**Universidad Nacional Mayor de San Marcos**

**Facultad de Ingeniería de Sistemas e Informática**

**E.A.P. de Ingeniería de Sistemas**



**Docente:** Dr. Vega Huerta, Hugo

**Alumna:** Santillán Rodríguez, Edgar Hugo

**Curso:** Proyecto deTesis II

**Proyecto:** SISTEMA BASADO EN MINERÍA DE DATOS PARA LA DETECCIÓN DE PATRONES COMERCIALES Y DE VENTA EN UNA EMPRESA COMERCIAL

# **RESUMEN**

**SISTEMA BASADO EN MINERÍA DE DATOS PARA LA DETECCIÓN DE PATRONES COMERCIALES Y DE VENTA EN UNA EMPRESA COMERCIAL**

 **UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS**

**Universidad del Perú, Decana de América**

**FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA**

**ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA DE SISTEMAS**

 

**SISTEMA BASADO EN MINERÍA DE DATOS PARA LA DETECCIÓN DE PATRONES COMERCIALES Y DE VENTA EN UNA EMPRESA COMERCIAL**

**Tesis para optar por el título profesional de:**

**INGENIERO DE SISTEMAS**

**Presentada por:**

**Santillán Rodríguez, Edgar Hugo**

Ciudad Universitaria, 10 de Mayo de 2018

**Santillán Rodríguez, Edgar Hugo**

**SISTEMA BASADO EN MINERÍA DE DATOS PARA LA DETECCIÓN DE PATRONES COMERCIALES Y DE VENTA EN UNA EMPRESA COMERCIAL**

“Tesis presentada a la Universidad Nacional Mayor de San Marcos, Lima, Perú, para obtener el Título de Ingeniero de Sistemas”

Orientador: Vega Huerta Hugo Froilán

UNMSM – LIMA

Mayo 2018

.

© Santillán Rodríguez, 2018.

Todos los derechos reservados.

Este trabajo está dedicado a mi familia por su constante apoyo hacia mi persona en todo momento hasta la actualidad.

**AGRADECIMIENTOS**

A mi tutor Vega Huerta Hugo Froilán, por su orientación, paciencia y confianza en la realización de esta obra.

A los profesores de la UNMSM, por ser modelos a seguir y parte importante en nuestra formación académica.

**SISTEMA BASADO EN MINERÍA DE DATOS PARA LA DETECCIÓN DE PATRONES COMERCIALES Y DE VENTA EN UNA EMPRESA COMERCIAL**

**RESUMEN**

El proyecto que se presenta en este documento tiene como objetivo la identificación de patrones comerciales y de venta en una Empresa comercial que podría predecir el comportamiento de la venta de algún producto por temporada; se realiza lo antes mencionado para que la Empresa comercial pueda tener un mayor conocimiento sobre las ventas que puede lograr a partir de información recopilada en los años anteriores, y con ello elaborar políticas que permitan sacar el máximo provecho al área comercial y pueda optimizar recursos en cierta medida.

Para lograr dichos objetivos se han de usar herramientas de software, herramientas de análisis de datos y metodologías que permitan realizar lo mencionado de acuerdo a lo esperado para ello se han de seleccionar herramientas estandarizadas y aceptadas internacionalmente en sus respectivos campos, en el caso de las herramientas de software, han de ser seleccionadas de acuerdo a la comparación de criterios y de acuerdo a los requerimientos que se necesiten , el mismo concepto aplica para las herramientas de análisis de datos.

En conclusión, el proyecto se lleva a cabo con éxito previniendo los efectos negativos o eventos inoportunos que puedan generarse durante su ejecución, además de mantener los objetivos generales y específicos con sus respectivos métodos para mantener la idea clara y concisa de lo que se pretende realizar desde el inicio de este.

