**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

**Alumno:** Santillán Rodríguez, Edgar Hugo

**Curso:** Proyecto deTesis II

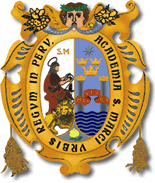
**Proyecto:** Sistema de minería de datos basado en patrones para un mayor aprovechamiento de la base de datos de una empresa.

**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 DE MINERÍA DE DATOS BASADO EN PATRONES PARA UN MAYOR APROVECHAMIENTO DE LA BASE DE DATOS DE UNA EMPRESA**

**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 DE MINERÍA DE DATOS BASADO EN PATRONES PARA UN MAYOR APROVECHAMIENTO DE LA BASE DE DATOS DE UNA EMPRESA**

“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

Junio 2018

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 DE MINERÍA DE DATOS BASADO EN PATRONES PARA UN MAYOR APROVECHAMIENTO DE LA BASE DE DATOS DE UNA EMPRESA**

# 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,Clustering, K-means.

# 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.

ÍNDICE

[RESUMEN 6](#_Toc516228959)

[INTRODUCCIÓN 7](#_Toc516228960)

[ÍNDICE DE FIGURAS 10](#_Toc516228961)

[ÍNDICE DE CUADROS 11](#_Toc516228962)

[CAPÍTULO I: 12](#_Toc516228963)

[VISIÓN DEL PROYECTO 12](#_Toc516228964)

[1.1. Planteamiento del problema 12](#_Toc516228965)

[1.1.1. El Negocio 12](#_Toc516228966)

[1.1.2. Los Procesos de Negocio 12](#_Toc516228967)

[1.2. Formulación del Problema 13](#_Toc516228968)

[1.2.1. Realidad Problemática 13](#_Toc516228969)

[1.2.2. Descripción del Problema 14](#_Toc516228970)

[1.3. Objetivos del Proyecto 15](#_Toc516228971)

[1.3.1. Marco Lógico 15](#_Toc516228972)

[1.3.2. Objetivo general 17](#_Toc516228973)

[1.3.3. Objetivos específicos 17](#_Toc516228974)

[1.4. Importancia del proyecto 17](#_Toc516228975)

[1.4.1. Justificación académica 17](#_Toc516228976)

[1.4.2. Beneficios tangibles 17](#_Toc516228977)

[1.4.3. Beneficios intangibles 17](#_Toc516228978)

[1.5. Alcance del proyecto 17](#_Toc516228979)

[CAPÍTULO II: 18](#_Toc516228980)

[MARCO TEÓRICO 18](#_Toc516228981)

[2.1. Data Mining 18](#_Toc516228982)

[2.1.1. Según (Jure Leskovec, Anand Rajaraman y Jeffrey David Ullman 2014) (ISBN: 978-1-1070-7723-2) 18](#_Toc516228983)

[2.1.2. Según (Matthew North 2012) (ISBN: 978-0-6156-8437-6) 20](#_Toc516228984)

[2.2. Agrupamiento o Clustering 24](#_Toc516228985)

[2.2.1. Según (Charu C. Aggarwal & Chandan K. Reddy 2014) (ISBN: 978-1-4665-5821-2) 24](#_Toc516228986)

[2.2.2. Según (Jiawei Han & Micheline Kamber) (ISBN 978-0-1238-1479-1) 28](#_Toc516228987)

[CAPÍTULO III: 33](#_Toc516228988)

[ESTADO DEL ARTE 33](#_Toc516228989)

[3.1. La minería de datos como un método innovador para la detección de patrones de deserción estudiantil en programas de pregrado en instituciones de educación superior (San Juan de Pasto, 2013) 33](#_Toc516228991)

[3.2. Implantación de un sistema de ventas que emplea una herramienta de Data Mining (Lima, 2012) 36](#_Toc516228992)

[BIBLIOGRAFÍA 39](#_Toc516228993)

# ÍNDICE DE FIGURAS

[**Ilustración 1: Uso de índices.** 19](#_Toc516230128)

[**Ilustración 2: Uso de índices.** 20](#_Toc516230129)

[**Ilustración 3: Modelo Conceptual CRISP-DM.** 21](#_Toc516230130)

[**Ilustración 4: Centralizar la información.** 22](#_Toc516230131)

[**Ilustración 5: Tipos de modelos de data mining.** 23](#_Toc516230132)

[**Ilustración 6: Datos agrupados mediante K-means.** 27](#_Toc516230133)

[**Ilustración 7: Agrupación de un conjunto de objetos utilizando el método k-means; para (b) actualizar centros de clúster y reasignar objetos en consecuencia (la media de cada clúster está marcada por un signo +) Pag. 453.** 32](#_Toc516230134)

[**Ilustración 8: Proceso KDD (UIAF, 2014).** 35](#_Toc516230135)

[**Ilustración 9: Técnicas de modelado de Data Mining (MOLINA & GARCIA, 2006).** 37](#_Toc516230136)

**FIGURA 3**No se encuentran elementos de tabla de ilustraciones.**0.** Ventana de Explorador de Soluciones………………………………...……94

