“Identificación de perfiles de rendimiento para alumnos de pregrado de Escuelas de Economía y Negocios”

Resumen

Proponemos una metodología de caracterización de perfiles de rendimiento para alumnos de pregrado de Escuelas de Economía y Negocios que permita apoyar las decisiones académicas de los estudiantes. Los objetivos específicos son identificar y describir a los distintos grupos de estudiantes agrupados por el semestre de avance académico y a diferentes variables de caracterización. Rendimiento académico es medido como porcentaje de reprobación del semestre anterior y total de créditos aprobados. Sobre carga académica, es medida como el número de créditos inscritos por sobre los recomendados para un determinado semestre. Además, se consideran variables de entorno, variables de comportamiento, y características socio-demográficas para mejorar la capacidad descriptiva de los perfiles. El marco de análisis es el Knowledge Discovery in Databases (KDD) siguiendo las guías de minería de datos de CRISP-DM.

El trabajo se organizó en dos etapas: i) exploración de patrones de agrupación usando algoritmos de clustering; ii) análisis confirmatorio a través de regresiones longitudinales en datos de panel. En este informe se reportan los principales resultados de la etapa i). Utilizamos una base de datos con n=16.269 observaciones que contiene alumnos de siete semestres entre otoño 2012 y otoño 2015. Los alumnos pertenecen a tres carreras de dos Escuelas de Economía y Negocios de la Universidad de Chile. Encontramos clústeres estadísticamente significativos de acuerdo a la medida de cohesión y separación silueta para cada semestre de avance y para cada carrera. Los resultados son confirmados utilizando la prueba F en un one-way ANOVA.

Observamos que las diferencias académicas de los primeros semestres tienden a desaparecer en los últimos semestres, a pesar de las grandes diferencias socio-económicas existentes entre los alumnos. En el futuro, dichos resultados pueden ser usados por un Sistema Experto para optimizar las recomendaciones a los estudiantes respecto a su itinerario formativo al interior de la Universidad.

Palabras Clave: Minería de Datos, Educación, Sistema Experto.

Introducción

La toma de decisiones en cualquier aspecto de la vida del ser humano, está ligada a las preferencias y a la racionalidad de éste. En el caso de la toma de decisiones de los alumnos en la educación superior, estas podrían estar influenciadas por la poca experiencia que los estudiantes poseen. Por ejemplo, un alumno podría no ponderar correctamente el tiempo, esfuerzo y el nivel intelectual necesario para enfrentar de manera exitosa el nivel de la carga académica que inscribirá en un semestre determinado. Estas decisiones pueden desencadenar consecuencias no deseadas a lo largo de su etapa universitaria, tales como constantes reprobaciones a lo largo de la carrera, llegando incluso a la deserción o eliminación académica de esta. Asimismo, estas decisiones no tan sólo pueden repercutir en su desempeño académico, sino también, en su estado de ánimo y salud psicológica.

En este sentido, las Facultades de Economía y Negocios, y en particular, las áreas encargadas del desarrollo y avance curricular, pueden tomar un rol preponderante en guiar a los alumnos hacia una buena elección, tanto en la cantidad de asignaturas que deben inscribir, como también en la variedad y dificultad de éstas mismas. Igualmente, si bien las decisiones de carga académica pueden ser extremadamente importantes, también existen situaciones fuera de norma en que los alumnos solicitan de manera excepcional una *sobre carga académica.* Esta corresponde a la inscripción por sobre la cantidad de créditos recomendados para un semestre de acuerdo a lo establecido en las mallas de avance de las carreras.

Con el fin de informar y apoyar la toma de decisiones por parte de las áreas de desarrollo y avance curricular, generalmente responsabilidades asignadas a una Secretaría de Estudios, este trabajo busca identificar perfiles de rendimiento para alumnos de pregrado de Escuelas de Economía y Negocios de acuerdo al nivel de carga y sobre carga académica, e identificar y describir dichos grupos a lo largo de cada semestre al momento de inscribir cursos en un semestre determinado.

Se busca encontrar patrones comunes y relevantes que identifiquen a estos grupos y permitan sugerir medidas de acción respecto a la cantidad de asignaturas que deben inscribir, como también la variedad y dificultad de éstas mismas. Asimismo, proponer recomendaciones respecto de si corresponde aceptar o rechazar solicitudes de inscripción de sobrecarga académica de acuerdo al rendimiento esperado del grupo.

En este trabajo se utilizan las herramientas de minería de datos en el marco de análisis de Knowledge Discovery in Databases [KDD] (Fayyad, Piatetsky-Shapiro & Smyth, 1996) y la metodología Cross-Industry Process for Data Mining [CRISP-DM] (Chapman, Clinton, Keber et all, 2000). El trabajo se organizó en dos etapas: i) exploración de patrones de agrupación usando algoritmos de clustering (two-steps k-means); ii) análisis confirmatorio a través de regresiones longitudinales en datos de panel. En este informe se reportan los principales resultados de la etapa i).

