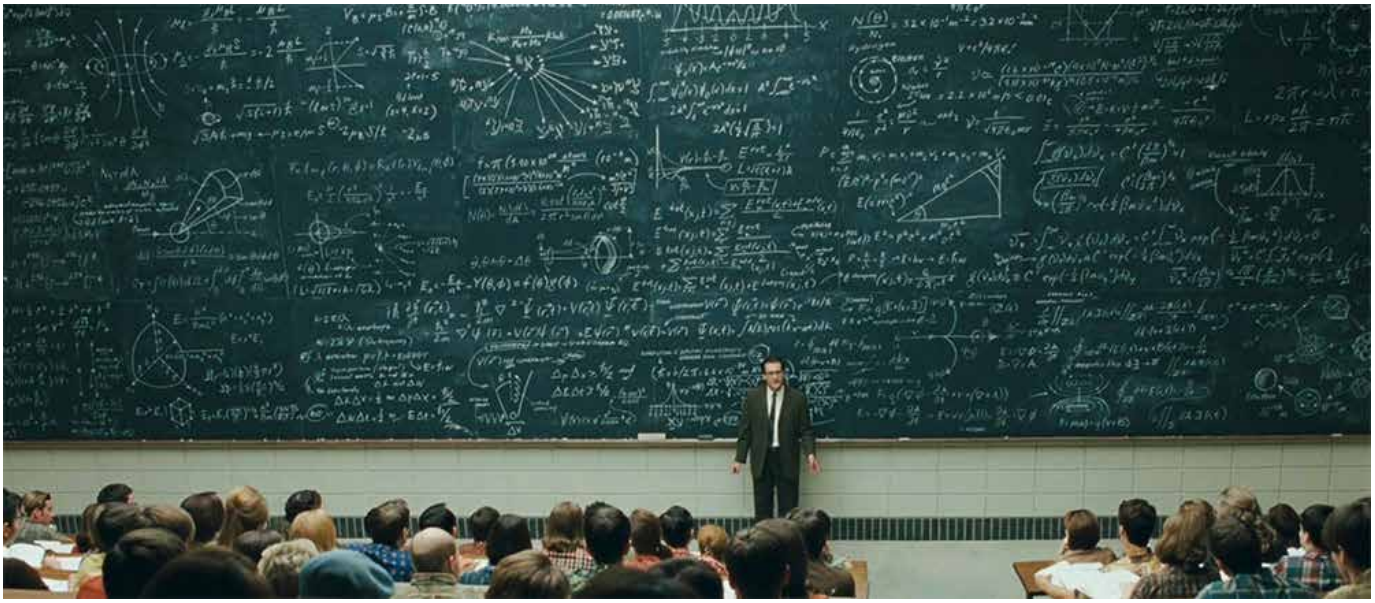

Detección de tendencias presentes y futuras en el aprovechamiento académico mediante aprendizaje profundo y árboles de decisión

Luis Enrique Vivanco Benavides[1], Abraham Jiménez Alfaro[2] y Edgar Corona Organiche[3]

Acerca de los autores...

[1] Estudiante de Maestría en Ingeniería en Sistemas Computacionales, Tecnológico de Estudios Superiores de Ecatepec.

[2, 3] División de Ingeniería en Sistemas Computacionales, Tecnológico de Estudios Superiores de Ecatepec.

[1] luis.bnavids@gmail.com,
[2] ajja_mx@yahoo.com, borganiche2003@yahoo.com.mx
[3] organiche2003@yahoo.com.mx