**FIGURA 31.** Editor de Dime**No se encuentran elementos de tabla de ilustraciones.**nsiones 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: Formatos de datos que deben ser aprovechados.** 23](#_Toc516230283)

[**Tabla 2: Métodos básicos de agrupamiento.** 30](#_Toc516230284)

[**Tabla 3: Cuadro comparativo de las herramientas de Data Mining.** 36](#_Toc516230285)

# 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

## 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

Información desorganizada

Reportes no muestran toda la información clave

Reportes creados sin usar toda la data disponible

**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

Demora para obtener información

Base de datos no analizada en su totalidad

#### Árbol de objetivos

Obtener información confiable y fácil de entender e interpretar

Reducir el riesgo de perder clientes

Analizar las enormes bases de datos

**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)

Obtención de patrones que pueden ser aprovechados por las gerencias

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

### 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: 978-1-1070-7723-2)

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.

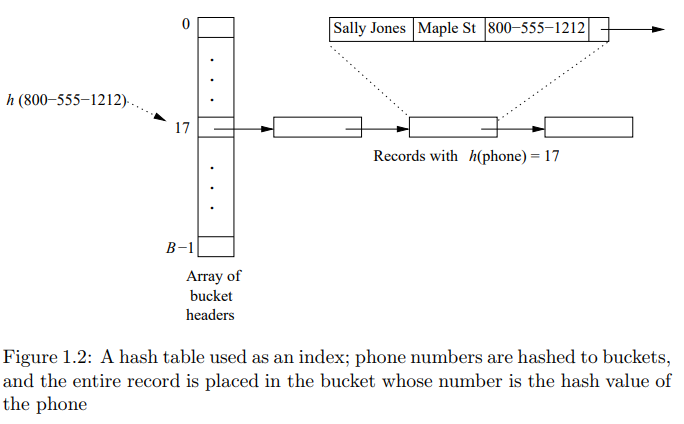
Por ejemplo, se podría dar el caso de tener un archivo con 3 campos (nombre, dirección y teléfono) y un índice en el campo del teléfono, dado un número de teléfono, el índice es el que nos va a permitir encontrar el registro de forma rápida.

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.

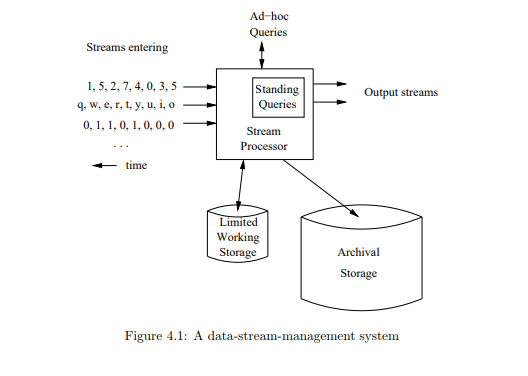
Luego, dado un valor de clave hash, podemos hacer hash, encontrar el cubo y necesitamos buscar solo ese cubo para encontrar los registros con ese valor para la clave hash. Si elegimos que la cantidad de cubos B sea comparable a la cantidad de registros en el archivo, habrá relativamente pocos registros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



**Ilustración 1: Uso de índices.**

.

Primero se explica la diferencia entre una secuencia 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.



**Ilustración 2: Uso de índices.**

### Según (Matthew North 2012) (ISBN: 978-0-6156-8437-6)

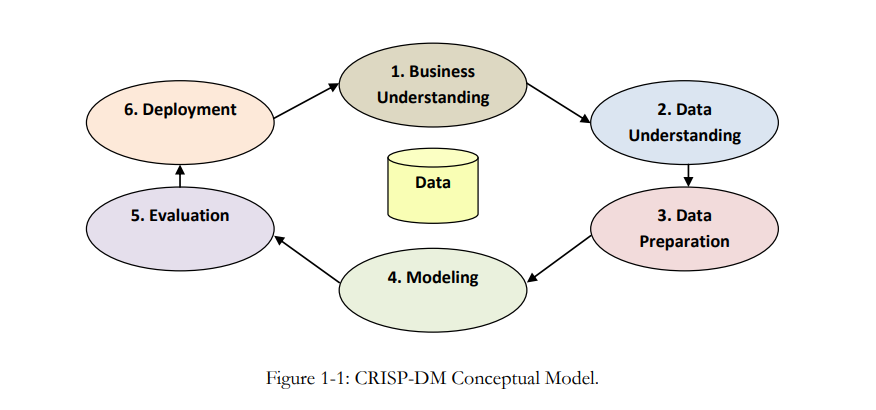
. El autor sostiene que “La minería de datos como disciplina es ampliamente transparente para el mundo”, es decir en la mayoría de los casos casi nunca percibimos o nos damos cuenta de lo que está sucediendo, pero que constantemente nosotros generamos datos con cada actividad que realizamos, por ejemplo 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.



