DESCUBRIMIENTO DE PATRONES SOCIO-ECONÓMICOS DE POBLACIÓN ESTUDIANTIL DE CARRERAS DE INGENERIA BASADO EN TECNOLOGÍAS DE EXPLOTACIÓN DE INFORMACIÓN

Laura Cecilia Díaz[1,3], Sebastian Martins[2,3], Ramón García-Martínez[3]

1. Doctorado en Administración y Política Pública del Instituto de Investigación y Formación en Administración Publica, y Facultad de Ciencias Exactas, Físicas y Naturales. Universidad Nacional de Córdoba.
2. Doctorado en Ciencias Informáticas de la Facultad de Informática. Universidad Nacional de La Plata.
3. Grupo de Investigación en Sistemas de Información. Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús.
http://sistemas.unla.edu.ar/sistemas/gisi/
lcd_ic@yahoo.com.ar, smartins089@gmail.com, rgm1960@yahoo.com

Resumen

La detección temprana de factores significativos que faciliten la mejora de los procesos de aprendizaje es un tema de relevancia en Educación Superior en contextos de masividad, en tanto contribuye al establecimiento de políticas que deriven en la mejora de la calidad de los profesionales egresados. En este artículo se presentan los resultados de la utilización de procesos de explotación de información orientados al descubrimiento de patrones de rendimiento académico en los primeros años de carreras de Ingeniería.

Palabras claves: Gestión de la Educación Superior, Carreras de Ingeniería, Estudiantes, Explotación de Información.

1. Introducción

En el multidisciplinar escenario de la Educación Superior en contextos de masividad se debaten cuestiones axiológicas, epistemológicas y metodológicas asociadas con la accesibilidad a los más altos niveles del conocimiento, con la construcción compleja de saberes, en un momento histórico caracterizado por vertiginosos cambios tecnológicos que impactan en las sociedades actuales [Juarros, 2006].

En este escenario, un tema de relevancia es la mejora, tanto en cantidad como en calidad, de profesionales egresados de las carreras denominadas TIC, en particular de las Ingenierías. Ello atento a satisfacer la demanda de especialistas en estas tecnologías. Actualmente la Secretaria de Políticas Universitarias orienta importantes acciones materializadas en Programas de becas a estudiantes como las Becas TIC de finalización de carrera, adquisición de recursos físicos como el Programa de Mejora a la Enseñanza de Grado (PAMEG) y otros más integrales que se dirigen a la carrera en su totalidad como el Programa de Mejora a las carreras de Informática (PROMINF).

La detección temprana de las capacidades de los estudiantes como factor significativo en la mejora del aprendizaje, despierta el interés en las investigaciones sobre tecnologías que permitan predecir su rendimiento académico, en particular en carreras de Ingeniería [Díaz et al., 2013].

La Explotación de Información es la subdisciplina de los Sistemas de Información que aporta las herramientas para la transformación de información en conocimiento [García-Martínez et al., 2015]. Ha sido definida como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información [Britos et al., 2005].

Un Proceso de Explotación de Información se define, como un grupo de tareas relacionadas lógicamente [Curtis et al., 1992] que, a partir

de un conjunto de información con un cierto grado de valor para la organización, se ejecuta para lograr otro, con un grado de valor mayor que el inicial [Ferreira et al., 2005]. Adicionalmente, existe una variedad de técnicas de minería de datos, en su mayoría provenientes del campo del Aprendizaje Automático [García-Martínez et al., 2013], susceptibles de ser utilizadas en cada uno de estos procesos.

La Ingeniería de Explotación de Información (IEI) entiende en los procesos y las metodologías utilizadas para: ordenar, controlar y gestionar la tarea de encontrar patrones de conocimiento en masas de información [Martins, 2014].

El uso de IEI, ofrece la oportunidad de descubrir comportamientos socioeconómicos, académicos, cognitivos, entre otros, de los sujetos en procesos de aprendizaje, que con otras metodologías no serían necesariamente detectados [Kuna et al., 2010].

En este contexto en este articulo se presentan las preguntas de investigación que se formularon los autores sobre rendimiento académico de alumnos de los primeros años de cursos de ingeniería (Sección 2), se describen los materiales y métodos utilizados para el descubrimiento de patrones de comportamiento (Sección 3), se muestran los resultados obtenidos y una interpretación tentativa (Sección 4), y se formulan conclusiones preliminares sobre los hallazgos y se plantean futuras líneas de investigación (Sección 5).

2. Preguntas de Investigación

En el contexto descripto en la Sección 1 se plantea la pregunta general:

> ¿Cómo se caracteriza el rendimiento académico de los estudiantes de Ingeniería en sus primeros años, tomando a la asignatura Informática como eje del análisis?

