TÉCNICAS DE MINERÍA DE DATOS COMO SOPORTE PARA LA GESTIÓN DE UN SISTEMA DE COMERCIALIZACIÓN DE ENERGÍA ELÉCTRICA

GESTIÓN DE UN SISTEMA DE COMERCIALIZACIÓN DE ENERGÍA ELÉCTRICA

AUTORES: Jorge Iván Pincay Ponce[1]
Navira Gissela Angulo Murillo[2]
Jorge Sergio Herrera Tapia[3]
Wilian Richart Delgado Muentes[4]

DIRECCIÓN PARA CORRESPONDENCIA: jorge.pincay@uleam.edu.ec

Fecha de recepción: 23-04-2020

Fecha de aceptación: 11-06-2020

RESUMEN

Tener un suministro de energía eléctrica suficiente es vital para la comunidad, lo que demanda mantenimiento y mejora continua del servicio por parte de las compañías prestatarias del servicio. Entre otros aspectos, estas compañías mantienen bases de datos que capturan el consumo de la energía y en tal sentido en la presente investigación se propone el uso de técnicas de Redes Neuronales Artificiales y Reglas de Asociación como soporte a la gestión del sistema de comercialización de la energía eléctrica en una empresa pública de la ciudad de Manta, a partir de una muestra de datos extraídos de las facturas de consumo residencial correspondientes al año 2015. Los algoritmos usados específicamente fueron el perceptrón multicapa a nivel de redes neuronales y PART como regla de asociación. En esta aplicación empírica de minería de datos, se demostró que las redes neuronales y reglas de asociación son alternativas viables para predecir los niveles de consumo y comprender los patrones de consumo de energía.

PALABRAS CLAVE: Redes neuronales; Reglas de Asociación; Datamining; minería de dato; WEKA; Consumo de energía eléctrica.

TECHNIQUES OF DATA MINING AS SUPPORT FOR THE MANAGEMENT OF A SYSTEM OF ELECTRIC ENERGY COMMERCIALIZATION

[1] Ingeniero en Sistemas por la Universidad Laica Eloy Alfaro de Manabí, Máster Universitario en Ingeniería de Software para la Web por la Universidad de Alcalá – España. Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador.
[2] Ingeniera en Sistemas por la Universidad Laica Eloy Alfaro de Manabí, Máster en Dirección Estratégica de las Tecnologías de la Información y Comunicación por la Universidad Nacional de Piura – Perú. Coordinadora de Planificación Estratégica y Operativa de la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. E-mail: navira.angulo@live.uleam.edu.ec
[3] Ingeniero en Sistemas Computacionales por la Universidad Técnica del Norte, Doctor en Informática por la Universidad Politécnica de Valencia – España. Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. E-mail: jorge.herrera@live.uleam.edu.ec
[4] Ingeniero en Sistemas Computacionales por la Universidad Técnica de Manabí, Máster en Informática de gestión y nuevas tecnologías por la Universidad Técnica Santa María – Chile. Docente titular en la carrera de Ingeniería en Sistemas de la Universidad Laica Eloy Alfaro de Manabí. Manta, Manabí, Ecuador. E-mail: wilian.delgado@live.uleam.edu.ec

ABSTRACT

Having a sufficient electrical power supply is vital for the community, which demands maintenance and continuous improvement of the service by the service companies. Among other aspects, these companies maintain data bases that capture energy consumption and in this sense the present research proposes the use of Artificial Neural Network techniques and Association Rules as support for the management of the marketing system of the electric power in a public company of Manta city, based on a sample of data extracted from residential consumption bills for the year 2015. The algorithms used specifically were the multilayer perceptron at the level of neural networks and PART as a rule of association. In this empirical application of data mining, it was shown that neural networks and association rules are viable alternatives to predict consumption levels and to understand energy consumption patterns.

KEYWORDS: Neural Networks; Association Rules; Datamining; Data Mining; WEKA; Electric Power Consumption.

INTRODUCCIÓN

La demanda de electricidad ha aumentado continuamente a lo largo de las últimas décadas, así como la atención que se presta a este consumo y sus impactos ambientales. Dicho crecimiento de consumo a nivel de los hogares y de la industria, en parte es motivado por el auge de dispositivos y servicios de tecnologías de la información y de la comunicación, así como el propio crecimiento demográfico, unido a factores que se pueden considerar tradicionales como lo son el clima, nivel socioeconómico, hábitos de consumo, entre otros (Ariza Ramírez, 2013, p. 17; Gönen, 1986; Van Heddeghem et al., 2014).

En los últimos años han habido esfuerzos por desarrollar métodos más precisos, confiables y computacionalmente eficientes para el pronóstico de la demanda de electricidad, pudiéndose clasificar por métodos matemáticos, estadísticos, de horizonte temporal y de Inteligencia Artificial (AI); dentro de la AI, tal como se lo ha hecho en el presente estudio, se han empleado técnicas de minerías de datos como lo son las Redes Neuronales Artificiales y las Reglas de Asociación (Ariza Ramírez, 2013, p. 32; Xiao & Fan, 2014a, p. 111); pese a esto, existen estudios que demuestran la existencia de brechas entre las investigaciones realizadas y la dirección de futuras investigaciones basadas en datos (Amasyali & El-Gohary, 2018a, p. 1193).

Las redes neuronales artificiales, son una de las técnicas más usadas en la predicción de consumos eléctricos (Amasyali & El-Gohary, 2018b) y consisten en un modelo computacional no lineal, inspirado en el cerebro humano. Generalmente incluyen tres capas secuenciales: la capa de entrada, la capa oculta y la capa de salida; cada capa tiene una cantidad de neuronas interconectadas, y cada neurona tiene una función de activación.

Normalmente, se utilizan tres tipos de parámetros para definir las redes neuronales: el patrón de interconexión entre las neuronas de las diferentes capas, el proceso de aprendizaje para actualizar los pesos de las interconexiones, y la función de activación que convierte la entrada ponderada de una neurona en su activación de salida (Wang & Srinivasan, 2015, p. 3340). En las redes neuronales, cada característica como por ejemplo el mes a pagar o el total a pagar se multiplica por su peso neuronal correspondiente y se resume con el sesgo. La función de activación se aplica para determinar la salida, por ejemplo, el mes en el que se paga.

Entre los tipos de redes neuronales artificiales se incluyen la propagación hacia atrás (BPNN), la función radial (RBFNN), la regresión general (GRNN), feed forward (FFNN), sistemas adaptativos de inferencia neuro - difusa (ANFIS), la mezcla jerárquica de expertos (HME), fuzzy c-means (FCC) y el perceptrón multicapa (MLP) (Amasyali & El-Gohary, 2018a, p. 1193). En esta investigación se emplea el perceptron multicapa, que es una clase de red neuronal de alimentación anticipada que consiste en al menos tres capas de nodos. Excepto por los nodos de entrada, cada nodo es una neurona que usa una función de activación no lineal. Un MLP utiliza una técnica de aprendizaje supervisada llamada backpropagation para el entrenamiento. Sus capas múltiples y activación no lineal distinguen al MLP de un perceptrón lineal (Rosenblatt, 1961; Rumelhart, Hinton, & Williams, 1985).

En esta investigación se empleó también reglas de asociación, que es un proceso de aprendizaje no supervisado, típicamente relacionado con el análisis de cesta de mercado, pero también aplicado en bioinformática y la sociología (Xiao & Fan, 2014b, 2014a, p. 112), estas buscan descubrir todas las reglas que satisfagan el mínimo especificado por el usuario mediante la denominada confianza mínima. Es necesario reconocer algunos conceptos relativos a las reglas de asociación, que de acuerdo con (Xiao & Fan, 2014a, p. 112) se resumen en:

- El soporte de una regla, que es la articulación de la probabilidad del antecedente y el consecuente.
- La confianza, que es la probabilidad condicional del consecuente, dado el antecedente.
- El soporte y la confianza se utilizan normalmente para determinar si la regla es estadísticamente significativa o no.
- El levantamiento, que es una medida de dependencia y correlación entre el antecedente y el consecuente. Si el levantamiento es igual a 1, indica que el antecedente y el consecuente son independientes entre sí, y, por lo tanto, lo descubierto tiene poco valor. Un levantamiento mayor que 1 indica una correlación positiva, lo que significa que la probabilidad del consecuente es positivamente afectada por la ocurrencia del antecedente.

