SOBRAPO

PODes
Revista Eletrônica Pesquisa Operacional para o Desenvolvimento

Rio de Janeiro, v.4, n.3, p. 257-270, setembro a dezembro de 2012

USO DE MINERÍA DE DATOS PARA DETERMINAR LA DISPONIBILIDAD DE UNA RED IP V.4 EN UNA CADENA DE TERMINALES DISTRIBUIDOS. ESTUDIO DE CASO EN UNA EMPRESA DE JUEGOS DE AZAR.

Lorena Pradenas Rojas
Magíster en Ingeniería Industrial, Facultad de Ingeniería
Universidad de Concepción. Casilla 160-C. Correo 3. Concepción, Chile
lpradena@udec.cl

Carlos Parra
Magíster en Ingeniería Industrial, Facultad de Ingeniería
Universidad de Concepción. Casilla 160-C. Correo 3. Concepción, Chile
carlosparra@udec.cl

Resumen

En este estudio se diseña e implementa un sistema de detección de patrones erráticos, en una red de terminales IP (puntos de venta), de una empresa de juegos de azar, mediante *minería de datos*, con el objetivo de mejorar el *"uptime"*. Para detectar los patrones se generó, una herramienta de recolección de datos de los terminales, y posteriormente se validaron, corrigieron e ingresaron a los algoritmos: ID3, C4.5 (J48) y NNGe de la *minería de datos* (en ambiente *open source Weka*). Los resultados alcanzados, mejoran en un 1% el *"uptime"* total de la red de terminales de venta, generando una apropiada optimización de recursos.

Palabras chave: Minería de datos, patrones erráticos, redes de terminales IP.

Abstract

In this study a system for detecting erratic patterns in a network of IP terminals (point of sales), of a game company, with data mining, in order to improve "*uptime*", is designed and implemented. To detect how the patterns are generated, a tool was developed for data collection terminals, and subsequently validated, edited and entered the algorithms: ID3, C4.5 (J48) and NNGe of data mining (in environment open source *Weka*). The results achieved, improved by 1% the total "*uptime*" network of sales terminals, generating an appropriate value for money.

Keywords: Data mining, erratic pattern, IP terminal network.

Versão inicial submetida em 16/11/2010. Versão final recebida em 31/08/2012.

1. INTRODUCCIÓN

En la última década muchas empresas usan la tecnología para ofrecer sus productos y servicios en línea. Algunas de las empresas que se encuentran en esta categoría son las dedicadas a la venta de juegos de azar (rubro considerado en este estudio), debido al gran número de puntos de venta que poseen, y que además se encuentran localizados geográficamente dispersos. En sus comienzos estas empresas utilizaban como medio de ventas, boletos con diseños impresos, entregando escasas opciones al cliente para elegir una combinación preferida de números. En nuestro caso, estos boletos eran despachados a otras regiones incrementando, así también, los costos de operación. Actualmente se usa la tecnología, entre otros para disminuir costos y proporcionar un servicio en línea tanto a clientes como también a revendedores (agencias). En esta empresa de jugos alrededor del 100 % de los procesos críticos del negocio están actualmente automatizados reduciendo los costos ya que, la tecnología distribuida permite centralizar y automatizar procesos de: venta, pedidos, despachos, pagos de apostadores y agentes.

Las agencias de ésta empresa de juegos de azar, se encuentran distribuidas en todo el país y realizan el proceso venta, requieren un terminal conectado por algún tipo de comunicación al sitio central, donde se concentra la información tanto de, ventas, apuestas y pagos. La conexión al sitio central es mediante una red IP, proporcionado y administrado por proveedores de comunicaciones y que dispone de distintos tipos de comunicación instalados en los puntos de venta. Los considerados en este estudio son, mediante línea dedicada (cable) e inalámbrico (GPRS, WILL). Estos enlaces de comunicación permiten a los terminales de venta conectarse a la planta de comunicaciones geográficamente más cercana. No obstante, estas comunicaciones no se encuentran exentas de problemas que lamentablemente afectan directamente al cliente final y al agente. Cualquier problema con la línea de comunicación, impide que el cliente pueda adquirir el producto y realizar cobros cuando su boleto ha obtenido un premio además, de no es posible imposibilidad vender, lo que afecta tanto a la empresa como al propio agente.

Todo lo anterior, hace necesario disponer de un sistema de detección de patrones de comportamiento errático, para así: Evitarlos en futuras instalaciones, corregir los puntos ya instalados y aumentar la disponibilidad de otros centros de venta (Kimball, 2008). Para identificar estos patrones en base a los atributos disponibles, en este estudio, se utilizó una metodología que incluye herramientas de minería de datos.

La minería de datos, es usada para descubrir conocimiento útil a partir de grandes cantidades de datos. Además, el descubrimiento del conocimiento es considerado un proceso que consta de varias etapas, tales como: Comprensión del dominio, preparación del conjunto de datos, descubrimiento de patrones, análisis de patrones descubiertos y utilización de resultados, permitiendo así negocios mas inteligentes desde el punto de vista estratégicos y tácticos (Abulkari y Job, 2003). La minería de datos es un proceso que combina métodos y herramientas de: Estadística, bases de datos, optimización y aprendizaje automático.

