APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS, PARA LA GENERACIÓN DE PRONÓSTICOS DE DISPONIBILIDAD DE AGUA EN LA CUENCA BAJA DEL RÍO SAN LORENZO

Ana Laura Herrera Prado[a], Diego Uribe Agundis[b], Arturo Ruíz Luna[c],

[a]Instituto Tecnológico de Mazatlán, anlauherrera@gmail.com, analaura@itmazatlan.edu.mx,
[b]Instituto Tecnológico de la Laguna, diegouribeagundis@gmail.com,
[c]Arturo Ruíz Luna, Centro de Investigación en Alimentación y Desarrollo, A.C., arluna@ciad.mx

RESUMEN

La Minería de Datos tiene un papel muy importante como tecnología de apoyo para explorar, analizar, comprender y aplicar el conocimiento adquirido de grandes volúmenes de datos, así como para identificar tendencias y comportamientos en la información, que faciliten una mejor comprensión de los fenómenos que nos rodean y ayudan en la toma de decisiones. En este estudio, se verificó la viabilidad del uso de diversas técnicas de minería de datos, como una herramienta computacional para predecir la disponibilidad de agua para fines agrícolas en la cuenca baja del río San Lorenzo. Este trabajo se basó en una revisión bibliográfica sobre estos temas y el análisis de repositorios disponibles en diversas instituciones mexicanas. Posteriormente, se definió, seleccionó y generó un dominio de datos. El cual se pre-procesó y transformó para entender el significado de los atributos, detectar errores de integración y lograr un dataset adecuado para construir modelos de series de tiempo. Las técnicas predictivas evaluadas fueron: Máquina de Soporte Vectorial, Regresión Lineal y Perceptrón Multicapas. El análisis y comparación de los patrones en series temporales arrojó como resultados: a) el escurrimiento del río San Lorenzo muestra dos comportamientos en periodos de tiempo distintos debido a la construcción y operación de la presa “José López Portillo” y b) las lluvias y evaporaciones registradas tienen un comportamiento estacional característico, reflejando los fenómenos meteorológicos esporádicos como huracanes o tormentas que ocurren en la región. Se obtuvieron correlaciones significativas al contrastar los resultados de las predicciones de las técnicas estudiadas, con los datos observados. Los coeficientes de correlación más altos se obtuvieron con la técnica Regresión Lineal. En conclusión, es viable el uso de técnicas de predicción para aplicarse en la generación de pronósticos de disponibilidad de agua hasta por un periodo de tres años, en la cuenca baja del río San Lorenzo.

1. INTRODUCCIÓN

En todos los sistemas hídricos donde el agua es limitada y tiene muchos usuarios, cada uno de ellos quisiera disponer siempre de cantidades suficientes de dicho recurso. Cualquier situación diferente, en muchas ocasiones produce inconformidades por parte de los usuarios, o en algunos casos, conflictos ambientales, sociales o productivos. Las autoridades encargadas de administrar el recurso, en el caso de México la Comisión Nacional del Agua (CONAGUA), tienen la dificultad de distribuir el agua de la manera más equitativa. Sin embargo, en cada periodo anual existe una disponibilidad finita de agua, la cual tiene que satisfacer a más usuarios cada vez. Para elaborar un análisis de disponibilidad realista del agua, se requiere contar con registros estadísticos de las entradas, que permitan conocer de manera adecuada la variación del volumen de lluvias, escurrimientos y almacenamiento (presas, embalses o acuíferos). Con esto se tendría la posibilidad de definir la manera de aprovechar el recurso agua, con la idea de impedir que el vital

líquido se termine de forma prematura. Cabe destacar que esta información se registra constantemente en repositorios oficiales de datos, mismos que pueden ser utilizados por la Minería de Datos, (en inglés data mining, DM). La Minería de Datos, es una de las técnicas más utilizadas actualmente para analizar y extraer la información útil de grandes bases de datos. Se fundamenta en varias disciplinas, como la estadística, las técnicas de visualización de datos, los sistemas para tomas de decisión, el aprendizaje automático o la computación paralela y distribuida, con la finalidad de extraer patrones, describir tendencias, predecir comportamientos y sobre todo, producir beneficios a aquellas entidades que posean amplias bases de datos de, aparentemente, escasa utilidad [1].