El pronóstico de la demanda de electricidad es una herramienta fundamental para la toma de decisiones operativas y estratégicas en las empresas eléctricas, cuya falta de precisión puede traer altos costos económicos (Ariza Ramírez, 2013). La problemática de la demanda y el consumo eléctrico tiene variadas connotaciones como las indicadas en la introducción de este documento. En esta investigación se emplearon técnicas de minería de datos típicas como lo son las redes neuronales y las reglas de asociación, para soporte en la gestión de comercialización de energía eléctrica. Con las redes neuronales se predijo los momentos de mayor demanda y con las reglas de asociación se encontró patrones de estos datos en la muestra de 1200 registros resultantes, sobre los cuales también es posible incorporar otras técnicas de minería de datos de clasificación, regresión, segmentación, asociación y análisis de secuencia.

Resulta notable la importancia de conocer a corto, mediano y largo plazo el crecimiento de la demanda de energía eléctrica, de una manera segura, confiable y cercana a la realidad. Para esto se requiere que las técnicas nombradas sean validadas teniendo siempre en cuenta que todas tienen ventajas y desventajas que hay que reconocer, ya que de esto depende garantizar el suministro de la energía eléctrica.

METODOLOGÍA

La aplicación de técnicas de minería de datos implica procesos que van desde la formulación de preguntas acerca de los datos hasta la creación e implementación de modelos empleables en diversos contextos. Un proceso Uno de los más documentados procesos para esta finalidad, es el de Microsoft, que se sigue en esta investigación y se detalla en las siguientes subsecciones, este proceso incluye seis pasos: definición del problema, preparación de los datos, exploración de los datos, generación, validación e implementación de los modelos (Microsoft, 2018). Investigación documental, basada en un estudio de campo.

Los datos facilitados por empresa pública de energía eléctrica, en adelante referida simplemente como empresa eléctrica, estaban alojados en una base de datos MySQL. Estos datos fueron analizados con el software de aprendizaje automático y de minería de datos WEKA (Waikato Environment for Knowledge Analysis) versión 3.8.1, Java Virtual Machine (JVM) versión 1.8, Java versión 9 y el conector MySQL - WEKA Connector/J versión 8.0.

Paso 1: Definición del problema

Conocido el esquema relacional de la base de datos de empresa eléctrica, los datos y tipos con los que se contaba, se procedió a la construcción de un archivo ARFF (Attribute Relation File Format) con datos extraídos a partir de la base de datos. Dado estos antecedentes, se definió el problema que fue abordado mediante predicciones y clasificaciones elaboradas empleando redes neuronales y reglas de asociación.

- Día del Mes: Se predice el día del mes en que más se generan cobros a partir de los datos reales en el archivo ARFF. Con respecto a los días de más cobro es de considerar que la mayoría ocurren luego del periodo de finalización de brindar el servicio eléctrico, guardado en el atributo “hasta”. Para que el modelo sea capaz de identificar el perfil de carga asociado a cada día del mes, se calculó e incluyó el atributo “DiaDelMes”.
- Día de la semana: Se predice el día de la semana en que más se generan cobros a partir de los datos reales en el archivo ARFF. Con respecto a los días de más cobro es de considerar que la mayoría ocurren en los días laborales de la semana, es decir de lunes a viernes, salvo casos donde empresa eléctrica atienda fines de semana, como por ejemplo en los denominados sábados de recuperación de feriados. Para que el modelo sea capaz de identificar el perfil de carga asociado a cada día de la semana, se incluyó una variable categórica que recoja este dato así: lunes = “1”, martes = “2”, … viernes = “5”.
- Mes del año: Se predice el mes del año en que más se generan cobros a partir de los datos reales que se tenga en el archivo ARFF. Para que el modelo sea capaz de identificar el perfil de carga asociado a cada día del mes, se incluyó una variable que recoja este número del día en el atributo “DiaDelMes”.
- Estaciones: La demanda diaria de energía varía significativamente entre el invierno y el verano, en los meses de invierno al margen de los aumentos de temperatura se evidencia más consumo por el mayor uso de dispositivos como ventiladores o acondicionadores de aire, así como por abarcar los meses de vacaciones en instituciones educativas de la costa de Ecuador. Esta variable categórica, a la que se la denominó estación, fue codificada considerando al invierno como los meses de enero a mayo y a los restantes meses se los consideró verano.

Adicionalmente se pueden extraer resultados por orden geográfico, pues se tiene los datos de ubicación de los clientes. Si bien los resultados esperados se han expresado en términos de consumo por Kilowatt (kW), también es posible expresarlos en términos monetarios, pues el valor a recaudar es proporcional al consumo del cliente, salvo en caso de beneficios concretos de ciertos clientes residenciales y que no se consideraron en esta investigación.

Paso 2: Preparación de datos

La base de datos de empresa eléctrica es extensa, pero para fines de esta investigación se ilustran las tablas que tienen campos implicados en las necesidades específicas del estudio (ver Ilustración 1). El archivo ARFF resultante, mencionado en el Paso 1, cuenta con 1200 registros aleatorios que corresponden a consumos eléctricos del sector residencial en el año 2015. Los datos en el archivo ARFF no se necesitaron limpiar, tampoco fue necesario agregar nuevos registros, pero si convertir ciertos tipos a texto o nominal y decimal o numeric, que son los tipos de datos que soporta WEKA (Witten, Frank, Hall, & Pal, 2016), esta conversión se hace empleando la función SQL CAST al momento de introducir la consulta SQL en el SQL Viewer de WEKA (ver Ilustración 2). Luego, los resultados se guardaron como archivo ARFF y en adelante fue posible operarlos y visualizarlos con la utilidad ARFF Viewer de WEKA tal como se muestra en la Ilustración 3.