Dos técnicas competitivas de algoritmos de clustering fueron utilizadas: k-means, donde el número de agrupaciones a encontrar en los datos es seleccionado por el investigador, y x-means donde el número de conglomerados es determinado automáticamente. La calidad de los clústers resultantes fue evaluada utilizando la medida de cohesión y separación Silueta (Kaufmann & Rousseeuw, 1990).

La contribución de este trabajo radica en la aplicación empírica de la minería de datos en el contexto de la Educación Superior en Latinoamérica, y en la propuesta de una metodología de análisis que permita la caracterización de perfiles de rendimiento para alumnos de pregrado de Facultades de Economía y Negocios. Se espera que este estudio entregue los cimientos para futuras investigaciones y que los resultados obtenidos sean utilizados para profundizar esta área de investigación. Por otro lado, se busca seguir creciendo en el uso de la herramienta de minería de datos en un campo muy importante como lo es la educación superior. Lo anterior con el

objetivo de encontrar resultados que generen impactos y mejoras tanto en los sistemas educacionales latinoamericanos, como en los resultados académicos de los alumnos.

Revisión Literaria

Existen numerosas definiciones de minería de datos. De acuerdo a Gartner Group, la minería de datos es el descubrimiento eficiente de patrones previamente desconocidos en grandes bases de datos (Gartner Group, 1994). Dunham el 2002 define minería de datos como: "La extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de los datos" (Dunham, 2002). Por otro lado, Fayyad et al 1996 describe la minería de datos como una etapa de descubrimiento en el proceso de KDD, la cual consiste en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre procesados. En particular, el marco de análisis KDD es entendido como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y entendibles en los datos (Fayyad, Piatetsky-Shapiro & Smyth, 1996). Profundizando en esta definición, "proceso" implica que el KDD comprende diversos pasos, los que incluyen la preparación de los datos, la búsqueda de patrones, la evaluación del conocimiento y el refinamiento, todos estos pueden repetirse en múltiples iteraciones. El concepto, "no trivial" hace referencia a que el proceso involucra investigación e inferencia, no es una simple y directa aplicación de cálculos sobre la data. Finalmente, por "válido" se entiende que los patrones descubiertos deben seguir siendo precisos, para integrar datos nuevos que representen un aporte sobre algo desconocido por el investigador.

CRISP-DM [Cross Industry Standard Process for Data Mining] fue creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, siendo esta una guía de referencia para el desarrollo sistemático de proyectos de minería de datos en el ámbito de negocios. Como metodología incluye las descripciones de fases normales de un proyecto, las tareas necesarias

de cada fase y una explicación de las relaciones entre las tareas. Por otro lado, como un modelo de proceso, éste ofrece un resumen del ciclo de vital de minería de datos.

Dentro de la minería de datos, además, se suelen organizar las tareas en dos grandes áreas o tipos de algoritmos de acuerdo a las técnicas de datos que se realizan. En este sentido, la minería de datos puede ser descriptiva, en la cual los algoritmos son utilizados para encontrar y describir patrones, tales como agrupaciones naturales, tendencias, trayectorias, correlaciones, co-ocurrencias de eventos, o secuencias de eventos interesantes; o puede ser predictivas o de pronóstico, cuando la tarea corresponde a describir la ocurrencia o valor de un fenómeno en el futuro, siendo éste de naturaleza categórica o continua (Dunham, 2002). Dentro de las tareas predictivas, además, se distingue la existencia de una variable "dependiente" que se pretende predecir y las "independientes" o "explicativas", que son los atributos usados para la predicción.

En este trabajo se ocupan principalmente algoritmos del tipo descriptivo, pertenecientes a una familia denominada *análisis de conglomerados*, también denominados *clustering*, por su nomenclatura en inglés (Jain, Murty, & Flynn, 1999). Éste corresponde al conjunto de técnicas que permiten encontrar similitudes entre filas de una base de datos, las cuales son organizadas en grupos según la homogeneidad de los mismos. A cada uno de éstos grupos se les denomina un clúster o conglomerado, donde cada uno posee elementos con características que los diferencias de los otros grupos. En contraste a las técnicas de predicción o pronóstico de una variable categórica, los grupos no han sido predefinidos por el usuario, y son encontrados de manera no supervisada por el algoritmo. Dado lo anterior, se requiere conocimiento del dominio o ámbito desde el cual los datos son extraídos, es decir, deben ser interpretados por un usuario experto en el contexto del problema.

De acuerdo a lo presentado por (Romero y Ventura, 2010) , existen 8 grandes grupos dentro de los cuales se clasifican las distintas referencias sobre minería de datos educacionales:

Educación Tradicional, Educación basada en la Web [E-learning], Sistemas de Gestión de Aprendizaje, Sistemas Tutoriales Inteligentes, Sistemas Educativos Adaptativos, Pruebas/Cuestionarios, Textos/Contenidos y otros. Asimismo, estos autores identifican once tipos de aplicaciones: a) Análisis y visualización de datos, b) Proporcionar información para apoyar a los instructores (Feedback), c) Recomendar cursos o actividades a los estudiantes, d) Predicción del rendimiento, e) Modelar el comportamiento del estudiante, f) Detección de comportamientos no deseados en estudiantes, g) Agrupación de estudiantes, h) Análisis de redes sociales, i) Generación automática de mapas conceptuales, j) Planificación y programación, y finalmente, k) Construcción de cursos. Los autores clasifican estas once tareas en cinco objetivos según los resultados generales a los que se llegan en cada una. Lo anterior se resume en la Tabla 1. Clasificación de Aplicaciones de Minería de Datos en Educación:

