Minería de Datos para la Caracterización de Recorridos Académicos

Hernán César Ahumada*, Oscar Eduardo Quinteros*, Claudia Alejandra Bazán*
*Universidad Nacional de Catamarca - Facultad de Tecnología y Ciencias Aplicadas
{hcahumada, oequinteros, cbazan}@tecno.unca.edu.ar

***Resumen*—A partir del registro histórico de actas de examen, se realizan a los datos las transformaciones necesarias para generar una vista minable que represente para cada alumno la secuencia temporal de aprobación de materias. A partir de la vista minable es posible reconstruir y caracterizar el recorrido académico del grupo de alumnos considerados en las materias de primer año de la carrera Ingeniería en Informática. Se llevan a cabo diversos análisis sobre la vista minable que revelan cuáles son las materias que plantean mayor dificultad en su aprobación. Además, se obtienen reglas de asociación que detectan correlaciones entre las materias aprobadas cuya interpretación permite encontrar interesantes interrelaciones entre las materias que figuran en las diferentes reglas. Se evalúan los niveles de cobertura y fortaleza de las reglas de asociación calculando e interpretando los valores de soporte, confianza y lift. Las conclusiones alcanzadas proveen conocimiento sobre el fenómeno analizado que resulta útil para determinar acciones que permitan elevar la eficiencia en el rendimiento académico.**

1. Introducción

La Minería de Datos Educativos (Educational Data Mining - EDM), [13], es una disciplina relacionada con el desarrollo de métodos para extraer información útil a partir de los datos que se generan en los entornos educativos, y utilizarla para mejorar dicho entorno. La información así obtenida se convierte en el insumo indispensable para la toma de decisiones. Jindal [9] menciona que uno de los principales objetivos de la EDM es el estudio, desde el punto de vista académico, del desempeño de los alumnos.

Las reglas de asociación ayudan a descubrir combinaciones de pares atributo-valor que ocurren con frecuencia en un conjunto de datos e inferir relaciones a partir de ellas que resultan bastante sencillas de interpretar [10].

La Minería de Secuencias [3] consiste en la búsqueda de patrones secuenciales frecuentes en una base de datos de eventos con fecha y hora [11] [7]. El proceso de descubrir patrones secuenciales involucra dos etapas: representar las secuencias y la aplicación del algoritmo que encontrará patrones frecuentes en las secuencias. El resultado de la primera etapa se denomina Vista Minable. Una Vista Minable es la consolidación en una única tabla de todas las observaciones y los atributos sobre los que se aplicarán los algoritmos de Minería de Datos.

Las actas de examen constituyen un registro histórico de la veces que cada alumno intenta aprobar las diferentes asignaturas del plan de estudio. Si se realiza el ordenamiento cronológico y agrupamiento de los datos de las actas de examen es posible reconstruir el recorrido académico llevado a cabo por los alumnos.

El presente trabajo pretende reconstruir y analizar, mediante técnicas de Reglas de Asociación y Minería de Secuencias, el recorrido académico de alumnos en asignaturas de primer año de la carrera Ingeniería en Informática que se dicta en la Facultad de Tecnología y Ciencias Aplicadas (FTyCA) de la Universidad Nacional de Catamarca (U.N.CA.). Se busca conocer y caraterizar el comportamiento y desempeño en los exámenes finales correspondientes al primer año de estudios. Para ello, primero se representan los datos bajo la forma de secuencias temporales, luego se analizan las características de dichas secuencias y finalmente se obtienen reglas de asociación.

En la Sección 2 se exponen los principales conceptos referidos a Reglas de Asociación. En la Sección 3 se explican los conceptos fundamentales relacionados a la Minería de Secuencias Temporales. La Sección 4 contiene el detalle de los datos y herramientas utilizadas. En la Sección 5, inicialmente se muestran resultados del análisis preliminar realizado al conjunto de datos original. Continúa con el proceso de transformación de los datos para generar la vista minable sobre la cual se realiza análisis de las características de las secuencias obtenidas. También se muestran e interpretan las reglas de asociación encontradas. Finalmente, la Sección 6 cierra el trabajo con las conclusiones y propuesta de trabajos futuros.

2. Reglas de Asociación

Reglas de Asociación es una técnica de minería de datos muy utilizada desde que Agrawal [1] presentara un algoritmo eficiente para generarlas automáticamente a partir del procesamiento de bases de datos. Una regla de asociación tiene dos partes, antecedente y consecuente. El antecedente está formado por un atributo o conjunción de atributos encontrados en los datos. El consecuente representa a otro atributo encontrado en combinación con el antecedente. Formalmente, una regla de asociación es una implicación

lógica de la forma $X \Rightarrow Y$, donde X es la premisa o antecedente que representa la condición o conjunción de condiciones que deben ser ciertas para que la conclusión o consecuente Y se cumpla. Por lo tanto, una regla de la forma $X \Rightarrow Y$, puede interpretarse de la siguiente forma: si sucede X, entonces sucede Y.