Tecnológico de Estudios Superiores de Ecatepec

Resumen

El aprendizaje automático puede definirse como aquel método dentro del cómputo científico donde un sistema software o algoritmo es capaz de extraer patrones ocultos de un conjunto de datos, identificando y aprendiendo los atributos existentes. Asimismo, el aprendizaje profundo consiste en una familia de técnicas que facilitan la búsqueda y clasificación de conocimiento por parte del aprendizaje automático. El presente artículo, que es de corte cuantitativo, expone las etapas de análisis y minería de datos posteriores al pre procesamiento, basadas en la metodología KDD a través de las cuales se pretende descubrir las relaciones ocultas entre los datos para poder encontrar tendencias presentes y futuras en el aprovechamiento académico de estudiantes de Ingeniería en Sistemas Computacionales del Tecnológico de Estudios Superiores de Coacalco (TESCo), en las materias que conforman el área de programación. Para ello, se aplicaron métodos de clasificación supervisada, tales como árboles de decisión y redes neuronales artificiales, las cuales se implementaron una vez que los datos fueron correctamente preparados y transformados para que los resultados obtenidos tengan validez.

Palabras Clave: Análisis y minería de datos, aprendizaje profundo, árboles de decisión, clasificación supervisada, redes neuronales artificiales.

Introducción

La gestión del conocimiento que actualmente se lleva dentro de algunas instituciones de educación superior (IES), ha mostrado deficiencias en el seguimiento y valoración de diversos asuntos de índole académico, dentro de los cuales destaca la medición del aprovechamiento del alumnado, donde ha faltado profundizar en los datos estadísticos de cada periodo, limitándose en identificar índices de reprobación en uno o varios periodos dados.

La aplicación de minería de datos para extracción de conocimiento del alumno, han sido factible desde que comenzaron a ser gestionados por las universidades, ya que desde hace varios años, algunas de estas instituciones estaban trabajando en la implementación de sistemas de gestión o transaccionales que integran procesos y áreas. Estos sistemas producen datos que se almacenan en sus bases con dimensiones considerablemente grandes y amplia diversidad de temas (Menéndez M., De Luján, M. 2012)

La Minería de Datos Educativos (EDM), según Quinteros, O., Funes y Ahumada (2016) es una disciplina relacionada con el desarrollo de métodos para extraer información útil a partir de los datos que se generan en los entornos educativos, y utilizarla para mejorarlos; asimismo, Rajni Jindal, R.y Dutta

Borah, M. (2013) mencionan que los métodos de minería de datos usados principalmente en la EDM, se dividen en dos grupos: los orientados a la verificación y los orientados al descubrimiento; dentro de estos últimos, tenemos a su vez métodos de Clasificación, Estadísticas, Agrupamiento, Predicción, Redes Neuronales, Minería de Reglas de Asociación y Minería Web.

El tratamiento adecuado de los bancos de información es fundamental para poder obtener datos estructurados que permitan utilizarse en el análisis enfocado a la extracción de conocimiento.

1. Análisis y modelado en minería de datos

El proceso de Extracción de Conocimiento a partir de Bases de Datos (KDD, del inglés Knowledge Discovery from Databases), podría definirse como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos (Usama, F. 1996). La implementación de esta metodología exige un conjunto de etapas rigurosas para que las técnicas seleccionadas puedan ofrecer resultados válidos. Sólo cuando se han finalizado exitosamente las etapas de pre procesamiento y transformación, es posible comenzar con la fase de análisis y modelado. Dentro de ella, existen distintas técnicas de análisis que habrán de seleccionarse con base en lo definido en la etapa de comprensión del dominio. Dichas técnicas se clasifican en predictivas y descriptivas.

Las técnicas predictivas especifican el modelo para los datos con base en un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos, antes de aceptarlo como válido (Pérez, M., 2015). En esta clasificación se incluyen algunas de ellas, que son:

- Modelos de regresión. Proceso estadístico para estimar las relaciones entre variables, que incluye técnicas para el modelado y análisis de variables, centrando la atención en la relación entre una variable dependiente y una o más variables independientes (Madrigal Espinoza, S. D., 2014, p. 12).

- Series temporales. Son una forma estructurada de representar datos en un cierto periodo. Sirven para generar pronósticos, extendiendo los valores conocidos a futuro, donde aún no hay mediciones disponibles.

- Árboles de decisión. Constituyen un método de segmentación enfocado en resolver problemas de discriminación en una población, segmentando de forma progresiva la muestra de interés para así obtener la correcta clasificación de grupos homogéneos (Vivanco, L., 2017).

