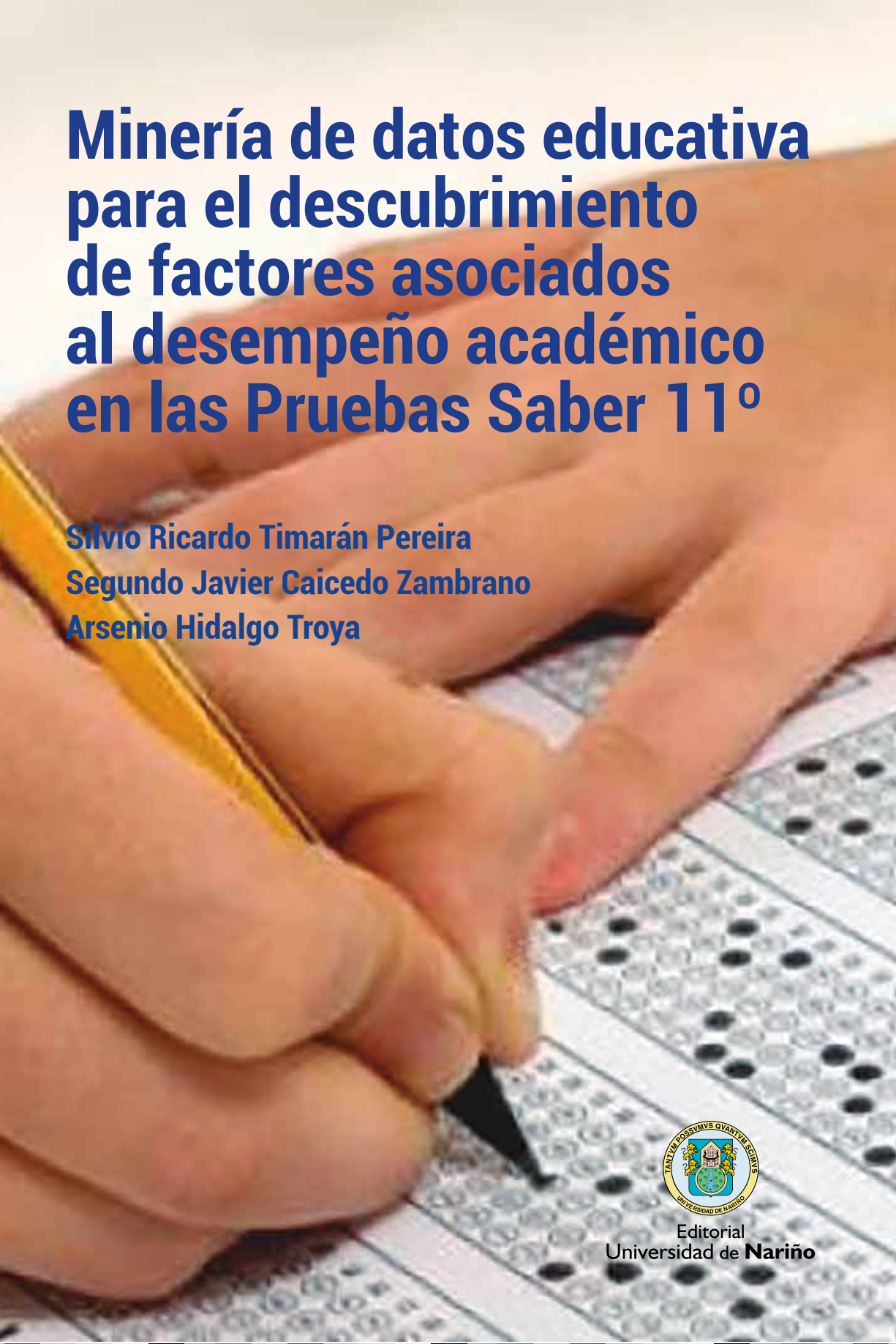
Minería de datos educativa para el descubrimiento de factores asociados al desempeño académico en las Pruebas Saber 11º

Silvio Ricardo Timarán Pereira
Segundo Javier Caicedo Zambrano
Arsenio Hidalgo Troya



Editorial
Universidad de Nariño

Editorial
Universidad de **Nariño**

MINERÍA DE DATOS EDUCATIVA PARA EL DESCUBRIMIENTO DE FACTORES ASOCIADOS AL DESEMPEÑO ACADÉMICO EN LAS PRUEBAS SABER 11º

MINERÍA DE DATOS EDUCATIVA PARA EL DESCUBRIMIENTO DE FACTORES ASOCIADOS AL DESEMPEÑO ACADÉMICO EN LAS PRUEBAS SABER 11º

Silvio Ricardo Timarán Pereira
Segundo Javier Caicedo Zambrano
Arsenio Hidalgo Troya

Editorial
Universidad de Nariño

Timarán, Silvio Ricardo

Minería de datos educativa para el descubrimiento de factores asociados al desempeño académico en las pruebas saber 11 / Silvio Ricardo Timarán, Segundo Javier Caicedo Zambrano, Arsenio Hidalgo Troya.- - 1ª.ed.- -San Juan de Pasto: Editorial Universidad de Nariño, 2021.

