

Aplicación de métodos de aprendizaje automático en un sistema basado en ontología

María-Isabel Castellanos¹, Ariam Rivas², Emilio Lucas¹

¹ Departamento de Ingeniería Informática, Universidad de Holguín, Cuba

² Departamento de Matemática, Universidad de Holguín, Cuba
mcastellanosd@uho.edu.cu, arivasm90@gmail.com,
elucas1@uho.edu.cu

Abstract. The ontology-based system for the management of environmental indicators in corporations (SIGCIA) allows the detection of an indicator alteration, if it exceeds a limit value. In this case, this system recommends the possible environmental impacts, the causes of the indicator alteration and the mitigation actions. In order to make these recommendations, the limit value for each indicator must be pre-defined in the software by the environmental management specialist. This means that the determination of limit values is done subjectively, based on the knowledge of the historical behavior of the indicator in a specific organization; so it is necessary to have an automatic forecast method. This research transits through all the phases of the process of Knowledge Discovery in Data (KDD). A selection of attributes in the dataset was made applying several selectors and a group of regression models were applied. Artificial Neural Networks with Multi-Layer Perceptron topology showed best performance. It allows the prediction of the limit value of the energy consumption indicator, dataset selected as study case. The prediction of limit values and the potential offered by the ontology-based recommendation system make it a powerful tool to support decision-making in the process of environmental management, with broad generalization possibilities in Cuban business sector.

Keywords: Environmental Indicators, Artificial Neural Networks, Forecast, Ontology-based System

Resumen. El sistema basado en ontología para la gestión de los indicadores ambientales empresariales (SIGCIA) permite detectar la alteración de un indicador, si sobrepasa un valor límite; recomendando los posibles impactos ambientales, las causas de la alteración del indicador y las acciones de mitigación. Para poder realizar estas recomendaciones, en el software debe estar predefinido por el especialista de gestión ambiental el valor límite para cada indicador. Esto trae como consecuencia que la determinación de valores límites se realice de forma subjetiva, en base al conocimiento del comportamiento histórico del indicador en una organización específica; por lo que se hace necesario contar con un método de pronóstico automático. Esta investigación transita por todas las fases del proceso de Descubrimiento del Conocimiento en Datos (KDD). Se realiza una selección de atributos en el conjunto de datos, donde se aplicaron

varios selectores y algoritmos de regresión. El de mejor rendimiento fue las Redes Neuronales Artificiales con topología *Multi-Layer Perceptron*; el cual permite la predicción del valor límite del indicador de consumo de energía que fue el conjunto de datos seleccionado como caso de estudio. La predicción de valores límites, unido a las potencialidades que ofrece el sistema de recomendación basado en ontología, convierten al SIGCIA en una potente herramienta de apoyo a la toma de decisiones en el proceso de Gestión Ambiental, con una amplia posibilidad de generalización en el sector empresarial cubano.

Palabras claves: Indicadores Ambientales, Redes Neuronales Artificiales, Predicción, Sistema Basado en Ontología

1. Introducción

La herramienta de gestión más importante para el control medioambiental son los indicadores¹. Los indicadores ambientales condensan la información y simplifican el acercamiento a los fenómenos medioambientales, a menudo complejos, lo que los hace muy útiles para la comunicación [10].

La utilidad de estos indicadores consiste en facilitar, tanto a la dirección de la organización como al resto de los miembros, información relevante y resumida, en forma de declaraciones concisas e ilustrativas para la toma de decisiones. Por lo tanto, aseguran una rápida evaluación de las principales mejoras y de los puntos débiles en la protección ambiental de la empresa; para aquellos que han de tomar las decisiones [13].

Por esta razón, es una necesidad el uso de indicadores ambientales para medir el comportamiento de la organización en esta área, facilitar la comunicación y condensar la información ambiental. El uso de indicadores, a su vez, contribuye a una mejora en la gestión del conocimiento ambiental.

Para gestionar el conocimiento es determinante su representación, lo cual no es más que el proceso de estructurar el conocimiento sobre un problema de manera que sea más fácil de resolver. En aras de potenciar la gestión y específicamente la representación del conocimiento, las Tecnologías Semánticas (TS) son cada vez más utilizadas [14][16]. Dentro de las TS, las ontologías son actualmente una de las Formas de Representación del Conocimiento (FRC) más empleadas [6][17].

