Aplicaciones empresariales de minería de datos usando software libre

Giovanni Francisco Acosta Henríquez
Máster en Dirección Estratégica de Ingeniería de Software
Docente investigador, Facultad de Ingeniería y Arquitectura
Universidad Católica de El Salvador, El Salvador
ghenriquez@catolica.edu.sv

Fecha de recepción: 18-12-2015 / **Fecha de aceptación:** 21-01-2016

Resumen

La minería de datos se ha transformado en la base de la toma de decisiones en las grandes empresas. Sin embargo, su utilización en las medianas y pequeñas empresas ha sido mínimo en el país. Entre las causas de este fenómeno están el bajo número de expertos en el área, los altos costos de asesoría y del software; y el factor mayor, la falta de conocimiento de los medianos y pequeños empresarios en la existencia de la minería de datos.

El artículo describe la metodología utilizada para determinar la herramienta de software libre de mejor aplicabilidad dentro de las empresas, y la elaboración de una guía metodológica para implementar aplicaciones empresariales de minería de datos, usando el software libre WEKA, como herramienta de apoyo informático para pequeñas y medianas empresas del país.

Palabras clave: minería de datos, WEKA, CRISP-DM, medianas y pequeñas empresas, software, asesoría

Abstract

Data mining has become the basis of decision making in large enterprises. However, their use in medium and small businesses has been minimal in the country. Among the causes of this phenomenon, there are the low number of experts in this field, the high costs of consulting and software, and the biggest factor, lack of knowledge of the medium and small entrepreneurs in the existence of data mining.

The article describes the methodology used to determine the free software tool better applicability within companies, and the development of a methodological guide for implementing business applications data mining, using free software WEKA as a tool for computer support for small and medium enterprises in the country.

Key words: data mining, WEKA , CRISP -DM , medium and small businesses , software , consulting

1. Introducción

El rápido desarrollo de las tecnologías de información y comunicación (TIC)[1], entre estos la innovación de los sistemas digitales, representan una revolución digital, que ha cambiado fundamentalmente la manera en que las personas piensan, actúan, comunican y trabajan. La revolución digital ha forjado nuevas modalidades de crear conocimiento y transmitir información. Ha reestructurado la forma en que los países hacen negocios y rigen su economía.

Debido al avance que existe en las TIC, las empresas se han tenido que enfrentar a nuevos desafíos que les permita analizar, descubrir y entender más allá de lo que las herramientas tradicionales reportan sobre información, al mismo tiempo que durante los últimos años el gran crecimiento de las aplicaciones disponibles en Internet han sido parte importante en las decisiones de negocio de las empresas.

De acuerdo con un estudio realizado por CISCO (2012), entre el 2011 y el 2016, la cantidad de tráfico de datos móviles crecerá a una tasa anual de 78%, así como el número de dispositivos móviles conectados a Internet excederá el número de habitantes en el planeta. Esta revolución digital ha provocado que los seres humanos creen y almacenen información constantemente y cada vez más en cantidades astronómicas. Esta generación masiva de datos se puede encontrar en diversas industrias: las empresas mantienen grandes cantidades de datos transaccionales, reúnen información sobre sus clientes, proveedores, operaciones, etc., de la misma manera sucede con el sector público. Si a todo esto se agrega la información generada diariamente en transacciones financieras en línea por dispositivos móviles, las redes sociales, ubicación geográfica, etc., lo que se estaría generando, según IBM, son alrededor de 2.5 quintillones de bytes diariamente en el mundo.

Muchas de las micro y pequeñas empresas del país, no se encuentran preparadas para manejar el volumen de datos que poseen hoy en día, por lo que necesitan herramientas con una mayor capacidad de análisis que les ayuden en la toma de decisiones. En ocasiones, además de un acceso rápido y eficiente a los datos, requieren capacidades analíticas que ayuden a sacar el máximo provecho a la información disponible.

Para hacer frente a la presente problemática, se propondrá a las empresas el uso de herramientas y técnicas de minería de datos sobre software libre, dado que la Minería de datos es una herramienta fundamental para analizarlos y explotarlos de forma eficaz, llegando al conocimiento a partir de los datos (MSDN, Microsoft, 2014).

2. Metodología

La investigación fue de tipo experimental, debido a que se exploraron, analizaron y probaron diferentes herramientas de software libre para minería de datos con el objeto de catalogar

1. A partir de esta aclaración, el autor también se referirá a ellas mediante estas siglas.

las aplicaciones empresariales de minería de datos, según las unidades económicas por sector que posee el país. Esto llevó a un nivel aplicativo que implicó la comparación de las herramientas estudiadas, y a un nivel explicativo que involucró la documentación e identificación de los beneficios, que obtendrán las empresas y el país con la implementación de las aplicaciones de minería de datos.