En Kotsiantis (2011), se establece que con el aumento de la cantidad de datos e información, la habilidad de aprendizaje incremental se hace cada vez más importante para el enfoque de *machine learning*.

En Cavalieri (2012), se analiza el sistema KNXnet/IP. El sistema de comunicación KNX, ha sido integrado al ambiente IP por medio de la definición de las especificaciones KNXnet/IP, que permite la integración de diferentes subredes KNX a través de una red IP, por medio de un dispositivo denominado Router KNXnet/IP.

Además, en Ur-Rahman y Harding (2012), se propone una metodología híbrida para la manipulación de datos con formato de texto, y se demuestra su efectividad con un caso de estudio real tomado de la industria. Para probar su utilidad se utilizan diferentes clasificadores, tales como: Arboles de decisión, 'Naive Bayesian Learner', clasificador K-NN y máquinas de soporte vectorial.

En Dreyer et al. (2009) se realiza una revisión de las técnicas de minería de datos encontradas en publicaciones SciELO y LILACS, entre 1999 y 2008. También recoge información de libros especializados. Se seleccionaron 32 artículos para la elaboración de esta publicación.

Por otro lado, es necesario exponer algunos estudios sobre la infraestructura tecnológica para: Medición, almacenamiento y correlación de los diferentes tipos de datos existentes en una red IPv.4, dentro del cual podemos mencionar el *"Measurement and Analysis of IP Network Usage and Behavior"* publicado por Cáceres et al. (2000) Enn este estudio se detallan los puntos más importantes para obtener una plataforma sólida de captura y procesamiento de información, por ejemplo: Generación de una arquitectura tecnológica para realizar las mediciones respectivas, herramientas de captura de datos, almacenamientos distribuidos, puntos de medición basados en protocolos y equipamiento de almacenamiento.

2. MÉTODOS

El objetivo de este estudio, es diseñar e implementar un sistema de detección de patrones de caídas de terminales de venta que permita, medir y mejorar el "*uptime*" de una red de ventas a lo menos en un 1%, utilizando las bondades de la minería de datos. Donde "*uptime"* es la disponibilidad que tienen los terminales de venta de enviar una transacción al "centro de procesamiento de datos" de la empresa, sin tener "caídas" de red. Para una mejor comprensión del estudio a continuación, se proporcionan algunas definiciones y temas relevantes.

2.1 ALGUNAS DEFINICIONES

- *Datamining*. Es una metodología que permite identificar tendencias, patrones y comportamientos, no solo para extraer información, sino también para descubrir las relaciones en bases de datos y entre otros, también, determinar comportamientos no evidentes. En: Witten y Frank (2005), Hand et al.(2001), Vercellis (2009),entre otros, se encuentran definiciones del concepto de minería de datos. Una de la más completa es:

> *"La minería de datos es el proceso de descubrir nuevas correlaciones, patrones y tendencias a través del análisis de grandes cantidades de datos almacenados en bases de datos, usando tecnología de patrones de reconocimiento, así como técnicas estadísticas y matemáticas"*

Los conocimientos extraídos de la minería de datos pueden ser en forma de: Relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa (es decir, un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Los modelos pueden ser de dos tipos: Predictivos y descriptivos.

Los modelos predictivos, estiman valores futuros o desconocidos de variables de interés, que se denominan variables objetivos o dependientes, usando otras variables independientes o predictivas.

Los modelos descriptivos, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, y no para predecir nuevos datos (Hernández et al., 2004).

-Descubrimiento del conocimiento (*KDD).* Se define como *"la extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos*". Existe una distinción clara entre el proceso de extracción de datos y el descubrimiento del conocimiento. Bajo sus convenciones, el proceso de descubrimiento del conocimiento considera los resultados tal como vienen en los datos (proceso de extraer tendencias o modelos de los datos). Adecuadamente y con precisión los transforma en información útil y entendible. Esta información no es fácilmente recuperable por las técnicas normales pero es descubierta a través del uso de técnicas de la Inteligencia Artificial. Diversos autores se refieren al proceso de minería de datos como la aplicación de un algoritmo para extraer patrones de datos y a KDD como el proceso completo (pre-procesamiento, minería, post-procesamiento).

El proceso de KDD, consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos. El "*Datamining*", como se mencionó, es una etapa del proceso de descubrimiento del conocimiento, KDD (descubrimiento del conocimiento), puede usarse como un medio para obtener conocimiento, de la misma manera que los agentes inteligentes realizan la recuperación de información en el Web. Nuevos modelos o tendencias en los datos pueden descubrirse usando estas técnicas. KDD también pueden ser usadas como una base para las interfaces inteligentes del futuro, agregando un componente del descubrimiento del conocimiento a una máquina de bases de datos o integrando KDD con las hojas de cálculo y visualizaciones.

