Academic Performance of University Students and its Relation with Employment

Laura Lanzarini
Instituto de Investigación en Informática LIDI
Facultad de Informática, Universidad Nacional de La Plata
La Plata, Buenos Aires, Argentina
laural@lidi.info.unlp.edu.ar

María Emilia Charnelli, Javier Díaz
Laboratorio de Investigación en Nuevas Tecnologías Informáticas
Facultad de Informática, Universidad Nacional de La Plata
La Plata, Buenos Aires, Argentina
{mcharnelli, jdiaz}@linti.unlp.edu.ar

***Abstract*—Educational Data Mining collects the various methods that allow extracting novelty and useful information from large data volumes in educational contexts. This paper describes the process used to, through Data Mining techniques, identify the most relevant characteristics in relation to student academic performance at the School of Computer Science of the National University of La Plata. The results obtained using the proposed method to process the information relating to regular and non-regular students at the UNLP allowed establishing interesting relationships in relation to student academic performance. Based on the obtained models it can be said that the fact that the student works does not mean that their academic performance decrease and young students that take several years to join the faculty have better performance if they express interest in getting a job.**

***Keywords*—Academic Performance, Educational Data Mining, Feature Selection, Data Visualization.**

I. Introducción

En la actualidad, la mayoría de los procesos, ya sean industriales, académicos, de negocios o de servicios, cuentan con información histórica almacenada. El avance de la tecnología ha permitido generar volúmenes de datos cada vez más grandes y difíciles de comprender y analizar. Distintas áreas han tratado de dar soluciones a este problema. La Minería de Datos reúne un conjunto de técnicas capaces de modelizar y resumir la información, facilitando su comprensión y ayudando a la toma de decisiones en situaciones futuras [1], [2].

El área educativa no escapa a esta realidad. Por lo general, los establecimientos disponen de información sumamente detallada de cada alumno pero carecen de modelos que les permitan describir de manera objetiva a sus estudiantes. Caracterizar a los estudiantes de una institución académica aporta información no trivial y de utilidad para la toma de decisiones, como por ejemplo, establecer políticas tendientes a mejorar el desempeño académico de los alumnos lo cual redundará en la reducción de la deserción universitaria.

El objeto de estudio presentado en este artículo es la Facultad de Informática de la UNLP. Dicha institución fue creada en 1999, aunque sus carreras de grado comenzaron en el año 1966 dentro de la Facultad de Ciencias Exactas. Actualmente consta con 3 carreras de grado, Licenciatura en Sistemas, Licenciatura en Informática y Analista Programador Universitario, y en conjunto con la Facultad de Ingeniería, dictan la carrera Ingeniería en Computación. Con un promedio anual de aproximadamente 800 ingresantes.

El presente trabajo se enmarca en lo que se conoce como proceso de Extracción de Conocimiento o KDD (Knowledge Discovery in Databases) a partir de la información disponible. El proceso de KDD tiene como objetivo la detección automática de patrones sin necesidad de contar con una hipótesis especificada a priori. Sin embargo, su aplicación requiere identificar, en base al problema a resolver, cuál es la información sobre la que se va a trabajar y cuál es el tipo de modelo que se desea obtener.

En referencia a la información sobre la que se va a trabajar, este artículo propone una metodología de trabajo que comienza con la identificación de los atributos que mejor caracterizan el avance académico de un alumno con el objetivo de reducir así la dimensión de la información a considerar. Esto permite enfocar el análisis en las características adecuadas y arribar a un perfil de alumno de fácil interpretación.

La UNLP utiliza el sistema SIU-Guaraní para la gestión académica de sus alumnos. Este sistema almacena los datos en una base de datos relacional. La información recolectada involucra a 5268 alumnos de la Facultad de Informática comprendido entre los años 2002 y 2012. Debido a la gran cantidad de datos que componen el dominio a trabajar se dificulta la identificación de patrones o relaciones existentes entre las opiniones de distintos sujetos. Por esto último, resulta de interés recurrir a técnicas objetivas que permitan identificar las características más relevantes. Sin embargo, antes de aplicar técnicas específicas de Minería de Datos y Visualización, es preciso verificar y preparar la información a fin de evitar inconsistencias.