- Algoritmos genéticos. Técnica de búsqueda iterativa inspirada en los principios de selección natural. Buscan modelar estrategias de optimización, generando poblaciones de individuos mediante la reproducción de las variables padre (Ponce, P., 2010).

- Redes bayesianas. Se basan en el teorema de Bayes. Éstos suponen que el efecto de un valor de atributo en una clase dada, es independiente de los valores de los otros atributos. Esta suposición es llamada relación de independencia condicional y permiten la representación de dependencias entre subconjuntos de atributos (Kamber, J., 2006).

- Redes neuronales. Conjunto de elementos de procesamiento de la información que se encuentran conectados entre sí, capaces de aprender con los datos que reciben en su capa de entrada. Resulta un modelo aplicable a diversos tipos de problemas como: reconocimiento de patrones, clasificación, discriminación, análisis y filtrado de datos, predicción, entre otros.

El proceso de extracción de patrones a través de la metodología KDD sugiere utilizar al menos dos técnicas para el minado de datos, esto para comparar los resultados de cada una, o bien, para la unificación de éstas adquiriendo resultados de mayor trascendencia.

1.1 Redes neuronales artificiales: Perceptron Multicapa

Las redes neuronales artificiales (RNA) son una clase de modelos discriminantes y de regresión lineal flexibles; son modelos de reducción de datos y sistemas dinámicos no lineales que constan de un número habitualmente grande de neuronas interconectadas por lo regular en forma compleja y que suelen organizarse en capas (Quintín, M., Santana, Y., 2007).

Existen distintos tipos de RNA, dentro de los cuales resaltan las perceptron multicapa (multilayer perceptron), que surgen con el fin de resolver problemas no lineales. La arquitectura de este tipo de RNA se caracteriza por tener sus neuronas agrupadas en capas de diferentes niveles, cada una formada por un conjunto de neuronas y se distinguen tres tipos diferentes de capas: la de entrada, cuyas neuronas se encargan únicamente de recibir las señales o patrones que proceden del exterior y propagar dichas señales a todas las neuronas de la siguiente capa; las ocultas, donde se realiza un procesamiento no lineal de los patrones recibidos, y por último la capa que actúa como salida de la red, proporcionando al exterior la respuesta de la red para cada uno de los patrones de entrada (Viñuela, P., Galván, I. 2004).

Con base en las propiedades mencionadas, este tipo de red neuronal se utilizó como primer paso para la clasificación supervisada de patrones. La Figura 1 representa gráficamente la RNA implementada, la cual consta de 19 neuronas en la capa de entrada, 3 capas ocultas, donde la primera está formada por 20 neuronas, la segunda por 15 y la tercera por 10. La capa de salida se conforma de una neurona.

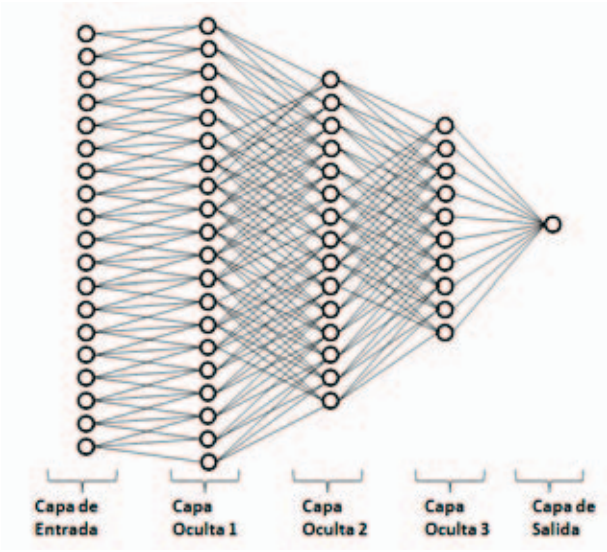

Figura 1

Modelo del perceptron multicapa implementado
Fuente: Elaboración propia.

