9983 INGENIERÍA DE EXPLOTACIÓN DE INFORMACIÓN APLICADA A LA GESTIÓN UNIVERSITARIA: CASO LICENCIATURA EN SISTEMA UNIVERSIDAD NACIONAL DE LANÚS

Santiago Bianco[(1)(2)(4)], Sebastián Martins[(2)(5)], Darío Rodríguez[(2)(3)(6)], Ramón García Martínez[(2)(3)]

[(1)]Maestría en Tecnología Informática Aplicada en Educación
Universidad Nacional de La Plata
[(2)]Grupo de Investigación en Sistemas de Información
Departamento de Desarrollo Productivo y Tecnológico
Universidad Nacional de Lanús
[(3)]Comisión de Investigaciones Científicas de la Provincia de Buenos Aires
http://sistemas.unla.edu.ar/sistemas/gisi/
[(4)]santiago.bianco.sb@gmail.com
[(5)]smartins089@gmail.com
[(6)]darodriguez@unla.edu.ar

Resumen: El abandono de los estudios universitarios en el nivel de pregrado, es un fenómeno global en el Sistema Universitario Argentino, que conlleva la necesidad de desarrollar políticas de retención de estudiantes. En este trabajo, se definen un conjunto de problemáticas comunes de interés orientadas a analizar el desgranamiento de los estudiantes en carreras de grado, con el objetivo definir un proceso estandarizado para extraer patrones de conocimiento que den soporte al proceso de toma de decisiones en materia de política educativa mediante la aplicación de procesos de Ingeniería de Explotación de Información, aplicando los mismos en el caso de la Licenciatura en Sistemas de la Universidad Nacional de Lanús.

Palabras claves: GESTIÓN DE LA EDUCACIÓN UNIVERSITARIA, INGENIERÍA DE EXPLOTACIÓN DE INFORMACIÓN APLICADA A LA EDUCACIÓN.

1. Introducción

En la mayoría de las Universidades públicas de la Argentina, el ingreso es irrestricto para las carreras de grado y pregrado. Cifras oficiales [SPU, 2014] indican que existen alrededor de 1.400.000 de estudiantes universitarios, ingresando por año al sistema alrededor de 330.000 alumnos y egresando por año aproximadamente 80.000 profesionales.

Se verifica un fenómeno muy preocupante a nivel global que es la deserción. Se puede definir a la deserción como el abandono por parte de un alumno de los estudios formales de una determinada carrera [Parrino, 2004]. Para algunos autores [Mansky, 1989], el abandono de una carrera por parte de un alumno no necesariamente es malo, ya que el paso de una persona por las aulas de una Universidad pudo significar un crecimiento personal, pudo agregar conocimientos útiles y aplicables en su vida, entre otros.

El propio concepto de deserción es muy discutido, pero lo que es claro es que para el estado implica un enorme costo, para el estudiante puede significar algún tipo de frustración y requiere de políticas de estado que la prevengan.

Muchos autores presentan distintas hipótesis que tratan de explicar los fenómenos de la deserción y retención de los estudiantes:

▪ Algunos explican estos fenómenos a partir de dos teorías sociológicas: “El modelo de integración del estudiante” [Spady, 1970; Tinto, 1975] donde la integración del estudiante al mundo académico afecta en forma directa a la determinación de abandonar o no los estudios, otro es el “Modelo de desgaste del estudiante” [Bean, 1980] que da relevancia a los factores externos a la institución educativa.

▪ Para [Hackman y Dysinger, 1970] el problema fundamental de la deserción tiene que ver con la ausencia de interés y no con la imposibilidad por parte del alumno de cumplir con los requisitos que la Universidad exige.

▪ Según [Braxton, 2000] de acuerdo a la relevancia que se le otorga a las variables que intentan explicar el fenómeno de la deserción y retención, sean familiares, individuales o institucionales, aborda distintas dimensiones de análisis: Psicológicos, Económicos, Sociológicos, Organizacionales y de Interacción.

▪ Para [Aparicio y Garzuzi, 2006] centra las causales del abandono en temas relacionados con procesos vocacionales.

Existen una variada cantidad de interpretaciones que intentan explicar el fenómeno de la deserción, sin embargo, al tratarse de una compleja problemática, se requiere de un análisis específico del dominio de interés.