La vocación agrícola del Estado de Sinaloa, además de su riqueza hidrológica basada en los once ríos que lo irrigan, requiere de una comprensión adecuada de los balances de agua. Por otro lado, la diversidad climática, topográfica y biológica con la que cuenta el Estado de Sinaloa hace necesaria la determinación de los requerimientos de agua para uso agrícola (riego) y el mantenimiento de ambientes naturales entre otros usos. Este proyecto logró verificar que la minería de datos, aplicada a las bases de datos disponibles, permite determinar la disponibilidad de agua de una cuenca en Sinaloa, como la del Río San Lorenzo. Esta información podría servir de referencia para realizar el balance hídrico de dicha cuenca en cualquier momento dentro de las fechas en las que se tenga registro, así como constituirse en una herramienta de previsión de eventos futuros como posibles crecidas o sequías.

2. TEORÍA

La Minería de Datos es una etapa del proceso de generación de conocimiento a partir de bases de datos (KDD, por las siglas en inglés de Knowledge Discovery in Databases), e incluye el análisis de grandes volúmenes de datos, con el objetivo de encontrar relaciones no conocidas y resumirlas de forma novedosa y útil para los dueños de la información [2]. Durante ese proceso, se aplican técnicas y herramientas para extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y comprensible. Uno de los objetivos de la minería de datos es que mediante el uso de sus herramientas se pueden predecir futuras tendencias y comportamientos, permitiendo tomar decisiones proactivas y conductivas por un conocimiento a partir de la información. Asimismo, dichas herramientas exploran las bases de datos en busca de patrones ocultos, descubriendo información predecible que un experto no puede llegar a hallar porque se encuentra fuera de sus expectativas [3].

El ciclo hidrológico determina los movimientos del agua en el sistema terrestre a través de tres procesos básicos: la precipitación, la evapotranspiración y la escorrentía. Con ellos modula el clima y la dinámica fluvial que hace posible la vida en la Tierra. Debido a la diversidad de culturas, climas, así como las actividades políticas y financieras en cada país, se requieren nuevos y mejores procedimientos para llevar a cabo una correcta gestión integrada de los recursos hídricos. Con el objeto de generar instrumentos de planeación que permitan un mayor acercamiento entre el sector privado y los usuarios comprometidos con la gestión y el aprovechamiento del agua; en los próximos años el manejo de los recursos hídricos, seguramente será cada día más complejo [4]. La clave para conocer la disponibilidad de agua en cualquier región del mundo son los balances hídricos.

Las técnicas de minería de datos se clasifican en supervisadas o predictivas y no supervisadas o descriptivas. Las técnicas predictivas se basan en análisis de patrones secuenciales, análisis de similitud en series temporales y en predicciones. El objetivo de una serie de tiempo es evaluar el comportamiento de un conjunto de datos en el pasado y en el presente con la finalidad de hacer un pronóstico o predicción del futuro o tendencia de comportamiento que seguirán los datos [5]. Los algoritmos de predicción que se probaron en el presente estudio, estuvieron basados en los

métodos de aprendizaje supervisado del tipo de: a) Máquina de Soporte Vectorial, b) Regresión Lineal y c) Perceptrón Multicapas.

