Sistema de minería de datos basado en patrones para mejorar el aprovechamiento de la base de datos de una empresa

Edgar Santillán Rodríguez edhu28one@gmail.com

Hugo Vega Huerta (asesor) hugovegahuerta@outlook.com

Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática Av. Germán Amezaga s/n, Ciudad Universitaria, Lima Julio 2018 – Perú

## Abstract

Currently, large amounts of data are produced in such a way that companies do not get to know all the information and therefore can’t get any benefit, which translates into little use of data.

The project presented in this document aims to identify commercial patterns and sales in a commercial company that could predict the behavior of the sale of a product by season; the aforementioned is done so that the commercial company can have a greater knowledge about the sales that can be obtained from information gathered in the previous years, and with this, elaborate policies that allow to take maximum advantage of the commercial area and can optimize resources in certain measure.

**Key words:** Data Mining, KDD - Discovery in database, Prediction, Data cleaning, Clustering, K-means.

## Resumen:

Actualmente se producen grandes cantidades de datos de tal forma que las empresas no llegan a conocer toda la información y por ende no logran sacar algún beneficio, lo que se traduce en el poco aprovechamiento de la data.

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.

**Palabras Claves:** Data Mining, KDD – Descubrimiento en base de datos, Predicción, Limpieza de datos, Clustering, K-means.

# INTRODUCCIÓN

En la actualidad las empresas generan muchos datos que no están clasificados, ordenados ni mucho menos estructurados, razón por la cual no se está 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.

# PRESENTACIÓN DEL PROBLEMA

* 1. **ANTECEDENTES DEL PROBLEMA**

Sin 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 (alrededor de un 15-20%) para la toma de decisiones y una demora (7-10 días) en la extracción y clasificación de la información.

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 (como máximo 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.

* 1. **FUNDAMENTACIÓN DEL PROBLEMA**

Este proyecto se desarrolla con el objetivo de aprovechar la base de datos de la empresa y así generar información que les permita realizar una mejor toma de decisiones.

* + 1. **PROBLEMA PRINCIPAL**

Mínimo aprovechamiento de la base de datos de la empresa.

* + 1. **PROBLEMAS SECUNDARIOS**
			1. Demora en la extracción y clasificación de la información.
			2. Insuficiente información histórica para realizar reportes.
			3. Distintos tipos de almacenamiento de la información.

# CONCEPTOS PREVIOS

* 1. **DATA MINING**

Según [1], 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.

# EXTRACCIÓN DE CARACTERÍSTICAS

Según [2], 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.

* 1. **CRISP-DM**

Según [1], 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.



**ILUSTRACIÓN 1: CRISP-DM.**

**(FUENTE: DATA MINING FOR THE MASSES WEB)**

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, es por ello que antes de comenzar con la implementación es necesario conocer o tener un entendimiento sobre la organización donde se va a realizar, conocer sus objetivos, observar su entorno, el ambiente donde se desenvuelve para así poder identificar sus problemas y poder generar una solución efectiva a la medida.

1. **Comprensión de datos.**

Actualmente los datos se encuentran dispersados en muchos ordenadores, es decir en la mayoría de casos se usan muchas fuentes de almacenamiento, lo que puede provocar información repetida, un difícil acceso para buscar y en términos generales no es una buena práctica que recomienden los expertos, lo que se necesita –en especial parta realizar un sistema de minería de datos- es que se recoja toda la información disponible que se encuentra dispersada en los distintos ordenadores y centralizarla en una gran base de datos, de esta forma podemos lograr un mayor entendimiento de los datos y por ende generar información productiva para analizar, ello nos garantizará un mejor proceso de extracción de datos que nos permitirá reflejar la realidad y poder generar patrones productivos.



**ILUSTRACIÓN 2: CENTRALIZAR LA INFORMACIÓN.**

**(FUENTE: DATA MINING FOR THE MASSES WEB)**

1. **Preparación de datos.**

Actualmente 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 nos permitan trabajar con dichos datos, es decir, buscar la manera de aprovechar la mayor cantidad de los datos posibles que tengamos a nuestra disposición, ello generará mucha más variedad de patrones y también logrará que estos sean más robustos.



**TABLA 1: FORMATOS DE DATOS QUE DEBEN SER APROVECHADOS.**

**(FUENTE: DATA MINING FOR THE MASSES WEB)**

1. **Modelado.**

Actualmente 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), en cualquier caso, el objetivo no cambia ya que nosotros estamos aprovechando la mayor cantidad de datos posibles para poder generar un sistema de minería de datos que se adapte a las necesidades de la empresa y le permita analizar la información con mayor precisión y claridad.



**ILUSTRACIÓN 3: MODELADO.**

**(FUENTE: DATA MINING FOR THE MASSES WEB)**

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, es en esta etapa donde comenzamos a evaluar lo generado en las etapas anteriores, esto con el fin de probar su validez antes de comenzar con la implementación.

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.

* 1. **CLUSTERING (AGRUPAMIENTO)**

Según [3], “la agrupación en clúster es el proceso de agrupar un conjunto de objetos de datos en múltiples grupos o clústeres para que los objetos dentro de un clúster tengan una gran similitud, pero son muy diferentes a los objetos en otros clústeres. Las diferencias y similitudes se evalúan en función de los valores de los atributos que describen los objetos y a menudo implican medidas de distancia.1 La agrupación como herramienta de minería de datos tiene sus raíces en muchas áreas de aplicación, como biología, seguridad, inteligencia comercial y búsqueda web.”

