DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO ACADÉMICO EN LAS COMPETENCIAS GENÉRICAS DE LAS PRUEBAS SABER PRO 2011-2

INFORME DE FINAL

Grupo de Investigación Aplicada en Sistemas-GRIAS
Universidad de Nariño

Grupo de Investigación Potencial Sináptico
Universidad Cooperativa de Colombia

San Juan De Pasto
Diciembre de 2014

DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO ACADÉMICO EN LAS COMPETENCIAS GENÉRICAS DE LAS PRUEBAS SABER PRO 2011-2

INFORME DE FINAL

Investigador Principal:
Ricardo Timarán Pereira, Ph.D.

Coinvestigadores:
Javier Caicedo Zambrano, Ph.D.
Arsenio Hidalgo Troya, Mg.
Isabel Hernandez Arteaga, Ph.D.
Juan Carlos Alvarado, Mg.

Grupo de Investigación Aplicada en Sistemas-GRIAS
Universidad de Nariño

Grupo de Investigación Potencial Sináptico
Universidad Cooperativa de Colombia

San Juan De Pasto
Diciembre de 2014

RESUMEN

En este documento se presentan los resultados del proyecto de investigación denominado "Descubrimiento de Patrones de Desempeño Académico en las Competencias Genéricas de las Pruebas Saber Pro 2011-2" financiado mediante contrato No. 61-2014, celebrado entre el Instituto Colombiano para la Evaluación de la Educación Superior- ICFES y la Universidad de Nariño con la participación de la Universidad Cooperativa de Colombia sede Pasto. El objetivo de este estudio fue descubrir patrones de desempeño académico en competencias genéricas de los estudiantes de programas profesionales en las pruebas Saber Pro 2011-2, a partir de los datos sociodemográficos, económicos, académicos e institucionales almacenados en las bases de datos del ICFES, utilizando técnicas de minería de datos. Se aplicó la metodología CRISP-DM y con los repositorios de datos, limpios y transformados para cada competencia, se descubrieron patrones asociados al buen o mal desempeño académico utilizando la técnica de clasificación basada en árboles de decisión. 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.

1. INTRODUCCIÓN

Uno de los objetivos del examen de estado de calidad de la educación superior Saber Pro, según el decreto 3963 del 14 de octubre de 2009, Ministerio de Educación Nacional -MEN (2009), 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 el Fomento de la Educación Superior –ICFES (2011), todos los estudiantes deben presentar los módulos de competencias genéricas sin importar el programa de formación que cursen, que incluye competencias de lectura crítica, comunicación escrita, razonamiento cuantitativo e inglés.

En la competencia de lectura crítica se evalúan los desempeños asociados a lectura, pensamiento crítico y entendimiento interpersonal (ICFES, 2012a). En comunicación escrita se evalúa la competencia para comunicar ideas por escrito referidas a un tema dado (ICFES, 2011), (ICFES, 2012a). En la competencia de razonamiento cuantitativo se evalúan los desempeños relacionados con uso de lenguaje cuantitativo y solución de problemas (ICFES, 2012a). 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 conocer cómo se asocian los factores socioeconómicos, académicos e institucionales del estudiante para obtener un determinado nivel de desempeño en estas competencias de las pruebas Saber Pro.

Los estudios que se han realizado hasta el momento (ICFES, 2010), (Zapata, 2011), (ICFES, 2012b) y (UNAL, 2012) 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. Mientras que la Estadística plantea hipótesis que deben ser validadas a partir de los datos disponibles, la minería de datos descubre patrones no previstos con la Estadística. Los resultados que se obtienen con las técnicas de minería de datos no se pueden utilizar para confirmar o rechazar hipótesis, porque puede conducir a errores críticos. Su función es explorar datos, darles sentido, convertir en

conocimiento un volumen de datos -que poco o nada aportan a la descripción- con el fin de interpretar fenómenos y adoptar decisiones de acuerdo a las necesidades.

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 (Timarán, Calderón, & Jiménez, 2013a) y (Timarán, Calderón, & Jiménez, 2013b). 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 (Valero, 2009), (Valero, Salvador, & García, 2010). 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, (Orea, Vargas, & Alonso, 2005).

El objetivo de este estudio fue descubrir patrones de desempeño académico en competencias genéricas de los estudiantes de programas profesionales en las pruebas Saber Pro 2011-2, a partir de los datos de los atributos sociodemográficos, económicos, académicos e institucionales almacenados en las bases de datos del ICFES, utilizando técnicas de minería de datos. Para alcanzar este objetivo, se requirió apropiar y analizar los elementos que constituyen cada una de las competencias genéricas; obtener un repositorio limpio y transformado de datos a partir de la selección de los atributos sociodemográficos, económicos, académicos e institucionales de las bases de datos del ICFES; generar el modelo de detección de patrones de desempeño académico mediante la aplicación de técnicas de minería de datos y, finalmente, generar conocimiento a partir de la interpretación de los patrones que se obtuvieron.

La presente investigación pretendió dar respuesta a la pregunta: ¿Cuáles son los patrones sociodemográficos, económicos, académicos e institucionales asociados al desempeño académico de los estudiantes en las competencias genéricas de los programas profesionales en las pruebas Saber Pro 2011-2?

2. 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 (Hernández, Ramírez, & Ferri, 2005).

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.

2.1 Comprensión del Negocio o 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.

El desempeño académico es un tema complejo en el campo de la educación superior, ya que, según Lerner (2012), tiene efecto en el proyecto educativo de los estudiantes, en su formación profesional y posterior desempeño laboral. El estudio realizado por Bodensiek (2010), concluye que factores como género, edad, trayectoria académica, características familiares, ocupación, nivel educativo de los padres, relaciones entre familia y escuela, cultura, infraestructura, modalidad de enseñanza y formas de evaluación, son elementos asociados al desempeño académico; con lo cual se pone en evidencia relaciones subjetivas, sociales y económicas complejas.

Si bien se puede entender el desempeño académico como la relación que existe entre el proceso de aprendizaje y los resultados obtenidos en valores predeterminados; autores como Lerner y Montes (2012), consideran que el desempeño académico no se explica únicamente por puntajes obtenidos; suponen la existencia de otros factores que aportan a su comprensión, los cuales se agrupan en las dimensiones: sociodemográfica, económica, académica, familiar, personal e institucional. Por su parte, Tonconi (2010), señala que el desempeño académico concebido como resultado, no siempre da cuenta de las competencias logradas por los estudiantes en el proceso de formación; su esfuerzo y la calidad de la formación no son directamente proporcionales con el resultado obtenido; por lo tanto, es necesario concebir un concepto de desempeño académico que corresponda e incluya el proceso del estudiante, sus condiciones sociodemográficas, económicas, académicas e institucionales.

Al respecto Castaño, Gallón, Gómez, y Vásquez (2004), encuentran cuatro conjuntos de factores determinantes del desempeño académico: las características sociodemográficas, académicas, socioeconómicas e institucional. Del mismo modo, Di Gresia y Porto (2004), toman

como variables de su estudio, las características del estudiante y su familia, encontrando que existen factores explicativos del desempeño académico, como: género, edad y educación de los padres. Siguiendo esta línea, Mizala, Romaguera, y Reinaga (2002), analizan los factores asociados al desempeño académico, identificando factores relacionados con las categorías: sociales, socioeconómicas e institucionales. Garbanzo (2014), expresa que en la actualidad, la tendencia económica y el desarrollo de la sociedad de la información y el conocimiento, requieren la revisión del desempeño académico del estudiante universitario, prioriza en su estudio la dimensión socioeconómica.

El concepto de desempeño académico que mejor enmarca esta investigación, se corresponde con el presentado por Lerner (2012), en el sentido de concebirlo como la relación entre el proceso de aprendizaje y el logro de competencias (genéricas), que comprende factores extrínsecos e intrínsecos al estudiante, y el producto derivado de él, expresado tanto en valores predeterminados por el contexto sociocultural (puntajes establecidos por el ICFES), como en las decisiones y acciones del estudiante en relación con las competencias genéricas logradas en la formación profesional.

Competencia es un término que en el campo de la educación ha generado una amplia gama de criterios respecto a su interpretación, que presenta significados diversos, y en ocasiones abstractos, en algunas elaboraciones teóricas; pues, se definen desde un enfoque conceptual más que científico. Según Hernández y Ojeda (2009), las competencias son el encadenamiento de saberes, orientados y articulados a la concepción del ser, del saber, saber hacer y del saber convivir. Por lo tanto, su desarrollo constituye un proceso de formación integral del estudiante como persona, ciudadano y profesional, a la hora de enfrentarse a problemas reales. Esta concepción implica cambios sustanciales, de tipo curricular, didáctico y actitudinal, para asumir el aprendizaje significativo, orientado por el pensamiento complejo. Este modelo de formación, demanda pasar de una lógica basada en el discurso teórico, a otra que implica solución de problemas en su contexto. Según Climent (2014), este enfoque está influenciado por la economía global; razón por la cual, la formación por competencias tiende hacia relaciones causa-efecto con los procesos globalizadores.

Este cambio de paradigmas en la formación profesional, según Valera (2010), implica cambios epistemológicos, axiológicos y actitudinales, no como efecto del aprendizaje memorístico, sino de un aprendizaje que acreciente capacidades humanas en el desarrollo integral

de las dimensiones cognoscitivas, como base de la toma de decisiones y la solución de problemas. Martínez, Cegarra y Rubio (2012), puntualizan que la formación por competencias pretende que la persona desarrolle capacidades amplias, aprendiendo, desaprendiendo y adaptándose a situaciones del mundo en transformación. Según el ICFES (2013), las competencias son el componente articulador de los niveles de formación, siendo referentes comunes sobre los que las personas deben saber y saber hacer durante su formación. Toda persona, independientemente de su profesión, debe desarrollar capacidades para aprender durante toda la vida, como: comprender contextos y situaciones, comunicarse con otros, seleccionar y analizar información, solucionar problemas de distinta naturaleza en un mundo cambiante.

En Colombia, la formación profesional contempla el desarrollo de competencias genéricas y específicas. Las genéricas identifican elementos comunes a cualquier profesión; Gonzáles y Wagenaar (2003), reconocen como tales: capacidad de aprender, toma de decisiones, diseño de proyectos, habilidades interpersonales, entre otras; ellas se complementan con las competencias específicas de cada profesión. Según Peña (2008), son transversales a todos los niveles educativos, transferibles a diversidad de funciones y se constituyen en respuesta a necesidades de la sociedad.

Para el ICFES (2011b), son competencias genéricas, aquellas que deben desarrollar todos los estudiantes de educación superior, independientemente del programa o nivel en el que se encuentren. En coherencia con esta concepción, las pruebas Saber Pro 2011-2, se estructuran en módulos, que evalúan no sólo las competencias consideradas fundamentales en un área del conocimiento, sino también, aquellas que deben desarrollar todos los estudiantes de educación superior. Las competencias genéricas del siglo XXI, según el ICFES (2013), son transversales a las áreas curriculares y del conocimiento; en un determinado momento y circunstancias pueden ser transferidas a otros contextos. Se consideran longitudinales, es decir, están presentes en todos los niveles de formación y su complejidad es progresiva a lo largo de toda su formación. Según el ICFES (2011b), la evaluación de competencias genéricas en las pruebas Saber Pro 2011-2 incluyó:

- Lectura Crítica. Evalúa la capacidad para leer de manera analítica y reflexiva, implica comprender planteamientos de un texto e identificar sus perspectivas y juicios de valor. Esto exige que el lector identifique y recupere información presente en uno o varios textos, construya su sentido global, establezca relaciones entre enunciados y evalúe su intencionalidad. Aborda

procesos como: ubicar información, relacionar, construir la representación global del texto, analizar y evaluar la relación entre procesos discursivos y el contexto sociocultural.

- Comunicación escrita. Evalúa la competencia para comunicar ideas por escrito referidas a un tema dado. Cada estudiante debe elaborar un escrito argumentativo según un tema planteado. Incluye acciones como: expresar adecuadamente la intención comunicativa, dar coherencia y cohesión al texto, expresarse mediante un lenguaje apropiado, aplicando las reglas que rigen el lenguaje escrito.

- Razonamiento cuantitativo. Evalúa comprensión de conceptos básicos de matemáticas para analizar, modelar y resolver problemas, aplicando métodos y procedimientos cuantitativos y esquemáticos, basados en las propiedades de los números y en las operaciones de las matemáticas; incluye, interpretación de datos, formulación y ejecución de problemas, evaluación y validación de resultados.

- Inglés. Evalúa la competencia para comunicarse efectivamente en inglés, se alinea con el Marco Común Europeo que permiten clasificar a los examinados en cuatro niveles de desempeño A1, A2, B1 y B2.

2.2 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 2011-2 y sobre los datos sociodemográficos, económicos, 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 encuentran almacenados los datos de todos los estudiantes de programas técnico profesionales, tecnológicos y profesionales que presentaron las pruebas Saber Pro 2011-2. Este repositorio inicial, el cual se denomina T153123A94, cuenta con 153.123 registros y 94 atributos. De este repositorio se seleccionaron únicamente los registros de estudiantes de programas profesionales, objeto de este estudio, quedando un repositorio con 97.068 registros y 94 atributos, identificado como T97068A94 y el cual sirvió de base para las subsiguientes fases. En la tabla 1 se encuentra el diccionario de datos de este repositorio, donde se describe cada uno de sus atributos.

Tabla 1. Diccionario de datos del repositorio inicial T97068A94

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 T97068A94 se clasificaron en cuatro dimensiones: sociodemográfica, económica, académica e institucional. En la tabla 2 se muestra esta clasificación.

Tabla 2. Clasificación de atributos en dimensiones

Con el fin de realizar un trabajo de exploración de los datos del repositorio T97068A94 y a través de un análisis preliminar, determinar que variables son potencialmente importantes para el estudio, se realizó un análisis de correlación entre las cuatro competencias genéricas, utilizando el coeficiente de correlación de Pearson, que establece en qué medida se asocian dichas competencias y un análisis de varianza (ANOVA) que establece las diferencias entre promedios de los datos de los atributos sociodemográficos, económicos, académicos e institucionales, en las cuatro competencias estudiadas. Los resultados del análisis de correlación se presentan en la tabla 3.

Tabla 3. Matriz de correlaciones de Competencias genéricas

De acuerdo a la tabla 3, se observa que las competencias en lectura crítica con razonamiento cuantitativo y lectura crítica con inglés, presentan correlaciones altas ($r>0,5$), (Cohen 1988); lectura crítica con comunicación escrita y razonamiento cuantitativo con inglés, presentan correlación moderada ($0,3<r<=0,5$); y que las competencias en comunicación escrita con razonamiento cuantitativo y comunicación escrita con inglés, presentan correlaciones bajas ($r<=0,3$). Por lo tanto, se espera que los estudiantes con buen desempeño en lectura crítica, también tengan buen desempeño en razonamiento cuantitativo, lo mismo que en inglés.

El ANOVA se realizó con algunas variables sociodemográficas, económicas, académicas e institucionales, a saber: género, estado civil, nivel educativo de los padres, estrato socioeconómico, tipo de institución, modalidad de estudio, acreditación institucional y de programas. Para el efecto se cruzaron los puntajes de las competencias genéricas obtenidos en la prueba con las variables mencionadas y se utilizaron las medidas estadísticas: promedio, desviación estándar, intervalos de confianza y el tamaño del efecto de la diferencia estandarizada de las medias *d* de Cohen.

Para establecer si las diferencias en puntajes en las cuatro competencias son estadísticamente significativas, se calcularon los intervalos de confianza de medias al 95%. Aquellos grupos que muestran rangos que no se traslapan presentan verdaderas diferencias, en caso contrario según los puntajes promedios se agrupan y se ordenan de acuerdo al nivel de desempeño de mayor a

menor. Para su interpretación se forman grupos simbolizados con las letras A, B, C, etc. para denotar puntajes promedios que van de los valores más altos a los más bajos. Igualmente para determinar el tamaño de las diferencias de los promedios entre los diferentes grupos de las variables analizadas se utilizó el estadístico *d* de Cohen con la siguiente escala propuesta por Cohen (1998): [0.0, 0.2] trivial o muy pequeña, (0.2, 0.5] pequeña, (0.5, 0.8] moderada y (0.8, infinito) grande. Este estadístico se obtiene dividiendo la diferencia de medias (en valor absoluto) de los dos grupos a comparar, entre la desviación estándar conjunta de los mismos. Para su cálculo se utilizó como referencia el grupo de más alto promedio en cada competencia.

A continuación se presenta el análisis de desempeño de los estudiantes en las pruebas Saber Pro 2011-2 según género, estado civil, nivel educativo de los padres, estrato socioeconómico, tipo de institución, modalidad de estudio, acreditación institucional y de programas.

En la tabla 4 se muestra el desempeño en las competencias genéricas según el género.

Tabla 4. Género y desempeño académico en competencias genéricas

De acuerdo a la tabla 4, y según el estadístico *d* de Cohen, en las pruebas Saber Pro 2011-2, en las competencias genéricas, los hombres presentan mejor desempeño que las mujeres, con diferencia de magnitud muy pequeña en lectura crítica y moderada en razonamiento cuantitativo e inglés y las mujeres presentan mejor desempeño en comunicación escrita, con diferencia muy pequeña.