-Mitos o conceptos sobre la minería de datos. Existe una serie de mitos acerca de la minería de datos, algunas de las cuales sn descritas por Jen Que Louie (2003) y son:

- *"La minería de datos es un proceso autónomo, que requiere poco o no requiere supervisión de un humano".*
- *"En la minería de datos la inversión utilizada se recupera muy rápidamente"*.
- *"Los paquetes de software de "datamining" son intuitivos y muy fáciles de usar"*.
- *"La minería de datos identifica la causa de los problemas o investiga el problema"*.
- *"La minería de datos puede limpiar los datos incorrectos o basura de la base de datos de manera automática"*.

Dado que el objetivo de este estudio es utilizar la minería de datos para mejorar el *"uptime"* de una red de terminales distribuidos, a continuación se detalla cada una de las actividades realizadas

en esta investigación para determinar los factores erráticos de un conjunto de datos que maneja una red de terminales IP de una empresa de juegos de azar.

-Aprendizaje del tema y estudio del problema. La problemática radica en descubrir y predecir el comportamiento de los terminales de venta, esto permitirá tomar mejores decisiones sobre la instalación de un nuevo terminal. El levantamiento de la problemática se puede tipificar como un problema de clasificación, ya que como resultado se espera obtener el subconjunto en el cual se encuentra el terminal a evaluar.

2.2. SELECCIÓN DEL CONJUNTO DE VARIABLES A ESTUDIAR Y RECOLECCIÓN DE DATOS

Esta etapa se refiere a la definición de variables independientes y mencionadas a continuación:

- ***Tipo de enlace de comunicación:*** Estas variables de enlace, puede tomar los siguientes valores: ADSL, WILL, GPRS.
- ***Tipo de Terminal:*** Esta variable puede tomar los siguientes valores: Olivetti, Ovation.
- ***Ubicación:*** Esta variable puede tomar los siguientes valores: Ciudades de Chile.
- ***Segmento de ventas al que pertenece:*** Esta variable binaria se refiere a si el terminal pertenece al conjunto de terminales que genera el 70% de las ventas totales o no. Puede tomar los valores: Si o No.

Por otro lado, la variable a explicar y predecir a través de las variables independientes, corresponde a la variable *"uptime"* , es una variable nominal, categórica y dependiente.

La categorización de la variable "*uptime"*. se realiza de acuerdo a la experiencia, definiendo rangos basados en umbrales. Ver Tabla 1, donde de acuerdo al porcentaje, se proporciona una calificación para el "*uptime*"

Tabla 1. Rangos basados en umbrales para *"uptime"*.

Porcentaje de "uptime"	Nivel de "uptime"
0 a 75	Muy Malo
76 a 85	Malo
85 a 95	Regular
96 a 100	Bueno

2.3 DISEÑO Y GENERACIÓN DE HERRAMIENTA DE CAPTURA DE DATOS

Es necesario diseñar e implementar una herramienta apropiada para la captura de datos, esto incluye el diseño de base de datos relacional y un software cliente servidor. A continuación se describe la operación de la herramienta (cliente-servidor), la cual se ilustra en la Figura 1 y conteniendo los elementos:

- ***Software cliente o agente:*** Este software extrae del terminal de venta los siguientes datos.
  - Número de terminal.
  - Fecha y hora.

- Identificador único de terminal.
- Identificador único de disco duro.
- Cantidad de impresoras térmicas que el terminal tiene encendidas.
- Versiones de software cargados.
- Nombre del terminal.

- ***Software servidor:*** Este software recepciona todas las transacciones enviadas por los terminales para luego almacenarlos en la *base de datos relacional*.

  En cuanto al volumen de los datos utilizados en este estudio se estiman alrededor de siete millones de datos mensuales con terminales operando 24 horas del día.

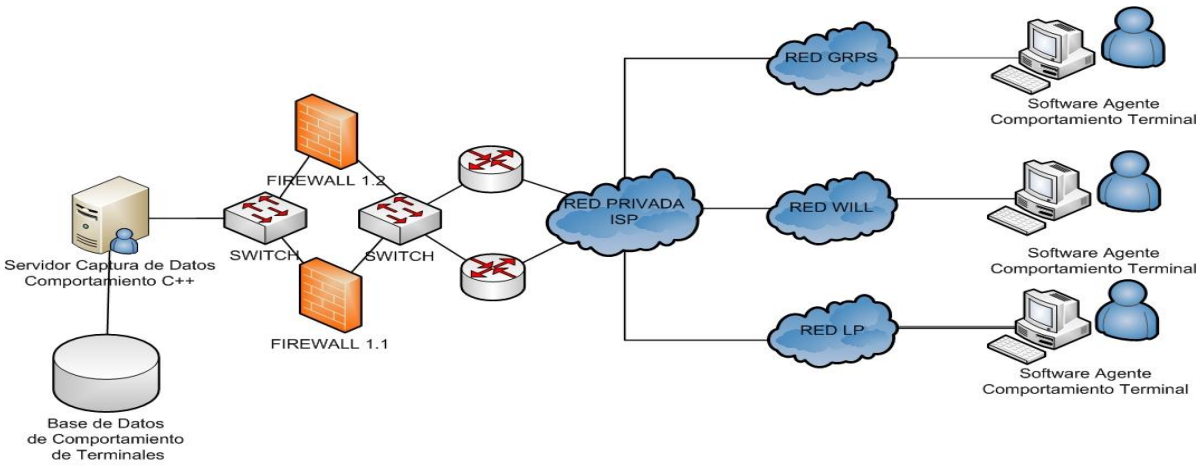

Figura 1. Esquema de funcionamiento del software cliente-servidor para la captura de datos.

