Pattern Discovery for Academic Performance on Critical Reading Competition

Ricardo Timarán Pereira, Ph.D.[1], Arsenio Hidalgo Troya, Mg.[1], Javier Caicedo Zambrano, Ph.D.[1]
Isabel Hernández Arteaga[2], Juan Carlos Alvarado[2]
[1]Universidad de Nariño, Colombia, ritimar@udenar.edu.co, arsenio.hidalgo@udenar.edu.co, jacaza@udenar.edu.co
[2]Universidad Cooperativa de Colombia sede Pasto, Colombia, isabel.hernadez@campusucc.edu.co, juan.alvarado@campusucc.edu.co

***Abstract*– The results of generic skill in Critical Reading of the Research Project that aims was to detect patterns of academic performance in generic skills of Colombian students of professional programs who took the Saber Pro test are presented. The socio-demographic, economic, academic and institutional information of these students was selected from the ICFES databases. A data repository was built, cleaned and transformed and from it, patterns associated with good or poor academic performance of students in this skill, were discovered using a classification model based on decision trees and the CRISP-DM methodology. Among the discovered patterns are highlighted, the institutional accreditation and the traditional learning, as two important factors associated with good academic performance. The discovered knowledge will be added to existing and it may be used in the decision making process of ICFES and government and academic institutions responsible for the quality of higher education.**

***Keywords*— Patters Discovery, Academic Performance, Critical Reading, Datamining.**

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2015.1.1.144
ISBN: 13 978-0-9822896-8-6
ISSN: 2414-6668

Descubrimiento de Patrones de Desempeño Académico en la Competencia de Lectura Crítica

Ricardo Timarán Pereira, Ph.D.[1], Arsenio Hidalgo Troya, Mg.[1], Javier Caicedo Zambrano, Ph.D.[1]
Isabel Hernández Arteaga[2], Juan Carlos Alvarado[2]
[1]Universidad de Nariño, Colombia, {ritimar,arsenio.hidalgo,jacaza}@udenar.edu.co
[2]Universidad Cooperativa de Colombia sede Pasto, Colombia,{isabel.hernadez, juan.alvarado}@campusucc.edu.co

***Abstract*– The results of generic skill in Critical Reading of the Research Project that aims was to detect patterns of academic performance in generic skills of Colombian students of professional programs who took the Saber Pro test are presented. The socio-demographic, economic, academic and institutional information of these students was selected from the ICFES databases. A data repository was built, cleaned and transformed and from it, patterns associated with good or poor academic performance of students in this skill, were discovered using a classification model based on decision trees and the CRISP-DM methodology. Among the discovered patterns are highlighted, the institutional accreditation and the traditional learning, as two important factors associated with good academic performance. The discovered knowledge will be added to existing and it may be used in the decision making process of ICFES and government and academic institutions responsible for the quality of higher education.superior.**

***Keywords*— Patters Discovery, Academic Performance, Critical Reading, Datamining.**

***Resumen*– En este artículo se presentan los resultados obtenidos en la competencia genérica de Lectura Crítica del proyecto de investigación cuyo objetivo fue detectar patrones de desempeño académico en las competencias genéricas de los estudiantes colombianos de programas profesionales que presentaron las pruebas de estado Saber Pro. Se seleccionó, de las bases de datos del ICFES, la información sociodemográfica, económica, académica e institucional de estos estudiantes. Se construyó, limpió y transformó un repositorio de datos y a partir de él, se descubrieron patrones asociados al buen o mal desempeño académico de los estudiantes en esta competencia, utilizando un modelo de clasificación basado en árboles de decisión y la metodología CRISP-DM. Entre los patrones descubiertos, se destacan la acreditación institucional y la modalidad de estudio presencial, como dos factores importantes asociados al buen desempeño académico. El conocimiento descubierto se incorporará al existente y podrá ser utilizado en los procesos de toma de decisiones del ICFES y de las instituciones gubernamentales y académicas que velan por la calidad de la educación superior.**

***Palabras Clave*—Descubrimiento de Patrones, Desempeño Académico, Lectura Crítica, Minería de datos.**

I. INTRODUCCIÓN

En Colombia, uno de los objetivos del examen de estado de calidad de la educación superior Saber Pro, es comprobar el grado de desarrollo de competencias de los estudiantes próximos a culminar los programas académicos de pregrado que ofrecen las instituciones de educación superior. El examen está compuesto por pruebas que evalúan competencias genéricas y específicas. De acuerdo a los lineamientos Saber Pro del Instituto Colombiano para la Evaluación de la Educación (ICFES) [1], todos los estudiantes deben presentar los módulos de competencias genéricas sin importar el programa de formación que cursen, que incluye competencias de lectura crítica, razonamiento cuantitativo, escritura e inglés.

En la competencia de lectura crítica se evalúan los desempeños asociados a lectura, pensamiento crítico y entendimiento interpersonal [2]. En la competencia de razonamiento cuantitativo se evalúan los desempeños relacionados con uso de lenguaje cuantitativo y solución de problemas [2]. En escritura se evalúa la competencia para comunicar ideas por escrito referidas a un tema dado [2], [3]. En inglés se evalúa la competencia del estudiante para comunicarse efectivamente en inglés.