Con el objeto de dar cumplimiento a los objetivos de la investigación se empleó la siguiente metodología:

- Definición del problema: se describió el problema, objetivos, justificación y alcance.
- Estado del arte: se estudió el origen y evolución de la minería de datos, usos, características, técnicas, herramientas y metodologías empleadas en los proyectos empresariales de minería de datos.
- Catalogación de aplicaciones: se identificaron aplicaciones empresariales de minería de datos de interés según las unidades económicas por sector que posee el país.
- Exploración de herramientas: se exploró el software libre disponible para minería de datos, y se realizó una comparativa sobre algunas métricas de calidad de software que poseen las herramientas más populares.
- Documentación: se documentó el proceso a seguir para implementar proyectos empresariales de minería de datos y las herramientas de software libre disponibles.
- Presentación de resultados: se elaboró una guía para implementar aplicaciones empresariales de minería de datos usando software libre, como herramienta de apoyo informático para pequeñas y medianas empresas del país, conclusiones y recomendaciones.

Con el desarrollo de la metodología descrita anteriormente, se logró conocer el estado del arte, proponer una metodología para la implementación de proyectos de minería de datos; comparar y sugerir herramientas de software libre para minería de datos y proporcionar una guía para la implementación de aplicaciones empresariales de minería de datos usando software libre.

3. Resultados

a. Catalogación de aplicaciones por sector

Según el directorio de unidades económicas 2011-2012, elaborado por la Dirección General de Estadística y Censos (DIGESTYC) del Ministerio de Economía (MINEC, s.f.), las unidades económicas del país se clasifican en los sectores: agroindustria, comercio, construcción, electricidad, industria, minas y canteras; servicios y transporte, con un total de 161,934 unidades económicas, las cuales se dividen así : comercio, 59.4%; servicios, 27.6%; industria,11.5%; transporte, 1.2%; construcción, 0.2%; y otros (agroindustria, electricidad y minas y canteras), 0.1%.

Para esta investigación se trabajó únicamente con los sectores que poseen el mayor núme-

ro de unidades económicas en el país, como: comercio, servicios, industria y transporte, realizando la siguiente catalogación de aplicaciones de minería de datos que pueden ser de interés para cada sector, mostrados en la siguiente figura:

Sector	Aplicación de minería de datos
Comercio	Análisis de mercado, distribución y comercio en general: Análisis de canasta de compras Evaluación de campañas publicitarias Análisis de fidelidad de los clientes. Reducción de fuga Segmentación de clientes Estimación de stocks, de costos y de ventas Aplicaciones financieras y banca: Obtención de patrones de uso fraudulento de tarjetas de crédito Determinación del gasto en tarjeta de crédito por grupos Cálculo de correlaciones entre indicadores financieros Identificación de reglas de mercado de valores a partir de datos históricos Análisis de riesgos en créditos Aplicaciones de seguros: Determinación de los clientes que podrían ser potencialmente caros Predicción de qué clientes contratan nuevas pólizas Identificación de patrones de comportamiento para clientes con riesgo Identificación de comportamiento fraudulento

Sector	Aplicación de minería de datos
Servicios	Medicina: Identificación de patologías. Diagnóstico de enfermedades Detección de pacientes con riesgo de sufrir una patología concreta Gestión hospitalaria y asistencial para el mejor uso de recursos Análisis de procedimientos médicos Recomendación priorizada de fármacos para una misma patología Tratamiento de imágenes médicas Telecomunicaciones: Establecimiento de patrones de llamadas Modelos de carga en redes Detección de fraude Energía: Estimación del modelo de demanda energética de cada cliente Predicción de anomalías Mantenimiento preventivo Ofertas personalizadas Biología, bioingeniería y otras ciencias: Análisis de secuencias de genes Análisis de secuencias de proteínas Predecir si un compuesto químico causa cáncer Clasificación de cuerpos celestes Predicción de recorrido y distribución de inundaciones Modelos de calidad de agua, indicadores ecológicos Turismo: Determinar características socioeconómicas de los turistas Identificar patrones de reservas Predicción de demanda de oferta turística Educación Selección o captación de estudiantes Detección de abandono y de fracaso Estimación del tiempo de estancia en la institución Web y correo: Análisis de comportamiento de los usuarios Detección de fraude en el comercio electrónico Análisis de log en un servidor web Clasificación y distribución automática de correo Detección de correo spam Análisis de empleo del tiempo en la web

Sector	Aplicación de minería de datos
Industria	Procesos industriales: Extracción de modelos sobre comportamiento de compuestos Detección de piezas con fallas. Modelo de calidad Predicción de fallos y accidentes Estimación de composiciones óptimas en mezclas Extracción de modelos de costos Extracción de modelos de producción
Transporte	Análisis transporte público y privado: Análisis del comportamiento de los pasajeros y las modalidades de viaje Predicción de rutas óptimas y tiempos de viaje Predecir la demanda de servicio de transporte (aéreo, terrestre, etc.) Predecir el impacto de un proyecto de transporte Predecir el estado del transito Realizar reportes operacionales con datos históricos Realizar predicciones creando modelos Realizar ofertas según demanda real de los pasajeros

Figura 1. Catalogación de aplicaciones de minería de datos por sector económico del país.

Todas estas aplicaciones muestran que la minería de datos puede ayudar a mejorar la toma de decisiones en diferentes actividades de las empresas, permitiendo conocer mejor el entorno en el que se desenvuelven.

b. Comparativa de herramientas de minería de datos

Actualmente, existen muchas herramientas enfocadas a la minería de datos que utilizan un gran número de técnicas para la solución de múltiples problemas en distintos campos. Entre las herramientas de software libre para minería de datos mejor evaluadas en cuanto a las métricas de calidad de software, según ISO (2014), se encuentran: Orange, RapidMiner, WEKA, JHepWork y KNIME.