Se ha desglosado la pregunta general en las siguientes preguntas específicas:

- ¿Qué similitudes socioeconómicas hay entre estos estudiantes? ¿cómo se caracterizan?
- ¿Qué similitudes en relación a su procedencia geográfica? ¿qué características se encuentran en estos grupos?
- ¿Qué distingue a los estudiantes de Córdoba capital, en el universo de los que proceden de la provincia?

3. Materiales y Métodos

Este trabajo no toma como variables representativas del rendimiento académico ni a las calificaciones de la asignatura ni al promedio de calificaciones con y sin aplazos, en razón del sesgo proveniente de las subjetividades de los evaluadores al generar esas calificaciones, y de las diversas normativas vigentes en las distintas unidades académicas. En su lugar, se utilizaron como variables representativas aspectos relativos al desempeño en el primer cuatrimestre académico y al cumplimiento del plan de carrera; siendo éstas más permeables al momento de realizar comparaciones o generar estándares.

En esta sección se describe la base de datos utilizada en la explotación de información (Sección 3.1), se presentan los procesos de explotación de información elegidos (Sección 3.2), y las tecnologías de minería de datos aplicadas en los procesos (Sección 3.3).

3.1. Descripción de la Base de Datos

Se ha utilizado la base de datos proveniente del sistema SIU_Guaraní de alumnos de las carreras de Ingeniería de la Universidad Nacional de Córdoba, inscriptos en la materia Informática en el primer cuatrimestre de los años 2012-2013 relevada en Julio de 2014. Cuenta con más de 1500 registros y contiene variada información del alumno tanto del tipo académico, como socio-económico y de situación geográfica. Del total de variables disponibles las siguientes trece fueron utilizadas en este trabajo:

- La fuente de ingresos del alumno: de su propio trabajo, de su familia y/o de beca.

- Los últimos estudios alcanzados por su padre y madre.
- El género.
- La ubicación de procedencia (generando tres variables booleanas, si es argentino, si es de la Provincia de Córdoba, y si es de Córdoba).
- Si el alumno aprobó Informática durante la cursada.
- Si el alumno realizo la cursada de Informática acorde a lo establecido en el plan de estudios.
- Dos variables que determinan el rendimiento del alumno en su primer año de ingreso y su desempeño en el total de años cursados respecto al plan de estudios.

Es relevante destacar que a partir de la base de datos original (con los datos en crudo) y la selección de las variables representativas para el dominio de interés, fueron realizadas distintas tareas de pre-procesado para adecuar la información a las necesidades y requerimientos específicos del proyecto.

3.2. Procesos de Explotación de Información

Los procesos de explotación de información definen las técnicas o algoritmos a utilizar en base a las características del problema de explotación. En [García-Martínez et al., 2013] se definen 5 tipos de procesos: descubrimiento de reglas de comportamiento, descubrimiento de grupos, descubrimiento de atributos significativos, descubrimiento de reglas de pertenencia a grupos y ponderación de reglas de comportamiento o de pertenencia a grupos. Acorde a los intereses de este trabajo, es relevante describir los siguientes dos procesos:

a) Descubrimiento de reglas de comportamiento:

El proceso de descubrimiento de reglas de comportamiento aplica cuando se requiere identificar cuáles son las condiciones para obtener determinado resultado en el dominio del problema. Para este proceso se propone la utilización de algoritmos de inducción TDIDT para descubrir las reglas de comportamiento de cada atributo clase. Este proceso y sus subproductos pueden ser visualizados gráficamente en la figura 1. Como resultado de la aplicación del algoritmo de inducción TDIDT al atributo clase se obtiene un conjunto de reglas que definen el comportamiento de dicha clase.

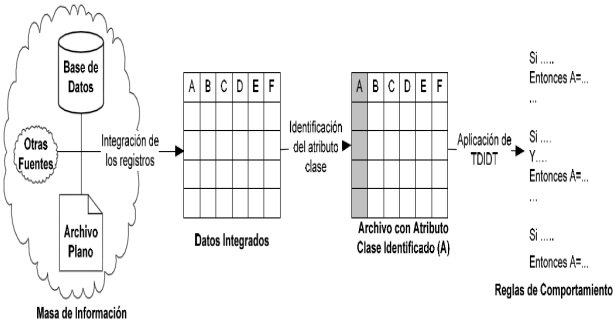

Fig. 1. Proceso de explotación de información: descubrimiento de reglas de comportamiento

b) Descubrimiento de reglas de pertenencia a grupos

El proceso de descubrimiento de reglas de pertenencia a grupos aplica cuando se requiere identificar cuáles son las condiciones de pertenencia a cada una de las clases en una partición desconocida "a priori", pero presente en la masa de información disponible sobre el dominio de problema. Para el descubrimiento de reglas de pertenencia a grupos se propone la utilización de algoritmos de agrupamiento (por ej.: SOM, K-MEANS) para el hallazgo de los mismos y; una vez identificados, la utilización de algoritmos de inducción (por ej.: de la familia TDIDT) para establecer las reglas de pertenencia a cada uno. Este proceso y sus subproductos pueden ser visualizados gráficamente en la figura 2.