A pesar que en la prueba Saber Pro no se pretende que los estudiantes de todas las formaciones desarrollen las competencias genéricas a un mismo nivel, ni aún las comunes a grupos de programas, sí es importante determinar cómo influyen los factores socioeconómicos, académicos e institucionales del estudiante para obtener un determinado nivel de desempeño de estas competencias en las pruebas Saber Pro. Los estudios que se han realizado hasta el momento [1][3][4][5] con respecto al análisis de los resultados de las pruebas Saber Pro se basan en información procesada mediante un análisis estadístico, donde fundamentalmente se consideran variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones, que por lo general están ocultas y que únicamente se pueden descubrir utilizando un tratamiento más complejo de los datos, que es posible con la minería de datos.

La minería de datos en la educación no es un tema nuevo, su estudio y aplicación ha sido muy relevante en los últimos años, se puede utilizar sus técnicas para explicar y/o predecir cualquier fenómeno dentro del campo educativo [6], [7]. Por ejemplo, utilizando las técnicas de minería de datos, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de deserción de cualquier estudiante [8], [9]. Las instituciones de educación pueden usar la minería de datos para hacer análisis comprensivos de las características de sus estudiantes, métodos evaluativos, develando procesos exitosos o por el contrario, detectando fraudes o inconsistencias, [10].

El resto del artículo se organiza de la siguiente manera: en la sección II se explica el proceso de detección de patrones de desempeño académico en Lectura Crítica utilizando la metodología CRISP-DM. En la sección III se evalúan e interpretan los patrones obtenidos. En la sección IV se

discuten los resultados y finalmente en la sección V se presentan las conclusiones y futuros trabajos.

II. METODOLOGÍA

La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Para el descubrimiento de patrones de desempeño académico, se aplicó la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining)*, el cual es uno de los modelos utilizados, principalmente, en los ambientes académico e industrial y la guía de referencia más ampliamente utilizada en el desarrollo de este tipo de proyectos [11].

Esta metodología contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación, que se describen a continuación.

A. Comprensión Del Problema

En esta fase, se realizaron las actividades que permitieron profundizar y apropiar de una manera completa el problema objeto de estudio, los objetivos y los requisitos de esta investigación, que posibilitaron la recolección de los datos correctos para interpretar adecuadamente los resultados. En esta fase, detectar patrones de rendimiento académico en las competencias genéricas de las pruebas Saber Pro, se convirtió en un problema de minería de datos.

B. Comprensión de los Datos

En esta fase, se identificó, recopiló y familiarizó con la información disponible en las bases de datos del ICFES, sobre los resultados de competencias genéricas en las pruebas Saber Pro aplicadas en el año 2011 semestre 2 y sobre los datos socioeconómicos, académicos e institucionales de los estudiantes de programas profesionales que presentaron esta prueba.

A partir de las bases de datos del ICFES, se construyó un repositorio inicial donde se almacenaron los datos de todos los estudiantes de programas técnico profesional, tecnológico y profesionales que presentaron las pruebas Saber Pro. Este repositorio inicial, el cual se denominó T153123A94, cuenta con 153.123 registros y 94 atributos. De este repositorio se seleccionaron únicamente los registros de estudiantes de programas profesionales y sus resultados en la competencia lectura crítica, quedando un repositorio con 97.055 registros y 94 atributos, identificado como T97055A94LEC y el cual sirvió de base para las subsiguientes fases. Además, con base en la conceptualización de desempeño académico y los antecedentes teóricos sobre los factores que intervienen en él, los cuales se asocian y colindan unos con otros, los 94 atributos del conjunto de datos T97055A94LEC se clasificaron en cuatro dimensiones: sociodemográfica, económica, académica e institucional.

C. Preparación de los Datos

En esta fase, los 94 atributos del repositorio base T97055A94LEC, considerados por el ICFES como los más importantes para capturar la información de las pruebas Saber Pro 2011-2, fueron depurados, teniendo en cuenta la calidad de los datos y las técnicas de minería de datos a aplicar, se limpiaron (eliminación de datos nulos y valores constantes) e integraron los datos, se generaron atributos adicionales a partir de los existentes por ganancia de información, y se realizaron transformaciones o cambios de formato a los valores de los atributos que se consideraron necesarios. Como resultado de esta fase se obtuvo un repositorio de datos limpio y transformado, con 32 atributos, listo para aplicarle las técnicas de minería de datos y al cual se le denominó T97055A32LEC. En el anexo I, que se encuentra al final de este artículo, está la descripción de este repositorio.

Con el fin de facilitar la detección de patrones de rendimiento académico se discretizaron los valores numéricos de ciertos atributos teniendo en cuenta un rango de valores o que las frecuencias por cada valor sean proporcionales, para evitar sesgos, al construir los modelos de minería de datos. En la tabla I se muestra la descretización del atributo edad. En la tabla II se presenta la generalización por zonas geográficas y en la tabla III la discretización del número de estudiantes por zona.

TABLA I
DISCRETIZACIÓN DEL ATRIBUTO EDAD

Valor	No. Estudiantes
Edad hasta 21	17785
Edad 22	15602
Edad 23	12133
Edad de 25 a 26	12022
Edad de 27 a 29	10766
Edad de 30 a 34	9915
Edad mayor igual que 35	9863
Edad 24	8900

TABLA II
GENERALIZACIÓN POR ZONAS GEOGRÁFICAS

Zonas	Departamentos
CARIBE	Atlántico, Bolívar, Cesar, Córdoba, Guajira, Magdalena, San Andrés, Providencia y Santa Catalina y Sucre.
CENTRO SUR	Amazonas, Caquetá, Huila, Putumayo y Tolima
CENTRO ORIENTE	Boyacá, Cundinamarca, Norte de Santander y Santander
EJE CAFETERO	Antioquia, Caldas, Risaralda y Quindío.
LLANO	Arauca, Casanare, Guainía, Guaviare, Meta, Vaupés y Vichada.
PACÍFICO	Cauca, chocó, Nariño y Valle del Cauca.
BOGOTÁ	Distrito Capital de Bogotá

D. Modelado

En esta fase se seleccionó la tarea de clasificación con árboles de decisión como la técnica de minería de datos más adecuada para solucionar el problema objeto de la investigación.