**Ilustración 3: Modelo Conceptual CRISP-DM.**

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.**

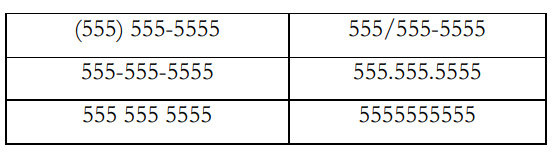
Actualmente los datos se encuentran dispersados en muchos ordenadores, es decir en la mayoría de casos se usan muchas fuentes de almacenamiento, lo que puede provocar información repetida, un difícil acceso para buscar y en términos generales no es una buena práctica que recomienden los expertos, lo 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 4: Centralizar la información.**

1. **Preparación de datos.**

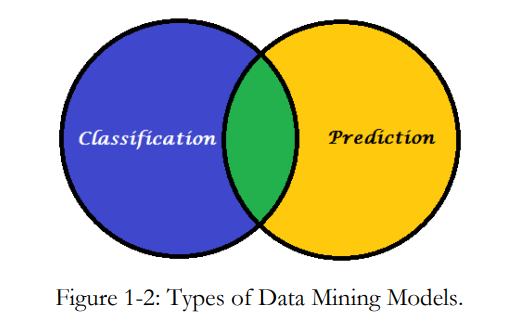
Los datos tienen distintas formas y formatos, pero no por eso vamos a descartar algunos de estos, eso reduciría nuestras opciones, lo que debemos hacer es usar herramientas que nos permitan trabajar con dichos datos, es decir, buscar la manera de aprovechar la mayor cantidad de los datos posibles que 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.**

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), 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 5: Tipos de modelos de data mining.**

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.

## Agrupamiento o Clustering

### Según (Charu C. Aggarwal & Chandan K. Reddy 2014) (ISBN: 978-1-4665-5821-2)

Uno de los mayores retos a los que se enfrenta la minería de datos siempre ha sido la agrupación, ya que de por si abarca una gran cantidad de aplicaciones como por ejemplo aprendizaje, segmentación, marketing objetivo entre otros.

Como la mayoría de información no viene identificada o etiquetada, la agrupación o clustering resulta ser una muy buena alternativa para poder para poder generar un modelo conciso de datos, es decir poder agrupar datos para que puedan ser interpretados.

El problema básico de la agrupación puede establecerse de la siguiente manera: “Given a set of data points, partition them into a set of groups which are as similar as possible.” (Dad un conjunto de puntos de datos, dividámoslo en un conjunto de grupos que tengan elementos lo más similares posibles).

Sin embargo, hay que tener en cuenta que esta es una definición aproximada más no exacta o definitiva, esto debido a que las variaciones que puedan ocurrir en la definición del problema pueden ser significativas, dependiendo del modelo que se vaya a usar.

Algunos escenarios donde podemos realizar la agrupación o clustering son los siguientes:

* Paso intermedio para otros problemas fundamentales de minería de datos.
* Filtrado colaborativo
* Segmentación del cliente.
* Resumen de datos.
* Detección dinámica de tendencias.
* Análisis de datos multimedia.
* Análisis de datos biológicos.
* Análisis de redes sociales.

Esto nos indica la gran diversidad de escenarios o problemas que podemos resolver mediante los algoritmos de agrupamiento o clustering.

#### Categorías de trabajo en el área de agrupamiento.

**Centrado en la técnica:**

Dado que la agrupación es un problema bastante popular, no es de extrañar que se usen muchos métodos, como técnicas probabilísticas, técnicas basadas en la distancia, técnicas espectrales, técnicas basadas en la densidad y técnicas basadas en la reducción de la dimensionalidad, proceso. Cada uno de estos métodos tiene sus propias ventajas y desventajas, y puede funcionar bien en diferentes escenarios y dominios problemáticos. Ciertos tipos de datos, como los datos de gran dimensión, los grandes datos o la transmisión de datos, tienen su propio conjunto de desafíos y, a menudo, requieren técnicas especializadas.

**Centrado en tipo de datos:**

Diferentes aplicaciones crean diferentes tipos de datos con diferentes propiedades. Por ejemplo, una máquina de ECG (electrocardiograma) producirá puntos de datos de series de tiempo que estén altamente correlacionados entre sí, mientras que una red social generará una mezcla de documentos y datos estructurales. Algunos de los ejemplos más comunes son datos categóricos, datos de series temporales, secuencias discretas, datos de red y datos probabilísticos. Claramente, la naturaleza de los datos tiene un gran impacto en la elección de la metodología utilizada para el proceso de agrupamiento. Además, algunos tipos de datos son más difíciles que otros debido a la separación entre diferentes tipos de atributos, como el comportamiento o los atributos contextuales.

**Información adicional de las variaciones de agrupamiento:**

También se diseñaron varias ideas para diferentes tipos de variaciones de agrupamiento. Por ejemplo, el análisis visual, el análisis supervisado, el análisis por conjuntos o el análisis de múltiples vistas se pueden usar para obtener información adicional. Además, la cuestión de la validación de clúster también es importante desde la perspectiva de obtener conocimientos específicos sobre el rendimiento de la agrupación.

