

MINERÍA DE DATOS

INGENIERÍA EN SISTEMAS COMPUTACIONALES

# PRESENTA

**ARANDA SOTO RAFAEL**

**COLUNGA GALVÁN DANIEL ENRIQUE**

**GÓMEZ SÁNCHEZ MAYRA NALLELY**

# ASESOR

## M.T.I. JORGE RAMÍREZ ORTEGA

### AGUASCALIENTES, AGS A 20 DE NOVIEMBRE DEL 2017

#### **Resumen.**

#### **Índice.**

[Introducción. 4](#_Toc498863031)

[Problemática. 9](#_Toc498863032)

[Preguntas de investigación. 12](#_Toc498863033)

[Hipótesis. 13](#_Toc498863034)

[Objetivo general. 14](#_Toc498863035)

[Objetivos específicos. 15](#_Toc498863036)

[Justificación. 16](#_Toc498863037)

[CAPÍTULO I 19](#_Toc498863038)

[Marco teórico. 19](#_Toc498863039)

[Marco conceptual. 26](#_Toc498863040)

[CAPÍTULO II 32](#_Toc498863041)

[Marco contextual. 32](#_Toc498863042)

[CAPÍTULO III 38](#_Toc498863043)

[Diseño de investigación. 38](#_Toc498863044)

[Metodología de la investigación. 40](#_Toc498863045)

[Población y muestra. 42](#_Toc498863046)

[CAPÍTULO IV 43](#_Toc498863047)

[Propuesta de instrumento. 43](#_Toc498863048)

[Alcance del instrumento. 44](#_Toc498863049)

[Resultados. 45](#_Toc498863050)

[Conclusiones. 47](#_Toc498863051)

[Recomendaciones. 48](#_Toc498863052)

# **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 es 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 donde 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 empezaron 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 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.

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 es factible 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 el único problema que se enfrenta, también existe la problemática de la inversión para nuevos proyectos, la mayoría de las empresas se niegan a esto, debido a que tienen ciertos criterios para hacer sus gastos, lo que en ocasiones los lleva a una zona de confort, como por ejemplo, cuando la señal de televisión abierta cambio de análoga a digital, una buena parte de la sociedad se oponía por temor ante diferentes situaciones, sin embargo, en la realidad los resultados fueron completamente diferentes a lo que se pensaba.

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 mejoró bastante, ya que con una simple antena de conejo se puede tener una muy buena señal.

Si bien esto en otros países no representó problema alguno, en nuestro país si, debido a que existen empresas que no confían en los propios cambios del país, por lo que buscan a personas de otros países para que vengan y nos digan qué hacer, por lo que también se requiere de un cambio necesario en nuestra sociedad, ya que somos un país con grandes ingenieros, los cuales solo buscan una pequeña oportunidad para realizar pequeños cambios que generen grandes ganancias.

Ahora bien, el caso a analizar es el de la Caja Popular Mexicana, que es una Cooperativa de ahorro y préstamo que ofrece servicios de ahorro, crédito e inversión. Aunque se sabe que este tipo de empresas son muy herméticas en cuanto a proporcionar información sobre sus sistemas y los cambios no son constantes y no es hasta que se dan cuenta que sus sistemas son totalmente obsoletos, cuando buscan hacer un cambio y es en este sentido que la importancia de las aplicaciones móviles está jugando un papel bastante importante en este sector, debido a que los usuarios evitan hacer filas y realizar transacciones desde su teléfono celular, lo que es muy cómodo, sin embargo representa un problema el mantener la seguridad, puesto que la delincuencia busca violar la seguridad y hacer mal uso de estas apps.

Otro problema es la comunicación de la empresa con el usuario, esto es, cuando se hace una llamada telefónica o se le envía un correo electrónico al usuario, normalmente le ofrecen otros servicios, más no le preguntan la opinión acerca del servicio que le están brindando. Sin duda, la ciberdelincuencia es un problema constante en todo el mundo y la banca, a nivel mundial, se ve amenazada por los delitos cibernéticos, entre los que destacan los robos de identidad; por lo que es necesario crear medidas de seguridad que garantice la confidencialidad de los datos de los usuarios.

