Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

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

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

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.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

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

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

Introduction

At present, information and communication technologies (ICT) have become one of the most significant tools for any organization that wishes to collect and analyze efficiently and efficiently the data generated every day (Ruiz, 2011). Indeed, ICTs allow information to be created, stored, exchanged and processed in different ways that can be used for their own benefit (Tello, 2007). Therefore, more and more organizations acquire or develop technological platforms to evaluate large volumes of data that serve to support decision making and respond to changes presented by the environment.

However, among the most relevant technologies in recent years to analyze this abundant information is data mining (in English data mining or DM), which has become indispensable for examining and obtaining results that until recently were hidden in profuse amounts of data. Effectively, thanks to this, behaviors can be identified on a specific good or service, which can be used to predict and extract information about certain patterns that would be impossible with other methods.

Data mining, therefore, is a new information management and analysis technology that takes advantage of existing capacity not only to process, store and transmit data at high speed and at low cost, but also to find specific content within the diversity of existing sources, which is very useful for organizations, because it allows them to make better informed decisions for their future (Altamiranda *et al.*, 2013).

In this context, climatology has long used statistical techniques and tools in a recurrent and systematic way to describe and predict climate behavior; however, the results with this type of methods are usually lower if they are purchased with other more sophisticated techniques such as data mining (Joya, Sistachs, Cabrera and Roura, 2014).

In fact, the application of data mining techniques in historical meteorological records allows us to give advance signals about eventual natural disasters caused by meteorological phenomena (Duque, Orozco and Hincapié, 2010), which makes it a useful tool to guide the way act in the face of an unexpected incident.

For this reason, the following study aims to provide technological solutions to identify the behavior or generate patterns around climatological data, such as rainfall

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. This is intended to generate measures and formulate prevention strategies in the face of any eventuality caused by the climate.

The article is structured in four sections: in the first one, some technological data mining tools are indicated; In the second section the CRISP-DM methodology (Cross Industry Standard Process for Data Mining) is explained, which has been used to analyze data; in the third, the results obtained with the Watson Analytics data mining tool are presented, and finally, in the last section, the conclusions are offered.

Technological tools of data mining

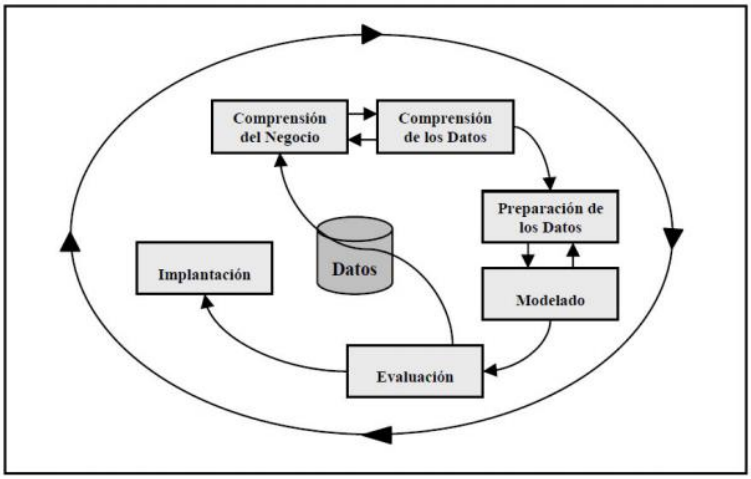
Today there are different technological tools of data mining that serve organizations as support for decision making, because they allow to identify behaviors or patterns on a specific good or service. Among the most popular tools include Clementine SPSS, Oracle Data Miner (Rodriguez and Diaz, 2007), although one of the most used in recent years is Watson Analytics because not only allows you to automatically perform data modeling, but also which also shows the most relevant facts, as well as the patterns and relationships that underlie them, through the exploration, prediction and presentation modules; In other words, it is a tool that allows you to get the most out of the data with minimal effort. Among its benefits are the following: automates predictive analysis, allows the formulation of interesting questions, facilitates the creation of dashboards and infographics, and provides an efficient exploration and data analysis.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

Materials and methods

Given the characteristics of the present study (to examine the climatological data recorded by the National Meteorological Service in the states of Chiapas, Oaxaca, Tabasco and Veracruz), the CRISP-DM methodology was used, which is one of the most used to analyze large volumes of data and discover valuable information. The CRISP-DM methodology, according to Chapman et al. (2000), consists of six phases: understanding the business, understanding the data, preparing the data, modeling, evaluation and implementation.