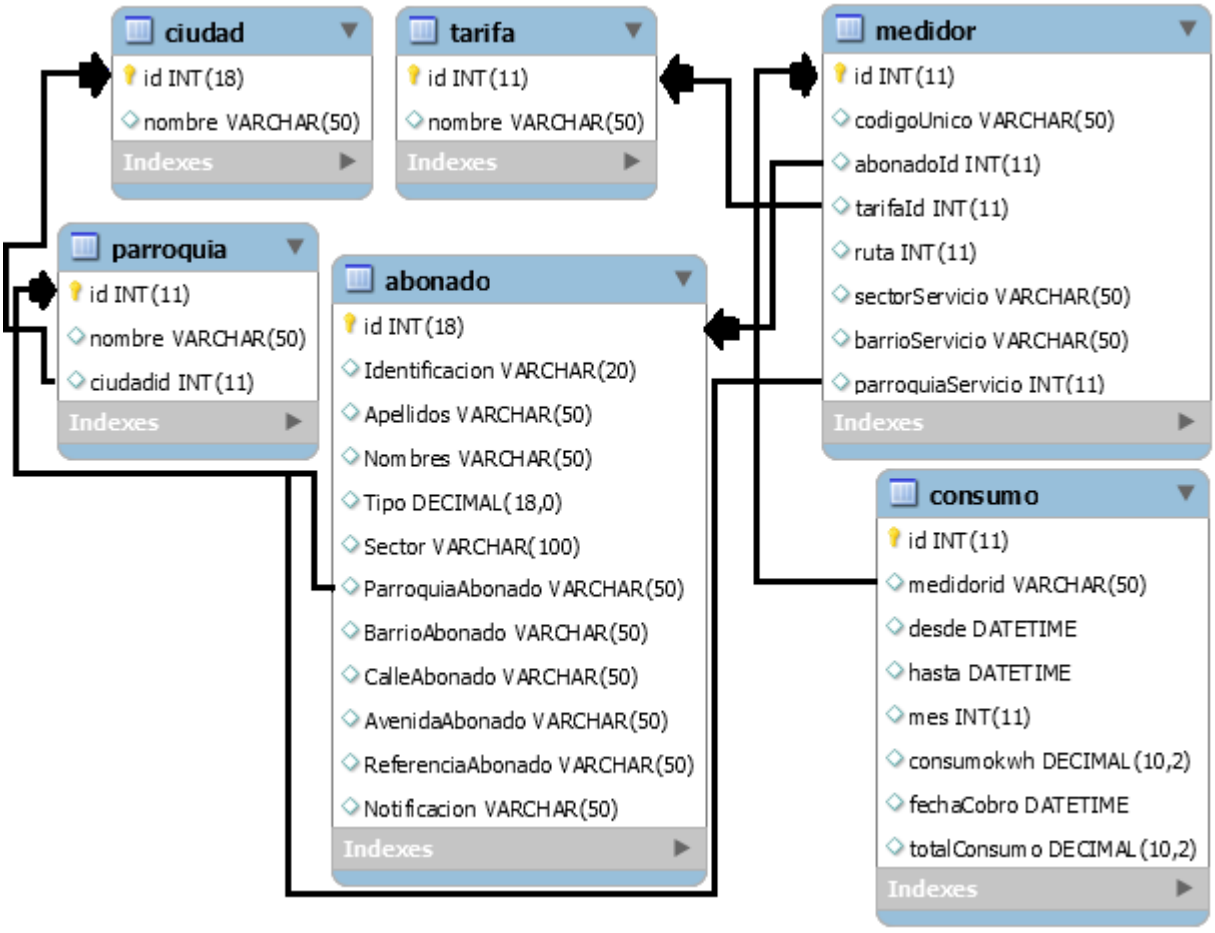
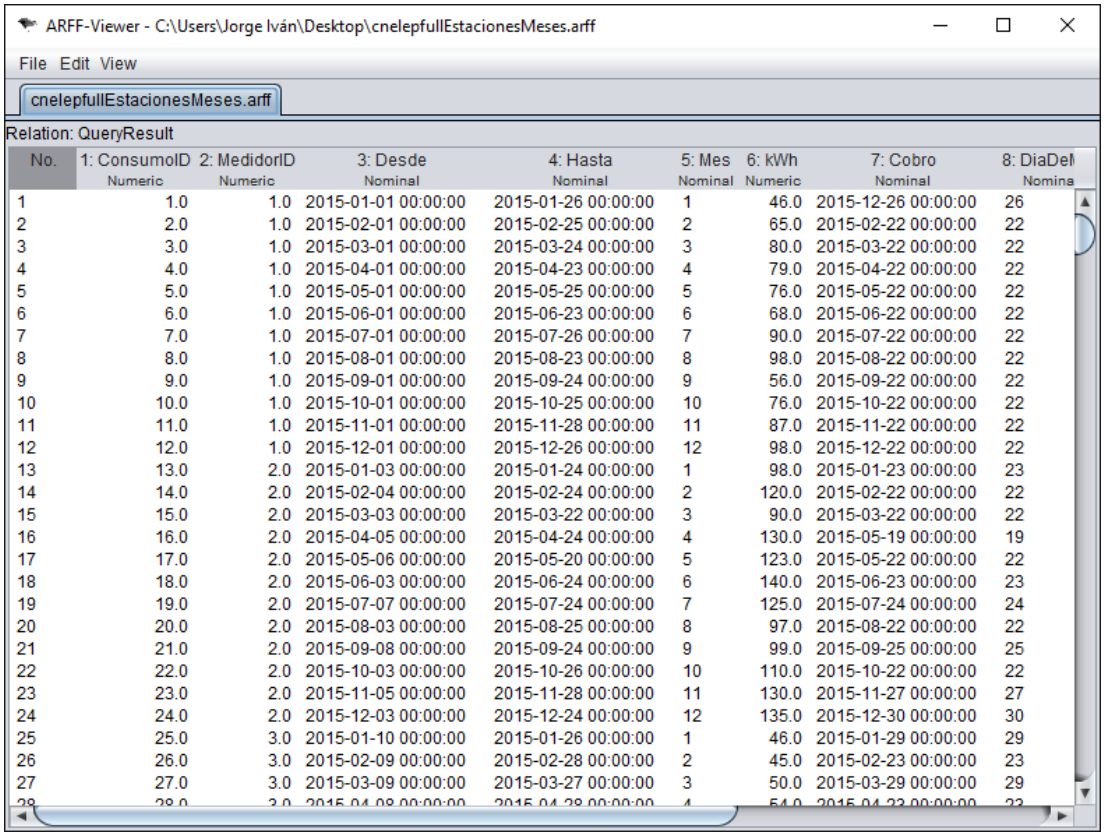


Ilustración 1: Diagrama Entidad Relación de las tablas empleadas para construir el archivo ARFF. Fuente: Investigación

```
SELECT  cast(consumo.id as decimal) as ConsumoID,
                cast(consumo.medidorid as decimal) as MedidorID,
                cast(consumo.desde as nchar(20)) as Desde,
                cast(consumo.hasta as nchar(20)) as Hasta,
                cast(consumo.mes as NCHAR(20)) as Mes,
                cast(consumo.consumokwh as decimal) as kWh,
                cast(consumo.fechaCobro as nchar(20)) as Cobro,
                cast(day(consumo.fechaCobro) as nchar(20)) as DiaDeMes,
                cast(dayofweek(consumo.fechaCobro) as nchar(10)) as DiaDeSemana,
                cast(monthname(consumo.fechaCobro) as nchar(20)) as MesdeAnio,
                CASE mes
                WHEN 1 THEN 'INVIERNO'
                WHEN 2 THEN 'INVIERNO'
                WHEN 3 THEN 'INVIERNO'
                WHEN 4 THEN 'INVIERNO'
                WHEN 5 THEN 'INVIERNO'
                ELSE 'VERANO'
                END as Estacion,
                cast(consumo.totalConsumo as decimal) as Total
FROM consumo
```

Ilustración 2: Consulta SQL para listar los datos de consumo desde WEKA, distinguiendo día de la semana, día del mes, mes del año y estación climática del cobro. Fuente: Investigación.



