

**TITULO DE TESINA**

**INGENIERIA EN SISTEMAS COMPUTACIONALES**

PRESENTA

**DANIEL ENRIQUE COLUNGA GALVÁN**

**MAYRA NALLELY GOMEZ SANCHEZ**

**RAFAEL ARANDA SOTO**

ASESOR

**M.T.I. JORGE RAMIREZ ORTEGA**

AGUASCALIENTES, AGS A 20 DE NOVIEMBRE DEL 2017

**Resumen**

**Índice**

**Introducción**

En la época actual, todos los días generamos datos, la revolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, almacenar, distribuir, se usa prácticamente día con día; muchas de las veces sin darnos cuenta, millones de bits se transmiten cada minuto a nivel mundial, esto también genera accesos y herramientas que controlamos desde dispositivos móviles, ya que el internet ahora está más al alcance de cualquier persona.

Con el importante progreso en informática y en las tecnologías relacionadas así como en la expansión de su uso en diferentes aspectos y enfoques de la vida cotidiana, se continúa recogiendo y almacenando gran cantidad de información en bases de datos situadas por toda la red dispersas por todo el mundo. Por lo que descubrir conocimiento de volumen de datos tan masivo representa un gran reto y cada vez más una necesidad latente.

La minería de datos es un intento de buscar sentido a la explosión de información que actualmente puede ser almacenada.

Hoy en día, los datos no están restringidos a tuplas representadas solamente con números o caracteres. El avance de la tecnología para la gestión de bases de datos hace posible integrar diferentes tipos de datos, tales como imagen, video, texto, y otros datos numéricos, en una base de datos de poca complejidad, facilitando el procesamiento multimedia. Como resultado, la mezcla tradicional de técnicas estadísticas y herramientas de gestión de datos ya no son adecuadas para analizar esta vasta colección de datos desiguales.

Actualmente, el Internet y su creciente demanda requiere del desarrollo de tecnologías de minería de datos más avanzadas para interpretar la información y poder así generar conocimiento de los datos distribuidos por todo el mundo.

En este siglo la demanda continuará en aumento, y el acceso a grandes volúmenes de datos multimedia traerá la mayor transformación para la sociedad. Por tanto, el desarrollo de la tecnología de minería de datos avanzada continuará siendo una importante área de estudio, y por lo tanto, se prevé que se haga una importante inversión en esta área de desarrollo en los próximos años.

Existen diversos dominios donde se almacenan grandes volúmenes de información en bases de datos centralizadas y distribuidas, tales como librerías digitales, archivos de imágenes, bioinformática, cuidados médicos, finanzas e inversión, fabricación y producción, negocios y marketing, redes de telecomunicación, etc.

La información forma parte importante de la piedra angular de diversas disciplinas actuales alrededor del mundo, siendo parte fundamental para obtener grandes ventajas, ya sea de manera competitiva o como forma de prevención.

Todo apunta a la necesidad de metodologías de análisis inteligente de datos, las cuales puedan descubrir conocimiento útil de los datos.

El término KDD (iniciales de Knowledge Discovery in Databases), acuñado en 1989 se refiere a todo el proceso de extracción de conocimiento a partir de una base de datos y marca un cambio de paradigma en el que lo importante es el conocimiento útil que seamos capaces de descubrir a partir de los datos.

Las tareas propias de la fase de minería de datos pueden ser descriptivas, (como descubrir patrones interesantes o relaciones describiendo los datos), o predictivas (por ejemplo: clasificar nuevos datos basándose en los anteriormente disponibles). Dicho de otro modo, es un campo interdisciplinar con el principal objetivo general de predecir las salidas y revelar las posibles relaciones existentes presentadas en los datos.

Para ello se utilizan herramientas automáticas que

* Emplean algoritmos sofisticados para descubrir principalmente patrones ocultos, asociaciones, anomalías, y/o estructuras de la gran cantidad de datos almacenados en los data warehouses u otros repositorios de información.
* Filtran la información necesaria de las grandes bases de datos.

Data warehousing, se refiere a las tendencias actuales en la recolección y limpieza de datos transaccionales para que estén disponibles para el análisis y la toma de decisiones. La minería de datos debe trabajar mano a mano con los almacenes de datos, sobre todo en los casos de volúmenes de datos muy grandes o de inter-relaciones entre los datos complejas, es decir, que no puedan ser expresadas en una tabla plana.

