UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

TEMA:

Diseño de un modelo predictivo a través de la técnica de minera de datos ‘Random Forest’ para la detección de fraude bypass en redes telefónicas en el Ecuador

AUTOR:

ALCIVAR LEÓN CRISTHIAN ROGER

Trabajo de titulación previo a la obtención del título de

INGENIERO EN SISTEMAS COMPUTACIONALES

TUTOR:

CORNEJO GOMEZ GALO ENRIQUE

Guayaquil, Ecuador

Guayaquil, a los 26 días del mes de Febrero del año 2020

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de titulación, fue realizado en su totalidad por **ALCIVAR LEÓN CRISTHIAN ROGER**, como requerimiento para la obtención del título de **Ingeniero en Sistemas Computacionales**.

TUTOR

Ing. Cornejo Gómez Galo Enrique, Mgs.

DIRECTOR DE LA CARRERA

Ing. Camacho Coronel Ana Isabel, Mgs.

Guayaquil, a los 26 días del mes de febrero del año 2020

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Yo, **Alcivar León Cristhian Roger**

DECLARO QUE:

El Trabajo de Titulación, **Diseño de un modelo predictivo a través de la técnica de minera de datos 'Random Forest' para la detección de fraude bypass en redes telefónicas en el Ecuador** previo a la obtención del título de **Ingeniero en Sistemas Computacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

Guayaquil, a los 26 días del mes de Febrero del año 2020

EL AUTOR

f. _____

Alcivar León Cristhian Roger

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

AUTORIZACIÓN

Yo, **Alcivar León Cristhian Roger**

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación, **Diseño de un modelo predictivo a través de la técnica de minera de datos 'Random Forest' para la detección de fraude bypass en redes telefónicas en el Ecuador**, cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 26 días del mes de Febrero del año 2020

EL AUTOR:

f. _____

Alcivar León Cristhian Roger

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

REPORTE URKUND

TUTOR

f. _____

Ing. Cornejo Gómez Galo Enrique, Mgs.

AGRADECIMIENTO

Quisiera extender mis agradecimientos a mis padres por el esfuerzo que día a día han realizado dentro de toda mi etapa de desarrollo personal. Deseando enfatizar el amor y perseverancia recibido por mi madre, Dra. Tanya León, que siempre estuvo pendiente de mi bienestar cada mañana para que yo pueda salir en busca de mis metas y por ser un pilar fundamental en mi crecimiento académico. Esto siempre estará marcado en mi corazón. Gracias.

Ing. Sally Parra, agradezco por tu constante apoyo a seguir adelante a lo largo de toda la carrera, también en lo personal y en lo profesional. Especialmente por tu paciencia y brindarme palabras de aliento en cada momento duro, y enseñarme a ser perseverante y motivarme a crecer cada día.

DEDICATORIA

Este trabajo está dedicado a mis padres, hermanos, compañera de vida y amigos cercanos, cada uno fue un apoyo fundamental en esta travesía y en la culminación de la carrera.

Una especial dedicatoria a mi yo del futuro, para que cada vez que leas este trabajo se restauren tus ganas de seguir adelante para alcanzar tus metas, recuerdes que el trabajo duro nunca termina y sepas lo capaz que eres.

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

TRIBUNAL DE SUSTENTACIÓN

f. _____

Ing. Camacho Coronel Ana Isabel, Mgs.

DIRECTOR DE CARRERA

f. _____

Ing. Toala Quimi Edison Jose, Mgs.

DOCENTE DE ÁREA

f. _____

Ing. Castro Aguilar Gilberto Fernando, Mgs.

OPONENTE

ÍNDICE GENERAL

RESUMEN

En las empresas de telecomunicaciones, dado a su creciente demanda de mantenernos comunicados, emplean mayores configuraciones y elementos que en ocasiones generan brechas que son aprovechadas por personas u organizaciones para realizar acciones ilícitas como es el caso del fraude por bypass. Es un hecho que el 5% de los ingresos de una empresa de telefonía móvil del país se pierdan por la falta de detección de este tipo de casos, el cual ocasiona fugas de ingresos para las organizaciones que ofrecen este servicio a nivel mundial y nacional. Este trabajo denominado “Diseño de un modelo predictivo a través de la técnica de minera de datos ‘Random Forest’ para la detección de fraude bypass en redes telefónicas en el Ecuador” plantea como objetivo la construcción de un modelo predictivo empleando minería de datos a través de la metodología KDD (“Descubrimiento de Conocimiento en Base de Datos”) de modo que contribuya a la eficacia en la detección de este tipo de fraude. La aplicación de la metodología se realiza mediante la herramienta de software KNIME implementando un flujo de trabajo con el uso de bosques aleatorios como técnicas de minería de datos del tipo clasificatorio supervisada donde se emplean los registros de llamadas como base para transformarlos a una vista minable apta para la construcción del modelo. Los resultados del trabajo indicaron que el uso de la minería de datos reportó mayor eficacia que el análisis de CDRs tradicional en la detección de casos de fraude bypass y plantea las bases para futuros estudios del tema en este modelo de negocios.

ABSTRACT

Telecommunications companies use greater configurations and elements in order to keep their growing demand of keep us communicated. These sometimes generates gaps that are used by organizations or people to carry out illegal actions such as bypass fraud. It fact, 5% of the mobile services companies's revenues in the country are lost due to the lack of detection of such events. Which causes revenue leaks for organizations that offer this service worldwide and nationally. This work called "Design of a predictive model through data mining technique 'Random Forest' for the detection of bypass fraud in telephone networks in Ecuador" aims to build a predictive model using data mining through of the KDD methodology ("Discovery of Knowledge in Database") in order to contributes to the effectiveness in the detection of this type of fraud. The methodology is applied using the KNIME software tool by implementing a workflow with the use of random forests as a supervised classification data mining technique where call records are used as basis for transforming them into a suitable mining view for the construction of the model. The results of the work indicated that data mining use reported greater efficiency than a traditional CDRs analysis in the detection of cases of bypass fraud and sets up the basis for future studies of the topic in this business.

INTRODUCCIÓN

El sector de las telefonías móviles ha dado pasos agigantados gracias a la evolución constante de la tecnología. Tanto en la manera de cómo comunicarnos y en los productos que ofrecen a la población. Este crecimiento también ha originado el surgimiento de distintas clases de fraude en este mercado, los cuales han dejado de ser sencillos, generando mejores técnicas y nuevos escenarios con distintos patrones fraudulentos llegando a afectar distintas compañías que ofrecen el servicio de telecomunicaciones alrededor del mundo.

Por esta razón, se demandan esfuerzos tecnológicos y económicos importantes a nivel mundial, dando paso a organizaciones como la asociación de control de fraude en comunicaciones (“CFCA” por sus siglas en ingles) que sin fines de lucro reúne información de proveedores de servicios de telecomunicaciones, socios expertos y organismos estatales con el fin de prever pérdidas de ingresos y controlar el fraude en este sector (Communications Fraud Control Association [CFCA], 2019).

No obstante, los tipos de fraudes que destacan de manera global a las empresas telefónicas son IRSF, interconexión bypass, arbitraje, robo de equipos y fraude en servicios Premium, entre otros; siendo el fraude de interconexión bypass uno de los más complejos debido a su alta capacidad de evasión, generando gran impacto económico a las empresas. (Koi-Akrofi et al., 2019).

En el Ecuador, a pesar del empeño en conjunto entre la Agencia de Regulación y Control de las Telecomunicaciones (ARCOTEL) con las distintas operadoras de servicios móviles; estas siempre estarán expuestas a nuevas modalidades de fraude telefónico que exigirán técnicas eficaces para mitigarlas.

Para este fin, las metodologías de minerías de datos son usadas por las empresas de telecomunicaciones para el ámbito de detección de fraudes ofreciendo mejores tiempos de respuesta en la lectura de patrones fraudulentos en los registros de llamadas. (Menéndez, 2008).

Es por ello, que el presente proyecto de titulación busca proponer el diseño de un modelo predictivo empleando las técnicas de minería de datos para la detección temprana de fraude de tipo ‘Bypass’ en telefonías móviles en el Ecuador. El cual se encuentra organizado de la siguiente manera: El capítulo I ofrece un contexto acerca de la problemática, la justificación, el alcance y los objetivos; En el capítulo II proporciona un marco teórico, conceptos y la legalidad para el presente proyecto; En el capítulo III se explica la metodología de la investigación usada; En el capítulo IV se detalla la metodología de minería de datos aplicadas y desarrollo; Por último serán presentadas las conclusiones y recomendaciones de la presente investigación.

CAPÍTULO I
EL PROBLEMA

En el presente capítulo se identifica el problema con el que las empresas de telefonías móviles se enfrentan, se plantean los objetivos cubiertos en el presente trabajo, así como el alcance del mismo y la propuesta que contribuirá a la solución del problema observado.

PLANTEAMIENTO DEL PROBLEMA

La telecomunicaciones a través en el mundo se ha convertido en un pilar fundamental para mantenernos conectados y actualizados a través de un primitivo teléfono móvil en 1947 hasta la actualidad donde los teléfonos celulares se han transformado en herramienta elemental para los negocios y personas brindándoles mayor productividad (Gámez et al., 2005). Las comunicaciones entre estos dispositivos son manejadas por las empresas de telefonía celular quienes tienen una fuerte posición en el mercado manteniendo valores de marca de hasta 108 mil millones de dólares como es en el caso de la empresa AT&T, líder a nivel mundial (Fernández, 2019).

Una alta evolución tecnológica junto al crecimiento exponencial de los suscriptores en las redes de telefonías móviles conlleva un incremento considerable en la cantidad de datos generados.

> En el 2017, ya había más de 5 mil millones de suscriptores móviles en todo el mundo, con una tasa de penetración del 66% de la población mundial, y un número total de suscripciones móviles que supera la población mundial en 7.79 mil millones. Además, la penetración de la telefonía móvil está en constante aumento y se prevé que alcance casi 6 mil millones de usuarios para el 2025, con 5 mil millones conectados a internet. (Lai et al., 2019)

Lo cual se traduce en una mayor cantidad de configuraciones, proveedores, elementos de interconexión que sean capaces de sobrellevar el aumento de la cantidad de registros de llamadas generadas por los mismos subscriptores. Que a su vez, producen ventanas o brechas que son aprovechadas por personas mal intencionadas que por distintos medios originan fraude, el cual es un gran problema para todas las empresas de telecomunicaciones a nivel mundial y es un factor importante en sus pérdidas de ingresos anuales (Koi-Akrofi et al., 2019).

Entonces resulta lógico que cada año aparezcan nuevos tipos de fraude en este sector llegando a tener al momento más de 40 tipos diferentes considerados como delitos, donde destacan el fraude por altos consumos e interconexión bypass, siendo este último el causante de pérdidas económicas de $4.3 miles de millones en el 2017 (CFCA, 2019) a nivel mundial y que han exigido medidas internacionales para la mitigación de afectaciones económicas que podrían incrementarse de continuar sin una solución a la problemática.

A nivel del Ecuador, la interconexión bypass ha sido el foco de atención para el combate de fraude en las telecomunicaciones que incluso ha llegado a generar pérdidas cuantiosas al estado ecuatoriano y a las operadoras autorizadas; Según la Agencia de Regulación y Control de las telecomunicaciones (ARCOTEL, 2019) en su informe de rendición de cuentas del 2018, en el Ecuador se reportó más de 460 mil números telefónicos

cursando tráfico no autorizado causado por fraude bypass. Así también por varias ocasiones se han realizado operativos de allanamientos a residencias que operan como centrales telefónicas clandestinas efectuando este tipo de fraude a las telecomunicaciones en el país (El Telégrafo, 2019).

Es destacable los esfuerzos que cada operadora y la ARCOTEL invierten para operaciones anti-fraude, por lo que actualmente para combatir este tipo de delitos en el país aplican sistemas que permitan hacer detecciones de este tipo de fraudes tales como Loop de llamadas, análisis de CDRs y perfilamiento (Camacho & González, 2016). Sin embargo, estos sistemas no emplean metodologías basadas en minería de datos ni modelos predictivos.

Es por ello que es necesario aplicar técnicas que impliquen el uso de la minería de datos que permitan hacer detecciones de manera preventiva e incluso predictiva que contribuyan a los sistemas ya existentes usados por las empresas de telefonía celular para combatir el fraude bypass en servicios móviles en el Ecuador.

OBJETIVOS

OBJETIVO GENERAL

Diseñar un modelo predictivo usando técnicas de minera de datos que pueda ser aplicado en redes telefónicas móviles del Ecuador para aportar a la detección de fraude bypass.

OBJETIVOS ESPECÍFICOS

- Identificar los procesos de detección de fraude bypass en telefonías celulares, con el fin de establecer las variables críticas que serán seleccionadas para el modelo predictivo.
- Aplicar un análisis estadístico descriptivo para la normalización de la información en la construcción de set de datos minables.
- Construir el modelo predictivo aplicando algoritmos de minería de datos para la detección de casos de fraude bypass.
- Evaluar el modelo propuesto utilizando datos generados en un ambiente controlado.

ALCANCE DEL PROBLEMA

El presente proyecto analizará el patrón de comportamiento del fraude bypass, a través del análisis de CDRs de voz, específicamente del tipo saliente, como sería en una de las empresas de telefónica móvil del Ecuador que contribuya al nivel de detección de este tipo de delitos.

Cabe indicar que para la creación de un set de datos minable que permita la creación de un modelo predictivo se analizará una base de datos de CDRs (registro de llamadas) en un periodo comprendido de 45 días obtenidos del repositorio académico de la Universidad de Harvard llamado "Dataverse", el cual será modelado de tal manera que cumpla con los lineamientos y parámetros fijados a través de entrevistas con especialistas en el tema de bypass y análisis de registros de llamadas junto a lo estudiado en el marco teórico del presente trabajo de titulación, con el fin de determinar irregularidades, patrones de comportamiento y comprender los datos para sean útiles para cumplir con el objetivo de este proyecto.

Los modelos predictivos a desarrollar y probar serán obtenidos a través de herramientas de minerías de datos, las cuales serán estudiadas y se extraerá entre ellas la más idónea para este tipo de estudios donde se muestre sistemáticamente la aplicación de la metodología de minería de datos seleccionada. Como resultado final de este proyecto se cumplirá con lo establecido en los objetivos propuestos del mismo.

El proyecto está categorizado dentro de la línea de investigación y desarrollo de nuevos servicios o productos de la carrera Ingeniería de Sistemas Computaciones de la Facultad de Ingeniería de la Universidad Católica Santiago de Guayaquil.

JUSTIFICACIÓN E IMPORTANCIA

El uso de la minería de datos está siendo empleado en la actualidad en diversas índoles dado a su aporte en la toma de decisiones al generar patrones de comportamiento con el fin de realizar predicciones y prevenciones en el caso que lo amerite.

En el medio de las telefonías móviles emplear minería de datos aporta con las herramientas para generar modelos que permitan entender el comportamiento de sus subscriptores con el afán de detectar patrones de fraudes que contribuyan a los sistemas de detección utilizados actualmente.

El resultado del presente trabajo presenta un modelo, que contribuya a la detección de casos de fraude bypass, y de esta manera poder minimizar los riesgos que corren las operadoras celulares en el país, brindando un apoyo además en la toma de decisiones.

Además, el presente trabajo de titulación espera establecer las pautas que pueden ser empleadas para la implementación de modelos predictivos más complejos a través del uso de las técnicas de minería de datos de modo de que permita la detección temprana de fraude de tipo bypass en el país de manera que beneficie a las empresas de telecomunicaciones y al estado.

HIPÓTESIS O PREGUNTA DE INVESTIGACIÓN

Por medio del presente trabajo de titulación se busca la confirmación de la hipótesis que se plantea:

Un modelo predictivo utilizando minería de datos contribuirá a la eficacia en la detección de casos de fraude tipo Bypass en las empresas de telecomunicaciones en el Ecuador.

VARIABLES DE LA INVESTIGACIÓN

- **Variable independiente:** modelo predictivo usando minería de datos
- **Variable dependiente:** eficacia en la detección de casos de fraude tipo Bypass.

CAPÍTULO II
MARCO TEÓRICO, CONCEPTUAL Y LEGAL

Para comprender el uso de las técnicas de minerías de datos y lo que conlleva ser aplicadas al manejo de grandes volúmenes de información para la construcción de un modelo predictivo que permita la detección de patrones de fraudes bypass en redes telefónicas móviles, es necesario exponer, en este capítulo, un aproximamiento a las diferentes técnicas, principios, clasificaciones y criterios expertos sobre el tema. De igual manera, diferentes conceptualizaciones de términos relaciones con el fraude en las telecomunicaciones y el ámbito de estudio; Y finalmente, un marco legal que sustenten lo contextualizado sobre las empresas que ofrecen un servicio de telefonía móvil en el Ecuador y sus normativas frente al fraude bypass en las telecomunicaciones.

MARCO TEÓRICO

Minería de datos

La minería de datos se puede conceptualizar como una ciencia para extraer información de grande volúmenes de datos también conocida como “pesca de datos” o “filtración de datos” (Camana, 2016). Por otro lado, se destaca que la minería de datos usa integradamente diferentes disciplinas como la estadística, el uso de base de datos y data-warehouse, reconocimiento de patrones, visualización de datos y procesamiento de información para el análisis de datos masivos (De Battista et al., 2016). Adicional, se menciona a la minería de datos como un conjunto de herramientas que generan contenido valioso a través del uso de modelos aplicando algoritmos computaciones (Kotu & Deshpande, 2014). Dado al volumen significativo de información que se puede generan en una red telefónica móvil, es indispensable el uso de la minería de datos para la construcción de modelos que vayan de la mano con los problemas generados en este tipo de negocios.

Según la estadística, un modelo estadístico se describe como la representación de la relación entre las variables de los datos, indicando como una o más variables se encuentran relacionadas con otras. Más allá, el proceso de modelamiento se define como la construcción de una abstracción representativa desde los datos observados. (M. H. Badii et al., 2017). Dentro de este contexto, la minería de datos es el proceso de construir un modelo representativo que este sujeto al comportamiento de la data observable del problema de estudio.

A su vez, las técnicas de minería de datos se encuentra clasificada según el modelo de aprendizaje en (Marin Castro, 2017):

- Supervisada o predictiva: Construye una función o relación basado en el comportamiento de data de entrenamiento previamente clasificada y usa está función para mapear futura data no clasificada, para ello se necesita grandes volúmenes de información para que el modelo “aprenda” de los datos. (García et al., n.d.)
- No supervisada o descriptiva: Ayuda a revelar patrones desconocidos permitiendo identificar propiedades de los datos examinados. No para predecir sino describir características importantes de la información. (Marin Castro, 2017)

Para efectos de este trabajo de titulación, nos basaremos en las técnicas supervisadas o predictivas las cuales se apegan más al objeto de estudio en este proyecto. Profundizando, las técnicas supervisadas son usadas para construir modelos a partir de set de datos que contienen ejemplos de los conceptos que deben ser aprendidos (Roiger, 2017). Estas técnicas de minería de datos supervisadas poseen su propia clasificación (Kotu & Deshpande, 2014):

- Predicción: En este grupo se predice un valor numérico.
- Clasificación: Este grupo se encarga de predecir variables de salida categorizadas (Si/No).