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, por lo que es necesario combatir estos casos con bastantes argumentos que necesitan ser explorados por algoritmos en la minería de datos.

# **Preguntas de investigación.**

1. ¿Se requiere de una gran inversión para implementar minería de datos en una pequeña empresa?
2. ¿Se requiere de equipo especializado para la implementación de minería de datos?
3. ¿Se requiere de gente especializada para la implementación de minería de datos?
4. ¿En que se podría ver beneficiada una pequeña empresa al implementar la minería de datos?
5. ¿Podría aplicarse una versión resumida de minería de datos?

# **Hipótesis.**

**H0**. Es viable implementar los procesos de minería de datos en pequeñas empresas

**H1.** No resulta factible la implementación de los procesos de minería de datos en pequeñas empresas

# **Objetivo general.**

Identificar la viabilidad de la implementación de los procesos de la minería de datos en las PYMES de Aguascalientes, así como los posibles beneficios que podrían presentarse a corto y largo plazo.

# **Objetivos específicos.**

1. Estimar el costo de inversión mínimo requerido para la instalación del equipo necesario orientado al funcionamiento de la minería de datos.
2. Definir las características computacionales mínimas indispensables para la implementación de minería de datos en una pequeña empresa.
3. Identificar las habilidades y características necesarias que debe tener el personal que monitoreará el proceso de minería de datos.
4. Describir cómo se puede beneficiar una pequeña empresa al emplear minería de datos.
5. Identificar las fases cruciales de la minería de datos que pueden sufrir modificaciones adecuadas a las necesidades de la empresa.

# **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, a nivel mundial, grandes cantidades de recursos 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 para 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, recurriendo a diversas tácticas y estrategias, proporcionando un sentido automatizado, con el propósito de 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 sectores se pueda asignar el personal 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 focalizar la atención, en Aguascalientes, a una sucursal de “Caja Popular Mexicana”, haciendo un análisis en sus datos, que permitan ayudar a dicha empresa a mejorar en varios aspectos.

Lo primordial será el poder detectar cualquier patrón que ayude a prevenir fraudes que se pudieran presentar, ya sea por empleados, por clientes o por personas externas a la empresa.

Con la minería de datos, se detectarán a tiempo características que puedan evaluar el comportamiento de los datos que se producen a diario en la institución financiera, se obteniendo un indicador que permita identificar malas prácticas.

Otro de los propósitos del uso de la minería de datos, es el poder evaluar la satisfacción del cliente, es decir, contar con indicadores en los datos que muestren las razones por las que los 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 cuentahabientes. 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 y mejorarlo.

Gracias también a la minería de datos, la institución podrá ver qué es lo que necesitan los cuentahabientes, las expectativas que tienen de un banco, entre otros aspectos. Todos estos datos ayudarán 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, 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 está generando cada día una cantidad enorme de datos, por lo que se requiere de tecnología que ayude en el proceso de búsqueda y comprensión de los mismos. Todos hemos podido observar como en las últimas décadas ha aumentado el volumen y diversidad de los datos almacenados; así la Minería de datos surge debido a la necesidad de utilizar la gran cantidad de datos almacenados por los sistemas de información de instituciones, empresas, gobiernos y particulares durante muchos años.

En su tesis doctoral Hasperué, W. (2014), menciona que los datos pasan a ser la materia prima que hay que explotar para obtener un nuevo producto, el conocimiento. Este conocimiento se convierte en un elemento muy valioso para la ayuda en la toma de decisiones sobre el ámbito en el que se han recopilado o extraído los datos.

Como lo señala Cutro (2008), la idea de Minería de Datos no es nueva, ya desde los años sesenta los estadísticos manejaban términos como Data Fishing, Data Mining (DM) o Data Archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido.

De acuerdo con, Hasperué, W. (2014). la Estadística es la primera ciencia que consideró a los datos como su materia prima, ante las nuevas necesidades y las nuevas características de los datos (gran volumen y tipología), han surgido otras disciplinas que comienzan a integrar lo que se conoce ahora como minería de datos.

Siguiendo con el desarrollo histórico, Cutro (2008) , señala que a principios de los años ochenta, Rakesh Agrawal, GioWiederhold, Robert Blum y Gregory Piatetsky-Shapiro entre otros, empezaron a consolidar los términos de Minería de Datos y Knowledge Discovery in Databases (KDD).