- **Orange:** es un programa informático para realizar minería de datos y análisis predictivo, desarrollado en la facultad de informática de la Universidad de Liubliana. Consta de una serie de componentes desarrollados en C++ que implementan algoritmos de minería de datos, así como operaciones de preprocesamiento y representación gráfica de datos. Los componentes de Orange pueden ser manipulados desde programas desarrollados en Python o a través de un entorno gráfico y se distribuye bajo licencia GPL (Licencia Pública General).

- **RapidMiner:** (anteriormente, YALE, Yet Another Learning Environment) es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación, educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales. La versión inicial fue desarrollada por el departamento de inteligencia artificial de la Universidad de Dortmund en 2001, y se distribuye bajo licencia AGPL (Licencia Pública General de Affero).
- **WEKA:** (Waikato Environment for Knowledge Analysis, en español «Entorno para Análisis del Conocimiento de la Universidad de Waikato»), es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato; es software libre distribuido bajo la licencia GNU-GPL (Licencia Pública General de GNU).
- **JHepWork:** es un marco de trabajo gratuito de análisis de datos para científicos, ingenieros y estudiantes escrito en Java. El programa está diseñado para áreas de gráficas científicas interactivas en 2D y 3D, y contiene bibliotecas numéricas científicas implementadas en Java para funciones matemáticas, números aleatorios, análisis estadístico, ajuste de curvas de regresión y otras actividades de minería de datos.
- **KNIME:** (Konstanz Information Miner) es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. Está construido bajo la plataforma Eclipse; fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania.

Para poder determinar la solución de software libre de minería de datos de mejor aplicabilidad, se realizó una tabla comparativa de las herramientas: Orange, WEKA, RapidMiner, JHepWork y KNIME, utilizando las buenas prácticas descritas en los factores de calidad del software de McCall, (Software Quality Attributes, s.f.); y la norma ISO-9126, se evaluó un subconjunto de los factores que integran el punto de vista externo del software; es decir, lo que le interesa al usuario, utilizando como criterios: alto (10 puntos), medio (5 puntos) y bajo (0 puntos), obteniendo los resultados que se presentan en la siguiente tabla:

Tabla 1. Comparativa de las herramientas en software libre para minería de datos

Software	Flexibilidad	Portabilidad	Interoperabilidad	Funcionalidad	Documentación	Facilidad de aprender	Facilidad de instalación	Facilidad de configurar	Actualizaciones	Soporte técnico	Total de puntos
Orange	Alto	Alto	Alto	Alto	Alto	Medio	Alto	Alto	Alto	Alto	95
WEKA	Alto	Alto	Alto	Alto	Alto	Alto	Alto	Alto	Alto	Alto	100
RapidMiner	Alto	Alto	Alto	Medio	Alto	Alto	Alto	Alto	Alto	Alto	95
JHepWork	Alto	Alto	Alto	Medio	Medio	Medio	Medio	Alto	Bajo	Bajo	60
KNIME	Alto	Alto	Alto	Medio	Medio	Medio	Alto	Alto	Medio	Alto	80

Por lo anterior, la herramienta de mejor aplicabilidad para el desarrollo de aplicaciones empresariales de minería de datos, según el subconjunto de factores evaluados, es el software libre WEKA, seguido de las herramientas Orange y RapidMiner.

c. Metodología para implementar proyectos de minería de datos

Existen diversos modelos de proceso que han sido propuestos para el desarrollo de proyectos de minería de datos tales como SEMMA (Sample, Explore, Modify, Model, Assess), DMAMC (Define, Measure, Analyze, Improve, Control) o CRISP-DM (Cross Industry Standard Process for Data Mining). Sin embargo, el modelo principalmente utilizado en los ambientes académico e industrial es el modelo CRISP-DM.

CRISP–DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos, como se puede constatar en la gráfica presentada en la siguiente figura. Esta gráfica, publicada el año 2007 y 2014 por kdnuggets.com, muestra el grado de uso de las principales guías de desarrollo de proyectos de minería de datos. (Ver figura 2).

CRISP-DM divide el proceso de minería de datos en seis fases principales:

1. **Comprensión del negocio:** esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición del problema de minería de datos, y un plan preliminar diseñado para alcanzar los objetivos.
2. **Comprensión de Datos:** esta fase comienza con una colección inicial de datos y procesos con actividades con el objetivo de familiarizarse con los datos, identificar la calidad de los problemas, para descubrir las primeras señales dentro de los datos y detectar temas interesantes para poder formular hipótesis de información oculta.

