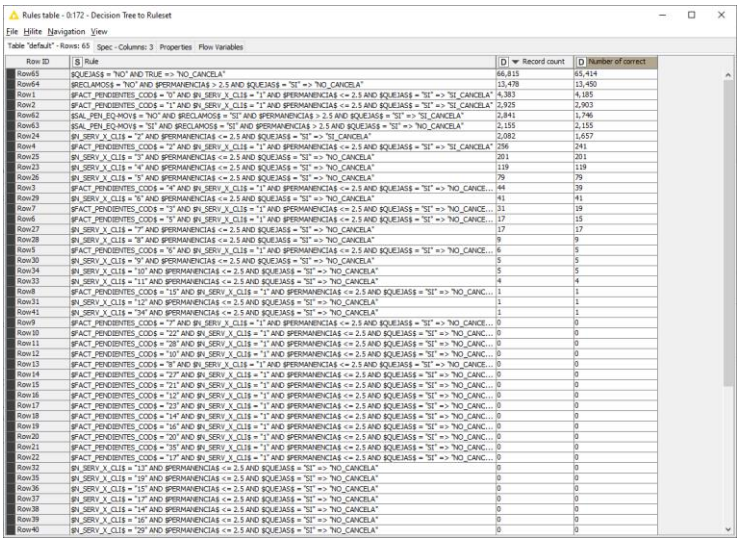
En los siguientes niveles 4 y 5 se ve el comportamiento sobre los clientes que tienen y no tienen ingresados reclamos de facturación por alguno de sus servicios contratados y en el último nivel que se muestra en la figura 64 se observa el comportamiento de los clientes para cancelar o no sus servicios contratados en base a si tienen o no un saldo pendiente por equipos móviles adquiridos.

En la Figura 69 se observa el contenido de la tabla con las reglas del árbol de decisión, misma que se genera en base al modelo del Árbol de decisión en base al entrenamiento de la data procesada.

En la Figura 70 se observa una muestra resultante de la data de prueba o verificación del modelo con los campos originales más campo de predicción y la debida segmentación en base a los colores configurados en el nodo Decision Tree to Ruleset.

Figura 69.

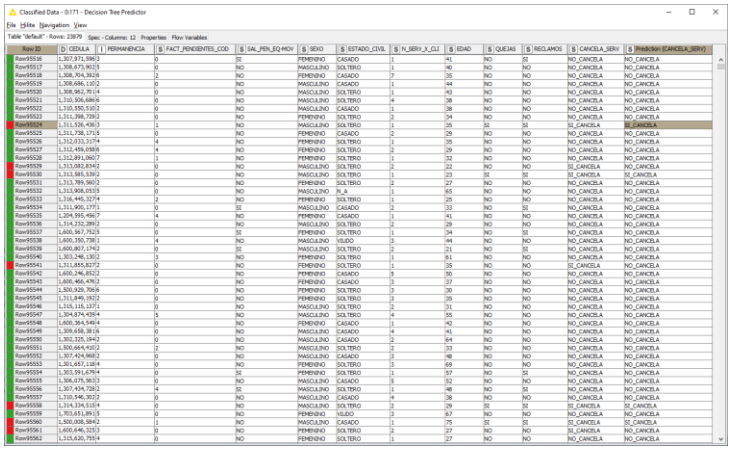
Reglas generada en el modelo Árbol de Decisión



Nota. La figura representa las Reglas generadas en el modelo Árbol de Decisión desarrollado en la herramienta Knime.

Figura 70.

Muestra resultante de la data de prueba (campos originales más campo de predicción, segmentación de colores)



Nota. La figura representa una tabla con una muestra del resultado del DataFrame luego de haber sido pasada por el modelo Árbol de Decisión desarrollado en la herramienta Knime.

Factores Posibles de Éxito del Modelo

Los factores posibles de éxito del modelo van encaminados a los beneficios o la gestión que se realice sobre los clientes con un alto grado de probabilidad de cancelación de sus servicios contratados en la empresa de telecomunicaciones.

Beneficios a otorgar a los clientes.

El o los beneficios que se otorguen a los clientes con un alto grado de probabilidad de cancelar sus servicios es uno de los ítems fundamentales a la hora de que los clientes tomen su decisión final de verdaderamente cancelar el servicio o no. Por ejemplo: si un cliente ha realizado varios reportes por fallas en el servicio de internet contratado y que esto representa o equivale a 3 días (10% del mes) sin servicio de internet, y este cliente es uno de los que se encuentran como resultado del modelo (es decir aun no cancela su servicio pero tiene un alto grado de probabilidad de hacerlo), es la oportunidad de llamarle o contactarle de alguna manera e indicarle que su problema va a ser resuelto de inmediato, pero adicionalmente ofrecerle uno o dos meses de servicio sin que el cliente deba pagar ningún valor por ello; este es uno de los posibles beneficios que se otorgan a los clientes para obtener una retención de clientes[12] eficiente en el transcurso del tiempo.

Gestión hacia los clientes.