#### Técnicas comunes utilizadas en el análisis de conglomerados

* **Métodos de selección de características (Feature Selection Methods)**

Este método es un paso de pre-proceso necesario para mejorar la agrupación subsiguiente, básicamente elimina las características ruidosas e irrelevantes en la fase de pre procesamiento.

“La selección de características y la reducción de dimensionalidad están estrechamente relacionadas. En la selección de características, se seleccionan subconjuntos originales de las características. En la reducción de dimensionalidad, se pueden usar combinaciones lineales de características en técnicas tales como el análisis de componentes principales [50] para mejorar aún más el efecto de selección de características. La ventaja de la primera es una mayor interpretabilidad, mientras que la ventaja de la segunda es que se requiere un número menor de direcciones transformadas para el proceso de representación.”

* **Modelos probabilísticos y generativos (Probabilistic and Generative Models)**

En este tipo de modelos la idea es modelar los datos de un proceso generativo, el conjunto de datos disponibles se usan para estimar los parámetros de manera “que tengan un ajuste de máxima verosimilitud con el modelo generativo”, en base a esto estimamos las probabilidades generativas de los puntos de datos subyacentes. Los puntos de datos que se ajusten a la distribución son los que tendrán más probabilidad de ajuste, mientras que las “anomalías” tendrán un ajuste muy bajo.

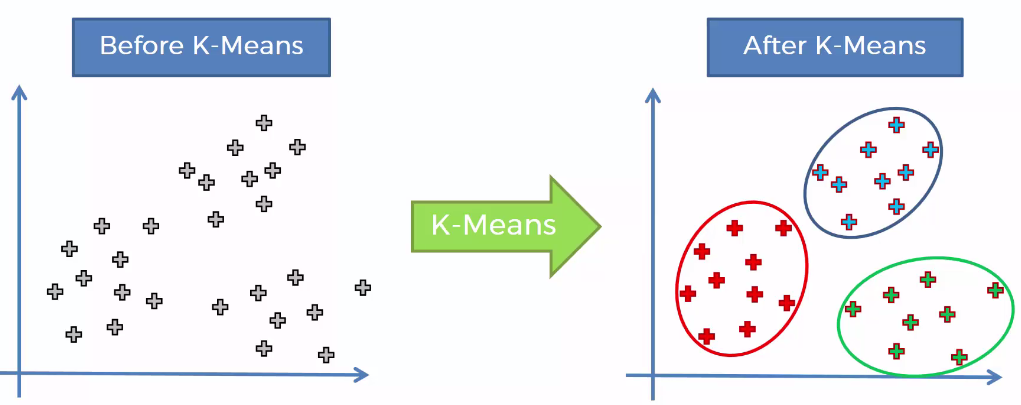
* **Algoritmos basados en distancia (Distance-Based Algorithms)**

La mayoría de modelos generativos pueden simplificarse en algoritmos basados en distancia, esto debido a que dicho modelo usan una función de distancia en su distribución de probabilidad, son los más usados debido a su relativa simplicidad y fácil implementación en una gran variedad de escenarios, incluyendo el nuestro (marketing-venta).

Los algoritmos basados en la distancia generalmente se dividen en dos tipos:

1. **Plano:** En este caso, los datos se dividen en muchos clústers en una sola toma, es necesario elegir un representante de la partición y la función de distancia para poder controlar el algoritmo que va a seguir, en cada interacción, los datos van a irse ajustando a sus representantes de particiones más cercanos, y luego el representante se va a ir ajustando a los puntos de datos asignados al clúster. Algunos métodos comunes para generar las particiones son los siguientes: **K-means, K-medians, K-medoids.**

**K-Means: “**En estos métodos, los representantes de particiones corresponden a la media de cada grupo. Tenga en cuenta que el representante de partición no se extrae del conjunto de datos original, sino que se crea como una función de los datos subyacentes. La distancia euclidiana se usa para calcular distancias. El método k-Means se considera uno de los métodos más simples y más clásicos para la agrupación de datos [46] y también es quizás uno de los métodos más ampliamente utilizados en implementaciones prácticas debido a su simplicidad.”



**Ilustración 6: Datos agrupados mediante K-means.**

1. **Jerárquico:** En estos métodos, los conglomerados se representan jerárquicamente mediante un dendograma, en “niveles variables de granularidad”, dependiendo de si esta representación es creada de forma ascendente o descendente, pueden considerarse aglomerativas o divisivas.

**Aglomerativas:** En este método, se usa un enfoque que va de abajo hacia arriba, comenzando con los puntos de datos individuales y luego fusionando los clusters de forma sucesiva hasta crear una estructura similar a un árbol

**Divisivos:** En estos métodos, se usa un enfoque de arriba hacia abajo, para dividir y formar una estructura también similar a la de un árbol, además se puede usar cualquier algoritmo de clúster plano para realizar la partición en cada paso, esto provoca una mayor flexibilidad en la estructura jerárquica del árbol como en el nivel de equilibrio de los diferentes conglomerados. En este caso no es necesario tener un árbol perfectamente equilibrado en profundidad de cada nodo o que el grado de cada rama sea exactamente dos.

