

MINERÍA DE DATOS (DATA MINING) EN LOS ABASTECIMIENTOS DE AGUA, CASOS HIPOTÉTICOS DE UTILIZACIÓN

José Luis Díaz Arévalo¹; Rafael Pérez García¹; Mario Alejandro Nudelman²;
Joaquín Izquierdo Sebastián¹

Resumen – En este trabajo se presenta la aplicabilidad de las técnicas de minería de datos de generación de reglas y árboles de decisión para la gestión de abastecimientos, teniendo en cuenta la importancia que representa en este momento la relación con el cliente. Se propone un ejercicio a partir de los datos que se suministran en la factura de abastecimiento en la ciudad de Valencia para la clasificación del cobro del cargo fijo, que se ve que no es proporcional con el consumo del periodo, penalizando de forma mayor los consumos más bajos. Un segundo ejercicio consiste en la clasificación del consumo medio en una ciudad del Norte de Argentina, en el municipio de Fontana provincia del Chaco, teniendo en cuenta variables tales como la cantidad de personas por vivienda, los ingresos medios por unidad familiar, el nivel de escolarización, y el tipo de vivienda en cuanto a su construcción.

Abstract – In this paper, we present the applicability of data mining techniques for decision trees and rules generation in the water supply management, in agreement with the customer relationships. We propose a pair of exercises that show the utility of these Data Mining techniques in the resolution of problems, as well as the understanding of patterns in the management of water business. The first one corresponds to data in the water invoice; this information shows that the consumption is penalized for low consumptions. The second exercise is the classification of water demand in a town of the north of Argentina in the Chaco province. The variables utilized correspond to number of persons by housing, incomes by familiar units, scholar levels, and the types of housings if they have deck or not have a deck for classification of the consumptions.

Palabras clave: Minería de datos, extracción de patrones, aprendizaje automático, sistemas de abastecimiento de agua.

Keywords: Data mining, pattern extraction applications, machine learning, water supply systems.

¹ Universidad Politécnica de Valencia – Departamento de Ingeniería Hidráulica y Medio Ambiente – Grupo Multidisciplinar de Modelación de Fluidos – Camino de Vera S/N – CP E46022 – Valencia (España) – Tel: (+34) 96 3879890 – Fax: (+34) 96 3877981 E-mail: jldiaz@gmmf.upv.es ; rperez@gmmf.upv.es

² Becario Doctoral del Centro Experimental de la Vivienda Económica / CONICET - Dirección postal de la institución: Igualdad 3585 (5003) Córdoba - Pcia. de Córdoba, Argentina – E-mail: marioalejandro_n@yahoo.com

INTRODUCCIÓN

En la actualidad los cambios estructurales que ha habido en el mundo en cuanto a la globalización, gobernabilidad, economías emergentes, entre otros, se presentan dos retos fundamentales para cualquier empresa, por una parte la relación que deben mantener con sus clientes, cada vez más informados y por tanto llamados a tomar decisiones, y por otra la innovación como pilar fundamental para el sostenimiento y crecimiento de las empresas.

En cuanto a la relación con los clientes es claro que las empresas cada día deben preocuparse más por estos, “El mercadeo ha evolucionado. El concepto ha cambiado de la búsqueda de participación de mercado a la participación de los clientes. Hoy, lo que importa es que tanta profundidad se tiene en la relación con esos clientes”. “La hipótesis es que mientras la empresa mantenga al cliente será más rentable para ella, pero creando valor para el cliente” (Robin Garzón, Revista Dinero, 2005). De acuerdo con lo anterior, ya sea que la gestión de los abastecimientos de agua sea pública o privada, al tratarse de un recurso natural se debe dar al usuario en condiciones óptimas. Esto no quiere decir que el agua deba ser entregada sin ningún tipo de costo, porque esta debe ser tratada y conducida para que llegue al usuario y esto tiene un valor, pero a su vez el usuario está en su derecho a exigir la calidad del producto ofrecido y un precio justo, que no se generen sobrecostos por fraudes entendidos como pérdidas en la red (que la empresa pueda darse cuenta que si mantiene las presiones en un cierto nivel las pérdidas en la red van a disminuir), quizás en una política de precios justa decirle al cliente que el cobro de la factura le puede disminuir de precio si tiene ciertos hábitos de consumo que lo van a conducir a esto.

“En el suministro de agua a ciudades e industrias uno de los principales problemas es el de las pérdidas en las tuberías de distribución que, en bastantes lugares, son más del 50% del agua producida. De hecho las restricciones de agua que se suelen dar en algunas ciudades en las épocas de escasez, no son tanto por el agua consumida por los habitantes, sino para evitar las pérdidas en las tuberías” (Tema 6, Alimentos y Agua para una Población Creciente, Agua en España).

Gran variedad de líderes mundiales y empresarios exitosos coinciden en que la innovación es el camino para llevar países y empresas hacia el crecimiento y el desarrollo, “La innovación tiene que ser el motor del cambio, las empresas tienen que encontrar su propia estrategia y ser las mejores. Las grandes fusiones no funcionan y el benchmarking y la imitación no tienen ningún sentido” (Tom Peters, Revista Dinero, 2005). “La única forma de crecer como empresa es por medio de la innovación, ya que es ahí donde esta el valor agregado. Las fusiones y las adquisiciones no son suficientes. En este nuevo mundo, el benchmarking ya no tiene sentido porque la base del conocimiento es totalmente diferente” (C.K. Prahalad, Revista Dinero, 2005).

No menos importante que lo anterior y ubicándonos en el tema central de la ponencia, esta el recurso agua, según el Fondo de la Población de Naciones Unidas, dentro de 25 años una de cada tres personas en la Tierra tendrá poca agua o nada (Claude, 2005). Solo el 2.5% del agua del planeta es dulce, y menos de la mitad tiene disponibilidad de uso. En la actualidad, más de 1200 millones de personas presentan algún grado de escasez del recurso, especialmente en América Latina, África y Asia.

Por tanto, la Minería de Datos presenta características propias como herramienta que permite abordar infinidad de casos de los anteriormente expuestos. Cabe anotar que en este trabajo se está realizando un desarrollo experimental a partir de investigaciones y experiencias

prácticas que se han desarrollado en diferentes áreas del conocimiento, y en las cuales se han obtenido buenos resultados, intentando innovar en la búsqueda de conocimiento en los tópicos que nos atañen de la gestión de los abastecimientos de agua, intentando generar inquietud, adquirir conocimiento y dar soluciones a problemáticas del día a día del manejo de los abastecimientos de agua.

GENERALIDADES

El concepto de minería de datos apareció hace más de una década. El interés en este campo y su explotación en diferentes especialidades (negocios, finanzas, ingeniería, banca, salud, sistemas de energía, meteorología....), se ha incrementado básicamente porque se cuenta con las herramientas necesarias para su utilización, tales como computadores potentes, medios de almacenamiento, abaratamiento de costos entre otros.

