

Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos

Discovery of profiles of student performance: a model of integration of academic and socioeconomic data

David L. La Red Martínez¹, Marcelo Karanik¹, Mirtha Giovannini¹,
María E. Báez¹, Juliana Torre¹

¹ Universidad Tecnológica Nacional, Argentina

lrmdavid@exa.unne.edu.ar , marcelo@frre.utn.edu.ar , meg_c51@yahoo.com.ar ,
maeugeniabaez@gmail.com , juliana.itorre@gmail.com

RESUMEN. Uno de los mayores problemas que enfrentan las universidades en Argentina, y que cada día toma mayor relevancia, es la alta tasa de deserción estudiantil, la cual se ve reflejada en el número de graduados, que en algunos casos no llega a la mitad de estudiantes. Para encontrar una solución a esta problemática se plantea la necesidad de estudiar sus causas, para lo cual se busca encontrar patrones entre las características de los estudiantes, y definir así perfiles que conduzcan al éxito o fracaso académico. Fundado en esto, este trabajo describe un modelo basado en técnicas de Data Mining para determinar los perfiles de rendimiento académico en la asignatura Algoritmos y Estructura de Datos de la carrera Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Resistencia (UTN-FRRe). Empleando los datos de los alumnos que cursaron la antedicha asignatura en el ciclo lectivo 2014, se procuró determinar en qué medida el desigual desempeño de los mismos es influenciado por otras variables de interés tales como los factores económicos, demográficos, sociales y culturales. En función a estas variables y a partir de técnicas de clasificación y determinación de patrones, se crearon perfiles de rendimiento académico con el objetivo principal de utilizar aquellos tendientes al fracaso o deserción como base a la determinación de futuras políticas de gestión académica que podrían implementarse para reducir dicho fenómeno.

ABSTRACT. One of the biggest problems faced by the universities in Argentina, and that each day takes greater relevance, is the high dropout rate, which is reflected in the number of graduates, which in some cases is less than half of students. To find a solution to this problem it is necessary to study its causes, for which it seeks to find patterns between student characteristics, and define profiles that lead to success or academic failure. Based on the above, this paper describes a system based on Data Mining model to determine the profiles of academic performance in the course Algorithms and Data Structure, of the Engineering in Information Systems, National Technological University, Resistencia Regional Faculty (UTN-FRRe). Using data from students who attended the above subject in the school year 2014, is aimed to determine to what extent the uneven performance of them is influenced by other variables of interest such as the economic, demographic, social, and cultural factors. Using these variables, and classification and determination of pattern techniques, profiles of academic performance were created with the main objective of using profiles that indicate trends of failure or dropping out, to determine future policies of academic management that could be implemented to reduce this phenomenon.

PALABRAS CLAVE: Rendimiento académico, Minería de datos, Reducción fracaso académico, Gestión académica.

KEYWORDS: Academic achievement, Data Mining, Reduced academic failure, Academic Management.

La Red, D. L.; Karanik, M.; Giovannini, M.; Báez, M. E.; Torre, J. (2016). Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos. *Campus Virtuales*, 5(2), 70-83.

1. Introducción

La calidad académica es un factor fundamental para el funcionamiento de las instituciones educativas. Por ello, las universidades analizan e implementan constantemente estrategias y herramientas que les permitan elevar el nivel académico. La determinación adecuada de éstas se basa en la evaluación correcta de varios aspectos como los contenidos dictados, las metodologías de enseñanza, las tecnologías disponibles, los recursos con los que se cuenta, las capacidades de los docentes y el rendimiento de los alumnos. Es decir, saber qué acciones llevar a cabo implica conocer cuáles son las debilidades y fortalezas como así también las amenazas y oportunidades. El rendimiento académico es la productividad de un alumno (Maletic et al., 2002) y está relacionado con la asimilación correcta de los contenidos, con las demás actividades que tiene y sus características personales (sociales e individuales), siendo un elemento fundamental de análisis ya que puede reflejar muchas características de las instituciones.

Dicho rendimiento se encuentra íntimamente ligado a las calificaciones obtenidas por los estudiantes cuando sus conocimientos, capacidades y aptitudes son evaluados. Pero esta evaluación no permite obtener información útil que permita detectar y corregir problemas cognitivos, de aprehensión, etc. Por este motivo hay que analizar aquellos factores que se encuentran detrás de las calificaciones, factores sociales, económicos, y de información previa del alumno, y de esta forma poder establecer perfiles de desempeño estudiantil.

En este trabajo se examina el rendimiento académico en la asignatura Algoritmos y Estructuras de Datos, del primer nivel de la carrera de Ingeniería en Sistemas de Información (ISI) de la UTN-FRRe, donde la tasa de alumnos libre, es decir, aquellos que deberán re-cursar la materia por no haber aprobado las instancias parciales, es muy alta. Por lo cual surge la necesidad de analizar la existencia de patrones de comportamiento que permitan distinguir diferentes perfiles de rendimiento académico.

Existen varias maneras de evaluar el desempeño académico: utilizando indicadores como las tasas de graduación, diferenciando por tipos de centros y analizando el rendimiento a partir de datos individuales (García & San Segundo, 2001); a través de las calificaciones obtenidas en los seminarios de ingreso a la Universidad, realizando el análisis de los datos mediante la técnica estadística de curva ROC (Receiver Operating Characteristic) (Vivo Molina et al., 2004); o aplicando el enfoque de función de producción para estimar los determinantes del rendimiento académico (Di Gresia, 2007). También existen varios estudios que utilizan técnicas matemáticas para la evaluación del desempeño. En este sentido, se ha estudiado la capacidad de la regresión lineal y de la regresión logística en la predicción del rendimiento y del éxito/fracaso académico, partiendo de variables como la asistencia y la participación en clase (García-Jiménez et al., 2000).