Figura 1. Modelo de proceso CRISP–DM



Fuente: Chapman *et al.* (2000)

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

Business understanding

In this phase, the scope and requirements were exhaustively analyzed; With this, it was possible to have an overview of the problem and develop a plan to offer a possible solution to the defined objective. The topic chosen in this research is the analysis of the variables rainfall and river expenditures within certain regions of the country where several natural disasters have been recorded, which could be prevented if a better analysis of the behavior of these two variables was made. The research is divided into two stages: the first consists of the creation of the model and its experimentation; the second (to be presented in a future study) is based on the implementation and demonstration of the results in real time.

Understanding the data

The data was collected, which were provided by the National Water Commission (Conagua) and the Mexican Institute of Water Technology (IMTA). In this phase we tried to have reliable and reliable data to be able to achieve the initial objective of the investigation. The records analyzed were 197 146, which included the period 2000-2010, as shown in Figure 2 and Figure 3.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

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

Id_Origen	IdEstacion	YEAR	MONTH	DIA1	DIA2	DIA3	DIA4	DIA5	DIA6	DIA7	DIA8	DIA9	DIA10	DIA11	DIA12
98697	20509	2002	09	2	56	1	1	0	1	0	0	30	0	0	1
98698	20509	2002	10	0	0	1	0	0	1	1	0	0	0	0	0
98699	20509	2002	11	1	73	58	0	0	0	0	0	13	0	0	0
98700	20509	2002	12	0	0	0	0	0	0	0	0	0	0	0	0
98701	20509	2003	01	0	0	0	0	0	0	0	0	0	0	0	0
98702	20509	2003	02	16	0	0	0	0	0	0	0	0	0	0	0
98703	20509	2003	03	0	0	0	0	0	0	0	0	0	0	0	0
98704	20509	2003	04	0	0	0	0	0	0	0	0	0	0	0	0
98705	20509	2003	05	0	0	0	0	0	0	0	0	0	0	0	0
98706	20509	2003	06	118	4	2	22	8	5	10	9	8	0	0	5
98707	7001	1970	10	0,5	7,5	2	16	0	0	0	0	2	1	0	17
98708	7001	1970	11	3	0	1	2,5	2	0,5	0	0	0	2	0	0,5
98709	7001	1970	12	0	2,5	2	2	0	0,5	0	4,5	0	0	0	0
98710	7001	1971	01	0,5	0	0	0	0	3	6	5	0	0	0	0
98711	7001	1971	02	0	1,5	0	0	0	0	0	1	8	0	0	7
98712	7001	1971	03	0	28	8	0	0	0	13,5	3	0	0	0	0
98713	7001	1971	04	0	0	0	0	2	4,5	4	0	0	0	0	0

Fuente: Conagua (2011)

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

pk_anio	pk_mes	ngasto_mxn	ndia_gmxm	nhora_gmxr	nlect_gmxm	ngasto_mir	ndia_gmim	nhora_gmin	nlect_gmim	nlect_mxm
2007	4	17,4	4		0,96	7,01	18		0,73	0,9
2007	5	14,6	4		0,91	7,01	7		0,73	0,9
2007	6	36,5	27		1,23	7,01	1		0,73	1,2
2007	7	154	20		2,1	7,86	9		0,75	2,
2007	8	97,4	22		1,75	11,8	4		0,85	1,7
2007	9	386	7		3,15	32,3	28		1,18	3,1
2007	10	28,1	10		1,29	12,6	13		1,13	1,2
2007	11	19,4	28		1,21	10,4	17		1,09	1,2
2007	12	10,4	1		1,09	8,42	21		1,05	1,0
2008	1	8,42	1		1,05	7,58	22		1,03	1,0
2008	2	7,58	1		1,03	6,73	18		1,01	1,0
2008	3	6,73	1		1,01	5,89	16		0,99	1,0
2008	4	6,45	1		1	5,61	23		0,98	1
2008	5	48	14		1,45	5,61	14		0,98	1,4
2008	6	69	8		1,6	5,61	4		0,98	1,(
2008	7	109	4		1,82	20,6	30		1,2	1,8
2008	8	58,5	12		1,52	16	23		1,15	1,5
2008	9	274	8		2,42	27,4	1		1,15	2,4
2008	10	47,1	7		1,43	25,3	31		1,11	1,4
2008	11	25.3	1		1.11	21	29		1.01	1.1

