**El uso de herramientas tecnológicas de minería de datos en el análisis de datos climatológicos**

***The use of technological tools for data mining in the analysis of climatological data***

***O uso de ferramentas tecnológicas para mineração de dados na análise de dados climatológicos***

**Jesús Abraham Castorena Peña**

Universidad Autónoma de Coahuila, México

jesuscastoenapena@uadec.edu.mx

**Alicia Elena Silva Ávila**

Universidad Autónoma de Coahuila, México

alicia.silva@uadec.edu.mx

**Alma Jovita Domínguez Lugo**

 Universidad Autónoma de Coahuila, México

almadominguez@uadec.edu.mx

**Diana Laura Rodríguez Montelongo**

 Universidad Autónoma de Coahuila, México

dyannardz\_16@hotmail.com

# Resumen

El análisis de grandes volúmenes de datos se está convirtiendo en un elemento clave para las organizaciones de cualquier sector debido al soporte que brinda en la toma de decisiones. El propósito de la siguiente investigación es aportar soluciones tecnológicas que permitan identificar el comportamiento o generar patrones en torno a datos climatológicos, como la precipitación registrada por el Servicio Meteorológico Nacional (SMN) en los estados de Chiapas, Oaxaca, Tabasco y Veracruz, contrastada con el nivel del gasto de los principales ríos de dichos estados. Para este estudio se utilizó la herramienta computacional de minería de datos Watson Analytics, la cual se caracteriza por realizar de forma automática la modelización de los datos y mostrar los hechos más relevantes, así como los patrones y relaciones que subyacen en ellos. En cuanto a la metodología de minería de datos empleada, se utilizó la metodología CRISP-DM propuesta por Chapman et al. (2000), debido a las características de esta investigación. La extracción de comportamiento y patrones de datos climatológicos proporcionaron información relevante para que organizaciones como el Servicio Meteorológico Nacional pueda tomar medidas y formular estrategias de prevención ante cualquier eventualidad ocasionada por el clima.

**Palabras clave:** datos climatológicos, metodología CRISP-DM, minería de datos, Watson Analytics.

# Abstract

The analysis of large volumes of data is becoming a key element for organizations in any sector, due to the support it provides in decision making. The purpose of the following research is to provide technological solutions of data mining (DM) that allow identifying the behavior or patterns presented by climatological data, such as precipitation, recorded by the National Meteorological Service (SMN) in the states of Chiapas. , Oaxaca, Tabasco and Veracruz, contrasted with the level of expenditure of the main rivers of these states. For the present investigation, the Watson Analytics data mining tool was used, which is characterized by the automatic modeling of the data, showing the most relevant facts, as well as the patterns and relationships that underlie them. Regarding the data mining methodology used, the one proposed by NCR, AG, SPSS, OHRA, Teradata and Daimer-Chrysler was used. CRISP-DM (Cross Industry Standard Process for Data Mining), due to the characteristics presented by the research was exploratory. The extraction of behavior and weather data patterns provide relevant information so that organizations such as the National Meteorological Service (SMN) can take measures and formulate prevention strategies in the face of any event caused by the weather.

**Key words:** climatological data, crisp methodology, data mining, Watson Analytics.

**Resumo**

A análise de grandes volumes de dados está se tornando um elemento chave para as organizações em qualquer setor devido ao suporte que ele fornece na tomada de decisões. O objetivo da pesquisa a seguir é fornecer soluções tecnológicas para identificar o comportamento ou gerar padrões em torno de dados climatológicos, como chuvas registradas pelo Serviço Meteorológico Nacional (SMN) nos estados de Chiapas, Oaxaca, Tabasco e Veracruz, em contraste com o nível de gastos dos principais rios dos referidos estados. A ferramenta de mineração de dados Watson Analytics foi utilizada para este estudo, que se caracteriza pela realização automática de modelagem de dados e mostra os fatos mais relevantes, bem como os padrões e relacionamentos que os fundamentam. Quanto à metodologia de mineração de dados utilizada, foi utilizado o proposto por NCR, AG, SPSS, OHRA, Teradata e Daimer-Chrysler. O CRISP-DM, devido às características apresentadas pela pesquisa, foi exploratório. A extração de padrões de comportamento e dados meteorológicos forneceu informações relevantes para organizações como o Serviço Meteorológico Nacional para tomar medidas e formular estratégias de prevenção em face de qualquer eventualidade causada pelo clima.