3. PARTE EXPERIMENTAL

Para verificar la viabilidad del uso de diversas técnicas de minería de datos, como una herramienta computacional para predecir la disponibilidad de agua en la cuenca baja del río San Lorenzo (Sinaloa) se aplicó la metodología propuesta por Pérez y Santín [6], para lo cual : a) Se delimitaron geográfica y conceptualmente las fronteras de la cuenca baja del río San Lorenzo, la superficie que se consideró como zona de producción agrícola en la cuenca; se definió el punto de referencia para el registro del caudal hídrico y el producto agrícola que se irriga (Maíz). b) Se realizó una revisión de las bases de datos puestas a disposición por CONAGUA, INEGI, CONABIO, IMTA y SAGARPA. Se prepararon los datos a utilizar, mediante un proceso de consolidación y limpieza según las necesidades de este estudio. Se utilizó el software de código abierto de la Universidad de Waikato (WEKA) y el PASW Statistics 18.0 como software especializado para explorar los datos y buscar inconsistencias. Se generaron los repositorios con los datos respecto al Volumen de Escurrimiento de agua en la estación hidrométrica y la Precipitación (lluvias). c) Se filtraron los atributos elegibles, se detectaron errores de integración de la base de datos, se analizaron los atributos e identificaron las herramientas de software o algoritmos de modelado. d) En este caso, fue necesario seleccionar un número efectivo de atributos, eliminar redundancias en los datos y se filtraron aquellos que eran considerados relevantes para el proceso de la minería de datos. La transformación de datos consistió básicamente en dejar los datos en el formato de entrada de la aplicación específica que se usó para realizar la minería y modelización (WEKA, PASW o EXCEL). e) Se generó un modelo propio para el cálculo de la disponibilidad de agua, basado en la Norma Oficial Mexicana NOM-011-CNA-2000 [7]. El modelo utilizado fue: Disponibilidad = $(Escurrimiento \times 0.7) + (P \times At)$, donde: $Escurrimiento$ = Volumen de Escurrimiento de Agua en la estación hidrométrica, P = Precipitación mensual y At = Área total cultivable destinada al maíz (objeto de este estudio). f) Las series temporales se realizaron empleando el módulo de predicciones del software WEKA y PASW. Primeramente se generó la serie temporal para la disponibilidad de agua en el río San Lorenzo durante un periodo de 18 años comprendidos entre 1994 y 2011. Esta tarea se realizó usando las técnicas: Máquina de Soporte Vectorial (SVM), Regresión Lineal (LR) y Perceptrón Multicapas (MLP). Posteriormente se calculó la predicción de la Disponibilidad de agua a futuro en el río San Lorenzo para el periodo 2009-2011.

Como principales resultados de la aplicación de las técnicas de minería de datos, se encontró que el escurrimiento del río San Lorenzo mostró dos comportamientos en periodos de tiempo distintos debido a la construcción y operación de la presa “José López Portillo” sobre el cauce del río (1976-1991). Además, las lluvias y evaporaciones registradas en los repositorios tienen un comportamiento estacional característico, reflejando los fenómenos meteorológicos esporádicos como huracanes o tormentas que ocurren en la región (figura 1).

Los resultados de las predicciones sobre las series temporales de la Disponibilidad de agua en el río San Lorenzo en el periodo 1994-2011, usando las técnicas predictivas SVM, LR y MLP se pueden observar en la figura 2. Cabe mencionar que cuando se generaban los pronósticos más allá de tres años, los resultados de las proyecciones eran más divergentes respecto a los datos observados.

En la Tabla 1 se presenta el análisis de significatividad de las predicciones obtenidas con WEKA, para cada uno de los atributos mencionados respecto a la disponibilidad de agua mediante cada una de las tres técnicas de predicción que ofrece dicho software.

