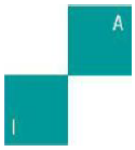
Inteligencia Artificial 21(62), (2018), 13-24
doi: 10.4114/intartif.vol21iss62pp13-24



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Rules for predicting compliance with the quality of wastewater in a treatment plant applying data mining

Reglas para predecir el cumplimiento de la calidad del agua residual en una planta tratadora con minería de datos

Facundo Cortés-Martínez, Alejandro Treviño-Cansino, María-Aracelia Alcorta-García*, Arturo Tadeo Espinoza-Fraire, José Armando Sáenz-Esqueda, Julio Gerardo Lozoya Vélez
facundo_cm@yahoo.com.mx, atrevino@ujed.mx, maaracelia@gmail.com, tadeo1519@gmail.com, jase1588@gmail.com, gerardo_lovelez@ujed.mx

Abstract A problem faced by the water operators is the compliance with the regulations on the quality of the treated wastewater. The most important thing is to implement strategies that favor compliance with the regulations. Data mining is a tool that allows the prediction of the water quality in the effluent of water treatment systems. In this research job, a criterion for nominal variables and data preprocessing is proposed. Subsequently, the data mining system (classification) was applied to define the prediction of water quality. The classificatory methodologies applied were: OneR, Decision Table, J48, single level decision tree, PART y LMT. Results show that, the best algorithm was the *Decision Table* with 85.45% of the instances correctly classified. The algorithm determined two rules for the regulation's achievement. It is important to mention that currently there are data mining procedures to predict water quality in the effluent of a treatment system, although, these procedures use strictly numeric variables; while in the current research, nominal variables were considered. Finally, results are discussed and industrial processes that generate organic waste and other pollutants are indicated.

Resumen Un problema que enfrentan los organismos operadores de agua, es el cumplimiento de la normatividad en la calidad del agua residual tratada. Por lo que es recomendable implementar estrategias que favorezcan el cumplimiento de las regulaciones. La minería de datos es una herramienta que permite predecir la calidad del agua en el efluente de los sistemas de tratamiento. En el presente estudio se propone un criterio para el pre procesado de datos donde se consideraron variables nominales. Posteriormente se aplicó el sistema de minería de datos (clasificación) para definir la predicción de la calidad del agua. Se utilizaron los siguientes clasificadores: *OneR*; *DecisiónTable*, J48, árbol de decisión de un solo nivel, *PART* y *LMT*. Los resultados muestran que el mejor algoritmo fue el *DecisiónTable* con el 85.45 % de instancias correctamente clasificadas. El algoritmo determinó dos reglas para el cumplimiento de la normatividad. Es importante indicar que a la fecha existen procedimientos con minería de datos para predecir la calidad del efluente de un sistema de tratamiento, pero utilizan estrictamente variables numéricas; mientras que en el presente trabajo se utilizaron variables nominales, finalmente se discuten los resultados y se indican los procesos industriales que generan materia orgánica y otros contaminantes.

Keywords: Biochemical oxygen demand, decision table, nominal variables, classification, data mining.
Palabras clave: Demanda bioquímica de oxígeno, tabla de decisión, variables nominales, clasificación, minería de datos.

ISSN: ISSN 1137-3601 (print), 1988-3064 (on-line)

1. Introducción

Un reto para las autoridades tanto municipales, estatales y federales, es la observancia de la calidad del agua residual tratada. Por lo que es necesario desarrollar esquemas de análisis para el cumplimiento de los límites máximos permitidos en el efluente de las plantas de tratamiento. De acuerdo con [23], [24] y [26] la minería de datos es una herramienta que procesa grandes cantidades de información, y además, permite monitorear el estado del sistema de tratamiento, la clasificación de las aguas residuales, el control de los procesos así como la predicción de la calidad del agua tratada. En [35] se define la minería de datos como las actividades necesarias para extraer información sintetizada y desconocida de manera de generar nuevos conocimientos. Es decir se refiere a la extracción de datos no triviales que se encuentran implícitos en la información proporcionada. De acuerdo con [1] el proceso se describe en términos generales en cinco partes: a) selección de la información; b) transformación de los datos: se refiere a organizar la información como se necesite; c) minería de datos, es decir una vez modificada la información se aplica el sistema; d) interpretación de los resultados y e) la incorporación del nuevo conocimiento.