**Palavras-chave:** dados climatológicos, metodologia nítida, mineração de dados, Watson Analytics.

**Fecha Recepción:** Abril 2017 **Fecha Aceptación:** Diciembre 2017

# Introducción

En la actualidad, las tecnologías de la información y la comunicación (TIC) se han convertido en una de las herramientas más significativas para cualquier organización que desee recopilar y analizar de forma eficaz y eficiente los datos generados día a día (Ruiz, 2011). En efecto, las TIC permiten crear, almacenar, intercambiar y procesar de distintas maneras información que puede ser usada en beneficio propio (Tello, 2007). Por ello, son cada vez más las organizaciones que adquieren o desarrollan plataformas tecnológicas para evaluar grandes volúmenes de datos que sirven para apoyar la toma de decisiones y responder a los cambios presentados por el entorno.

Ahora bien, entre las tecnologías más relevantes de los últimos años para analizar esa abundante información se encuentra la minería de datos (en inglés *data mining* o DM), la cual se ha vuelto indispensable para el examinar y obtener resultados que hasta hace poco se hallaban ocultos en profusas cantidades de datos. Efectivamente, gracias a esta se pueden identificar comportamientos sobre un determinado bien o servicio, lo cual se puede emplear para predecir y extraer información en torno a determinados patrones que con otros métodos sería imposible de realizar.

La minería de datos, por ende, es una nueva tecnología de manejo y análisis de información que aprovecha la capacidad existente no solo para procesar, almacenar y transmitir datos a gran velocidad y a bajo costo, sino también para hallar un contenido específico dentro la diversidad de fuentes existentes, lo cual resulta muy útil para las organizaciones, pues les permite tomar decisiones mejor fundamentadas para su futuro (Altamiranda *et al.*, 2013).

En este contexto, la climatología ha utilizado durante mucho tiempo las técnicas y herramientas estadísticas de manera recurrente y sistemática para describir y predecir el comportamiento del clima; sin embargo, los resultados con este tipo de métodos suelen ser menores si se compran con otras técnicas más sofisticadas como la minería de datos (Joya, Sistachs, Cabrera y Roura, 2014).

En efecto, la aplicación de técnicas de minería de datos en registros meteorológicos históricos permite dar señales anticipadas sobre eventuales desastres naturales ocasionados por fenómenos meteorológicos (Duque, Orozco e Hincapié, 2010), lo cual la convierte en una herramienta útil para orientar la manera de actuar ante un incidente inesperado.

Por tal motivo, el siguiente estudio tiene como objetivo aportar soluciones tecnológicas que permitan identificar el comportamiento o generar patrones en torno a datos climatológicos, como la precipitación registrada por el Servicio Meteorológico Nacional (SMN) en los estados de Chiapas, Oaxaca, Tabasco y Veracruz, contrastada con el nivel del gasto de los principales ríos de dichos estados. Con esto se procura generar medidas y formular estrategias de prevención ante cualquier eventualidad ocasionada por el clima.

El artículo está estructurado en cuatro secciones: en la primera se señalan algunas herramientas tecnológicas de minería de datos; en la segunda sección se explica la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual ha sido empleada para analizar datos; en la tercera se presentan los resultados obtenidos con la herramienta de minería de datos Watson Analytics, y finalmente, en la último apartado, se ofrecen las conclusiones.

# Herramientas tecnológicas de minería de datos

Hoy en día existen diferentes herramientas tecnológicas de minería de datos que sirven a las organizaciones como soporte para la toma de decisiones, pues permiten identificar comportamientos o patrones sobre un determinado bien o servicio. Entre las herramientas más populares destacan SPSS Clementine, Oracle Data Miner (Rodríguez y Díaz, 2007), aunque una de las más empleadas en los últimos años es Watson Analytics debido a que no solo permite realizar de forma automática la modelización de los datos, sino que también muestra los hechos más relevantes, así como los patrones y relaciones que subyacen en ellos, a través de los módulos de exploración, predicción y presentación; en otras palabras, es una herramienta que permite sacar el máximo provecho de los datos con el mínimo esfuerzo. Entre sus beneficios se encuentran los siguientes: automatiza los análisis predictivos, permite la formulación de preguntas de interés, facilita la creación de *dashboards* e infografías, y proporciona una exploración y análisis de datos eficiente.