En la tabla 5 se muestra el desempeño en las competencias genéricas según el estado civil. En esta tabla se puede observar que los solteros(as) presentan el mejor desempeño en todas las pruebas (grupo A), mientras que casados(as) y separado(as) y/o divorciados(as) pertenecen al segundo grupo (grupo B) igualmente en las cuatro competencias. Los grupos de Viudos(as) y Unión libre presentan el más bajo desempeño en Lectura crítica y Escritura, mientras que en el módulo de Razonamiento Cuantitativo los de Unión Libre están en el grupo intermedio B y los viudos(as) en el último grupo C y finalmente en el módulo de Inglés, los viudos(as) pertenecen al grupo B y los de Unión Libre están en el último grupo C. Según el estadístico *d* de Cohen todos los grupos presentan diferencias pequeñas frente a los solteros, excepto los viudos que alcanzan un tamaño moderado en los módulos de lectura crítica y razonamiento cuantitativo. En comunicación escrita las diferencias se ubican entre muy pequeñas (casados, separados) y pequeñas (unión libre y viudos) frente al grupo de solteros con mejor desempeño. En el módulo de inglés los estudiantes casados, en unión libre y viudos alcanzan diferencias moderadas en los

puntajes promedios frente a los solteros, mientras que las diferencias de los separados/divorciados son pequeñas frente a estos últimos.

Tabla 5. Estado civil y desempeño académico en competencias genéricas

En la tabla 6 se muestra el desempeño en las competencias genéricas según el nivel educativo de los padres.

De acuerdo a esta tabla, la tendencia en el desempeño académico en las cuatro competencias genéricas está asociado directamente con el nivel educativo de los padres; a mayor nivel educativo de estos, mejor desempeño académico del estudiante. Tales resultados se observan en la tabla 5, donde el promedio del desempeño en las cuatro competencias genéricas, crece en la medida que aumenta el nivel de escolaridad de los padres, pues, cada uno de los seis niveles constituye un grupo, así: posgrado (grupo A), profesional (grupo B), técnico (grupo C), secundaria (grupo D), primaria (grupo E) y sin escolaridad (grupo F); los cuales se presentan en la tabla en orden consecutivo según la media. El valor *d* de Cohen indica que en lectura crítica y comunicación escrita las diferencias, en el promedio de puntajes, de los estudiantes cuyos padres alcanzan niveles de posgrado frente a los estudiantes cuyos padres tiene estudios profesionales y técnicos, son pequeñas, mientras que, frente a los estudiantes con padres con niveles de secundaria y primaria, son moderadas y frente a los estudiantes con padres sin ningún nivel educativo, son grandes, en el módulo de lectura crítica y moderadas en comunicación escrita. En razonamiento cuantitativo, frente al grupo de mejor desempeño, los estudiantes cuyos padres tienen estudios de profesionales, presentan diferencias muy pequeñas, los estudiantes con padres con estudios técnicos y de secundaria, presentan diferencias pequeñas y finalmente los de padres con estudios de primaria o ninguno presentan diferencias moderadas. En inglés, las diferencias son pequeñas, cuando los padres tienen estudios profesionales, moderadas, si tienen estudios técnicos y grandes si alcanzan máximo estudios de secundaria, frente a los estudiantes con padres que tienen estudios de posgrado.

Tabla 6. Nivel educativo de los padres y desempeño académico en competencias genéricas

En la tabla 7 se muestra el desempeño en las competencias genéricas según el estrato socioeconómico.

Tabla 7. Estrato socioeconómico y desempeño académico en competencias genéricas.

Según la tabla 7, el desempeño académico en las cuatro competencias, se asocia con el estrato socioeconómico del estudiante; marcando la tendencia, que a mayor estrato, mejor

desempeño académico. Los resultados muestran que el promedio del desempeño asciende en la medida que sube el estrato social del estudiante. Sin embargo el estadístico *d* de Cohen sugiere que las diferencias del promedio de puntajes, frente al estrato 6, van desde muy pequeñas (estratos 5 y 4) hasta moderadas (estrato 1 y rural) en el módulo de lectura crítica, desde muy pequeñas (estrato 5) a grandes (estrato 1) en los módulos de comunicación escrita y razonamiento matemático (estrato 1) y en el módulo de Inglés, los estudiantes de todos los estratos presentan diferencias grandes, excepto los estudiantes de estrato 5 que presentan diferencias pequeñas.

En la tabla 8 se muestra el desempeño en las competencias genéricas según la modalidad de estudio

Tabla 8. Modalidad de estudio y desempeño académico en competencias genéricas

Se observa en la tabla 8, que los estudiantes que se forman en modalidad presencial (grupo A), obtienen mejor desempeño académico que aquellos que estudian en programas a distancia (grupo B). Las diferencias en todos los módulos son moderadas, excepto en comunicación escrita, donde las diferencias son moderadas, según el estadístico *d* de Cohen.

En la tabla 9 se muestra el desempeño en las competencias genéricas según el tipo de institución.

Tabla 9. Tipo de institución y desempeño académico en competencias genéricas

La tabla 9 muestra dos grupos, las IES oficiales (grupo A) y privadas y de régimen especial (grupo B). Se puede observar que los estudiantes del grupo A tienen mejor desempeño que los del grupo B. En la competencia de comunicación escrita, no hay diferencia significativa en los desempeños académicos de estudiantes de IES oficiales, privadas y de régimen especial. En razonamiento cuantitativo, se distinguen dos grupos: IES oficiales y de régimen especial (grupo A) y privadas (grupo B), presentando mejor desempeño los estudiantes del grupo A. Finalmente, en inglés, se notan dos grupos: IES privadas (grupo A) y oficiales y de régimen especial (grupo B), presentando mejor desempeño los del grupo A, lo cual puede ocurrir por tener mejores posibilidades de acceso a materiales y cursos adicionales. Sin embargo, teniendo en cuenta el estadístico *d* de Cohen, en todas las competencias genéricas, las diferencias de los promedios de los puntajes, son muy pequeñas o pequeñas.

En la tabla 10 se muestra el desempeño en las competencias genéricas según la acreditación institucional y en la tabla 11 el desempeño en las competencias genéricas según la acreditación de los programas académicos.

Según la tabla 10, se evidencia que estudiantes de IES acreditadas (grupo A), tienen mejor desempeño académico en las cuatro competencias objeto estudio, en comparación con estudiantes que se forman profesionalmente en IES no acreditadas (grupo B), lo cual se muestra en la tabla 10. Según el estadístico d de Cohen las diferencias de medias entre los dos grupos anteriores son pequeñas en comunicación escrita, moderadas en lectura crítica y razonamientos cuantitativos y grandes en el módulo de inglés.

El mismo comportamiento anterior en el desempeño académico se encuentra en los programas acreditados (grupo A) y no acreditados (grupo B), tal como se muestra en la tabla 11, donde el mejor desempeño lo presentan los estudiantes de programas acreditados. Según el estadístico d de Cohen las diferencias de medias entre los dos grupos anteriores, son pequeñas en lectura crítica, comunicación escrita y razonamientos cuantitativos y moderadas en inglés.

Tabla 10. Acreditación Institucional y desempeño académico en competencias genéricas

Tabla 11. Acreditación del programa y desempeño académico en competencias genéricas

2.3 Preparación de los Datos

En esta fase, los 94 atributos del repositorio base T97068A94, 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, listo para aplicarle las técnicas de minería de datos.

El proceso de limpieza y transformación de este repositorio se realizó de la siguiente manera: Se realizó una exploración inicial de los datos almacenados en el repositorio T97068A94 con el fin de determinar la calidad de los datos. En la tabla 12 se describen los atributos del repositorio con un porcentaje alto de valores nulos. En la tabla 13 se muestran los atributos con valores constantes y los atributos identificadores.

Tabla 12. Atributos con porcentaje alto de datos nulos

Tabla 13. Atributos con valores constantes y atributos identificadores

Se efectuó una primera selección de atributos y se descartaron aquellos atributos con un alto porcentaje de valores nulos, por la imposibilidad de encontrar sus valores a través de fuentes

externas de datos (ver tabla 12); tampoco se tomaron en cuenta los atributos con valores constantes y aquellos que servían de identificadores de cada estudiante (ver tabla 13).

La alta dimensionalidad es un problema para el descubrimiento de patrones con minería de datos (Hernández, Ramirez & Ferri, 2005). Uno de los criterios utilizados para resolver este problema, es reducir el número de atributos a analizar, a través de la transformación de estos, en nuevos atributos que generalicen los datos y que brinden mayor información. Teniendo en cuenta este criterio, en el repositorio de datos T97068A94, se seleccionaron los atributos que por sí mismos no tenían mayor significado, pero que si se consolidaban en un nuevo, adquirían mayor semántica; por esta razón, se crearon nuevos atributos en reemplazo de éstos, obteniendo atributos más representativos para el estudio y reduciendo el número de variables a tener en cuenta en la investigación. En la tabla 14 se describe el proceso sobre cómo se construyeron estos nuevos atributos y los atributos que reemplazan. Los atributos reemplazados fueron eliminados.

Tabla 14. Nuevos atributos del repositorio de datos

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 15 se detalla este proceso para el atributo estu_edad.

Tabla 15. Valores discretizados del atributo estu_edad_examen

En esta investigación, se consideró conveniente crear un nuevo atributo denominado inst_programa_zona, con el cual se reorganizarían las IES en zonas geográficas, de manera que la información suministrada aportara al descubrimiento de patrones de desempeño académico, ya que el atributo estu_zona de la base de datos del ICFES (ver tabla 1), que ubica la zona cardinal de la ciudad de residencia del estudiante, no aporta información relevante para este estudio. Durante el proceso se consideraron varias opciones de clasificación de dichas zonas, entre las que se cuentan, la reorganización por zonas geográficas como tal, zonas político-administrativas, zonas culturales y zonas de los Órganos Colegiados de Administración y Decisión (OCAD); estas últimas, creadas y utilizadas en el marco del Sistema General de Regalías (SGR), que busca beneficiar a todos los departamentos y municipios del país, no únicamente a aquellos en los cuales se lleva a cabo la explotación de recursos naturales no renovables, por lo cual se constituyen en una propuesta que se fundamenta no sólo con argumentos de justicia social, sino,

con la finalidad de distribuir de manera equitativa los recursos de regalías en todas las zonas del país. Todos los recursos del SGR financiarán proyectos de inversión presentados a los OCAD por las entidades territoriales, según lo reglamentado en el decreto 1075 de 2012. Los OCAD se constituyen a nivel municipal, departamental, regional, nacional y el OCAD de Ciencia Tecnología e Innovación, al cual pertenecen las universidades.

Estas zonas OCAD, en la actualidad constituyen un reordenamiento territorial del país, frente a las regiones geográficas que ha manejado históricamente y que ha permitido marcar diferencias significativas entre el centro y la periferia del país. La ventaja de esta organización, es que cada zona OCAD agrupa departamentos completos, con sus respectivos municipios y gracias a esto es posible predecir cuál sería el rendimiento académico en las competencias genéricas en estas zonas con la información de las pruebas Saber Pro y, en futuros estudios, cómo las regalías han influido en el desarrollo de las IES y en el rendimiento académico de los estudiantes de estas zonas. Los OCAD, se encuentran distribuidos en 6 zonas que cubre la totalidad del país. Teniendo en cuenta que Bogotá es la ciudad donde se concentra la gran mayoría de universidades, se consideró para esta investigación, separarla como una nueva zona. Razones por las cuales las zonas geográficas en las que se agrupan las IES, para este estudio son siete, como se detallan en la tabla 16, con los departamentos que agrupa.

Tabla 16. Valores del atributo inst_programa_zona

Clasificadas las IES por zonas, se procedió a contar el número de estudiantes e IES por zonas y crear dos nuevos atributos: num_estudiantes_zona y num_instituciones_zona, para luego discretizarlos.

Para los valores del atributo num_estudiantes_zona, se considera ALTO si el número de estudiantes es mayor que 20000, MEDIO si el número de estudiantes esta entre 10.000 y 20.000 y BAJO si el número de estudiantes es menor que 10.000. En la tabla 17 se muestra este proceso.

Tabla 17. Valores discretizados del atributo num_estudiantes_zona

Para los valores del atributo num_instituciones_zona, se considera ALTO si el número de IES es mayor que 70, MEDIO si el número de IES está entre 60 y 70 y BAJO si el número de IES es menor que 60. En la tabla 18 se muestra este proceso.

Tabla 18. Valores discretizados del atributo num_instituciones_zona

Para la nueva variable eco_condicion_hogar se tuvieron en cuenta los materiales de los pisos de la vivienda. En la tabla 19 se muestran estos valores y el material de los pisos.

Tabla 19. Valores del atributo eco_condicion_hogar

Para asignar los valores a la nueva variable eco_condicion_hogar se creó un índice con los valores de los atributos que de los electrodomésticos que esta variable reemplazó. El índice es el resultado de la sumatoria de los valores de la presencia (1) o ausencia (0) de cada electrodoméstico en la vivienda. Si el índice está entre 4 y 5 la condición del hogar es BUENA.

Si el índice está entre 2 y 3 la condición del hogar es MEDIA y finalmente, si el índice está entre 0 y 1 la condición del hogar es MALA. En la tabla 20 se muestran las variables que intervienen en el cálculo del índice condición_hogar.

Tabla 20. Cálculo del índice de condición_hogar

De igual manera se procedió con los valores de la nueva variable eco_condicion_tic para los valores BUENA, MEDIA y MALA pero teniendo en cuenta los servicios con que cuenta el hogar. En la tabla 21 se muestran los atributos que intervienen en el cálculo del índice de condición_tic.

Tabla 21. Cálculo del índice de condición_tic

Para los valores de la variable eco_condicion_transporte se consideraron dos valores: PARTICULAR si se contaba en el hogar con un automóvil y PÚBLICO si no se contaba, teniendo en cuenta la variable econ_sn_automovil que reemplazó.

Para asignar los valores a la nueva variable eco_condicion_vive se calculó el índice de hacinamiento. El hacinamiento refiere a la relación entre el número de personas que habitan una vivienda o casa y el espacio o número de cuartos disponibles (Spicker, Alvarez, & Gordon, s. f.). Generalmente se aceptan los valores: hasta 2.4 - sin hacinamiento; de 2.5 a 4.9 - hacinamiento medio y de 5.0 o más - hacinamiento crítico.

Teniendo en cuenta estos conceptos, el índice de hacinamiento para cada estudiante se obtuvo dividiendo los valores de los atributos fami_num_pers_fam entre infa_dormitorios. Variables que fueron reemplazadas por el atributo eco_condicion_vive. Los valores de este nuevo atributo se asignaron teniendo en cuenta los valores aceptados para el hacinamiento.

Para obtener los valores de la nueva variable area_grupo_referencia, se agruparon los valores del atributo estu_grupo_referencia en Ciencias Humanas, Ciencias Sociales y Ciencias Naturales y Técnicas. Este proceso se muestra en la tabla 22.

Tabla 22. Asignación de valores del atributo area_grupo_referencia

Como resultado de todos estos procesos se obtuvo un nuevo conjunto de datos, limpio y transformado, listo para aplicarle las técnicas de minería de datos denominado T97068A35. La descripción del repositorio T97068A35 con 97068 registros y 35 atributos, se muestra en la tabla 23, organizados por los factores sociodemográficos, económicos, institucionales, institucionales y clases.

Tabla 23. Diccionario de datos del repositorio final T97068A35

A partir del repositorio T97068A34 y con el fin de descubrir patrones asociados al rendimiento académico en competencias genéricas, se construyó por cada competencia (clase) un repositorio de datos. Para cada repositorio se tomaron los 31 atributos descritos en la tabla 23 y el respectivo atributo clase dependiendo de la competencia. Se tuvieron que eliminar, en cada repositorio, los registros con valores nulos en el atributo clase (competencia). La descripción de cada repositorio por competencia se muestra en la tabla 24.

Tabla 24. Repositorios por cada competencia

2.4 Modelado

En esta fase se seleccionaron las técnicas de modelado más apropiadas para el proyecto de Minería de Datos. Teniendo en cuenta que el problema a solucionar es descubrir los factores sociodemográficos, económicos, académicos e institucionales asociados al rendimiento académico en las competencias genéricas de los estudiantes que presentaron las pruebas SaberPro 2011-2, se exploraron y evaluaron tanto tareas de minería de datos predictivas como descriptivas.

Las predictivas se aplican a problemas en los que hay que predecir nuevos datos para uno o más ejemplos que van acompañados de una salida denominada clase (Hernández et al., 2004). Para este tipo de tareas se evaluó la tarea de clasificación utilizando la técnica de árboles de decisión.

Las descriptivas se aplican a problemas en los que hay que describir los valores de los datos existentes (Hernández et al., 2005). Para este tipo de tareas se evaluó la tarea de asociación utilizando el algoritmo Apriori (Agrawal & Srikant, 1994).

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

cuatro competencias genéricas, evaluadas en las pruebas SaberPro 2011-2. Con asociación se pretende identificar cuales factores sociodemográficos, económicos, académicos e institucionales están sucediendo frecuentemente juntos y que están relacionados o asociados con el rendimiento académico en las competencias genéricas de SaberPro 2011-2. Con esta tarea no fue posible obtener para la competencia de inglés, patrones asociados al desempeño académico para valores sobre la media y para la competencia de lectura crítica patrones asociados al desempeño académico para valores bajo la media. Por esta razón se escogió la tarea de clasificación basada en árboles de decisión.

2.4.1 Descubrimiento de patrones de desempeño académico con árboles de decisión.

La técnica de clasificación utilizada para el descubrimiento de patrones de desempeño académico en las competencias genéricas de las pruebas SaberPro 2011-2 fue árboles de decisión. El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender (Han & Kamber, 2001), (Sattler & Dunemann, 2001), (Timarán & Millán, 2006). 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 (Hernández & Lorente, 2009).

