Minería de datos para generación de reglas de tendencia de precipitación pluvial en el estado de Morelos

Miguel Angel Vazquez Zavaleta*

Instituto Mexicano de Tecnología del Agua

Fecha de recepción: 10 de abril de 2018 – Fecha de aprobación: 31 de julio de 2018

RESUMEN

El observar los eventos climatológicos ha sido considerado como parte fundamental de prácticas culturales, socioeconómicas y de investigación científica. Se ha logrado adquirir una gran cantidad de datos del medio ambiente, con la dificultad que conlleva el manipularlos, analizarlos e interpretarlos. Las contribuciones recientes a las ciencias exactas, sociales y humanas han permitido un avance extraordinario en tecnologías de cómputo y análisis. De estas contribuciones se han desarrollado técnicas de gestión y análisis de datos como son la minería de datos. El objeto de este estudio es "aplicar técnicas de minería de datos para la creación de reglas de tendencia de precipitación pluvial en el estado de Morelos", utilizando datos de temperatura, unidades de calor y precipitación pluvial de 29 estaciones climatológicas situadas dentro del estado. Las series de datos de temperatura y precipitación pluvial se evaluaron mediante mecanismos de control de calidad y homogeneidad. Las técnicas de minería de datos que se utilizaron son: el particionado k-means y los árboles de decisión C4.5; de la primera técnica se generaron 58 pares de agrupaciones de temperatura los cuales representan la climatología de la zona para cada estación, estas agrupaciones se utilizaron como umbrales de discretización de la serie de temperatura para simplificar los 29 conjuntos de reglas de tendencia de precipitación pluvial generados con el algoritmo C4.5.

Palabras clave: minería de datos, control de calidad, Morelos, precipitación, temperatura.

Data Mining for the Generation of Trend Rules of Rain Precipitation in the State of Morelos

ABSTRACT

The observation of climatological events has been considered as a fundamental part of cultural socioeconomic and scientific research practices. It has been managed to acquire a large amount of environmental data, with the difficulty involved in manipulating, analyzing and interpreting them. Recent contribution to the exact, social and human sciences have allowed an extraordinary advance in computer and analysis. From these contributions, data management and data analysis techniques such as data mining have been developed. The purpose of this study is to "apply data mining techniques for the creation of rainfall precipitation trend rules in the state of Morelos", using temperature data, heat units and rainfall from 29 climatological stations located within the

*miguel.leta@gmail.com

Nota: Este artículo de investigación es parte de Ingeniería–Revista Académica de la Facultad de Ingeniería, Universidad Autónoma de Yucatán, Vol. 22, No. 2, 2018, ISSN: 2448-8364.

state. The series of temperature and rainfall data were evaluated through quality and homogeneity control mechanisms. The data mining techniques that were used are: the k-means partitioning and decision trees C4.5; From the first technique, 58 pairs of temperature groups were generated, which represent the climatology of the area for each station, these grouping were used as discretization thresholds of the temperature series to simplify the 29 sets of rainfall precipitation trend rules generated with the algorithm C4.5.

Keywords: data mining, quality control, Morelos, rainfall, temperature.

1. Introducción

El planeta a lo largo del tiempo se ha visto afectado por distintos acontecimientos climatológicos, es conocido que los huracanes en el territorio nacional mexicano no solo afectan la zona donde tiene un impacto directo, si no por el contrario en zonas que se encuentran a su alrededor con precipitaciones pluviales inusuales y cambios repentinos de temperatura.

De acuerdo con la historia, el estado de Morelos ha registrado el 60% de inundaciones del país *(CONAGUA et al. 2010)*. Siendo un área de estudio factible para la implementación de técnicas de minería de datos sobre información de precipitación pluvial y temperatura con periodos de datos de 50 años (1960 – 2010), información obtenida de la base de datos CLICOM (por sus siglas en inglés, Climate Computer Project) administrada por el Servicio Meteorológico Nacional.

El analizar información histórica meteorológica ha permitido identificar cambios y variaciones importantes en el comportamiento del clima, como se ha realizado en diversos estudios enfocados al cambio climático, estos estudios generan la pauta para crear líneas de mitigación y adaptación a corto y mediano plazo dentro de todos los sectores socioeconómicos y sociales, en los últimos años la red de estaciones climatológicas han sido un insumo importante para este tipo de análisis, en el año 2011 se realizó un estudio sobre el estado de Guerrero utilizando la red de estaciones agroclimatológicas de México y obteniendo una variación visible en el clima de esa área *(Mendoza et al. 2011)*.

El territorio del estado de Morelos se encuentra localizado dentro de la gran cuenca del río Balsas, la cual ha sido catalogada como la 18ª región hidrológica administrativa en el país y como una de las más grandes a nivel nacional, ya que cuenta con una extensión territorial del 117,405 km^2 *(CONAGUA 2010)*.

El volumen de información que se genera a diario con respecto a las condiciones climatológicas, atmosféricas e hidrológicas crece continuamente, generando grandes repositorios de datos que contiene información esencial para el análisis, sin embargo, el analizar estos repositorios con técnicas de análisis estadísticas comunes no es una tarea fácil, es aquí donde las técnicas de minería de datos ayudan a agilizar este proceso, aplicando modelos computacionales basado en aprendizaje inductivo y procesos de control de calidad y homogeneidad para obtener reglas de tendencia para la predicción de la precipitación pluvial en el estado de Morelos, transformando la información global de bajo nivel en conocimiento de alto nivel, el cual es comprensible al entendimiento humano *(Riquelme et al. 2006)*.