Objetivo	Tarea(s)
Proporcionar información para los instructores	A y B
Proporcionar información para los estudiantes	C
Revelan caracterícticas de los estudiantes	D, E, F y G
Estudian gráficos y relaciones entre los estudiantes y los conceptos	H y I
Ayudan a crear y/o planear los cursos	J y K

Tabla 1. Extraído del artículo "Educational Data Mining: A Review of the State of the Art" (2010), de los autores Romero & Ventura. Elaboración propia.

De acuerdo a esto, nuestro trabajo se posicionaría dentro de las referencias de Educación Tradicional, Sistemas de Gestión de Aprendizaje, y dentro de las tareas c), d) e) y g). Asimismo, la contribución de este trabajo radica en la aplicación empírica de la minería de datos en el contexto de la Educación Superior en Latinoamérica, y en la propuesta de una metodología de análisis que permita la caracterización de perfiles de rendimiento para alumnos de pregrado de la Facultad de Economía y Negocios.

Metodología y Datos

La metodología de análisis de la etapa 1 consistió en la aplicación de técnicas de conglomerado o clustering (Jain, Murty, & Flynn, 1999). En particular, dos técnicas competitivas de

algoritmos de clustering fueron utilizadas, k-means, donde el número de agrupaciones a encontrar en los datos es seleccionado por el investigador, y x-means, también conocido como two-steps k-means, donde el número de conglomerados es determinado automáticamente de acuerdo a un proceso de selección de características estadísticas deseadas de los clústers, i.e., homogeneidad intra-grupal, heterogeneidad inter-grupal, y ratio de tamaño de grupos. La calidad de los clústers, además, fue evaluada utilizando la medida de cohesión y separación de clústers Silueta (Kaufmann & Rousseeuw, 1990).

La base de datos utilizada correspondió a alumnos pertenecientes a dos Escuelas de Pregrado de la Facultad de Economía y Negocios de la Universidad de Chile. La base contiene n=16.269 observaciones correspondiente a alumnos de siete semestres entre otoño 2012 y otoño 2015, los que son seguidos a través del tiempo (datos de panel). La Tabla 2 presenta el número de observaciones para cada semestre:

Semestre	Número de alumnos
2015-1	2.696
2014-2	2.344
2014-1	2.520
2013-2	2.160
2013-1	2.312
2012-2	2.039
2012-1	2.198
Total	16.269

Tabla 2. Número de alumnos en el estudio por semestre. Elaboración propia.

Los datos provienen de dos fuentes de información: (1) el Sistema de Administración Docente Facultad de Economía y Negocios (FEN) y (2) la Base del Departamento de Evaluación, Medición y Registro Educacional (DEMRE) dependiente de la Universidad de Chile. La primera fuente de información contiene datos académicos de los estudiantes en su paso por la FEN. Mientras que la segunda fuente de información entrega los datos socio-demográficos que declaran los estudiantes al momento de rendir la Prueba de Selección Universitaria (PSU), previo al ingreso a la Facultad.

Respecto de las variables descriptivas en estudio, se consideraron los resultados encontrados en un trabajo anterior (Ortega, Lee, Silva & Vásquez, 2015). Este trabajo buscaba identificar las variables claves que influyen el nivel de la carga académica que inscriben los estudiantes en FEN encontrando diferentes variables claves para una carga académica alta, media y baja. La Tabla 3 presenta un resumen de los atributos seleccionados, los cuales son agrupados de acuerdo a su fuente de origen:

Variables	
Variables Académicas	**Variables Socioeconómicas**
Cod_Alumno	Grupo Dependencia
Periodo	Tiene_Trabajo_rem
% Reprob Semestre anterior	Horas_que_dedica_trabajo
Semestre_Avance	Grupo _familiar
Tramos	Ingreso_bruto_familiar
Factor_Rend Sem Anterior	¿Vive_sus_padres?
Tipo Carrera	Ed_Padre
	Ed_madre
	Sexo
	Régimen

Tabla 3. Atributos seleccionados. Elaboración propia.

Dentro del primer grupo, "Cod_Alumno" representa un atributo que hace único al estudiante dentro de la Facultad. El "Periodo" se refiere al semestre en el cual se inscriben los cursos. El atributo "% _Reprob Sem Anterior" es el porcentaje de cursos que reprobó un determinado alumno el semestre anterior del periodo de inscripción de cursos correspondiente. "Sem_Avance" entrega el semestre en el que se encuentra un alumno al momento de realizar la inscripción establecida (periodo), este semestre se obtiene del total acumulado de UDs/créditos cursados y aprobados hasta dicho periodo. "Tramos" es una variable que se definió por conveniencia, la cual mide el grado de avance del alumno en términos del número de UDs o créditos que ha aprobado del total éstos. Se definieron 7 tramos. Los tramos aumentan de manera ascendente en diez UDs o seis créditos según la malla en la que se encuentre determinado alumno. El "Factor_Rend Sem Anterior" es un porcentaje que representa la cantidad de cátedras cursadas y aprobadas sobre la cantidad total de cátedras cursadas,

aprobadas y reprobadas. Por último, el atributo "Tipo Carrera" indica la carrera que el alumno está estudiando.

