Modelo para la Clasificación de los Actores en el Ecosistema de Emprendimiento utilizando Minería de Datos

Diana Nayeli González Vieyra [1, 2], Maricela Quintana López [1], Victor Manuel Landassuri Moreno [1], José Rafael Molina López [2]

[1] Centro Universitario UAEM Valle de México, Maestría en Ciencias de la Computación, Blvd. Universitario s/n Predio San Javier, Atizapán de Zaragoza, Estado de México, México. CP 54500.

[2] Instituto Tecnológico de Tlalnepantla, Dirección Av. Instituto Tecnológico s/n, La Comunidad, Tlalnepantla de Baz, Estado de México, México. CP 54070.

Resumen

El ecosistema de emprendimiento es una opción de desarrollo para los alumnos que se encuentran cursando una carrera, sin embargo, decidir el rol a desempeñar dentro del ecosistema no es una decisión fácil, ya que no tienen clara la diferencia en habilidades y actitudes que posee un emprendedor, un innovador y un investigador. Por ello es necesario contar con una herramienta que ayude al estudiante a elegir el rol, basándose en sus habilidades y actitudes. En este artículo, se presenta la manera en que se utilizó la minería de datos para generar un modelo que permite clasificar al estudiante con un 93.5% de éxito. El modelo muestra que a lo más se requieren 6 preguntas acerca de sus características para orientar al alumno acerca del rol a desempeñar dentro del ecosistema.

Abstract

An entrepreneurship environment is a developmental option for the students who are pursuing a career, however, deciding the role to play within the ecosystem is not an easy decision, since they do not have clear the difference in skills and attitudes between an Entrepreneur, an Innovator, and a Researcher. Thus, a tool that helps the student to choose the role, based on their abilities and attitudes is necessary. In this article, we present the way in which data mining was used to generate a model that allows classifying the student with a 93.5% success. The model shows that at most 6 questions about its characteristics are required to guide the student about the role to play within the environment.

Palabras claves: árboles de decisión, clasificación, ecosistema de emprendimiento, minería de datos.

1. INTRODUCCIÓN

Pertenecer al ecosistema de emprendimiento resulta cada vez más viable como una opción de desarrollo para los egresados de una carrera, por ello las instituciones incluyen en sus planes de estudio materias relacionadas con el desarrollo de proyectos y realizan concursos de innovación o emprendimiento donde pueden mostrarlos y probar lo que el ecosistema ofrece (Cruz, 2017).

Decidir el rol a desempeñar dentro de este ecosistema no es una decisión fácil para los alumnos, sobre todo porque no tienen clara la diferencia en habilidades y actitudes que posee un emprendedor, un innovador y un investigador. La elección representa una decisión importante, ya que si su decisión es la correcta podrían vivir de las actividades que el proyecto genere (Villatoro, 2015).

Por lo anterior, se considera importante tener una herramienta de apoyo que ayude al estudiante a elegir el rol a desempeñar dentro del ecosistema, basándose en sus habilidades y actitudes. La propuesta que se presenta se basa en utilizar técnicas de minería de datos para generar un modelo que sirva para este propósito.
La estructura de este trabajo es la siguiente, en la Sección 2 se presenta la metodología de extracción del conocimiento, mientras que en la Sección 3 los algoritmos de minería de datos empleados son mostrados. En la Sección 4, se presentan los experimentos conducidos, sus resultados e interpretación. Finalmente, la Sección 5, las conclusiones alcanzadas.

2. METODOLOGÍA

La metodología empleada está basada en la de extracción del conocimiento usando minería de datos, KDD (Knowledge Discovery Data Mining) la cual se muestra en la Figura 1. Consiste de 3 grandes etapas: la preparación de los datos, la generación del modelo y la validación e interpretación del mismo (M. Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996).

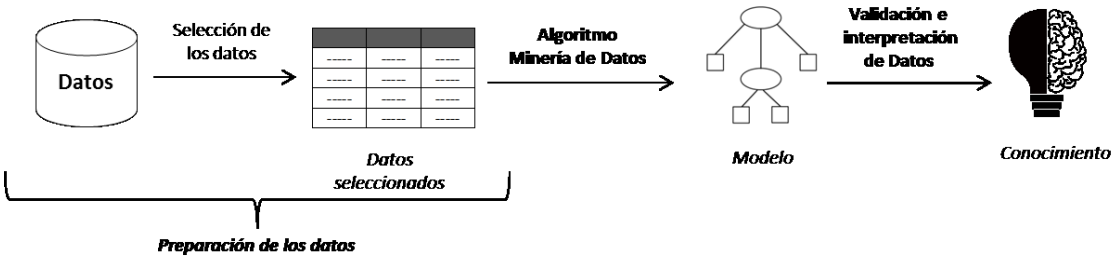

Figura 1. Metodología para la extracción del conocimiento

Preparación de los datos: en esta etapa, se aplicó un cuestionario desarrollado por expertos del Tecnológico Nacional de México (TecNM) que consiste de 21 preguntas, a los participantes del Evento Nacional Estudiantil de Innovación Tecnológica ENEIT 2016, y derivado de su experiencia asignaron un rol a cada uno. De las respuestas recolectadas, se eliminaron aquellas encuestas en las que solo se registraron, pero no respondieron las preguntas, al igual que aquellas en las que los encuestados respondieron a todo la misma respuesta. El total de encuestas seleccionadas fue de 464.