El algoritmo de la herramienta Weka (Hall et al., 2011) utilizado para obtener el modelo de clasificación con árboles de decisión fue J48, el cual implementa al algoritmo C.45, (Quinlan, 1993). 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 (Hernández & Lorente, 2009). El parámetro más importante que se debe tener en cuenta para la poda es 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 (García & Álvarez, 2010). Otra forma de variar el tamaño del árbol es a través del parámetro M que especifica el mínimo número de instancias o registros por nodo del árbol, si

bien es menos importante puesto que depende del número absoluto de instancias en el conjunto de datos de partida (Hall, Frank, & Witten, 2011).

Antes de construir un modelo se debe definir un procedimiento para probar la calidad del modelo y su validez. Por tanto, para entrenar y probar un modelo de clasificación, el diseño de prueba específica divide los datos en dos conjuntos: entrenamiento y prueba. Existen diferentes medidas de evaluación del clasificador en Weka (Hall et al., 2011):

-Usar el conjunto de datos de entrenamiento (*Use training set*): se emplea todo el conjunto de datos para entrenar el modelo y después se prueba (esta técnica puede ser muy buena para ese conjunto de datos pero puede ser poco precisa para nuevos datos).

-Proveer un conjunto de datos de prueba (*Supplied test set*): se emplea un conjunto de datos para entrenar y otro conjunto independiente al universo de los datos con los que se está trabajando para prueba (corriendo el riesgo que el conjunto de prueba no refleje o se corresponda con las características de los datos que se emplearon para entrenar el modelo).

-Porcentaje de Partición (*Percentage Split*): se emplea un % aleatorio de datos para entrenar y otro % para probar, este método difiere del anterior en que ambos conjuntos pertenecen al universo de datos con el que se está trabajando por lo que se elimina el riesgo que corre el anterior.

-Validación cruzada (*Cross validation*): Este mecanismo permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición (Hernández, Ramirez y Ferri, 2004). Para este caso particular se utiliza el método de evaluación validación cruzada con n pliegues (n-fold cross validation). Esta es la opción por defecto y la más comúnmente utilizada. 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*) (Hernández et al., 2004).

Por otra parte, es bastante sencillo evaluar o estimar el coste de un clasificador para un determinado conjunto de ejemplos si se dispone 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) (Fernández, 2009).

Teniendo en cuenta los parámetros de evaluación anteriores y los repositorios de datos descritos en la tabla 24 se procedió a construir los diferentes árboles de decisión con el algoritmo J48. Se escogió como clase el desempeño en la competencia respectiva y para evaluar la calidad del modelo y su validez se escogió el método de validación cruzada y específicamente la validación cruzada con 10 pliegues por los mejores resultados que se obtienen.

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=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%.

Figura 1. Árbol para lectura crítica podado

Figura 2. Precisión y matriz de confusión del árbol de lectura crítica

Figura 3. Reglas más representativas de lectura crítica

2.4.1.1 Descubrimiento de patrones de desempeño en lectura crítica.

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en la competencia de lectura crítica de los estudiantes de programas profesionales que presentaron las pruebas SaberPro 2011-2, se utilizó el conjunto de datos T97055A32LEC. El árbol construido

con los parámetros M=971 (1%) y C=0.25 para la prepoda y confianza mayor o igual al 60% y soporte mayor o igual al 0.5% se muestra en la figura 1. En la figura 2 se muestra la precisión del árbol y su matriz de confusión y en la figura 3 se muestran las reglas más representativas con un soporte>=0.5% y confianza>=60%.

2.4.1.2 Descubrimiento de patrones de desempeño en comunicación escrita

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en la competencia de comunicación escrita de los estudiantes de programas profesionales que presentaron las pruebas SaberPro 2011-2, se utilizó el conjunto de datos T95337A32ESC. El árbol construido con los parámetros M=954 (1%) y C=0.25 para la prepoda y confianza mayor o igual al 60% y soporte mayor o igual al 0.5% se muestra en la figura 4. En la figura 5 se muestra la precisión del árbol y su matriz de confusión y en la figura 6 se muestran las reglas más representativas con un soporte>=0.5% y confianza>=60%.

2.4.1.3 Descubrimiento de patrones de desempeño en razonamiento cuantitativo

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en la competencia de razonamiento cuantitativo de los estudiantes de programas profesionales que presentaron las pruebas SaberPro 2011-2, se utilizó el conjunto de datos T97057A32CUA. El árbol construido con los parámetros M=971 (1%) y C=0.25 para la prepoda y confianza mayor o igual al 60% y soporte mayor o igual al 0.5% se muestra en la figura 7. En la figura 8 se muestra la precisión del árbol y su matriz de confusión y en la figura 9 se muestran las reglas más representativas con un soporte>=0.5% y confianza>=60%.

Figura 4. Árbol para comunicación escrita podado

Figura 5. Precisión y matriz de confusión del árbol de comunicación escrita

Figura 6. Reglas más representativas de comunicación escrita

Figura 7. Árbol para razonamiento cuantitativo podado

Figura 8. Precisión y matriz de confusión del árbol de razonamiento cuantitativo

Figura 9. Reglas más representativas de razonamiento cuantitativo

2.4.1.4 Descubrimiento de patrones de desempeño en inglés

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en la competencia de inglés de los estudiantes de programas profesionales que presentaron las pruebas SaberPro 2011-2, se utilizó el conjunto de datos T96946A32ING. El árbol construido con los parámetros M=970 (1%) y C=0.25 para la prepoda y confianza mayor o igual al 60% y soporte

mayor o igual al 0.5% se muestra en la figura 10. En la figura 11 se muestra la precisión del árbol y su matriz de confusión y en la figura 12 se muestran las reglas más representativas con un soporte>=0.5% y confianza>=60%.

Figura 10. Árbol para inglés podado

Figura 11. Precisión y matriz de confusión del árbol de inglés

Figura 12. Reglas más representativas de inglés

2.5 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 el capítulo de resultados.

2.6 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 de los estudiantes de programas profesionales en las Pruebas SaberPro, será posible analizar los resultados y determinar sus efectos.

3. RESULTADOS

En esta sección se evalúan los modelos de clasificación basados en árboles de decisión obtenidos por cada competencia y se interpretan los resultados obtenidos en la etapa de modelamiento.

3.1 Evaluación e interpretación de resultados para la competencia lectura crítica

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 2).

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 (ver figura 3), los patrones más representativos descubiertos, teniendo en cuenta el soporte y la confianza, son:

Regla 2. 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 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 OCAD a la que pertenece la institución está entre 60 y 70, el número de estudiantes de programas profesionales en la zona OCAD 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 4. 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 7. 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 9. 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 OCAD 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 8. 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 OCAD 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 12. 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 OCAD 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.

3.2 Evaluación e interpretación de resultados para la competencia comunicación escrita

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos T95337A32ESC (ver figura 4), en el cual se almacenan los datos válidos sobre los factores sociodemográficos, económicos, académicos e institucionales de 95.337 estudiantes de programas profesionales que presentaron las pruebas SaberPro 2011-2 en la competencia genérica de comunicación escrita, donde se escogió el atributo *mod_comunica_escrita_desemp* como clase, se puede observar que este clasifica correctamente a 56.669 instancias, que corresponde a un porcentaje de precisión del 59,4% y 38.668 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 40,5% (ver figura 5).

Teniendo en cuenta la distribución de los valores del atributo clase *mod_comunica_escrita_desemp* del repositorio T95337A32ESC que es de 46.986 registros para el valor "sobre la media" y 48.351 registros para el valor "bajo la media" y evaluando el modelo con la matriz de confusión de la figura 5, este clasifica correctamente a 25.435 casos de estudiantes cuyos resultados en comunicación escrita están sobre la media y a 31.234 casos que están bajo la media. Por otra parte, clasifica incorrectamente a 21.551 casos de estudiantes cuyos resultados en comunicación escrita están sobre la media y 17.117 casos que están bajo la media. Esto significa que el modelo clasifica correctamente al 54,13 % de los estudiantes que están sobre la media en comunicación escrita y al 64,6% de las estudiantes que están bajo la media en esta competencia.

De acuerdo a las reglas de clasificación obtenidas con este modelo (ver figura 6), los patrones más representativos descubiertos, teniendo en cuenta el soporte y la confianza, son:

Regla 3. Si la IES no está acreditada y la metodología de formación es a distancia entonces el desempeño de los estudiantes en comunicación escrita tiene mayor probabilidad de estar bajo la media. El 12.36% de los 95.337estudiantes evaluados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 62.42% de los 11.787 (12.36%) estudiantes se clasifican correctamente y el 24.38% de los 48.351 estudiantes que están bajo la media cumplen este patrón.

Regla 1. Si la IES no está acreditada, la metodología de formación es presencial, los estudiantes son de género masculino y la zona OCAD en la que se ofrece el programa es la

Caribe entonces el desempeño de los estudiantes en comunicación escrita tiene mayor probabilidad de estar bajo la media. El 5.44% de los 95.337 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 64.92% de los 5.188 (5.44%) estudiantes se clasifican correctamente y el 10.73% de los 48.351 estudiantes que están bajo la media cumplen este patrón.

Regla 5. Si la IES está acreditada y el número de estudiantes de programas profesionales en la zona OCAD en la que se ofrece el programa es mayor que 20.000 entonces el desempeño de los estudiantes en comunicación escrita tiene mayor probabilidad de estar sobre la media. El 7.82% de los 95.337 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 71.06% de los 7.458 (7.82%) estudiantes se clasifican correctamente y el 15.87% de los 46.986 estudiantes que están sobre la media cumplen este patrón.

Regla 6. Si la IES está acreditada y el número de estudiantes de programas profesionales en la zona OCAD en la que se ofrece el programa está entre 10.000 y 20.000 y los estudiantes son de género femenino entonces el desempeño de los estudiantes en comunicación escrita tiene mayor probabilidad de estar sobre la media. El 5.04% de los 95.337 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 62.67% de los 4.801 (5.04%) estudiantes se clasifican correctamente y el 10.22% de los 46.986 estudiantes que están sobre la media cumplen este patrón.

Regla 8. Si la IES no está acreditada, la metodología de formación es presencial, los estudiantes son de género femenino y el programa es acreditado entonces el desempeño de los estudiantes en comunicación escrita tiene mayor probabilidad de estar sobre la media. El 4.32% de los 95.337 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 60.51% de los 4.115 (4.32%) estudiantes se clasifican correctamente y el 8.76 % de los 46.986 estudiantes que están sobre la media cumplen este patrón.

3.3 Evaluación e interpretación de resultados para la competencia razonamiento cuantitativo

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos T97057A32CUA (ver figura 7), en el cual se almacenan los datos válidos sobre los factores sociodemográficos, económicos, académicos e institucionales de 97.057 estudiantes de

programas profesionales que presentaron las pruebas SaberPro 2011-2 en la competencia genérica de razonamiento cuantitativo, donde se escogió el atributo *mod_razona_cuantitativo_desemp* como clase, se puede observar que este clasifica correctamente a 64.549 instancias, que corresponde a un porcentaje de precisión del 66,5% y 32.508 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 33,5% (ver figura 8).

Teniendo en cuenta la distribución de los valores del atributo clase *mod_razona_cuantitativo_desemp* del repositorio T97057A32CUA que es de 48.144 registros para el valor "sobre la media" y 48.913 registros para el valor "bajo la media" y evaluando el modelo con la matriz de confusión de la figura 8, este clasifica correctamente a 29.093 casos de estudiantes cuyos resultados en razonamiento cuantitativo están sobre la media y a 35.456 casos que están bajo la media. Por otra parte, clasifica incorrectamente a 19.051 casos de estudiantes cuyos resultados en razonamiento cuantitativo están sobre la media y 13.457 casos que están bajo la media. Esto significa que el modelo clasifica correctamente al 60,43 % de los estudiantes que están sobre la media en razonamiento cuantitativo y al 72,49% de las estudiantes que están bajo la media en esta competencia.

De acuerdo a las reglas de clasificación obtenidas con este modelo (ver figura 9), los patrones más representativos descubiertos, teniendo en cuenta el soporte y la confianza, son:

Regla 1. Si la IES no está acreditada, el área del grupo de referencia del programa es Ciencias Sociales y la metodología de formación es a distancia entonces el desempeño de los estudiantes en razonamiento cuantitativo tiene mayor probabilidad de estar bajo la media. El 10.23% de los 97.057 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 73.88% de los 9.927 (10.23%) estudiantes se clasifican correctamente y el 20.30% de los 48.913 estudiantes que están bajo la media cumplen este patrón.

Regla 5. Si la IES no está acreditada, el área del grupo de referencia del programa es Ciencias Sociales, la metodología de formación es presencial y el área del conocimiento al que pertenece el programa es Ciencias dela Educación entonces el desempeño de los estudiantes en razonamiento cuantitativo tiene mayor probabilidad de estar bajo la media. El 7.24% de los 97.057 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de

esta manera. El 70.66% de los 7.029 (7.24%) estudiantes se clasifican correctamente y el 14.37% de los 48.913 estudiantes que están bajo la media cumplen este patrón.

Regla 6. Si la IES no está acreditada, el área del grupo de referencia del programa es Ciencias Sociales, la metodología de formación es presencial y el área del conocimiento al que pertenece el programa es Ciencias Sociales y Humanas entonces el desempeño de los estudiantes en razonamiento cuantitativo tiene mayor probabilidad de estar bajo la media. El 6.13% de los 97.057 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 68.35% de los 5.953 (6.13%) estudiantes se clasifican correctamente y el 12.17% de los 48.913 estudiantes que están bajo la media cumplen este patrón.

Regla 8. Si la IES no está acreditada, el área del grupo de referencia del programa es Ciencias Humanas, la institución es de carácter privado y los ingresos familiares de los estudiantes es de tres salarios mínimos entonces el desempeño de los estudiantes en razonamiento cuantitativo tiene mayor probabilidad de estar bajo la media. El 4.29% de los 97.057 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 66.54% de los 4.166 (4.29%) estudiantes se clasifican correctamente y el 8.52% de los 48.913 estudiantes que están bajo la media cumplen este patrón.

Regla 17. Si la IES no está acreditada, el área del grupo de referencia del programa es Ciencias Naturales y Técnicas, la metodología de formación es presencial, el programa no es acreditado y el género de los estudiantes es masculino entonces el desempeño de los estudiantes en razonamiento cuantitativo tiene mayor probabilidad de estar sobre la media. El 9.26% de los 97.057 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 70.52% de los 8.988 (9.26%) estudiantes se clasifican correctamente y el 18.67% de los 48.144 estudiantes que están sobre la media cumplen este patrón.

Regla 12. Si la IES está acreditada y el área del conocimiento al que pertenece el programa es Ingeniería, Arquitectura, Urbanismo y Afines entonces el desempeño de los estudiantes en razonamiento cuantitativo tiene mayor probabilidad de estar sobre la media. El 4.21% de los 97.057 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 90.05% de los 4.090 (4.21%) estudiantes se clasifican correctamente y el 8.50% de los 48.144 estudiantes que están sobre la media cumplen este patrón.

Regla 16. Si la IES está acreditada y el área del conocimiento al que pertenece el programa es Economía, Administración, Contaduría y Afines entonces el desempeño de los estudiantes en razonamiento cuantitativo tiene mayor probabilidad de estar sobre la media. El 3.91% de los 97.057 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 76.37% de los 3.791 (3.91%) estudiantes se clasifican correctamente y el 7.87% de los 48.144 estudiantes que están sobre la media cumplen este patrón.

3.3 Evaluación e interpretación de resultados para la competencia inglés

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos T96946A32ING (ver figura 10), en el cual se almacenan los datos válidos sobre los factores sociodemográficos, económicos, académicos e institucionales de 96.946 estudiantes de programas profesionales que presentaron las pruebas SaberPro 2011-2 en la competencia genérica de inglés, donde se escogió el atributo *mod_ingles_desemp* como clase, se puede observar que este clasifica correctamente a 69.600 instancias, que corresponde a un porcentaje de precisión del 71,8% y 27.346 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 28,2% (ver figura 11).

Teniendo en cuenta la distribución de los valores del atributo clase *mod_ingles_desemp* del repositorio T96946A32ING que es de 40.037 registros para el valor “sobre la media” y 56.909 registros para el valor “bajo la media” y evaluando el modelo con la matriz de confusión de la figura 11, este clasifica correctamente a 20.764 casos de estudiantes cuyos resultados en inglés están sobre la media y a 48.836 casos que están bajo la media. Por otra parte, clasifica incorrectamente a 19.273 casos de estudiantes cuyos resultados en inglés están sobre la media y 8073 casos que están bajo la media. Esto significa que el modelo clasifica correctamente al 51.86% de los estudiantes que están sobre la media en inglés y al 85,81% de las estudiantes que están bajo la media en esta competencia.

De acuerdo a las reglas de clasificación obtenidas con este modelo (ver figura 12), los patrones más representativos descubiertos, teniendo en cuenta el soporte y la confianza, son:

Regla 3. Si la IES no está acreditada, la metodología de formación es presencial y los estudiantes están clasificados en el nivel 2 del SISBEN entonces el desempeño de los estudiantes en inglés tiene mayor probabilidad de estar bajo la media. El 14.39% de los 96.946 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta

manera. El 74.31% de los 13.951 (14.39%) estudiantes se clasifican correctamente y el 24.51% de los 56.909 estudiantes que están bajo la media cumplen este patrón.