El Consorcio SIU [SIU, 2016], es el organismo encargado de desarrollar sistemas informáticos utilizados en la gestión de distintas áreas de las instituciones que componen el sistema universitario nacional argentino. Conformado por más de 40 Universidades Nacionales de la Argentina, provee de un conjunto de herramientas informáticas que acompañan las políticas de Estado y colaboran en la mejora de la gestión. Dichas herramientas permiten realizar estudios analíticos y estadísticos de la información almacenada. La Ingeniería de Explotación de Información, ofrece la oportunidad de extraer información oculta en los datos, novedosa y de interés, como por ejemplo: el descubrimiento de 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]. La Ingeniería de Explotación de Información (IEI), es la sub-disciplina de los Sistemas de Información, que 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]. Esta aporta las herramientas para la transformación de información en conocimiento [García-Martínez et al., 2015] con el objetivo de dar soporte al proceso de toma de decisiones.

En [Romero y Ventura, 2007] se señalan algunos desafíos importantes que diferencian la implementación de estas técnicas, para el dominio específico de la educación, identificándose un en la disciplina un esfuerzo sostenido en el tiempo, en el estudio e implementación de técnicas de IEI aplicadas a la educación (EDM, de sus siglas en inglés Educational Data Mining). EDM [IEDMS, 2017] se define como “una disciplina emergente, relacionada con el desarrollo de métodos para explorar los tipos únicos de datos que provienen del entorno educativo y el uso de esos métodos para entender

mejor a los estudiantes y al entorno en el que aprenden.”. Las principales categorías en las cuales las líneas de investigación se han centrado son [Romero y Ventura, 2010]: análisis y visualización de los datos, detección de comportamientos no deseados, modelado del estudiante, proveer recomendaciones a los docentes, administrativos y/o responsables académicos, predicción del rendimiento de los estudiantes, entre otros, señalándose que las investigaciones en los últimos años, se han centrado principalmente en el estudio del comportamiento de los estudiantes en sistemas educativos (aulas tradicionales, ambientes e-learning, etc), existiendo una incipiente tendencia hacia el análisis de la información para ayudar los sistemas educativos, y la potencial mejora de algunos aspectos de la calidad de la educación y de los procesos de aprendizaje, y el análisis del comportamiento de los estudiantes en cursos y carreras universitarios [Baker e Yacef, 2009].

Además, en [Romero y Ventura, 2010], se señalan como líneas de investigación de interés, el diseño de procesos o métodos estandarizados que faciliten a educadores y/o usuarios no expertos en el área de IEI, la implementación de las técnicas de extracción de conocimiento, las cuales permitan a los usuarios abstraerse de los aspectos específicos de este tipo de algoritmos.

En este contexto, es de interés en este trabajo, la identificación de problemáticas estándares y relevantes para la gestión de carreras universitarias, las cuales permitan identificar un proceso automatizable o semi-automatizable de implementación de técnicas de explotación de información, las cuales permitan evaluar el desempeño de los estudiantes en sus carreras.

En este artículo se presentan las preguntas de investigación que se formularon los autores sobre rendimiento académico de alumnos (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

El abandono de los estudios universitarios en el nivel de pregrado, es un fenómeno global en el Sistema Universitario Argentino, que conlleva la necesidad de desarrollar políticas de retención de estudiantes. Estas políticas requieren la identificación de las posibles causas de deserción y desgranamiento.

En relación a esta problemática, se desprenden dos preguntas de investigación generales:

I. ¿Cuáles son los factores académicos y socioeconómicos que influyen sobre el desgranamiento que afecta a la carrera durante los primeros años de cursada?

II. ¿Cuáles son las características que diferencian a los estudiantes que logran terminar la carrera por sobre aquellos que, si bien tienen la condición de regularidad, no pueden avanzar?