### Según (Jiawei Han & Micheline Kamber) (ISBN 978-0-1238-1479-1)

“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:

|  |  |
| --- | --- |
| Método | Características generales |
| 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). |

**Tabla 2: Métodos básicos de agrupamiento.**

**K-means, una técnica basada en centroide:**

Vamos a suponer que tenemos un conjunto de datos **D** que contiene **n** objetos en el espacio Euclidiano.

Además, los métodos de particionamiento distribuyen de **D** en k clústers (c1, c2,…c**K**) es decir, Ci ⊂ D y Ci **∩** Cj = ∅ para (1 ≤ i, j ≤ k), tenemos que usar una función objetivo que permita evaluar la calidad de la partición de modo que los elementos dentro de un clúster sean similares entre sí, pero diferentes al de los demás clústers.

Una técnica de partición basada en un centroide es la que usa el centroide de un clúster Ci para representar ese clúster. “El centroide se puede definir de varias maneras, como por la media o medoide de los objetos (o puntos) asignados al clúster. La diferencia entre un objeto p ∈ Ci y ci, el representante del grupo, se mide por dist (p, ci), donde dist (x, y) es la distancia euclidiana entre dos puntos xey. La calidad del clúster Ci se puede medir mediante la variación conincluso, que es la suma del error al cuadrado entre todos los objetos en Ci y el centroide ci, definido como E donde la suma del error al cuadrado para todos los objetos en el conjunto de datos; p es el punto en el espacio que representa un objeto dado; y ci es el centroide del grupo Ci (tanto p como ci son multidimensionales). En otras palabras, para cada objeto en cada grupo, la distancia desde el objeto a su centro de grupo se cuadra, y las distancias se suman. Esta función objetivo intenta hacer que los clusters k resultantes sean tan compactos y tan separados como sea posible. La optimización de la variación dentro del cluster es computacionalmente desafiante. En el peor de los casos, tendríamos que enumerar una cantidad de posibles divisiones que son exponenciales a la cantidad de clústeres y verificar los valores de variación dentro del clúster. Se ha demostrado que el problema es NP-hard en el espacio euclidiano general incluso para dos clusters (es decir, k = 2). Además, el problema es NP-hard para un número general de clusters k incluso en el espacio euclidiano 2-D. Si el número de conglomerados k y la dimensionalidad del espacio d son fijos, el problema se puede resolver en el tiempo ), donde n es el número de objetos. Para superar el costo computacional prohibitivo para la solución exacta, los enfoques codiciosos a menudo se usan en la práctica. Un buen ejemplo es el algoritmo k-means, que es simple y de uso común.”

El algoritmo k-means para particionar, donde el centro de cada cluster está representado por el valor medio de los objetos en el cluster.

Entrada:

**K**: la cantidad de clusters,

**D**: un conjunto de datos que contiene n objetos.

**Salida**: un conjunto de **k** clusters.

Método:

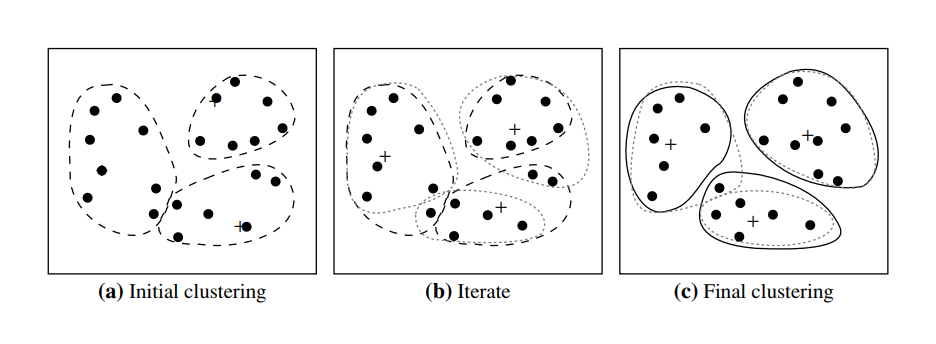
(1) elija arbitrariamente **k** objetos de **D** como centro inicial del clúster;

(2) repetir

(3) (re) asigne cada objeto al clúster al que el objeto sea más similar, basado en el valor medio de los objetos en el cluster;

(4) actualice el clúster significa, es decir, calcule el valor medio de los objetos para cada grupo;

(5) hasta que no haya cambios;



**Ilustración 7: Agrupación de un conjunto de objetos utilizando el método k-means; para (b) actualizar centros de clúster y reasignar objetos en consecuencia (la media de cada clúster está marcada por un signo +) Pag. 453.**

# CAPÍTULO III:

# ESTADO DEL ARTE

En este capítulo vamos a revisar distintos trabajos que se han venido realizando a nivel mundial, cuyas contribuciones en el ámbito de la minería de datos ha producido resultados favorables que han de ser mencionados en esta sección como casos de éxito.