Regla 1. Si la IES no está acreditada y la metodología de formación es a distancia entonces el desempeño de los estudiantes en inglés tiene mayor probabilidad de estar bajo la media. El 12.45% de los 96.946 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 87.16% de los 12.074 (12.45%) estudiantes se clasifican correctamente y el 21.22% de los 56.909 estudiantes que están bajo la media cumplen este patrón.

Regla 2. Si la IES no está acreditada, la metodología de formación es presencial y los estudiantes están clasificados en el nivel 1 del SISBEN entonces el desempeño de los estudiantes en inglés tiene mayor probabilidad de estar bajo la media. El 10.69% de los 96.946 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 79.66% de los 10.366 (10.69%) estudiantes se clasifican correctamente y el 18.22% de los 56.909 estudiantes que están bajo la media cumplen este patrón.

Regla 4. Si la IES no está acreditada, la metodología de formación es presencial y los estudiantes no están clasificados en el SISBEN y los estudiantes tienen personas a cargo entonces el desempeño de los estudiantes en inglés tiene mayor probabilidad de estar bajo la media. El 8.90% de los 96.946 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 67.79% de los 8.627 (8.90%) estudiantes se clasifican correctamente y el 18.22% de los 56.909 estudiantes que están bajo la media cumplen este patrón.

Regla 10. Si la IES está acreditada, y los estudiantes no tienen personas a cargo entonces el desempeño de los estudiantes en inglés tiene mayor probabilidad de estar sobre la media. El 15.0% de los 96.946 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 81.72% de los 14.539 (15.0%) estudiantes se clasifican correctamente y el 36.31% de los 40.037 estudiantes que están sobre la media cumplen este patrón.

Regla 12. Si la IES no está acreditada, la metodología de formación es presencial, los estudiantes no están clasificados en el SISBEN, no tienen personas a cargo, son de estrato 4 y el número de estudiantes de la zona en la cual se ofrece el programa es mayor que 20.000 entonces el desempeño de los estudiantes en inglés tiene mayor probabilidad de estar sobre la media. El

2.30% de los 96.946 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 68.18% de los 2.232 (2.30%) estudiantes se clasifican correctamente y el 5.57% de los 40.037 estudiantes que están sobre la media cumplen este patrón.

Regla 13. Si la IES no está acreditada, la metodología de formación es presencial, los estudiantes no están clasificados en el SISBEN, no tienen personas a cargo y son de estrato 5 entonces el desempeño de los estudiantes en inglés tiene mayor probabilidad de estar sobre la media. El 2.05% de los 96.946 estudiantes analizados en esta competencia en la prueba SaberPro 2011-2 se clasifican de esta manera. El 68.64% de los 1.990 (2.05%) estudiantes se clasifican correctamente y el 4.97% de los 40.037 estudiantes que están sobre la media cumplen este patrón.

4. DISCUSION

El objetivo de esta investigación fue "descubrir patrones de desempeño académico en competencias genéricas de los estudiantes de programas profesionales en las pruebas Saber Pro 2011-2, a partir de los datos sociodemográficos, económicos, académicos e institucionales almacenados en las bases de datos del ICFES, utilizando técnicas de minería de datos", el cual se logró en su totalidad.

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; este procedimiento se realizó para cada una de las cuatro competencias genéricas. En los patrones descubiertos, se destacan la acreditación institucional y la modalidad de estudio como dos atributos importantes asociados al desempeño académico de los estudiantes en las pruebas SaberPro, en las cuatro competencias genéricas.

En efecto, de acuerdo a los modelos de clasificación basados en árboles de decisión, construidos para cada competencia genérica, muestran que el patrón de IES acreditada se asocia con el buen desempeño académico de los estudiantes de programas profesionales en las cuatro competencias genéricas objeto de estudio; resultado que está en línea con la fundamentación conceptual expresada en la política pública sobre calidad de la educación superior, particularmente, sobre el tema de acreditación institucional; pues de hecho, el cumplimiento de los doce factores de calidad señalados por el CNA (2013), dan cuenta de las condiciones de

favorabilidad de todos los procesos al interior de las IES, que permiten un mejor desempeño académico de los estudiante.

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 (2011), 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 las competencias genéricas de las pruebas Saber Pro 2011-2.

La modalidad del programa, según resultados de la investigación de Parra et al. (2009), realizada con los estudiantes de primer semestre de pregrado de la Facultad de Ingeniería de la Universidad de Antioquia, 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 (2008), 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 (2007), 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 (2009), 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 (2002), 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 (2011), 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.

Particularmente para cada competencia se encuentra que la acreditación institucional y la modalidad de estudio se asocian al buen o bajo desempeño académico de los estudiantes que presentaron las pruebas Saber Pro 2011-2.

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 económica), 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 (2007), 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 (2000), 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.

Analizando el modelo de la figura 4, en la competencia de comunicación escrita el buen desempeño se asocia al hecho que la IES esté acreditada y que haya un alto número de estudiantes en la zona geográfica de la IES. En un orden inferior de jerarquía, se encuentra que también tienen buen desempeño las mujeres en general y los hombres que estudien carreras de las Ciencias Humanas. En el caso de que la IES no esté acreditada y la modalidad sea presencial, también es un patrón de buen desempeño, el hecho de que el programa sea acreditado.

Por su parte, el bajo desempeño en comunicación escrita tiene relación con IES no acreditada y la modalidad a distancia. En el caso de la modalidad presencial, el bajo desempeño se asocia con el género masculino de las zonas geográficas Caribe, Centro Sur y Pacífico.

En este contexto, están los hallazgos de Gómez y Soares (2013), quienes en su análisis sobre la diferencia de género en relación con el desempeño académico, expresan que tales diferencias no son significativas; que no se puede afirmar de modo definitivo que exista una relación directa entre el rendimiento académico y el sexo; sin embargo, a decir de Garbanzo (2007), las mujeres si tienen un ligero mejor desempeño académico que los hombres; y en el mismo sentido, Montero y Villalobos (2004), señalan que existe una relación significativa entre el sexo y el desempeño académico, a favor de las mujeres.

Analizando el modelo de la figura 7, en la competencia de razonamiento cuantitativo el buen desempeño se asocia al hecho que la IES esté acreditada y que el área del grupo de referencia sea Ciencias Naturales y Técnicas. En un orden inferior de jerarquía, se encuentra que también tienen buen desempeño los estudiantes de IES no acreditadas, de modalidad presencial y de Ciencias

Naturales y Técnicas. Por su parte, el bajo desempeño en la competencia de razonamiento cuantitativo tiene relación con IES no acreditadas y el área del grupo de referencia de Ciencias Sociales y Humanas.

En efecto, los estudiantes de las áreas de ciencias sociales y humanas, y ciencias de la educación, presentan bajo desempeño en esta competencia; realidad que es reconocida por las IES, a tal punto, que como alternativa de solución vienen implementando cursos adicionales a sus planes de estudio, con el fin de proporcionar una formación básica en matemáticas, en áreas que, al margen de las Pruebas SaberPro, habría sido impensable la implementación de dichos cursos. No obstante, lo ideal sería que aquellos cursos fueran constitutivos de los currículos de todos los programas de formación profesional, en la perspectiva de recuperar su aporte a la formación y estructuración del pensamiento matemático, de modo que, como señalan Davidson y McKinney (2001), se desarrolle la competencia para la interpretación e inferencia a partir de modelos; la representación de la información matemática de diversas formas; la utilización de métodos aritméticos, algebraicos, geométricos y estadísticos para solucionar problemas; la estimación y chequeo de las respuestas a problemas matemáticos con miras a establecer su racionabilidad, identificar alternativas y seleccionar los mejores resultados; y el reconocimiento de los límites de los métodos matemáticos y estadísticos, en la vía de contribuir al fomento y desarrollo de la competencia de razonamiento cuantitativo.

Es en este sentido que el MEN (s.f), concibe la alfabetización cuantitativa como el dominio del lenguaje y los métodos matemáticos como indispensables para lograr la comprensión de problemas y discusiones cotidianas contemporáneas, sin lo cual las personas se encontrarían parcialmente incomunicadas. En la misma línea, haciendo referencia al pensamiento matemático como una competencia genérica emergente, los autores Toro y Villaveces (2008, 1), plantean que "podemos pensar que hay un sustrato mínimo de competencias matemáticas, sin el cual un profesional puede quedar parcialmente incomunicado del mundo moderno que lo rodea"; afirmación que hace notar que la competencia del razonamiento cuantitativo es indispensable y, por lo mismo, inaplazable para todo profesional.

A pesar del reconocimiento de la importancia del razonamiento cuantitativo en la formación de todo profesional, se encuentran currículos de programas de humanidades, ciencias sociales y afines, atienden poco o nada su fomento y desarrollo, hecho que podría asociarse con el bajo

desempeño académico en la competencia de razonamiento cuantitativo, tal como se encontró en esta investigación.

Analizando el modelo de la figura 10, en la competencia de inglés el buen desempeño se asocia al hecho que la IES esté acreditada y que no tenga personas a cargo. En el caso de que la IES no esté acreditada y la modalidad sea presencial, también es un patrón de buen desempeño, el hecho de que los estudiantes pertenezcan a estratos altos (4, 5 y 6).

Por su parte, el bajo desempeño en inglés tiene relación con IES no acreditada y la modalidad a distancia. En el caso de la modalidad presencial, el bajo desempeño se asocia con niveles SISBEN bajos (1 y 2) y con estratos socioeconómicos bajos (1 y 2).

Estos hechos coinciden con Garbanzo (2007), Seibold (2000) y Montero & Villalobos (2004), 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, en cuanto que forma parte de los patrones descubiertos, particularmente en la competencia genérica de inglés.

En resumen, en los patrones de desempeño académico de las competencias genéricas, descubiertos y considerados en esta discusión, la acreditación institucional se encuentra asociada con otros atributos, tanto en el desempeño por encima de la media como por debajo de ella, excepto en lectura crítica, en la cual constituye en sí mismo, un patrón de desempeño académico por encima de la media. 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.

Por otra parte, también se encontró que existe asociación para el buen desempeño académico en las competencias genéricas de lectura crítica y comunicación escrita, entre el alto número de estudiantes e instituciones por zona y otros atributos. En este contexto conviene destacar, que las variables denominadas número de IES y número de estudiantes por zona, se crearon en el marco de esta investigación, y se constituyen en atributos importantes para el desarrollo de la educación superior en las zonas y regiones a las cuales pertenecen; por lo cual, deben ser consideradas en

las políticas públicas de mejoramiento de la calidad en la formación de profesionales y su cobertura.

Así mismo, el área de los grupos de referencia de los programas académicos, está asociada con el buen y mal desempeño en la competencia genérica de razonamiento cuantitativo. Otros atributos en los patrones descubiertos en esta investigación asociados al buen y mal desempeño académico en las competencias de lectura crítica e inglés, es el estrato socioeconómico y el nivel SISBEN al que pertenece el estudiante

Finalmente, otro de los hallazgos de este estudio, es la identificación del atributo género en los patrones de desempeño académico. Así pues, el género femenino se asocia, junto con otros atributos, al buen desempeño en la competencia de comunicación escrita; y el género masculino con un bajo desempeño; no obstante, se destaca que tales diferencias no son significativas.

Teniendo en cuenta los resultados de esta investigación, 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.

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. De igual manera, el desconocimiento del documento de identificación del estudiante, solicitado al ICFES, dificultó la obtención de otros posibles atributos que se podrían obtener o complementar de las bases de datos de la Registraduría Nacional del Estado Civil y del SISBEN.

REFERENCIAS

Agrawal, R. & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In: 20th International Conference on Very Large Data Bases, VLDB 1994, (12-15/09/1994). Santiago de Chile (Chile): VLDB. Proceedings. p. 487-499. ISBN: 1-55860-153-8.

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

Bodensiek, A. (2010). Estudio sobre los factores que influyen en el rendimiento escolar. Secretaria de Educación Municipal.

Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2004). Deserción estudiantil universitaria: una aplicación de modelos de duración. *Lecturas de economía*, *60*(60), 39–65.

Climent-Bonilla, J. B. (2014). Origen, desarrollo y declive de las competencias individuales en tiempos de incertidumbre. *Educación y Educadores*, *17*(1). Recuperado a partir de http://educacionyeducadores.unisabana.edu.co/index.php/eye/article/view/3499

Cohen, J. (1988). Análisis de poder estadístico para las Ciencias del comportamiento (2ª ed.). Nueva Jersey: Lawrence Erlbaum.

Congreso de Colombia. Ley 30 (1992). Por la cual se organiza el servicio de la educación superior. Disponible en: http://www.bienestarbogota.unal.edu.co/pdf/egresados/ley30de1992.pdf

Colombia. Consejo Nacional de Acreditación – CNA. (2013). Lineamientos para la acreditación de programas de pregrado. Bogotá. CNA.

Colombia. Consejo Nacional de Acreditación – CNA. (s. f.). Acreditación Institucional. Disponible en: http://www.cna.gov.co/1741/article-186371.html

Colombia. Presidencia de la República. (2012). Decreto 1075 -2012. Por el cual se reglamenta la organización y funcionamiento de los Órganos Colegiados de Administración y Decisión y las secretarías técnicas, de acuerdo con lo establecido en el artículo 6° de la Ley número 1530 de 2012. Disponible en: http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=47530

Colombia. Ministerio de Educación Nacional MEN. (s. f.). Propuesta de lineamientos para la formación por competencias en educación superior. Disponible en: http://www.mineducacion.gov.co/1621/articles-261332_archivo_pdf_lineamientos.pdf

Davidson, M. y McKinney, G. (2001), "Quantitative Reasoning: An Overview". Western Washington University. Disponible en: http://www.wwu.edu/vpue/documents/issue8.pdf

Di Gresia, L., & Porto, A. (2004). Características y Rendimiento de Estudiantes Universitarios. El Caso de la Facultad de Ciencias Económicas de la Universidad Nacional de La Plata. Recuperado a partir de http://sedici.unlp.edu.ar/handle/10915/3505

Fernández, G. (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

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

Garbanzo, G. (2014). Factores asociados al rendimiento académico tomando en cuenta el nivel socioeconómico: Estudio de regresión múltiple en estudiantes universitarios. *Revista Electrónica Educare*, *18*(1), 119-154.

García, M., & Álvarez, A. (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

Gómez, G. Soares, A. (2013). Diferencias de género con relación al desempeño académico en estudiantes de nivel básico. *Alternativas en Psicología*, XVII, 28, 106-118.

González, J., & Wagenaar, R. (2003). Structures in Europe". Informe Final Proyecto Piloto-Fase 1. Informe realizado por más de 100 universidades, coordinado por la Universidad de Deusto (España) y la Universidad de Groningen (Países Bajos) y apoyado por la Comisión Europea. *Editado por Julia González y Robert Wagenaar.-Bilbao*.

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

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

Hernández, E., & Lorente, R. (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

Hernández, I., & Ojeda, B. (2009). Competencias: un reto en la formación de profesionales. *Memorias*, *5*(12), 102–119.

Hernández, J., Ramírez, M. J., & Ferri, C. (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

Instituto Colombiano para el Fomento de la Educación Superior –ICFES. (2012a). Examen Saber Pro, junio de 2012–I. Módulos de competencias genéricas y especificas disponibles. Evaluación de la calidad de la educación superior. ICFES. Recuperado a partir de www.icfes.gov.co/examenes/.../151-saber-pro-modulos-de-competencias

Instituto Colombiano para el Fomento de la Educación Superior –ICFES. (2012b). Saber Pro: Principales resultados en Competencias Genéricas. Santa Marta, Colombia. ICFES. Recuperado a partir de www.icfes.gov.co/examenes/.../151-saber-pro-modulos-de-competencias

Instituto Colombiano para el Fomento de la Educación Superior –ICFES. (2010). Exámenes de Estado de calidad de la educación superior. Saber Pro (ECAES). Análisis de resultados del período 2004-2008. ICFES. Recuperado a partir de http://54.208.2.57/datos/Informe%20final%20resultados%20SABER%20PRO%202004%22008.pdf

Instituto Colombiano para el Fomento de la Educación Superior –ICFES. (2011). Lineamientos Saber Pro. Bogotá. ICFES. Recuperado a partir de http://aprendeenlinea.udea.edu.co/lms/moodle/file.php/532/Lineamientos_SABER_PRO_2011_2_30_08_1_.pdf

Instituto Colombiano para la Evaluación de la Educación -ICFES. (2011b). Examen saber pro noviembre de 2011-2. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior. ICFES. Recuperado a partir de http://acofartes.org.co/docsweb/documento/ICFES%202011,%20M%C3%93DULOS%20COMPETENCIAS%20GEN%C3%89RICAS%20Y%20ESPEC%C3%8DFICAS.pdf

Instituto Colombiano para la Evaluación de la Educación -ICFES. (2013). Estructura del Saber Pro. ICFES. Recuperado a partir de http://ead.uis.edu.co/empresarial/images/stories/doc/SABER_PRO_2013_ESTRUCTURA.pdf

Lerner, J. (2012). *Rendimiento académico de los estudiantes de pregrado de la Universidad EAFIT*. perspectiva cuantitativa.

Lerner, J., & Montes, I. (2012). Rendimiento académico de los estudiantes de pregrado de la Universidad EAFIT*. *Cuadernos de Investigación*, (91). Recuperado a partir de http://publicaciones.eafit.edu.co/index.php/cuadernos-investigacion/article/view/1254

Martínez Martínez, A., Cegarra Navarro, J., & Rubio Sánchez, J. A. (2012). Aprendizaje basado en competencias, una propuesta para la autoevaluación docente. Recuperado a partir de http://digibug.ugr.es/handle/10481/23026

Ministerio de Educación Nacional. (2009). Decreto 3963 del 14 de octubre de 2009, por el cual se reglamenta el Examen de Estado de Calidad de la Educación Superior. Ministerio de Educación Nacional. Recuperado a partir de www.mineducacion.gov.co/1621/article-205955.html

Mizala, A., Romaguera, P., & Reinaga, T. (2002). *Factores que inciden en el rendimiento escolar en Bolivia*. Recuperado a partir de http://ideas.repec.org/p/edj/ceauch/61.html

Moncada, L. y Rubio M. (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.