El concepto de minería de datos también se relaciona con los conceptos de aprendizaje automático y de estadística. En general, la estadística es la primera ciencia que históricamente extrae información de los datos básicamente mediante metodologías procedentes de las matemáticas.

Cuando se empezó a usar los ordenadores como apoyo para esta tarea surgió el concepto de Machine Learning, traducido como Aprendizaje Automático.

Posteriormente con el incremento del tamaño y con la estructuración de los datos es cuando se empieza a hablar de minería de datos.

Un claro ejemplo de lo la importancia que juegan los datos en las empresas y organizaciones, es la información que se genera a diario en cualquier banco del mundo. Toda esta información debe tener un seguimiento riguroso, para organizarse y almacenarse adecuadamente previo a su distribución.

En el primer trimestre del año 2017, las reclamaciones en materia de comercio por internet presentadas ante la propia institución bancaria de acuerdo a cifras recaudadas por la CONDUSEF (Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros) ascendieron a un total de 579,649 teniendo un total del 93% de resoluciones favorables en un tiempo de respuesta promedio de 7 días, por lo que al contar con un número tan elevado para un período de 3 meses nos hace que surjan las preguntas ¿Se puede prever este tipo de fraudes? ¿Podría haber una forma de disminuir o erradicar los mismos?

El fenómeno del lavado de dinero en México genera ganancias superiores a los 25 mil millones de dólares al año, y representa una décima parte del dinero obtenido por el crimen organizado, mismo que resulta muy difícil de detectar y comprobar con técnicas y herramientas convencionales, sin embargo este tipo de operaciones fraudulentas suelen tener patrones característicos que permiten con cierto grado de probabilidad distinguirlos de los movimientos legales.

Para detectar estos patrones se necesita el uso de la minería de datos. Otro de los usos que se le puede dar la información como se comentaba, es el marketing.

En Aguascalientes existen empresas pequeñas de financiamiento que pueden ser beneficiadas en el área de marketing y subir en grandes proporciones su número de clientes. Teniendo acceso a las necesidades del cliente, así como a las deficiencias que no les agradan, se puede elevar exponencialmente el número de clientes.

Con la minería de datos se pretende que una microempresa de financiamiento se convierta en una mediana empresa ofreciendo más y mejores servicios que la competencia para poder pasar de ser una pequeña empresa a una mediana.

**Problemática**

La resistencia al cambio siempre será un conflicto en cualquier ámbito, científicamente está probado que el usuario promedio se encuentra en una zona de confort, si bien el cambio es bueno el usuario se negará siempre, esto es algo que tenemos que cambiar pues países de primer mundo constantemente buscan la mejora, y el usuario se adapta sin inconveniente, esto principalmente se vuelve la problemática en nuestro país.

Ese no es únicamente al problema que nos enfrentamos, también tenemos la problemática de la inversión para nuevos proyectos, la mayoría de las empresas se niegan a esto, pues tienen cierto criterio para hacer gastos, esto desafortunadamente no siempre es lo correcto, pues como bien lo decimos el cambio es bueno, pero la zona de confort juega un papel bastante importante, ejemplifiquémoslo de la siguiente forma: cuando la señal de televisión abierta cambio de análoga a digital ¿Qué ocurrió? … la mayoría de la sociedad en nuestro país estaba en contra por diferentes situaciones, por cambiar los televisores, por el temor al cambio, pero la realidad es completamente diferente.

La señal mejoró bastante en cuanto a imagen y pixeles, se añadieron más canales de los que normalmente se tenían, se contamina menos con televisores modernos, la señal mejoro bastante es decir con una simple antena de conejo puedes tener una muy buena señal.

Si bien esto en otros países no fue ningún problema, la economía en nuestro país juega un papel muy importante, pues si lo analizamos detalladamente la mayoría de las empresas no confían en los propios cambios del país, buscan a que otras personas de otros países vengan y nos digan que hacer, esto también es un cambio necesario en nuestra sociedad, somos un país con grandes ingenieros, los cuales solo buscan una pequeña oportunidad en ocasiones con pequeños cambios generan grandes ganancias.

Ahora bien el cambio que tratamos de realizar es en la banca específicamente en la caja popular mexicana, principalmente sabemos que este tipo de empresas son muy cerradas en cuanto a sus sistemas, los cambios no son constantes son más necesarios cuando se dan cuenta que su sistemas es totalmente obsoleto entonces buscan hacer un cambio, la importancia de las aplicaciones móviles está jugando un papel bastante importante en este sector pues evitar filas y realizar transacciones desde nuestro celular es algo muy cómodo, desafortunadamente es un gran problema mantener una seguridad, pues la delincuencia busca erradicar la seguridad y hacer mal uso de estas apps.

