Predicción de áreas de investigación de tesistas de carreras de informática de la Universidad de Morón

Iris Sattolo, Gastón Álvarez, Nicolás Armilla, Matías García, Javier Lafont, Mariuz Gabriel, Lucila Mira, Marisa Panizzi

Universidad de Morón. Facultad de Informática, Ciencias de la Comunicación y Técnicas Especiales

iris.sattolo@gmail.com; gaston_alvarez19@hotmail.com; nicolasarmilla@hotmail.com; matias@clustersistemas.com; lafontjavier@hotmail.com; gmariuz91@gmail.com; lucilamira@gmail.com; marisapanizzi@outlook.com

Resumen

En trabajos anteriores se analizó el problema que se les presenta a los tesistas de las carreras de informática de la Universidad de Morón, en el momento de la elección del tema de su tesina de grado, aplicando distintos algoritmos de minería de datos para obtener modelos de clasificación. En este desarrollo, se planteó la obtención de un modelo de predicción mediante redes bayesianas, ya que éstas tienen una semántica muy rica y son fácilmente interpretables. Se presentan los resultados del mapeo sistemático de literatura (SMS), cuyo propósito ha sido identificar los problemas que se resuelven con la minería de datos en el nivel de educación superior. Luego se realizó la comparación de los algoritmos J4.8 y Naïve Bayes como algoritmos de clasificación aplicados a los datos actuales, y se analizó la red obtenida para predecir evidencias futuras. Los resultados expuestos no son concluyentes ya que se sigue recolectando información, pero la descripción del estudio realizado pone en valor el interés de la técnica empleada y sienta las bases para mejorar el alcance de la investigación en trabajos futuros.

Palabras claves: Predicción de Área de elección de tesis, carreras de grado de informática, Minería de datos Educacional, redes bayesianas.

1. Introducción

Los procesos de minería de datos permiten a las organizaciones descubrir conocimiento para la toma de sus decisiones. En los últimos años, se comenzó a aplicar la minería de datos en el dominio educacional con el propósito de resolver diferentes tipos de problemas como, por ejemplo, la deserción y desgranamiento, el rendimiento académico de los alumnos, la clasificación de perfiles de estudiantes, entre otros.

Nuestro grupo de investigación trabajó con los datos obtenidos de las tesis ya rendidas de las materias de fin de carrera en el área de informática, aplicando algoritmos de minería de datos para: caracterizar perfiles de tesistas, trabajo desarrollado en [1] y [2]. La correlación de las áreas de investigación de las tesis, que eligen los tesistas, con las características relevantes sobre los perfiles obtenidos [3]. Al inicio del trabajo se indagó sobre las investigaciones realizadas en el mundo académico en las cuales se utiliza la minería de

datos para la resolución de problemas [13] [14]. Para contrastar nuestra propuesta se desarrolló un mapeo sistemático de la literatura (en inglés Systematic Mapping Studies o SMS), por el cual se identificaron preguntas como: ¿Qué problemas resuelve la Minería de Datos Educacional, que metodologías se utilizan al aplicar minería de datos en instituciones académicas, cuales herramientas, lenguajes de programación y algoritmos se utilizan con mayor frecuencia?

Éste SMS permitió cotejar y sistematizar la evidencia empírica de la aplicación de la minería de datos, en el contexto educacional de Nivel Superior [4].

En vista a los resultados del SMS como también cubrir el objetivo final de la investigación, se decidió trabajar en un modelo predictivo utilizando en esta instancia el algoritmo de Naïve Bayes. Se utilizaron distintas herramientas evaluando las mismas. En esta comunicación se muestran los resultados obtenidos con la herramienta Elvira [7].

2. Contexto

Desde las cátedras de tesis, se ha observado que el mayor inconveniente que posee el alumno al comenzar la materia es la definición del tema, ocasionando un retraso en la finalización de sus estudios, y en algunos casos el abandono de la carrera en su última materia.

Este problema, dio origen al Proyecto de Investigación titulado "Aplicación de tecnologías inteligentes de explotación de información para el análisis de perfiles de tesistas de grado de carreras informáticas de la UM" (Código 17/01-MP-001) financiado por la Secretaria de Ciencia y Tecnología de la Universidad de Morón.