173p.: fig; tab.
Referencias Bibliográficas

ISBN: 978-958-5123-81-6 impreso
ISBN: 978-958-5123-80-9 digital

1. Análisis de datos - educación 2. Metodología CRISP-DM-Técnica Minería de Datos (Educación Nariño) 3. Desempeño académico (pruebas Saber 11 2015-2016) - Nariño. 4. Pruebas-meidciones sistema educativo (Nariño - Colombia)

373.126 T582 – SCDD-Ed.22 Biblioteca Alberto Quijano Guerrero

Minería de datos educativa para el descubrimiento de factores asociados al desempeño académico en las Pruebas Saber 11º

ISBN: 978-958-5123-81-6 impreso
ISBN: 978-958-5123-80-9 digital
Primera edición

Impresión:
Graficolor Pasto SAS
Calle 18 No. 29-67 Tel. 7310652
graficolorpasto@hotmail.com

Fecha de publicación: Junio de 2021
San Juan de Pasto, Nariño, Colombia

Agradecimientos

Al Sistema de Investigaciones de la Universidad de Nariño por financiar esta investigación y la publicación de este libro.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

Actualmente, el Ministerio de Educación Nacional (MEN) concibe el objetivo de la educación como el desarrollo de determinadas competencias y, en consecuencia, a estas como el objeto de la evaluación. Dentro de las diferentes competencias que pueden desarrollarse a lo largo del proceso educativo hay una categoría que merece especial atención: la de las competencias genéricas, entendidas como aquellas que resultan indispensables para el desempeño social, laboral y cívico de todo ciudadano, independientemente de su oficio o profesión. Contrastan con las competencias (no-genéricas) propias de oficios o actividades laborales particulares, que resultan de un entrenamiento especializado (Icfes, 2013).

Según Fernández (2005), es función principal de la evaluación en la educación, orientar y apoyar las acciones de mejoramiento de la calidad mediante la obtención, análisis e interpretación de información válida y confiable. En efecto, una adecuada evaluación, que tome en consideración los avances de las ciencias de la cognición, de la pedagogía y de la administración, aporta elementos para una acertada toma de decisiones en los distintos ámbitos educativos tales como: los procesos de enseñanza-aprendizaje, la formulación de políticas, programas y proyectos, la asignación de recursos y el perfeccionamiento de los procesos curriculares, pedagógicos y de gestión (Timarán, Caicedo & Hidalgo, 2019).

El Instituto Colombiano para la Evaluación de la Educación - Icfes es un organismo encargado de la evaluación de la educación en todos sus niveles y de adelantar investigaciones sobre factores que inciden en la calidad educativa con la finalidad de ofrecer información que contribuya al mejoramiento de esta. Actualmente el Icfes diseña y aplica las pruebas Saber 3º, Saber 5º, Saber 9º, Saber 11º, con las cuales evalúa la Educación Básica y Media; y Saber Pro, para evaluar la Educación Superior.

El Examen de Estado de la educación media, Saber 11º, deben presentarlo estudiantes que se encuentren finalizando el grado undécimo, con el fin de obtener resultados oficiales para efectos de ingreso a la educación superior. También pueden presentarlo quienes ya hayan obtenido el título de bachiller o hayan superado el examen de validación del bachillerato, de conformidad con las disposiciones vigentes. En esta investigación únicamente se tuvo en cuenta los primeros.

Según el Decreto 869 de 2010, los objetivos de esta prueba son: seleccionar estudiantes para la educación superior; monitorear la calidad de la formación que ofrecen los establecimientos de educación media y producir información para la estimación del valor agregado de la educación superior (Icfes, 2014).

El examen evalúa cinco componentes basados en las aptitudes que deben desarrollar los educandos según los estándares básicos de competencias (MEN, 2006): lectura crítica, matemáticas, sociales y ciudadanas, inglés y ciencias naturales.

En el estudio de Posada & Mendoza (2014) se concluye que los resultados de pruebas nacionales e internacionales muestran que Colombia posee un sistema educativo con bajos logros académicos de sus estudiantes, en cada uno de los niveles de estudio. Esta situación es crítica, pues de continuar persistiendo esos rendimientos académicos en la mayor parte del estudiantado colombiano, los rendimientos asociados a las economías de escala entre el capital físico y el capital humano seguirán llevando al país por una senda de desarrollo restringido y bajo crecimiento económico.

El desempeño académico, como un fenómeno complejo y multidimensional, está documentado por diferentes autores; por ejemplo, Nieto (2008), explica que este tema amerita que se estudie mediante diferentes tipos y modelos de investigación: exploratorio, descriptivo y explicativo, en los cuales se relacionan dimensiones que agrupan muchas variables. Según el autor, a pesar de la gran cantidad de información existente relacionada con el tema, ésta no es concluyente.

Montes y Lerner (2012), consideran que el desempeño académico no se explica únicamente por las calificaciones o puntajes que obtienen los estudiantes; asumen la existencia de otros factores que se pueden agrupar en cinco dimensiones, a saber: académica, económica, familiar, personal e institucional.

Ahora bien, Tonconi (2010), indica que el desempeño académico concebido como resultado, no siempre da cuenta de las competencias que logran los estudiantes en el proceso de su formación; pues, en general, el esfuerzo del estudiante y la calidad del proceso de su formación, no se relacionan directamente con los resultados que obtienen; según el autor, es necesario un concepto que relacione e incluya el proceso del estudiante y sus condiciones socioeconómicas.

Por su parte, según Castaño (2004), existen cuatro grupos de variables que determinan el desempeño académico: sociodemográficas, académicas, socioeconómicas, e institucionales. Para Porto y Di Gresia (2004), constituyen factores explicativos del desempeño académico: el género, las mujeres tienen mejor desempeño que los hombres; la edad, los jóvenes se desempeñan mejor que los adultos; y el nivel educativo de los padres, a mayor nivel, mejor desempeño.