La minería de datos (Data Mining) corresponde a una fase de un proceso llamado Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Data Bases, KDD), que nos permite “*identificar patrones válidos, novedosos, potencialmente útiles y comprensibles de los datos*” (Fayyad, 1996). El propósito general de la minería de datos es como herramienta que permite procesar información de gran cantidad de datos almacenados, desarrollando procedimientos para manejar estos datos y tomar decisiones futuras.

Cabe destacar que el Descubrimiento de Conocimiento en Bases de Datos (KDD) involucra la investigación de diferentes áreas tal cómo: el aprendizaje automático, el reconocimiento de patrones, las bases de datos, la estadística, la inteligencia artificial, la adquisición de conocimiento y la visualización de datos.

Dentro del descubrimiento de conocimiento, el aprendizaje automático es reconocido comúnmente como la aplicación de algoritmos de inducción, es decir, el descubrimiento de conocimiento en bases de datos (Minería de Datos) y la parte aprendizaje automático tienen que ver con el aprendizaje a partir de ejemplos propuestos para los algoritmos de inducción utilizados. La Figura 1 muestra los pasos seguidos en el proceso de Descubrimiento de Conocimiento en grandes bases de datos (KDD).

Es difícil dar una definición precisa para el aprendizaje automático (Cios et al. 2000). En sentido amplio se puede presentar como la habilidad de un programa de computador para mejorar su desempeño en alguna especialidad basado en experiencias anteriores, también podemos definir el aprendizaje automático cómo la modelación del ambiente, o que es la habilidad de un programa para generar una nueva estructura de datos que es diferente de unos viejos, o como la generación de reglas *si...entonces* de datos numéricos o nominales.

La ventaja de los algoritmos de aprendizaje automático en contraste con las *Cajas Negras* de las redes neuronales, es que las reglas que se generan son explícitas y pueden ser modificadas, *aprendiendo de*, o utilizadas por usuarios en formación en una especialidad.

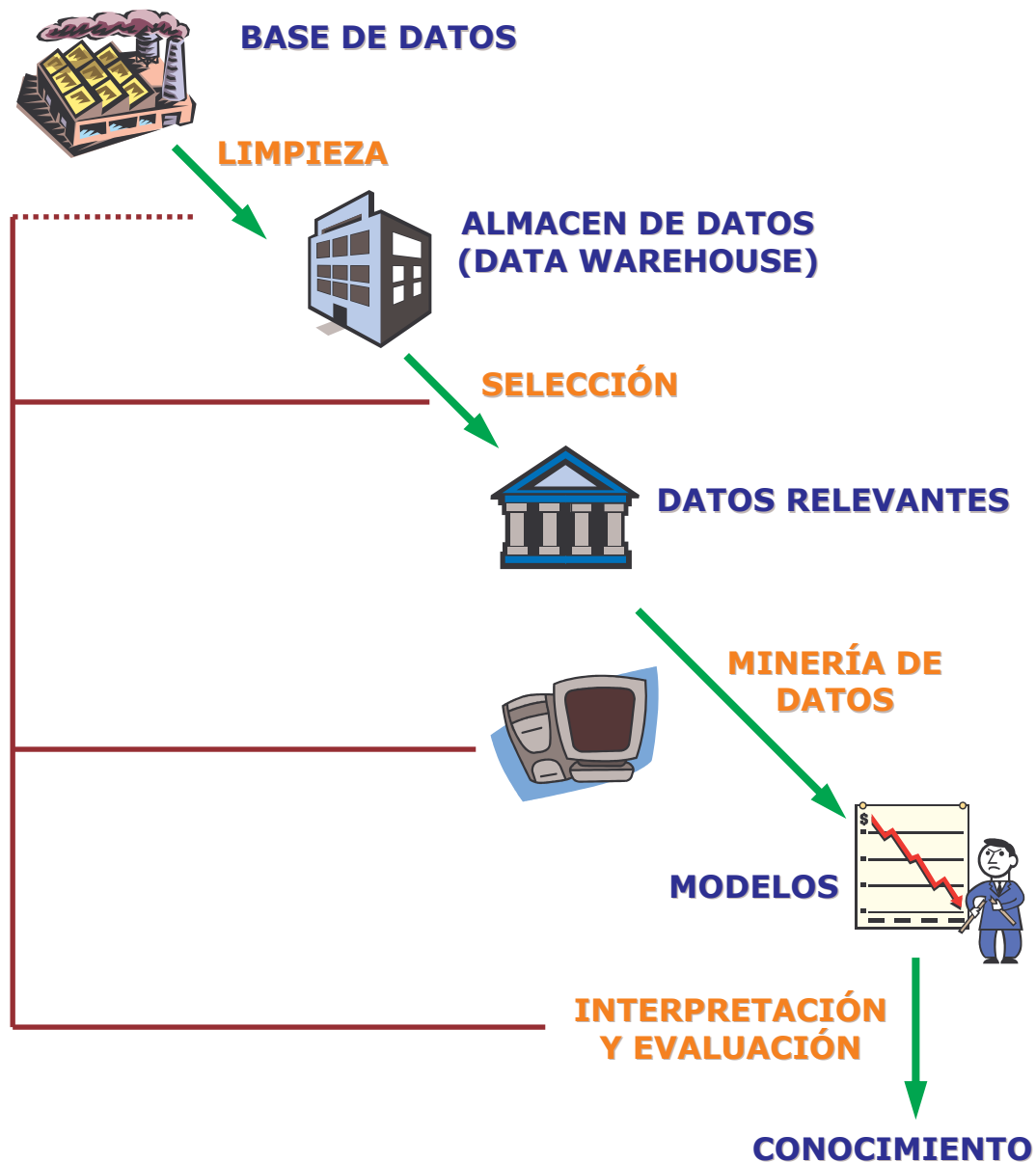


Figura 1 – Pasos del descubrimiento de conocimiento KDD.

APLICACIONES DE MINERÍA DATOS A LAS REDES DE ABASTECIMIENTO

El tema ha sido investigado y aplicado con éxito en diferentes ramas de la ciencia y el conocimiento, especialmente en temas de telecomunicaciones, banca y finanzas, mercados, medicina entre otros, pero en temas de ingeniería Civil y en especial los concernientes a los sistemas de abastecimiento de agua es muy escasa la información y referencias con las que se cuenta y la explotación del potencial de estas técnicas de análisis de datos está por desarrollarse. Algunos autores (An et al, Fleming, Jain, Stark) han aplicado algunas de las técnicas de minería de datos para desarrollar investigaciones tendientes a la determinación del caudal de consumo, básicamente las técnicas de Conjuntos Aproximados (Rough Sets) y

Redes Neuronales (Artificial Neural Networks) con base en mediciones de parámetros diferentes que consideran pueden afectar el consumo y han obtenido resultados bastante precisos y aproximados a la realidad y a los obtenidos por técnicas tradicionales de manejo de datos. Otros autores (Babovic, Savic) han trabajado modelos basados en técnicas de minería de datos para optimizar el mantenimiento y rehabilitación de activos (tuberías, bombas, válvulas).