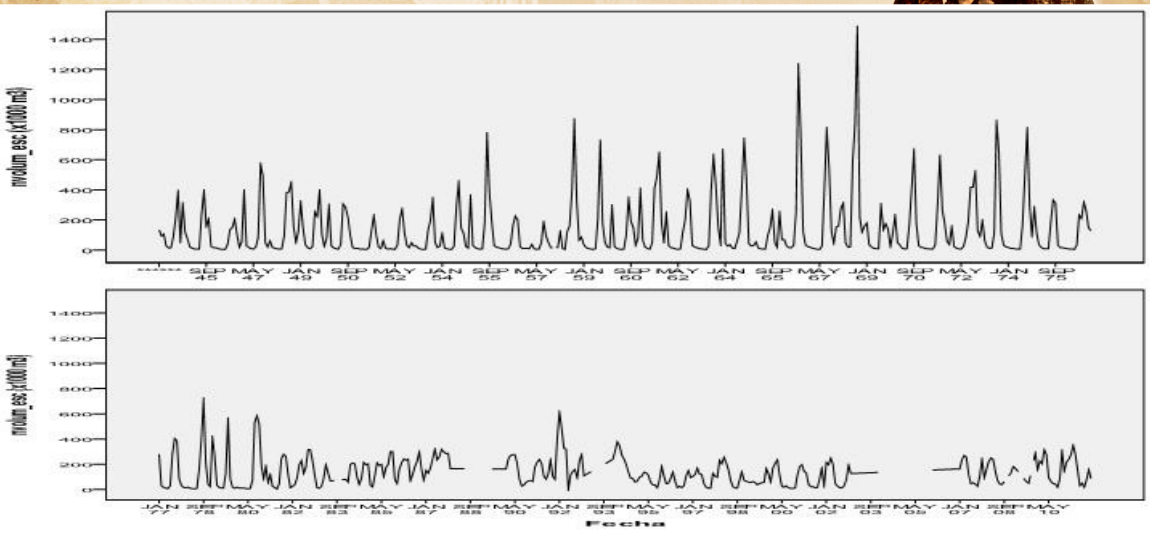

Figura 1. Comportamiento del volumen mensual de escurrimiento del río San Lorenzo en los periodos 1944-1976 y 1977-2011.© Ana-Laura Herrera-Prado.

