San Marcos

MIEMBRO DE LA RED ILUMNO

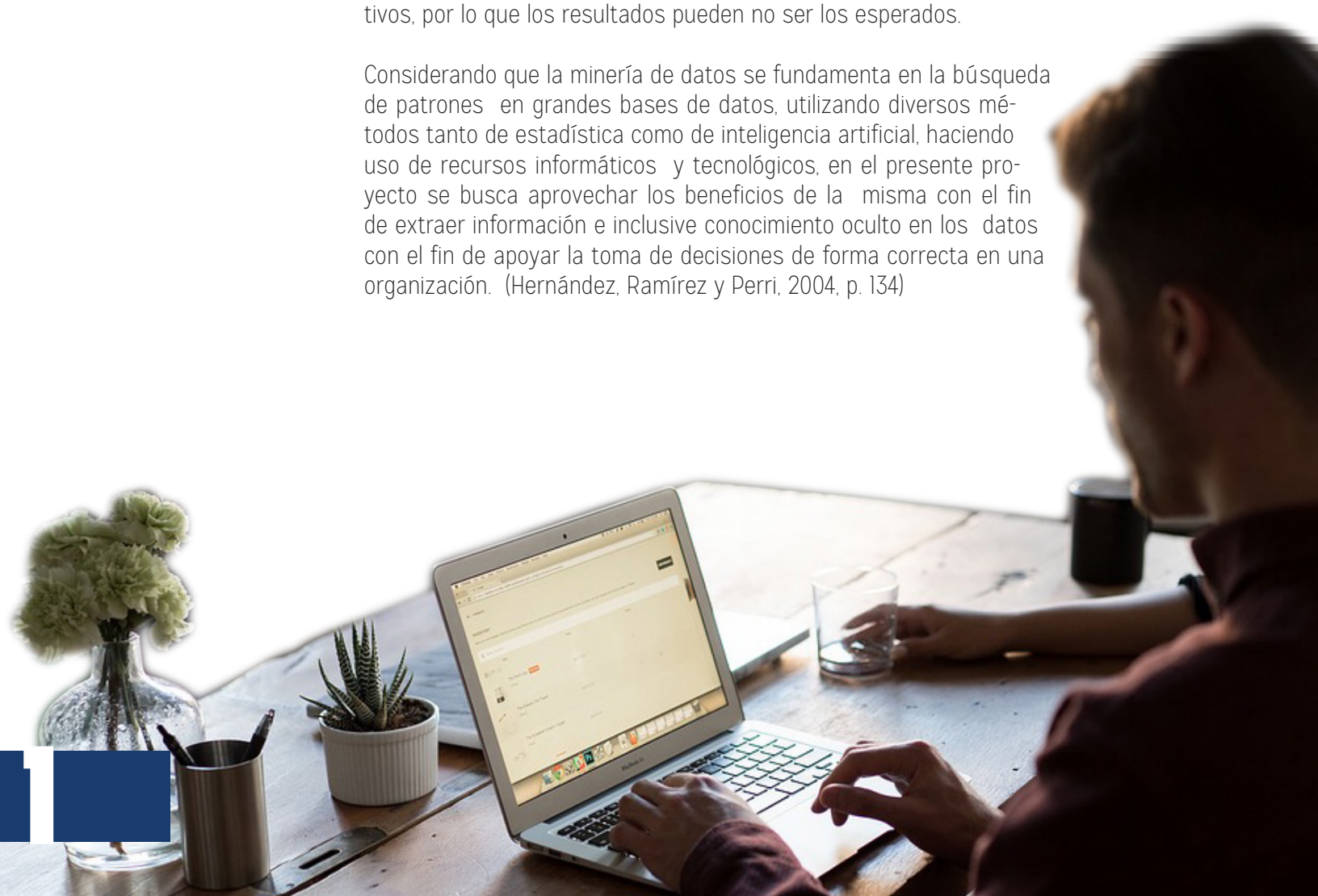
PRESENTACIÓN DE PROYECTOS

PRESENTACIÓN DE PROYECTOS

EVALUACIÓN, ADMINISTRACIÓN Y PRESENTACIÓN DE MODELOS DE MINERÍA DE DATOS EN PROYECTOS FINALES

Se puede decir que la recolección y almacenamiento de datos ha sido una de las tareas más basicas en todo tipo de empresas, puesto que se hace necesario contar con un histórico de los movimientos comerciales que realizan en una organización para poder llegar a controlar dichos movimientos, pero hasta ahora en muchas de estas empresas este control se lleva de una manera muy arcaica o con métodos que no son muy efectivos, por lo que los resultados pueden no ser los esperados.