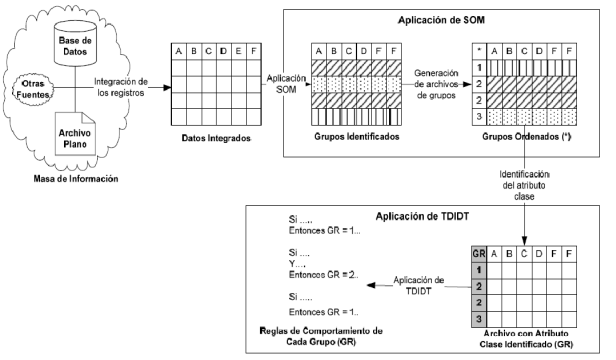

Fig. 2. Proceso de explotación de información: descubrimiento de reglas de pertenencia a grupos

3.3. Tecnologías de Minería de Datos Aplicadas en los Procesos

Los tipos de algoritmos de minería de datos relevantes para el caso de aplicación, se identifican a partir de los procesos de explotación de información aplicados. De dicho análisis se identifican dos tipos de algoritmos: de la familia TDIDT y de clustering.
El algoritmo de clasificación mediante arboles de decisión utilizado es C4.5 [Quinlan, 1993], el cual es descendiente de los algoritmos CLS e ID3. El algoritmo c4.5 clasifica el conjunto de datos mediante la generación de árboles de decisiones, los cuales consisten en una lista de reglas de la forma "si A y B y … entonces clase X" a partir de las cuales se pueden identificar todas las reglas que describen a una clase.
Se utilizaron distintos algoritmos de clustering pertenecientes a distintas familias o tipología de algoritmos, con el objetivo de poder identificar distintas características complementarias a partir del modo en el cual cada tipología de algoritmo comprende el conjunto de datos. Los algoritmos utilizados son: Mapas Auto-Organizados (SOM), K-Means y Hierarchical Aglomerative Clustering (HAC).
El algoritmo SOM [Kohonen, 1995], perteneciente a la familia de algoritmos de agrupamiento basado en modelos, es un tipo de red neuronal la cual utiliza una función de cercanía (o vecindario) de registros con el objetivo de determinar las propiedades topológicas del espacio de entrenamiento.
El algoritmo K-means [MacQueen, 1967], perteneciente a la familia de algoritmos de particionamiento, es un método iterativo simple para particional un conjunto de datos en un numero K de clusters, donde K es un valor definido por el usuario. La idea principal del algoritmo es definir un conjunto K de centroides (uno por cada cluster a identificar), a los cuales se les asocia los puntos (un punto representa un registro) más cercanos. Una vez asignada toda la base de datos, se recalculan la ubicación de los centroides como baricentros de los registros asignados a cada uno, y se vuelve a repetir el proceso de agrupamiento hasta que no se produzcan más cambios en la ubicación de los centroides.
El algoritmo HAC, perteneciente a la familia de algoritmos Jerárquicos, cuyo método de agrupamiento es Bottom-up (de abajo hacia arriba) donde cada registro representa un cluster por sí mismo. Posteriormente, cada cluster es agrupado en clusters más generales de forma sucesiva hasta el nivel deseado, generando un dendograma cuyas agrupaciones se encuentran a una altura similar.

4. Resultados e Interpretación

El objetivo general en esta instancia consiste en realizar un análisis exploratorio de las características socioeconómicas de los estudiantes de Ingeniería, representativa de asignaturas en contextos de masividad en la UNC, en relación a su rendimiento académico.
En la sección 4.1 se presenta una descripción estadística de las variables consideradas para las preguntas de investigación previamente identificadas. Posteriormente se presentan los resultados obtenidos de aplicar los procesos de explotación de información para determinar patrones respecto al desempeño académico a partir de las variables socioeconómicas (sección 4.2), las variables socioeconómicas y de procedencia geográfica (sección 4.3) de la población global y las variables socioeconómicas y de procedencia geográfica para los estudiantes cordobeses. Adicionalmente, se considera relevante nuevamente destacar que se han utilizado distintos tipos de algoritmos de agrupamiento con el objetivo de incorporar visualizaciones complementarias de los datos a partir de las características distintivas de cada algoritmo considerado.

4.1. Análisis Estadístico del Dominio

Inicialmente se proporcionan los resultados estadísticos generales (tabla 1), para todas las variables involucradas en el desarrollo, con el fin de facilitar la lectura posterior de los resultados obtenidos.