Ambas categorías se basan en un modelo construido por un set de datos previamente tratado y conocido y a su vez poseen varios algoritmos que son usados dependiendo el objeto de estudio y la complejidad del problema. De

modo que, se pueda predecir si un servicio de telefónica móvil es un fraude por bypass o no; Se emplearía la técnica de minería de datos supervisada de clasificación, esto basado en los criterios indicados anteriormente.

En la figura 1 se muestra las distintas técnicas de minería de datos ya antes mencionadas junto a los algoritmos empleados según sus distintas clasificaciones.

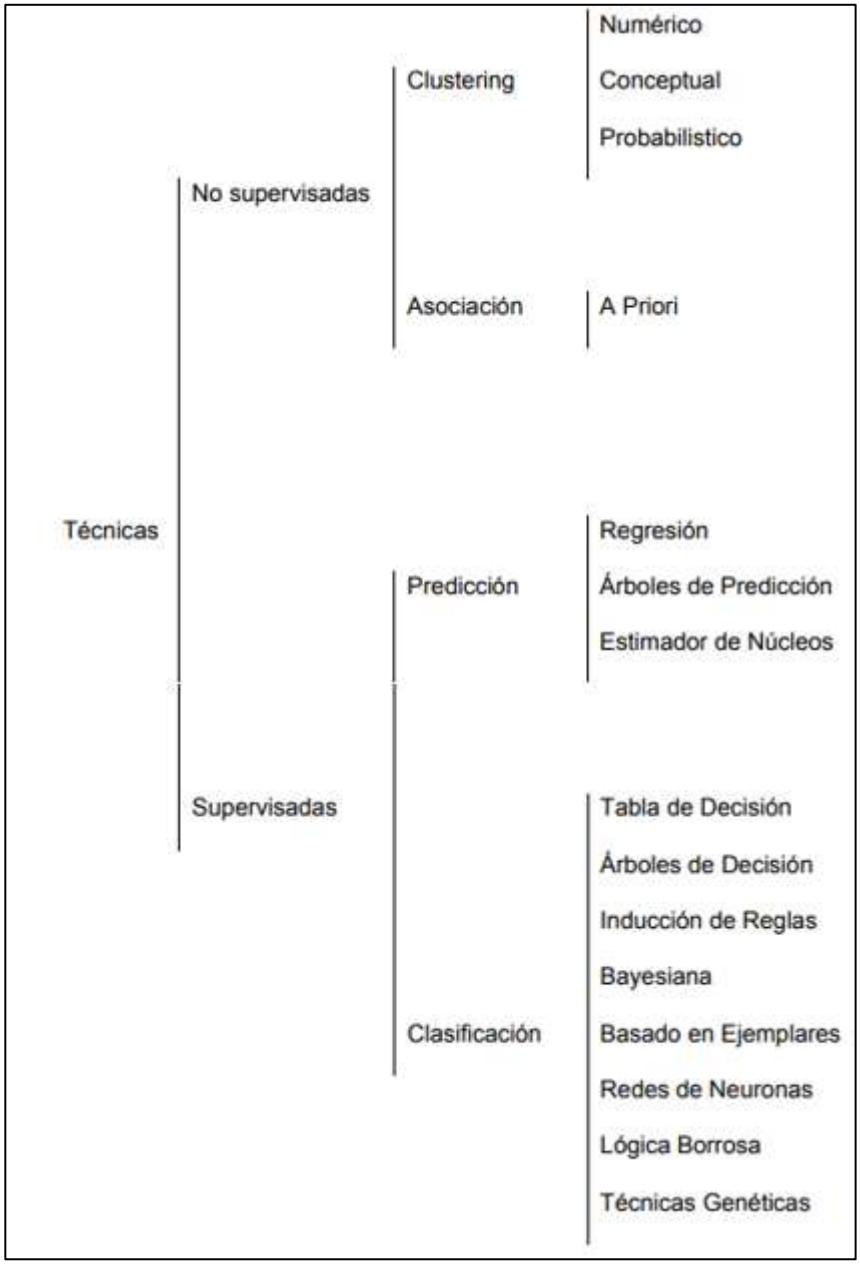

Figura 1. Técnicas de minería de datos, tomado de (Molina & García, 2017)

La elección del algoritmo definido como un procedimiento paso a paso para resolver un problema, depende del tipo de set de datos, objetivo de la minería de datos, número de registros, potencia de procesamiento, entre otras (Witten et al., 2016). Dentro de los algoritmos de clasificación utilizados en minería de datos se pueden mencionar (Molina & García, 2017):

- Tabla de decisión: Esta algoritmo se base en seleccionar subconjuntos de atributos y calcular su precisión para predecir o clasificar los ejemplos.
- Arboles de decisión: Pueden definirse como una serie de reglas que son tomadas de acuerdo a un conjunto de ejemplos que se representaran de forma de árbol.
- Inducción de reglas: Se trata de darle a un árbol de decisión un cierto número de reglas basadas en patrones a partir de los datos de entrada, a través del uso de métodos estadísticos que permitan podar el árbol.
- Basado en ejemplares: Tiene como principio el almacenamiento de ejemplos, los cuales son los más representativos y son tomados en cuenta al momento de realizar la predicción a través de una función de proximidad o parecido.
- Redes neuronales: Es una técnica que pretende simular el aprendizaje a través del funcionamiento de las neuronas capaces de detectar complejos patrones basados en experiencias del pasado. Requiere un alto poder de adiestramiento.
- Lógica borrosa: Se basa en una técnica de categorización por rangos donde es tomado en cuenta la disyunción de sus límites.
- Técnicas genéticas: Representan un modelado simulando a los cromosomas en un contexto evolucionista, donde en cada “cepa” se constituirá en un algoritmo más óptimo.

Así mismo, pueden existir combinaciones entre los algoritmos de clasificación antes mencionados, siendo uno de ellos los bosques aleatorios o “Random Forest”, los cuales reúnen características principales de los árboles de decisión y otros algoritmos, permitiéndoles ser versátiles en su uso frente a distintos problemas de minería de datos. Según su estudio (Subudhi et al., 2020) definen:

Un bosque aleatorio puede definirse como un método supervisado en donde son empleados múltiples arboles de decisión para crear un bosque, el cual trabaja mejor con set de datos de gran tamaño. Donde este es dividido de manera aleatoria en varias partes con la misma estructura. Los bosques aleatorios son un algoritmo eficiente con alta precisión en las predicciones realizadas donde la decisión final es tomada con la colección de decisiones individuales para obtener la mejor precisión de clasificación.

En la figura 2 se muestra el diagrama funcional del algoritmo de bosques aleatorios donde se muestra de manera gráfica el comportamiento del mismo.

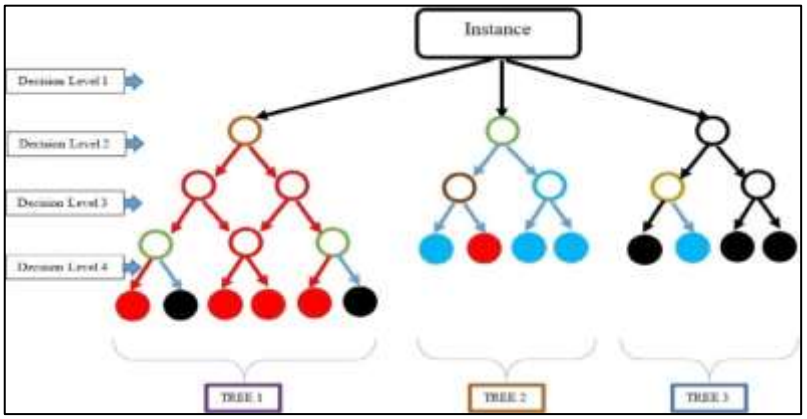

Figura 2. Diagrama funcional de los bosques aleatorios o "Random Forest", tomado de (Subudhi et al., 2020)

Los múltiples árboles de un bosque aleatorio según indica Belgiu & Dragut (2016):

Son construidos a través de subset de datos constituidos por muestras que pueden ser elegidas varias veces, mientras que otras muestras no pueden ser elegidas nunca. Alrededor de dos tercios de la muestra es usada para entrenar a los árboles (“in-bag samples” por su término en inglés), mientras que el tercio restante es usado para la técnica interna de validación cruzada que permite evaluar que tan buenos son los resultados que el modelo desempeñó. Este error de estimación es conocido como “Out-of-bag error (OBB)”.

Además, cada árbol es producido independientemente, sin ninguna poda previa, y cada uno es dividido usando un número definido de variables seleccionadas aleatoriamente. Cada árbol muestra su resultado independiente y la decisión de clasificación final es tomada por la clase con más votos que recibe de cada árbol generado, luego de haber sido sometidos a nuevos datos sin clasificar (Belgiu & Drăguţ, 2016). Lo anterior, puede ser visualizado en la figura 3 a continuación.

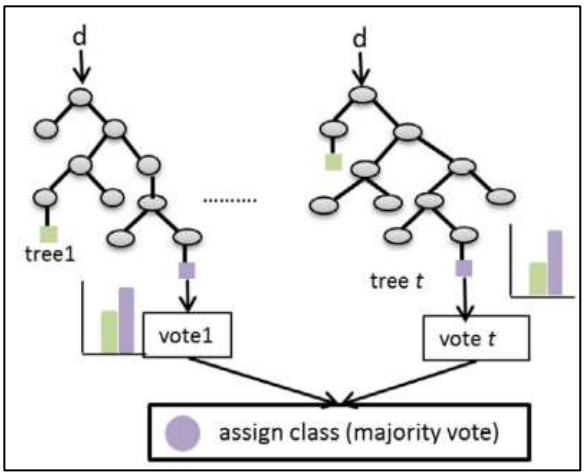

Figura 3. Fase de clasificación de bosques aleatorios t = número de árboles y d = nuevos datos, tomado de(Belgiu & Drăguţ, 2016)

Dado a que los bosques aleatorios son computacionalmente eficientes y no se sobreentrenan, el número de árboles en el bosque puede ser tan grande como sea posible (Guan et al., 2013). Algunos estudios reportan que un número eficiente de árboles en el bosque aleatorio es de 500 debido a que el error se estabiliza antes de alcanzar este número. (Lawrence et al., 2006). Otra razón es que este número de 500 es el más común usado en las implementaciones en R de bosques aleatorios. (Belgiu & Drăguţ, 2016). Basados en estos dos criterios, este número de 500 es referente para definir número de árboles para el bosque aleatorio o “random forest” empleado en este trabajo de titulación. En cuanto, al número de variables que tomará cada subset de datos para cada árbol es definido por la raíz cuadrada del número total de variables del set de datos o vista minable (Gislason et al., 2006).

El resultado de los algoritmos de minería de datos deben ser aplicados a través de un proceso de minería de datos, el cual en su estudio (Bhatia, 2019) lo ilustra en seis fases, también detalladas en la figura 4:

- Definición de problema: Se trata de definir los requerimientos y el objetivo del proyecto a implementar.
- Entendimiento de los datos: Empieza con la recolección de datos, donde se determina la relevancia de cada variable y sus distintas fuentes.
- Preparación de los datos: Tratamiento de los datos para prepararlos para el uso del algoritmo de minería de datos.
- Modelamiento: Elección y aplicación del algoritmo de minería de datos que contribuirá para la construcción del modelo.
- Evaluación: Pruebas de los resultados mostrados por cada modelo implementado.
- Despliegue: Puesta en producción del modelo elegido que será utilizado con datos nuevos.

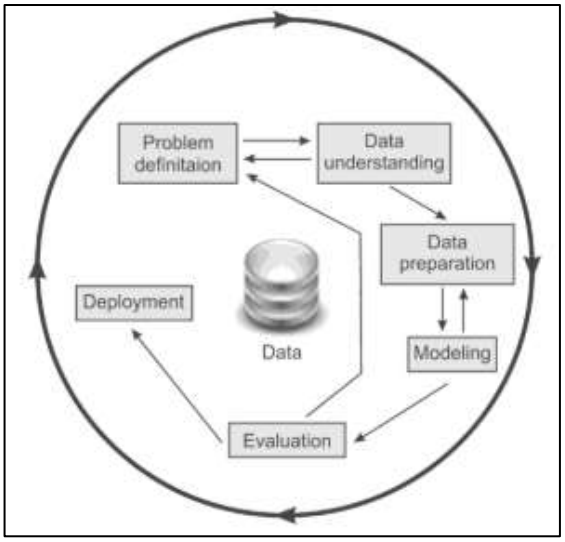

Figura 4. Proceso de minería de datos, tomado de (Bhatia, 2019)

De tal manera, para cumplir con este proceso se necesitan metodologías, que permitan identificar patrones y comportamientos que puedan ser útiles. Para ello se emplea la extracción del conocimiento, relacionado con el “proceso de descubrimiento en base de datos” conocido como Knowledge Discovery in Databases (KDD) (Camana, 2016). El cual es

definido como el proceso que combina descubrimiento y análisis que consiste en extraer patrones en formas de reglas o funciones, a partir de los datos, para que el usuario las analice (Pereira et al., 2016). Por otro lado, KDD es interpretado como el proceso de uso de la base de datos junto a acciones que constituyen a varias etapas para identificar patrones que pueden ser considerados como nuevos conocimientos. (Lagla et al., 2019).

Este proceso es interactivo e iterativo, resumido en 5 etapas: Selección, pre-procesamiento/limpieza, transformación/reducción, minería de datos e interpretación/evaluación (Pereira et al., 2016) las cuales serán descritas en el capítulo III del presente trabajo de titulación. Se muestra en la figura 5.

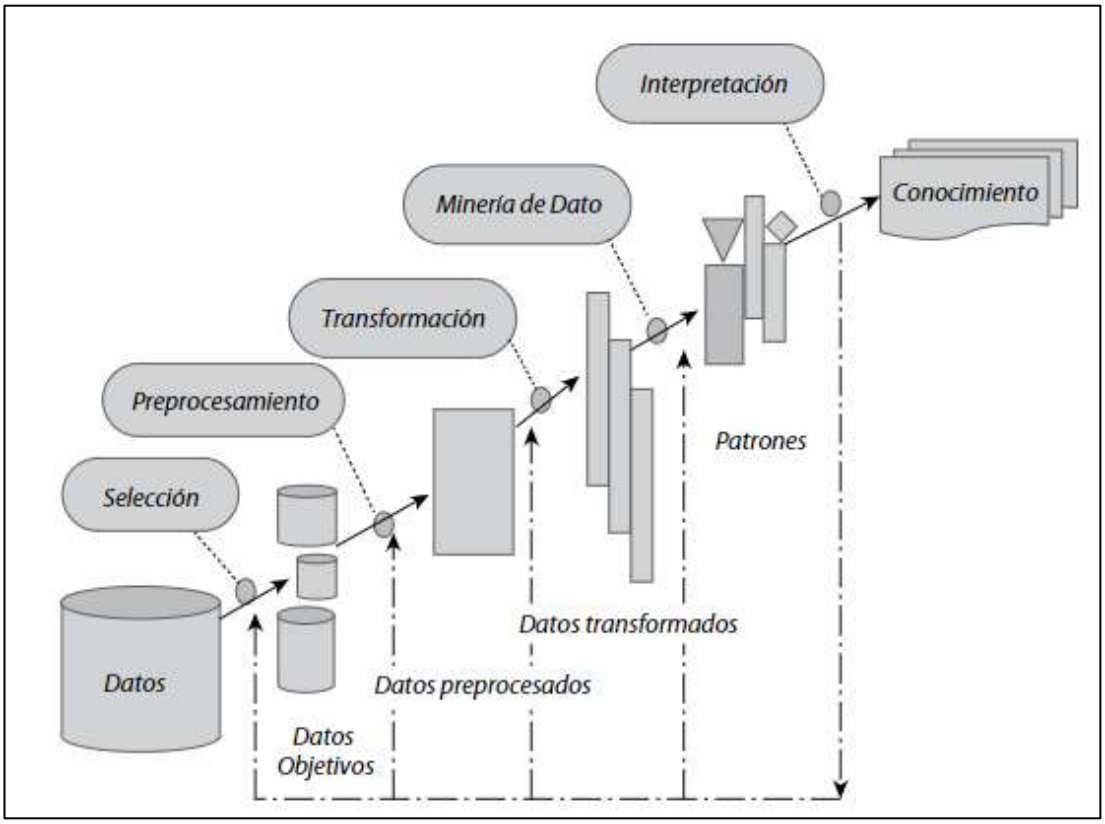

Figura 5. Etapas del proceso KDD, tomado de (Pereira et al., 2016)

Varias metodologías existen para el proceso de minería de datos, tales como KDD, SEMMA (proceso de selección, exploración, modelamiento de datos, evaluación) y CRISP-DM (proceso estándar, jerárquico, modelo de negocio) (Pacherres Gutiérrez, 2018). Las cuales son comparadas en la figura 6 a continuación.

KDD	SEMMA	CRISP – DM
Pre – KDD	----------	Comprensión del Negocio
Selección	Muestreo	Comprensión de los Datos
Preprocesamiento	Exploración	
Transformación	Modificación	Preparación de los Datos
Minería de Datos	Modelado	Modelado
Interpretación/Evaluación	Evaluación	Evaluación
Post KDD		Implementación

Figura 6. Cuadro comparativo metodología de minería de datos, tomado de (Pacherres Gutiérrez, 2018)

El proceso KDD será implementado en este trabajo de titulación al tratarse de una metodología más completa en relación a sus semejantes al contar con más etapas donde se analiza paso a paso la información hasta llegar al resultado. Se puede indicar que incluso las otras metodologías parten sus bases sobre el proceso KDD con ciertas variantes según sus necesidades.

El fraude y detección en servicios móviles en telecomunicaciones

En términos simples, en telecomunicaciones, cualquier actividad donde se obtenga un servicio sin intenciones de pagar por ello es fraude. En otras palabras, cualquier intento de beneficiarse de los servicios de las telecomunicaciones sin pagar por ello o pagar menos que el precio oficial es fraude en telecomunicaciones (Koi-Akrofi et al., 2019). Por otro lado, el fraude en telecomunicaciones es el hurto de los servicios de telecomunicación o el uso de los servicios de telecomunicación para cometer otras formas de fraude. (Nassau County Spin, 2018). Este último concepto se refiere a que se puede usar el servicio de las telecomunicaciones para afectar a otro tipo de grupos.