Considerando que la minería de datos se fundamenta en la búsqueda de patrones en grandes bases de datos, utilizando diversos métodos tanto de estadística como de inteligencia artificial, haciendo uso de recursos informáticos y tecnológicos, en el presente proyecto se busca aprovechar los beneficios de la misma con el fin de extraer información e inclusive conocimiento oculto en los datos con el fin de apoyar la toma de decisiones de forma correcta en una organización. (Hernández, Ramírez y Perri, 2004, p. 134)



RECALIBRACIÓN DE MODELOS

> Cuando a los datos previamente recolectados y almacenados se les da un trato adecuado, es posible aplicar sobre estos diversas metodologías, entre las cuales se encuentran las técnicas de minería de datos, de tal forma que estas permitan conocer el comportamiento de los inventarios o las posibles relaciones que se presenten entre dos o más productos. (Morand et ál, 2004, p123)

Básicamente la presentación de un proyecto persigue un fin, pero no siempre es el mismo en cada organización. Puede ser una presentación a clientes, a inversores, a colaboradores, otros. Lógicamente la preparación de la presentación depende de esa circunstancia porque a cada uno de estos le puede interesar un aspecto diferente del proyecto y debemos adaptar el discurso para ofrecerles lo que les interesa. Partimos de la base de que un proyecto debe estructurarse en "el qué" (lo que vamos a hacer durante el proceso), "el cómo" (cómo vamos a hacerlo), "el quién" (quien está detrás del proyecto), "el cuánto" (cuánto nos va a costar y los beneficios que obtendremos). A la hora de preparar una presentación debemos plantearnos cuáles de estos aspectos interesan a nuestro auditorio para que nuestro mensaje cubra sus expectativas. Por supuesto también es relevante el formato. Podemos tener que realizar una presentación durante un tiempo limitado o tener libre disponibilidad de tiempo y también es importante saber de qué medios de apoyo podemos disponer para reforzar nuestro discurso. La aplicación de este modelo ayuda de tal forma que se puede encontrar a partir de los datos información que hasta el momento había sido desconocida, además de que dicha información obtenida ayuda en la toma de decisiones o al desarrollo de algún proceso a nivel empresarial. (Padmanabhan, 1999, p. 423)

Desde un punto de vista más pragmático y asociándolo directamente a las actividades de negocios, la minería de datos es el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información no estructurada (interna y externa a la compañía) en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento y así dar soporte a la toma de decisiones sobre el negocio.