Entre los algoritmos para encontrar reglas de asociación se destacan los denominados Apriori [2] y Eclat [15]. Ambos algoritmos se usan para identificar eventos frecuentes, con la ventaja de que los resultados obtenidos con Apriori pueden usarse para generar reglas de asociación.

Para cuantificar el grado de validez de una regla de asociación, se utilizan principalmente dos medidas denominadas *Soporte* y *Confianza* [8]. Soporte se puede calcular tanto para un evento en particular como para una regla.

El soporte de un evento es la frecuencia relativa de ocurrencia de tal evento en la muestra de datos. Dicha frecuencia puede ser vista como la probabilidad marginal (o no condicional) de que ocurra el evento. En símbolos:

$$sop(X) = P(X) \tag{1}$$

Mientras que el soporte de una regla de asociación se entiende como la proporción de veces que el antecedente X y el consecuente Y se presentan de manera simultánea en los datos. Es, por lo tanto, la probabilidad conjunta de ambos eventos. En símbolos:

$$sop(X \Rightarrow Y) = P(X \cap Y) \tag{2}$$

El soporte de una regla (Ecuación 2) se calcula como el cociente del número de casos a los cuales la regla se aplica y predice correctamente entre el número total de casos.

En tanto que, Confianza se define como el cociente del soporte de la regla (Ec. 2) sobre el soporte del antecedente (Ec.1). En símbolos:

$$conf(X \Rightarrow Y) = \frac{sop(X \Rightarrow Y)}{sop(X)} \tag{3}$$

$$conf(X \Rightarrow Y) = \frac{P(X \cap Y)}{P(X)} \tag{4}$$

$$conf(X \Rightarrow Y) = P(Y \setminus X) \tag{5}$$

La confianza se calcula como el cociente entre el número de secuencias en las cuales la regla se cumple sobre el número de secuencias en las que aparece el antededente X de la regla (Ec. 3). Es decir, representa la proporción de secuencias que contienen a X en las cuales también de observa la ocurrencia de Y (Ec. 4). Por lo tanto, puede interpretarse como la probabilidad del consecuente Y dado que ocurre X (Ec. 5).

Debido a que un algoritmo de reglas de asociación tiende a encontrar una gran cantidad reglas, es usual establecer valores de umbral mínimos tanto para el soporte como para la confianza. Como el soporte de una regla mide la frecuencia de ella y la confianza representa la fortaleza de la regla, es de interés encontrar reglas frecuentes (soporte alto) y fuertes (confianza alta). En cambio, una regla de asociación con bajo soporte significa que la misma se cumple en pocos casos. En tanto que, si la confianza de la regla es baja implica que la ocurrencia del antecedente influye débilmente sobre el consecuente.

Para una regla de asociación, el indicador *lift* permite medir la ganancia de información, entendida como la reducción en la incertidumbre por el aporte que supone la ocurrencia del antecedente X de la regla [6]. Se define como el cociente entre la probabilidad condicional de Y dado que ocurre X (confianza de la regla) y la probabilidad no condicional del consecuente Y (soporte del consecuente). Se calcula dividiendo la confianza de la regla en el soporte del consecuente (Ecuación 6).

$$lift(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{sop(Y)} \tag{6}$$

Valores de lift próximos a 1 pueden interpretarse como que la regla establecida prácticamente no reduce la incertidumbre ya que numerador y denominador son valores cercanos. Cuando el lift sea mayor a 1, se deberá al hecho de que la probabilidad condicional del consecuente Y (confianza de la regla) es mayor que la probabilidad no condicional (soporte del consecuente Y). Cuanto mayor sea el lift, diremos que más fuertemente relacionados están el antecedente y el consecuente. Dando a entender que el hecho de que ocurra el antecedente eleva la probabilidad de que también se presente el consecuente de la regla.

3. Minería de Secuencias Temporales

La Minería de Secuencias Temporales consiste en encontrar patrones de secuencias, generalmente bajo la forma de asociaciones del tipo: *cuando ocurre A, entonces ocurre B dentro de algún lapso de tiempo* [4].

La formulación del problema de minería de secuencias frecuentes involucra los siguientes elementos básicos [14]:

- Alfabeto: conjunto de ítems. $I = \{i_1, i_2; ...; i_m\}$
- Evento: n-upla no ordenada de ítems. $\alpha_i = (i_1; i_2; ...; i_n)$
- Secuencia: lista ordenada de eventos $\alpha = (\alpha_1 \rightarrow \alpha_2 \rightarrow ... \rightarrow \alpha_k)$

Una secuencia α está compuesta de uno o más eventos α_i que a su vez incluyen n ítems del alfabeto I.

El tamaño de un evento está dado por la cantidad n de ítems que lo integran.

El ancho de la secuencia lo determina el máximo tamaño de los eventos α_i que componen dicha secuencia. El largo de una secuencia está determinado por la cantidad k de eventos que componen la misma.

El primer paso para aplicar un método de minería de secuencias es representar los datos a procesar bajo el formato de *transacciones*. Una transacción o secuencia se identifica mediante un número (Id de transacción) y está compuesta por una serie de eventos.