3. Desarrollo

Para el desarrollo del SMS, se utilizó el proceso propuesto por Kitchenham et al. [5] [6]. Este estudio permitió identificar los grupos, institutos o laboratorios de investigación que trabajan en minería de datos para resolver problemas en el contexto de la Educación Superior, en qué tipo de problemas se focaliza, las metodologías o procesos más empleados y las herramientas o lenguajes de programación que se utilizan. Las preguntas de investigación (PI), su motivación (MO) que se plantearon para el SMS se encuentran en la tabla 1.

Preguntas de Investigación (PI)	Motivación (MO)
PI1: ¿Qué se intenta resolver con EDM (Educational Data Mining)	**MO1:** Descubrir que se resuelve con minería de datos en el contexto educacional
PI2: ¿Qué metodologías usan para aplicar minería de datos en instituciones académicas de nivel superior?	**MO2:** Identificar las metodologías más utilizadas en las instituciones académicas de nivel superior.
PI3: ¿Qué herramientas de trabajo y lenguajes de programación se utilizan para realizar minería de datos?	**MO3:** Descubrir las herramientas y lenguajes de programación más empleados.
PI4: ¿Qué algoritmos se aplican?	**MO4:** Determinar qué algoritmos de minería de datos son los más utilizados para resolver problemas en educación de nivel superior.

Tabla 1. Preguntas de investigación (PI) y Motivación (MO).

En los gráficos 1, 2, 3 y 4, se da respuesta a cada una de las preguntas de investigación