Otro de los ítems o factor clave de éxito a la hora de que los clientes tomen su decisión final en cuanto a cancelar o no sus servicios contratados es la gestión o trato que reciben por parte de los asesores de telecomunicaciones en el momento en que se contactan o comunican con ellos; es decir, no es suficiente con que se comuniquen con los clientes en una fase temprana de alta intención de cancelar sus servicios y decirle 'Sr, XYZ no cancele su servicio de zzzz, y a cambio no le vamos a cobrar la factura de uno o dos meses', sino más bien es la empatía[13] que se le debe transmitir al cliente, de esta forma el cliente sentirá que nos hacemos eco de su malestar, y que vamos a solucionar o corregir el problema o queja reportado en el menor tiempo posible, apersonándonos y dando el respectivo seguimiento para la solución y a su vez transmitiendo o informando los respectivos avances del caso en cuestión. Y adicionalmente a todo lo anteriormente indicado se complementa esta

[12] La retención de clientes es la capacidad de una empresa para retener a sus clientes. Con el tiempo, la retención se mide por el porcentaje de clientes que una empresa mantiene sobre su número total de clientes dentro de un período de tiempo específico.

[13] La empatía es una habilidad fundamental en la atención al cliente, es la capacidad de entender el problema del otro, casi sentirlo, vivirlo y saber responder poniéndonos en la piel del otro. Cuando se consigue ser empático, podemos lograr cosas extraordinarias como convertir a nuestros clientes en embajadores de marca.

gestión con un beneficio para en algo tratar de minorar la molestia o impacto que pudiere haber ocasionado el problema con el servicio contratado y que fue reportado por el cliente.

Capítulo V

Conclusiones y Recomendaciones

Conclusiones

Luego de haber concluido el presente proyecto de tesis, se concluye lo siguiente:

- Posterior a realizar un estudio de literatura de los artículos relacionados al tema de investigación, se identifica que en varios de los artículos revisados se realiza un análisis y comparación de los métodos de predicción utilizados, pero no se encontró estudios similares para empresas de telecomunicaciones, es por eso que se optó por realizar 3 de las técnicas de modelamiento más usadas para variables categóricas y numéricas, estas fueron: Árboles de Decisión, Naive Bayes y Redes Neuronales.
- La causa más fuerte o potente por la que los clientes optan o deciden cancelar sus servicios contratados con la empresa de telecomunicaciones objeto de estudio es o son las fallas técnicas presentadas en el servicio y que se traducen o reflejan en la variable quejas.
- Luego de las validaciones realizadas a los 3 modelos propuestos se concluye que la técnica de Árbol de Decisión es la más exacta con un 98.59%, la que menor porcentaje de error posee (1.40%), la que mejor se adapta y predice la data ingresada para nuestro caso de estudio específico, permitiéndonos obtener e identificar patrones o reglas del comportamiento que tienen en común los clientes previo a cancelar sus servicios contratados con la empresa de telecomunicaciones objeto de estudio.
- En base a los resultados obtenidos en el desarrollo del presente proyecto se concluye que la hipótesis propuesta es positiva ya que el modelo analítico construido permite alertar de manera temprana, indicando quienes son los clientes que tienen una alta probabilidad o intención de cancelar los servicios contratados con la empresa de telecomunicaciones objeto de estudio.

Recomendaciones

- El modelo analítico desarrollado debe ser empleado como insumo en el diseño del plan de inteligencia de negocios de la empresa de Telecomunicaciones objeto de estudio en su fase preliminar.
- El realizar la evaluación de resultados del modelo predictivo nos permitió conocer que la permanencia de los clientes en la empresa objeto de estudio se determina en base a la aplicación de una atención y beneficio atractivo a los clientes que tienen una gran intención de cancelar sus servicios.
- El presente estudio dio como resultado que el modelo predictivo propuesto es de carácter multidisciplinar, pero enfocado al uso en empresas de telecomunicaciones; el modelo desarrollado puede ser utilizado para todos los tipos de servicios o productos en el área.

Bibliografía:

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Step-by-step data mining guide*. https://www.the-modeling-agency.com/crisp-dm.pdf

Dolatabadi, S. H., & Keynia, F. (2017). Designing of customer and employee churn prediction model based on data mining method and neural predictor. *IEEE*, 74–77. https://ieeexplore.ieee.org/document/8075270

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54.

Gallardo Arancibia, J. A. (2009). *Metodología para la definición de requisitos en proyectos de data mining* [UNIVERSIDAD POLITÉCNICA DE MADRID]. http://oa.upm.es/1946/

Mishra, A., & Reddy, U. S. (2017). A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. *IEEE*, 721–725. https://ieeexplore.ieee.org/document/8365230

Mishra, K., & Rani, R. (2017). Churn prediction in telecommunication using machine learning. *IEEE*, 2252–2257. https://doi.org/10.1109/ICECDS.2017.8389853.

Molina López, J. M., & García Herrero, J. (2006). *TECNICAS DE ANALISIS DE DATOS*. http://matema.ujaen.es/jnavas/web_recursos/archivos/weka master recursos naturales/apuntesAD.pdf

Piatetsky, G. (2020). *Leaders, Changes, and Trends in Gartner 2020 Magic Quadrant for Data Science and Machine Learning Platforms*. https://www.kdnuggets.com/2020/02/gartner-mq-2020-data-science-machine-learning.html

Qaisi, L. M., Rodan, A., Qaddoum, K., & Al-Sayyed, R. (2018). Customer churn prediction using data mining approach. *IEEE*, 348–352. https://doi.org/10.1109/CTIT.2018.8649494.