Puede observarse que las preguntas requieren enfocarse en dos momentos distintos de la carrera de los estudiantes. Para contestar la pregunta I es necesario prestar atención a los primeros años de cursada, mientras que para contestar la pregunta II se

precisa un análisis transversal a todos los años de la carrera, de manera que se contemple el avance del estudiante.
Para dar respuesta a dichas problemáticas, se identifican un conjunto de preguntas específicas, a partir de las cuales es posible responder la problemática general. Para la pregunta I, se identifican los siguientes interrogantes:

a. ¿Qué características destacadas poseen los estudiantes regulares con respecto a aquellos que no lo son?
b. ¿Qué características poseen aquellos estudiantes que han podido rendir al menos 2 finales por año en sus primeros años en la universidad?
c. ¿Cómo se caracterizan los estudiantes que han rendido la máxima cantidad de finales posibles en sus primeros años en la carrera?
d. Tomando a aquellos estudiantes que hayan perdido la regularidad o no puedan aprobar al menos una materia por año, ¿qué factores afectan este bajo rendimiento?

De la pregunta II, se desprenden las siguientes preguntas más específicas:

a. En el caso de que la carrera tenga título intermedio, ¿Qué características destacables tienen aquellos estudiantes que consiguen el título de grado con respecto a los que sólo consiguen el título intermedio?
b. ¿Qué factores afectan a aquellos estudiantes regulares pero que no pueden conseguir el título intermedio con respecto a aquellos que sí lo logran?
c. ¿Qué características tienen aquellos estudiantes regulares que no pueden conseguir el título de grado?

3. Materiales y Métodos

Con el objetivo de evaluar la viabilidad de las preguntas de investigación previamente definidas, se utiliza un caso de estudio correspondiente a la carrera de Licenciatura en Sistemas de la Universidad Nacional de Lanús.
En esta sección se describe la base de datos utilizada en la explotación de información (Sección 3.1), se introducen las actividades de procesamiento de los datos realizadas (Sección 3.2) y se presentan las técnicas de IEI utilizadas (Sección 3.3).

3.1. Descripción de la Base de Datos

La base de datos, perteneciente al sistema del sistema SIU Guaraní, cuenta con 1441 registros de estudiantes de la carrera desde el año 2008 al 2016, cuya información está dividida en cuatro tablas que bridan información acerca de los estudiantes:

- **Datos personales**: principalmente domicilio, estudios de los padres, horarios de trabajo, colegio secundario, título secundario, discapacidades, sexo, fecha de nacimiento, año de egreso del secundario.
- **Datos académicos**: plan de estudio, cohorte, regularidad, etc.
- **Exámenes y equivalencias**: materia rendida, tipo de evaluación, fecha de examen, nota y resultado (Aprobado, no aprobado).
- **Datos de cursadas**: materia, nota, resultado (abandonó, regular, libre), fecha de cierre de cursada.

A partir del conjunto de datos disponibles, se identifican aquellas variables relevantes para comprender el comportamiento de la población de estudio. En la tabla 1, describe los campos sociales disponibles.

3.2. Procesamiento de los datos

La implementación de las técnicas de IEI requieren de una etapa de transformación de los datos de acuerdo a las necesidades intrínsecas de las herramientas y para mejorar la comprensión y la calidad de los resultados finales (por ejemplo: creación de variables significativas, imputación de valores, etc.).
En consecuencia, se requiere integrar la información que se encuentra en distintas tablas, representando a cada estudiante como un registro (figura 1).

Tabla 1. Descripción de variables sociales de la base de datos original

Variable	Tipo	Valores	Distribución
Sexo	Discreto	Masc.	86,75% (1250)
		Fem.	13,25% (191)
Nacionalidad	Discreto	Argentino	98,96%(1426)
		Extranjero	0,97% (14)
		Naturalizado	0,07% (1)
Estado civil	Discreto	Soltero	95,48%(1376)
		Casado	2,81% (41)
		Divorciado	0,27% (4)
		Separado	0,34% (5)
		Sin datos	1,03% (15)
Vive con	Discreto	Núcleo familiar	60,65 (874)
		Solo/a	15,4% (222)
		Otros familiares	2,98% (43)
		No familiares	0,48% (7)
		Sin datos	20,48% (295)
Residencia	Discreto	Propia	22,83% (329)
		Alquilada	47,81% (689)
		Familiar	5,84% (84)
		Sin datos	23,52% (339)
Discapacidad visual	Booleano	Sí	1,11% (16)
		No	98,89%(1425)
Discapacidad auditiva	Booleano	Sí	0,07% (1)
		No	99,93%(1440)
Discapacidad motriz inferior	Booleano	Sí	0,07% (1)
		No	99,93%(1440)
Discapacidad motriz superior	Booleano	Sí	0,14% (2)
		No	99,86%(1439)
Discapacidad concentración	Booleano	Sí	0,76% (11)
		No	99,24%(1430)
Discapacidad comunicación	Booleano	Sí	0,49% (7)
		No	99,51%(1434)
Sector trabajo	Discreto	Público	7,16% (103)
		Privado	23,73% (342)
		No trabaja	46,7% (673)
		Sin datos	22,41% (323)
Fecha de nacimiento	Discreto	Con datos	100% (1441)
Cantidad de hijos	Discreto	Con datos	38,1% (549)
		Sin datos	61,9% (892)
Año de ingreso en el colegio secundarios	Discreto	Con datos	100% (1441)
Cohorte	Nominal	Con datos	100% (1441)