CASOS HIPOTÉTICOS

Gestión del cobro

Este ejercicio está basado en información proporcionada por los recibos de cobro de servicio de abastecimiento en la ciudad de Valencia durante los últimos 5 años. La información corresponde a una vivienda ocupada por dos personas y los datos utilizados son los que se dan en la factura de cobro, tal como se muestra en la Figura 2.

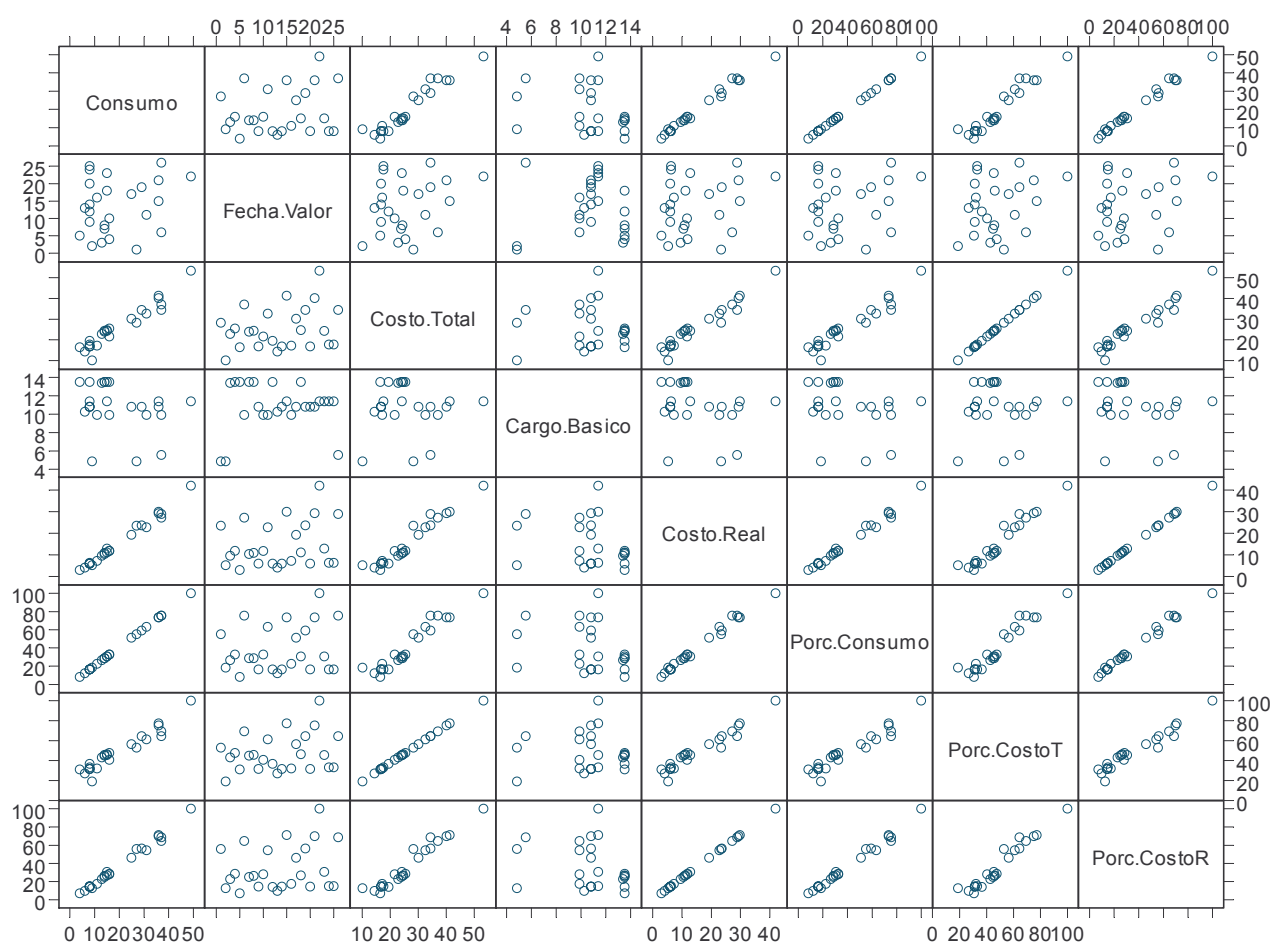


Figura 2. Dispersión de Variables

Los datos corresponden al Consumo (m^3), en este aspecto cabe anotar que en la ciudad de Valencia las lecturas se realizan en algunas zonas cada tres meses y en otras cada dos, además una de estas lecturas se hace estimada y la otra es real tomada del contador. La fecha valor corresponde a la fecha en que se factura, el costo total corresponde al valor total de la

factura, el costo real es el costo total menos el cargo básico (fijo) y los porcentajes de consumo y costos corresponden a valores con respecto al mayor de cada uno de ellos.

En este grafico se puede observar la relación existente entre cada una de las variables, notándose que la fecha valor y el cargo básico no muestran un comportamiento característico o relación de dependencia con ninguna de las variables.

En la Figura 3 se puede observar como en la mayoría de los casos cuando la lectura se realiza de forma estimada el cobro total es inferior a cuando la lectura se realiza leyendo el contador.