El mismo autor divide a la evolución de sus herramientas en el transcurso del tiempo en cuatro etapas principales:

* Colección de Datos (1960).
* Acceso de Datos (1980).
* Almacén de Datos y Apoyo a las Decisiones (principios de la década de 1990).
* Minería de Datos Inteligente. (finales de la década de 1990).

Actualmente Hasperué, W. (2014) establece que la minería de datos tiene dos retos fundamentales:

* Trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos, confidencialidad, etc.).
* Usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil.

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 donde se utilizan 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 con 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.

# **Marco conceptual.**

**Datos:** Cifra, letra o palabra que se suministra a la computadora como entrada y la máquina almacena en un determinado formato.

**Digital:** Que suministra los datos mediante dígitos o elementos finitos o discretos.

**Bits:** Un bit es un dígito del [sistema de numeración](https://es.wikipedia.org/wiki/Sistema_de_numeraci%C3%B3n) [binario](https://es.wikipedia.org/wiki/Sistema_binario). La capacidad de almacenamiento de una memoria digital también se mide en bits, pues esta palabra tiene varios significados.

**Dispositivos Móviles:** (mobile device), también conocido como [computadora de bolsillo](https://es.wikipedia.org/wiki/Computadora_de_bolsillo) computadora de mano (palmtop o [handheld](https://es.wikipedia.org/wiki/Handheld" \o "Handheld)), es un tipo de [computadora](https://es.wikipedia.org/wiki/Computadora) de tamaño pequeño, con capacidades de procesamiento, con [conexión a Internet](https://es.wikipedia.org/wiki/Conexi%C3%B3n_a_Internet) , con [memoria](https://es.wikipedia.org/wiki/Memoria_(inform%C3%A1tica)), diseñado específicamente para una [función](https://es.wikipedia.org/wiki/Funci%C3%B3n_(programaci%C3%B3n)), pero que pueden llevar a cabo otras funciones más generales. Estrictamente hablando, muchos de los llamados dispositivos móviles no tienen la capacidad de moverse. Más bien son dispositivos que pueden ser fácilmente transportados por sus usuarios.

**Internet:** Red informática de nivel mundial que utiliza la línea telefónica para transmitir la información.

**Informática:** Conjunto de conocimientos técnicos que se ocupan del tratamiento automático de la información por medio de computadoras.

**Base de datos:** es una colección de información organizada de forma que un programa de ordenador pueda seleccionar rápidamente los fragmentos de datos que necesite. Una base de datos es un sistema de archivos electrónico. Las bases de datos tradicionales se organizan por campos, registros y archivos.

**Minería de datos:** La minería de datos o exploración de datos (es la etapa de análisis de “Knowledge Discovery in Databases” o KDD) es un campo de la estadística y las [ciencias de la computación](https://es.wikipedia.org/wiki/Ciencias_de_la_computaci%C3%B3n) referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

**Tupla:** La tupla es un tipo de dato secuencial. Sirve para agrupar, como si fueran un único valor, varios valores. El tipo de datos que representa a las tuplas se llama tuple, y es inmutable: una tupla no puede ser modificada una vez que ha sido creada.

**Carácter:** En terminología informática y de telecomunicaciones, un carácter es una unidad de información que corresponde aproximadamente con un grafema o con una unidad o símbolo parecido, como los de un alfabeto o silabario de la forma escrita de un lenguaje natural.

**Multimedia:** El término multimedia se utiliza para referirse a cualquier objeto o sistema que utiliza múltiples medios de expresión físicos o digitales para presentar o comunicar información.

**Librerías Digitales:** Se refiere a los recursos informáticos documentales, a los que se accede mediante diferentes dispositivos de comunicación que usan los servicios de Internet.

**Bioinformática:** La bioinformática es una disciplina que permite aplicar herramientas de la informática al estudio y la gestión de datos de la biología.