En la investigación realizada por Navarro (2003), se destaca que el desempeño académico constituye una red dinámica de articulaciones cognitivas, cuyos rasgos característicos diferencian los resultados de cualquier proceso de enseñanza y de aprendizaje.

El interés por conocer los factores que determinan el rendimiento académico de los estudiantes tuvo relevancia con la publicación

del Informe Coleman et al. (1966), en el cual se concluyó que el rendimiento escolar en los Estados Unidos estaba influenciado en gran medida por las características socioeconómicas de los estudiantes y por el hecho que, poco o nada tenían que ver en el desempeño académico, las variables asociadas a la institución educativa. Estos resultados generaron controversia, ya que muchos críticos del tema no concebían que las variables asociadas a la institución educativa no tuvieran influencia en el desempeño académico.

Uno de esos estudios que controvierten lo expuesto por Coleman et al. (1996) es el realizado en Colombia por Correa (2004) en la ciudad de Cali; allí, el autor en el estudio incluye características individuales y aspectos asociadas a la institución. El autor concluye que las variables asociadas al plantel son muy significativas a la hora de determinar el rendimiento académico de los estudiantes. Por su parte, Gamboa, Casas y Piñeros (2003), realizaron mediciones que intentan explicar los niveles de aporte de los distintos factores y agentes involucrados en el desempeño del estudiante, entre los cuales están la familia, la institución educativa, los docentes, el curso, los compañeros, entre otros.

En el Sistema de Universidades Estatales del Caribe Colombiano, los autores Villalba y Salcedo (2008), realizaron la investigación "Rendimiento académico en el nivel de educación media, como factor asociado al rendimiento académico en la universidad". Se buscó establecer la relación entre el desempeño académico en el nivel medio versus el desempeño académico en la universidad. Es un estudio descriptivo correlacional, donde se estudiaron las siguientes variables: desempeño académico en educación media, indicadores de desempeño y el desempeño en la universidad. Los resultados permitieron determinar que existe una relación estadísticamente positiva ($\rho = 0{,}417$) entre el rendimiento académico en educación media y en la universidad. Si bien no se puede afirmar que el rendimiento académico en la educación media, constituye un predictor del desempeño en la universidad, si se puede afirmar que hay una tendencia que las estudiantes mantengan o mejoren el rendimiento académico en la universidad con relación al obtenido en educación media.

La Secretaria de Educación Distrital de Bogotá, adelantó el estudio liderado por Bodensiek (2010), sobre los factores que influyen en el rendimiento escolar, cuyo objetivo es discutir sobre los factores asociados al desempeño académico exitoso, los cuales se han encontrado en diferentes estudios a nivel internacional, con el fin de identificar los que influyen en los resultados del examen de Estado para ingreso a la Educación Superior. El estudio se realizó con estudiantes de grado once de 55 colegios públicos y 71 colegios privados de Bogotá, donde se contemplan variables endógenas y exógenas. Dentro de las primeras está el género, la edad, la frecuencia de estudio y hábitos, y trayectoria de la vida académica; en las segundas, se relacionan la comunidad, es decir, el entorno inmediato, la familia, su composición, clima familiar, nivel de ingresos económicos, la ocupación y el nivel educativo de los padres, seguridad, infraestructura, recursos disponibles para el aprendizaje, entre otros. Entre las variables que sobresalen por su asociación al desempeño académico, están las siguientes: estrato social del estudiante, ingresos económicos, nivel educativo de los padres, características profesionales de los docentes y su vinculación a la institución.

Según Rodríguez, Benavides y Riascos (2019), los estudios que se han realizado hasta el momento en Colombia para determinar el rendimiento, se han centrado en el uso de técnicas estadísticas tradicionales, dejando información valiosa sin explotar y que por general está oculta, que se puede descubrir utilizando un tratamiento complejo de los datos, que es posible con la minería de datos. De acuerdo a Timarán et al. (2016), la estadística plantea hipótesis que deben ser validadas a partir de los datos disponibles y la minería de datos descubre patrones no previstos desde la estadística. La minería de datos no se puede utilizar para confirmar o rechazar hipótesis, su objetivo es explorar datos, darles sentido, convertir en conocimiento un volumen de datos que por sí solos no aportan a la toma de decisiones. En este contexto, la minería de datos emerge como el siguiente paso evolutivo en el proceso de análisis de datos.

Para los autores Pérez y Santin (2007), la minería de datos es el proceso de descubrimiento de nuevas y significativas relacio-

nes, patrones y tendencias utilizando diferentes tareas a partir de grandes volúmenes de datos. Estas tareas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando técnicas como clasificación, clustering, patrones secuenciales y asociaciones, entre otras.

Al proceso de minería de datos se le conoce también como el proceso de descubrimiento de conocimiento en bases de datos KDD (del inglés Knowledge Discovery in Databases) (Fayard, Piatetsky-Shapior & Smith,1996), (Timarán, 2009). KDD es un proceso interactivo e interactivo compuesto por varias fases de las cuales una de ellas es la minería de datos. Implica no solo obtener los modelos o patrones (en la fase de minería de datos), sino seleccionar, limpiar, transformar los datos e interpretar y evaluar los patrones para convertirlos en conocimiento y de esta manera puedan ser útiles para ayudar a la toma de decisiones efectivas en las organizaciones (Agrawal & Srikant, 1994), (Chen, Han, & Yu, 1996), (Piatetsky-Shapiro, Brachman, Khabaza, Kloesgen, & Simoudis, 1996), (Han, Kamber & Pei, 2012).