En este sentido los autores [3] desarrollaron un Sistema basado en ontología para la gestión del conocimiento de indicadores ambientales (SIGCIA); con el objetivo de gestionar el conocimiento ambiental que se infiere del almacenamiento histórico de los indicadores ambientales empresariales. En esta investigación se propone el uso de las Redes Neuronales Artificiales (RNA) para predecir el valor límite del indicador a partir de su almacenamiento histórico. Entre las potencialidades que las RNA ofrecen está que no necesitan un experto humano al cual extraerle el conocimiento.

¹ Un dato que ha sido seleccionado a partir de un conjunto estadístico más amplio por poseer una significación y una representatividad particular [10].

2. Sistema basado en ontología de indicadores ambientales

Una de las formas de representación del conocimiento más empleadas actualmente son las ontologías; las cuales ofrecen disímiles ventajas para la modelación, generación, distribución y uso del conocimiento producido y acumulado en las organizaciones [1][4]. Dadas estas ventajas para la gestión del conocimiento, las ontologías son ampliamente usadas para gestionar los amplios volúmenes de información ambiental que devienen de este proceso, principalmente del almacenamiento histórico de los indicadores ambientales.

El sistema SIGCIA se basa en la ontología OntoEnvironmental, en la cual se modelan los indicadores ambientales que rigen el proceso de gestión ambiental empresarial. En la Figura 1 se muestra el sistema de indicadores al que se hace referencia.

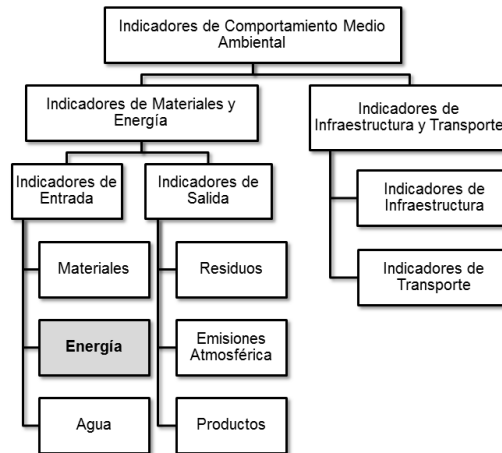


Fig. 1. Indicadores medioambientales para empresas.

Este software basado en este sistema de indicadores, permite el cálculo de estos teniendo en cuenta que el indicador debe tener predefinido su valor límite (el cual es definido por el Especialista de Gestión Ambiental de forma manual, en correspondencia al concomitamiento tácito sobre el comportamiento histórico del indicador). En respuesta a esta acción el sistema se encarga de comparar el valor y el valor límite del indicador. En caso de que el valor del indicador resulte mayor que el valor límite, el sistema declara que el indicador está alterado y mediante la máquina de inferencia recomienda las posibles causas, los posibles impactos ambientales y las acciones de mitigación [3].

No obstante a las potenciales mejoras que la implantación del sistema SIGCIA ofrece para el correcto desempeño del proceso de gestión ambiental en las

organizaciones, aún existe la siguiente deficiencia. Al Especialista de Gestión Ambiental se le dificulta el establecimiento de valores límites debido a que los indicadores reflejan diversas área de la entidad (por ejemplo área energética, transporte), por lo tanto se establece el valor de una forma subjetiva. Esto trae como consecuencia que una mala decisión al establecer un valor límite restrinja las potencialidades que el sistema ofrece, al no hacer las recomendaciones de forma oportuna.

3. Proceso de descubrimiento de conocimiento en datos

Los grandes volúmenes de datos e información que se manejan en la actualidad han traído como consecuencia la necesidad de desarrollar técnicas y herramientas que permitan asistir al hombre a extraer información útil, conocimiento y patrones de los datos almacenados. Para suplir esta necesidad surge el Proceso de Descubrimiento de Conocimiento en Datos (*KDD*).