Este trabajo está organizado de la siguiente forma: la sección 2 describe trabajos de otros autores relacionados con esta temática, la sección 3 describe el preprocesamiento efectuado sobre los datos originales, la sección 4 muestra la selección de atributos relevantes a través de un método wrapper e incluye una visualización que permite apreciar su incidencia en la condición de regularidad del alumno; la sección 5 muestra la construcción un árbol y un conjunto de reglas utilizando sólo los atributos seleccionados a partir de los cuales se detecta la relación entre el avance académico y la situación laboral del alumno. Finalmente en la sección 6 se presentan las conclusiones de este trabajo.

978-1-4673-9143-6/15/$31.00

II. Trabajos relacionados

Actualmente la problemática de deserción en las carreras de Informática forma parte de una situación a la cual se enfrentan tanto autoridades como docentes. Existen distintas herramientas, desde becas, programas de tutorías y seguimientos por parte de los gabinetes pedagógicos para trabajar con alumnos que abandonan la carrera. Si bien se considera que estas herramientas son útiles e importantes, actúan en instancias en las cuales el alumno ya tomó la decisión de abandonar sus estudios. La comunidad educativa coincide en la necesidad de hacer esfuerzos para revertir esta situación y cualquier tipo de medidas que se adopten deben estar basadas en información útil para la rápida toma de decisiones.

Distintos autores han propuesto diferentes enfoques relacionados con la captación de estudiantes como en el análisis y detección de abandonos y también con la estimación de la duración de la carrera. La Red et al. [3], aplican técnicas de clustering según diferentes criterios como la situación académica y la situación laboral de alumnos de la FACENA de la UNNE en la asignatura Sistemas Operativos con el fin de de instrumentar medidas de apoyo especiales a los alumnos con perfil de alto riesgo de fracaso. Otros aplican técnicas de árboles de decisión C4.5 y el algoritmo de k vecinos más próximos. Valero et al. [4] [5], predicen la deserción escolar en la Universidad Técnica de Izúcar de Matamoros, mientras que Rodallegas et al. [6], construyen un modelo que ayuda a predecir, desde que los alumnos ingresan a la Universidad, las causas que los llevarán a reprobar, así como las materias con mayor riesgo en la Universidad Popular del Estado de Puebla. Formia et al. [7] utilizan técnicas de Minería de Datos para caracterizar la deserción universitaria en la UNRN. Este trabajo presenta la aplicación de un método de selección de características basado en proyecciones SOAP para obtener los atributos más relevantes relacionados con la situación académica, personal y laboral.

En particular, en lo que se refiere al tema central de este artículo, en [7] se analiza la situación laboral del alumno tanto en lo que se refiere al trabajo actual como a sus intenciones de trabajar en el futuro. Como resultado, los métodos de selección de atributos utilizados han coincidido en considerar la situación laboral como información relevante para construir el modelo.

En lo que se refiere al rendimiento académico, hay autores que han estudiado cómo evoluciona el progreso de los alumnos durante sus estudios. En [8], [9] se han utilizado técnicas de clustering específicamente aplicadas al rendimiento de los alumnos en cada uno de los años. Mientras que en [10] se aplicaron reglas de asociación sobre las calificaciones obtenidas para analizar la relación intrínseca entre los diversos cursos analizados.

III. Preparación de los datos

Las primeras etapas del proceso de KDD involucran la comprensión del dominio y la recopilación de los datos. Los datos de los alumnos de la Facultad de Informática fueron recolectados del sistema SIU-Guaraní. Generalmente en la mayoría de las unidades académicas de la UNLP, la información personal de los alumnos es cargada por los mismos a la hora de inscribirse en una carrera universitaria. El cuestionario del sistema contiene información tanto sobre datos personales como laborales y se organiza como se observa en la Tabla I.

