FACULTAD DE POSTGRADO

TESIS DE POSTGRADO

FACTORES QUE CARACTERIZAN LÍNEAS TELEFÓNICAS UTILIZADAS PARA TRÁFICO GRIS EN EMPRESA DE TELEFONÍA MÓVIL

SUSTENTADO POR:

GINA MARÍA VALLADARES HERNÁNDEZ
ALEJANDRO JOSUÉ CALDERÓN TORRES

PREVIA INVESTIDURA AL TÍTULO DE

MÁSTER EN GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN

TEGUCIGALPA, FRANCISCO MORAZÁN, HONDURAS, C.A.
ABRIL 2014

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA

UNITEC

FACULTAD DE POSTGRADO

AUTORIDADES UNIVERSITARIAS

RECTOR

LUIS ORLANDO ZELAYA MEDRANO

SECRETARIO GENERAL

JOSÉ LÉSTER LÓPEZ

VICERRECTOR ACADÉMICO

MARLON ANTONIO BREVÉ REYES

DECANO DE LA FACULTAD DE POSTGRADO

DESIRÉ TEJADA

FACTORES QUE CARACTERIZAN LÍNEAS TELEFÓNICAS UTILIZADAS PARA TRÁFICO GRIS EN EMPRESA DE TELEFONÍA MÓVIL

TRABAJO PRESENTADO EN CUMPLIMIENTO DE LOS REQUISITOS EXIGIDOS PARA OPTAR AL TÍTULO DE MÁSTER EN GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN

ASESOR METODOLÓGICO

JUAN MARTÍN HERNÁNDEZ

ASESOR TEMÁTICO

HÉCTOR EMILIO GUEVARA PINTO

MIEMBROS DE LA TERNA

CARLOS PÉREZ

CARLOS ORDÓÑEZ

MIGUEL RAMÍREZ

FACULTAD DE POSTGRADO

FACTORES QUE CARACTERIZAN LÍNEAS TELEFÓNICAS UTILIZADAS PARA TRÁFICO GRIS EN EMPRESA DE TELEFONÍA MÓVIL

AUTORES:
Gina María Valladares Hernández
Alejandro Josué Calderón Torres

Resumen

Actualmente en Honduras existe una deficiencia en la detección de líneas telefónicas utilizadas para tráfico gris en redes GSM, debido a la variabilidad de los factores principales considerados y la rigidez de los métodos tradicionales de detección. El presente estudio tiene como propósito conocer factores característicos de líneas telefónicas utilizadas para tráfico gris en Honduras, con la finalidad de utilizar el conocimiento de dicho perfil para la detección de casos de fraude mediante una red neuronal con aprendizaje supervisado. Los resultados del estudio se presentan de forma separada para el enfoque cualitativo y cuantitativo, de forma que se comprendan los factores más significativos encontrados en cada análisis. Con la caracterización tanto del perfil de línea telefónica de tráfico gris como de línea telefónica legal, se procedió a construir un modelo de red neuronal con un 99% de efectividad de detección, y para efectos prácticos de aplicabilidad se describen las bases para la posterior construcción de un modelo de detección de tráfico gris.

Palabras Clave: perceptrón multicapa, red GSM, red neuronal, tráfico gris.

GRADUATE SCHOOL

FACTORS THAT DETERMINE TELEPHONE LINES USED FOR BYPASS IN MOBILE PHONE COMPANY

AUTHORS:
Gina María Valladares Hernández
Alejandro Josué Calderón Torres

Abstract

Honduras currently suffers a deficiency in the detection of mobile telephone lines used for bypass in GSM networks, due to the variable nature of the main factors considered, and the rigidity of traditional detection methods. The present study aims to find characteristic factors of telephone lines used for bypass, in order to use this knowledge for detection of bypass cases through the use of a neural network with supervised learning. The results of this study are presented separately for the qualitative and quantitative approach, for a better understanding of the most significant factors discovered. With the characterization of both the bypass telephone line and the legal telephone line, the study proceeds to build a neural network model with 99% detection efficiency, and for applicability matters, the basics are described for the subsequent construction of a bypass detection model.

Keywords: multilayer perceptron, GSM network, neural network, bypass.

DEDICATORIA

Dedicamos este documento de tesis a Dios, por concedernos todo lo que somos, y a nuestros padres, por su ejemplo invaluable.

AGRADECIMIENTO

A las personas que aportaron sus conocimientos y su tiempo para enriquecer este trabajo, nuestro asesor temático Héctor Guevara, nuestros asesores metodológicos Juan Martín Hernández y Cinthia Cano, a los expertos en el área de aseguramiento de ingresos por compartir su experiencia, y muy especialmente a nuestros queridos amigos, compañeros y permanente equipo de trabajo: Erick, José y Servio.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO I. PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 INTRODUCCIÓN

En la investigación presentada se expone la problemática del tráfico gris en el rubro de la telefonía móvil en Honduras, con el propósito de comprender a profundidad los escenarios de fraude a través de éste método, y así poder exponer los puntos de enfoque en aras de minimizar las incidencias del mismo.

La presente investigación está estructurada de forma en que inicialmente se explican las bases sobre las cuales se hace posible el tráfico gris, para luego exponer la situación actual de dicha problemática en las empresas de telefonía móvil en Honduras. Posteriormente, se consideran las bases financieras y legales en las que se enmarca la problemática de tráfico gris para determinar la magnitud del impacto a los operadores de telefonía móvil. Finalmente, mediante estudio científico, se analizan métodos existentes para identificar éste tipo de fraude y así evitar sus reincidencias.

Otras investigaciones han brindado resultados prometedores, ofreciendo soluciones al problema de tráfico gris mediante el uso de tecnologías informáticas, combinando métodos inteligentes con sistemas de procesamiento de información, para brindar finalmente una solución que abarca los aspectos más relevantes para la industria de las telecomunicaciones.

La investigación puede tomarse como base para la aplicación práctica de métodos de detección de fraude, considerando los factores que en ésta investigación se incluyen, para facilitar los primeros pasos en la caracterización del tráfico gris en cada caso específico por operador móvil.

1.2 ANTECEDENTES DEL PROBLEMA

En Honduras existen más de siete millones de líneas de telefonía móvil, estas se distribuyen entre las tres compañías proveedoras de dichos servicios: Claro, Hondutel y CELTEL (CONATEL, 2011, p.4). Las tres compañías de telecomunicaciones en Honduras cuentan con el servicio de llamadas de voz internacionales que permiten la comunicación rápida y efectiva entre habitantes de regiones geográficas distintas.

Debido a la demanda del servicio de llamadas internacionales, éste representa importantes utilidades para los proveedores, dado que el precio por minuto de una de estas llamadas es mayor al de una llamada local, y las tarifas varían de acuerdo al país u operador destino de la llamada.

Un factor que afecta de forma negativa la percepción de utilidades por parte de las empresas de telecomunicaciones debido a las llamadas internacionales, es el llamado tráfico gris, el cual disfraza las llamadas internacionales como locales, evadiendo así los costos asociados y por ende, haciendo fraude al proveedor. El servicio de telefonía móvil tiene cobertura en los dieciocho departamentos del país, tanto en zonas rurales como urbanas, permitiendo acceder al servicio de llamadas internacionales desde cualquiera de estas zonas.

Ocurre en ocasiones que un suscriptor recibe una llamada que es identificada por la red como llamada local, fácilmente reconocible a través de su teléfono celular que reconoce un número telefónico local, pero al responder se da cuenta que está recibiendo una llamada de una línea internacional. Este es un ejemplo típico de tráfico gris, en el que se puede reconocer entonces que existió un mecanismo de adulteración del transporte de la llamada para la evasión de costos.

Las empresas de telecomunicaciones cuentan con iniciativas para mitigar las pérdidas monetarias en las que incurren debido al tráfico gris, y cuentan con diferentes mecanismos de detección de dicho tráfico, los cuales deben mantener y actualizar constantemente. No existe una solución definitiva para la problemática abordada, pero

si existen factores sobresalientes entre los diferentes mecanismos existentes, los cuales hacen más efectivas unas soluciones sobre las demás.

1.2.1. TRÁFICO GRIS EN HONDURAS

El fraude en la telefonía móvil afecta en gran medida a las compañías proveedoras de este servicio a nivel global, debido a que en éstos casos se afecta de forma monetaria y también se ve afectada la imagen de la empresa, como consecuencia de verse comprometida la integridad y seguridad de la red de servicios. Los reportes del 2003 de la Asociación del Control del Fraude en las Telecomunicaciones (CFCA, por sus siglas en inglés), indican que: “el estimado de pérdidas anuales en el mundo por concepto de fraude está entre treinta y cinco y cuarenta billones de dólares estadounidenses” (Baluja y Llanes, 2005, p.1). Estudiaremos el caso de Honduras, para poder conocer en qué medida los operadores locales están sufriendo estas pérdidas que se reportan a nivel mundial.

Las vulnerabilidades que presenta el servicio de telefonía móvil varía desde aspectos de carácter físico, relacionados al tema de movilidad de los equipos terminales, hasta aspectos de seguridad en las redes de servicio. Los operadores de telefonía móvil en Honduras operan desde el 1996, y hasta fines del 2005 las empresas debían recurrir a Hondutel, la empresa estatal, para cursar los servicios de larga distancia internacional de sus respectivos clientes, debido a la existencia de un régimen de exclusividad que favorecía a dicha empresa (Tribunal Superior de Cuentas, 2005, p.3). En el 2006, se vence el tiempo de exclusividad, lo que significa que las empresas privadas pueden instalar sus propios medios de transporte de llamadas internacionales.

Con la expiración de la exclusividad de Hondutel y debido al rápido crecimiento de suscriptores de telefonía móvil, la mayor parte de las llamadas internacionales se dirigían entonces hacia teléfonos móviles. Con este nuevo medio de transporte, se puede enviar directamente tráfico gris hacia el operador de telefonía móvil, ya que en las llamadas internacionales habitualmente no se identifica el teléfono origen de la

llamada, por lo que el cargo a realizar se efectúa con un costo mucho menor al que se cobraría por ser una terminación de llamada proveniente de un destino internacional (Tribunal Superior de Cuentas, 2005, p.8). En la Figura 1 se muestra los involucrados en el transporte de llamadas de tráfico gris hacia empresas de telefonía móvil en Honduras.

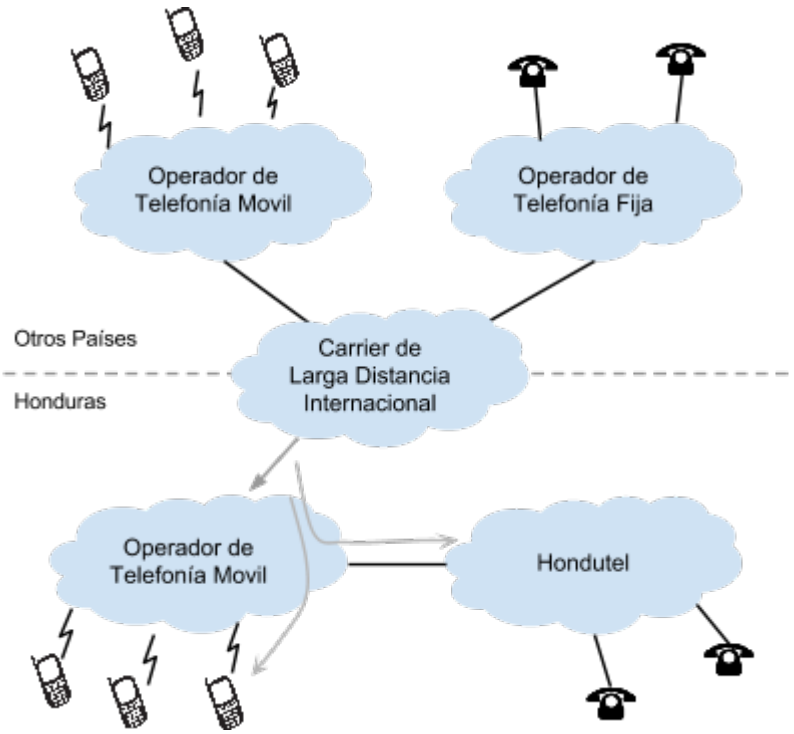

Figura 1. Flujo de Tráfico Gris Mediante un Operador de Telefonía Móvil
Fuente: Tribunal Superior de Cuentas, 2005

Para la estimación de la magnitud del tráfico gris, se debe analizar el tráfico internacional de entrada registrado en Honduras, ya que el tráfico gris sigue rutas no habituales para entrar al país, de modo que no es registrado por los operadores locales. Sin embargo, el mismo tráfico puede ser considerado como legal en otros países, de modo que en el origen del tráfico entrante a Honduras sí es registrado por los carriers y reportado ante los entes reguladores. De esta manera, una aproximación a la cantidad de tráfico gris corresponde a la diferencia entre el tráfico reportado por un país externo, versus el tráfico reportado por Honduras. En la Tabla 1 se muestra la información del Tribunal Superior de Cuentas al año 2005, tomando los Estados Unidos de América como ejemplo.

Tabla 1. Estimación de la Magnitud del Tráfico Gris desde EUA (millones de min)

Tipo de Tráfico	2002	2003	2004	2005
Internacional de Entrada	325.4	404.2	513.4	611.4
Internacional desde los EUA	296.1	367.8	467.2	556.4
Registrado por la FCC (regulador EUA)	370.7	423.6	580.4	711.4
Pérdida (tráfico gris, minutos)	74.6	55.8	113.2	155.0
Pérdida (tráfico gris, en porcentaje)	20.1%	13.2%	19.5%	21.8%

Fuente: Tribunal Superior de Cuentas (2005)

Tal como lo establece el Tribunal Superior de Cuentas de Honduras (2005), se concluye que “Un volumen de tráfico gris equivalente al 20% de tráfico enviado desde EUA, significó pérdidas del orden de USD 35 millones anuales” (p.19). Estas pérdidas afectan tanto al operador estatal como a los operadores de telefonía móvil.

1.3 DEFINICIÓN DEL PROBLEMA

1.3.1 ENUNCIADO

Debido a la variedad de servicios y cantidad en alza de suscriptores de servicios de telefonía móvil, han surgido diferentes mecanismos fraudulentos para realizar llamadas internacionales, lo que representa una pérdida en la percepción de ingresos por parte del proveedor de servicios. Para el operador en estudio, es de suma importancia lograr identificar de forma efectiva las líneas móviles utilizadas para tráfico gris, ya que esto incrementa la posibilidad de bloquear de forma adecuada las llamadas fraudulentas, y así evitar riesgos de fuga de ingresos para la organización.

1.3.2 FORMULACIÓN DEL PROBLEMA

Existe una deficiencia en la detección de líneas telefónicas utilizadas para tráfico gris, dado que los factores principales estudiados son de carácter variable y los métodos tradicionales de detección carecen de la flexibilidad requerida para tal efecto.

1.3.3 PREGUNTAS DE INVESTIGACIÓN

1. ¿Qué factores son determinantes para la detección de líneas telefónicas utilizadas para tráfico gris?
2. ¿Cómo se caracteriza el perfil de una línea telefónica legal?
3. ¿Cómo se caracteriza el perfil de una línea telefónica que realiza tráfico gris?
4. ¿Cómo puede identificarse automáticamente el comportamiento ilegal de una línea telefónica?

1.4 OBJETIVOS DEL PROYECTO

1.4.1 OBJETIVO GENERAL

Identificar factores significativos para la detección de líneas telefónicas utilizadas para tráfico gris mediante el análisis de líneas identificadas como fraudulentas para la detección de nuevos casos de fraude.

1.4.2 OBJETIVOS ESPECÍFICOS

- Conocer las características que definen el comportamiento de una línea telefónica legal.
- Analizar las características que definen el comportamiento de una línea telefónica que realiza tráfico gris.
- Determinar un mecanismo automatizado que permita diferenciar los perfiles previamente establecidos en base a comportamientos reconocidos como ilegales.

1.5 HIPÓTESIS Y VARIABLES DE INVESTIGACIÓN

1.5.1 HIPÓTESIS

Con el propósito de enmarcar el problema que se desea abordar, y para dar más claridad a los objetivos de la investigación, se formula a continuación la hipótesis:

H_i: Mediante la selección de los factores apropiados, la efectividad de detección de tráfico gris será mayor a 90%.

H_o: Mediante la selección de los factores apropiados, la efectividad de detección de tráfico gris será menor o igual a 90%.

1.5.2 VARIABLES DE INVESTIGACIÓN

En el presente estudio se consideran múltiples variables a incluir para la caracterización de un perfil de usuario, las cuales se detallan en la Figura 2:

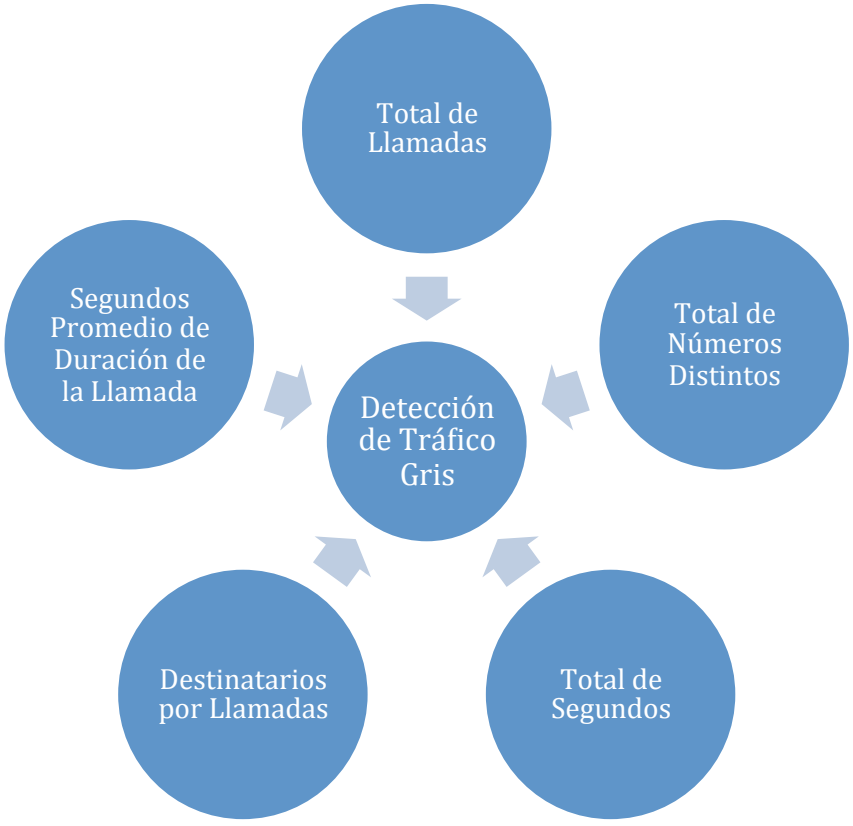

Figura 2. Variables de Estudio
Fuente: Elaboración propia según investigación

Las variables que se determinaron para la investigación se definen en la Tabla 2 para dar validez a la hipótesis sostenida en la presente investigación.

Tabla 2. Definición de variables

Variable	Definición Conceptual	Unidad de Análisis y Medición	Indicador
Detección de Tráfico Gris	La efectividad de detección de una línea utilizada para tráfico gris basada en los factores de estudio.	Análisis de datos con perceptrón multicapa.	Porcentaje de efectividad.
Total de Llamadas	El promedio del número total de llamadas diarias realizadas por una línea telefónica en el período de tiempo estudiado.	Procesamiento de CDRs.	El número de llamadas.
Total de Números Distintos	El total de números telefónicos distintos llamados por una línea telefónica en el período de tiempo estudiado.	Procesamiento de CDRs.	La cantidad de números distintos.
Total de Segundos	El total de segundos de llamada de una línea telefónica en el período de tiempo estudiado.	Procesamiento de CDRs.	La cantidad de segundos.
Destinatarios por Llamadas	La relación entre la cantidad de destinatarios distintos llamados versus el total de llamadas salientes de una línea telefónica en el período de tiempo estudiado.	Procesamiento de CDRs.	La razón porcentual de destinatarios distintos /total de llamadas.
Segundos Promedio de Duración de la Llamada	El promedio de segundos de duración de las llamadas efectuadas por una línea telefónica en el período de tiempo estudiado.	Procesamiento de CDRs.	La cantidad de segundos promedio por llamada.

Fuente: Elaboración propia según investigación.

1.6 JUSTIFICACIÓN

La detección de tráfico gris puede representar la diferencia entre percibir o dejar de percibir una importante suma de dinero para el proveedor de servicios de telefonía móvil, dado que la tarifa de llamadas internacionales es mayor a una llamada local, y el bloqueo de tráfico gris permite que las líneas deban recurrir al método tradicional de llamada internacional para poder efectuar la misma sin ser detectado como tráfico fraudulento, y por ende ser bloqueado.

Efectuando algunos cálculos sencillos, se puede estimar una cantidad de pérdidas debido al tráfico gris, en donde se deja de percibir lo que pudo haber sido una llamada internacional, pero en realidad se utilizaron medios alternos para realizar la llamada.

La razón principal que justifica los esfuerzos por detectar y reducir el tráfico gris es reducir las utilidades que se dejan de percibir debido a la realización de llamadas internacionales por vías alternas a las del operador, las cuales pueden afectar tanto al operador destino como operador origen de la llamada, puesto que existen diferentes escenarios en los que se realiza una llamada internacional, ya sea saliente de un usuario local o saliente de un usuario visitante, el cual pertenece a una red distinta a la visitada y cuya llamada representa costos a cubrir por parte de su operador, y también representa la percepción de ingresos del operador destino de la llamada debido a acuerdos de roaming entre operadores.