Otro problema es la comunicación de la empresa con el usuario, pues normalmente cuando se hace una llamada o nos llega un correo normalmente nos ofrecen otros servicios, pero nunca nos pregunta cómo estamos en cuanto a nuestro servicio, solo nos dan la opción de reportar, la ciberdelincuencia es un problema constante en todo el mundo es por ello que es necesario crear medidas de seguridad para estos casos

Consideremos también que la banca a nivel mundial se ve amenazada por los delitos cibernéticos, robos de identidad entre muchos otros.

En una entrevista realizada a  Anthony Caterino, líder global de Servicios Financieros de EY e Ignacio Aldonza, socio líder del Sector Financiero de la firma en México, “En México es parecido, va siguiendo la estela, el robo de identidad a través del sistema financiero mexicano es un problema grave y creciente, una violación a los derechos fundamentales de las personas”

De tal forma que no solo afecta el ámbito de sistemas pues se requiere también apoyo del gobierno así como cambio de leyes o adaptaciones para dicho problema, es necesario competir a estos casos con bastantes argumentos que necesitan ser explorados por algoritmos en la minería de datos

**Preguntas de investigación**

1. ¿A través del uso de minería de datos, se disminuirá la probabilidad de fraudes en la empresa de financiamiento “Caja Popular Mexicana”?
2. ¿A través del uso de minería de datos, se disminuirá la frecuencia de fraudes en la empresa de financiamiento “Caja Popular Mexicana”?
3. ¿A través del uso de minería de datos se podría encontrar alguna estrategia para la retención de clientes en la empresa de financiamiento “Caja Popular Mexicana”?
4. ¿A través de la minería de datos se podrán encontrar fallos en la atención al cliente en la empresa de financiamiento “Caja Popular Mexicana”?
5. ¿A través de la minería de datos se puede predecir los gastos por tipo de clientes en la empresa de financiamiento “Caja Popular Mexicana”?
6. ¿A través de la minería de datos se pueden predecir la periodicidad de fraudes en la empresa de financiamiento “Caja Popular Mexicana”?
7. ¿Con el uso de minería de datos se incrementará la seguridad en la integridad de los datos en la empresa de financiamiento “Caja Popular Mexicana”?
8. ¿Con el uso de minería de datos se puede incrementar el número de clientes que usen los servicios de la empresa de financiamiento “Caja Popular Mexicana”

**Hipótesis**

**H0**. El uso de minería de datos en cajas de ahorro permitirá identificar patrones de fraude, así como medir la satisfacción del cliente.

**H1.** El uso de minería de datos en aplicaciones o sistemas bancarios no permitirá al usuario poder tener un control más elevado sobre sus clientes, así como ayudar a identificar patrones de fraude, medir la satisfacción del cliente

**OBJETIVO GENERAL:**

Identificar la viabilidad de la implementación de la minería de datos en una caja de ahorro.

**OBJETIVOS ESPECÍFICOS:**

1. Identificar la factibilidad de uso de la minería de datos para la prevención de fraudes para las cajas de ahorro
2. Identificar la posibilidad de medir la satisfacción del cliente mediante la minería de datos.

**Justificación**

A nivel mundial la minería de datos tiene diferentes usos, gracias a esta técnica de análisis de datos, se puede descubrir información que no se tiene o no se espera obtener.

Con la minería de datos se ahorran grandes cantidades en empresas a nivel mundial y de igual manera abre muchas oportunidades de negocios.

De igual manera trabajar con esta técnica implica cuidar cada detalle debido a que el resultado final del análisis involucra tomar una decisión que será importante que definirá el rumbo del negocio al que se aplicó.

El uso de minería de datos abre paso en el mercado contribuyendo a la toma de decisiones como se mencionaba anteriormente, tácticas así como estrategias proporcionando un sentido automatizado para identificar información clave en grandes volúmenes de datos que generan los procesos tradicionales en un negocio o empresa.

De igual manera permite a los usuarios dar prioridades en las acciones mostrando factores de riesgo.

Gracias al uso de minería de datos se pueden dividir los problemas en distintos sectores y esto provoca que en esos diversos sectores se pueda asignar el personal adecuado así como especializado en el ámbito de cierto problema para así poder optimizar tiempo y recursos

Se pretende que esta técnica de análisis se implemente en este país ya que implicaría un cambio significativo en los sectores de economía, logística, empresas de las que depende la atención al cliente, en fin, México tiene gran potencial así como los recursos necesarios para llevar a la práctica la minería de datos.