TABLE I: Estructura de la encuesta en SIU-Guaraní

1	Datos Personales (estado civil, familiares a cargo, con quien vive, etc.).
2	Financiamiento de estudios (familia, beca, trabajo).
3	Situación laboral (si busca trabajo, cuántas horas trabaja, relación con la carrera.
4	Situación padres (si viven, nivel de estudios y su actividad profesional).
5	Otros estudios.
6	Tecnología (si dispone de PC, acceso a Internet, etc.)
7	Nivel de idiomas.

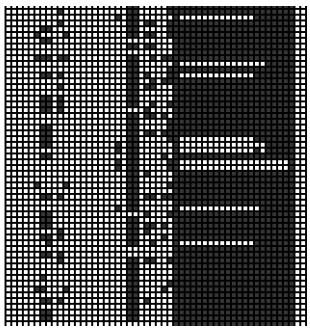

Fig. 1: Patrón de las preguntas no contestadas

Completadas las primeras etapas del KDD, se continúa con la etapa de preparación de los datos. Es necesario seleccionar y preparar el subconjunto de datos a utilizar. Esta fase cubre todas las actividades para construir el conjunto final de los datos que serán utilizados por las técnicas de modelado.

A. Atributos con datos inconsistentes

En la figura 1 se muestra una visualización que representa un diagrama de dispersión con las respuestas de los alumnos. Cada fila de la matriz representa a un alumno y cada columna una pregunta del cuestionario. El color oscuro significa que un alumno no respondió la pregunta. Se observa el patrón que siguen las preguntas no contestadas del SIU-Guaraní en los últimos años. La mayoría de las respuestas nulas se corresponden con las últimas preguntas del cuestionario web, algunas de las cuáles se tratan de otros estudios realizados, nivel de idiomas, uso de la PC e Internet, etc. En las primeras preguntas, se destacan las respuestas nulas de cantidad de familiares a cargo y relación del trabajo con la carrera, ya que más del 50 por ciento respondió que no trabaja y que vive con la familia.

Los atributos con más de un 80 por ciento de nulos fueron descartados para el análisis

B. Atributos con datos no generalizables

Se eliminaron atributos no generalizables como el nombre del estudiante, el número de documento, el número de legajo y el número de inscripción.

Se redujo la cardinalidad de algunos atributos utilizando categorías más genéricas incrementando así su capacidad pre-

dictiva. Por ejemplo, en los datos originales se registra colegio secundario con 453 valores diferentes, que representan los nombres de los colegios en los que los alumnos cursaron el nivel medio. Por lo que el nombre de la escuela fue reemplazado por dos atributos, uno que indica si la institución es pública o privada, y otro que indica si la institución es técnica o no.

Algo similar se hizo con el lugar de procedencia del alumno, reemplazándolo con el atributo que indica si es del interior o no, para este caso se consideró si las localidades estaban a más de 60 km de la ciudad de La Plata, lugar donde se encuentra dicha unidad académica.

Se redujo la cardinalidad de la actividad laboral de los padres con los valores más representativos en cada caso. En la Facultad de Informática-UNLP ser empleado representa la mayoría de los valores que toma el atributo actividad laboral tanto de la madre como del padre. Se redujo la cardinalidad a si tiene relación de dependencia o no, ya que tener personas a cargo, ser ama de casa o bien ser independiente se pueden unificar como actividades que no son dependientes.

C. Transformaciones realizadas

La creación de características consiste en generar nuevos atributos con el objetivo de mejorar la calidad y comprensión del conocimiento extraído. En esta dirección, se transformó la fecha de nacimiento por la edad de los alumnos. Por otro lado, a partir del año de egreso del secundario y del año de ingreso a la facultad, se generó un nuevo atributo que calcula esta diferencia.

Otras transformaciones que se realizaron tienen que ver con si trabaja y busca trabajo, cómo costea sus estudios, con quién vive. Se realizó la numerización de algunos atributos y la posterior normalización de su rango, de acuerdo a los requerimientos de las técnicas de Minería de Datos a utilizar. Se numerizó el máximo nivel de estudios de los padres y la cantidad de horas semanales que trabaja el alumno.

Además de los datos censales, se dispone de toda la información académica de los estudiantes.