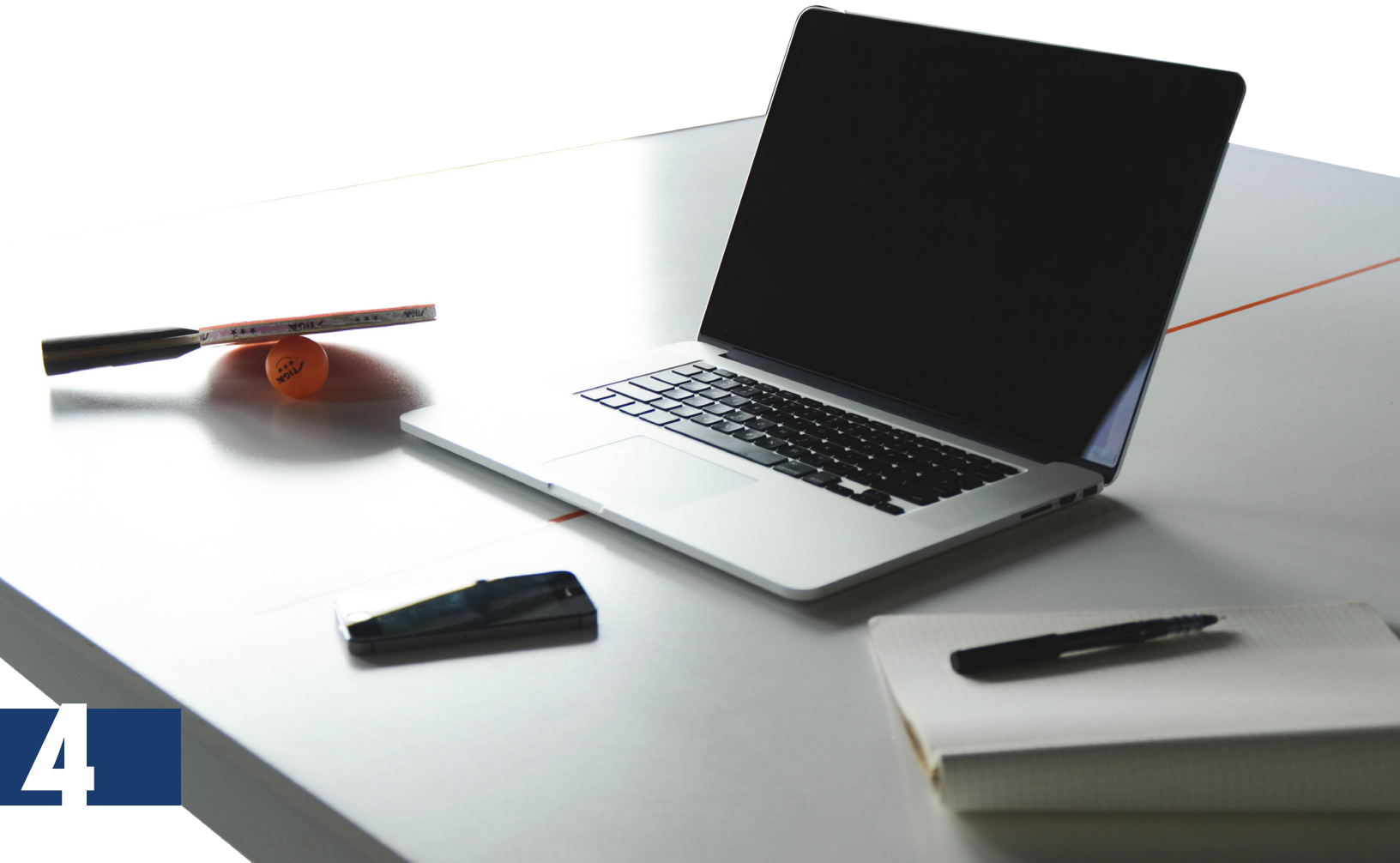
Ahora bien, tomando a Piatesky y Shapiro (1991, p. 234).

> Destacan que desde un punto de vista más teórico, la minería de datos se define como el proceso completo de extracción de información, que se encarga además de la preparación de los datos y de la interpretación de los resultados obtenidos, a través de grandes cantidades de datos, posibilitando de esta manera el encuentro de relaciones o patrones entre los datos procesados. (2004, p. 123)

Por su parte, tomando a Molina y García

> (...) explican que los datos tal cual se almacenan en las bases de datos no suelen proporcionar beneficios directos; su valor real reside en la información que podamos extraer de ellos, es decir, información que nos ayude a tomar decisiones o a mejorar la comprensión de los fenómenos que nos rodean. Ejemplos de ello pueden ser: contrastar que todo va bien, analizar diferentes aspectos de la evolución de la empresa, presentar información de forma más intuitiva, comparar información en diferentes períodos de tiempo, comparar resultados con previsiones, identificar comportamientos y evoluciones excepcionales, confirmar o descubrir tendencias e interrelaciones, entre otras acciones. (p. 143)

Para tal razón, se deben definir medidas cuantitativas para los patrones obtenidos (precisión, utilidad y beneficio obtenido), para establecer medidas de interés que consideren la validez y simplicidad de los patrones obtenidos mediante alguna de las técnicas de Minería de Datos. El objetivo final de todo esto es incorporar el conocimiento obtenido en algún sistema real, tomar decisiones a partir de los resultados alcanzados, o simplemente registrar la información conseguida y suministrársela a quien esté interesado. (BBVA Con tu empresa, 2013, párr. 11).



EVALUACIÓN DE RESULTADOS

Los analistas entenderán los fundamentos y conceptos básicos necesarios para participar en un proyecto de minería de datos, así como la clasificación de los sistemas y los tipos de modelos de minería de datos existentes.