Generación del modelo: para generar el modelo se aplicaron los algoritmos ID3 y J48 que se encuentran disponibles en el software WEKA de la Universidad de Waikato de Nueva Zelanda (Eibe , Mark A. , & H., 2016). La salida es un árbol de decisión que puede examinarse y utilizarse para clasificar nuevas instancias. Los algoritmos se explicarán en la sección 3.

Validación e interpretación del modelo: los datos recolectados se dividieron en dos conjuntos, uno de entrenamiento para generar el modelo, y otro de prueba para validarlo; en algunos casos se utilizó la validación cruzada que genera varias pruebas de manera aleatoria. De los resultados de las pruebas se calcula el porcentaje de éxito y de error.

3. ALGORITMOS DE MINERÍA DE DATOS

Los algoritmos empleados son ID3 y J48, que se describirán sucintamente a continuación.

El algoritmo ID3 fue desarrollado por J. Ross Quilan en 1983, su principal característica es que utiliza la entropía para hacer la separación de las clases y genera un modelo en forma de árbol de decisión. En cada nodo se verifica un atributo y se elige, basado en su valor, el camino que debe seguir, se continúa así hasta llegar a las hojas donde se le asigna la clasificación. Este algoritmo utiliza un enfoque inductivo para generar el modelo para clasificar. (Hernández Orallo, Ramírez Quintana, & Ferri Ramirez, 2004)

El algoritmo C4.5 es una versión mejorada del algoritmo ID3, fue desarrollada en 1993, este algoritmo nos permite utilizar valores numéricos para los atributos, y utiliza la información del radio de ganancia para no favorecer a los que son altamente ramificados. Incorpora una poda del árbol de clasificación una vez que éste ha sido inducido, por lo que los árboles son más pequeños. La implementación en WEKA de este algoritmo se conoce como J48 (Chapman & Hall, 2009).

4. EXPERIMENTOS, RESULTADOS Y ANÁLISIS

Los experimentos consistieron en aplicar los algoritmos ID3 y J48 a las 464 instancias, para generar de manera inductiva un modelo que las clasificará, utilizando la herramienta WEKA. En ambos casos, se utilizaron diferentes porcentajes de partición (entrenamiento-prueba) y también la validación cruzada (con diferentes particiones) con el fin de determinar qué algoritmo resulta mejor al momento de clasificar. Los resultados se muestran en las tablas 1-4.

En la tabla 1, se muestran los resultados del algoritmo ID3 con diferentes porcentajes de partición para el entrenamiento y prueba del modelo. En la primera y segunda columnas se observa el porcentaje de la muestra que se utilizó para entrenamiento (E) y para prueba (P). Los resultados muestran que la mejor opción es utilizar 80% de la muestra para el entrenamiento y 20% para la prueba obteniendo un acierto en la clasificación de 63%.

% E	% P	% Éxito	% Error
80	20	63	37
85	15	60	40
90	10	61	39

Tabla1. Resultados usando ID3 - Particiones

En la tabla 2, se utilizó también el algoritmo ID3 pero usando la validación cruzada, esto es formando las diferentes particiones (k-folds) para entrenamiento y muestra, de forma aleatoria y repitiendo el proceso k veces. El mejor modelo con un 72.25% de éxito se obtuvo con 5 particiones (equivalentes a 80% entrenamiento y 20% para prueba).

ISSN: 2007-4786
Innovación y Desarrollo Tecnológico
Volumen 9 – Número 2
Abril – Junio 2017

Folds	% Éxito	% Error
5	72.25	27.75
7	70	30
10	68.3	31.7

Tabla 2. Resultados usando ID3-Validación Cruzada

Se observa que, en ambos casos, los mejores resultados se obtienen al utilizar 80% instancias para entrenamiento y 20% para prueba.

De manera análoga, se realizaron los experimentos utilizando ahora el algoritmo C4.5 en su versión J48 del WEKA donde los resultados se muestran en las tablas 3 y 4. En ellas se puede observar que los mejores modelos se generaron con una partición de 90% de instancias para entrenamiento y 10% para prueba. En el primer caso, con 93.5% de éxito y en la validación cruzada con un 78.23%.