No.	1: ConsumoID Numeric	2: MedidorID Numeric	3: Desde Nominal	4: Hasta Nominal	5: Mes Nominal	6: kWh Numeric	7: Cobro Nominal	8: DiaDel Nomina
1	1.0	1.0	2015-01-01 00:00:00	2015-01-26 00:00:00	1	46.0	2015-12-26 00:00:00	26
2	2.0	1.0	2015-02-01 00:00:00	2015-02-25 00:00:00	2	65.0	2015-02-22 00:00:00	22
3	3.0	1.0	2015-03-01 00:00:00	2015-03-24 00:00:00	3	80.0	2015-03-22 00:00:00	22
4	4.0	1.0	2015-04-01 00:00:00	2015-04-23 00:00:00	4	79.0	2015-04-22 00:00:00	22
5	5.0	1.0	2015-05-01 00:00:00	2015-05-25 00:00:00	5	76.0	2015-05-22 00:00:00	22
6	6.0	1.0	2015-06-01 00:00:00	2015-06-23 00:00:00	6	68.0	2015-06-22 00:00:00	22
7	7.0	1.0	2015-07-01 00:00:00	2015-07-26 00:00:00	7	90.0	2015-07-22 00:00:00	22
8	8.0	1.0	2015-08-01 00:00:00	2015-08-23 00:00:00	8	98.0	2015-08-22 00:00:00	22
9	9.0	1.0	2015-09-01 00:00:00	2015-09-24 00:00:00	9	56.0	2015-09-22 00:00:00	22
10	10.0	1.0	2015-10-01 00:00:00	2015-10-25 00:00:00	10	76.0	2015-10-22 00:00:00	22
11	11.0	1.0	2015-11-01 00:00:00	2015-11-28 00:00:00	11	87.0	2015-11-22 00:00:00	22
12	12.0	1.0	2015-12-01 00:00:00	2015-12-26 00:00:00	12	98.0	2015-12-22 00:00:00	22
13	13.0	2.0	2015-01-03 00:00:00	2015-01-24 00:00:00	1	98.0	2015-01-23 00:00:00	23
14	14.0	2.0	2015-02-04 00:00:00	2015-02-24 00:00:00	2	120.0	2015-02-22 00:00:00	22
15	15.0	2.0	2015-03-03 00:00:00	2015-03-22 00:00:00	3	90.0	2015-03-22 00:00:00	22
16	16.0	2.0	2015-04-05 00:00:00	2015-04-24 00:00:00	4	130.0	2015-05-19 00:00:00	19
17	17.0	2.0	2015-05-06 00:00:00	2015-05-20 00:00:00	5	123.0	2015-05-22 00:00:00	22
18	18.0	2.0	2015-06-03 00:00:00	2015-06-24 00:00:00	6	140.0	2015-06-23 00:00:00	23
19	19.0	2.0	2015-07-07 00:00:00	2015-07-24 00:00:00	7	125.0	2015-07-24 00:00:00	24
20	20.0	2.0	2015-08-03 00:00:00	2015-08-25 00:00:00	8	97.0	2015-08-22 00:00:00	22
21	21.0	2.0	2015-09-08 00:00:00	2015-09-24 00:00:00	9	99.0	2015-09-25 00:00:00	25
22	22.0	2.0	2015-10-03 00:00:00	2015-10-26 00:00:00	10	110.0	2015-10-22 00:00:00	22
23	23.0	2.0	2015-11-05 00:00:00	2015-11-28 00:00:00	11	130.0	2015-11-27 00:00:00	27
24	24.0	2.0	2015-12-03 00:00:00	2015-12-24 00:00:00	12	135.0	2015-12-30 00:00:00	30
25	25.0	3.0	2015-01-10 00:00:00	2015-01-26 00:00:00	1	46.0	2015-01-29 00:00:00	29
26	26.0	3.0	2015-02-09 00:00:00	2015-02-28 00:00:00	2	45.0	2015-02-23 00:00:00	23
27	27.0	3.0	2015-03-09 00:00:00	2015-03-27 00:00:00	3	50.0	2015-03-29 00:00:00	29

Ilustración 3: Vista parcial de los 1200 registros devueltos por la consulta ejecutada en WEKA

Fuente: Investigación

Los datos agregados como atributos calculados que constituyeron variables categóricas del modelo de redes neuronales son: día de la semana, día del mes, mes de año y estación climática. Las nuevas columnas que se han indicado son adecuadas para el análisis por medio de redes neuronales. Durante la preparación de los datos se verificó y eliminó inconsistencias, por ejemplo, que una factura haya sido registrada como cobrada antes de su emisión, lo cual se refleja en la fecha de cobro. Se determinó qué columnas eran más relevantes para el modelo, siendo estas las fechas de cobro y los kW consumidos.

Paso 3: Exploración de datos

Los datos nominales disponibles en primera instancia son: *fecha desde*, *fecha hasta, mes, fecha de cobro, día de la semana, mes de año* y *estación*; los numéricos son: *código del consumo*, *código del medidor, consumo en kilowatt* y *número del día en el mes*. Estos datos resultan de la consulta mostrada en la Ilustración 2. Las decisiones se extraen principalmente de los atributos: *consumo en kilowatt, cobro, día del mes, día de la semana, mes del año* y *estación*; por tanto se ha establecido un perfil de datos de cada uno de estos atributos, empleando técnicas de exploración que según (Microsoft, 2018), son calcular los valores mínimos y máximos, calcular la media, desviaciones estándar, y examinar la distribución de los datos. A continuación, se muestran los perfiles de algunos datos luego de analizarlos con la herramienta WEKA:

- Atributo *Desde*: No es relevante para las predicciones y clasificaciones buscadas porque las predicciones se basan en la descomposición de la *fecha de cobro* y del *mes del cobro*. En la *ilustración 4* se muestran 143 fechas distintas entre las 1200 analizadas, es decir el 11% son únicas.
- Atributo *Mes:* Este atributo tiene un dominio que va desde 1 = enero hasta 12 = diciembre, mismo que es relevante para las predicciones y clasificaciones buscadas, porque refleja la categoría o mes que se debe pagar entre cada fecha. Existen 12 meses distintos, como era de suponerse y en todos los casos hay más de un registro para cada mes, por eso unique es 0%, como se muestra en la Ilustración 5.
- Atributo *kW* consumido en cada periodo: Siguiendo el proceso aplicado con los dos atributos anteriores, este atributo es relevante para las predicciones y clasificaciones pese a la cercanía entre la desviación estándar y el promedio, pues según la herramienta WEKA un 49% tienen por lo menos dos registros, lo que significa que el 49% de los abonados tiene al menos un consumo similar con otro abonado.
- Atributo de *fecha* en la que se hace el Cobro*:* Este atributo es relevante porque a partir de ahí se sacan variables categóricas para hacer las predicciones y clasificaciones; sin embargo, directamente el dato no es utilizable. En todo caso hay 160 fechas de cobro entre los 1200 registros y el 64% de estos tienen por lo menos más de un cobro asociado.
- Atributo *Dia del Mes* de cada fecha en la que se hace el cobro: Este atributo es relevante porque a partir de él se sacan las variables categóricas para hacer las predicciones y clasificaciones. Se registran 19 números de días distintos entre los 1200 registros y el 96% de ellos tienen por lo menos más de un cobro asociado, por ejemplo, en los días 26 se registran 402 de 1200 cobros.
- *Atributo Día de la semana* de cada fecha en la que se hace el cobro: Este atributo es relevante porque a partir de ahí se sacan las variables categóricas para hacer las

predicciones y clasificaciones. Cada uno de los siete días semanales tienen al menos dos cobros asociados. Se registran 176 de 1200 cobros los días 7 = domingo, lo que significa que muchas personas hacen sus pagos fuera de las oficinas de la empresa eléctrica.

- Atributo *Estación climática* de cada fecha en la que se hace el cobro: Este atributo es relevante porque a partir de ahí se sacan variables categóricas para hacer las predicciones y clasificaciones. Se registran cobros en ambas estaciones climáticas de Ecuador.
- Atributo *Mes de año* de cada fecha en la que se hace el cobro: Este atributo es relevante porque a partir de ahí se sacan variables categóricas para hacer las predicciones y clasificaciones. Se registran todos los meses entre los 1200 registros y cada mes (100%) tiene por lo menos dos cobros asociados. Diciembre registra 101 cobros.
- Atributo *Total de dinero* recaudado en cada cobro: Atributo relevante para las predicciones y clasificaciones buscadas, aunque la desviación estándar es relativamente alta dada la diferencia entre consumo entre los usuarios.

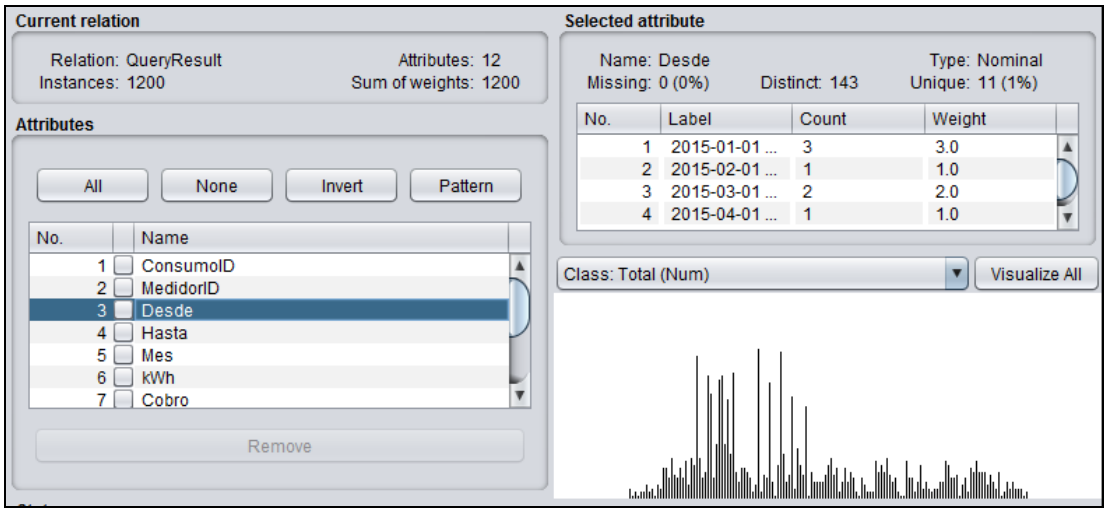


Ilustración 4: Estadísticas del atributo Desde el periodo que se factura.