La minería de datos y la obtención de modelos se pueden concebir como aprendizaje a partir de datos. El aprendizaje puede ser supervisado y no supervisado. Dentro de la minería de datos se encuentran diferentes tipos de tareas, cada una de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos (Hernández, Ramírez & Ferri, 2005). Se clasifican en tareas predictivas y descriptivas.

Las tareas predictivas pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables o atributos denominadas variables independientes o predictivas. Son procesos de aprendizaje supervisado porque se parte de un conocimiento previo de los datos. Se supone que el usuario conoce con certeza las categorías de cada registro con el que se cuenta. En el aprendizaje supervisado, se pretende clasificar a los registros de datos en alguna de las categorías predefinidas. En el desarrollo de un modelo predictivo, la meta es crear un clasificador que pueda predecir la categoría de un registro basándose en los datos con los

que cuenta. Estas tareas se fundamentan en la identificación de relaciones entre variables en eventos pasados, para luego explotar dichas relaciones y predecir posibles resultados en futuras situaciones. Al proceso de transferir el conocimiento al modelo se le conoce como entrenamiento. Los datos utilizados en este proceso se llaman conjunto de entrenamiento. Son ejemplos de este tipo de tareas de minería de datos la regresión, la clasificación y las series de tiempos entre otras.

Las tareas descriptivas identifican patrones que explican o resumen los datos. Sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Son procesos de aprendizaje no supervisados, ya que se buscan automáticamente grupos de valores para que después el usuario intente encontrar las correspondencias entre esos grupos seleccionados automáticamente y las categorías que le puedan ser de interés. En el aprendizaje no supervisado se cuenta con un conjunto de datos, pero no existe información sobre cómo agrupar los datos en conglomerados. Basándose en similitudes entre los datos, se empiezan a desarrollar conglomerados o clústeres entre los datos hasta que comienzan a aparecer un conjunto de patrones diferenciables. Aquí no hay diferenciación entre variables independientes o predictores y variables dependientes. Entre las tareas descriptivas de minería de datos se pueden citar asociación, agrupamiento o clustering, patrones secuenciales y correlaciones.

La aplicación de las técnicas y herramientas de la minería de datos en los diferentes contextos educativos se conoce como minería de datos educativa EDM (del inglés Educational Data Mining).

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 et al., 2013). Las instituciones de educación pueden usar la minería de datos en la educación para hacer análisis comprensivos de las características de sus estudiantes, métodos evaluativos, develando procesos exitosos o, por el contrario, detectando fraudes o inconsistencias (Valero et al, 2010).

Entre los estudios más recientes de la minería de datos aplicados al ámbito del desempeño académico, está el realizado por Osmanbegović y Suljić (2012) quienes comparan diferentes métodos y técnicas de predicción a partir de variables sociodemográficas de estudiantes del departamento de Economía de la Universidad de Tuzla (Bosnia and Herzegovina), sus resultados demuestran que el clasificador Naive Bayes arroja resultados superiores en la predicción del desempeño académico. Por su parte, Ahmad, Ismail, y Aziz (2015) utilizan las técnicas de árboles de decisión, Naive Bayes y reglas de clasificación para predecir el rendimiento académico en el primer año de estudiantes de pregrado en Ciencias computacionales. Las variables que se tuvieron en cuenta corresponden a variables demográficas y familiares, obteniendo una tasa de precisión de 71%. En la investigación de Khobragade y Mahadik (2015), con el algoritmo Naive Bayes y utilizando datos socioeconómicos y académicos de los estudiantes logran identificar 11 características más importantes en el desempeño académico, alcanzando una tasa de precisión en la predicción del 87.12%. Por otro lado, Badr et al. (2016), utilizan métodos de clasificación basados en reglas de asociación para construir un clasificador que permita evaluar el desempeño de estudiantes universitarios en cursos de programación, logrando tasas de precisión del 63%-67%. Los investigadores Hamsa, Indiradevi, y Kizhakkethottam (2016) desarrollaron un modelo de predicción del éxito académico, usando árboles de decisión y el algoritmo genético fuzzy. Asimismo, Costa et al. (2016) evalúan el poder de predicción de las técnicas de minería de datos Support Vector Machine, árboles de decisión, redes neuronales y Naive Bayes, con el fin de predecir el fracaso académico en cursos de programación. Sus resultados muestran que las técnicas utilizadas fueron capaces de identificar tempranamente aquellos estudiantes con riesgo de fracasar, especialmente, la técnica de Support Vector Machine supera en términos de error de prueba a los otros métodos mencionados.

En Colombia, una aplicación de la EDM en el área de desempeño académico en las pruebas Saber, fue el estudio realizado por Timarán et al. (2016), cuyo objetivo fue descubrir patrones

asociados al desempeño académico en las competencias genéricas de los estudiantes de programas profesionales que presentaron las pruebas Saber Pro 2011-2. En el estudio utilizaron la técnica *clasificación* basada en árboles de decisión. En los patrones descubiertos se destaca que, la acreditación institucional y la modalidad de estudio, son dos factores importantes asociados al desempeño académico de los estudiantes en las pruebas Saber Pro 2011-2.

Por otra parte, Blanco (2015) en su tesis de maestría, aplica la minería de datos en la educación con el fin de analizar el desempeño académico de los estudiantes del departamento del Cesar que presentaron las Pruebas Saber 11º en el año 2012-2, para el ingreso a la Educación Superior utilizando la técnica de *clustering*.