**Marketing:** Conjunto de técnicas y estudios que tienen como objeto mejorar la comercialización de un producto.

**Redes de Telecomunicación:** Se entiende por red de telecomunicación al conjunto de medios (transmisión y conmutación), tecnologías (procesado, multiplexación, modulaciones), protocolos y facilidades en general, necesarios para el intercambio de información entre los usuarios de la red. La red es una estructura compleja.

**Algoritmo:** Conjunto ordenado de operaciones sistemáticas que permite hacer un cálculo y hallar la solución de un tipo de problemas.

**Datawarehouse**: Un Datawarehouse es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.

**Machine Learning:** El aprendizaje automático o aprendizaje de máquinas (del inglés, “Machine Learning”) es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender.

**Confort:** Condiciones materiales que proporcionan bienestar o comodidad.

**Banca:** Pequeña empresa bancaria.

**Obsoleto:** Adjetivo [cosa] Que no se usa en la actualidad, que ha quedado claramente anticuado.

**Aplicaciones Móviles:** Una aplicación móvil es un programa que usted puede descargar y al que puede acceder directamente desde su teléfono o desde algún otro aparato móvil.

**App:** Es un programa que puede ser instalado en dispositivos móviles y computadores para que el usuario realice distintos tipos de tareas, como por ejemplo jugar, recibir noticias actualizadas y descargar archivos multimedia.

**Ciberdelincuencia:** Por ciberdelincuencia se entienden las actividades delictivas realizadas con ayuda de redes de comunicaciones y sistemas de información electrónicos o contra tales redes y sistemas por ejemplo los ataques contra los sistemas informáticos, la denegación de servicio y la piratería.

**Periodicidad:** Frecuencia con la que aparece, sucede o se realiza una cosa repetitiva.

**Optimizar:** En matemáticas e informática, determinar los valores de las variables que intervienen en un proceso o sistema para que el resultado que se obtenga sea el mejor posible.

**Fidedigna:** Que es digno de ser creído o que merece crédito.

**Computación:** El término computación tiene su origen en el vocablo en latín computatio. Esta palabra permite abordar la noción de cómputo como cuenta o cálculo, pero se usa por lo general como sinónimo de informática (del francés informatique). De esta manera, puede decirse que la computación nuclea a los saberes científicos y a los métodos. Estos sistemas automatizados de información se consiguen a través de herramientas determinadas que han sido creadas para dicho fin, los ordenadores o computadoras.

**Ciencia de Datos:** La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la [estadística](https://es.wikipedia.org/wiki/Estad%C3%ADstica), la [minería de datos](https://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos), el [aprendizaje automático](https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico) y la [analítica predictiva](https://es.wikipedia.org/wiki/An%C3%A1lisis_predictivo).

**Inteligencia Artificial:** La inteligencia artificial (IA), también llamada inteligencia computacional, es la inteligencia exhibida por máquinas.

**Big Data:** (en español, grandes datos o grandes volúmenes de datos) es un término evolutivo que describe cualquier cantidad voluminosa de datos estructurados, semiestructurados y no estructurados que tienen el potencial de ser extraídos para obtener información.

**Datos Abiertos:** Los datos abiertos son datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen.

**Modelado:** El modelado de sistemas software es una técnica para tratar con la complejidad inherente a estos sistemas. El uso de modelos ayuda al ingeniero de software a “visualizar” el sistema a construir. Además, los modelos de un nivel de abstracción mayor pueden utilizarse para la comunicación con el cliente.

**Herramientas Informáticas:** Las herramientas informáticas (tools, en inglés), son programas, aplicaciones o simplemente instrucciones usadas para efectuar otras tareas de modo más sencillo. En un sentido amplio del término, podemos decir que una herramienta es cualquier programa o instrucción que facilita una tarea.

# **CAPÍTULO II**

# **Marco contextual.**

Según la revista UNAM volumen 12, publicada en el sitio web, <http://www.revista.unam.mx/vol.12/num8/art77/index.html>, la exploración de información en espacios heterogéneos requiere métodos de minería de datos tan eficientes como sus interfaces visuales. Pero para que la minería de datos sea efectiva se necesita involucrar la mente humana ya que con la gran capacidad que posee, puede hacer exploraciones visuales para identificar los procesos o realizar los algoritmos necesarios para poder tener un mejor análisis de los datos en cuestión.