1.1. Minería de datos por aprendizaje inductivo

La minería de datos se ha descrito como la extracción no trivial de información implícita, previamente desconocida y con elevado potencial de utilidad *(Joseph L. Fleiss y Joseph Zubin 1992)*. Permite descubrir relaciones significantes, patrones y tendencias al explorar y analizar grandes cantidades de datos *(Pérez y Santín 2008).*

Aplicar técnicas de minería de datos a la información de las redes de estaciones climatológicas, permite obtener conocimiento que se encuentra oculto y que es de gran utilidad para la toma de decisiones, al generar reglas de tendencia basadas en la precipitación pluvial y temperatura ambiental, mostrando aplicación con estudios para calcular la severidad de daños en zonas afectadas por incendios, inundaciones u otro tipo de evento climatológico *(Fernández-Manso, A 2008)* y para estudios referentes a la telemedición satelital *(Pech 2002)*.

La minería de datos es solo un paso de todo un proceso de extracción de conocimiento (KDD por sus siglas en inglés: Knowledge Discovery in Databases), utilizando algoritmos que permiten realizar el análisis de los datos a fin de producir un conjunto de patrones o modelos sobre ellos *(Fayyad 1996)*.

1.1.1. Algoritmo de árboles de decisión C4.5

Desarrollado en el año de 1993 por J.R. Quinlan como una extensión de mejora para el algoritmo ID3 (por sus siglas en inglés, Induction Decision Trees), perteneciente a la familia de algoritmos descendientes (TDIDT, por sus siglas en inglés: Top Down Induction Tress). El algoritmo genera un árbol de decisión a partir de los datos suministrados mediante particiones que se realizan de manera recursiva, maximizando la homogeneidad entre los conjuntos, utilizando la función de la entropía creada por Rudolph Clausius en 1854 y publicada por Claude E. Shannon en el año de 1948 *(Shannon 1948)*.

Las particularidades de los árboles de decisión C4.5 con respecto a su antecesor el algoritmo ID3 son la siguientes:

1. Permite el manejo de valores continuos o desconocidos.
2. Los árboles se crean con un menor número de ramas, debido a que cada rama o subárbol puede representar a un atributo o a un conjunto de atributos.
3. Distribución de la información y selección de atributos con base a pruebas heurísticas combinadas.
4. La búsqueda para crear el árbol de decisión se basa en la estrategia de búsqueda en profundidad (DFS, por sus siglas en inglés, Depth-First Search) *(Moreno 2009)*.

El modelo C4.5 implementa diversos métodos heurísticos en el proceso de elección y discriminación de los atributos en la búsqueda recursiva, debido a su enfoque en lograr una homogeneidad entre los conjuntos de datos, comenzando por el cálculo de la entropía, seguida por la función de la ganancia y finalmente la razón de la ganancia.

La entropía permite medir la incertidumbre que existe dentro de un sistema; con el fin de medir la probabilidad de que ocurra cada uno de los posibles resultados.

$$Entropía(S) = info(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \quad (1)$$

Siendo p_i la proporción de ejemplos de la clase C_i en el conjunto S.

La función de ganancia dentro del algoritmo permite conocer la diferencia entre la entropía de un nodo con respecto a sus descendientes.

$$Ganancia(S,A) = Entropía(S) - \sum_{v=1}^{n} \frac{|S_v|}{|S|} - Entropía(S_v) \quad (2)$$

Siendo *A* un atributo del conjunto *S* y *v* un subconjunto de ejemplos definidos del conjunto *S* con respecto al atributo *A*.

Asimismo, el algoritmo implementa la razón de ganancia (Gain Ratio) para llevar acabo la elección de los atributos, ya que esta ecuación evita que las variables con mayor número de categorías sean beneficiadas de manera arbitraria en la selección.

$$RazónGanancia(S,A) = \frac{Ganancia(S,A)}{-\sum_{i=1}^{c} p_i \log_2 p_i} \quad (3)$$

$$donde \;\; p = \frac{n_c}{n};$$

Sea n_c el número de muestras de clase *c*, y *n* el número de muestras totales del universo en proceso.

1.1.2. Algoritmo de agrupamiento K-means

En el año 1967 fue presentado como un algoritmo de agrupamiento por MacQueen, clasificado como un algoritmo de método de particionado y de recolección de elementos, siendo que emplea técnicas no jerárquicas o de partición *(MacQueen 1967; Garre et al. 2007)*.

K-means conglomera un conjunto de datos del universo en un número de clústeres o grupos definido a priori a la implementación como se muestra en la sección de metodología. Este algoritmo comienza con un conjunto de centroides obtenidos de manera arbitraria para cada clúster, durante el proceso estos centroides son reasignados a otros elementos del conjunto hasta que los clústeres convergen, es decir, hasta que no existe diferencia entre el centroide previo y el sucesor de acuerdo al cálculo de la ecuación de distancia (ver, ecuación 4).

$$\boldsymbol{c_j} = \frac{1}{|c_j|} \sum_{\forall x \in c_j} x \quad (4)$$

siendo z el elemento del conjunto de c_j

Para el cálculo de la distancia entre los elementos del conjunto y los centroides se utilizó la función de distancia euclidiana; ecuación que calcula el valor de la distancia entre un par de coordenadas dadas por dos conjuntos, el conjunto $x = \{x_1,\ldots,x_n\}$ y el conjunto $y = \{y_1,\ldots,y_n\}$ *(Torrente 2007)*.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

2. Metodología

Es enfocada al tratamiento de datos climatológicos y en la implementación de técnicas de minería de datos para clasificación y generación de reglas. Se divide en cinco fases incorporando la metodología CRISP-DM (por sus siglas en inglés, Cross Industry Standard Process for Data Mining):

1. Extracción de los datos climatológicos
2. Depuración de los datos
3. Ejecución de mecanismos de control de calidad
4. Ejecución de mecanismos de homogeneización de datos
5. Aplicación de técnicas de minería de datos

2.1. Extracción de los datos climatológicos

Se enfoca en la recolección de los datos de la base de datos CLICOM; los datos que son recolectados deben comprender al menos un periodo de información histórica de 50 años, para poder obtener series en procesos posteriores con un mínimo de 30 años de acuerdo a las recomendaciones de la Organización Meteorológica Mundial (OMM) para el análisis climatológico.