En el estudio de Rodríguez et al. (2018), se pretende establecer cuáles son los factores que más contribuyen a predecir el éxito académico de los estudiantes admitidos en la Universidad de los Andes en el período 2015-2017, utilizando los datos de la prueba Saber 11º de esos años e información adicional sobre características sociodemográficas de los estudiantes y variables sobre los establecimientos educativos.

En la región, no se han planteado investigaciones que analicen el desempeño de los estudiantes, de las diferentes instituciones educativas de educación media del departamento de Nariño (Colombia) en las pruebas Saber 11º, utilizando técnicas de minería de datos educativa, por lo cual, el estudio que se reporta en este texto, resultó ser muy pertinente.

El objetivo de la investigación que se reporta en este libro, fue descubrir factores asociados al desempeño académico en las competencias que evalúan las pruebas Saber 11º de los estudiantes de instituciones de educación media del departamento de Nariño (Colombia), que presentaron este examen, en los años 2015 y 2016, estudio que se realizó a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del Icfes, utilizando técnicas de minería de datos educativa. La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental y siguiendo como metodología CRISP-DM, la guía de referencia ampliamente utilizada en el

desarrollo de proyectos de minería de datos (Chapman et al., 2000). Se descubrieron patrones asociados al buen o mal desempeño académico en las pruebas Saber 11º. Se documentó y socializó el conocimiento generado que puede ser de utilidad para soportar la toma de decisiones de las instituciones educativas y gubernamentales acerca del mejoramiento de la calidad de la educación secundaria en el departamento de Nariño.

La presente investigación proyectó dar respuesta a la pregunta: ¿Cuáles son los factores socioeconómicos, académicos e institucionales asociados al desempeño académico de los estudiantes colombianos que presentan las pruebas Saber 11º?

Este libro se organiza por capítulos. En el capítulo I, se explica en que consiste las pruebas Saber 11º, sus objetivos y competencias que evalúa. En el capítulo II, se conceptualiza acerca de la minería de datos y su relación con la minería de datos educativa. En el capítulo III, se presenta un análisis exploratorio de los datos y se muestran algunas tendencias sobre el desempeño académico con relación a las competencias genéricas. En el capítulo IV, se muestra cómo se desarrolló el proyecto aplicando las diferentes fases de la metodología CRISP-DM. En el capítulo V, se evalúan, interpretan y discuten los resultados obtenidos; finalmente, en el capítulo VI, se presentan conclusiones y recomendaciones de acuerdo a los resultados obtenidos.

Capítulo I

PRUEBAS SABER 11º Y ESTUDIOS RELACIONADOS

Según la Constitución Política Nacional (Art. 67, C.P.C. de 1991), la educación es un derecho y un servicio público que tiene una función social, a través de la cual, se busca el acceso al conocimiento, a la ciencia, a la técnica, y a los demás bienes y valores de la cultura; así que, corresponde al Estado regular y ejercer la inspección y vigilancia de la educación, con el fin de garantizar su calidad, el cumplimiento de los objetivos, en la perspectiva de mejorar la formación moral, intelectual y física de los colombianos.

La Ley 1324 le confiere al Instituto Colombiano para Evaluación de la Educación (Icfes) la misión de evaluar, mediante exámenes externos estandarizados, la formación que se ofrece en el servicio educativo en los distintos niveles. Estos exámenes están estandarizados, en la medida en que las condiciones de aplicación y el procesamiento de los resultados son uniformes. También establece que MEN define lo que debe evaluarse en estos exámenes (Icfes, 2014).

Teniendo en cuenta el Plan Nacional Decenal de Educación 2006-2016, el Icfes ha avanzado en la alineación del Sistema Nacional de Evaluación Externa Estandarizada (SNEE), a través

de la reestructuración de los exámenes: en 2009 con un nuevo diseño de Saber 3º, 5º y 9º; en 2010 con el rediseño de Saber Pro; en 2014 con los cambios en Saber 11º. La alineación posibilita la comparación de los resultados en distintos niveles educativos, ya que los diferentes exámenes evalúan unas mismas competencias en algunas de las áreas que los conforman, a saber, las competencias genéricas (Icfes, 2014).

En el caso particular del examen Saber 11º, que se aplica desde el segundo semestre de 2014, la alineación consiste en que se introdujo una prueba de competencias ciudadanas; se distinguió en la prueba de matemáticas entre lo que es genérico y lo que no lo es y, finalmente, se fusionaron diferentes pruebas en torno a las competencias genéricas que evalúan en común, así: Lenguaje y Filosofía se fusionaron en una prueba de Lectura Crítica; Física, Química y Biología se fusionaron en una prueba de Ciencias Naturales (que incluye el componente de Ciencia, Tecnología y Sociedad establecido en los Estándares); y las competencias ciudadanas se evalúan mediante una prueba de Sociales y Ciudadanas (Icfes, 2014).

El MEN orienta el diseño de los planes de estudio, la enseñanza en el aula y establece, en su conjunto, expectativas de calidad sobre lo que deben aprender los estudiantes a lo largo de su formación. Con base en esto, el Icfes busca evaluar las competencias desarrolladas durante su formación básica y media a través del examen Saber 11º.