Con clasificación se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes de programas profesionales, los factores socioeconómicos, académicos e institucionales asociados a un probable buen o mal desempeño académico en la competencia genérica de

lectura crítica, evaluada en las pruebas SaberPro 2011-2. El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender [12], [13], [14]. La importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación. La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y sólo una hoja, asignando una única clase a la predicción [15].

TABLA III
DISCRETIZACIÓN DE NÚMERO RO DE ESTUDIANTES POR ZONA

Zona	No. Estudiantes	Valor
BOGOTA	33544	ALTO
EJE CAFETERO	16500	MEDIO
CARIBE	15312	MEDIO
CENTRO ORIENTE	13519	MEDIO
PACIFICO	11766	MEDIO
CENTRO SUR	5283	BAJO
LLANO	1144	BAJO

El algoritmo de la herramienta Weka [16] utilizado para obtener el modelo de clasificación con árboles de decisión fue J48, el cual implementa al algoritmo C.45, [17]. El algoritmo J48 se basa en la utilización del criterio ratio de ganancia (*gain ratio*). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido [15]. El parámetro más importante que se tuvo en cuenta para la poda fue el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25% y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños [18]. Otro parámetro utilizado para variar el tamaño del árbol fue a través de M que especifica el mínimo número de instancias o registros por nodo del árbol [16].

Antes de construir un modelo se definió el procedimiento para probar la calidad del modelo y su validez. Teniendo en cuenta que para entrenar y probar un modelo de clasificación, se divide los datos en dos conjuntos: entrenamiento y prueba [16], se utilizó el método de validación cruzada (*Cross validation)* por ser la opción por defecto y la más comúnmente utilizada. Este mecanismo permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición [11]. Para este caso particular se utiliza el método de evaluación validación cruzada con *n* pliegues *(n-fold cross validation*). Este método consiste en dividir el conjunto de entrenamiento en n subconjuntos disjuntos de similar tamaño llamados pliegues (*folds*) de forma aleatoria. El número de subconjuntos se puede introducir en el campo *Folds*. Posteriormente se realizan *n* iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes *n-1* (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones. Otra ventaja de la validación cruzada es que la varianza de los *n* errores de muestra parciales, permite estimar la variabilidad del método de aprendizaje con respecto al conjunto de datos. Comúnmente, se suelen utilizar 10 particiones (10-fold cross validation) [11].

Por otra parte, se evaluó o estimó el coste del clasificador para el repositorio T97055A32LEC a través de la matriz de confusión. La matriz de confusión (*Confusion Matrix*) representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila *i, i =1…n* constituyen el número de instancias que realmente pertenecen a la clase *i.* Similarmente la sumatoria de los ejemplos o registros en cada columna *j, j = 1…n* son las instancias que ha predicho el algoritmo al valor *j* de la clase. Los valores en la diagonal son los aciertos y el resto son los errores de clasificación (ejemplos que pertenecían a la clase *i* de la fila *i* y fueron clasificados incorrectamente en otra) [19].

Teniendo en cuenta los parámetros de evaluación anteriores, se procedió a construir los diferentes árboles de decisión con el algoritmo J48. Se escogió como clase el puntaje en la competencia de lectura crítica, el cual fue discretizado en los valores: "por encima" o "por debajo" de la media.

Con el fin de obtener diferentes modelos de árboles por competencia y reglas de clasificación generalizadas hasta reglas más detalladas, se establecieron cuatro porcentajes de prepoda del árbol para el factor M igual a 10%, 5%, 1% y 0.5% del total de registros del repositorio de datos, manteniendo constante el factor confianza C en el 25%. Se escogió el árbol construido con los parámetros M=975 (1%) y C=25% por los mejores resultados obtenidos y por la facilidad de análisis de los patrones. Una vez construido los árboles se aplicó un proceso de pospoda para dejar las ramas y por ende las reglas más representativas, que son aquellas que sobrepasan un mínimo soporte del 0.5% y una confianza del 60%. . El árbol construido con estos parámetros se muestra en el anexo II al final del artículo. En la figura 1 se muestra la precisión del árbol y su matriz de confusión.

E. Evaluación

En esta fase se evaluaron los patrones descubiertos con el fin de determinar su validez, remover los patrones redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario. La evaluación e interpretación de los patrones descubiertos se describe en la sección de resultados.

F. Implementación

En esta fase, el conocimiento descubierto se incorporará al existente y se podrá integrar a los procesos de toma de decisiones del ICFES y de las instituciones gubernamentales y académicas que velan por la calidad de la educación superior. Una vez estas instituciones intervengan los factores asociados al desempeño académico en la competencia genérica de Lectura Crítica, de los estudiantes de programas profesionales que presentaron las Pruebas SaberPro, será posible analizar los resultados y determinar sus efectos.