1.1 Revisión de la Literatura

Algunos estudios similares para la predicción de la calidad del agua: en [38] se inspeccionaron aguas residuales ácidas con minería de datos, luego en [48] por medio de modelos de simulación se reportó la predicción del caudal en el efluente del sistema de tratamiento. Tiempo después en [40] los autores llevaron a cabo estudios con el fin de identificar patrones ocultos con inducción de reglas, se utilizaron 18 variables; es decir, diferentes parámetros en el influente del sistema de tratamiento y las mediciones se aplicaron en la entrada y salida de cada proceso. Algunos parámetros considerados fueron la Demanda Bioquímica de Oxígeno (DBO); Demanda Química de Oxígeno (DQO); Gasto y Sólidos Suspendidos. Enseguida algunos estudios publicados acerca del uso de redes neuronales para predecir la calidad del agua residual, tanto en el influente como en el efluente de sistemas de tratamiento: [42], [46], [16], [37], [34], [36] y [2]. Un trabajo acerca de la identificación de patrones en los datos puede consultarse en [31] los autores publicaron un estudio para identificar patrones en el flujo: predicción de gasto en el influente del sistema de tratamiento. Luego [8] realizó estudios para definir el grado de contaminación en un río con minería de datos; enseguida [29] reportaron un estudio para predecir la calidad del agua en el influente del sistema de tratamiento, con el fin de ahorrar electricidad donde utilizaron series temporales para la predicción. Luego [44] realizaron una investigación para predecir los sólidos suspendidos totales con minería de datos y series temporales. Por ese tiempo [28] realizó un estudio de clasificación con muestras de agua contaminada, donde se incluyeron 30 industrias. Los resultados demostraron que las industrias presentaron un patrón de descarga similar. El análisis fue realizado con variables numéricas y los parámetros fueron: Demanda Bioquímica de Oxígeno; potencial de hidrógeno (pH); Demanda Química de Oxígeno, Sólidos en Suspensión, grasas y aceites. Enseguida [3] mencionaron que la técnica de agrupación de datos es muy útil en la aplicación de minería de datos y que esta técnica es muy usada en diferentes disciplinas: ingeniería, economía, geología, electrónica, estadística y psicología. Luego [9] monitoreó la eficiencia de depuración en una planta de tratamiento de lodos activados donde se aplicó un análisis estadístico multivalente y los parámetros analizados fueron: potencial de hidrógeno, conductividad, gasto, sólidos en suspensión DBO, DQO, nitrógeno y fósforo.

Todos los trabajos mencionados han contribuido satisfactoriamente con modelos de predicción. No obstante, en ningún estudio se han aplicado variables nominales para la predicción de la calidad del agua en sistemas de tratamiento. La aportación del presente estudio con respecto a los autores indicados es la manera en la que fueron pre procesados los datos con el uso de variables nominales. Se incluyó como base lo indicado en la normatividad y se agregó una variable nominal con dos posibles resultados: “NO_CUMPLE” y “SÍ_CUMPLE”.

La DBO se refiere al oxígeno disuelto necesario para que los microorganismos, presentes en el agua residual, puedan procesar la materia orgánica. Lo anterior en un lapso de tiempo de 5 días a 20 °C. [11], [12], [33]. El método de análisis en laboratorio de los Sólidos Sedimentables mide la cantidad aproximada de lodos que serán generados en un proceso de sedimentación primaria. La prueba se realiza con un *cono (Imhoff)* en 60 minutos [11]. Solidos Suspendidos: se refiere a la medición de partículas sólidas y suspendidas que contiene el agua residual. Este parámetro bloquea los rayos solares en los sistemas de agua y en ocasiones los sólidos suspendidos contienen metales [11].

Se utilizó la herramienta (*Waikato Environment for Knowledge Analysis WEKA* por sus siglas en inglés). Los algoritmos utilizados llevan a cabo el aprendizaje computacional, estadísticas, asociaciones, anomalías, eventos en los datos, visualización, inteligencia artificial y reconocimiento de patrones [1], [22]. Los procesos desarrollados por los sistemas de minería de datos tienen como propósito ajustar modelos o definir patrones.

El objetivo de presente artículo fue definir un criterio de pre procesado de datos incluyendo una variable nominal con dos posibles resultados, lo anterior considerando algunos parámetros tanto en el influente como en el efluente de un sistema de tratamiento de aguas residuales de lodos activados. Lo indicado con el propósito de predecir la calidad del agua residual en la salida de la planta aplicando la herramienta *WEKA.* Después de obtener los resultados, se realizarán recomendaciones para favorecer el control de la contaminación en los procesos industriales.

1.2 Procesos Industriales que Generan Materia Orgánica

Con el propósito de regular los límites de concentración en la entrada del sistema de tratamiento, y favorecer el cumplimiento de las condiciones que se indican en la normatividad, la CNA e IMTA en [11]; además en [19], [20], [21] y [43] sugieren se controle a las empresas que generan contaminantes. En la Tabla 1 se indican algunos de los procesos industriales que deben verificarse.

Tabla 1. Procesos que incluyen alto grado de contaminantes. Fuente: [11].

	Descripción	CE	pH	GyA	SS	SST	DBO	DQO	SAAM
3111	Industria de la carne			X		X	X	X	
3112	Fabricantes de lácteos			X		X	X	X	
3113	Procesamiento de alimentos enlatados	o		X	X		X	X	X
3114	Beneficio y molienda de cereales y otros productos agrícolas			X	X	X			
3115	Pastelerias			X	X			X	
3116	Molienda de maíz	o	o		X	X	X	X	
3117	Fabricantes de aceites y grasas			o				X	
3118	Industria de azucar	o	o			X	X	X	o
3119	Fabricación de cacao, chocolate y productos de confitería			X	X	X	X	X	
3121	Fabricación de otros productos alimenticios para el consumo humano			X	X	X	X	X	

El control de los contaminantes de las industrias y comercios tiene como propósito la protección del sistema de drenaje y alcantarillado sanitario, equipos de bombeo y la correcta operación de la planta de tratamiento municipal [11]. Algunos sistemas de tratamiento que pueden implementar las empresas previamente al vertido del agua residual de su proceso, son las siguientes: trampas de grasas, aceites y sólidos, balances de masa de procesos industriales, plantas de tratamiento y precipitación química entre otros.