A continuación, se listan los pasos que se utilizan para la extracción de los datos y estructuración:

1. Determinar los datos importantes para el estudio: Se seleccionan los datos de precipitación, temperatura máxima, temperatura mínima y unidades de calor.
2. Realizar la conexión con la base de datos CLICOM del Servicio Meteorológico Nacional.
3. Generar la consulta para las estaciones localizadas dentro del estado de Morelos.
4. Estructurar los datos de acuerdo al formato de entrada de las herramientas de control de calidad utilizadas, en este caso para el paquete de software RClimdex.

2.2. Depuración de los datos

La depuración de datos tiene como objetivo eliminar el ruido y la información que proveen significado poco relevante al análisis, en esta fase se incorporan dos procesos de la metodología CRISP-DM; compresión de los datos y preparación de los datos.

Los pasos del proceso de depuración de los datos son:

1. Seleccionar las series estacionales (precipitación, temperatura y unidades de calor).
2. Selección de series de datos de acuerdo al porcentaje de años total y efectivos de la serie.
3. Localizar variables serán analizadas con control de calidad y homogeneidad
4. Integración de nuevas variables a partir de las ya existentes si es requerido, sin alterar su significado.

2.3. Ejecución de mecanismos de control de calidad

Con el objetivo de utilizar series de información confiables y una vez realizada la extracción y depuración de los datos del estudio, se procede a la implementación de los controles de calidad con el paquete de software RClimdex, utilizando los índices de cambio climático sugeridos por ETCCDI (por sus siglas en inglés, Experts Team in Climate Change Detection and Indices) para verificar que no existan valores anómalos generados por eventos relacionados con errores de medición, transmisión, cambios de instrumentación entre otros. A continuación, se muestra el proceso para aplicar los mecanismos de control de calidad:

1. Ejecutar el paquete de software RClimdex.
2. Cargar los archivos con las series de datos en formato diario de la información a verificar; datos previamente estructurados y depurados.
3. Establecer el valor de desviación estándar para el análisis de control de calidad.
4. Aplicar el control de calidad y analizar los resultados obtenidos en los archivos de bitácora de error (log files) con información histórica de la climatología de la zona.
5. Realizar la corrección de las series de datos si es necesario y efectuar nuevamente el proceso de control de calidad hasta que la serie apruebe los controles de calidad.

2.4. Ejecución de mecanismos de homogeneización de datos

Los mecanismos de homogeneidad permiten agregar seguridad y representatividad a los datos que se analizan, permitiendo únicamente fluctuaciones que representen cambios por efectos naturales.

El mecanismo de homogeneización se realiza con la implementación del programa RHTest, a fin de encontrar los puntos de cambio o de quiebre dentro de la serie. A continuación, se muestran los pasos para aplicar dicho mecanismo:

1. Ejecutar el paquete de software RHTest.

2. Transformar las series de datos diarias a series mensuales, estas series deberán tener la estructura utilizada por el paquete de software RClimdex.
3. Cargar los archivos generados anteriormente para su análisis, aplicar los controles de homogeneidad y analizar los archivos de bitácoras de error (log files) con información histórica de la climatología de la zona.
4. Realizar la corrección de los puntos sugeridos en las bitácoras de error si es necesario y efectuar nuevamente el proceso de homogeneidad.

2.5. Aplicación de técnicas de minería de datos

El aplicar las técnicas de minería de datos permite culminar la implementación de la metodología, esta incluye los últimos tres procesos de la metodología CRISP-DM; modelado, evaluación y visualización. A continuación, se enlistan los pasos para efectuar esta última fase:

1. Aplicar el algoritmo K-means sobre las series de temperatura máxima y mínima.
2. Extraer resultados del algoritmo K-means de cada una de las series de datos.
3. Discretizar las series de datos de temperatura máxima y mínima de acuerdo a los resultados.
4. Estructurar las series de datos con las variables de precipitación, temperatura máxima y mínima (previamente discretizadas) y unidades de calor a formato de entrada del algoritmo C4.5 del paquete de software WEKA.
5. Aplicar el algoritmo de árboles de decisión C4.5 con las series de datos.
6. Extraer y analizar los resultados obtenidos con el algoritmo de modelos de decisión.

3. Resultados

En el estado de Morelos existen instaladas aproximadamente setenta estaciones climatológicas situadas en puntos estratégicos, para medir las condiciones meteorológicas de una región específica.

De acuerdo a los resultados de obtenidos derivados de las primeras dos fases de la metodología: extracción de los datos climatológicos y depuración de los datos, fueron seleccionadas veintinueve estaciones de las setenta disponibles las cuales se muestran en la tabla 1.