CAPÍTULO II. MARCO TEÓRICO

2.1 ANÁLISIS DE LA SITUACIÓN ACTUAL

El proveedor de telecomunicaciones objeto de estudio sufre la problemática de fraude en base a diversos métodos existentes, debido a que presta el servicio de llamadas internacionales hacia múltiples países y operadores a nivel global, las cuales son realizables desde líneas prepago o postpago por igual (W. Baluja & S. Llanes, 2005, p.46). La empresa telefónica también brinda el servicio de roaming internacional, disponible en más de 214 operadores de telefonía móvil distribuidos a lo largo de 124 países a nivel de los 5 continentes.

La diversidad de tipos de fraude se relaciona estrechamente con los métodos de detección a utilizar, dado que se requiere conocer factores determinantes y patrones de conducta tanto de uso legal como ilegal, como punto de partida para inferir sobre la forma en la que se identificará y subsecuentemente bloqueará el uso fraudulento del servicio de llamadas. Los métodos actuales utilizados son poco flexibles, dado que son basados en umbrales de tolerancia, los cuales se fijan de forma estática (W. Baluja & S. Llanes, 2005, p.50). Este enfoque es poco adaptativo y no considera márgenes de variación para adecuarse a escenarios con cambios condicionales, por lo que tienden a desfasarse respecto al comportamiento real del tráfico de llamadas.

2.2 TEORÍAS

2.2.1 REDES GSM

El Sistema Global para Comunicaciones Móviles (GSM, por sus siglas en inglés), es una de las tecnologías celulares más utilizadas alrededor del mundo, con una amplia variedad de proveedores por país y una numerosa cantidad de dispositivos celulares disponibles para la utilización de una red GSM. GSM se origina en 1982, cuando un grupo llamado Group Special Mobile (GSM) fue creado por la Conferencia Europea de Administraciones Postales y de Telecomunicaciones (CEPT, por sus siglas en inglés), con el propósito de diseñar una tecnología de comunicación móvil. También existe la

Asociación GSM (GSMA), la cual representa los intereses de la industria global de comunicaciones móviles. (Redl, Weber, 1998)

Cada dispositivo registrado en la red GSM es identificado y autorizado a través de un Módulo de Identidad de Suscriptor (SIM, por sus siglas en inglés), el cual es un circuito integrado que almacena la Identidad Internacional de Suscriptor Móvil (IMSI, por sus siglas en inglés) y la llave relacionada al mismo. Un circuito SIM se empotra en una tarjeta plástica removible, y es llamado tarjeta SIM, la cual puede transferirse y utilizarse entre distintos dispositivos móviles. (Redl, Weber, 1998, p.303)

GSM funciona de acuerdo a sitios en donde tiene cobertura, ya que se despliega en forma de red a través de la geografía de un país por medio de antenas, y la disponibilidad de utilización del servicio depende principalmente de encontrarse físicamente en un área de cobertura de red con un dispositivo capaz de registrarse en la red GSM.

Las llamadas realizadas en una red GSM pueden ser locales o de larga distancia, y debido a ésta última existen servicios de llamadas internacionales desde el operador local, en las que típicamente se realizan hacia proveedores diferentes fuera de la red local del operador. Este tipo de llamadas tradicionalmente incurren en cobro diferenciado por no pertenecer el destinatario al mismo operador, por lo que el costo puede ser muy alto en comparación a una llamada local (CONATEL, 2005, p.8).

En Honduras, los proveedores de telefonía móvil no tienen cobro diferenciado por regiones del país, es decir, si el destinatario de una llamada es suscriptor de un proveedor local, la llamada no se considera de larga distancia internacional, aun cuando el destinatario pueda encontrarse en una ubicación diferente al originante de la llamada, pero aún dentro de los límites del país. Debido a esto, el único tipo de llamadas con costo diferenciado son las que terminan en un número telefónico externo al país, el cual puede ser una línea fija o móvil, en las que el proveedor del destinatario es uno diferente del proveedor local.

Para minimizar los puntos de no cobertura en una red GSM, los proveedores de telecomunicaciones han determinado acordar contratos con otros proveedores para una cobertura conjunta de sus redes, con lo que se logra abarcar mayor espacio geográfico, el cual puede variar de cobertura nacional hasta internacional. A dicha extensión del servicio se le denomina roaming, lo cual se define como la extensión de la conectividad hacia una ubicación distinta a la red local donde el servicio fue suscrito, aplicable tanto a redes GSM como a redes basadas en GSM.

El término GSM roaming se define como la habilidad de un cliente móvil para automáticamente realizar o recibir llamadas de voz, enviar o recibir datos, o tener acceso a otros servicios, incluyendo servicios locales de red, al desplazarse hacia áreas de cobertura geográfica no cubiertas por su red local, mediante el uso de una red visitante. (Redl, Weber, 1998, p.303)

Los aspectos legales del negocio de roaming entre proveedores de servicio para el cobro de las utilidades obtenidas usualmente se estipulan mediante Acuerdos de Roaming. La GSMA cuenta con una forma estandarizada para la negociación del contenido de dichos acuerdos para sus miembros, entre los que se incluyen aspectos de seguridad financiera, garantía de procedimientos y de la forma en la que se realiza la actualización de ubicación de un suscriptor.

Típicamente las tarifas de roaming son en base a cargos por minuto, de acuerdo a los precios estipulados por el proveedor de servicios local. Las tarifas pueden variar dependiendo del país u operador que se visita, debido a diferencias en Acuerdos de Roaming entre proveedores de países distintos (GSMA, 2012). En Honduras, las tarifas por roaming de los proveedores locales son bastante altas en comparación al precio de una llamada local, y los suscriptores pueden revisar los precios a pagar de acuerdo al país visitante, visitando los portales web de los proveedores locales.

Dado que el servicio de roaming comprende negociaciones con otros proveedores de servicio, los costos en los que incurre un operador de telefonía móvil para ofrecer el

servicio de llamadas en roaming pueden ser muy altos. En Honduras este tipo de llamadas no se encuentran incluidas en los planes de voz tradicionales, por lo que el proveedor transfiere estos costos de infraestructura y acuerdos de roaming hacia el consumidor final.

En aras de evitar los costos, tanto de llamadas internacionales como del servicio de roaming, existen diversas técnicas fraudulentas para realizar llamadas de forma menos costosa, e incluso de forma gratuita. Este tipo de tecnologías representa una pérdida de ingresos para los proveedores de servicios móviles en especial, dado que el número de suscriptores nuevos va en incremento, y debido a esto los operadores no tienen datos históricos de cada usuario en el que se puedan basar para determinar legitimidad del uso.

2.2.2 FRAUDE EN TELEFONÍA MÓVIL

La diversidad de los tipos de fraude es tan amplio como el rango del impacto dañino que puede causar este tipo de tecnología, ya que puede llegar a afectar a más de un operador en países distintos, y por ende acuerdos de pago debido a contratos preestablecidos. En estos casos, la pérdida es del tipo moneda “fuerte”, dado que se requiere un pago entre operadores, al contrario de fraudes ocurridos a lo interno de la red de un operador, donde se involucra moneda “débil”, ya que el operador no necesita pagar a otro proveedor, el fraude solo ocurre debido al uso de su red sin percibir los ingresos asociados (Trevisan, 2000, p.61).

Un tipo de fraude consiste en la suscripción de servicios para hacer uso de roaming internacional, para el cual se realiza un cobro basado en los registros de archivos transferidos desde el operador visitado hacia el operador local, llamados archivos de procedimiento de cuenta transferida (TAP, por sus siglas en inglés). Basados en los registros de los archivos TAP, los operadores llegan a acuerdos de pago para compensar el uso de la infraestructura de red visitada como su propia, lo cual hace posible brindar el servicio de roaming a los suscriptores. El fraude por roaming

internacional es un problema de moneda fuerte, ya que aun cuando el suscriptor no pague la factura de consumo, el operador local tiene la obligación legal de pagar por el servicio prestado del operador visitado. (Trevisan, 2000, p. 61)

Otro tipo de fraude es el consistente en la clonación, en la cual la identidad de un dispositivo móvil GSM puede ser duplicada si se llega a obtener la llave de su tarjeta SIM (Trevisan, 2000, p. 61). Una vez obtenida la llave de la SIM, la seguridad se ve comprometida, ya que un suscriptor usurpador se puede autenticar en la red para realizar llamadas fraudulentas en las que el cobro se realizará a la víctima, sin haber sido éste el autor de la llamada.

En el caso de clonación de SIM, se requiere un acceso físico a la tarjeta, el cual es suficiente para poder acceder a la llave, mediante un protocolo criptográfico utilizado por la red que se encarga de exigir a la SIM la comprobación del conocimiento de la llave. Los códigos criptográficos utilizados en las tarjetas SIM no son lo suficientemente robustos para resistir ataques de este tipo, ya que a través de fuerza bruta el atacante puede interactuar repetidamente con la SIM para obtener la llave a través de técnicas matemáticas.

Los diferentes métodos de fraude pueden utilizarse en conjunto y forman un nuevo tipo de fraude, más sofisticado y complejo. Por ejemplo, cuando el método de clonación de SIM es utilizado en conjunto con el servicio de roaming internacional, la pérdida potencial de ingresos se vuelve más grande para el proveedor, y también aumenta en complejidad para la detección, dada la dispersión de este tipo de líneas a través de diferentes países y operadores a nivel global.

Con el surgimiento de nuevas tecnologías de comunicaciones, también se crean nueva formas de fraude. Este es el caso del fraude por SIM box, el cual se basa en la tecnología de Voz Sobre IP (VoIP). Las investigaciones existentes sobre fraudes se enfocan generalmente en los fraudes por suscripciones, los cuales son los tipos dominantes de fraude a nivel global en la industria de las telecomunicaciones. Sin

embargo, existe otro tipo de fraude llamado SIM box bypass, el cual se ha convertido en un amenaza que representa un reto para numerosas compañías de telefonía móvil (Elmi, Ibrahim, Sallehuddin, 2013, p.576). En Honduras y en otros países de Centro y Sur América, éste tipo de fraude es muy común.

El éxito del fraude por SIM box depende en la obtención de grandes cantidades de tarjetas SIM, por lo que los efectos varían dependiendo de las regulaciones correspondientes a cada país. En países donde las tarjetas SIM no registradas no son permitidas y las leyes de gobierno reconocen a los dispositivos SIM box como equipo ilegal, el efecto es mucho menor comparado a países donde la obtención de tarjetas SIM por clientes es barata e incluso puede llegar a ser gratuita, y dónde las leyes del gobierno no prohíben usuarios no registrados. En Honduras, la ley requiere que todo usuario de telefonía móvil se encuentre registrado (CONATEL, 2012, p.13).

La existencia de este tipo de fraude es un hecho, sin embargo, los métodos de detección del mismo no son de interés para todos los operadores de telefonía móvil a nivel global. El hecho que, debido a variaciones en leyes gubernamentales y de telecomunicaciones, este tipo de fraude representa un problema únicamente para ciertas compañías de telecomunicaciones a nivel mundial, podría ser la razón por la cual existe muy poca documentación pública de este tipo de fraude.

El fraude por SIM box toma lugar cuando individuos u organizaciones adquieren miles de tarjetas SIM, ofreciendo llamadas gratuitas o de bajo costo a líneas móviles. Las tarjetas SIM son utilizadas para canalizar llamadas nacionales o internacionales fuera del operador de red móvil y entregarlas como llamadas locales (Elmi, Ibrahim, Sallehuddin, 2013, p.576). Esta práctica fraudulenta motiva a los suscriptores ofreciéndoles un servicio de bajo costo, por lo que el suscriptor evade los costos de una llamada internacional, o los cargos de roaming en su factura. Incluso puede ocurrir que el consumidor final no se entere que su llamada está siendo transportada mediante una SIM box.

2.2.3 TÉCNICAS DE INTELIGENCIA ARTIFICIAL

La naturaleza dinámica y cambiante del fraude en las telecomunicaciones ha llevado a las empresas a abordar el problema utilizando diversas técnicas avanzadas de análisis y reconocimiento de información. Estas técnicas se basan en el estudio de los registros generados por la actividad de los suscriptores de telefonía y en la detección de patrones anormales en su comportamiento.

Debido a que los modelos tradicionales basados en reglas estáticas son muy rígidos, se han creado métodos de detección más flexibles basados en técnicas de inteligencia artificial. Las técnicas más utilizadas para este fin se basan en conceptos como las redes neuronales artificiales, la minería de datos y la lógica difusa. El presente estudio está orientado a la utilización de un modelo basado en redes neuronales artificiales, de forma más específica, una red perceptrón multicapa.

2.2.3.1. REDES NEURONALES ARTIFICIALES

Una de las técnicas de inteligencia artificial más utilizadas en la detección de patrones son las redes neuronales artificiales. La definición de red neuronal artificial propuesta por Graupe (2007) es la siguiente:

> Las redes neuronales artificiales son, como su nombre lo indica, redes computacionales que intentan simular, de forma básica, las redes de células nerviosas (neuronas) del sistema biológico central (humano o animal) (...) permiten el uso de operaciones de cálculo muy simples (adiciones, multiplicación y elementos lógicos fundamentales) para resolver problemas matemáticos complejos, problemas no lineales o problemas estocásticos (p.1).

Las redes neuronales artificiales sirven como apoyo en la resolución de problemas complejos, sin embargo, no todo tipo de problema puede ser adaptado a una solución con redes neuronales. Heaton (2008) afirma:

> Las redes neuronales a menudo no son adecuadas para los problemas en los que se debe saber exactamente cómo se obtuvo la solución. Una red neuronal

> puede ser muy útil para resolver el problema para la que fue entrenada, pero la red neuronal no puede explicar su razonamiento. La red neuronal sabe algo porque fue entrenada para saberlo. La red neuronal no puede explicar cómo se siguió una serie de pasos para obtener la respuesta" (p.43).

Sin embargo, a pesar de que las redes neuronales no son adaptables a cualquier tipo de problema, el reconocimiento de patrones es una de las áreas en las que más se utiliza este enfoque:

> El reconocimiento de patrones es quizá el uso más común de las redes neuronales. Para este tipo de problema, se presenta a la red neuronal un patrón. Este puede ser una imagen, un sonido o cualquier otro tipo de dato. La red neuronal entonces trata de determinar si el dato de entrada coincide con el patrón que ha sido entrenada para reconocer (Heaton, 2008, p.44).

Para la resolución de problemas de reconocimiento de patrones utilizando redes neuronales, se debe entrenar la red previamente para reconocer el patrón de interés. Es decir, la red neuronal primero debe aprender a reconocer el patrón. "En el contexto de redes neuronales, el aprendizaje está definido como un proceso por el cual los parámetros libres de una red neuronal se adaptan a través de un proceso continuo de la estimulación por el medio ambiente" (Yeung, Cloete, Shi y Ng, 2010, p.5). En el entrenamiento de redes neuronales, existen dos tipos de aprendizaje:

- Aprendizaje supervisado: Durante la sesión de entrenamiento, se aplica una entrada a la red neuronal y se obtiene una respuesta. La respuesta obtenida es posteriormente comparada con una respuesta esperada. Si la respuesta obtenida difiere de la respuesta esperada, la red neuronal genera una señal de error, a partir de la cual se calcula el ajuste que debe hacerse a los pesos de la red de manera que la respuesta coincida con la respuesta esperada (Yeung et al., 2010, p.5).

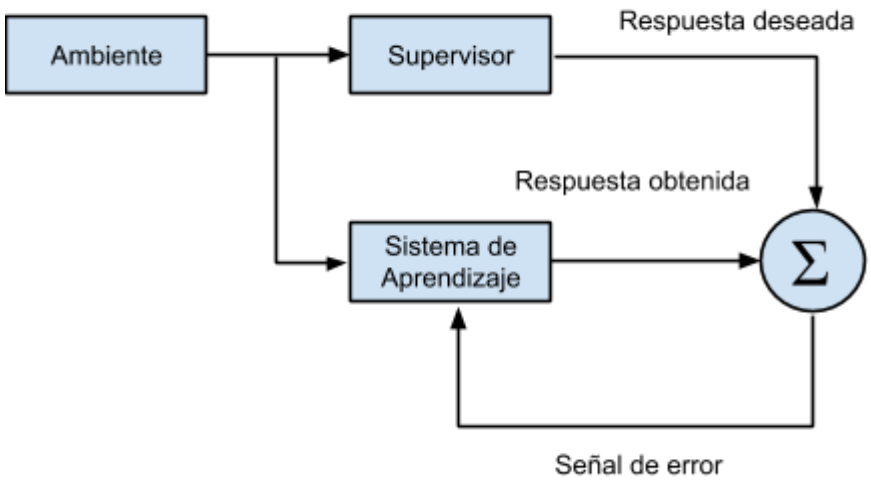

Figura 3. Modelo de aprendizaje supervisado.
Fuente: Yeung et al. (2010)

- Aprendizaje no supervisado: En el aprendizaje no supervisado no se utiliza una respuesta esperada. Durante la fase de entrenamiento, la red neuronal recibe patrones de entrada los cuales organiza arbitrariamente en categorías. Luego, cuando se aplica una entrada a la red, ésta provee una respuesta de salida indicando a qué categoría corresponde la entrada (Yeung et al., 2010, p.5).

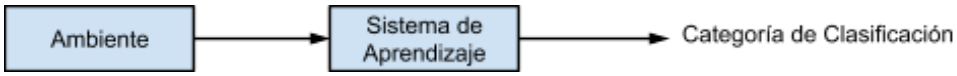

Figura 4. Modelo de Aprendizaje No Supervisado.
Fuente: Yeung et al. (2010)

El aprendizaje de la red neuronal tiene que realizarse de acuerdo con la estructura de la misma. Yeung (2010) caracteriza la estructura de una red neuronal de la siguiente forma:

> Una red neuronal artificial es una arquitectura que consiste de varias neuronas artificiales, las cuales trabajan en conjunto para dar respuesta a las entradas. Algunas veces se considera una red neuronal como una función de caja negra. Aquí, el mundo exterior presenta sus entradas en las neuronas de entrada y recibe sus salidas de las neuronas de salida. Las neuronas intermedias no son

vistas externamente, por eso comúnmente se les conoce como unidades ocultas (p.3).

Cada neurona artificial es un elemento que posee un estado interno conocido como nivel de activación, el cual recibe señales que le permiten cambiar de estado. Las neuronas poseen una función denominada función de transición de estado o función de activación, la cual les permite cambiar de nivel de activación a partir de las señales que reciben. Una neurona puede recibir señales del exterior o de otras neuronas conectadas a la misma. La estructura básica de una neurona artificial básica se muestra en la Figura 5 (Galván, 2004, p.5).

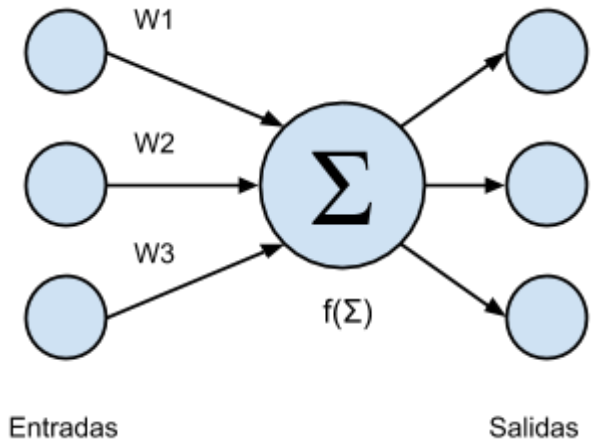

Figura 5. Esquema de Neurona Artificial.
Fuente: Galván (2004)

En esta estructura, la neurona recibe entradas desde otras neuronas de la red con pesos específicos (W1, W2 y W3). La neurona calcula el valor de salida aplicando la función de activación a los valores de entrada. El resultado de dicha función es propagado a través de todas las salidas de la neurona. A la forma en que se conectan entre sí las neuronas artificiales para formar redes se le denomina patrón de conectividad o arquitectura de la red. La arquitectura básica de una red neuronal artificial es la red multicapa, la cual se muestra en la Figura 6 (Galván, 2004, p.7).

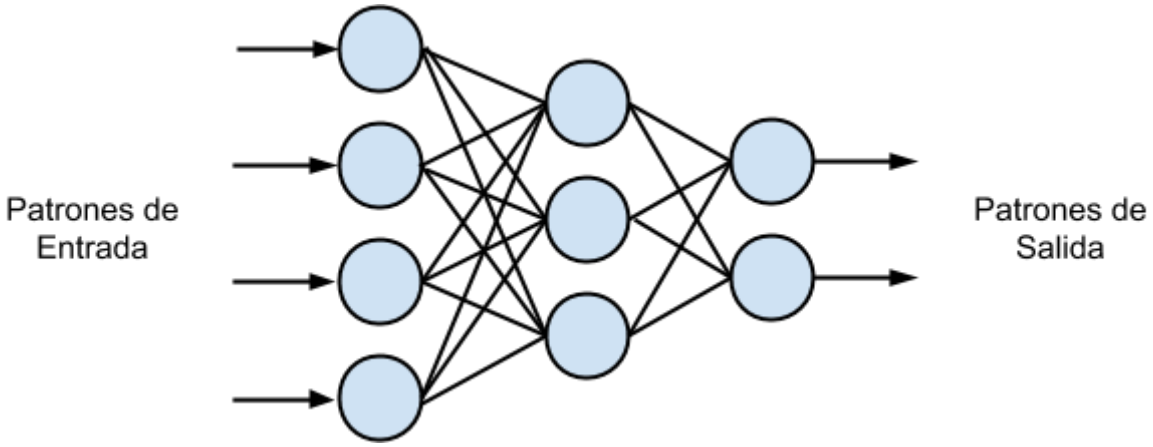

Figura 6. Arquitectura de Red Neuronal Multicapa.
Fuente: Galván (2004)

2.2.4 MÉTODOS DE DETECCIÓN DE FRAUDE

Algunas teorías de fraude se enfocan a la monitorización a través de un sistema especializado para los elementos de red del proveedor. Este sistema debe ser lo suficientemente flexible para poder adaptarse a los emergentes sistemas de fraude, y posiblemente contar con monitorización a nivel de protocolos estándar de red, tales como SS7, el cual se utiliza para la señalización en todas las llamadas de voz. El sistema propuesto por dichas teorías, debería ser capaz de proveer información de las llamadas en tiempo real, incluyendo detección de roaming, estudio de perfiles de usuario, análisis en base a zonas geográficas, y otros factores varios (Trevisan, 2000, p. 62).