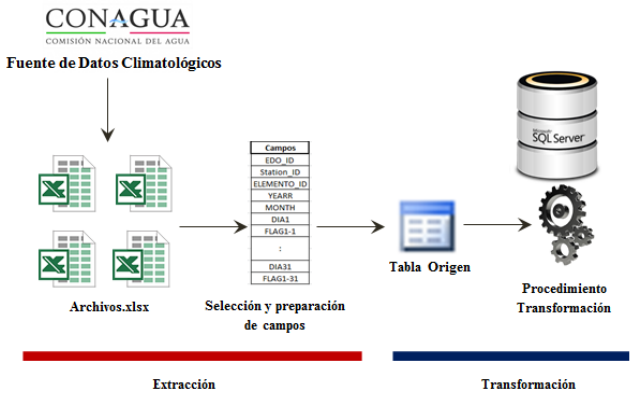
Fuente: IMTA (2017)

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

Preparation of the data

In this phase the process of selection, cleaning and transformation of data was carried out in order to clearly and precisely define the data set that best characterized the problem; this to be able to feed the data model built in the next phase, as shown in figure 4.

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



Fuente: Elaboración propia

Modeling

It is very important to make an appropriate data model so that the information load is as simple as possible, as well as ensuring that the processes do not break easily. Within the modeling, three types were reviewed: star scheme, snowflake and constellation.

Star scheme

It is a type of relational database scheme that consists of a central table of facts, which is surrounded by tables of dimensions that form a kind of "star" (Poblete and Zambrano, 2013). This can have any number of dimension tables.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

Snowflake

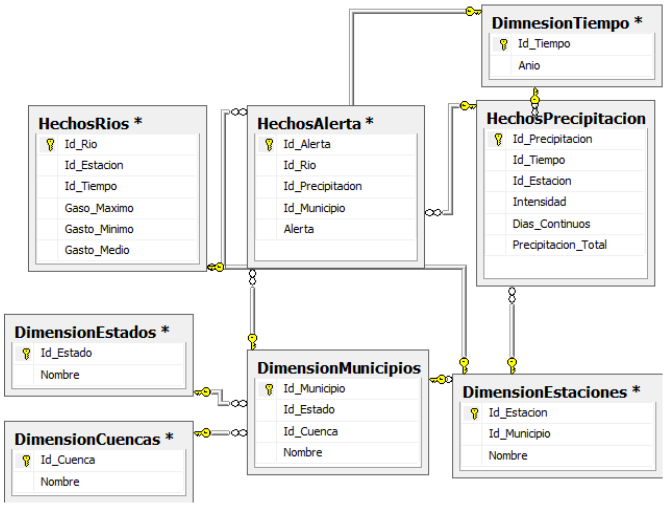
It is an extension of the star schema, which consists of a fact table that is connected to many dimension tables, which can be linked to other dimension tables through a many-to-one relationship (Cedeño, 2006). The most important advantage of this is that it occupies less storage space.

Constellation

This model is a combination between the star and snowflake scheme, in which several tables of facts have dimension tables. The main objective of this is to take advantage of both schemes.

Once each of the schemes was analyzed, it was determined that the constellation model was the most appropriate, according to the characteristics of this study (see Figure 5). After elaborating the data model, it was fed with the data of the previous phase, so that it could be evaluated and validated for its correct functioning.

Figura 5. Modelo de datos constelación DM



Fuente: Elaboración propia

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

Evaluation

In this phase, each of the generated models was explored, through experimentation, to verify its correct functioning and efficiency, so that the best results for the problem and the data in question could be ensured. The evaluation consisted in feeding each one of the models with real data, in such a way that a series of consultations could be carried out. Once the results were obtained, they were analyzed in detail to verify that the model was fulfilling the main objective of the investigation. Likewise, the documentation of the results obtained from the data model was made to promote decision making.