**Palabras claves:**  Data Mining, KDD – Descubrimiento en base de datos, Predicción, Limpieza de datos.

# **INTRODUCCIÓN**

El propósito de esta investigación tiene como objetivo principal la implementación de un sistema basado en Minería de Datos, aplicando la metodología KDD para poder generar patrones comerciales y de ventas en una Empresa Comercial.

En la actualidad las empresas generan muchos datos que no están clasificados, ordenados ni mucho menos estructurados, razón por la cual no se esta aprovechando toda la información y el conocimiento que este puede brindar para la toma de decisiones, los responsables de tomar las decisiones y elaborar políticas en la empresa necesitan la mayor cantidad de información posible al momento de tomar una decisión.

Hoy en día la información que puede generar un organismo crece de forma exponencial, tal es la magnitud del crecimiento que se están buscando formas de optimizar el almacenamiento de estos, sin embargo, si estos datos pudieran organizarse, clasificarse y estructurarse podrían generar información valiosa para una empresa que le permitiría tomar decisiones a futuro en base a la información recolectada durante años, es por ello la razón de este proyecto que busca mejorar la toma de decisiones de los responsables y generar políticas que permitan optimizar sus ventas e intereses comerciales.

Contenido

[**RESUMEN** 2](#_Toc513727707)

[**INTRODUCCIÓN** 9](#_Toc513727708)

[**ÍNDICE DE FIGURAS** 12](#_Toc513727709)

[**ÍNDICE DE CUADROS** 13](#_Toc513727710)

[CAPÍTULO I: 15](#_Toc513727711)

[VISIÓN DEL PROYECTO 15](#_Toc513727712)

[1.1. Planteamiento del problema 15](#_Toc513727713)

[1.1.1. El Negocio 15](#_Toc513727714)

[1.1.2. Los Procesos de Negocio 16](#_Toc513727715)

[1.2. Formulación del Problema 16](#_Toc513727716)

[1.2.1. Realidad Problemática 16](#_Toc513727717)

[1.2.2. Descripción del Problema 17](#_Toc513727718)

[1.3. Objetivos del Proyecto 18](#_Toc513727719)

[1.3.1. Marco Lógico 18](#_Toc513727720)

[1.3.2. Objetivo general 19](#_Toc513727721)

[1.3.3. Objetivos específicos 20](#_Toc513727722)

[1.4. Importancia del proyecto 20](#_Toc513727723)

[1.4.1. Justificación académica 20](#_Toc513727724)

[1.4.2. Beneficios tangibles 20](#_Toc513727725)

[1.4.3. Beneficios intangibles 20](#_Toc513727726)

[1.5. Alcance del proyecto 20](#_Toc513727727)

[CAPÍTULO II: 21](#_Toc513727728)

[MARCO TEÓRICO 21](#_Toc513727729)

[2.1. Data Mining 21](#_Toc513727730)

[2.1.1. Según (Jure Leskovec, Anand Rajaraman y Jeffrey David Ullman 2014) (ISBN: 9781107077232) 21](#_Toc513727731)

[2.1.2. Según (Matthew North 2012) (ISBN: 978-0615684376) 25](#_Toc513727732)

[CAPÍTULO III: 28](#_Toc513727733)

[ESTADO DEL ARTE 28](#_Toc513727734)

[REFERENCIAS BIBLIOGRÁFICAS 29](#_Toc513727735)

# **ÍNDICE DE FIGURAS**

[Ilustración 2Niveles Jerárquicos de la Toma de Decisiones en las Organizaciones. Fuente: Peña (2006) 24](file:///C%3A%5CUsers%5Cjruiton%5CDownloads%5C20180426_Tesis_BusinessIntelligenceGSS%20-%20Joel%20Ruiton.docx#_Toc512597189)

[Ilustración 3Metodología Rapid Warehousing 29](file:///C%3A%5CUsers%5Cjruiton%5CDownloads%5C20180426_Tesis_BusinessIntelligenceGSS%20-%20Joel%20Ruiton.docx#_Toc512597190)