Estudios Padre	Ordinal	Con datos	79,04% (1139)
		Sin datos	20,96% (302)
Estudios Madre	Ordinal	Con datos	78,97% (1138)
		Sin datos	21,03% (303)

En adición, se derivaron atributos de mayor valor a partir de las variables disponibles:

- Edad al primer año de cursada: calculada a partir de la fecha de nacimiento y el año de ingreso a la universidad.
-

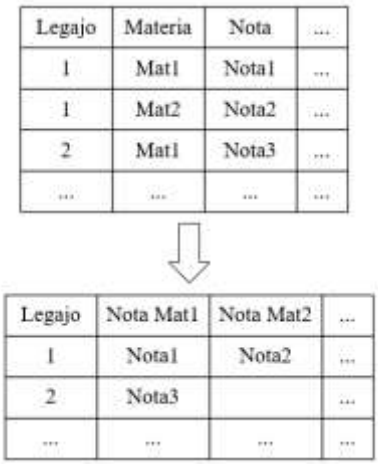

Fig. 1. Conversión de formato de tablas de cursadas y finales.

- Discapacidad: esta es una reducción de distintas variables que indicaban si el estudiante tenía algún tipo de discapacidad (motriz, visual, etc.). Se optó por generalizar todas ellas en una variable booleana.
- Diferencia entre el egreso del secundario e ingreso a la carrera: calculada a partir de los años de egreso del colegio secundario y cohorte.
- Tipo de colegio secundario: puede tomar tres valores: comercial, bachiller o técnico. Se genera a partir del título obtenido en el secundario.
- Trabaja: valor booleano calculado a partir de la existencia del horario de trabajo.
- Categoría de últimos estudios del padre y madre: es un valor continuo de 0 a 5 calculado según los valores de la tabla 2.

La descripción detallada de estas nuevas variables y su distribución en la población pueden observarse en la Tabla 3.

Tabla 2. Categorías consideradas para los estudios de los padres

Estudios del padre/madre	Valor
No registra	0
Hasta secundario incompleto	1
Hasta universitario incompleto	2
Título superior o universitario completo	3
Posgrados finalizados o cursados	4

Tabla 3. Descripción de variables socioeconómicas generadas a partir de las existentes

Variable	Tipo	Valores	Distribución
Edad al primer año de cursada	Discreto	18 a 64	μ=25,31 σ=4,64
Discapacidad	Booleano	Sí	97,78%(1409)
		No	2,22% (32)
Diferencia entre el egreso del secundario e ingreso a la carrera	Discreto	0 a 37	μ=3,54 σ=4,16
Tipo de colegio secundario	Nominal	Técnico	12,77% (184)
		Bachiller	55,59% (801)
		Comercial	31,37% (452)
		Sin datos	0,28% (4)
Trabaja	Booleano	Sí	35,39% (510)
		No	64,61% (931)
Categoría últimos estudios del padre	Ordinal	0	27,34% (394)
		1	38,03% (548)
		2	27,62% (398)
		3	6,18% (89)
		4	0,83% (12)
Categoría últimos estudios de la madre	Ordinal	0	24,64% (355)
		1	29,49% (425)
		2	30,19% (435)
		3	14,84% (185)
		4	2,85% (41)

Finalmente, se crearon variables relativas a la situación académica de los estudiantes en la carrera, los cuales se analizan para el caso de estudio en la Tabla 4:

- *Posible Licenciado:* si ha aprobado todas las materias correspondientes a la carrera de grado con una tolerancia de 3 materias.
- *Posible Analista Programador Universitario (APU):* si ha aprobado todas las materias necesarias para obtener el título intermedio con una tolerancia de ±3 materias.