El fraude en las telecomunicaciones pueden ser clasificados en cuatro grupos (Gosset & Hyland, 2018):

- Fraude contractual: Esta categoría se trata del uso del servicio sin intenciones de pagar por ello. Entre ellos se identifican el fraude subscriptor.
- Fraude por hackeo: Este fraude es caracterizado por la violación a los sistemas de las empresas de telecomunicaciones aprovechándose de las brechas de seguridad para explotar o revender las funcionalidades. Entre ellos se identifica el hackeo de centrales PBX
- Fraude técnico: Todos los fraudes en esta categoría involucran ataques en contra las debilidades en la tecnología de los sistemas móviles. Este tipo involucra cierto conocimiento técnico y entre ellos se encuentra la clonación de celulares.
- Fraude en procedimientos: Los fraudes en esta categoría involucra ataques en contra los procedimientos empleados para minimizar el fraude, atacando las debilidades en los procedimientos definidos del negocio. Entre ellos tenemos fraude de roaming.

El fraude causa un impacto negativo en todos, incluyendo clientes finales, dado a que las pérdidas incrementan los costos de operación de las empresas de telecomunicaciones. En los últimos tiempos, han surgido nuevos tipos y métodos de fraude donde existe una necesidad de actualizar los conocimientos para combatir y evitar pérdidas económicas (*Communications Fraud Control Association*, 2019).

En el año 2017, a nivel mundial las pérdidas económicas de todos los tipos de fraude combinado en las telecomunicaciones llegan a $29,3 miles de millones clasificados en los distintos tipos de fraude como se muestra en la figura 7, destacando entre ellos el fraude de tipo IRSF (“Fraude de ingresos internacionales”) y Fraude por bypass (*Communications Fraud Control Association*, 2019).

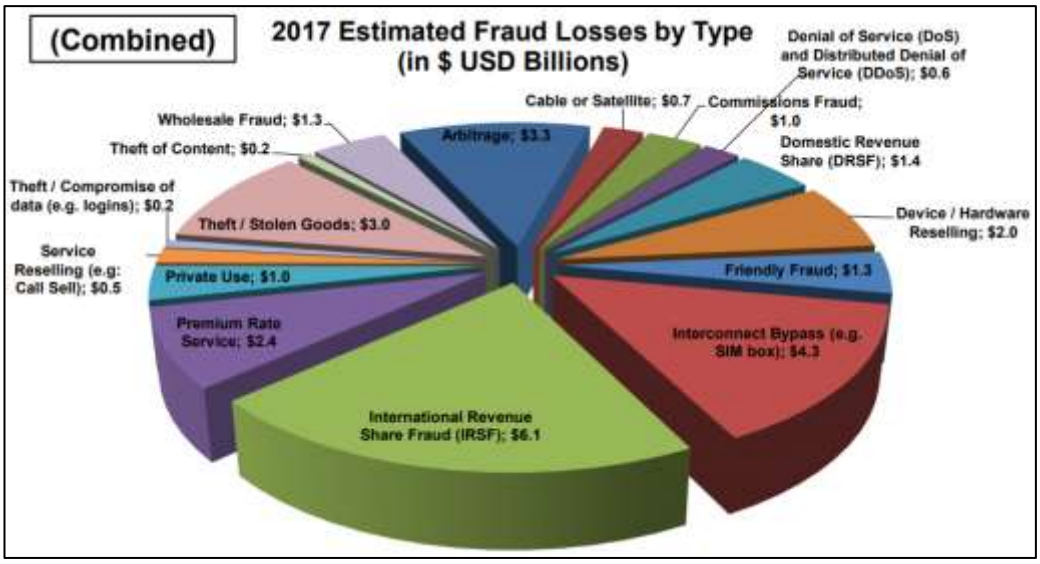

Figura 7. Pérdidas económicas por tipo de fraude en el 2017, tomado de (Communications Fraud Control Association, 2019)

Este último, fraude por bypass ha causado cuantiosas pérdidas para todas las operadoras a nivel del mundo y en el estado ecuatoriano. El mismo consiste en el procedimiento por el cual operadoras no autorizadas penetran llamadas entrantes internacionales en la red de una operadora legal como si fueran llamadas locales, evitando los cargos que cobran las mismas por prestar los servicios de larga distancia internacional (Camacho & González, 2016). Por otro enfoque, es la práctica de enrutamiento ilegal de tráfico telefónico que afecta a países donde el monto recibido por llamadas internacionales es un monto significativo de ingresos para las empresas de telefonía autorizada para cursar este tipo de llamadas (Román & Mauricio, 2008).

Inicialmente, el fraude por bypass se cometía con telefonía fija, hasta que los delincuentes notaron el uso de la telefonía móvil, por el cual se complicó la detección dado a la facilidad de movilidad de estos dispositivos. (Camacho & González, 2016). A pesar de esto, en el país se ha logrado detectar exitosamente ciertas centrales clandestinas donde se cursaba tráfico internacional no autorizado “bypass”, en donde la fiscalía y policía nacional ecuatoriana intervino desmantelando estos lugares, que afectaban económicamente a la operadora Movistar Ecuador (El Comercio, 2012).

En el Ecuador, el fraude por bypass se realiza en inmuebles pequeños, en sectores con alta demanda de servicios de telecomunicaciones con la finalidad de camuflar su presencia; además, con equipos de menor capacidad tecnológica y costos bajos en comparación de los equipos usados en las empresas de telecomunicaciones, los cuales son operados por pocas personas. La manera de ser rentable este tipo de fraude, es a través de negociaciones con los carriers de modo que se curse tráfico telefónico a menor valor que las telefónicas habilitadas para prestar el servicio (Román & Mauricio, 2008). Los personas que practican este fraude cursan las llamadas provenientes de destinos internacionales simulándolas como llamadas locales, de modo de que se cobren a un precio local impuestos por las operadoras locales de telefonía móvil (Camacho & González, 2016).

El bypass también es conocido como fraude por interconexión que utiliza rutas alternas para cursar las llamadas, evadiendo las rutas legales definidas por la operadora legal del país, con el fin de evadir las tasas legales de comercialización. Estas rutas alternas implican también un costo menor para cursar el tráfico telefónico debido al uso de equipos y enlaces de bajo costo, y por no pagar el valor de concesión del estado ni los impuestos que implica dar el servicio. Para este tipo de rutas no legales se utilizan los siguientes elementos de red de menor costo: modems, routers, Gateways, simcards, y simbox. Más relevante aún, el fraude bypass de tipo internacional entrante es el más importante en nuestro país debido al volumen de tráfico generado (Román & Mauricio, 2008).

En la figura 8 podemos apreciar una comparativa entre ruta legal y no legales por método bypass. Mientras que en la figura 9 se visualizan los equipos básicos para realizar bypass en una red móvil.

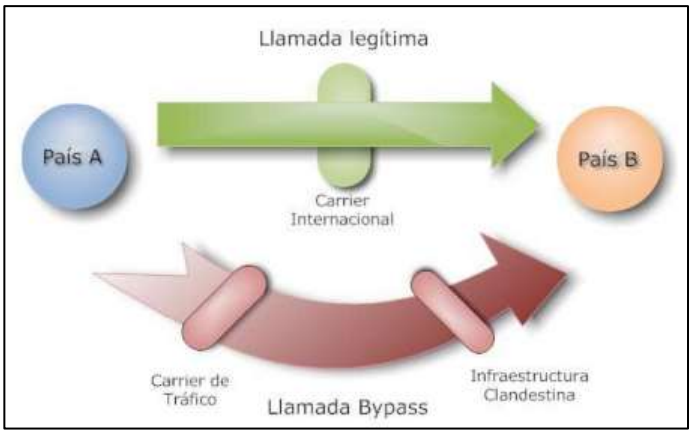

Figura 8. Llamada entrante internacional bypass, tomado de (Román & Mauricio, 2008)

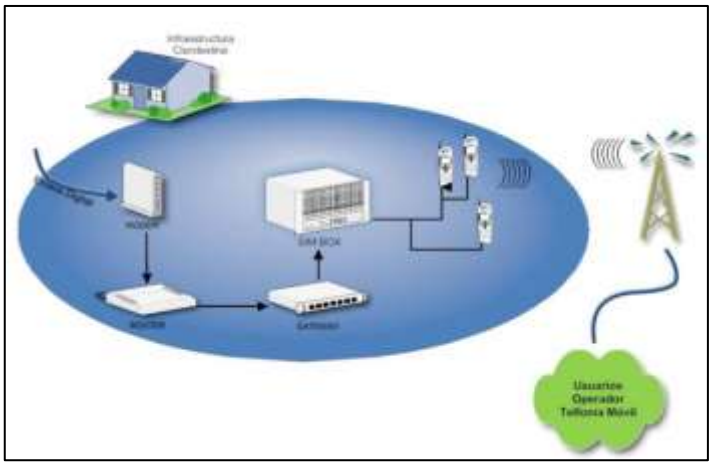

Figura 9. Equipos utilizados para realizar bypass, tomado de (Román & Mauricio, 2008)

El dispositivo simbox o Gateway GSM que contiene varias tarjetas SIMCARDS, une un enlace de internet con líneas telefónicas móviles. De modo que el fraude ocurre cuando una llamada internacional que proviene de un Gateway VoIP como paquetes de datos es transformado en señales de voz. En estos paquetes de datos viene información del número B, el cual es interpretado por el dispositivo simbox que realiza un remarcado automático al destino final, lo cual se muestra en la figura 10. En consecuencia, el tráfico es facturado como una llamada local en lugar de la internacional traduciéndose en pérdidas de ingresos para la empresa (Román & Mauricio, 2008).

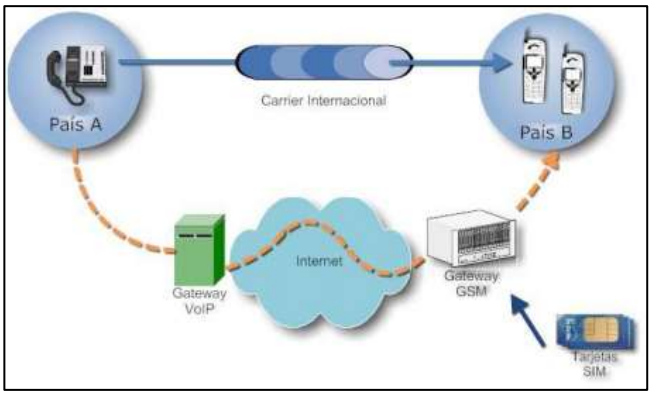

Figura 10. Bypass utilizando un dispositivo simbox, tomado de (Román & Mauricio, 2008)

Existen varios mecanismos de detección empleados de acuerdo al tipo de fraude en telecomunicaciones de los cuales se mencionan (Reaves et al., 2015):

- Llevar un registro de los equipos (IMEIS) utilizados para fraude.
- Identificar celdas donde se realiza el fraude.
- Llevar una bitácora de casos anteriores para buscar relaciones.
- Monitorear en tráfico local como el de larga distancia internacional.
- Estudiar las características físicas de los equipos utilizados para fraude.
- Verificación de información comercial de los clientes.
- Evaluar los reclamos ingresados por los clientes.
- Realizar pruebas de calidad de servicio desde el exterior y localmente.

Con respecto al fraude del tipo bypass se intenta la detección de dos frentes tanto a los casos de simcards empleadas para realizar bypass y la ubicación de estas infraestructuras clandestinas. Para ello el análisis es basado en una revisión y monitorio minucioso de lo obtenido en los CDRs (registros de llamadas) donde se tratan de observar patrones o comportamientos inusuales que sugieran este tipo de fraude, basados en indicadores como (Elmi et al., 2013):

- Incremento considerable de llamadas locales y la vez decremento de las llamadas internacionales entrantes.
- Simcards que solo realizan llamadas y no reciben ninguna.
- Llamadas prolongadas en horarios inusuales.
- Grado de interacción entre celdas muy bajo (movilidad).

- Mayor dispersión en entre las llamadas salientes y números llamados.
- Marcación de primera llamada a números especiales o número en particular.

Todos estos lineamientos serán tomados en cuenta para el objeto de estudio en este trabajo de titulación, tanto a la alineación simulada de registros de llamadas (CDRs) y para la construcción del modelo predictivo de detección de casos de fraude bypass.

Sin duda, es relevante para una empresa de telecomunicaciones contar con un área de aseguramiento de ingresos que a través de un equipo calificado se dispongan de métodos para la detección en este tipo de fraudes. Más allá, se conceptualiza el aseguramiento de ingresos como la disciplina asociada a empresas de telecomunicaciones que tiene como objetivo identificar y eliminar las causas tecnológicas o de procesos que dan origen a fugas de ingresos (Castro Aguilar et al., 2016). De hecho, esta disciplina refuerza sus conocimientos constantemente a través de la Asociación Global de Profesionales de Aseguramiento de Ingresos GRAPA, que promueve las mejores prácticas donde se resaltan la maximización de ingresos y la reducción de costos como pilares básicos de la disciplina (Asociación Global de Profesionales de Aseguramiento de Ingresos -GRAPA, 2020). Y a su vez, asociaciones como TMForum quienes proponen una transformación digital a las empresas proveedoras de comunicaciones a través de un ambiente colaborativo en el ámbito tecnológico, entre ellos el aseguramiento de ingreso, con el afán de mantenerse a la vanguardia del mundo digital (*TM Forum*, 2020).

Refiriéndose particularmente al fraude por bypass en las telecomunicaciones, TM Forum enfatiza el uso del análisis de los CDRs a través de técnicas de minería de datos para mejorar la detección de este tipo de fraude (TM Forum, 2015). Por su parte, GRAPA con el objetivo de mantenerse actualizados se organizan foros regionales donde se comparten información y métodos sobre la detección de fraudes en las telecomunicaciones. (GRAPA 2020).

MARCO CONCEPTUAL

A continuación se hace referencia a algunos términos utilizados en este trabajo de titulación que fueron de gran importancia para la construcción del modelo predictivo y para la comprensión del ámbito de estudio. Entre ellos tenemos:

Los **registros de llamadas** (Conocidos como “Call Data Record” o CDR por sus siglas en inglés) es un registro sobre las llamadas realizadas por los usuarios de una telefonía móvil. Estos registros contienen información como la duración, el origen, el destino de las mismas (Camacho & González, 2016). Adicional, se definen como registros de actividad de los usuarios de una red celular, los cuales contienen las coordenadas geográficas en tiempos aproximados de un usuario por cada llamada realizada la cual conserva cada empresa de telecomunicaciones para propósitos de cobro y de estudios (Ficek & Kencl, 2012).

Dentro de estos CDRs se puede encontrar información de **celdas telefónicas**, definidas como el área de cobertura definida para transmisores y receptores dentro de una misma estación base. Una red celular se encuentra dividida en varias de estas celdas, las cuales están representadas de manera hexagonal conteniendo un grupo único de frecuencias asignadas. Físicamente están representadas como un torre celular y un cuarto pequeño que contiene los equipos de transmisión de radio (Iguasnia et al., 2017). A través de estas celdas que tienen asignadas un identificador único se puede conocer una ubicación aproximada de donde se realiza un evento de conexión a la red como una llamada.

Otros términos importantes son las simcards, simbox e IMEI. La **simcard** es la tarjeta de módulo de identidad del subscriptor (Subscriber Identity Module Card) que contiene información de la subscripción del usuario de una red móvil desde protocolos básicos de conexión hasta los contactos del subscriptor, la cual se inserta un equipo terminal con el que se requiere servicio en la red de una empresa de telecomunicaciones (Naranjo et al., 2016). Actualmente estas tarjetas han venido disminuyendo su tamaño para adaptarse a las nuevas tecnologías y diseños de las nuevas gamas de

teléfonos móviles. Mientras que un dispositivo **Simbox**, es un equipo que permite conectar más de una simcard a la vez, que por lo general es usado para finalizar llamadas internaciones transformándolas a llamadas locales de modo fraudulento. Por su parte, un **IMEI** (International Mobile Station Equipment Identity) es un código único que cada terminal móvil posee para ser identificado a nivel mundial constituido de 15 dígitos donde los 8 primeros identificar el modelo del terminal. (Duarte, 2017). A través de este código se puede identificar los equipos simbox que hayan sido utilizados para cometer fraudes de tipo bypass en una red móvil.

Con lo que respecta a la minera de datos, es relevante aclarar ciertos conceptos. Un **set de datos** es una colección de información con una estructura definida por atributos que pueden ser numéricos o continuos, y categóricos o nominales. Estos set de datos están constituidos por registros que fueron previamente tratados con diferentes técnicas donde es usada generalmente la estadística descriptiva, con el fin de poderse identificar patrones representativos. **Outliers o datos atípicos** son anomalías en un set de datos, por lo general de gran tamaño. Estos datos deben ser comprendidos y requieren un tratamiento especial, donde comúnmente son eliminados de modo que se obtenga una información más representativa que permita generalizar un patrón (Kotu & Deshpande, 2014).

Adicional, es necesario realizar una comparativa entre los procesos de minería de datos y **Aprendizaje de máquinas (Machine Learning)**, siento estos procesos que van de la mano teniendo un rol importante en analizar y entender los datos con el fin de apoyar a la toma de decisiones en un modelo de negocios. A pesar de ello, estos conceptos cuentan con ciertas diferencias entre su significado, objetivo y naturaleza detalladas en la tabla 1:

Tabla 1. *Comparativa entre Minería de datos y Machine Learning*

	Minería de datos	Machine Learning
Significado	Representa extraer útil conocimiento de grandes cantidades de información	Introduce nuevos algoritmos de información basada en experiencias pasadas.
Objetivo	Examinar patrones en la información existente, los cuales pueden ser usados para definir reglas.	Enseñar a la computadora a aprender y entender las reglas definidas.
Naturaleza	Requiere de interacción e intervención humana para su desarrollo.	Es automático una vez que es diseñado se auto implementa donde es requerido poco esfuerzo humado.

Nota. Fuente de (Bhatia, 2019). Elaboración propia.

En base a lo planteado, este trabajo de titulación se basa solamente en el proceso de minería de datos al requerirse una interacción humana debido al problema planteado donde es necesario entender y preparar la información e interpretar patrones de comportamiento, específicamente el comportamiento bypass.

MARCO LEGAL

En este apartado se indicarán las leyes, normas y reglamentos que rigen en el Ecuador en donde se enmarcan los datos e investigaciones del presente trabajo de titulación.