El Decreto 869 de 2010 establece como objetivos del examen Saber 11º:

- Comprobar el grado de desarrollo de las competencias de los estudiantes que están por finalizar el grado undécimo de la educación media.
- Proporcionar elementos al estudiante para la realización de su autoevaluación y el desarrollo de su proyecto de vida.
- Proporcionar a las instituciones educativas información pertinente sobre las competencias de los aspirantes a ingresar a programas de educación superior, así como sobre las de quienes

son admitidos, que sirva como base para el diseño de programas de nivelación académica y prevención de la deserción en este nivel.

- Monitorear la calidad de la educación de los establecimientos educativos del país, con fundamento en los estándares básicos de competencias y los referentes de calidad emitidos por el Ministerio de Educación Nacional.
- Proporcionar información para el establecimiento de indicadores de valor agregado, tanto de la educación media como de la educación superior.
- Servir como fuente de información para la construcción de indicadores de calidad de la educación, así como para el ejercicio de la inspección y vigilancia del servicio público educativo.
- Proporcionar información a los establecimientos educativos que ofrecen educación media para el ejercicio de la autoevaluación y para que realicen la consolidación o reorientación de sus prácticas pedagógicas.
- Ofrecer información que sirva como referente estratégico para el establecimiento de políticas educativas nacionales, territoriales e institucionales.

El examen Saber 11º se compone de cinco pruebas (Icfes, 2018): lectura crítica, matemáticas, sociales y ciudadanas, inglés y ciencias naturales. Estas pruebas evalúan competencias, entendidas como las habilidades necesarias para aplicar de manera flexible los conocimientos en diferentes contextos. En este sentido, para realizar este examen no se requiere solamente saber conceptos o datos, sino saber cómo emplearlos para resolver problemas en situaciones de la vida cotidiana.

La prueba de Lectura Crítica evalúa las competencias necesarias para comprender, interpretar y evaluar textos que pueden encontrarse en la vida cotidiana y en ámbitos académicos no especializados. Se espera que los estudiantes que culminan la educación media cuenten con las capacidades lectoras para tomar posturas críticas frente a esta clase de textos (Icfes, 2016). Esta

prueba evalúa tres competencias que recogen, de manera general, las habilidades cognitivas necesarias para leer de manera crítica: identificar y entender los contenidos locales que conforman un texto; comprender cómo se articulan las partes de un texto para darle un sentido global; reflexionar en torno a un texto y evaluar su contenido. Las dos primeras competencias se refieren a la comprensión, ya sea a nivel local o global, del contenido de un texto, y la tercera a la aproximación propiamente crítica frente a este (Icfes, 2018).

La prueba de Matemática evalúa las competencias de los estudiantes para enfrentar situaciones que pueden resolverse con el uso de algunas herramientas matemáticas (Icfes, 2016). Tanto las competencias definidas para la prueba como los conocimientos matemáticos que el estudiante requiere para resolver las situaciones planteadas se contemplan en las definiciones de los Estándares Básicos de Competencias de Matemáticas del MEN. En esta prueba, se integran competencias y contenidos en distintas situaciones o contextos, en las cuales las herramientas matemáticas cobran sentido y son un importante recurso para la comprensión, la transformación, la justificación y la solución de los problemas involucrados (Icfes, 2018). De acuerdo con lo anterior, se integran competencias y contenidos en distintas situaciones o contextos, en los cuales las herramientas matemáticas cobran sentido y son un importante recurso para la comprensión de situaciones, la transformación de información, la justificación de afirmaciones y la solución de problemas. En esta prueba se definen tres competencias que recogen los elementos centrales de los procesos que se describen en los estándares básicos de competencias: interpretación y representación; formulación y ejecución; argumentación (MEN, 2006).

La prueba de Sociales y Ciudadanas evalúa los conocimientos y habilidades del estudiante que le permiten comprender el mundo social desde la perspectiva propia de las ciencias sociales y situar esta comprensión como referente del ejercicio de su papel como ciudadano. Evalúa también su habilidad para analizar distintos eventos, argumentos, posturas, conceptos, modelos, dimensiones y contextos, así como su capacidad de reflexionar y emitir juicios críticos sobre estos (Icfes, 2016). En esta prueba se evalúan tres

competencias que están alineadas con lo propuesto en los estándares básicos de competencias en ciencias sociales y competencias ciudadanas: pensamiento social; interpretación y análisis de perspectivas; pensamiento reflexivo y sistémico (MEN, 2006).

La prueba de inglés, evalúa la competencia para comunicarse efectivamente en inglés. A su vez, en relación con el Marco Común Europeo de Referencia para las lenguas (MCER), se clasifican a los evaluados en 5 niveles de desempeño: A-, A1, A2, B1 y B+. Teniendo en cuenta que, en Colombia, existe población que se encuentra por debajo del primer nivel del MCER (A1), se incluyó en la prueba de inglés un nivel inferior a A1, denominado A-, que corresponde a aquellos desempeños mínimos que involucran el manejo de vocabulario y estructuras básicos. De igual forma, se incluye un nivel superior al B1 para aquellos estudiantes que superan lo evaluado en este nivel, denominado B+. La prueba busca que el estudiante demuestre sus habilidades comunicativas a nivel de lectura y uso del lenguaje (Icfes, 2018). La prueba busca que el estudiante demuestre sus habilidades comunicativas a nivel de lectura y uso del lenguaje. El MEN propuso como meta para el año 2019 alcanzar el nivel B1 en la población de educación media (Icfes, 2016).