En contraste a algunas soluciones de detección de fraude vía monitorización, existen soluciones basadas en el post-procesamiento de los datos provenientes de los elementos de red del operador. En base a toda la información recopilada, es posible estudiar, analizar y generar patrones y perfiles de comportamiento, a través del procesamiento de grandes cantidades de archivos los cuales se ingresan a un sistema central para su posterior evaluación. Esta práctica es conocida como minería de datos, en la cual es necesario un tiempo considerable de procesamiento, por lo que existe cierta debilidad en cuanto al tiempo de respuesta que pueda tener una solución como

ésta, dado que se los resultados se obtienen en base a una serie de procesos a ejecutar a través de diversos elementos de red (Trevisan, 2000, p. 62).

Para la detección de comportamiento fraudulento en compañías telefónicas, la minería de datos debe basarse en la información generada por los consumidores al hacer uso de los servicios. Las compañías de telecomunicaciones almacenan grandes cantidades de registros de llamadas de los consumidores, los cuales se conocen como CDRs. “Cuando un suscriptor realiza una llamada en la red del operador, se prepara un ticket que contiene información completa de la llamada realizada, incluyendo la identificación del suscriptor, el número al que llamó, la duración de la llamada, la hora, el destino y otros” (Elmi, Ibrahim, Sallehuddin, 2013, p.577). De esta forma, las bases de datos de CDRs contienen información muy valiosa sobre el uso de los servicios de telefonía y pueden ser muy útiles para estudiar el comportamiento de los consumidores y detectar patrones de fraude.

Rodríguez y Sánchez (2005) afirman que las bases de datos utilizadas como fuente de información para la detección de fraude telefónico deben cumplir las siguientes condiciones: “una, ser lo suficientemente amplias para permitir la acertada caracterización del fenómeno y atenuar la influencia de cualquier anomalía temporal; y la otra, estar actualizadas para garantizar que la caracterización evolucione al ritmo del fenómeno” (p.48).

La combinación de las teorías es la que está siendo adoptada por empresas de telefonía móvil, dado que la flexibilidad de un sistema en tiempo real de la mano con la robustez de la minería de datos provee los tiempos de respuestas requeridos para detectar efectivamente actos fraudulentos en la red móvil, y por tanto cumple con el requisito primordial de brindar detección en tiempo real.

El enfoque para detección de fraude propuesto por Rodríguez y Sánchez (2005), establece como fuentes de información un conjunto de CDRs, los flujos de información generados por los protocolos SS7 y los registros de información de roaming. Este

enfoque propone el almacenamiento inicial de los flujos de información en una base de datos centralizada para su posterior clasificación de acuerdo a ciertos atributos previamente establecidos. El resultado de ésta clasificación debe ser procesado por un sistema experto empleando reglas y umbrales, a partir de los cuales se deben generar alertas en caso de detectarse situaciones o comportamientos atípicos. El flujo de éste modelo de detección de fraude se ilustra en la Figura 7.

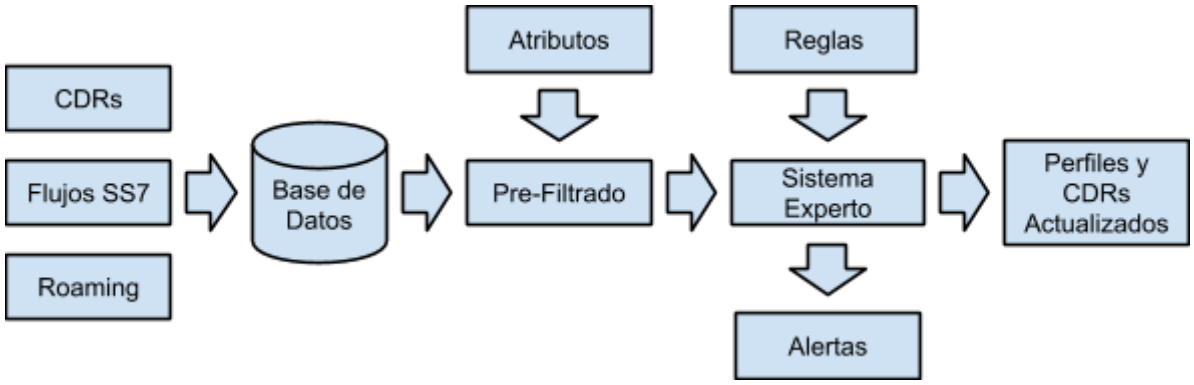

Figura 7. Modelo de Detección de Fraude.
Fuente: Rodríguez y Sánchez (2005)

Un detalle muy importante a tomar en cuenta en este enfoque es el nivel de complejidad del sistema experto que toma las decisiones. Rodríguez y Sánchez (2005) afirman:

> Los métodos convencionales de detección de fraude, basados en simples umbrales, son usados para combatir la mayoría de los fraudes. Sin embargo, algunos problemas permanecen sin solución ante ellos, al desplazarse de su comportamiento habitual, escapando así de la rigidez de sus filtros. Las técnicas de inteligencia artificial (IA), pueden ser exitosamente usadas en la lucha contra estos problemas. El término inteligencia cubre muchas habilidades conocidas, incluyendo la capacidad de solucionar problemas, de aprender y de entender métodos de solución. (p.49)

Tomando en cuenta lo anterior, se puede establecer una categorización de los sistemas expertos que realizan el reconocimiento del fraude. En primer lugar, están los sistemas rígidos que funcionan en base a reglas y umbrales preestablecidos y en segundo lugar están los sistemas flexibles, los cuales se basan en técnicas de inteligencia artificial y permiten la adaptación de la solución a las características cambiantes del problema.

Existen diversas técnicas de inteligencia artificial que pueden ser utilizadas para brindar flexibilidad, automatización y capacidad de aprendizaje a los sistemas de detección de fraude. Algunas de éstas pueden utilizarse de forma individual o pueden utilizarse en combinación con otras técnicas para la detección del fraude telefónico. Rodríguez y Sánchez (2005) proponen las siguientes técnicas:

- Sistemas de aprendizaje de máquina automatizado: son sistemas que, a partir de una base de casos inicial crean reglas que caracterizan los casos analizados. Estas reglas luego pueden ser ingresadas en un sistema de inferencia adaptativo neurodifuso (ANFIS) o pueden servir como base para la construcción de un sistema de inferencia difusa (DIS) para luego inferir.

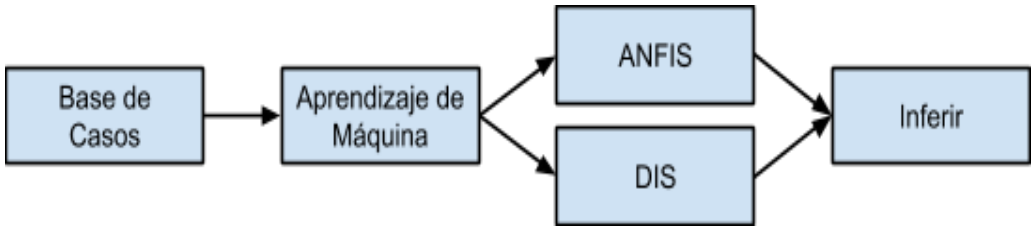

Figura 8. Sistema de Aprendizaje de Máquina Automatizado.
Fuente: Rodríguez y Sánchez (2005)

- La creación de un conjunto de reglas difusas por parte de personas expertas, las cuales pueden ser utilizadas para la construcción de un DIS o pueden ser mejoradas por medio de un ANFIS para luego inferir.

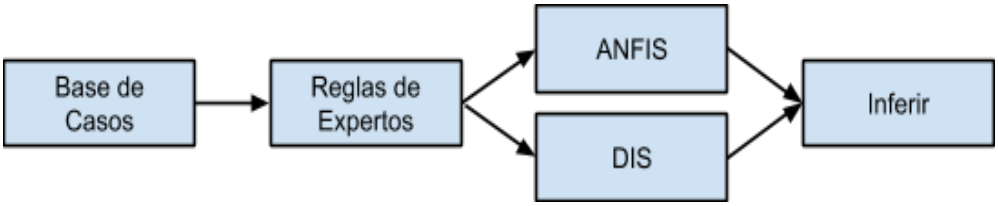

Figura 9. Sistema Basado en Reglas Difusas de Expertos.
Fuente: Rodríguez y Sánchez (2005)

- Redes neuronales artificiales, en especial, el perceptrón multicapa (MLP). En este caso, se brinda a la red neuronal una base de casos clasificada, para que ésta encuentre correlaciones entre los casos y aprenda a detectar irregularidades en nuevos casos proporcionados.

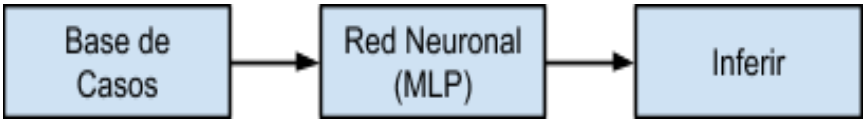

Figura 10. Modelo Basado en Perceptrón Multicapa.
Fuente: Rodríguez y Sánchez (2005)

2.2.5 BASES LEGALES

2.2.5.1 REGULACIONES EMITIDAS POR CONATEL

El ente regulador de las telecomunicaciones en Honduras es la Comisión Nacional de Telecomunicaciones CONATEL, la cual es una entidad desconcentrada de la Secretaría de Estado en el Despacho de Finanzas, respecto de la cual funciona con independencia técnica, administrativa y presupuestaria. (CONATEL, 1995, p.4)

CONATEL toma la atribución de promover la universalización de los servicios de telecomunicaciones y procurar su más alta calidad y menor costo posible, adoptando las medidas necesarias para que dichos servicios se brinden de forma eficiente, ininterrumpida, sin interferencia y sin discriminaciones. También es facultad de CONATEL la regulación de las tarifas de las empresas de telecomunicaciones, así

también el establecimiento de las bases para la interconexión de las redes de telecomunicaciones en Honduras.

En el caso que una empresa de telecomunicaciones incurra en incumplimiento de las leyes de CONATEL, recibirá una sanción correspondiente a la criticidad de la infracción. CONATEL realiza supervisiones a las empresas de telefonía móvil para determinar el cumplimiento de los artículos establecidos en la Ley Marco del Sector de Telecomunicaciones, entre los cuales existen prohibiciones de las prácticas fraudulentas en los servicios telefónicos:

> Artículo 47: se prohíbe la práctica de llamadas revertidas que involucren servicios telefónicos prestados dentro del territorio nacional, que son sistemáticamente originados fuera del país como resultado directo de llamadas internacionales no completadas, originadas dentro del territorio nacional, a menos que el proveedor de dichos servicios tenga una concesión debidamente otorgada por CONATEL para la prestación de servicios finales públicos (CONATEL, 1995, p.12).

Las supervisiones de CONATEL velan por el cumplimiento de todos los artículos de la Ley Marco del Sector de Telecomunicaciones, y la empresa tiene la obligación de presentar la información requerida por CONATEL para demostrar el cumplimiento del Marco Legal respectivo.

CONATEL, a través de su Reglamento de Interconexión en el artículo 15, regula el acceso a operadores de Gestión y Establecimiento de Comunicaciones de larga distancia, indicando que los operadores deberán permitir a todos sus suscriptores, mediante la presuscripción, y a todos sus usuarios, mediante marcación directa de prefijos de acceso asignados por CONATEL, elegir al operador autorizado que prefieran para gestionar y establecer sus comunicaciones de larga distancia (CONATEL, 2003).

2.2.6 BASES FINANCIERAS

La justificación financiera para enfocar esfuerzos en sistemas de detección de fraude es basada en las pérdidas en las que incurre el operador de telefonía móvil debido a la evasión de los métodos tradicionales de acceso, con lo cual se evita el pago

correspondiente y por ende se deja de percibir ingresos por parte del proveedor .Los sistemas de detección de fraude son una oportunidad de brindar más libertad al usuario o a la organización en general, en términos de restricciones de uso que puedan limitar al consumidor final o a la empresa, debido a posibles limitaciones en productos nuevos en aras de evitar prácticas fraudulentas.

Una red segura es un factor importante para la buena imagen de los operadores de telefonía móvil, y convertirse en una ventaja competitiva respecto a sus competidores. Con la adopción de un sistema de detección de fraudes, el proveedor de servicios puede evitar fuga de ingresos, ya sea en moneda dura o suave, es decir, operaciones que requieren realizar un pago entre operadores u operaciones que no incurren en ningún pago entre dichos operadores. Como beneficio adicional de ser capaz de detectar fraudes en la red, las relaciones con otros proveedores de servicios también pueden verse mejoradas, dado que un socio proveedor confiable en términos de seguridad y control de fraude es apreciable para múltiples operadores a nivel global (Trevisan, 2000, p. 61).

Los operadores de telecomunicaciones comprenden el valor de optimizar los ingresos en servicios de telefonía móvil, así como el mejoramiento de la calidad y seguridad de los servicios, en aras de maximizar los márgenes de ganancia e incrementar los niveles de satisfacción de los clientes, por tanto, buscan adoptar las medidas y prácticas recomendadas de la industria para mejorar estos aspectos.

Considerando que la empresa de telefonía móvil en estudio tuviese 2 millones de suscriptores con un ingreso promedio por usuario (ARPU, por sus siglas en inglés) de $5 mensual, representaría $10 millones de ingreso mensual. Si asumimos de acuerdo a los estudios en la materia, que el porcentaje de pérdidas debido a fraude es de un 7%, el costo de fraude al año representa la importante suma de total de $8.4 millones anual. Si un sistema de detección de fraude puede disminuir el porcentaje de pérdida debido a estos actos, el retorno de la inversión asumida para la implementación de dicho sistema

puede obtenerse fácilmente al cabo de pocos meses a partir de su integración, incluso si el porcentaje de casos de fraude bloqueado es relativamente bajo.

El fraude puede ocasionar pérdidas tanto en escenarios locales como en escenarios de roaming, pudiendo afectar los ingresos percibidos por llamadas internacionales, o en llamadas realizadas en escenarios de roaming, entrantes o salientes.

2.2.7 MARCO REFERENCIAL

La aproximación de los estudios existentes hacia el fraude por medio de tráfico gris y sus diversos métodos de ejecución presenta una serie de descriptores que pueden ser utilizados en la detección de tarjetas SIM utilizadas en dispositivos especiales para fraude, llamados SIM Box. Este tipo de aplicaciones de redes neuronales ha generado expectativas superiores de soluciones efectivas, ya que son capaces de aprender de patrones complejos y tendencias a partir de datos no estructurados.

Las redes neuronales han mostrado un rendimiento superior respecto a otras técnicas en el ámbito del fraude en las telecomunicaciones. En el estudio referencial se ha aplicado el método de aprendizaje asistido utilizando el Perceptrón Multicapa como clasificador. El conjunto de datos utilizado se obtiene de una red de comunicaciones móvil real y contiene suscriptores y tarjetas SIM que han sido probadas y aprobadas por el operador como utilizadas para el fraude por SIM box, o tarjetas SIM normales.

Los operadores de telefonía móvil almacenan grandes volúmenes de información respecto a las llamadas realizadas a través de la red, a través de archivos CDRs. Cada vez que un usuario genera una llamada a través de la red móvil, se genera un CDR, en el que se incluye una variedad de información de la llamada, como ser el id de suscriptor, el número destino, duración de la llamada, hora, ubicación del destinatario, etc. En el escenario de fraude estudiado, la base de datos de CDRs sirve como una importante fuente de donde se puede extraer información para obtener conocimientos útiles sobre los usuarios, y así identificar llamadas fraudulentas realizadas por los mismos.

El estudio referencial es basado en una red GSM, orientado específicamente a CDRs de usuarios prepago. El conjunto de datos utilizado para éste experimento contiene 234,324 llamadas realizadas por un total de 6,415 usuarios en una celda de la red. El conjunto de datos consiste en 2,126 suscriptores fraudulentos y 4,289 suscriptores normales, lo que equivale a 66.86% usuarios legítimos y 33.14% de usuarios fraudulentos. Los registros obtenidos de estas transacciones de llamadas se obtuvieron en un periodo total de dos meses.

En el estudio referencial se utiliza una red neuronal con prealimentación, con el propósito de alimentarla a través de un algoritmo de propagación hacia atrás, minimizando así los errores a través del entrenamiento. Un total de 9 factores han sido identificados como útiles para la detección de fraude por SIM box. Los resultados del experimento revelan que la red neuronal artificial (ANN) ha mostrado una precisión en la clasificación de un 98.71%, el cual representa una detección altamente efectiva.

CAPÍTULO III. METODOLOGÍA

En el presente capítulo, se detalla el esquema de diseño del estudio, describiendo los enfoques y métodos que conforman la metodología de investigación utilizada, con el objetivo de contar con una guía a seguir para realizar el trabajo investigativo y analítico.

3.1 ENFOQUE Y MÉTODOS

El estudio se enmarca dentro de un enfoque mixto, caracterizado por su composición variada de criterios de investigación, procesos sistemáticos y empíricos en donde resulta necesario analizar de forma estructurada los datos recolectados de manera que se pueda buscar comportamientos y relaciones entre elementos, para lograr demostrar la hipótesis propuesta, mediante la integración y discusión de los datos cualitativos y cuantitativos recolectados.

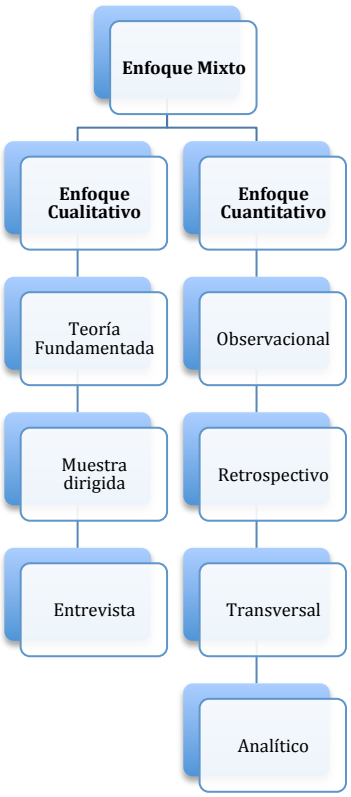

Figura 11. Enfoque y Métodos de Estudio
Fuente: Elaboración propia según investigación

En la Figura 11 se detalla el enfoque utilizado en la investigación, de carácter mixto pero con dominancia cuantitativa, ya que ésta es complementada por el enfoque cualitativo, el cual aporta valiosos puntos de vista para el estudio. El enfoque cualitativo marca la pauta sobre la aproximación a tomar para el estudio cuantitativo, ya que es en el estudio cualitativo donde se analiza la teoría que fundamenta la investigación, y luego se selecciona a un grupo de expertos en el área de aseguramiento de ingresos, específicamente en el estudio de tráfico gris, para ser entrevistados en relación a factores determinantes, caracterización de perfiles, métodos existentes, y otros aspectos.

En el enfoque cuantitativo se recolectan datos para probar la hipótesis propuesta, en base a mediciones numéricas y análisis con métodos de aprendizaje supervisado, con los cuales se establecen patrones de comportamiento y se predicen futuros casos positivos o negativos de utilización de tráfico gris en líneas móviles. El enfoque cuantitativo en este estudio es de carácter observacional, ya que las variables independientes no se manipulan en ningún aspecto de forma deliberada, por lo contrario, las acciones se limitan a observar los datos en su comportamiento natural. El estudio es transversal, la recolección de información toma lugar en una única ocasión.

3.2 DISEÑO DE LA INVESTIGACIÓN

La clasificación del estudio de acuerdo a la intervención de los investigadores puede considerarse como cuasi experimental, debido a que los factores de estudio no serán elegidos al azar, y son de carácter externo a este estudio. Los métodos observacionales se llevan a cabo en situaciones no controladas, por las que no es factible realizar el presente estudio con ese tipo de aproximación por parte de los investigadores.

Tal como es mostrado en la Figura 12, el proceso de investigación se desarrolla en las siguientes fases:

Fase 1: Identificación de características utilizadas por expertos para la caracterización de líneas telefónicas móviles utilizadas para tráfico gris.

Fase 2: Identificación de técnicas comúnmente utilizadas para la detección de líneas telefónicas móviles que cursan tráfico gris.

Fase 3: Indagación sobre el comportamiento promedio de un usuario regular (no fraudulento) en una empresa de telefonía móvil.