- *Estudiante regular:* ha rendido al menos dos finales por año.
- *Estudiante activo:* si ha cursado al menos una materia por año.

Tabla 4. Descripción de indicadores de rendimiento académico

Variable	Valores	Distribución
Posible Licenciado	Sí	1,67% (24)
	No	98,33% (1417)
Posible APU	Sí	3,33% (48)
	No	96,67% (1393)
Es Regular	Sí	16,38%(236)
	No	83,62% (1205)
Es activo	Sí	48,44%(698)
	No	51,56% (743)

3.3. Procesos de Explotación de Información

Los Procesos de Explotación de Información (PEI) 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. A partir de las preguntas planteadas en la sección 2, se identifican la aplicación de los siguientes: para las preguntas I.a, II.a y II.b el proceso Descubrimiento de Reglas de Comportamiento (aplicable cuando se requiere identificar cuáles son las condiciones para obtener determinado resultado en el dominio del problema), mientras que para las problemáticas I.b, I.c, I.d y II.c el proceso Descubrimiento de Reglas de Pertenencia a grupos (aplicable cuando se requiere identificar cuáles son las condiciones de pertenencia a cada una de las clases en una partición desconocida “a priori”). Para la valoración de la performance de los resultados, se utiliza el método cross-validation con la configuración de 10 subconjuntos y repeticiones, y como estrategia de optimización del modelo “Grid Search”.

4. Resultados e Interpretación

En esta sección se presentan los resultados obtenidos a partir de la aplicación de los distintos procesos de explotación de información mencionados para resolver las preguntas de investigación relevantes para este caso de estudio.

Pregunta I.a: *¿Qué características destacadas poseen los estudiantes regulares con respecto a aquellos que no lo son?*
Para el análisis de la pregunta, se seleccio-naron a aquellos estudiantes con condición de regularidad y activos en la carrera, excluyendo aquellos registros sin datos de los estudios de los padres, ya que se la considera una variable de interés para el análisis, obteniéndose un total de 545 registros. Las variables utilizadas son:

- Posee alguna discapacidad
- Tipo de secundario
- Edad al primer año de la cursada
- Estudios del padre

- Estudios de la madre
- Diferencia en años desde que terminó el secundario y comenzó la carrera
- Trabaja

A partir del PEI identificado, se identifica al algoritmo C4.5, con la siguiente configuración como optima:

- *Atributo clase:* ¿Es alumno regular?
- *Tamaño mínimo de las hojas del árbol de decisión:* 15
- *Nivel de confianza:* 0.25

Como resultado se obtuvieron las reglas de la figura 2 con la precisión establecida en la matriz de confusión mostrada en la Tabla 5.

SI el estudiante fue a un colegio técnico

Y la diferencia entre el egreso del colegio secundario y la Universidad es menor a 1 año

ENTONCES NO es alumno regular (77% de 22 casos)

SI el estudiante fue a un colegio técnico

Y la diferencia entre el egreso del colegio secundario y la Universidad es mayor o igual a 1 año

ENTONCES SI es un alumno regular (54% de 35 casos)

SI el estudiante se recibió de bachiller

ENTONCES NO es alumno regular (74% de 314 casos)

SI el estudiante fue a un colegio comercial

Y la diferencia entre el egreso del colegio secundario y la Universidad es mayor a 2 años y medio

ENTONCES NO es alumno regular (77% de 114 casos)

SI el estudiante fue a un colegio comercial

Y la diferencia entre el egreso del colegio secundario y la Universidad es menor a 2 años y medio

Y la madre tiene estudios secundarios completos o superiores

ENTONCES SI es alumno regular (56% de 25 casos)

SI el estudiante fue a un colegio comercial

Y la diferencia entre el egreso del colegio secundario y la Universidad es menor a 2 años y medio

Y la madre tiene estudios secundarios completos o inferiores

ENTONCES NO es alumno regular (66% de 35 casos)

Fig. 2. Reglas generadas para la pregunta I.a

Tabla 5. Matriz de confusión para el proceso de la pregunta I.a.