2.4 ALMACENAMIENTO Y LIMPIEZA DE DATOS

En esta etapa del proceso se realizan las siguientes operaciones:

- ***Creación de base de datos de trabajo:*** Los datos son almacenados en forma automática en la base de datos que existe en el servidor de captura de datos.

- ***Limpieza y pre-procesamiento de los datos:*** Los datos son exportados a planillas Excel para su mejor manejo y detección de errores. Los datos erróneos son corregidos de acuerdo a la experiencia, ya que muchos son evidentes y fáciles de corregir, los más frecuentes son:

  - Atributos sin valor corregidos, revisando archivos de exportación de datos.
  - Atributos con valores fuera de rango, corregidos revisando archivos de exportación de datos.

- Atributos con datos inconsistentes, son eliminados del estudio.
- Atributos con datos incompletos, corregidos revisando archivos de exportación de datos.
- Atributos con caracteres inconsistentes, corregidos consultando a la fuente de información.
- Atributos duplicados, son eliminados los considerados sobrantes.

- ***Reducción de datos y proyección:*** Debido a la cantidad de datos existentes se realizan reducciones a nivel de base de datos, los que no presentan utilidad, como por ejemplo números de serie del equipamiento, cantidad de impresoras entre otros.

2.5 SELECCIÓN DE ALGORITMO DE MINERÍA DE DATOS

El algoritmo que permite clasificar y predecir el futuro comportamiento, usa árboles de decisión y ha sido mencionado en diversas aplicaciones (Vitt et al., 2003). La característica principal para el uso de árboles de decisión es su facilidad para la interpretación debido a que el resultado será entregado como procedimiento al personal operativo encargado de las reparaciones e instalación de nuevos terminales. Específicamente, los algoritmos utilizados en este estudio son **ID3, C4.5 (J48) y NNGe**. Los algoritmos **ID3** y **C4.5 (J48)** se conocen como TDIDT *(Top-Down Induction of Decision Trees)*.

Por otro lado, en los algoritmos mencionados la función utilizada es representada mediante un árbol de decisión, con reglas "*if-then"* para una mejor legibilidad. Los resultados que entrega este tipo de algoritmo, son ideales para la de toma de decisiones, permitiendo a la unidad de nuevas instalaciones de puntos de ventas, decidir el tipo de comunicación a usar en las nuevas instalaciones y así mejorar en forma substancial el *"uptime"* del punto de venta.

Cada nodo interno del árbol corresponde a un *"test"* de valor de algún atributo de la instancia a clasificar, y las ramas que descienden de este nodo son rotuladas con los valores posibles del *test*. Cada nodo hoja de un árbol de decisión especifica el valor retornado en caso de que la hoja sea alcanzada. Estos valores corresponden a posibles clasificaciones de una instancia, una instancia se clasifica comenzando en el nodo raíz del árbol verificando el atributo especificado por este nodo.

Un resumen del método utilizado para resolver el problema se presenta en la Figura 2 y las etapas previas descritas, corresponden a las seis primeras de la Figura.