Esta revista propone que por las experiencias que se tienen se propone que se integren en los proyectos de Data Mining un Modelo de Datos Analítico (MDA), que, sin ser un Data Warehouse, pueda ser útil a los usuarios finales para una mejor explotación de lo que existe oculto en las bases de datos. Sin embargo, es importante adquirir las tecnologías y metodologías que Big Data ofrece para poder alcanzar esto.

El objetivo número uno proveer al usuario de “toda” la información que existe, que sea fácil de explotar, y que ayude a tomar decisiones más asertivas.

Uno de los hechos relevantes es que Data Mining ya ha evolucionado para venderse más como concepto que como producto. Por lo tanto, ahora se vende como mejora en la utilidad, propensión de fuga del cliente, perfilamiento del comportamiento de los defraudadores de tarjetas de crédito o como una mejora para el ajuste de parámetros en las herramientas de calificación de riesgo.

Según la revista CIO, publicada en el sitio web <http://cio.com.mx/del-data-mining-al-big-data/> en los años 90, fue cuando la minería de datos se consolidó como una tecnología de apoyo para la toma de decisiones, se han logrado grandes avances sobre todo en la parte de las comunicaciones y el manejo de grandes cantidades de información, ya que hoy en día surgen nuevos retos sobre todo en la parte de integrar información de los dispositivos móviles a la ya gran Base de Datos.

Sin embargo la revista Cubana de Ciencias Informáticas volumen 3, publicada en el sitio web <http://www.redalyc.org/pdf/3783/378343637009.pdf> menciona que la minería de datos no tiene una definición como tal, ya que la definición puede variar ya que tiene mucho que ver entre los investigadores y los fines que se le den, por ejemplo, la definición y los fines serán muy diferentes para un estadístico o para un analista de datos. La definición que se acerca más a lo general es la de: análisis de archivo y bitácoras de transacciones que trabaja a nivel de conocimiento con el fin de descubrir padrones, reglas, relaciones o excepciones útiles para la toma de decisiones.

La revista Cubana de Ciencias Informáticas, menciona también la importancia de tener los objetivos claros de búsqueda, así como concretos. Menciona también que muchas veces hay información corrompida, para la cual lo más recomendable será aplicar técnicas estadísticas las cuales medirán el nivel de confiabilidad de las relaciones halladas.

La minería de datos es un proceso de extracción de información y búsqueda de patrones de comportamiento que simple vista se ocultan entre grandes cantidades de información, existen varias técnicas y algoritmos que ayudan a obtener la información.

Según la revista Cubana de Ciencias Informáticas hay dos tipos de algoritmos:

* Supervisado o predictivos.
* No supervisados o del descubrimiento del conocimiento.

Los supervisados o predictivos, como su nombre lo dice, predicen el valor de un atributo o un conjunto de datos, que ayudan a conocer otros atributos, Mientras que los, no supervisados o del descubrimiento del conocimiento son los que descubren patrones y tendencias en los datos actuales, el descubrimiento de esa información sirve para llevar a cabo acciones y obtener beneficios de dichas acciones.

También se menciona en la revista la existencia de varias técnicas de recopilación de datos, herramientas que son usadas por minería de datos. Entre estas técnicas de recopilación se tiene:

* Almacenamiento de datos (Data Warehousing): El almacenamiento de datos se define como un proceso de organización de grandes cantidades de datos de diversos tipos guardados con organización de tal manera que esa organización facilite la recuperación con fines analíticos. El almacenamiento de datos tiene una gran importancia en el proceso de minería de datos pues en cierta medida, permite la recuperación o al menos la referencia a determinados conjuntos de datos de importancia para un proceso de toma de decisión dado. En la actualidad existe gran variedad de sistemas comerciales para el almacenamiento de datos entre los que se destacan Oracle, Sybase, MS SQL Server, entre otros.
* Análisis exploratorio de datos (Exploratory Data Analysis (EDA)): Las técnicas de análisis exploratorio de datos juegan un papel muy importante en la minería de datos. Las mismas tienen como objetivo determinar las relaciones entre las variables cuando no hay o no está totalmente definida la naturaleza de estas relaciones. Las técnicas exploratorias tienen un fuerte componente computacional abarcando desde los métodos estadísticos simples a los más avanzados como o las técnicas de exploración de multivariables diseñadas para identificar patrones en conjunto de datos multivariables.