De acuerdo a Fayyad [8] es definido como: *El proceso no trivial de identificación en los datos, de patrones válidos, novedosos, comprensibles y potencialmente útiles*. El KDD es un área de la computación que intenta explotar la enorme cantidad de información mediante el descubrimiento de patrones útiles y representativos, extrayendo conocimiento que pueda asistir a un humano para llevar a cabo tareas de forma más eficiente y satisfactoria. A continuación se muestran de manera general las fases por las que atraviesa este proceso:

- Selección: desarrolla un entendimiento del dominio del problema y de los datos que serán utilizados en la tarea de descubrimiento de conocimiento.
- Pre-procesamiento y transformación: cubre el total de actividades para la construcción de conjunto de datos final. En esta tarea se incluye la selección de registros, atributos, limpieza de los datos, tratamiento de los valores ausentes, entre otros. También se realiza la transformación de los datos en el formato requerido por la herramienta de minería de datos seleccionada. Esta tarea consume entre el 35% y 20% del tiempo [5].
- Minería de datos (MD): es la determinación de la tarea de descubrimiento a realizar (clasificación, regresión, agrupamiento) y la aplicación de uno o varios algoritmos, de dicha tarea, con el fin de descubrir patrones ocultos en los datos. Esta tarea ocupa entre el 15% y 20% del tiempo de realización del proyecto.
- Interpretación y evaluación: se interpretan y evalúan los patrones descubiertos, por lo que en ocasiones es necesario regresar a los pasos anteriores, lo que implica repetir el proceso, tal vez con otros datos, algoritmos, metas y estrategias. Este paso puede ser auxiliado por visualizaciones y contribuye a eliminar patrones redundantes o irrelevantes.

3.1. Selección de los datos

Todo proyecto KDD tiene sus orígenes en la petición de un cliente el cual quiere mejorar algunos de sus procesos, haciendo uso de los datos históricos de

dichos procesos. Para sacar el máximo provecho de estos datos se hace necesario que los que realicen de este tipo de proyecto conozcan y comprendan dichos datos.

El almacenamiento histórico de cada indicador ambiental constituye un conjunto de datos. A este indicador es necesario calcularle su valor límite para delimitar cuando está alterado y recomendar las posibles causas, los posibles impactos ambientales y las acciones de mitigación. Para determinar el valor límite se aplicó el esquema KDD, con la idea de encontrar un modelo que permita obtener dicho valor de la manera más exacta posible.

Para la realización de esta investigación se cuenta con los datos referentes al indicador de consumo de energía mensual de la Empresa Constructora del Poder Popular (ECOPP) de la provincia de Holguín, Cuba. Esta información contiene un historial de aproximadamente 5 años (del 1 de diciembre del 2010 hasta 1 de noviembre del 2015). La ECOPP para registrar en sus archivos el consumo de electricidad mensual maneja cinco parámetros como se muestra en la tabla 1.

Table 1. Descripción del conjunto de datos.

Atributo	Valor
potencia global activa (<i>kilowatt</i>)	real
potencia global reactiva (<i>kilowatt</i>)	real
voltaje (<i>volt</i>)	real
intensidad global (<i>ampere</i>)	real
consumo (<i>watt/hora</i>)	real

La figura 2 muestra el comportamiento de las instancias del conjunto de datos. Se puede observar como en la mayoría de los meses el consumo de energía se encuentra entre 1600-2445 (watt/hora).

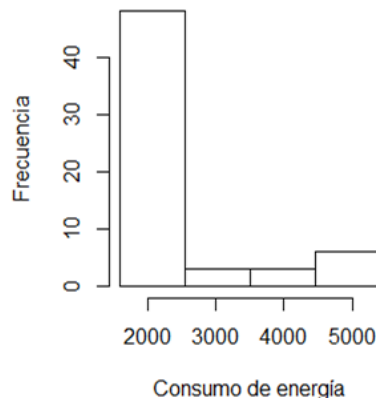


Fig. 2. Distribución por instancia (atributo consumo).

3.2. Selección de atributos

La Selección de Atributos (SA) se puede definir como el proceso de obtener los n atributos más representativos de los N originales a partir de la eliminación de los redundantes e irrelevantes. De una manera más formal, el objetivo es seleccionar un subconjunto de atributos S del espacio original A con clase C , tal que $P(C|S) \approx P(C|A)$, es decir, obtener mejor o igual desempeño predictivo mediante la eliminación de atributos ruidosos y redundantes.

Existen varios criterios para agrupar los algoritmos de SA. Uno de ellos lo constituye el modo de aplicación, según el cual pueden ser vistos como de filtro (*filter*) o envolvente (*wrapper*). En esta investigación el modo de aplicación de los algoritmos de selección de atributos fue mediante el criterio de *wrapper*.

