

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.**

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. 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.

Con el importante progreso en la 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 obtener 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 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 suficientes para analizar esta vasta colección de datos heterogéneos.

Actualmente, se 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 y el internet de las cosas 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, por lo que se prevé una importante inversión en esta área de desarrollo en los próximos años.

La información es 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 acuerdo con Martínez-Luna (2011) las instituciones y empresas privadas coleccionan bastante información (ventas, clientes, cobros, pacientes, tratamientos, estudiantes, calificaciones, fenómenos meteorológicos, etcétera, según su giro), aprovechando que las computadoras y los discos de almacenamiento se han abaratado, y las comunicaciones son también baratas y confiables.

Si esta información no se analiza de manera adecuada se corre el riesgo de afectar la toma de decisiones.

Descubrir conocimiento de este enorme volumen de datos es un reto en sí mismo. La minería de datos es un intento de buscarle sentido a la explosión de información que actualmente puede ser almacenada.

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.

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 las empresas pequeñas y medianas (PYMES) se fortalezcan, ofreciendo más y mejores servicios que la competencia, contribuyendo al desarrollo del estado de Aguascalientes y del país. .

# **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, puesto que 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 México sí, debido a que existen empresas que no confían en los propios cambios del país, por lo que buscan a personas de más allá de las fronteras 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 las empresas pequeñas y medianas (PYMES), mismas que cumplen con un importante papel en la economía de todos los países. De acuerdo con el INEGI, (2009), las empresas micro, pequeñas y medianas, representan a nivel mundial el segmento de la economía que aporta el mayor número de unidades económicas y personal ocupado, llegando al 90%, de las unidades económicas totales.

Según datos del INEGI (2014) existían en el país 5’654,014 establecimientos, de los cuales el 4% corresponden a las PYMES.

La misma fuente registra para el estado de Aguascalientes, un total de 47,449 PYMES (5.4% del total de las empresas en el estado) y muchas de ellas no están aprovechando todo el potencial que les ofrece la Minería de Datos.

Las PYMES tienen grandes ventajas, como su capacidad de adaptabilidad gracias a su estructura pequeña, su posibilidad de especializarse en cada nicho de mercado, ofreciendo un tipo de atención directa y finalmente su capacidad comunicativa.

La mayor ventaja de una PYME es su capacidad de cambiar rápidamente su estructura productiva en el caso de variar las necesidades de mercado, lo cual es mucho más difícil en una gran empresa, con un importante número de empleados y grandes sumas de capital invertido. Sin embargo, el acceso a mercados tan específicos o a una cartera reducida de clientes, aumenta el riesgo de quiebra de estas empresas, por lo que es importante que amplíen su mercado o sus clientes.

En este contexto, la minería de datos se constituye en una herramienta para fortalecerlas.

# **Preguntas de investigación.**

1. ¿Las PYMES en Aguascalientes están generando grandes volúmenes de información?
2. ¿Se requiere de una gran inversión para implementar minería de datos en una pequeña empresa?
3. ¿Se requiere de equipo especializado para la implementación de minería de datos?
4. ¿Se requiere de personal especializado para la implementación de minería de datos?
5. ¿En qué se podrían ver beneficiadas las pequeñas y medianas empresas al implementar la minería de datos?

# **Hipótesis.**

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

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

# **Objetivo general.**

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

# **Objetivos específicos.**

1. Identificar si las pequeñas y medianas empresas de Aguascalientes generan grandes volúmenes de datos.
2. Identificar las necesidades de estructura de los datos de las pequeñas y medianas empresas en Aguascalientes.
3. Identificar la generación de diferentes tipos de datos en las pequeñas y medianas empresas en Aguascalientes.
4. Identificar el uso de software especializado para el manejo de datos.

1. Identificar las habilidades y características necesarias que debe tener el personal que monitoreará el proceso de minería de datos.
2. Describir cómo se puede beneficiar una pequeña empresa al emplear 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, 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.

Con la minería de datos, se detectarán a tiempo características que puedan evaluar el comportamiento de los datos que se producen a diario en las PYMES en Aguascalientes, obteniendo indicadores que permitan 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 empresa y cambiar a otra.

Con esto se pretende que la institución mejore en su atención al cliente, decidir cómo si es necesario ofrecer más. 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 clientes, las expectativas que tienen de la empresa, entre otros aspectos. Todos estos datos ayudarán también a conseguir nuevos clientes y a elevar su satisfacción.