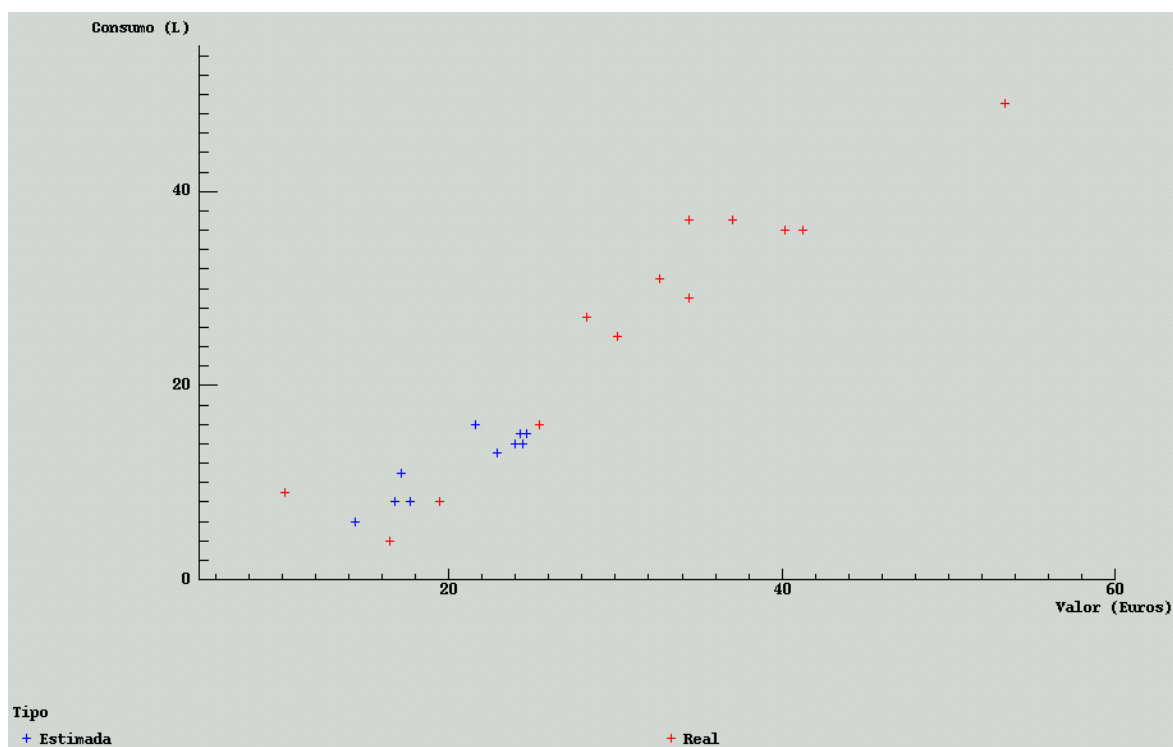


Figura 3. Dispersión de Variables

En la Figura 4 se puede observar cómo el cargo básico penaliza en mayor proporción los consumos bajos que los altos, algo que bajo el punto de vista de una estructura tarifaria que quiera promover el ahorro y no desperdicio del agua no parece muy lógico, aunque bien es cierto que estos costos corresponden al mínimo que se debe cobrar por el suministro generado por los gastos que conlleva éste, la presentación al cliente no se ve clara, aunque es cierto que podemos estar lejos de un mercado libre del agua, más lógico pareciera que si hay un cargo fijo el precio del agua se valla incrementando con el consumo.

Se puede observar como para consumos bajos, la diferencia en porcentaje de lo correspondiente al cargo básico y al cobro total de la factura puede llegar a un 20%, mientras que para valores de consumos más altos esta diferencia se reduce significativamente (menos del 1%).

Por tanto, se propone un primer modelo generado a partir de estos datos para clasificar el valor del consumo fijo haciendo uso del algoritmo C5.0 implementado en el programa Clementine de SPSS tal como se muestra en la Figura 5.

Se puede observar como la variable que tiene mayor ganancia es el costo total que se está cobrando en la factura, para valores de menos de 13 euros se propone un costo fijo de 5 euros, incrementándose para valores superiores. Sin embargo, también se puede ver que este valor de cargo fijo es alto para consumos bajos, lo cual confirma que este valor está penalizando de forma mayor los consumos menores.

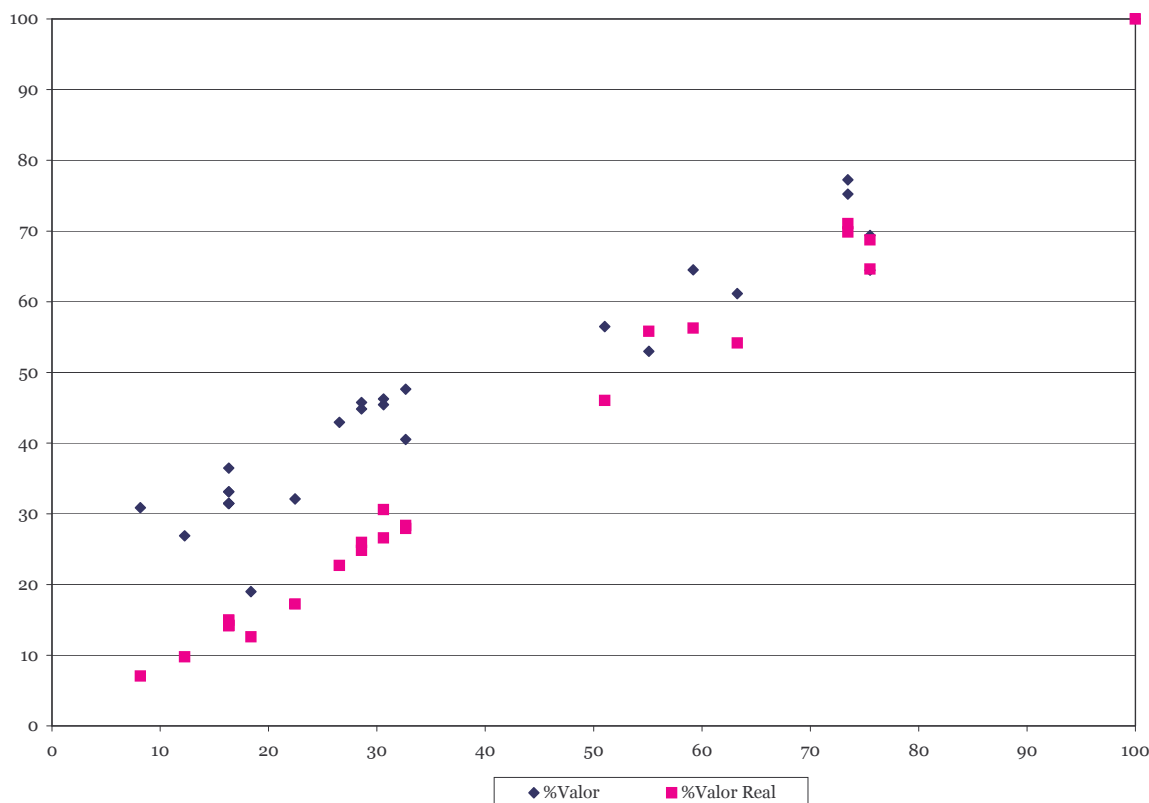


Figura 4. Comparación entre el porcentaje de consumo, el valor cobrado y el valor sin cargo básico

También se puede observar como otra variable que tiene importancia al momento de generar las reglas, es si la lectura corresponde a un valor estimado o si es la lectura del contador. En general las lecturas estimadas corresponden a valores muy inferiores a las lecturas tomadas del contador, por tanto se está penalizando al momento de no tomar la lectura al cliente. En este caso es importante anotar que en todas las facturas se está cobrando un canon de arrendamiento del medidor, por tanto el cliente está en su derecho de que se le cobre por lo que está consumiendo realmente.