## La minería de datos como un método innovador para la detección de patrones de deserción estudiantil en programas de pregrado en instituciones de educación superior (San Juan de Pasto, 2013)

El trabajo realizado por Ricardo Timarán Pereira y Andrés Calderón Romero, en la universidad de Nariño, Colombia, logra identificar que en Latinoamérica, la educación superior presenta altas tasas de deserción estudiantil, especialmente en los primeros semestres académicos, conllevando a efectos de tipo financiero, académico y social tanto para las Instituciones de Educación Superior (IES) como para el estudiante, la región, el país y el Estado. En este artículo se presenta uno de los resultados del proyecto de investigación financiado por el Ministerio de Educación Nacional cuyo objetivo fue detectar patrones de deserción estudiantil a partir de los datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado de la Universidad de Nariño e Institución Universitaria CESMAG, dos instituciones de educación superior (la primera pública y la segunda privada) de la ciudad de Pasto (Colombia), utilizando técnicas de Minería de Datos. Inicialmente, se seleccionaron, de las bases de datos de estas instituciones, los datos socioeconómicos, académicos, disciplinarios e institucionales de los estudiantes que ingresaron en los años 2004, 2005 y 2006 a los diferentes programas de pregrado, con el fin de hacerles un seguimiento hasta el año 2011, para determinar la deserción de los estudiantes. Se construyó un repositorio de datos utilizando el SGBD PostgreSQL en el cual se aplicaron técnicas para obtener datos limpios a los cuales se les aplicó las técnicas de minería de datos. Obteniendo al final un resultado favorable los cuales sirvieron para definir políticas para la retención estudiantil.

Metodología : Descubrimiento De Conocimiento En Bases De Datos (KDD).

Técnica : La técnica de clasificación basada en árboles de decisión.

Algoritmo : Algoritmo de árboles de decisión C4.5. b (r)

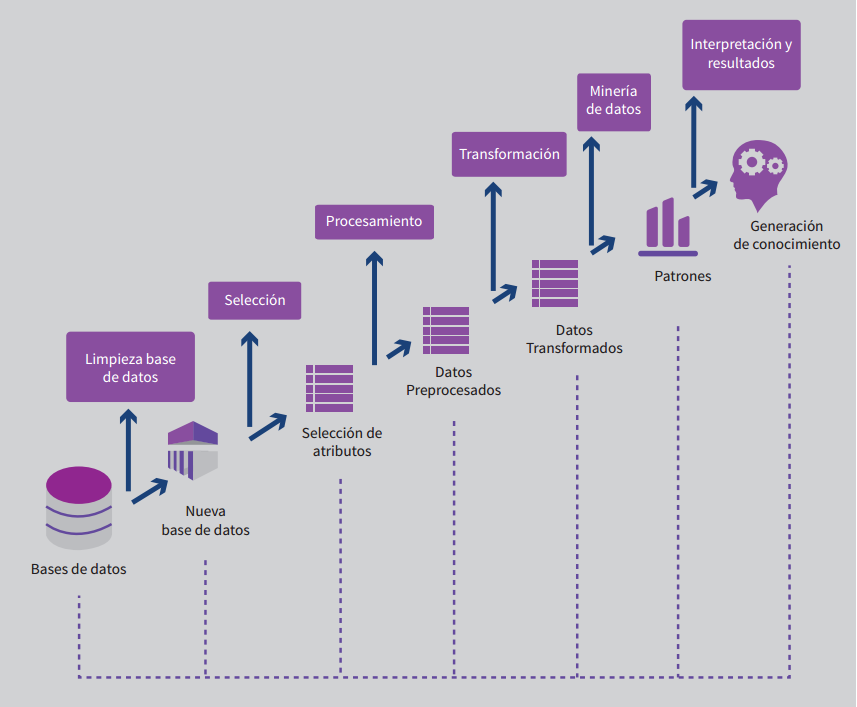
Trabajos futuros: Se ha obtenido un patrón general de deserción estudiantil en las dos IES determinado por un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera. Se han determinado factores socioeconómicos y académicos asociados a la deserción estudiantil. La evaluación, análisis y utilidad de estos patrones permitirá soportar la toma de decisiones eficaces de las directivas universitarias enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil que actualmente se encuentran establecidos. Como trabajos futuros están el continuar con el estudio de deserción estudiantil en la Universidad de Nariño e Institución Universitaria CESMAG aplicando otras técnicas de minería de datos tales como asociación y clustering con el fin de determinar afinidades, similitudes y relaciones entre los factores socioeconómicos y académicos de las estudiantes que desertan.