La revista también menciona las Herramientas empleadas en el proceso de extracción de minería de datos las cuales se clasifican en:

* Técnicas de verificación: El sistema se limita a comprobar hipótesis suministradas.
* Método de descubrimiento: Es en la que se encuentran patrones de interesantes, así como técnicas de predicción, todo esto mediante la exploración.

La revista concluye que la minería de datos se presenta como tecnología de apoyo, para poder explorar, analizar, comprender y aplicar el conocimiento obtenido usando grandes volúmenes de datos y descubrir patrones que ayuden a la identificación de estructuras de datos. Los productos a comercializar son costosos y requieren de mucha experiencia en su utilización. Es muy fácil hallar patrones equívocos o no interesantes. La aplicación de estas herramientas ayuda en el proceso de toma de decisión en de las organizaciones.

Según la revista Proceso, publicada en el sitio web, <http://www.proceso.com.mx/404432/el-data-mining-invade-mexico-a-costa-de-los-usuarios>, los gobiernos y empresas recurren cada vez más al análisis masivo de datos, y suele hacerse ese análisis en función a sus datos específicos, por lo que practicar estas técnicas pueden afectar la privacidad de los datos de las personas.

La información solicitada por empresas o corporaciones es algo que se hace diariamente. Dichas empresas en México utilizan esta información para distintos fines uno de ellos es la publicidad, a manera de búsqueda de patrones, pueden visualizar los gustos y preferencias de los consumidores.

De igual manera, la revista Proceso menciona que pueden obtener las empresas de manera muy fácil, es decir que la tiene a la mano para sus fines laborales. Por lo que se menciona también en la revista la importancia de entrar en territorio fiscal para poder penalizar todas las incidencias que presenten.

# **CAPÍTULO III**

# **Diseño de investigación.**

La presente investigación busca que las PYMES del estado de Aguascalientes en México, conozcan las nuevas técnicas de generación, procesamiento e interpretación de información como lo es la minería de datos, teniendo así la posibilidad de implementarla con facilidad en un futuro cercano y beneficiarse así de los resultados que se proveen a través de los algoritmos y procesos diseñados especialmente para la empresa.

Esta investigación emplea los conocimientos básicos de una rama importante dentro de la ciencia de datos: la minería de datos; esto para averiguar si las PYMES promedio de Aguascalientes cuentan con los requisitos e infraestructura necesaria para poner en marcha la misma.

Cabe destacar que la investigación presentada es una investigación exploratoria con enfoque cualitativo, ya que se ofrecen un primer acercamiento al problema que se pretende estudiar y conocer. Así mismo, se realiza para conocer el tema que se abordará, lo que nos permita “familiarizarnos” con algo que hasta el momento desconocíamos. Los resultados de la investigación pretenden darnos un panorama o conocimiento superficial del tema.

El alcance que tiene esta investigación es descriptiva, ya que a través de los resultados se pretenderá detallar

Este trabajo no realiza manipulación de variables, por ende, se afirma que la investigación posee un diseño no experimental.

# **Metodología de la investigación.**

Como técnica para la recolección de datos se optó por realizar una encuesta diseñada para conocer la perspectiva del empleado acerca de la situación actual de la empresa en cuanto a recopilación, procesamiento y tratado de información se refiere.

Para esto el instrumento recaba la información descriptiva del usuario con el fin de realizar una interpretación de los resultados más acertada encontrando posibles patrones o tendencias en cuanto a la carrera procedente del trabajador, el tiempo que tiene laborando en la empresa o su edad.

La escala de Likert de acuerdo a Sánchez (1993), es una escala psicométrica comúnmente utilizada en cuestionarios y es la escala de uso más amplio en encuestas para la investigación. Al responder a una pregunta de un cuestionario elaborado con la técnica de Likert, se especifica el nivel de acuerdo o desacuerdo con una declaración (elemento, ítem o reactivo o pregunta). A pesar de esta limitación la escala de Likert tiene la ventaja de ser más sencillo de construir y de aplicar, con la ventaja de que proporciona una base para una primera ordenación de los individuos.