Denominación de la variable	Media	DesvSt
'Aprobó Inf. en cursada'	0,31	0,46
'Ritmo Inicial'	1,68	0,95
'Cumple Plan'	3,13	0,94
'Demora en Cursarla'	0,38	0,74
'Beca'	0,07	0,25
'Trabaja'	0,14	0,35
'Familia'	0,92	0,37
'Sexo'	0,74	0,44
'Madre últimos Estudios'	2,42	0,87
'Padre últimos Estudios'	2,20	0,94
'Argentino'	0,98	0,13
'Córdoba'	0,59	0,49
'Capital Córdoba'	0,29	0,45

Tabla. 1. Resultados del procesamiento estadístico de las Variables usadas para la descripción del problema

En la tabla 1, se presentan la media, en adelante Media del Universo (MdU), y el desvió estándar para cada variable seleccionada. Se observa que, del total de la población estudiantil, **aproximadamente**: el 40% cursa Informática el año en que ingresa, el 31% la aprueba en ese período de cursada, la mayoría cursa simultáneamente entre 3 y 4 materias, el 7% accede a beca, el 14% trabaja, el 98% es de nacionalidad argentina, el 59% es de Córdoba y el 29% de Córdoba Capital, el 92% vive con su familia, el 74% es de sexo masculino, el nivel de estudios de la madre es superior al del padre, ambos oscilan entre dos límites: estudios secundarios completos o universitarios incompletos y, estudios universitarios o superiores completos, teniendo en cuenta que cuando se acercan a ambos límites pueden pertenecer a la categoría anterior, estudios secundarios incompletos, o a la posterior, estudios de postgrado.

4.2. Respuesta académica en relación con variables socioeconómicas

Los resultados de la aplicación de HAC clasifican al universo de estudiantes en cuatro conjuntos (C1, C2, C3 y C4):

C1 (105 individuos): Conformado por **estudiantes que no viven con su familia de origen y no tienen beca**. En su mayoría trabajan, cumplen un poco más lento el plan de carreras que la MdU, el nivel de estudios de ambos padres es el más bajo, alcanzan menor porcentaje de aprobación que la MdU durante la cursada, no son ingresantes del año en curso, y poseen la mayor composición masculina.

C2 (284 individuos): **Están cursando la asignatura en un período posterior al año en que ingresaron, viven con su familia y no poseen beca.** Tienen el ritmo más bajo de plan de carrera, ninguno aprueba, el nivel de estudios de ambos padres es levemente inferior a la MdU.

C3 (105 individuos): Casi la totalidad (97%) de **los beneficiarios de beca.** Son los que mejor llevan al día la carrera, y alcanzan el máximo porcentaje de aprobación (57%), sin embargo su ritmo inicial es bueno pero no es el más alto. Es el grupo con mayor componente Femenino (F), el nivel de estudios de la Madre es superior a la media, mientras que el del Padre es inferior a la MdU; el 85% vive con su familia y el 7% trabaja.

C4 (1036 individuos): **Viven con su familia y cursan informática el año en que ingresan.** Prácticamente es nulo el número de becarios, llevan muy bien su plan de carrera, alto porcentaje de aprobación (39%), pocos trabajan (7%) y ambos padres tienen el nivel más alto de estudios.

El análisis de los datos poblacionales utilizando el algoritmo SOM, permite caracterizar grupos (C1_1, C1_2, C2_1, C2_2) que se detallan a continuación:

C1_1 (112 individuos): Conformado por aquellos que **cursan la materia en el primer año, trabajan y casi en su totalidad no alcanza a aprobarla**. No viven con su familia, ambos padres tienen el nivel más bajo de estudios y hay preeminencia de componente Masculino (M). Evidencian un buen ritmo inicial y cumplimiento del plan de carrera, aunque inferior a la media de la población beneficiaria de becas.

C1_2 (328 individuos): **Demoran en cursarla.** Se destacan por tener un alto porcentaje de trabajadores (22%), no aprobar la materia y demorarse en el plan de carrera. Además, en relación a la MdU, acceden a menos becas, es menor el porcentaje de los estudiantes que

vive con su familia e inferior el nivel de estudios alcanzados por ambos padres.

C2_1 (442 individuos): **Cursan la asignatura el primer año, un gran número (el mayor) aprueba en la cursada y no trabajan**. Se destacan por llevar al día la carrera, el ritmo inicial más alto, el mayor porcentaje de becarios, el nivel más alto de estudios alcanzados por ambos padres y viven con la familia.

C2_2 (571 individuos): **Cursan la materia en su primer año, no aprueban durante la cursada, no trabajan y no poseen beca.** Viven con su familia y tienen el menor componente M. Además cumplen mejor el plan de carrera que la MdU y los niveles de estudios alcanzados por ambos padres son inferiores a la MdU.