[Ilustración 4Ciclo de Vida de la Metodología de Ralph Kimball 32](file:///C%3A%5CUsers%5Cjruiton%5CDownloads%5C20180426_Tesis_BusinessIntelligenceGSS%20-%20Joel%20Ruiton.docx#_Toc512597191)

**FIGURA 30.** Ventana de Explorador de Soluciones………………………………...……94

**FIGURA 31.** Editor de Dimensiones para DIM ESPECIFICACION………………..……94

**FIGURA 32.** Editor de Dimensiones para DIM ZONA ………………………………..….95

**FIGURA 33.** Editor de Dimensiones para DIM TIEMPO……………………………..…..95

# **ÍNDICE DE CUADROS**

[Tabla 1 Cuadro comparativo entre las metodologías de Ralph Kimball y Bill Inmon 35](#_Toc512597185)

[Tabla 2Cuadro comparativo entre las herramientas para el desarrollo de BI 39](#_Toc512597186)

[Tabla 3Puntuación de los parámetros de evaluación de las herramientas de BI. 40](#_Toc512597187)

[Tabla 4Resultados del cuadro comparativo entre las herramientas de BI 40](#_Toc512597188)

# CAPÍTULO I:

# VISIÓN DEL PROYECTO

## Planteamiento del problema

### El Negocio

Una Empresa Comercial está dedicada a vender productos y servicios, por lo tanto posee una extensa área geográfica donde ejerce dichas venta y servicios. Además, cuenta con una sede central y oficinas de anexo, y constantemente busca expandir su área de ventas para así poder incrementar sus ganancias, esto en base a las decisiones y políticas que elaboran para competir en el mercado actual.

**Misión**

Ayudar a nuestros clientes a optimizar la gestión de clientes, las ventas, la atención y la experiencia de sus clientes con sus marcas. (Pendiente)

**Visión**

La visión de GSS es liderar en el ámbito internacional la oferta de soluciones de gestión de clientes en tiempo real, de manera integral y multicanal; mediante la innovación continua en nuevas soluciones, tecnologías y procesos. (Pendiente)

### Los Procesos de Negocio

Las principales áreas que se encuentran implicadas en el planteamiento del problema y su solución son: la gerencia del área de ventas, el área de tecnología y la gerencia del área de marketing.

El área de ventas es la encargada de vender los productos y servicios que ofrece la empresa comercial para poder generar ganancias y obtener un mercado que confíen en los productos y servicios ofrecidos.

Una vez realizada el proceso de venta, se genera una gran cantidad de información producto de la venta, donde el cliente brinda detalles de la operación muy aparte de la generada durante la venta, esta información sirve al área de marketing para poder generar campañas y políticas que le permitan acaparar el mercado y obtener nuevos clientes.

El área de tecnología son los responsables de implementar soluciones tecnológicas que faciliten y permitan aprovechar dicha información obtenida durante las ventas

* 1. **Formulación del Problema**

### Realidad Problemática

La mayoría de las instituciones están comenzando a organizar, clasificar y usar correctamente la información, ya que conforme avanza la tecnología resulta más fácil almacenar grandes cantidades de datos, esta tendencia continua a un ritmo acelerado y permite obtener ventajas en campos tales como el marketing, criminalística, educación, y muchos más.

“La cantidad de información que nos llega cada día es tan inmensa que nos resulta difícil asimilarla”.1

En España, la minería de datos se aplica de forma exitosa en los procesos industriales, lo cual ha permitido agilizar el proceso en la producción del acero formando una red neuronal basada en la información obtenida.

“Los resultados obtenidos han servido para resaltar las principales ventajas que aportan las redes neuronales en el modelado de procesos industriales, versatilidad para adaptarse a diversas funciones, efectividad para modelar procesos no lineales aplicación intuitiva y facilidad de implementación”2

En algunas instituciones educativas de Colombia también se han logrado buenos resultados que permiten obtener nuevas posibilidades de su explotación, además de permitir desarrollar patrones generales de acuerdo al problema que se intenta solucionar.