Para nuestro caso, la escala de Likert utilizada fue la siguiente.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Totalmente de acuerdo | De acuerdo | Ni de acuerdo ni en desacuerdo | En desacuerdo | Totalmente en desacuerdo |

# **Población y muestra.**

Para el caso de la investigación, el universo se encuentra compuesto por todas las pequeñas y medianas empresas del estado de Aguascalientes, que de acuerdo con los censos económicos realizados por el INEGI (2014), éste está conformado por un total de 2,536.

En estadística el tamaño de la muestra es el número de sujetos que componen la muestra extraída de una población, necesarios para que los datos obtenidos sean representativos de la población (Wiley, 1962). El tamaño de la muestra se calculó con 90% de nivel de confianza con un margen de error del 10%. Por lo tanto, de la fórmula para el cálculo del tamaño de la muestra:

Substituyendo los valores:

Obtenemos que:

# **CAPÍTULO IV**

# **Propuesta de instrumento.**

Para hacer una aproximación de la realidad en torno a la información en las PYMES de Aguascalientes, el instrumento está dirigido a los estudiantes de ingeniería en sistemas computacionales o ingeniería en computación inteligente que trabajen en alguna pequeña o mediana empresa del estado, ya que al tener una sólida información en el área tecnologías de la computación, pueden brindar un panorama más acertado de la situación actual en la organización.

Así mismo recaba las opiniones acerca de si la empresa emplea software especializado para la limpieza y tratamiento de sus datos, si el hardware empleado en la empresa facilita el tratamiento de los datos, si se cuenta con un almacén apropiado para los datos, si la empresa le interesa guardar la información importante, si la empresa maneja cotidianamente un gran volumen de datos, si el volumen de datos que se tratan en la empresa son manejados adecuadamente, si el manejador de datos que se utiliza en la empresa es el apropiado, si la empresa maneja una gran variedad de tipos de datos, si los datos de la empresa se encuentran con una estructura adecuada, si la información se encuentra respaldada.

# **Alcance del instrumento.**

Al diseñar el presente instrumento, tuvo que tomarse en cuenta hacia qué tipo de usuarios iba dirigido, y hacia qué sector se deseaba enfocar, para poder así redactar las preguntas necesarias que ayudaran a interpretar de manera intuitiva los patrones y tendencias que pudieran presentarse.

# **Resultados.**

A continuación, se presentarán los resultados de la encuesta aplicada. Comenzando por mostrar las gráficas correspondientes a la información descriptiva del usuario que pondrá en contexto las poblaciones que intervinieron en el instrumento.

Primero, se muestra el resultado de las proporciones de las carreras que participaron

|  |  |  |
| --- | --- | --- |
| Carrera | Frecuencia | Porcentaje |
| Ingeniería en Sistemas Computacionales | 50 | 75.7575758% |
| Ingeniería en Computación Inteligente | 16 | 24.2424242% |
| Total | 66 | 100% |

a continuación, se muestran los rangos de edades obtenidas

|  |  |  |
| --- | --- | --- |
| Rango | Frecuencia | Porcentaje |
| 18 - 21 | 10 | 15.1515152 |
| 21 - 24 | 37 | 56.0606061 |
| 24 - 27 | 16 | 24.2424242 |
| 27 - 30 | 2 | 3.03030303 |
| 30+ | 1 | 1.51515152 |
| Total | 66 | 100 |

# **Conclusiones.**

La tendencia de generar grandes cantidades de datos, así como los formatos de los mismos es un hecho en la sociedad dependiente de los sistemas computacionales es cada vez más evidente,

Introducción del tema, breve recapitulación

Hablar de los resultados

Hacer hincapié que se hizo entorno a Aguascalientes.

Consideramos que los resultados obtenidos en la presente investigación la situación de las pequeñas empresas.

Hablar de la importancia de los resultados de la tesina con los empresarios u organizaciones

Esperamos que el resultado de la presente investigación contribuya a entender mejor los sistemas de explotación de bases de datos, además esperamos que sirva para futuras investigaciones ya que somo se mencionó a lo largo del trabajo la importancia.

# **Recomendaciones.**