Fuente: Investigación.

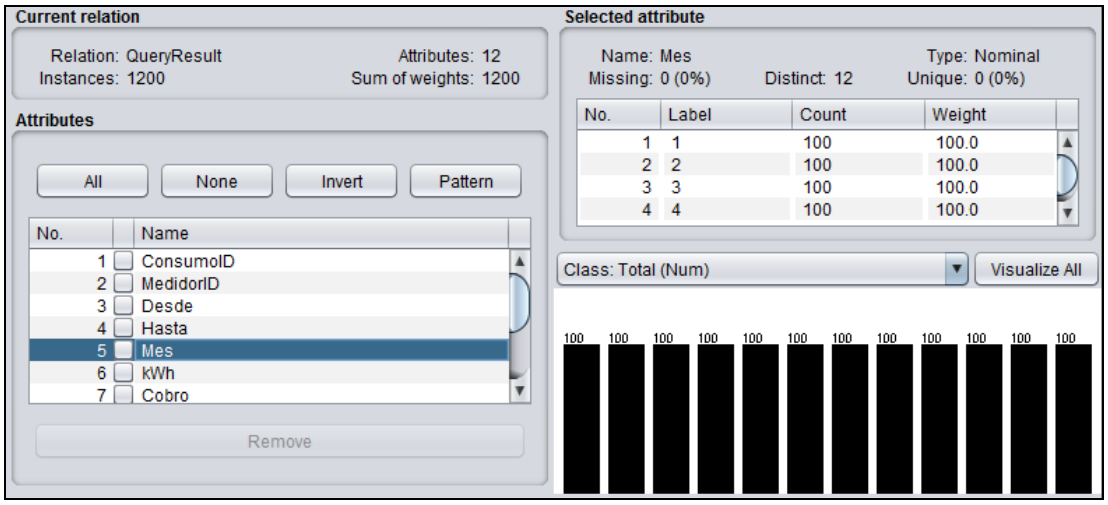


Ilustración 5: Estadísticas del atributo Mes en el periodo que se factura. Fuente: Investigación.

Paso 4: Generación de modelos

- El primer modelo generado fue una Red Neuronal de clase Perceptron Multicapa, que usa backprogation para clasificar las 1200 instancias. Para este ejercicio se tiene inicialmente se tiene:
- Número de instancias: 1200.
- Atributos de entrada: Mes a pagar y total a pagar.
- Atributos de salida: mes en que se paga.
- Entrenamiento: Training set.
- Número de capas: 3 ocultas con 6 neuronas cada una (ver hidden layer en la Ilustración 6)
- Número de épocas: 500, lo que significa que los 1200 registros se introducen 500 veces hasta procurar que el error cuadrático medio sea el menor posible.

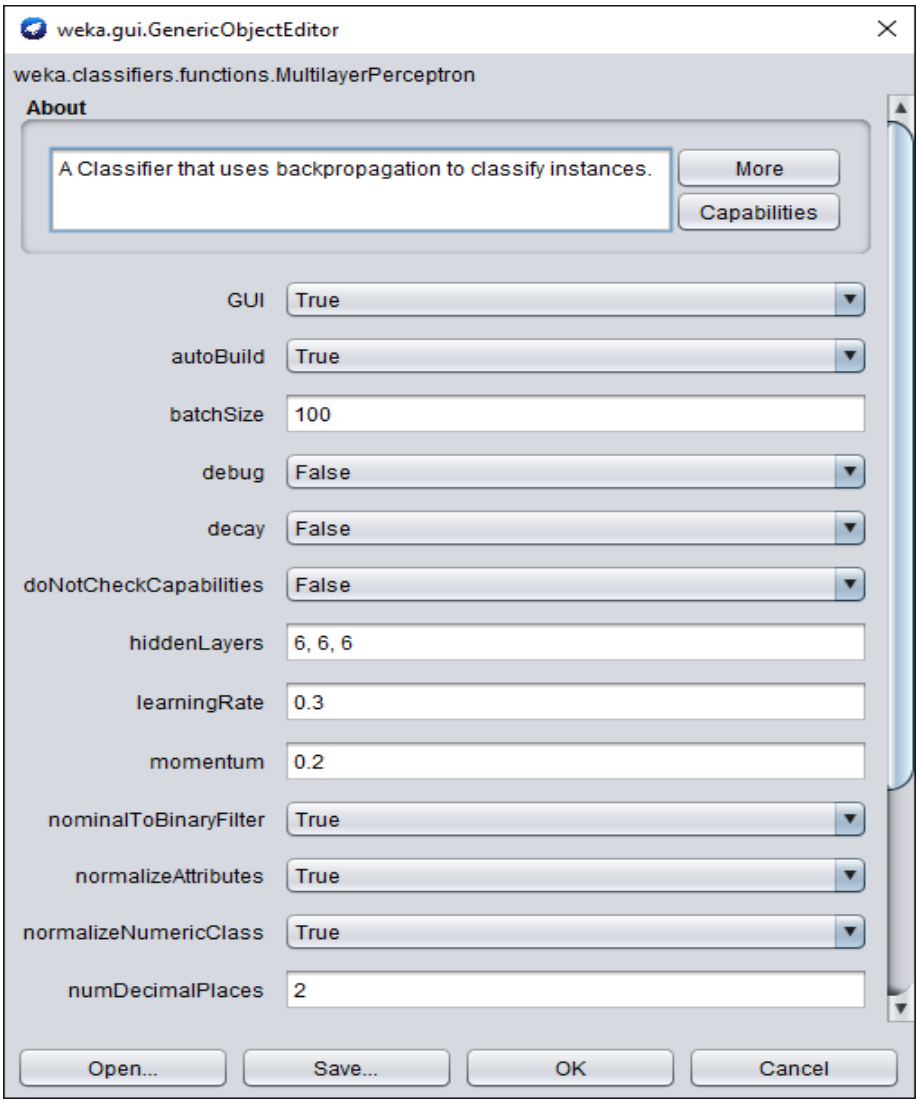


Ilustración 6: Configuración del perceptron multicapa. Fuente: Investigación.