Montero, E., Villalobos, J. (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.

Orea, S. V., Vargas, A. S., & Alonso, M. G. (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.

Parra, C., Mejía, L., Valencia, A., Castañeda, E., Restrepo, G., Usuga, O. y Mendoza, R. (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.

Peña, L. (2008). La competencia oral y escrita en la educación superior. Recuperado a partir de http://mineducacion.gov.co/1621/articles-189357_archivo_pdf_comunicacion.pdf

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

Quinlan, J. R. (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

Robles, N. (2013). Desarrollo de Competencias Genéricas Mediante el Aprendizaje Basado en Problemas. Eleventh LACCEI Latin American and Caribbean Conference for Engineering and Technology (LACCEI'2013) "Innovation in Engineering, Technology and Education for Competitiveness and Prosperity" .August 14 - 16, 2013. Cancun, Mexico.

Rubio, M. J. (2009). Nuevas Orientaciones y Metodología para la Educación a Distancia Loja - Ecuador: Editorial de la Universidad Técnica Particular de Loja.

Salas, R. (2000). Educación Médica Superior. La calidad en el desarrollo profesional: avances y desafíos. Educ Med Super, 14(2). Recuperado de http://scielo.sld.cu/scielo.php?pid=S0864-21412000000200003&script=sci_arttext

Sattler, K.-U., & Dunemann, O. (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

Sarramona, J. (2002). Evaluación de programas de educación a distancia. Barcelona: Universidad Autónoma de Barcelona.

Seibold, J. (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

Spicker, P., Alvarez, S., & Gordon, D. (s. f.). hH: Hacinamiento Hambruna. Recuperado a partir de http://biblioteca.clacso.edu.ar/gsdl/collect/clacso/index/assoc/D9393.dir/h.pdf

Timarán, R., Calderón, A., & Jiménez, J. (2013a). 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

Timarán, R., Calderón, A., & Jiménez, J. (2013b). 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

Timarán, R., & Millán, M. (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

Tonconi, J. (2010). Factores que influyen en el rendimiento académico y la deserción de los estudiantes de la facultad de Ingeniería Económica de la UNA-PUNO. *Revista electrónica de Educación y Desarrollo*, *2*(11).

Toro, J. y Villaveces, J. (2008). El pensamiento matemático: una competencia genérica emergente. Disponible en: http://www.mineducacion.gov.co/1621/articles-189357_archivo_pdf_matematica_1B.pdf),

Universidad Nacional de Colombia –UNAL. (2012). Análisis de los resultados obtenidos por la Universidad Nacional de Colombia sede Bogotá en las pruebas Saber Pro 2011–2. Universidad Nacional de Colombia –UNAL. Recuperado a partir de www.unal.edu.co/diracad/evaluacion/SaberPro_2012/analisis_de_resultados.pdf

Valera, R. (2010). El proceso de formación del profesional en la educación superior basado en competencias: el desafío de su calidad, en busca de una mayor integralidad de los egresados. *Civilizar*, 117.

Valero Orea, S. (2009). Aplicación de técnicas de minería de datos para predecir deserción. *Aplicación de técnicas de minería de datos para predecir deserción. México*.

Valero, S., Salvador, A., & García, M. (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.

Zapata, L. (2011). Factores Académicos asociados al bajo rendimiento en Inglés en la Pruebas ECAES presentada por los estudiantes de la Facultad de Educación en el año 2009. Fundación Universitaria Luis Amigó, Facultad de Educación, Licenciatura en Educación Básica con Énfasis en Inglés. Medellín, Colombia.

Tabla 1. Diccionario de datos del repositorio inicial T97068A94

No	Atributo	Descripción	Valores
1	estu_consecutivo	Código identificador del estudiante	
2	estu_cod_aplicacion	Nombre de la aplicación	EK+año-semestre
3	estu_genero	Género.	M - Masculino F - Femenino
4	estu_nacimiento_dia	Día de nacimiento del estudiante.	
5	estu_nacimiento_mes	Mes de nacimiento del estudiante	
6	estu_nacimiento_anno	Año de nacimiento del estudiante	
7	estu_pais_reside	País de residencia	
8	estu_estado_civil	Estado civil	Ver tabla 1a
9	estu_disc_invidente	Discapacidad Invidente	
10	estu_disc_sordo_con_interprete	Discapacidad Sordo - requiere intérprete	
11	estu_disc_sordo_sin_interprete	Discapacidad Sordo - no requiere intérprete	
12	estu_disc_motriz	Discapacidad Motriz	
13	estu_disc_sordo_ceguera	Discapacidad sordo-ceguera.	
14	estu_zona	Código identificador de la zona de la ciudad donde reside	
15	estu_anno_egreso	Año en que salió de bachiller.	
16	econ_area_vive	Área donde vive	(1) Cabecera municipal (2) Área rural
17	estu_reside_codmpio	Código DANE del municipio de residencia.	Códigos DANE
18	estu_exam_codmpio_presentacion	Código del Municipio donde presentó el examen	Códigos DANE
19	estu_exam_mpio_presentacion	Nombre municipio donde presentó el examen	
20	estu_exam_dpto_presentacion	Nombre del departamento donde presentó el examen	
21	estu_exam_cod	Código identificador del examen / prueba que está presentando el estudiante	Código 548
22	estu_exam_nombre	Nombre del examen que presentó el estudiante	Examen 2011-2
23	inst_cod_institucion	Código SNP de la institución de educación superior	
24	inst_nombre_institucion	Nombre de la institución de educación superior	
25	inst_origen	Origen Institución Educativa	Ver tabla 1b
26	inst_caracter_academico	Carácter Académico	Ver tabla 1c
27	estu_prac_id_prgrm_academico	Código identificador programa académico	
28	estu_prgm_academico	Nombre programa académico	
29	estu_nivel_prgm_academico	Nivel del programa académico que cursa el estudiante	Universitario
30	estu_metodo_prgm	Metodología del programa académico bajo la cual cursa el estudiante	Presencial Semipresencial Distancia (tradicional) A distancia (virtual)
31	dipo_codigomunicipio	Código municipio del programa	Código DANE
32	inst_cod_jornada	Código de la jornada de clases	Ver tabla 1d
33	estu_area_conoc	Nombre del área de conocimiento a la que pertenece el programa académico del estudiante	Ver tabla 1e
34	estu_nucleo_pregrado	Nombre del núcleo de pregrado al que pertenece el programa académico del estudiante	Ver tabla 1f
35	estu_grupo_referencia	Nombre del grupo de referencia al que pertenece el programa académico del	Ver tabla 1g

		estudiante	
36	estu_cod_grupo_ref	Código del grupo de referencia al que pertenece el programa académico del estudiante	Ver tabla 1g
37	estu_semestre_cursa	Semestre que cursa actualmente el estudiante	1 a 12 semestres 12 Significa 12 o más.
38	estu_pje_creditos	Porcentaje de créditos cursados y aprobados por el estudiante	Ver tabla 1h
39	inst_vlr_matricula_ant	Valor anual de la matrícula el año anterior	Ver tabla 1i
40	estu_sn_matricula_propio	Si canceló la matrícula año anterior con recursos propios	Si(1) No(0)
41	estu_sn_matricula_padres	Si canceló la matrícula año anterior con recursos de los padres o familiares	Si(1) No(0)
42	estu_sn_matricula_beca	Si canceló la matrícula año anterior con recursos de beca o subsidio	Si(1) No(0)
43	estu_sn_matricula_credito	Si canceló la matrícula año anterior con recursos de crédito	Si(1) No(0)
44	estu_titulo_bto	Título de bachiller obtenido por el estudiante	A-Académico N -Normalista T-Técnico
45	estu_exam_semestre_pretacion	Semestre en el que presentó el examen de estado	1 o 2 semestre
46	estu_exam_anno_presentacion	Año en el que presentó el examen de estado	
47	estu_hogar_actual	Tipo de hogar actual donde reside el estudiante	1. Es el habitual-permanente 2. Es temporal por razones de estudio u otra razón
48	fami_num_pers_grup_fam	Número total de personas que conforman el grupo familiar	
49	estu_sn_cabeza_fmlia	Si el estudiante es cabeza de familia.	Si (1) No (2)
50	fami_num_pers_cargo	Número de personas a cargo (cuando es cabeza de familia).	0 a 5 personas 5 representa 5 o más
51	fami_cod_educa_padre	Código nivel educativo más alto alcanzado por el padre del estudiante	Ver tabla 1j
52	fami_cod_educa_madre	Código nivel educativo más alto alcanzado por la madre del estudiante	Ver Tabla 1j
53	fami_sn_lee_escribe_padre	Si puede leer y escribir el padre del estudiante	Si (1) No (2)
54	fami_sn_lee_escribe_madre	Si puede leer y escribir la madre del estudiante	Si (1) No (2)
55	fami_cod_ocup_padre	Ocupación actual del padre del estudiante (o última si falleció)	Ver Tabla 1k
56	fami_cod_ocup_madre	Ocupación actual de la madre del estudiante (o última si falleció)	Ver tabla 1k
57	estu_estrato	Estrato socioeconómico de la vivienda donde reside actualmente el estudiante según el recibo del servicio de energía eléctrica	Ver tabla 1l
58	fami_num_hermanos	Número de hermanos del estudiante	
59	fami_nivel_hermano	Nivel educativo más alto alcanzado por los hermanos mayores del estudiante	Ver tabla 1m
60	econ_cuartos	Número total de cuartos que dispone la residencia del estudiante	
61	fami_nivel_sisben	Nivel de clasificación del SISBEN	Ver Tabla 1n

62	econ_material_pisos	Material de los pisos que predomina en la vivienda donde reside el hogar habitual o permanente del	Ver Tabla 1o
63	econ_sn_televisor	Si hay televisor en el hogar habitual del estudiante	Si (1) No (0)
64	econ_sn_motocicleta	Si hay una motocicleta en el hogar habitual del estudiante	Si (1) No (0)
65	econ_sn_energia	Si posee el servicio de energía eléctrica en el hogar habitual del estudiante	Si (1) No (0)
66	econ_sn_acueducto	Si posee el servicio de acueducto en el hogar habitual del estudiante	Si (1) No (0)
67	econ_sn_alcantarillado	Si posee alcantarillado en el hogar habitual del estudiante	Si (1) No (0)
68	econ_sn_aseo	Si posee el servicio de aseso en el hogar habitual del estudiante	Si (1) No (0)
69	econ_sn_estufa	Si posee estufa a gas o eléctrica en el hogar habitual del estudiante	Si (1) No (0)
70	econ_sn_telefonia	Si posee el servicio de telefonía en el hogar habitual del estudiante	(1): Si (0): no
71	econ_sn_internet	Si posee el servicio de Internet en el hogar habitual del estudiante	(1): Si (0): no
72	econ_sn_servicio_tv	Si posee el servicio de servicio de televisión por cable, satélite o parabólica en el hogar habitual del estudiante	(1): Si (0): no
73	econ_sn_computador	Si hay un computador en el hogar habitual del estudiante	(1): Si (0): no
74	econ_sn_celular	Si hay un teléfono celular en el hogar habitual del estudiante	(1): Si (0): no
75	econ_sn_dvd	Si hay un reproductor DVD en el hogar habitual del estudiante	(1): Si (0): no
76	econ_sn_lavadora	Si hay una máquina lavadora de ropas en el hogar habitual del estudiante	(1): Si (0): no
77	econ_sn_microondas	Si hay un horno microondas en el hogar habitual del estudiante	(1): Si (0): no
78	econ_sn_automovil	Si hay un automóvil particular en el hogar habitual del estudiante	(1): Si (0): no
79	econ_sn_horno	Si hay un horno eléctrico o a gas en el hogar habitual del estudiante	(1): Si (0): no
80	econ_sn_nevera	Si hay una nevera o enfriador en el hogar habitual del estudiante	(1): Si (0): no
81	infa_dormitorios	Número de dormitorios en el hogar habitual del estudiante	1 a 10 dormitorios 10 representa 10 o mas
82	fami_ing_fmliar_mensual	ingresos mensuales familiares en salarios mínimos	Ver tabla 1p
83	estu_trabaja	Si estudiante trabaja o no	Ver tabla 1q
84	estu_horas_trabajo	Nro. de horas / semanal si trabaja	
85	estu_otro_idioma_lee	Si lee otro idioma	Ver tabla 1r
86	estu_otro_idioma_habla	Si habla otro idioma	Ver tabla 1r
87	estu_nivel_postgrado	Si el estudiante posee un postgrado	05 -Especialización 06 -Maestría 07 -Doctorado 98 -Ninguno
88	estu_etnia	Si el estudiante pertenece a una etnia	Ver tabla 1s
89	mod_lectura_critica	Puntaje asignado al Módulo de Lectura Critica	

90	mod_comunica_escrita_punt	Puntaje asignado al módulo de Comunicación Escrita
91	mod_comunica_escrita_desem	Desempeño asignado al módulo de Comunicación Escrita
92	mod_razona_cuantitativo_punt	Puntaje asignado al módulo de Razonamiento Cuantitativo
93	mod_ingles_punt	Puntaje asignado al módulo de Ingles
94	mod_ingles_desem	Desempeño asignado al módulo de Ingles

Fuente: ICFES

Tabla 1a. Estado Civil

Código	Descripción
01	Soltero(a)
02	Casado(a)
03	Viudo(a)
04	Separado(a) y/o Divorciado
05	Unión Libre

Fuente: ICFES

Tabla 1b. Origen institución

Descripción
No Oficial – Corporación
No Oficial - Fundación
No Oficial
Oficial
Oficial Departamental
Oficial Municipal
Oficial Nacional
Régimen Especial

Fuente: ICFES

Tabla 1c. Carácter académico

Descripción
Académico
Escuela Tecnológica
Institución Tecnológica
Institución Universitaria
Normalista
Técnica Profesional
Técnica Profesional
Universidad

Fuente: ICFES

Tabla 1d. Código jornada

Código	Descripción
1	Diurno
2	Nocturno
3	Educación a Distancia
4	Mixta
5	Registro Calificado
6	Reg.Calif Acred.Voluntaria
9	No aplica
10	Registro Simple
11	Registro Alta Calidad
12	Jornada Única

Fuente: ICFES

Tabla 1e. Área del Conocimiento

Descripción
Agronomía, Veterinaria y afines
Bellas Artes
Ciencias de la Educación
Ciencias de la Salud
Ciencias Sociales y Humanas
Economía, Administración, Contaduría y Afines
Ingeniería, Arquitectura, Urbanismo y Afines
Matemáticas y Ciencias Naturales

Fuente: ICFES

Tabla 1f. Núcleo pregrado

Código	Descripción
1	Administración
2	Agronomía
3	Antropología, artes liberales
4	Arquitectura
5	Artes representativas
6	Artes plásticas, visuales y afines
7	Bacteriología
8	Bibliotecología, otros de ciencias sociales y humanas
9	Biología, microbiología y afines
10	Ciencia política, relaciones internacionales
11	Comunicación social, periodismo y afines
12	Contaduría publica
13	Deportes, educación física y recreación
14	Derecho y afines
15	Disenio
16	Diseño
17	Economía
18	Educación
19	Enfermería
20	Filosofía, teología y afines
21	Física
22	Formación relacionada con el campo militar o policial
23	Geografía, historia
24	Geología, otros programas de ciencias naturales
25	Ingeniería administrativa y afines
26	Ingeniería agrícola, forestal y afines
27	Ingeniería agroindustrial, alimentos y afines
28	Ingeniería Ambiental, sanitaria y afines
29	Ingeniería biomédica y afines
30	Ingeniería civil y afines
31	Ingeniería de minas, metalurgia y afines
32	Ingeniería de sistemas, telemática y afines
33	Ingeniería eléctrica y afines
34	Ingeniería electrónica, telecomunicaciones y afines
35	Ingeniería industrial y afines
36	Ingeniería mecánica y afines
37	Ingeniería química y afines
38	Instrumentación quirúrgica

39	Lenguas modernas, literatura, lingüística y afines
40	Matemáticas, estadística y afines
41	Matemáticas, estadística y afines
42	Medicina
43	Medicina Veterinaria
44	Música
45	Nutrición y dietética
46	Odontología
47	Optometría, otros programas de ciencias de la salud
48	Otras ingenierías
49	Otros programas asociados a bellas artes
50	Psicología
51	Publicidad y afines
52	Química y afines
53	Salud pública
54	Sin clasificar
55	Sin especificar
56	Sociología, trabajo social y afines
57	Terapias
58	Zootecnia

Fuente: ICFES

Tabla 1g. Grupo de Referencia

Código	Detalle
1	Arquitectura y Urbanismo
2	Artes, diseño, comunicación
3	Bellas artes y diseño
4	Ciencias agropecuarias
5	Ciencias económicas y administrativas
6	Ciencias militares y navales
7	Ciencias Naturales y exactas
8	Ciencias Sociales
9	Comunicación, periodismo y publicidad
10	Derecho
11	Educación
12	Grupo referencia nacional
13	Humanidades
14	Ingeniería
15	Medicina
16	Militar y policial
17	Recreación y deportes
18	Salud