Las transacciones o secuencias temporales incorporan además información sobre el momento de ocurrencia de

cada evento (variable tiempo). Por ello en una secuencia temporal, la serie de eventos se ordena según la variable tiempo y se consigna el tiempo de ocurrencia del evento.

El total de transacciones constituye la base de datos D para realizar la minería de secuencias. Por lo tanto D es una colección de secuencias de entrada. Cada secuencia de entrada tiene un identificador único (S_{id}) y, a su vez, cada evento de una secuencia tiene también un identificador único (E_{id}). Además, cada evento de una secuencia tiene asociado un valor de la variable tiempo ($t(E_{id})$), donde no puede haber 2 eventos con el mismo valor de tiempo. Por lo tanto, $t(E_{id})$ se puede usar como identificador del evento. Dentro de una transacción, los eventos se ordenan de manera ascendente según la variable tiempo. Es decir, si el evento α_i ocurre antes que el evento α_j, entonces $t(\alpha_i) < t(\alpha_j)$.

4. Materiales y Métodos

Se realiza un trabajo de investigación de tipo cuantitativo, sobre los exámenes finales desde Abril-2004 hasta Marzo-2015 de una población de 370 alumnos de las cohortes 2004-2013 en la carrera de Ingeniería en Informática que se dicta en la Facultad de Tecnología y Ciencias Aplicadas de la Universidad Nacional de Catamarca.

El plan de estudios de Ingeniería en Informática, contempla en primer año siete materias que fueron numeradas de la siguiente manera:

1. Fundamentos de Informática.
2. Química.
3. Física I.
4. Sistemas de Representación.
5. Álgebra.
6. Geometría Analítica.
7. Análisis Matemático I.

La información extraída del sistema SIU-GUARNI fue almacenada en un archivo en formato Microsoft Excel con la totalidad de exámenes rendidos por los alumnos de Ingeniería en Informática. Se obtienen 3310 registros (exámenes), con los siguientes campos:

- IdAlumno: número de legajo que identifica unívocamente a cada alumno.
- Cohorte: año de ingreso a la Carrera del alumno.
- Materia: código alfanumérico que identifica a cada materia del CCA.
- Fecha: fecha del examen en la cual se inscribió el alumno.
- Resultado: resultado obtenido por el alumno en un examen (Aprobado, Reprobado o Ausente).

En el presente trabajo, se parte del conjunto inicial de datos que es transformado en un conjunto D de secuencias de exámenes aprobados que posea las variables (S_{id}), (E_{id}) y $t(E_{id})$, tal como lo requiere una vista minable para minería de secuencias temporales. Luego, como primer estudio de las secuencias, se realizan diversos análisis sobre el conjunto de secuencias D. Finalmente se buscan las reglas de asociación más relevantes del fenómeno estudiado.

Para realizar el procesamiento y análisis de las secuencias temporales, se utilizó el software *R* [12], ejecutando funciones de las librerías *arules* [8] y *arulesSequences* [5] sobre la vista minable construida previamente.

5. Resultados

5.1. Exploración inicial de los datos.

Los 3310 registros de examen corresponden a 370 alumnos. Del total de inscripciones a exámenes, se tiene que en 1207 de ellos el resultado es *Aprobado*, 1068 con resultado *Ausente* y 1035 con resultado *Reprobado*. El detalle, según el resultado del examen, para cada una de las materias de primer año se muestra en la Tabla 1.

Tabla 1. CANTIDAD Y PORCENTAJE DE EXÁMENES POR MATERIA

Materia	Aprobado	Ausente	Reprobado	Total
1	234 (45 %)	132 (25 %)	157 (30 %)	523 (16 %)
2	114 (23 %)	206 (41 %)	184 (36 %)	504 (15 %)
3	121 (43 %)	122 (43 %)	41 (14 %)	284 (9 %)
4	136 (66 %)	42 (20 %)	29 (14 %)	207 (6 %)
5	246 (37 %)	173 (25 %)	260 (38 %)	679 (21 %)
6	256 (45 %)	119 (21 %)	195 (34 %)	570 (17 %)
7	100 (19 %)	274 (50 %)	169 (31 %)	543 (16 %)

En la Tabla 1, los porcentajes de la última columna indican para cada materia la proporción de exámenes registrados. La materia 5 (Álgebra) es la que registra mayor cantidad de inscripciones a examen. Representa el $21\,\%$ de los 3310 registros analizados. Las asignaturas 3 (Física I) y 4 (Sistemas de Representación) tienen los porcentajes más bajos de inscripción a examen final, $9\,\%$ y $6\,\%$ respectivamente. Las restantes materias poseen porcentajes de participación de entre el $15\,\%$ y $17\,\%$.