III. RESULTADOS

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos T97055A32LEC (ver figura 1), en el cual se almacenan los datos válidos sobre los factores sociodemográficos, económicos, académicos e institucionales de 97.055 estudiantes de programas profesionales que presentaron las pruebas SaberPro 2011-2 en la competencia genérica de lectura crítica , donde se escogió el atributo *mod_lectura_critica_desemp* como clase, se puede observar que este clasifica correctamente a 62.781 instancias , que corresponde a un porcentaje de precisión del 64,7% y 34.274 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 35,3% (ver figura 1).

```
=== 10 Fold Cross Validation ===
=== Summary ===
Correctly Classified Instances          62781               64.686 %
Incorrectly Classified Instances        34274               35.314 %
Kappa statistic                          0.2831
Mean absolute error                      0.442
Root mean squared error                  0.4704
Relative absolute error                 89.1798 %
Root relative squared error             94.4938 %
Coverage of cases (0.95 level)          99.999 %
Mean rel. region size (0.95 level)      99.9995 %
Total Number of Instances               97055

=== Confusion Matrix ===
a      b <-- classified as
37498   15548 | a = Y
18726   25283 | b = N
```

Fig. 1. Precisión y Matriz de Confusión

Teniendo en cuenta la distribución de los valores del atributo clase *mod_lectura_critica_desemp* del repositorio T97055A32LEC que es de 53.046 registros para el valor "sobre la media" y 44.009 registros para el valor "bajo la media" y evaluando el modelo con la matriz de confusión de la figura 2, este clasifica correctamente a 37. 498 casos de estudiantes cuyos resultados en lectura crítica están sobre la media y a 25.283 casos que están bajo la media. Por otra parte, clasifica incorrectamente a 15.548 casos de estudiantes cuyos resultados en lectura crítica están sobre la media y 18.726 casos que están bajo la media. Esto significa que el modelo clasifica correctamente al 70.7 % de los estudiantes que están sobre la media en lectura crítica y al 57.4% de las estudiantes que están bajo la media en esta competencia.

De acuerdo a las reglas de clasificación obtenidas con este modelo, los patrones más representativos descubiertos, teniendo en cuenta el soporte y la confianza, son:

Regla 1. Si la IES no está acreditada y la metodología de formación es a distancia entonces el desempeño de los estudiantes en lectura crítica tiene mayor probabilidad de estar bajo la media. El 12.48% de los 97.055 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 69.11% de los 12.112 (12.48%) estudiantes se clasifican correctamente y el 27.52% de los 44.009 estudiantes que están bajo la media cumplen este patrón.

Regla 2. Si la IES no está acreditada, la metodología de formación es presencial, el carácter académico de la institución es Universidad, el número de IES en la zona a la que pertenece la institución está entre 60 y 70, el número de estudiantes de programas profesionales en la zona en la que se ofrece el programa está entre 10.000 y 20.000, los estudiantes pertenecen a programas no acreditados y además tienen personas a cargo, entonces el desempeño de los estudiantes en lectura crítica tiene mayor probabilidad de estar bajo la media. El 4.85% de los 97.055 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 62% de los 4.710 (4.85%) estudiantes se clasifican correctamente y el 10.70% de los 44.009 estudiantes que están bajo la media cumplen este patrón.

Regla 3. Si la IES no está acreditada, la metodología de formación es presencial, el carácter académico de la institución es Institución Universitaria y los estudiantes están clasificados en el nivel 1 del SISBEN entonces el desempeño de los estudiantes en lectura crítica tiene mayor probabilidad de estar bajo la media. El 2.07% de los 97.055 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 67.35% de los 2.012 (2.07%) estudiantes se clasifican correctamente y el 4.57% de los 44.009 estudiantes que están bajo la media cumplen este patrón.

Regla 4. Si la IES está acreditada entonces el desempeño de los estudiantes en lectura crítica tiene mayor probabilidad de estar sobre la media. El 16.80% de los 97.055 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 78.87% de los 16.306 (16.80%) estudiantes se clasifican correctamente y el 30.74% de los 53.046 estudiantes que están sobre la media cumplen este patrón.

Regla 5. Si la IES no está acreditada, la metodología de formación es presencial, el carácter académico de la institución es Universidad, el número de IES en la zona a la que pertenece la institución está entre 60 y 70, el número de estudiantes de programas profesionales en la zona en la que se ofrece el programa es mayor que 20.000 entonces el desempeño de los estudiantes en lectura crítica tiene mayor probabilidad de estar sobre la media. El 13.93% de los

97.055 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 62.80% de los 13.524 (13.93%) estudiantes se clasifican correctamente y el 25.49 % de los 53.046 estudiantes que están sobre la media cumplen este patrón.

Regla 6. Si la IES no está acreditada, la metodología de formación es presencial, el carácter académico de la institución es Universidad, el número de IES en la zona a la que pertenece la institución es mayor que 70 entonces el desempeño de los estudiantes en lectura crítica tiene mayor probabilidad de estar sobre la media. El 8.37% de los 97.055 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 69.43% de los 8.126 (8.37%) estudiantes se clasifican correctamente y el 15.32 % de los 53.046 estudiantes que están sobre la media cumplen este patrón.

Regla 7. Si la IES no está acreditada, la metodología de formación es presencial, el carácter académico de la institución es Universidad, el número de IES en la zona a la que pertenece la institución está entre 60 y 70, el número de estudiantes de programas profesionales en la zona en la que se ofrece el programa está entre 10.000 y 20.000, los estudiantes pertenecen a programas no acreditados, no tienen personas a cargo, no están clasificados en el SISBEN y la institución es de tipo oficial entonces el desempeño de los estudiantes en lectura crítica tiene mayor probabilidad de estar sobre la media. El 2.80% de los 97.055 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 61.11% de los 2.718 (2.80%) estudiantes se clasifican correctamente y el 5.12 % de los 53.046 estudiantes que están sobre la media cumplen este patrón.