# Materiales y métodos

Dadas las características del presente estudio (examinar los datos climatológicos registrados por Servicio Meteorológico Nacional en los estados de Chiapas, Oaxaca, Tabasco y Veracruz), se empleó la metodología CRISP-DM, la cual es una de las más usadas para analizar grandes volúmenes de datos y descubrir información valiosa. La metodología CRISP-DM, según Chapman *et al.* (2000), está formada por seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

**Figura 1.** Modelo de proceso CRISP–DM



Fuente: Chapman *et al.* (2000)

# Comprensión del negocio

En esta fase se analizaron de forma exhaustiva los alcances y requerimientos; con ello, se pudo tener una visión general del problema y elaborar un plan para ofrecer una posible solución al objetivo definido.El tema elegido en esta investigación es el análisis de las variables precipitación pluvial y gasto de ríos dentro de ciertas regiones del país donde se han registrado diversos desastres naturales, los cuales se podrían prevenir si se realizara un mejor análisis del comportamiento de estas dos variables. La investigación está divida en dos etapas: la primera consiste en la creación del modelo y en su experimentación; la segunda (que se presentará en un próximo estudio) se basa en la implementación y demostración de los resultados en tiempo real.

# Comprensión de los datos

Se realizó la recopilación de los datos, los cuales fueron proporcionados por la Comisión Nacional del Agua (Conagua) y por el Instituto Mexicano de Tecnología del Agua (IMTA). En esta fase se procuró contar con datos fidedignos y confiables para poder conseguir el objetivo inicial de la investigación. Los registros analizados fueron 197 146, los cuales comprendieron el periodo 2000-2010, como se muestra en la figura 2 y la figura 3.

**Figura 2.** Datos de la variable *precipitación* en los estados Chiapas, Oaxaca, Tabasco y Veracruz



 Fuente: Conagua (2011)

**Figura 3.** Datos de ríos con datos anuales de gasto máximo, mínimo y medio



Fuente: IMTA (2017)

# Preparación de los datos

En esta fase se llevó a cabo el proceso de selección, limpieza y transformación de datos con el fin de definir de manera clara y precisa el conjunto de datos que mejor caracterizaban al problema; esto para poder alimentar el modelo de datos construido en la siguiente fase, como se enseña en la figura 4.

**Figura 4.** Metodología del proceso de extracción, limpieza y transformación de datos



Fuente: Elaboración propia

# Modelado

Resulta de suma importancia realizar un apropiado modelo de datos para que la carga de la información sea lo más sencilla posible, así como asegurar que los procesos no se rompan con facilidad. Dentro del modelado se revisaron tres tipos: esquema estrella, copo de nieve y constelación.

### Esquema estrella

Es un tipo de esquema de base de datos relacional que consiste en una tabla central de hechos, la cual está rodeada de tablas de dimensiones que forman una especie de “estrella” (Poblete y Zambrano, 2013). Este puede tener cualquier número de tablas de dimensiones.

### Copo de nieve

Es una extensión del esquema estrella, el cual consta de una tabla de hechos que está conectada a muchas tablas de dimensiones, que pueden estar unidas a otras tablas de dimensiones a través de una relación de muchos a uno (Cedeño, 2006). La ventaja más relevante de este se halla en que ocupa un menor espacio de almacenamiento.

### Constelación

Este modelo es una combinación entre el esquema estrella y copo de nieve, en el cual varias tablas de hechos comporten tablas de dimensiones. El principal objetivo de este es aprovechar las ventajas de ambos esquemas.

Una vez analizados cada uno de los esquemas, se determinó que el modelo constelación era el más adecuado, según las características de este estudio (ver la figura 5). Luego de elaborar el modelo de datos, se alimentó con los datos de la fase anterior, de modo que pudiera ser evaluado y validado para su correcto funcionamiento.

**Figura 5.** Modelo de datos constelación DM



Fuente: Elaboración propia

# Evaluación

En esta fase se exploraron, por medio de experimentación, cada uno de los modelos generados para comprobar su correcto funcionamiento y eficacia, de forma que se pudieran asegurar los mejores resultados para el problema y los datos en cuestión. La evaluación consistió en alimentar cada uno de los modelos con datos reales, de tal manera que se pudieran realizar una serie de consultas. Obtenidos los resultados, se analizaron de forma detallada para verificar que el modelo estuviera cumpliendo con el objetivo principal de la investigación. Asimismo, se realizó la documentación de los resultados obtenidos del modelo de datos para promover la toma de decisiones.