También en la Tabla 1, los porcentajes de las columnas **Aprobado**, **Ausente** y **Reprobado** indican, por fila, la proporción de exámenes según el resultado de los mismos para cada materia. En este aspecto se destaca la asignatura 4 (Sistemas de Representación) que si bien tiene un bajo porcentaje de inscripción a examen, posee una alta tasa de aprobación (66 %). Análisis Matemático I (materia 7) tiene el mayor nivel de ausentismo en examen y la menor tasa de aprobación. Niveles similares de ausentismo lo tienen Química y Física I. En cuanto a los porcentajes de Reprobados, las materias 3 (Física I) y 4 (Sistemas de Representación) muestran los menores valores ($14\,\%$). En el resto de materias, los reprobados en examen es de entre $30\,\%$ y $38\,\%$.

Para conocer la proporción de alumnos que aprobaron cada materia, se calculó la frecuencia relativa dividiendo la cantidad de aprobados (columna 2 de la Tabla 1) sobre el total de 370 alumnos. Las frecuencias relativas obtenidas se grafican en la Figura 1. En dicha figura se observa que las materias 1 (Fundamentos de Informática), 5 (Álgebra) y 6 (Geometría Analítica) tienen proporciones de aprobación semejantes puesto que alrededor del $66\,\%$ de los alumnos tienen aprobadas esas materias, es decir que 2 de cada 3

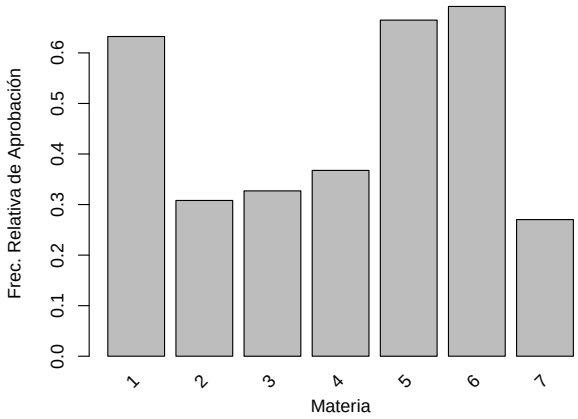

Figura 1. Frecuencia relativa de aprobación por materia

alumnos las aprobaron. Mientras que las materias 2 (Química), 3 (Física I) y 4 (Sistemas de Representación) exhiben porcentajes de aprobación que rondan el 33 %, dando a entender que fueron aprobadas por sólo 1 de cada 3 alumnos. En tanto que la materia 7 (Análisis Matemático I) muestra la menor proporción de aprobación ya que cada 4 alumnos sólo 1 la aprobó.

5.2. Modelado de la Vista Minable

En minería de secuencias temporales se requiere que los datos a analizar sean presentados con un estructura específica que constituye la vista minable del problema. Una vista minable organiza los datos en formato de transacciones. Una transacción se identifica mediante un número y está compuesta por una serie de eventos con uno o más ítems. Además, se necesita que la vista minable tenga un atributo que exprese el tiempo del evento.

En la presente sección, se realizan las operaciones y transformaciones de los datos originales para dejar conformada la vista minable. Se construye una secuencia por cada alumno, donde los ítems que las componen son las materias que haya aprobado, indicado como tiempo del evento el semestre académico en el que fueron aprobadas.

Para determinar el tiempo $t(E_{id})$ de cada evento, se considera como un semestre académico al período Abril-Septiembre y siguiente semestre comprende los meses Octubre-Marzo. Para cada alumno, el valor del semestre académico se determina calculando la cantidad de semestres que transcurrieron desde su ingreso a la carrera. De este modo, el tiempo de cada evento expresa en qué semestre de su recorrido académico aprobó las diferentes materias de primer año. En cada evento de una secuencia, el orden en el que aparecen los ítems refleja la cronología de aprobación de materias en el semestre académico dado por el tiempo del evento. Un extracto de la vista minable con las secuencias (transacciones) obtenidas se muestra en la Tabla 2

Tabla 2. EJEMPLO DE SECUENCIA TEMPORAL DE MATERIAS APROBADAS POR CADA ALUMNO.

IdSecuencia (S_{id})	**Tiempo del evento** $t(E_{id})$	**Ítems del evento**
6	1	$\{6\}$
6	2	$\{1,4,5\}$
12	2	$\{6,4,5,1\}$
12	3	$\{3,2\}$
12	4	$\{7\}$
21	2	$\{6\}$
25	5	$\{1,6\}$
25	10	$\{5\}$
31	2	$\{5,4,6,1\}$
31	3	$\{3\}$
70	2	$\{6,1\}$
90	2	$\{6,1,3,5\}$
90	3	$\{7\}$
90	5	$\{2\}$

Si en las secuencias de la Tabla 2 se suprime la variable *Tiempo del Evento*, se obtiene una vista minable apropiada para la búsqueda de reglas de asociación. En este caso las secuencias o transacciones adoptan una estructura como se muestra en la Tabla 3. En este formato cada secuencia contiene las materias aprobadas por un alumno, pero no informa el orden cronológico en el que fueron aprobadas ni el tiempo en el que se aprobó cada materia. A partir de este formato de vista minable el algoritmo APriori construye las reglas de asociación.

Tabla 3. TRANSACCIONES CON MATERIAS APROBADAS POR CADA ALUMNO.