IV. Discusión

Para efectos de la discusión de los resultados, se escogieron los patrones de mayor confianza, tanto los que se ubican por encima de la media, como aquellos que se sitúan por debajo de ella. En los patrones descubiertos, se destacan la acreditación institucional y la modalidad de estudio como dos factores importantes asociados al desempeño académico de los estudiantes en las pruebas SaberPro, en la competencia genérica de Lectura Crítica.

En efecto, de acuerdo a los modelos de clasificación basados en árboles de decisión, construidos para esta competencia genérica, muestran que el patrón de IES acreditada se asocia con el buen desempeño académico de los estudiantes de programas profesionales que presentaron las pruebas Saber Pro; resultado que está en línea con la fundamentación conceptual expresada en la política pública sobre calidad de la educación superior en Colombia, particularmente, sobre el tema de acreditación institucional; pues de hecho, el cumplimiento de los doce factores de calidad señalados por el Consejo Nacional de Acreditación CNA [20], dan cuenta de las condiciones de favorabilidad de todos los procesos al interior de las IES, que permiten un mejor desempeño académico de los estudiantes.

En la investigación sobre compromiso estudiantil y desempeño académico universitario, realizada en siete universidades colombianas con acreditación institucional de alta calidad, llevada a cabo por Pineda [21], los hallazgos desde la perspectiva de los estudiantes, señalan que estas universidades se preocupan por proponer actividades y tareas que implican la aplicación de conceptos y teorías en problemas prácticos, le asignan importancia al trabajo en equipo, utilización de medios electrónicos con fines académicos y a procesos de realimentación del aprendizaje de los estudiantes, fomentan espacios de socialización que fortalecen la calidad de las relaciones estudiante y comunidad educativa; estrategias que favorecen el desempeño académico de los estudiantes en el desarrollo de las competencias genéricas.

En el mismo orden de ideas, si la IES no está acreditada, este hecho junto con la modalidad de estudio a distancia, se constituye en patrón que se asocia con el bajo desempeño académico de los estudiantes de carreras profesionales en la competencia genérica de Lectura Crítica de las pruebas Saber Pro.

La modalidad del programa, según resultados de la investigación de Parra et al. [22], realizada con los estudiantes de primer semestre de pregrado de la Facultad de Ingeniería de la Universidad de Antioquia (Colombia), cohorte 2012-2, hace referencia a la metodología mediante la cual se oferta y desarrolla el programa, la cual puede ser: presencial, semi-presencial, a distancia y virtual. En dicho estudio se encontró que la modalidad del programa se asocia de manera importante al desempeño académico. Por ejemplo, en la condición de insuficiente quedaron el 17% de los matriculados en la modalidad presencial y el 50% de los de la modalidad virtual; cifras que indican que los estudiantes de programas presenciales tienen mejor desempeño que aquellos de modalidad a distancia.

Por su parte, Artunduaga [23], en relación con la modalidad de formación, considera que es factor clave del desempeño académico la forma cómo se orienta al estudiante en el aprendizaje; al respecto, Vásquez y Rodríguez [24], evidencian el papel del estudiante de modalidad a distancia, que es él, el responsable de su éxito educativo, le corresponde auto-dirigirse y auto-organizarse, a fin de alcanzar los objetivos propuestos en el plan de estudio. Dichos autores argumentan que si no tiene desarrolladas estas capacidades de independencia, tendería al bajo desempeño, pues, su característica principal está dada por la separación entre estudiante y docente, que puede conllevar de manera paradójica las ventajas y desventajas de este modelo y, potenciar igualmente efectos desfavorables si los programas no están bien estructurados y diseñados.

En igual sentido, Rubio [25], considera que la modalidad a distancia, a diferencia de la presencial, tiene sus propias características: autonomía en el aprendizaje, utilización de tecnología, apoyo-tutorial, comunicación masiva, entre otros, que influyen en el desempeño académico; Sarramona [26],

señala que la adaptación del material para la educación a distancia con todos los principios pedagógicos que ésta requiere, constituye una tarea difícil, dada la falta de especialistas en el tema; a menudo se confunde un recurso para la autoformación con un simple material informativo, lo cual, posteriormente se convierte en factor negativo en el desempeño académico del estudiante y fuente de fracaso de los programas. El autor afirma que el bajo desempeño académico en la modalidad a distancia puede provenir de varias fuentes: capacidad intelectual insuficiente, débil compromiso académico y otros factores provenientes de las dimensiones sociodemográfica y económica, que caracterizan al estudiante de modalidad a distancia; en los sociodemográficos destaca la edad, el grupo familiar, las personas a cargo, el nivel cultural, el estado civil, la ubicación geográfica, entre otros; en los factores económicos subraya el estrato social, los ingresos económicos individuales y familiares; por su parte, Moncada y Rubio [27], exponen, además, que estos estudiantes tienen características particulares que dependen de la dispersión geográfica, edad, estado civil, motivación, intereses profesionales, falta de tiempo, el aislamiento, la falta de recursos y manejo de la tecnología, la mayoría posee obligaciones laborales y familiares.