El algoritmo KMeans Strengthening nos ha dado otra perspectiva del análisis de la formación de grupos, tomando como característica objetivo a la variable "Cumple Plan", como una expresión del desempeño general del estudiante. Los grupos identificados con objetivo 'Cumple Plan' son cinco los cuales se describen a continuación ordenados de manera decrecientes acorde al cumplimiento del plan de carrera:

C5 (462 individuos): **Aprueban durante la cursada y viven con su familia**. Se destacan por tener el porcentaje mínimo de trabajadores, el máximo de becados (12%), no demoran en cursar y ambos padres poseen el más alto nivel de estudios.

C4 (654 individuos): **Cursan el primer año pero no aprueban la asignatura y viven con su familia**. Se destacan por tener el menor porcentaje de becas (4%). Además el ritmo inicial y el componente F son superiores a MdU. El nivel de estudios de ambos padres y el porcentaje de trabajadores son inferiores a la MdU.

C1 (103 individuos): **No aprueban en Cursada y no viven con su familia**. Se destacan por el máximo porcentaje de trabajadores (90%) y el más bajo nivel de estudios de ambos padres. Además demoran en cursar, y tienen un porcentaje de becas ligeramente superior s la MdU (10%). Tienen menor componente F que la MdU y su ritmo inicial es levemente más bajo

C2 (218 individuos): **No cursan la asignatura conforme al plan de carrera y además no aprueban durante la Cursada, viven con su familia y son varones.** Se destacan por tener el ritmo inicial más bajo. Además tienen levemente menor porcentaje de trabajadores (12%), ídem para nivel de estudios de ambos padres y el % de becas (4%), con respecto a la MdU.

C3 (77 individuos): **No cursan la asignatura conforme al plan de carrera y además no la aprueban en la Cursada, viven con su familia, todas mujeres.** Se destacan por el contener el mínimo porcentaje de becas (4%), conjuntamente con C4 y cursar a lo sumo dos materias al inicio de la carrera, conjuntamente con C2. Además, y muy levemente, tienen menor porcentaje de trabajadores e inferior nivel de estudios de ambos padres, en relación a la MdU.

Se realizaron intentos de profundizar el análisis para la población que trabaja y la que es beneficiaria de becas, obteniendo pocos hallazgos significativos.

- Para la población de becarios: Los que llevan mejor la carrera, tuvieron un ritmo inicial mayor a la MdU y aprobaron en cursada.
- Para la población que trabaja: Si cursan la materia con demora de a lo sumo un año y no alcanzan la aprobación, su nivel de cumplimiento de plan es intermedio.
- Además, para ambos grupos, se escogió al atributo 'Aprobó Informática en cursada' como representativo del desempeño académico, obteniéndose algunas reglas más significativas.
- Para los beneficiarios de beca (107 individuos), si cursaron informática sin demoras y actualmente llevan la carrera al día, entonces aprobaron en la cursada (80% de 65). Si demoraron en cursarla, no la aprobaron en cursada o si no demoraron pero no llevan la carrera al día, tampoco aprobaron (67% de 30).

- Para la población que trabaja: si no cumple el plan entonces no aprobó informática en la cursada (es la relación más fuerte).

Otras relaciones halladas, poco significativas, involucran nivel de estudios de los padres, si vive o no con su familia y otros atributos que sería interesante su consideración para realizar indagaciones con metodologías cualitativas de investigación y profundizar desde el paradigma interpretativo.

De lo anterior, se puede observar que en los algoritmos intervienen atributos diferentes para sus procesos de clasificación, esto complejiza la interpretación pero enriquece el análisis. *A través de HAC se logra discriminar el universo de becarios, SOM separa a los que demoran en cursarla, y KMeans a los que aprobaron en la cursada*.

Por otra parte, en las reglas de pertenencia del algoritmo HAC participaron 'Beca', 'Vive con la familia' y 'Demora en cursar la asignatura', mientras que SOM involucró a 'Demora en cursarla', 'Aprobó Informática en cursada', 'Beca' y 'Trabaja', por último, para Kmeans 'Aprobó en cursada', 'Vive con la familia', 'Demora en cursarla' y 'Mujer/Varón'.

4.3. Respuesta académica en relación con variables socioeconómicas y de procedencia geográfica

En este apartado se muestran los resultados e interpretaciones, de los procesos con la misma metodología que en el apartado anterior, incorporando los tres atributos de procedencia geográfica identificados en la sección 3.1.

Los resultados de la aplicación de HAC clasifican al universo de estudiantes en cinco conjuntos (C1, C2, C3, C4 y C5). Las principales variables que participan en la caracterización de los grupos son: 'Beca' (si/no), 'Vive con la familia' (si/no), 'Ritmo inicial' (si/no), 'Trabaja' (si/no) y 'Argentino' (si/no). Este algoritmo clasifica grupos muy eficientemente, e invita a profundizar en su composición. A través del procesamiento estadístico se realizan comparaciones con la misma metodología utilizada en el apartado anterior.