% E	% P	% Éxito	% Error
80	20	75.5	20.4
85	15	78.6	21.4
90	10	93.5	6.5

Tabla 3. J48 Porcentaje de Partición

Folds	% Éxito	% Error
5	77.15	22.84
7	77.15	22.84
10	78.23	21.76

Tabla 4. J48 Validación Cruzada

Derivado de los resultados, emplear el algoritmo J48 con porcentaje de validación 90% entrenamiento y 10% para prueba, es la mejor opción para generar un modelo para clasificar a los alumnos dentro del ecosistema de emprendimiento.

El modelo, en forma de árbol de decisión, que se obtiene al emplear el algoritmo J48 con una partición 90-10, con éxito de 93.5%. (Figura 2).

En la tabla 5, se muestran los resultados al aplicar el modelo al conjunto de prueba (matriz de confusión). Se observa que, los 15 investigadores fueron clasificados de manera correcta, de los 27 emprendedores se clasifican bien 25, y de los 4 innovadores se clasifican de manera correcta 3 (diagonal de la matriz). Dando un porcentaje de clasificados correctamente de 43 de 46 instancias.

	Investigador	emprendedor	innovador	total		Error
investigador	**15**	0	0	15	100%	0%
emprendedor	1	**25**	1	27	93%	7%
innovador	0	1	**3**	4	75%	25%
	matriz de confusión 43 en la diagonal			46	93.5%	6.5%

Tabla 5. Matriz de confusión

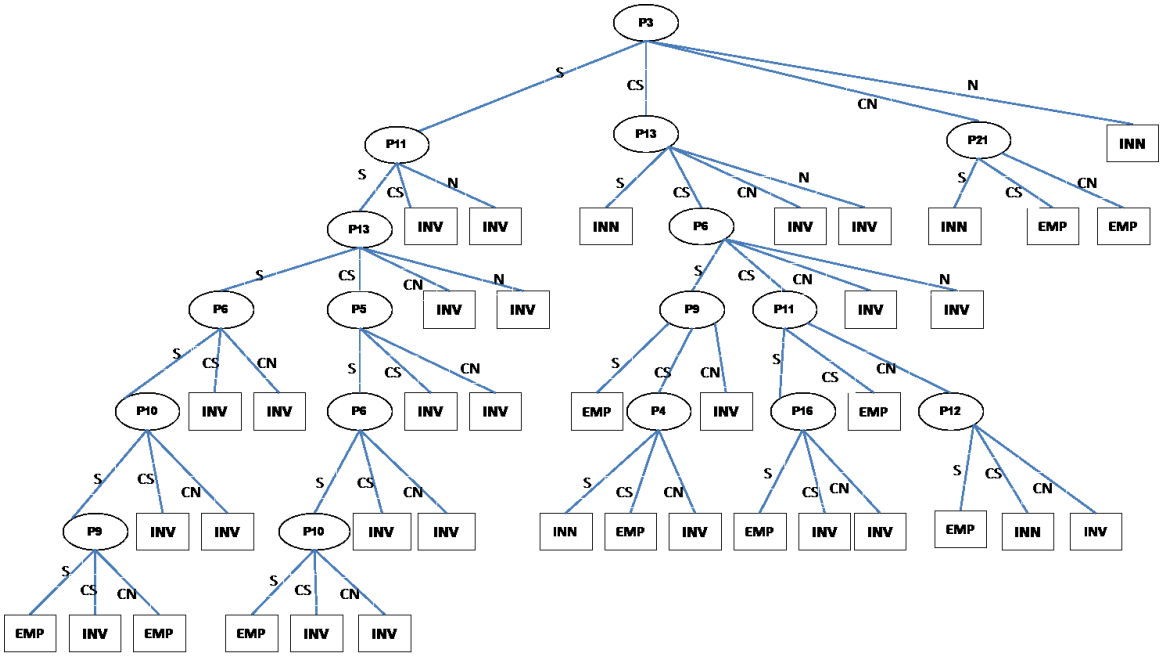

Figura 2. Árbol de decisión usando J48 con partición 90%-10%

Analizando el modelo obtenido, este presenta ciertas ventajas ya que el modelo realiza la clasificación empleando solo 11 preguntas de las 21 (ver tabla 6). También considerando el recorrido más largo, con solo 6 de esas 11 preguntas es posible clasificar al estudiante, lo cual es bueno tomando en consideración la capacidad de retención y el tiempo de estancia del usuario (Fuenmayor & Villasmil, 2008) (Corral, 2010). De las 39 hojas, 10 de ellas clasifican como emprendedor, 24 como investigador y solo 5 como innovador.