Por otro lado, los atributos seleccionados provenientes de la base de datos proporcionada por el DEMRE son diez. Inicialmente los atributos relacionados a la educación del alumno encontramos "Grupo_Dependencia" que nos indica si el alumno proviene de un colegio Particular Pagado (1), Particular Subvencionado (2) o de uno Municipal (3). Por su parte, el atributo "Régimen" indica si el alumno proviene de un colegio Masculino (1), Femenino (2) o Coeducacional (3). Además, dentro de los atributos propios del alumno encontramos, "Sexo" que representa si el alumno es Hombre (1) o Mujer (2). "Tiene_trabajo_rem" indica si un alumno tiene trabajo o no (1), y si este es permanente (3) o temporal (2). De manera complementaria a este último atributo, "Horas_que_dedica_trabajo" entrega un número de dos dígitos indicando cuantas horas semanales el alumno le dedica al trabajo. En cuanto a información sobre la estructura familiar del alumno, se seleccionó "Grupo_familiar" que indica el número de personas que componen el grupo familiar. "Ingreso_bruto_familiar" por su parte muestra el tramo en el que se encuentra la familia, pudiendo clasificarse en doce tramos, con una diferencia de CLP$144.000 entre los tramos (aproximadamente USD200). El atributo "¿Vive_sus_padres?" indica si ambos padres viven (1), sólo lo hace la madre (2), sólo lo hace el padre (3) o ambos están fallecidos (4). Por último, se incluyó la educación que tenían los padres, teniendo trece categorías de acuerdo a la cantidad de años dedicados al estudio.

Como resultados se reportan distintos clústeres por semestre de avance y escuela. Se obtuvo además, la media, la desviación estándar, el error estándar de la media y la cantidad de estudiantes pertenecientes a cada clúster obtenido para cada una de las variables en estudio. Asimismo, para comprobar si efectivamente los centroides de grupos encontrados son estadísticamente distintos entre sí, se utilizó un análisis de one-way ANOVA para muestras independientes, evaluado de acuerdo al test F.

Resultados

A continuación se detallaran para dos semestres de avance, los distintos clústeres que se formaron para las dos escuelas que conforman FEN. Mayores detalles de los semestres restantes están disponibles a solicitud del lector. El tamaño del clúster en número de alumnos es entregado entre paréntesis.

Semestre de Avance 2 - Escuela de Economía y Administración. Tabla 4. Panel A.

Clúster 1 (202): Poseen el mayor factor de rendimiento alcanzando un 91,7%, además poseen el menor porcentaje de reprobación del semestre anterior con un 9,3%. Este grupo está conformado sólo por mujeres, las que provienen generalmente de colegios coeducacionales y particulares pagados. Poseen el segundo ingreso bruto familiar más alto, encontrándose en promedio entre el tramo 10 y 11. La educación de ambos padres es en promedio alta, superando los 15 años de educación.

Clúster 2 (90) y Clúster 3 (328): Ambos clúster son bastante similar al clúster 1, sin embargo poseen un factor de rendimiento un poco menor (89% y 87,2% respectivamente) y un porcentaje de reprobación del semestre anterior un poco más alto (13% y 14% respectivamente). Sin embargo, la principal diferencia es que el clúster 2 se compone en su totalidad de hombres los cuales provienen de colegios masculinos, mientras que el clúster 3 está compuesto 100% de hombres quienes provienen exclusivamente de colegios coeducacionales.

Clúster 4 (257): Este grupo de alumnos es el que posee menor factor de rendimiento (87,3%) similar al del clúster 2. A la vez poseen la mayor tasa de reprobación del semestre anterior (15% aproximadamente). Su principal diferencia con los otros clústeres es que generalmente estos alumnos provienen de colegios municipales. Por otro lado, son el grupo que en promedio tienen el menor ingreso bruto ubicándose en el tramo 3. Además sus padres poseen los menores

años de educación en promedio, alcanzando los 12 años. Este grupo está compuesto por hombres y mujeres, sobresaliendo la presencia masculina.

Existen variables que poseen promedios muy similares para los cuatro clúster que se formaron, las más homogéneas son: Tramo que se mueve entre 4,6 y 4,7, Grupo familiar que varía entre 4 y 5, la variable de Viven sus padres la cual se encuentra entre 4 y 5.

Semestre de Avance 6 - Escuela de Economía y Administración. Tabla 4. Panel B.

Clúster 1 (172): Su desempeño académico es promedio. Este grupo está compuesto sólo por mujeres las que tienden a provenir de colegios particulares subvencionados o municipales. Son uno de los grupos que poseen los menores ingresos familiares promedio (tramo 3), junto al clúster 5. La cantidad de años promedio de educación de los padres no alcanza la universitaria completa, teniendo como resultado un promedio de 12 años.