En la Figura 6, se presenta el resultado de generación de reglas sin tener en cuenta el costo total de la factura, es decir se está tomando el valor sin incluir el cargo fijo, en este caso la variable que presenta mayor ganancia es el consumo, cómo parece ser más lógico, aunque como se puede observar se siguen penalizando los consumos bajos con respecto a los altos. Para consumos menores de 8 m^3 se tienen valores de casi 14 euros de cargo básico (1.75 euros por m^3) siendo que el valor promedio del agua en Valencia es de 1 euro por m^3 .

Para consumos más altos, superiores a 26 m^3 se obtienen valores de 4.85 euros por m^3 de cargo básico (0.19 euros por m^3).


```

Valor (Euros) < 12.25 [Prom: 4.85, Efecto: -5.978 | (1, 1.0) -> 4.85
Valor (Euros) >= 12.25 [Prom: 11.068, Efecto: +0.239 | (25)
  Consumo (L) < 26 [Prom: 11.885, Efecto: +0.818 | (17)
    Valor (Euros) < 22.275 [Prom: 11.236, Efecto: -0.649 | (18)
      Tipo [Real] [Prom: 13.52, Efecto: +2.284 | (2, 1.0) -> 13.52
      Tipo [Estimada] [Prom: 10.665, Efecto: -0.571 | (8)
        Consumo (L) < 9.5 [Prom: 10.92, Efecto: +0.255 | (6)
          Valor (Euros) < 17.235 [Prom: 10.68, Efecto: -0.24 | (4)
            Consumo (L) < 7 [Prom: 10.26, Efecto: -0.42 | (1, 1.0) -> 10.26
            Consumo (L) >= 7 [Prom: 10.82, Efecto: +0.14 | (3, 1.0) -> 10.82
          Valor (Euros) >= 17.235 [Prom: 11.4, Efecto: +0.48 | (2, 1.0) -> 11.4
            Consumo (L) >= 9.5 [Prom: 9.9, Efecto: -0.765 | (2, 1.0) -> 9.9
        Valor (Euros) >= 22.275 [Prom: 12.813, Efecto: +0.928 | (7)
          Valor real (Euros) < 12.375 [Prom: 13.494, Efecto: +0.681 | (5)
            Consumo (L) < 13.5 [Prom: 13.39, Efecto: -0.104 | (1, 1.0) -> 13.39
            Consumo (L) >= 13.5 [Prom: 13.52, Efecto: +0.026 | (4, 1.0) -> 13.52
          Valor real (Euros) >= 12.375 [Prom: 11.11, Efecto: -1.703 | (2)
            Tipo [Real] [Prom: 10.82, Efecto: -0.29 | (1, 1.0) -> 10.82
            Tipo [Estimada] [Prom: 11.4, Efecto: +0.29 | (1, 1.0) -> 11.4
        Consumo (L) >= 26 [Prom: 9.33, Efecto: -1.738 | (8)
          Valor (Euros) < 34.43 [Prom: 6.767, Efecto: -2.563 | (3)
            Valor real (Euros) < 23.085 [Prom: 9.9, Efecto: +3.133 | (1, 1.0) -> 9.9
            Valor real (Euros) >= 23.085 [Prom: 5.2, Efecto: -1.567 | (2)
              Consumo (L) < 32 [Prom: 4.85, Efecto: -0.35 | (1, 1.0) -> 4.85
              Consumo (L) >= 32 [Prom: 5.55, Efecto: +0.35 | (1, 1.0) -> 5.55
            Valor (Euros) >= 34.43 [Prom: 10.868, Efecto: +1.538 | (5)
              Valor (Euros) < 40.69 [Prom: 10.513, Efecto: -0.355 | (3)
                Consumo (L) < 36.5 [Prom: 10.82, Efecto: +0.307 | (2, 1.0) -> 10.82
                Consumo (L) >= 36.5 [Prom: 9.9, Efecto: -0.613 | (1, 1.0) -> 9.9
              Valor (Euros) >= 40.69 [Prom: 11.4, Efecto: +0.532 | (2, 1.0) -> 11.4

```

Figura 5. Reglas obtenidas para el cargo básico

```

Consumo (L) < 26 [Prom: 11.494, Efecto: +0.666 | (18)
  Valor real (Euros) < 5.62 [Prom: 9.543, Efecto: -1.951 | (3)
    Consumo (L) < 7.5 [Prom: 11.89, Efecto: +2.347 | (23)
      Tipo [Real] [Prom: 13.52, Efecto: +1.63 | (1, 1.0) -> 13.52
      Tipo [Estimada] [Prom: 10.26, Efecto: -1.63 | (1, 1.0) -> 10.26
    Consumo (L) >= 7.5 [Prom: 4.85, Efecto: -4.693 | (1, 1.0) -> 4.85
  Valor real (Euros) >= 5.62 [Prom: 11.885, Efecto: +0.39 | (15)
    Consumo (L) < 12 [Prom: 11.24, Efecto: -0.645 | (7)
      Tipo [Real] [Prom: 13.52, Efecto: +2.28 | (1, 1.0) -> 13.52
      Tipo [Estimada] [Prom: 10.86, Efecto: -0.38 | (6)
        Consumo (L) < 9.5 [Prom: 11.052, Efecto: +0.192 | (5)
          Valor real (Euros) < 6.125 [Prom: 10.82, Efecto: -0.232 | (3, 1.0) -> 10.82
          Valor real (Euros) >= 6.125 [Prom: 11.4, Efecto: +0.348 | (2, 1.0) -> 11.4
        Consumo (L) >= 9.5 [Prom: 9.9, Efecto: -0.96 | (1, 1.0) -> 9.9
    Consumo (L) >= 12 [Prom: 12.449, Efecto: +0.564 | (8)
      Valor real (Euros) < 11.45 [Prom: 13.488, Efecto: +1.039 | (4)
        Consumo (L) < 13.5 [Prom: 13.39, Efecto: -0.098 | (1, 1.0) -> 13.39
        Consumo (L) >= 13.5 [Prom: 13.52, Efecto: +0.032 | (3, 1.0) -> 13.52
      Valor real (Euros) >= 11.45 [Prom: 11.41, Efecto: -1.039 | (4)
        Valor real (Euros) < 11.815 [Prom: 9.9, Efecto: -1.51 | (1, 1.0) -> 9.9
        Valor real (Euros) >= 11.815 [Prom: 11.913, Efecto: +0.503 | (3)
          Valor real (Euros) < 12.375 [Prom: 13.52, Efecto: +1.607 | (1, 1.0) -> 13.52
          Valor real (Euros) >= 12.375 [Prom: 11.11, Efecto: -0.803 | (2)
            Tipo [Real] [Prom: 10.82, Efecto: -0.29 | (1, 1.0) -> 10.82
            Tipo [Estimada] [Prom: 11.4, Efecto: +0.29 | (1, 1.0) -> 11.4
    Consumo (L) >= 26 [Prom: 9.33, Efecto: -1.498 | (8)
      Consumo (L) < 28 [Prom: 4.85, Efecto: -4.48 | (1, 1.0) -> 4.85
      Consumo (L) >= 28 [Prom: 9.97, Efecto: +0.64 | (7)
        Valor real (Euros) < 29.1 [Prom: 9.043, Efecto: -0.928 | (4)
          Valor real (Euros) < 28 [Prom: 10.207, Efecto: +1.164 | (3)
            Consumo (L) < 30 [Prom: 10.82, Efecto: +0.613 | (1, 1.0) -> 10.82
            Consumo (L) >= 30 [Prom: 9.9, Efecto: -0.307 | (2, 1.0) -> 9.9
          Valor real (Euros) >= 28 [Prom: 5.55, Efecto: -3.493 | (1, 1.0) -> 5.55
        Valor real (Euros) >= 29.1 [Prom: 11.207, Efecto: +1.237 | (3)
          Valor real (Euros) < 29.58 [Prom: 10.82, Efecto: -0.387 | (1, 1.0) -> 10.82
          Valor real (Euros) >= 29.58 [Prom: 11.4, Efecto: +0.193 | (2, 1.0) -> 11.4