Número	Estación	Latitud	Longitud	Altitud (m.)
17012	OAXTEPEC	18°54’23’’	98°58’13’’	1380
17013	TEMILPA	18°42’21’’	99°5’38’’	1135
17014	TEMIXCO	18°51’16’’	99°13’38’’	1283
17015	TEPALCINGO	18°35’47’’	98°50’37’’	1160
17016	TEQUESQUITENGO	18°36’40’’	99°15’35’’	932
17018	TICUMAN	18°45’33’’	99°7’16’’	970
17019	TILZAPOTLA	18°29’16’’	99°18’22’’	1303
17020	TLACOTEPEC	18°48’48’’	98°45’0’’	1754
17021	TLACUALERA	18°37’0’’	98°56’37’’	1250
17022	TRES CUMBRES	19°2’12’’	99°15’29’’	2639
17024	YAUTEPEC	18°51’16’’	99°1’18’’	1343
17026	C.A.E. LA VICTORIA	18°38’12’’	99°12’3’’	1364

17028	JONACATEPEC	18°41’30’’	98°49’29’’	1350
17031	JOJUTLA (DGE)	18°35’2’’	99°11’3’’	959
17033	XICATLACOTLA (CFE)	18°27’0’’	99°6’0’’	1095
17036	LAGUNILLAS	18°29’0’’	98°43’0’’	1010
17038	NEXPA	18°31’12’’	99°8’42’’	800
17039	TLACOTENCO	19°2’21’’	99°5’38’’	2836
17043	YECAPIXTLA	18°53’30’’	98°51’30’’	1600
17044	E.T.A. 040 AMACUZAC	18°35’55’’	99°22’10’’	1278
17047	HUITZILAC	19°3’30’’	99°16’27’’	2801
17054	MOYOTEPEC	18°40’15’’	98°58’31’’	1154
17056	SAN PABLO HIDALGO	18°34’55’’	99°2’41’’	925
17057	EL LIMON	18°31’52’’	98°56’15’’	1248
17058	CUENTEPEC	18°51’37’’	99°19’35’’	1487
17060	ALPONOCAN	18°55’52’’	98°41’23’’	2769
17061	APANCINGO	18°40’48’’	99°27’49’’	1152
17072	ALPUYECA	18°44’6’’	99°15’57’’	1025
17076	PUENTE DE IXTLA	18°37’45’’	99°19’33’’	903

Tabla 1. Estaciones climatológicas utilizadas del estado de Morelos.

El proceso de control de calidad de los datos fue aplicado a las veintinueve estaciones climatológicas seleccionadas para las variables meteorológicas de precipitación y temperatura de sus respectivas series de datos. A continuación, se muestra el desarrollo de la metodología y los resultados sobre la estación 17024 “Yautepec, latitud: 18°51’16’’, longitud: 99°1’18’’ y altitud: 1343 m.”:

El paquete de software RClimdex realiza el control en bloques de 10 años para verificar que no existan datos anómalos con la implementación de los índices de cambio climático de ETCCDI, como se puede apreciar para esta estación climatológica (ver, Figura 1, Figura 2 y Figura 3) las series de datos sobre temperatura máxima (periodo de 1955 a 1964), temperatura mínima (periodo de 1985 a 1994) y sobre la precipitación (periodo de 1995 a 2004) muestra en las series de datos valores que no generan cambios por valores anómalos durante el tiempo.

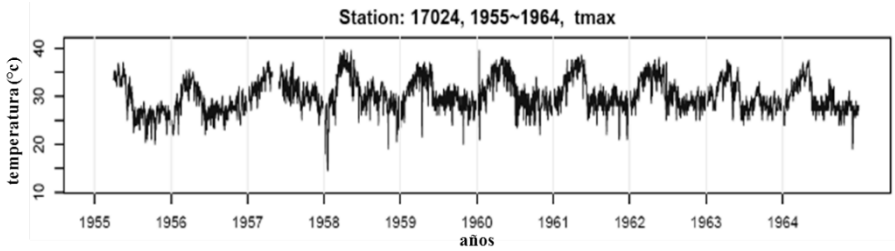

Figura 1. Resultado del control de calidad para la estación 17024, sobre la temperatura máxima durante el rango de 1955 a 1964. Fuente: RClimdex.

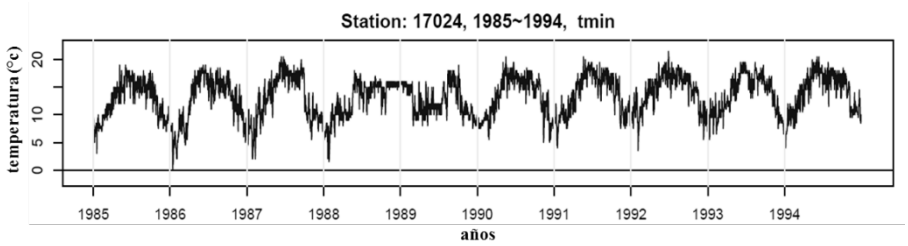

Figura 2. Resultado del control de calidad para la estación 17024, sobre la temperatura mínima durante el rango de 1985 a 1994. Fuente: RClimdex.

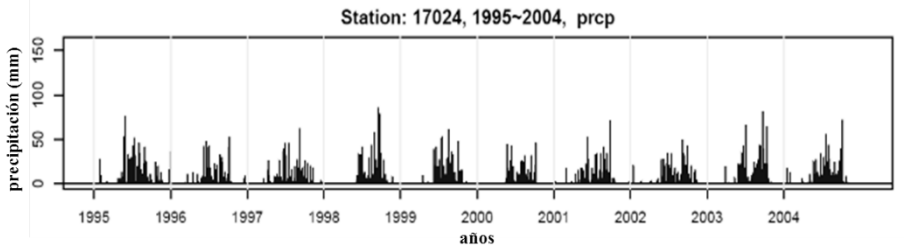

Figura 3. Resultado del control de calidad para la estación 17024, sobre la precipitación, durante el rango de 1995 a 2004. Fuente: RClimdex.

El resultado del control de calidad también muestra los valores extremos posibles para la serie de temperatura diaria indicando la fecha de ocurrencia, como se puede ver en la tabla 2 para la estación 17024, mismos que corresponden con eventos y la climatológica de la zona.