 “Se ha obtenido un patrón general de deserción estudiantil en las dos IES determinado por un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera. Se han determinado factores socioeconómicos y académicos asociados a la deserción estudiantil.” 3

En Perú, en el ámbito criminalístico se ha intentado desarrollar avances en cuanto al buen uso de la información, permitiendo implementar sistemas de tipo alerta que basados en la información de los datos permitieron hacer visible los conocimientos que tenían algunas comisarías de la ciudad de Chiclayo, lo que se tradujo en la reducción del tiempo y facilidad de acceso a sistemas informáticos colaborativos que pudieron hacer más eficiente su labor.

“Tuvo como hipótesis que el desarrollo de un modelo de minería de datos como herramienta de apoyo podría contribuir a la caracterización de perfiles delictivos.”4

### Descripción del Problema

Actualmente muchas empresas comerciales generan las políticas y decisiones en base a reportes estadísticos en Excel que no usan toda la información disponible de la base de datos ya que esta no se encuentra ordenada, clasificada ni estructurada, por ello la base de datos no se está usando al máximo para generar estos reportes, lo que conlleva a un mínimo aprovechamiento de la información almacenada en la base de datos para la toma de decisiones (variable 1: uso de la base de datos, valor 1: 20%; variable 2: demora en la extracción y clasificación de la información valor 2: 7 días).

Del mismo modo, ya que tratan la información en Excel no se toma la data de años anteriores, provocando una toma de decisiones con información histórica insuficiente (variable 3: rango de información histórica, valor 3: 6 meses), esto conlleva a una toma de decisiones de las áreas de marketing y ventas en base a recursos limitados, debido a que no se está explotando dicha información.

## Objetivos del Proyecto

### Marco Lógico

#### Árbol de problemas

Reportes creados sin usar toda la data disponible

Reportes no muestran toda la información clave

Información desorganizada

**Mínimo aprovechamiento de la información almacenada en la base de datos para la toma de decisiones** (variable 1: uso de la base de datos, valor 1: 20%; variable 2: demora en la extracción y clasificación de la información valor 2: 7 días) (variable 3: rango de información histórica, valor 3: 6 meses)

Poco aprovechamiento de la data

Base de datos no analizada en su totalidad

Demora para obtener información

#### Árbol de objetivos

Obtener información confiable y fácil de entender e interpretar

Analizar las enormes bases de datos

Reducir el riesgo de perder clientes

**Mayor aprovechamiento de la información almacenada en la base de datos para la toma de decisiones** (variable 1: uso de la base de datos, valor 1: 60%; variable 2: demora en la extracción y clasificación de la información valor 2: 2 días) (variable 3: rango de información histórica, valor 3: 9 años)

Poder generar políticas, campañas y estrategias que ayuden a incrementar las ventas

Obtención de patrones que pueden ser aprovechados por las gerencias

### Objetivo general

Para mejorar la situación actual indicada se analizará la base de datos y se emitirán los reportes generados a las gerencias implicadas, esto para poder lograr un mayor aprovechamiento de la información almacenada en la base de datos para la toma de decisiones(variable 1: uso de la base de datos, valor 1: 60%; variable 2: demora en la extracción y clasificación de la información valor 2: 2 días) (variable 3: rango de información histórica, valor 3: 1 año), lo que permitirá a las gerencias de ventas y marketing poder tomar decisiones y campañas en base a dichos reportes.

Para poder elaborar los reportes y mostrar información sólida y confiable, usaremos información histórica insuficiente (variable 3: rango de información histórica, valor 3: 9 años).

### Objetivos específicos

1. Analizar la base de datos para determinar la magnitud de información que podemos obtener.
2. Detección de clústers, para poder reconocer los patrones durante el análisis de los datos
3. Aplicar la técnica Regresión, para poder predecir algunos resultados a futuro en base a la información obtenida.
4. Implementar algoritmos que permitan analizar los datos y clasificarlos.