La prueba de Ciencias Naturales evalúa la capacidad que tiene el estudiante de comprender y usar nociones, conceptos y teorías de las ciencias naturales en la solución de problemas. Evalúa también la habilidad del estudiante para explicar cómo ocurren algunos fenómenos de la naturaleza basado en observaciones, patrones y conceptos propios del conocimiento científico. La prueba, además, involucra en la evaluación el proceso de indagación, que incluye observar y relacionar patrones en los datos para derivar conclusiones de fenómenos naturales. La prueba de ciencias naturales no pretende evaluar conocimientos científicos en sentido estricto, sino la capacidad de los estudiantes para reconstruir significativamente el conocimiento existente, razonar, tomar decisiones, resolver problemas, pensar con rigurosidad y valorar de manera crítica el conocimiento y sus consecuencias en la sociedad y en el ambiente (Icfes, 2018).

Los estudios de factores asociados tienen como objetivo identificar las variables que más influencian el rendimiento escolar de los estudiantes, con el fin de avanzar en la construcción y entendimiento de los aspectos que inciden en la calidad de la educación. En Colombia se han realizado varios estudios que buscan determinar los factores que influyen en el rendimiento académico de los estudiantes en las pruebas Saber 11º. En este sentido, Piñero y Rodríguez (1998) realizaron un estudio acerca del efecto de los insumos escolares (aspectos familiares, entorno escolar y aptitud personal) en la educación secundaria en Colombia, y su efecto sobre el rendimiento académico de los estudiantes, medido por medio de las pruebas de estado Icfes (hoy Saber 11º), donde el principal resultado fue que el nivel socioeconómico del estudiante tiene un efecto directamente proporcional sobre el rendimiento académico de este. Por tal motivo, mencionan la importancia de concientizar a las personas, con la responsabilidad compartida que tiene la familia, la comunidad y la escuela en el proceso educativo.

En el estudio efectuado por Gaviria y Barrientos (2001), los autores analizaron los resultados de las pruebas de estado de 1999, en el cual, encontraron que las características asociadas a la institución educativa afectan de manera importante el rendimiento académico y que lo hacen en mayor medida que las variables socioeconómicas; además, en él, se reconoce que el nivel educativo de los padres tiene un efecto importante en el desempeño académico. Encontraron, además, que existe una brecha importante entre los resultados de instituciones oficiales y privadas. Estos hallazgos ponen en cuestión los resultados de Coleman et al (1966), en lo atinente a la influencia de las variables institucionales en el desempeño académico de los estudiantes.

Dichos autores también estudiaron estos aspectos en la ciudad de Medellín (Barrientos, 2008), donde analizaron las pruebas Icfes para el periodo 2004 y 2006. Encontraron que el efecto del colegio parece cada vez ser menos fuerte que variables individuales; además, evidenciaron que las variables relacionadas con el colegio afectan más a los estudiantes en instituciones privadas que en las públicas. Comparan además sus resultados con el estudio realizado en Bogotá, concluyendo que los resultados son similares, en cuanto a la baja calidad de la educación pública es general.

En este mismo sentido, los resultados obtenidos en el estudio "Determinantes del rendimiento académico en Colombia: pruebas Icfes Saber 11º, 2009" (Chica, Galvis y Ramírez; 2009), muestran la relevancia que tienen las variables socioeconómicas en el desempeño en las áreas de matemáticas y lenguaje, variables que son consideradas fundamentales en el aprendizaje; en particular, las variables nivel de ingreso y nivel de escolaridad de los padres, que presentan un impacto positivo y significativo en el resultado de las pruebas. Igualmente se encontró un impacto significativo de la jornada académica, pues, los bachilleres de jornada completa obtienen puntajes más altos comparados con los estudiantes de otras jornadas. Estos resultados son coincidentes con el estudio de Santín (2001), el cual pone en evidencia que determinados factores socioeconómicos se asocian al desempeño académico, al igual que ciertas características familiares, destacándose particularmente, el nivel educativo de los padres; y también son coincidentes con los estudios de Espínola y Martínez (1996), quienes evidenciaron que la familia y las prácticas instruccionales del profesor, son claramente las variables más importantes en cuanto a su efecto sobre el logro educativo de los estudiantes de educación básica. Dichos hallazgos destacan la importancia que tienen las variables individuales y las institucionales en los estudios relacionados con el desempeño académico de los estudiantes.

En la investigación de Sánchez y Otero (2012) denominada "Educación y Reproducción de la desigualdad en Colombia", se realizó una comparación socioeconómica de los resultados obtenidos por los estudiantes de educación media en las pruebas Saber 11º, entre los años 2008 al 2011 en el componente matemático. Se observó que los de estratos altos obtienen mejores resultados en las pruebas académicas estandarizadas que sus pares de estratos bajos. El puntaje medio en el área de matemática de la prueba Saber 11º es proporcional al estrato. Estos resultados muestran que los estudiantes con condiciones socioeconómicas favorables, dado que obtienen mejores puntajes, son los que pueden acceder a una educación superior de mayor calidad, pues las mejores universidades, tanto públicas como privadas, exigen altos puntajes en las pruebas Saber 11º para el ingreso. En contraste, los estudiantes con condiciones socioeconómicas desfavorables no

sólo obtienen una educación de menor calidad, sino que esta no parece estar mejorando a través del tiempo.

Corsi et al. (2012), en su trabajo titulado "Factores asociados a desempeños destacados y no destacados en las pruebas Saber 11 (2009-2)", utilizaron un modelo de regresión logística para identificar cuáles eran los factores que podrían tener un nivel de influencia importante en el desempeño académico. Los resultados sugieren que los estudiantes que obtuvieron resultados más altos fueron hombres, con mayores ingresos familiares, no trabajan, no pertenecen a una etnia, han estudiado en colegio privado de jornada completa por seis años o más, no perdieron ningún grado en primaria y tienen acceso a internet. Por el contrario, las mujeres fueron las que obtuvieron puntajes más bajos y estaban en riesgo por tener un ingreso familiar bajo, probablemente pertenecen a una etnia, trabajan y no tienen acceso a internet.