Sin embargo, uno de los mejores sectores en los que podría ayudar más, sería usar la minería de datos para aplicaciones y sistemas bancarios, ya que gracias a esta técnica, se podrían detectar oportunamente fraudes, evasión de impuestos, lavado de dinero, enriquecimiento ilícito, entre otras.

Lo que se pretende es empezar en Aguascalientes, empezar con una sucursal de “Caja Popular Mexicana”. Se pretende comenzar haciendo un análisis en sus datos que permitan ayudar a dicha empresa a mejorar en varios aspectos.

Los principal, como se comentaba anteriormente será poder detectar cualquier patrón que ayude a prevenir fraudes que se pueda presentar ya sea por empleados, por clientes o por personas externas a la empresa.

Con la minería de datos, se detectaran a tiempo características que puedan evaluar el comportamiento de los datos que se producen a diario en la institución bancaria, se evaluaran y se obtendrá un indicador que permita identificar malas prácticas.

Otra de los propósitos que tendrá el uso de la minería de datos será poder evaluar la satisfacción del cliente, es decir, se podrá ver algún indicador en los datos por los que clientes han decidido abandonar la institución bancaria y cambiar a otra.

Con esto se pretende que la institución mejore en su atención al cliente, decidir cómo si es necesario ofrecer más servicios o mejores tasas de interés a sus cuentahabiente.

También se podrá recabar información de los comentarios positivos así como de los negativos que se le hacen a la institución, y mediante esto poder mantener el servicio o mejorarlo.

Gracias también a la minería de datos, la institución podrá ver que es lo que necesita los cuentahabientes, es decir que están buscando, las expectativas que se esperan de un banco, entre otras. Todos estos datos, ayudaran también a conseguir nuevos clientes, a elevar el número de persona incorporadas a la institución.

Después de este análisis de datos, la institución bancaria será una empresa más sólida, una empresa que podrá ofrecer mayores beneficios, una empresa bancaria que tendrá la posibilidad de ofrecer mejores servicios que otras, por lo que podrá convertirse de una pequeña a una mediana empresa.

**CAPÍTULO I**

**Marco teórico**

La sociedad actual es fiel testigo de los grandes cambios que hemos tenido en un periodo de tiempo muy corto, como se señala en la obra de El Shock del futuro (Toffler, 1970), estamos expuestos continuamente a una sobrecarga de información “information overload”; lo que presenta retos como almacenar, administrar y procesar la información, a fin de acceder al conocimiento, para actuar con sabiduría. Ahora, el reto que se presenta es encontrar información fidedigna, veraz y con la mayor calidad posible, entre un gran volumen de datos, en donde muchos de ellos son sólo “basura”, que afectan la calidad de la información.

La Cadena Datos-Información-Conocimiento, denominada “Jerarquía de la Información” o “Pirámide del Conocimiento” (Cadenas Figueredo, 2015), es uno de los modelos más importantes en la Gestión de la Información y la Gestión del Conocimiento; y este modelo permite el abstraer y principalmente estructurar diferentes temas, con sus respectivos contenidos.

El concepto de “Jerarquía del Conocimiento” ó “Pirámide del Conocimiento” (Ackoff, 1989), es un referente teórico que permite dar sustento al orden que se debe seguir en la gestión de los datos. La secuencia de datos, información, conocimiento y sabiduría, DIKW, por sus siglas en inglés (data, information, knowledge, wisdom), es la que da estructura a los trabajos dedicados a la obtención de resultados, a partir de grandes volúmenes de datos.

La explotación, manejo, administración y gestión de los datos, ha dado como resultado el desarrollo de una serie de teorías provenientes de muchos campos, que basan su desarrollo en diversas técnicas, incorporando elementos de las matemáticas, estadística, computación, ingenierías e incluso de las ciencias sociales, que tienen como propósito fundamental el reconocimiento de patrones, recurriendo a una informática avanzada, modelando la realidad y desarrollando algoritmos que automaticen los procesos, para llegar a óptimos resultados.

La Ciencia de los Datos (Figuerola, 2015), es un concepto relativamente nuevo que se utiliza a menudo de manera intercambiable con inteligencia o análisis de negocios competitivos. La ciencia de datos busca utilizar todos los datos disponibles y relevantes para contar efectivamente una historia que pueda ser fácilmente comprendida por los no practicantes. Un practicante de la ciencia de datos se denomina un científico de datos.