El artículo está organizado de la siguiente manera: en el apartado 2 se indican los materiales y métodos, en 3 en forma detallada el criterio para el pre procesado de los datos; en 4 en filtrado de los datos, en 5 se muestran los resultados de la aplicación de los diferentes clasificadores; y por último en 6, las conclusiones.

2. Materiales y Métodos

La información de las mediciones del agua residual tanto en el influente como del efluente del sistema de tratamiento, fueron tomados de una Planta de Tratamiento de Aguas Residuales (PTAR) de lodos activados localizada en Manresa cerca de Barcelona. El sistema recibe un gasto en el influente de 30,000 m^3/día de una población de 75,000 habitantes como se establece en [30]. Esta base datos ha sido utilizada para otros estudios: [4], [6], [14], [17]. Debido al volumen de información no se incluye en el presente documento; sin embargo pueden verificarse en la cita ya señalada.

3. Pre procesado de Datos

En esta etapa se llevó a cabo un conjunto de operaciones con la finalidad de preparar los datos de análisis, con el objetivo de adaptarlos para aplicar la técnica de minería de datos que mejor se adapte al problema. Para alimentar el modelo se consideraron los siguientes parámetros: gasto en el efluente (Qe); ZNe, potencial de hidrógeno en el efluente (PHe): DBOe, DQOe, sólidos suspendidos en el efluente (SSe); sólidos suspendidos volátiles en el efluente (SSVe); sólidos sedimentables en el efluente (SSEDe) y conductividad.

Adicionalmente, como ya se indicó, se agregó un atributo nominal si cumple o no con la norma NOM-SEMARNAT-001-1996 [18]. Enseguida se muestra el pre procesado de datos realizado.

```
@relation 'DATOS_ULTIMO_NOM_001_Modificado-
weka.filters.unsupervised.attribute.StringToNominal-Rlast-
weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,15-
weka.filters.unsupervised.attribute.StringToNominal-Rlast-
weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,8,9,10,11,12,13,15,16,14-
weka.filters.unsupervised.attribute.StringToNominal-Rlast-
weka.filters.unsupervised.attribute.Reorder-
R1,2,3,4,5,6,7,8,9,10,11,12,14,15,16,13-
weka.filters.unsupervised.attribute.StringToNominal-Rlast-
weka.filters.unsupervised.attribute.Remove-R10-12,14-16'
@attribute Q-E numeric
@attribute ZN-E numeric
@attribute PH-E numeric
@attribute DBO-E numeric
@attribute DQO-E numeric
@attribute SS-E numeric
@attribute SSV-E numeric
@attribute SED-E numeric
@attribute COND-E numeric
@attribute CUMPLE_CON_LA_NORMA_NOM-SEMARNAT-001 {NO_CUMPLE,SI_CUMPLE}
```

4. Filtrado de Datos

Se cuenta con 526 instancias y 10 atributos, enseguida se realizó una limpieza de datos: a) existen valores perdidos en los atributos por lo que fueron sustituidos con la aplicación: (*WEKA-Filter: Unsupervised-Attribute-Replace Missing Value*, filtrar WEKA: atributo no supervisado-reemplazar valor perdido*)* La inclusión citada de los valores perdidos se realizó considerando la media para los valores continuos y la moda para los valores discretos; b) luego se identificaron los valores atípicos y extremos con (*WEKA-Filter*: *Unsupervised-Attribute-IntercuartileRange*, filtrar WEKA: atributo no supervisado-rango intercuartil) enseguida se eliminaron usando (*WEKA-Filter: Unsupervised-Instance-RemoveWithValues*, filtrar WEKA: instancia no supervisada-remover con valores). Finalmente se obtuvieron 493 instancias [15]. Para determinar el atributo nominal se consideró el diagrama mostrado en la Figura 1, el cual indica el límite máximo permitido por la norma para Sólidos Sedimentables, Sólidos Suspendidos Totales y DBO en el efluente.

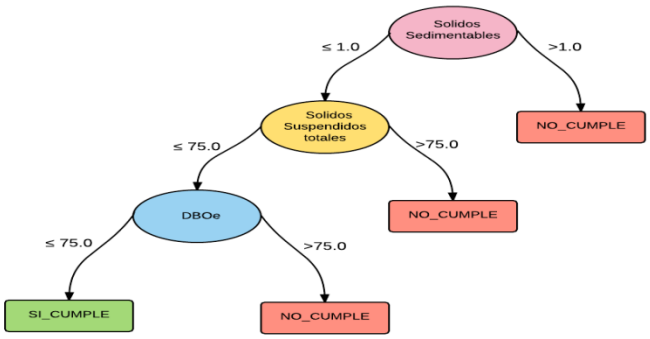


Figura 1. Pre procesado de datos para variable nominal “SÍ_CUMPLE” y “NO_CUMPLE”.