**Conclusión**

Se ha logrado encontrar muchas causantes de deserción estudiantil, uno de los más notorios de los indicadores encontrados es el promedio bajo que obtienen los estudiantes durante los primeros semestres de la carrera, también se logró encontrar otro indicador muy marcado, los factores socioeconómicos, con esto se ha logrado obtener un patrón general de deserción estudiantil a través de la minería de datos, que luego de ser clasificada y tratada, va a brindar un panorama más detallado que deberán de ser tomadas en cuenta por las autoridades respectivas. Gracias a la minería de datos se ha logrado un avance en cuanto a la organización y buen uso de la información en Colombia, como se observa en esta publicación, si los datos son analizados, seleccionados y filtrados correctamente pueden ser una fuente muy rica de conocimiento, lo que va a permitir realizar acciones correctivas y preventivas en dichas instituciones de Colombia.

**Aportes:**

El tema mencionado previamente utiliza la metodología del Descubrimiento De Conocimiento En Bases De Datos (KDD), que se define como un “Proceso no trivial de identificar patrones válidos, novedosos potencialmente útiles y, en última instancia, comprensibles a partir de los datos”. (Fayad, Piateski-Shapiro y Smyth, 1996, citado en Camargo y Silva, 2010 p. 12), además según UIAF (2014), el proceso de Knowledge Discovery in DataBases (KDD) surte cuatro pasos para la generación de conocimiento. Estas etapas pueden ser recursivas, es decir, que se retorna a ellas una y otra vez (proceso iterativo) a medida que se obtienen resultados preliminares que requieren replantear las variables iniciales.



**Ilustración 8: Proceso KDD (UIAF, 2014).**

**Utilidad para el proyecto de tesis**

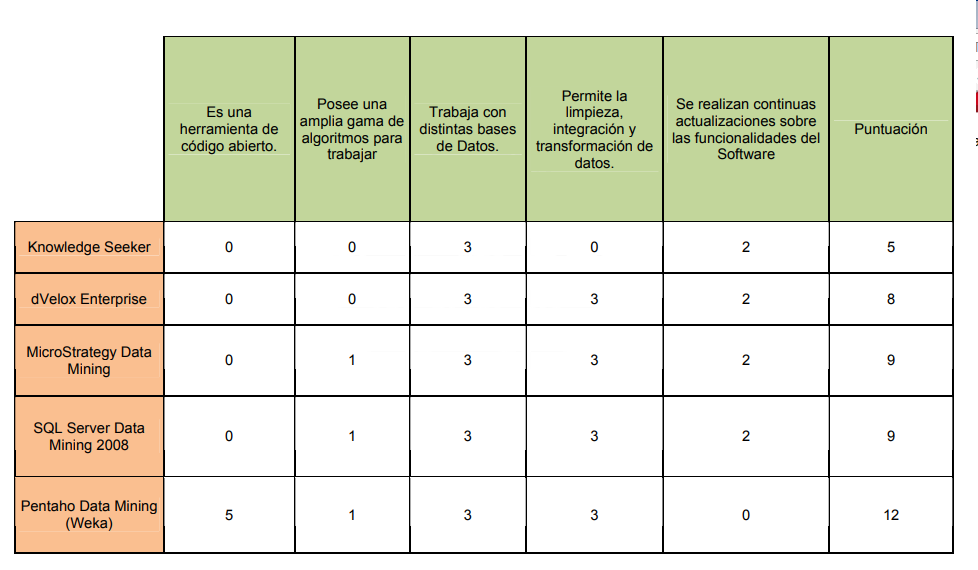
Se tomará en cuenta el trabajo realizado para la elaboración y estructura de la base de datos utilizando la técnica del proceso KDD para lograr un mayor aprovechamiento de la base de datos.

## Implantación de un sistema de ventas que emplea una herramienta de Data Mining (Lima, 2012)

En esta Tesis sustentada por Miguel Angel Berrospi Ramirez, se puede observar que data mining es una ventaja competitiva que añade valor agregado en una organización, para este caso, el rubro textil, es importante el estar informado constantemente sobre el comportamiento del mercado, poder realizar la atención de una forma personalizada, es por ello la importancia de una herramienta de data mining que le permita obtener dichos beneficios para poder mantenerse en el mercado.

La empresa en mención muestra una pérdida de clientes por la constante competencia y una falta de organización estratégica de la empresa para poder retenerlos, es aquí donde la herramienta implementada va a jugar un papel importante, ya que le va a permitir no solo aprovechar la base de datos, si no poder generar clusters que le permitan una mejor toma de decisiones.

Se evaluaron varias herramientas de data mining, considerando las herramientas de paga y las herramientas de código abierto, a continuación un cuadro comparativo de las herramientas, ventajas y desventajas según Angel.