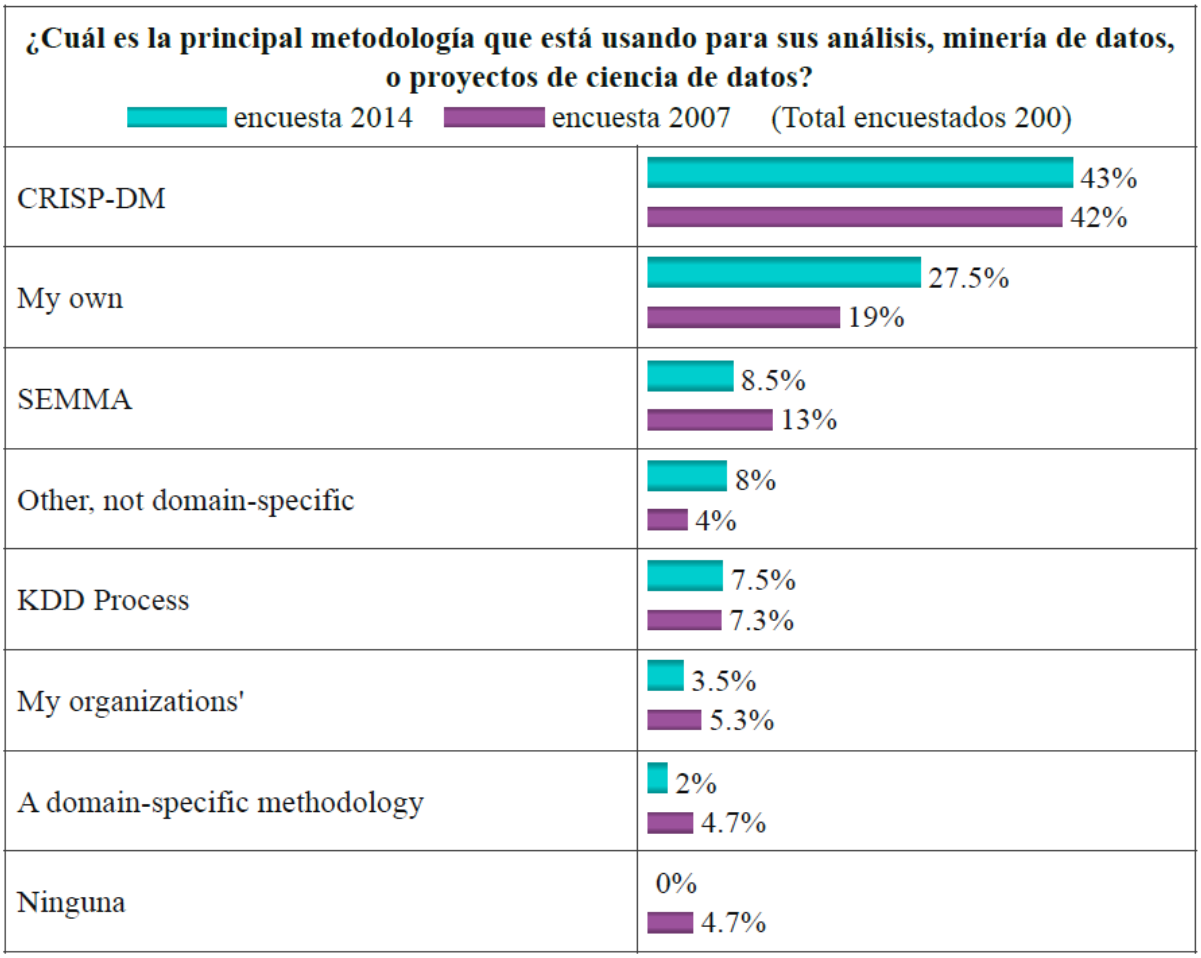

Figura 2. Resultado de encuesta sobre uso de metodología para minería de datos.

Fuente: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-scienceprojects.html, 2014.

3. **Preparación de datos:** esta fase cubre todas las actividades para construir el conjunto de datos. Estas tareas son ejecutadas en múltiples oportunidades y sin orden. Las tareas incluyen selección y transformación de tablas, registros y atributos y limpieza de datos para las herramientas de modelado.
4. **Modelado:** en esta fase se seleccionan y aplican varias técnicas de modelado y se calibran los parámetros para obtener óptimos resultados. Hay varias técnicas que tienen requerimientos específicos para la forma de los datos, por lo que frecuentemente es necesario volver a la fase de preparación de datos.
5. **Evaluación:** en esta etapa del proyecto se ha construido un modelo (o modelos) que parece tener gran calidad, desde una perspectiva de análisis de datos.
6. **Despliegue:** esta fase depende de los requerimientos, logrando ser simple como

la generación de un reporte o compleja como la implementación de un proceso de explotación de información que atraviese a toda la organización.

En muchos casos, será el cliente, no el analista de datos, que llevará a cabo los pasos de implementación. Incluso si el analista despliega el modelo es importante para el cliente para entender por adelantado las acciones que deberán llevarse a cabo, con el fin de hacer realidad el uso de los modelos creados.

d. Guía para implementar aplicaciones empresariales de minería de datos

Con base en los resultados de la comparativa de las herramientas de software libre para minería de datos realizada para la presente investigación, se seleccionó el software libre WEKA, por ser la de mejor aplicabilidad. Este software fue la herramienta utilizada en la elaboración de la guía para implementar aplicaciones empresariales de minería de datos desarrollada en la presente investigación.

WEKA soporta varias tareas estándar de minería de datos, especialmente, preprocesamiento de datos, clustering, clasificación, regresión, visualización y selección. Todas las técnicas de WEKA se fundamentan en la asunción de que los datos están disponibles en un fichero plano (Attribute-Relation File Format, ARFF); o una relación, en la que cada registro de datos está descrito por un número fijo de atributos.