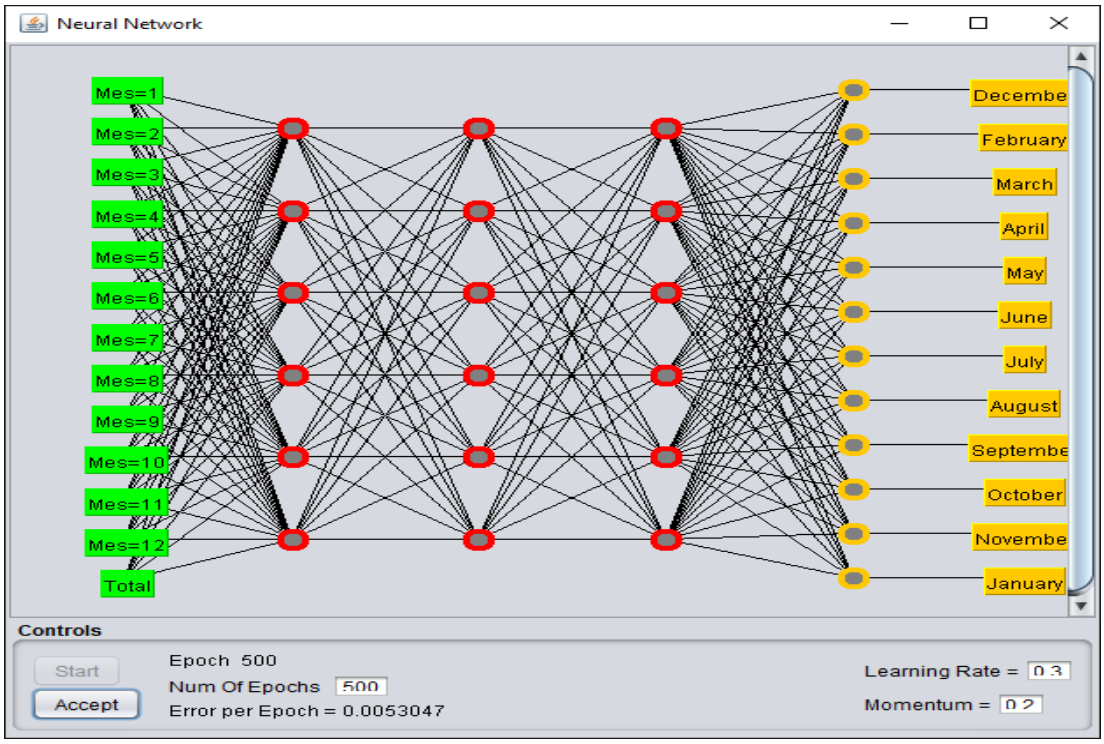


Ilustración 7: Modelo de la red neuronal con tres capas ocultas de seis neuronas cada una, configurado con 500 épocas. Para generar el modelo es necesario que GUI en la Ilustración 6 este fijado a true. Una vez que haga clic en el botón Aceptar se puede leer el texto de detalle del gráfico. Fuente: Investigación.

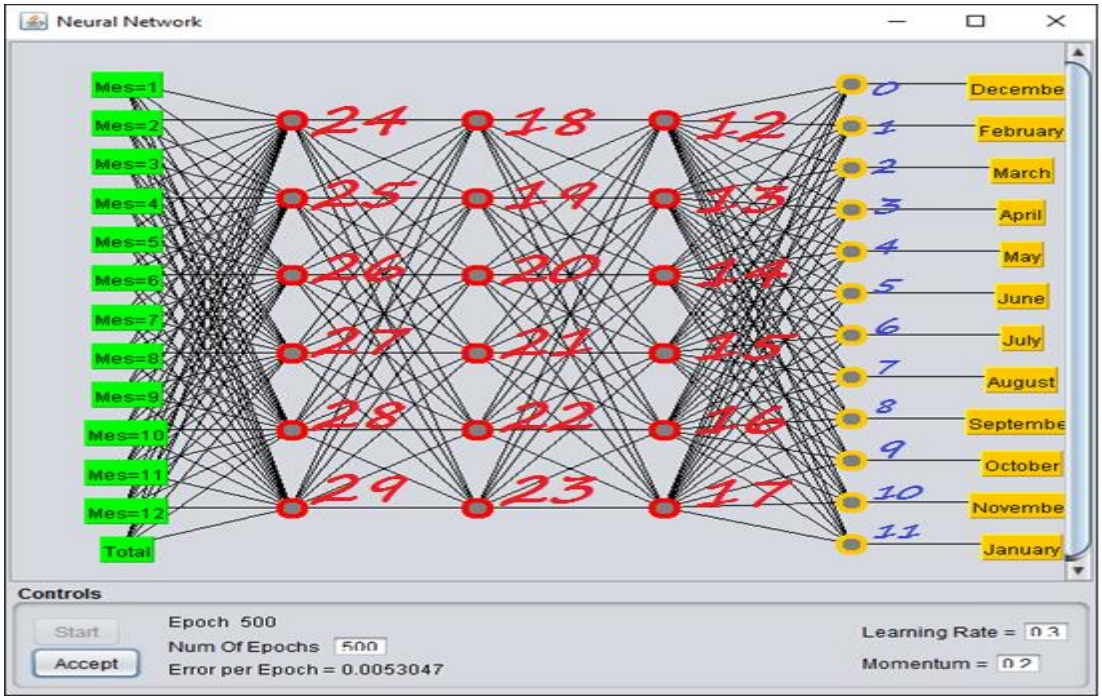
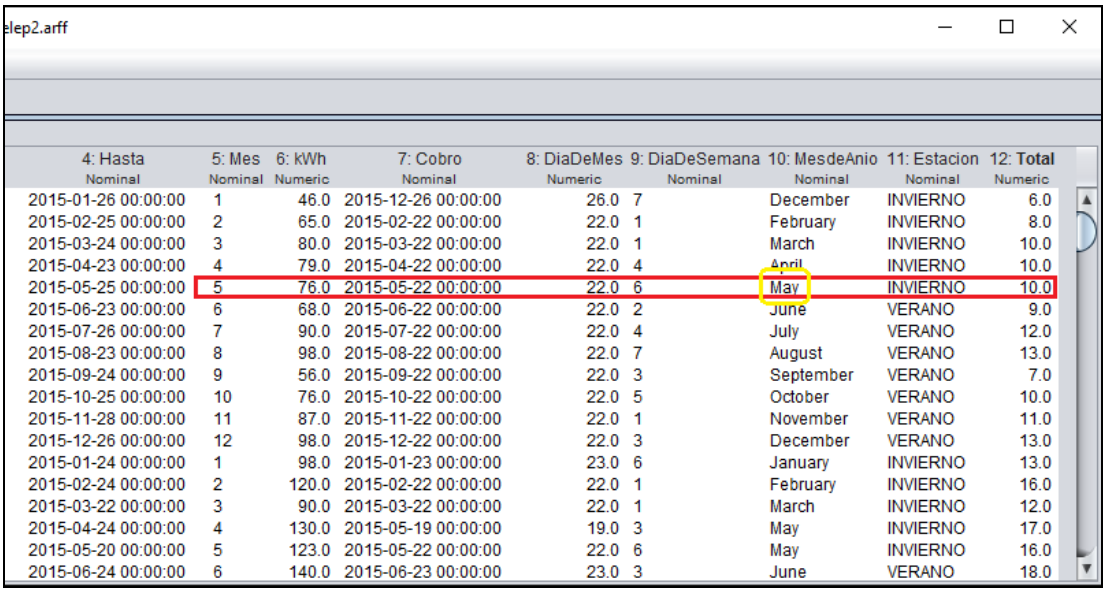


Ilustración 8: A modo de ilustración, WEKA identifica los 29 nodos, donde 12, es decir del 0 al 11 son clases de salida y 18 se corresponden con las tres capas ocultas de seis neuronas cada una. El orden sucede porque se aplica el algoritmo de backpropagation. Fuente: Investigación.