Clúster 2 (395): Desempeño académico promedio. Este grupo está compuesto únicamente por hombres, quienes mayoritariamente provienen de colegios particulares pagados y en un 100% son colegios con régimen masculinos. Son uno de los grupos que posee el ingreso familiar mayor, ubicándose en el tramo 10. Sumado a lo anterior, poseen los padres con la mayor educación junto al clúster 3, superando en promedio los 15 años.

Clúster 3 (298): Grupo que posee el mejor desempeño académico, ya que tienen el mayor factor de rendimiento (94%) y además poseen la menor tasa de reprobación del semestre anterior (6,9%). Este grupo está compuesto solo por mujeres, las que generalmente provienen de colegios particulares pagados y de tipo coeducacional. Poseen los más altos ingresos brutos (Tramo 10) y a la vez tiene a los padres con la mayor cantidad de años educativos, alcanzando los 16 años promedio.

Clúster 4 (195): Poseen un desempeño académico promedio. Este grupo está compuesto sólo por hombres, los cuales provienen generalmente de colegios particulares subvencionados con

régimen masculinos. El ingreso familiar está en el promedio y la educación de los padres tiende a ser media-alta, ya que alcanza los 14 años.

Clúster 5 (123): Grupo con el menor desempeño académico, a pesar de que sigue siendo relativamente bueno. Su factor de rendimiento es de un 89,7% y la tasa de reprobación alcanza un 11,3%. Este grupo está compuesto solamente por hombres, los cuales provienen generalmente de colegios subvencionados o municipales y a la vez tienden a ser coeducacionales.

En el sexto semestre podemos apreciar clúster muy similares tanto en el factor rendimiento como en la tasa de reprobación. La variable Tramo arroja promedios muy similares para todos los clústeres, variando de 4,7 a 4,8. Los clústeres siguen diferenciándose principalmente por las variables socio-demográficas, como lo son el ingreso, el grupo de dependencia y la educación de los padres. Es importante mencionar la similitud entre clústeres, los cuales tienden a diferenciarse únicamente por el género y el grupo de dependencia.

Tabla 4. Clúster y Test de Medias Escuela de Economía y Administración
Panel A. Escuela de Economía y Administración

	Clúster 1				Clúster 2				Clúster 3				Clúster 4				One-way ANOVA			
Variable	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**F-Test**	**gl1**	**gl2**	**p-value**
FACTOR REND SEM ANTERIOR	0.917	0.113	0.008	202	0.873	0.156	0.016	90	0.89	0.127	0.007	328	0.872	0.134	0.008	257	4,944	3	873	0.2%
% _REPROB SEM ANTERIOR	0.093	0.151	0.011	202	0.144	0.202	0.021	90	0.131	0.172	0.009	328	0.146	0.179	0.011	257	4,014	3	873	0.8%
TRAMOS	4,629	0.75	0.053	202	4,700	0.694	0.073	90	4,659	0.69	0.038	328	4,658	0.701	0.044	257	0.218	3	873	88.4%
GRUPO_DEPENDENCIA	18,639	3,492	0.246	202	17,333	4,983	0.525	90	19,223	2,807	0.155	328	6,887	2,428	0.151	257	860,297	3	873	0.0%
TIENE_TRABAJO_REM	0.011	0.103	0.007	189	0	0	0	80	0.007	0.118	0.007	289	0.012	0.109	0.007	251	0.309	3	805	81.9%
HORAS_QUE_DEDICA_TRABAJO	0.099	1,274	0.09	202	0	0	0	90	0.146	2,650	0.146	328	0.226	3,138	0.196	257	0.226	3	873	87.8%
GRUPO_FAMILIAR	4,851	2,531	0.178	202	4,711	2,781	0.293	90	4,598	2,641	0.146	328	4,054	1,568	0.098	257	4,945	3	873	0.2%
INGRESO_BRUTO_FAM	10,510	2,660	0.187	202	9,889	3,272	0.345	90	10,793	2,479	0.137	328	3,261	1,934	0.121	257	531,796	3	873	0.0%
?VIVEN_SUS_PADRES?	4,899	0.541	0.04	179	4,922	0.48	0.055	77	4,755	0.867	0.052	273	4,360	1,310	0.083	250	14,813	3	775	0.0%
EDU_PADRE	16,179	1,707	0.132	168	16,260	1,491	0.175	73	16,312	1,591	0.098	266	12,234	3,239	0.221	214	164,295	3	717	0.0%
EDU_MADRE	15,860	1,840	0.14	172	15,493	2,075	0.24	75	15,770	1,832	0.112	269	11,963	3,167	0.203	244	142,551	3	756	0.0%
SEXO	2,000	0	0	202	1,000	0	0	90	1,000	0	0	328	1,358	0.48	0.03	257	671,877	3	873	0.0%
REGIMEN	2,827	0.379	0.027	202	1,000	0	0	90	3,000	0	0	328	2,549	0.739	0.046	257	506,634	3	873	0.0%