Fase 4: Recopilación e investigación de casos históricos reales de tráfico gris y casos de tráfico legal en una empresa de telefonía móvil.

Fase 5: Construcción de un modelo para la detección de tráfico gris, basado en redes neuronales (perceptrón multicapa).

Fase 6: Entrenamiento de la red neuronal con los casos históricos de fraude y legales previamente obtenidos.

Fase 7: Ejecución de las pruebas de detección de casos de tráfico gris, utilizando la red neuronal entrenada en la fase previa.

Fase 8: Análisis de los resultados y conclusiones del estudio.

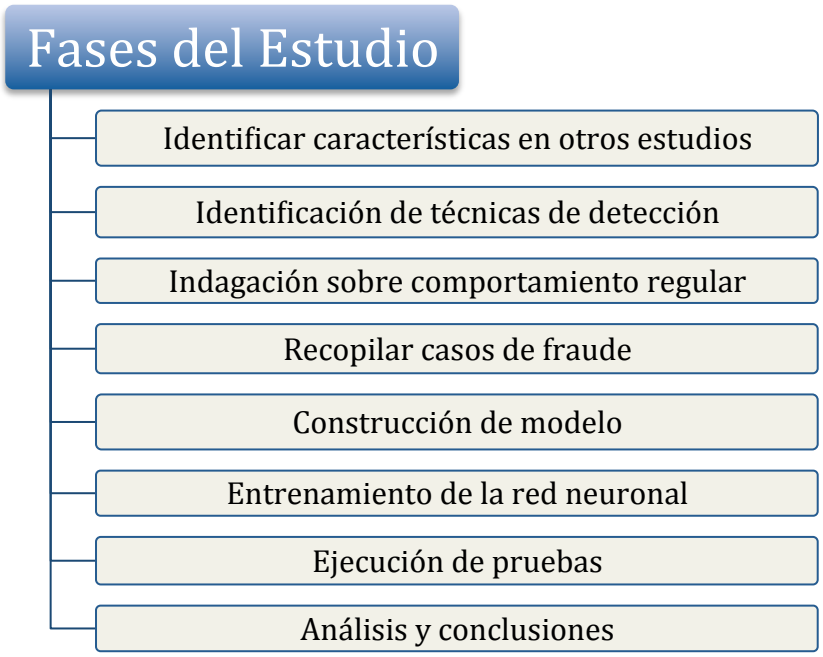

Figura 12. Diseño del Estudio
Fuente: Elaboración propia según investigación

3.2.1 POBLACIÓN Y MUESTRA

Como parte de la población elegida para el presente estudio, se consideran los registros de casos donde se ha identificado positivamente la utilización de las líneas telefónicas para tráfico gris, es decir, se recopilaron datos históricos reales sobre casos positivos de fraude de este tipo. Para efectos de muestreo y de comprobación de las herramientas utilizadas, se utiliza una población que reproduzca de la mejor manera los factores significativos, y por ello debe elegirse con un muestreo intencional-discrecional, caracterizado por el esfuerzo deliberado de obtener muestras representativas, incluyendo casos típicos de tráfico gris.

3.2.2 UNIDAD DE ANÁLISIS Y RESPUESTA

La entidad mayor o representativa del presente estudio son todos los suscriptores del servicio de telefonía en una empresa de telefonía móvil. Es importante comprender que se considera como unidad de análisis la totalidad de los suscriptores debido a que el estado de un suscriptor puede variar en el tiempo, es decir, una línea telefónica

utilizada para tráfico legal podría luego ser utilizada para comportamiento fraudulento, por ejemplo, en tráfico gris. La unidad de respuesta consiste en el porcentaje de efectividad de detección de tráfico gris en base a los factores identificados en el estudio.

3.3 TÉCNICAS E INSTRUMENTOS APLICADOS

Con el objetivo de recopilar datos de interés para la investigación, se ha solicitado a la empresa en estudio brindar acceso a registros dentro de sus herramientas de procesamiento de datos. Por medio de una entrevista al personal experto del área, se obtuvo información sobre los perfiles de los usuarios de telefonía móvil, los factores que comúnmente son utilizados para la detección de tráfico gris y las técnicas que hasta ahora han sido utilizadas para esta tarea.

3.3.1 INSTRUMENTOS

Los instrumentos utilizados en la investigación para la recopilación de los datos necesarios para el análisis son la entrevista y las bases de datos.

3.3.1.1 ENTREVISTA

Por medio de una entrevista al personal experto del área se logró conocer de forma más cercana los criterios aplicados para el análisis de casos de tráfico gris. Aun cuando las técnicas de detección más comunes son basadas en métodos estructurados, siempre el criterio del experto analista será un factor contribuyente a la efectividad del análisis realizado.

Por medio de la entrevista se obtuvieron opiniones valiosas sobre los posibles factores que caracterizan un perfil de tráfico gris en telefonía móvil, de primera mano con expertos que han trabajo en diversas actividades relacionadas a la detección de tráfico gris. La entrevista se aplicó a expertos en el área de telefonía móvil con experiencia comprobada en el área de estudio.

3.3.1.2 BASES DE DATOS

Las bases de datos proveen una fuente confiable, integral y completa de la información necesaria para contar con datos representativos de la población total del estudio. Dentro del marco de las bases de datos podemos destacar que la masividad de los registros almacenados es tal, que resulta necesario encontrar métodos que permitan descubrir el conocimiento a partir de la gran cantidad de información, por lo que se aplican métodos matemáticos de análisis para encontrar los registros de valor para el estudio. En resumen, la minería de datos es aplicada a la información suministrada por la base de datos, de forma que el análisis pueda ser más expedito comparado a tomar el total del contenido de la base de datos.

3.3.2 TÉCNICAS

El presente estudio tiene como fuentes de información las bases de datos en las que se recopilan registros de llamadas mediante un procesamiento de archivos CDRs. Con la centralización de dicha información, se procede a tomar una muestra de la base total para hacer un análisis de comportamiento sobre ella, y así poder proponer una base de comportamiento legal. No obstante, para efectos de un análisis objetivo y fundamentado, se requiere contar con los aportes de los especialistas en el área, por lo que el estudio considera como una técnica importante la entrevista a personas expertas en el área de aseguramiento de ingresos, como sustento al enfoque de análisis utilizado sobre la técnica de recolección de datos en la base de datos.

3.3.3 PROCEDIMIENTOS

Para efectos de recolección de datos cualitativos, se realizaron entrevistas con expertos calificados en el área de aseguramiento de ingresos, específicamente en la detección y prevención del tráfico gris. La entrevista se formuló de manera escrita, como un cuestionario abierto, para que los expertos pudiesen expresar sus opiniones libremente y sin limitarse a respuestas preestablecidas. Posterior a la formulación de la entrevista, ésta se digitalizó y fue enviada por correo electrónico a los expertos para ser completada. Una vez se recibió confirmación de parte de cada experto de haber

completado la entrevista, se consolidaron los resultados para analizar y comparar cada respuesta brindada.

Para el análisis cuantitativo, se obtuvo acceso a la base de datos de una empresa de telefonía móvil de Honduras, de la cual se extrajo una porción de registros de dos meses calendario, para su posterior procesamiento y análisis.

3.4 FUENTES DE INFORMACIÓN

Para enriquecer la investigación, se utilizaron diversas fuentes de información, dado que existen numerosas investigaciones, libros de texto, revistas científicas y expertos en la materia que aportaron significativamente a comprender mejor el ambiente de estudio, en este caso, los servicios de telefonía móvil, tráfico gris y redes neuronales. Se describen a continuación las fuentes primarias y secundarias utilizadas en el presente estudio.

3.4.1 FUENTES PRIMARIAS

La principal fuente de datos para el estudio está conformada por las entrevistas a los expertos, y los registros de casos reales de líneas telefónicas utilizadas para tráfico gris. Estos registros fueron obtenidos mediante la recolección de archivos CDRs en las plataformas de telefonía móvil, centralizados en un sistema de procesamiento central de archivos para su posterior extracción, transformación y carga en cada uno de los sistemas adecuados para obtener la información relevante requerida.

Las fuentes primarias utilizadas brindan datos de primera mano, y en este estudio se consideran resultados de estudios anteriores sobre el uso de redes neuronales en la detección de tráfico fraudulento, se cuenta con testimonio de expertos y libros de texto de los temas en estudio.

3.4.2 FUENTES SECUNDARIAS

Las fuentes secundarias utilizadas en esta investigación son:

- Reportes elaborados por diferentes entidades nacionales del rubro de las telecomunicaciones.
- Visitas en línea realizadas a los portales web de información sobre la telefonía móvil en Honduras.
- Discusiones personales con personas conocedoras de la industria de las telecomunicaciones.

CAPÍTULO IV. RESULTADOS Y ANÁLISIS

4.1 ANÁLISIS CUALITATIVO

Los resultados del análisis cualitativo se explicarán a detalle en esta sección. Se ha recopilado información de expertos en el área de aseguramiento de ingresos en telefonía móvil, específicamente con experiencia en detección de tráfico gris, participando en actividades de recolección, procesamiento y análisis de datos y uso de herramientas especializadas, lo cual da una mayor amplitud al estudio en las áreas de conocimiento relacionadas a procesos de detección de tráfico gris.

La herramienta utilizada para la recopilación de datos cualitativos es la encuesta, la cual en este caso específico consta de 10 preguntas abiertas, para dar una mayor amplitud a la experiencia que puede compartir el experto al momento de ser entrevistado. Se analizarán las respuestas a cada una de las preguntas de la entrevista, para concluir destacando los hallazgos principales de la investigación.

El fundamento teórico indica que existen múltiples técnicas conocidas utilizadas por los individuos que realizan tráfico gris, y los expertos entrevistados señalan que las más utilizadas en Honduras varían entre dispositivos especializados tales como los SIM Box, simulación de comportamiento humano, automatización de actividades originadoras de tráfico, adquisición masiva y rotación de tarjetas SIM por parte de un mismo suscriptor, e incluso suplantación de líneas de usuarios genuinos. La simulación del comportamiento humano está vinculada a la automatización de las actividades que originan tráfico, enmarcadas en el contexto de usuario de telefonía móvil. Las actividades pueden ser recargas electrónicas de saldo, consulta de saldo vía SMS o USSD, llamadas de voz, navegación en red de datos móvil, etc. La Figura 13 ilustra las técnicas que resaltan los expertos entrevistados.

Figura 13. Técnicas Relacionadas al uso de Tráfico Gris en Telefonía Móvil
Fuente: Elaboración propia según investigación

En aras de modelar un perfil de usuario de una línea telefónica utilizada para tráfico gris, se consulta a los expertos sobre su experiencia en la caracterización de dicho perfil. Por medio de la entrevista, es notable que existen múltiples factores a considerar, que varían dependiendo de la amplitud del estudio a realizar y el enfoque hacia el tipo de red, en este caso, una red de telefonía móvil. Las características de usuario que resaltan en la entrevista son ilustradas en la Figura 14, y son descritas a continuación:

- Alta cantidad de llamadas: específicamente, puede ser que se den muchas llamadas en un corto periodo de tiempo, que difiere del comportamiento regular de un usuario no fraudulento.
- Mayor duración de llamadas: Este puede significar un factor irregular, cuando ocurren duraciones sustanciosas de llamadas salientes.
- Relación desproporcionada de llamadas entrantes versus llamadas salientes: la línea telefónica utilizada para tráfico gris presenta un patrón de consumo anómalo entre su tráfico de llamadas entrantes y salientes. La variación entre cada tipo de llamada suele ser mucho mayor cuando la línea se utiliza para éste tipo de fraude, especialmente notable en el aumento significativo de llamadas salientes.
- Desvío significativo de una actividad específica: Un usuario regular varía sus actividades entre los diferentes productos que ofrece el operador de telefonía móvil, por ejemplo, llamadas de voz, mensajes de texto (SMS), navegación en

red de datos Móvil o Internet, llamadas a servicios USSD, llamadas a códigos cortos de servicios como consulta de saldo, pagos vía operador móvil, etc. Sin embargo, una línea telefónica utilizada para tráfico gris tiene un comportamiento que genera una cantidad irregular de transacciones de un solo tipo en un corto periodo de tiempo, dada la repetición de un patrón específico de comportamiento por tiempo prolongado.

- Ubicación típicamente invariable: la línea telefónica utilizada para tráfico gris tiende a ubicarse en la misma zona geográfica la mayor parte de tiempo, como consecuencia de utilizarse en SIM Box. El estudio de este comportamiento se realiza con la obtención de la ubicación de usuario que provee los registros de la llamada, y es notable la poca movilidad que tienen las líneas fraudulentas.

Figura 14. Características de Líneas Telefónicas Fraudulentas
Fuente: Elaboración propia según investigación

Como complemento para el modelamiento del perfil de línea telefónica utilizada para tráfico gris, también se consulta a los expertos sobre la caracterización de un perfil de una línea telefónica legal. Las características de usuario legal que resaltan en la entrevista son ilustradas en la Figura 15, y se describen a continuación:

- Comportamiento congruente con el promedio de uso de otros usuarios de la red: el comportamiento general de los usuarios legales se mantiene cerca de los márgenes promedio para todos los servicios: llamadas de voz, mensajes de

texto, etc., por lo que los casos considerados legales son aquellos que se aproximan aceptablemente al comportamiento global de los usuarios de la red.

- Movilidad constante: El usuario legal no se mantiene fijo en una única ubicación geográfica, por lo contrario, tiene hábitos de vida que lo obligan a desplazarse por diferentes áreas, lo que se traduce a el registro de múltiples zonas de cobertura en la generación de sus registros de llamadas.
- Variedad de transacciones: El usuario regular varía sus actividades entre los diferentes productos que ofrece el operador de telefonía móvil, por ejemplo, llamadas de voz, mensajes de texto (SMS), navegación en red de datos Móvil o Internet, llamadas a servicios USSD, llamadas a códigos cortos de servicios como consulta de saldo, pagos vía operador móvil, etc.
- Relación proporcionada de llamadas entrantes versus salientes: El usuario legal tiene una relación aceptable de llamadas entrantes versus llamadas salientes, por ejemplo, un 20% de las llamadas totales pueden ser llamadas entrantes.
- Llamadas de voz y mensajes de texto recurrentes con los mismos destinatarios: Aun cuando el número de llamadas salientes puede ser muy alta, el usuario legal no llama a una cantidad muy grande de números telefónicos distintos, por lo que se ven llamadas y mensajes de texto intercambiados regularmente con un numero reducido de destinatarios.

Figura 15. Características de Líneas Telefónicas Legales
Fuente: Elaboración propia según investigación

Para efectos de la presente investigación, es importante conocer las características descritas anteriormente: características de líneas telefónicas fraudulentas y de líneas telefónicas legales. Tomando como base el aporte de los expertos mediante la entrevista, puede caracterizarse el perfil de usuario que se busca, considerando además que el usuario fraudulento siempre se encuentra en búsqueda de maneras de evadir la detección de su comportamiento ilegal. Los expertos indican que los perfiles de usuario que realizan tráfico gris tienden a ser muy dinámicos, cambian su comportamiento constantemente para evitar la detección. Las líneas telefónicas de usuarios fraudulentos buscan la manera de ser detectados de forma tardía, llegando incluso a simular comportamiento humano para evitar ser detectados.

Las principales dificultades en el trabajo de detección de fraude, de acuerdo al criterio de los expertos, son de diversa naturaleza. Por un lado, la disponibilidad de tiempo y esfuerzo requerido para la investigación resulta bastante demandante debido a la limitación de recursos en las oficinas de aseguramiento de ingresos. La investigación de tendencias actuales de detección y noticias sobre hallazgos en dicha área requieren dedicación y tiempo, para mantener al día los conocimientos y permanecer en constante aprendizaje de nuevos comportamientos fraudulentos. Los individuos que realizan el fraude tienden a encontrar nuevas formas ingeniosas para despistar a los algoritmos de detección, por ejemplo, reciclando las tarjetas SIM periódicamente, o cambiándose de ubicación en cortos periodos de tiempo. El análisis de los datos se debe realizar luego de procesada la información requerida, por ejemplo, los registros de llamadas de los usuarios de tráfico gris, y éste análisis resulta tedioso.

El conocimiento de las técnicas de detección utilizadas en otros países resulta muy importante para la constante mejora de las técnicas empleadas, y los entrevistados indican que la comunicación entre operadoras hermanas es constante, y que los operadores de telefonía móvil monitorean frecuentemente las noticias del foro global de fraude de la GSMA, para mantenerse informados. Entre las técnicas de detección utilizadas en países diferentes a Honduras, existen los proveedores de pruebas de llamadas internacionales, y el análisis de CDRs. Los entrevistados indican además que

están técnicas pueden ser muy efectivas, pero no para la detección inmediata de casos de fraude, dado que requieren procesamiento de la información recolectada en cada caso.

La adaptación de técnicas de detección de tráfico gris utilizadas en otros países puede o no ser adaptable al caso específico de la telefonía móvil en Honduras, por un lado, un par de expertos entrevistados indican que las técnicas empleadas en Honduras han sido mucho más complejas que las técnicas compartidas por otros países, y que otros métodos pueden resultar menos efectivos que el utilizado actualmente, mediante el análisis de CDRs. Sin embargo, otro experto entrevistado indica que si podría ser conveniente la utilización de técnicas alternas, dada la complejidad y tamaño mayor de las operaciones de países con mayor cobertura geográfica, siempre adecuando la herramienta a los casos específicos de Honduras.

Respecto a la disposición de las empresas de telefonía móvil a invertir en nuevas soluciones para la detección de tráfico gris, los expertos entrevistados coinciden en que estos sí están dispuestos a invertir, debido a que tienen consciencia del riesgo al que se exponen, debido a las pérdidas en las que incurren por casos de fraude, por tanto, en la medida en que se impacten sus ganancias, invertirán para mitigar dicha pérdida. Los expertos indican que el proceso de aseguramiento de ingresos es primordial en el área de las telecomunicaciones para blindar y evitar fugas innecesarias de ingresos, siempre considerando que es fundamental tener una relación de costo – beneficio. Mientras el fraude se mantenga en un margen aceptable, las empresas se limitan a continuar invirtiendo en nuevas tecnologías de detección de tráfico gris y casos de fraude.

Para el presente estudio es muy importante conocer la confianza que los expertos tienen en métodos estadísticos avanzados, y para conocer su opinión, se les entrevistó respecto a la aplicabilidad de métodos de reconocimiento de patrones, los cuales son fundamentales para los métodos de aprendizaje asistido. Los expertos entrevistados coinciden sin excepción en que se puede mejorar la efectividad de detección de métodos tradicionales, aplicando técnicas basadas en reconocimiento de patrones,

dado que estas técnicas han probado ser altamente efectivas, y recalcan especialmente que cuando existe una detección temprana es bastante favorable, ya que la detección oportuna es sumamente relevante al medir efectividad de una herramienta de detección. Los expertos indican que, con la tecnología actual, estas técnicas de reconocimiento de patrones pueden aplicar procedimientos computarizados complejos, mejorando así la efectividad de detección.

Concluyendo en base a las respuestas brindadas en la entrevista a expertos, se logran resaltan factores fundamentales a considerar para la caracterización de un perfil de tráfico gris, según lo indican los expertos en la materia. Debido a la delimitación del alcance de este estudio, se estudiarán 5 variables importantes de acuerdo al criterio de expertos, y estas se resumen en la Figura 16, en la que se correlaciona cada variable con la característica indicada por los expertos entrevistados, ya sea que parta de un perfil legal o de fraude.

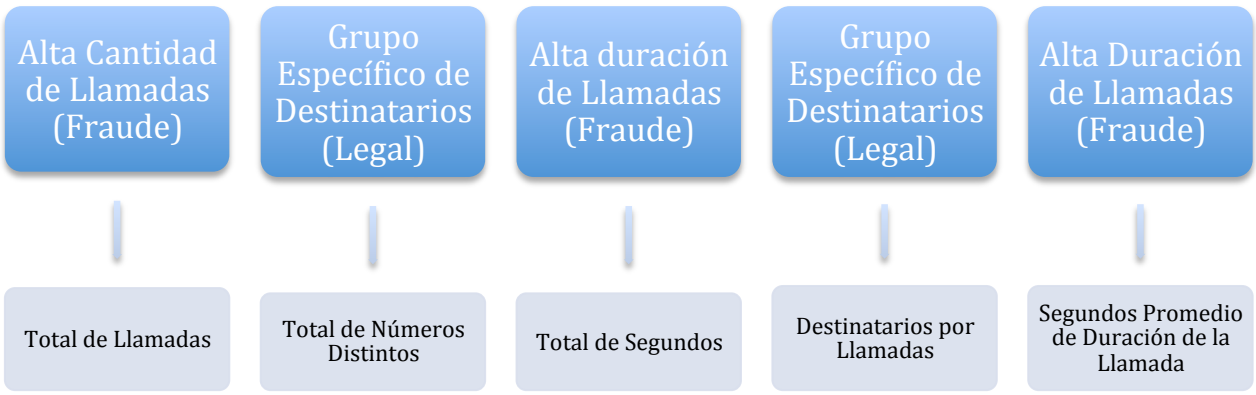

Figura 16. Opinión de Expertos Entrevistados versus Variables de Estudio
Fuente: Elaboración propia según investigación

4.2 ANÁLISIS CUANTITATIVO

Los criterios que fueron tomados en cuenta para llevar a cabo el análisis cuantitativo se basan en los resultados obtenidos a partir del análisis cualitativo, el cual contempla la opinión de expertos en el área de telefonía móvil con experiencia comprobada en el tema de estudio. La información obtenida de los expertos a través de entrevistas

contribuyó de manera significativa a la selección de factores determinantes en la detección de fraude y a la caracterización de los perfiles de uso.