elep2.arff

4: Hasta	5: Mes	6: kWh	7: Cobro	8: DiaDeMes	9: DiaDeSemana	10: MesdeAnio	11: Estacion	12: Total
Nominal	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal	Nominal	Numeric
2015-01-26 00:00:00	1	46.0	2015-12-26 00:00:00	26.0	7	December	INVIERNO	6.0
2015-02-25 00:00:00	2	65.0	2015-02-22 00:00:00	22.0	1	February	INVIERNO	8.0
2015-03-24 00:00:00	3	80.0	2015-03-22 00:00:00	22.0	1	March	INVIERNO	10.0
2015-04-23 00:00:00	4	79.0	2015-04-22 00:00:00	22.0	4	April	INVIERNO	10.0
2015-05-25 00:00:00	5	76.0	2015-05-22 00:00:00	22.0	6	May	INVIERNO	10.0
2015-06-23 00:00:00	6	68.0	2015-06-22 00:00:00	22.0	2	June	VERANO	9.0
2015-07-26 00:00:00	7	90.0	2015-07-22 00:00:00	22.0	4	July	VERANO	12.0
2015-08-23 00:00:00	8	98.0	2015-08-22 00:00:00	22.0	7	August	VERANO	13.0
2015-09-24 00:00:00	9	56.0	2015-09-22 00:00:00	22.0	3	September	VERANO	7.0
2015-10-25 00:00:00	10	76.0	2015-10-22 00:00:00	22.0	5	October	VERANO	10.0
2015-11-28 00:00:00	11	87.0	2015-11-22 00:00:00	22.0	1	November	VERANO	11.0
2015-12-26 00:00:00	12	98.0	2015-12-22 00:00:00	22.0	3	December	VERANO	13.0
2015-01-24 00:00:00	1	98.0	2015-01-23 00:00:00	23.0	6	January	INVIERNO	13.0
2015-02-24 00:00:00	2	120.0	2015-02-22 00:00:00	22.0	1	February	INVIERNO	16.0
2015-03-22 00:00:00	3	90.0	2015-03-22 00:00:00	22.0	1	March	INVIERNO	12.0
2015-04-24 00:00:00	4	130.0	2015-05-19 00:00:00	19.0	3	May	INVIERNO	17.0
2015-05-20 00:00:00	5	123.0	2015-05-22 00:00:00	22.0	6	May	INVIERNO	16.0
2015-06-24 00:00:00	6	140.0	2015-06-23 00:00:00	23.0	3	June	VERANO	18.0

Ilustración 9: El modelo dice que, si el cobro vence un 25 de mayo, el margen de error o de que no se pague en mayo es de 0,0192 (positivo, es decir un poco después de la fecha, pero en el mes de mayo mismo). Fuente: Investigación.

El segundo modelo generado fue el de Reglas de Asociación con el algoritmo PART, empleado para determinar los días de la semana en que más se cobra el servicio eléctrico. Sus configuraciones fueron:

- Número de instancias: 1200.
- Atributos de entradas: Mes a pagar, kW consumidos, día del mes del cobro, mes del año del cobro, la estación climática.
- Atributos de salida: día de la semana del cobro.
- Entrenamiento: Use training set.
- Algoritmo: PART.
- Número de épocas: 500, lo que significa que los 1200 registro se introducen 500 veces hasta procurar que el error cuadrático medio sea lo menor posible.

Las siguientes son algunas reglas de asociación obtenidas con el Algoritmo PART, con sus respectivos significados:

- Cuando tocaba pagar septiembre, los abonados que pudieron hacerlo entre los días 20 y 22 (11 en total) prefirieron pagar un miércoles.
- Cuando tocaba pagar mayo, los abonados que pudieron hacerlo entre los días 24 y 26 (30 en total) prefirieron pagar un miércoles.
- Cuando tocaba pagar abril, los abonados que pudieron hacerlo los días 25 o 26 (33 en total) prefirieron pagar un lunes.
- Cuando tocaba pagar algún mes del invierno, los abonados que pudieron hacerlo entre los días 22 y 24 (12 en total) prefirieron pagar un domingo.

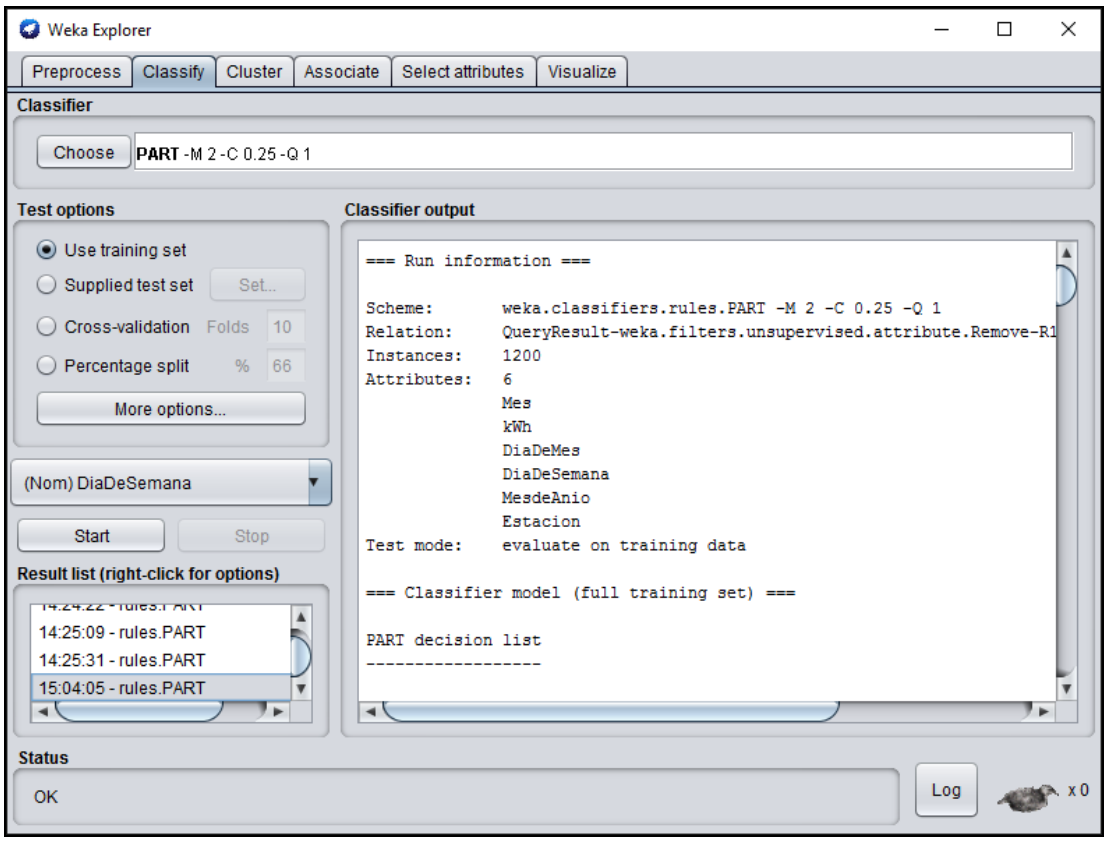


Ilustración 10: Vista general de los parámetros del modelo PART para el listado de reglas de predicción del día de la semana en que se cobró. *Fuente: Investigación.*

Paso 5: Validación de modelos

En la *Ilustración 11*, se muestra la Matriz de confusión correspondiente al Perceptron Multicapa, que evidencia la aceptable clasificación de los datos, por ejemplo, en mayo (fila) se registraron 101 valores cobrados, de los cuales el modelo ha clasificado correctamente como *e* (e=mayo) a 100 e incorrectamente clasificó 1 caso como *d* (d = abril). En noviembre no hay errores de clasificación.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   <-- classified as
 100   0   0   0   0   0   0   0   0   0   0   1 |   a = December
   0  99   0   0   0   0   0   0   0   0   0   0 |   b = February
   0   0 100   0   0   0   0   0   0   0   0   0 |   c = March
   0   0   0  99   0   0   0   0   0   0   0   0 |   d = April
   0   0   0   1 100   0   0   0   0   0   0   0 |   e = May
   0   0   0   0   0 100   0   0   0   0   0   0 |   f = June
   0   0   0   0   0   0 100   0   0   0   0   0 |   g = July
   0   0   0   0   0   0   0 100   0   0   0   0 |   h = August
   0   0   0   0   0   0   0   0 100   0   0   0 |   i = September
   0   0   0   0   0   0   0   0   0 100   1   0 |   j = October
   0   0   0   0   0   0   0   0   0   0  99   0 |   k = November
   0   1   0   0   0   0   0   0   0   0   0  99 |   l = January
```

Ilustración 11: Matriz de confusión correspondiente al Perceptron Multicapa.

Fuente: Investigación.