Panel B. Escuela de Economía y Administración

	Clúster 1				Clúster 2				Clúster 3				Clúster 4				Clúster 5				One-way ANOVA			
Variable	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**F-Test**	**gl1**	**gl2**	**p-value**
FACTOR REND SEM ANTERIOR	0.901	0.098	0.007	172	0.922	0.085	0.004	395	0.941	0.082	0.005	298	0.92	0.083	0.006	195	0.897	0.099	0.009	123	8,290	4	1178	0.0%
% _REPROB SEM ANTERIOR	0.111	0.165	0.013	172	0.078	0.134	0.007	395	0.069	0.125	0.007	298	0.075	0.127	0.009	195	0.113	0.173	0.016	123	4,139	4	1178	0.2%
TRAMOS	4,750	0.879	0.067	172	4,701	0.769	0.039	395	4,822	0.728	0.042	298	4,882	0.761	0.054	195	4,886	0.791	0.071	123	2682	4	1178	3.0%
GRUPO_DEPENDENCIA	8,750	4,228	0.322	172	19,519	2,143	0.108	395	19,664	1,804	0.104	298	13,359	7,007	0.502	195	8,780	3,093	0.279	123	457,720	4	1178	0.0%
TIENE_TRABAJO_REM	0.013	0.114	0.009	153	0.021	0.179	0.01	340	0.018	0.177	0.011	284	0.024	0.217	0.017	169	0	0	0	123	0.448	4	1064	77.4%
HORAS_QUE_DEDICA_TRABAJO	0.035	0	0.025	172	0.111	1070	0.054	395	0.178	2,626	0.152	298	0.097	1	0.069	195	0	0	0	123	0.416	4	1178	79.7%
GRUPO_FAMILIAR	3,698	2,041	0.156	172	4,253	2,600	0.131	395	4,899	2,420	0.14	298	4,179	2,666	0.191	195	4,309	1,505	0.136	123	7,491	4	1178	0.0%
INGRESO_BRUTO_FAM	3,971	2,739	0.209	172	10,322	2,748	0.138	395	10,379	2,679	0.155	298	7,744	4,070	0.291	195	3,675	2,125	0.192	123	255,704	4	1178	0.0%
?VIVEN_SUS_PADRES?	4,463	1183	0.097	149	4,838	0.679	0.038	315	4,900	0.539	0.033	271	4,688	1	0.075	160	4,695	1	0.084	118	8,157	4	1008	0.0%
EDU_PADRE	13,812	3,063	0.266	133	16,349	1,383	0.078	312	16,408	1,915	0.118	265	14,743	2,874	0.233	152	12,619	3,455	0.325	113	81,999	4	970	0.0%
EDU_MADRE	13,430	2,656	0.223	142	15,773	1,850	0.105	309	16,071	1,636	0.1	266	14,168	3,055	0.245	155	12,595	2,984	0.277	116	75,664	4	983	0.0%
SEXO	2,000	0	0	172	1,000	0	0	395	2,000	0	0	298	1,000	0	0	195	1,000	0	0	123	n.a.	n.a.	n.a.	n.a.
REGIMEN	2,587	0.494	0.038	172	3,000	0	0	395	2,826	0.389	0.023	298	1,000	0	0	195	2,992	0.09	0.008	123	2,000,299	4	1178	0.0%

Semestre de Avance 2 - Escuela de Sistemas de Información y Auditoría Tabla 5. Panel A

Clúster 1 (144): Está conformado sólo por mujeres, las que generalmente provienen de colegios con régimen coeducacional. Con un factor de rendimiento del 84,4% y una tasa de reprobación entorno al 20%.

Clúster 2 (51): Grupo que en general proviene de colegio municipal. Pertenecen al tramo 2 de ingreso bruto familiar. Además sus padres son los que tienen en promedio la menor cantidad de años en la educación (10 años). Está compuesto tanto por hombres como por mujeres, los que generalmente provienen colegios con régimen coeducacional. Además, muestra una factor de avance y tasa de reprobación del semestre anterior levemente menor al clúster 1 sin ser muy distinto a los otros grupos.

Clúster 3 (162): Grupo que posee el mayor factor de rendimiento (87%). Generalmente provienen de colegios particulares subvencionados o particulares pagados. Sumado a lo anterior, son el grupo que posee en promedio el mayor ingreso bruto familiar (Tramo 7). Está compuesto únicamente por hombres, los que provienen en su totalidad de colegios coeducacionales.

Clúster 4 (64): Es el segundo grupo con menor factor de rendimiento (83,7%) y el que posee la mayor tasa reprobación del semestre anterior (20%). Generalmente provienen de colegios particulares subvencionados. Son los únicos que no trabajan. Está compuesto sólo hombres, los que vienen en un 100% de colegios masculinos.

Todos los clústeres en general poseen un factor de rendimiento entre el 82% y el 87% y una tasa de reprobación del semestre anterior sobre el 15%, estando todos bastante cercanos en los resultados académicos.

Semestre de Avance 6 - Escuela de Sistemas de Información y Auditoría. Tabla 5. Panel B.

Clúster 1 (103): Poseen la menor tasa de reprobación (11%) y un factor de rendimiento de 90%. Son únicamente hombres, los que generalmente provienen de colegios particulares subvencionados pero de tipo coeducacional. Son los que poseen mayor cantidad de integrantes en su familia y a la vez el mayor ingreso bruto familiar.

