Revista Logos, Ciencia & Tecnología
ISSN: 2145-549X
revistalogoscyt@gmail.com
Policía Nacional de Colombia
Colombia

Pedraza, Evelín; Herrera, Fabián; Díaz, Daissy; Gaona, Paulo; Montenegro, Carlos; Castro, Mario
Variables más influyentes en la calidad del agua del río Bogotá mediante análisis de datos
Revista Logos, Ciencia & Tecnología, vol. 7, núm. 2, enero-junio, 2016, pp. 32-39
Policía Nacional de Colombia
Bogotá, Colombia

Evelín Pedraza*
Fabián Herrera**
Daissy Díaz***
Paulo Gaona****
Carlos Montenegro*****
Mario Castro******

Variables más influyentes en la calidad del agua del río Bogotá mediante análisis de datos

Most influential people in the water quality of the river Bogota variables using data analysis

Variáveis mais influentes sobre a qualidade de água do Rio Bogota através da análise de dados

Resumen

En este documento se analiza la calidad del agua del río Bogotá correspondiente al periodo 2008-2015, la muestra incluye datos proporcionados por la Corporación Autónoma Regional (CAR) de Cundinamarca, según fases de la minería de datos, con el fin de comprobar patrones de comportamiento y la definición de variables de mayor impacto en la calidad del agua de la Cuenca.

Palabras clave: análisis de datos, árbol de decisión, índice de calidad del agua (ICA), minería de datos variables de impacto.

Abstract

In this paper the analysis of data on water quality in the Bogotá river is performed of the 2008-2015 period provided by the Regional Autonomous Corporation (CAR) of Cundinamarca, by applying the different phases of data mining in order to check whether the identification of patterns of behavior and defining variables of greatest impact on water quality in the basin is possible.

Fecha de recepción del artículo: 2 de diciembre de 2015
Fecha de aceptación del artículo: 14 de junio de 2016
DOI: http://dx.doi.org/10.22335/rlct.v7i2.258

* Estudiante Ingeniería de Sistemas, Universidad Distrital Francisco José de Caldas. Contacto: evelin.pedraza.5@gmail.com http://orcid.org/0000-0002-0828-1041

** Estudiante Ingeniería de Sistemas, Universidad Distrital Francisco José de Caldas. Contacto: fcherrerar@correo.udistrital.edu.co. http://orcid.org/0000-0002-1871-543X

*** Estudiante de Maestría en Desarrollo Sustentable y Gestión Ambiental, Universidad Distrital Francisco José de Caldas. Contacto: daissymdiaz@gmail.com http://orcid.org/0000-0002-7467-5620

**** Ingeniero de sistemas y Magíster en Ciencias de la Información y las Comunicaciones de la Universidad Distrital Francisco José de Caldas, Doctor en Informática de la Universidad de Alcalá de Henares España. Profesor de la Universidad Distrital Francisco José de Caldas. Contacto: pagaonag@udistrital.edu.co. http://orcid.org/0000-0002-8758-1412

***** Ingeniero de Sistemas y Magíster en Ciencias de la Información y las Comunicaciones de la Universidad Distrital Francisco José de Caldas, Máster en Dirección e Ingeniería de Sitios Web de la UNIR en la Rioja, España, y Doctor en Sistemas y Servicios Informáticos por la Universidad de Oviedo España. Profesor de la Universidad Distrital Francisco José de Caldas. Contacto: cemontenegrom@udistrital.edu.co. http://orcid.org/0000-0002-3608-7158

****** Biólogo, Universidad de los Andes. Doctor en Biología, Ecología y Etología Universite De Lyon I, docente investigador de la Universidad Cooperativa de Colombia. Contacto: mario.castrof@campus.ucc.edu.co. http://orcid.org/0000-0001-6328-8994

ISSN 2145–549X | ISSN 2422-4200, Vol. 7, No. 2, Enero - Junio 2016

Keywords: data analysis, data mining, decision tree, impact variables, water quality index.