Los envoltentes son aquellos que se auxilian de la precisión del clasificador para evaluar a los subconjuntos del espacio. Esta estrategia ofrece mejores resultados, ya que en un paso previo a la clasificación el algoritmo de aprendizaje escoge a los atributos que mejor representen el conocimiento para su construcción; sin embargo es altamente costosa. Se emplearon cinco algoritmos de estrategias envoltentes: *LinearRegression* (*LR*), *MultilayerPerceptron* (*MLP*), *M5P*, *K vecinos más cercanos* (*K-nn*) y *M5Rules* (*M5R*).

LinearRegression El análisis regresivo es una técnica utilizada para inter y extra polar las observaciones, las cuales pueden clasificarse como regresión lineal o no lineal. Se habla de modelo de regresión cuando la variable de respuesta y las variables explicativas son todas cuantitativas. Además permite determinar el modelo o ecuación matemática que mejor representa la relación existente entre las variables que se analizan [18].

MultilayerPerceptron Una Red Neuronal Artificial (RNA) es un modelo computacional que pretende simular el funcionamiento del cerebro. El proceso de aprendizaje de una RNA de topología *MultilayerPerceptron* consiste en determinar los pesos que permitan codificar el conocimiento subyacente en los datos [2]. Esto consiste en hacer variar los pesos en función de alguna regla de aprendizaje hasta que los mismos sean constantes, con lo que se dice que la red ha aprendido. El buen funcionamiento predictivo del mismo viene dado por la alta tolerancia al ruido de los datos y la habilidad de capturar relaciones complejas entre los atributos y la clase.

M5P En el caso del algoritmo *M5P*, se trata de obtener un árbol de modelos (un modelo lineal que predice el valor de la clase), si bien se puede utilizar para obtener un árbol de regresión, por ser este un caso específico de árbol de modelos [12].

K vecinos más cercanos Es un algoritmo simple que almacena todos los casos disponibles y clasifica los nuevos casos sobre la base de una medida de

similitud (funciones de distancia). K-nn se ha utilizado en el reconocimiento de patrones estadísticos, estimación y ya en el inicio de década de 1970 como una técnica no paramétrica. Un caso se clasifica por un voto de la mayoría de sus vecinos, con el caso que se asigna a la clase más común entre sus vecinos más cercanos K, medidos por una función de la distancia. Si $K = 1$, entonces el caso es simplemente asignado a la clase de su vecino más cercano [15].

M5Rules El algoritmo M5R [11] es un algoritmo de inducción simple aunque los conjuntos de reglas de los árboles de modelos son precisos. El algoritmo trabaja construyendo modelos de árboles repetidamente y selecciona la mejor regla en cada iteración. Produce conjuntos de reglas que son en general tan precisos y a la vez más pequeños que los árboles de modelos contruidos con el conjunto de datos completo.

3.3. Minería de datos

En esta investigación se ha preparado un estudio experimental para explorar el comportamiento de las RNA, en conjuntos de datos donde su tipo de clase es continua. Se ha utilizado para la ejecución de los experimentos la herramienta Weka [9]. Desarrollada en la Universidad de Waikato (Nueva Zelanda) bajo licencia GNU (General Public License), se caracteriza por la independencia de arquitectura.

Es utilizado un esquema de experimentación basado en validación cruzada que garantiza una mayor robustez estadística. Esta propuesta consiste en un procedimiento de validación cruzada con diez particiones con una corrida como lo propone [7]. Como parámetro de evaluación se emplea: coeficiente de correlación y error medio absoluto.

La correlación es la técnica estadística que estudia el problema de medir la intensidad o el grado de relación que existe entre las variables que se investigan. El Coeficiente de Correlación, es un valor comprendido entre -1 y 1, indica la relación lineal que existe entre dos variables. El error medio absoluto mide la magnitud media de los errores en un conjunto de pronósticos, sin tener en cuenta su dirección. Mide la precisión para las variables continuas.

Se realizó una experimentación utilizando los algoritmos de selección de estrategia *Wrapper* mencionados anteriormente y como algoritmos de regresión: *LR*, *MLP*, *M5P*, *K-nn* y *M5R*. Los resultados arrojados se muestran en la Tabla 2 y Tabla 3.