```

Figura 6. Reglas obtenidas para el cargo básico

Influencia de variables socioeconómicas en la Gestión de un abastecimiento

El ejercicio que se ha realizado como aplicación de la inducción de reglas, consiste en una serie de datos de encuestas realizadas en Argentina en el municipio de Fontana Provincia del Chaco.

Se cuenta con un registro de consumos mensuales medidos por la empresa suministradora de agua en el municipio, 2229 usuarios con registros no continuos desde el año 1993 al año 2003. Además se realizaron 950 encuestas directas a residentes para obtener información que pueda influir en el funcionamiento del sistema de abastecimiento, tales como, la ubicación de la vivienda, el número de habitantes por vivienda, el nivel de escolarización, los ingresos mensuales del núcleo familiar y el tipo de construcción cuantificada en la superficie total y cuanto de esta está cubierta y cuanto no lo está.

Para desarrollar el ejercicio se utilizó el programa Clementine 6.0 y el nodo árbol de clasificación y regresión C&R que corresponde a un método de clasificación y predicción basado en árboles. El modelo trabaja dividiendo la muestra basándose en el campo que de mayor ganancia de información. Cada sub-muestra definida por la primera partición es dividida de nuevo, generalmente basándose en un campo diferente y el proceso se repite hasta que las sub-muestras obtenidas no puedan ser divididas. Finalmente se revisa el nivel más bajo del árbol obtenido, y aquellos que no contribuyen significativamente al valor del modelo son removidos o podados. Es importante mencionar que este modelo es bastante robusto a datos faltantes y a grandes números de campos.

El primer paso consistió en realizar una selección de datos ya que la cantidad de información proveniente de la empresa suministradora de agua en cuanto a consumos es mayor que el número de viviendas encuestadas, así que se realizó un cotejamiento de la información para asignarle a cada vivienda encuestada su registro de consumos mensuales.

Sobre esta selección de datos se realizó un preprocesamiento básicamente para descartar datos que podrían considerarse espurios, los objetos que tienen falta de alguno de sus atributos fueron descartadas, al igual que los datos de consumos con valor faltante o cero. Así resultó un conjunto de 200 registros aproximadamente.

```
CantPers < 3.5 [Prom: 18.348, Efecto: +0.91 ] (23)
  CantPers < 2.5 [Prom: 16.871, Efecto: -1.477 ] (9)
    CantPers < 1.5 [Prom: 18.11, Efecto: +1.239 ] (3, 1.0) -> 18.11
    CantPers >= 1.5 [Prom: 16.252, Efecto: -0.619 ] (6, 1.0) -> 16.252
  CantPers >= 2.5 [Prom: 19.298, Efecto: +0.95 ] (14, 1.0) -> 19.298
CantPers >= 3.5 [Prom: 17.064, Efecto: -0.374 ] (86)
  CantPers < 5.5 [Prom: 16.658, Efecto: -0.407 ] (38)
    CantPers < 4.5 [Prom: 17.218, Efecto: +0.561 ] (24, 1.0) -> 17.218
    CantPers >= 4.5 [Prom: 15.696, Efecto: -0.961 ] (14, 1.0) -> 15.696
  CantPers >= 5.5 [Prom: 17.923, Efecto: +0.858 ] (18)
    CantPers < 6.5 [Prom: 18.68, Efecto: +0.757 ] (14, 1.0) -> 18.68
    CantPers >= 6.5 [Prom: 15.273, Efecto: -2.65 ] (4)
      CantPers < 8.5 [Prom: 15.875, Efecto: +0.603 ] (2)
        CantPers < 7.5 [Prom: 16.08, Efecto: +0.205 ] (1, 1.0) -> 16.08
        CantPers >= 7.5 [Prom: 15.67, Efecto: -0.205 ] (1, 1.0) -> 15.67
      CantPers >= 8.5 [Prom: 14.67, Efecto: -0.603 ] (2, 1.0) -> 14.67
```

Figura 7. Clasificación de la demanda de acuerdo al número de habitantes por vivienda

Un primer ejercicio consistió en clasificar las demandas de agua de acuerdo a la al número de habitantes por vivienda, tal como se muestra en la Figura 7, cuyos resultados muestran como está es mayor cuanto menos número de habitantes se tienen por vivienda.

En segundo lugar, se planteó utilizar el nivel de escolaridad de la cabeza de unidad familiar para ver como es el patrón de comportamiento relacionado con el consumo. Tal como puede observarse en la Figura 8 se nota cierta tendencia a niveles de escolarización menores mayor consumo promedio.

```

scolaridad ['Terciaria Completa'] [Prom: 8.967, Efecto: -1.585 ] -> 8.967
scolaridad ['Terciaria Incompleta' 'Primaria Incompleta' 'Secundaria Incompleta' 'Primaria Completa' 'Secundaria Completa'] [Prom: 10.762, |
Escolaridad ['Secundaria Completa'] [Prom: 10.319, Efecto: -0.443 ] -> 10.319
Escolaridad ['Terciaria Incompleta' 'Primaria Incompleta' 'Secundaria Incompleta' 'Primaria Completa'] [Prom: 11.074, Efecto: +0.313 ]
Escolaridad ['Secundaria Incompleta' 'Primaria Completa'] [Prom: 10.967, Efecto: -0.108 ]
Escolaridad ['Primaria Completa'] [Prom: 10.874, Efecto: -0.093 ] -> 10.874
Escolaridad ['Secundaria Incompleta'] [Prom: 11.093, Efecto: +0.126 ] -> 11.093
Escolaridad ['Terciaria Incompleta' 'Primaria Incompleta'] [Prom: 11.612, Efecto: +0.538 ]
Escolaridad ['Primaria Incompleta'] [Prom: 11.57, Efecto: -0.042 ] -> 11.57
Escolaridad ['Terciaria Incompleta'] [Prom: 11.75, Efecto: +0.138 ] -> 11.75