Fuente: ICFES

Tabla 1h. Porcentaje de créditos cursados y aprobados

Código	Descripción
0	No sigue el sistema de créditos
1	Menos del 40% (AÑOS 2008-1, 2008-2 Y 2009,2)
1	Menos del 75% (AÑO 2010-2)
2	Entre el 40% y el 59% (AÑOS 2008-1, 2008-2 Y 2009,2)
2	Entre el 75% y el 80% (AÑO 2010-2)
3	Entre el 60% y el 79% (AÑOS 2008-1, 2008-2 Y 2009,2)
3	Entre el 81% y el 90% (AÑO 2010-2)
4	El 80% ó más (AÑOS 2008-1, 2008-2 Y 2009,2)
4	Más del 90% (AÑO 2010-2

Fuente: ICFES

Tabla 1i. Valor anual de la matricula año anterior

Código	Descripción
1	No pagó matrícula
2	Menos de 500 mil
3	Entre 500 mil y menos de 1 millón
4	Entre 1 millón y menos de 3 millones
5	Entre 3 millones y menos de 5 millones
6	Más de 5 millones

Fuente: ICFES

Tabla 1j. Educación del padre o madre

Código	Descripción
0	Ninguno
9	Primaria incompleta
10	Primaria completa
11	Bachillerato incompleto
12	Bachillerato completo
13	Educación técnica o tecnológica sin titulo
14	Educación técnica o tecnológica con titulo
15	Educación profesional sin titulo
16	Educación profesional con titulo
17	Postgrado

Fuente: ICFES

Tabla 1k. Ocupación del padre o madre

Código	Descripción
13	Empresario(a) dueño(a) de empresa industrial, comercial, agropecuaria -de servicios de más de10 trabajadores.
14	Pequeño empresario(a) dueño(a) de microempresa -pequeño-negocio-familiar -de finca -parcela, que vive de su explotación
15	Empleado-con cargo-como-director(a) -gerente general de empresa privada -entidad pública
16	Empleado(a) de nivel directivo (con personas a cargo) en empresa privada -entidad pública
17	Empleado(a) de nivel técnico/ profesional de empresa privada -en entidad pública
18	Empleado(a) de nivel auxiliar -administrativo de empresa privada -en entidad pública
19	Obrero u operario empleado(a) de empresa privada -en entidad pública
20	Profesional independiente ejerce su profesión sin vinculación laboral permanente en empresa privada -entidad pública
21	Trabajador por cuenta propia ejerce un oficio sin vinculación laboral: comerciantes por cuenta propia; obreros u operarios independientes; jornaleros; trabajadores independientes que prestan servicios
22	Hogar personas dedicadas principalmente a las labores del hogar
23	Pensionado(a) persona que vive de una pensión por concepto de jubilación
24	Rentista persona que vive de ingresos -rentas generados de su patrimonio
25	Estudiante persona que la mayor parte del tiempo se dedica al estudio
26	Otra actividad u ocupación
99	No sabe

Fuente: ICFES

Tabla 1l. Estrato socioeconómico

Código	Descripción
1	Estrato1
2	Estrato2
3	Estrato3
4	Estrato4
5	Estrato5
6	Estrato6
8	Vive en una zona rural donde no hay estratificación socio económica

Fuente: ICFES

Tabla 1m. Educación de los hermanos mayores

Código	Descripción
0	ninguno
9	primaria incompleta
10	primaria completa
11	bachillerato incompleto
12	bachillerato completo
13	educación técnica o tecnológica sin titulo
14	educación técnica o tecnológica con titulo
15	educación profesional sin titulo
16	educación profesional con titulo
17	postgrado
88	No tiene hermanos mayores
99	no sabe

Fuente: ICFES

Tabla 1n. Nivel de SISBEN

Código	Descripción
1	Nivel 1
2	Nivel 2
3	Nivel 3
4	Está clasificada en otro nivel
5	No está clasificada por el SISBEN

´ Fuente: ICFES

Tabla 1o. Material de los pisos

Código	Descripción
1	Tierra, arena
2	Cemento, gravilla, ladrillo
3	Madera burda, tabla -tablón
4	Madera pulida, baldosa, tableta, mármol, alfombra
5	Madera pulida, mármol, alfombra -tapete de pared a pared

Fuente: ICFES

Tabla 1p. Ingresos familiares mensuales

Código	Descripción
1	Menos de 1 SM
2	Entre 1 y menos de 2 SM
3	Entre 2 y menos de 3 SM
4	Entre 3 y menos de 5 SM
5	Entre 5 y menos de 7 SM
6	Entre 7 y menos de 10 SM
7	10 -más SM

Fuente: ICFES

Tabla 1q. Trabajo del estudiante

Código	Descripción
0	No
3	Si, para contribuir a pagar su matrícula y/-los gastos del hogar
4	Si, por ser práctica obligatoria del programa de estudios
5	Si, para adquirir experiencia y/-recursos para sus gastos personales

Fuente: ICFES

Tabla 1r. Lee o habla otro idioma

Código	Descripción
01	Inglés
02	Francés
03	Alemán
04	Italiano
05	Portugués
06	Japonés
99	Otro

Fuente: ICFES

Tabla 1s. Etnia estudiante

Código	Descripción
01	Comunidades negras
02	Raizal (isleño)
03	Paez
04	Sikuani
05	Arhuaco
06	Emberá
07	Guambiano
08	Pijao
09	Wayúu
10	Zenú
11	Pasto
12	Cancuamo
13	Inga
14	Tucano
15	Huitoto
16	Cubeo
17	Comunidad Rom(gitana)
99	Otro

Fuente: ICFES

Tabla 2. Clasificación de atributos en dimensiones

Dimensión	Atributo
Sociodemográfica	estu_consecutivo, estu_cod_aplicación, estu_genero, estu_nacimiento_dia, estu_nacimiento_mes, estu_nacimiento_ anno, estu_pais_ reside, estu_estado_civil, estu_disc_invidente, estu_disc_sordo_con_interprete, estu_disc_sordo_sin_interprete, estu_disc_motriz, estu_disc_sordo_ceguera, estu_zona, econ_area_vive, estu_reside_codmpio, estu_hogar_actual, fami_num_pers_grup_fam, estu_sn_cabeza_fmlia, fami_num_pers_cargo, fami_cod_educa_padre, fami_cod_educa_madre, fami_sn_lee_escribe_padre, fami_sn_lee_escribe_madre, fami_cod_ocup_padre, fami_cod_ocup_madre, fami_num_hermanos, fami_nivel _hermano, estu _etnia
Económica	inst_vlr_matricula_ant, estu_sn_matricula_propio, estu_sn_matricula_padres, estu_sn_matricula_beca, estu_sn_matricula_credito, estu_estrato, econ_cuartos, fami_nivel_sisben, econ_material_pisos, econ_sn_televisor, econ_sn_motocicleta, econ_sn_energía, econ_sn_acueducto, econ_sn_alcantarillado, econ_sn_aseo, econ_sn_estufa, econ_sn_telefonia econ_sn_internet, econ_sn_servicio_tv, econ_sn _computador, econ_sn _celular , econ_ sn _dvd, econ_sn_lavadora, econ_sn _microondas, econ_sn_automovil, econ_sn_horno, econ_sn _nevera, infa_dormitorios, fami_ing _fmliar_mensual, estu_trabaja, estu_horas_trabajo
Académica	estu_anno_egreso, estu_exam_codmpio_presentacion, ,estu_exam_mpio_presentacion , estu_exam_dpto_presentacion, estu_exam_cod, estu_exam_nombre, estu_prac_ id_ prgrm_academico, estu_prgm_academico, estu_nivel_prgm_academico, estu_metodo_progr, dipo_codigomunicipio, inst_cod_jornada, estu_area_conoc, estu_nucleo_pregrado, estu_grupo _referencia, estu_cod_grupo_ref, estu_semestre_cursa, estu_pje_creditos, estu_titulo_bto, estu_exam_semestre_pretacion, estu_exam_anno_presentación, estu_otro_idioma _lee, estu_otro _idioma _habla, estu_nivel_postgrado
Institucional	inst_cod_institucion, inst_nombre_institucion, inst_origen, inst_caracter_academico, dipo_codigo_municipio
Clases	mod_lectura _critica, mod_comunica_escrita_punt, mod_comunica _escrita_desem, mod_razona_cuantitativo_punt, mod_ingles_punt, mod_ingles_desem

Fuente: esta investigación

Tabla 3. Matriz de correlaciones de Competencias genéricas

Competencias	Coeficiente	Lectura Crítica	Comunicación Escrita	Razonamiento Cuantitativo	Inglés
Lectura Crítica	Coeficiente de Correlación	1	,355**	,535**	,508**
	p valor		0,000	0,000	0,000
	N		95337	97068	97068
Comunicación Escrita	Coeficiente de Correlación		1	,233**	,295**
	p valor			0,000	0,000
	N			95337	95337
Razonamiento Cuantitativo	Coeficiente de Correlación			1	,481**
	p valor				0,000
	N				97068
Inglés	Coeficiente de Correlación				1
	p valor				
	N				

**. La correlación es significativa al nivel 0,01 (bilateral).

Fuente: elaboración propia

Tabla 4. Género y desempeño académico en competencias genéricas

Competencia Genérica	Género	N	Media	Error típico	Intervalo de confianza para la media al 95%		Grupo	d
					Límite inferior	Límite superior		
Lectura crítica	Femenino	56824	10,199	0,0041	10,191	10,207	B	0,11
	Masculino	40231	10,308	0,0049	10,298	10,318	A	
Comunicación escrita	Femenino	55948	10,216	0,0041	10,208	10,225	A	
	Masculino	39389	10,058	0,0052	10,048	10,068	B	0,16
Razonamiento cuantitativo	Femenino	56825	9,994	0,0039	9,987	10,002	B	0,46
	Masculino	40232	10,457	0,0055	10,447	10,468	A	
Inglés	Femenino	56755	10,222	0,0049	10,213	10,232	B	0,23
	Masculino	40191	10,507	0,0066	10,494	10,520	A	

Fuente: elaboración propia

Tabla 5. Estado civil y desempeño académico en competencias genéricas

Competencia genérica	Estado civil	N	Media	Error típico	Intervalo de confianza para media al 95%		Grupo	d
					Límite inferior	Límite superior		
Lectura crítica	Soltero	77951	10,328	0,0035	10,321	10,335	A	
	Casado	10482	9,939	0,0089	9,922	9,957	B	0,40
	Unión Libre	7175	9,846	0,0112	9,824	9,868	C	0,49
	Separado/Divorciado	1283	9,931	0,0255	9,881	9,980	B	0,41
	Viudo	164	9,707	0,0751	9,559	9,856	C	0,64
Comunicación escrita	Soltero	76671	10,194	0,0036	10,187	10,201	A	
	Casado	10257	10,023	0,0095	10,004	10,042	B	0,17
	Unión Libre	6993	9,896	0,0118	9,873	9,919	C	0,30
	Separado/Divorciado	1262	10,056	0,0269	10,003	10,109	B	0,14
	Viudo	154	9,830	0,0765	9,679	9,981	C	0,37
Razonamiento cuantitativo	Soltero	77951	10,267	0,0037	10,259	10,274	A	
	Casado	10484	9,876	0,0087	9,859	9,894	B	0,38
	Unión Libre	7175	9,847	0,0107	9,826	9,867	B	0,41
	Separado/Divorciado	1283	9,822	0,0251	9,773	9,871	B	0,43
	Viudo	164	9,543	0,0665	9,411	9,674	C	0,70
Inglés	Soltero	77875	10,474	0,0046	10,465	10,483	A	
	Casado	10463	9,831	0,0086	9,815	9,848	B	0,52
	Unión Libre	7163	9,730	0,0093	9,711	9,748	B	0,60
	Separado/Divorciado	1282	9,864	0,0263	9,813	9,916	B	0,48
	Viudo	163	9,726	0,0738	9,580	9,871	B	0,58

Fuente: elaboración propia

Tabla 6. Nivel educativo de los padres y desempeño académico en competencias genéricas

Competencia genérica	Nivel educativo de los padres	N	Media	Error típico	Intervalo de confianza para la media al 95%		Grupo	d
					Límite inferior	Límite superior		
Lectura crítica	Posgrado	10964	10,675	0,0091	10,657	10,693	A	
	Profesional	17915	10,459	0,0073	10,445	10,473	B	0,22
	Técnico	10207	10,331	0,009	10,313	10,349	C	0,37
	Secundaria	26765	10,218	0,0057	10,207	10,229	D	0,49
	Primaria	21664	10,017	0,0063	10,004	10,029	E	0,70
	Ninguno	9538	9,842	0,0096	9,823	9,861	F	0,88
Comunicación escrita	Posgrado	10817	10,399	0,01	10,379	10,419	A	
	Profesional	17668	10,247	0,0077	10,232	10,262	B	0,15
	Técnico	10022	10,205	0,0099	10,185	10,224	C	0,19
	Secundaria	26272	10,124	0,0061	10,112	10,136	D	0,27
	Primaria	21250	10,042	0,0067	10,029	10,055	E	0,36
	Ninguno	9307	9,949	0,0101	9,929	9,969	F	0,45
Razonamiento cuantitativo	Posgrado	10964	10,599	0,0108	10,577	10,62	A	
	Profesional	17914	10,383	0,0082	10,367	10,399	B	0,19
	Técnico	10207	10,253	0,0099	10,233	10,272	C	0,32
	Secundaria	26768	10,165	0,0061	10,153	10,177	D	0,42
	Primaria	21663	9,979	0,0063	9,966	9,991	E	0,62
	Ninguno	9539	9,803	0,0092	9,785	9,821	F	0,77
Ingles	Posgrado	10958	11,231	0,0147	11,203	11,26	A	
	Profesional	17898	10,818	0,0106	10,798	10,839	B	0,28
	Técnico	10196	10,379	0,0112	10,357	10,401	C	0,63
	Secundaria	26736	10,214	0,0066	10,201	10,227	D	0,83
	Primaria	21642	9,901	0,0061	9,889	9,913	E	1,15
	Ninguno	9514	9,730	0,0083	9,713	9,746	F	1,20

Fuente: elaboración propia

Tabla 7. Estrato socioeconómico y desempeño académico en competencias genéricas.

Competencia genérica	Estrato Social	N	Media	Error típico	Intervalo de confianza para la media al 95%		Grupo	d
					Límite inferior	Límite superior		
Lectura crítica	Estrato 6	2605	10,777	0,0193	10,739	10,815	A	
	Estrato 5	5067	10,646	0,0134	10,62	10,673	B	0,14
	Estrato 4	13617	10,510	0,0082	10,494	10,527	C	0,28
	Estrato 3	36353	10,301	0,0049	10,291	10,31	D	0,51
	Estrato 2	29271	10,101	0,0055	10,09	10,111	E	0,72
	Zona rural	112	9,996	0,1014	9,795	10,196	F	0,79
	Estrato 1	10042	9,746	0,0099	9,726	9,765	G	1,04
Comunicación escrita	Estrato 6	2571	10,463	0,0196	10,424	10,501	A	
	Estrato 5	5004	10,363	0,0145	10,335	10,391	B	0,10
	Estrato 4	13430	10,309	0,0088	10,292	10,326	C	0,15
	Estrato 3	35789	10,192	0,0053	10,181	10,202	D	0,27
	Estrato 2	28705	10,070	0,0058	10,058	10,081	E	0,40
	Estrato 1	9730	9,836	0,0098	9,817	9,855	F	0,64
	Zona rural	108	9,791	0,0885	9,615	9,966	F	0,68
Razonamiento cuantitativo	Estrato 6	2605	10,747	0,0229	10,702	10,792	A	
	Estrato 5	5067	10,573	0,0159	10,541	10,604	B	0,15
	Estrato 4	13617	10,426	0,0095	10,407	10,444	C	0,29
	Estrato 3	36353	10,232	0,0053	10,221	10,242	D	0,50
	Zona rural	112	10,024	0,0867	9,852	10,196	E	0,62
	Estrato 2	29271	10,049	0,0056	10,038	10,06	E	0,71
	Estrato 1	10042	9,748	0,0092	9,73	9,766	F	1,02
Inglés	Estrato 6	2605	12,219	0,031	12,158	12,28	A	
	Estrato 5	5067	11,455	0,022	11,412	11,498	B	0,49
	Estrato 4	13617	10,913	0,0124	10,888	10,937	C	0,89
	Estrato 3	36353	10,306	0,0061	10,294	10,318	D	1,60
	Zona rural	112	10,232	0,1284	9,978	10,487	D	1,26
	Estrato 2	29271	9,953	0,0059	9,942	9,965	E	2,12
	Estrato 1	10042	9,644	0,0095	9,625	9,662	F	2,32

Fuente: elaboración propia

Tabla 8. Modalidad de estudio y desempeño académico en competencias genéricas

Competencias genéricas	Modalidad de estudio	N	Media	Error típico	Intervalo de confianza para la media al 95%		Grupo	d
					Límite inferior	Límite superior		
Lectura crítica	Presencial	84272	10,326	0,0033	10,32	10,333	A	
	A distancia	12783	9,702	0,0079	9,686	9,718	B	0,66
Comunicación escrita	Presencial	82888	10,193	0,0035	10,186	10,2	A	
	A distancia	12449	9,870	0,0085	9,853	9,887	B	0,32
Razonamiento cuantitativo	Presencial	84272	10,259	0,0036	10,252	10,266	A	
	A distancia	12785	9,706	0,0074	9,692	9,721	B	0,54
Inglés	Presencial	84202	10,451	0,0044	10,442	10,459	A	
	A distancia	12744	9,613	0,0063	9,601	9,625	B	0,69