Una alternativa es desarrollar los métodos de evaluación aprovechando las tecnologías de información disponibles y sus beneficios. En este sentido las técnicas de Data Warehouse (DW) y Data Mining (DM) son herramientas de suma utilidad para la obtención de conocimiento en grandes volúmenes de datos y aún en volúmenes reducidos (Jackson, 2002), (Natek y Zwilling, 2013), (Maindonald, 2016), (Dell Software, 2016).

Un DW es una colección de datos orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar a tomar decisiones (Salton, 1989). Los DW surgieron por dos razones: primero, la necesidad de proporcionar una fuente única de datos limpia y consistente para propósitos de apoyo para la toma de decisiones; y segundo, la necesidad de hacerlo sin afectar a los sistemas operacionales (Molina López y García Herrero, 2012).

DM es un área de las ciencias de la computación referido al proceso que intenta descubrir patrones en conjuntos de datos (Witten & Eibe, 2005), (Hand et al., 2001), (Pang-Ning et al., 2006), (Maimon y Rokach, 2010). Utiliza los métodos de Inteligencia Artificial y del Aprendizaje de Máquinas para extraer patrones y tendencias de forma que posibiliten la descripción y comprensión de los datos y también la predicción de comportamientos futuros.



En este trabajo se propone la utilización de técnicas de DW y DM sobre la información del desempeño de los alumnos de la cátedra Algoritmos y Estructura de Datos de la Carrera de Ingeniería en Sistemas de Información de la Facultad Regional Resistencia de la UTN, durante el ciclo lectivo 2014. El artículo se encuentra estructurado de la siguiente manera: en la Sección 2 se explican los conceptos básicos sobre DM haciendo hincapié en la DM en la Educación. En la Sección 3 se exhibe el diseño del modelo propuesto. En la Sección 4 se muestran los resultados obtenidos y finalmente, en la Sección 5, las conclusiones obtenidas.

2. Minería de Datos

La DM es un proceso que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores aparentemente caóticos que tienen una explicación que pueden descubrirse mediante diversas técnicas de esta herramienta (Fayyad et al., 1996). El objetivo fundamental es aprovechar el valor de la información localizada y usar los patrones preestablecidos para que los directivos tengan un mejor conocimiento de su negocio y puedan tomar decisiones más confiables (Gil y Rao, 1996).

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los siguientes pasos básicos (Microsoft, 2016): a) definir el problema, b) preparar los datos, c) explorar los datos, d) generar modelos, e) explorar y validar los modelos, f) implementar y actualizar los modelos.

En la figura 1 se puede visualizar a cada uno de los pasos del proceso, el cual es cíclico, lo que significa que la creación de un modelo de DM es un proceso dinámico e iterativo (Microsoft, 2016).

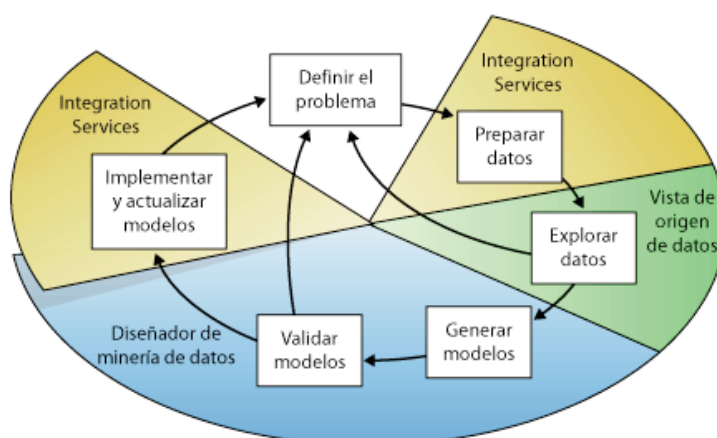


Figura 1. Proceso MD (Microsoft, 2016).

2.1. Algoritmos de Data Mining

Un algoritmo de DM es un conjunto de cálculos y reglas heurísticas que permite crear un modelo de DM a partir de los datos (Wu et al., 2008). Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de DM. A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas. Algunos de los algoritmos más utilizados en tal sentido son los de árboles de decisión y los de clustering demográfico; se han utilizado estos algoritmos debido a que permiten crear grupos de alumnos con perfiles concretos en función de atributos definidos:

- Árboles de decisión, este algoritmo genera un modelo de DM mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. El aprendizaje de árboles de decisión está englobado como una metodología del aprendizaje supervisado, que suele ser más robusto frente al ruido y conceptualmente sencillo, aunque los sistemas que han resultado del perfeccionamiento y de la evolución de los más antiguos se complican con los procesos que incorporan para ganar fiabilidad. La mayoría de los sistemas de aprendizaje de árboles suelen ser no incrementales, pero existe alguna excepción (Quinlan, 1986), (Rokach y Maimon, 2008), (Molina López y García Herrero, 2012), (Letham et al., 2015).