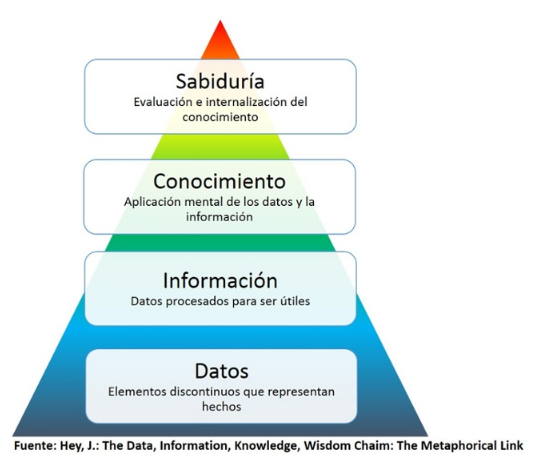
# **CAPÍTULO I**

# **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” (Navarro, 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 es un concepto relativamente nuevo que se utiliza a menudo de manera intercambiable con inteligencia o análisis de negocios competitivos. Liu, (2012), menciona que 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.

El mismo autor define a la ciencia de datos como 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).

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.

KDD es el proceso general y la minería de datos forma parte de él. 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 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.

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.**

En este apartado se abordan los términos sustantivos que dan soporte al desarrollo del presente trabajo, buscando dejar en claro los conceptos utilizados tanto en el ámbito de la Minería de Datos, así como en el de las Pequeñas y Medianas Empresas (PYMES).

Minería de Datos (Data Mining)

La Minería de Datos es una disciplina de carácter multidisciplinar, ya que recurre a la estadística en virtud de que es una ciencia que comprende una serie de métodos y procedimientos destinados a la recopilación, tabulación, procesamiento, análisis e interpretación de datos cuantitativos y cualitativos; de igual manera trabaja con bases de datos, en donde almacena, estructura, estandariza y explota una gran cantidad de información, recurriendo también a la Inteligencia Artificial, que representa un conjunto de disciplinas de software, lógica, informática y filosofía que están destinadas a hacer que las computadoras realicen funciones que anteriormente se pensaba que eran exclusivamente humanas.

Al revisar las fuentes documentales se observa que la definición de Minería de datos 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 citada por Rodríguez Suarez (2009) “*La minería de datos es el análisis de habitualmente grandes, series de datos para encontrar relaciones inesperadas y resumir la información de nuevas maneras que sean entendibles y útiles por el propietario de los datos”.*

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. Se ha llegado a decir que sin la estadística no existiría la Minería de Datos.

# Otro término íntimamente ligado a la minería de datos es la Inteligencia Artificial, que es una rama de la informática, encargada de la simulación de procesos de inteligencia humana por parte de máquinas, especialmente sistemas informáticos. Estos procesos incluyen el aprendizaje (la adquisición de información y reglas para el uso de la información), el razonamiento (usando las reglas para llegar a conclusiones aproximadas o definitivas) y la autocorrección; y Machine Learning, que es una rama de la Inteligencia Artificial, cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender.

Así mismo es preciso hacer mención del término Big Data, de los que Maté (2014) menciona que según el diccionario Oxford:  *“datos de tamaño muy grande, típicamente hasta el extremo de que su gestión presenta retos logísticos significativos”*, y cita que Gartner (2012) definió como *“activos de información caracterizados por su volumen elevado, velocidad elevada y alta variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y la toma de decisiones en las organizaciones”.* Esta definición destaca las 3 “V” de los Big Data: Volumen, Velocidad y Veracidad.

# La Minería de Datos recurre al Big Data, cuando realiza procesos de análisis, con grandes volúmenes de datos. Con ayuda de Data Mining, se realiza la extracción de grandes volúmenes de datos, buscando que estos sean no triviales, es decir, que en realidad aporten elementos para el análisis; que sean implícitos, al contener información persé; previamente desconocidos, porque no se tenía conciencia de que existían; además de potencialmente útiles, ya que su simple obtención puede redundar en un beneficio para quien logro extraer el dato.

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.

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.

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.

Rodríguez Suárez (2009) define el Data warehousing como *“un proceso de organización de grandes cantidades de datos de diversos tipos guardados en la organización con el objetivo de facilitar la recuperación de la misma 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.

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.

En la actualidad existe una fuerte tendencia a trabajar con datos abiertos, también conocidos como Open Data, que son todos aquellos datos primarios (sin procesar) que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos.