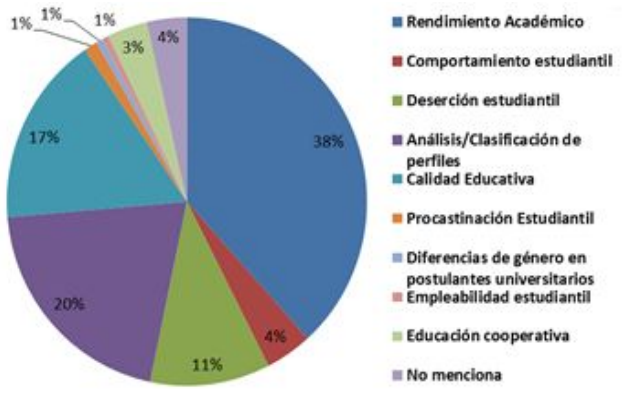

Gráfico 1. Contribuciones de la Minería de Datos en el Dominio Educacional

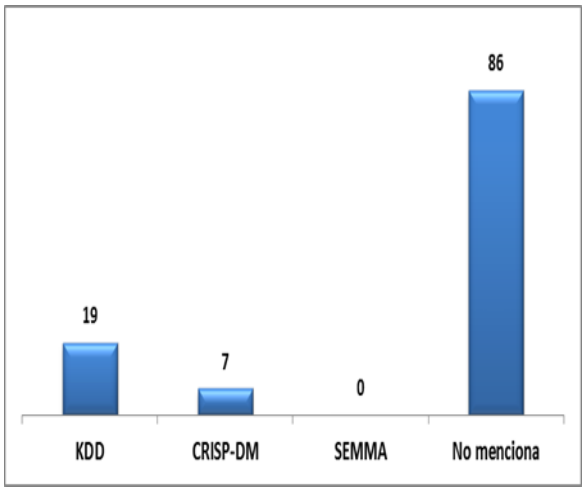

Gráfico 2. Metodologías utilizadas para proyectos de minería de datos

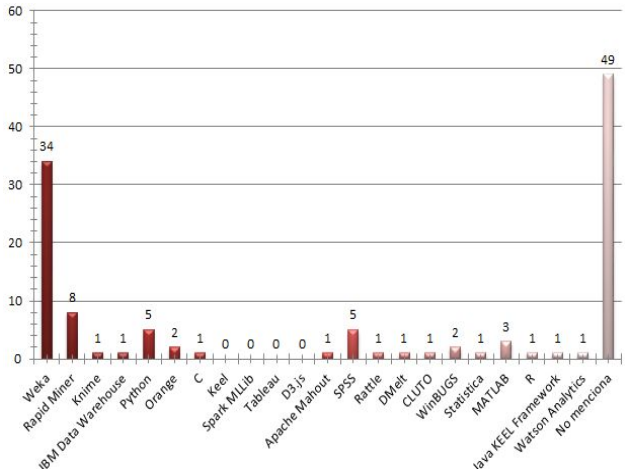

Gráfico 3. Herramientas y lenguajes de programación utilizados

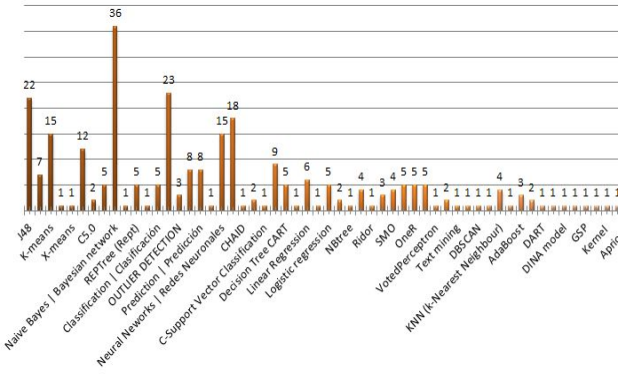

Gráfico 4. Algoritmos utilizados en los artículos analizados

Descubrimiento de la información

De los 112 artículos primarios, se extrajo la siguiente información:

- Los problemas que se resuelven con la aplicación de procesos de minería de datos son de rendimiento académico en un primer lugar, con el objetivo de estudiar el desempeño de un estudiante universitario, ya sea durante en el transcurso o final de una carrera. En un segundo lugar, se aplica con el propósito de clasificar a los estudiantes y poder realizar un análisis sobre los distintos grupos por sus similitudes y diferencias entre ellos. En un tercer lugar, se utiliza para analizar políticas institucionales universitarias, a partir del conocimiento obtenido de la institución como herramienta para la toma de decisiones estratégicas.
- De las metodologías empleadas, se denota que el proceso KDD es el más utilizado y en segundo orden, CRISP-DM.
- De las herramientas y lenguajes de programación que se utilizan para minería de datos, la herramienta más utilizada es Weka. A continuación, Rapid Miner y SPSS. En el grupo de los lenguajes de programación, se

destaca el uso de Python para este tipo de proyecto.

- Del uso de los algoritmos para la construcción de los modelos, se denota un uso significativo de las redes bayesianas, en particular el modelo Naïve Bayes. De esto se deduce una fuerte tendencia hacia la construcción de modelos predictivos. En un segundo lugar, se observa el uso de algoritmos de clustering en general (sin especificar los algoritmos en particular). En un tercer lugar, se menciona el uso del algoritmo J4.8 y los árboles de decisión (sin especificar los algoritmos). El uso de algoritmos de clustering y de árboles de decisión, también muestran la tendencia hacia la construcción de modelos descriptivos.

A partir de los resultados del SMS, correlacionándolos con nuestros desarrollos, se obtuvieron las siguientes conclusiones:

PI1 ¿Qué se intenta resolver con EDM?

No hay en las investigaciones investigadas en el SMS, propuestas sobre caracterización de perfiles de tesistas de carreras de Informática.

PI2 ¿Qué metodologías se utilizan?

En nuestros trabajos se utilizó el proceso KDD el cual es el más empleado en la búsqueda realizada.

PI3 Sobre herramientas y lenguajes de programación, nos encontramos a la par de las mayores utilizadas. Se trabajó con Weka [8], y Rapid Miner [9], pero no se utilizó Python.

PI4 Sobre los algoritmos utilizados para la construcción de modelos descriptivos, nuestras investigaciones se encuentran al mismo nivel que los trabajos reflejados en el SMS, (con algoritmos para selección de atributos [2], árboles de decisión- J4.8, ID3 y CART). Hasta el momento anterior de realizar esta presentación, no se plantearon modelos predictivos.