Analizando el modelo de la figura 1, en la competencia de lectura crítica el buen desempeño se asocia al hecho que la IES esté acreditada. En el caso de que la IES no lo sea, junto con la modalidad presencial, también es un patrón de buen desempeño. Siguiendo el modelo jerárquicamente desde el nodo de modalidad presencial, se encuentra que el buen desempeño académico en esta competencia, está asociado además, con altos ingresos familiares (dimensión socioeconómica), un número alto y medio de IES y estudiantes en la zona y con programas acreditados (dimensión institucional).

Por su parte, el bajo desempeño en lectura crítica tiene relación con IES no acreditada y la modalidad a distancia. En el caso de la modalidad presencial, el bajo desempeño se asocia a la dimensión económica bajo nivel SISBEN y con la dimensión institucional, tipo de institución: instituciones tecnológicas o técnicas profesionales. En este contexto, Garbanzo [28], indica que factores como la pobreza y la falta de apoyo social están relacionados con el desempeño académico; no obstante, advierte que, no hay correspondencia estricta entre las desigualdades sociales y educativas; indica que existen otros factores tales como la familia, el sistema educativo, la institución, que pueden incidir en tal desigualdad. Seibold [29], señala que, si bien el contexto socioeconómico afecta el nivel de calidad educativa, de ningún modo lo determina; reconoce que se presenta asociación entre estas variables, pero no una relación causa–efecto.

Estos hechos coinciden con Garbanzo [28], Seibold [29] y Montero & Villalobos [30], en el sentido que un resultado generalmente aceptable en el desempeño académico, es la existencia de una asociación significativa entre el nivel socioeconómico del estudiante y su desempeño académico, hecho que coincide con lo encontrado en la presente investigación.

V. Conclusiones y Trabajos futuros

Los resultados obtenidos con el modelo de clasificación por árboles de decisión, indican que este es capaz de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos que se encuentran almacenados en las bases de datos del ICFES.

En los patrones de desempeño académico descubiertos en la competencia genérica de Lectura Crítica, la acreditación institucional se encuentra asociada con otros factores, tanto en el desempeño por encima de la media como por debajo de ella. Entre dichos atributos se destacan: la modalidad del programa, el área del conocimiento, el número de instituciones en la zona, el número de estudiantes por zona, el estrato socioeconómico, el número de personas a cargo del estudiante y el género, entre otros; atributos que, de alguna manera, son considerados en los resultados de investigaciones de diversos autores en relación con el desempeño académico.

Entre las dificultades presentadas en el desarrollo de la investigación están la mala calidad de los datos de las bases de datos del ICFES, que se tuvieron que descartar ciertos atributos por la imposibilidad de obtener sus valores en otras fuentes, y que de alguna manera, podrían influir en el descubrimiento de los patrones objeto de este estudio, además del gran consumo de recursos que implicó el proceso de limpieza y transformación de datos.

Se plantea como trabajos futuros complementar este estudio utilizando otras técnicas de minería de datos que permitan relacionar que atributos se presentan juntos asociados al desempeño académico en las pruebas Saber Pro y cómo se agrupan los individuos de acuerdo a su rendimiento en dichas pruebas, y la realización de estudios sobre el desempeño académico en las pruebas Saber Pro en programas de formación tecnológica con técnicas de minería de datos.

Además, sería recomendable realizar estudios sobre el análisis relacional entre las pruebas Saber Once, el desempeño académico en las IES en la formación profesional y las pruebas SaberPro; la relación que pudiera existir entre el número de estudiantes y el número de IES en una zona geográfica con el desempeño académico; investigaciones comparadas sobre desempeño académico entre las diferentes modalidades de estudio, entre otras iniciativas.

Agradecimientos

Este proyecto de investigación se financió con recursos del Instituto Colombiano para la Evaluación de la Educación Superior ICFES y con recursos de contrapartida de la Universidad de Nariño y la Universidad Cooperativa de Colombia sede Pasto.

REFERENCIAS

[1] ICFES, 2011, Lineamientos Saber Pro. Bogotá: 2011. [online]: http://aprendeenlinea.udea.edu.co/lms/moodle/file.php/532/Lineamientos_SABER_PRO_2011_2_30_08_1_.pdf. [consulta: 26/11/2013].

[2] ICFES, 2012, Examen Saber Pro, junio de 2012–I. Módulos de competencias genéricas y especificas disponibles. Evaluación de la calidad de la educación superior. Bogotá: 2012. [on line]: www.icfes.gov.co/examenes/.../151-saber-pro-modulos-de-competencias. [consulta:26/11/2013].

[3] ICFES, 2012, Saber Pro: Principales resultados en Competencias Genéricas. Santa Marta, Colombia: 2012. [on line]: www.icfes.gov.co/examenes/.../151-saber-pro-modulos-de-competencias. . [consulta:28/11/2013].

[4] L. Zapata, 2011, Factores académicos asociados al bajo rendimiento en inglés en las pruebas ECAES presentadas por los estudiantes de la Facultad de Educación en el año 2009. (Trabajo de grado de pregrado). Fundación Universitaria Luis Amigó, Facultad de Educación, Licenciatura en Educación Básica con Énfasis en Inglés. Medellín, Colombia, 2011.

[5] UNAL, 2012, Universidad Nacional de Colombia. Análisis de los resultados obtenidos por la Universidad Nacional de Colombia sede Bogotá en las pruebas Saber Pro 2011–2. Bogotá: Universidad Nacional de Colombia. [on line]: www.unal.edu.co/diracad/evaluacion/SaberPro_2012/analisis_de_resultados.pdf.[[consulta:28/11/2013].