En el país, el servicio de telecomunicaciones se encuentra administrado a través de concesiones, amparado en la Constitución de la República del Ecuador (2008) indica:

> Art. 249.- Será responsabilidad del Estado la provisión de servicios públicos (…) telecomunicaciones y otros de naturaleza similar. Podrá prestarlos directamente o por delegación a empresas mixtas o privadas, mediante concesión, asociación, capitalización, traspaso de la propiedad accionaria o cualquier otra forma contractual, de acuerdo

con la ley. Las condiciones contractuales acordadas no podrán modificarse unilateralmente por leyes u otras disposiciones (…).

Particularmente los servicios telecomunicaciones de telefonía móviles actualmente se encuentran con una concesión otorgada por El Estado a dos empresas privadas y una pública. Siendo estas las únicas empresas autorizadas para cursar tráfico móvil internacional y nacional en el país.

Cuando uno de estos tres operadores telefónicos móviles autorizados detecta tráfico telefónico irregular en su red puede acogerse a causales dispuestos en el Reglamento de Interconexión que les permite realizar la suspensión de la comunicación de estos servicios, basados en el art. 49 de este Reglamento el cual indica (CONATEL, 2007a):

> Art. 49.- Causales para la desconexión.- Una vez registrado el acuerdo de interconexión por la Secretaría Nacional de Telecomunicaciones, la interconexión entre redes públicas sólo podrá ser interrumpida o terminada de conformidad con las causales establecidas en los respectivos acuerdos de interconexión, previa comunicación enviada a la Secretaría Nacional de Telecomunicaciones y autorización de la Superintendencia de Telecomunicaciones.

Entre estos causales se encuentran las actividades ilícitas como el fraude bypass, donde cada operadora móvil está en la potestad de suspender los servicios, confirmado dentro la Ley Orgánica de Telecomunicaciones (2015) que indica:

> Artículo 25.- Derechos de los prestadores de servicios de telecomunicaciones. Son derechos de los prestadores de servicios de telecomunicaciones, con independencia del título habilitante del cual se derive tal carácter, los siguientes: (…) 2. Suspender el servicio provisto por falta de pago de los abonados o clientes con uso ilegal del servicio calificado por autoridad competente, previa notificación al abonado o cliente.

La reafirmación que las únicas operadoras autorizadas para prestar servicios telefónicos de larga distancia internacional son las que cuentan una

concesión con el estado, así como la prohibición del reoriginamiento o enmascaramiento de este tipo de tráfico quedan expuestos queda reforzando en el Reglamento del Servicio Telefónico de Larga Distancia Internacional que en sus artículos 4 y 13 mencionan (CONATEL, 2007):

> Art. 4.- La autorización para la prestación y explotación del Servicio Telefónico de Larga Distancia Internacional (STLDI), será parte integrante de un contrato de concesión suscrito con la Secretaría Nacional de Telecomunicaciones, por autorización previa del CONATEL, para la prestación del servicio final de telefonía fija o de servicios móviles, de conformidad con el ordenamiento jurídico vigente.
>
> Art. 13.- Se prohíbe expresamente el reoriginamiento o enmascaramiento del tráfico internacional entrante o saliente con los operadores extranjeros o entre operadores nacionales con los cuales se mantengan relaciones de interconexión.

Por lo consiguiente, actividades como el fraude bypass donde un servicio es ofrecido sin estar autorizado a realizar esta actividad queda legalmente prohibido con sanciones de privación de libertad de uno a tres años tal como se indica en el artículo 188 del Código Orgánico Integral Penal del Ecuador (2014):

> Artículo 188.- Aprovechamiento ilícito de servicios públicos.- (…) La persona que ofrezca, preste o comercialice servicios públicos de luz eléctrica, telecomunicaciones o agua potable sin estar legalmente facultada, mediante concesión, autorización, licencia, permiso, convenios, registros o cualquier otra forma de contratación administrativa, será sancionada con pena privativa de libertad de uno a tres años.

Las acciones de allanamientos a instalaciones clandestinas donde ya fue detectado actividades ilícitas por fraude bypass están amparadas dentro de los artículos 35 y 194 del Código de Procedimiento Penal del Ecuador (2001):

> Art. 35.- Actos urgentes.- En los casos de acción pública, la Fiscal o el Fiscal podrá realizar los actos urgentes que impidan la consumación

del delito o los necesarios para conservar los elementos de prueba pero sin afectar los derechos del ofendido.

Art. 194.- Casos.- La vivienda de un habitante del Ecuador no puede ser allanada sino en los casos siguientes:
(…) 3. Cuando se trate de impedir la consumación de un delito que se está cometiendo (…).

A través de la aplicación de estos artículos se ha permitido la detención de personas tras los allanamientos de varios casos en el país realizando actividades fraudulentas del tipo bypass, como los que se expusieron anteriormente en este trabajo de titulación.

De los aspectos legales en este apartado, es necesario recalcar que es considerado ilegal generar, cursar, o enmascarar tráfico internacional sin autorización, considerado dentro de ello el fraude bypass dado a su naturaleza de hacer pasar tráfico internacional como local sin autorización para realizar enriquecimiento ilícito aprovechándose de los servicios de telecomunicaciones legalmente autorizados en el país. Además, que este tipo de delitos es penado con privación de libertad desde uno a tres años, y las empresas de telecomunicaciones están en su facultad de suspender los servicios de casos detectados como bypass amparados en los causales por uso inapropiado del servicio según lo estipula la ley. Así además de los posibles allanamientos una vez triangulado la ubicación de las instalaciones ilícitas de modo que sean capturados en delito flagrante.

Dado a la gravedad expuesta en estos artículos del cometimiento de estas actividades ilícitas en el Ecuador, es importante mejorar las detecciones de este tipo de fraudes que perjudican a las empresas privadas de telecomunicaciones móviles autorizadas y al estado ecuatoriano al generar cuantiosas pérdidas como las ya expuestas con anterioridad. La aplicación de técnicas como la minería de datos como la propuesta en el presente trabajo de titulación contribuirá a la detección temprana de este tipo de fraude bypass.

CAPÍTULO III
METODOLOGÍAS Y RESULTADOS

En este capítulo, se describirán las metodologías a implementar para obtener la información de modo que permitan cumplir el desarrollo de los objetivos planteados en el presente trabajo de titulación.

El presente capitulo se encuentra dividido en tres apartados, en el primero se presenta la metodología de investigación a usar, para luego mostrar en el segundo la metodología de minería de datos y por último la aplicación de los instrumentos de recolección de datos tomadas.

METODOLOGÍA DE LA INVESTIGACIÓN

Dentro de los enfoque de la investigación tenemos el cualitativo definido como subjetivo donde predomina la interpretación de los hechos basados en las opiniones de la gente, patrones culturales e interacción con el medio. Y por otro lado, el cuantitativo, el cual parte de un problema y objetivos definidos donde se plantea una hipótesis que a través de técnicas estadísticas para el análisis de la información buscará comprobarla o descartarla. Para ello se emplean varios instrumentos de recolección y mediciones de variables estructuradas (González, 2016).

El presente trabajo de titulación tiene un enfoque cuantitativo debido a que la investigación parte de una problemática enmarcada en el fraude bypass en el Ecuador, definiendo objetivos de diseñar un modelo predictivo a través del uso de minería de datos que permita llegar a conclusiones con el fin de comprobar o descartar la hipótesis que se plantea sobre si este modelo contribuirá a la eficacia en la detección de casos bypass en el país.

El enfoque cuantitativo permite el uso de varios tipos de investigación. En efecto, para el presente trabajo se usará según su profundidad la investigación exploratoria dado a que se tratará de encontrar patrones representativos en los datos que serán analizados y a través de la interpretación de los resultados plantear pautas sobre la detección de casos bypass a través de técnicas de minería de datos.

Definiéndose el uso de la investigación exploratoria en este trabajo dado al concepto que enmarca analizar e investigar aspectos concretos que aún no han sido analizados en profundidad, ofreciendo a investigaciones futuras una guía para realizar análisis en el tema tratado (Torres, 2015).

A través del uso del método deductivo se podrá llegar a emitir conclusiones particulares en base a la observación del comportamiento general de los datos de registros de llamadas analizados. Lo anterior basado en la definición del método deductivo que indica el estudio de un conjunto de premisas o proposiciones para extraer conclusiones lógicas particulares (Martínez, 2014).

Una vez definidos estos parámetros en la presente investigación, se logrará un estudio comprensible, con las respuestas esperados. Esto debido al seguimiento que aportan a los distintos niveles de desarrollo, de modo que se cumplan los objetivos planteados de una manera adecuada.

Para dar cumplimiento a cada uno de los objetivos planteados en este trabajo de titulación se planteará un seguimiento de cada aporte en las diferentes partes del desarrollo de la investigación que brinda el enfoque, tipo y método de investigación planteados, apoyando por las técnicas de recolección de información que en este caso serán:

- El análisis de contenido a través de documentación basada en fraudes de tipo bypass en telefonías móviles como en la aplicación de minería de datos a través del algoritmo de bosques aleatorios o random forest ya revisados en el capítulo II del presente trabajo.
- Entrevistas llevadas a cabo con expertos en telefonía y redes móviles donde se discutirán acerca de patrones de fraude, análisis de CDRs, casos de detección, etc.

Vale indicar que este proyecto no utilizará como técnica de recolección de información a la encuesta por lo que no estará definido una población y muestra. Lo anterior se debe a la naturaleza de la investigación basada en la revisión y análisis de datos encontrados en los registros de llamadas cifrados.

METODOLOGÍA DE MINERÍA DE DATOS

La metodología KDD (“Knowledge Discovery in Databases) será la utilizada en esta investigación debido a ser un proceso interactivo e iterativo, que aplica la minería de datos dentro de sus fases de implementación, además de especializarse en set de datos de gran tamaño con algoritmos de aprendizaje y clasificatorios muy eficientes (Guzmán, 2016), esenciales para la construcción de modelos predictivos que consideren todas las variables de estudio y manejar gran cantidades de información sobre el registro de llamadas como el deseado en este trabajo de titulación.

Dentro de los beneficios por el cual este proyecto apunta a trabajar con esta metodología es por su versatilidad al manejar múltiples variables de entrada sin exclusiones dándoles su respectiva importancia según el caso de estudio (Guzmán, 2016).

Estas etapas del proceso KDD se describen a continuación (Camana, 2016):

En la etapa de selección de datos se recopila los datos relevantes y se identifica el tipo y fuente de la información para ser posteriormente usada en el pre-procesamiento, esto de acuerdo a los objetivos y metas definidos para el estudio.

La etapa de pre-procesamiento/limpieza consiste en una exploración y corrección donde se evalúa la calidad de los datos, esto debido a que la información suele venir de distintas fuentes que requieren ser tratadas para ser consolidada. Además se discriminan ciertos registros que no son relevantes para el objeto de estudio y los registros nulos o duplicados.

La etapa de transformación/reducción consiste en la construcción de la vista minable, es decir, los datos sufren una transformación a partir de la información existente, así como la creación de nuevas variables e incluso la extracción de variables que no aportan al estudio. Además, se realiza la normalización de datos para la posterior etapa.

La etapa de minería de datos consiste en la búsqueda de patrones de interés a través del uso de diferentes algoritmos de minería de datos, en función de encontrar la solución al problema planteado. Estos pueden ser predictivos o descriptivos de acuerdo a la problemática.

En la etapa de interpretación/evaluación se evalúan los patrones obtenidos y en el caso de ser necesario se itera nuevamente a etapas previas. Además se puede incluir la visualización de los patrones obtenidos y la traducción de estos para que sea entendible para el usuario final.

INSTRUMENTOS DE RECOLECCIÓN DE DATOS

En el presente trabajo de titulación se utilizó como herramienta de recolección de datos a la entrevista del tipo abiertas dado a que es prioritario recabar la mayor cantidad de información específica del comportamiento de casos de fraude bypass en redes móviles y además de análisis de CDRs (registros de llamadas) a realizarse para la detección del fraude.

RESULTADOS Y ANÁLISIS DE LA ENTREVISTA

Las entrevistas fueron realizadas a dos especialistas del área del control de riesgo de una empresa de telecomunicaciones del país, actualmente cesantes. Las mismas fueron realizadas de manera personal se desarrolló una guía que fue entregada previamente a los entrevistados.

Tabla 2. *Codificación empleada para los entrevistados*

Entrevistado	Tema tratado	Codificación
Especialista 1. Ingeniero en sistemas computaciones.	Fraude por bypass	E1
Especialista 2. Ingeniero en estadística.	Análisis de CDRs (Registros de llamadas)	E2

Nota. Para facilitar la manipulación de la información en la investigación se codifican los entrevistados (Entrevistado = E). Elaboración propia.

Como se indica en la tabla 2, cada entrevistado tiene su tópico en el que cuenta con más experiencia, es por ello que se trató específicamente el tema de acuerdo a su especialidad y conocimiento obtenido de cada uno de ellos.

Tabla 3. *Guía de entrevista para cada especialista*

Pregunta	**E1**	**E2**
P1	¿Cómo se caracteriza el comportamiento de los casos de fraude bypass?	¿Cómo se caracteriza el comportamiento de los casos de fraude bypass y clientes normarles?
P2	¿Cuál es la ocurrencia de este tipo de fraudes?	¿Qué tipos de servicios móviles son mayormente detectados como casos de fraude bypass?
P3	¿Qué criterios de detección o métodos son utilizados actualmente en el país?	En breve ¿En qué consiste el proceso de perfilamiento?
P4	¿Se utilizan actualmente procesos de minería de datos para la detección de estos casos?	¿Se utilizan actualmente procesos de minería de datos para la detección de estos casos?

Nota. Para facilitar la manipulación de la información en la investigación se codifican las preguntas (P = pregunta). Elaboración propia.

Detalle de entrevistas

Como conclusiones de las entrevistas realizadas, se destacan los siguientes puntos:

1. El comportamiento de un caso de fraude bypass está regido a altos consumos en llamadas locales en un periodo corto de tiempo.
2. Anteriormente, las ocurrencias de este tipo de fraude se presentaban en la madrugada, sin embargo, desde la implementación de procesos automáticos de detección ocasionó que la ocurrencia se disperse a

cualquier momento del día, sin remarcarse un momento del día en particular. Incluso algunos casos ocurren en la mañana/tarde para enmascararse como servicios corporativos que realizan sus actividades en jornada laboral.

3. Los servicios más recurrentes para este tipo de fraude es el prepago, dado a la facilidad de obtención del servicio, pudiendo ser adquirido el cualquier local comercial.
4. Dentro de los criterios usados para la detección se analizan eventos como el aumento de llamadas locales en comparación con la disminución de llamadas internacionales, servicios que realicen un gran número de llamadas pero que nunca reciban una, una baja movilidad tras una conexión a una misma celda todo el tiempo, y así como altos consumos en un corto periodo de tiempo.
5. Se conoció acerca del concepto de dispersión que consiste en un valor porcentual obtenido de la división entre la cantidad de números llamados sobre la cantidad total de llamadas realizadas. Para los casos de fraude mayormente indicaron que el valor de dispersión es más cercano a uno.
6. Se trató acerca del concepto de perfilamiento, que consiste en estudiar el comportamiento del uso del servicio de cada uno de los clientes con el afán de monitorearlos y en su defecto se puedan asociar con casos fraudulentos.
7. Se reconoció que al momento no se están utilizando en sus operaciones, métodos que involucren técnicas de minerías de datos como la tratada en el presente trabajo de titulación.

El resultado de estas entrevistas contribuyó al modelamiento de la información utilizada para la construcción del set de datos minable, llevándola gracias a la ayuda de los entrevistados a un contexto nacional con respecto al comportamiento del consumo, tanto para servicios considerados como normales y también los casos de fraude bypass.

CAPÍTULO IV
PROPUESTA

El capítulo IV presenta la construcción del modelo predictivo propuesto, a través de la implementación de la metodología KDD con el uso de los algoritmos de bosques aleatorios o “Random Forest” como técnica de minería de datos. Se describirá como se fue desarrollando el proceso en cada una de sus etapas, así como sus resultados a través del uso de la herramienta de minería de datos escogida tras el respectivo análisis según sus características expuesto a continuación.

SELECCIÓN DE LA HERRAMIENTA DE MINERÍA DE DATOS PARA LA IMPLEMENTACIÓN DE METODOLOGÍA KKD

La elección de la herramienta de minería de datos se realizó a través de la elaboración de un benchmark de acuerdo a 5 diferentes herramientas activas en el mercado donde se analizaron parámetros como el tipo de software (libre o pagado), las plataformas que soportan, los algoritmos en los que se especializan, la disponibilidad de foros de ayuda oficiales que se encuentren activos, si cuentan con una interfaz gráfica que facilite el desarrollo y por último si cuentan con una herramienta de pre visualización de resultados.

Tabla 4. *Comparación de herramientas de minería de datos.*

HERRAMIENTA	TIPO DE SW	PLATAFORMA	ESPECIALIDAD	FORO ACTIVO	INTERFAZ GRAFICA	PRE VISUALIZACION
RapidMiner	PAGADA	2	Redes neuronales/Clustering/Machine Learning	SI	SI	SI
Weka	LIBRE	3	Machine Learning	NO	SI	NO
Knime	LIBRE	3	Segmentación/Arboles de decisión/Random Forest/SVM/Redes neuronales	SI	SI	SI
SPSS	PAGADA	2	Arimas/Arboles de regresión/Clustering/Regresiones	SI	SI	SI
JHepWork	LIBRE	3	Redes neuronales/Regresiones/Ecuaciones paramétricas	NO	SI	NO

Nota. Información tomada de las páginas oficiales de cada una de las herramientas. Elaboración propia.

La herramienta de minería de datos seleccionada para la desarrollo de la metodología KDD en este trabajo de titulación fue KNIME, al tratarse de un software de uso libre que cuenta soporte multiplataforma, un foro oficial activo donde se pueden realizar consultas/discusiones acerca de implementaciones, interfaz gráfica interactiva, y la pre visualización de los resultados, entre otras características que sus similares difieren. En especial, KNIME es especialista en la implementación de árboles de decisiones y bosques aleatorios “Random Forest” tal como se puede apreciar en la tabla 4.

Adicional, con respecto al uso y habilidad de ejecución de las herramientas de minería de datos, KNIME destaca como líder según cuadrante mágico de Gartner según la empresa Gartner (2019):

Figura 11. Cuadrante mágico de Gartner para plataformas de ciencia de datos. Tomado de (Gartner, 2019)

DESARROLLO DE LA METODOLOGIA

1. SELECCIÓN DE DATOS

Para la obtención de la información se procedió con la simulación del comportamiento de servicios móviles mediante la obtención de una base de datos de registros de llamadas del repositorio “Dataverse” perteneciente a la Universidad de Harvard (2019). A través de esta base se obtuvieron 1’609.108 millones de registros de llamadas realizados por 22.800 servicios móviles en un periodo de 45 días. Los mismos fueron adaptados a la realidad del Ecuador siguiendo los lineamientos del comportamiento de líneas consideradas como bypass y líneas de uso normal discutidas con los especialistas y estudiadas en el marco teórico del presente trabajo de titulación. Dado al gran volumen de información, esta fue dividida en dos archivos con formato xls (Formato de Excel). Para la comprensión de los campos que contiene esta base de datos se presenta tabla 5 con la descripción de cada uno de ellos a continuación.