Al sistema *WEKA* se alimentaron los datos de la entrada de la planta y la variable nominal, la cual representa los datos en la salida, lo anterior para que el sistema encuentre la relación que existe entre estos datos y genere el posible resultado conforme a la norma; ya que no tiene caso alimentar el sistema con los datos de salida dado que la relación sería obvia para predecir la variable nominal: se obtendrían solamente los parámetros ya indicados por la norma y ninguna relación entre la entrada y la variable nominal.

De acuerdo con [27], [32], [33], [5] y [47] la Demanda Química de Oxígeno, la Demanda Bioquímica de Oxígeno, el potencial de Hidrógeno y Sólidos Suspendidos Totales, son indicadores muy usados para definir la calidad del agua residual. En la Tabla 2 se muestran los límites máximos permitidos para la descarga de agua residual tratada al cuerpo receptor, como se observa los valores son similares a los indicados en la Figura 1.

Tabla 2. Límites máximos permisibles para contaminantes básicos.

Parámetros	Aguas costeras Recreación (B). Promedio mensual
Sólidos sedimentables (mL/L)	1.0
Sólidos Suspendidos Totales (mg/L)	75.00
Demanda Bioquímica de Oxígeno (mg/L)	75.00

Fuente: [18].

4.1. **Algoritmos de Clasificación.**

Se ingresaron los datos pre procesados al *WEKA* y posteriormente se dividieron las instancias para entrenar el modelo (66%) y para medir su precisión (33%) [41].

Una vez que se delimitó el porcentaje para entrenar el modelo, se utilizaron seis clasificadores incluidos en el sistema *WEKA*: a) árboles de decisión de un nivel; b) clasificador IR (*OneR*); c) tablas de decisión; d) J48; e) *Projective Adaptiva Resonance Theory (PART)* teoría de la resonancia adaptativa proyectiva por sus siglas en inglés y f) *Logistic Model Trees (LMT)* modelo logístico de árboles por sus siglas en inglés. Finalmente se realizó un comparativo para determinar el mejor clasificador. Como ya se indicó el número de instancias, una vez aplicado el tratamiento de datos, fue de 493. Según [45] la instancia se refiere a la muestra que se realiza cada día en donde se miden todos los parámetros, por lo que las instancias corresponden al número de muestras que se tomaron del agua residual en un periodo de tiempo (por día). En la Tabla 3, se muestran las variables que alimentarán al modelo de pronóstico, de igual forma se observan los diferentes clasificadores que serán aplicados, para que posteriormente el sistema WEKA lleve a cabo la predicción.

Tabla 3. Clasificadores y parámetros para la predicción del modelo propuesto

Clasificadores y parámetros					
J48	DecisionStump	OneR	DecisionTable	LMT	PART
Qe	Qe	Qe	Qe	Qe	Qe
ZNe	ZNe	ZNe	ZNe	ZNe	ZNe
PHe	PHe	PHe	PHe	PHe	PHe
DBOe	DBOe	DBOe	DBOe	DBOe	DBOe
DQOe	DQOe	DQOe	DQOe	DQOe	DQOe
SSe	SSe	SSe	SSe	SSe	SSe
SSVe	SSVe	SSVe	SSVe	SSVe	SSVe
SSEDe	SSEDe	SSEDe	SSEDe	SSEDe	SSEDe
Conductividad	Conductividad	Conductividad	Conductividad	Conductividad	Conductividad
Variable nominal	Variable nominal	Variable nominal	Variable nominal	Variable nominal	Variable nominal
Sí cumple	Sí cumple	Sí cumple	Sí cumple	Sí cumple	Sí cumple
No cumple	No cumple	No cumple	No cumple	No cumple	No cumple

(*OneR*): es un algoritmo sencillo, no obstante desarrolla en parte el criterio utilizado por los árboles de decisión. Una característica de este algoritmo es que define el atributo que explica de mejor forma la clase de salida (ver [27]).

Tabla de decisión: se refiere a una matriz de renglones y columnas en donde se mencionan condiciones y acciones. Las reglas de decisión incluidas constituyen el procedimiento que debe seguirse cuando existe un número determinado de condiciones.

Árbol de decisión J48: es una versión afinada del modelo que incluye el *OneR*. El clasificador J48 utiliza el algoritmo C4.5. Éste es uno de los algoritmos más utilizados en los trabajos publicados en minería de datos.

Árbol de decisión de un nivel (*DecisiónStump*) algunas características de esta versión: a) es posible analizar datos categóricos; b) genera árboles de decisión binarios, es decir de un solo nivel ya sea con datos categóricos o numéricos y c) puede ser utilizado para mejorar los métodos [10].
Projective Adaptiva Resonance Theory (PART) teoría de la resonancia adaptativa proyectiva por sus siglas en inglés: no considera el trabajo de optimización global que se incluye en el algoritmo C4.5 y además, define una lista de decisión, pero sin tomar en cuenta restricciones (establece el procedimiento de divide y vencerás). Una característica del *PART* es que determina un árbol de decisión en forma parcial con el fin de determinar una regla. Otro punto importante es que para podar el árbol es necesario que se conozcan todas las consideraciones [10].
Logistic Model Trees (LMT) modelo logístico de árboles por sus siglas en inglés: genera un análisis detallado de los datos, el algoritmo utilizado, construye un árbol de decisión e incluye funciones de regresión [39].