Implementation

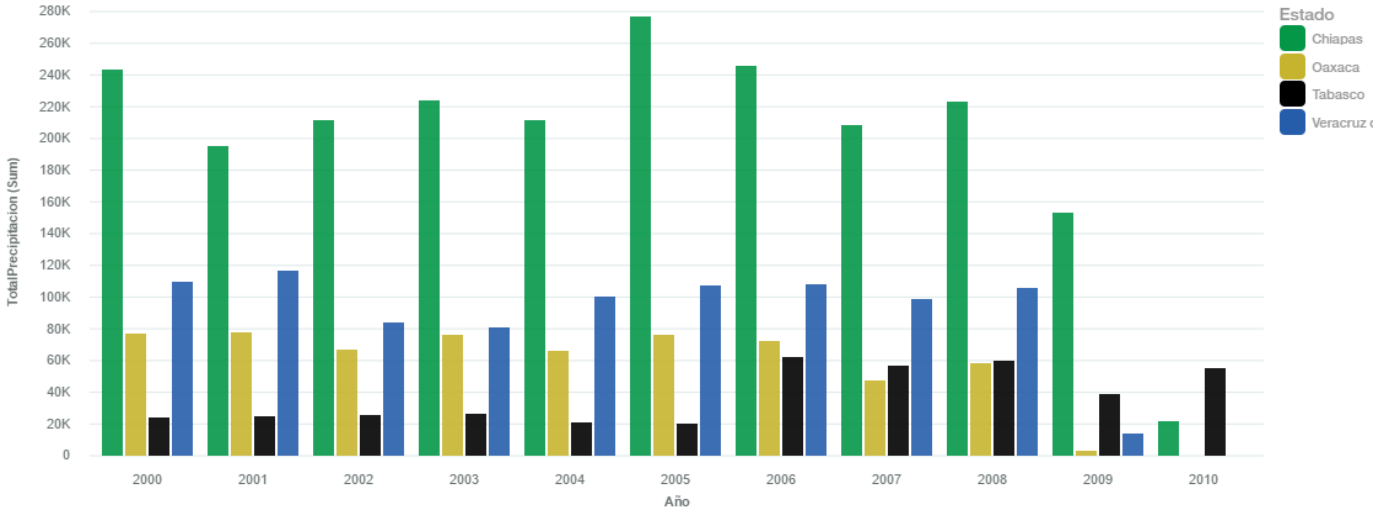
Given the nature of the research and the definition of the problem, the results were not implemented within the organization, since in this first stage -as mentioned above- the study was exclusively designed to create, experiment and validate the data model.

Results and Discussion

Below are the findings found after analyzing the climatological data with the Watson Analytics technology tool. According to the analysis of the rainfall variable, it was clearly and easily identified that the state with the highest level of rainfall was the state of Chiapas, which registered a maximum rainfall of 276 857 mm in the year 2005. it is a variable of interest to alert about possible flood risks in that state, as shown in figure 6.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