Tabla 5. *Diccionario de datos de base de registros de llamadas.*

Campo	Descripción
Fecha	En formato aaaa-mm-dd se registra la fecha en que se realiza una llamada.
numero_A	Se identifica como el número saliente, es decir el que origina la llamada.
numero_B	Se identifica como el número entrante, es decir como el que recepta la llamada.
celda_I	Se trata de la celda de donde se origina la llamada. Hace referencia la ubicación.
celda_O	Se trata de la celda donde se recepta llamada. Hace referencia la ubicación.
Duracion	Expresada en segundos la duración de la llamada realizada.

tiempo_inicio	En formato hh:mm:ss se registra la hora en que se realiza la llamada.
tiempo_fin	En formato hh:mm:ss se registra la hora en que finaliza la llamada.
codi_pais	Código del país al cual es dirigida la llamada. Valor ‘LO’ hace referencias a llamadas hacia destinos locales.
desc_dest	Descripción del destino al cual es dirigida la llamada.
tipo_cdr	El tipo de registro de llamadas que pueden ser del tipo saliente (1) y entrante (2).
prod_I	Se identifica al tipo de producto el cual origina la llamada, es decir si es un servicio prepago (PPA) o pospago (POS).
prod_O	Se identifica al tipo de producto el cual origina la llamada, es decir si es un servicio prepago o postpago.
central_I	Se trata de la celda de donde se origina la llamada.
Imsi	Se trata de la IMSI de donde se origina la llamada.
Imei	Se trata del IMEI (Equipo) de donde se origina la llamada.
Bypass	Se identifica a la llamada como comportamiento bypass (1) o normal (0)

Nota. Elaboración propia.

Así mismo, fue necesario para las siguientes etapas de la metodología realizar una definición de las variables con la que cuenta la base de datos original, de modo que sea más factible su lectura en el posterior análisis. Lo indicado se resume en la siguiente tabla.

Tabla 6. *Tipos de variables de base de registros de llamadas.*

Campo	Tipo de dato	Escala	Tipo de Variable
Fecha	Date time	Cuantitativa	Continua
numero_A	String	Cualitativa	Nominal
numero_B	String	Cualitativa	Nominal
celda_I	String	Cualitativa	Nominal
celda_O	String	Cualitativa	Nominal
Duracion	Number	Cuantitativa	Discreta
tiempo_inicio	String	Cuantitativa	Continua
tiempo_fin	String	Cuantitativa	Continua
codi_pais	String	Cualitativa	Nominal
desc_dest	String	Cualitativa	Nominal
tipo_cdr	Number	Cuantitativa	Categórica
prod_I	String	Cualitativa	Nominal
prod_O	String	Cualitativa	Nominal
central_I	Number	Cuantitativa	Categórica
Imsi	String	Cualitativa	Nominal
Imei	String	Cualitativa	Nominal
bypass	Number	Cuantitativa	Categórica

Nota. Elaboración propia.

En esta etapa de la metodología, se procede con la carga de la información de la base de datos a la herramienta KNIME. Para ello se hace uso del nodo denominado ‘Excel Reader (xls), el cual lee los datos del archivo

seleccionado y los carga en una tabla dentro del software. Debido al gran volumen de información, la información se encuentra dividida en dos archivos, razón por la cual se utilizó el nodo ‘Concatenate’, el mismo une la información de los dos archivos en una sola tabla para su posterior uso.

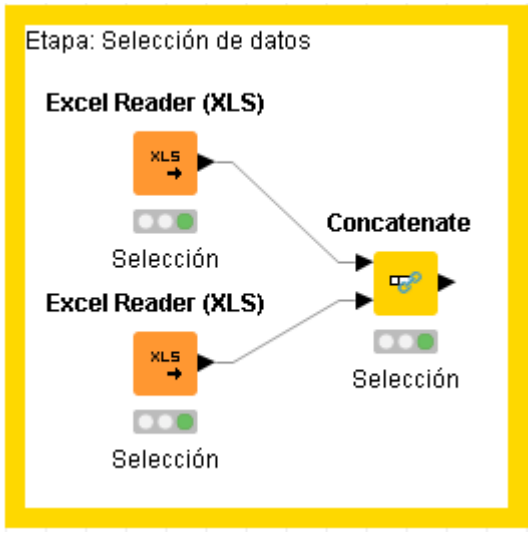

Figura 12. Etapa de selección de datos. Tomado de la herramienta KNIME. Elaboración propia.

De tal manera que los datos son visualizados de forma preliminar en la herramienta KNIME de la siguiente manera.

Columns: 17	Column Type	Column Index
FECHA	Local Date Time	0
NUMERO_A	String	1
NUMERO_B	String	2
CELDA_I	String	3
CELDA_O	String	4
DURACION	Number (inte…	5
TIEMPO_INICIO	String	6
TIEMPO_FIN	String	7
CODI_PAIS	String	8
DESC_DEST	String	9
TIPO_CDR	Number (inte…	10
PROD_I	String	11
PROD_O	String	12
CENTRAL_I	Number (inte…	13
IMSI	String	14
IMEI	String	15
BYPASS	Number (inte…	16

Figura 13. Estructura de la base de datos original. Tomado de la herramienta KNIME. Elaboración propia.

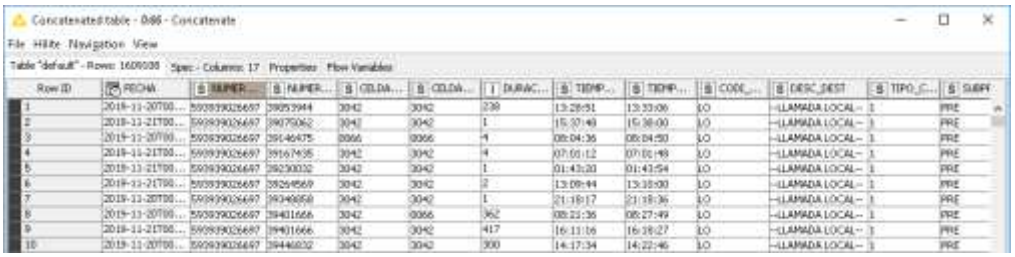

Figura 14. Previsualización de la tabla cargada en el nodo Concatenate. Tomado de la herramienta KNIME. Elaboración propia.

2. PRE-PROCESAMIENTO/LIMPIEZA

Una vez que la información ya se encuentra cargada en la herramienta, se procedió con el pre-procesamiento y limpiezas requeridos en esta etapa de la metodología.

Como primer paso en esta etapa, se realizó una detección de registros duplicados a través del nodo ‘Duplicate Row Filter’, donde todos los campos fueron seleccionados para realizar la revisión y en el caso de hallar casos de duplicidad se configuró el nodo para que los remueva de la tabla.

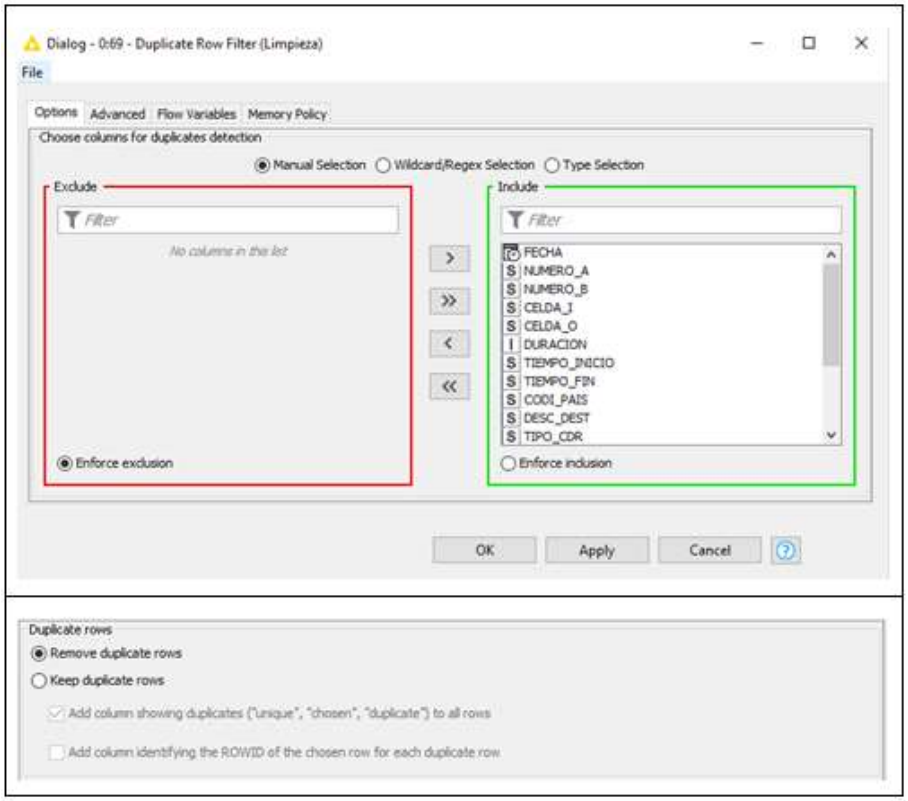

Figura 15. Configuración del nodo 'Duplicate Row Filter'. Tomado de la herramienta KNIME. Elaboración propia.

Luego de ello, se realizó la revisión de valores nulos o vacíos a través del nodo ‘Missing Value’. Así mismo como estrategia de configuración y siguiendo los fundamentos de esta etapa, fueron considerados todos los campos para el análisis de blancos/nulos y en el caso de ser detectados se configuró el nodo para que proceda a removerlos. De esta manera conseguimos obtener una información más pura para las siguientes etapas.

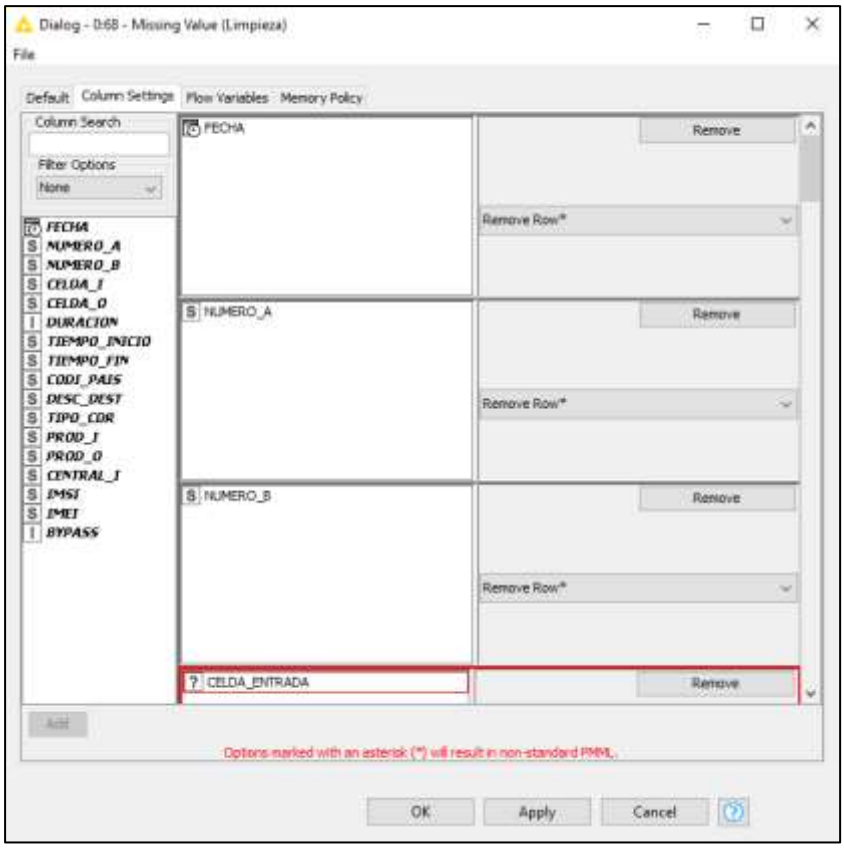

Figura 16. Configuración del nodo 'Missing Value'. Tomado de la herramienta KNIME. Elaboración propia.

Por último en esta etapa, se realizó un filtrado de la información en base a dos criterios, el primero en base a lo estudiado en el marco teórico acerca del comportamiento del fraude bypass que se enmascara como llamadas locales a servicios móviles de la operadora de destino de las llamadas internacionales, motivo por el cual se realizó el filtro de aquellos registros que en el campo ‘CODI_PAIS’ sea igual a ‘LO’ haciendo referencia a lo indicado según el diccionario de datos de la base. El segundo criterio de filtrado de información se basa en lo rescatado de las entrevistas con los especialistas, donde ambos concordaban que la mayor ocurrencia del fraude bypass se da en producto del tipo prepago, en lugar del pospago. Razón por la cual se aplicó el filtro solo de los servicios móviles que realizan llamadas sean prepago (Campo PROD_I = “PPA”).

Como resultado final de esta etapa, la información sufrió una reducción en sus registros la cual se detalla a continuación según el escenario.

Tabla 7. *Detalle de limpieza de datos*

Escenario de limpieza	Cantidad de registros removidos	Porcentaje
Registros duplicados	0	0%
Registros vacíos/nulos	0	0%
Registros filtrados	163.148	10%

Nota. La cantidad de registros removidos está presentada en porcentaje en relación a los registros de la base de datos original (1’609.108 registros). Elaboración propia.

Tras la implementación de esta etapa de pre-procesamiento y limpieza, la base de registros de llamadas se redujo a 1’445.960 filas para la siguiente fase.

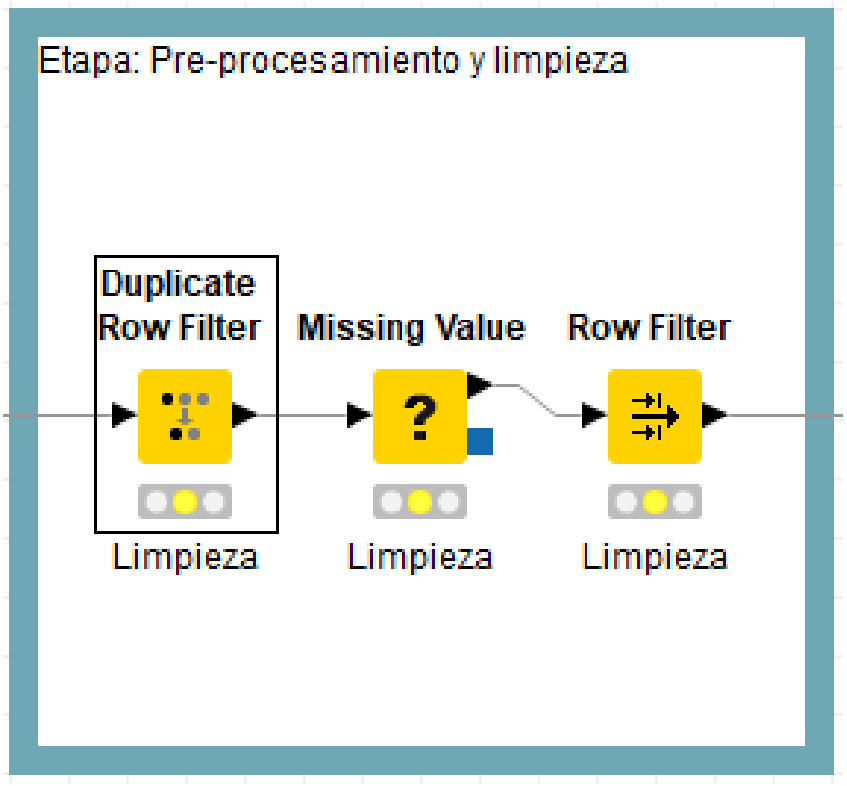

Figura 17. Etapa de pre-procesamiento/limpieza. Tomado de la herramienta KNIME. Elaboración propia.

3. TRANSFORMACIÓN/REDUCCIÓN

Esta etapa se caracteriza por la generación de la vista minable que será usada en la posterior minería de datos. Para ello, se subdividió el proceso en dos partes descritas a continuación

En la primera parte de esta etapa, se procedió con la eliminación de variables de la base de datos que no contribuyen al análisis. Para dicho fin, fue utilizado el nodo ‘Column filter’ que nos permitió excluir estas variables. Los motivos de la exclusión se detallan a continuación.

Tabla 8. *Detalle de las variables excluidas en la etapa de reducción*

Variable eliminada	Motivos
celda_O	Para el objeto de estudio no es de gran aporte conocer la celda donde se receptan las llamadas dado a su alta variabilidad.
codi_pais	Esta variable ya cumplió su función al momento de ser utilizada para el filtrado de llamadas con solo destinos locales realizado en la etapa previa. Razón por la cual ya no es necesario mantenerla.
desc_dest	Dado a que ya se cuenta con la codificación del destino al cual se realiza la llamada, este campo se convierte en redundante y no será usado en el procesamiento de datos.
tipo_cdr	Se trata de un valor fijo dado a que la base de datos cuenta con un solo tipo de registros de llamadas siendo este del tipo saliente, por lo cual es redundante.
prod_I	Esta variable ya cumplió su función al momento de ser utilizada para el filtrado de llamadas de solo servicios móviles prepago (PPA) realizado en la etapa previa. Razón por la cual ya no es necesario mantenerla.

prod_O	Para el objeto de estudio no es de gran aporte conocer el producto de destino donde se receptan las llamadas dado a su alta variabilidad.
Imsi	Debido a que ya poseemos la identificación del número_A, el valor de la IMSI se vuelve redundante para el estudio y no fue usado en el procesamiento de los datos.

Nota. Elaboración propia.