C1 (179 individuos): **Son argentinos, no tienen beca y trabajan**. La mayoría (130) comenzó con buen ritmo, sin embargo 36 individuos, que trabajan y no tienen beca, comenzaron con un bajo ritmo inicial y no vivían con su familia. El máximo porcentaje es de Córdoba y de Capital. El nivel de estudios de ambos padres es el más bajo entre los estudiantes de nacionalidad argentina. Evidencian un nivel muy bajo de aprobación de Informática durante la cursada.

C2 (107 individuos): Conformado por **argentinos, beneficiarios de becas**. Poseen el máximo porcentaje de aprobados en cursada y de cumplir actualmente el plan de carrera. Son altos los niveles de estudios alcanzados por sus padres.

C3 (950 individuos): **Argentinos, que no trabajan y no tienen beca, comenzaron con el más alto ritmo su carrera, sin embargo el grupo anterior mejores porcentajes de aprobación y de mantener el plan de carrera.** En su casi totalidad, viven con su familia. Ambos padres tienen el nivel más alto de estudios.

C4 (26 individuos): Conformado por **los estudiantes extranjeros.** Al mismo porcentaje que la MdU, viven con su familia y trabajan, tienen el máximo componente masculino y el padre con el nivel de estudios más alto, incluso que el de la madre para el mismo grupo y el de todos los grupos. Demoran mucho en cursar la materia (aunque menos que C5). El nivel de estudios de la madre es casi el más bajo (C1 es el menor).

C5 (244 individuos): **Son argentinos, no trabajan ni tienen beca, comenzaron con el más bajo ritmo.** Los más demorados en su plan de carrera, ninguno aprobó Informática en cursada y son los que más demoraron en cursarla con respecto a su año de ingreso. Hay un porcentaje más alto de estudiantes que no son de Córdoba, sin embargo no es significativo con respecto a C3. Casi la totalidad vive con su familia

El análisis de los datos poblacionales utilizando el algoritmo SOM, permite caracterizar seis grupos (C1_1, C1_2, C1_3, C2_1, C2_2, C2_3). Si bien no resulta tan

discriminante como el anterior, se evidencian aportes novedosos. Participan dos variables de rendimiento académico en la caracterización de los grupos: 'Demora en Cursarla' y 'Aprobó Informática en cursada'. Además las variables socioeconómicas participantes resultaron: 'Córdoba capital' (si/no) y 'Trabaja' (si/no). La caracterización de cada grupo se detalla a continuación:

C1_1 (99,70% de 336 individuos): C**ursan la asignatura el primer año de la carrera, aprueban durante la cursada, no son procedentes de Córdoba Capital y no trabajan**

C 1_2 (99,23% de 259 individuos): C**ursan el primer año, son de Córdoba capital y no trabajan.**

C2_1 (98,91% de 460 individuos): También **Cursan el primer año pero no aprueban durante la cursada, no son de Capital y no trabajan.**

C2_2 (132 individuos)**: Cursan el primer año, no aprueban en la cursada, no son de Capital y trabajan** (88,89% de 45 individuos) o **la cursan el primer año y son de Capital** (66,55% de 87 individuos).

C3_1 (97.33% de 262 individuos): **Demoran en cursarla y no trabajan.**

C 3_2 (98,61% de 72): **Demoran en cursarla y trabajan**.

El algoritmo KMeans Strengthening ha proporcionado otra perspectiva del análisis de la formación de grupos, tomando como atributo objetivo a la variable 'Cumple Plan'. Las características que participan en la asociación son: 'Familia' (si/no), 'Demora en cursarla' (si/no) y 'Aprueba en cursada' (si/no), ninguna relativa a la procedencia geográfica. El algoritmo ha identificado 5 grupos (de K_1 a K_5), los cuales se presentan de forma decreciente acorde al cumplimiento del plan de carrera:

k_5 (462 individuos): Todos **aprobaron informática en cursada, con el más alto ritmo inicial de carrera, viven con su familia y ambos padres alcanzaron el máximo nivel de estudios**. Poseen un porcentaje de beca superior a la MdU y el mínimo porcentaje de trabajadores.

K_4 (654 individuos): **Viven con su familia, no demoran en cursarla pero ninguno aprobó.** El nivel estudios de ambos padres y el porcentaje de becarios es inferior a la MdU.

K_1 (121 individuos): (por debajo de la MdU) Compuesto con el **máximo porcentaje de trabajadores, de Capital y de Córdoba y de becados**. El más bajo nivel de estudios alcanzados por ambos padres, no viven con su familia, el ritmo inicial de carrera y el porcentaje de aprobados es inferior a la MdU.