Los datos, una vez que se han vuelto públicos, se convierten en un gran reto para la Minería de Datos, ya que es factible el identificar patrones y obtener conclusiones, que permitan llegar a resultados, para una adecuada toma de decisiones.

En esencia, la minería de datos analiza datos en bruto, obteniendo conclusiones a partir de la aplicación de algoritmos desarrollados para el análisis de grandes volúmenes de información, identificando patrones encubiertos y estableciendo relaciones escondidas, todo ello a partir de inferencias.

Para llevar a cabo el análisis de los datos, se recurre al desarrollo de algoritmos que permiten, mediante un proceso automatizado, la identificación de patrones a partir de los cuales es posible extraer datos potencialmente útiles. Estos algoritmos pueden ser predictivos, descriptivos, de segmentación o exploratorios

El término algoritmo, se define como una secuencia de instrucciones que representan un modelo de solución para determinado tipo de problemas; es decir, corresponde a un conjunto de instrucciones que, realizadas en orden, conducen a obtener la solución de un problema (Leiserson, 2009).

La Minería de Datos requiere de un hardware robusto, entendiendo el concepto de Hardware como la parte “dura” de la informática, es decir, la maquinaria real utilizada para el procesamiento electrónico de los datos, requiere de variados y robustos equipos y es aquí donde se hace presente el concepto de cluster, que es un conjunto de máquinas funcionando como unidad y trabajando juntas para tratar una única tarea.

De igual forma, el Data Mining recurre a software, que es un conjunto de programas y rutinas que permiten a la computadora realizar determinadas tareas; resaltando que existe software especializado en explotar grandes volúmenes de datos a una gran velocidad de procesamiento.

Entre las herramientas informáticas más utilizadas en la Minería de Datos, estan: SAS Analytics, Clementine, Rapminder, SQL Server, OLAP, Query Tools, por mencionar sólo algunas. Con ellas es posible procesar grandes volúmenes de información, filtrando diversos tipos de datos y obteniendo conclusiones a partir de los patrones descubiertos.

Es imprescindible incluir un concepto sumamente abstracto, en virtud de los avances de la tecnología y a la tendencia cada vez mayor de realizar procesos remotos, con ayuda del internet y de robustos servidores, con los que se pretende evitar la saturación de memoria de los equipos de las empresas y organizaciones; y ese concepto es la nube, Cloud Computing, por su nombre en inglés, que la empresa International Business Machines (IBM) define como un nuevo modelo de consumo y prestación de servicios, adaptable a una demanda no previsible y que ofrece al consumidor la posibilidad de seleccionar y operar, de forma directa, a una serie de servicios a los que es posible acceder de manera remota. Sin duda, la Minería de Datos recurre cada vez más con mayor frecuencia a la Nube.

Una parte crucial en la Minería de Datos es el proceso que se sigue, para extraer información relevante y obtener un conocimiento que se encontraba escondido dentro de un gran volumen de información; y cuando se habla de un proceso, se hace referencia a una serie de pasos sistematizados para lograr un fin. En el siguiente esquema se muestra el proceso que sigue la Minería de Datos, para extraer valiosa información para sus propios fines.

**Imagen que contiene captura de pantalla

Descripción generada con confianza alta**

La información puede ser obtenida mediante asociaciones, secuencias, clasificaciones, agrupamientos o bien para realizar pronósticos.

El proceso de la Minería de Datos debe cumplir con los siguientes pasos: selección, pre procesado, selección de características, extracción de conocimientos, evaluación y conocimiento.

Es importante hacer una evaluación en donde se obtengan medidas de precisión, de explicación o de integración, a fin de que se analice si los datos obtenidos son: comprensibles, válidos, con certidumbre y potencialmente útiles.

La Minería de Datos tiene aplicaciones en diversos ámbitos, como en el Gobierno, en la Mercadotecnia, como parte de la Inteligencia de Negocios, en la Academia, en la Banca y en diversas empresas, que manejan grandes volúmenes de información.

El presente trabajo orienta la aplicación de la Minería de Datos hacia las pequeñas y medianas empresas, identificadas con el acrónimo de PYMES, quienes tienen la necesidad de trabajar grandes volúmenes de datos.