Proceso KDD

El proceso de descubrimiento de la información está dividido en fases, según Hernández Orallo et al. [7] se divide en:

1) Fase de integración y recopilación de datos
2) Fase de selección, limpieza y transformación
3) Fase de minería de datos
4) Fase de evaluación e interpretación
5) Fase de difusión

Para el avance de nuestro trabajo, en la fase de *integración y recopilación de datos*, se determinaron las fuentes útiles para extraer información. Se utilizó una planilla de cálculo que la cátedra posee con datos de las tesis defendidas de las carreras Licenciatura en Sistemas e Ingeniería en Informática. Para completar la información se construyó un instrumento de recolección de datos denominado TESISTA-UM, su construcción se encuentra en [2]. En la fase de *selección, limpieza y transformació*n, se detectaron valores erróneos o faltantes, corrigiendo los datos incorrectos y decidiendo sobre las estrategias que se aplicarán sobre los datos incompletos. Esta fase se explica con detalle en [2] y [3].

En la fase 3, *fase de minería de datos*, se construyó un modelo preliminar de carácter descriptivo, en cual el objetivo no es predecir nuevos datos, sino describir los existentes. Este modelo permitió identificar las áreas de investigación seleccionadas por los tesistas y su relación con otros atributos que definen al mismo. Para la construcción del modelo, se buscó un conjunto de reglas de asociación entre los atributos (carrera, edad, grupo familiar, área de trabajo) y el atributo objetivo (área de tesis

seleccionada). Se experimentó con los algoritmos de árboles de decisión: J4.8 de Weka [8], ID3 en RapidMiner [9] y CART en Knime [10].

Como es conocido, uno de los problemas a los que se enfrenta la minería de datos es: ¿cómo ocuparse de la incertidumbre? Éste hecho, no representa un problema al trabajar con métodos y técnicas bayesianas, ya que una de sus principales características es el uso explícito de la teoría de probabilidad.

Para concluir con nuestro trabajo, se propuso utilizar un modelo predictivo, utilizando redes bayesianas, las cuales permiten doble uso: descripción y predicción. El algoritmo utilizado fue Naïve Bayes.

Naïve Bayes

El fundamento principal del clasificador Naïve Bayes realizado por Duda & Hart en el año 1973 [6] es la suposición de que todos los atributos son independientes conocido el valor de la variable clase. Por cierto, es que, asumir esta suposición es bastante fuerte y poco realista, pero el clasificador Naïve Bayes (NB), en la mayoría de los casos, como se vio en el SMS, es uno de los clasificadores más utilizados.

La hipótesis de independencia asumida por el clasificador NB da lugar a un modelo gráfico probabilístico en el que existe un único nodo raíz (la clase), y en la que todos los atributos son nodos hojas que tienen como único padre a la variable clase.

En cualquier sistema de clasificación de patrones se tiene: un conjunto de datos representados por atributos y valores, donde el problema consiste encontrar una función que clasifique dichos ejemplos. La idea de usar el teorema de Bayes en cualquier problema de aprendizaje automático es que se puede estimar la probabilidad a posteriori de cualquier hipótesis consistente con el conjunto de datos.

Una red bayesiana [11] es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres, la variable a la que apunta el arco es dependiente (causa-efecto) de la que está en el origen de éste. La topología o estructura de la red nos da información sobre las dependencias probabilísticas entre las variables, pero también sobre las independencias condicionales de una variable (o conjunto de variables) dada otra u otras variables, dichas independencias, simplifican la representación del conocimiento (menos parámetros) y el razonamiento (propagación de las probabilidades).

Existen dos maneras de justificar los enlaces que se introducen u omiten al construir la red. La primera es de naturaleza teórica: se forma un modelo causal a partir de la experiencia de un especialista y se trazan los arcos correspondientes al modelo. El otro camino para justificar la red consiste en realizar una comprobación empírica a partir de un conjunto suficientemente amplio de casos, utilizando las herramientas estadísticas que se emplean para detectar correlaciones. La estructura de la red, por si misma aporta gran cantidad de información cualitativa. En efecto, un arco XY indica, ya antes de conocer el valor concreto de probabilidad condicional, que hay una correlación entre ambas variables: el valor que toma X influye sobre la probabilidad de Y, y viceversa.