Cabe resaltar que los efectos sobre la eficiencia de los resultados en la minería de datos como apoyo a la toma de decisiones está generando miles de opiniones desde diferentes perspectivas, entre ellas se puede destacar: la imposibilidad de encontrar conclusiones únicas referentes a la evaluación de un resultado obtenido con esta técnica. Sin embargo, la consecuencia de los datos arrojados es interesante, por algo más que su precisión.

EL ATRACTIVO QUE OFRECE LA MINERÍA DE DATOS EN LA IMPLEMENTACIÓN DE LAS ACTIVIDADES DE NEGOCIO: TALES COMO LA BONDAD, APLICABILIDAD, LA RELEVANCIA Y LA NOVEDAD; INDICADORES QUE APORTAN UNA IDEA DE LAS IMPLICACIONES Y UTILIDADES QUE PROPORCIONA ESTA PRÁCTICA.

De acuerdo con García y Acevedo (2010), el atractivo que ofrece la minería de datos en la implementación de las actividades de negocio: tales como la bondad, aplicabilidad, la relevancia y la novedad; indicadores que aportan una idea de las implicaciones y utilidades que proporciona esta práctica.

LOS PRINCIPALES INDICADORES PARA LA EVALUACIÓN DE LOS RESULTADOS SON LOS SIGUIENTES:

1. Indicadores de la bondad del resultado

"Estos indicadores tratan de aportar una idea acerca del error que se comete al emplear un modelo para realizar una tarea. Tal como manifiestan Padmanabhan y Tuzhilin (1999) esta es una medida de la fortaleza estadística del resultado. Para este indicador se utilizan las siguientes medidas: precisión, ratio de error, varianza y matriz de confusión" (Marcano y Talavera, 2007, párr. 14), siendo las dos últimas derivaciones de las anteriores. La precisión se utiliza cuando el resultado se presenta en forma de clasificación o estimación, la cual se mide a través del porcentaje de predicciones que son correctas. Para efectos de la clasificación, se emplea el porcentaje de casos bien clasificados y para la estimación del porcentaje de registros, se emplea una estimación que el decisor considere correcta. Para medir la precisión se puede emplear el coeficiente de confianza, el cual no es más que la probabilidad condicionada de un hecho con respecto a otro. (Vallespir, W, 2009, p. 17)

La distancia es otra técnica de minería de datos empleada cuando se disponen de variables continuas y numéricas, mediante la raíz cuadrada de la suma al cuadrado de las distancias en cada eje. Una medida que complementa a la precisión es el Ratio de error, que mide el porcentaje de casos en los que el resultado no coincide con la realidad.

2. Indicadores de relevancia del resultado

Los indicadores más representativos en este grupo son el coeficiente de cobertura, el coeficiente de apoyo y el coeficiente de significación. Estos indicadores tienen que ver directamente con la importancia que tiene el resultado arrojado por las técnicas de minería y miden la aportación a la situación actual y la frecuencia de utilidad del resultado, cuando la presentación de estos se hace en forma de reglas.

El coeficiente de cobertura mide el porcentaje de registros en los cuales se puede aplicar la regla. Por otro lado, el coeficiente de apoyo permite mostrar el porcentaje de ocasiones en que globalmente aparece la relación descrita por la regla, se recomienda representar el resultado en porcentaje. Por último, el coeficiente de significación sirve para medir el grado de importancia de la regla a través de la aportación que supone respecto a la pura probabilidad.

3. Indicadores de novedad del resultado

Cuando la información es excesivamente abundante y obvia, puede presentarse el problema al generar reglas. Para ello, existe el coeficiente de novedad, creado para indicar si una regla es interesante o no en función del número de reglas ya generadas, para un área de conocimiento concreta. Su objetivo es ayudar a evitar las redundancias en su obtención. Autores como Marcano y Talavera (2007) entre otros, apoyan por la inclusión del conocimiento previo del negocio, e intuición que detentan las decisiones para de esta manera: restringir el espacio de búsqueda, obtener conocimiento más preciso y eliminar aquel que resulte no interesante. (p. 17)

4. Indicadores de aplicabilidad del resultado

La actividad de las organizaciones actuales demanda cada vez más, tiempos de respuesta más rápidos, por lo cual es necesario que tanto la creación o generación de modelos como los resultados del mismo, deben estar disponibles en el menor tiempo posible. Para llegar a este resulatado, hay que buscar la simplicidad de los modelos y de la forma de representar la salida o resultados del análisis, para transformar el conocimiento obtenido y poder aplicarlo al negocio; para lograr esto, se cuenta con el coeficiente de simplicidad, la tasa interna de retorno y el valor actual neto. (Faggiano, Q. 2008, p. 71)

APLICACIÓN DE MÉTODOS DE APROXIMACIÓN CON BASE A UNA FUNCIÓN DE CERCANÍA Y MÉTODOS DE AGREGACIÓN.