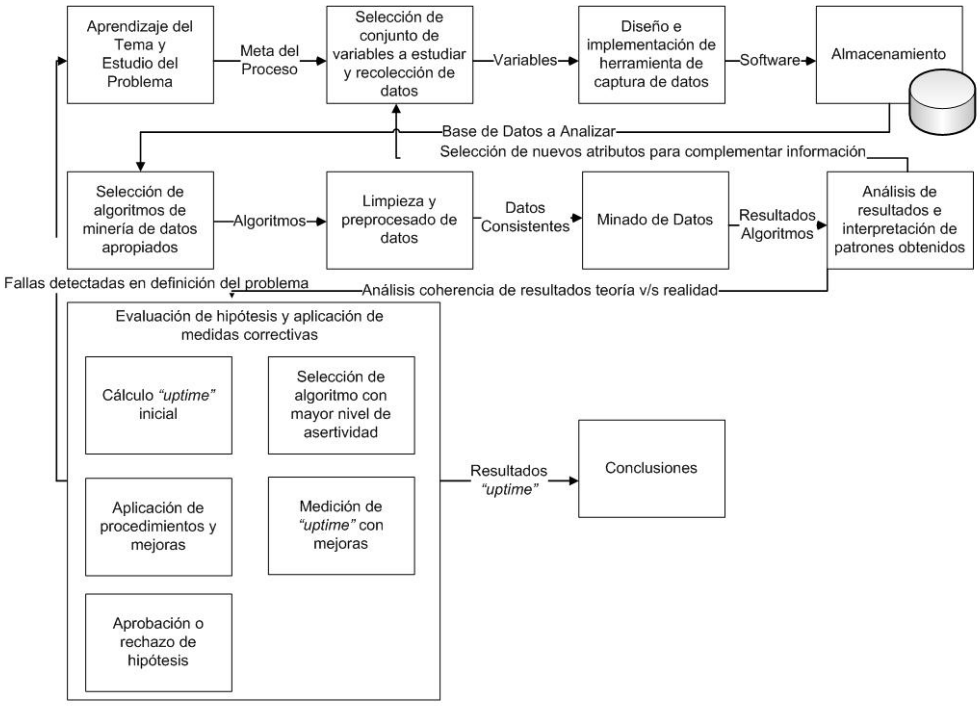

Figura 2. Esquema de metodología utilizada

3. RESULTADOS

3.1 MINADO DE DATOS

La obtención de resultados, se inicia en la etapa de *minado de datos* correspondientes al sexto rectángulo de la Figura 2. Específicamente en esta investigación se usaron 1621 registros extraídos de la siguiente forma:

- **Selección de la muestra de entrenamiento:** La muestra utilizada para el entrenamiento fue obtenida en forma aleatoria con la función "*random"* del motor de la base de datos, obteniendo como resultado 866 registros de información, correspondientes al mes de noviembre del año 2009.

- **Selección de la muestra para validación del aprendizaje:** La muestra utilizada para la validación del aprendizaje fue obtenida en forma aleatoria con la función *random* del motor de la base de datos, corresponde a 735 registros de información correspondientes del mes de diciembre del año 2009.

Utilizando el algoritmo **ID3 (*Induction Decision Trees*)** del entrenamiento, se obtienen los resultados presentados, en la matriz de *confusión* de aprendizaje supervisado, conteniendo en las columnas, lo casos "predichos" y en las filas, los casos "acertados", con las instancias clasificadas correctamente por el algoritmo **ID3**, ver Tabla 2. De un total de 866 instancias se clasificaron, correctamente 694 instancias; 596 puntos de venta fueron clasificados con *"uptime"* Bueno, 67 con *"uptime"* Regular, 17 con *"uptime"* Malo y 14 con *"uptime"* Muy malo. En las columnas, de la Tabla 2: A es Bueno, B es Regular, C es Malo y D es Muy malo.

Tabla 2. Matriz de confusión ID3.

A	B	C	D
596	5	1	1
74	**67**	0	0
58	4	**17**	0
26	2	1	**14**

En la Tabla 2, incorrectamente son clasificados 5 puntos de venta, que fueron clasificados como Regular y eran Bueno. Un punto de venta fue clasificado como Malo cuando era Regular, un punto de venta fue clasificado como Muy malo cuando era Malo. Además, 74 puntos de venta, fueron clasificados como Bueno y eran Regular. También, 58 puntos de venta, fueron clasificados como Bueno y eran Regulares, 4 fueron clasificados como Regular y eran Malos. En la última fila, se observa que 26 puntos de venta, fueron clasificados como Bueno y eran Regular, 2 puntos de venta fueron clasificados como Regular y eran Malos y un punto de venta fue clasificado como Malo y era Muy malo.

En resumen, aplicando el algoritmo **ID3**, hubo 694 instancias clasificadas correctamente, correspondientes a la suma de la diagonal de la Tabla 2 o equivalente a un 80 % de asertividad. 172 instancias clasificadas incorrectamente, correspondiendo a la suma de los elementos fuera de la diagonal principal o con asertividad sólo del 20 % del total de instancias de la muestra de entrenamiento.

Por otro lado, usando el algoritmo **C4.5 (J48),** en la etapa de entrenamiento, se obtienen los resultados de la Tabla 3, al igual que en la Tabla 2, se tiene que en las columnas: A es Bueno, B es Regular, C es Malo y D es Muy malo

Tabla 3. Matriz de confusión J48.

A	B	C	D
603	0	0	0
141	**0**	0	0
79	0	**0**	0
43	0	0	**0**

De un total de 866 instancias, se clasificó correctamente 603 que es la suma de la diagonal de la Tabla 3, es decir, un 70 % de asertividad. Incorrectamente fueron clasificados 141 puntos de venta, calificados como Bueno a pesar que eran Regular, 79 clasificados como Bueno y era Malos y 43 como Bueno a pesar que eran Muy malos o sea 30% de asertividad.

Utilizando el algoritmo **NNGe (*Nearest neighbor*)** en la etapa de entrenamiento se obtuvo los resultados de la Tabla 4. Al igual que en las tablas 2 y 3, en las columnas: A es Bueno, B es Regular, C es Malo y D es Muy malo

Tabla 4. Matriz de confusión NNGe.