K_2 (219 individuos): Todos **varones que viven con su familia, ninguno aprobó en cursada y su ritmo inicial era cero**. El nivel estudios ambos padres, el porcentaje de trabajadores, de becados y de procedentes de Capital y de Córdoba, son inferiores a la MdU.

K_3 (77 individuos): **Los que más demoran en cursarla, viven con su familia**, **ninguno aprobó informática, todas mujeres, ritmo inicial más bajo.** El nivel estudios de ambos padres y el porcentaje de becados son inferiores a la MdU**.**

Tal vez el mayor aporte en la interpretación de estos resultados, es la caracterización de los extranjeros.

4.4. Respuesta académica en relación con variables socioeconómicas y de procedencia geográfica para los estudiantes cordobeses

Basado en el interés de indagar las características de los estudiantes del interior de Córdoba y de Capital, se repitió el proceso con once atributos: 'Aprobó Inf. en cursada', 'Ritmo Inicial', 'Cumple Plan', 'Demora en Cursarla', 'Beca', 'Trabaja', 'Familia', 'Sexo', 'Madre últimos Estudios', 'Padre últimos Estudios' y 'Capital Córdoba'.

Los resultados de la aplicación de HAC clasifican al universo de estudiantes en cuatro conjuntos (C1, C2, C3, C4):

C_1 (172 individuos): **Demoran en cursar Informática, no tienen beca y unos pocos que no viven con su familia cumplen el plan lentamente (19), los que viven con su familia totalizan 153.**

C_2 (97% de 71 individuos): **La cursan el primer año y tienen beca.**
C_3 (60 individuos): **La cursan el primer año, no tiene beca y no vive con la familia.**
C_4 (590 individuos): **Cursan la asignatura el primer año de la carrera, no poseen beca y viven con la familia.**

El análisis de los datos poblacionales utilizando el algoritmo SOM, permite caracterizar seis grupos (C1_1, C1_2, C1_3, C2_1, C2_2, C2_3), de los cuales a continuación se muestran los que evidencian reglas significativas.

C1_2 (219 individuos): Los que **la cursan el primer año, no tienen beca, no trabajan y la aprueban en cursada.**
C1_3 (70 individuos): Los que **la cursan el primer año y tienen beca.**
C2_1 (123 individuos): **Los más rezagados**.
C2_2 (99% de 313 individuos): La **cursan el primer año, no poseen beca, no la aprueban en cursada y no trabajan.**
C2_3 (93% de 20 individuos): Compuesto por los que la **cursan el primer año, no poseen beca, trabajan, independientemente que aprueben o no la materia** durante la cursada, aunque son muchos más los que no la aprueban.
De KMeans Strengthening con target 'Cumple Plan' presentados en orden decreciente al atributo objetivo:
K_5 (273 individuos): **Aprobaron durante la cursada y, casi la totalidad, vive con su familia**. Se distinguen porque ambos padres alcanzaron el más alto nivel de estudios y por haber cursado la asignatura sin demora alguna.
K_4 (375 individuos): **No aprobaron durante la cursada pero cursaron el primer año, viven con su familia**. El nivel de estudios de ambos padres es el segundo más alto.
K_1 (92 individuos): **No aprobaron en cursada y no viven con su familia (83) o aprobaron, no viven con su familia y trabajan (9)**; Se distinguen por tener el porcentaje más elevado de Estudiantes de Capital. Aunque levemente inferior a los demás, sus padres tienen el nivel de estudios más bajo.
K_2 (114 individuos): **No aprobaron Informática en cursada, viven con su familia, no la cursaron el año de ingreso, y son varones.** El nivel estudios de ambos padres similar a la MdU.
K_3 (44 individuos): **No aprobaron durante la cursada, viven con su familia, no la cursaron el año de ingreso y son mujeres**. El nivel de estudios de ambos padres similar a la MdU.

5. Conclusiones

Múltiples son las interpretaciones que podrían hacerse a partir de los resultados presentados. Las que se realizan en esta presentación responden al objetivo de una primera caracterización de la población en estudio, a instancias de obtener los primeros resultados que dan luz a los procesos decisionales de las políticas públicas destinadas a este universo de estudiantes. En este sentido, no se realizan otras interpretaciones también de relevancia que serían de utilidad para: la base de conocimiento de sistemas tutores inteligentes aplicados en Educación Superior, los actores directamente involucrados en los procesos de enseñanza aprendizaje como docentes, estudiantes, potenciales estudiantes, entre otros.
En el apartado 4.2 se pudo apreciar que cada algoritmo separaba distintos grupos de interés: HAC al universo de becarios, SOM a los que demoran en cursar la Informática, escogida como representativa de asignaturas en contextos de masividad en el UNC, y KMeans a los que la aprueban conforme a lo esperado en el plan de estudios de la carrera. Todos ellos dan respuesta a la primera pregunta específica de investigación, y es posible adecuar su profundidad y extensión al propósito que persigue el actor que se interroga.
En la sección 4.3, al incorporarse variables relacionadas con la procedencia de los estudiantes, los algoritmos caracterizaron a los extranjeros y, además descubrieron relaciones significativas para los estudiantes que proceden de Córdoba que trabajan o no y que aprueban o no la asignatura conforme a lo esperado. Además, tanto para éste como para