## Importancia del proyecto

### Justificación académica

Poder dar a conocer las ventajas que te ofrece la minería de datos brindándote patrones, de este modo poder elaborar campañas de marketing y de ventas basándose en la información obtenida.

### Beneficios tangibles

1. Mayor aprovechamiento de la base de datos
2. Mejorar los reportes e informes brindados a las gerencias de Marketing y Ventas

### Beneficios intangibles

1. Mejorar las ventas
2. Mejorar la toma de decisiones de las gerencias de marketing y ventas.

## Alcance del proyecto

El alcance de la tesis es la elaboración y detección de patrones de venta y comerciales que permitirán realizar una mejor toma de decisiones, así como la elaboración de reportes sofisticados, que permitirán mostrar información de fácil interpretación para poder generar campañas de marketing y de ventas.

# CAPÍTULO II:

# MARCO TEÓRICO

## Data Mining

### Según (Jure Leskovec, Anand Rajaraman y Jeffrey David Ullman 2014) (ISBN: 9781107077232)

Los autores mencionados definen a la minería de datos como “El descubrimiento de modelo de datos”

Extracción de características:

Un modelo basado en características siempre busca en un fenómeno los ejemplos más resaltantes y en base a ello representa los datos.

Los tipos de extracción de características de datos a gran escala que muestra este libro son:

Conjunto de elementos frecuentes, donde nos indican que este modelo tiene sentido para datos que consisten en cestas de mercados, como por ejemplo la hamburguesa y el kétchup, productos que siempre se compran juntos cuando vamos al supermercado, por ello las promociones y ofertas de estos productos

Artículos similares, en esta parte, los autores nos dan a entender que el objetivo es encontrar un par de ejemplos con características similares y muchos elementos en común, a este proceso se le llama “filtrado colaborativo”, un ejemplo de uso de este proceso se da en la plataforma Amazon, cuando un cliente busca un producto, además del producto buscado el resultado le muestra productos con características similares, logrando así un incremento en sus ventas y preferencia por parte de los clientes.

El uso de índices, esta estructura de datos resulta de mucha utilidad al momento de recuperar objetos dependiendo del valor de uno o mas elementos de los objetos a buscar, generalmente la mayoría son registros, donde en uno de esos campos se encuentra el índice.

Por ejemplo, se podría dar el caso de tener un archivo con 3 campos (nombre, dirección y teléfono) y un índice en el campo del teléfono, dado un número de teléfono, el índice es el que nos va a permitir encontrar el registro de forma rápida.

En el libro también se indica la manera de implementar índices mediante una Hash Table, es decir el campo o los campos en los que se basa el índice a partir de una clave Hash para una función Hash.

“Los registros tienen la función hash aplicada al valor de la clave hash, y el registro en sí mismo se coloca en el cubo cuyo número está determinado por la función hash.

El depósito podría ser una lista de registros en la memoria principal, o un bloque de disco, por ejemplo.

Luego, dado un valor de clave hash, podemos hacer hash, encontrar el cubo y necesitamos buscar solo ese cubo para encontrar los registros con ese valor para la clave hash. Si elegimos que la cantidad de cubos B sea comparable a la cantidad de registros en el archivo, habrá relativamente pocos registros en cualquier segmento, y la búsqueda de un cubo llevará poco tiempo.”



 Flujos de minería de datos

Primero se explica la diferencia entre una secuencias y bases de datos.



Es decir, el hecho de que la velocidad de llegada de los elementos implica tener un sistema que controle la velocidad con que se leen los datos del disco, y no necesariamente tiene que ser uniforme, una vez controlado este proceso, el Sistema de Gestión de Base de Datos se encargará de administrar los datos sin sufrir ninguna pérdida.