4.2. Precisión por Clase de un Clasificador

De acuerdo con [7], se indican los parámetros de exactitud que incluye el *WEKA*:
Verdadero positivo (*TP* por sus siglas en inglés) se refiere a la magnitud de ejercicios o ejemplos que fueron clasificados como una clase x considerando todos los ejercicios que verdaderamente incluyen la clase x.
Falso positivo (*FP* por sus siglas en inglés) es una parte de ejemplos clasificados de la clase x pero que realmente son de otra clase.
La precisión: también es una parte de ejemplos que en verdad cuentan con la clase x considerando todas las clasificaciones de la clase x.
Medida: se refiere a una medición combinada entre la precisión y la cobertura.

5. Resultados y Discusión

5.1 Prueba con Diferentes Clasificadores

En la Tabla 4 se muestran los resultados obtenidos con los distintos clasificadores aplicados a las mediciones de la calidad del agua en el efluente de la planta de tratamiento.

Tabla 4. Resultados de la ejecución de clasificadores (33% instancias para la predicción)

Algoritmos	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)
J48	83.6364	16.3636
DecisionStump	82.4242	17.5758
OneR	82.4242	17.5758
DecisionTable	**85.4545**	**14.5455**
LMT	83.0303	16.9697
PART	83.6364	16.3636

De acuerdo a la Tabla 4, el algoritmo que mejor clasificó fue la tabla de decisión: 85.45 % de probabilidad de aciertos. Se observa que los resultados de los algoritmos (*DecisionStump y OneR)* son similares; de igual forma el J48 y *PART*. Según [45] es motivo para desechar estos resultados. El algoritmo *LMT* resultó con el segundo lugar en la clasificación. Se analizarán solamente los resultados del algoritmo (*DecisionTable)*.

5.2 Precisión del Clasificador (*DecisionTable)*

La Tabla 5 muestra la precisión del clasificador (*DecisionTable)* por clase, se tiene entonces una media ponderada de 85.50 % para las instancias que resultaron correctamente clasificadas (verdaderos positivos) con

respecto a lo real; para los falsos positivos se obtuvo una media ponderada de 65.50 % de instancias incorrectamente catalogadas y una precisión del clasificador de 85.50 %.

Tabla 5. Precisión detallada por clase

	TP Rate	FPRate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.993	0.793	0.854	0.993	0.918	0.377	0.763	0.911	NO_CUMPLE
	0.207	0.007	0.857	0.207	0.333	0.377	0.740	0.394	SÍ_CUMPLE
Weighted Avg.	0.855	0.655	0.855	0.855	0.816	0.377	0.759	0.821	

La matriz de confusión (Tabla 6) muestra las instancias clasificadas por clase, así como cuantas son clasificadas correcta e incorrectamente. Se han clasificado 158 instancias como “NO_CUMPLE”, de las cuales 135 fueron correctamente clasificadas y 23 incorrectas. Así mismo se muestra que fueron clasificadas 7 instancias como “SI_CUMPLE”, de las cuales solamente 6 instancias son correctas: la otra en realidad son “NO_CUMPLE”.

Tabla 6. Matriz de confusión con el clasificador Tabla de decisión

Predicción			
No cumple		Sí cumple	
Correcta	135	Incorrecta	1
Incorrecta	23	Correcta	6
Suma	158	Suma	7

5.3 Reglas del clasificador (*DecisionTable)*

La Tabla 7 indica las reglas definidas con el clasificador (*DecisionTable)*, la cual muestra el rango para cada contaminante en la entrada de la planta (ZN-E, DBO-E y SS-E) y el resultado obtenido para la variable nominal (Sí Cumple o No Cumple).

Tabla 7. Reglas para los parámetros de entrada en la planta de tratamiento de aguas residuales. Clasificador *DecisionTable*.

Regla	ZN-E		DBO-E		SS-E		Resultado
	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo	
1	-INF	0,475	153,5	INF	149	INF	NO_CUMPLE
2	0,475	1,715	153.5	INF	149	INF	NO_CUMPLE
3	1,715	INF	153.5	INF	149	INF	NO_CUMPLE
4	-INF	0,475	-INF	153.5	149	INF	SÍ_CUMPLE
5	0,475	1,715	-INF	153.5	149	INF	NO_CUMPLE
6	1,715	INF	-INF	153.5	149	INF	NO_CUMPLE
7	-INF	0,475	153.5	INF	-INF	149	NO_CUMPLE
8	0,475	1,715	153.5	INF	-INF	149	NO_CUMPLE
9	1,75	INF	153.5	INF	-INF	149	NO_CUMPLE
10	-INF	0,475	-INF	153.5	-INF	149	SÍ_CUMPLE
11	1,715	INF	-INF	153.5	-INF	149	NO_CUMPLE
12	0,475	1,715	-INF	153.5	-INF	149	SÍ_CUMPLE

Según *WEKA* el clasificador (*DecisionTable*) analiza el subconjunto de atributos con el algoritmo (*Best First)*, y selecciona los tres atributos mostrados en la Tabla 7. Se descarta el resto de los atributos que se alimentaron al inicio, de esta manera se depuran los datos en la entrada para simplificar las reglas resultantes. La Tabla 8 muestra las tres reglas con el resultado Sí Cumple. El resto de los casos resultaron No Cumple.