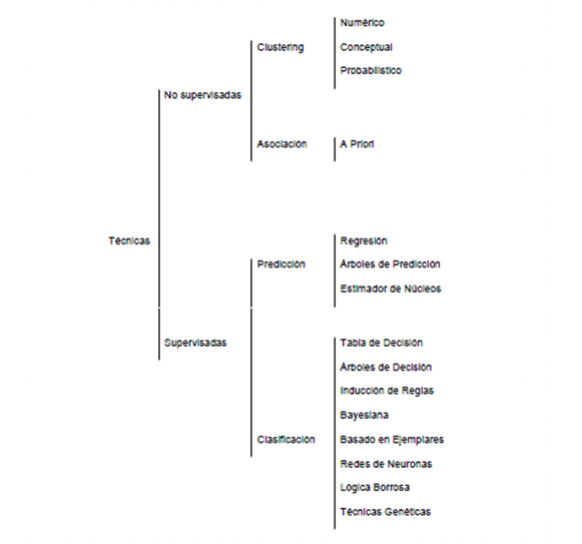
**Tabla 3: Cuadro comparativo de las herramientas de Data Mining.**

De todas las indicadas, se escoge a la herramienta Pentaho Data Mining (Weka), principalmente por ser software de código abierto, lo que permite un acceso rápido a la herramienta y accesorios, además de estar integradas con las otras soluciones de la plataforma Pentaho, Weka posee soportes para la integración de datos, análisis, dashboards y reportes, con las principales características:

* Motor poderoso para manejo de grandes volúmenes de datos.
* Posee una gran colección de algoritmos confiables y robustos.
* Relaciones eficientes y capacidad de descubrir patrones de datos.
* Integración simple y acelerada de datos.

Con esta implementación, el usuario, además de obtener un mayor aprovechamiento de la base de datos, va a tener acceso a nuevos niveles de información, donde no solo va a encontrar la clasificación de sus clientes, sino también el poder comprender como han ido evolucionando en el tiempo, y en base a ello evaluar y predecir el comportamiento de compra de muchos de estos.

Existen muchas técnicas de modelado para lograr lo mencionado previamente, a continuación un pequeño esquema de las técnicas según su clasificación:



**Ilustración 9: Técnicas de modelado de Data Mining (MOLINA & GARCIA, 2006).**

Como podemos observar, en este proyecto se requiere y busca predecir el comportamiento que tienen los clientes, entonces resulta viable usar una técnica supervisada de predicción, en este caso Árboles de decisión, para ello implementaron el algoritmo J48, que es el nombre como se le llama en Weka al algoritmo C4.5 (árbol para generar una decisión).

Los resultados de este proyecto saltan a la luz, se logró identificar el porcentaje de deserción de clientes, las posibles razones y además de ellos los posibles indicios de cuando un cliente va a dejar de consumir los productos ofrecidos por la empresa textil, si bien la herramienta logró resultados esperados, es casi seguro que se obtendrán resultados ligeramente diferentes de usarse otro algoritmo u herramienta, pero que pueden llegar a ser tan confiables como el que se trabajó en este proyecto.

**Utilidad para el proyecto de tesis**

Se tomará en cuenta el trabajo realizado para construcción del modelo (técnica de modelado y software) utilizando la herramienta Pentaho Weka para lograr obtener patrones de comportamiento en los clientes de la empresa.

# BIBLIOGRAFÍA

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

<https://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf>

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

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

* **[A. Charu, R. Chandan, 2014]:** Charu C. Aggarwal & Chandan K. Reddy (editors): Data Clustering: Algorithms and Applications. Chapman & Hall / CRC Press, 2014. ISBN 1466558210.

<http://charuaggarwal.net/clusterbook.pdf>

* **[T. Pang-Ning, M. Steinbach, V. Kumar, 2006]:** Pang-Ning Tan, Michael Steinbach & Vipin Kumar: Introduction to Data Mining Addison-Wesley, 2006. ISBN 0321321367.
* **[H. Jiawei, M. Kamber, 2006]:** Jiawei Han & Micheline Kamber: Data Mining: Concepts and Techniques Morgan Kaufmann, 2006. ISBN 1558609016.

<http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>

* **[M. Berrospi, 2012]:** Miguel Angel Berrospi Ramirez: Implantación de un Sistema de Ventas que emplea una herramienta de Data Mining, 2012.

<http://tesis.pucp.edu.pe/repositorio/bitstream/handle/123456789/5002/BERROSPI_MIGUEL_IMPLANTACION_SISTEMA_VENTAS_HERRAMIENTA_DATA_MINING.pdf?sequence=1>

* **[A. Gonzales, 2007]:** Ana González Marcos: Desarrollo de técnicas de minería de datos en procesos industriales, 2007. Tesis doctoral de la Universidad de la Rioja, España.

<https://dialnet.unirioja.es/servlet/tesis?codigo=1166>

* **[J. Jiménez, 2013]:** Javier Jiménez Toledo: La Minería de Datos como un Método innovador para la detección de Patrones de deserción estudiantil en Programas de Pregrado en instituciones de educación superior, 2013. Artículo de la Universidad de Nariño, Colombia.

<http://www.acofipapers.org/index.php/acofipapers/2013/paper/viewFile/211/112>

* **[L. Jiménez, 2015]:** Lesly Haymet Jiménez Berríos: Aplicación de un Sistema de alerta temprana basada en la Minería de Datos para identificar patrones delictivos en la Ciudad de Chiclayo, 2015.Tesis para optar por el título de Ingeniero de Sistemas y Computación, Universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú.

<http://tesis.usat.edu.pe/handle/usat/543>