Por lo pronto, en la configuración del nodo ‘Column Filter’ se excluyen las variables mencionadas, tal como se muestra a continuación:

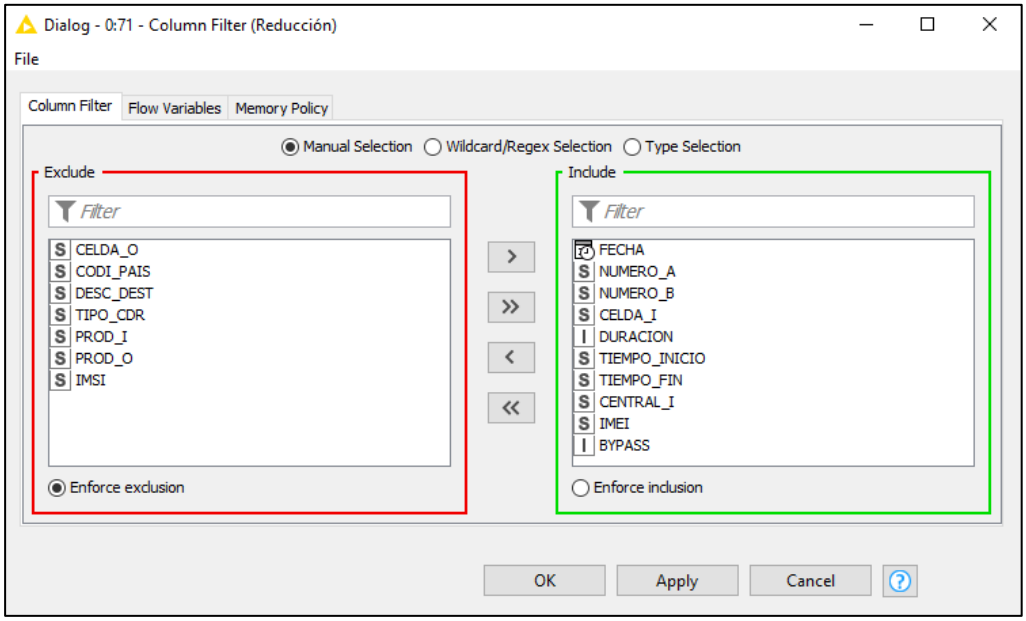

Figura 18. Configuración del nodo 'Column Filter' en la etapa de Reducción. Tomado de la herramienta KNIME. Elaboración propia.

Luego de este paso se procedió con la primera transformación dentro de la base de datos a través del nodo ‘Math Formula’ donde se configuro de modo que el campo duración que originalmente estaba expresado en segundos, sea transformado a minutos a través de la fórmula “round($DURACION$/60,2)”. Lo anterior ayuda a reducir el rango de esta variable que facilitará en el estudio y lectura de la misma. A partir de este punto, la variable ‘DURACION’ pasó de ser de tipo discreta a continua, con la inclusión de dos decimales.

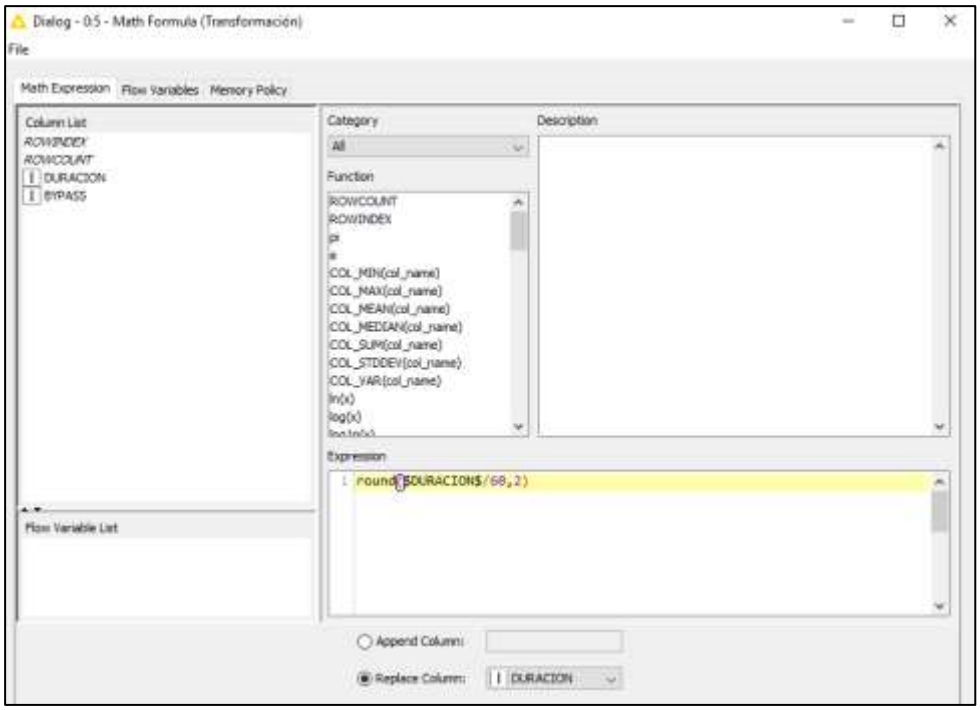

Figura 19. Configuración del nodo 'Math Formula' en la etapa de Transformación. Tomado de la herramienta KNIME. Elaboración propia.

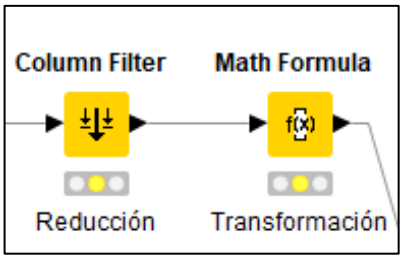

Figura 20. Flujo del proceso con nodos 'Colunm Filter' y 'Math Formula'. Etapa de Reducción y Transformación. Tomado de la herramienta KNIME. Elaboración propia

Como tercer paso dentro de esta etapa, se realizó la construcción de dos variables nuevas, las cuales ayudan para posteriormente discretizar las variables del tiempo de inicio ("TIEMPO_INICIO") y fin ("TIEMPO_FIN") de la llamada en una variable de ocurrencia del evento. Para la construcción de estas variables se usaron los nodos 'String Manipulation' y 'String to Number'. El primer nodo me permite substraer de las variables "TIEMPO_INICIO" y "TIEMPO_FIN" (ambas en formato hh:mm:ss) solo la hora del evento de registro de llamada a través de las fórmulas substr($TIEMPO_INICIO$,0 ,2) y substr($TIEMPO_FIN$,0 ,2). Las nuevas variables se denominarán "HORA_A" y "HORA_B" respectivamente.

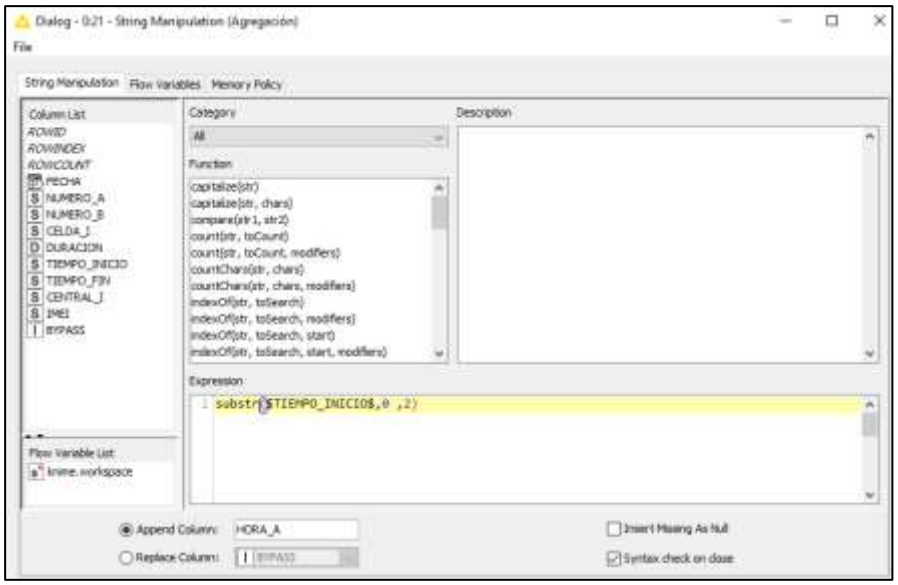

Figura 21. Configuración del nodo 'String Manipulation' para el campo HORA_A. Etapa de Agregación. Tomado de la herramienta KNIME. Elaboración propia

Tras la creación de las nuevas variables, se debió transformar las mismas a un formato numérico para su posterior uso en el nodo próximo, el cual solo admite variables de este tipo. Esta transformación se realizó a través del nodo ‘String To Number’.

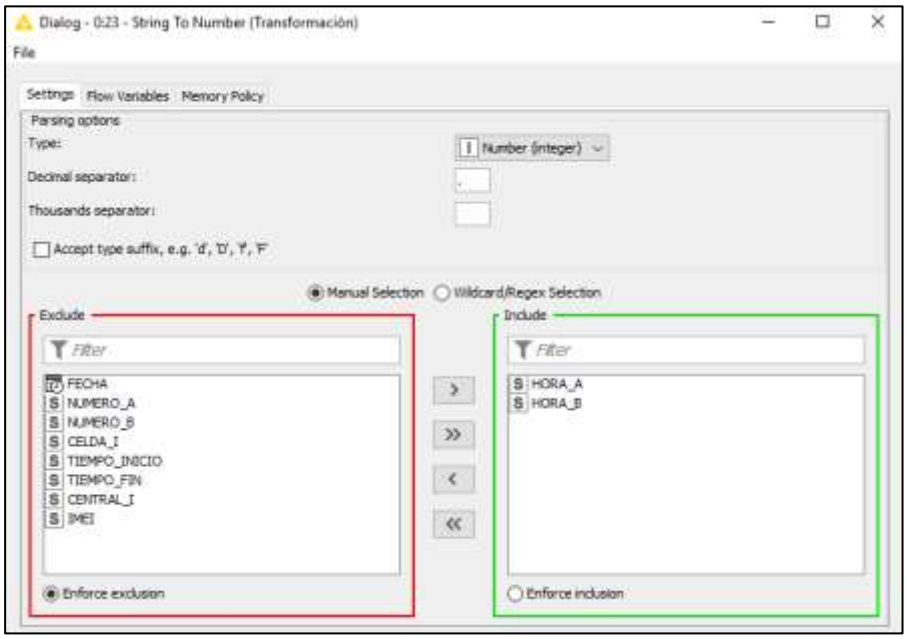

Figura 22. Configuración del nodo 'String to Number’. Etapa de Transformación. Tomado de la herramienta KNIME. Elaboración propia

Tras la aplicación de los nodos mencionados el flujo se visualiza de la siguiente manera en la figura 22.

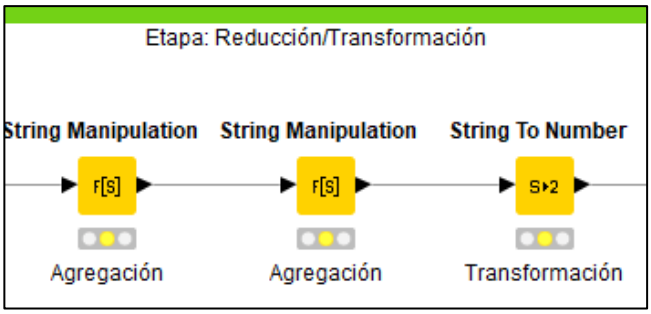

Figura 23. Flujo del proceso con nodos 'String Manipulation' y 'String To Number. Etapa de Agregación y Transformación. Tomado de la herramienta KNIME. Elaboración propia

El siguiente paso dentro de esta etapa es la creación de la variable denominada "Ocurrencia" que no es más que la variable discretizada en relación al tiempo que se realiza la llamada de los servicios móviles, el cual se detalla a continuación.

Tabla 9. *Definición de la variable "Ocurrencia"*

Ocurrencia	**Reglas**
Madrugada	HORA_A >= 0 && HORA_B <= 5
Mañana	HORA_A >= 6 && HORA_B <= 11
Tarde	HORA_A >= 12 && HORA_B <= 17
Noche	HORA_A >= 18 && HORA_B <= 23
NC	NC – No Ocurrencia, usado en el caso de las llamadas que inician dentro de un periodo y terminan en otro distinto.

Nota. La ocurrencia son definidas en base a las variables "HORA_A" y HORA_B" definidas por el tiempo de inicio y fin del registro de llamada de cada servicio móvil. Elaboración propia.

El objetivo de esta variable es mapear el periodo del día donde ocurre más actividad en llamadas por parte de cada uno de los servicios móviles encontrados en nuestra base de datos. De esta manera posteriormente es posible determinar el periodo de tiempo donde más ocurren eventos de fraude bypass a través de la frecuencia absoluta de la ocurrencia.

Para la creación de esta variable se usó el nodo 'Java Snippet', el cual nos permite el ingreso de estas reglas a través de código Java para tratar las

condiciones definidas para las variables “HORA_A” y “HORA_B”, se muestra a continuación:

Figura 24. Configuración del nodo 'Java Snippet'. Etapa de Transformación. Tomado de la herramienta KNIME. Elaboración propia

Tras las transformaciones y agregaciones realizadas al momento, la información queda constituida de la siguiente manera.

Columns: 13	Column Type	Column Index
FECHA	Local Date Time	0
NUMERO_A	String	1
NUMERO_B	String	2
CELDA_I	String	3
DURACION	Number (double)	4
TIEMPO_INICIO	String	5
TIEMPO_FIN	String	6
CENTRAL_I	String	7
IMEI	String	8
BYPASS	Number (integer)	9
HORA_A	Number (integer)	10
HORA_B	Number (integer)	11
OCURRENCIA	String	12

Figura 25. Pre visualización de los datos tras las reducción/agregación/transformación realizada. Tomado de la herramienta KNIME. Elaboración propia

A partir de esta tabla generada por el último nodo (“Java Snippet”) se realizó una transformación por medio de la agrupación de los registros a través de las columnas “NUMERO_A” y “BYPASS”, de modo a que se puedan realizar agregaciones nuevas de acuerdo el comportamiento de cada uno de los servicios móviles. Estas agrupaciones son consideradas en la siguiente tabla.

Tabla 10. *Detalle de agrupación por los campos "número A" y "Bypass".*

Lineamientos	Descripción
NUMERO_B – Unique Count	Cantidad de números B únicos al cual el número A ha llamado.
DURACION – Count	Cantidad de llamadas realizadas por número A.
DURACION – COUNT	Total de minutos realizados en llamadas del número A.
CELDA – UNIQUE COUNT	Cantidad de celdas únicas de las cuales el número A ha llamado
IMEI – UNIQUE COUNT	Cantidad de equipos únicos en el cual en el número A realizó las llamadas.
DISPERSION	NUMERO_B – Unique Count / DURACION – Count Es la relación porcentual entre la cantidad de números B llamados / cantidad de llamadas realizadas

Nota. Elaboración propia.

La agrupación es realizada a través del nodo "Group By", el cual es configurado de tal manera que sean considerados los lineamientos explicados en la tabla anterior. Vale indicar que para el último campo "DISPERSION" se utilizó al nodo "Math Formula" a través de la fórmula ($Unique count*(NUMERO_B)$/$Count*(DURACION)$)*100.

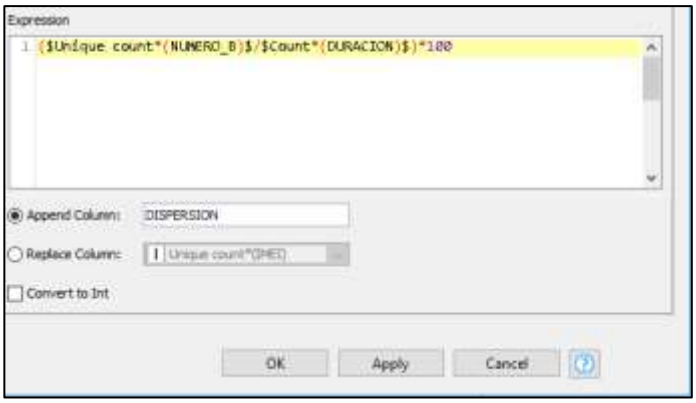

Figura 26. Configuración del nodo “Math Formula” para el campo “DISPERSION”. Tomado de la herramienta KNIME. Elaboración propia

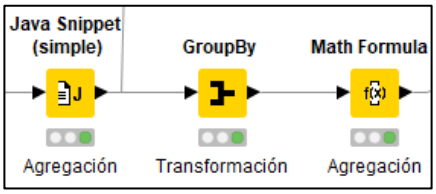

Figura 27. Flujo del proceso con los nodos “Java Snippet” y “Group By” en la tabla de Transformación. Tomado de la herramienta KNIME. Elaboración propia

Figura 28. Configuración del nodo “Group By” para las agrupaciones por “Numero_A” y “Bypass”. Tomado de la herramienta KNIME. Elaboración propia

Así mismo, se realizó una agrupación individual y en paralelo de la columna “Ocurrencia” para determinar a través de la frecuencia relativa el periodo de mayor actividad de cada uno de los servicios móviles. Esta agrupación adicional se realizó a través del uso de dos nodos “Group By”, el primero realizó un conteo de los registros por cada uno de los servicios móviles (“NUMERO_A”) y la ocurrencia de registros (“OCURRENCIA). En el segundo se realizó una selección por “NUMERO_A” de la ocurrencia con mayor cantidad de registros por cada uno de los servicios. Las configuraciones de este grupo de nodos se detallan a continuación.

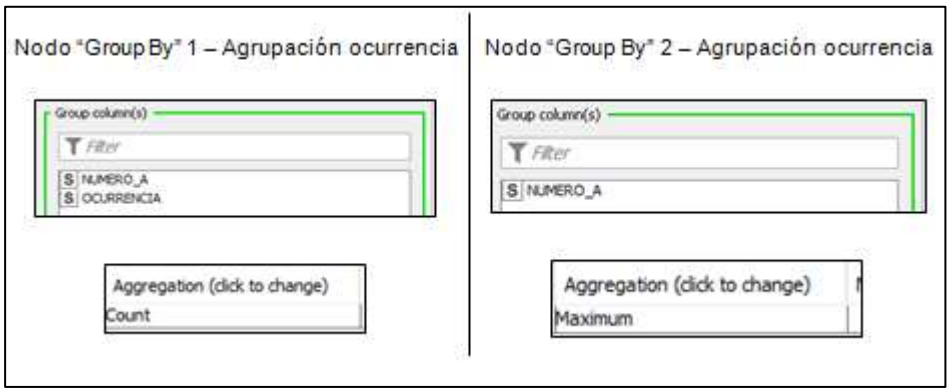

Figura 29. Configuración de los nodos “Group By” 1 y 2 de la Agrupación por Numero_A y Ocurrencia. Tomado de la herramienta KNIME. Elaboración propia

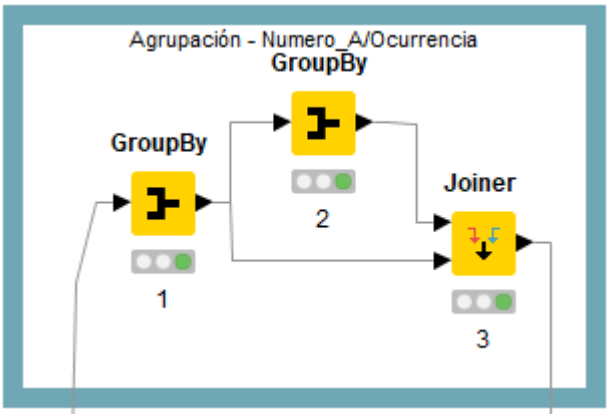

Figura 30. Configuración de los nodos “Group By” 1 y 2 de la Agrupación por Numero_A y Ocurrencia. Tomado de la herramienta KNIME. Elaboración propia

Una vez finalizado, se procedió a través del nodo “Joiner” a la unión de los datos de la agrupación por “NUMERO_A” y “BYPASS” con la agrupación de “OCURRENCIA” a través del campo “NUMERO A”.