Tabla 8. Reglas con resultado Sí Cumple

Regla 4	Regla 10	Regla 12
Si:	Si:	Si:
ZN-E≤0,475	ZN-E≤0,475	0,475≤ZN-E≤1,715
DBO-E≤153,5	DBO-E≤153,5	DBO-E≥153,5
SS-E≥149	SS-E≤149	SS-E≤149
Entonces: Sí Cumple	Entonces: Sí Cumple	Entonces: Sí Cumple

En la Tabla 8 se observa que la regla 4 y 10 incluyen las mismas condiciones para ZN-E y DBO-E y además, la condición para SS-E es inversa entre ambas reglas, por lo tanto, se puede eliminar esta última condición. La Tabla 9 muestra las reglas resultantes.

Tabla 9. Reglas simplificadas con resultado Sí Cumple

Regla 4 y 10	Regla 12
Si:	Si:
ZN-E≤0,475	0,475≤ZN-E≤1,715
DBO-E≤153,5	DBO-E≥153,5
	SS-E≤149
Entonces: Sí Cumple	Entonces: Sí Cumple

Es importante aclarar que lo indicado en la Tabla 2 es el criterio para el pre procesado de los datos, y las reglas presentadas en la Tabla 9 tienen la finalidad de determinar si se cumple con los criterios indicados. Según los resultados de la Tabla 5 al aplicar la predicción se cumple con un 85.5% de precisión en el pronóstico de la variable nominal. Es importante indicar que las reglas mostradas en la Tabla 9 no corresponden a ecuaciones de predicción numérica, sino nominal, por lo que no es posible comparar la norma de la Tabla 2 con las reglas para la variable nominal.

De acuerdo con [24] cuando las reglas que conducen a una variable (sí o no) las condiciones no entran en conflicto, por lo que no existe ambigüedad en la interpretación de las reglas.

En la revisión de literatura realizada no se encontraron antecedentes de estudios similares al presente; es decir, donde fueran consideradas el uso de variables nominales. Por tal motivo en la discusión se consideraron trabajos con diferentes metodologías pero que también predicen la calidad del agua residual.

[29], [28] y [44] utilizaron para la predicción análisis estadísticos, series de tiempo y minería de datos; mientras que en el presente estudio se aplicó exclusivamente la clasificación de minería de datos. Una diferencia importante respecto a la investigación de [29], [34], [37] y [44] fue la determinación de reglas para favorecer el cumplimiento de los límites máximos permitidos, para la descarga de agua residual a los cuerpos receptores. Los criterios y resultados son diferentes, no obstante lo anterior, los estudios indicados son capaces de predecir la calidad del agua residual en el efluente de los sistemas de tratamiento. [25] publicó un estudio para predecir la demanda bioquímica de oxígeno en el efluente de una planta de tratamiento de aguas residuales, utilizando redes neuronales con 5 parámetros: DQO, SS, Conductividad Eléctrica (CE), Temperatura y Potencial de Hidrógeno. El estudio consistió en una comparación de resultados: redes neuronales y regresión lineal múltiple. Se obtuvieron mejores resultados con las redes neuronales; mientras que en el estudio propuesto sólo se aplicó la minería de datos considerando seis algoritmos, donde se obtuvieron tres importantes condiciones para la predicción de la calidad del agua residual. Es importante

destacar la simplicidad del modelo propuesto respecto a los indicados en el apartado de revisión de la literatura.

6. Conclusiones

Con el criterio del pre procesado de datos y el uso de variables nominales fue posible definir reglas para la predicción de la calidad en el efluente del sistema de tratamiento. Además, es posible utilizar el modelo para identificar los principales contaminantes que intervienen en la calidad final del agua, y de esta forma localizar a las empresas contaminantes con el fin de realizar un adecuado control de descargas. Es decir contar con elementos para la toma de decisiones: si es prudente activar o no el programa de control de descargas de aguas residuales de procesos industriales, comerciales y de servicios.

Los resultados del presente trabajo pueden tomarse como fundamento para la toma de decisiones. Por ejemplo, el citado programa de control ambiental necesita de grandes recursos económicos para la verificación del total de los procesos industriales que utilizan agua en su proceso. Las reglas obtenidas identifican únicamente los contaminantes críticos que deberán supervisarse, lo anterior restringe la utilización de recursos económicos en la verificación de las descargas de aguas residuales.

Es importante indicar que la aportación del presente trabajo fue el uso de variables nominales para el pronóstico de la calidad del agua residual con una precisión del 85.5%.

La minería de datos es una herramienta que favorece la generación de reglas para predecir la calidad del agua residual en la salida del sistema de tratamiento, por lo cual puede aplicarse en trabajos futuros a sistemas de lagunas de oxidación y como una herramienta auxiliar para el diseño de plantas de tratamiento y toma de decisiones.

Los resultados del proceso dependerán de los límites máximos permitidos indicados en la Tabla 2. La predicción que arroja el modelo es útil cuando se cuenta con los datos de entrada y se puede saber de antemano si se cumplirá o no con la norma.

El criterio de análisis propuesto en el presente estudio puede usarse en cualquier planta de tratamiento de aguas residuales, siempre y cuando se cuente con la base de datos del sistema.

Reconocimientos

Se agradece al Programa para el Fortalecimiento de la Calidad Educativa (PFCE) 2017 por los recursos asignados para la realización del presente estudio.

Referencias

[1] Arévalo1, J. L. y García, R. P. Estado del arte en la utilización de técnicas avanzadas para la búsqueda de información no trivial a partir de datos en los sistemas de abastecimiento de agua potable, 2008. www.lenhs.ct.ufpb.br/html/downloads/serea/trabalhos/A15_15.pdf. (accessed December, 29, 2016).

[2] Atasoy, A.D., Babar, B., Sahinkaya, E. Artificial neural network prediction of the performance of upflow and downflow fluidized bed reactors treating acidic mine drainage water. Mine water and the environment. New York, Springer, vol.1, 222–228, 2013.

[3] Ay, M., and Kisi, O. Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. Journal of Hydrology, 511, 279-289, 2014.

[4] Belanche, L., Sànchez, M., Cortés, U., and Serra, P. A knowledge-based system for the diagnosis of wastewater treatment plants. In Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. Springer Berlin Heidelberg, 324-336, 1992.

[5] Bernard, O., Hadj-Sadok, Z., Dochain, D., Genovesi, A., and Steyer, J. P. Dynamical model development and parameter identification for an anaerobic wastewater treatment process. Biotechnology and Bioengineering, 75(4), 424–438, 2001.

[6] Béjar, J., Cortés, U., and Poch, M. Linneo, A classification methodology for illstructured domains. Facultat d'Informtica de Barcelona, Tech. Rep., 1993.

[7] Bouckaert, R. R., Frank, E., and Hal, M. WEKA manual for version 3-7-8, 2013.

[8] Cărbureanu, M. Pollution Level Analysis of a Wastewater Treatment Plant Emissary using Data Mining. Petroleum-Gas University of Ploieşti Bulletin Mathematics-Informatics-Physics Series, 62(1), 69-78, 2010.

[9] Carpe Cánovas, J. Monitorización de la eficiencia de depuración de una planta de tratamiento de agua residual urbana mediante el empleo de técnicas de análisis estadístico multivariante. Tesis of Master degree, 2015.

[10] Césari, M. Aprendizaje automático con WEKA, 2007.

[11] Comisión Nacional del Agua. Guía para el control de descargas a los sistemas de alcantarillado urbano o municipal. México2000a.

[12] Comisión Nacional del Agua. Manual de diseño de agua potable, alcantarillado y saneamiento. Manual de diseño de lagunas de estabilización. Instituto Mexicano de Tecnología del Agua, Jiutepec, México. 234, 2007b. http://www.conagua.gob.mx/conagua07/publicaciones/publicaciones/Libros/10DisenoDeLagunasDeEstabilizacion.pdf

[14] Cortés-Martínez, F., Treviño-Cansino, A., Sáenz-López, A., and Narayanasamy, R. Statistical analysis for the characterization of the wastewater in the influent of a treatment plant (Case of study). International Journal of Engineering and Technical Research, Vol. 4 (1). 103-110, 2016.

[15] Cortés-Martínez, F., Espinoza-Fraire, A.T., Sáenz-López. A. y Narayanasamy. Limpieza y descripción gráfica de datos con WEKA, *Memorias del Congreso Internacional de Investigación Academia Journals CICS*, Tuxpan 2017, Vol. 9, No. 4, 405-410, 2017.

[16] Chen, W.C., Chang, N.B., Shieh, W.K. Advanced hybrid fuzzy-neural controller for industrial wastewater treatment. Journal of Environmental Engineering 127 (11), 1048–1059, 2001.

[17] De Garcia Blanco, J. Avaluació de tècniques de classificació per a la gestió de bioprocessos: aplicació a un reactor de fangs activats, 1993.

[18] Diario Oficial de la Federación, (DOF). Norma Oficial Mexicana NOM-001-ECOL-1996, que establece los límites máximos permisibles de contaminantes en las descargas de aguas residuales a los sistemas en aguas y bienes nacionales, 1997. https://books.google.com.mx/books?id=N-5WAAAAMAAJ

[19] Enviromental Protection Agency. Guidance Manual for Preventing Interference at POTWs. USA, 1987.

[20] EPA 1991 Environmental Protection Agency Control de Descargas Irregulares Hacia las POTWs. USA.

[21] EPA 1994 Environmental Protection Agency, Guía, procedimientos y pautas recomendadas para establecer e implementar un programa de pretratamiento. USA.

[22] Grossman, R., Kasif, S., Moore, R., Rocke, D., Ullman, J., Data mining research: opportunities and challenges, A report of three NSF workshops on mining large, massive, and distributed data, September 18, 1998.

[23] Guclu, D. and Dursun, S. Artificial Neural network modelling of a large-scale wastewater treatment plant operation. Bioprocess Biosystems Engineering, vol 33, no. 9, 1051-1058, 2010.

[24] Hall, M., Witten, I., and Frank, E. , Data mining: Practical machine learning tools and techniques. Kaufmann, Burlington, 2011.

[25] Heddam, S., Lamda, H., and Filali, S. Predicting effluent biochemical oxygen demand in a wastewater treatment plant using generalized regression neural network based approach: a comparative study. Environmental Processes, 3(1), 153-165, 2016.

[26] Henze, M., Harremes, P., Jansen, J. L. C., and Arvin, E. Wastewater treatment: biological and chemical processes. New York: Springer, 2001.

[27] Jiménez, M. G., and Álvarez, A. Análisis de datos en WEKA–pruebas de selectividad, 2010. http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

[28] Kaur I. Cluster Analysis on various samples of water pollution. American International Journal of Research in Science, Technology, Engineering and Mathematics, 4(2), 149-153, 2013.

[29] Kusiak, A., Verma, A., and Wei, X. A data-mining approach to predict influent quality. Environmental monitoring and assessment, 185(3), 2197-2210, 2012.

[30] Lichman, M. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2013. https://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant

[31] Lindqvist, J., Wik, T., Lumley, D., Aijala, G. Influent Load Prediction Using low Order Adaptive Modeling. In: Proceedings of the 2nd IWA Conference on Instrumentation, Control and Automation, Busan, South Korea, 2005.

[32] Mamais, D., Jenkins, D., and Pitt, P. A rapid physical–chemical method for the determination of readily biodegradable soluble COD in municipal wastewater. Water Research, Biological Wastewater Treatment, IWA Publisher, 27(1), 195–197, 1993.

[33] Metcalf and Eddy, Inc.,Wastewater Engineering: Treatment, Disposal, Reuse. McGraw-Hill, New York, 1991.

[34] Mjalli, F. S., Al-Asheh, S., and Alfadala, H. E. Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. Journal of Environmental Management, 83(3), 329-338, 2007.

[35] Olaru, C.; Wehenkel, L. “Data Mining”. IEEE Computer Applications in Power, Volume 12, Number 3, 19-25, 1999.

[36] Pai, T. Y., Yang, P. Y., Wang, S. C., Lo, M. H., Chiang, C. F., Kuo, J. L., and Chang, Y. H. Predicting effluent from the wastewater treatment plant of industrial park based on fuzzy network and influent quality. Applied Mathematical Modelling, 35(8), 3674-3684, www.journals.elsevier.com/applied-mathematical-modelling. 2011.

[37] Patricia, K., Esquerrea, O., Seborgb, D.E., Moria, M., Bruns, R.E. Application of steady state and dynamic modeling for the prediction of the BOD of an aerated lagoon on a pulp and paper mill. Part II-nonlinear approaches. Chem. Eng. J. Nova Science Publishers, Inc, New York. 105 (5), 61–69, 2004.

[38] Qasim, S. R. Wastewater treatment plants: planning, design, and operation. New York, CRC Press, 1998.

[39] Romero Beltrán C.A. Interfaces de usuario “Explorer” y “Knowledge” Flow en Weka” Universidad Nacional de Colombia, p. 39, 2014.

[40] Sànchez-Marrè, M., Gibert, K., and Rodríguez-Roda, I. GESCONDA: A tool for knowledge discovery and data mining in environmental databases. Research on Computing Science. Centro de Investigación en Computación, Instituto Politécnico Nacional, México DF, México, 11, 348-364, 2004.

[41] Smith, T. C., and Frank, E. Introducing Machine Learning Concepts with WEKA. Statistical Genomics: Methods and Protocols, 353-378, 2016.

[42] Tay, J.H., Zhang, X. Neural fuzzy modeling of anaerobic biological wastewater treatment systems, J. Environ. Eng. ASCE, 125 (12) 1149–1159, 1999.

[43] Universidad Nacional Autónoma de México. Curso uso eficiente del agua y control de calidad de las descargas de aguas residuales en la industria. México, 2000.

[44] Verma, A., Wei, X., and Kusiak, A. Predicting the total suspended solids in wastewater: a data-mining approach. Engineering Applications of Artificial Intelligence, 26(4), 1366-1372, 2013.

[45] Waikato Environment for Knowledge Analysis (WEKA) (undated). Minería de datos. Recovered from: http://www.it.uc3m.es/~jvillena/irc/practicas/03-04/18.mem.pdf

[46] Yu, H. Q., and Fang, H. P. Acidogenesis of gelatin-rich wastewater in an up flow anaerobic reactor: influence of pH and temperature. Water Research, 37(1), 55–66, 2003.

[47] Zhang, Q. and Stanley, S. Real-Time Water Treatment Process Control with Artificial Neural Networks. J. Environ. Eng., 153-160, 1999.

[48] Zhu, J., Zurcher, J., Rao, M., Meng, M.Q.H. An on-line wastewater quality prediction system, based on a time delay neural network. Eng. Appl. Artif. Intell. 11 (2), 747–758, 1998.