- Clustering Demográfico: Proporciona una agrupación rápida y natural de las bases de datos de gran tamaño. El número de agrupaciones que se deben crear se determina automáticamente, y éstas se caracterizan por las distribuciones de valor de sus miembros. Se calcula la similitud de cada registro con cada una de las agrupaciones existentes en la actualidad. Si la similitud mayor calculada está por encima de un determinado umbral, el registro se añade a la agrupación relevante. Las características de esta agrupación cambian de acuerdo a esto. Si la similitud calculada no está por encima del umbral o en el caso de que no haya agrupación (como el caso inicial), se crea una agrupación nueva que contiene sólo el registro. Se puede especificar el número máximo de agrupaciones, así como el umbral de similitud (Michaud, 1999), (Grabmeier y Rudolph, 2002), (IBM Knowledge Center, 2016).

2.2. Data Mining en la Educación

Existen varios campos de aplicación del DM, uno de ellos es la Educación (Educational DM, EDM). Su objetivo principal es el desarrollo de métodos para la exploración de tipos de datos únicos provenientes de plataformas educativas, y usando estos métodos, entender mejor a los estudiantes en el proceso de aprendizaje (Baker y Yaceff, 2009). En tal sentido se han aplicado gran variedad de técnicas de DM a diversos contextos educativos. En cada caso, el objetivo es traducir los datos en bruto en información significativa sobre el proceso de aprendizaje con el fin de tomar mejores decisiones sobre el diseño y la trayectoria de un ambiente de aprendizaje (Baker, 2010), (Romero y Ventura, 2010).

Existen algunas diferencias y/o ventajas entre aplicar DM con respecto a sólo utilizar modelos estadísticos (Aluja, 2001):

- La DM es más amplia ya que es un proceso completo formado por varias etapas y que incluye muchas técnicas, entre ellas, las estadísticas. Este proceso de descubrimiento de información está formado por las etapas de pre-procesado, la aplicación de técnicas de DM (una de ellas puede ser estadística) y la evaluación e interpretación de los resultados.

- En las técnicas estadísticas (análisis de datos) se suele utilizar como criterio de calidad la verosimilitud de los datos dado el modelo. En DM se suele utilizar un criterio más directo, por ejemplo, el porcentaje de datos bien clasificados.

- En estadística la búsqueda se suele realizar mediante modelización basada en un algoritmo de ascenso de colinas (hill-climbing) en combinación con un test de hipótesis basado en razón de verosimilitud. En DM se suele utilizar una búsqueda basada en meta-heurísticas.

- La DM está orientada a trabajar con cantidades muy grandes de datos (millones y billones de datos), en cambio la estadística no suele funcionar tan bien en bases de datos de tan gran tamaño y alta dimensionalidad.

3. Modelo Propuesto

Para la implementación del modelo propuesto se utilizaron las herramientas incluidas en el IBM Data Warehouse Edition (DWE) V.9.5 y el IBM Info Sphere Warehouse (ISW) V.9.7 (IBM, 2005), que se compone de varios productos, pero a los fines de este trabajo se han usado los siguientes componentes:

- El Sistema de Gestión de Base de datos DB2 Enterprise Server Edition (DB2 ESE), que permite



La Red, D. L.; Karanik, M.; Giovannini, M.; Báez, M. E.; Torre, J. (2016). Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos. *Campus Virtuales*, 5(2), 70-83.

conexiones multiusuarios y presenta una alta escalabilidad.

- Design Studio (DS): Una plataforma para diseñar las reglas de las cargas de trabajo, los flujos de transformación de datos y los flujos analíticos para la DM y el análisis de textos.
- Intelligent Miner (IM): Conjunto de funcionalidades que permiten realizar análisis de información de acuerdo a conceptos de Business Intelligence (BI).

3.1. Proceso de Captura, Almacenamiento y Análisis de Datos

Para el proceso de captura de datos se requirió la participación de los estudiantes, pues era necesario conocer los aspectos personales que no se podían obtener de otra manera si no era a través de respuestas directas por parte de los alumnos. A tal fin, se utilizó una aplicación web que permitió contar con una Encuesta On-Line compuesta por preguntas relacionadas a situación socio-económica familiar, historial de estudios secundarios, entre otras cuestiones, incluyendo aspectos actitudinales del alumno respecto del estudio y de las TIC.

Los datos obtenidos de la manera antedicha pasaron por un proceso de ETL (Extract, Transform and Load) el cual extrae los datos de los sistemas fuentes, los cuales pasan por una secuencia de transformaciones antes de ser almacenados en el DW para luego utilizarlos para los procesos de DM.

La estructura del modelo de DW utilizado se compone de la tabla de hechos Alumno y varias tablas de dimensiones asociadas a la misma, que incluyen características que se desean estudiar. La tabla de hechos incluye información específica del alumno y su rendimiento académico, en tanto que las tablas de dimensiones contienen información que hace a la descripción del entorno socio-económico del alumno y familia, sus antecedentes académicos en los estudios secundarios y su actitud hacia el estudio y hacia las TIC (situación laboral de la madre, del padre y propia; importancia que da al estudio; importancia que da a las TICs; residencia actual; lugar de procedencia; tipo de escuela secundaria donde cursó sus estudios; horas dedicadas al estudio) (figura 2).



Figura 2. Datos específicos del alumno utilizados para la realización del modelo de DW.

En la etapa de DM se seleccionaron las técnicas a utilizar, creándose los flujos de minería correspondiente, en los cuales se parametrizan los respectivos algoritmos.