[6] R. Timarán, A. Calderón, and J. Jiménez, 2013. Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil [Application of data mining in extracting student dropout profiles]. *Ventana Informática*, (28). Recuperado a partir de http://revistasum.umanizales.edu.co/ojs/index.php/ventanainformatica/article/view/181.

[7] R. Timarán, A. Calderón, and J. Jiménez, 2013. La minería de datos como un método innovador para la detección de patrones de deserción estudiantil en programas de pregrado en instituciones de educación superior. En *WEEF 2013 Cartagena*. Recuperado a partir de http://www.acofipapers.org/index.php/acofipapers/2013/paper/view/211.

[8] S. Valero 2009. Aplicación de técnicas de minería de datos para predecir deserción. *Aplicación de técnicas de minería de datos para predecir deserción. México*.

[9] S. Valero, A. Salvador, and M. García, 2010. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene*, *779*(73), 33.

[10] S. Orea, A. Vargas, and M. Alonso, 2005. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene*, *779*(73), 33.

[11] J. Hernández, M. Ramírez, and C. Ferri, 2005. Introducción a la Minería de Datos. *Editorial Pearson Educación SA, Madrid*. Recuperado a partir de http://dspace.ucbscz.edu.bo/dspace/handle/123456789/526

[12] J. Han, and M. Kamber, 2001. *Data Mining: Concepts and Techniques, Third Edition* (3 edition.). Burlington, MA: Morgan Kaufmann.

[13] K. Sattler, and O. Dunemann, 2001. SQL database primitives for decision tree classifiers. En *Proceedings of the tenth international conference on Information and knowledge management* (pp. 379–386). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=502650

[14] R. Timarán, and M. Millán, 2006. New algebraic operators and SQL primitives for mining classification rules. En *Computational Intelligence* (pp. 61–65). Recuperado a partir de http://www.actapress.com/PaperInfo.aspx?PaperID=29048&reason=500

[15] E. Hernández, and R. Lorente, 2009. *Minera de datos aplicada a la detección de Cáncer de Mama*. Universidad Carlos III de Madrid. Recuperado a partir de http://tps5to-utn-frre.googlecode.com/svn/trunk/BI/Cancer%20de%20Mama/14.pdf

[16] M. Hall, E. Frank, and I. Witten, 2011. Practical Data Mining: Tutorials. University of Waikato. Recuperado a partir de http:// www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf

[17] J. Quinlan, 1993. *C4. 5: programs for machine learning* (Vol. 1). Morgan kaufmann. Recuperado a partir de http://books.google.com.co/books?hl=es&lr=&id=HExncpjbYroC&oi=fnd&pg=PR7&dq=Programs+for+Machine+Leraning&ots=nLkbbRq2Yj&sig=Y5h5CQUdtbZjs1Fjd8ilbJfyRLE

[18] M. García, and A. Álvarez, 2010. Análisis de datos en WEKA–pruebas de selectividad. *línea] disponible en http://www. it. uc3m. es/jvillena/irc/practicas/06-07/28. pdf*. Recuperado a partir de http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

[19] G. Fernández, 2009. Extracción de Información de la Web usando Técnicas de Minería de Datos. Recuperado a partir de Disponible en: http://www.tdg-seville.info/Download.ashx?id=48

[20] Consejo Nacional de Acreditación – CNA. 2013. Lineamientos para la acreditación de programas de pregrado. Bogotá. CNA.

[21] C. Pineda, and A. Pedraza, 2011. Persistencia y graduación. Hacia un modelo de retención estudiantil para Instituciones de educación superior. Bogotá: ARFO Editores e Impresores Ltda

[22] C. Parra, L. Mejía, A. Valencia, E. Castañeda, G. Restrepo, O. Usuga, and R. Mendoza, 2013. Rendimiento académico de los estudiantes de primer semestre de pregrado de la Facultad de Ingeniería de la Universidad de Antioquia: cohorte 2012-2. Medellín: Ingeniería y Sociedad. Disponible en: http://www.udea.edu.co/portal/page/portal/bibliotecaSedesDependencias/unidadesAcademicas/FacultadIngenieria/Diseno/Archivos/Tab/Rendimiento%20acad%C3%A9mico%20de%20los%20estudiantes%5B1%5D.pdf.

[23] M. Artunduaga, 2008. Variables que influyen en el rendimiento académico en la Universidad. Recuperado de http://es.slideshare.net/1234509876/variables-del-rendimiento-acadmico-universidad.

[24] C. Vásquez, C. and M. Rodríguez, 2007. La deserción estudiantil en educación superior a distancia: perspectiva teórica y factores de incidencia. Revista Latinoamericana de Estudios Educativos, XXXVII(3 y 4), 107-122.

[25] M. Rubio, 2009. Nuevas Orientaciones y Metodología para la Educación a Distancia Loja - Ecuador: Editorial de la Universidad Técnica Particular de Loja.

[26] J. Sarramona, 2002. Evaluación de programas de educación a distancia. Barcelona: Universidad Autónoma de Barcelona.

[27] L. Moncada, and M. Rubio, 2011. Determinantes inmediatos del rendimiento académico en los nuevos estudiantes matriculados en el sistema de educación superior a distancia del Ecuador: caso Universidad Técnica Particular de Loja. RIED, 14 2, 77-95.