Tasa de Error: 27,71%			
¿Es regular?	NO	SI	TOTAL
NO	361	27	388
SI	124	33	157
TOTAL	485	60	545

Según las reglas las variables más influyentes para determinar la regularidad de un estudiante son los años que pasaron entre el egreso del secundario y el ingreso a la universidad y el tipo de colegio segundario al que asistieron.
Los resultados sugieren una asociación negativa del tiempo entre que el estudiante egreso del secundario e ingreso a la universidad y mantener la regularidad. A su vez, el hecho de haber estudiado en un colegio técnico está asociado con el mantenimiento de la regularidad de los estudiantes.

Pregunta I.b: *¿Qué características poseen aquellos estudiantes que han podido rendir al menos 2 finales por año en sus primeros años en la universidad?*

Para este problema, la población de interés son aquellos estudiantes con condición de regularidad, incluyendo a aquellos con posibilidad de recibirse. Se excluyeron a los registros con datos faltantes en cuanto a los estudios de los padres, obteniéndose un total de 158 casos para analizar. Las variables consideradas son:

- ¿Posee alguna discapacidad?
- Tipo de secundario
- Edad al primer año de la cursada
- Estudios del padre
- Estudios de la madre
- Diferencia en años desde que terminó el secundario y comenzó la carrera
- ¿Trabaja?
- ¿Es posible Licenciado?
- ¿Es posible APU?

En primera instancia, se implementó el algoritmo de agrupamiento K-Means, obteniendo la siguiente configuración como óptima:

- *Número de clusters:* 6
- *Método de normalización de distancia:* varianza

A partir de los resultados obtenidos, se implementa el algoritmo C4.5, con la configuración óptima:

- *Atributo clase:* Marcado del algoritmo K-Means
- *Tamaño mínimo de las hojas del árbol de decisión:* 10
- *Nivel de confianza:* 0.25

Como resultado, se obtiene una descripción de las características de cada grupo que componen la partición original de estudiantes regulares, con una tasa de error del 0,08%. Estos grupos se describen a continuación:

Grupo 1 (43 estudiantes): compuesto por alumnos con menos de 30 años al momento de ingresar a la carrera, que aún no están en condiciones de licenciarse, con ambos padres con estudios secundarios completos o superiores y con título secundario de bachiller o técnico. Es uno de los grupos más grandes e indica que aquellos jóvenes que ingresan con menos de 30 años tienen menos dificultades a la hora de mantener la regularidad. Se observa también una posible influencia de los padres para que esta condición se mantenga.

Grupo 2 (10 estudiantes): compuesto principalmente por estudiantes mayores a 30 años al momento del ingreso.

Grupo 3 (14 estudiantes): es integrado por posibles licenciados. Evidencia el hecho de que es difícil que aquellos que estén por recibirse pierdan la condición de regularidad.

Grupo 4 (43 estudiantes): compuesto por menores a 30 años al momento del ingreso, no son posibles licenciados, tienen título de bachiller o técnico y madres con estudios secundarios completos o inferiores. Refuerza lo obtenido por el Grupo 1, en cuanto a la edad de los ingresantes.

Grupo 5 (25 estudiantes): está compuesto por menores a 30 años en el ingreso, no son posibles licenciados, tienen título comercial y madres con estudios secundarios completos o superiores.

Grupo 6 (23 estudiantes): integrado por menores a 30 años en el ingreso, no son posibles licenciados, tienen título comercial y madres con estudios secundarios completos o inferiores.

Del análisis de todos los grupos se desprende que la población de alumnos regulares está caracterizada principalmente por la edad que los estudiantes tenían al momento de ingresar a la carrera, además de la posible influencia de los estudios alcanzados por la madre.

Pregunta II.a: *¿Qué características destacables tienen aquellos estudiantes que consiguen el título de grado con respecto a los que sólo consiguen el título intermedio?*

La población de interés son aquellos estudiantes cercanos a obtener el título de intermedio, y aquellos recibidos o cerca de recibirse del título de grado, removiendo nuevamente los registros sin datos de los estudios de los padres, resultando en 24 casos.