Asociados a la Ciencia de Datos, se encontrarán otros conceptos cuyo principal propósito es el aprovechar grandes volúmenes de información, almacenada en enormes Bases de Datos. Es así que conceptos como Inteligencia Artificial, Big Data, Datos abiertos (Open Data), Fishing data, Data Warehouse, Minería de datos (Data Mining); entre otros conceptos surgidos en décadas recientes, empiezan a ser cada vez más utilizados por los “científicos de los datos”.

El tema central que nos ocupa es el de la Minería de Datos, que se define, de acuerdo con Mena (1999), como « *Iterative process of extracting hidden predictive patterns from large data-bases, using AI technologies as well as statistics techniques* », o de otro modo más simple, en convertir datos sin aparente utilidad, en conocimiento. Sus aproximaciones son varias, siendo el término inglés «Knowledge Discovery in Databases» (Proceso iterativo de extracción de patrones predictivos escondidos de grandes bases de datos, usando tecnologías de inteligencia artificial así como técnicas estadísticas) -representado por las siglas KDD, o «descubrimiento de conocimiento en bases de datos» es el que más se utiliza para asimilarlo a esta disciplina.

Aunque técnicamente debemos saber que KDD es el proceso general y la minería de datos forma parte de él. Es así como a partir la Minería de Datos se pretende extraer información o conocimiento que sea válido y novedoso, es decir que aporte algo nuevo, potencialmente útil, ya que debe derivar hacia la toma de decisiones y que éstas sean comprensibles para los usuarios.

En síntesis “Data Mining” es el tratamiento de datos masivos para extraer conclusiones e información relevante de ellos.

En el ámbito de la investigación es fundamental el análisis, clasificación, segmentación y explotación de los datos para la formulación de hipótesis; por lo que se recurre, ya sea al método descriptivo, mediante el cual se tratan de identificar elementos interpretables, a través de la búsqueda de reglas de asociación y patrones secuenciales, o bien, se utilizan **Métodos predictivos, en el que se** usan variables para predecir valores futuros o desconocidos de otras variables.

La Minería de datos cuenta con su propia metodología, que consta de las siguientes fases:

* Análisis del problema
* Comprensión de los datos
* Preparación de los datos
* Modelamiento
* Evaluación
* Despliegue funcional

El análisis del problema es una fase crucial del trabajo, ya que permite contextualizar la situación de la cual se parte y conocer cuáles son las necesidades de contar con determinados datos y saber de cuáles de ellos se dispone y cuáles son imprescindibles de obtener.

En la comprensión de los datos es preciso identificar si son continuos o discretos, qué tipos de valores adquieren, si están bien capturados y la utilidad que se les puede dar a futuro.

En la fase de preparación de los datos se tiene que evaluar la forma en cómo se organizarán, la interrelación que tienen entre sí y la forma en cómo se les puede dar el mayor aprovechamiento.

Durante el Modelamiento, se requieren definir los algoritmos que se van a utilizar para dar tratamiento a los datos.

En la evaluación se analizan los resultados obtenidos y que tienen que ser sometidos a comprobación, verificando que estén libres de errores y ratificando que son útiles para los objetivos que se persiguen.

Finalmente, en el despliegue funcional, se desarrollan herramientas informáticas para automatizar los procesos, que permitan identificar alertas, generar estadísticas y reportes, que sean de utilidad para la adecuada toma de decisiones.

La Minería de Datos es de gran valía para todas aquellas personas u organizaciones que requieren explotar grandes volúmenes de información en lapsos sumamente cortos, comparados con los tiempos que invertirían si los trabajos se hicieran de forma artesanal.

Grandes empresas privadas, la banca, las Universidades e incluso el sector gubernamental tienen en la Minería de Datos una poderosa herramienta para el aprovechamiento de los grandes volúmenes de información de la que son depositarios.

De acuerdo a Aluja (2001), un campo privilegiado de aplicación de las técnicas de minería de datos es el marketing, concretamente todo aquello que se agrupa bajo el nombre de CRM (Costumer Relationship Management), donde el objetivo es conocer lo mejor posible los clientes para poder satisfacerlos mejor y asegurar así la rentabilidad de las empresas. Problemas tales como estimar el potencial económico de los clientes, modelizar la probabilidad de baja, medir la satisfacción por el servicio, descubrir nuevos segmentos de clientes potenciales etc.