La principal fuente de datos para el análisis cuantitativo fue extraída a partir de la base de datos existente del operador de telefonía móvil, la cual almacena datos extraídos de los archivos CDR. Cuando un suscriptor realiza una llamada por medio del operador de telefonía móvil, se genera un registro con información completa te la llamada, incluyendo detalles como el número del suscriptor, el número de destino, la fecha y hora de la llamada, la duración de la llamada, el identificador de las celdas de telefonía celular y otra gran cantidad de detalles. Estos datos registran el comportamiento de un suscriptor a lo largo del tiempo, por tanto, representan una fuente muy importante para este tipo de estudio.

El conjunto de datos utilizado para el presente análisis está formado por los datos de 1,473,510 llamadas realizadas por 6,031 suscriptores de telefonía móvil a lo largo de 2 meses calendario. De este total de suscriptores, 2,210 son casos reconocidos de fraude y 3,821 son suscriptores legales.

Los casos fueron almacenados en una base de datos MySQL y posteriormente fueron procesados para calcular las variables necesarias para el estudio. Estas variables son el promedio de llamadas que un suscriptor realiza por día, el promedio de distintos números destino a los que un suscriptor llama por día, el promedio de segundos que un suscriptor llama por día, el promedio de duración de las llamadas de un suscriptor y la razón de la cantidad de números distintos a los que llama un suscriptor sobre el total de las llamadas que realiza.

4.2.1 CONSTRUCCIÓN DEL MODELO

El análisis de datos está basado en la utilización de un modelo de red neuronal artificial para la clasificación automática de casos de fraude y casos legales. De forma más específica, se utilizó el modelo de red perceptrón multicapa, el cual consta de una capa

de neuronas de entrada, una o varias capas de neuronas ocultas y una capa de neuronas de salida.

Para construir el perceptrón multicapa se utilizó el módulo de redes neuronales de la herramienta IBM SPSS Statistics versión 22. Los 6,031 casos de estudio fueron divididos aleatoriamente en tres particiones de la siguiente forma:

- Partición de entrenamiento: Es el conjunto de casos utilizados para entrenar la red neuronal. Para el presente estudio se utilizó el 50% de los casos para el entrenamiento de la red.

- Partición de prueba: Es un conjunto independiente de casos que se reservan para realizar un seguimiento de los errores durante el entrenamiento, con la finalidad de evitar un exceso de entrenamiento (obtener una red muy estricta). Para el presente estudio se utilizó el 20% de los casos como muestra de prueba.

- Partición de reserva: Es un conjunto independiente de casos que se utilizan para evaluar la red neuronal final. Esta evaluación ofrece resultados más confiables ya que estos casos no son utilizados durante el entrenamiento de la red. Para el presente estudio se utilizó el 30% restante de los casos para la evaluación final.

Utilizando como referencia el estudio realizado por Elmi et al. (2013), se utilizó una arquitectura de red con dos capas ocultas y la función sigmoide como función de activación.

Para efectos de simplificación, en la elaboración del conjunto de datos se consideraron los siguientes nombres cortos para las variables de estudio:

Tabla 3. Nombres de Variables Utilizadas en el Conjunto de Datos

Nombre de Variable	Nombre Corto
Segundos Promedio de Duración de la Llamada	duracion_llamada
Destinatarios por Llamadas	numeros_llamadas
Total de Segundos	segundos_dia
Total de Llamadas	llamadas_dia
Total de Números Distintos	numeros_dia
Detección de Tráfico Gris	caso_bypass

Fuente: Elaboración propia según investigación

4.2.2 RESULTADOS

En esta sección se presentan los resultados obtenidos durante la fase de análisis. Primero se estudia el comportamiento de cada una de las variables independientes y su nivel de influencia en la predicción de valores correctos para la variable dependiente, con la finalidad de determinar la importancia de cada variable para el modelo final. Posteriormente, se analiza el modelo con todas las variables independientes en conjunto y la capacidad del modelo final de clasificar correctamente los casos de fraude y los casos legales.

4.2.2.1 SEGUNDOS PROMEDIO DE DURACIÓN DE LA LLAMADA

La primera variable independiente considerada para efectos de análisis es la variable segundos promedio de duración de la llamada, la cual representa la duración promedio en segundos de las llamadas de un suscriptor. Para determinar el impacto de esta variable sobre la variable dependiente se creó un modelo de red utilizando solamente estas dos variables. El resumen del modelo creado es el siguiente:

Tabla 4. Resumen Modelo de Segundos Promedio de Duración de la Llamada

Entrenamiento	Error de suma de cuadrados	521.570
	Porcentaje de pronósticos incorrectos	25.0%
	Regla de parada utilizada	1 pasos consecutivos sin disminución del error
	Tiempo de preparación	0:00:00.25
Pruebas	Error de suma de cuadrados	211.074
	Porcentaje de pronósticos incorrectos	23.8%
Reserva	Porcentaje de pronósticos incorrectos	26.8%

Fuente: Elaboración propia según investigación

Como se muestra en la Tabla 4, el entrenamiento de la red se detuvo debido a que se alcanzó un punto en el cuál ya no había disminución del error. Utilizando este modelo para evaluar los casos de reserva, se obtuvo un 26.8% de pronósticos incorrectos, esto significa que este modelo falla aproximadamente en 1 de cada 4 casos.

Tabla 5. Clasificación Modelo de Segundos Promedio de Duración de la Llamada

		Pronosticado			
Ejemplo	Observado	Legal	Bypass	Datos Reales	Porcentaje correcto
Entrenamiento	Legal	1519	371	1890	80.4%
	Bypass	378	730	1108	65.9%
	% Global	63.3%	36.7%	100%	75.0%
Pruebas	Legal	606	137	743	81.6%
	Bypass	148	308	456	67.5%
	% Global	62.9%	37.1%	100%	76.2%
Reserva	Legal	926	262	1188	77.9%
	Bypass	229	417	646	64.6%
	% Global	63.0%	37.0%	100%	73.2%

Fuente: Elaboración propia según investigación

La Tabla 5 muestra los resultados de clasificación utilizando los casos de entrenamiento, prueba y reserva. Los resultados de mayor interés son los resultados de

la muestra de reserva ya que son datos que no fueron utilizados durante el entrenamiento. Se puede observar que este modelo clasificó incorrectamente 262 falsos positivos y 229 falsos negativos, lo cual representa un 73.2% de efectividad de detección.

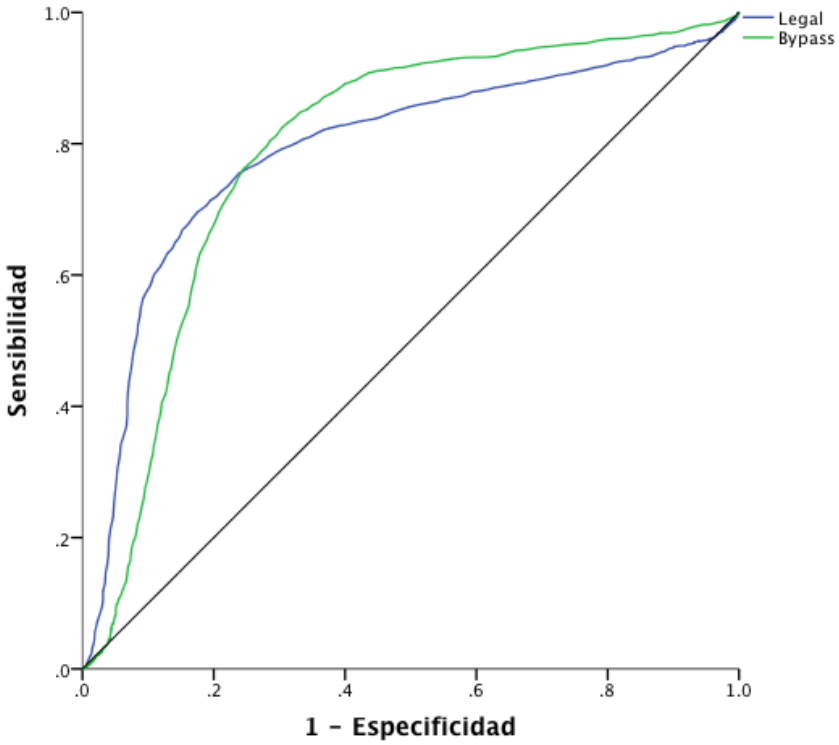

Figura 17. Curva COR Modelo de Segundos Promedio de Duración de la Llamada
Fuente: Elaboración propia según investigación

La curva COR (Característica Operativa del Receptor) se utiliza para determinar qué tan óptimo es un modelo de clasificación binario. Entre más se alejan las curvas de la diagonal hacia la esquina superior izquierda más óptimo es el modelo. Los valores de área bajo la curva pueden estar en el rango de 0.5 a 1.0. En este caso específico, se obtuvo un área bajo la curva de 0.79, lo que categoriza al modelo como “bueno”.

4.2.2.2 DESTINATARIOS POR LLAMADAS

En esta sección se analiza el comportamiento de la variable destinatarios por llamadas, la cual es cantidad de números destino distintos sobre el total de llamadas. Para determinar el impacto de esta variable sobre la variable dependiente se creó un modelo de red utilizando solamente estas dos variables. El resumen del modelo creado es el siguiente:

Tabla 6. Resumen Modelo de Destinatarios por Llamadas

Entrenamiento	Error de suma de cuadrados	61.551
	Porcentaje de pronósticos incorrectos	2.4%
	Regla de parada utilizada	1 pasos consecutivos sin disminución del error
	Tiempo de preparación	0:00:00.10
Pruebas	Error de suma de cuadrados	24.479
	Porcentaje de pronósticos incorrectos	2.3%
Reserva	Porcentaje de pronósticos incorrectos	2.1%

Fuente: Elaboración propia según investigación

Como se muestra en la Tabla 6, el entrenamiento de la red se detuvo debido a que se alcanzó un punto en el cuál ya no había disminución del error. Utilizando este modelo para evaluar los casos de reserva, se obtuvo solamente un 2.1% de pronósticos incorrectos, lo cual indica que esta variable puede tener un buen impacto en el modelo de detección.

Tabla 7. Clasificación Modelo de Destinatarios por Llamadas

		Pronosticado			
Ejemplo	Observado	Legal	Bypass	Datos Reales	Porcentaje correcto
Entrenamiento	Legal	1828	62	1890	96.7%
	Bypass	9	1099	1108	99.2%
	% Global	61.3%	38.7%	100%	97.6%
Pruebas	Legal	721	22	743	97.0%
	Bypass	5	451	456	98.9%
	% Global	60.6%	39.4%	100%	97.7%
Reserva	Legal	1153	35	1188	97.1%
	Bypass	3	643	646	99.5%
	% Global	63.0%	37.0%	100%	97.9%

Fuente: Elaboración propia según investigación

La Tabla 7 muestra los resultados de la clasificación utilizando el modelo actual. Se puede observar que al utilizar la muestra de reserva el modelo clasificó incorrectamente 35 falsos positivos y 3 falsos negativos, lo cual representa un 97.9% de efectividad de detección.

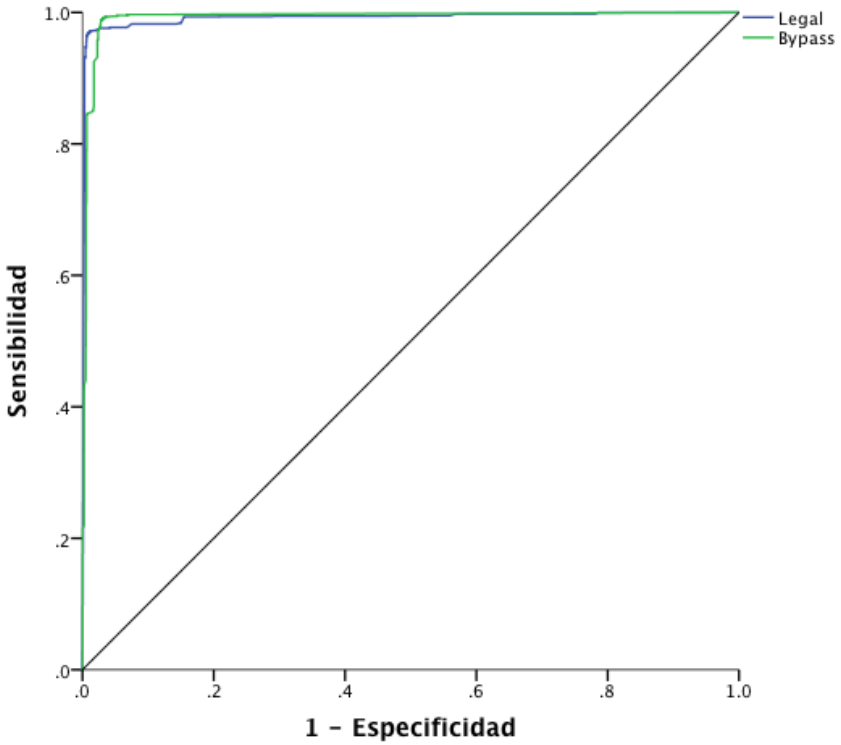

Figura 18. Curva COR del Modelo de Destinatarios por Llamadas
Fuente: Elaboración propia según investigación

El valor del área bajo la curva COR para el presente modelo de clasificación es de 0.992, lo que categoriza al modelo como “excelente”. Se puede notar este comportamiento en la forma de la curva de la Figura 18, la cual se acerca mucho a la esquina superior izquierda.

4.2.2.3 TOTAL DE SEGUNDOS

En esta sección se analiza el comportamiento de la variable total de segundos, la cual es el promedio de segundos que un suscriptor consume en un día. Para determinar el impacto de esta variable sobre la variable dependiente se creó un modelo de red utilizando solamente estas dos variables. El resumen del modelo creado es el siguiente:

Tabla 8. Resumen Modelo Total de Segundos

Entrenamiento	Error de suma de cuadrados	58.806
	Porcentaje de pronósticos incorrectos	2.5%
	Regla de parada utilizada	1 pasos consecutivos sin disminución del error
	Tiempo de preparación	0:00:00.04
Pruebas	Error de suma de cuadrados	28.938
	Porcentaje de pronósticos incorrectos	3.1%
Reserva	Porcentaje de pronósticos incorrectos	2.5%

Fuente: Elaboración propia según investigación

Como se muestra en la Tabla 8, el entrenamiento de la red se detuvo debido a que se alcanzó un punto en el cuál ya no había disminución del error. Utilizando este modelo para evaluar los casos de reserva, se obtuvo solamente un 2.5% de pronósticos incorrectos, lo cual indica que esta variable puede tener un buen impacto en el modelo de detección.

Tabla 9. Clasificación Modelo Total de Segundos

		Pronosticado			
Ejemplo	Observado	Legal	Bypass	Datos Reales	Porcentaje correcto
Entrenamiento	Legal	1860	30	1890	98.4%
	Bypass	46	1062	1108	95.8%
	% Global	63.6%	36.4%	100%	97.5%
Pruebas	Legal	733	10	743	98.7%
	Bypass	27	429	456	94.1%
	% Global	63.4%	36.6%	100%	96.9%
Reserva	Legal	1171	17	1188	98.6%
	Bypass	29	617	646	95.5%
	% Global	65.4%	34.6%	100%	97.5%

Fuente: Elaboración propia según investigación

La Tabla 9 muestra los resultados de la clasificación utilizando el modelo actual. Se puede observar que al utilizar la muestra de reserva el modelo clasificó incorrectamente 17 falsos positivos y 29 falsos negativos, lo cual representa un 97.5% de efectividad de detección.

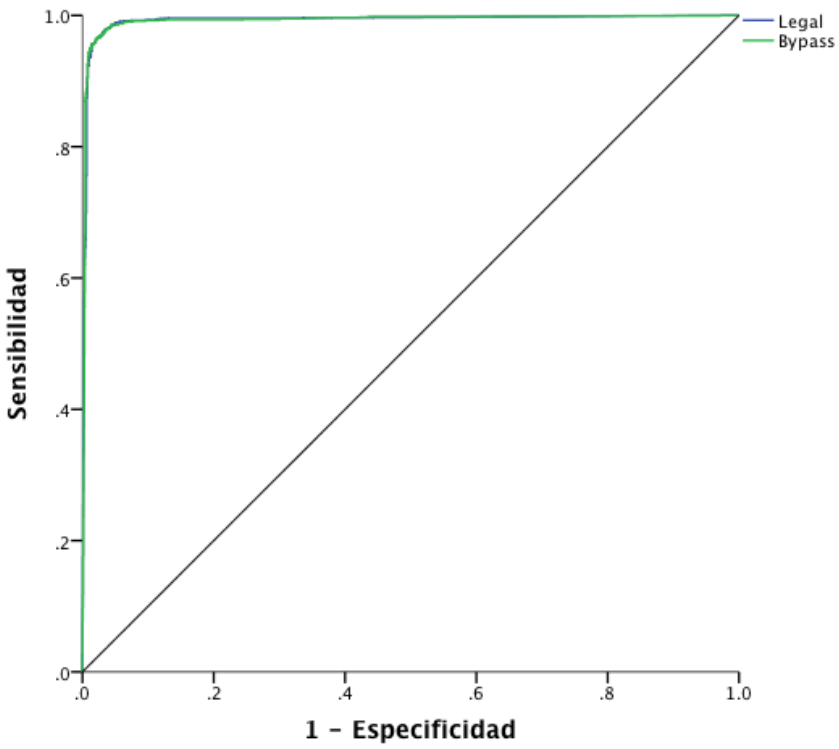

Figura 19. Curva COR del Modelo de Total de Segundos
Fuente: Elaboración propia según investigación

El valor del área bajo la curva COR para el presente modelo de clasificación es de 0.993, lo que categoriza al modelo como “excelente”. Se puede notar este comportamiento en la forma de la curva de la Figura 19, la cual se acerca mucho a la esquina superior izquierda.

4.2.2.4 TOTAL DE LLAMADAS

En esta sección se analiza el comportamiento de la variable total de llamadas, la cual es el promedio de llamadas que un suscriptor hace en un día. Para determinar el impacto de esta variable sobre la variable dependiente se creó un modelo de red utilizando solamente estas dos variables. El resumen del modelo creado es el siguiente:

Tabla 10. Resumen Modelo Total de Llamadas

Entrenamiento	Error de suma de cuadrados	9.360
	Porcentaje de pronósticos incorrectos	0.3%
	Regla de parada utilizada	1 pasos consecutivos sin disminución del error
	Tiempo de preparación	0:00:00.09
Pruebas	Error de suma de cuadrados	4.550
	Porcentaje de pronósticos incorrectos	0.3%
Reserva	Porcentaje de pronósticos incorrectos	0.1%

Fuente: Elaboración propia según investigación

Como se muestra en la Tabla 10, el entrenamiento de la red se detuvo debido a que se alcanzó un punto en el cuál ya no había disminución del error. Utilizando este modelo para evaluar los casos de reserva, se obtuvo solamente un 0.1% de pronósticos incorrectos, lo cual indica que esta variable puede tener un buen impacto en el modelo de detección.

Tabla 11. Clasificación Modelo Total de Llamadas

		Pronosticado			
Ejemplo	Observado	Legal	Bypass	Datos Reales	Porcentaje correcto
Entrenamiento	Legal	1882	8	1890	99.6%
	Bypass	0	1108	1108	100.0%
	% Global	62.8%	37.2%	100%	99.7%
Pruebas	Legal	740	3	743	99.6%
	Bypass	0	456	456	100.0%
	% Global	61.7%	38.3%	100%	99.7%
Reserva	Legal	1186	2	1188	99.8%
	Bypass	0	646	646	100.0%
	% Global	64.7%	35.3%	100%	99.9%

Fuente: Elaboración propia según investigación

La Tabla 11 muestra los resultados de la clasificación utilizando el modelo actual. Se puede observar que al utilizar la muestra de reserva el modelo clasificó incorrectamente sólo 2 falsos positivos, lo cual representa un 99.9% de efectividad de detección.

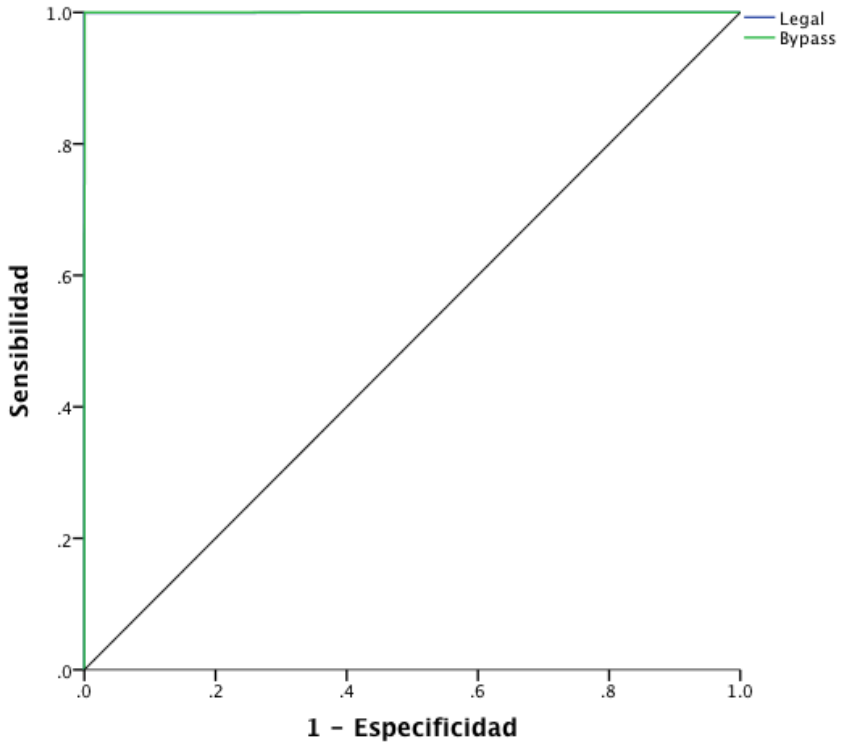

Figura 20. Curva COR del Modelo Total de Llamadas
Fuente: Elaboración propia según investigación

El valor del área bajo la curva COR para el presente modelo de clasificación es de aproximadamente 1.0, lo que categoriza al modelo como “excelente”, ilustrado en la Figura 20. Este resultado indica que esta es una de las variables que puede tener mayor impacto positivo en el modelo final.