Clúster 2 (169): Poseen el mayor factor de rendimiento, cercano al clúster anterior (90,7%). Está formado sólo por mujeres, las que en su mayoría provienen de colegios particulares subvencionados. Son el grupo que posee menor ingreso bruto familiar, acercándose al tramo 5.

Clúster 3 (48): Grupo con el menor rendimiento académico, ya que son los que poseen un factor de rendimiento menor (86,9%) y una tasa de reprobación mayor (16,1%). Este grupo está compuesto solo por hombres, los cuales tienden a provenir de colegios particulares subvencionados con régimen masculinos.

Las variables Tramo y educación de padres (13 años aproximadamente) son similares en los tres clústeres que se formaron. Un resultado a destacar es que lo que diferencia al clúster 1 del clúster 3 es el factor de rendimiento del semestre anterior. Además es importante mencionar que las diferencias socio-demográficas que se tendían a repetir en los clústeres anteriores, como lo eran las variables grupo de dependencia y educación de los padres.

Tabla 5. Clúster y Test de Medias Escuela de Sistemas de Información y Auditoria

Panel A. Escuela de Sistemas de Información y Auditoría

	Clúster 1				Clúster 2				Clúster 3				Clúster 4				One-way ANOVA			
Variable	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**F-Test**	**gl1**	**gl2**	**p-value**
FACTOR REND SEM ANTERIOR	0.844	0.154	0.013	143	0.821	0.142	0.02	51	0.87	0.156	0.012	162	0.837	0.163	0.02	64	1,707	3	416	16.5%
% _REPROB SEM ANTERIOR	0.194	0.22	0.018	143	0.184	0.209	0.029	51	0.161	0.215	0.017	162	0.204	0.239	0.03	64	1	3	416	46.1%
TRAMOS	4,597	0.796	0.066	144	4,392	0.603	0.084	51	4,586	0.745	0.059	162	4,391	0.633	0.079	64	2091	3	417	10.1%
GRUPO_DEPENDENCIA	12,465	6,173	0.514	144	7,255	2,513	0.352	51	14,012	5,706	0.448	162	10,625	5,669	0.709	64	20,767	3	417	0.0%
TIENE_TRABAJO_REM	0.023	0.194	0.017	132	0.02	0.143	0.02	49	0.052	0.25	0.02	154	0	0	0	59	1164	3	390	32.3%
HORAS_QUE_DEDICA_TRABAJO	0.222	1,938	0.161	144	0.392	2801	0.392	51	0.889	5,467	0.43	162	0	0	0	64	1252	3	417	29.1%
GRUPO_FAMILIAR	4,146	1,843	0.154	144	4,059	1,475	0.207	51	4,043	1,843	0.145	162	3,875	1,732	0.217	64	0	3	417	79.4%
INGRESO_BRUTO_FAM	6,243	3,820	0.318	144	2,647	1,180	0.165	51	7,340	3,916	0.308	162	6,719	3,574	0.447	64	22,204	3	417	0.0%
¿VIVEN_SUS_PADRES?	4,600	1097	0.096	130	4,061	1547	0.221	49	4,400	1204	0.1	145	4,702	1	0.132	57	3,327	3	377	2.0%
EDU_PADRE	14,246	3,032	0.279	118	10,652	2,877	0.424	46	14,624	2,476	0.215	133	14,054	2,706	0.362	56	24,940	3	349	0.0%
EDU_MADRE	13,786	2,930	0.261	126	10,936	3,010	0.439	47	14,280	2,636	0.22	143	14,263	2,539	0.336	57	18,465	3	369	0.0%
SEXO	2,000	0	0	144	1,549	0.503	0.07	51	1,000	0	0	162	1,000	0	0	64	980,811	3	417	0.0%
REGIMEN	2,701	0.489	0.041	144	2,804	0.401	0.056	51	3,000	0	0	162	1,000	0	0	64	635,674	3	417	0.0%