Esta red trabaja bajo un modelo clásico de aprendizaje, denominado regla de aprendizaje de Hebb. Se le llama aprendizaje hebbiano a aquellas formas de aprendizaje que involucran una modificación de los pesos proporcionales al producto de una entrada *j* por la salida *i* de la neurona (Martín del Brío, B. S. 2007). A continuación se ejemplifican las expresiones que sirven para calcular las activaciones de las neuronas de la red de cada una de las capas.

Esta RNA con *C* capas y n_c neuronas en la capa *c*; la expresión w^c_{ij} representa el peso de la conexión de la neurona *i* de la capa *c* para *c=3,..., C.*

La activación de las neuronas de la capa de entrada (a^1_c) se representa de la siguiente manera:

$$a_i^1 = x_i \; para \; i = 1, 2, ..., n_1 \qquad \textbf{(1)}$$

Donde X = (x_1, x_2..., x_{n1}), el patrón de entrada de la red.

La activación de las neuronas de cada capa oculta *c*(a^c_i), aplican la función de activación *f* a la suma de los productos de las activaciones que percibe por sus pesos correspondientes.

Lo anterior se representa de la siguiente manera:

$$a_i^c = f(\sum_{j=1}^{n_{c-1}} w_j^{c-1} a_j^{c-1} + u_i^c) \; para \; i = 1, 2, ..., n_c \; y \; c = 2, 3, ..., C-1 \qquad \textbf{(2)}$$

Donde a^{c-1}_j representa las activaciones de las neuronas *C-1.*

La activación de la neurona de la capa de salida viene dada por la función de activación *f* aplicada a la suma de los productos de las entradas que obtiene de sus pesos. A continuación se muestra la expresión correspondiente:

$$y_i = a_i^C = f(\sum_{j=1}^{n_{C-1}} w_j^{C-1} a_j^{C-1} + u_i^C) \; para \; i = 1, 2, ..., n_c \qquad \textbf{(3)}$$

Donde Y = (y_1, y_2..., y_{nC}), representa el vector de salida de la red.

El modelo de RNA utilizado puede considerarse como aprendizaje profundo, debido a la cantidad de capas ocultas y al número de neuronas que las conforman. Dicha red presenta una precisión estimada del 93.234% al momento de ser entrenada con el conjunto de datos disponibles, previamente preparados y transformados. El modelo de RNA entrenado que se ha obtenido, será utilizado posteriormente.

2. Integración de técnicas de clasificación supervisada

Como se comentó anteriormente, al implementar la minería de datos utilizando la metodología KDD se recomienda aplicar varias técnicas para comparar y/o complementar resultados. Ya que el presente análisis es de

índole predictivo, todos los métodos de clasificación ocupados son de tipo supervisado, partiendo de variables dependientes para poder obtener la segmentación requerida.

Cuando el perceptron multicapa se encuentra entrenado, se dice que este modelo ha aprendido del conjunto de datos y que ha logrado alcanzar una clasificación de los mismos. El modelo de aprendizaje profundo fue determinado con base en la cantidad de datos preparados y transformados que se tienen a disposición, de los atributos de éstos y de la complejidad que presentan sus relaciones, por lo cual se requiere una RNA de clasificación con un alto nivel de exactitud.

2.1 Multilayer perceptron y árbol de decisión CHAID

Los árboles de decisión son un método de clasificación supervisada, ya que si se tiene una variable o clase dependiente, el objetivo del clasificador va a ser averiguar dicha clase para casos nuevos (Sierra, B. 2006). El algoritmo de árboles de decisión CHAID (Chi-square Automatic Interaction Detector) se caracteriza por realizar una búsqueda de relaciones estadística entre las variables del conjunto de datos donde se aplica. Otras propiedades que lo distinguen, son el hecho de que no maneja una fase de post-poda para evitar que el modelo resulte sobreentrenado, deteniendo al algoritmo en la misma etapa en la que se construye el árbol. La Figura 2 muestra parte del árbol resultante, aplicado a la RNA.