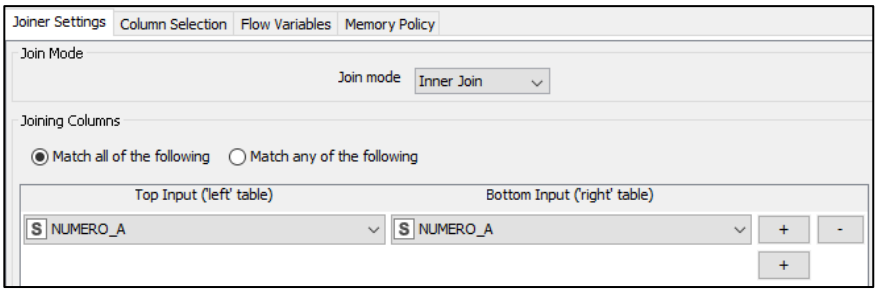

Figura 31. Configuración de del nodo “Joiner”. Tomado de la herramienta KNIME. Elaboración propia.

Una vez unidos, se hizo el uso del nodo “Column Rename” para cambiar el nombre de algunas variables para mayor facilidad de entendimiento y el tipo para el manejo en etapas posteriores. Tras la aplicación, los campos que constituyen a la vista minable primitiva quedaron de la siguiente manera.

Tabla 11. *Detalle de agrupación por los campos “número A” y “Bypass”.*

Campo	Descripción
NUMERO_A	Número saliente que origina las llamadas
BYPASS	Variable clasificatoria. Se identifica al “NUMERO_A” como caso bypass (1) o no (0).
NUMEROS_B	Variable numérica. Cantidad de “NUMEROS_B” únicos llamados por el “NUMERO_A”
LLAMADAS	Variable numérica. Cantidad de llamadas realizadas por el “NUMERO_A”.
MINUTOS	Variable decimal. Duración en minutos del total de llamadas realizadas por el “NUMERO_A”.
CANT_CELDA	Cantidad de celdas únicas del cual el “NUMERO_A” ha llamado.
CANT_EQUIPO	Cantidad de equipos del cual el “NUMERO_A” realizó las llamadas.
DISPERSION	NUMEROS_B / LLAMADAS Es la relación porcentual entre la cantidad de “NUMEROS_B” únicos llamados sobre la cantidad de llamadas realizadas
OCURRENCIA	Periodo de mayor actividad del “NUMERO_A”.

Nota. Elaboración propia.

En su totalidad, la primera parte de la etapa de Reducción/Transformación se encontró constituida por 13 nodos de la herramienta KNIME donde se realizaron 1 reducción de 7 variables, 5 transformaciones y 4 agregaciones.

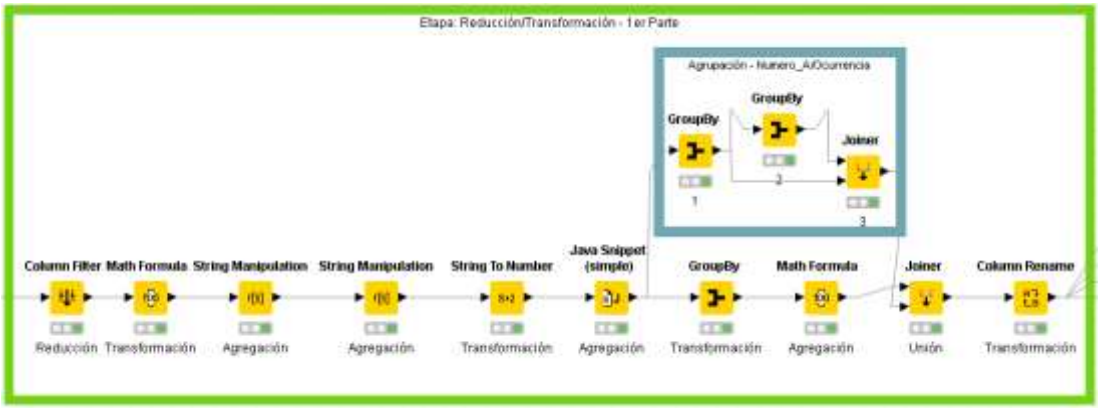

Figura 32. Primera parte de la etapa de Reducción/Transformación. Tomado de la herramienta KNIME. Elaboración propia.

Vale indicar que tras la primera parte de la esta etapa quedaron los registros de 21.533 servicios móviles con un registro por cada uno de ellos, con 9 columnas ya identificadas anteriormente.

rows: 21533 | Spec - Columns: 9 | Properties | Flow Variables

NUMER...	BYPASS	NUMEROS_B	LLAMADAS	MINUTOS	CANT_CELDA	CANT_EQUIPO	DISPERSION	OCURRENCIA
593985877538	0	8	19	16.42	2	2	42.105	NOCHE
593995956513	0	18	88	96.65	12	1	20.455	MAÑANA
593985874091	0	1	3	1.77	1	1	33.333	TARDE
593985872425	0	12	26	26.1	7	1	46.154	MAÑANA
593985871936	0	6	12	73.37	3	1	50	MAÑANA
593985868588	0	5	9	5.57	3	1	55.556	TARDE
593985868133	0	7	16	16.62	4	1	43.75	NOCHE

Figura 33. Pre-visualización de la tabla de datos en la primera parte de la etapa de Reducción/Transformación. Tomado de la herramienta KNIME. Elaboración propia.

Para la segunda parte de esta etapa, se realizó la eliminación de outliers (datos atípicos/ruido) sobre la data resultante de la primera parte tras la aplicación de estadística descriptiva. Para este fin, se determinaron 3 variables a ser tomadas en cuenta para el análisis debido a un alto grado de correlación entre ellas.

La correlación se determinó tras el cálculo del coeficiente de correlación entre las variables LLAMADAS-MINUTOS y NUMEROS_B-LLAMADAS mediante el uso del nodo “Linear Correlation” con los siguientes resultados.

Tabla 12. *Resultado del nodo "Linear Correlation". Tomado de la herramienta KNIME.*

LLAMADAS-MINUTOS

Row ID	LLAMADAS	MINUTOS
LLAMADAS	1.0	0.4608387599908...
MINUTOS	0.4608387599...	1.0

NUMEROS_B-LLAMADAS

Row ID	NUMEROS_B	LLAMADAS
NUMEROS_B	1.0	0.93328139356...
LLAMADAS	0.93328139356...	1.0

Nota. Elaboración propia.

Para ambos casos, los coeficientes de correlación son mayores a cero y cercanos a uno, lo que se demuestra que estas variables se encuentran fuertemente correlacionadas entre sí y se interpreta como "a mayor cantidad de llamadas, mayor cantidad de minutos" (coeficiente de 0.46) y "a mayor cantidad de números_B, mayor cantidad de llamadas" (coeficiente de 0.93).

Adicionalmente, se construyeron tres diagramas de cajas para cada uno de estas variables para identificar visualmente la distribución que poseen y la presencia de datos atípicos gráficamente.

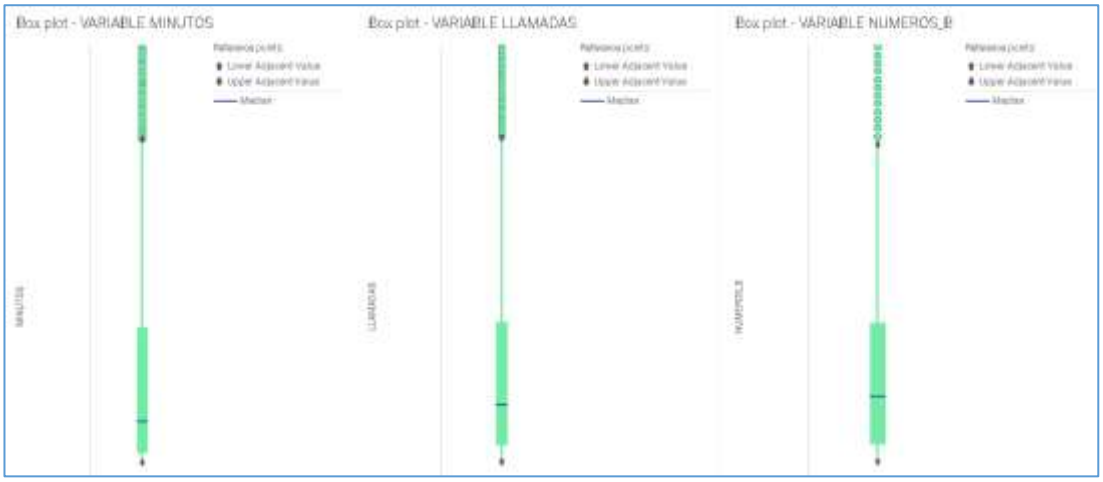

Figura 34. Diagrama de cajas para variables "MINUTOS", "LLAMADAS" y "NUMEROS_B". Tomado de la herramienta SPOTFIRE. Elaboración propia.

Por medio de la visualización del diagrama de cajas se evidenció una presencia de sesgo a la derecha en las tres variables estudiadas y una alta presencia de outliers. Así mismo, se construyó una tabla estadística que define el comportamiento de estos datos a través de la descripción de las variables estudiadas, descritos de la siguiente manera.

Tabla 13. *Análisis estadístico. Tomado de la herramienta SPOTFIRE.*

VARIABLE	DESCRIPCIÓN ESTADÍSTICA
MINUTOS Avg 87,8356 Median 30,3 StdDev 213,277 Max 13383,5 Min 0 Outliers 1845 UAV 242,4	El promedio de duración de llamadas es de 88 minutos con una desviación de +/- 213 minutos y datos medios de 30 minutos. Existieron 1.845 datos atípicos sobrepasando el umbral superior de 242 minutos.
LLAMADAS Avg 46,8172 Median 22 StdDev 118,837 Max 4370 Min 1 Outliers 1562 UAV 122	Existe un promedio 47 llamadas realizadas con una desviación de +/- 118 llamadas y datos medios de 22 llamadas. Existieron 1.562 datos atípicos sobrepasando el umbral superior de 122 llamadas.
NUMEROS_B Avg 24,008 Median 9 StdDev 108,238 Max 4307 Min 1 Outliers 1601 UAV 40	El promedio de números marcados (NUMEROS_B) es de 24, con una desviación de +/- 108 y datos medios de 9. Existieron 1.600 datos atípicos sobrepasando el umbral superior de 40 números.

Nota. Elaboración propia.

Luego de haberse identificado los datos atípicos, a través del nodo “Numeric Outliers” se procedió con la eliminación de 3.147 outliers que generaban ruido a la información.

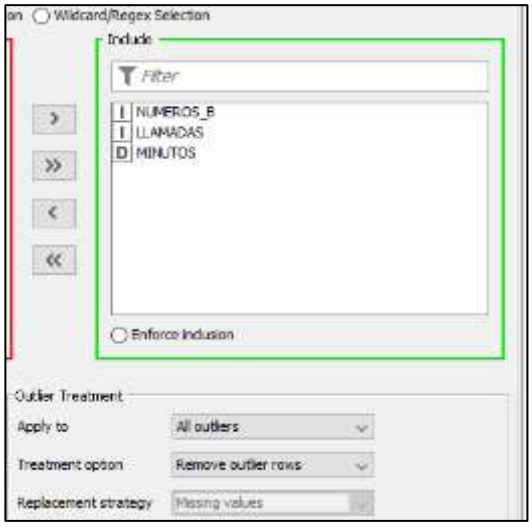

Figura 35. Configuración del nodo “Numeric Outliers”. Tomado de la herramienta SPOTFIRE. Elaboración propia.

Tras este paso se dio por finalizada la 2da etapa de Reducción/Transformación a través del siguiente flujo.

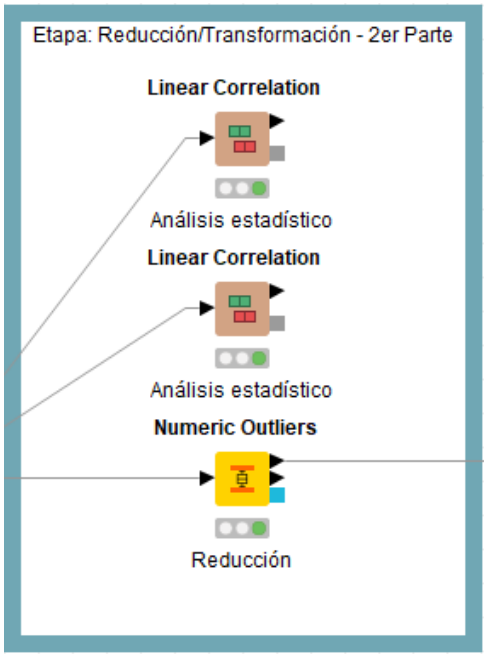

Figura 36. Segunda parte de la etapa de Reducción/Transformación. Tomado de la herramienta KNIME. Elaboración propia.

Vista minable

El resultado de esta etapa fue la construcción de la vista minable utilizada en el proceso de minería de datos compuesta de los registros de 18.386 servicios móviles en 9 campos enlistados y descriptos en la tabla 11. La construcción de la misma involucró el uso de 19 nodos de la herramienta KNIME desde la etapa de selección de datos hasta la etapa previa a la implementación de la minería de datos.

DATASET ORIGINAL

Columns: 17	Column Type	Column Index
FECHA	Local Date Time	0
NUMERO_A	String	1
NUMERO_B	String	2
CELDA_I	String	3
CELDA_O	String	4
DURACION	Number (inte...	5
TIEMPO_INICIO	String	6
TIEMPO_FIN	String	7
CODI_PAIS	String	8
DESC_DEST	String	9
TIPO_CDR	Number (inte...	10
PROD_I	String	11
PROD_O	String	12
CENTRAL_I	Number (inte...	13
IMSI	String	14
IMEI	String	15
BYPASS	Number (inte...	16

VISTA MINABLE

Columns: 9	Column Type	Column Index
NUMERO_A	String	0
BYPASS	String	1
NUMEROS_B	Number (integer)	2
LLAMADAS	Number (integer)	3
MINUTOS	Number (double)	4
CANT_CELDA	Number (integer)	5
CANT_EQUIPO	Number (integer)	6
DISPERSION	Number (double)	7
OCURRENCIA	String	8

Figura 37. Estructura de la base de datos original. Tomado de la herramienta KNIME. Elaboración propia.

Tabla 14. *Tipos de variables de la vista minable*

Campo	Tipo de dato	Escala	Tipo de Variable
numero_A	String	Cualitativa	Nominal
Bypass	String	Cualitativa	Nominal
numeros_B	Number	Cuantitativa	Discreta
Llamadas	Number	Cuantitativa	Discreta
Minutos	Number	Cuantitativa	Continua
cant_celda	Number	Cuantitativa	Discreta
cant_equipo	Number	Cuantitativa	Discreta
dispersion	Number	Cuantitativa	Continua
ocurrencia	String	Cualitativa	Nominal

Nota. Elaboración propia.

Mientras que los rangos y posibles valores se describen a continuación.

Tabla 15. *Rangos y posibles valores en la vista minable.*

Campo	Posibles valores
numero_A	18.386 numeros_A únicos
bypass	0: normal - 1: bypass
numeros_B	Min: 1 – Max: 40
llamadas	Min: 1 – Max: 122
minutos	Min: 0.1 – Max: 242.31
cant_celda	Min: 1 – Max: 73
cant_equipo	Min: 1 – Max: 18
dispersion	Min: 2.0 – Max: 100
ocurrencia	Madrugada – Mañana – Tarde - Noche

Nota. Elaboración propia.

Vale indicar que la vista minable es un set de datos ya tratado, sin valores vacíos, nulos, duplicados y sin registros que causen ruido tras la eliminación de los datos atípicos/outliers.

4. MINERIA DE DATOS

Para la construcción del modelo predictivo se utilizó una técnica de minería de datos **supervisada** debido a que se cuenta con un set de datos de entrenamiento previamente tratado y de **clasificación** dado a que se requiere predecir una variable de salida categorizada, en este caso si un servicio móvil es bypass (1) o no (0). No se usarán del tipo regresión dado a que no se requieren estimar valores numéricos, tan solo una categoría.

Dentro de los algoritmos de clasificación, en este trabajo de titulación fue utilizado son los bosques aleatorios o “random forest”, explicado a continuación.

Para la implementación de esta técnica de minería de datos, se utilizaron 3 nodos, el primero llamado “Partitioning” realiza la partición de la información en una relación 75% para el aprendizaje del modelo y el 25% restante para las pruebas del modelo construido.

El segundo nodo llamado “Random Forest Learner” realiza la construcción del modelo en base a la configuración de ciertos parámetros, tales como:

- La columna objetivo, el cual es la variable a estudiar, en este caso “Bypass”.
- Las variables a ser tomadas en cuenta para la construcción del modelo.
- El criterio de división, basados en el algoritmo de ganancia de información.
- La profundidad de cada árbol, la cual se constituyó con la mitad de la cantidad total de las variables del set de datos. Para este estudio, se configuró en 5.
- El número de árboles que tendrá el bosque aleatorio.

Vale indicar que para el presente trabajo de titulación se consideraron varias configuraciones en la cantidad de número de árboles tomados para la construcción del modelo. Los cuales fueron tomados los más representativos y analizados en la etapa de interpretación de resultados.

Por último, el nodo “Random Forest Predictor” es utilizado para probar el modelo construido con el 25% de la información del set de datos, de modo que pueda ser analizado, en la siguiente etapa de esta metodología.

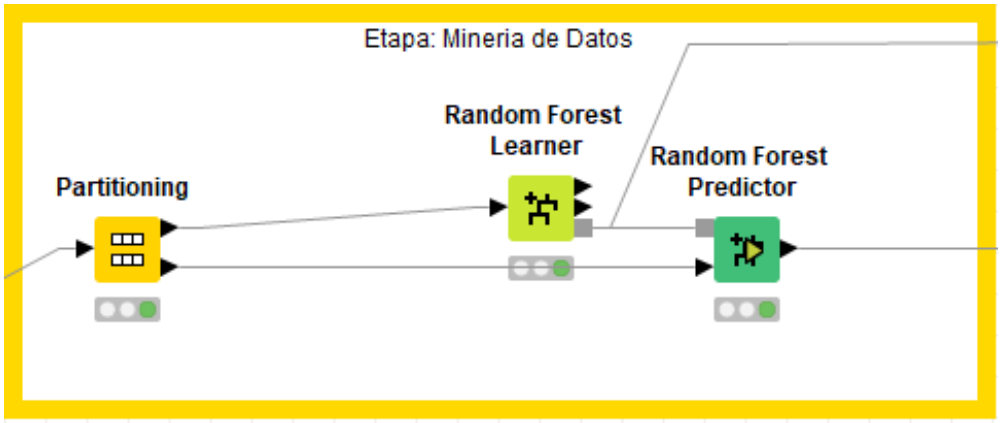

Figura 38. Etapa de minería de datos. Tomado de la herramienta KNIME. Elaboración propia.