```

Figura 8. Reglas obtenidas para el nivel de escolarización

Como segundo ejercicio se utilizó el nivel de ingresos medio mensual por unidad familiar, en este caso puede observarse tal como se muestra en la Figura 9 que ingresos mensuales bajos presentan un consumo promedio un poco más alto.

```

Ingreso Mensual < 425 [Prom: 10.36, Efecto: -0.773 ] (119)
Ingreso Mensual < 362.5 [Prom: 10.812, Efecto: +0.452 ] (96)
Ingreso Mensual < 325 [Prom: 10.74, Efecto: -0.072 ] (92)
Ingreso Mensual < 290 [Prom: 11.018, Efecto: +0.278 ] (70)
Ingreso Mensual < 275 [Prom: 10.585, Efecto: -0.433 ] (62, 1.0) -> 10.585
Ingreso Mensual >= 275 [Prom: 14.376, Efecto: +3.358 ] (8, 1.0) -> 14.376
Ingreso Mensual >= 290 [Prom: 9.856, Efecto: -0.884 ] (22, 1.0) -> 9.856
Ingreso Mensual >= 325 [Prom: 12.457, Efecto: +1.646 ] (4, 1.0) -> 12.457
Ingreso Mensual >= 362.5 [Prom: 8.472, Efecto: -1.888 ] (23)
Ingreso Mensual < 395 [Prom: 5.875, Efecto: -2.597 ] (4, 1.0) -> 5.875
Ingreso Mensual >= 395 [Prom: 9.018, Efecto: +0.547 ] (19, 1.0) -> 9.018
Ingreso Mensual >= 425 [Prom: 12.343, Efecto: +1.21 ] (76)
Ingreso Mensual < 1075 [Prom: 12.641, Efecto: +0.298 ] (69, 1.0) -> 12.641
Ingreso Mensual >= 1075 [Prom: 9.406, Efecto: -2.937 ] (7)
Ingreso Mensual < 1462.5 [Prom: 6.75, Efecto: -2.656 ] (4, 1.0) -> 6.75
Ingreso Mensual >= 1462.5 [Prom: 12.947, Efecto: +3.541 ] (3, 1.0) -> 12.947

```

Figura 9. Reglas obtenidas para el ingreso mensual promedio por unidad familiar

En cuanto al tipo de construcción, cuantificada en la superficie cubierta de unidad familiar o vivienda, se aprecia que a mayor superficie cubierta es mayor el consumo, tal como se aprecia en la Figura 10.

```

SupCub < 45 [Prom: 6.681, Efecto: -3.969 ] (17, 1.0) -> 6.681
SupCub >= 45 [Prom: 11.016, Efecto: +0.367 ] (184, 1.0) -> 11.016

```

Figura 10. Reglas obtenidas para el tipo de construcción

En la Figura 11, se presentan un conjunto de reglas obtenidas teniendo en cuenta todas las variables mencionadas anteriormente, encontrándose que la variable con mayor significancia corresponde a la superficie del predio.

```

Sup Predio < 350 [Prom: 10.563, Efecto: -0.166 | 383]
Ingreso Mensual < 425 [Prom: 9.608, Efecto: -0.955 | 377]
SupCub < 45 [Prom: 4.783, Efecto: -4.825 | 434]
  Escolaridad ['Primaria Completa' N/E] [Prom: 5.518, Efecto: +0.735 | 473]
  Superf Libre < 167.5 [Prom: 5.16, Efecto: -0.353 | 424, 1.0] -> 5.16
  Superf Libre >= 167.5 [Prom: 6.473, Efecto: +0.955 | 42, 1.0] -> 6.473
  Escolaridad ['Terciaria Completa' 'Secundaria Completa'] [Prom: 3.571, Efecto: -1.212 | 428, 1.0] -> 3.571
SupCub >= 45 [Prom: 10.398, Efecto: +0.789 | 434]
  Sup Predio < 287.5 [Prom: 11.875, Efecto: +1.478 | 428]
  Sup Predio < 265.5 [Prom: 10.925, Efecto: -0.95 | 432]
  Sup Predio < 110 [Prom: 21.75, Efecto: +10.825 | 45, 1.0] -> 21.75
  Sup Predio >= 110 [Prom: 10.392, Efecto: -0.532 | 433]
  Superf Libre < 155.5 [Prom: 9.445, Efecto: -0.947 | 425, 1.0] -> 9.445
  Superf Libre >= 155.5 [Prom: 14.006, Efecto: +3.613 | 438, 1.0] -> 14.006
  Sup Predio >= 265.5 [Prom: 25.911, Efecto: +14.036 | 413, 1.0] -> 25.911
  Sup Predio >= 287.5 [Prom: 7.852, Efecto: -2.545 | 433]
  Escolaridad ['Secundaria Incompleta' 'Primaria Completa'] [Prom: 9.931, Efecto: +2.079 | 431]
  Ingreso Mensual < 230 [Prom: 8.696, Efecto: -1.236 | 424, 1.0] -> 8.696
  Ingreso Mensual >= 230 [Prom: 12.402, Efecto: +2.471 | 434, 1.0] -> 12.402
  Escolaridad ['Terciaria Completa' 'Secundaria Completa' 'Primaria Incompleta'] [Prom: 6.293, Efecto: -1.559 | 428, 1.0] -> 6.293
Ingreso Mensual >= 425 [Prom: 11.496, Efecto: +0.933 | 434]
  Escolaridad ['Terciaria Completa' 'Secundaria Incompleta' 'Secundaria Completa' N/E] [Prom: 10.833, Efecto: -0.663 | 427]
  SupCub < 130 [Prom: 10.637, Efecto: -0.195 | 433]
  Escolaridad ['Secundaria Incompleta' 'Secundaria Completa' N/E] [Prom: 10.832, Efecto: +0.194 | 428]
  Superf Libre < 203.5 [Prom: 10.623, Efecto: -0.208 | 433]
  Ingreso Mensual < 725 [Prom: 10.389, Efecto: -0.235 | 422, 1.0] -> 10.389
  Ingreso Mensual >= 725 [Prom: 12.294, Efecto: +1.671 | 434, 1.0] -> 12.294
  Superf Libre >= 203.5 [Prom: 15.625, Efecto: +4.793 | 412, 1.0] -> 15.625
  Escolaridad ['Terciaria Completa'] [Prom: 8.564, Efecto: -2.073 | 427]
  SupCub < 85 [Prom: 6.11, Efecto: -2.454 | 428, 1.0] -> 6.11
  SupCub >= 85 [Prom: 13.473, Efecto: +4.909 | 435, 1.0] -> 13.473
  SupCub >= 130 [Prom: 15.957, Efecto: +5.125 | 435, 1.0] -> 15.957
  Escolaridad ['Terciaria Incompleta' 'Primaria Incompleta' 'Primaria Completa'] [Prom: 15.171, Efecto: +3.675 | 429]
  Superf Libre < 203.5 [Prom: 13.608, Efecto: -1.562 | 425, 1.0] -> 13.608
  Superf Libre >= 203.5 [Prom: 18.218, Efecto: +3.047 | 428, 1.0] -> 18.218
Sup Predio >= 350 [Prom: 16.023, Efecto: +5.293 | 424]
  Escolaridad ['Secundaria Completa' 'Primaria Incompleta'] [Prom: 23.113, Efecto: +7.091 | 429, 1.0] -> 23.113
  Escolaridad ['Terciaria Completa' 'Secundaria Incompleta' 'Primaria Completa'] [Prom: 11.768, Efecto: -4.255 | 416, 1.0] -> 11.768