3.4. Interpretación y evaluación

En la Tabla 2 y en la Tabla 3 los valores resaltados son los algoritmos de mejor coeficiente de correlación y error absoluto medio respectivamente. Se puede observar como el algoritmo de regresión *MLP* con el selector de atributos de estrategia *Wrapper MLP* es el de mayor coeficiente de correlación en la Tabla 2 y el de menor error medio absoluto en la Tabla 3.

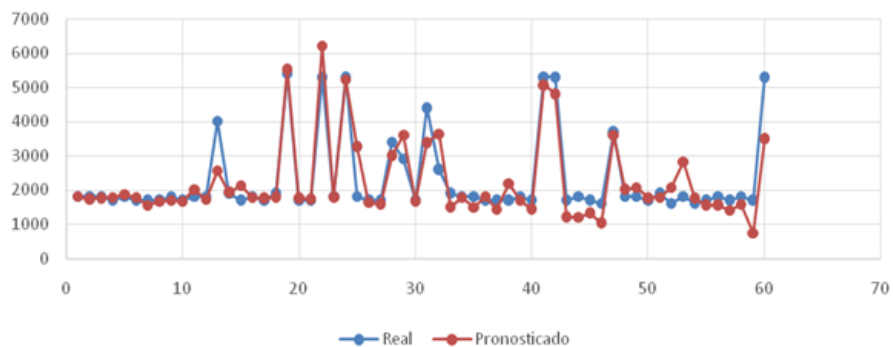
Table 2. Resultado de coeficiente de correlación.

	Wrapper					
	LR	MLP	M5P	K-nn	M5R	-
LR	0,598	0,631	0,598	0,164	0,636	0,640
MLP	0,796	0,901	0,796	0,226	0,888	0,885
M5P	0,876	0,863	0,876	0,718	0,874	0,863
K-nn	0,705	0,806	0,705	0,674	0,683	0,7
M5R	0,800	0,860	0,800	0,779	0,882	0,86

Table 3. Resultado de error medio absoluto.

	Wrapper					
	LR	MLP	M5P	K-nn	M5R	-
LR	733	739	733	825	726	723
MLP	436	329	436	795	349	351
M5P	396	397	396	573	401	397
K-nn	390	295	390	370	350	368
M5R	420	373	420	364	352	373

En la Figura 3 se muestra el consumo de energía eléctrica en la ECOPP durante cinco años. En dicha gráfica de tendencia los consumos están presentados por meses. El color azul representa el consumo real, mientras que el color rojo es el consumo pronosticado por el algoritmo MLP. Se puede apreciar como el error del clasificador es bajo.

**Fig. 3.** Pronóstico del consumo de energía.

4. Integración del algoritmo de predicción de valores límites en el SIGCIA

Con el modelo óptimo seleccionado, MLP, se dio paso a la integración del algoritmo de predicción de valores límites, Figura 4, en el modelo SIGCIA. En este diagrama el primer paso es que el usuario introduzca los datos, que constituyen el valor del indicador para ese mes. Luego el modelo de RNA seleccionado analiza estos datos de entrada y realiza su pronóstico, que para el SIGCIA va ser el valor límite. Cuando se obtenga el valor real del indicador este se compara con el pronóstico (valor límite). En caso de ser mayor el real, mediante la máquina de inferencia, que posee la arquitectura del SIGCIA, esta recomienda las posibles causas de alteración del indicador, los posibles impactos ambientales y las acciones de mitigación.

Para lograr la integración del algoritmo de predicción de valores límites al SIGCIA, se crea en el sistema un grupo de clases. Una clase (Modelo_Clasificación) la cual se encarga de construir el Dataset haciendo uso de los datos que se encuentran en la tabla (tb_consumo_energía) incluida en la base de datos db_indicadores. Para la construcción del modelo y clasificar la instancia se hace uso de la librería WEKA ver Figura 5.

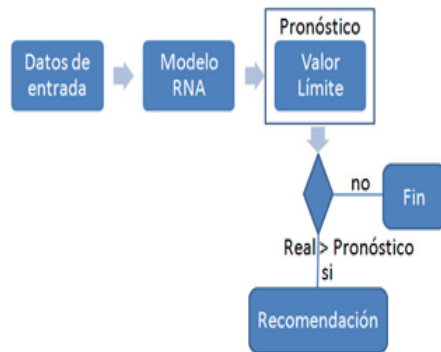


Fig. 4. Diagrama de flujo predicción de valor límite.