4.2.2.5 TOTAL DE NÚMEROS DISTINTOS

En esta sección se analiza el comportamiento de la variable total de números distintos, la cual es el promedio de números distintos a los que un suscriptor llama en un día.

Para determinar el impacto de esta variable sobre la variable dependiente se creó un modelo de red utilizando solamente estas dos variables. El resumen del modelo creado es el siguiente:

Tabla 12. Resumen Modelo Total de Números Distintos

Entrenamiento	Error de suma de cuadrados	1.683
	Porcentaje de pronósticos incorrectos	0.1%
	Regla de parada utilizada	1 pasos consecutivos sin disminución del error
	Tiempo de preparación	0:00:00.11
Pruebas	Error de suma de cuadrados	1.677
	Porcentaje de pronósticos incorrectos	0.2%
Reserva	Porcentaje de pronósticos incorrectos	0.1%

Fuente: Elaboración propia según investigación

Como se muestra en la Tabla 12, el entrenamiento de la red se detuvo debido a que se alcanzó un punto en el cuál ya no había disminución del error. Utilizando este modelo para evaluar los casos de reserva, se obtuvo solamente un 0.1% de pronósticos incorrectos, lo cual indica que esta variable puede tener un buen impacto en el modelo de detección.

Tabla 13. Clasificación Modelo Total de Números Distintos

Ejemplo	Observado	Pronosticado			
		Legal	Bypass	Datos Reales	Porcentaje correcto
Entrenamiento	Legal	1890	0	1890	100.0%
	Bypass	2	1106	1108	99.8%
	% Global	63.1%	36.9%	100%	99.9%
Pruebas	Legal	743	0	743	100.0%
	Bypass	2	454	456	99.6%
	% Global	62.1%	37.9%	100%	99.8%
Reserva	Legal	1188	0	1188	100.0%
	Bypass	1	645	646	99.8%
	% Global	64.8%	35.2%	100%	99.9%

Fuente: Elaboración propia según investigación

La Tabla 13 muestra los resultados de la clasificación utilizando el modelo actual. Se puede observar que al utilizar la muestra de reserva el modelo clasificó incorrectamente sólo 1 falso negativo, lo cual representa un 99.9% de efectividad de detección.

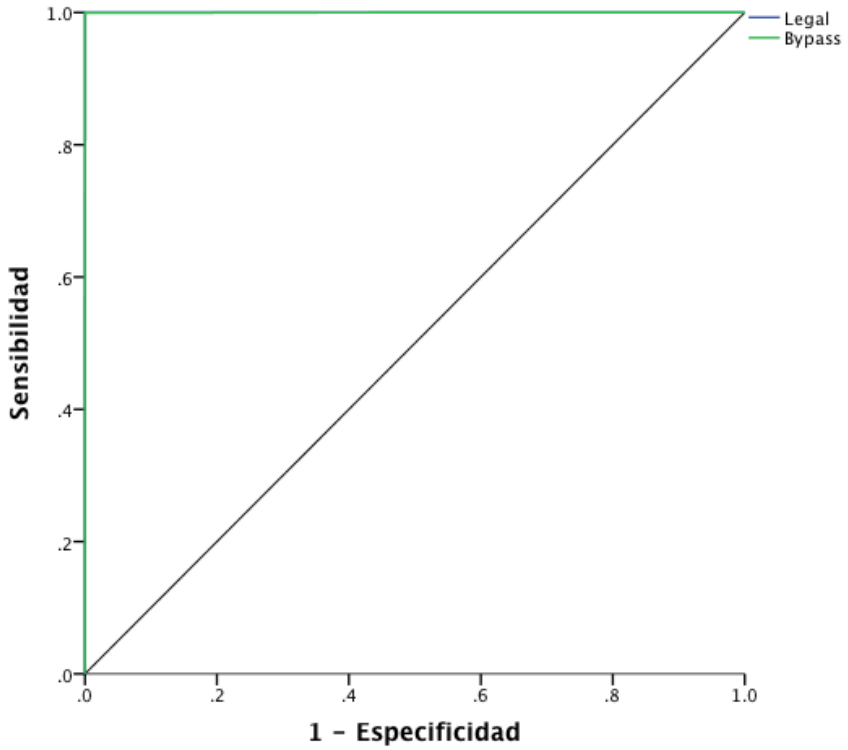

Figura 21. Curva COR del Modelo Total de Números Distintos
Fuente: Elaboración propia según investigación

El valor del área bajo la curva COR para el presente modelo de clasificación es de aproximadamente 1.0, lo que categoriza al modelo como “excelente”. Este resultado indica que esta es una de las variables que puede tener mayor impacto positivo en el modelo final.

4.2.2.6 MODELO FINAL

En esta sección se presenta el modelo de perceptrón multicapa construido utilizando todas las variables de estudio en conjunto. Este modelo permite una detección más realista ya que al utilizar varias variables se pueden detectar patrones de comportamiento más complejos. La arquitectura de red utilizada para el presente análisis se muestra en la Figura 22:

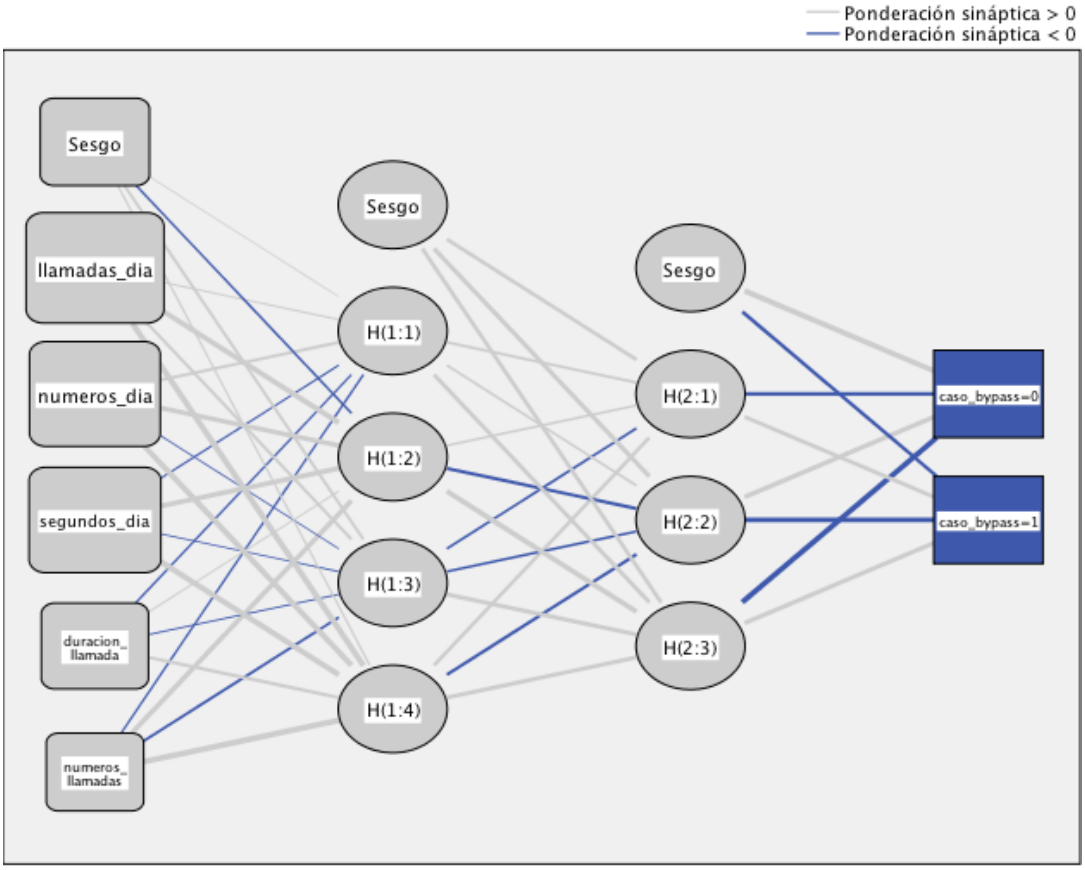

Figura 22. Arquitectura de Red del Modelo Final
Fuente: Elaboración propia según investigación

El perceptrón multicapa presenta cinco entradas (una para cada variable independiente) las cuales están conectadas a las cuatro neuronas de la primera capa oculta. Esta capa

está conectada a una segunda capa oculta que contiene tres neuronas, las cuales a su vez están conectadas a las dos salidas de clasificación. Debido a que el aprendizaje es supervisado, el modelo también presenta tres unidades de sesgo, las cuales realizan los ajustes necesarios durante el proceso de aprendizaje.

Tabla 14. Clasificación Modelo Final

		Pronosticado			
Ejemplo	Observado	Legal	Bypass	Datos Reales	Porcentaje correcto
Entrenamiento	Legal	1890	0	1890	100.0%
	Bypass	1	1107	1108	99.9%
	% Global	63.1%	36.9%	100%	100.0%
Pruebas	Legal	742	1	743	99.9%
	Bypass	1	455	456	99.8%
	% Global	62.0%	38.0%	100%	99.8%
Reserva	Legal	1187	1	1188	99.9%
	Bypass	1	645	646	99.8%
	% Global	64.8%	35.2%	100%	99.9%

Fuente: Elaboración propia según investigación

La Tabla 14 muestra los resultados obtenidos en la etapa de clasificación, después de haber entrenado la red. Se puede observar que al utilizar la muestra de reserva el modelo final clasificó incorrectamente sólo 1 falso positivo y 1 falso negativo, lo cual representa un 99.9% de efectividad de detección. Este resultado comprueba la validez de la hipótesis de investigación, ya que utilizando los factores seleccionados se logró crear un modelo de clasificación con una efectividad de detección mayor al 90%.

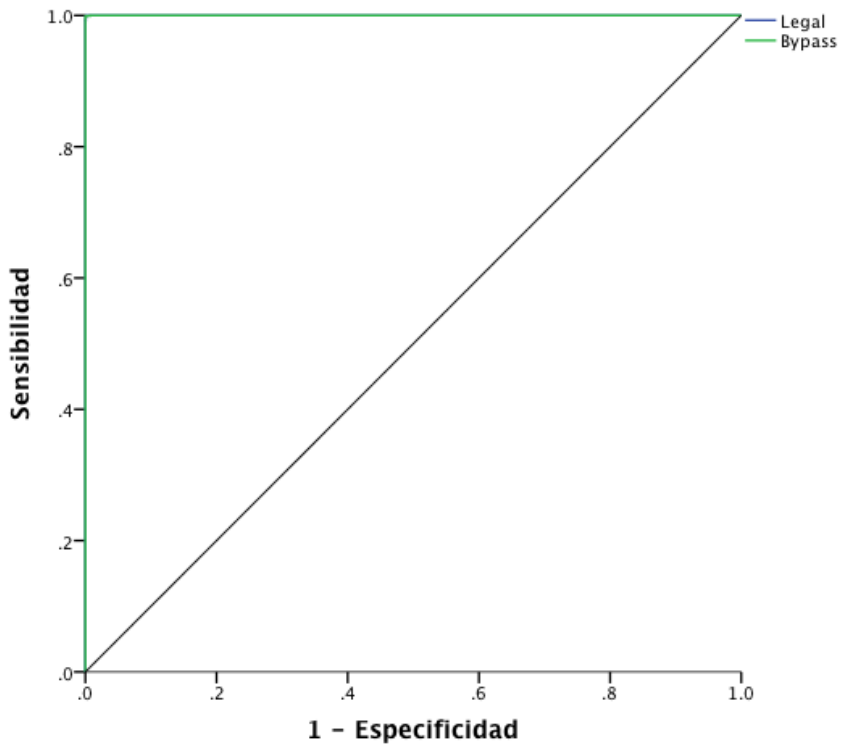

Figura 23. Curva COR del Modelo Final
Fuente: Elaboración propia según investigación

El valor del área bajo la curva COR para el modelo de clasificación final es de aproximadamente 1.0, lo que categoriza al modelo como “excelente”, ilustrado en la Figura 23.

Para finalizar con el análisis, se muestra en la Figura 24 el gráfico de importancia normalizada de las variables independientes. La importancia normalizada indica el nivel de participación que tiene cada una de las variables independientes en la forma en que la red clasifica los casos. Se puede observar que las tres variables independientes que ejercen mayor influencia en el modelo final son el total de llamadas, el total de segundos y el total de números distintos. Para futuros estudios, se puede considerar el reemplazo de las variables menos significativas por otras variables que se consideren de interés.

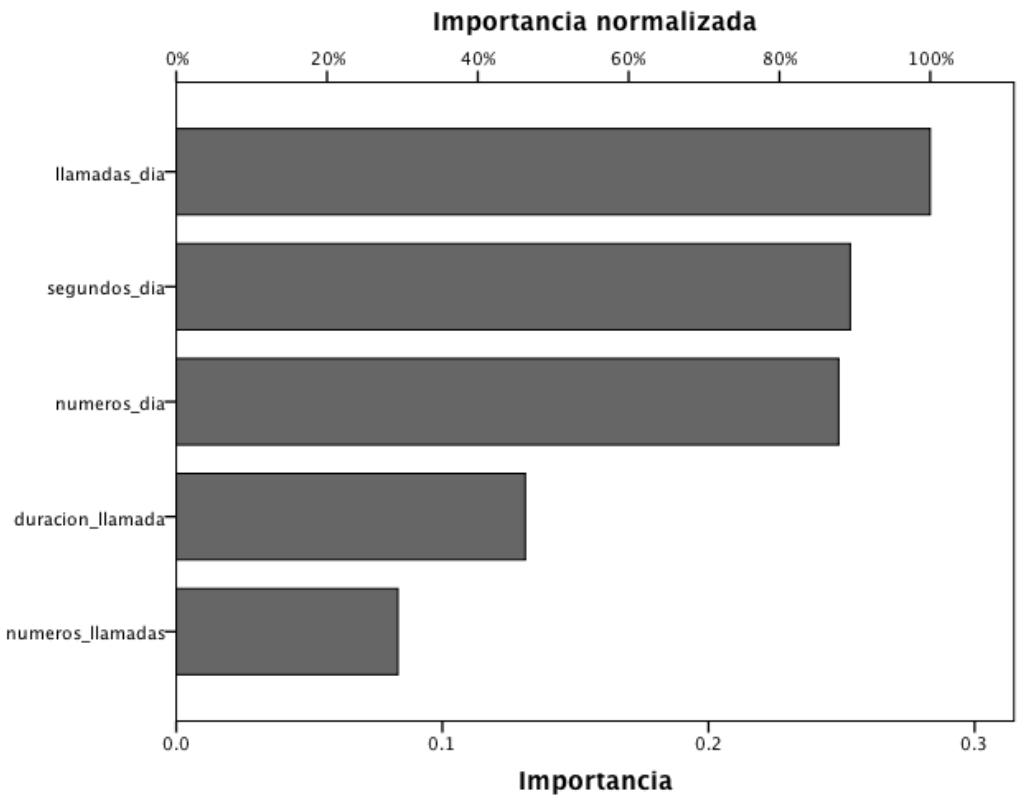

Figura 24. Importancia de Variables Independientes
Fuente: Elaboración propia según investigación

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- En base a la entrevista realizada a expertos, se concluye que las características que definen una línea telefónica legal son el cumplimiento del perfil promedio de la red, la movilidad constante del usuario, la variedad de transacciones, una relación proporcionada de llamadas entrantes versus salientes, y la ocurrencia recurrente de llamadas hacia un grupo específico de destinatarios.

- De acuerdo a la experiencia compartida por los expertos mediante la entrevista, se concluye que las características que definen una línea telefónica utilizada para tráfico gris son una alta cantidad de llamadas, una alta duración de llamadas, la relación desproporcionada de llamadas entrantes versus salientes, la ejecución de una actividad específica de forma recurrente, y la ubicación invariable del usuario.

- Basados en los resultados de la investigación y del análisis cualitativo y cuantitativo del estudio, se concluye que los mecanismos de reconocimiento de patrones resultan efectivos como apoyo a la detección de líneas telefónicas utilizadas para tráfico gris. En el caso específico de la presente investigación, el perceptrón multicapa demostró una alta efectividad en la clasificación de líneas fraudulentas y líneas legales en base al estudio de su comportamiento.

- Mediante el análisis de casos previamente identificados como fraudulentos y la opinión de expertos, se logró identificar algunos de los factores determinantes de una línea telefónica utilizada para tráfico gris, y se construyó un modelo que puede ser utilizado posteriormente para la detección de nuevos casos de fraude.

5.2 RECOMENDACIONES

- Para estudios futuros que pretendan modelar un perfil de usuario, se recomienda considerar factores adicionales de comportamiento, ya que en los archivos CDRs se registra una amplia variedad de campos que pueden ser utilizados para mayor precisión. La inclusión de más características puede llevar a una mayor efectividad, siempre analizando la relación de los factores a considerarse y su relevancia en el estudio.

- Para efectos prácticos, se recomienda aplicar al modelo conjuntos de datos más extensos, tanto en la cantidad de registros como en el rango de tiempo de la muestra. De esta forma, el modelo será capaz de aprender nuevos comportamientos de los suscriptores estudiados para mejorar la detección de cada perfil.

- Al utilizar este modelo en un ambiente de producción, se recomienda la revisión periódica del mismo para mantener los perfiles actualizados con respecto a nuevos comportamientos, a manera de mejorar continuamente el modelo para efectos de mantener la efectividad deseada.

- Aun cuando la automatización de la detección aporta mucho valor, es recomendable siempre referirse a la opinión de expertos que aporten mayor visibilidad a cada escenario, que propongan nuevas ideas y soluciones, para mejorar los métodos utilizados.

CAPÍTULO VI. APLICABILIDAD

GUÍA PARA LA CREACIÓN Y APLICACIÓN DE UN MODELO DE DETECCIÓN DE LÍNEAS TELEFÓNICAS UTILIZADAS PARA TRÁFICO GRIS EN UNA EMPRESA DE TELEFONÍA CELULAR

6.1 INTRODUCCIÓN

La presente guía tiene como objetivo servir de referencia para la creación y aplicación de un modelo basado en redes neuronales, el cual permita la clasificación automática de suscriptores de telefonía celular de acuerdo a su perfil de utilización del servicio de llamadas en una red GSM. Basado en las variables seleccionadas, el modelo será capaz de diferenciar entre líneas legales y líneas utilizadas para tráfico gris, con un alto nivel de efectividad.

Para poder poner en práctica esta guía, el lector debe tener acceso a una base de datos con los registros de las llamadas de los suscriptores que desea clasificar. También, para la fase de entrenamiento de la red, es necesario tener acceso a los registros de llamadas de suscriptores reconocidos como legales y de suscriptores que previamente hayan sido identificados como tráfico gris, ya que a partir de estos casos es que la red “aprenderá” las características de cada perfil.

Para la preparación de los datos es necesario tener acceso a un sistema de gestión de bases de datos relacionales (en este estudio se utilizó MySQL). Para la creación del modelo es indispensable contar con la herramienta IBM SPSS Statistics y su módulo de redes neuronales. Se asume que el lector tiene conocimientos sobre creación y manipulación de bases de datos relacionales.

El estudio que acompaña a esta guía ha demostrado la efectividad que puede lograr el modelo con los factores seleccionados, sin embargo, se recomienda al lector ampliar ésta lista de factores y comprobar el rendimiento del modelo con la finalidad de identificar nuevas características que sean de interés en la detección de tráfico gris.