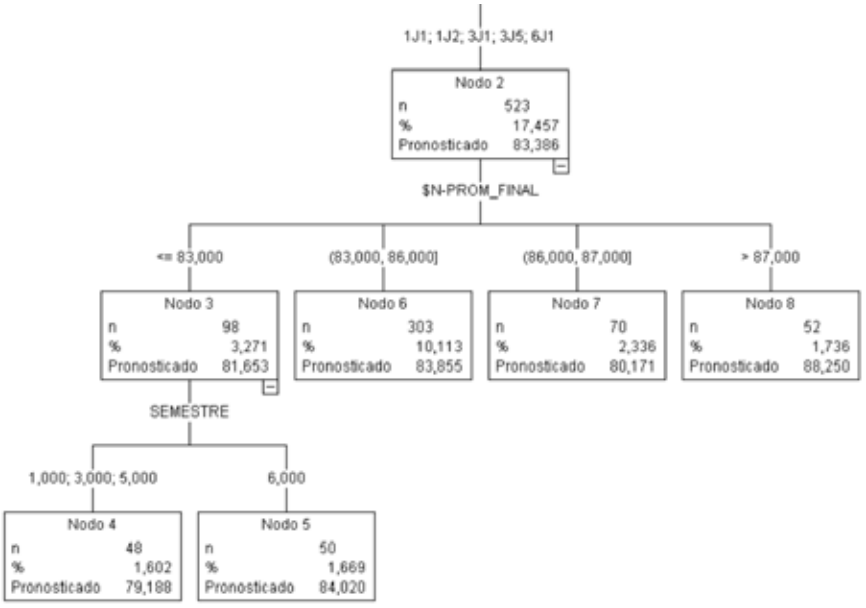

Figura 2

Fragmento del árbol de decisión CHAID aplicado al modelo de RNA
Fuente: Elaboración propia

La descripción de los resultados se realizará más adelante.

2.2 Multilayer perceptron y árbol de decisión C&RT

El algoritmo de árboles de decisión C&RT (Classification and Regression Trees) es un método especializado en trabajar con variables dependientes categóricas.

El método comienza dividiendo la muestra en subconjuntos y evaluando cada predictor cuantitativo para encontrar el mejor punto de corte o cada predictor categórico y las mejores agrupaciones de categorías. A continuación se comparan también los predictores, seleccionándose el predictor y la división que produce la mayor bondad de ajuste. Para predictores cuantitativos suele utilizarse la minimización del error cuadrático o de la desviación media absoluta respecto de la mediana (Pérez, M. 2015). Este algoritmo es utilizado para árboles de clasificación cuando la variable dependiente es de tipo cualitativo y para árboles de regresión cuando la variable dependiente es de tipo cuantitativa. La Figura 3 muestra parte del árbol C&RT aplicado a la RNA utilizada.