```
=== Summary ===

Correctly Classified Instances        1196               99.6667 %
Incorrectly Classified Instances         4                0.3333 %
Kappa statistic                          0.9964
Mean absolute error                      0.0192
Root mean squared error                  0.0637
Relative absolute error                 12.5856 %
Root relative squared error             23.0315 %
Total Number of Instances             1200
```

Ilustración 12: Resumen de los 1200 registros analizados con el Perceptron Multicapa. El root mean squared error es del 0,0637 en tanto que el error absoluto es muy pequeño pero positivo.

Fuente: Investigación.

Las siguientes validaciones se corresponden a las reglas de asociación generadas con el algoritmo *PART*:

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 172   1   0   1   1   0   1 |   a = 7
   1 172   0   1   1   2   0 |   b = 1
   1   0 138   0   0   1   0 |   c = 4
   0   1   0 151   0   0   1 |   d = 6
   0   0   0   0 173   2   1 |   e = 2
   2   0   2   1   1 153   0 |   f = 3
   1   1   1   2   1   0 213 |   g = 5
```

Ilustración 13: Matriz de confusión reportada al aplicar PART. Reporta 28 errores, por ejemplo, para el miércoles (fila f=3) 153 registros se clasificaron correctamente y 6 no. Fuente: Investigación.

En general, *PART* generó 127 reglas a partir de los 1200 registros con un error medio cuadrático de 0,069. Apenas 28 registros se clasificaron incorrectamente, tal cual se detalló en la matriz de confusión.

Paso 6: Implementación y actualización de modelos

La propuesta de esta investigación es alternativa, y respecto a este sexto paso del denominado Proceso de Microsoft para construir modelos de minería de datos, la empresa eléctrica puede usar los modelos para crear predicciones como soporte a sus decisiones comerciales, tal cual este estudio sugiere. En general, realizar modificaciones constantes puede mejorar la efectividad de las estrategias de implementación, además de crear consultas de contenido para recuperar estadísticas, reglas o fórmulas del modelo o embeber la funcionalidad de los modelos en una aplicación.

RESULTADOS

Luego de haber analizado los datos con los algoritmos mencionados se obtuvieron los siguientes resultados:

- El perceptron multicapa, con su algoritmo *backpropagation* funciona bastante bien, con 3 capas ocultas de 6 neuronas cada una, pues el error medio absoluto indica la calidad de la medida del modelo al ser apenas del 0,02 después de que se revisó 500 veces (épocas) cada uno de los 1200 registros. La diagonal de la matriz de confusión representada en la Ilustración 11, que mide el acuerdo inter evaluador para las variables nominales que en este caso son los meses en que se cobró, clasificó correctamente 1196 registros, lo que guarda concordancia con el resultado reflejado en la medida estadística del *Coeficiente de Kappa* que alcanza el 0,9964 sobre 1.
- Las reglas de asociación con el algoritmo PART, también reportan datos interesantes, pues el error medio absoluto del modelo es apenas del 0,0095 una vez que se revisó las 1200 instancias que generaron un total de 127 reglas, el modelo clasificó correctamente cerca del 98% de las instancias, lo que se respalda con la matriz de confusión representada en la Ilustración 13 que mide el acuerdo inter evaluador para las variables nominales analizadas y concuerda con la medida estadística del Coeficiente de Kappa que alcanza el 0,9727 sobre 1. Adicionalmente, el algoritmo PART resulta menos complejo de actualizar por parte del personal de TI de la empresa eléctrica, en comparación con el perceptron multicapa dado que la cantidad de configuraciones que se requiere es menor.

DISCUSIÓN

Coincidiendo con (Ariza Ramírez, 2013, p. 24), el pronóstico de la demanda de energía eléctrica es un procedimiento sistemático que permite definir cuantitativamente la demanda futura procurando la exactitud de la información, pero sin olvidar la presencia de incertidumbres; como las reflejadas en las matrices de confusión resultantes en este estudio; donde también se puede cruzar información y pronosticar la demanda en forma de energía (*Wh*) y/o potencia (*W*).

A pesar de la importancia de los enfoques basados en minería de datos, como en este trabajo, y pese a los resultados favorables, por ejemplo, en la predicción de los días de mayor pago, los algoritmos presentados pueden tener limitaciones como el hecho de que los modelos de redes neuronales o de reglas de asociación, según (Li & Wen, 2014) pueden no funcionar adecuadamente fuera de sus datos de entrenamiento o sí se generaliza o no, mucho más allá del

rango de entrenamiento. Aun así, las reglas de asociación han sido incluidas por la IEEE International Conference on Data Mining, entre los diez primeros algoritmos de minería de datos más influyentes en la comunidad de investigación (Wu et al., 2008, p. 2), en tanto que las redes neuronales son en concreto una de las técnicas más usadas en la predicción de consumos eléctricos (Amasyali & El-Gohary, 2018b).

Respecto al presente estudio, y más en particular sobre la construcción del archivo ARFF extraído a partir de la base de datos MySQL de la empresa eléctrica, y que contó con 1200 registros que corresponden a consumos eléctricos del sector residencial en el año 2015, en la práctica se debe contar por lo menos con registros de 10 años para el pronóstico de demanda de energía eléctrica.

CONCLUSIÓN

No hay un modelo de minería de datos o combinación de algoritmos de aprendizaje automático único para todos los conjuntos de datos, por lo tanto, es esencial considerar caso por caso los aspectos discutidos en este documento, incluidos los datos disponibles y las propiedades de estos algoritmos, en favor de mejoras entre las cuales resalta el análisis de la eficiencia energética. Aunque la verdadera importancia del pronóstico de la demanda se incrementa en la medida que el cumplimiento de los objetivos trazados dependa lo menos posible del azar, incluso es recomendable que en el caso de los perceptrones multicapas se realicen simulaciones paramétricas que determinen combinaciones más precisas en cuanto a número de capas y neuronas por capas, disipando la posible incertidumbre sobre los resultados de las decisiones tomadas a partir de los modelos.

REFERENCIAS BIBLIOGRÁFICAS

Amasyali, K., & El-Gohary, N. M. (2018a). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*, 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095

Amasyali, K., & El-Gohary, N. M. (2018b). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*, 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095

Ariza Ramírez, A. M. (2013). *Métodos utilizados para el pronóstico de demanda de energía eléctrica en sistemas de distribución*. Universidad Tecnológica de Pereira, Pereira - Colombia. Retrieved from https://tinyurl.com/y7akrz7z

Gönen, T. (1986). *Electric power distribution system engineering*. New York, New York, USA: McGraw-Hill.

Li, X., & Wen, J. (2014). Review of building energy modeling for control and operation. *Renewable and Sustainable Energy Reviews*, *37*, 517–537. https://doi.org/10.1016/j.rser.2014.05.056

Microsoft. (2018). Data Mining Concepts. Retrieved August 1, 2018, from https://tinyurl.com/yay5hjqt

Rosenblatt, F. (1961). *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms*. Buffalo, NY: Cornell Aeronautical Lab Inc. Retrieved from https://tinyurl.com/yb8qk6zz

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science.

Van Heddeghem, W., Lambert, S., Lannoo, B., Colle, D., Pickavet, M., & Demeester, P. (2014). Trends in worldwide ICT electricity consumption from 2007 to 2012. *Computer Communications*, *50*, 64–76. https://doi.org/10.1016/j.comcom.2014.02.008

Wang, Z., & Srinivasan, R. S. (2015). A review of artificial intelligence based building energy prediction with a focus on ensemble prediction models. In *Winter Simulation Conference (WSC), 2015* (pp. 3438–3448). IEEE. https://doi.org/10.1109/WSC.2015.7408504

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Burlington, MA: Morgan Kaufmann. Retrieved from

http://www.cs.waikato.ac.nz/~ml/weka/book.html

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., … Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/DOI 10.1007/s10115-007-0114-2

Xiao, F., & Fan, C. (2014a). Data mining in building automation system for improving building operational performance. *Energy and Buildings*, *75*, 109–118. https://doi.org/10.1016/j.enbuild.2014.02.005

Xiao, F., & Fan, C. (2014b). Data mining in building automation system for improving building operational performance. *Energy and Buildings*, *75*(3), 109–118. https://doi.org/10.1016/j.enbuild.2014.02.005