# Implementación

Dada la naturaleza de la investigación y la definición del problema, no se implementaron los resultados dentro de la organización, ya que en esta primera etapa —como se mencionó anteriormente— el estudio estuvo diseñado exclusivamente para crear, experimentar y validar el modelo de datos.

# Resultados y discusión

A continuación se presentan los hallazgos encontrados una vez analizados los datos climatológicos con la herramienta tecnológica Watson Analytics. De acuerdo con el análisis de la variable *precipitación*, se logró identificar de manera clara y sencilla que la entidad federativa con un mayor nivel de lluvia fue el estado de Chiapas, el cual registró una precipitación máxima de 276 857 mm en el año 2005. Esta es una variable de interés para alertar sobre posibles riesgos de inundaciones en dicho estado, tal como se muestra en la figura 6.

**Figura 6.** Precipitación por estados y años



Fuente: Elaboración propia

A continuación, en la figura 7, se muestra cómo en la mayor parte de los estados de Chiapas, Oaxaca, Tabasco y Veracruz la cantidad de lluvia registrada resulta ser más intensa en verano, con su punto más alto de precipitación en el mes de septiembre. Esta información resulta significativa para los municipios de cada uno de esos estados, pues permitiría tomar medidas preventivas ante posibles inundaciones. Téngase en cuenta que 22.2 % de la precipitación acumulada anual tiende a escurrirse por los ríos o arroyos (CONAGUA, 2014), lo cual podría provocar algún tipo de desbordamiento que genere inundaciones.

**Figura 7.** Precipitación total por estado y mes (periodo 2000-2010)



Fuente: Elaboración propia

Por otra parte, los municipios que registraron los niveles más altos de precipitación fueron Tapachula y Unión Juárez, lo cual indica que estas zonas tienen mayores probabilidades de sufrir algún tipo de desastre debido a la gran acumulación de precipitación. De hecho, contrastando esta información con un estudio de Vergara, Ellis, Cruz, Alarcón y Galván (2011), se puedo verificar que el municipio de Tapachula del estado de Chiapas sufrió grandes inundaciones en el año 2005, lo cual dejó un número considerable de dignificados.

También se debe destacar que durante la explotación de los datos se constató que en el año 2005 el estado de Chiapas, y específicamente el municipio de Tapachula, registraron los niveles más altos de precipitación, como se muestra en la figura 8.

**Figura 8.** Municipios con mayor precipitación en el periodo 2000-2010 

Fuente: Elaboración propia

En cuanto a los organismos de cuenca, de acuerdo con los datos históricos y las variables de estudio definidas, los resultados reflejan que la cuenca Golfo Sur tiene un alto riesgo de inundación, mientras que las cuencas de Golfo Centro y Pacífico Sur tienen un grado medio y bajo de peligro, respectivamente (ver figura 9[[1]](#footnote-1)). Por tanto, resulta necesario que los estados pertenecientes a estos organismos de cuenca tomen medidas preventivas para este tipo de desastres meteorológicos.

**Figura 9.** Nivel de alerta por organismo de cuenca



Fuente: Elaboración propia

# Conclusiones

Las nuevas herramientas tecnológicas de minería de datos son una de las herramientas más valiosas tanto para las organizaciones actuales, en general, como para el campo de la meteorología, en particular, debido a la gran cantidad de datos que se pueden generar, analizar y comprender para descubrir patrones en datos climatológicos. Esto sirve para orientar el proceso de toma de decisiones, así como la formulación de estrategias de prevención ante posibles desastres naturales.

A través de herramientas tecnológicas de minería de datos (como Watson Analytics) se pueden obtener de manera rápida y eficiente patrones y comportamientos sobre grandes volúmenes de información, que permiten aportar algún tipo de valor para las organizaciones. Asimismo, el empleo de este tipo de tecnologías facilita la toma de decisiones gracias a los diversos análisis predictivos que ofrece mediante la creación de cuadros de mando e infografías.