# **CAPÍTULO II**

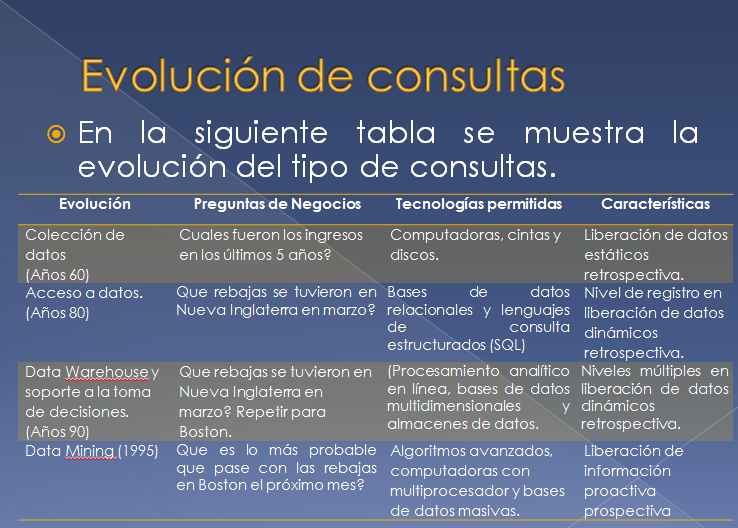
# **Marco contextual.**

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.

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 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.

La evolución de sus herramientas en el transcurso del tiempo puede dividirse 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.

Existen numerosas áreas donde la minería de datos se puede aplicar, prácticamente en todas las actividades humanas que generen datos, Riquelme Santos (2006):

* Comercio y banca: segmentación de clientes, previsión de ventas, análisis de riesgo.
* Medicina y Farmacia: diagnóstico de enfermedades y la efectividad de los tratamientos.
* Seguridad y detección de fraude: reconocimiento facial, identificaciones biométricas, accesos a redes no permitidos, etc.
* Recuperación de información no numérica: minería de texto, minería web, búsqueda e identificación de imagen, video, voz y texto de bases de datos multimedia.
* Astronomía: identificación de nuevas estrellas y galaxias.
* Geología, minería, agricultura y pesca: identificación de áreas de uso para distintos cultivos o de pesca o de explotación minera en bases de datos de imágenes de satélites
* Ciencias Ambientales: identificación de modelos de funcionamiento de ecosistemas naturales y/o artificiales (p.e. plantas depuradoras de aguas residuales) para mejorar su observación, gestión y/o control.
* Ciencias Sociales: Estudio de los flujos de la opinión pública. Planificación de ciudades: identificar barrios con conflicto en función de valores sociodemográficos.

Ejemplos de la aplicación de la minería de datos en las PYMES, Martínez- Luna (2011)**:**

* Mejorar los servicios o productos que se ofrecen. Esto es posible si se registra en la bodega el detalle de la respuesta a la compra por parte de los clientes al haber cambios en los productos o servicios, en cuanto a si se incrementa o se disminuye la venta.
* Evitar situaciones no deseadas, como la de perder clientes en servicios contratados. Estas situaciones se pueden prevenir, ya que se tiene el historial de la facturación de un servicio contratado, como el teléfono, al igual que los clientes que tienen el antecedente de que se han quejado por el servicio, los periodos de tiempo en que su número de llamadas decrece, y los que han cancelado su contrato en condiciones similares. También se debe tener datos de clientes que se han logrado retener y con qué estrategias se logró, al igual que el costo de cada estrategia. Se busca retener clientes, dado que es más barato mantenerlos que ganar nuevos clientes.
* No manufacturar productos que en un futuro ya no se venderán. Se pueden predecir cambios en los gustos de los consumidores, dado que con el historial de ventas se detectan las características de los productos que se dejan de vender.
* Detectar productos de temporada. Una tienda comercial vende sus productos y registra la fecha de venta. Al revisar sus ventas por largos periodos, puede saber con precisión el intervalo de fechas en que algunos de estos productos tienen un alto volumen de ventas, y con esta información tomar una serie de decisiones alrededor de este comportamiento: cuáles productos comprar y ofrecer, cuándo pedir los productos para tenerlos disponibles, qué cantidad solicitar y almacenar para esas ventas con el fin de no tener sobrantes, realizar la publicidad apropiada para su venta, y en qué lugares ofrecer los productos o servicios.
* Conocer productos o servicios que se pueden vender en forma conjunta. Al revisar el historial de las ventas se identificarán los productos que coinciden en su venta conjunta, y con las estadísticas se seleccionarán los conjuntos de productos que coinciden en alto porcentaje, definido por el usuario interesado