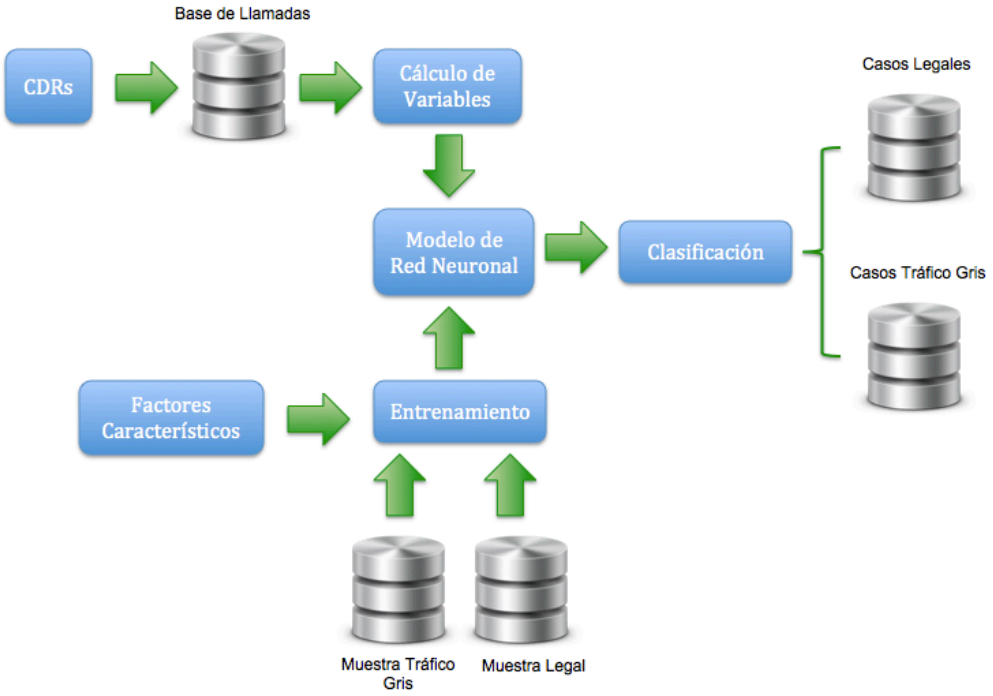

Figura 25. Diagrama de Modelo de Clasificación
Fuente: Elaboración propia según investigación

6.2 FASES PARA LA CREACIÓN DEL MODELO

A continuación se presenta una serie de fases ordenadas para la creación y aplicación del modelo de clasificación.

6.2.1 FASE 1: PREPARACIÓN DE DATOS

La primera fase consiste en la recolección de los datos para el análisis y la posterior transformación de los mismos a un formato apropiado para la creación del modelo. En esta fase se debe tener acceso a una base de datos con los registros de llamadas de los suscriptores que desea analizar. En esta guía, se asume que los casos comprobados de fraude están separados de los casos comprobados legales.

Los campos utilizados para el presente estudio son el número de teléfono, la fecha de la llamada, el número destino y la duración en segundos de la llamada. Si se desea agregar más factores al modelo, debe seleccionar los campos extra que sean necesarios. Para el estudio se recomienda crear dos tablas, una para cargar los registros de los casos de fraude previamente identificados y otra para cargar los registros de los casos legales.

Para la creación de las tablas se recomienda utilizar la siguiente sintaxis SQL:

```
/*Creación de la tabla muestra_legal para almacenar casos legales*/
CREATE TABLE muestra_legal (
  telefono varchar(8) COLLATE latin1_spanish_ci NOT NULL,
  fecha_llamada datetime NOT NULL,
  destino varchar(8) COLLATE latin1_spanish_ci NOT NULL,
  segundos decimal(10,0) NOT NULL);

/*Creación de la tabla muestra_bypass para almacenar casos de fraude*/
CREATE TABLE muestra_bypass (
  telefono varchar(8) COLLATE latin1_spanish_ci NOT NULL,
  fecha_llamada datetime NOT NULL,
  destino varchar(8) COLLATE latin1_spanish_ci NOT NULL,
  segundos decimal(10,0) NOT NULL);
```

Una vez creadas las tablas, es necesario cargar en ellas los registros de llamadas previamente identificados. Si los registros están en un archivo externo, por ejemplo un documento CSV, se recomienda utilizar la siguiente sintaxis SQL para cargarlos en las tablas:

```
/*Carga de registros de llamadas legales desde archivo CSV*/
LOAD DATA LOCAL INFILE 'registros_legales.csv'
INTO TABLE muestra_legal
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
(telefono, fecha_llamada, destino, segundos);

/*Carga de registros de fraude desde archivo CSV*/
LOAD DATA LOCAL INFILE 'registros_bypass.csv'
INTO TABLE muestra_bypass
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
(telefono, fecha_llamada, destino, segundos);
```

Hasta este punto, se tienen dos tablas con información a nivel de llamadas individuales, sin embargo, para el análisis es necesario resumir esta información a nivel de un registro por suscriptor. Esto significa que es necesario calcular los valores de las variables de estudio para cada suscriptor. Para la creación de los resúmenes se recomienda utilizar la siguiente sintaxis SQL:

```
/*Creación de la tabla resumen para casos legales*/
CREATE TABLE resumen_legal AS
SELECT DISTINCT teléfono,
COUNT(*) AS total_llamadas,
COUNT(DISTINCT(DATE(fecha_llamada))) AS dias_distintos,
SUM(segundos) AS total_segundos,
COUNT(DISTINCT(destino)) AS numeros_distintos
FROM muestra_legal
GROUP BY telefono;

/*Creación de la tabla resumen para casos de fraude*/
CREATE TABLE resumen_bypass AS
SELECT DISTINCT telefono,
COUNT(*) AS total_llamadas,
COUNT(DISTINCT(DATE(fecha_llamada))) AS dias_distintos,
SUM(segundos) AS total_segundos,
COUNT(DISTINCT(destino)) AS numeros_distintos
FROM muestra_bypass
GROUP BY telefono;
```

A partir de las tablas resumen, el cálculo de las variables para cada suscriptor es más directo. Para crear las tablas con las variables de estudio, se recomienda utilizar la siguiente sintaxis SQL:

```
CREATE TABLE casos_legales AS
SELECT telefono, CEIL(total_llamadas/dias_distintos) AS llamadas_dia,
CEIL(numeros_distintos/dias_distintos) AS numeros_dia,
CEIL(total_segundos/dias_distintos) AS segundos_dia,
CEIL(total_segundos/total_llamadas) AS duracion_llamada,
numeros_distintos/total_llamadas AS numeros_llamadas,
0 AS caso_bypass
FROM resumen_legal;

CREATE TABLE casos_bypass AS
SELECT telefono,
CEIL(total_llamadas/dias_distintos) AS llamadas_dia,
CEIL(numeros_distintos/dias_distintos) AS numeros_dia,
CEIL(total_segundos/dias_distintos) AS segundos_dia,
CEIL(total_segundos/total_llamadas) AS duracion_llamada,
numeros_distintos/total_llamadas AS numeros_llamadas,
1 AS caso_bypass
FROM resumen_bypass;
```

Es importante notar que, aparte de las variables calculadas, se introduce una nueva variable llamada caso_bypass la cual toma el valor de 0 para los suscriptores legales y 1 para los casos de fraude. Este campo es utilizado como variable dependiente en el modelo y sirve para identificar a los casos legales y los casos de fraude.

En este punto, los datos de cada suscriptor están debidamente resumidos y categorizados de manera uniforme en ambas tablas. Ahora, es necesario crear una sola tabla con todos los casos para su posterior exportación a SPSS. Para la creación de ésta tabla única se recomienda utilizar la siguiente sintaxis SQL:

```
CREATE TABLE casos_modelo AS
SELECT * FROM casos_legales
UNION
SELECT * FROM casos_bypass
ORDER BY RAND();
```

De esta forma se crea una tabla única con todos los casos ordenados de forma aleatoria para el entrenamiento del modelo en SPSS.

6.2.2 FASE 2: CREACIÓN DEL MODELO

El objetivo de esta fase es la creación de un modelo de red neuronal, específicamente perceptrón multicapa, a partir de los casos preparados en la fase anterior. Para la creación del modelo se debe cargar en SPSS la tabla casos_modelo, para esta tarea se debe seleccionar el menú Archivo – Abrir base de datos – Nueva consulta, y seleccionar la tabla desde la conexión con la base de datos.

Para la creación de un modelo perceptrón multicapa en SPSS se debe seleccionar el menú Analizar – Redes neuronales – Perceptrón multicapa. Se mostrará la siguiente ventana:

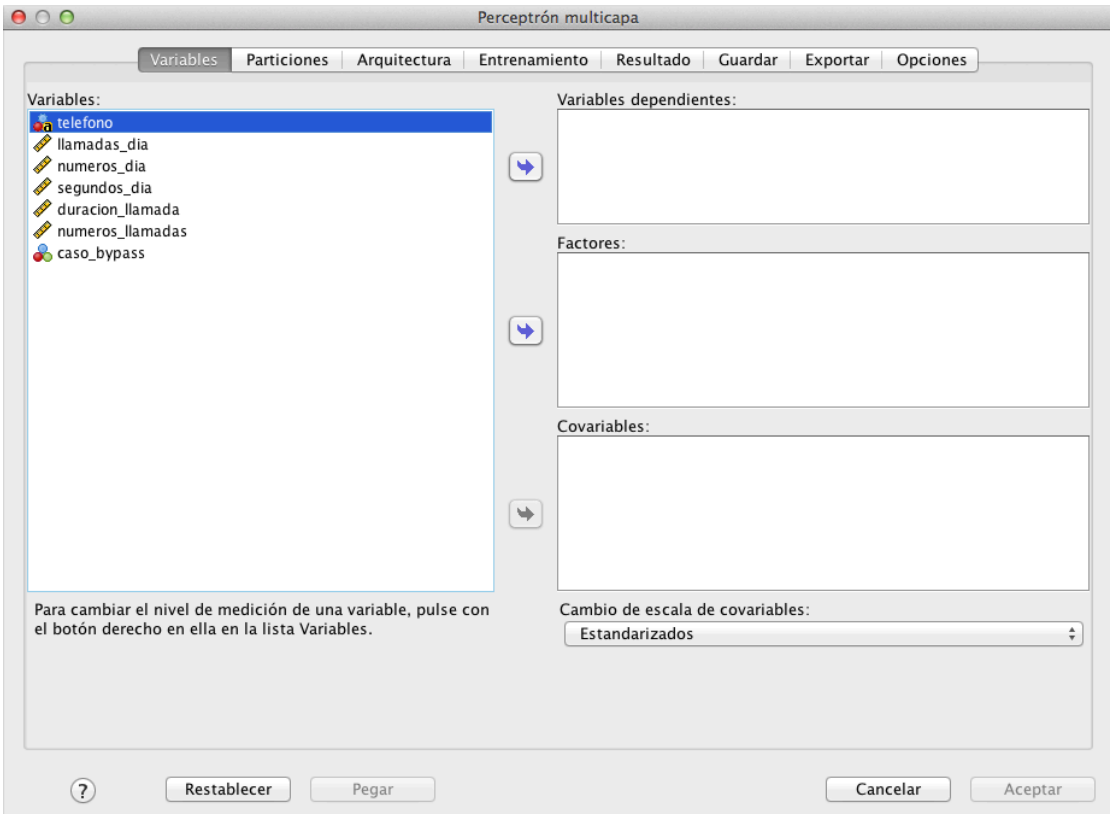

Figura 26. Pestaña de Selección de Variables
Fuente: Definición propia según investigación

En la parte izquierda aparecen todas las variables de estudio, se debe seleccionar la variable caso_bypass y asignarla como variable dependiente. Si en el estudio existieran variables categóricas éstas tendrían que ser asignadas como factores. Las variables de estudio restantes son variables de escala, por tanto deben ser asignadas como covariables:

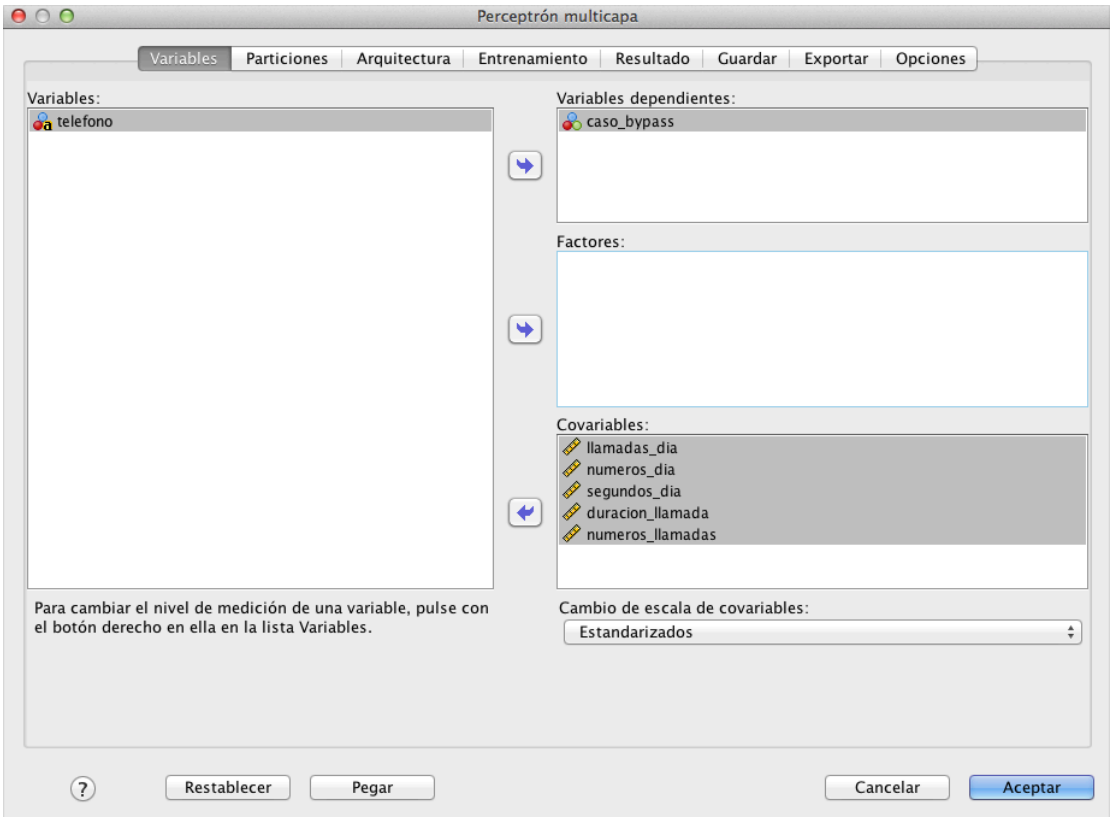

Figura 27. Asignación de Variables
Fuente: Definición propia según investigación

En la pestaña de partición se debe seleccionar qué porcentaje de los casos se asignará a cada partición para entrenamiento, prueba y reserva. Se debe asignar los porcentajes que se consideren necesarios, teniendo en cuenta que lo ideal es tener una partición de entrenamiento mayor que las de prueba y reserva:

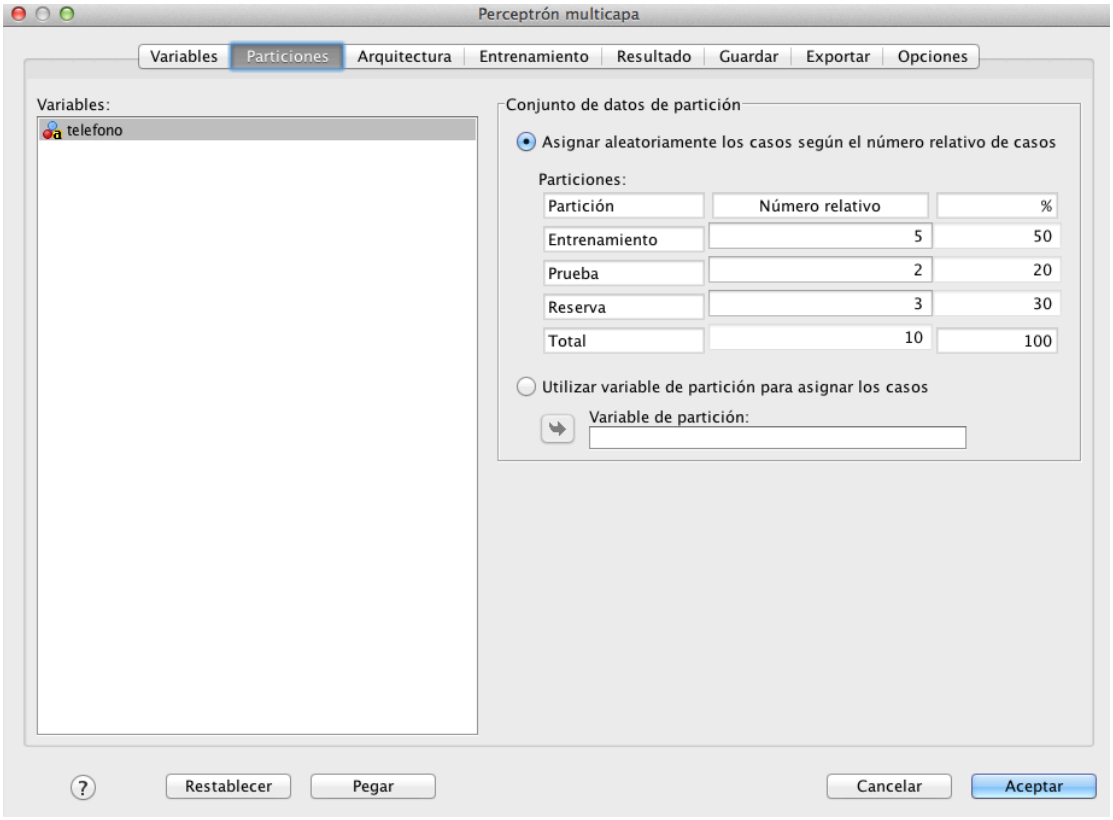

Figura 28. Pestaña de Particiones
Fuente: Definición propia según investigación

En la pestaña de arquitectura se debe seleccionar la arquitectura de la red. Según el estudio de referencia se han obtenido mejores resultados con una arquitectura de dos capas y utilizando la función sigmoide como función de activación. Sin embargo, se recomienda experimentar con nuevas arquitecturas para distintos conjuntos de datos y decidir qué arquitectura brinda los mejores resultados en cada caso:

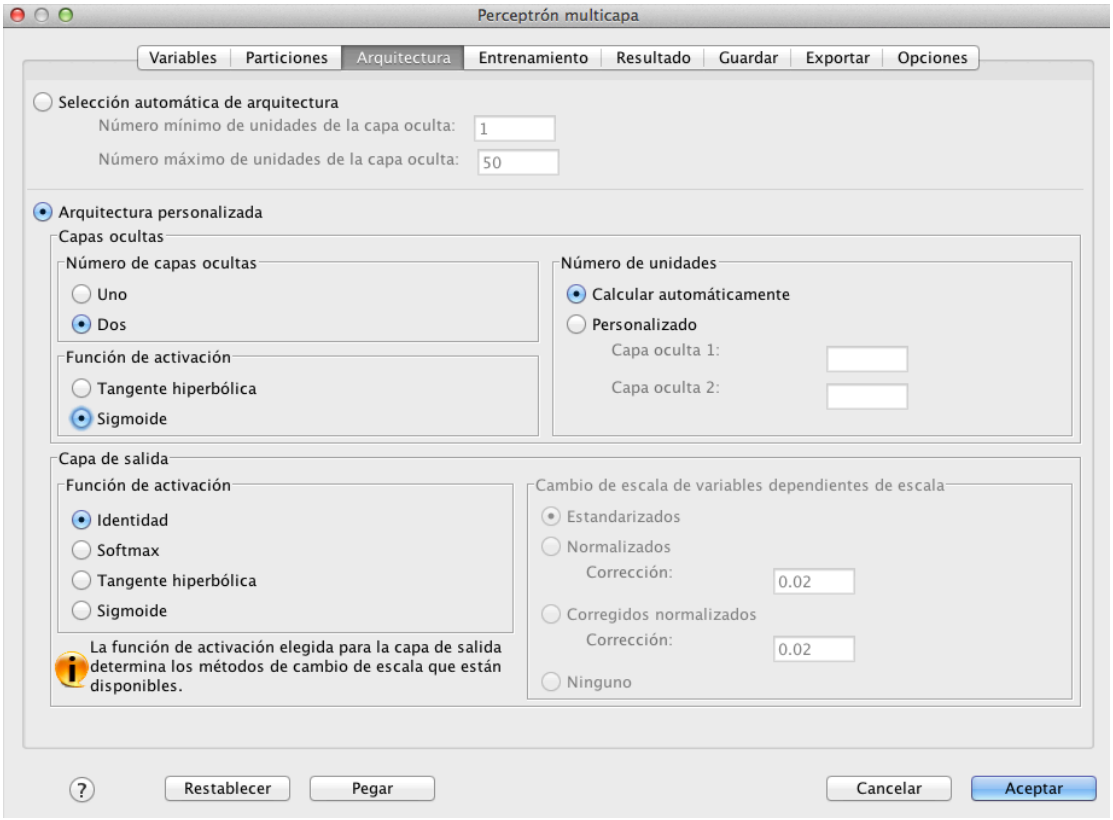

Figura 29. Pestaña de Arquitectura
Fuente: Definición propia según investigación

En la pestaña de entrenamiento se debe seleccionar el tipo de entrenamiento que la red utilizará para procesar los registros. El entrenamiento por lote actualiza las ponderaciones sinápticas de la red sólo al pasar todos los registros de entrenamiento. El entrenamiento en línea actualiza las ponderaciones sinápticas después de procesar cada registro en el entrenamiento. El entrenamiento por mini lote divide los registros en pequeños grupos y actualiza las ponderaciones sinápticas cada vez que procesa un grupo.