Se trabajó con técnicas de Clusterización (Cluster Demográfico) y especialmente con la técnica supervisada de Clasificación con Árboles de Decisión, cuyo objetivo es realizar clasificaciones sobre los datos conocidos y así crear modelos que luego puedan ser utilizados para predecir o clasificar valores nuevos o desconocidos.

El análisis de los resultados se basó en considerar como parámetro de minería la variable relacionada a la Situación Final del alumno, la cual reflejó su estado en la materia al cierre del Ciclo Lectivo.

4. Pruebas y Resultados

Con el objetivo de encontrar perfiles de rendimiento académicos, se realizaron pruebas sobre los datos de los alumnos del ciclo lectivo 2014. Se han considerado los diferentes factores culturales, sociales, económicos y actitudinales teniendo en cuenta su condición final al cierre del año, es decir, si el alumno se encontraba regular, promovido o libre. Se consideró en situación de “Libre” al alumno que no aprobó ni los exámenes parciales ni los recuperatorios, “Regular” a quien logró aprobar los 3 exámenes parciales con nota mayor o igual al 60% pero sin alcanzar el 75% en todos ellos, finalmente los alumnos en estado “Promocionado”, son aquellos que aprobaron todos los parciales con nota igual al 75% o superior.

Para definir los perfiles se requirió tanto las encuestas completadas por los alumnos, así como las notas obtenidas por ellos; la cátedra de Algoritmos y Estructuras de Datos de la Facultad Regional Resistencia facilitó esta información. En las pruebas se consideró como parámetro principal de minería la variable Condición Final del Alumno al cierre del ciclo lectivo, habiéndose obtenido los siguientes resultados: 74% de alumnos en condición de libre, 18% de alumnos regulares y por último 8% de alumnos promovidos.

Analizando cada uno de los factores sociales, económicos, culturales y actitudinales, y comparándolos con las notas finales obtenidas por los alumnos, luego de aplicar algoritmos de Árboles de Decisión y Clustering Demográfico se obtuvieron los resultados que se comentarán a continuación.

En la figura 3 se puede visualizar los porcentajes en el criterio “tipo de escuela secundaria”, así como su distinción según la condición final del alumno. El mejor rendimiento académico fue alcanzado por alumnos provenientes de escuelas privadas particulares, el cual asciende a un 38% entre regularizados y promocionados siguiendo con las escuelas provinciales y municipales que alcanzan un 30%. Esto implica que aquellos alumnos que provengan de escuelas privadas particulares tienen mejor rendimiento académico frente a las demás entidades educativas, donde el mayor porcentaje de alumnos libres, 86% provienen de escuelas públicas nacionales.

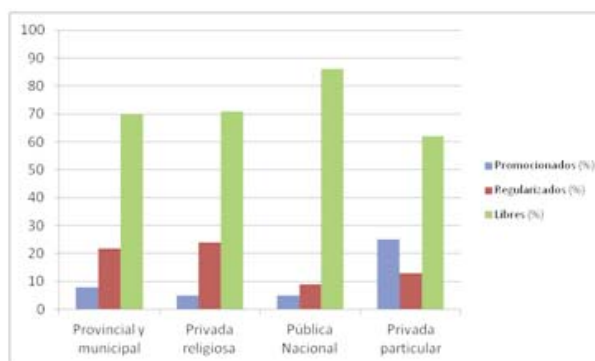


Figura 3. Diferenciación según condición final y tipo de escuela secundaria.

En el ítem visualizado en la figura 4, “horas de estudio semanales”, el mayor porcentaje de promocionados y regularizados se observa en el grupo de quienes estudian más de 10 y hasta 20 horas semanales, el cual alcanza un 28%. En los alumnos libres el mayor porcentaje se da en el grupo que dice estudiar más de 20 horas semanales con un 79%, siguiendo aquellos que estudian hasta 10 horas semanales con un 73%.



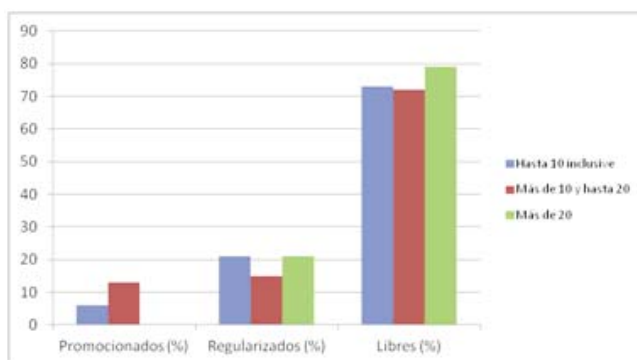


Figura 4. Distinción de la condición final y las horas de estudio semanales.

En cuanto a las “horas trabajadas semanalmente por los alumnos”, cuyos resultados se muestran en la figura 5, se pudo observar que los promocionados no trabajan. En el grupo de alumnos que trabajan hasta 20 horas semanales, el porcentaje de alumnos libres asciende a un 83%. Se puede visualizar que aquellos alumnos que no trabajan tienen mayor rendimiento.