IdSecuencia (S_{id})	**Ítems**
6	$\{1,4,5,6\}$
12	$\{1,2,3,4,5,6,7\}$
21	$\{6\}$
25	$\{1,5,6\}$
31	$\{1,3,4,5,6\}$
70	$\{1,6\}$
88	$\{\}$
90	$\{1,2,3,5,6,7\}$

5.3. Análisis de la Vista Minable

La vista minable de la Tabla 2 refleja para cada valor de IdSecuencia (S_{id}) el recorrido académico de un alumno en lo referido a la aprobación de materias de primer año. De este modo se puede conocer en qué semestres académicos el alumno rinde con éxito las materias de primer año. A partir de dichas secuencias temporales de aprobación de materias, se puede determinar en qué año de su historia académica los alumnos aprueban las diferentes asignaturas de primer año. Se considera como año 1 el año de ingreso a la carrera y comprende los semestres 1 y 2, el año 2 abarca los semestres 3 y 4, y así sucesivamente. En la Tabla 4 se muestran, por fila, los porcentajes de aprobación de cada materia en los respectivos años académicos. Se puede observar que las materias 1 (Fundamentos de Informática), 5 (Álgebra) y 6

(Geometría Analítica) se caracterizan por el alto porcentaje (entre 73 % y 79 %) de alumnos que aprueban dichas materias lo hacen durante el año académico 1. En tanto que las materias 2 (Química), 3 (Física I) y 4 (Sistemas de Representación) alrededor del 35 % de los alumnos que las aprueban antes que finalice el primer año académico. En el caso de la materia 7 (Análisis Matemático I) los porcentajes evidencian que el mayor nivel de aprobación ocurre en el transcurso del segundo año académico.

Tabla 4. PORCENTAJE DE APROBACIÓN DE MATERIAS SEGÚN AÑO ACADÉMICO.

	1	**2**	**3**	**4**	**5**	> 5
1	73,1 %	17,9 %	7,3 %	1,3 %	0,4 %	0,0 %
2	35,1 %	21,9 %	20,2 %	12,3 %	4,4 %	6,1 %
3	34,7 %	29,8 %	10,7 %	15,7 %	3,3 %	5,8 %
4	35,3 %	34,6 %	15,4 %	8,1 %	3,7 %	2,9 %
5	78,5 %	16,3 %	1,2 %	3,3 %	0,8 %	0,0 %
6	79,7 %	15,6 %	4,3 %	0,4 %	0,0 %	0,0 %
7	26,0 %	32,0 %	15,0 %	15,0 %	6,0 %	6,0 %

La longitud de una transacción está determinada por la cantidad de elementos (ítems) que la componen. En base al detalle de las transacciones de la Tabla 3, se determina la longitud de cada una de ellas realizando el conteo de la cantidad de transacciones con una determinada longitud. Los valores obtenidos se muestran en la Tabla 5 y representan la cantidad de alumnos con una determinada cantidad de materias de primer año aprobadas. La Tabla 5 informa que de los 370 alumnos considerados, el 17,3 % de ellos no aprobó ninguna materia del primer año de estudios. Mientras que alrededor del 20 % de la población bajo estudio tiene aprobadas las 7 materias de primer año de Ingenieía en Informática.

Tabla 5. CANTIDAD Y PORCENTAJE DE TRANSACCIONES SEGÚN CANTIDAD DE ÍTEMS.

Cant. Ítems	**Cant. Transacciones**	**% Transacciones**
0	64	17,3 %
1	55	14,9 %
2	46	12,4 %
3	53	14,3 %
4	31	8,4 %
5	24	6,5 %
6	22	5,9 %
7	75	20,3 %

5.4. Reglas de Asociación

A partir de la vista minable consignada en la Tabla 3 y mediante la aplicación del algoritmo Apriori se buscaron reglas de asociación cuyo soporte sea mayor o igual a $0,20$ y confianza no menor a $0,30$. Para ello se utilizó la librería *arules* [8] del software *R* [12]. En la Tabla 6, se listan las reglas de asociación cuyos antecedentes constan de una sola materia. Esto es, las asociaciones encontradas entre pares de materias. Se indican para cada una de ellas los valores de Soporte, Confianza y Lift que son indicadores para cuantificar la relevancia y la significancia de una regla de asociación. Las reglas de asociación están ordenadas según el valor decreciente del lift de cada una de ellas.

Tabla 6. SOPORTE, CONFIANZA Y LIFT DE REGLAS CON 1 ÍTEM EN EL ANTECEDENTE.