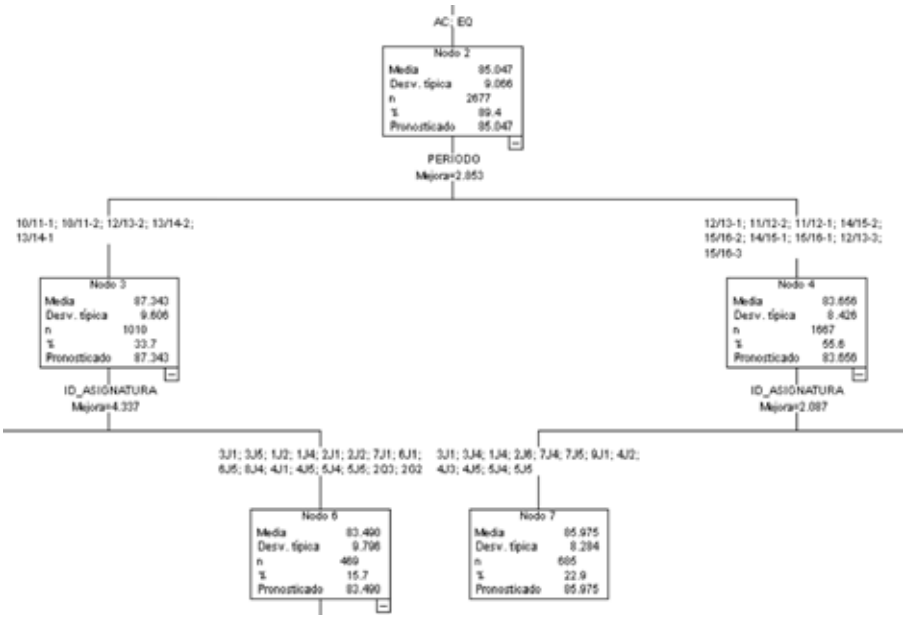

Figura 3

Fragmento del árbol de decisión C&RT aplicado al modelo de RNA Fuente: Elaboración propia

3. Resultados

La evaluación de los modelos obtenidos, así como su análisis y comparación, representan una etapa de vital importancia para el descubrimiento del conocimiento. Es en ésta donde se interpretan y evalúan los resultados alcanzados a fin de comprobar que los objetivos planteados han sido cumplidos.

El primer algoritmo en ser evaluado es el árbol C&RT, el cual muestra lo siguiente:

- Se encontró que en los periodos 10/11/1, 10/11/2, 12/13/2, 13/14/1 y 13/14/2 se logró obtener una relación entre las asignaturas de probabilidad y estadística, investigación de operaciones, métodos numéricos, tópicos avanzados de programación y sistema programables con el aprovechamiento académico más alto de estas materias en todos los periodos.

- Se comprueba la estabilidad relativa hallada por el método anterior en los periodos 10/11/1, 10/11/2 y 13/14/2.

- Se descubre que en los periodos 10/11/1, 10/11/2, 12/13/2, 13/14/1 y 13/14/2 existió una integración mayor en el aprovechamiento académico general, y se comprobó que en dichos periodos se dio el aprovechamiento más alto.

En la Figura 4 se muestra de forma resumida las tendencias encontradas.

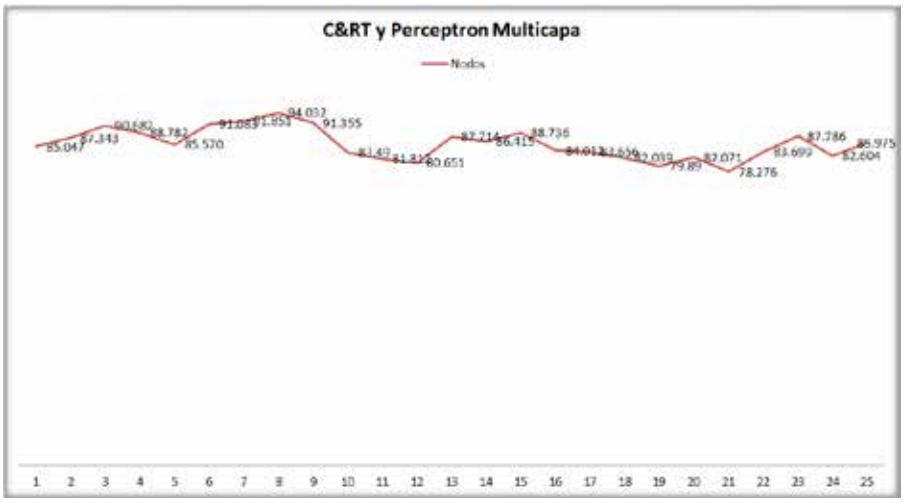

Figura 4

Gráfica estadística de tendencias según árbol C&RT y modelo de RNA. Fuente: Elaboración propia

El siguiente algoritmo a evaluar, es el árbol CHAID, con el cual se obtuvo lo siguiente:

- Se comprueba que el aprovechamiento académico general para los semestres 1, 2, 3, 5 y 6 es impactado directamente por las asignaturas: Cálculo Diferencial, Fundamentos de Programación, Cálculo Vectorial, Sistemas Operativos y Lenguajes y Autómatas I.