Figura 6. Precipitación por estados y años

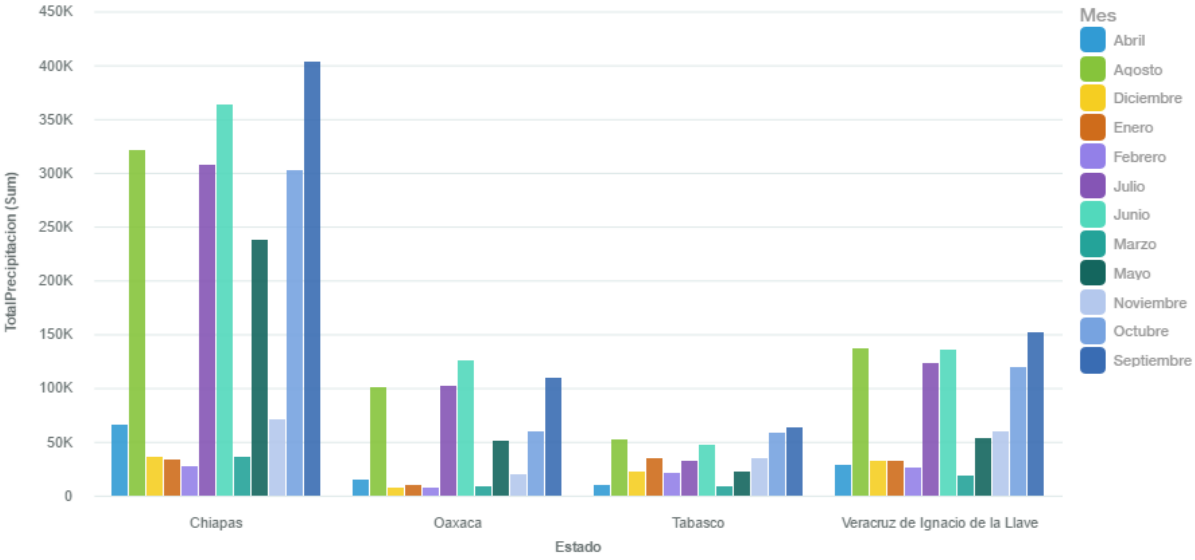


Fuente: Elaboración propia

Then, in figure 7, it shows how in most of the states of Chiapas, Oaxaca, Tabasco and Veracruz the amount of rainfall recorded turns out to be more intense in summer, with its highest point of precipitation in the month of September . This information is significant for the municipalities of each of these states, because it would allow to take preventive measures against possible floods. Bear in mind that 22.2% of annual accumulated precipitation tends to run off rivers or streams (CONAGUA, 2014), which could cause some type of overflow that generates flooding.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