Nro.	Regla	Soporte	Confianza	Lift
1	$\{2\} \Rightarrow \{7\}$	0,22	0,73	2,69
2	$\{3\} \Rightarrow \{7\}$	0,24	0,73	2,69
3	$\{2\} \Rightarrow \{3\}$	0,25	0,80	2,44
4	$\{7\} \Rightarrow \{4\}$	0,23	0,86	2,34
5	$\{3\} \Rightarrow \{4\}$	0,28	0,85	2,32
6	$\{2\} \Rightarrow \{4\}$	0,26	0,83	2,27
7	$\{3\} \Rightarrow \{1\}$	0,32	0,99	1,57
8	$\{4\} \Rightarrow \{1\}$	0,35	0,96	1,52
9	$\{1\} \Rightarrow \{2\}$	0,29	0,46	1,50
10	$\{7\} \Rightarrow \{5\}$	0,27	0,99	1,49
11	$\{2\} \Rightarrow \{5\}$	0,30	0,98	1,48
12	$\{3\} \Rightarrow \{5\}$	0,32	0,98	1,47
13	$\{7\} \Rightarrow \{1\}$	0,25	0,92	1,45
14	$\{4\} \Rightarrow \{5\}$	0,35	0,96	1,45
15	$\{3\} \Rightarrow \{6\}$	0,32	0,99	1,43
16	$\{4\} \Rightarrow \{6\}$	0,35	0,96	1,39
17	$\{6\} \Rightarrow \{2\}$	0,29	0,42	1,37
18	$\{6\} \Rightarrow \{7\}$	0,25	0,37	1,36
19	$\{1\} \Rightarrow \{6\}$	0,58	0,91	1,32
20	$\{1\} \Rightarrow \{5\}$	0,55	0,87	1,30
21	$\{5\} \Rightarrow \{6\}$	0,57	0,85	1,23

En la Tabla 6, por ejemplo, para calcular el soporte de la regla nº 1, $\{2\} \Rightarrow \{7\}$, se contabiliza la cantidad de veces que se dan simultáneamente los dos eventos, es decir alumnos que aprobaron ambas materias. De la base de las transacciones de la vista minable surge que 83 de los 370 alumnos tienen aprobadas las materias 2 y 7 (Química y Análisis Matemático I). El cociente representa la probabilidad conjunta. Por lo tanto:

$$sop(\{2\} \Rightarrow \{7\}) = \frac{83}{370} = 0,22. \tag{7}$$

El valor de soporte obtenido en la Ec.7 significa que la regla en cuestión se verifica en el 22 % de los casos. Se interpreta que ese porcentaje de alumnos tienen aprobadas las materias que figuran tanto en el antecedente como en el consecuente de la regla de asociación.

En tanto que, la Confianza de una regla de asociación se determina según lo expresado en la Ecuación 3. Es decir, dividiendo el soporte de regla en el soporte del antecendente. En el caso de la regla 1 de la Tabla 6, el soporte de la regla está calculado en la Ecuación 7. Para determinar el soporte del antecedente se contabilizan las transacciones donde aparezca el ítem 2. Según la Tabla 1, la materia 2 fue aprobada por 114 alumnos. Por lo tanto:

$$sop(\{2\}) = \frac{114}{370} \tag{8}$$

En consecuencia la Confianza de la regla $\{2\} \Rightarrow \{7\}$ es:

$$conf(\{2\} \Rightarrow \{7\}) = \frac{sop(\{2\} \Rightarrow \{7\})}{sop(\{2\})} \tag{9}$$

$$conf(\{2\} \Rightarrow \{7\}) = \frac{\frac{83}{370}}{\frac{114}{370}} = \frac{83}{114} = 0,73 \tag{10}$$

La fracción en la Ecuación 10 plantea que de los 114 alumnos que aprobaron la materia 2 (Química), 83 de ellos también aprobaron la materia 7 (Análisis Matemático I). Dado que la confianza de una regla se interpreta como la probabilidad condicional del consecuente dado que ocurre el antecedente, entonces, el valor de confianza obtenido representa que existe una probabilidad de $0,73$ de que un alumno haya aprobado la asignatura 7 si es que tiene aprobada la materia 2.

Como está definido en la Ecuación 6 el lift es el cociente entre la confianza de la regla y el soporte del consecuente de la misma. En la Ecuación 11 se calcula el soporte del consecuente como la cantidad de alumnos que aprobaron la materia 7 (100, según la Tabla 1) sobre el total de alumnos (370).

$$sop(\{7\}) = \frac{100}{370} \tag{11}$$

Entonces el lift de la regla de asociación $\{2\} \Rightarrow \{7\}$, se calcula dividiendo el resultado de la Ecuación 10 en el valor de la Ecuación 11. Es decir:

$$lift(\{2\} \Rightarrow \{7\}) = \frac{conf(\{2\} \Rightarrow \{7\})}{sop\{7\}} \tag{12}$$

$$lift(\{2\} \Rightarrow \{7\}) = \frac{\frac{83}{114}}{\frac{100}{370}} = 2,69 \tag{13}$$

El valor del lift de una regla de asociación es un indicador de la significancia estadística de la misma. Como en el ejemplo el lift es mayor a uno $(2,69)$ significa que existe dependencia entre el antecedente y el consecuente, esto es que si el alumno aprueba la materia 2 se incrementa la probabilidad de que también apruebe la materia 7.