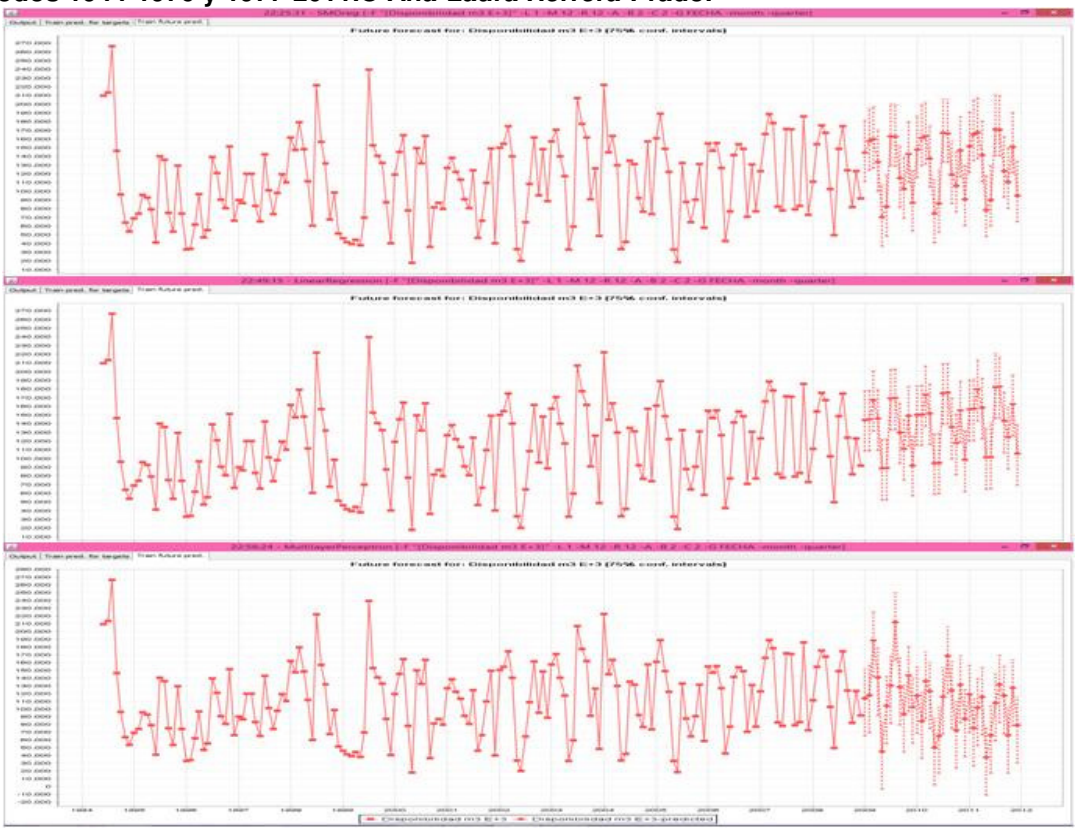

Figura 2. Predicción de la Disponibilidad de agua a futuro en el río San Lorenzo para el periodo 2009-2011, usando las técnicas a) SVM, b) LR y c) MLP. Datos observados en línea continua y predicción a futuro en línea punteada con intervalos de confianza al 75%. © Ana-Laura Herrera-Prado.

Las técnicas de predicción son comparadas por el coeficiente de correlación que se obtuvo al correlacionar las predicciones hechas por cada modelo, vs. los valores observados contenidos en el repositorio. Al aplicar cada técnica y deduciendo que todas reflejan una correlación significativa ($p<0.05$), se concluye que las predicciones que proyecta a futuro el software WEKA pueden ser consideradas como significativas para la Disponibilidad del agua.

Tabla 1. ANÁLISIS DE SIGNIFICATIVIDAD DE LAS PREDICCIONES REALIZADAS CON WEKA. © Ana-Laura Herrera-Prado.

ATRIBUTO	TÉCNICA DE PREDICCIÓN	COEFICIENTE DE CORRELACIÓN	CORRELACIÓN SIGNIFICATIVA α = 0.05
Disponibilidadm3+E3	SVM	0.4865	si
	LR	0.4922	si
	MLP	0.3447	si

4. CONCLUSIONES

De acuerdo con los resultados obtenidos en este estudio, el pre-procesamiento de las bases de datos seleccionadas permite obtener repositorios con los atributos útiles para el cálculo de la disponibilidad de agua. La transformación de los datos da lugar a un solo repositorio con atributos integrados de diversos dataset, además de atributos adicionales producto de diversos cálculos y transformaciones de la información original. Dicha transformación se realiza con el propósito de adecuarlos a los programas y algoritmos que se aplican para la predicción de las series de tiempo. Por lo expuesto en este documento se concluye que es viable el uso de técnicas de predicción para aplicarse en la generación de pronósticos de disponibilidad de agua hasta por 3 años, con fines preventivos y predictivos en la cuenca baja del río San Lorenzo (Sinaloa). Donde, todas las técnicas de predicción tienen una correlación significativa con los datos observados y el coeficiente de correlación más alto se obtiene con la técnica de Regresión Lineal (LR).

BIBLIOGRAFÍA

[1] J. García, “RemoteMining: Aplicando minería de datos a teledetección sobre LIDAR”. Ph.D..Tesis, Universidad de Sevilla, 2008, pp. 77.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases”.AI MAGAZINE, American Association for Artificial Intelligence.1996. pp. 37-54.

[3] C.C. Presser. “*Data mining*”, El Cid Editor, e-libro Corp. Argentina, 2009. pp. 11.

[4] J. Aparicio, J. Lafragua, A. Gutiérrez, R. Mejía y E. Aguilar, “Evaluación de los recursos hídricos. Elaboración del balance hídrico integrado por cuencas hidrográficas”. UNESCO. Uruguay, 2006.

[5] J. Rodríguez, A. Pierdant y C. Rodríguez, “Estadística aplicada II. Estadística en la administración para la toma de decisiones”, Grupo editorial patria, S.A. de C.V. México, 2010, pp. 370.

[6] C. Pérez and D. Santín, “Data Mining, Soluciones con Enterprise Miner”. Alfaomega. México. 2006. pp. 546.

[7] NOM-011-CNA-2000. NORMA Oficial Mexicana NOM-011-CNA-2000, “Conservación del recurso agua-Que establece las especificaciones y el método para determinar la disponibilidad media anual de las aguas nacionales”. CNA, SEMARNAT. En: DOF- 17 de abril de 2002.