Fuente: elaboración propia

Tabla 9. Tipo de institución y desempeño académico en competencias genéricas

Competencias genéricas	Tipo de institución	N	Media	Error típico	Intervalo de confianza para la media al 95%		Grupo	d
					Límite inferior	Límite superior		
Lectura crítica	Oficial	38864	10,269	0,0051	10,259	10,279	A	
	Privada	58144	10,228	0,0039	10,220	10,236	A	0,04
	Régimen especial	47	9,996	0,1197	9,755	10,237	B	0,27
Comunicación escrita	Oficial	38099	10,137	0,0052	10,126	10,147	A	
	Privada	57191	10,161	0,0042	10,152	10,169	A	0,02
	Régimen especial	47	10,155	0,1508	9,852	10,459	A	0,02
Razonamiento cuantitativo	Oficial	38867	10,251	0,0055	10,241	10,262	A	
	Privada	58143	10,143	0,0041	10,135	10,151	B	0,11
	Régimen especial	47	10,296	0,134	10,026	10,565	A	0,04
Inglés	Oficial	38810	10,219	0,0059	10,207	10,230	B	0,16
	Privada	58089	10,422	0,0054	10,412	10,433	A	
	Régimen especial	47	9,847	0,0973	9,651	10,043	C	0,44

Fuente: elaboración propia

Tabla 10. Acreditación Institucional y desempeño académico en competencias genéricas

Competencias genéricas	Acreditación Institucional	N	Media	Error típico	Intervalo de confianza para la media al 95%		Grupo	d
					Límite inferior	Límite superior		
Lectura crítica	Acreditada	16306	10,835	0,0072	10,821	10,849	A	
	No acreditada	80749	10,125	0,0033	10,118	10,131	B	0,76
Comunicación escrita	Acreditada	16152	10,547	0,0082	10,531	10,563	A	
	No acreditada	79185	10,070	0,0035	10,063	10,077	B	0,48
Razonamiento cuantitativo	Acreditada	16307	10,833	0,009	10,815	10,850	A	
	No acreditada	80750	10,056	0,0034	10,049	10,062	B	0,78
Inglés	Acreditada	16300	11,508	0,0118	11,485	11,531	A	
	No acreditada	80646	10,104	0,0037	10,097	10,112	B	1,23

Fuente: elaboración propia

Tabla 11. Acreditación del programa y desempeño académico en competencias genéricas

Acreditación programa		N	Media	Error típico	Intervalo de confianza para la media al 95%		Grupo	d
					Límite inferior	Límite superior		
Lectura crítica	Acreditado	14036	10,628	0,0081	10,612	10,644	A	
	No acreditado	83019	10,179	0,0033	10,173	10,186	B	0,47
Comunicación escrita	Acreditado	13839	10,383	0,0087	10,366	10,400	A	
	No acreditado	81498	10,112	0,0035	10,105	10,118	B	0,27
Razonamiento cuantitativo	Acreditado	14037	10,615	0,0098	10,596	10,634	A	
	No acreditado	83020	10,114	0,0034	10,107	10,120	B	0,50
Inglés	Acreditado	14021	11,027	0,0124	11,003	11,051	A	
	No acreditado	82925	10,224	0,0041	10,216	10,232	B	0,65

Fuente: elaboración propia

Tabla 12. Atributos con porcentaje alto de datos nulos

Atributo	Nulos (%)
estu_disc_ sordo_ ceguera, econ_area_vive, fami_sn_lee_escribe_padre, fami_sn_lee_escribe_madre, fami_num_hermanos, fami_nivel_hermano, econ_cuartos, econ_sn_televisor, econ_sn_motocicleta, econ_sn_energia, econ_sn_acueducto, econ_sn_alcantarillado, econ_sn_aseo, econ_sn_estufa, estu_otro_idioma_lee, estu_otro_idioma_habla, estu_nivel_postgrado, estu_etnia	100
estu_disc_invidente, estu_disc_sordo_con_interprete, estu_disc_sordo_sin_interprete, estu_zona	99.99
estu_disc motriz	99.92
estu_anno egreso	97.77

Fuente: esta investigación

Tabla 13. Atributos con valores constantes e identificadores

Atributo	Valor
estu_consecutivo	SBPRO+identificador
estu_cod_aplicacion	EK20112
estu_pais_reside	COLOMBIA
estu_exam_nombre	EXAMEN ESPECIFICO 2011-2
estu_exam_cod	548
estu_nivel_prgm_academico	UNIVERSITARIO

Fuente: esta investigación

Tabla 14. Nuevos atributos del repositorio de datos

Atributo	Descripción	Acción	Valores
estu_nacimiento_fecha	Fecha de nacimiento del estudiante	Reemplaza a los atributos: estu _nacimiento_día estu_nacimiento_mes estu_nacimiento_año	Valores con formato DD/MM/YYYY
estu_edad	Edad del estudiante en el momento de presentar la prueba	Reemplaza al atributo: estu_nacimiento_fecha	Valores numéricos
estu_discapacidad	Tipo de discapacidad del estudiante	Reemplaza a los atributos: estu_disc_invidente estu_disc_sordo_con_interprete estu_disc_sordo_sin_interprete estu_disc_motriz estu_disc_sordo_ceguera	I= invidente R= sordo con interprete S= sordo sin interprete M=discapacidad motriz C= sordo ceguera N = sin discapacidad
estu_reside_coddpto	Departamento de residencia del estudiante	Reemplaza al atributo: estu_reside_codmpio	Códigos DANE
estu_pers_cargo	Si el estudiante tiene personas a cargo	Reemplaza al atributo: fami_num_pers_cargo	SI NO
estu_financiacion_matricula	Forma de financiar el pago de la matrícula	Reemplaza a los atributos: estu_sn_matricula_propio estu_sn_matricula_padres estu_sn_matricula_beca estu_sn_matricula_credito	COMBINACIONES EN CODIGO BINARIO DE 4 BITS
fami_nivel_educa_padres	Máximo nivel educativo completo entre el padre y la madre	Reemplaza a los atributos: fami_cod_educa_padre fami_cod_educa_madre	PRIMARIA SECUNDARIA TÉCNICO/TECNOLOGICO PROFESIONAL POSTGRADO NINGUNO
fami_ocup_padre	Ocupación del padre	Reemplaza al atributo: fami_cod_ocup_padre	DIRECTIVO EMPLEADO EMPRESARIO HOGAR INDEPENDIENTE OTRA PENSIONADO PROFESIONAL
fami_ocup_madre	Ocupación de la madre	Reemplaza al atributo: fami_cod_ocup_madre	Los mismos valores de la ocupación del padre
inst_tipo	Tipo de institución del estudiante	Reemplaza a los atributos: inst_origen	OFICIAL PRIVADA RÉGIMEN ESPECIAL
inst_acreditada	Determina si la institución donde pertenece el	Reemplaza a los atributos: inst_cód_institución inst_nombre_institucion	ACREDITADA NO ACREDITADA

	estudiante es acreditada o no según CNA		
inst_prog_acreditado	Determina si el programa al cual pertenece el estudiante es acreditado o no según CNA	Reemplaza a los atributos: estu_prac_id_progr_académico estu_progr_académico	ACREDITADO NO ACREDITADO
area_grupo_referencia	Áreas de los grupos de referencia de los programas	Reemplaza al atributo: estu_grupo_referencia	CIENCIAS HUMANAS CIENCIAS SOCIALES CIENCIAS NATURALES Y TÉCNICAS
inst_cod_dpto	Código DANE del departamento donde se ofrece el programa	Reemplaza al atributo: dipo_cod_municipio	Códigos departamentos según DANE
inst_dpto_programa	Nombre del departamento donde se ofrece el programa académico	Reemplaza al atributo: inst_cod_dpto	Nombres departamentos de Colombia según DANE
inst_programa_zona	Zona geográfica donde se ofrece el programa	Reemplaza al atributo: inst_dpto_programa	BOGOTA EJE CAFETERO CARIBE CENTRO ORIENTE PACÍFICO CENTRO SUR LLANO
num_estudiantes_zona	Número de estudiantes por zona geográfica	Nuevo atributo generado a partir del atributo: inst_programa_zona	ALTO MEDIO BAJO
num_instituciones_zona	Número de Instituciones de Educación Superior (IES) por zona geográfica	Nuevo atributo generado a partir del atributo: inst_programa_zona	ALTO MEDIO BAJO
eco_condicion_vivienda	Condición de la vivienda del estudiante	Reemplaza al atributo: econ_material_pisos	BUENA MALA REGULAR
eco_condicion_hogar	Condición del hogar del estudiante	Reemplaza a los atributos: econ_ sn _dvd econ_sn_lavadora econ_sn _microondas econ_sn_horno econ_sn _nevera	BUENA REGULAR MALA
eco_condicion_transporte	Condición de transporte en el hogar del estudiante	Reemplaza al atributo: econ_sn_automovil	PARTICULAR PÚBLICO

eco_condicion_tic	Condición de uso de TIC en el hogar del estudiante	Reemplaza a los atributos: econ_sn_telefonia econ_sn _celular econ_sn_internet econ_sn_servicio_tv econ_sn _computador	BUENA REGULAR MALA
eco_condicion_vive	Condición de vida del estudiante	Reemplaza a los atributos: infa_dormitorios fami_num_pers_fam	SIN HACINAMIENTO HACINAMIENTO MEDIO HACINAMIENTO CRITICO
mod_lectura_critica_desemp	Desempeño del estudiante en el Módulo de lectura crítica	Reemplaza al atributo: mod_lectura_critica	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA (< la media)
mod_comunica_escrita_desemp	Desempeño del estudiante en el Módulo de comunicación escrita	Reemplaza a los atributos: mod_comunica_escrita_punt mod_comunica_escrita_desem	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA (< la media)
mod_razona_cuantitativo_desemp	Desempeño del estudiante en el Módulo de razonamiento cuantitativo	Reemplaza al atributo: mod_razona_cuantitativo_punt	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA (< la media)
mod_ingles_desemp	Desempeño del estudiante en el Módulo de inglés	Reemplaza a los atributos: mod_ingles_punt mod_ingles_desem	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA (< la media)

Fuente: esta investigación

Tabla 15. Valores discretizados del atributo estu_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
Sin Dato	82
	97068

Fuente: esta investigación

Tabla 16. Valores del atributo inst_programa_zona

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

Fuente: Zonas OCAD y esta investigación.

Tabla 17. Valores discretizados del atributo num_estudiantes_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
	97068	

Fuente: esta investigación

Tabla 18. Valores discretizados del atributo num_instituciones_zona

Zona	No. Instituciones	Valor
BOGOTA	73	ALTO
EJE CAFETERO	78	ALTO
CARIBE	64	MEDIO
CENTRO ORIENTE	56	BAJO
PACIFICO	64	MEDIO
CENTRO SUR	21	BAJO
LLANO	12	BAJO
	368	

Fuente: esta investigación

Tabla 19. Valores del atributo eco_condicion_hogar

Material de los Pisos de la Vivienda	Condición Vivienda
Tierra, Arena	MALA
Cemento, Gravilla, Ladrillo	MALA
Madera Burda, Tabla – Tablon	REGULAR
Madera Pulida, Baldosa, Tableta, Marmol, Alfombra	BUENA
Madera Pulida, Marmol, Alfombra - Tapete De Pared A Pared	BUENA

Fuente: esta investigación.

Tabla 20. Calculo del índice de condición_hogar

Electrodomésticos	Si	No
econ_sn_dvd	1	0
econ_sn_lavadora	1	0
econ_sn_microondas	1	0
econ_sn_horno	1	0
econ_sn_nevera	1	0

Fuente: esta investigación.

Tabla 21. Cálculo del índice de condición_tic

Servicios	Si	No
econ_sn_internet	1	0
econ_sn_servicio_tv	1	0
econ_sn_telefonia	1	0
econ_sn_celular	1	0
econ_sn_computador	1	0

Fuente: esta investigación.

Tabla 22. Asignación de valores del atributo area_grupo_referencia

Grupo de Referencia Programas	Área Grupo Referencia Programas
Ciencias Económicas y Administrativas	CIENCIAS SOCIALES
Ingeniería	CIENCIAS NATURALES Y TÉCNICAS
Educación	CIENCIAS SOCIALES
Derecho	CIENCIAS HUMANAS
Salud	CIENCIAS HUMANAS
Ciencias Sociales	CIENCIAS SOCIALES
Comunicación, Periodismo y Publicidad	CIENCIAS SOCIALES
Medicina	CIENCIAS HUMANAS
Bellas Artes y Diseño	CIENCIAS HUMANAS
Ciencias Agropecuarias	CIENCIAS NATURALES Y TÉCNICAS
Ciencias Naturales Y Exactas	CIENCIAS NATURALES Y TÉCNICAS
Arquitectura Y Urbanismo	CIENCIAS SOCIALES
Humanidades	CIENCIAS HUMANAS
Militar y Policial	CIENCIAS SOCIALES
Recreación y Deportes	CIENCIAS HUMANAS
Ciencias Militares y Navales	CIENCIAS NATURALES Y TÉCNICAS
Artes - Diseño - Comunicación	CIENCIAS SOCIALES

Fuente: esta investigación.

Tabla 23. Diccionario de datos del repositorio final T97068A35

No	Atributo	Descripción	Valores
		Sociodemográficos	
1	estu_genero	Género	M ,F
2	estu_edad	Edad del estudiante en el momento de presentar la prueba	Ver tabla 15
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	Ver tabla 1p
19	estu_trabaja	Si estudiante trabaja o no	Ver tabla 1q
		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	Ver tabla 1e

		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 OCAD donde se ofrece el programa	BOGOTA, EJE CAFETERO CARIBE, CENTRO ORIENTE, PACÍFICO,CENTRO SUR, LLANO
30	num_estudiantes_zona	Número de estudiantes por zona OCAD	ALTO, MEDIO, BAJO
31	num_instituciones_zona	Número de Instituciones de Educación Superior (IES) por zona OCAD	ALTO, MEDIO,BAJO
		Clases	
32	mod_lectura_critica_desemp	Desempeño del estudiante en el Módulo de lectura crítica	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA (< la media)
33	mod_comunica_escrita_desemp	Desempeño del estudiante en el Módulo de comunicación escrita	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA (< la media)
34	mod_razona_cuantitativo_desemp	Desempeño del estudiante en el Módulo de razonamiento cuantitativo	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA (< la media)
35	mod_ingles_desemp	Desempeño del estudiante en el Módulo de inglés	SOBRE LA MEDIA (>= la media) BAJO LA MEDIA (< la media)

Fuente: esta investigación.

. **Tabla 24**. Repositorios por cada competencia

Repositorio	Descripción
T97055A32LEC	Repositorio para análisis de la competencia de lectura crítica. Contiene 97055 registros y 32 atributos.
T95337A32ESC	Repositorio para análisis de la competencia de comunicación escrita. Contiene 95337 registros y 32 atributos.
T97057A32CUA	Repositorio para análisis de la competencia de razonamiento cuantitativo. Contiene 97057 registros y 32 atributos.
T96946A32ING	Repositorio para análisis de la competencia de inglés. Contiene 96946 registros y 32 atributos.

Fuente: esta investigación.