Vale indicar que el resultado de ejecución facilita el modelo predictivo, el cual puede ser visualizado gráficamente, y también a nivel de reglas. Adicional, se agregan 4 campos adicionales al set de datos con el valor de la predicción obtenida para cada uno de los casos, así como el nivel de confidencialidad obtenido.

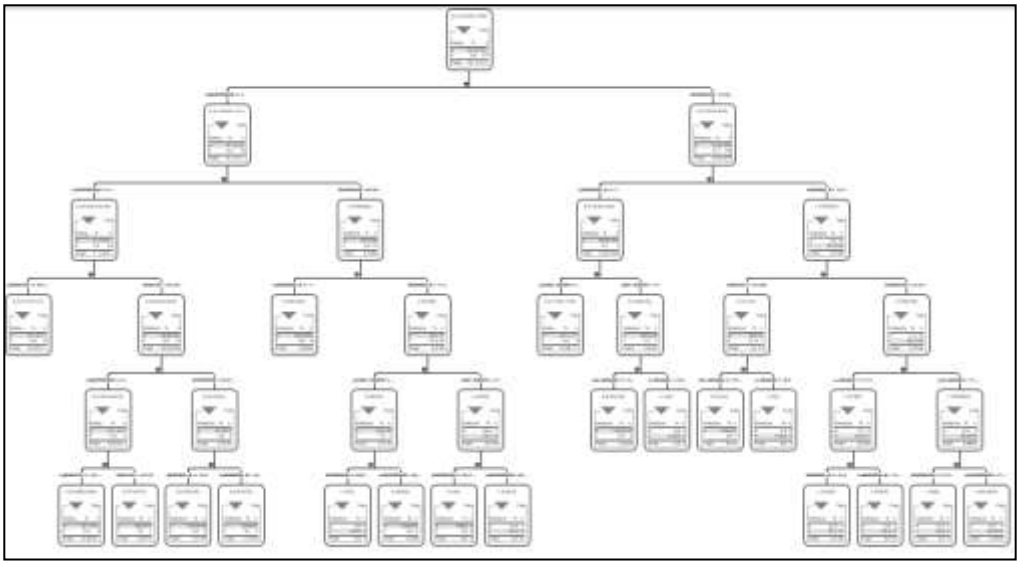

Figura 39. Representación gráfica de uno de los árboles resultantes. Tomado de la herramienta KNIME. Elaboración propia.

Se realizaron 45 corridas con varías configuraciones donde se determinaron tres puntos representativos con 10, 100 y 500 árboles, las cuales marcaron una diferencia en el nivel de precisión del modelo.

5. INTERPRETACIÓN Y EVALUACIÓN

Para la evaluación del modelo se utilizó el nodo “Scorer”, el cual realiza una matriz de confusión donde se determinan la cantidad de predicciones correctas e incorrectas con el fin de calcular el nivel de precisión de cada uno de los modelos obtenidos. De los cuales se realiza un comparativa de los modelos construidos con 10, 100 y 500 árboles como se muestra a continuación.

Tabla 16. *Comparativa de algoritmos usados.*

Algoritmo	Resultados
Random Forest con 10 árboles	Precisión: 87,862% - Error: 12,138%
Random Forest con 100 árboles	Precisión: 95,258% - Error: 4,742%
Random Forest con 500 árboles	Precisión: 99,782% - Error: 0,218%

Nota. Elaboración propia.

Se concluyó con la opción a tomar en el random forest con 500 árboles al contar con la mayor cantidad de precisión obtenida tras la evaluación.

Este modelo, arrojó como resultado un total de 19 reglas, donde 8 de ellas determinan si un caso es fraude bypass mientras que las 11 restantes indican que se trata de un caso con comportamiento normal. Dentro de las variables más significativas que indicó el modelo predictivo, se destacan por orden de relevancia las siguientes:

Tabla 17. *Relevancia de variables del modelo predictivo.*

Algoritmo	Resultados
1. Dispersión	> = 76%
2. Cantidad de equipos	1 - 2
3. Cantidad de números	>= 13
4. Cantidad de minutos	> 23,40
5. Cantidad de llamadas	>= 18

Nota. Elaboración propia.

El resultado del modelo propuesto sobre la partición del set de datos utilizado como prueba en este trabajo de titulación determinó que el 5,70% de las líneas se consideran como caso de fraude bypass.

Esto en contraste con la información original, donde el 4,65% de los casos fueron considerados como bypass, por lo que se puede determinar que la hipótesis planteada en este trabajo de titulación es cierta, demostrando que un modelo predictivo utilizando técnicas de minería de datos contribuye con un mayor eficacia a la detección de casos de fraude bypass, el cual puede ser utilizado en empresas de telefonía móvil en el Ecuador.

CONCLUSIONES

Para el presente trabajo de titulación se presentan las siguientes conclusiones obtenidas:

Se identificó que el análisis de CDRs y el Perfilamiento son procesos de detección de fraude bypass primordiales que han tenido que adaptarse a la evolución de las técnicas fraudulentas.

Se estableció que la detección de fraude bypass se realiza principalmente en función de la ‘dispersión de las llamadas’ y alto consumo realizado en un pequeño periodo de tiempo.

Se determinó la aplicación de la metodología KDD para procesar la información cruda y normalizar los datos en la etapa de Reducción/Transformación para obtener de manera exitosa una vista minable.

Se consideró factible el uso de la técnica de minería de datos “Random Forest” para la construcción del modelo predictivo de detección de casos Bypass debido a su alta precisión en clasificación.

Se generó un modelo de predicción con una precisión del 99,7% y margen error del 0,3%, por lo que se concluye que el mismo cumple con un nivel de eficacia aceptable para contribuir a la detección de fraude Bypass.

RECOMENDACIONES

Para el presente trabajo de titulación se presentan las siguientes recomendaciones a ser contempladas para futuros trabajos:

Se recomienda complementar con registros históricos la fuente de información para que el modelo pueda extraer nuevos patrones y reglas de detección, extendiendo el número de días y así mismo la cantidad de casos a estudiar.

Se recomienda el estudio por separado de los casos outliers obtenidos en la aplicación de la metodología de minería de datos. Dado a que no constituían dentro del alcance del presente trabajo de titulación fueron descartados como requisito para la construcción del modelo predictivo, sin embargo, es considerable su estudio dado a que podrían ser casos de fraude bypass.

Se sugiere el uso de las técnicas de minería de datos en las empresas de telefonía móvil del Ecuador como parte de sus herramientas para detección de fraudes conocidos.

Se recomienda el uso de las variables seleccionadas en este estudio como base para la construcción de nuevos modelos predictivos para detección de casos Bypass.

Es recomendable que las operadoras de telefonía móvil del país se mantengan estudiando nuevos patrones de fraude y evaluar su detección a través del uso de la minería de datos.

Se recomienda el uso e implementación de la minería de datos en otros modelos de negocio de modo que aprovechen su potencial para la toma de decisiones y aseguramiento de ingresos.

BIBLIOGRAFÍA

Aseguramiento de Ingresos para el Profesional de Aseguramiento de Ingresos—GRAPA. (2020). http://www.grapatel.com/espanol/

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bhatia, P. (2019). *Data Mining and Data Warehousing: Principles and Practical Techniques*. Cambridge University Press.

Camacho, K., & González, M. (2016). *Diseño de un sistema de detección y control del fraude en la prestación de los servicios de telefonía fija y servicio móvil avanzado en el ecuador* [Escuela Superior Politécnica del Litoral]. https://www.dspace.espol.edu.ec/retrieve/98770/D-103331.pdf

Camana, R. G. (2016). Potenciales Aplicaciones de la Minería de Datos en Ecuador. *Revista Tecnológica - ESPOL*, *29*(1). http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/464

Castro Aguilar, G. F., Pérez Pupo, I., Piñero Pérez, P. Y., & García Vacacela, R. (2016). Método para el aseguramiento de ingresos en entornos de desarrollo de software. *Revista Cubana de Ciencias Informáticas*, *10*, 43–57.

Communications Fraud Control Association. (2019). https://www.cfca.org/

CONATEL. (2007a). *Reglamento de interconexion*. https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2012/11/Reglamento-de-Interconexion.pdf

CONATEL. (2007b). *Reglamento del servicio telefonico de larga distancia internacional*.

Duarte, P. R. B. (2017). *Estudio sobre la vulnerabilidad de los servicios de telefonía sobre ip inteligentes a través de internet y su repercusión sobre las redes de telefonía por conmutación de circuitos*. 190.

El Comercio, D. (2012). *La Supertel intervino un bypass telefónico en Quito*. El Comercio. https://www.elcomercio.com/actualidad/negocios/supertel-intervino-bypass-telefonico-quito.html

El Telégrafo, D. (2019, June 7). *Arcotel detectó un 'bypass' clandestino en Quito*. El Telégrafo - Noticias del Ecuador y del mundo. https://www.eltelegrafo.com.ec/noticias/judicial/12/arcotel-bypass-quito

Elmi, A. H., Ibrahim, S., & Sallehuddin, R. (2013). Detecting SIM Box Fraud Using Neural Network. In K. J. Kim & K.-Y. Chung (Eds.), *IT Convergence and Security 2012* (pp. 575–582). Springer Netherlands. https://doi.org/10.1007/978-94-007-5860-5_69

Fernández, R. (2019). *Telecomunicaciones: Ranking de empresas por valor de marca 2019*. Statista.

https://es.statista.com/estadisticas/600868/ranking-de-las-principales-marcas-de-telecomunicaciones--por-valor-de-marca/

Ficek, M., & Kencl, L. (2012). Inter-Call Mobility model: A spatio-temporal refinement of Call Data Records using a Gaussian mixture model. *2012 Proceedings IEEE INFOCOM*, 469–477. https://doi.org/10.1109/INFCOM.2012.6195786

Gámez, O. R., Perdomo, R. H., Hidalgo, L. T., Escalona, L. G., & Romero, R. R. (2005). Telefonía móvil celular: Origen, evolución, perspectivas. *Ciencias Holguín*, *XI*(1), 1–8.

Gartner. (2019). https://www.gartner.com/en

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters*, *27*(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

González, H. D. L. (2016). *Metodología de la investigación: Propuesta, anteproyecto y proyecto*. Ecoe Ediciones.

GRAPA. (2020). *Revenue Assurance Standards*. http://www.grapatel.com/A-GRAPA/05-Standards/standards.asp

Guan, H., Li, J., Chapman, M., Deng, F., Ji, Z., & Yang, X. (2013). Integration of orthoimagery and lidar data for object-based urban thematic mapping using random forests. *International Journal of Remote Sensing*, *34*(14), 5166–5186. https://doi.org/10.1080/01431161.2013.788261

Guzmán, E. (2016). *Minería de Datos*.

Iguasnia, F. O., Salazar, M. A., Iguasnia, J. O., & Sánchez, M. V. (2017). Uso de tecnología HSPA (HSPA +) y su evolución con la generación de los celulares. *Revista Científica y Tecnológica UPSE*, *4*(1), 164–173. https://doi.org/10.26423/rctu.v4i1.246

Koi-Akrofi, G., Koi-Akrofi, J., Odai, D., & Twum, E. (2019). Global telecommunications fraud trend analysis. *International Journal of Innovation and Applied Studies*, *25*, 940–947.

Kotu, V., & Deshpande, B. (2014). *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann.

Lagla, G. A. F., Moreano, J. A. C., Arequipa, E. E. Q., & Quishpe, M. W. V. (2019). Minería de datos como herramienta estratégica. *RECIMUNDO: Revista Científica de la Investigación y el Conocimiento*, *3*(1), 955–970.

Lai, S., Farnham, A., Ruktanonchai, N. W., & Tatem, A. J. (2019). Measuring mobility, disease connectivity and individual risk: A review of using mobile phone data and mHealth for travel medicine. *Journal of Travel Medicine*, *26*(3). https://doi.org/10.1093/jtm/taz019

Lawrence, R. L., Wood, S. D., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*, *100*(3), 356–362. https://doi.org/10.1016/j.rse.2005.10.014

M. H. Badii, Castillo, J., Landeros, L., & Cortez, K. (2017). Papel de la estadística en la investigación científica. *Innovaciones de Negocios*,

4(7).

http://revistainnovaciones.uanl.mx/index.php/revin/article/view/180

Marin Castro, H. M. (2017). *Minería de Datos*. Universidad Politécnica de Victoria. https://www.tamps.cinvestav.mx/~hmarin/Mineria/EC2.pdf

Martinez, H. (2014). *Metodología de la investigación*. Cengage Learning Editor.

Menéndez, J. Á. (2008). *Minería de Datos: Aplicaciones en el sector de las telecomunicaciones*. 8.

Molina, J., & García, J. (2017). *Técnicas de Análisis de Datos*.

Naranjo, D., Carrión, D., & Mejía, I. (2016). *Evolución de la tecnología móvil. Camino a 5G*. http://www.eumed.net/rev/cccss/2016/04/5G.html

Nassau County Spin. (2018). *Telecommunications Fraud*. 2.

Pacherres Gutiérrez, J. J. (2018). Propuesta de una metodología de extracción de conocimientos a partir de datos de las prestaciones del seguro integral de salud en la región Piura en el año 2016. *Universidad Católica Los Ángeles de Chimbote*. http://repositorio.uladech.edu.pe/handle/123456789/2868

Pereira, S. R. T., Arteaga, I. H., Zambrano, S. J. C., Troya, A. H., & Pérez, J. C. A. (2016). Descubrimiento de patrones de desempeño académico. *Ediciones Universidad Cooperativa de Colombia*. http://dx.doi.org/10.16925/9789587600490

Reaves, B., Shernan, E., Bates, A., Carter, H., & Traynor, P. (2015). *Boxed Out: Blocking Cellular Interconnect Bypass Fraud at the Network Edge*. 17.

Román, C., & Mauricio, C. (2008). *El bypass en redes telefónicas celulares, técnicas de detección de números celulares implicados y de infraestructuras ilegales*. http://bibdigital.epn.edu.ec/handle/15000/1028

Subudhi, A., Dash, M., & Sabut, S. (2020). Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier. *Biocybernetics and Biomedical Engineering*, *40*(1), 277–289. https://doi.org/10.1016/j.bbe.2019.04.004

TM Forum. (2015). *Neural Technologies implements 'most effective bypass fraud solution to date.'* TM Forum. https://www.tmforum.org/press-and-news/neural-technologies-implements-most-effective-bypass-fraud-solution-to-date/

TM Forum. (2020). TM Forum. https://www.tmforum.org/

Torres, B. (2015). *Metodologia de la investigación: Para administración, economía, humanidades y ciencias sociales*. Prentice Hall.

DECLARACIÓN Y AUTORIZACIÓN

Yo, ALCIVAR LEÓN CRISTHIAN ROGER, con C.C: # 0930283296, autor del trabajo de titulación: **Diseño de un modelo predictivo a través de la técnica de minera de datos 'Random Forest' para la detección de fraude bypass en redes telefónicas en el Ecuador,** previo a la obtención del título de INGENIERO EN SISTEMAS COMPUTACIONALES en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, 26 de Febrero del 2020

f. _____
Nombre: Alcivar León Cristhian Roger
C.C: 0930283296

Presidencia de la República del Ecuador

Plan Nacional de Ciencia, Tecnología, Innovación y Saberes

SENESCYT
Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA

FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN

TEMA Y SUBTEMA:	**Diseño de un modelo predictivo a través de la técnica de minera de datos 'Random Forest' para la detección de fraude bypass en redes telefónicas en el Ecuador**		
AUTOR(ES)	**Alcivar León, Cristhian Roger**		
REVISOR(ES)/TUTOR(ES)	**Cornejo Gómez, Galo Enrique**		
INSTITUCIÓN:	**Universidad Católica de Santiago de Guayaquil**		
FACULTAD:	**Faculta de Ingeniería**		
CARRERA:	**Ingeniería en sistemas computacionales**		
TITULO OBTENIDO:	**Ingeniero en sistemas computacionales**		
FECHA DE PUBLICACIÓN:	26 de Febrero del 2020	**No. DE PÁGINAS:**	**80**
ÁREAS TEMÁTICAS:	**Minería de datos, Aseguramiento de ingresos, Fraude en telecomunicaciones**		
PALABRAS CLAVES/ KEYWORDS:	**Minería de datos, Metodología KDD, Fraude Bypass, Random Forest, Bosques aleatorios, KNIME**		

RESUMEN/ABSTRACT: En las empresas de telecomunicaciones, dado a su creciente demanda de mantenernos comunicados, emplean mayores configuraciones y elementos que en ocasiones generan brechas que son aprovechadas por personas u organizaciones para realizar acciones ilícitas como es el caso del fraude por bypass. Es un hecho que el 5% de los ingresos de una empresa de telefonía móvil del país se pierdan por la falta de detección de este tipo de casos, el cual ocasiona fugas de ingresos para las organizaciones que ofrecen este servicio a nivel mundial y nacional. Este trabajo denominado "Diseño de un modelo predictivo a través de la técnica de minera de datos 'Random Forest' para la detección de fraude bypass en redes telefónicas en el Ecuador" plantea como objetivo la construcción de un modelo predictivo empleando minería de datos a través de la metodología KDD ("Descubrimiento de Conocimiento en Base de Datos") de modo que contribuya a la eficacia en la detección de este tipo de fraude. La aplicación de la metodología se realiza mediante la herramienta de software KNIME implementando un flujo de trabajo con el uso de bosques aleatorios como técnicas de minería de datos del tipo clasificatorio supervisada donde se emplean los registros de llamadas como base para transformarlos a una vista minable apta para la construcción del modelo. Los resultados del trabajo indicaron que el uso de la minería de datos reportó mayor eficacia que el análisis de CDRs tradicional en la detección de casos de fraude bypass y plantea las bases para futuros estudios del tema en este modelo de negocios.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:** +593-9-83821557	E-mail: cristhian.alcivar93@gmail.com
CONTACTO CON LA INSTITUCIÓN (COORDINADOR DEL PROCESO UTE)::	**Nombre: Toala Quimi, Edison José**	
	Teléfono: +593-4-220263 ext. 1025	
	E-mail: edison.toala@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA	
Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	