WEKA, también proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (Java Database Connectivity) y puede procesar el resultado devuelto por una consulta hecha a la base de datos. No puede realizar minería de datos multi-relacional, pero existen aplicaciones que pueden convertir una colección de tablas relacionadas de una base de datos en una única tabla que ya puede ser procesada con WEKA.

Entre las principales ventajas de WEKA, están: la libre disponibilidad bajo la licencia pública general de GNU, es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma. Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado, y es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.

De acuerdo a Morate (s.f.), WEKA posee los siguientes componentes principales:

Simple CLI: es una abreviación de Simple Client. Esta interfaz proporciona una consola para poder introducir comandos. A pesar de ser en apariencia muy simple es extremadamente potente porque permite realizar cualquier operación soportada por WEKA de forma directa. No obstante, es muy complicada de manejar, ya que es necesario un conocimiento completo de la aplicación.

Explorer: el modo Explorador es el más usado y más descriptivo; permite realizar

operaciones sobre un sólo archivo de datos, y permite realizar tareas de: preprocesado de los datos y aplicación de filtros, clasificación, Clustering, búsqueda de asociaciones, selección de atributos y visualización de datos.

Experimenter: esta opción permite definir experimentos más complejos, con objeto de ejecutar uno o varios algoritmos sobre uno o varios conjuntos de datos de entrada, y comparar estadísticamente los resultados.

KnowledgeFlow: esta opción permite llevar a cabo las mismas operaciones de Explorer, con una configuración totalmente gráfica, inspirada en herramientas de tipo data-flow, para seleccionar componentes y conectarlos en un proyecto de minería de datos. Desde que se cargan los datos, se aplican algoritmos de tratamiento y análisis, hasta el tipo de evaluación deseada.

Aprovechando las bondades de WEKA como herramienta de software libre para minería datos, se ha elaborado una guía para implementar aplicaciones empresariales de minería de datos, con el siguiente contenido: introducción, definición y conceptos principales de minería de datos, aplicaciones de minería de datos, algoritmos, metodología CRISP-DM, uso de WEKA, requisitos de instalación, descarga, instalación, ejecución, Tutorial1: entorno de trabajo de WEKA, Tutorial2: problemas de regresión, Tutorial3: problemas de clasificación, Tutorial4: problemas de agrupación, Tutorial5: crear un ARFF desde datos de EXCEL y Tutorial6: conectar WEKA a la base de datos MySQL y Referencias.

e. Algunos ejemplos de implementación de aplicaciones de minería de datos

El dueño de una pequeña empresa familiar se plantea la siguiente pregunta ¿De cuánto serán los ingresos de la empresa para el año 2025?, utilizando minería de datos se puede pronosticar la respuesta. Para ello será necesario crear un pequeño conjunto de datos, con los atributos: año e ingreso anual; dicho conjunto de datos puede ser elaborado en un sencillo editor como el bloc de notas con extensión .arff, para luego ser procesado con el software de minería de datos Weka.

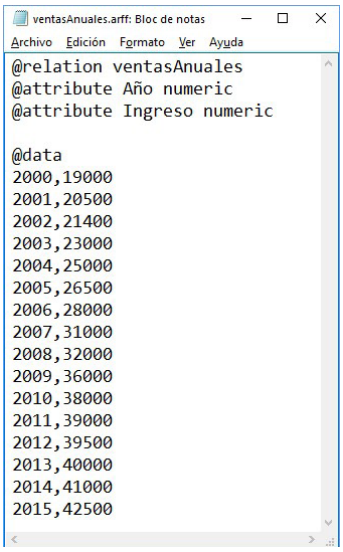

Figura 3. Conjunto de datos (dataset) elaborado en el bloc de notas.

En la herramienta Explorer de Weka, se carga el conjunto de datos creado anteriormente, obteniendo la siguiente vista (Figura 4).