Con el objetivo de analizar el avance de los alumnos en sus estudios y por cuestiones de simplicidad sólo se trabajó con la cantidad de finales aprobados por alumno al finalizar cada año durante los primeros 5 años de su vida universitaria. Se consideró que esta cantidad de años resulta representativa y coincide con la duración de las carreras de la Facultad. Los valores de estos atributos se obtienen al calcular para cada alumno la proporción de finales aprobados desde el inicio de su carrera hasta el final de cada año lectivo en relación a la cantidad total de materias según cada carrera como se observa en la siguiente ecuación

$$avance_i = \frac{f_i}{F} \qquad i = 1..5 \tag{1}$$

donde f_i es la cantidad total de materias que el alumno registra como aprobadas al finalizar el i-ésimo año, F es la cantidad total de materias de la carrera y $avance_i$ es un valor entre 0 y 1 que representa el avance que el alumno tiene en su carrera al finalizar el i-ésimo año.

Con el objetivo de analizar el avance de los alumnos en cada uno de los primeros 5 años de su vida universitaria se utilizaron los atributos creados según la ecuación (1) para formar tres grupos: ALTO, MEDIO y BAJO desempeño académico. Esto permitió agregar un nuevo atributo que se denominó Ritmo y que permite indicar a través de un único valor si el alumno registra un avance adecuado, regular o lento en sus estudios.

Por último, se creó un campo que resume el estado académico del alumno, cuyo valor indica si se trata de un alumno regular o no, teniendo en cuenta las condiciones de regularidad establecidas en la Facultad de Informática indicada a continuación:

> *"Establecer la condición de regularidad para todos los alumnos de la Facultad (incluyendo ingresantes) con la Aprobación de 1 Examen Final o de 1 Cursada de Trabajos Prácticos (1 actividad positiva) durante el transcurso de los últimos 3 ciclos lectivos (1 de Marzo a 28 de febrero)".*

IV. Selección de Características

Las técnicas de Minería de Datos aplicadas sobre información estructurada compuesta por un número importante de características dan como resultado modelos complejos. Dependiendo de la técnica utilizada, datos con una dimensión alta producen o bien árboles enormes o conjuntos de reglas con alta cardinalidad y antecedentes formados por un número importante de conjunciones [11] o funciones discriminantes difíciles de interpretar.

Para resolver este problema es preciso analizar, en forma previa a la construcción del modelo, cuáles son los atributos más representativos de la información disponible. Una vez seleccionados los atributos más relevantes, la técnica a utilizar verá simplificada su tarea y ofrecerá como resultado un modelo más sencillo y fácil de interpretar [12].

En el caso particular del problema a resolver en este artículo, la selección de características juega un rol fundamental ya que se espera poder identificar los atributos adecuados que permitan construir un modelo del avance académico de los alumnos por tratarse de una métrica estrechamente relacionada con la condición de regularidad.

Las dos técnicas principales para la selección de características son: los métodos de filtros y los métodos wrappers [13]. Los métodos de filtro se basan en características generales del conjunto de entrenamiento para seleccionar algunas características sin utilizar ningún algoritmo de aprendizaje. Los métodos wrappers requieren un algoritmo de aprendizaje predefinido para realizar la selección, y utiliza su rendimiento para evaluar y determinar cuáles características serán seleccionadas. [14].

Para establecer un ranking entre los atributos se utilizó el método Chi2 propuesto por Liu et al en [15]. Este método es uno de los más utilizado a la hora de seleccionar atributos y se basa en un método estadístico para comparar proporciones. Para medir el desempeño de los atributos se utiliza la medida χ^2 que permite determinar un valor proporcional a la relación que existe entre una clase c y una característica f que puede tomar r valores posibles.