El estudio referencial obtuvo mejores resultados utilizando entrenamiento en línea, sin embargo, se recomienda experimentar con otros tipos de entrenamiento para distintos conjuntos de datos, con el objetivo de seleccionar el modo de entrenamiento que más se apegue al caso específico que se está estudiando:

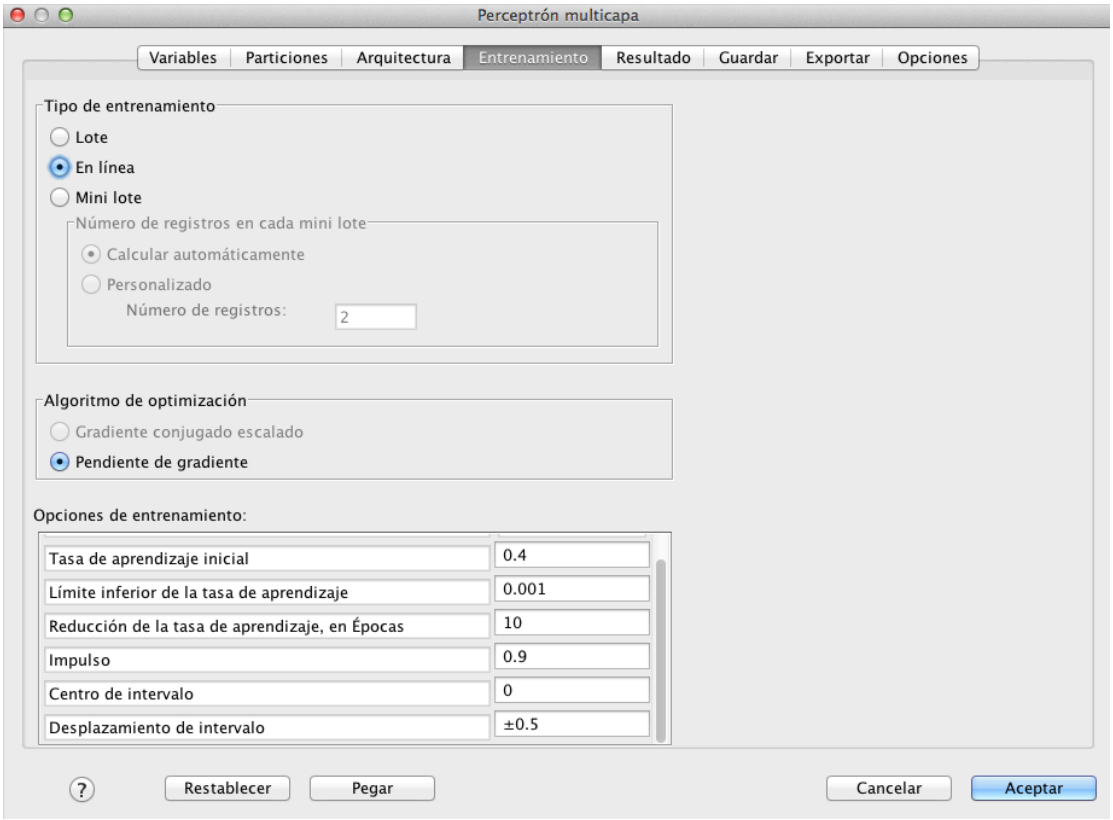

Figura 30. Pestaña de Entrenamiento
Fuente: Definición propia según investigación

En la pestaña de resultados se debe seleccionar el conjunto de gráficos o reportes de interés para el estudio. Se recomienda seleccionar todos estos elementos la primera vez que se use la herramienta, esto con el objetivo de conocer todos los detalles que ofrece SPSS:

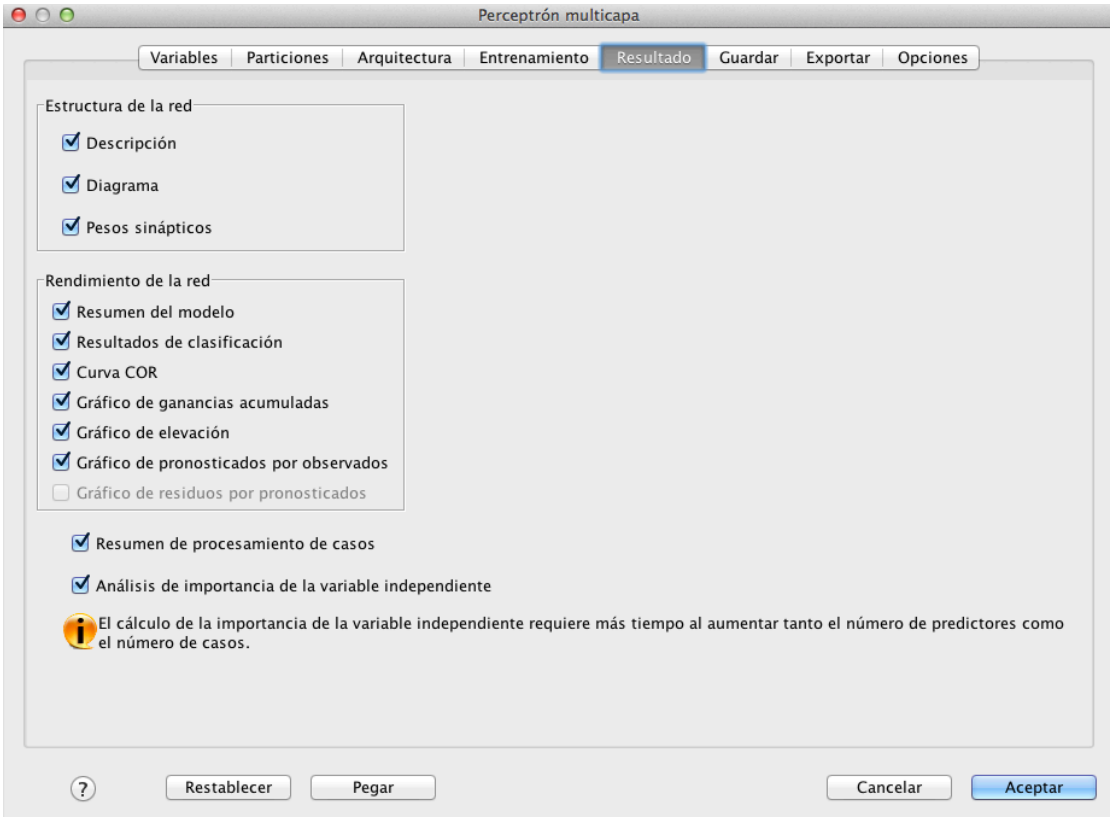

Figura 31. Pestaña de Resultados
Fuente: Definición propia según investigación

Una vez seleccionados todos los detalles del modelo, se debe presionar el botón "Aceptar" para la generación del perceptrón multicapa. SPSS mostrará los detalles del modelo construido.

6.3 APLICACIÓN DEL MODELO

Una vez identificados los valores óptimos para la configuración del modelo, se puede proceder a aplicar esa configuración específica para la clasificación de una base de datos de suscriptores de interés. En esta fase, se debe preparar la información de los casos de estudio de la misma forma que se prepararon los datos que sirvieron para entrenar la red.

Una vez cargados los datos en SPSS, se debe identificar cuáles son los datos de entrenamiento y cuáles son los datos de prueba. Para ello se debe crear una variable

numérica de partición en SPSS y asignar un valor positivo a esta variable en los casos de entrenamiento y un valor cero en los casos de prueba. La muestra de reserva estará conformada por los casos a los que se quiera aplicar el modelo para clasificación, para ello, asigne un valor negativo en la variable de partición de estos casos.

El modelo debe crearse con los valores de configuración óptimos identificados en la fase anterior, sin embargo, para asignar las particiones ahora se debe utilizar la variable de partición:

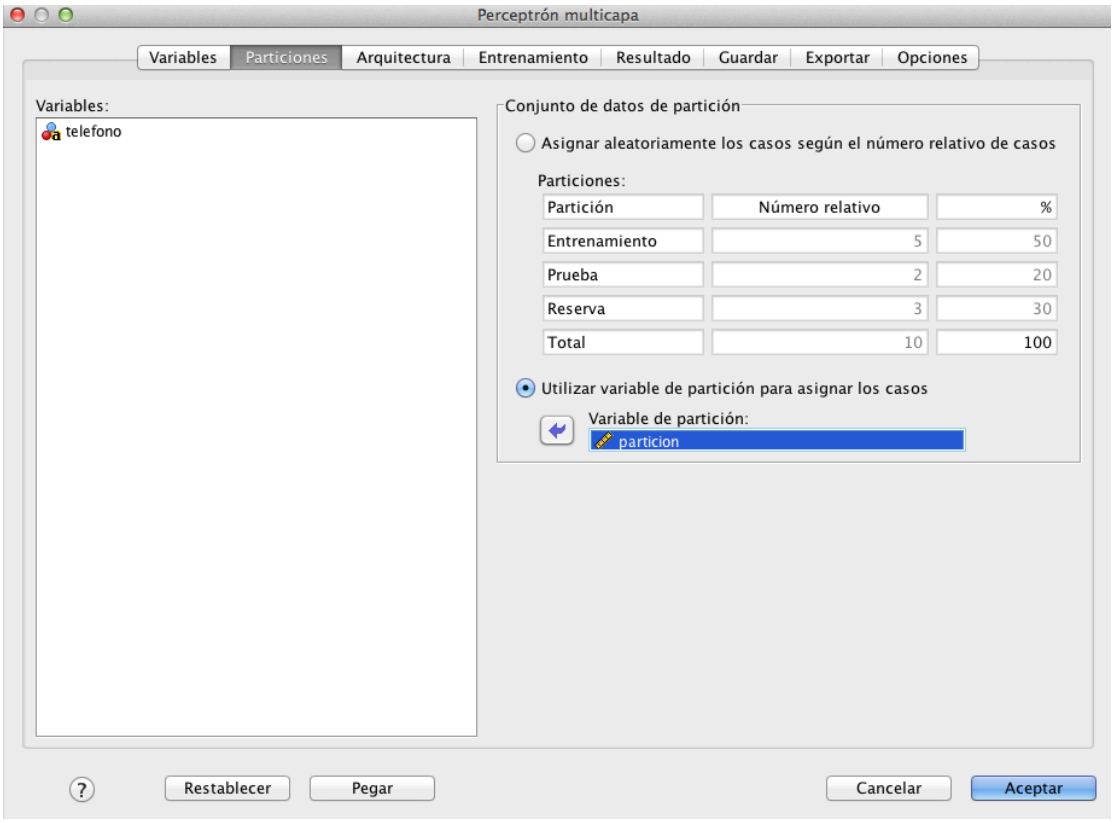

Figura 32. Asignación de Variable de Partición
Fuente: Definición propia según investigación

Una vez seleccionados todos los detalles del modelo, se debe presionar el botón "Aceptar" para la generación del perceptrón multicapa. SPSS mostrará los detalles del modelo construido y la clasificación de los casos de interés.

6.4 CRONOGRAMA DE EJECUCIÓN

En la Figura 33 se presenta el cronograma de actividades para la construcción y aplicación del modelo de tráfico gris.

Se ha considerado una etapa inicial de capacitación en SPSS, para poder asegurar la base de conocimientos requerida para una implementación efectiva.

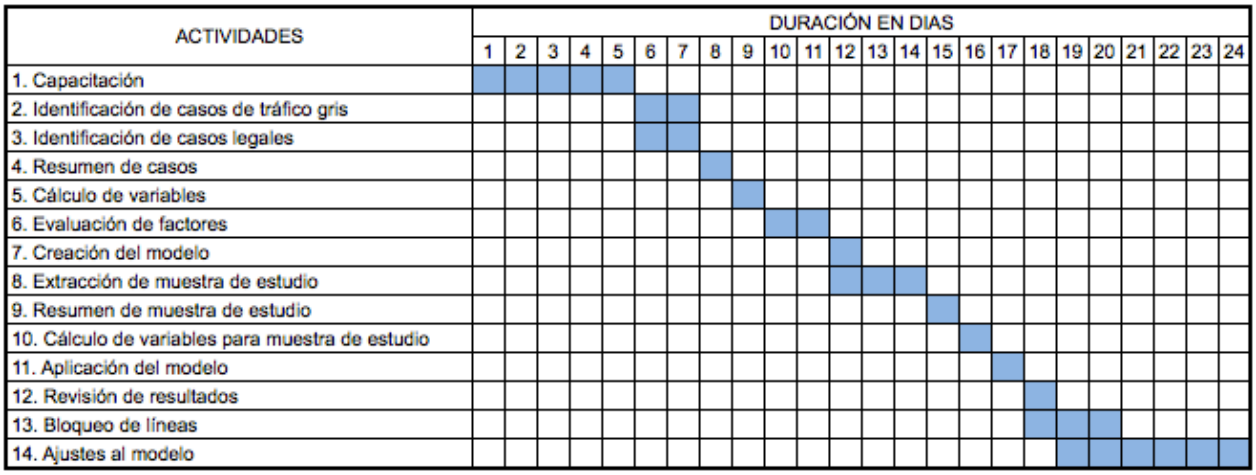

Figura 33. Cronograma de Ejecución
Fuente: Definición propia según investigación

6.5 PRESUPUESTO

PRODUCTO	CANTIDAD	PRECIO UNITARIO	PRECIO TOTAL
Licencia SPSS Base	2	$ 1,000.00	$ 2,000.00
Licencia SPSS Neural Networks	2	$ 1,000.00	$ 2,000.00
Capacitación SPSS (5 días)	2	$ 500.00	$ 1,000.00
		TOTAL:	$ 5,000.00

Figura 34. Estimación de Presupuesto
Fuente: Definición propia según investigación

En la Figura 34 se presenta el presupuesto requerido para la construcción y aplicación del modelo de trafico gris. Los montos considerados en la elaboración del presupuesto son estimados, ya que los precios pueden variar de acuerdo a la disponibilidad de contratos existentes con los proveedores.

BIBLIOGRAFÍA

3GPP (2013). TR 21.905 - Vocabulary for 3GPP Specifications.

Cahners In-Stat & MDR (2002). The Mobile Glossary.

CONATEL (1995). Ley Marco del Sector de Telecomunicaciones.

CONATEL (2003). Reglamento de Interconexión.

CONATEL (2011). Informe de Actividades y Logros Realizadas por parte de CONATEL durante el año 2011.

CONATEL (2012). Diario La Gaceta, Resolución NR002/12

Elmi, A. H., Ibrahim, S., & Sallehuddin, R. (2013). Detecting SIM Box Fraud Using Neural Network. Lecture Notes in Electrical Engineering, p. 575.

Galvan I. (2004). Redes de Neuronas Artificiales: Un Enfoque Práctico. Madrid: Pearson.

Grabosky, Peter N., Smith, Russell G. Crime in the Digital Age: Controlling Telecommunications and Cyberspace Illegalities, (3 ed. 2009), p. 71.

Graupe D. (2007). Principles of Artificial Neural Networks. (2 ed.). Chicago, USA: World Scientific Publishing.

GSMA (2012). Explicación del Roaming Móvil

GSMA (2013). The Mobile Economy

Heaton J. (2008). Introduction to Neural Networks with Java. (2 ed.). St. Louis, USA: Heaton Research.

Heckmann, O. M. (2007). The Competitive Internet Service Provider: Network Architecture, Interconnection, Traffic Engineering and Network Design. John Wiley & Sons., p. 20 – 23.

Heine, Gunnar (1999). Gsm Networks: Protocols, Terminology, and Implementation.

IBM Corp. (2013). IBM SPSS Neural Networks 22., p.11

Mansell, R. K. (1986). The Telecommunication Bypass Threat: Real or Imagined?. Journal Of Economic Issues (Association For Evolutionary Economics), 20(1), p. 145.

Nedjah, Nadia & Macedo Mourelle, Luiza de (2005). Evolvable Machines: Theory & Practice.

Redl S., & Weber M. (1998). GSM and Personal Communications Handbook

Rodríguez López, C. A., & Sánchez, S. M. (2005). Sistema para la Detección de Patrones de Fraude en las Redes de Telecomunicaciones. (2 ed.), p. 44–50.

Sivanandam, S. N. & Paulraj, M. (2009). Introduction to Artificial Neural Networks, p.120.

Trevisan, P. (2000). Monitoring mobile fraud. Telecommunications, 34(8), 61-64.
Tribunal Superior de Cuentas (2005). Investigación Técnica Practicada al Control del Tráfico Gris en la Empresa Hondureña de Telecomunicaciones.

W. Baluja & A. Llanes (2005). Estado Actual y Tendencias del Enfrentamiento del Fraude en las Redes de Telecomunicaciones.

Yeung D.S., Cloete I., Shi D. & Ng W.W.Y. (2010). Sensitivity Analysis for Neural Networks. Berlin: Springer-Verlag.

Yegnanarayana, B. (2006). Artificial Neural Networks.

ANEXOS

Entrevista: Fraude en Telefonía Móvil

1. En su tiempo trabajando en el campo de aseguramiento de ingresos, ¿De qué actividades ha participado relacionadas con resolver la problemática de tráfico gris?

2. ¿Cuáles son las técnicas conocidas más utilizadas por los individuos que realizan tráfico gris?

3. Según su experiencia, ¿Cuáles características básicas determinan un comportamiento fraudulento?

4. Según su experiencia, ¿Cuáles son las características básicas de una línea telefónica legal?

5. Según su experiencia, ¿Qué tan dinámico es el perfil de un individuo que realiza tráfico gris en Honduras?

6. Según su experiencia, ¿Cuáles han sido las principales dificultades en el trabajo de detección de fraude?

7. ¿Está usted al tanto de las técnicas de detección que se utilizan en otros países? Si la respuesta es sí, ¿Sabe qué tan efectivas son?

8. Según su experiencia, ¿Considera que las técnicas aplicadas para detección de tráfico gris en otros países sean fácilmente adaptables a Honduras? ¿Por qué sí o por qué no?

9. Según su experiencia, ¿Las empresas están dispuestas a invertir para implementar nuevas técnicas de detección? ¿Por qué sí o por qué no?

10. ¿Considera que puede mejorar la efectividad de detección de métodos tradicionales aplicando técnicas basadas en reconocimiento de patrones? ¿Por qué sí o por qué no?

GLOSARIO

Acuerdo de Roaming: Negociación entre dos operadores en el que se tratan los aspectos técnicos y comerciales que son necesarios para permitir este servicio. (GSMA, 2012, p.1)

Archivo TAP: Archivos de Procedimiento de Cuenta Transferida, utilizados para facturar las llamadas cuando se hace roaming, enviados a un centro de intercambio que se reenvían al operador base. (GSMA, 2012, p.2)

Área de Cobertura de Red: Un área donde los servicios de telefonía celular móvil se proveen por un sistema celular móvil al nivel requerido por dicho sistema. (3GPP, 2013, p.11)

ARPU: Ingreso Promedio Por Usuario. Se refiere a la cantidad de ingreso bruto que un operador de telefonía puede esperar, en promedio, de sus clientes. Normalmente, se calcula en base mensual, trimestral y anual (Cahners In-Stat/MDR, 2002, p. 2).

Carrier: proveedor de telecomunicaciones que cuenta con infraestructura propia para transporte y conectividad entre operadores y países. Hay del tipo locales y de larga distancia (Heckmann, 2007, p. 20-23).

CDR: Registro de Datos de Cobro, una colección de información en un formato específico sobre un evento tarificable, utilizado para cobro. (3GPP, 2013,p.10)

Clonación de SIM: Método de suplantación de identidad de una SIM, generalmente con objetivos fraudulentos. (Redl & Weber, 1998, p.462)

GSM: Sistema Global para Comunicaciones Móviles. Es una de las tecnologías celulares más utilizadas alrededor del mundo (Cahners In-Stat/MDR, 2002, p. 6).

IMSI: Identidad Internacional de Usuario Móvil, combinación del código de la SIM con los códigos de red asociado (3GPP, 2013, p.29).

Perceptrón Multicapa: Arquitectura de red neuronal constituida por una capa de entrada, una o más capas ocultas y una capa de salida las cuales están formadas por unidades de procesamiento (Nedjah & Macedo, 2005).

Ponderación Sináptica: Estimación de coeficiente que muestra la relación entre las unidades de una capa determinada con las unidades de la capa siguiente. (IBM, 2013, p. 11)

Red GSM: Red de telecomunicaciones que utiliza como base elementos de tecnología GSM (Heine,1999).

Red Local: La red base del suscriptor, responsable por el aprovisionamiento y control de los servicios del usuario. (3GPP, 2013, p.15)

Red Neuronal con Prealimentación: tipo de red neuronal que consiste en varias capas de unidades de procesamiento, cada capa alimenta las entradas de la capa siguiente hacia adelante a través de conexiones (Yegnanarayana, 2006, p.88).

Red Visitante: La red visitante o red sirviente, provee al usuario el acceso a los servicios de la red local (3GPP, 2013, p.29).

Retropropagación: Una red de retropropagación es una red multicapa con prealimentación en la cual las capas contienen neuronas de sesgo para ajuste de error (Sivanandam & Paulraj, 2009, p.120).

Roaming: La habilidad de un usuario de funcionar en una red de servicios diferente a la red local. La red visitante podría ser una red compartida operada por dos o más operadores de red (3GPP, 2013, p.26).

SIM Box: Dispositivo que mapea una llamada VoIP a una tarjeta SIM del mismo operador que el número destino (Elmi, Ibrahim, Sallehuddin, 2013, p.576).

Tarjeta SIM: Tarjeta de Módulo de Identidad del Suscriptor. Se utiliza en teléfonos GSM para almacenar el número del teléfono y otra información. Puede ser removida e insertada en otros teléfonos, permitiendo a los usuarios conservar sus números y realizar o recibir llamadas (Cahners In-Stat/MDR, 2002, p. 11).

Tráfico Gris: Se refiere a la creciente tendencia de gran cantidad de usuarios de telefonía hacia la evasión de los servicios de interconexión del proveedor local de telefonía, utilizando rutas alternas para llamadas internacionales (Mansell, 1986, p. 145).

VoIP: Tecnología Voz sobre IP. Consiste en el transporte de llamadas de voz sobre redes de datos.