A	B	C	D
505	59	21	18
40	**91**	6	4
25	17	**32**	5
13	6	0	**24**

De un total de 866 instancias, se clasificó correctamente a 652, que es la suma de la diagonal de la Tabla 4, es decir, un 75 % de asertividad, 505 puntos de venta fueron clasificados correctamente con *"uptime"* Bueno, 91 con *"uptime"* Regular, 32 con *"uptime"* Malo, y 24 con *"uptime"* Muy malo. Incorrectamente fueron clasificados 59 puntos de venta tratados, como Regular y eran Bueno, 21 punto de venta fueron clasificados como Malo y eran Regular, 18 puntos de venta fueron clasificados como Muy malo y eran Malos. Además, 40 puntos de venta, fueron clasificados como Bueno a pesar que eran Regular, 6 fueron clasificados como Malo y eran Regular, 4 fueron clasificados como Muy malo y eran Malo. También 25 puntos de venta, fueron clasificados como Bueno y eran Regulares, 17 fueron clasificados como Regular y eran Malo, 5 fueron clasificados como Muy malos y que eran Malo. Finalmente en la última fila, 13 puntos de venta, fueron clasificados como Bueno y eran Regular, 6 fueron clasificados como Regular y eran Malo.

Una tabla comparativa con los resultados de la aplicación de los algoritmos de minería de datos se observa en la Tabla 5.

Tabla 5. Resultados entrenamientos. ID3 - J48 - NNGe.

Algoritmo	Correctamente Clasificados	Incorrectamente Clasificados	% Tasa de Asertividad
ID3	694	172	80
J48	603	263	70
NNGe	652	214	75

El algoritmo que mejor nivel de predicción presenta para este tipo de problemas es el algoritmo de **ID3**, clasificando con un porcentaje de asertividad de un 80 %, el algoritmo tiene un 10 % más de asertividad que **J48**, por ende, si aplicamos el árbol de decisión obtenido por **ID3** en la instalaciones de nuevas agencias, podemos predecir el 80 % del comportamiento. Esto permitirá tomar decisiones con mayores niveles de información y seleccionar el enlace de comunicación y atributos que mejor se comporten en una comuna determinada. En el caso del algoritmo **NNGe** el porcentaje de asertividad es de un 75 % el cual se encuentra por sobre **J48**.

3.2 RESULTADOS CON GRUPO DE DATOS DE PRUEBA

Una vez realizada el aprendizaje en el "minado de datos", se procede a validar el aprendizaje.

Tabla 6. Resultados de validación ID3 - J48 - NNGe.

Algoritmo	Correctamente Clasificados	Incorrectamente Clasificados	% Tasa de Asertividad
ID3	585	150	79,6
J48	507	228	69,0
NNGe	560	175	76,2

En la Tabla 6, se muestran los resultados utilizando 735 datos para validación del aprendizaje, en el cual se aprecia una disminución leve, en el nivel de asertividad del algoritmo **ID3** para la clasificación de estos datos, al igual para el algoritmo **J48** que disminuyó a un 69,0% de asertividad, no es el caso de **NNGe** el cual aumento un poco mas de 1% su nivel de asertividad.

3.3 ANÁLISIS DEL TIEMPO CONSTRUCCIÓN DE MODELOS

En el "minado de datos", se presentan en la Tabla 7, el tiempo resultante de la ejecución de los algoritmos.

Tabla 7. Tiempos ejecución de algoritmos ID3 - J48 - NNGe.

Algoritmo	Tiempos usado en construcción del Modelo (seg.)
ID3	0,06
J48	0,03
NNGe	0,022

El algoritmo que peor desempeño tiene en la construcción del modelo es el algoritmo **J48** y el mejor es el algoritmo **NNGe**. El algoritmo que posee el mejor nivel de predicción construye el modelo en tan solo 0.03 segundos (**ID3**).

3.4 ANÁLISIS DE RESULTADOS E INTERPRETACIÓN DE PATRONES OBTENIDOS

En la octava etapa del procedimiento (Figura 2), los patrones de comportamiento erróneos alcanzados por el algoritmo, se observan en la Tabla 8, que presenta la combinación de atributos que producen como resultado un *"Uptime"* deficiente (Malo o Muy malo), entonces deben ser evitados para nuevas instalaciones de terminales de venta. Esta Tabla fue construida a partir de los resultados entregados por el algoritmo **ID3.**

Tabla 8. Resultados clasificación combinación atributos *"uptime"* en ciudades estudiadas.