Por lo tanto, los valores de soporte, confianza y lift de la regla $\{2\} \Rightarrow \{7\}$, representan que la misma se verifica en el 22 % de los casos, y que si un alumno tiene aprobada en la materia 2 existe una probabilidad de $0,73$ de que también apruebe la materia 7. Como la regla posee un lift de $(2,69)$ significa que es alta la significancia estadística de ella, es decir que el hecho de aprobar la materia 2 (Química) eleva fuertemente la probabilidad de aprobar la materia 7 (Análisis Matemático I).

En la Tabla 7 se detallan las reglas de asociación cuyos antecedentes constan de dos materias (ítems). Las reglas están ordenadas en forma decreciente con respecto al valor de lift. También se indican los valores de soporte y confianza de cada regla de asociación.

Las primeras tres reglas en la Tabla 7, se caracterizan por los altos valores de confianza y lift que poseen. El soporte de cada una de ellas es superior al 20 %. Además, tienen la misma materia en el consecuente (Análisis Matemático I). En el antecedente de las mismas figuran de a pares las materias 2 (Química) , 3 (Física I) y 4 (Sistemas de Representación). Por lo tanto, estas reglas indican que aprobando el par de materias del antecedente se eleva fuertemente la probabilidad de aprobar Análisis Matemático I (consecuente).

En la Tabla 8, se muestran las reglas de asociación cuyos antecedentes lo componen tres materias.

Tabla 7. SOPORTE, CONFIANZA Y LIFT DE REGLAS CON 2 ÍTEMS EN EL ANTECEDENTE.

Nro.	Regla	Soporte	Confianza	Lift
1	{2,3} ⇒ {7}	0,21	0,85	3,13
2	{3,4} ⇒ {7}	0,23	0,82	3,02
3	{2,4} ⇒ {7}	0,21	0,80	2,96
4	{1,7} ⇒ {3}	0,24	0,96	2,92
5	{6,7} ⇒ {3}	0,24	0,94	2,86
6	{3,5} ⇒ {7}	0,24	0,75	2,76
7	{1,7} ⇒ {2}	0,21	0,85	2,75
8	{2,5} ⇒ {7}	0,22	0,74	2,74
9	{2,4} ⇒ {3}	0,23	0,89	2,74
10	{1,2} ⇒ {3}	0,25	0,84	2,58
11	{2,6} ⇒ {3}	0,25	0,84	2,58
12	{1,7} ⇒ {4}	0,23	0,93	2,54
13	{3,5} ⇒ {2}	0,25	0,77	2,50
14	{6,7} ⇒ {4}	0,23	0,91	2,49
15	{4,5} ⇒ {7}	0,23	0,66	2,43
16	{1,4} ⇒ {3}	0,28	0,79	2,40
17	{4,5} ⇒ {3}	0,28	0,79	2,40
18	{4,6} ⇒ {3}	0,28	0,79	2,40
19	{1,2} ⇒ {4}	0,26	0,88	2,39
20	{2,6} ⇒ {4}	0,26	0,88	2,39
21	{4,5} ⇒ {2}	0,25	0,72	2,33

Tabla 8. SOPORTE, CONFIANZA Y LIFT DE REGLAS CON 3 ÍTEMS EN EL ANTECEDENTE.

Nro.	Regla	Soporte	Confianza	Lift
1	{2,3,4} ⇒ {7}	0,20	0,88	3,26
2	{2,4,5} ⇒ {7}	0,21	0,81	2,99
3	{5,6,7} ⇒ {3}	0,24	0,95	2,89
4	{2,4,5} ⇒ {3}	0,23	0,90	2,77
5	{1,2,5} ⇒ {3}	0,25	0,85	2,60
6	{2,5,6} ⇒ {3}	0,25	0,85	2,60
7	{5,6,7} ⇒ {4}	0,23	0,92	2,52
8	{1,4,6} ⇒ {3}	0,28	0,80	2,46
9	{4,5,6} ⇒ {3}	0,28	0,80	2,46
10	{1,4,5} ⇒ {3}	0,28	0,80	2,44
11	{1,5,6} ⇒ {3}	0,32	0,61	1,86
12	{1,5,6} ⇒ {2}	0,29	0,55	1,79
13	{1,5,6} ⇒ {4}	0,34	0,65	1,78

La regla Nº3 de la Tabla 8, se destaca por los valores de confianza y lift que posee. El valor de confianza (0,95), indica que si el alumno aprueba las tres materias del antecedente, es altamente probable que apruebe también la materia del consecuente. En tanto que el valor de lift igual a 2,89 confirma que si se cumple el antecedente se incrementa la probabilidad de que ocurra el consecuente y que tal evento no es fruto del azar.

6. Conclusiones y Trabajos Futuros

Los resultados obtenidos en este trabajo, mediante técnicas de minería de datos, permiten conocer patrones de comportamiento académico de los alumnos en base a las actas de exámenes finales de materias de primer año de las carrera de Ingeniería en Informática.

Se pudo determinar los porcentajes de aprobación, ausentismo y reprobación de cada una de las materias de primer año (Tabla 1). Dicha información es útil para distinguir diferentes comportamientos al momento de afrontar

exámenes finales, y amerita que se le preste atención por parte de las diferentes cátedras. Principalmente aquellas con altos niveles de ausentismo y de reprobación.