De los mismos, de consideraron para el análisis las siguientes variables:

- ¿Posee alguna discapacidad?
- Tipo de secundario
- Edad al primer año de la cursada
- Estudios del padre
- Estudios de la madre
- Diferencia en años desde que terminó el secundario y comenzó la carrera

Para la ejecución se utilizó el algoritmo C4.5, obteniendo la siguiente configuración:

- *Atributo clase:* ¿Es un posible licenciado?
- *Tamaño mínimo de las hojas del árbol de decisión:* 4
- *Nivel de confianza:* 0.25

Como resultado se obtuvieron las reglas de la figura 3 con la tasa de error establecida en la matriz de confusión mostrada en la Tabla 6.

Si bien los registros resultantes son reducidos, se identifican como variable más influyente entre aquellos que terminan la carrera o sólo obtienen el título intermedio, los estudios de la madre.

SI el estudiante era menor de 23 años al ingresar a la universidad
Y la madre tiene estudios secundarios incompletos o inferiores
ENTONCES NO es posible licenciado (71,43% de 7)
SI el estudiante era menor de 23 años al ingreso
Y la madre tiene estudios secundarios completos o superiores
ENTONCES SI es un posible licenciado (84,62% de 13)
SI el estudiante era mayor de 23 años al ingreso
ENTONCES NO es un posible licenciado (100% de 4)

Tabla 6. Matriz de confusión para el proceso de la pregunta II.a.

Tasa de Error: 16,67%			
¿Es posible grauado?	NO	SI	TOTAL
NO	9	2	11
SI	2	11	13
TOTAL	11	13	24

De los resultados expuestos, puede observarse que aquellos estudiantes que hayan obtenido el título de analista programador universitario, hayan comenzado la carrera con menos de 23 años y tengan madres con estudios secundarios completos son más propensos a terminar la carrera de grado una vez alcanzado el título intermedio.

Pregunta II.b: *¿Qué factores afectan a aquellos estudiantes regulares pero que no pueden conseguir el título intermedio?*

La población de interés son aquellos estudiantes cercanos a obtener el título de intermedio que manteniendo la condición de regularidad. Se removieron los registros sin datos de los estudios de los padres, resultando en 158 registros. Las variables de interés son:

- ¿Posee alguna discapacidad?
- Tipo de secundario
- Edad al primer año de la cursada
- Estudios del padre
- Estudios de la madre
- Diferencia en años desde que terminó el secundario y comenzó la carrera

Para la ejecución se utilizó el algoritmo C4.5 con la siguiente configuración óptima:

- *Atributo clase:* ¿Es un posible Analista Programador Universitario?

- *Tamaño mínimo de las hojas del árbol de decisión:* 4
- *Nivel de confianza:* 0.25

Para esta pregunta, no se obtuvieron resultados de interés, identificando como variable descriptiva la moda del atributo clase (tabla 7).

Tabla 7. Matriz de confusión para el proceso de la pregunta II.b.

Tasa de Error: 16,46%			
¿Es posible analista?	NO	SI	TOTAL
NO	132	0	132
SI	26	0	26
TOTAL	158	0	158

En [Bianco et al., 2017] se presentan los resultados obtenidos para las problemáticas I.c, I.d y II.c, a partir de los cuales se fortalecen la asociación entre el rendimiento académico de los estudiantes, los estudios académicos de los padres, el ingreso de la carrera y el tiempo transcurrido entre que terminaron el secundario y comenzaron a estudiar en la universidad.