#### ¿Qué es el análisis de conglomerados?

Es un proceso que nos permite realizar la partición de un conjunto de datos en subconjuntos (clúster), cada subconjunto de clúster posee objetos con características similares, y cada clúster tiene características diferentes, a esto podemos llamarlo agrupamiento, cabe resaltar que esta división o agrupamiento es realizada por los algoritmos, lo que puede ser provechoso ya que se pueden descubrir nuevos grupos que estaban previamente desconocidos antes de agrupar los datos.

El análisis de estos datos se ha realizado en distintos escenarios, en los que se incluye la inteligencia empresarial, el reconocimiento de patrones de imágenes la búsqueda web, biología entre otros.

Especialmente en la inteligencia empresarial, la agrupación en clústers puede utilizarse para poder agrupar a los clientes de diversas formas posibles, siempre y cuando estos tengan ciertas características similares que les permitan realizar dicha agrupación.

#### Requisitos para el análisis de conglomerados:

Escalabilidad, Capacidad de tratar con diferentes tipos de atributos, Descubrimiento de clusters con forma arbitraria, Requisitos para el conocimiento del dominio para determinar los parámetros de entrada, Capacidad de tratar con datos ruidosos, Agrupamiento incremental e insensibilidad al orden de entrada, Capacidad de agrupar datos de alta dimensionalidad, Agrupación basada en restricciones, Interpretabilidad y usabilidad.

#### Descripción general de los métodos básicos de agrupamiento

En general, los principales métodos de agrupamiento fundamentales se pueden clasificar en las siguientes categorías:

**Métodos de particionamiento:**

* + - Encontrar grupos mutuamente exclusivos de forma esférica.
		- Basado en la distancia.
		- Puede usar k-means o k-medoid (etc.) para representar el centro del clúster.
		- Eficaz para conjuntos de datos de tamaño pequeño a mediano.

**Métodos jerárquicos.**

* + - Agrupamiento es una descomposición jerárquica (es decir, niveles múltiples).
		- No se pueden corregir fusiones o divisiones erróneas.
		- Puede incorporar otras técnicas como micro-clustering o considerar los "enlaces" del objeto.

**Métodos basados en densidad.**

* + - Puede encontrar clusters de forma arbitraria.
		- Los clusters son regiones densas de objetos en el espacio que son separados por regiones de baja densidad.
		- Densidad de cluster: cada punto debe tener un número mínimo de puntos dentro de su "vecindario".
		- Puede filtrar valores atípicos.

**Métodos basados en grillas.**

* + - Usar una estructura de datos de cuadrícula multi-resolución.
		- Tiempo de procesamiento rápido (típicamente independiente del número de objetos de datos, aunque dependen del tamaño de la cuadrícula).

# MODELADO DEL SISTEMA

* 1. **CASOS DE USO DEL NEGOCIO**



**ILUSTRACIÓN 4: CASO DE USO DEL NEGOCIO**

**(FUENTE: ELABORACIÓN PROPIA)**

* 1. **REQUERIMIENTOS FUNCIONALES**



**TABLA 2: REQUERIMIENTOS FUNCIONALES**

**(FUENTE: DATA MINING FOR THE MASSES WEB)**

* 1. **DIAGRAMA DE CASOS DE USO DEL SISTEMA**



 **ILUSTRACIÓN 5: CASO DE USO DEL SISTEMA**

**(FUENTE: ELABORACIÓN PROPIA)**

* 1. **MODELO CONCEPTUAL**

****

**ILUSTRACIÓN 6: DIAGRAMA DE CLASES**

# METODOLOGÍA PARA EL SISTEMA

* 1. **ESTÁNDARES DE DOCUMENTACIÓN**

Para el modelado del sistema es importante mantener un estándar que permita el fácil uso y entendimiento de la información que se va a tratar.



**TABLA 3: ESTÁNDAR PROPUESTO**

**(FUENTE: ELABORACIÓN PROPIA)**

* 1. **GESTIÓN DEL PROYECTO**

A continuación se muestra el organigrama del proyecto, la cual representa de qué manera se distribuyen las personas que lo dirigen.



**ILUSTRACIÓN 7: ORGANIGRAMA**

**(FUENTE: ELABORACIÓN PROPIA)**

# CONCLUSIONES

* Al implementar el proyecto se optimizaría el uso de la base de datos aprovechando la información para la toma de decisiones.
* La minería de datos le permite a la empresa obtener información a través de procesos en las bases de datos que le pueden permitir elaborar políticas de manejo en la empresa.

# RECOMENDACIONES

* Se debe tener en cuenta que mientras más información se pueda tratar, se podrá aprovechar mejor la base de datos y se podrán generar más patrones y una segmentación más robusta de los clientes.

# REFERENCIAS BIBLIOGRÁFICAS

[1] (M. North, 2012) (ISBN: 9780615684376): Dr. Matthew A. North: Data mining for the masses.

[2] (J. Leskovec, A. Rajaraman, J. Ullman 2014) (ISBN: 9781107077232) Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman: Mining of Massive Datasets.

[3] (H. Jiawei, M. Kamber 2006) (ISBN: 1558609016) Jiawei Han & Micheline Kamber: Data Mining: Concepts and Techniques Morgan Kaufmann.

***Sobre los autores:***

******

***SANTILLÁN RODRÍGUEZ, Edgar:*** Nació en Lima - Perú el 28/12/90. Estudió la carrera de Ingeniería de Sistemas e Informática en la Universidad Nacional Mayor de San Marcos. Actualmente se desempeña como Analista RPA en Interbank.



***VEGA HUERTA, Hugo:*** Nació en Ancash - Perú el 14/09/64. Estudió la carrera de Computación en la Universidad Nacional Mayor de San Marcos. Actualmente es catedrático en la UNMSM y en la Universidad Ricardo Palma.