# **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 |

A continuación, se muestran la frecuencia de sexo en los encuestados.

|  |  |  |
| --- | --- | --- |
| Sexo | Frecuencia | Porcentaje |
| Hombre | 42 | 63.6363636 |
| Mujer | 24 | 36.3636364 |
| Total | 66 | 100 |

Finalmente, para el contexto de los usuarios, se despliega la información respecto al tiempo en la empresa del usuario

|  |  |  |
| --- | --- | --- |
| Tiempo en la empresa | Frecuencia | Porcentaje |
| 1 - 3 Meses | 15 | 22.7272727 |
| 3 - 6 Meses | 12 | 18.1818182 |
| 6 Meses - 1 Año | 8 | 12.1212121 |
| Más de 1 Año | 31 | 46.969697 |
|  | 66 | 100 |

Para la pregunta “La empresa emplea software especializado para la limpieza y tratamiento de sus datos”, se obtuvo el siguiente recuento.

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 10 | 15.15151515 |
| De acuerdo | 24 | 36.36363636 |
| Ni de acuerdo ni en desacuerdo | 15 | 22.72727273 |
| En desacuerdo | 11 | 16.66666667 |
| Totalmente en desacuerdo | 6 | 9.090909091 |
| Total | 66 | 100 |

Por lo que se puede observar, más de la mitad de los encuestados se inclina por la percepción de que la empresa donde se encuentran cuenta con software especializado para la limpieza y tratamiento de la información, mientras que poco más del 22% se mantiene neutro y más del 25% de la población tiene cierto nivel de desacuerdo, por lo que podemos inferir que la mayoría de las PYMES en Aguascalientes cuentan con los programas adecuados para el procesamiento necesario de la información

El hardware empleado en la empresa facilita el tratamiento de los datos

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 15 | 22.72727273 |
| De acuerdo | 25 | 37.87878788 |
| Ni de acuerdo ni en desacuerdo | 16 | 24.24242424 |
| En desacuerdo | 4 | 6.060606061 |
| Totalmente en desacuerdo | 6 | 9.090909091 |
| Total | 66 | 100 |

se puede apreciar en los datos obtenidos, que la mayoría de los usuarios encuentra el hardware utilizado por la empresa una herramienta de utilidad y no como un obstáculo que dificulte las tareas cotidianas; hecho que es de vital importancia si se planea implementar los procesos de minería de datos en la misma a pesar de que se observa que un pequeño porcentaje de los usuarios encuentra el hardware actual como un impedimento.

Se cuenta con un almacén apropiado para los datos.

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 20 | 30.3030303 |
| De acuerdo | 19 | 28.78787879 |
| Ni de acuerdo ni en desacuerdo | 12 | 18.18181818 |
| En desacuerdo | 7 | 10.60606061 |
| Totalmente en desacuerdo | 8 | 12.12121212 |
| Total | 66 | 100 |

podemos observar que de acuerdo a los usuarios encuestados, la mayor parte aprecia que el almacén designado por la empresa para el respaldo de los datos es el adecuado, punto fundamental para el funcionamiento óptimo de la minería de datos, mientras que una pequeña parte de la misma opina que no es la mejor opción y podría complicar el progreso.

La empresa le interesa guardar la información importante

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 27 | 40.90909091 |
| De acuerdo | 31 | 46.96969697 |
| Ni de acuerdo ni en desacuerdo | 3 | 4.545454545 |
| En desacuerdo | 1 | 1.515151515 |
| Totalmente en desacuerdo | 4 | 6.060606061 |
| Total | 66 | 100 |

se puede observar que las opiniones acerca de que la empresa guarda la información sensible es dominante, siendo este un puto vital en la planeación para implementar minería de datos ya que uno de los requisitos más importantes es que se tenga la información completa, ya que en caso contrario se podrían generar malas interpretaciones de la misma, perjudicando a la organización el vez de beneficiarla.

La empresa maneja cotidianamente un gran volumen de datos.

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 26 | 39.39393939 |
| De acuerdo | 23 | 34.84848485 |
| Ni de acuerdo ni en desacuerdo | 14 | 21.21212121 |
| En desacuerdo | 3 | 4.545454545 |
| Totalmente en desacuerdo | 0 | 0 |
| Total | 66 | 100 |

observamos que el punto de vista de los encuestados se inclina en gran medida a que la presente empresa maneja grandes volúmenes de información cotidianamente, por lo que ayuda a reforzar la idea de que puede implementarse minería de datos en las PYMES, ya que entre mayor sea la cantidad de datos que sean manejados por la empresa, se obtendrán modelados e interpretaciones más exactas de la realidad.