Año	Mes	Día	tmaxb	tmax	tmaxa	tminl	tmin	tmina
1957	7	28	20.37	35.5	38.23	7.09	9	24.09
1958	1	6	19.64	18	37.85	-3.46	9	19.04
1958	1	19	14.76	14.5	41.84	-2.98	11	18.63
1958	1	20	16.07	15	40.52	-0.99	8	16.24
1959	4	14	23.69	22	44.55	4.35	16	22.86
1959	4	16	23.35	21.5	44.37	4.87	11	22.79
1966	9	21	19.18	19	37.54	7.47	16	24.3
1967	1	10	15.31	13	40.8	-2.9	7	19.13
1967	1	11	15.29	12	40.76	-1.84	9	17.84
1967	9	23	18.92	23	37.72	7.4	7	23.76
1976	4	2	24.65	23	43.53	4.29	11	20.87
1976	5	12	22.87	22	45.06	6.25	16	25.53
1976	5	13	20.83	20.5	46.03	6.23	17.5	26.16
1976	5	18	20.3	20	45.45	7.04	15.5	24.82
1976	5	26	19.77	20	45.41	8.27	18	24.97
1980	1	26	20.06	20	38.72	-1.83	10.5	18.91
1983	2	25	20.59	20	41.34	0.27	8	18.72
1983	3	13	21.15	19.5	43.9	2.32	12.5	19.91

1992	1	28	21.43	22	37.55	-1.74	13.5	18.54
1995	12	31	19.54	20.5	37.37	-1.59	11	17.57
2010	2	18	20.44	21.5	41.11	1	13	17.31

Tabla 2. Valores extremos de temperatura y precipitación detectados por RClimdex en la estación 17024.

Siendo “tmaxb, tmaxa y tmax” la temperatura máxima menor, mayor, registrada; “tminb, tmina y tmin” la temperatura mínima menor, mayor y registrada.

Los resultados del proceso de homogeneidad con el paquete de software RHtest permitieron comprobar que las series de datos climatológicas mantuvieran una uniformidad durante el periodo de tiempo de la muestra verificando la tendencia estacional interanual. A continuación, se muestra el resultado obtenido para la estación 17024:

Para la variable de precipitación el resultado del proceso de homogeneidad no encontró registros de datos que generen cambios que cambiaran la tendencia estacional de la serie con respecto a la climatología de la región (ver, Figura 4).

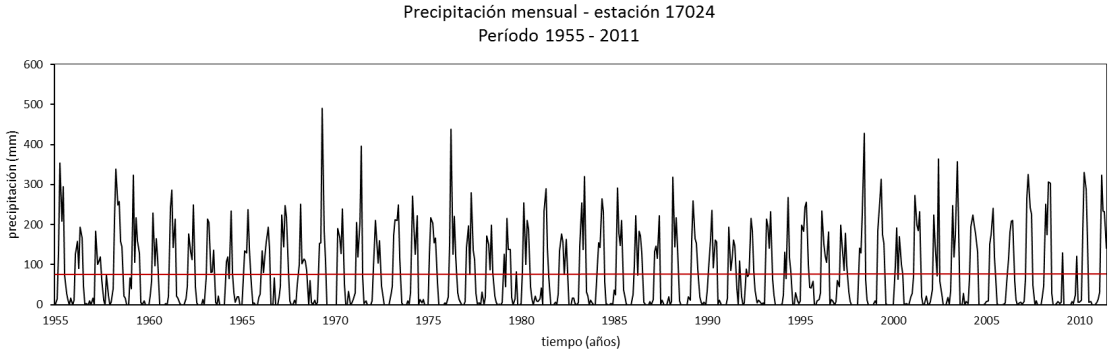

Figura 4. Precipitación pluvial mensual - estación 17024 (1955 – 2011). Fuente: Elaboración propia.

Para la variable meteorológica de temperatura máxima el proceso de homogeneidad localizo ocho valores mensuales que requieren verificación, los cuales se encuentran en las fechas: 1956-11, 1977-06, 1978-11, 1985-09, 1988-05, 1989-03, 1992-12 y 2004-03. Una vez realizado el análisis de estos valores y la corrección necesaria, la serie de datos muestra una tendencia positiva la cual describe un comportamiento estable para la estación y para el tipo de variable como se muestra en la figura 5, con lo que se determina que esta serie no interrumpe las variaciones interanuales y los datos pueden ser utilizados para las técnicas de minería.

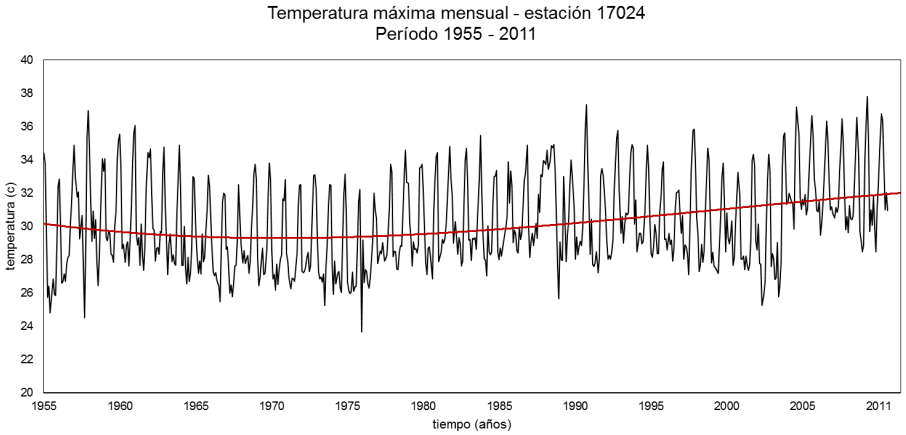

Figura 5. Temperatura máxima mensual - estación 17024 (1955 – 2011). Fuente: Elaboración propia.