El primer objetivo que se plantea consiste en determinar el número de clases que van a poseer los datos de entrenamiento.

CUANDO LA PARTICIÓN TIENE UN NÚMERO DE CLASES BAJO, SIGNIFICA QUE EXISTEN MUCHAS MUESTRAS QUE SON ASIGNADAS A CLASES DE UNA FORMA MUY FORZADA, LO QUE SE TRADUCE EN QUE LAS MISMAS SE SITÚAN LEJOS DE LOS CENTROS DE LOS CLÚSTERES Y POR TANTO GENERAN VALORES DE LOS GRADOS DE PERTENENCIA BAJOS, SEGÚN EL MÉTODO DE AGRUPAMIENTO BORROSO.

Debido a que las muestras utilizadas son muy heterogéneas fue posible realizar distintas pruebas, comprobando el correcto funcionamiento de los algoritmos en conjuntos de datos que contenían de dos a siete clases.

Cuando la partición tiene un número de clases bajo, significa que existen muchas muestras que son asignadas a clases de una forma muy forzada, lo que se traduce en que las mismas se sitúan lejos de los centros de los clústeres y por tanto generan valores de los grados de pertenencia bajos, según el método de agrupamiento borroso.

En el caso de un número de clases alto, existen muestras que se asignan a un clúster pero que en realidad podrían estar asignadas a cualquier otro clúster próximo. Esto se traduce en el hecho de que existirán muestras con elevados grados de pertenencia para clases diferentes.

Si hablásemos de 1-NN existe un problema de ruido cuando una muestra debe pertenecer a una clase, y tiene como más cercano un vecino de una clase diferente, como sucede en el caso de las clases solapadas. Esto hace que automáticamente se asigne la muestra a la clase equivocada.

En el otro extremo está el caso de que de forma empírica, no puede utilizarse el algoritmo de vecinos más cercanos Fuzzy con más vecinos que el número total de muestras, aunque esto significa que utilizar tantos vecinos como muestras en la evaluación dará como porcentaje de aciertos el porcentaje de muestras de la clase mayoritaria. Significa que cualquier muestra será asociada a la clase mayoritaria del conjunto, y en consecuencia esto conlleva a un resultado nada deseable.

Por esta razón, se ha buscado un número adecuado de vecinos mediante el proceso de prueba y error. Las pruebas se realizaron para estimar el comportamiento de la técnica de clasificación y así predecir nuevas situaciones; para ello se utilizó la fórmula de la precisión de la ecuación. (Hernández, Ramírez y Perri, 2004, p. 143).

LOS PRINCIPALES PROCESOS DE ESTA APLICACIÓN SON:

- **Visualización**

Los modelos de visualización pueden ser bidimensionales, tridimensionales o incluso multidimensionales. Se han desarrollado varias herramientas de visualización para integrarse con las bases de datos ofreciendo una visualización de forma interactiva a la minería de datos. Las tecnologías de la visualización son excelentes para ubicar patrones en un conjunto de datos y pueden ser usadas al comienzo de un proceso de *Data Mining*, para tomar un *feeling* de la calidad del conjunto de datos.

- **Procesamiento paralelo**

Esta técnica ha sido utilizada durante mucho tiempo. El área se ha desarrollado significativamente, desde sistemas con un único procesador hasta sistemas multiprocesadores. Los sistemas de multiprocesamiento pueden estar formados por sistemas distribuidos o por sistemas centralizados de multiprocesadores con memoria compartida o con multiprocesadores sin memoria compartida. Estos sistemas no fueron comercializados hasta el desarrollo del *Data Warehouse*, ya que ellos emplean el procesamiento paralelo para acelerar el proceso de las consultas.

Estos sistemas se han empezado a utilizar recientemente para las aplicaciones comerciales, debido en parte a la explosión del *Data Warehouse* y de las técnicas de minería de datos, donde el rendimiento de los algoritmos de consulta es crítico.