Comuna	Enlace	TipoTerminal	Ventas	Uptime
Ancud	DIGI	N/R	N/R	Muy malo
Calama	N/R	Olivetti	NO	Malo
Calbuco	N/R	N/R	N/R	Malo
Colbún	N/R	N/R	N/R	Muy malo
Conchalí	N/R	Olivetti	SI	Malo
Estación Central	DIGI	N/R	N/R	Muy malo
Huechuraba	N/R	N/R	N/R	Muy malo
Independencia	ADSL	N/R	NO	Malo
LaReina	DIGI	Ovation	N/R	Malo
PedroAguirreCerda	Will	N/R	NO	Muy malo
PuenteAlto	DIGI	Olivetti	N/R	Malo
Purén	N/R	N/R	N/R	Malo
Quillota	N/R	Olivetti	NO	Muy malo
QuintaNormal	WILL	N/R	N/R	Muy malo
Recoleta	DIGI	N/R	SI	Malo
Renca	DIGI	N/R	SI	Malo
SanJoaquin	ADSL	N/R	SI	Malo
SanMiguel	ADSL	Olivetti	N/R	Malo
Talca	N/R	Olivetti	SI	Malo
Taltal	N/R	N/R	N/R	Malo
TierraAmarilla	N/R	N/R	N/R	Malo
Valdivia	ADSL	N/R	NO	Malo
Valparaíso	DIGI	N/R	SI	Muy malo
Providencia	N/R	Olivetti	NO	Malo

A partir de la Tabla 8, generada por el algoritmo **ID3,** se puede comentar que:

- En la comuna de Ancud y Estación Central, el tipo de comunicación denominada "DIGI" no debe ser utilizada para nuevas instalaciones por que no funciona correctamente, se

recomienda el uso de comunicación “ADSL”, además se sugiere cambiar el tipo de enlace de los terminales de venta que actualmente están funcionando con el sistema “DIGI” a “ADSL”, se informa también el mal comportamiento de la red inalámbrica en esas comunas.

- En las comunas de Calbuco, Colbún, Huechuraba, Purén, TalTal y TierraAmarilla se debe informar al proveedor de servicio de comunicaciones que los diferentes tipos de comunicación instalada no funcionan correctamente, se recomienda realizar las solicitudes respectivas para identificar el origen del problema.
- En Conchalí y Talca el tipo de comunicación no es relevante para determinar el comportamiento del *“uptime”*, el atributo que define el comportamiento es tipo de terminal en este caso “Olivetti” y el nivel de ventas “SI”, el atributo terminal de preferencia debiera ser “Ovation” para evitar el mal comportamiento del *“uptime”* y el nivel de ventas debiera ser “NO”.
- En Recoleta, Renca y Valparaíso para un nivel de ventas “SI” o alto más el atributo comunicación “DIGI” el comportamiento es Malo, se recomienda la instalación de los nuevos terminales de venta con enlaces de comunicación “ADSL” o “WILL”.
- En Independencia y Valdivia para un nivel de ventas “NO” o bajo el tipo de enlace “ADSL”, no es recomendable ya que existe un comportamiento errático bajo estas condiciones.
- En Calama y Quillota el tipo de terminal “Olivetti” y el nivel de ventas “NO” o bajo, sin importar el enlace provoca un nivel de comportamiento bajo en el atributo *“uptime”*.
- En el caso de la comuna de La Reina para nuevas instalaciones no se recomienda la instalación de la combinación de enlace de tipo DIGI con la marca de terminal “Ovation”.
- En Pedro Aguirre Cerda, la instalación de enlace de tipo WILL para terminales con un nivel de ventas NO o bajo, no es recomendable instalar este tipo de combinación debido al mal comportamiento de *“uptime”*.
- En Puente Alto, no se recomienda la instalación de enlace de comunicación DIGI y terminal de marco “Olivetti” debido a que el comportamiento del terminal utilizando es Malo.
- Para la comuna de Quinta Normal no es recomendable la instalación de enlaces de comunicación WILL.
- En la comuna de San Joaquín la instalación de enlace de tipo ADSL para terminales los cuales tienen un nivel de ventas alto no es recomendable.
- En San Miguel para terminales que tienen ventas altas SI no es recomendable utilizar enlaces ADSL.
- En la comuna de San Miguel no es recomendable utilizar enlaces de tipo ADSL con tipos de terminal “Olivetti”.
- Para la comuna de Providencia no es recomendable utilizar terminales de tipo “Olivetti”.

3.5 EVALUACIÓN DE HIPÓTESIS Y APLICACIÓN DE MEDIDAS CORRECTIVAS

En la novena etapa del procedimiento (ver Figura 2) se obtienen los siguientes resultados:

- Cálculo de *“uptime”* inicial (antes de aplicar procedimientos de instalación en base a los resultados obtenidos en el proceso de minado de datos).

  Durante el mes de noviembre, en el 92,97% del tiempo disponible los terminales estaban en condiciones de enviar una transacción (administrativa o de venta al *site* central, o sea el enlace de comunicación estaba operativo).

- Selección de algoritmo que presenta un mejor porcentaje de asertividad en cuanto a la clasificación y predicción de *“uptime”* en terminales.

El que mejor comportamiento obtuvo es el algoritmo ID3 con un valor aproximado de 80%.

- Aplicación del procedimiento de mejora e instalaciones nuevas obtenido durante el proceso de minado de datos para nuevas instalaciones e instalaciones que presenten un *"uptime"* deficiente (*"Malo"* o *"Muy malo")*.

  Se realizan los cambios respectivos para 28 terminales que presentaban *"uptime"* deficiente para luego calcular el nuevo *"uptime"* de la red.