Los resultados obtenidos para la variable meteorológica de temperatura mínima localizaron dos valores mensuales que requieren verificación, los cuales se encuentran en las fechas: 1959-10 y 1966-09. Al igual que con la variable temperatura máxima una vez realizada la corrección necesaria, como se aprecia en la figura 6 la serie de datos muestra una tendencia sinusoidal la cual describe un comportamiento estable sobre la línea del tiempo y manteniendo los ciclos interanuales presentes, con lo que se determina que la serie de datos también puede ser utilizada.

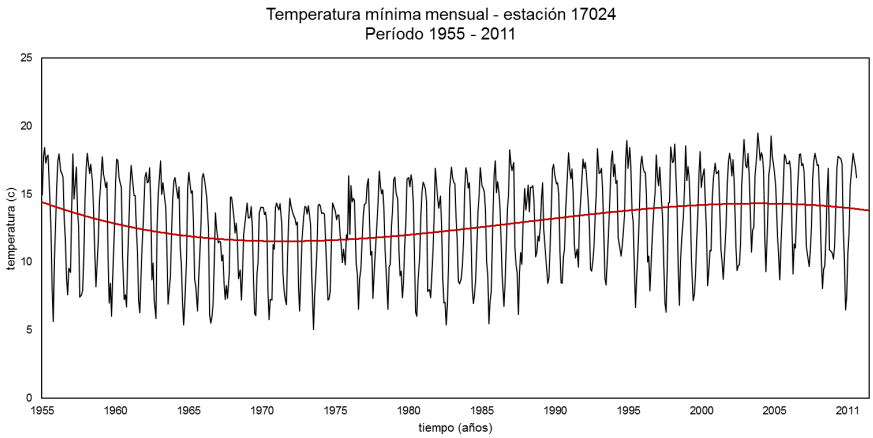

Figura 6. Temperatura mínima mensual - estación 17024 (1955 – 2011). Fuente: Elaboración propia.

Realizados los procesos anteriores se procede a la aplicación de las técnicas de minería de datos.

El algoritmo K-means fue aplicado con WEKA utilizando un universo de datos de 18,013 registros diarios, los resultados obtenidos para la estación 17024 de las variables de temperatura máxima y mínima se muestran en las tablas 3 y 4:

	Registros	Clúster 0	Clúster 1	Clúster 2	Clúster 3
Datos por grupo	18013	3747	7291	4119	2856
Umbrales de temperatura	N/A	12° a 27.3 °	27.4 ° a 30.3 °	30.4° a 33.4°	33.5° a 40 °
Etiqueta de rango	N/A	baja	media	alta	muy alta

Tabla 3. Resultado del algoritmo k-means en temperatura máxima de estación 17024.

	Registros	Clúster 0	Clúster 1	Clúster 2	Clúster 3
Datos por grupo	18013	3750	5352	5092	3819
Umbrales de temperatura	N/A	-1° a 9.3 °	9.4 ° a 13.1 °	13.2° a 16.2°	16.3° a 21.5 °
Etiqueta de rango	N/A	baja	media baja	media	alta

Tabla 4. Resultado del algoritmo k-means en temperatura mínima de estación 17024.

Con los umbrales de temperatura generados por el algoritmo K-means se realiza la discretización de las series de temperatura máxima y mínima de la estación 17024, para obtener valores nominales, esto permite al algoritmo de árboles de decisión C4.5 simplificar las reglas.

Cada estación genera sus umbrales de temperatura de acuerdo a la ejecución propia del algoritmo K-means y a sus series de datos.

La discretización de los datos de precipitación pluvial para cada una de las estaciones utiliza valores que engloban a siete condiciones de la precipitación, que son: nula, baja, ligera, moderada, fuerte, intensa y torrencial (ver, Figura 7).

[0 - 0] = nula
[0.1 – 0.5) = baja
[0.5 – 5) = ligera
[5 – 20) = moderada
[20 – 70) = fuerte
[70 – 150) = intensa
[> 150) = torrencial

Figura 7. Rangos de precipitación para discretización.

Una vez obtenidos los umbrales de temperatura y haber discretizado cada una de las series de datos de la estación 17024, se configura y se crea el árbol de decisión C4.5 con el entorno de análisis WEKA, y finalmente se procesan los resultados para

generar las reglas de precipitación pluvial. A continuación, se muestran las reglas generadas para la estación 17024:

Sección 1. Si la temperatura mínima y la temperatura máxima son bajas,

a. Y las unidades de calor son menores o iguales a 5.12, la precipitación pluvial tiende a ser nula.

b. Y las unidades de calor son menores o iguales a 5.12, pero mayores a 4.02, la precipitación pluvial tiende a ser fuerte.

c. Y las unidades de calor son menores o iguales a 4.02, la precipitación pluvial tiende a ser nula.

d. Y las unidades de calor son mayores a 5.12, la precipitación pluvial tiende a ser nula.

Sección 2. Si la temperatura mínima es baja o media baja, y la temperatura máxima es alta o muy alta, la precipitación pluvial tiende a ser nula.

Sección 3. Si la temperatura mínima es media baja, y la temperatura máxima es baja,

a. Y las unidades de calor son menores o iguales 8.8, la precipitación pluvial tiende a ser moderada.

b. Y las unidades de calor son menores o iguales a 8.8, pero mayores a 3, la precipitación pluvial tiende a ser ligera.

c. Y las unidades de calor son mayores a 8.8, la precipitación pluvial tiende a ser nula.

Sección 4. Si la temperatura mínima es media baja, y la temperatura máxima es media, alta o muy alta, la precipitación pluvial tiende a ser nula.