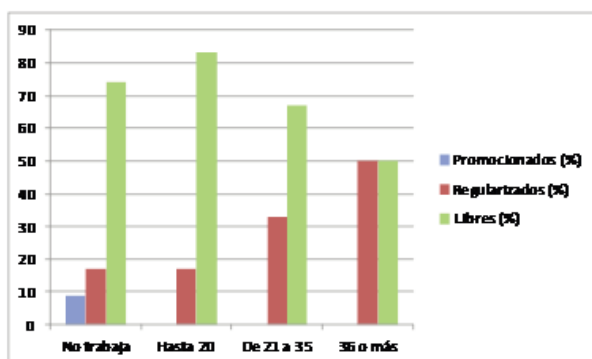


Figura 5. Distinción según condición final del alumno y horas trabajadas.

En la variable “horas trabajadas por la madre semanalmente”, que se muestra en la figura 6, se observa que el mayor porcentaje de éxito académico (promocionado y regularizado) corresponde al grupo cuyas madres trabajan de 21 a 35 horas semanales con un 30%, seguido por las que trabajan más de 36 horas semanales con un 28%. Los porcentajes más altos de alumnos libres corresponden a madres que trabajan hasta 20 horas semanales el cual alcanza un 91% y a madres que no trabajan que obtienen un 74%.

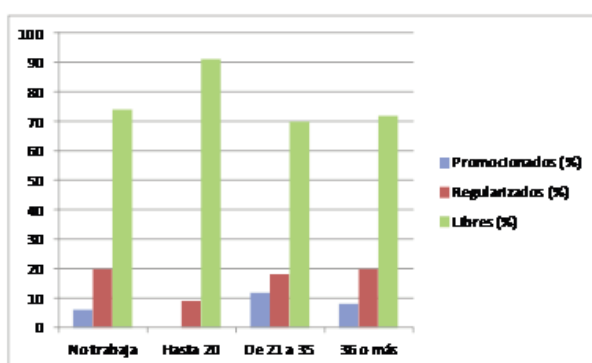


Figura 6. Distinción según condición final del alumno y horas trabajadas por la madre.

En cuanto a las “horas trabajadas por el padre semanalmente”, que se muestra en la figura 7, se ve el mayor porcentaje de éxito académico (promocionado y regularizado) correspondiente al grupo cuyos padres trabajan de 21 a 35 horas semanales, el cual alcanza un 42%. El porcentaje más alto de alumnos libres corresponde al grupo cuyos padres trabajan hasta 20 horas semanales, con un 83%.

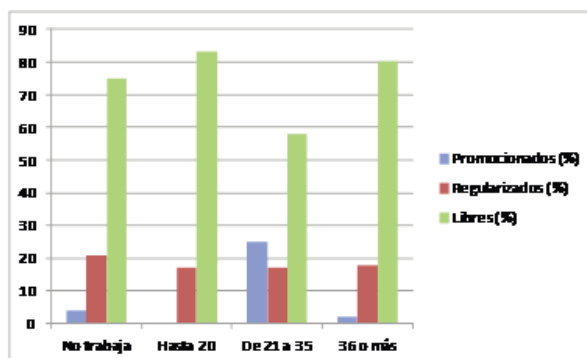


Figura 7. Distinción según condición final del alumno y horas trabajadas por el padre.

En la dimensión de “importancia otorgada al estudio”, que se visualiza en la figura 8, se aprecia un mayor porcentaje de éxito académico (promocionados y regularizados) correspondiente al grupo que informa otorgar más importancia al estudio que a la familia, el cual asciende a un 43%. El porcentaje más alto de fracaso académico (alumnos en condición de libres) corresponde al grupo que informó otorgar más importancia al estudio que al trabajo, llegando a un 79%, seguido por los que dicen otorgar más importancia al estudio que a la diversión, o los cuales corresponde un 77%.

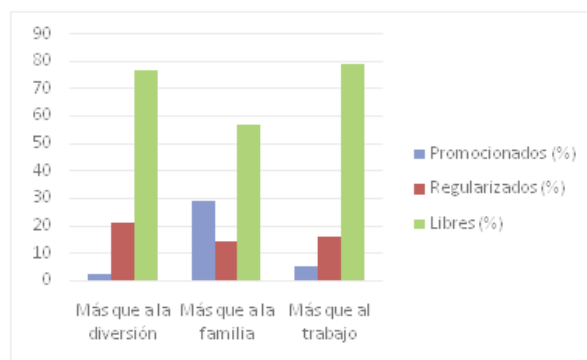


Figura 8. Distinción según condición final del alumno e importancia dada al estudio.

En cuanto a la dimensión “relación del trabajo del alumno con el estudio”, que se grafica en la figura 9, se observa que el mayor porcentaje de éxito académico (promocionados y regularizados) corresponde al grupo que trabaja en temas relacionados con la carrera que estudia, alcanzando un 27%. Los porcentajes más altos de alumnos libres corresponden a quienes trabajan en temas parcialmente relacionados con el estudio que realizan, representando un 80% de dicho grupo, en tanto que los alumnos en condición de libre representan un 75% del grupo de aquellos que trabajan en temas no relacionados con el estudio.

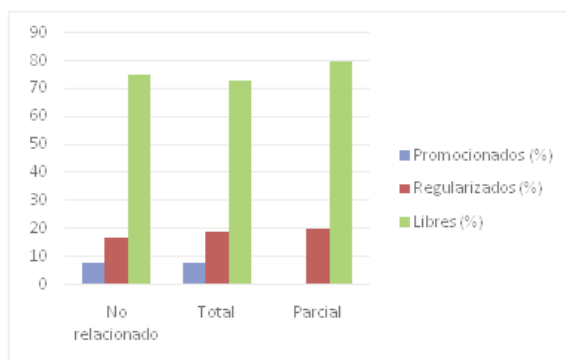


Figura 9. Distinción según condición final del alumno y relación del trabajo con el estudio.