el apartado anterior, el mejor desempeño en el plan de carreras lo alcanzan los estudiantes que aprobaron la asignatura durante la cursada, con un bajo porcentaje de trabajadores y alto de beneficiarios de beca, con un alto nivel de estudios de ambos padres. El desempeño más bajo mostró estar relacionado con demoras en el inicio de la carrera, la no aprobación de la asignatura durante la cursada, un bajo porcentaje de beneficiarios de beca y un nivel más bajo de estudios alcanzados por los padres. En la sección 4.4 se pretendió dar respuesta a la última pregunta de investigación. No se encontraron hallazgos novedosos con respecto a los apartados anteriores, con excepción del cluster K_1 cuyo rendimiento es intermedio, compuesto en su mayoría por estudiantes de Córdoba Capital que se caracterizan por no vivir con su familia de origen y además, porque sus padres tienen un nivel de estudios levemente inferior a la MdU, esto es, secundario incompleto.

En apretada síntesis, para todos los grupos descubiertos, el nivel de estudios de la madre es superior al del padre, los estudiantes que poseen beca muestran una tendencia favorable a mejorar el rendimiento académico, el género de los estudiantes no parece tener mayor relevancia en su desempeño y a los estudiantes que trabajan se les dificulta más sostener el plan de carrera al día.

Las líneas de trabajo sugeridas se relacionan con la necesidad de identificar los atributos de índole socioeconómicos que mayor impactan en el rendimiento académico de los estudiantes, incorporar aspectos relativos a sus formas de vida e indagar en otras poblaciones estudiantiles de la Universidad Nacional de Córdoba que resulten también significativas de la Educación Superior en contextos de masividad. A tal fin, se tiene previsto focalizarse en la asignatura Gestión Gubernamental de la carrera de Contador Público de la Facultad de Ciencias Económicas, que reúne las características requeridas y para la cual se cuenta con la disponibilidad a la base de datos del sistema SIU_Guaraní.

6. Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el proyecto SECyT 05/M257 de la Universidad Nacional de Córdoba; y por el proyecto 33A205 de la Universidad Nacional de Lanús.

7. Referencias

Britos, P., Hossian, A., García Martínez, R. y Sierra, E. 2005. *Minería de Datos Basada en Sistemas Inteligentes*. 876 páginas. Editorial Nueva Librería. ISBN 987-1104-30-8.

Curtis, B., Kellner, M., Over, J. 1992. *Process Modelling*. Communications of the ACM, 359: 75-90.

Diaz, L., Algorry, A., Eschoyez, M., Barto, C., Marangunic, R. 2013. *Actions towards the application of intelligent systems in computer education*. IEEE Latin America Transactions, 11(1): 591-595. ISSN 1548-0992.

Ferreira, J., Takai, O., Pu, C. 2005. *Integration of Business Processes with Autonomous Information Systems: A Case Study in Government Services*. Proceedings Seventh IEEE International Conference on E-Commerce Technology. Pág. 471-474.

Garcia-Martinez, R., Britos, P. Martins, S., Baldizzoni, E. 2015. *Ingeniería de Proyectos de Explotación de Información*. Nueva Librería. ISBN 987-1871-34-1.

García-Martínez, R., Britos, P., Rodríguez, D. 2013. *Information Mining Processes Based on Intelligent Systems*. Lecture Notes on Artificial Intelligence, 7906: 402-410. ISSN 0302-9743.

Juarros, M. F. 2006. *¿Educación superior como derecho o como privilegio?: Las políticas de admisión a la universidad en el contexto de los países de la región*. Andamios, 3(5): 69-90. ISSN 1870-0063.

Kohonen, T. 1982. Self Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics. Vol 43.

Kuna, H., García Martínez, R. Villatoro, F. 2010. *Pattern Discovery in University Students Desertion Based on Data Mining*. Advances and Applications in Statistical Sc. J., 2(2): 275-286. ISSN 0974-6811.

MacQueen, J. 1967. *Some methods for classification and analysis of multivariate observations*. 5th Berkeley Symposium on mathematics, Statistics and Probability, 1, S. 281-298.

Martins, S. 2014. *Derivación del Proceso de Explotación de Información Desde el Modelado del Negocio*. Revista Latinoamericana de Ingeniería de Software, 2(1): 53-76. ISSN 2314-2642.

Quinlan, J. R. 1993. C4.5: *Programs for Machine Learning*. San Mateo: Morgan Kaufmann.