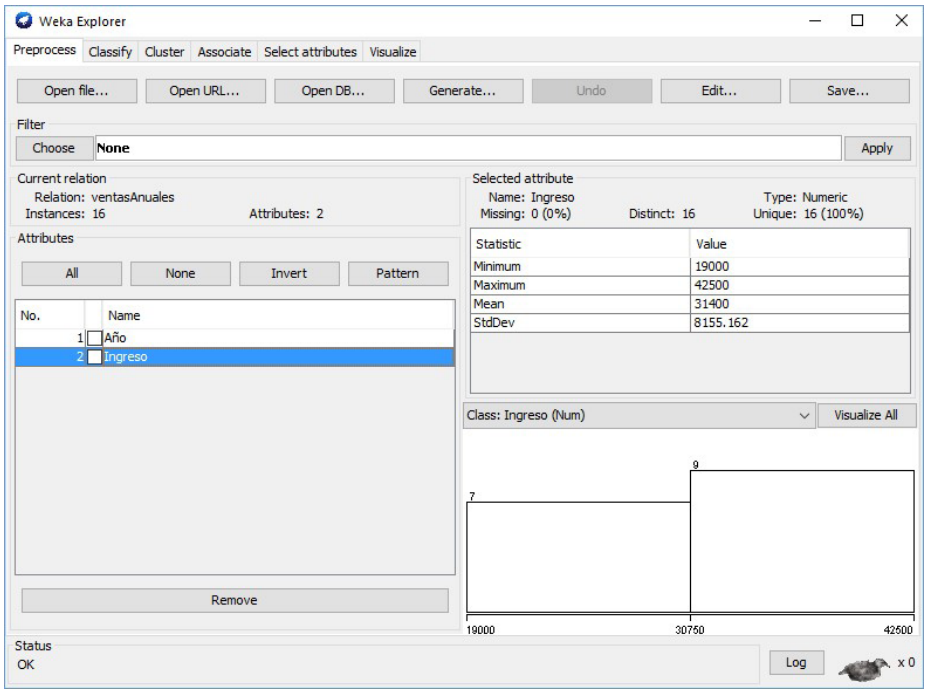

Figura 4. Conjunto de datos cargado en la herramienta Explorer de Weka.

Para responder ¿de cuánto serán los ingresos de la empresa para el año 2025?, se debe aplicar una regresión lineal. Para ello se selecciona la ficha Clasificación (Classify), luego el filtro de Regresión lineal (LinearRegression) en la carpeta de funciones y se inicia el proceso (Figura 5)

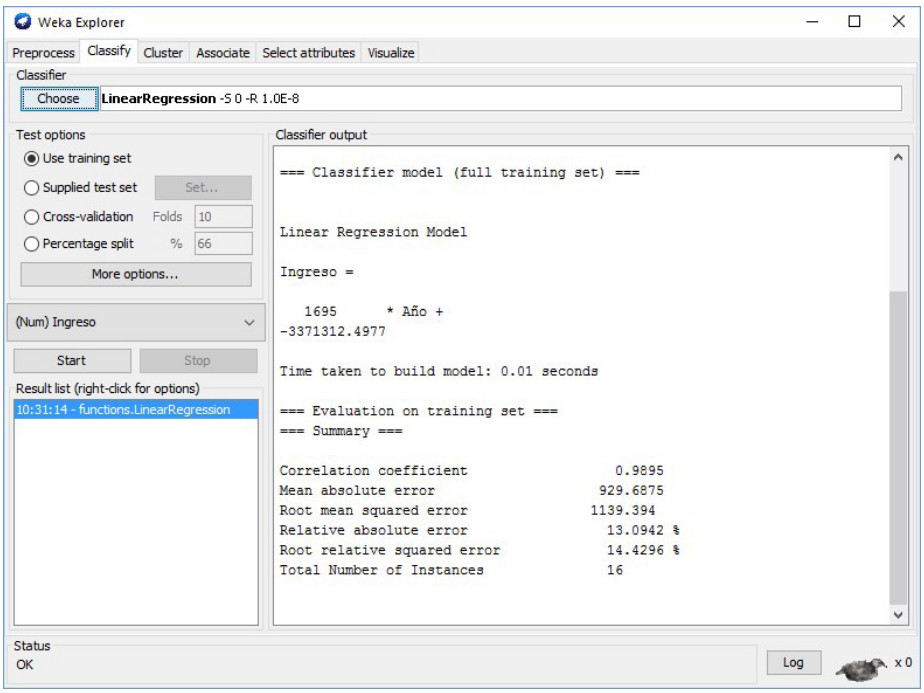

Figura 5. Aplicación de filtro de regresión lineal en WEKA.

Aplicada la regresión, proporciona datos importantes, como el coeficiente de correlación entre el año y el ingreso obtenido que es del 98% con un error relativo del 13.1% y el modelo de regresión aplicado, provee la siguiente fórmula para pronosticar el ingreso para un determinado año: Ingreso = 1695 * año - 3371312.4977. Calculando el ingreso para el año 2025, sustituyendo en la fórmula del modelo, se tiene: Ingreso = 1695 * 2015 - 3371312.4977, que equivale a $61,062.50 para el año 2025.

Otro caso a resolver será para una empresa que desea conocer el perfil de los clientes que compraron productos o un producto en específico para identificarlos, y realizar un marketing más efectivo. El conjunto de datos a evaluar posee los siguientes atributos: sexo, estado civil, número de hijos; si posee casa y qué producto compró entre dos productos a evaluar para este ejemplo. El conjunto de datos a trabajar posee 100 instancias (registros o filas de datos).

Se carga el conjunto de datos a la herramienta Explorer de WEKA. Con el atributo sexo seleccionado, se puede observar que 49 clientes son del sexo masculino y 51 del sexo femenino, de igual forma se podría clasificar los clientes con el resto de atributos (Figura 6)