Introducción

A partir de técnicas de minería de datos se busca seleccionar la información relevante de sobre la calidad del agua del río Bogotá en el período correspondiente a 2008-2015, con información sobre parámetros que corresponden a muestreos realizados dos (2) veces al año en cada una de las 81 estaciones de monitoreo ubicadas a lo largo de la cuenca del río Bogotá.

Esta información fue suministrada por la CAR de Cundinamarca, autoridad ambiental competente encargada del manejo de la cuenca en estudio. La cuenca del río Bogotá tiene un área de drenaje de 5,907 km^2, riega el departamento de Cundinamarca en sentido noreste-sureste, desde su nacimiento en el municipio de Villapinzón a 3.300 m. s.n.m., hasta su desembocadura en el río Magdalena en el municipio de Girardot a una altura 600 m. s.n.m (Subdirección Ambiental & Rep, 2011).

La cuenca se encuentra dividida en tres (3) tramos:

- La Cuenca Alta desde su nacimiento hasta el norte de la zona urbana del Distrito Capital, con una longitud de 165 km.
- La Cuenca Media, desde el inicio de la zona urbana de Bogotá hasta el Salto de Tequendama, con una longitud de 90 km.
- La Cuenca Baja, aguas abajo del Salto de Tequendama hasta su desembocadura en el río Magdalena, con una longitud de 55 km (Subdirecci *et al.*, 2011).

A partir de la extracción del conocimiento útil se pretende llegar a metas generales que puedan mostrar una tendencia del comportamiento de los parámetros estudiados; asimismo, se busca aportar información para futuras investigaciones con temas relacionados con el desarrollo de *software* basado en conocimiento, y el recurso hídrico según los diferentes resultados arrojados por el estudio realizado en este proyecto.

Estado del arte

Chang, Tsai, Chen, Coynel, & Vachaud (2015) manifiestan que el muestreo ambiental es complicado, laborioso, costoso y requiere de mucho tiempo (Hernández, 2011). También es poco probable que tenga datos de series de tiempo de la calidad del agua continuos a largo plazo con propiedades completas en todos los lugares de muestreo en un sistema fluvial.

Dado esto, Roland, E., Uhrmacher, A. & Saha (2009) comentan que en algunos trabajos se opta por utilizar algoritmos de predicción de la minería de datos que ha sido de gran utilidad para la extracción de conocimiento en muchos ámbitos gracias a su función, ya que es una de las formas más sofisticadas de extraer información importante y relevante a partir de una base de datos, al utilizar técnicas para encontrar patrones y crear modelos con dicha información (Rosado, 2015), para superar la escasez de datos y, simultáneamente, aumentar la fiabilidad de los modelos (Chang *et al.*, 2015).

Un ejemplo de la aplicación de la minería de datos en el campo de la investigación forestal es el artículo: "Investigación de indicadores generales que influyen sobre los incendios forestales y su modelado de susceptibilidad utilizando diferentes técnicas de minería de datos" (Marín, Quintero y Medina, 2013), en el que se trabajó la búsqueda de indicadores que más influyen en la aparición de incendios forestales y mapas de susceptibilidad con base en el árbol de regresión impulsado (BRT), el modelo aditivo generalizado (GAM), y los modelos de conjunto de árboles al azar (RF) en los municipios de *Minudasht*, Provincia de *Golestan*, Irán.

Para el estudio se utilizaron 15 atributos, con el fin de identificar las ubicaciones de incendios forestales y los registros históricos. De manera que, en los resultados finales se encontraron tres (3) de las 15 cualidades, a saber, precipitación anual, distancia a las carreteras, y los factores de uso de la tierra (Pourtaghi, Pourghasemi, Aretano, & Semeraro, 2016). Otro artículo indica el procedimiento que se siguió después de recopilar información biológica y socioeconómica sobre la cuenca alta del río Otún en Risaralda (Colombia).

Con la información recopilada se realiza un análisis con el fin de identificar patrones que relacionen ambos tipos de información con el fin de utilizar estos patrones para tomar decisiones sobre el manejo y conservación de la fauna en esta cuenca del río, estos patrones se identificaron mediante el sistema de información geográfica "ArgGis". Usando minería de datos con base en árboles de relaciones; obtenidos con el algoritmo J-48 en WEKA y utilizando relaciones en Matlab. En conclusión, el estudio determinó tres (3) variables físicas significativas como: altitud, precipitación y temperatura, y una variable socioeconómica (Coronel-Picón, Obregón-Neira, & Jiménez-Romero, 2012).

En el artículo modelado de la calidad del agua en un río urbano, se utilizan datos de factores hidrológicos, proponen un esquema de análisis sistemático (SAS) para evaluar la interrelación espacio-temporal de la calidad del agua en un río urbano y la construcción de modelos de estimación de la calidad del agua al utilizar dos redes neuronales artificiales estáticas y una red neuronal artificial (RNA) dinámica, junto con la prueba Gamma (GT) basada en la calidad del agua, datos hidrológicos y datos económicos, así como la concentración de clorofila en el depósito Daechung, Corea. Para encontrar una correlación entre estas variables y establecer un modelo predictivo de la concentración de clorofila en el depósito Daechung se utilizaron modelos estadísticos como R y la minería de datos, que implementa algoritmos como modelos de árbol, redes neuronales artificiales (ANN), y la función de base radial (RBF).

Una herramienta óptima para la minería de datos y puesta en práctica en este artículo es Weka (Entorno para análisis del conocimiento de la Universidad de Waikato). Con el uso de una base de datos de 10 años de información y usando una selección de atributos en Weka, se encontró que las mejores variables para la predicción de la concentración de clorofila en el depósito Daechung son el COD, T-P, y PO4-P (Chang *et al.*, 2015).

Método

Pérez (2015) afirma que "la minería de datos se define como un conjunto de técnicas encaminadas al descubrimiento de la información contenida en grandes conjuntos de datos" y Vergel, & Martínez (2013) señalan que existen técnicas de las cuales se escogió modelado originado por los datos, en la que los modelos se crean automáticamente partiendo del reconocimiento de patrones siguiendo las fases de selección de los datos, limpieza de los datos, codificación de los datos procesados, minería de los datos transformados, modelo y, finalmente, interpretación y evaluación del conocimiento, las cuales se explicarán y desarrollarán en el presente estudio.

Al inicio, se tienen 34.000 registros en documentos tipo PDF con variables que no tienen el papel de dependientes o independientes, contenidas en el agua en diferentes cantidades y distribuidas en diversas zonas por las que pasa la fuente hídrica.

Fase de selección de los datos

El proceso de análisis de datos se establece inicialmente con el proceso de definición de las estructuras de datos. A partir de la recopilación documental de los Boletines de Calidad Hídrica del río Bogotá se procedió a establecer la conversión del formato de documentos portátiles (PDF) a un formato de texto plano con el fin de facilitar su introducción en el motor de base de datos MySQL. Para ello, se realizó un programa convertidor bajo el lenguaje Java y el uso de la librería PdfBox.

Se cuenta con una base de datos de 58 variables que son parámetros físico-químicos contenidos en el agua del río Bogotá. Todas estas variables tienen un porcentaje de afectación sobre la calidad del agua, es decir, la existencia en mayor o menor medida de algunas de estas variables definen si el agua en ese punto de monitoreo es buena o mala para los diferentes usos como el consumo doméstico, uso en agricultura y ganadería, como fuente de energía, entre otros.

Para el cálculo del índice de calidad del agua (ICA) en Colombia, se ha medido un conjunto de cinco (5) variables, por saber: oxígeno disuelto, sólidos suspendidos totales, demanda química de

ISSN 2145–549X | ISSN 2422-4200, Vol. 7, No. 2, Enero - Junio 2016

oxígeno, conductividad eléctrica y pH total. Este cálculo se explica en el artículo (Castro, Almeida, Ferrer, y Díaz, 2014).

Fases de exploración y limpieza

En la fase de exploración se hace uso de herramientas informáticas gráficas, *rapid miner*, *ableau* (Tunjano y Calvo, 2011), con el objetivo de observar gráficamente las tendencias que pueden tener los datos iniciales y a partir de esto ver posibles caminos por los que se puede llevar la investigación. Al observar que los datos se encuentran ordenados alfabéticamente y no en el orden geográfico en el que deberían estar, se opta por realizar un ordenamiento en tablas, primero por periodos: 2008-1, 2008-2, 2009-1 ,2009-2, 2010-1, 2010-2, 2011-1, 2012-1, 2012-2, 2013-1, 2013-2, 2014-1, 2014-2, 2015-1, los cuales dan un total de 14 tablas en las que falta el período 2011-2 debido a que en este período existen datos de varias variables pero no se encuentran los datos completos de las variables utilizadas para calcular el ICA, que en este estudio es la variable principal, por esto se excluye tal periodo. Después en cada tabla se clasifican los 81 puntos de monitoreo en su respectiva Cuenca Alta, Media o Baja, lo que muestra que en la totalidad de los períodos hay más puntos de la Cuenca Alta en comparación con la Media y la Baja; asi mismo, se observa que la Cuenca Alta cuenta con mayor calidad en el agua ya que es donde nace la fuente hídrica.

En cuanto a la limpieza de los datos, se evidencia que de un período a otro, se han dejado de monitorear algunos puntos que tienen una calidad del agua atípica en comparación con los demás puntos de la misma Cuenca, algo que sucede en las tres (3) cuencas, lo cual conlleva a una vaga comparación de los puntos irregulares encontrados.

Organización de datos en un modelo relacional

A partir de la conversión de datos, se procedió a realizar un esquema lógico de la base de datos bajo el modelo relacional, para lo cual se realizó la identificación de las diferentes relaciones entre entidades como los puntos de monitoreo a lo largo de la fuente hídrica, las mediciones realizadas en un período específico, entre otras (*Figura 1*).

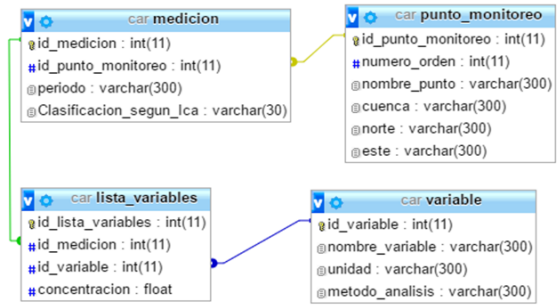

Figura 1. Esquema lógico del modelo relacional de la base de datos. (Fuente propia).

El esquema lógico basado en un modelo relacional, está compuesto de cuatro (4) tablas: medición, lista_variables, punto_monitoreo y variable.

Depuración de datos

Con base en el modelo relacional propuesto, a través de un servidor web Apache y su utilidad de administración, se realizó una depuración de datos a partir de la identificación de errores a nivel de codificación de caracteres como la existencia de registros duplicados asociados a las variables de medición, la identificación de valores inconsistentes y su rectificación de acuerdo con los valores límite determinados dentro de los documentos técnicos, mediante sentencias SQL.

Clasificación del ICA

Se establece una calificación de la calidad del agua, definido como "un número o una clasificación descriptiva de varios parámetros, cuyo propósito principal es simplificar la información para que pueda ser útil en la toma de decisiones de las autoridades" (IDEAM., 2013). Para ello, documentos como la hoja metodológica del indicador índice de calidad del agua (IDEAM., 2013), determinan rangos de clasificación, de acuerdo a valores optativos del indicador. Este esquema de calificación se expone en la Tabla 1.

Tabla 1

Calificación de la calidad del agua según los valores del ICA.

Categoría de valores que puede tomar el Indicador	Calificación de la calidad del agua	Señal de alerta
0,00-0,25	Muy mala	Rojo
0,26-0,50	Mala	Naranja
0,51-0,70	Regular	Amarillo
0,71-0,90	Aceptable	Verde
0,91-1,00	Buena	Azul

(IDEAM. 2013).

Resultados

Fase de transformación en minería de datos.

Para el análisis de datos se optó por la herramienta *rapidminer* ya que es una aplicación eficiente, rápida y fácil de emplear para la minería de datos. Con el apoyo de los datos proporcionados por la CAR y la herramienta mencionada, se desarrollaron diversos métodos para el análisis de datos que permitieron hacer inferencias acerca del comportamiento del ICA y demás variables pertenecientes a la base de datos.

Para empezar, se optó por la elaboración de un árbol de decisiones, con el objetivo de observar como se ve afectado el ICA por los diferentes factores que influyen en la calidad del agua del río Bogotá.

En la elaboración del árbol de decisión, fueron necesarios diversos operadores proporcionados por *rapidminer,* cuyo funcionamiento se detalla a continuación:

El operador *retrieve* permite leer los datos pertenecientes a una base de datos o un archivo en *excel*, que se haya importado previamente. Este operador no tiene un puerto de entrada, puesto que lo único que requiere es un conjunto de datos válidos para trabajar.

El operador *validation* permitirá saber la precisión con la que se construirá el árbol de decisiones, el cual está compuesto de dos (2) subprocesos: *training subprocess* y *testing subprocess*.

Debe importarse el operador *retrieve* a la zona de trabajo, después unir su salida *out* con la entrada *tra* del operador *validation*. Posteriormente, unir las salidas *mod* y *ave* con las salidas *res* del proceso. Sin embargo, dentro del operador *validation,* hay dos (2) subprocesos que podemos utilizar para obtener el modelo de árbol de decisión que deseamos, además de una verificación de la veracidad de este modelo.

En el subproceso *training* se construye el modelo con el que se van a tratar los datos. En este caso, el modelo por aplicar sólo consta del operador *decision tree*, el cual a partir de un conjunto de datos de entrada, entre los que se encuentra la variable dependiente, crea un árbol de decisiones basado en los valores de dicho conjunto de datos. Debe unirse la salida *tra* del subproceso *training* a la entrada *tra* del operador *decisión tree,* y, de este último, unir su salida *mod* a la salida *mod* del subproceso *training.*

Posteriormente en el subproceso *testing*, cuyo propósito es la medición del rendimiento del modelo, se aplica el operador *apply model* para aplicar el modelo de árbol de decisión del subproceso *testing* en la prueba de la veracidad del árbol, puesto que si se usara el operador *decision tree* sin verificar la veracidad en la construcción del árbol, se podría llegar a hacer el análisis de un árbol totalmente erróneo.

Debe enlazarse las salidas *mod* y *tes*, con las entradas *mod* y *uni* del operador *apply model*. De este operador debe enlazarse su salida *lab* con la entrada *lab* del operador *performance.* Este operador se utiliza para la evaluación del desempeño y proporciona una lista de los valores de los criterios de rendimiento. Estos criterios de rendimiento se determinan automáticamente con el fin de ajustarse al tipo de tarea de aprendizaje.

ISSN 2145–549X | ISSN 2422-4200, Vol. 7, No. 2, Enero - Junio 2016

Y, de este último, debe enlazarse su salida *per* con la salida *ave* del subproceso *testing*.

El árbol se dividió en varias imágenes, puesto que por su magnitud, no puede ser expuesto en una sola imagen.

En las siguientes imágenes se ilustrará la información correspondiente al árbol generado, en diferentes partes para su explicación:

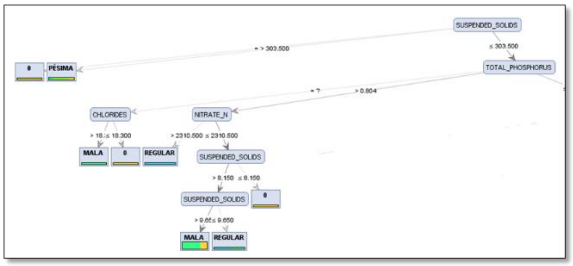

Figura 2. Parte izquierda del árbol de decisión. (Fuente propia aplicación rapidminer).

De acuerdo con la *figura 2*, el nodo principal es la variable *suspended solids*, de manera que si los sólidos suspendidos son mayores a 303,500, la calidad del agua será pésima. Por otro lado, si los sólidos suspendidos son menores o iguales que 303,500 y el fósforo total es mayor que 0.804 y el nitrato es menor o igual que 2310,500 y los sólidos suspendidos son mayores que 0,150 pero mayores a 9,650, la calidad del agua será mala. Si los sólidos son aún menores que 9,650, la calidad del agua será regular.

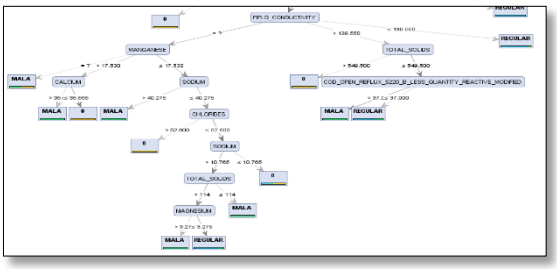

Figura 3. Rama total_phosporus del árbol de decisión. (Fuente propia aplicación rapidminer).

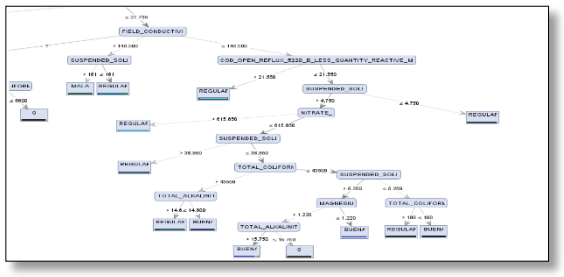

Figura 4. Rama field_conductivity del árbol de decisión. (Fuente propia aplicación rapidminer).

De las dos (2) imágenes figura 3 y figura 4, anteriormente mostradas, se puede observar que si los sólidos suspendidos son menores o iguales a 303,500, y el fosforo total es menor que 0,804, y el cod reflujo abierto 5220 b menor calidad del reactivo modificado es mayor que 34,050, y el magnesio es menor o igual que 52,105, y el calcio es menor o igual que 2,985, la calidad del agua será regular.

De manera similar se puede abordar la interpretación del contenido restante del árbol mostrado en las siguientes figuras 5 y 6.

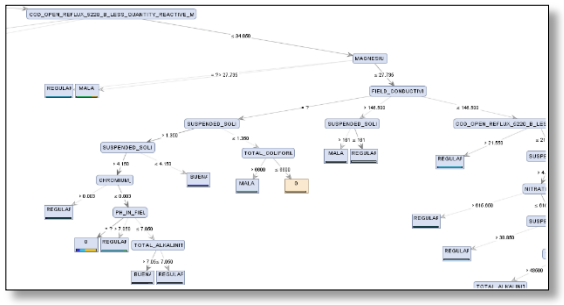

Figura 5. Rama magnesiu del árbol de decisión. (Fuente propia aplicación rapidminer).

Para finalizar, se desarrolló una matriz de correlación para tener una observación cuantitativa de la influencia de las variables entre sí. El proceso es sencillo puesto que sólo implica el uso de dos (2) operadores; se debe hacer uso del operador *retrieve*, uniendo su salida *out* con la entrada *exa*, del operador *correlation matrix* cuya función es determinar la correlación existente entre los distintos atributos que pertenecen a la base de datos que se trabajó, y de este último unir su salida *mat* con la salida *res* del proceso.

Como resultado de la acción de estos operadores, se obtuvo una relación clara entre todas las variables que la CAR registró en su base de datos; sin embargo, aquí solo se revelará la correlación existente entre el ICA y las demás variables, por medio de la siguiente tabla 2.

Tabla 2

Correlaciones ICA-variables

Variables	Correlación ICA
Field_conductivity	0.5250970498276808
Total_alkalinity	0.2890202097916725
Cod_open_reflux_5220_b_less_quantity_reactive_modified	0.2820786762575087
Total_phosphorus	0.24845797477672457
Ph_in_field	0.23194799617066356
Magnesium	0.19851896527002771
Suspended_solids	0.13798815069209802
Total_solids	0.09068316528937045
Total_coliforms	0.08073515331281593
Calcium	0.07869230009850271
Chromium_6	0.045601606121833275
Manganese	0.015759929863442525
Chlorides	0.002444330270115761
Sodium	-0.008089169933836198
Nitrate_n	-0.05115976857490208

(Fuente propia aplicación *rapidminer*).

En la tabla se relacionan los diferentes atributos que componen la base de datos, en orden descendente. Esta relación está dada por un número entre -1 y 1, donde -1 denota la más baja correlación y 1 denota la más alta correlación.

Discusión

Se puede observar que la mayoría de las variables presentes en el árbol de decisión tienen una correlación mayor a cero (0). Las que no la tienen pertenecerán a una rama que no es la más eficaz para describir un modelo que permita determinar el comportamiento del ICA. Así se deberá perfeccionar el modelo de árbol de decisión para optimizar sus aplicaciones, teniendo en cuenta las variables que tengan mayor correlación con el ICA y estén presentes en el árbol de decisión, puesto que aunque todas las variables van a tener cierta influencia en la calidad del agua, solo se debe tomar en cuenta las que tengan un impacto realmente significativo sobre él.

De la observación del árbol, se aprecia el impacto que tienen los sólidos suspendidos y se puede visualizar cómo no solamente son el nodo padre, sino que, también, forman parte de otras ramas, especificando en qué proporciones los sólidos suspendidos afectan la calidad del agua. De esta forma con la medición de esta variable es posible descartar muchas otras ramas del árbol.

Adicionalmente, se puede ver cómo la falta de datos afectaron los resultados de este estudio, puesto que algunas ramas llegan a una conclusión de ICA igual a cero (0); es decir, no se tienen suficientes datos para medir la calidad del agua como muy mala, mala, regular, aceptable o buena. Para lo que se propone en este proyecto, estos datos vienen a ser irrelevantes, puesto que se busca la identificación de las variables que permitan determinar el ICA, y una rama que no proporciona esta información, no es de mucha relevancia.

Conclusiones

Este estudio ilustró el uso de la minería de datos y sus fases en el manejo de información correspondiente a parámetros relacionados con la calidad del agua, a través del uso de un *software* libre como lo es *rapidminer*, el cual facilita el tratamiento de datos y los medios necesarios para el manejo adecuado de la base de datos correspondiente a las variables del agua del río Bogotá.

ISSN 2145–549X | ISSN 2422-4200, Vol. 7, No. 2, Enero - Junio 2016

El estudio reveló que en el ICA solo las 13 variables que se muestran en la tabla 2 son relevantes, puesto que se evidencia una mayor influencia de estas sobre el mismo, dejando de lado la gran mayoría de datos expuestos en los informes de la CAR de Cundinamarca, lo que permite suponer que si se avanza en el estudio de variables de mayor influencia en la calidad del agua u otros recursos naturales, se podrán generar modelos óptimos que permitirán disminuir los costos y el tiempo en el monitoreo de tales recursos.