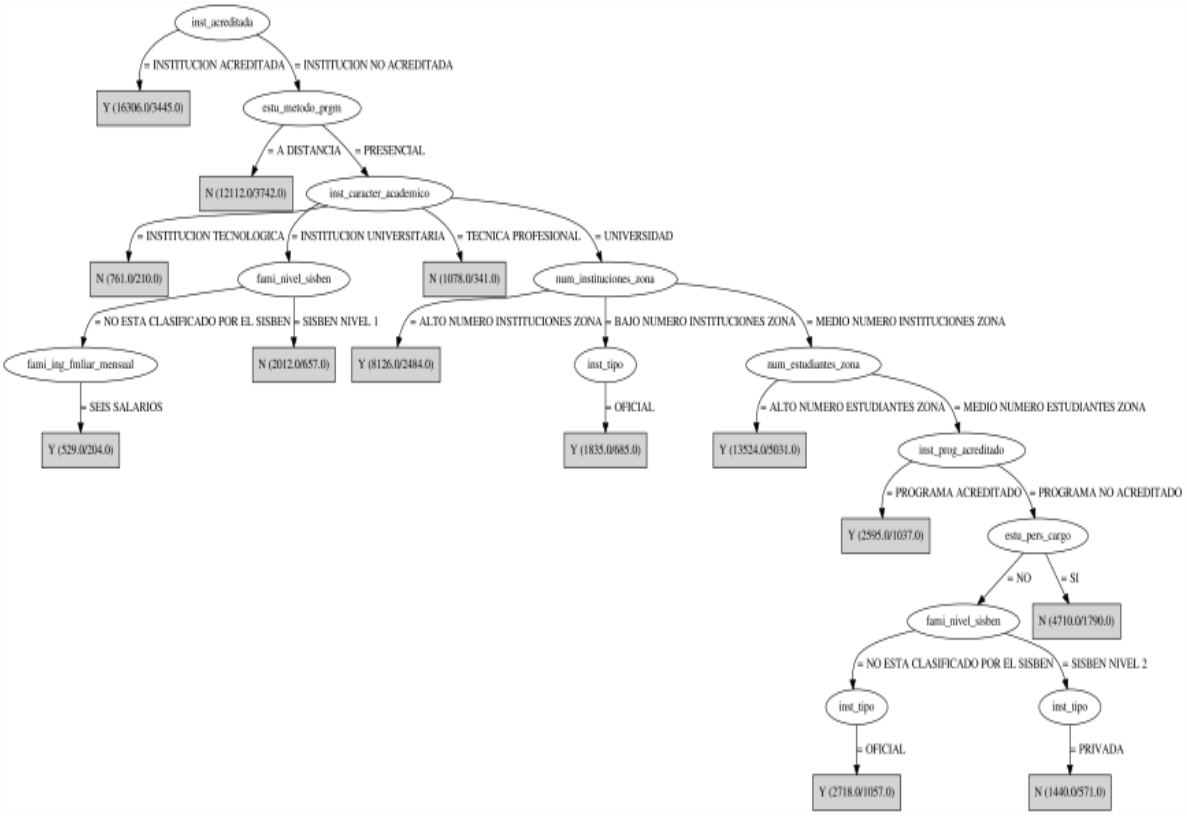

Figura 1. Árbol para lectura crítica podado

```
=== 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
```

Figura 2. Precisión y matriz de confusión del árbol de lectura crítica

	Antecedente	Consecuente	Confianza (%)	Soporte (%)
1	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = INSTITUCION TECNOLOGICA	N	72.40	0.78
2	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = A DISTANCIA	N	69.11	12.48
3	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = TECNICA PROFESIONAL	N	68.37	1.11
4	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = INSTITUCION UNIVERSITARIA y fami_nivel_sisben = SISBEN NIVEL 1	N	67.35	2.07
5	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_pers_cargo = SI	N	62.00	4.85
6	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_pers_cargo = NO y fami_nivel_sisben = SISBEN NIVEL 2 y inst_tipo = PRIVADA	N	60.35	1.48
7	inst_acreditada = INSTITUCION ACREDITADA	Y	78.87	16.80
8	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = ALTO NUMERO INSTITUCIONES ZONA	Y	69.43	8.37
9	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	62.80	13.93
10	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = BAJO NUMERO INSTITUCIONES ZONA y inst_tipo = OFICIAL	Y	62.67	1.89
11	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = INSTITUCION UNIVERSITARIA y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y fami_ing_fmliar_mensual = SEIS SALARIOS	Y	61.44	0.55
12	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_pers_cargo = NO y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y inst_tipo = OFICIAL	Y	61.11	2.80
13	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_prog_acreditado = PROGRAMA ACREDITADO	Y	60.04	2.67

Figura 3. Reglas más representativas de lectura crítica

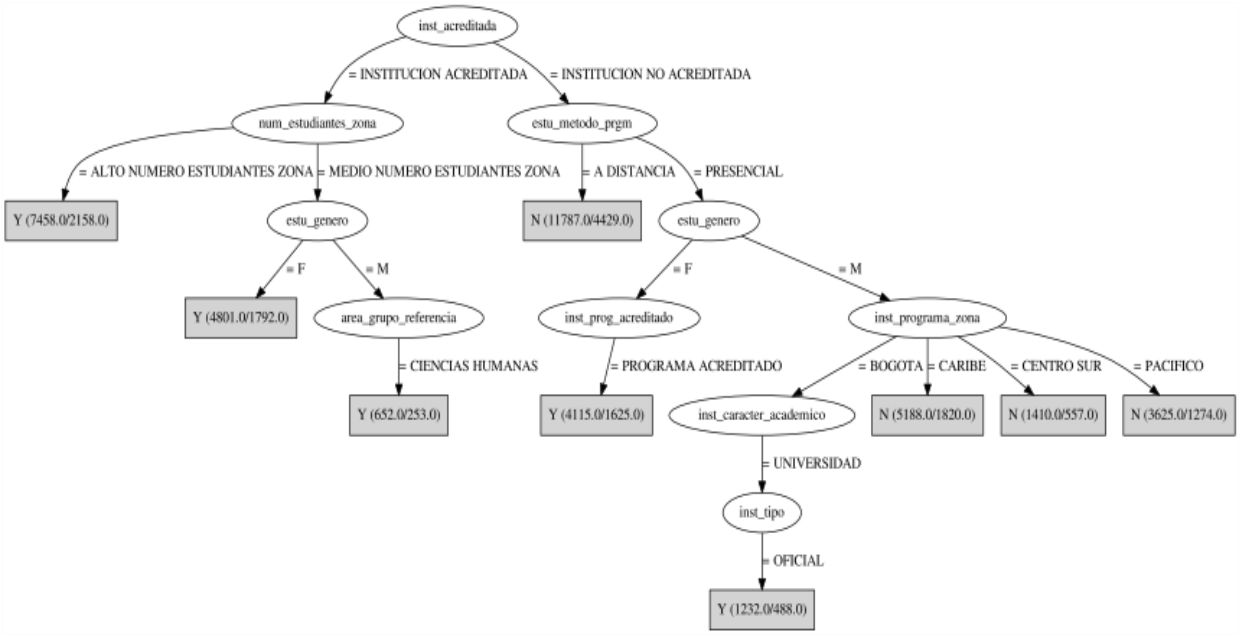

Figura 4. Árbol para comunicación escrita podado

```
=== 10 Fold Cross Validation ===
=== Summary ===
Correctly Classified Instances          56669          59.4407 %
Incorrectly Classified Instances        38668          40.5593 %
Kappa statistic                          0.1876
Mean absolute error                      0.4762
Root mean squared error                  0.4882
Relative absolute error                 95.2592 %
Root relative squared error             97.6493 %
Coverage of cases (0.95 level)          100 %
Mean rel. region size (0.95 level)      99.9969 %
Total Number of Instances               95337

=== Confusion Matrix ===
a       b <-- classified as
31234   17117 | a = N
21551   25435 | b = Y
```

Figura 5. Precisión y matriz de confusión del árbol de comunicación escrita

	Antecedente	Consecuente	Confianza (%)	Soporte (%)
1	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = M y inst_programa_zona = CARIBE	N	64.92	5.44
2	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = M y inst_programa_zona = PACIFICO	N	64.86	3.80
3	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = A DISTANCIA	N	62.42	12.36
4	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = M y inst_programa_zona = CENTRO SUR	N	60.50	1.48
5	inst_acreditada = INSTITUCION ACREDITADA y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	71.06	7.82
6	inst_acreditada = INSTITUCION ACREDITADA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y estu_genero = F	Y	62.67	5.04
7	inst_acreditada = INSTITUCION ACREDITADA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y estu_genero = M y area_grupo_referencia = CIENCIAS HUMANAS	Y	61.20	0.68
8	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = F y inst_prog_acreditado = PROGRAMA ACREDITADO	Y	60.51	4.32
9	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = M y inst_programa_zona = BOGOTA y inst_caracter_academico = UNIVERSIDAD y inst_tipo = OFICIAL	Y	60.39	1.29

Figura 6. Reglas más representativas de comunicación escrita

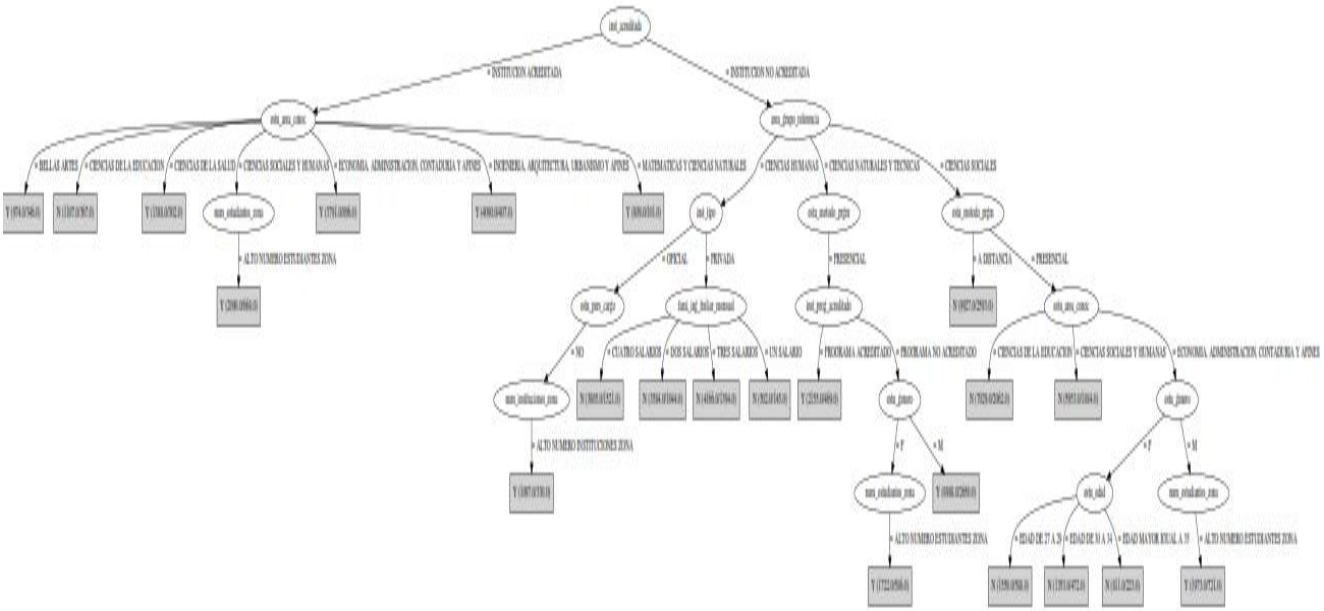

Figura 7. Árbol para razonamiento cuantitativo podado

```
=== 10 Fold Cross Validation ===
=== Summary ===
Correctly Classified Instances          64549          66.5063 %
Incorrectly Classified Instances        32508          33.4937 %
Kappa statistic                          0.3295
Mean absolute error                      0.4273
Root mean squared error                  0.4626
Relative absolute error                 85.4605 %
Root relative squared error             92.5244 %
Coverage of cases (0.95 level)          100 %
Mean rel. region size (0.95 level)      99.9923 %
Total Number of Instances                97057

=== Confusion Matrix ===
a       b <-- classified as
29093   19051 | a = Y
13457   35456 | b = N
```

Figura 8. Precisión y matriz de confusión del árbol de razonamiento cuantitativo

	Antecedente	Consecuente	Confianza (%)	Soporte (%)
1	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = A DISTANCIA	N	73.88	10.23
2	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES y estu_genero = F y estu_edad = EDAD MAYOR IGUAL A 35	N	72.50	0.84
3	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = PRIVADA y fami_ing_fmliar_mensual = UN SALARIO	N	71.12	0.52
4	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = PRIVADA y fami_ing_fmliar_mensual = DOS SALARIOS	N	70.87	3.69
5	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = CIENCIAS DE LA EDUCACION	N	70.66	7.24
6	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = CIENCIAS SOCIALES Y HUMANAS	N	68.35	6.13
7	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = CIENCIAS DE LA EDUCACION	N	66.85	1.14
8	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = PRIVADA y fami_ing_fmliar_mensual = TRES SALARIOS	N	66.54	4.29
9	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES y estu_genero = F y estu_edad = EDAD DE 30 A 34	N	65.11	1.39
10	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES y estu_genero = F y estu_edad = EDAD DE 27 A 29	N	62.28	1.61
11	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = PRIVADA y fami_ing_fmliar_mensual = CUATRO SALARIOS	N	60.03	3.92
12	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = INGENIERIA, ARQUITECTURA, URBANISMO Y AFINES	Y	90.05	4.21
13	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = MATEMATICAS Y CIENCIAS NATURALES	Y	87.52	0.83
14	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = CIENCIAS DE LA SALUD	Y	78.24	1.43
15	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS NATURALES Y TECNICAS y estu_metodo_prgm = PRESENCIAL y inst_prog_acreditado = PROGRAMA ACREDITADO	Y	78.24	2.22
16	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES	Y	76.37	3.91
17	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS NATURALES Y TECNICAS y estu_metodo_prgm = PRESENCIAL y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_genero = M	Y	70.52	9.26
18	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = OFICIAL y estu_pers_cargo = NO y num_instituciones_zona = ALTO NUMERO INSTITUCIONES ZONA	Y	69.92	1.13
19	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = CIENCIAS SOCIALES Y HUMANAS y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	68.04	2.15

20	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS NATURALES Y TECNICAS y estu_metodo_prgm = PRESENCIAL y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_genero = F y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	65.97	1.77
21	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = BELLAS ARTES	Y	64.48	1.00
22	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES y estu_genero = M y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	63.46	2.03

Figura 9. Reglas más representativas de razonamiento cuantitativo

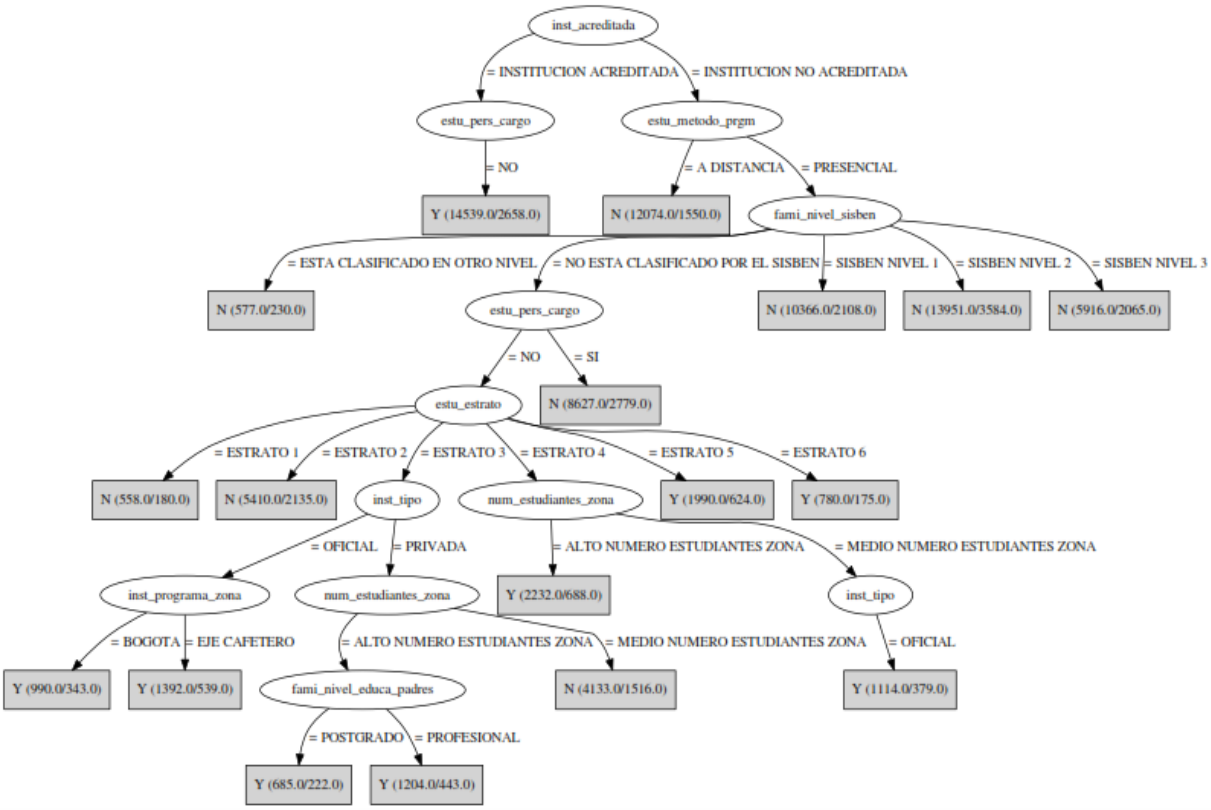

Figura 10. Árbol para inglés podado

```
=== 10 Fold Cross Validation ===
=== Summary ===
Correctly Classified Instances          69600          71.7925 %
Incorrectly Classified Instances        27346          28.2075 %
Kappa statistic                          0.3931
Mean absolute error                      0.3815
Root mean squared error                  0.437
Relative absolute error                 78.6852 %
Root relative squared error             88.7538 %
Coverage of cases (0.95 level)          99.9979 %
Mean rel. region size (0.95 level)      99.999 %
Total Number of Instances               96946

=== Confusion Matrix ===
a       b <-- classified as
48836   8073 | a = N
19273   20764 | b = Y
```

Figura 11. Precisión y matriz de confusión del árbol de inglés

	Antecedente	Consecuente	Confianza (%)	Soporte (%)
1	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = A DISTANCIA	N	87.16	12.45
2	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = SISBEN NIVEL 1	N	79.66	10.69
3	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = SISBEN NIVEL 2	N	74.31	14.39
4	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = SI	N	67.79	8.90
5	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 1	N	67.74	0.58
6	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = SISBEN NIVEL 3	N	65.09	6.10
7	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = PRIVADA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA	N	63.32	4.26
8	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 2	N	60.54	5.58
9	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = ESTA CLASIFICADO EN OTRO NIVEL	N	60.14	0.60
10	inst_acreditada = INSTITUCION ACREDITADA y estu_pers_cargo = NO	Y	81.72	15.00
11	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 6	Y	77.56	0.80
12	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 4 y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	69.18	2.30
13	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 5	Y	68.64	2.05
14	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = PRIVADA y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA y fami_nivel_educa_padres = POSTGRADO	Y	67.59	0.71
15	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 4 y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_tipo = OFICIAL	Y	65.98	1.15
16	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = OFICIAL y inst_programa_zona = BOGOTA	Y	65.35	1.02
17	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = PRIVADA y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA y fami_nivel_educa_padres = PROFESIONAL	Y	63.21	1.24
18	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = OFICIAL y inst_programa_zona = EJE CAFETERO	Y	61.28	1.44

Figura 12. Reglas más representativas de inglés