En la dimensión “género”, mostrada en la figura 10, el mayor porcentaje de éxito académico (promocionados y regularizados) corresponde al grupo femenino obteniendo un 29%. Los porcentajes más altos de alumnos libres corresponden al grupo masculino con un 75%.

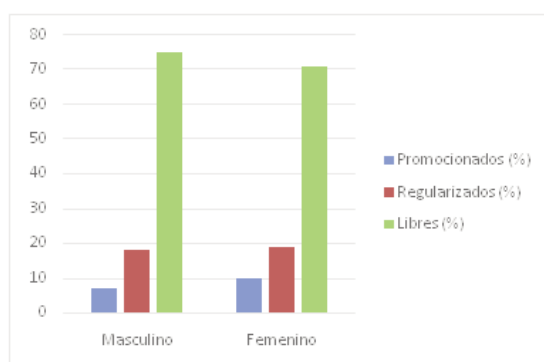


Figura 10. Distinción según condición final del alumno y género.

En cuanto a la dimensión “residencia del alumno”, mostrada en la figura 11, el mayor porcentaje de éxito académico (promocionados y regularizados) corresponde al grupo que reside con familiares, siendo el 28%. Los porcentajes más altos de alumnos libres corresponden al grupo que reside en forma independiente alcanzando un 78% o en alguna situación no prevista el cual asciende a un 80%, esto implica que no residen con familiares, ni en residencia universitaria, ni de manera independiente.

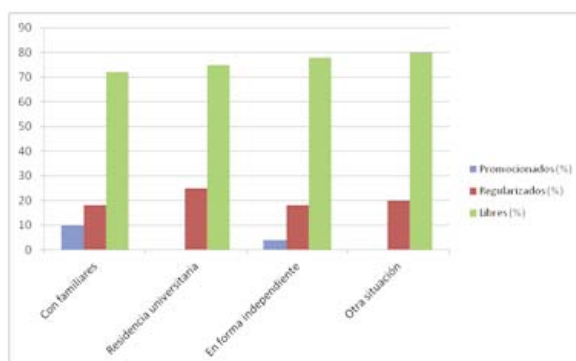


Figura 11. Distinción según condición final del alumno y residencia.

En cuanto a la dimensión de “estudios de la madre”, que se visualiza en la figura 12, el mayor porcentaje de éxito académico (promocionado y regularizado) corresponde al grupo cuyas madres tienen estudios

universitarios de posgrado, con un 50%. En contraposición, los porcentajes más altos de alumnos libres corresponden al grupo cuyas madres no hicieron estudios formales, representando un 100% y al grupo cuyas madres tienen estudios universitarios incompletos alcanzando un 92%. Es de destacar que en el grupo de alumnos cuyas madres tienen estudios de primaria completa, el 100% consiguió la condición final de regularizado.

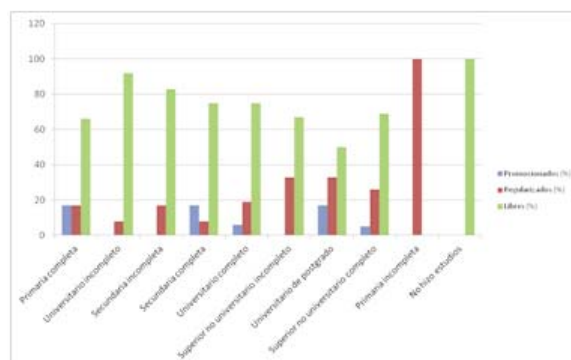


Figura 12. Distinción según condición final del alumno y estudios de la madre.

En la dimensión de “estudios del padre”, que se grafica en la figura 13, el mayor porcentaje de éxito académico (promocionados y regularizados) corresponde al grupo cuyos padres tienen estudios universitarios de posgrado, representando el 50% de dicho grupo. Los porcentajes más altos de alumnos libres corresponden al grupo cuyos padres no hicieron estudios formales, el cual asciende al 100%, y al grupo cuyos padres tienen primaria incompleta, obteniendo también un 100%.

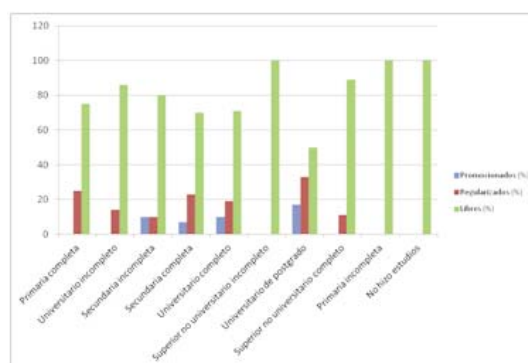


Figura 13. Distinción según condición final del alumno y estudios del padre.

Finalmente, en cuanto a la dimensión “consideración acerca de las TIC”, que se muestra en la figura 14, el mayor porcentaje de éxito académico (promocionados y regularizados) corresponde al grupo que considera que el dominio de las TIC será imprescindible para el ejercicio profesional, obteniendo un 30%. El porcentaje más alto de alumnos libres pertenece al grupo que considera que las TIC facilitan el proceso de enseñanza, alcanzando un 77%.