Para escalar las técnicas de minería de datos se necesita *hardware y software* apropiado, por lo que los fabricantes de bases de datos están empleando computadores con procesamiento paralelo para llevar a cabo las actividades de minería.

- **Apoyo a la toma de decisiones**

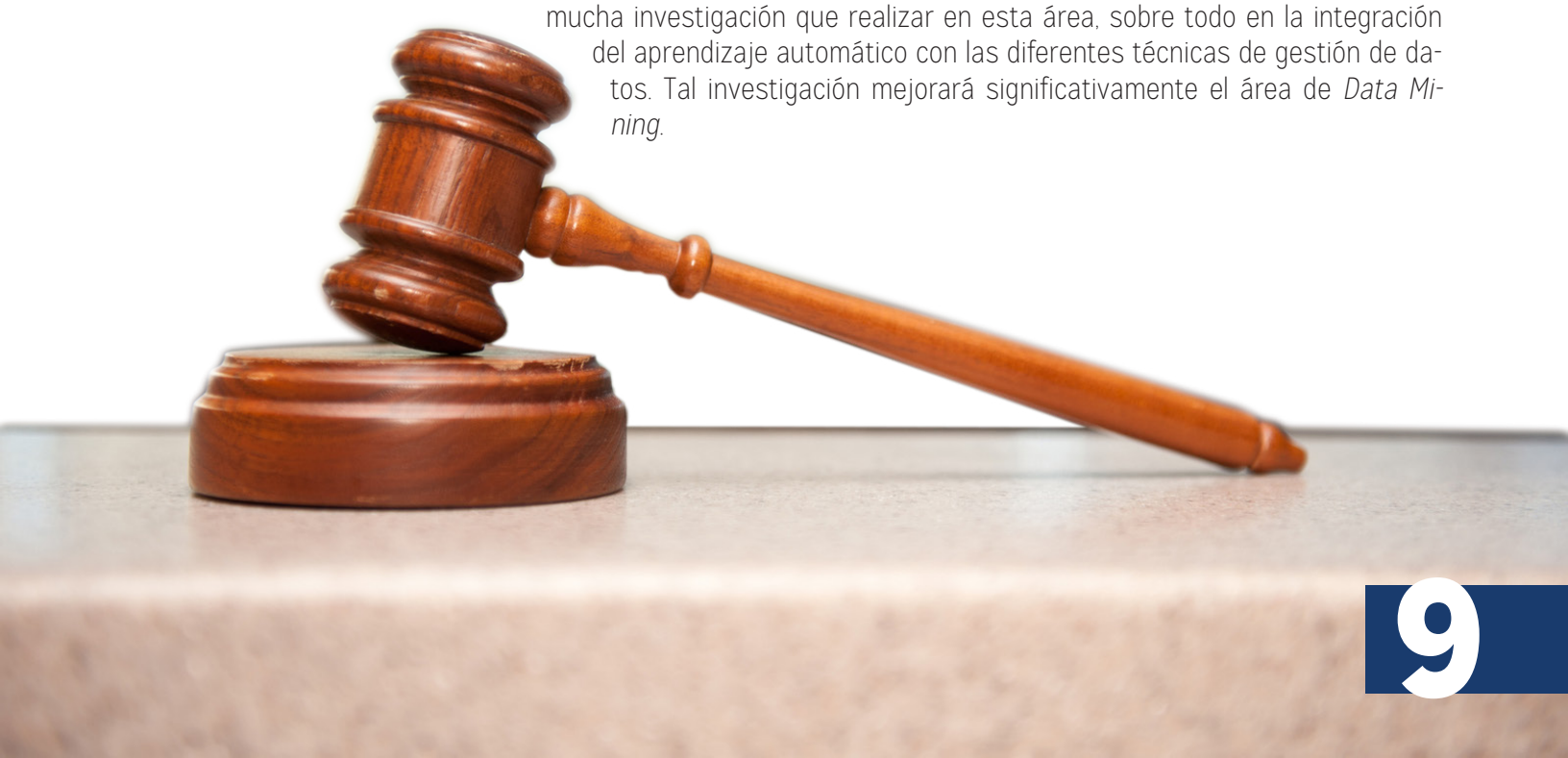
Los sistemas de apoyo a la toma de decisiones son las herramientas que usan los directivos para tomar decisiones eficaces, basándose en la teoría de la decisión. Por su parte, se puede considerar a las herramientas de minería de datos como tipos especiales de herramientas de apoyo a la toma de decisiones.