El volumen de datos que se tratan en la empresa son manejados adecuadamente.

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 13 | 19.6969697 |
| De acuerdo | 23 | 34.84848485 |
| Ni de acuerdo ni en desacuerdo | 19 | 28.78787879 |
| En desacuerdo | 5 | 7.575757576 |
| Totalmente en desacuerdo | 6 | 9.090909091 |
| Total | 66 | 100 |

a pesar de que se percibe que más de la mitad de los usuarios consideran que el volumen de datos tratado por la empresa es el adecuado, una importante parte de los encuestados opinan que el trato de los datos ya no se adecúa a las necesidades, por lo que la minería de datos podría ayudar en esta parte tan crucial al brindar un mejor entendimiento de un volumen tan masivo de datos.

El manejador de datos que se utiliza en la empresa es el apropiado (emplea el paradigma adecuado, es estable, permite la facilidad de expansión etc).

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 17 | 25.75757576 |
| De acuerdo | 19 | 28.78787879 |
| Ni de acuerdo ni en desacuerdo | 20 | 30.3030303 |
| En desacuerdo | 5 | 7.575757576 |
| Totalmente en desacuerdo | 5 | 7.575757576 |
| Total | 66 | 100 |

se puede observar que los trabajadores perciben en su mayoría que el manejador de datos se adecúa a las necesidades para las que fue diseñado originalmente, pero también un gran porcentaje de los usuarios opina un cierto descontento, que puede ser causado por un enfoque mal planteado para la base de datos, la poca estabilidad o dificultad de tratar con los datos, por lo que, si se desea implementar minería de datos en un futuro, podría complicar su funcionamiento al adecuarlo a todas las limitantes de manejo o acceso de datos presentes.

La empresa maneja una gran variedad de tipos de datos (numéricos, cadenas de texto, videos, imágenes, etc.)

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 24 | 36.36363636 |
| De acuerdo | 22 | 33.33333333 |
| Ni de acuerdo ni en desacuerdo | 16 | 24.24242424 |
| En desacuerdo | 4 | 6.060606061 |
| Totalmente en desacuerdo | 0 | 0 |
| Total | 66 | 100 |

se percibe que la mayoría de los encuestados considera que la base de datos con la que trabaja la organización donde se encuentra maneja un catálogo de datos muy heterogéneo, por lo que, si se desea implementar minería de datos, debe de considerarse el tipo de análisis que se desea realizar y ver así los tipos de datos que se verán involucrados con el fin de generar resultados puntuales y específicos.

Los datos de la empresa se encuentran con una estructura adecuada.

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 15 | 22.72727273 |
| De acuerdo | 24 | 36.36363636 |
| Ni de acuerdo ni en desacuerdo | 17 | 25.75757576 |
| En desacuerdo | 10 | 15.15151515 |
| Totalmente en desacuerdo | 0 | 0 |
| Total | 66 | 100 |

de acuerdo con los resultados obtenidos a partir de las encuestas se percibe que más de la mitad de los usuarios consideran que los datos manejados por la empresa se tienen organizados con la estructura apropiada, mientras que una parte importante opina que dicha aseveración no se cumple del todo, por lo tanto, al implementarse la minería de datos en un futuro, debe de ser considerado ya que como se mencionó anteriormente, para un desempeño ideal de la misma, se debe contar con un buen diseño de la base de datos.

La información se encuentra respaldada.

|  |  |  |
| --- | --- | --- |
| Respuesta | Frecuencia | Porcentaje |
| Totalmente de acuerdo | 23 | 34.84848485 |
| De acuerdo | 28 | 42.42424242 |
| Ni de acuerdo ni en desacuerdo | 11 | 16.66666667 |
| En desacuerdo | 4 | 6.060606061 |
| Totalmente en desacuerdo | 0 | 0 |
| Total | 66 | 100 |

finalmente, se puede apreciar que los encuestados en su gran parte consideran que la información de la empresa se encuentra respaldada, siendo de vital importancia no solo para la minería de datos en si, sino que para la seguridad, disponibilidad y confiabilidad de los datos de la misma.

# **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.**