5. Conclusiones

En el presente artículo se introdujeron dos problemáticas generales de interés, y siete preguntas de investigación específicas, orientadas a analizar el desgranamiento de los estudiantes en carreras de grado, con el objetivo de estandarizar las problemáticas que pueden abordarse mediante la tecnología IEI que den soporte al proceso de toma de decisiones en materia de política educativa. Mediante el caso de la carrera de Licenciatura en Sistemas de la Universidad Nacional de Lanús, se evaluó la viabilidad y calidad de los resultados obtenidos, derivándose las siguientes conclusiones con respecto a las dos preguntas generales: I) la edad con la que los estudiantes ingresan a la carrera está asociada con el mantenimiento de regularidad por lo menos durante los primeros años de la cursada. Entre menor sea, menos posibilidades hay de que abandone; II) los últimos estudios que hayan obtenido los padres, particularmente la madre, presentan una asociación fuerte con respecto al rendimiento académico, disminuyendo la deserción y fomentando la culminación de la carrera. Aquellos estudiantes cuyos padres posean estudios secundarios completo o superior, presentan una tendencia positiva en su desempeño; III) entre más tiempo tarde el alumno en ingresar al sistema universitario luego de finalizar los estudios secundarios será más propenso a abandonar la carrera o demorarse respecto al cumplimiento del plan de estudios; IV) el hecho de haber asistido a un colegio secundario técnico parece influir positivamente en el rendimiento académico de los estudiantes (teniendo en consideración que la carrera analizada es del tipo técnico).

Como futuro trabajo, se prevé refinar el modelado de aquellas problemáticas cuyos resultados no fueron exitosos (II.b) y ampliar el conjunto de variables y procesos utilizados.

En adición a las líneas de investigación definidas en [Romero y Ventura, 2010], las problemáticas definidas en este trabajo, sientan las bases para definir un modelo estandarizado, que faciliten a educadores y/o usuarios no expertos en el área de IEI, la

implementación de las técnicas de extracción de conocimiento, las cuales permitan a los usuarios abstraerse de los aspectos específicos de este tipo de algoritmos.

6. Reconocimientos

En memoria del Prof. Dr. Ramón García-Martínez.

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por el proyecto 33A205 de la Universidad Nacional de Lanús.

7. Referencias

Aparicio, M. Garzuzi, V (2006): Dinámicas Identitarias, Procesos Vocacionales y su Relación con el Abandono de los Estudios. Un Análisis en Alumnos Ingresantes a la Universidad. Revista De Orientación Educacional V20. Pp 15-36

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM-Journal of Educational Data Mining, 1(1), 3-17.

Bean, J. P. (1980): Student Attrition, Intensions and Confidence. In: Research in Higher Education 17. Pp 291-320.

Bianco, S., Martins, S., Rodriguez, D., García-Martínez, R. (2017). Ingeniería de explotación de información aplicada a la gestión universitaria: Caso Sistemas Universidad Nacional de Lanús. http://sistemas.unla.edu.ar/sistemas/gisi/papers/Reporte de Tareas RT-UNLa-DDPyT-GISI-2017-04

Braxton, R. 2000. Reworking the student departure puzzle. Vanderbilt University Press.

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

Hackman, J. Dysinger, W. S. (1970): Commitment To College as a Factor in Student Attrition. Sociology of Education, 1970, 43 (3), 311-324.

IEDMS (2017). International Educational Data Mining Society. www.educationaldatamining .org Página vigente al 05/04/2017.

IESALC-UNESCO (2005). Datos para Colombia. SNIES, Ministerio de Educación Nacional. Colombia.

Kuna, H., García Martínez, R. Villatoro, F. (2010). Pattern Discovery in University Students Desertion Based on Data Mining. Advances and Applications in Statistical Sciences Journal, 2(2): 275-286.

Mansky, C. (1989): Anatomy of The Selection Problem. Journal of Human Resources 24, 343-360.

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.

Parrino, M. (2004): De la Reflexión a la Acción Política para Disminuir los Procesos de Deserción Universitaria. IV Coloquio Internacional sobre Gestión Universitaria en America Del Sud. Floreanopolis.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

SIU (2016). Sistema Inter Unversitario. http://www.siu.edu.ar/. Página vigente al 20/12/16.

Spady, W. (1970): Dropouts From Higher Education: An Interdisciplinary Review And Synthesis. Intechange 1. Pp.64-85.

SPU (2014). Anuario de Estadísticas Universitarias. Secretaria de Políticas Universitarias. Ministerio de Educación. Argentina. http://portales.educacion.gov.ar/spu/investigacion-y-estadisticas/anuarios/. Página vigente al 10/02/17.

Tinto (1975). Dropout From Higher Education: A Theoretical Synthesis of Recent Research. Review of Educational Research 45. Pp. 89-125.