TABLE II: Atributos seleccionados

Busca trabajo
Edad
Tiempo en ingresar a la facultad
Ritmo

Dado un conjunto de datos D de n ejemplos, la medida χ^2 se calcula mediante la siguiente fórmula:

$$\chi^2(D,c,f) = \sum_{i=1}^{r} \frac{(n_{i_{pos}} - \mu_{i_{pos}})^2}{\mu_{i_{pos}}} + \frac{(n_{i_{neg}} - \mu_{i_{neg}})^2}{\mu_{i_{neg}}} \quad (2)$$

donde $n_{i_{pos}}$ y $n_{i_{neg}}$ representan la cantidad de ejemplos positivos y negativos para el valor i de la característica f, respectivamente, y $\mu_{i_{pos}}$ y $\mu_{i_{neg}}$ son los valores esperados si los datos estuvieran uniformemente distribuidos.

El puntaje obtenido al evaluar $\chi^2(D,c,f)$ sigue la distribución χ^2 y el objetivo del algoritmo de selección es simplemente elegir un subconjunto de características entre las que posean los puntajes más elevados dado que estas serán las más relevantes al momento de tener que realizar una discriminación entre las clases.

Para determinar cuantos atributos seleccionar se evaluó la performance de distintos clasificadores a medida que se van incorporando una a una las características ordenadas por puntaje en forma decreciente. Es decir que, para cada clasificador se mide su desempeño utilizando en su construcción sólo la primera característica; luego se repite este mismo proceso para las dos características con mayor puntaje y se continúa de la misma forma, incorporando las características de a una y evaluando la performance alcanzada hasta que no se produzcan cambios durante un cierto número de iteraciones [16].

Para el criterio de performance del clasificador se tuvo en cuenta la cantidad de aciertos obtenidos al momento de predecir si un alumno ha dejado de ser regular, es decir la performance de los verdaderos negativos (el atributo construido indica si es regular). Esto se debe al desbalance entre las clases ya que el 71% de los alumnos cumplen con la condición de regularidad establecida por la Facultad; por lo tanto, la precisión del clasificador en general, teniendo en cuenta aciertos en ambas clases, puede ser buena aunque no tenga un buen desempeño sobre la clase de interés: los alumnos NO regulares.

V. Resultados

La selección de atributos se realizó utilizando un método wrapper cuyo clasificador se basó en una Máquina de Vectores de Soporte [17]. En ambos casos los atributos seleccionados fueron los indicados en la tabla II y constituyen el 8% de la cantidad total de atributos.

El haber utilizado una Máquina de Vectores de Soporte para seleccionar los atributos no permite contar con un modelo descriptivo de la información disponible. Por tal motivo se utilizaron los atributos seleccionados para construir tres tipos de modelos diferentes que permiten clasificar a los alumnos en Regulares y No Regulares utilizando los siguientes métodos: C4.5, PART y un multiperceptrón entrenado con el algoritmo de backpropagation [18], [19], [20]. Los dos primeros dan como resultado un modelo descriptivo mientras que el tercero permite verificar la precisión obtenida.

TABLE III: Resultados obtenidos utilizando todos los atributos

Método	Tasa de Acierto	Precision(pos) Regulares	Precision(neg) No Regulares
C4.5	81.4%	88.9%	63.1%
PART	81.1%	88.4%	63.5%
BPN	81.2%	87.9%	64.3%

TABLE IV: Resultados obtenidos utilizando sólo los atributos seleccionados

Método	Tasa de Acierto	Precision(pos) Regulares	Precision(neg) No Regulares
C4.5	79.46%	85.47%	64.82%
PART	79.46%	85.47%	64.82%
BPN	79.50%	85.70%	64.67%

Para medir el desempeño de cada modelo se utilizó la tasa de acierto y la precisión de cada clase los cuales se calculan de la siguiente forma

$$tasa\_de\_aciertos = \frac{t\_pos + t\_neg}{pos + neg} \quad (3)$$

$$precision(pos) = \frac{t\_pos}{t\_pos + f\_pos} \quad (4)$$

$$precision(neg) = \frac{t\_neg}{t\_neg + f\_neg} \quad (5)$$

Donde

- t_pos y t_neg corresponden a la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares) correctamente clasificados por el método respectivamente.
- f_pos y f_neg representan la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares) incorrectamente clasificados por el método respectivamente.
- pos y neg son la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares) reales del problema (las respuestas esperadas).

Las tablas III y IV resumen los resultados obtenidos.