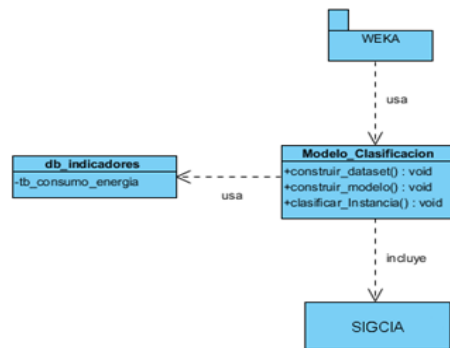
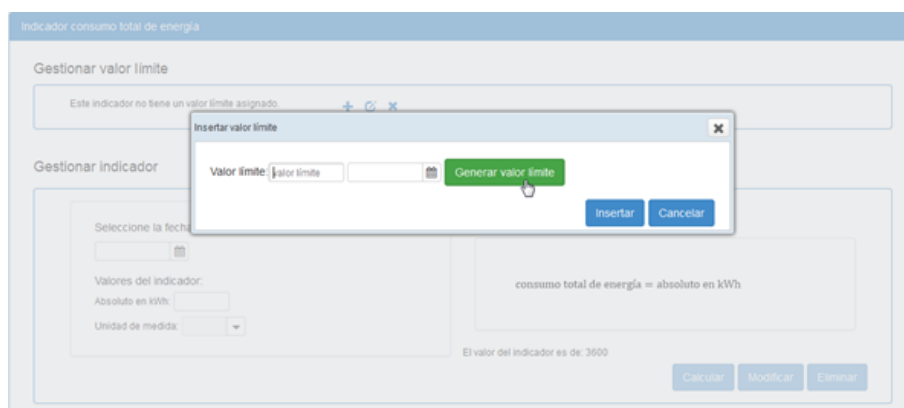


Fig. 5. Integración de Weka al SIGCIA.

4.1. Funcionalidades del SIGCIA una vez integrado

Cuando el usuario (Especialista de Gestión Ambiental) desea gestionar el valor límite del indicador, en este caso consumo total de energía, una vez seleccionado el indicador hace click en el botón adicionar y se muestra una ventana de diálogo donde se tienen dos opciones, introducir el valor manualmente o generar el valor como se muestra en la Figura 6. Si el usuario selecciona la opción de generar el valor límite, se le muestra un formulario el cual debe llenar para el correcto funcionamiento del algoritmo (ver Figura 7). Luego de llenar los datos del

formulario correctamente el usuario hace click en el botón insertar, acto seguido se ejecuta el método generarValorlímite() el cual crea primeramente una instancia temporal con los datos introducidos por el usuario en el formulario, construye el modelo de regresión, clasifica la instancia y propone el valor límite. Una vez definido el valor límite del indicador; el usuario calcula el valor real del mismo para el mes en curso. Este valor es guardado en la base de datos y se actualiza el dataset. Posteriormente si el valor del indicador real sobrepasa el valor límite el sistema ofrece un conjunto de recomendaciones, fruto de las consultas realizadas a la ontología OntoEnvironmental.



The screenshot shows a web interface for managing energy consumption indicators. The main window is titled 'Indicador consumo total de energía'. It has two main sections: 'Gestionar valor límite' and 'Gestionar indicador'. The 'Gestionar valor límite' section shows a message: 'Este indicador no tiene un valor límite asignado'. The 'Gestionar indicador' section includes a date selector, a list of indicator values, and a unit of measurement dropdown. A modal dialog box titled 'Insertar valor límite' is open, showing a 'Valor límite' input field, a 'Generar valor límite' button, and 'Insertar' and 'Cancelar' buttons. The background interface shows a value of 3600 and buttons for 'Calcular', 'Modificar', and 'Eliminar'.

Fig. 6. Generar valor límite.



The screenshot shows a dialog box titled 'Insertar Indicador'. It contains four input fields: 'Fecha' with the value '1/06/2016', 'Potencia Global Reactiva' with the value '0.282', 'Voltaje' with the value '237.14', and 'Intensidad Global' with the value '14.2'. At the bottom right, there are 'Insertar' and 'Cancelar' buttons.