### Según (Matthew North 2012) (ISBN: 978-0615684376)

. El autor sostiene que “La minería de datos como disciplina es ampliamente transparente para el mundo”, es decir en la mayoría de los casos casi nunca percibimos o nos damos cuenta de lo que está sucediendo, pero que constantemente nosotros generamos datos con cada actividad que realizamos, por ejemplo cuando vamos a realizar las comprar a un mercado y pagamos con tarjetas, o cuando navegamos por la web buscando información de acuerdo a nuestros intereses.

“Estos datos se almacenan en grandes conjuntos en potentes computadoras propiedad de las compañías con las que trabajamos todos los días. Estos conjuntos de datos son patrones indicadores de nuestros intereses, nuestros hábitos y nuestros comportamientos. La minería de datos permite a las personas ubicar e interpretar esos patrones, ayudándoles a tomar decisiones mejor informadas” y servir mejor a sus clientes.

Sin embargo, esto genera cierta preocupación sobre la privacidad, ya que las organizaciones acumulan cada vez más cantidad de datos tanto en volumen como en diversidad de estos, muchos de los cuales tienen naturaleza personal y privada.

El proceso de Minería de datos.

El proceso de la minería de datos se ha ido refinando conforme pasa el tiempo, pero no fue hasta que “la automotriz Daimler-Benz, el proveedor de seguros OHRA, el fabricante de hardware y software NCR Corp. y el fabricante de software estadístico SPSS, Inc. comenzaron a trabajar juntos para formalizar y estandarizar un enfoque de minería de datos” el resultado de lo mencionado fue CRISP-DM el proceso estándar de CRoss-Industry para la minería de datos, si bien es cierto que maneja ciertos softwares, este modelo fue desarrollado de manera conceptual, es decir que puede ser aplicado en cualquier software o tipo de datos. Este proceso consta de 6 pasos o etapas.



* + - 1. Comprensión empresarial u organizacional.

Todas las empresas u organismos son capaces de usar la minería de datos para responder preguntas y resolver problemas, sin embargo, es importante establecer y definir que se quiere saber, esto con el fin de orientar el esfuerzo y lograr una exitosa extracción de datos que permitan responder dichas preguntas.

* + - 1. Comprensión de datos.

Actualmente los datos se encuentran dispersados en muchos ordenadores, lo que se necesita es que se recoja toda la información disponible y centralizarla para lograr un mayor entendimiento de los datos, ello nos garantizará un mejor proceso de extracción de datos.

* + - 1. Preparación de datos

Los datos tienen distintas formas y formatos, pero no por eso vamos a descartar algunos de estos, eso reduciría nuestras opciones, lo que debemos hacer es usar herramientas que permitan trabajar dichos datos para así poder generar más patrones.



* + - 1. Modelado

En esta parte es donde implementamos los algoritmos para buscar, identificar y mostrar cualquier patrón o mensaje en los datos a tratar, hay dos tipos de modelos en la minería de datos, los que clasifican y los que permiten predecir (árboles de decisión)



* + - 1. Evaluación

“La evaluación se puede lograr utilizando una serie de técnicas, tanto de naturaleza matemática como lógica.”

Esto con el fin de identificar si el análisis tiene un falso positivo o si el modelo no encuentra patrones interesantes

* + - 1. Despliegue o implementación

En esta etapa se automatiza el modelo que hemos generado, integrarse con los sistemas de gestión o de información operacional existentes, alimentar el aprendizaje del modelo para mejorar su precisión y rendimiento, y monitorear y medir los resultados del uso del modelo, si bien al principio no puede mostrar los resultados esperados, conforme vaya procesando la información llegará a mostrar los patrones y las tendencias esperadas.

# CAPÍTULO III:

# ESTADO DEL ARTE

# REFERENCIAS BIBLIOGRÁFICAS

Data mining for the masses. Dr. Matthew North 2012 <https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf>

Mining of Massive Datasets Jure Leskovec, Anand Rajaraman y Jeffrey David Ullman 2014

<http://www.mmds.org/>