La frecuencia relativa de aprobación de cada materia (Figura 1), pone en evidencia cuáles son las materias que los alumnos tienen mayor dificultad de aprobar. Ese grupo de materias lo encabeza *Análisis Matemático I* y lo completan *Química*, *Fisica I* y *Sistemas de Representación*.

La representación de los datos de exámenes bajo el formato de secuencias con eventos en diferentes tiempos, permite reconstruir el recorrido académico de aprobación de materias (Tabla 2). El conjunto de secuencias constituye lo que se denomina vista minable a partir de la cual se realizaron diversos análisis. Contando la cantidad de ítems que componen cada secuencia se determinó cuántas materias aprobó cada alumno y se contabilizaron los alumnos según la cantidad de materias aprobadas (Tabla 5). Se llega así a conocer que el 80 % de lo alumnos no completó el primer año de la carrera.

La determinación del año académico en el cual se aprueba cada una de las materias de primer año (Tabla 4) hace posible distinguir las materias que son aprobadas en primer lugar y aquellas cuya aprobación es postergada. Permite conocer que la gran mayoria de los alumnos que aprueban Fundamentos de Informática, Álgebra y Geometría Analítica lo hacen dentro de primer año académico. Mientras que la aprobación de las restantes materias les insume hasta más de 5 años académicos. Nuevamente *Química*, *Fisica I* y *Análisis Matemático I* aparecen como las asignaturas que más tiempo les demanda a los alumnos aprobar.

Estos resultados permiten identificar cuáles son las asignaturas o grupos de asignaturas que presentan a los estudiantes mayores dificultades en su aprobación. En base a ello se podrán diseñar y tomar medidas didácticas destinadas a subsanar dichas dificultades para mejorar los indicadores en la aprobación del primer año de la carrera de Ingeniería en Informática.

También se obtuvieron reglas de asociación cuya relevancia y significancia fue cuantificada por indicadores de soporte, confianza y lift (Tablas 6, 7 y 8). Mediante esta técnica de minería de datos se realizó un análisis novedoso del recorrido académico en las asignaturas del primer año de la carrera. Se encontraron así grupos de materias cuya aprobación es mas frecuente, dando a entender que el hecho de aprobar la/s materia/s del antecedente hace más probable la aprobación de la materia que figura en el consecuente de cada regla de asociación.

La vista minable generada en el presente trabajo (Tabla 2) habilita su procesamiento, en futuros trabajos, mediante algoritmos de Minería de Secuencias con el objetivo de conocer patrones frecuentes en los recorridos académicos bajo la forma de reglas que involucren a la variable tiempo.

Agradecimientos

Los autores expresan su agradecimiento a las autoridades de la Facultad de Tecnología y Ciencias Aplicadas (FTyCA) de la Universidad Nacional de Catamarca (U.N.CA.) por autorizar el uso de los datos analizados. En especial al Lic. Marcelo Ríos, encargado del sistema SIU-Guaraní.

La presente publicación se realizó en el marco del proyecto "*Extracción de Conocimiento sobre Recorridos Académicos en asignaturas del CCA*" que se lleva a cabo en el Departamento de Informática de la FTyCA-U.N.CA.

Referencias

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE, 1995.

[4] Jussi Ahola. Mining sequential patterns. 2001.

[5] Christian Buchta, Michael Hahsler, and Daniel Diaz. arulessequences: Mining frequent sequences. *R package version 0.2-4*, 2013.

[6] Emiliano Carreno and Guillermo Leguizamón. Balance entre confianza, soporte y comprensibilidad en la evolucion de reglas de clasificacion. In *VIII Workshop de Investigadores en Ciencias de la Computación*, 2006.

[7] Dante Conti and Fco Javier Martínez de Pisón Ascacíbar. Reglas de asociación en series temporales: panorama referencial y tendencias.

[8] Michael Hahsler, Bettina Grün, and Kurt Hornik. Introduction to arules: Mining association rules and frequent item sets. *SIGKDD EXPLORATIONS*, 2:0–4, 2007.

[9] Rajni Jindal and Malaya Dutta Borah. A survey on educational data mining and research trends. *International Journal of Database Management Systems*, 5(3):53, 2013.

[10] Brett Lantz. *Machine learning with R*. Packt Publishing Ltd, 2013.

[11] Prof Pinkal Shah and AK Dua. Algorithm for sequence mining using gap constraints. *International Journal of Engineering Research and Development*, pages 37–49, 2014.

[12] R Core Team. *R: A Language and Environment for Statistical Compu ting*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[13] Cristóbal Romero and Sebastián Ventura. Educational data mining: A review of the state of the art. *Trans. Sys. Man Cyber Part C*, 40(6):601–618, November 2010.

[14] Mohammed J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Mach. Learn.*, 42(1-2):31–60, January 2001.

[15] Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, Wei Li, et al. New algorithms for fast discovery of association rules. In *KDD*, volume 97, pages 283–286, 1997.