En la tabla III la tasa de acierto de los métodos C4.5 y PART es de 81.4% y 81.1% respectivamente mientras que al utilizar los atributos seleccionados esta tasa baja a 79.46% en ambos casos tal como se observa en la tabla IV. Esto, como ya se mencionó anteriormente, se debe al desbalance de clases. Nótese que la diferencia reside en la precisión con la que se predice si un alumno es regular (columnas Precision(pos)

```
tiempo_ingresar_facultad≤2.5 años: SI (2798/419)
tiempo_ingresar_facultad>2.5 años
| edad≤26 años
| | busca_trabajo= SI: SI (25/2)
| | busca_trabajo= NO: NO (1532/539)
| edad>26 años: SI (913/122)
```

Fig. 2: Arbol obtenido utilizando el método C4.5 aplicado a los atributos seleccionados

```
tiempo_ingresar_facultad≤2.5 años: SI(2798/419)

edad ≤ 26 años AND
busca_trabajo = NO: NO(1532/539)

:SI (938/124)
```

Fig. 3: Reglas obtenidas con el método PART aplicado a los atributos seleccionados

de ambas tablas). Sin embargo, si se observa la precisión de la clase de interés (columnas Precision(neg)) los valores obtenidos utilizando los atributos seleccionados resultan ligeramente superiores a los que se obtienen trabajando con todos los atributos.

Las figuras 2 y 3 muestran la simplicidad de los modelos obtenidos. Allí se observa que la decisión por parte de los alumnos de buscar trabajo tiene una estrecha relación con su condición de ser alumno regular. Según los modelos obtenidos, cuando se trata de alumnos que demoran en ingresar a la Facultad más de 2,5 años y son menores a 26 años si no manifiestan tener intenciones de buscar trabajo tienen un alto riesgo de no ser alumnos regulares.

Es importante remarcar que entre los atributos originales existe información relacionada a la situación laboral que no ha sido seleccionada por el método wrapper. En particular se conoce si trabaja o no y en caso de que lo haga, también se encuentra registrada la cantidad de horas que le dedica al trabajo y si tiene relación con la carrera o no. Es decir que la simplicidad del modelo obtenido también permite afirmar que la situación laboral del alumno no incide en el avance de sus estudios. Esto coincide con lo afirmado en [21] donde los autores afirman que no existe evidencia de que el tiempo de estudio y la cantidad de horas de trabajo incidan en el rendimiento académico de los alumnos universitarios.

VI. Conclusiones

En este artículo se ha desarrollado un caso de estudio que muestra cómo detectar características de un problema en un dominio específico en forma clara y precisa. En este caso particular se pudieron obtener, a partir de la información de los alumnos de la Facultad de Informática de la UNLP, los atributos más representativos para la construcción de un modelo de clasificación que permite describir y caracterizar a los alumnos según su condición de regularidad.

De los modelos obtenidos se puede afirmar que los atributos seleccionados son adecuados para predecir la condición de no regularidad de un estudiante. También dejan de manifiesto la no incidencia de la situación laboral actual de los alumnos en lo que se refiere a su rendimiento académico.

El alumnado de la Facultad de Informática de la UNLP es muy joven. Si bien el rango de edades se encuentra entre 17 y 64 años, la edad promedio es de 21,383 años con una desviación estandar de 4,486 años. Es por esto que la relación, obtenida a partir de los modelos, entre el tiempo que tarda un alumno en ingresar a la Facultad, su edad inferior a 26 años y su deseo de conseguir trabajo resulta relevante. Es preciso continuar analizando a los alumnos que se encuentran ubicados en este segmento a fin de determinar si su necesidad de buscar trabajo se relaciona con su situación económica actual o con una decisión personal de dedicar al estudio únicamente una parte del día.

Actualmente se está trabajando con la información de los alumnos de la UTN Regional La Plata con el objetivo de establecer similitudes y diferencias entre las poblaciones de alumnos. Los resultados de esta comparación permitirán medir con mayor precisión la incidencia los atributos referidos a la edad y la situación laboral del alumno ya que la UTN se caracteriza por tener alumnos que trabajan y por lo general poseen una edad algo superior a los de la UNLP.