# Bibliografía

Altamiranda, L., Peña, A. M., Ospino, M., Volpe, I., Ortega, D. y Cantillo, E. (2013). “Minería de datos como herramienta para el desarrollo de estrategias de mercadeo B2B en sectores productivos, afines a los colombianos: una revisión de casos”, en *Sotavento mba*, 22, 126-136. Recuperado de <http://revistas.uexternado.edu.co/index.php/sotavento/article/viewFile/3709/3841>.

Cedeño, A. (2006). Modelo multidimensional. Industrial, 27(1), 15-18. Recuperado de <https://dialnet.unirioja.es/descarga/articulo/4786663.pdf>.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). CRISP-DM 1.0. Step-by-step data mining guide. Retrieved from <https://www.the-modeling-agency.com/crisp-dm.pdf>.

CONAGUA. (2014). Estadísticas del Agua en México. México. Recuperado de http://www.conagua.gob.mx/CONAGUA07/Publicaciones/Publicaciones/EAM2014.pdf

Duque, N., Orozco, M. e Hincapié, L. (2010). Minería de datos para el análisis de datos meteorológicos. *Tendencias en Ingeniería de Software e Inteligencia Artificial, 4,* 105-114. Recuperado de <http://www.docentes.unal.edu.co/morozcoa/docs/Duque2011.pdf>.

Joya, G., Sistachs, V., Cabrera, M. A. y Roura, P. (2014). Aplicación de la técnica de minería de datos SOM utilizando el lenguaje R en datos climáticos. Ciencias de la Tierra y el Espacio, 15(2), 113-123. Recuperado de <http://www.iga.cu/publicaciones/revista/assets/2.mineria.datos.meteorologia.pdf>.

Poblete, G. y Zambrano, C. (2013). Bases de datos multidimensionales para datos educacionales. Jornadas Chilenas de Computación. Temuco, Chile. Recuperado de <http://jcc2013.inf.uct.cl/wp-content/proceedings/ECC/Bases%20de%20Datos%20multidimensionales%20para%20datos%20educacionales.pdf>.

Rodríguez, Y. y Díaz, A. (2009). Herramientas de minería de datos. Revista Cubana de Ciencias Informáticas, *3*(3-4), 73-80. Recuperado de <http://www.redalyc.org/pdf/3783/378343637009.pdf>.

Ruiz, J. (2011). Herramientas para la investigación en tecnologías de la información y la comunicación. Casos de estudio. Revisa de Currículum y Formación del Profesorado, *15*(1), 139-149. Recuperado de <http://www.ugr.es/~recfpro/rev151ART9.pdf>.

Tello, E. (2007). Las tecnologías de la información y comunicaciones (TIC) y la brecha digital: su impacto en la sociedad de México. Revista de Universidad y Sociedad del Conocimiento, 4(2). Recuperado de https://dialnet.unirioja.es/descarga/articulo/2521723.pdf.

Vergara, M., Ellis, E., Cruz, J., Alarcón, L. y Galván, U. (2011). La conceptualización de las inundaciones y la percepción del riesgo ambiental. *Política y Cultura*, (36), 15-69. Recuperado de <http://www.redalyc.org/pdf/267/26721226003.pdf>.

|  |  |
| --- | --- |
| Rol de Contribución | Autor(es) |
| **Conceptualización** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo Grado de contribución: Igual |
| **Metodología** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo Grado de contribución: Igual |
| **Software** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo Grado de contribución: Igual |
| **Validación** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo Grado de contribución: Igual |
| **Análisis Formal** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo Grado de contribución: Igual |
| **Investigación** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo Grado de contribución: Igual |
| **Recursos** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo Grado de contribución: Igual |
| **Curación de datos** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo Grado de contribución: Igual |
| **Escritura - Preparación del borrador original** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo, Grado de contribución: Igual |
| **Escritura - Revisión y edición** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo, Grado de contribución: Igual |
| **Visualización** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo, Grado de contribución: Igual |
| **Supervisión** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo, Grado de contribución: Igual |
| **Administración de Proyectos** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo, Grado de contribución: Igual |
| **Adquisición de fondos** | Jesús Abraham Castorena Peña, Alicia Elena Silva Ávila, Alma Jovita Domínguez Lugo y Diana Laura Rodríguez Montelongo , Grado de contribución: Igual |

1. La información utilizada solo fue tomada de algunos estados donde se pudo constatar y comprobar la existencia de los ríos. [↑](#footnote-ref-1)