En el modelo obtenido (Figura 2), la raíz es P3 (atributo relacionado con el interés de que su nombre se recuerde), dependiendo de qué tan importante es para la persona, será el recorrido a seguir para obtener una de las tres clasificaciones posibles, emprendedor (EMP), innovador (INN) o investigador (INV).

El árbol muestra la confusión que existe entre los actores del ecosistema de emprendimiento, los atributos P6 (que hacen referencia a qué le importa más a la persona si el reconocimiento o el dinero) y P13 (que invertiría su dinero de años de ahorro en un negocio familiar) son atributos que influyen más a la hora de clasificar a una persona como investigador o emprendedor. Si la persona muestra más interés por el reconocimiento, el

modelo lo clasifica como un investigador, de lo contrario como un emprendedor; situación que pasa cuando un experto lo realiza de manera presencial.

Al momento de emplear el modelo se observa que los estudiantes tienen más claro qué camino se debe tomar para realizar investigación o emprendurismo; pero aún desconocen qué es realizar una innovación, es por ello que solo 5 recorridos nos clasifican como innovador.

P3. ¿Has estado interesado en que tu nombre se recuerde en la historia por los avances o descubrimientos que hayas hecho en tu vida?
P4. ¿Tienes actitud reflexiva?
P5. ¿Has pensado ganar un Premio en alguna categoría?
P6. ¿Te importa más el reconocimiento, que el dinero?
P9. ¿Eres capaz de visualizar tus objetivos fácilmente?
P10. ¿Eres capaz de identificar soluciones a las necesidades de tu entorno?
P11. ¿Te gustaría ser tu propio jefe?
P12. ¿Te consideras una persona perseverante?
P13. ¿Invertirías tu dinero de años de ahorro en un negocio familiar?
P16. ¿Tomas riesgos generalmente cuando se trata de cuestiones laborales?
P21. ¿Desarrollas soluciones para un problema?

Tabla 6. Preguntas utilizadas

A pesar de ser 4 las opciones de las respuestas: Siempre (S), Casi Siempre (CS), Casi Nunca (CN), Nunca (N), existen nodos en los que definitivamente no aparece la opción Nunca (N), esto se debe a que ninguno de los encuestados la eligió como respuesta, motivo por el cual, es posible replantear si esa pregunta debe ser modificada o eliminada la opción de respuesta.

5. CONCLUSIÓN

En este trabajo se presentó la metodología de extracción del conocimiento, los algoritmos de minería de datos empleados y los resultados obtenidos en relación al problema que presentan alumnos de nivel licenciatura a la hora de decidir si son Innovadores, Emprendedores o Investigadores (ecosistema de emprendimiento).

Se ha comprobado que la minería de datos empleando el algoritmo J48 es la mejor opción, para determinar el modelo mediante el cual podemos clasificar a un estudiante dentro del ecosistema de emprendimiento con un 93.5% de éxito.

De las 21 preguntas del cuestionario, el modelo generado solo utiliza 11. En el recorrido más largo del árbol, es posible clasificar a un alumno con máximo 6 preguntas.

El modelo generado, refleja que los estudiantes tienen más claro lo que es ser un investigador y un emprendedor (34 hojas), pero en cuanto a la innovación son pocos los que saben al respecto (5 hojas).

Como trabajo futuro se aplicará el modelo obtenido de este trabajo para clasificar a los participantes del ENEIT 2017; también utilizando la misma metodología se obtendrá un nuevo modelo, que se espera incluya más características de los innovadores y se realizará un análisis comparativo entre los dos modelos.

REFERENCIAS

Chapman, & Hall. (2009). *The Top Ten Algoritms in Data Mining.* U.S.A.: CRC Press.

Corral, Y. (2010). Diseño de cuestionarios para recolección de datos. *Revista Ciencias de la Educación, 20*(36), 152-168.

Cruz, A. (25 de 02 de 2017). *La matrícula del Súper Tecnológico TecNM rebasó los 581 mil estudiantes*. Recuperado el 28 de 02 de 2017, de http://www.cronica.com.mx: http://www.cronica.com.mx/notas/2017/1011693.html

Eibe , F., Mark A. , H., & H., I. (2016). *The WEKA Workbench.* Obtenido de Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques": http://www.cs.waikato.ac.nz/ml/weka/book.html

Fuenmayor , G., & Villasmil, Y. (2008). La percepción, la atención y la memoria. *Revista de Artes y Humanidades UNICA, 9*(22), 187-202.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramirez, C. (2004). *Introducción a la Minería de Datos.* España: Editorial Pearson.

M. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* England: The MIT Press.

Villatoro, P. (07 de Septiembre de 2015). *Cómo entender el ecosistema de emprendimiento e inversión en México*. Recuperado el 22 de Enero de 2016, de http://www.forbes.com.mx.

***** Correo autor: *diana_vieyra14@yahoo.com*