References

[1] E. Charnelli, L. Lanzarini, G. Baldino, and J. Díaz, "Determining the profiles of young people from buenos aires with a tendency to pursue computer science studies," in *Proceedings del XX Congreso Argentino de Ciencias de la Computación CACIC*. Red UNCI, 2014.

[2] P. Cabena, *Discovering data mining: from concept to implementation*, ser. An IBM Press Book Series. Prentice Hall PTR, 1998.

[3] D. L. La Red Martínez, J. C. Acosta, L. A. Cutro, V. E. Uribe, and A. R. Rambo, "Data warehouse y data mining aplicados al estudio del rendimiento académico y de perfiles de alumnos," in *Proceedings XII Workshop de Investigadores en Ciencias de la Computación - WICC*. Red UNCI, 2010, pp. 162–166.

[4] S. Valero and A. Salvador, "Predicción de la deserción escolar usando técnicas de minería de datos," in *Simposio Internacional en Sistemas Telemáticos y Organizaciones Inteligentes SITOI*, 2009, pp. 332–340.

[5] S. Valero, A. Salvador, and M. GarcÃa, "MinerÃa de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos," in *II Conferencia Conjunta Iberoamericana sobre Tecnologías para el aprendizaje CcITA*, 2010.

[6] E. Rodallegas, A. Torres, B. Gaona, E. Gastelloú, R. Lezama, and S. Valero, "Modelo predictivo para la determinación de causas de reprobación mediante minería de datos," in *Proceedings de la II Conferencia Conjunta Iberoamericana sobre Tecnologías para el aprendizaje CcITA*, 2010.

[7] S. Formia, "La deserción en cursos universitarios. construcción de modelos sobre datos de la universidad nacional de río negro," Master's thesis, Tesis de Magister en Tecnología Informática aplicada en Educación, RedUNCI, 3 2014.

[8] R. Asif, A. Merceron, and M. K. Pathan, "Investigating performance of students: A longitudinal study," in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, ser. LAK '15. ACM, 2015, pp. 108–112.

[9] A. J. Bower, "Analyzing the longitudinal k-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis," *Practical Assessment, Research and Evaluation*, vol. 5, no. 7, 2010.

[10] J. Wang, Z. Lu, W. Wu, and Y. Li, "The application of data mining technology based on teaching information," in *In Computer Science Education ICCSE*, 2012.

[11] M. Sebban, R. Nock, J. H. Chauchat, and R. Rakotomalala, "Impact of learning set quality and size on decision tree performances," *Int. Journal of Computers, Systems and Signals*, vol. 1, pp. 85–105, 2000.

[12] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. D. Jong, S. Dzeroski, S. E. Fahlman, D. Fisher, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R. Michalski, T. Mitchell, P. Pachowicz, Y. Reich, H. Vafaie, W. V. D. Welde, W. Wenzel, J. Wnek, and J. Zhang, "The monk's problems a performance comparison of different learning algorithms," Tech. Rep., 1991.

[13] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML 2001. Williamstown, MA, USA: Williams College, 2001.

[14] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 1, pp. 1–14, Jan 2013.

[15] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, ser. TAI '95. Washington, DC, USA: IEEE Computer Society, 1995, pp. 88–. [Online]. Available: http://dl.acm.org/citation.cfm?id=832245.832359

[16] M. P. O'Mahony, P. Cunningham, and B. Smyth, "An assessment of machine learning techniques for review recommendation," in *Artificial Intelligence and Cognitive Science*. Springer, 2010, pp. 241–250.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[18] J. Quinlan, *C4.5: Programs for Machine Learning*, ser. Morgan Kaufmann series in machine learning. Morgan Kaufmann, 1993.

[19] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[20] F. Rosenblatt, *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*, ser. Report (Cornell Aeronautical Laboratory). Spartan Books, 1962.

[21] S. Nonis and G. Hudson, "Academic performance of college students: Influence of time spent studying and working," *Journal of Education for Business*, vol. 81, no. 3, pp. 151–159, 2006.