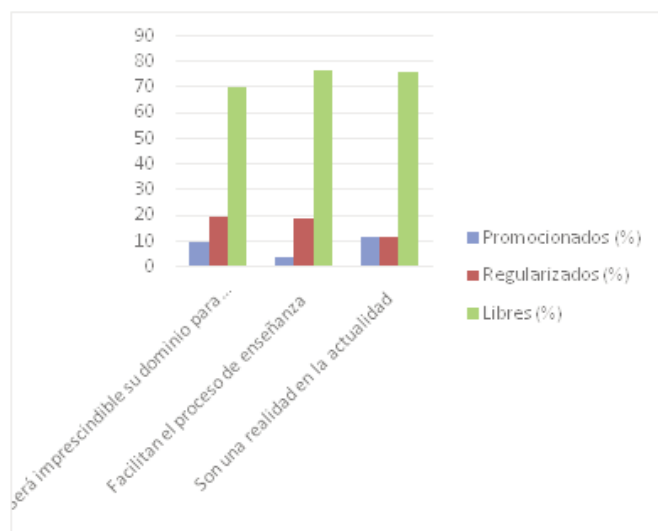


Figura 14. Distinción según condición final del alumno y consideración acerca de las TIC.

5. Conclusiones

El trabajo realizado permitió definir perfiles según el rendimiento académico de los alumnos con la técnica de Data Mining, esto se basó en los resultados obtenidos en el ciclo lectivo 2014, observando tanto la condición final del alumno como la influencia que tiene el entorno socio económico, cultural y actitudinal frente a sus estudios.

Los resultados obtenidos demostraron que los alumnos que dedican menos horas al estudio y le dan mayor importancia tienen mejor rendimiento, lo cual sugiere que la calidad del estudio que presentan los alumnos libres frente a los regulares y promovidos no es la adecuada, vislumbrando una necesidad de mejorar la forma de estudio de los mismos.

Con respecto al trabajo es notorio que aquellos alumnos que no trabajan tienen mayor rendimiento ya que tienen mayor tiempo para dedicarle al estudio, así como también aquellos cuyo empleo se encuentra relacionado con el estudio, esto muestra que el hecho de llevar a la práctica lo estudiado ayuda en el desempeño académico.

Desde la perspectiva de los padres se visualiza que el alumno promovido y regularizado es consciente del esfuerzo que ponen en su trabajo dándole importancia a sus estudios universitarios, vislumbrándose la necesidad de brindarles mayor atención a aquellos alumnos cuyos padres trabajan menos de 20 horas semanales o no realizaron estudios universitarios.

Del análisis de los resultados obtenidos de los procesos de minería de datos, surgieron perfiles característicos de alumnos exitosos académicamente y de aquellos que no lograron dicho éxito (fracaso académico), que se resumen en la tabla 1.

Éxito académico	Fracaso académico
Los alumnos provienen de escuelas privadas particulares.	Los alumnos provienen de escuelas públicas.
Estudian más de 10 y hasta 20 horas semanales.	Estudian más de 20 horas semanales.
No trabajan o trabajan 36 o más horas semanales.	Trabajan hasta 20 horas semanales.
Sus madres trabajan de 21 a 35 horas semanales.	Sus madres trabajan hasta 20 horas semanales o no trabajan.
Sus padres trabajan de 21 a 35 horas semanales.	Sus padres trabajan hasta 20 horas semanales.
Otorgan más importancia al estudio que a la familia.	Otorgan más importancia al estudio que al trabajo y que a la diversión.
Disponen de obra social propia.	No poseen obra social o poseen obra social de los padres.
Trabajan en temas relacionados con la carrera que estudian.	Trabajan en temas parcialmente relacionados o no relacionados con el estudio que realizan.
El mayor porcentaje de éxito académico corresponde al grupo femenino.	Los porcentajes más altos de alumnos libres corresponden al grupo masculino.
Residen con familiares.	Residen en forma independiente o en alguna situación no prevista de residencia (no residen con familiares, ni en residencia universitaria, ni de manera independiente).
Sus madres tienen estudios universitarios de posgrado.	Sus madres no hicieron estudios formales, tienen estudios universitarios incompletos o estudios secundarios incompletos.
Sus padres tienen estudios universitarios de posgrado.	Sus padres no hicieron estudios formales, tienen primaria incompleta o estudios superiores no universitarios incompletos.
Consideran que el dominio de las TIC será imprescindible para el ejercicio profesional.	Consideran que las TIC facilitan el proceso de enseñanza.

Tabla 1. Principales características mayoritarias según perfil de rendimiento académico.

Con la determinación de estos perfiles se pueden tomar medidas correctivas sobre los alumnos que presentan un perfil de fracaso académico, y de esta forma lograr bajar los índices de deserción en la asignatura (y en la universidad).

Claramente esta herramienta presentada aquí, permitirá disminuir la deserción temprana de los alumnos de la materia de Algoritmos y Estructura de Datos de la Universidad Tecnológica Nacional.



Reconocimientos

Este trabajo se ha realizado en el marco del proyecto de investigación acreditado “Determinación de perfiles de estudiantes y de rendimiento académico mediante la utilización de minería de datos”, código 25/L059– UTI1719, continuado en el proyecto “Diseño de un modelo predictivo de rendimiento académico mediante la utilización de minería de datos”, código UTI3808TC, ambos acreditados por la Secretaría de Ciencia, Tecnología y Postgrado de la Universidad Tecnológica Nacional.