Una ventaja de las redes bayesianas es que un mismo nodo puede ser fuente de información u

objeto de predicción dependiendo de cuál sea la evidencia disponible.

La información que se obtiene de una red se puede obtener de dos maneras:

- Obteniendo las probabilidades *a posteriori* de las variables de interés, dado que se conoce el valor que toman algunas otras variables observadas. Este tipo de razonamiento se suele utilizar en sistemas donde se desee realizar un diagnóstico o una predicción.
- Buscando la configuración de las variables que maximicen la probabilidad conjunta dada la evidencia observada. Este proceso se conoce como *abducción* y se utiliza para explicar la evidencia observada.

Para este estadío del proyecto se utilizó la herramienta Elvira [12]. Ésta, permite construir una red a partir de la experiencia del especialista y también provee una interfaz que posibilita la introducción del set de datos, el rellenado de campos vacíos y admite la elección de distintos algoritmos (Naïve Bayes, TAN, KDB).

Experimentación

Es importante aclarar, que durante todo el desarrollo de la investigación se continuó con la recolección de datos a través de nuestro instrumento TESISTAS-UM, lo cual permitió incrementar la cantidad de la muestra. De 114 obtenidas hasta el trabajo en [3] se logró un set de datos de 178. Con este set de datos se realizaron los siguientes experimentos:

Experimento 1

La pregunta que nos orientó en este punto fue: *¿Cómo cambiaron los resultados al incorporar nuevos datos?*

Se trabajó con el software Weka obteniendo nuevamente un modelo de clasificación utilizando el algoritmo J4.8, con cross-validation con 10 carpetas en ambos casos.

Con cross-validation se calcula el porcentaje de aciertos esperado, haciendo una validación cruzada de k hojas (podemos seleccionar el k, que por omisión es de 10)

Los resultados obtenidos se muestran en la tabla 2. En dicha tabla se comparan los resultados entre los 114 y 178 datos.

	Valores con 178	**Valores con 114**
Correctamente clasificados	**48.31%**	48.24 %
Incorrectamente clasificadas	**51.68%**	51.75 %
Tamaño del árbol (nodos)	**53**	33
Tiempo de evaluación	**0,02 seg**	0.02 seg

Tabla 2. Comparación de resultados utilizando el algoritmo J4.8

Cuando aún los valores correctamente clasificados fueron en un porcentaje 0.07% mejores, se observó en este experimento que el algoritmo J4.8 aumentó significativamente la cantidad de nodos que mostraba en su clasificación. De 33 nodos que arrojó la ejecución del algoritmo, se incrementó a 53, dificultando su comprensión.

Experimento 2

Se trabajó con los datos actuales, la pregunta de investigación propuesta: *¿Existe alguna diferencia en la clasificación de datos, pero utilizando esta vez el algoritmo Naïve Bayes?*

Para la respuesta a este cuestionamiento, se utilizó Weka en modo experimenter, donde se realizó una corrida con el mismo set de datos, pero con los algoritmos J4.8 y Naïve Bayes, los dos en cross-validaton con 10 carpetas. El resultado de este se muestra en la tabla 3. Se denota que el algoritmo Naïve Bayes mejora la clasificación.

Datos	**Algoritmo J4.8**	**Algoritmo Naïve Bayes**
178	47,46 %	48,06%

Tabla 3. Comparación de rendimiento de los algoritmos J4.8 y Naïve Bayes.

Experimento 3

La pregunta motivadora en este caso fue: *¿Qué diferencias se encuentran entre los datos?*

De las 178 muestras se eliminaron los atributos no relevantes, detectados en nuestro primer trabajo, luego, utilizando el software Elvira, se aplicó el algoritmo Naïve Bayes obteniéndose la red que se muestra en el gráfico 5, donde la variable clase es el Área de elección de tesis y los nodos hijos son las variables independientes: carrera, edad, grupo familiar, área de trabajo.

El software Elvira da la posibilidad de visualizar la distribución conjunta de probabilidades, al pasar al modo inferencia. En el gráfico 6, se muestran las mismas.

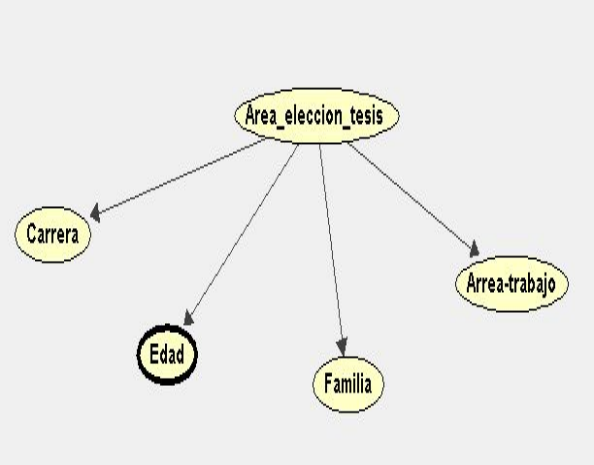

Gráfico 5. Red Bayesiana mostrada en Elvira

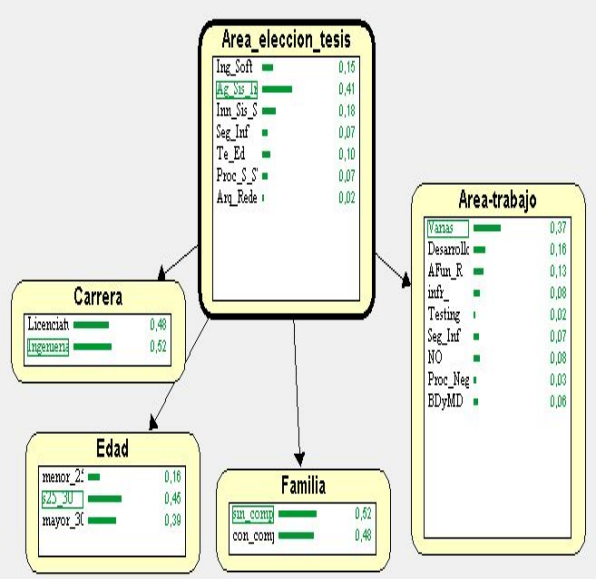

Gráfico 6. Distribución conjunta de probabilidades

Para evaluar si se mantuvo la tendencia que mostraron nuestros datos en la primera comprobación, se confeccionó la tabla 4. En la misma se muestran los nodos (atributos) con los valores actualizados de nuestros datos y los valores que se obtuvieron en nuestro trabajo anterior.

Nodo	Valores con 178 instancias	Valores con 114 instancias
Carrera		
Ingeniería	**52 %**	60 %
Licenciatura	**48 %**	40 %
Edad		
Menores a 25	**16 %**	21%
Entre 25 y 30	**45 %**	46%
Mayores a 30	**39 %**	33%
Familia		
C-compromiso	**48 %**	42%
S-compromiso	**52 %**	58%
Área Trabajo		
Varias	**37%**	47%
Desarrollo	**16%**	17,50%
Analista Func-Req	**13%**	9%
Infraestructura	**8%**	7%
Testing	**2%**	2%
Seguridad-Infor.	**7%**	7%
No-trabaja	**8%**	7%
Proc-Negocio	**3%**	1,5%
BDyMD	**6%**	5%
Á-elección-tesis		
Ag-sis-Intelig	**41%**	40,50%
Inn-sis-Soft	**18%**	16%
Ing-sofware	**15%**	12%
Te-Ed	**10%**	14%
Seg-Inf	**7%**	5%
Proc-Señales	**7%**	8,5%
Arq-Redes	**2%**	3%

Tabla 4. Valores de los nodos con 114 y 178 muestras.

De la tabla 4, se desprende que las tendencias en las áreas elegidas para el desarrollo de la tesis conservan una distribución semejante con los dos conjuntos de datos (Agentes y sistemas Inteligentes, Innovación en Sistemas de Software e Ingeniería de Software).

Experimento 4

Nuestra última pregunta de investigación fue: *¿Qué predicción realiza la red bayesiana?*

La red que muestra el software Elvira puede visualizarse en modo inferencia. En este modo se puede cargar un nuevo caso que actúa como evidencia disponible, desconociendo el valor de la clase; la red predice cual sería la probabilidad de que esa persona elija un área de tesis específica.