- Se encontró una relación estrecha entre las asignaturas Probabilidad y Estadística, Tópicos Avanzados de Programación y Graficación, misma que arroja el aprovechamiento académico más alto de todas las materias, independientemente del periodo.

- El aprovechamiento académico más bajo se dio en la relación de asignaturas de Programación Orientada a Objetos, Ingeniería de Software, Lenguajes y Autómatas II, Programación Lógica y Funcional, Programación Web con las materias de Ecuaciones Diferenciales, Cálculo Diferencial, Integral y Vectorial, debido a que en dichas materias hubo un mayor grado de reprobación y recursamiento, en los periodos 12/13/1, 12/13/2, 13/14/1, 13/14/2, 14/15/1, 14/15/2, 15/16/1 y 15/16/2.

A continuación, en la Figura 5 se muestran las tendencias obtenidas por el árbol CHAID.

Figura 5

Gráfica estadística de tendencias según árbol CHAID y modelo de RNA.
Fuente: Elaboración propia

Conclusiones

Cada una de las técnicas implementadas se caracteriza por tener un nivel de eficiencia y eficacia bastante aceptable, siempre y cuando la base de datos haya sido preparada correctamente en las fases de pre procesamiento. Durante el proceso de clasificación y predicción, se descubrieron relaciones lineales y no lineales entre diferentes asignaturas, evidenciando su impacto en el aprovechamiento académico en distintos periodos y semestres.

El conocimiento obtenido por la RNA y árbol C&RT muestra un pico de aumento en el aprovechamiento académico en los nodos 8 y 9, donde el valor más alto fue de 94.032, mientras que en los resultados de la RNA y árbol CHAID se ejemplifica algo similar en los nodos 26 y 27, donde el valor más alto fue 91.725. En ambos casos, a partir de esos picos, se observa una estabilidad relativa con ligera tendencia de declive.

De acuerdo con los hallazgos obtenidos mediante técnicas basadas en inteligencia artificial para el reconocimiento de patrones, que al compararse demostraron resultados similares, se concluye que existe estabilidad con una ligera tendencia de declive en el aprovechamiento académico en el área de programación.

Referencias

Menéndez, M. y De Luján, M. (s/m, 2012). Informática en el Estado Argentino (presidencia), Sistemas para la toma de decisiones en el ámbito universitario. 6o Simposio, Argentina.

Quinteros, O.; Funes, A.; Ahumada, H. (2016). Extracción de Conocimiento en el Cursado del Ciclo Común de Articulación de Carreras de Ingeniería. XVIII Workshop de Investigadores en Ciencias de la Computación.

Rajni Jindal, R. y Dutta Borah, M. (2013): "A survey on educational data mining and research trends". *International Journal of Database Management Systems.*

Usama, F. (1996). *The kdd process for extracting useful knowledge from volumes of data.* (Eds): Commum ACM, (pp. 27-34).

Pérez, M. (2015). *Minería de datos a través de ejemplos.* México: Alfaomega.

Madrigal Espinoza, S. D. (2014). *Modelos de regresión para el pronóstico de series temporales con estacionalidad creciente.* Computación y Sistemas, (pp 12).

Vivanco, L. (2017). Búsqueda de conocimiento a través de métodos de reconocimiento de patrones. *Technological Innovation Journal in Electromechanical Engineering*, (46-50).

Ponce, P. (2010). *Inteligencia Artificial con aplicaciones en la ingeniería*. México: Alfaomega.

Kamber, J., (2006). *Data Mining: Concepts and Techniques.* USA: Morgan Kaufmann.

Quintín, M., y Santana, Y. (2007). *Aplicación de las redes neuronales artificiales a la regresión*. Madrid: La Muralla.

Viñuela, P., y Galván, I. (2004). *Redes de Neuronas Artificiales. Un Enfoque Práctico.* Madrid: Pearson Educación.

Martín del Brío, B. S. (2007). *Redes neuronales y sistemas borrosos*. Madrid: Alfaomega.

Sierra, B. (2006). *Aprendizaje automático: Conceptos básicos y avanzados.* Pearson Prentice Hall.