Fig. 7. Atributos del indicador.

5. Conclusiones

En el estudio de los algoritmos de selección de atributos con la estrategia Wrapper para seleccionar los métodos a aplicar en la investigación, permitió constatar que existían atributos redundantes e irrelevantes, debido a que el de mejores resultados fue Wrapper (MLP) eliminando el atributo Intensidad. El estudio realizado sobre los cinco modelos de regresión de distintas ramas, arrojó que el algoritmo de regresión MultiLayerPerceptron en cuanto a los parámetros medidos (coeficiente de correlación y error absoluto medio) fue el de mejores resultados.

La integración del algoritmo de regresión MultiLayerPerceptron al sistema SIGCIA permite la predicción del valor límite del indicador consumo de energía que fue el conjunto de datos seleccionado en la investigación actual. Lo anterior facilita el trabajo del Especialista de Gestión Ambiental debido a que el sistema realiza las recomendaciones de forma oportuna y favorece la toma de decisiones al respecto. Los resultados obtenidos con la aplicación del algoritmo de regresión MultiLayerPerceptron al conjunto de datos tomado de la Empresa Constructora del Poder Popular acerca del indicador consumo de energía demuestra que para este indicador en otras organizaciones el algoritmo antes mencionado se puede generalizar.

Referencias

1. Arp, R., Smith, B., Spear, A.D.: Building ontologies with basic formal ontology. Mit Press (2015)
2. Bishop, C.: Extremely well-written, up-to-date. Requires a good mathematical background, but rewards careful reading, putting neural networks firmly into a statistical context. Neural Networks for Pattern Recognition (1995)
3. Castellanos Domínguez, M.I., Quevedo Castro, C.M., Vega Ramírez, A., Grangel González, I., Moreno Rodríguez, R.: Sistema basado en ontología para el apoyo a la toma de decisiones en el proceso de gestión ambiental empresarial. In: II International Workshop of Semantic Web. vol. 1797. CEUR-WS, La Habana, Cuba (2016), <http://ceur-ws.org/Vol-1797/>
4. Cimiano, P., McCrae, J., Buitelaar, P.: Lexicon model for ontologies: Community report. W3C Ontology-Lexicon Community Group (2016)
5. Cios, K.J., Kurgan, L.A.: Trends in Data Mining and Knowledge Discovery (Dm), 1–26 (2000)
6. Davis, E.: Representations of commonsense knowledge. Morgan Kaufmann (2014)
7. Demsar, J.: Comparison of Classifiers over Multiple Data Sets. Journal of Machine Learning Research vol. 7: 31 (2006)
8. Fayyad, U.: The KDD Process for Extracting Useful Knowledge from Volumes of Data 39(11), 27–34 (1996)
9. Frank, I.H.W., E: Practical Machine Learning Tools and Techniques (2005)
10. Heink, U., Kowarik, L.: What are indicators? On the definition of indicators in ecology and environmental planning. Ecological Indicators ELSEVIER 10, 584–593 (2010), www.elsevier.com/locate/ecolind
11. Holmes, G., Hall, M., Frank, E.: Generating Rule Sets from Model Trees (1999)

12. Holte, R.C.: Very Simple Classification Rules Perform Well on Most Commonly Used Datasets (1988), 63–91 (1993)
13. IHOBE: Guía de Indicadores Medioambientales para la Empresa (1999)
14. Kharlamov, E., Solomakhina, N., Özçep, O.L., Zheleznyakov, D., Hubauer, T., Lamparter, S., Roshchin, M., Soyly, A., Watson, S.: How semantic technologies can enhance data access at siemens energy. In: International Semantic Web Conference. pp. 601–619. Springer (2014)
15. Khelifi J., F..J.: K-NN Regression to Improve Statistical Feature Extraction for Texture Retrieval. IEEE Transactions on Image Processing pp. 20, 293–298. (2011)
16. Meroño Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., Van Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web 6(6), 539–564 (2015)
17. Tiropanis, T., Davis, H.C., Cerri, S.A.: Semantic Technologies and Learning. In: Encyclopedia of the Sciences of Learning, pp. 3029–3032. Springer (2012)
18. Torres-Reyna: Regression Analysis. Data Analysis 101. University Princeton (2009)