Sección 5. Si la temperatura mínima es media, y la temperatura máxima es baja,

a. Y las unidades de calor son menores o iguales a 11.52, la precipitación pluvial tiende a ser nula.

b. Y las unidades de calor son menores o iguales a 11.52, pero mayores a 6.7, la precipitación pluvial tiende a ser moderada.

c. Y las unidades de calor son mayores a 11.52, la precipitación pluvial tiende a ser nula.

Sección 6. Si la temperatura mínima es media, y la temperatura máxima es media, alta o muy alta, la precipitación pluvial tiende a ser nula

Sección 7. Si la temperatura mínima es media, y la temperatura máxima es baja,

a. Y las unidades de calor son menores o iguales a 10.53, la precipitación pluvial tiende a ser ligera.

b. Y las unidades de calor son menores o iguales a 10.53, pero mayor a 8.37, la precipitación pluvial tiende a ser nula.

c. Y las unidades de calor son mayores a 10.53, la precipitación pluvial tiende a ser moderada.

Sección 8. Si la temperatura mínima y máxima son alta,

a. Y las unidades de calor son menores o iguales a 14.75, la precipitación pluvial tiende a ser nula.

b. Y las unidades de calor son menores o iguales a 15.55, pero mayores a 14.75, la precipitación pluvial tiende a ser nula.

c. Y las unidades de calor son menores o iguales a 15.89, pero mayores a 15.55, la precipitación pluvial tiende a ser ligera,

d. Y las unidades de calor son mayores a los valores del conjunto 15.89, la precipitación pluvial tiende a ser moderada.

Sección 9. Si la temperatura mínima es muy alta, y la temperatura máxima es alta,

a. Y las unidades de calor son menores o iguales a 14.59, la precipitación pluvial tiende a ser nula.
b. Y las unidades de calor son menores o iguales a 14.59, pero mayores a 14.02, la precipitación pluvial tiende a ser ligera.
c. Y las unidades de calor son menores o iguales a 14.59, pero mayores 14.54, la precipitación pluvial tiende a ser moderada.
d. Y las unidades de calor son mayores a 14.59, la precipitación pluvial tiende a ser nula.

Sección 10. Si la temperatura mínima es alta, y la temperatura máxima es muy alta, la precipitación pluvial tiende a ser nula.

A continuación en la tabla 5 se muestra el resultado de la comparación de 10 registros de datos tomados aleatoriamente de la estación 17024 contra el conjunto de reglas generado por el algoritmo C4.5, donde se puede apreciar que realiza una asignación de precipitación pluvial que representa al valor real registrado y en algunos casos muestra un valor equiparable al mismo, teniendo como referencia que en los distintos tipos de modelos de pronóstico la precipitación es considerada como una variable meteorológica difícil de pronosticar el conjunto de reglas obtuvo aproximadamente un 90% de precisión en la clasificación para este conjunto de datos muestra.

Fecha	Precipitación	Temperatura máxima	Temperatura mínima	Unidades de calor	Clasificación	Resultado
1975/06/22	2.5 mm ligera	26 °C baja	13 °C media baja	10.25 ° día	Sección 3 índice c	nula
1984/03/12	0 mm nula	33 °C alta	10 °C media baja	11.87 ° día	Sección 4	nula
1967/01/10	47 mm fuerte	13 °C baja	7° C baja	4.14 ° día	Sección 1 índice b	fuerte
1999/11/28	0 mm nula	26.5 °C baja	13.56 °C baja	8.76 ° día	Sección 1 índice d	nula
2002/05/02	0 mm nula	35 °C muy alta	16.5 °C alta	14.56 ° día	Sección 10	nula
1967/09/23	43 mm fuerte	23 °C baja	7 °C baja	4.14 ° día	Sección 1 índice b	fuerte
1979/06/11	0 mm nula	32 °C alta	15.5 °C media	14.89 ° día	Sección 6	nula
2004/04/07	0 mm nula	33.5 °C muy alta	19 °C alta	14.41 ° día	Sección 10	nula
1961/01/11	2 mm ligera	27 °C baja	10 °C media baja	7.29 ° día	Sección 3 índice b	ligera
2001/07/06	0 mm nula	30 °C alta	15 °C media	12.25 ° día	Sección 6	nula

Tabla 5. Resultados de clasificación de datos muestra de la estación 17024.

En la tabla 6 se muestra el resumen de los resultados obtenidos por el algoritmo C4.5 para las veintinueve estaciones climatológicas utilizadas, mostrando el tamaño de árbol generado y el número de reglas encontradas con dicha técnica.

Estación	Nombre	Instancias clasificadas	Tamaño de árbol	Número de reglas
17012	OAXTEPEC	17,012	125	67
17013	TEMILPA	19,341	99	53
17014	TEMIXCO	19,412	39	24
17015	TEPALCINGO	18,483	37	22
17016	TEQUESQUITENGO	13,737	51	30
17018	TICUMAN	17,828	123	68
17019	TILZAPOTLA	20,212	37	23
17020	TLACOTEPEC	19,714	127	68
17021	TLACUALERA	17,281	65	37
17022	TRES CUMBRES	8,826	195	103
17024	YAUTEPEC	18,013	73	32
17026	C.A.E. LA VICTORIA	15,621	145	77
17028	JONACATEPEC	15,395	117	63
17031	JOJUTLA (DGE)	10,073	103	55
17033	XICATLACOTLA (CFE)	14,614	31	19
17036	LAGUNILLAS	11,995	37	22
17038	NEXPA	12,582	75	43
17039	TLACOTENCO	9,800	153	82
17043	YECAPIXTLA	11,714	125	68
17044	E.T.A. 040 AMACUZAC	11,526	99	55
17047	HUITZILAC	17,753	221	115
17054	MOYOTEPEC	8,989	51	30
17056	SAN PABLO HIDALGO	9,655	33	20
17057	EL LIMON	11,496	61	34
17058	CUENTEPEC	10,130	75	42
17060	ALPONOCAN	8,168	127	69
17061	APANCINGO	10,667	59	34
17072	ALPUYECA	10,175	109	60
17076	PUENTE DE IXTLA	10,160	31	19

Tabla 6. Resultados del algoritmo C4.5 de las 29 estaciones climatológicas.