Cómo citar este artículo / How to cite this paper

La Red, D. L.; Karanik, M.; Giovannini, M.; Báez, M. E.; Torre, J. (2016). Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos. *Campus Virtuales*, 5(2), 70-83. (www.revistacampusvirtuales.es)

Referencias

- Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Quaderns d'Estadística i Investigació Operativa*, 25(3), 479-498.
- Baker, R. (2010). Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education* (3rd edition), 7, 112-118. Oxford, UK: Elsevier.
- Baker, R.; Yaceff, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1).
- Dell Software. (2016). *Data Mining Techniques – Statistics Textbook*. (<http://documents.software.dell.com/statistics/textbook/data-mining-techn>)
- Di Gresia, L. (2007). *Rendimiento Académico Universitario*. Tesis Doctoral. Universidad Nacional de La Plata, Argentina.
- Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. (1996). From data mining to knowledge discovery: An overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Thurusamy, R.: *Advances in Knowledge Discovery and Data Mining*, pp. 1-34. AAAI Press/MIT Press, Cambridge, Massachusetts.
- García, M. M.; San Segundo, M. J. (2001). El Rendimiento Académico en el Primer Curso Universitario. *X Jornadas de la Asociación de Economía de la Educación*. Libro de Actas, 435-445. España.
- García-Jiménez, M. V.; Alvarado Izquierdo, J. M.; Jiménez Blanco, A. (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema*, 12(2).
- Gil, H. S.; Rao, P. C. (1996). *The Official Client / Server Computing Guide to Data Warehousing*. Que Pub., Simon & Schuster Company.
- Grabmeier, J.; Rudolph, A. (2002). Techniques of Cluster Algorithms in Data Mining. *Data Mining and Knowledge Discovery*, 6, 303-360. Editors: Fayyad; Mannila; Ramakrishnan. Kluwer Academic Publishers. Netherlands.
- Hand, D. J.; Heikki, M.; Padhraic, S. (2001). *Principles of data mining*. MIT Press.
- Jackson, J. (2002). Data Mining: A Conceptual Overview. *Communications of the Association for Information Systems*, 8, 267-296.
- IBM (2005). *DB2 Universal Database Data Warehouse Edition*. (ftp://public.dhe.ibm.com/ps/products/dwe/V8.2.1/es_ES/getstart821.pdf)
- IBM Knowledge Center (2016). Demographic Clustering. (https://www.ibm.com/support/knowledgecenter/SSEPGG_10.5.0/com.ibm.im.model.doc/c_distribution_based_clustering.html)
- Letham, B.; Rudin, C.; McCormick, T. H.; Madigan, D. (2015). Interpretable Classifiers Using Rules and Bayesian Analysis: Building A Better Stroke Prediction Model. *The Annals of Applied Statistics*. Institute of Mathematical Statistics, 9(3), 1350-1371. USA.
- Maindonald, J. (2016). *Data Mining from a Statistical Perspective*. Statistical Consulting Unit of the Graduate School Australian National University. (<http://maths-people.anu.edu.au/~johnm/dm/dmpaper.html>)
- Maimon, O.; Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. Springer, New York. ISBN 978-0-387-09823-4.
- Maletic, J. I.; Collard, M. L.; Marcus, A. (2002). Source Code Files as Structured Documents, in *Proceedings 10th IEEE International Workshop on Program Comprehension (IWPC'02)*, pp. 289-292, Paris, France, June 27-29.
- Michaud, P. (1999). Clustering Techniques. *Future Generation Computer Systems*, 13(2).
- Microsoft (2016). Conceptos de minería de datos. ([https://msdn.microsoft.com/es-ES/library/ms174949\(v=sql.120\).aspx](https://msdn.microsoft.com/es-ES/library/ms174949(v=sql.120).aspx))
- Molina López, J.; García Herrero, J. (2012). *Técnicas de Análisis de Datos*. Madrid, España.
- Natek, S.; Zwillig, M. (2013). Data Mining For Small Student Data Set – Knowledge Management System For Higher Education Teachers. *Make Learn. Active Citizenship by Management, Knowledge Management & Innovation Knowledge and Learning*. International Conference 2013, 19-21 June, Zadar, Croatia.
- Pang-Ning, T.; Steinbach, M.; Kumar, V. (2006). *Introduction to data mining*. Library of Congress.
- Quinlan, J. R. (1986). *Induction of Decision Trees*. *Machine Learning* 1, 81-106, Kluwer Academic Publishers.
- Rokach, L.; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub. Co. Inc. ISBN 978-

La Red, D. L.; Karanik, M.; Giovannini, M.; Báez, M. E.; Torre, J. (2016). Descubrimiento de perfiles de rendimiento estudiantil: un modelo de integración de datos académicos y socioeconómicos. *Campus Virtuales*, 5(2), 70-83.

9812771711.

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*, 40(6), 601-618.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley.

Vivo Molina, J. M.; Franco Nicolás, M.; Sánchez de la Vega, M. del M. (2004). Estudio del rendimiento académico universitario basado en curvas ROC. *Revista de Investigación Educativa, RIE*, 22(2), 327-340. España.

Witten, I. H.; Eibe, F. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, X.; Kumar, V.; Ross Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Yu, P. S.; Zhou, Z. H.; Steinbach, M.; Hand, D. J.; Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge Information Systems*, 14, 1-37, Springer-Verlag London Limited.