Acorde con los resultados del presente trabajo, resulta conveniente que en futuros estudios de este tipo, se tenga en cuenta desde el inicio de la investigación, el impacto resultante de contar con datos faltantes, para hacer un tratamiento óptimo y así evitar la aparición de ramas con datos iguales a cero (0) o irrelevantes.

Referencias bibliográficas

Castro, M., Almeida, J., Ferrer, J., & Diaz, D. (2014). Indicadores de la calidad del agua: evolución y tendencias a nivel global. *Ingeniería Solidaria; Vol. 10, Núm. 17* http://doi.org/10.16925/in.v9i17.811

Chica, O., Galvis, N. & Madrid, J. (2007). *Validación Métodos Analíticos en Aguas. Medellín.*, Validación.

Coronel-Picón, Y. R., Obregón-Neira, N., & Jiménez-Romero, G. L. (2012). *Relationship Patterns between Biological Information and Physical and Socioeconomic Information. The Otun River Basin in Risaralda*, Bogotá: UAN. 16.

Hernández Gómez, J. (2011). Inteligencia económica. Revista Logos Ciencia & Tecnología, 3(1), 37-55. doi:http://dx.doi.org/10.22335/rlct.v3i1.105

IDEAM. (2013). *Hoja metodológica del indicador índice de calidad del agua*. Bogotá: IDEAM, 11.

Marín Fonseca, R., Quintero Montenegro, D., & Medina Valencia, J. (2013). El rol de la gestión del conocimiento en la implementación de un Sistema Integrado de Gestión. Revista Logos Ciencia & Tecnología, 4(2), 33-41. doi:http://dx.doi.org/10.22335/rlct.v4i2.188

Medina Bejarano, R., & Pineda Torres, N. (2012). Globalización, tecnociencias y culturas relacionales. Revista Logos Ciencia & Tecnología, 4(1), 107-120. doi:http://dx.doi.org/10.22335/rlct.v4i1.173

Pourtaghi, Z. S., Pourghasemi, H. R., Aretano, R., & Semeraro, T. (2016). Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. *Ecological Indicators*, *64*, 72–84. http://doi.org/10.1016/j.ecolind.2015.12.030

Roland, E., Uhrmacher, A. & Saha, K. (2009). Data mining for simulation algorithm selection. *Proceeding Simutools '09 Proceedings of the 2nd International Conference on Simulation Tools and Techniques*, *14*.

Rosado Gómez, A., & Verjel Ibáñez, A. (2015). Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander. *Revista Tecnura*, 19(45), 101-114. doi:http://dx.doi.org/10.14483/udistrital.jour.tecnura.2015.3.a08

Subdirecci C. A. R., Ambiental, D., & Rep, S. (2011). Boletín de calidad de las cuencas de la jurisdicción car 2011. Bogotá:CAR, 36p.

Tunjano Huerta, C., & Calvo Valencia, D. (2011). Evaluación de sustancias fitoprotectoras usadas como estrategia de neutralización de la acción del glifosato sobre cultivos de erythroxylum coca. *Revista Logos Ciencia & Tecnología*, 2(2), 26-31. doi:http://dx.doi.org/10.22335/rlct.v2i2.79

Vachaud, G., Chang, F.-J., Tsai, Y.-H., Chen, P.-A., & Coynel, A. (2015). Modeling water quality in an urban river using hydrological factors – Data driven approaches. *Journal of Environmental Management,* 151, 87–96. http://doi.org/http://dx.doi.org/10.1016/j.jenvman.2014.12.014

Vergel, M. Martínez, J. (2015). Filosofía gerencial seis sigma en la gestión universitaria. *Revista Face*, 15(2). 99-106