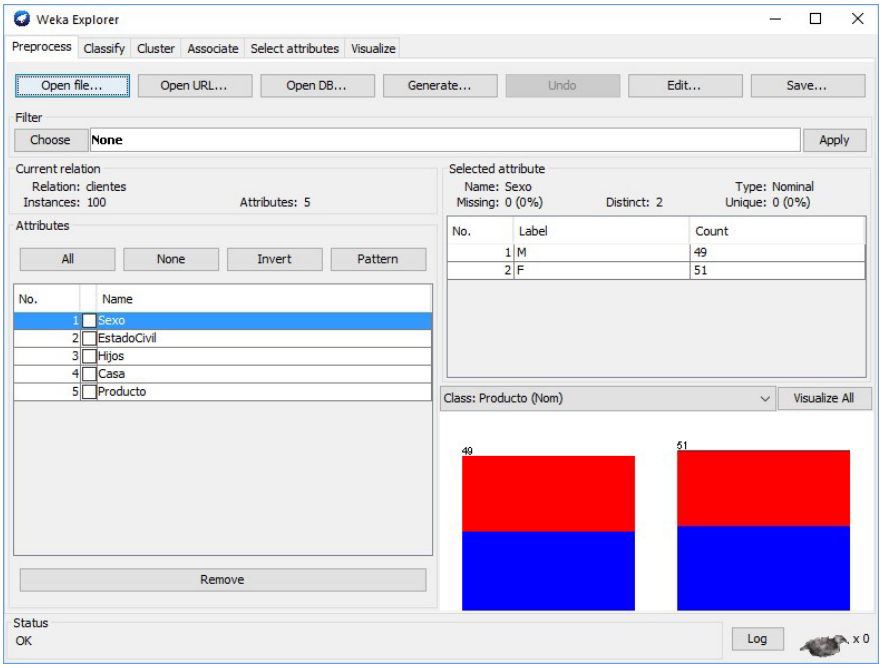

Figura 6. Vista de la herramienta Explorer con clasificación de clientes por sexo.

Para este tipo de problema se utiliza el algoritmo de árbol de decisión, que constituye una de las técnicas de toma de decisiones más empleada en minería de datos. Para construirlo se selecciona la ficha Clasificación (Classify), de la herramienta Explorer; se selecciona el atributo «producto» y la opción «árbol J48», es un algoritmo para la generación del árbol de decisión, elaborado en lenguaje Java. Para visualizar el árbol de decisión de manera gráfica, se hace clic con el botón secundario del mouse sobre la lista del resultado, y se escoge la opción Visualizar Árbol (Visualze tree) (Figura 7)

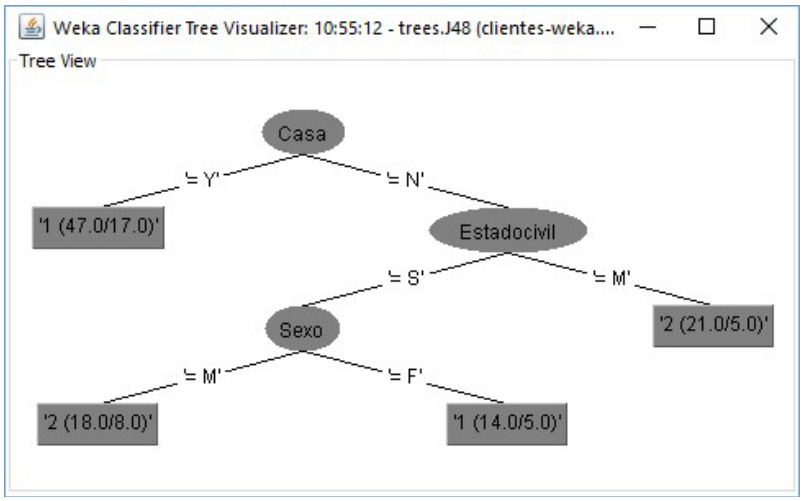

Figura 7. Árbol de decisión para conocer el perfil de clientes.

Para responder la interrogante sobre el perfil de los clientes que compraron productos o un producto en específico, el árbol de decisión se interpreta de la siguiente manera: de los clientes que poseen casa propia, 17 de 47 compraron el producto 1. De los clientes que no poseen casa propia y estado civil casado, 5 de 21 compraron el producto 2. De los mismos clientes, pero solteros y sexo masculino, 8 de 18 compraron el producto 2; y de los mismos clientes, pero de sexo femenino, 5 de 14 compraron el producto 1. A partir de estas reglas se pueden tomar decisiones sobre el tipo de promociones que se pueden dirigir a un determinado perfil de cliente.

f. Presupuesto de implementación

Para la implementación de las aplicaciones empresariales de minería de datos utilizando el software libre WEKA, se requieren los componentes detallados en la Tabla 2.

Tabla 2. Implementación de aplicaciones empresariales con WEKA

No	Concepto	Precio
1	Computadora con sistema operativo Windows, Linux o Mac	$ 1,000
2	WEKA (ver. 3.6) para Windows, Linux o Mac	$ 0
3	OpenOffice Calc para Windows, Linux o Mac (ver. 4.1.2)	$ 0
4	MySQL Community Server 5.7.10 para Windows, Linux o Mac	$ 0
5	Guía para implementar aplicaciones de minería de datos	*
		$ 1,000

(1) La computadora puede ser de escritorio o laptop; el tipo de sistema puede ser 32 o 64 bits; la velocidad de procesamiento y memoria incidirá en el rendimiento de la aplicación.

(2) WEKA es software libre disponible para Windows, Linux o Mac en la siguiente URL: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

(3) OpenOffice Calc es software libre disponible para Windows Linux o Mac en la siguiente URL: https://www.openoffice.org/es/descargar/index.html

(4) MySQL Community Server es software libre disponible para Windows Linux o Mac en la siguiente URL: http://dev.mysql.com/downloads/mysql

(5) El otorgamiento de uso de la guía para implementar aplicaciones de minería de datos usando el software libre WEKA será concedido por la Universidad Católica de El Salvador.