Panel B. Escuela de Sistemas de Información y Auditoría

	Clúster 1				Clúster 2				Clúster 3				One-way ANOVA			
Variable	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**Media**	**Desv. Est.**	**Error Est.**	**n**	**F-Test**	**gl1**	**gl2**	**p-value**
FACTOR REND SEM ANTERIOR	0.9	0.105	0.01	103	0.907	0.095	0.007	169	0.869	0.105	0.015	48	2,674	2	317	7.1%
% _REPROB SEM ANTERIOR	0.111	0.156	0.015	103	0.128	0.182	0.014	169	0.161	0.178	0.026	48	1,366	2	317	25.7%
TRAMOS	4,583	0.799	0.079	103	4,615	0.824	0.063	169	4,646	0.956	0.138	48	0.103	2	317	90.2%
GRUPO_DEPENDENCIA	12,864	6,007	0.592	103	10,562	5,861	0.451	169	10,521	6,461	0.933	48	5,182	2	317	0.6%
TIENE_TRABAJO_REM	0.02	0.14	0.014	101	0	0	0	154	0.143	0.417	0.064	42	11001	2	294	0.0%
HORAS_QUE_DEDICA_TRABAJO	0.485	4,456	0.439	103	0	0	0	169	2667	10,769	1554	48	5652	2	317	0.4%
GRUPO_FAMILIAR	4,583	1,332	0.131	103	3,834	1,782	0.137	169	3,688	1,764	0.255	48	7,987	2	317	0.0%
INGRESO_BRUTO_FAM	6,155	3,816	0.376	103	4,882	3,592	0.276	169	5,292	3,087	0.446	48	4,021	2	317	1.9%
¿VIVEN_SUS_PADRES?	4,820	0.716	0.072	100	4,653	1030	0.084	150	4,302	1282	0.196	43	4,211	2	290	1.6%
EDU_PADRE	13,879	3,008	0.302	99	13,182	3,803	0.325	137	13,625	3,295	0.521	40	1,193	2	273	30.5%
EDU_MADRE	13,768	3,047	0.306	99	13,336	2,955	0.247	143	13,714	2,472	0.381	42	1	2	281	48.5%
SEXO	1,000	0	0	103	2,000	0	0	169	1,000	0	0	48	n.a.	n.a.	n.a.	n.a.
REGIMEN	3,000	0	0	103	2,586	0.494	0.038	169	1,000	0	0	48	521,357	2	317	0.0%

Discusión y Conclusiones

Se observa que el porcentaje de reprobación va a la baja a medida que avanzan los semestres. Además, se puede observar que la mayor disminución en la tasa de reprobación se encuentra en el grupo compuesto mayoritariamente de estudiantes provenientes de colegios municipales, el cual es un 15% en el segundo semestre y que en los últimos 4 semestres se estabiliza entorno al 10%. Se puede observar que el factor de rendimiento, en la Escuela de Economía y Administración, mejora para ambos clúster a medida que avanza el tiempo de estadía en la Facultad. Para el clúster compuesto mayoritariamente de estudiantes provenientes de colegio particular este valor se estabiliza entorno al 94% en los últimos tres semestres. Mientras que para el clúster compuesto mayoritariamente de estudiantes provenientes de colegio municipal este valor se estabiliza entorno al 90%.

Las principales conclusiones que nos entregó el análisis de los resultados obtenidos, fue que en la Escuela de Economía y Administración, las variables que tienden a diferenciar a los estudiantes en los primeros semestres son atributos del tipo socio-demográficos y académicos, mientras que para los últimos semestres las diferencias entre los grupos de estudiantes se generan únicamente por las variables socio-demográficas. Los resultados académicos tienden a ser levemente mejores para los estudiantes que provienen mayoritariamente de colegios particulares en los primeros semestres. Sin embargo, esta diferencia comienza a desaparecer a medida que avanzan los semestres, llegando a diferencias académicas muy pequeñas en el sexto semestre manteniendo la diversidad en las variables de socio-demográficas.

Por otra parte, en la Escuela de Sistemas de Información y Auditoria se puede observar que las variables socio-demográficas son similares entre los grupos. Las principales diferencias se pueden ver por el lado del desempeño académico en los primeros semestres, las que desaparecen a medida que avanzan los semestres. En esta Escuela no solo aumenta el rendimiento del 84% al 89% a medida que avanzan los semestres, sino que además disminuye la diferencia inicial con la Escuela de Economía y Administración en más de la mitad.

Vemos que es importante identificar y conocer los distintos grupos de alumnos existentes en las Escuelas de Pregrado, ya que en el futuro se puedan aplicar medidas preventivas y particulares a un determinado grupo de alumnos dependiendo del clúster al que pertenecen. Ello con el fin de potenciar y continuar el apoyo entregado a los estudiantes por la Facultad, evitando situaciones no deseadas por los estudiantes y aumentando las tasas de retención de ambas Escuelas de Pregrado, lo que va en beneficio de toda la comunidad universitaria. Finalmente, se debe destacar que frente a una gran diversidad socio-demográficas en los grupos de estudiantes, se logra disminuir la tasa de reprobación y el factor de rendimiento en la Facultad para todos los grupos de estudiantes a medida que avanzan los semestres.

Referencias

- Chapman, P., Clinton, J., Keber y otros (2000). "CRISP-DM 1.0 Step by step guide". SPSS (www.crisp-dm.org/CRISPWP-0800.pdf).
- Dunham, M. (2002)." Data Mining, Introductory and Advanced Topics", Prentice Hall.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence.
- Gartner Group 1994. Data mining: The next generation of business intelligence? ATG Research Note T-517-246, Gartner Group Inc., Stamform, CT.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM Computing Surveys. doi:10.1145/331499.331504
- Kaufmann, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis [Hardcover]. New York: Wiley-Interscience; 99 edition. Retrieved from http://www.amazon.com/Finding-Groups-Data-Introduction-Analysis/dp/0471878766
- Ortega, C., Lee, M., Silva, D. & Vásquez, J. 2015. "Carga Académica: Identificación de factores claves en una escuela de Economía y Negocios", Viña del Mar, Chile CLADEA, pp 157
- Romero, C., Ventura, S. (2010). "Educational Data Mining: A Review of the State of the Art". IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews.