Descubrimiento de las áreas de investigación seleccionadas por los tesistas de las carreras de informática de la UM mediante árboles de decisión

Iris Sattolo[1], Gaston Alvarez[1], Matias Garcia[1], Javier Lafont[1,] Lucila Mira[1] , Gabriel Mariuz[1],Nicolás Armilla[1], Marisa Panizzi[1]
[1] Universidad de Morón. Facultad de Informática, Ciencias de la Comunicación y Técnicas Especiales. Instituto de Investigación, Desarrollo e Innovación de Sistemas de Información.
Cabildo 134. Morón. Argentina
iris.sattolo@gmail.com; gaston_alvarez19@hotmail.com; matias@clustersistemas.com; lafontjavier@hotmail.com; gmariuz91@gmail.com; lucilamira@gmail.com; nicolasarmilla@hotmail.com; marisapanizzi@outlook.com

Resumen. Las cátedras de las materias de tesis, en las carreras de informática de la Universidad de Morón, cuentan con diferentes herramientas, que brindan apoyo académico a los que comienzan el cursado. Se ha observado que los alumnos tienen dificultad para seleccionar el tema de tesis y, en algunos casos, motiva el abandono o retraso en su carrera. Contar con un registro de tesis, de los graduados, resultó ser útil para obtener patrones sobre la relación entre: el área de investigación abordada en dicha tesis y características que definen al tesista. Para nuestro trabajo se utilizó "el proceso de descubrimiento de conocimiento en base de datos". Finalmente, se presentan los resultados de la utilización de los algoritmos J4.8 en WEKA, ID3 en RapidMiner y CART en Knime y los distintos modelos generados, con los datos recopilados.

Palabras clave: árboles de decisión, WEKA, RapidMiner, Knime, áreas de investigación de tesistas, Carreras de informática UM.

1 Introducción

Las carreras de grado de Informática en la Universidad de Morón (en adelante UM), se encuentran organizadas de manera diferente en sus asignaturas de tesis. La carrera Ingeniería en Informática consta de una asignatura anual, denominada "Tesis de Grado – Proyecto Final Integrador" [1]. La carrera Licenciatura en Sistemas posee dos asignaturas cuatrimestrales, la primera de ellas denominada "Trabajo de Diploma" y la segunda, "Tesis de Grado" [2]. Las cátedras de tesis, de ambas carreras, no cuentan con un instrumento formal que permita a los docentes identificar los saberes que poseen los alumnos, su experiencia laboral, sus tiempos de dedicación a la academia, sus características personales, las líneas de investigación, desarrollo e innovación preferidas, tiempos de permanencia para el desarrollo de sus trabajos de tesis, entre otros. En la actualidad, las cátedras de tesis poseen una planilla de cálculo que contiene la información sobre datos del alumno (apellido, nombre, matrícula y carrera, fecha de defensa de la tesis, docente-tutor o docente-director del trabajo,

título del trabajo, línea de investigación en la cual se inserta el trabajo, resumen, objetivos y futuros trabajos). Este archivo permitió a las cátedras llevar un registro sistemático de un total de 290 tesis, desde el año 2005 hasta de marzo del 2018 [3].

Durante el transcurso de los últimos años, los docentes de las cátedras de las dos carreras observaron que el mayor inconveniente que manifiesta el alumno al comenzar la asignatura es la definición del tema ocasionando un retraso en la finalización de sus estudios y en algunos casos el abandono de la carrera.

En este contexto es que se plantearon las siguientes preguntas de investigación, las cuales impulsaron este trabajo. ¿Son suficientes los datos recolectados de los graduados para determinar patrones de elección del tema de tesis? ¿Se puede correlacionar el área de trabajo de tesis seleccionada, con características tales como área de trabajo, tiempo que disponen para el desarrollo de la tesis, edad al comenzar la tesis, entre otros?

La respuesta a la primera pregunta dio origen a la construcción de un instrumento de recolección de datos denominado TESISTAS-UM y así se completó la información de la cátedra. Se probó su calidad y, por último, se logró una muestra de 114 graduados, compuesta por 46 de la carrera Licenciatura en Sistemas y 68 de la carrera Ingeniería en Informática [4].

Para la segunda pregunta, se decidió utilizar el proceso de "descubrimiento de conocimiento en bases de datos" (en inglés, Knowledge Discovery in Databases, KDD) considerado como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" [5]. Este proceso comprende diversas etapas, reflejadas en la Figura 1, que van desde la obtención de los datos hasta la aplicación del conocimiento adquirido en la toma de decisiones. Entre esas etapas, se encuentra la que puede considerarse como el núcleo del proceso KDD y que se denomina Minería de Datos (en inglés, Data Mining) [6].

Fig. 1. Fases del Proceso KDD [6]

Los problemas tratados en este trabajo se enmarcan en el esquema general del KDD aplicando tareas de clasificación, para obtener un modelo descriptivo, incluidas en la fase de Minería de Datos. En la misma se aplican métodos de aprendizaje para la obtención de modelos y patrones. El aprendizaje siempre será entendido como supervisado, donde los casos pertenecientes al conjunto de datos tienen a priori asignada una clase o categoría, siendo el objetivo encontrar patrones o tendencias de los casos pertenecientes a una misma clase.

La siguiente sección presenta el desarrollo del trabajo, en la cual se describe la metodología empleada y se enuncian los materiales utilizados. Posteriormente, se

presentan los resultados obtenidos de las pruebas realizadas con los algoritmos de árbol de decisión utilizados en diferentes herramientas de trabajo. Se comparten las conclusiones obtenidas y se enuncian futuras líneas de trabajo.

2 Desarrollo

Para el descubrimiento de la información se aplicaron las fases propuestas en Hernández Orallo et al. [6], las cuales se desarrollan en las secciones siguientes.

2.1 Fase de Integración y recopilación

Para la realización de esta fase, se utilizó la planilla de cálculo que la cátedra posee, la cual permitió un registro sistemático de las tesis defendidas y el instrumento de recolección de datos construido [4]. En este estadio del trabajo, se contó con una muestra compuesta por 46 de graduados de la carrera Licenciatura en Sistemas y 68 graduados de la carrera Ingeniería en Informática, totalizando una muestra de 114.

El grupo de investigación reconoce que la muestra no es la aconsejada para el proceso de minería de datos, pero es el resultado de la recolección realizada en un Universidad de gestión privada.

2.2 Fase de selección, limpieza y transformación

Esta fase, conlleva el problema de identificar el conjunto representativo de características adecuadas para construir el modelo. Para la definición de los atributos relevantes a utilizar en la construcción del modelo, se decidió utilizar los resultados de la evaluación realizada en nuestro trabajo anterior y expuesto en TE&ET 2018. Los atributos seleccionados fueron: Carrera, Área de trabajo, Grupo familiar y Edad [4]. Para el atributo clase, se utilizó: área de tesis.

En una primera instancia, para la definición de las áreas de trabajo de las tesis, se realizó un catálogo basado en las áreas de investigación propuestas en el Congreso Argentino de Ciencias de la Computación (CACIC), el cual es organizado por la Red de Universidades con Carreras en Informática (RedUNCI) [7].

Del análisis de datos recolectados, se observó que en el área de Computación Gráfica, Imágenes y Visualización había solamente dos tesis. Dada la escasa elección del área, estas tesis se reagruparon en el área de Innovación en Sistemas de Software por las temáticas abordadas en los mismos. Los temas de tesis que pertenecían a las áreas Innovación en Educación en Informática y Tecnología Informática aplicada a la Educación se reagruparon en un área a la que se denominó Tecnología y Educación por tener las mismas incumbencias propuestas para el Congreso Nacional de Tecnología en Educación y Educación en Tecnología (TE&ET), el cual es organizado por la Red de Universidades con Carreras en Informática (RedUNCI) [7].

El atributo edad originalmente contaba con cuatro rangos (Menor de 25, de 25 a 30, de 31 a 35, Mayor de 35), se reagrupó considerando los siguientes rangos

(Menores de 25 años, Entre 25 y 30 años, Mayores a 30 años) para lograr una distribución más balanceada.

En la recolección de datos, el atributo grupo familiar se definió con cuatro valores (vive-con-padres, solo, hijos y pareja), del análisis de este atributo se logró una reagrupación en dos valores (Con compromiso, Sin compromiso). Esta decisión se fundamenta en la dedicación que el tesista cuenta para con la tesis, el tesista que vive solo (sin hijos) o con sus padres tiene más tiempo respecto a los que son casados o viven en pareja y/o tienen hijos.

La tabla 1, presenta los atributos a utilizar en la fase de DM con sus valores asociados luego de la transformación y limpieza de los datos.

Tabla 1. Atributos y valores para ser utilizados en la fase de Minería de Datos

Atributos	Valores
Área de tesis	Agentes y sistemas inteligentes/ Ingeniería de Software/ Base de Datos y Minería de datos/ Innovación en Sistemas de Software/ Arquitectura, Redes y Sistemas Operativos/ Seguridad Informática/ Tecnología y Educación/ Procesamiento de señales y sistemas en tiempo real
Carrera	Licenciatura en Sistemas/ Ingeniería en Informática
Área de trabajo	Análisis funcional y requerimientos/ Bases de datos y minería de datos/ Desarrollo/Infraestructura/ Procesos de negocio/ Seguridad informática/Testing/Varios/ No trabaja
Edad	Menores de 25 años/ Entre 25 y 30 años/ Mayores a 30 años
Grupo familiar	Con compromiso/ Sin compromiso

2.3 Fase de minería de datos

En este trabajo, se construyó un modelo preliminar de carácter descriptivo, en el cual el objetivo no es predecir nuevos datos, sino describir los existentes [6]. Este modelo debe permitir identificar las áreas de investigación de las tesis seleccionadas por los tesistas y su relación con otros atributos que definen al mismo. Para la construcción del modelo, se trabajó con la tarea descriptiva, reglas de asociación [8] dado que se buscó el hallazgo de reglas de asociación entre los atributos (carrera, edad, grupo familiar, área de trabajo) y el atributo objetivo (área de tesis). Se experimentó con los algoritmos de árboles de decisión: J4.8 de WEKA [9], ID3 en RapidMiner [10] y CART en Knime [11].

2.3.1 Análisis de los algoritmos de árboles de decisión

Los árboles de decisión son una de las formas más sencillas de representación del conocimiento adquirido. Dentro de los sistemas basados en árboles de decisión, habitualmente denominados TDIDT (Top Down Induction of Decision Trees) [12], se pueden destacar dos familias o grupos: la familia ID3, el más representativo es el propio algoritmo ID3 propuesto por Quinlan [12] y la familia de árboles de regresión,

cuyo exponente más significativo es Cart, desarrollado por Breiman et al. [13]. Los TDIDT se caracterizan por utilizar una estrategia de divide y vencerás descendente, es decir, partiendo de los descriptores hacia los ejemplos, dividen el conjunto de datos en subconjuntos siguiendo un determinado criterio de división. A medida que el algoritmo avanza, el árbol crece y los subconjuntos de ejemplos son menos numerosos. [12].

Los algoritmos generan reglas de decisión que son presentadas como un árbol, donde la población total (nodo raíz) es sucesivamente dividida (ramas-nodos intermedios) hasta obtener segmentos de similar comportamiento (nodos hojas) en relación con la variable objetivo. Las hojas contienen la predicción. En cada división se selecciona al predictor que mejor separa a la población con respecto a la variable objetivo. Si la variable objetivo es categórica, se llaman árboles de clasificación, en cambio, si es continua se llaman árboles de regresión [14].

Los diversos algoritmos se diferencian por razones como: naturaleza de los datos a clasificar, número de ramas que pueden dividir, criterios utilizados para la división, administración de los valores faltantes y métodos de poda (simplificación del árbol).

CART genera solo árboles binarios, es decir de cada nodo se desprende exactamente dos ramas. Mientras que J4.8 e ID3 pueden generar más de dos ramas. Si el predictor es de tipo nominal, CART agrupa las categorías en dos y genera solo dos ramas. J4.8 e ID3 por defecto generan una rama por cada categoría, pero además presentan la opción de agrupar categorías para generar menos ramas. Si el predictor es de tipo continuo, CART, J4.8 e ID3 buscan un valor de división y generan solo dos ramas. Si el predictor es ordinal, se pueden ingresar a los algoritmos CART, J4.8 e ID3 como si fuera continuo.

Para definir los criterios de división, se utilizan medidas de pureza para seleccionar los atributos que mejor dividen a las instancias. J4.8 e ID3 utilizan la medida de Razón de Ganancia por defecto (basado en entropía). En WEKA solo se dispone de esta medida, mientras que RapidMiner provee de Information-gain, Gain-ratio, Gini Index, Accuracy. CART utiliza el Índice de Gini por defecto.

Los Métodos de poda utilizados en los algoritmos son los siguientes; J4.8 e ID3 cuenta por defecto con el método de Error Pesimista y dos etapas de poda: local y global. WEKA adicionalmente ofrece el método de Error Reducido. En WEKA se cuenta con la opción de ejecutar ambas etapas, mientras que en RapidMiner solo presenta la etapa local. CART tiene por defecto el método de Costo-Complejidad.

En la tabla 2, se presenta una síntesis de la comparativa de los algoritmos empleados para la construcción del modelo.

Tabla 2. Comparativa de características de los algoritmos de árboles de decisión utilizados

Algoritmo	Variables predictorias	Tipo de división	Criterio de división	Método de poda	Implementación
J4.8	Continuas/Discretas	Binaria/ n-aria	Gain ratio (Entropía)	Pre- /Post-	Libre (Weka)
ID3	Discretas	n-aria	Ganancia de información (Entropía)	No	Comercial

CART	Continuas/Discretas	Binaria	Impureza (índice de Gini)	Post-	Libre Comercial

2.3.2 Construcción del modelo preliminar.

Para la construcción del modelo se utilizaron los algoritmos de árboles de decisión, J4.8 (de WEKA), ID3 (en RapidMiner) y CART (en Knime). Se utilizaron los atributos descriptos en la Tabla 1. Se seleccionó como atributo meta o clase, el área de tesis y se trabajó sin poda.

De la construcción del primer modelo en las tres herramientas, se observaron los siguientes tamaños de árboles: CART (59), J4.8 (45), ID3 (54), además de la compleja visualización e interpretación para la solución a nuestro problema planteado. Los tres algoritmos utilizan como nodo raíz, el atributo Área de trabajo. El atributo carrera en J48 e ID3 es el segundo nodo elegido, mientras que en CART es utilizado en las terceras o cuartas líneas de división.

Luego de estas comprobaciones y en correspondencia con la decisión ya tomada de investigar el problema a través de métodos de clasificación, sumado al interés de las cátedras de poder diferenciar entre ambas carreras, si existen diferencias entre las elecciones de los tesistas se dividió el conjunto de datos en dos. El primero para la carrera Licenciatura en Sistemas y el segundo para la carrera Ingeniería en Informática. La Figura 2, presenta el modelo logrado para la carrera Licenciatura, con el algoritmo J4.8 en WEKA. La Figura 3, presenta el modelo logrado para la carrera Licenciatura en Sistemas, con el algoritmo ID3 utilizando la herramienta RapidMiner La Figura 4, presenta el modelo logrado para la carrera Licenciatura en Sistemas, con el algoritmo CART utilizando la herramienta Knime (los tres modelos sin poda).

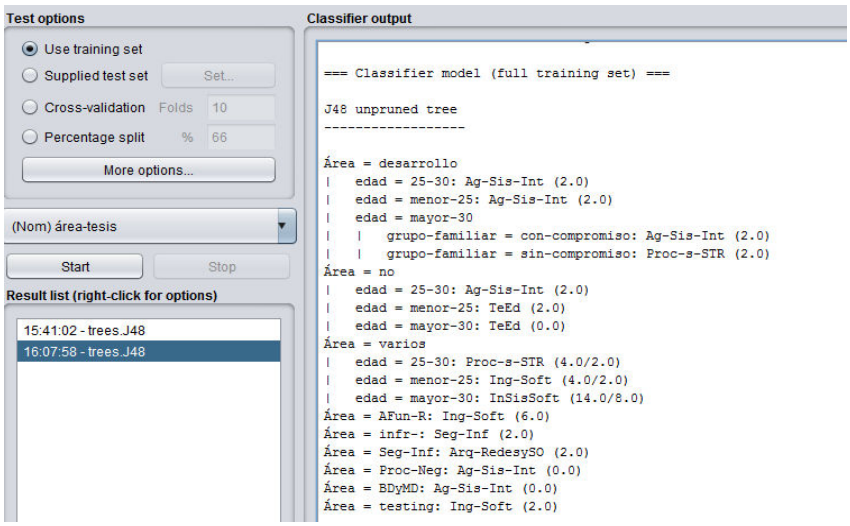

Fig. 2. Modelo para la carrera Licenciatura en Sistemas utilizando el algoritmo J4.8 en la herramienta WEKA.

```
Área = Anal-func-Requ: Ing-Soft (Ing-Soft=6, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0)
Área = Seg-Inf: Arq-RedesySO (Ing-Soft=0, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=2, Proc-Señales-STR=0)
Área = desarrollo
|   edad = 25-30: Ag-Sis-Int (Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0)
|   edad = mayor30
|   |   Grupo-familiar = con-compromiso: Ag-Sis-Int (Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=0)
|   |   Grupo-familiar = sin-compromiso: Proc-Señales-STR (Ing-Soft=0, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-
Inf-Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=2)
|   edad = menor-25: Ag-Sis-Int (Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0)
Área = infraestructura: Seg-Inf (Ing-Soft=0, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=2, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0)
Área = no
|   edad = 25-30: Ag-Sis-Int (Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0)
|   edad = menor-25: Tec-Inf-Aeducacion (Ing-Soft=0, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=2,
Arq-RedesySO=0, Proc-Señales-STR=0)
Área = testing: Ing-Soft (Ing-Soft=2, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-RedesySO=0,
Proc-Señales-STR=0)
Área = varios
|   edad = 25-30
|   |   Grupo-familiar = con-compromiso: Proc-Señales-STR (Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-
Inf-Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=2)
|   edad = mayor30
|   |   Grupo-familiar = con-compromiso: In-Sis-Soft (Ing-Soft=2, Ag-Sis-Int=2, In-Sis-Soft=4, Seg-Inf=2, Tec-Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=0)
|   |   Grupo-familiar = sin-compromiso: In-Sis-Soft (Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=2, Seg-Inf=0, Tec-Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=0)
|   edad = menor-25
|   |   Grupo-familiar = sin-compromiso: Ing-Soft (Ing-Soft=2, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=0)
```

Fig. 3. Modelo para la carrera Licenciatura en Sistemas utilizando el algoritmo ID3 utilizando RapidMiner

```
CART Decision Tree

Area=(varios)|(no)|(infraestructura)|(desarrollo)|(Seg-Inf)
| Area=(varios)|(no)|(desarrollo)|(Anal-func-Requ)|(testing)
| | edad=(mayor30)
| | | Area=(varios)|(Anal-func-Requ)|(Seg-Inf)|(infraestructura)|(no)|(testing)
| | | | Grupo-familiar=(sin-compromiso): Ag-Sis-Int(2.0/2.0)
| | | | Grupo-familiar!=(sin-compromiso): In-Sis-Soft(4.0/6.0)
| | | Area!=(varios)|(Anal-func-Requ)|(Seg-Inf)|(infraestructura)|(no)|(testing)
| | | | Grupo-familiar=(sin-compromiso): Proc-Seales-STR(2.0/0.0)
| | | | Grupo-familiar!=(sin-compromiso): Ag-Sis-Int(2.0/0.0)
| | edad!=(mayor30)
| | | edad=(menor-25)|(mayor30)
| | | | Area=(no): Tec-Inf-Aeducacion(2.0/0.0)
| | | | Area!=(no)
| | | | | Area=(varios)|(Anal-func-Requ)|(Seg-Inf)|(infraestructura)|(no)|(testing): Ag-Sis-Int(2.0/2.0)
| | | | | Area!=(varios)|(Anal-func-Requ)|(Seg-Inf)|(infraestructura)|(no)|(testing): Ag-Sis-Int(2.0/0.0)
| | | edad!=(menor-25)|(mayor30)
| | | | Grupo-familiar=(sin-compromiso): Ag-Sis-Int(4.0/0.0)
| | | | Grupo-familiar!=(sin-compromiso): Ag-Sis-Int(2.0/2.0)
| Area!=(varios)|(no)|(desarrollo)|(Anal-func-Requ)|(testing)
| | Grupo-familiar=(sin-compromiso): Seg-Inf(2.0/0.0)
| | Grupo-familiar!=(sin-compromiso): Arq-RedesySO(2.0/0.0)
Area!=(varios)|(no)|(infraestructura)|(desarrollo)|(Seg-Inf): Ing-Soft(8.0/0.0)

Number of Leaf Nodes: 12

Size of the Tree: 23
```

Fig. 4. Modelo para la carrera Licenciatura en Sistemas utilizando el algoritmo CART utilizando Knime

Si bien el algoritmo J4.8 en la herramienta WEKA es más claro de interpretar, no clasifica correctamente todas las instancias. Por ejemplo: Área=Varios→ edad=25-30→: Proc-s-Str (4.0/2.0) clasifica 4 instancias de las cuales 2 son incorrectas.

ID3 en RapidMiner en Área= Varios → edad= 25-30 → de las 4 instancias distribuye 2 instancias en Procesamiento de Señales y las otras 2, en Agentes y Sistemas Inteligentes.

En CART la regla es más compleja ya que realiza divisiones binarias y para el caso de atributos con varios valores dificulta su lectura. Sin embargo, entre la edad 25-30 en los que trabajan en varias áreas, solo clasifica las instancias de Agentes y Sistemas Inteligentes, dejando sin clasificar las instancias de Procesamiento de Señales. Se concluye que este algoritmo no es recomendable para atributos con varios valores.

2.4 Fase de evaluación e interpretación

En la Tabla 3, se presenta un análisis comparativo del porcentaje que representan las áreas de trabajo de mayor incidencia en la selección del área de tesis.

Tabla 3. Comparativa de la incidencia del atributo área de trabajo en la selección del área de tesis (ambas carreras).

	Total, alumnos	**Ingeniería**	**Licenciatura**
Área Tesis	114	68	46
Innovación en Sistemas de Software	16%	18%	13%
Tecnología y Educación	14%	20%	
Agentes y Sistemas Inteligentes	40,50%	44%	35%
Ingeniería de Software	12%		26%
Total de las 3 áreas	83%	82%	74%
Área de Trabajo			
Desarrollo	17,50%	12%	17%
Varios	47%	47%	48%
Total de las 2 área de trabajo	64,50%	59%	65%

De la tabla 3, se visualiza que el área de tesis más seleccionada es Agentes y Sistemas Inteligentes para ambas carreras. Los tesistas se desempeñaban en desarrollo o en distintas áreas en el momento de inicio del trabajo de tesis.

Con los árboles de decisión obtenidos, se confirma que el área de Sistemas Inteligentes es la más seleccionada con una distribución porcentual no significativa para ambas carreras.

Las tendencias generales encontradas al evaluar los árboles de decisión son las siguientes:

- De la carrera Licenciatura en Sistemas, los árboles de decisión obtenidos muestran que las personas que trabajaban en el rubro Varios, no se inclinan por el área de tesis: Agentes y Sistemas Inteligentes. Tendencia mayoritaria entre todos los tesistas. En cambio, sí la seleccionan los tesistas pertenecientes al área de Desarrollo. Los tesistas que trabajaban como Analistas funcionales, todos se inclinan por el área Ingeniería de Software. Los tesistas que eligen Innovación en Sistemas de Software trabajaban todos en el rubro Varios.

- De la carrera Ingeniería en Informática, se observa que el área de tesis Agentes y Sistemas Inteligentes es elegida tanto por los que trabajaban en Desarrollo, como los que lo hacían en el rubro Varios. Los tesistas que trabajaban en Desarrollo se inclinan también por Innovación en Sistemas de Software. Los que seleccionan Tecnología y Educación trabajaban en las áreas: Analista Funcional, Infraestructura, y Seguridad en Informática.

Los modelos resultantes para cada una de las carreras (Licenciatura en Sistemas e Ingeniería en Informática) formalizan en gran medida la realidad que perciben los docentes de las cátedras de tesis. No obstante, para lograr una mayor cobertura de la realidad respecto a la elección del área de investigación, se debería considerar ampliar la cantidad de atributos que describen al tesista y experimentar con ellos.

3. Conclusiones y futuros trabajos

Los logros alcanzados son solo la etapa preliminar de la investigación que puede realizarse utilizando técnicas de DM. La continuación de este trabajo permitirá avanzar en la utilización de otras técnicas sobre el mismo conjunto de datos y datos futuros que surjan del incremento de la muestra.

Se aplicó el proceso metodológico sugerido en KDD, se resolvieron los problemas encontrados respecto a los datos y a la construcción de los modelos, esto nos permitió comprobar la naturaleza iterativa del proceso.

Se consiguió describir los perfiles de los tesistas aportando información útil, en relación con la incidencia del campo laboral en que se desempeña al inicio de su trabajo de tesis y el área de investigación seleccionada para la tesis.

Como trabajos futuros se identifican: a) el refinamiento de los modelos logrados mediante la incorporación de nuevos atributos para el posterior análisis de su incidencia en la solución del problema; b) la obtención de nuevos modelos, utilizando técnicas de DM aún no abordadas que permitan avanzar en la determinación de las variables incidentes en la selección del área de tesis.

Agradecimientos

La investigación que se en este artículo ha sido financiada por el 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) de la Secretaria de Ciencia y Tecnología de la Universidad de Morón.

Referencias

1. Plan de Estudios de la Carrera Ingeniería en Informática. Universidad de Morón. Facultad de Informática, Ciencias de la Comunicación y Técnicas Especiales. Página web: www.unimoron.edu.ar/area/informatica/stream/af40023e0-ingenieria-en

2. Plan de Estudios de la Carrera Licenciatura en Sistemas. Universidad de Morón. Facultad de Informática, Ciencias de la Comunicación y Técnicas Especiales. Página web: www.unimoron.edu.ar/area/informatica/stream/af40023e1-licenciatura-en
3. 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 ISSN: 2531-0607.
4. 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 en trámite.
5. Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. "From Data Mining to Knowledge Discovery: An Overview" Advances in Knowledge Discovery and Data Mining, pp:1-34, AAAI/MIT. Press.1996.
6. Hernández Orallo José, Ramírez Quintana Maria José, Ferri Ramírez César. Introducción a la Minería de Datos. Ed. Pearson Educacion S.A. Madrid. (2004).
7. Red de Universidades con Carreras en Informática (RedUNCI). Página Web: http://redunci.info.unlp.edu.ar. Disponible online en junio 2018.
8. Waldo Hasperué. Tesis Doctoral en Ciencias Informáticas: ¨Extracción de Conocimiento en Grandes Bases de Datos Utilizando Estrategias Adaptativas¨. Universidad Nacional de La Plata. Facultad de Informática. Marzo 2012. http://sedici.unlp.edu.ar/handle/10915/4215
9. Weka. University of Waikato. Machine Learning Group. Página web: www.cs.waikato.ac.nz/ml/weka/downloading.html. Disponible online en junio 2018.
10. RapidMiner Management Team (S/A). RapidMinerStudio. Página Web: https://rapidminer.com/products/studio/. Disponible online en junio de 2018.
11. Kmine Analytics Platform versión 3.5.3. Página web: www.kmine.com. Disponible online en junio de 2018.
12. J. Quinlan. Induction of decision trees. Machine Learning, 1:81_106, 198. Página web: http://hunch.net/~coms-4771/quinlan.pdf. Disponible online en mayo 2018.
13. L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classi_cation and regresión trees. Wadsworth Int. Group, Belmont, CA, 1984.
14. Álex Moreno-Salazar, M. Puricación Vicente-Galindo & M. Puricación Galindo-Villardón. Aprendizaje basado en árboles de decisión: un estudio crítico desde Weka, RapidMiner y SPSS ModelerXXVI Simposio Internacional de Estadística 2016. Sincelejo, Sucre, Colombia, 8 al 12 de agosto de 2016.