4. Discusión

El estudio presentado en este proyecto de investigación permitió catalogar las aplicaciones de minería de datos por sector económico que posee el país, realizar una comparativa de las herramientas de software libre de esta rama y conocer sobre la metodología más utilizada para el desarrollo de proyectos empresariales de esta índole; llegando a las siguientes conclusiones y recomendaciones:

La minería de datos es un tema que no ha sido aprovechado por las micro y pequeñas empresas del país, debido al desconocimiento de los beneficios que este provee y a la falta de personal informático en estas empresas, capacitado para implementar aplicaciones de este tipo a bajo costo.

En la actualidad, existen diversos paquetes de software que implementan diversas técnicas, tanto comerciales (Oracle Data Mining, Clementine, SQL Server etc.), y otras alternativas de software libre (WEKA, Orange, RapidMiner, KNIME, etc.). En el caso de las comerciales son muy costosas para ser implementadas por las micro y pequeñas empresas del país, siendo entonces una buena alternativa el software libre para resolver los problemas

que requieren el uso de minería de datos, ya que no poseen ningún costo por licencia.

CRISP-DM es la guía metodológica más utilizada en el desarrollo de proyectos de minería de datos, la cual aportará a las empresas las fases y actividades a realizar, al incursionar en la implementación de aplicaciones de este tipo.

En este trabajo se ha expuesto la herramienta de software libre WEKA como la mejor alternativa para la implementación de aplicaciones empresariales de minería de datos, ya que es un herramienta de fácil aprendizaje y muy eficiente a la hora de aplicar diversas técnicas dentro de esta rama.

La guía de implementación de aplicaciones de minería de datos, utilizando el software libre WEKA, elaborada como producto de la presente investigación, se espera que sirva como herramienta de apoyo para las micro y pequeñas empresas que desean iniciar en el tema, pero deberá ser ejecutada por personal técnico con conocimientos de bases de datos.

En resumen, la minería de datos se presenta como una tecnología innovadora, que ofrece una serie de beneficios, como: el ahorro de grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios, facilitando la toma de decisiones en un negocio.

5. Referencias

Azevedo, A., y Santos, M. F. (2015). Integration of data mining in business intelligence systems. USA: IGI Global.

Azevedo, A., y Santos M. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. IADIS European Conference Data Mining. USA: IGI Global.

Baker, E. (2010). Handbook of Educational Data Mining. USA: CRC Press.

Barranco F., R. (2012). ¿Qué es Big data? IBM Bluemix. Recuperado de: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Bramer, M. (2013). Principles of Data Mining. New York: Springer.

Coronel, M., R. (2011). Bases de datos, diseño, implementación y administración. México: Cengage Learning.

García H., J. y Molina L., J. (2012). Técnicas de análisis de datos Aplicaciones prácticas utilizando Microsoft Excel y weka. Madrid: Universidad Carlos III de Madrid.

Han, J., Kamber, M. y Pei, J. (2012). Data Mining. Concepts and Techniques. USA: Morgan Kaufmann.

Hernández O., J., Ramirez Q., M. J., y Ferri R., C. (2004). Introducción a la Minería de Datos. Madrid: Pearson.

Hofmann, M. y Klinkenberg, R. (2014). RapidMiner data mining use cases and business analytics applications. Florida: CRC Press.

ITU (s.f.) Cumbre Mundial sobre la Sociedad de la Información (UN–UIT) 2003- 2005. Recuperado de: http://www.itu.int/wsis/basic/faqs.asp?lang=es

Leskivec, J., Rajaraman, A. y Ullman, J. (2014). Mining of Massive Datasets. United Kingdom: Cambridge Univerity Press.

Marcano, Y. y Talavera, R. (s.f.) Minería de Datos Como soporte a la toma de decisiones empresariales. Recuperado de http://www.scielo.org.ve/scielo.php?pid=S101215872007000100008&script=sci_arttext

Marquéz P., M. (2014). Minería de datos a través de ejemplos. Madrid: RC Libros.

Pérez L., C., Santín G., D. (2007). Minería de Datos. Técnicas y Herramientas. Madrid: Thomson.

Sierra A., B. (2006). Aprendizaje automático: conceptos básicos y avanzados: aspectos prácticos utilizando el software WEKA. España: PRENTICE-HALL.

Soman, K. P., Diwakar, S., y Ajay, V. (2006) Insight into data mining theory and practice. India: Prentice – Hall.

Sumathi, S. y Sivanandam, S. (2006). Introduction to Data Mining and its Applications. Berlin: Springer.

Suh, S. (2012). Practical applications of data mining. Canada: Jones & Barlett Learning.

Shmueli, G., Patel, N. y Bruce, P. (2010). Data mining for business intelligence, concepts, techniques, and applications in Microsoft Office Excel. USA: John Wiley.

Tan, P., Steinbach, M. y Kumar, V. (2006). Introduction to Data Mining. Addison-Wesley.

Witten, I., Frank, E. y Hall, M. (2011). Data Mining. Practical machine learning tools and techniques. USA: Morgan Kaufmann.

Witten, I., Frank, E. y Hall, M. (2011). Weka: Practical machine learning tools and techniques with java implementations. USA: Morgan Kaufmann.