6. CONCLUSIONES

Este estudio presentó la metodología y resultados para la extracción de las reglas de precipitación pluvial en el estado de Morelos, dichas reglas fueron extraídas con la implementación de técnicas de minería de datos en registros diarios brindados por la base de datos CLICOM, en específico a las estaciones climatológicas tradicionales localizadas dentro de las 7 subcuencas hidrológicas de Morelos pertenecientes a la cuenca Balsas.

Los mecanismos de control de calidad y homogeneidad fueron aplicados a las series de datos de las variables meteorológicas precipitación pluvial y temperatura a 29 de 70 estaciones climatológicas disponibles, es decir, el 41% de estaciones totales. Los controles de calidad utilizaron los índices

propuestos por el Grupo de Expertos en Detección e Índices de Cambio Climático, estos permitieron detectar en el universo de datos de aproximadamente 400,372 registros diarios por variable los valores anómalos que fueron generados por errores propios del instrumento, de transmisión y de medición, y los valores considerados extremos generados por fenómenos naturales como frentes fríos, tormentas tropicales, huracanes entre otros; la homogeneidad permitió analizar la tendencia de las series en valores mensuales a fin de localizar interrupciones entre los ciclos estacionales, las estaciones mostraron un comportamiento estable con una tendencia sinusoidal en temperatura mínima y ligeramente creciente en temperatura máxima.

El algoritmo k-means permitió generar agrupaciones de las variables de temperatura máxima y mínima para las 29 estaciones climatológicas, obteniendo un total de 58 conjuntos de rangos o umbrales de dichas variables, estos rangos o umbrales de temperatura permiten representar la climatología propia de la región donde se encuentra situada cada una de la estaciones, debido a la diversidad de climas que existen en el estado de Morelos; teniendo zonas donde las temperaturas superan los 30° centígrados durante el día y zonas que durante la noche el termómetro desciende por debajo de los 0° centígrados.

Finalmente, la creación de las reglas se alcanzó con la aplicación de cada uno de los procesos anteriormente descritos en combinación con las técnicas de minería de datos para generar árboles de decisión C4.5 de Quinlan, utilizando un periodo mínimo de 30 años de datos efectivos (series de datos limpios, depurados y discretizados), logrando construir 29 conjuntos de reglas de asociación para sistemas de predicción de precipitación pluvial; el número promedio de reglas generadas para cada una de las estaciones es de 50, siendo la estación 17047 localizada en el municipio de Huitzilac la que generó el mayor número de reglas con un valor de 115, mientras que al estación 17076 localizada en el municipio de Puente de Ixtla generó el menor número con un tamaño de 19 reglas.

4. RECONOCIMIENTOS

Los datos del CLICOM fueron proporcionados por la Coordinación General del Servicio Meteorológico Nacional a través del CICESE.

5. REFERENCIAS

CONAGUA. (2010). Programa Hídrico Visión 2030 del Estado de Morelos. México: Secretaría de Medio Ambiente y Recursos Naturales, Capitulo II ¿Cómo planeamos?, Secretaría del Medio Ambiente y Recursos Naturales (eds), 9-13, Tlalpan, México.

CONAGUA, SEMARNAT & OCB. (2010). Estadísticas del Agua en la Cuenca del Río Balsas, 2010. Secretaría del Medio Ambiente y Recursos Naturales (eds), Tlalpan, México.

Fayyad, U., Piatetsky-Shapiro, G., & Padhraic, S. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence AI Magazine, 37-54, Volume 17, number 3.

Joseph L. Fleiss & Joseph Zubin (2010) On The Methods And Theory Of Clustering, Multivariate Behavioral Research, 235-250, 4:2.

Garre, M., Cuadrado, J. J., Silicia, M. A., Rodríguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. Revista Esañola de Innovación, Calidad e Ingeniería de Software, 6-22, Vol. 3, No. 1.

Fernández-Manso, A. 'Minería de Datos' (Data mining) aplicada a imágenes de satélite para el análisis y la cuantificación de daños por incendios forestales en Castilla y León, Junta de Castilla y León, Universidad de León, España.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297, Volume 1: Statics.

Mendoza-Uribe, I., Sánchez-Lobato, R., Cruz-Enríquez, O. Z., & Pineda-Flores, N. A. (2011). Análisis de tendencias de cambio climático y capacitación en el uso de la información climática en el estado de Guerrero. Jiutepec, Morelos.

Moreno-Montiel, B. (2009). Minería sobre grandes cantidades de datos. Tesis de maestría, Posgrado en Ciencias y Tecnologías de la Información, Universidad Autónoma Metropolitana D.F., México.

Pech-Palacio, M. A. (2002). Adaptación y Uso de Minería de Datos Espaciales y no Espaciales. Tesis de maestría, Universidad de las Américas Puebla, Puebla, México.

Pérez-López, C., & Santín-González, D. (2008). Minería de Datos, Técnicas y Herramientas. España: Thomson Ediciones Parainfo S.A.

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias, Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, 10 (29)., 11-18

Shannon, C. E. (1948). A Mathematical Theory of Communication. The Bell system Technical Journal, 379-423,623-656.

Torrente-Orihuela, A. (2007). Métodos de clustering en datos de expression génica, Tesis doctoral, Universidad Carlos II de Madrid, Madrid, España.