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



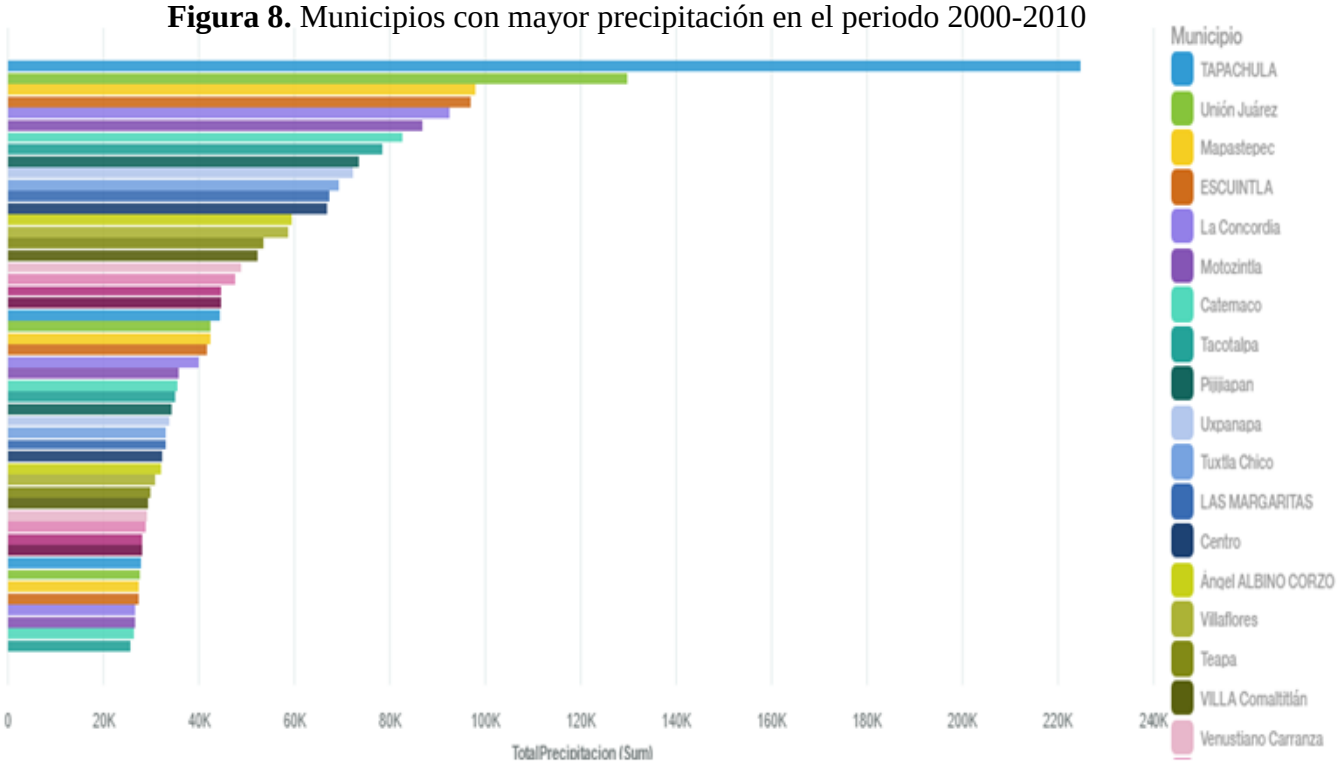
Fuente: Elaboración propia

On the other hand, the municipalities that registered the highest levels of precipitation were Tapachula and Unión Juárez, which indicates that these areas are more likely to suffer some type of disaster due to the large accumulation of precipitation. In fact, contrasting this information with a study by Vergara, Ellis, Cruz, Alarcón and Galván (2011), we can verify that the municipality of Tapachula of the state of Chiapas suffered great floods in 2005, which left a considerable number of dignified

It should also be noted that during the exploitation of the data it was found that in 2005 the state of Chiapas, and specifically the municipality of Tapachula, recorded the highest levels of precipitation, as shown in figure 8.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

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



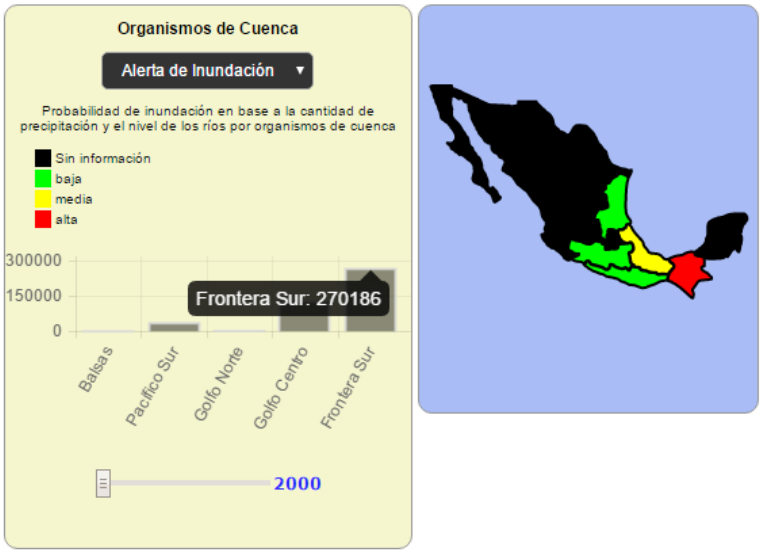
Fuente: Elaboración propia

Regarding the basin organizations, according to the historical data and the defined study variables, the results show that the Golfo Sur basin has a high risk of flooding, while the Central and South Pacific basins have a medium degree and under danger, respectively (ver figura 9[1]). Therefore, it is necessary that the states belonging to these basin organizations take preventive measures for this type of meteorological disasters.

[1] La información utilizada solo fue tomada de algunos estados donde se pudo constatar y comprobar la existencia de los ríos.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

Figura 9. Nivel de alerta por organismo de cuenca



Fuente: Elaboración propia

Conclusions

The new technological tools of data mining are one of the most valuable tools both for current organizations, in general, and for the field of meteorology, in particular, due to the large amount of data that can be generated, analyzed and understood to discover patterns in climatological data. This serves to guide the decision-making process, as well as the formulation of prevention strategies against possible natural disasters.

Through data mining technology tools (such as Watson Analytics), patterns and behaviors can be obtained quickly and efficiently on large volumes of information, which provide some kind of value for organizations. Likewise, the use of this type of technologies facilitates the decision making thanks to the diverse predictive analysis that it offers through the creation of dashboards and infographics.

Bibliography

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.

Revista Iberoamericana de las Ciencias Computacionales e Informática
ISSN: 2007 - 9915

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