En el gráfico 6, se muestra la evidencia cargada en los nodos marcados de color gris y con línea roja el valor del atributo seleccionado, donde: carrera es Licenciatura, edad mayor a 30 años, Familia: con compromiso familiar y trabaja en desarrollo.

La inferencia (predicción) que realiza la red es: la persona puede elegir Agentes y Sistema Inteligentes en un 47% o Innovación en sistemas de software en un 47%

.

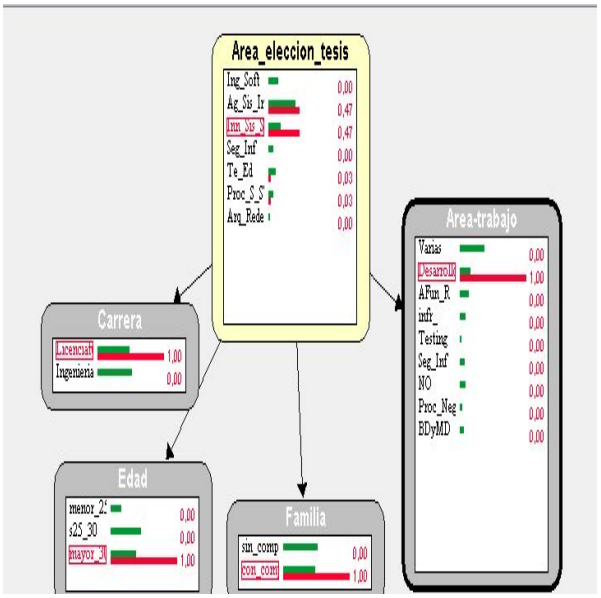

Gráfico 6. Probabilidad a posteriori de una persona que trabaja en desarrollo.

Si cambiamos la evidencia en donde: carrera Licenciatura, edad mayor a 30 años, compromiso familiar, pero trabaja de Analista funcional y Requerimientos, resulta que la red

infiere en el área de elección de tesis es Ingeniería de software en 82% (gráfico 7).
En el gráfico 7, los cuadros en gris muestran los nodos que cargan las evidencias, con las líneas rojas en el valor que toma el atributo seleccionado. La red infiere el atributo Área de elección de tesis (nodo amarillo).

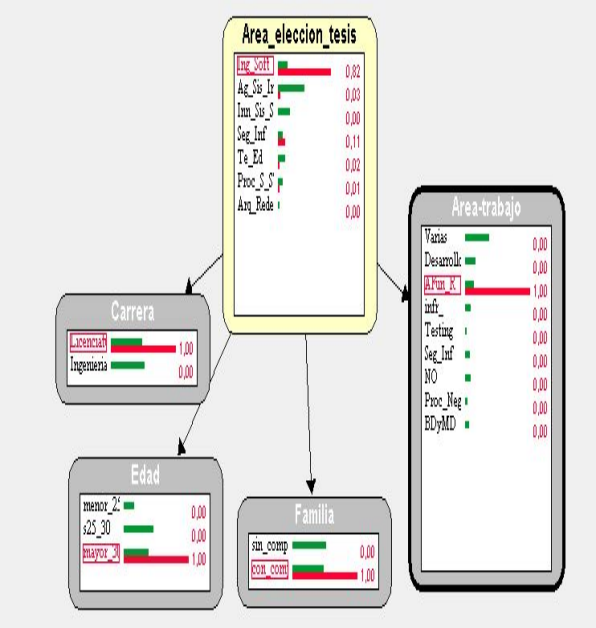

Gráfico 7. Probabilidad a posteriori de una persona que trabaja como analista funcional.

Evaluación e Interpretación

Al incrementar el set de datos (34 %) y evaluarlos con el algoritmo J4.8 se incrementó la dificultad en la interpretación del modelo obtenido ya que se obtuvieron 20 nodos más en el árbol obtenido.
Las redes bayesianas aportan ventajas, gracias a su rica semántica, que permite al usuario entender fácilmente los resultados.
La red obtenida a través de la evaluación empírica de los datos utilizados refleja la realidad que los docentes de las cátedras de tesis de ambas carreras han observado.