En general, las herramientas de apoyo a la toma de decisiones podrían utilizarse también como herramientas para eliminar los resultados innecesarios e irrelevantes obtenidos de la minería de datos. Igualmente pueden ser consideradas de este tipo, herramientas tales como las hojas de cálculo, sistemas expertos, sistemas de hipertexto, sistemas de gestión de información de *Web* y cualquier otro sistema que ayude a analistas y gestores a manejar eficazmente grandes cantidades de datos e información. Recientemente ha aparecido un área nueva llamada gestión del conocimiento, la cual trata de manejar eficazmente los datos, la información y el conocimiento de una organización. (Laudon, 2008, p. 123)

- **Aprendizaje automático**

En muchos casos, el aprendizaje automático consiste fundamentalmente en el aprendizaje de reglas a partir de los datos y por eso muchas de las técnicas de aprendizaje automático son utilizadas en la actualidad en las actividades de minería.

Así pues, se han desarrollado distintas técnicas para el aprendizaje automático, incluyendo el aprendizaje conceptual donde se aprende los conceptos desde diferentes ejemplos de entrenamiento, haciendo uso de las redes neuronales, los algoritmos genéticos, los árboles de decisión y la programación de la lógica inductiva. Hay todavía mucha investigación que realizar en esta área, sobre todo en la integración del aprendizaje automático con las diferentes técnicas de gestión de datos. Tal investigación mejorará significativamente el área de *Data Mining*.



Igualmente Molina y García (2004) afirman que,

> **Que la aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, razón por la cual esta técnica es mucho más eficiente que el análisis dirigido a la verificación, cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad elevada. Esta técnica consiste en aprender de las experiencias del pasado con respecto a alguna medida de rendimiento. Molina y García (2004, p. 123)**

En definitiva, Molina y García (2004) dicen que:

> (...) la utilidad de aplicaciones futuras en KDD es de largo alcance. KDD puede usarse como un medio de recuperación de información, de la misma manera que los agentes inteligentes realizan la recuperación de información en la Web. El KDD también puede usarse como una base para las interfaces inteligentes del mañana, agregando un componente del descubrimiento del conocimiento a un sistema de bases de datos o integrando KDD con las hojas de cálculo y visualizaciones. Nuevos modelos o tendencias en los datos podrán descubrirse usando estas técnicas. (p. 112)

Resumiendo a Díaz y Pérez (2004) mencionan la importante labor que tienen los algoritmos de Minería de Datos en la exactitud de determinados conjuntos de datos numéricos. Estos son: los algoritmos de Redes Neuronales e Híbridos de Aprendizaje (inducción de reglas y árboles de decisión), entre otros. (p. 143)

REFERENCIAS BIBLIOGRÁFICAS

Benzecri, J. (2015). Gestión del conocimiento y minería de datos. París: Dunod.

Berry, M. y Linoff, G. (2014). Data Mining Techniques for Marketing Sales and Customer Support. USA: John Wile& Sons

Chatfield, C. y Collins, A.J. (1999). Introduction to multivariate analysis. London: Chapman and Hall.

Kamber M. (2006). Data mining: concepts and techniques. Morgan Kaufmann.

Hoaglin D.C., Mosteller F., Tukey J.W. (2005). Exploring Data tables. Trends and Shapes, Wiley, N.Y.

Ester, M., Kriegel, H., y Sander, J (1999). Knowledge discovery in spatial databases. KI-99. Advanc Artif Intellig.

Jambu, M. (2000). Classification Automatique pour l'Analyse des données. París: Dunod.
Johnson Dallas, E. (2013). Métodos multivariados aplicados al análisis de datos. México: Thomson editores.

Johnson, R.A. y Wichern Dean, W. (2010). Applied Multivariare Statistical Analysis. 3rd De. USA: Prentice Hall Inc.

Wichern Dean, W. (2008). Sistemas de información gerencial: Administración de la información digital. México: Pearson.

Lebart, L., Morineau, A. y Tabard, N. (2000). Téchniques de la description statistique. París: Dunod.

Lebart, L., Morineau, A. y Piron, M. (2005). Statistique exploratoire multidimensionelle. París: Dunod.