```

Figura 11. Reglas obtenidas para el conjunto de todas las variables

Aquí se puede observar que básicamente, lo obtenido a escala individual para cada variable se refleja de forma similar al realizar el análisis conjunto, y esta es una de las ventajas de utilizar los algoritmos de la generación de las reglas de decisión, ya que son bastante robustos tanto a datos faltantes como a datos erróneos. En este caso al hacer el análisis conjunto no se encontró información de todas las variables para cada uno de los predios, pero el algoritmo permite trabajar estos registros incompletos sin problema. La variable que presenta mayor ganancia es el tamaño del predio, notándose que para superficies de predio entre 155 m² y 290 m² e ingresos mensuales menores de 25 es donde se presenta el mayor consumo.

CONCLUSIONES

Se han presentado un par de aplicaciones de la Minería de Datos, en la gestión de los sistemas de abastecimiento de agua teniendo en cuenta la relación o la afectación que tienen los clientes al considerar la influencia que tiene el cargo fijo sobre el valor total a pagar en la factura y otro que tiene en cuenta factores socioeconómicos en los patrones de consumo.

Los dos ejercicios son importantes al intentar establecer relaciones con los clientes, ya que en el primero de ellos se puede concluir que no se incentiva el ahorro de agua y en el segundo se puede actuar para mejorar los hábitos de consumo.

Se quiere hacer énfasis en que los resultados mostrados en este trabajo no son decisivos o influyentes científicamente, pero se pretende generar la inquietud hacia la utilización de las herramientas de la minería de datos, ver que la información de la que se dispone nos puede desvelar conocimiento más allá del que se percibe a simple vista.

AGRADECIMIENTOS

El presente trabajo se ha realizado dentro del proyecto de investigación titulado “SIBA: Aplicación de técnicas de minería de datos en la planificación, operación, mantenimiento y gestión de sistemas de abastecimiento de agua” (referencia DPI 2004-04330), financiado por el Ministerio de Educación y Ciencia de España.

REFERENCIAS BIBLIOGRÁFICAS

- An, A.; Shan, N.; Chan, C.; Cercone, N.; Ziarko, W. 1997, “*Applying knowledge discovery to predict water-supply consumption*”, IEEE Intelligent Systems & Their Applications, Volume 12, Number 4, pages 72-78.
- Babovic, V.; Drècourt, J.; Keijzer, M.; Hansen, P. 2001, “*Modelling of Water Supply Assets: A Data Mining Approach*”, D2K Technical Report 1000-1, 2000, February 6, 2001.
- Buchheit, R.B.; Garrett, J.H. JR; Lee, S.R.; Brahme, R. 2000, “*A Knowledge Discovery Framework for City Civil Infrastructure: A Case Study of the Intelligent Workplace*”, Engineering with Computers, Number 16, Pages 264-274, 2000.
- Claude, M. (2005). “*La silenciosa guerra por el agua*”, Periódico El Tiempo. http://eltiempo.terra.com.co/opinion/colopi_new/columnas_del_dia/ARTICULO-WEB-NOTA_INTERIOR-2556936.html, recuperado el 4 de octubre de 2005.
- Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (Editors). 1996. “Advances in Knowledge Discovery and Data Mining”, AAAI Press / The MIT Press, 611pp.
- Fleming, N.S. 1994, “Forecasting water consumption in the Northern Adelaide Plains, SA, using artificial neural networks, regression and time series methods”, Research Report No. R115, March 1994, The University of Adelaide, Department of Civil and Environmental Engineering, 64pp.
- Jain, A.; Ormsbee, L.E., 2002, “Short-Term Water Demand Forecast Modelling Techniques – Conventional Methods Versus AI”, Journal AWWA, Volume 94, Number 7, July 2002, pages 64-72.
- Jain, A.; Varshney, A.K.; Joshi, U.C. 2001, “Short-Term Water Demand Forecast Modelling at IIT Kanpur Using Artificial Neural Networks”, Water Resources Management, Volume 15, pages 299-321.
- “*Lealtad, el verdadero tesoro*”, (s.f.). Revista Dinero, ed. 239. Recuperado el 5 de octubre de 2005, de <http://www.dinero.com/dinero/ArticuloView.jsp?id=22101>.
- Savic, D.A.; Davidson, J.W.; Davis, R.B. (1999) “*Data Mining and Knowledge discovery for the water industry*”, Water Industry Systems, modelling and optimisation applications”, Volume 2, Edited by Dragan A. Savic and Godfrey A. Walters, Research Studies Press Ltd, August, 1999.
- Stark, H.L.; Stanley, S.J.; Buchanan, I.D., 1999, “Water Demand Forecasting Using Artificial Neural Networks”, WWA and OMWA Joint Annual Conference, Ottawa, Ont., Canada., May 16-19,1999, 10pp.

“Tema 6: Alimentos y Agua para una población creciente. Agua en España”. Libro Electrónico Ciencias de la Tierra y El Medio Ambiente. Recuperado el 26 de octubre de 2005, <http://www.esi.unav.es/asignaturas/ecologia/Hipertexto/06Recursos/110AguaEsp.htm>.

“Una nueva forma de administrar”, (s.f). Revista Dinero, ed. 239. Recuperado el 5 de octubre de 2005, de <http://www.dinero.com/dinero/ArticuloView.jsp?id=22100>.