3. Conclusiones

El SMS permitió sistematizar la evidencia empírica de la aplicación de minería de datos educacional en el Nivel de Educación Superior. Este sirvió para comparar nuestro trabajo en desarrollo y fue la base para la propuesta del modelo predictivo utilizando el algoritmo Naïve Bayes.
Se ha construido una red que modeliza las relaciones de dependencia e independencia condicional de algunas variables relevantes, que posibilita predecir en un cierto grado que área de tesis elegirá un tesista con algunas características dadas. El modelo obtenido es teóricamente posible y coincide con la evaluación diagnóstica que realizan los docentes de las cátedras en base a su experiencia.

Bibliografía

[1] Marisa Panizzi, Iris Sattolo, Oscar Bravo, Javier Lafont and Nicolas Armilla. Aplicación de tecnologías inteligentes de explotación de información para el análisis de perfiles de tesistas de las carreras de grado de Informática de la Universidad de Morón. Actas de las XXIV Jornadas sobre la Enseñanza Universitaria de la Informática (JENUI 2018), Universitat Oberta de Catalunya, Barcelona.4 al 6 de julio 2018. ISSN: 2531-0607.

[2]Iris Sattolo, Gastón Alvarez, Nicolás Armilla, Oscar Bravo, Matias García, Javier Lafont, Gabriel Mariuz, Lucila Mira, Marisa Panizzi. Hacia la caracterización de perfiles de tesistas de Carreras de Informática de la Universidad de Morón. XIII Congreso Nacional de Tecnología en Educación y Educación en Tecnología (TE&ET 2018). Universidad Nacional de Misiones. Posadas, Misiones. Argentina. 14 y 15 de junio 2018. ISBN 978-950-766-124-2

[3] Iris Sattolo, Gaston Alvarez, Matias Garcia, Javier Lafont, Lucila Mira, Gabriel Mariuz, Nicolás Armilla, Marisa Panizzi.
Descubrimiento de las áreas de investigación seleccionadas por los tesistas de las carreras de informática de la UM mediante árboles de decisión. XXIV Congreso Argentino de Ciencias de la computación (CACIC 2018) Tandil Universidad Nacional del Centro de la Pcia. De Bs.As. ISBN 978-950-658-472-6

[4] Panizzi Marisa, Establecimiento del estado del Arte sobre la Minería de Datos Educacional en el nivel Superior: Un Estudio de Mapeo Sistemático. Revista de Investigaciones Científicas de la Universidad de Morón Nro. 4 Año 2. 2019. ISSN 2591-5444

[5] Kitchenham, B. and Charters, S. (2007) Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE 2007-001, Keele University and Durham University Joint Report

[6] Genero Bocco Marcela, Cruz-Lemus José Antonio y Piattini Velthuis Mario. (2014). Métodos de investigación en ingeniería del software. Madrid, España: Editorial Ra-Ma.

[7] Hernández Orallo José, Ramírez Quintana Maria José, Ferri Ramírez César. Introducción a la Minería de Datos. 1ª Ed. Alhambra. (2004).

[8] Weka. University of Waikato. Machine Learning Group. Página web:
www.cs.waikato.ac.nz/ml/Weka/downloading.html. Disponible online en junio 2018.

[9] RapidMiner Management Team (S/A). RapidMinerStudio. Página Web:
https://rapidminer.com/products/studio/. Disponible online en junio de 2018

[10] Kmine Analytics Platform versión 3.5.3. Página web: www.kmine.com. Disponible online en junio de 2018.

[11] Castillo, Gutiérrez y Hadi. Sistemas Expertos y Modelos de Redes Probabilísticas
https://personales.unican.es/gutierjm/papers/BookCGH.pdf

[12] Elvira software. Página web:
http://www.ia.uned.es/investig/proyectos/elvira/

[13] Lin, S. H. Data mining for student retention management (2012). Journal of Computing Sciences in Colleges, 27(4), 92-99.

[14] Luan, J. (2002). Data mining and its applications in higher education. Article in New Directions for Institutional Research 2002 (113): 17-36.