- Medición para verificar el porcentaje de aumento de *"uptime"* en la red de terminales.

  El aumento de *"uptime"* fue de 92,97% a 94,13%, lo que representa una mejora de 1,16%.

- Aprobación o rechazo de hipótesis de investigación.

  "Es posible utilizar minería de datos para determinar y mejorar la disponibilidad de una red IP v.4 en una cadena de terminales distribuidos en al menos 1 %".

  De acuerdo a las mediciones realizadas luego de aplicar correcciones a los terminales con *"uptime"* deficiente se logró mejorar un 1,16% el *"uptime"* de la red de terminales de venta distribuidos, por lo tanto se aprueba la hipótesis de la investigación.

4. CONCLUSIONES

La metodología utilizada permite obtener resultados coherentes para la resolución del problema.

La principal dificultad que se presenta al realizar análisis de datos con *"datamining"*, es la obtención y estandarización de los datos que consume gran parte del tiempo esto es descrito en la literatura y así fue comprobado, en este caso correspondió al 75% del tiempo total.

El primer paso fue la programación de una herramienta de captura de estado de los terminales de ventas a través de todo el país, la cual fue instalada en la totalidad de los terminales, para luego proceder a almacenar estos datos en una base de datos y posteriormente procesados y estandarizados.

En la etapa de análisis de resultados, mediante la comparación de las técnicas de *"datamining"* utilizadas, se puede concluir que el uso de esta técnica es aplicable a problemas asociados con gran volumen de datos.

Respecto de los algoritmos utilizados para la resolución de la problemática, el de mejor desempeño en cuanto al porcentaje de clasificación correcta, es el algoritmo **ID3** pertenecientes a la técnica de árboles de decisión, el cual entregó alto nivel de asertividad con un 80 % comparado contra 70 % de **J48** y un 75 % de **NNGe.**

Sobre el tiempo en la construcción del modelo, el algoritmo que obtuvo el menor tiempo es el algoritmo de **NNGe** con solo 0.022 segundos.

El algoritmo **ID3,** entrega una serie de resultados y reglas que pueden ser fácilmente interpretadas para mejorar el nivel de *"uptime"* de los puntos de venta, en base a esta información se

generaron procedimientos para las futuras instalaciones en terreno, estos procedimientos incluyen que tipos de combinaciones de atributos no deben ser utilizados, ya que perjudicaría el *"uptime"* total de la red de ventas, la aplicación de estas herramientas entrega la información necesaria para realizar mejoras periódicas y la corrección de fallas dentro del proceso global de mejora continua.

Una vez interpretados los resultados obtenidos del *minado de datos*, se realizan las mejoras respectivas en 28 puntos de venta y en 10 nuevos puntos, esto entrega una mejora de "*uptime*" de la red de terminales de un 1,16%, por lo tanto y se prueba la hipótesis de investigación**.**

REFERENCIAS BIBLIOGRÁFICAS

ABULKARI, K. & JOB, V. (2003). Business Intelligence in Action. CMA Management, 77 (1), 15.

CAVALIERI, S. (2012). Modelling and analysing congestion in KNXnet/IP. Computer Standards & Interfaces, 34 (3), 305-313.

CÁCERES, R. & DUFFIELD, N. & FELDMANN, A. & FRIEDMANN, J.D. & GREENBERG, A. & GREER, R. & JOHNSON, T. & KALMANEK, C.R. & KRISHNAMURTHY, B. & LAVELLE, D. & MISHRA, P.P. & REXFORD, J. & RAMAKRISHNAN, K.K. & TRUE, F.D. & VAN DER MEMIE, J.E. (2000). Measurement and Analysis of IP Network Usage and Behavior. Communications Magazine, IEEE, 38 (5), 144-151.

DREYER, N. & DE FÁTIMA, H. (2009). Data mining: a literatura review. Acta Paul Enferm, 22 (5), 680-690.

HAND D. & MANNILA H. & SMYTH P. (2001). Principles of Data Mining. MIT Press, Cambridge.

HERNÁNDEZ, J. & RAMÍREZ, M.J. & FERRI, C. (2004). Introducción a la Minería de Datos. Pearson Prentice Hall.

KIMBALL R. (2008). The Data Warehouse Lifecycle Toolkit. John Wiley & Sons, Nueva York.

KOTSIANTIS, S.B. (2011). An incremental ensemble of classifiers. Artificial Intelligence Review, 36 (4), 249-266.

UR-RAHMAN, N. & HARDING J.A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Systems with Applications, 39 (5), 4729-4739.

VERCELLIS C. (2009). Business Intelligence: Data Mining and Optimization for Decision Making, John Willey & Sons.

VITT E. & LUCKEVICH M. & MISNER S. (2003). Business Intelligence: Técnicas de análisis para la toma de decisiones estratégicas. McGraw-Hill, Madrid, España.

WITTEN H. & FRANK E. (2005). Data Mining: Practical Machine Learning Tools and Techniques Morgan Kaufman Publishers, Inc., San Francisco, California.