[28] G. Garbanzo, 2007. Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde calidad de la educación superior pública. Revista Educación, 31(1):43-63.

[29] J. Seibold, 2000. La calidad integral en educación. Reflexiones sobre un nuevo concepto de calidad educativa que integre valores y equidad educativa. Revista Iberoamericana de Educación. Recuperado de http://www.rieoei.org/rie23a07.htm

[30] E. Montero, and J. Villalobos, 2004. Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico y a la repetición estudiantil en la Universidad de Costa Rica. San José, Costa Rica: Universidad de Costa Rica.

ANEXO I

DICCIONARIO DE DATOS DEL REPOSITORIO FINAL T97055LEC

No	Atributo	Descripción	Valores
		Sociodemográficos	
1	estu_genero	Género	M ,F
2	estu_edad	Edad del estudiante en el momento de presentar la prueba	
3	estado_civil	Estado civil del estudiante.	SOLTERO, CASADO

			SEPARADO/DIVORCIADO, UNION LIBRE, VIUDO
4	estu_hogar_actual	Tipo de hogar actual donde reside el estudiante	HABITUAL/PERMANENTE TEMPORAL
5	estu_sn_cabeza_fmlia	Si el estudiante es cabeza de familia o no	SI, NO
6	estu_pers_cargo	Si el estudiante tiene personas a cargo o no	SI, NO
7	fami_nivel_educa_padres	Máximo nivel educativo completo entre el padre y la madre	PRIMARIA, SECUNDARIA, TÉCNICO/TECNOLOGICO, PROFESIONAL,POSTGRADO NINGUNO
8	fami_ocup_padre	Ocupación del padre	DIRECTIVO,EMPLEADO, EMPRESARIO,HOGAR, INDEPENDIENTE,OTRA PENSIONADO,PROFESIONAL
9	fami_ocup_madre	Ocupación de la madre	Los mismos valores de la ocupación del padre
Económicos			
10	estu_financiacion_matricula	Forma de financiar el pago de la matrícula	PROPIOS, CRÉDITO, PADRES BECA Y TODAS LAS COMBINACIÓNES POSIBLES ENTRE ESTOS VALORES
11	estu_estrato	Estrato socioeconómico del estudiante	ESTRATOS 1 a 6, ZONA RURAL SIN ESTRATO
12	fami_nivel_sisben	Nivel de clasificación en el SISBEN al que pertenece el estudiante	NIVELES 1, 2, 3, OTRO NIVEL, NO ESTA EN SISBEN
13	econ_condicion_vivienda	Condición de la vivienda del estudiante	BUENA, MALA,REGULAR
14	eco_condicion_hogar	Condición del hogar del estudiante	BUENA, MALA,REGULAR
15	eco_condicion_transporte	Condición de transporte en el hogar del estudiante	PARTICULAR, PÚBLICO
16	eco_condicion_tic	Condición de uso de TIC en el hogar del estudiante	BUENA, REGULAR, MALA
17	eco_condicion_vive	Condición de vida del estudiante	SIN HACINAMIENTO, HACINAMIENTO MEDIO, HACINAMIENTO CRITICO
18	fami_ing_fmliar_mensual	Ingresos mensuales familiares en salarios mínimos	
19	estu_trabaja	Si estudiante trabaja o no	
Académicos			
20	estu_metodo_prgm	Metodología del programa académico bajo la cual cursa el estudiante	A DISTANCIA, PRESENCIAL
21	estu_area_conoc	Nombre del área de conocimiento a la que pertenece el programa académico del estudiante	
22	area_grupo_referencia	Área del grupo de referencia de los programas	CIENCIAS HUMANAS, CIENCIAS SOCIALES, CIENCIAS NATURALES Y TÉCNICAS
23	estu_pje_creditos	Porcentaje de créditos cursados por el estudiante al realizar la prueba	MAS DEL 90 ENTRE 81 Y 90 ENTRE EL 75 Y EL 80 MENOS DEL 75 NO SIGUE EL SISTEMA DE CRÉDITOS
24	estu_titulo_bto	Tipo de título de bachillerato obtenido	ACADEMICO, NORMALISTA, TÉCNICO
Institucionales			
25	inst_tipo	Tipo de institución del estudiante	OFICIAL,PRIVADA, RÉGIMEN ESPECIAL
26	inst_caracter_academico	Carácter académico de la IES	ESCUELA TECNOLÓGICA, INSTITUCIÓN TECNOLÓGICA, INSTITUCIÓN UNIVERSITARIA, TÉCNICA PROFESIONAL, UNIVERSIDAD
27	inst_acreditada	Si la institución donde pertenece el estudiante es acreditada o no según CNA	ACREDITADA, NO ACREDITADA
28	inst_prog_acreditado	Si el programa al cual pertenece el estudiante es acreditado o no según CNA	ACREDITADO, NO ACREDITADO
29	inst_programa_zona	Zona geográfica donde se ofrece el programa	BOGOTA, EJE CAFETERO CARIBE, CENTRO ORIENTE, PACÍFICO,CENTRO SUR,

			LLANO
30	num_estudiantes_zona	Número de estudiantes por zona	ALTO, MEDIO, BAJO
31	num_instituciones_zona	Número de Instituciones de Educación Superior (IES) por zona	ALTO, MEDIO,BAJO
Clase			
32	mod_lectura_critica_desemp	Desempeño del estudiante en el Módulo de lectura crítica	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA(< la media)

ANEXO II

ÁRBOL DE DECISIÓN PARA LECTURA CRÍTICA