En el trabajo de Muñoz (2017) sobre los determinantes de las diferencias en los resultados de las pruebas académicas de estado en la educación media oficial en Bogotá (Colombia) muestra, mediante un análisis descriptivo de los resultados de las pruebas Saber 11º que presentaron los estudiantes de Bogotá en el segundo semestre de año 2016, que los factores de género, jornada, estrato económico, el acceso en el hogar a internet y un computador presentan un significativo distanciamiento entre los grupos de estudiantes que asisten a un colegio de naturaleza oficial y los que lo hacen en colegios de naturaleza privada, así mismo menciona que aquellas instituciones que son de carácter privado son aquellas instituciones que presentan mejores resultados.

Morales (2019) en su trabajo denominado "Factores Determinantes en el Rendimiento de los Estudiantes de la Región Pacífico-colombiana en las Pruebas Saber-11", por medio de una regresión múltiple lineal buscó hallar cuáles eran las características que hacían que un estudiante de la región tuviera un mayor desempeño en las pruebas Saber 11º. Los resultados sugieren que la dedicación a la lectura, al internet, la ubicación geográfica del colegio y si este es bilingüe están entre las principales características que implican buenos resultados en esta prueba.

Capítulo II

EFECTOS DEL DESEMPEÑO ACADÉMICO EN LAS PRUEBAS SABER 11º

Se presenta en este capítulo un análisis estadístico preliminar de los datos con el fin de identificar las características sociodemográficas de los estudiantes de educación media del departamento de Nariño que presentaron las pruebas SABER 11º en el período 2015-2016 y establecer un comparativo del desempeño académico de estos estudiantes, en las cinco competencias de dichas pruebas, con los municipios de todo el departamento, lo que se denominó el "efecto localidad" y también el rendimiento específico en dichas competencias según las subregiones y las variables socioeconómicas: género, edad, estrato social, ingreso familiar, tipo de colegio, jornada, condición en la que vive, condición de la vivienda, condición de las TIC, educación de la madre y educación del padre, para establecer como son las diferencias en las diferentes categorías de dichas variables, lo que hemos denominado el efecto de cada variable.

Para determinar el tamaño del efecto en las diferentes variables se establecen las diferencias estandarizadas de las medias con el estadístico d de Cohen. Con base en este estadístico se estableció

un ranking, en estas competencias, de los municipios y las subregiones en donde están ubicadas las instituciones de educación secundaria de todo el departamento en el período analizado.

Mediante análisis estadístico descriptivo con la calificación media obtenida en las pruebas por estudiantes de cada institución de educación secundaria del departamento de Nariño se explica el rendimiento académico que presentan en las diferentes competencias de las pruebas SABER 11 en el período 2015 a 2016, los efectos "localidad" y "subregión", determinados en gran parte por la calidad de las instituciones en cada municipio y subregión, el relativamente débil poder explicativo que tiene el nivel socioeconómico y las brechas de género encontradas en las diferentes áreas estudiadas. Se utiliza para la clasificación, las diferencias estandarizadas, tomando como referente el más alto promedio alcanzado de cada competencia para establecer el "efecto localidad" y el "efecto subregión" para ubicar la posición de cada municipio y subregión en el contexto departamental. Y de otra parte tomar el promedio de los resultados para cada una de las categorías de las variables sociodemográficas para establecer el "efecto de cada variable".

Analizar los resultados del desempeño académico de las instituciones de educación secundaria constituye un factor relevante, debido a que permite tener evidencias sólidas de la calidad de la educación que brindan estas instituciones en cada municipio y subregión y en la medida que identifican elementos que afectan el rendimiento de los estudiantes, los cuales son importantes de considerar en el momento de diseñar políticas públicas en pro del mejoramiento de los procesos educativos.

En este contexto, el presente trabajo aporta evidencia empírica acerca de los factores que determinan el rendimiento académico de los estudiantes de secundaria del departamento de Nariño. Se utilizan datos de corte transversal concernientes a la prueba SABER 11º - 2015 a 2016 disponibles en las bases de datos del Instituto Colombiano para la Evaluación de la Educación (Icfes)

Inicialmente se describen las características socioeconómicas, las cuales pueden generar posibles brechas de rendimiento académico de los estudiantes de secundaria del departamento de Nariño que presentaron las pruebas SABER 11º en el período 2015-2016, a partir de la información contenida en los formularios de inscripción. Posteriormente, y con el fin de tener una comprensión preliminar de la relación entre los datos, se realiza un análisis de correlación entre las cinco competencias y su resultado global. Finalmente, se establece el efecto que tienen en el desempeño de dichas competencias, las variables: localidad, subregión y las variables socioeconómicas: género, edad, estrato social, ingreso familiar, tipo de colegio, jornada, condiciones en la que vive, condiciones de la vivienda, condiciones de TIC, nivel educativo de la madre y nivel educativo del padre.

2.1 CARACTERÍSTICAS SOCIOECONÓMICAS

En el período 2015-2016, un total de 33.438 estudiantes de Educación Secundaria del departamento de Nariño presentaron las pruebas SABER11º. La Tabla 1 presenta las características socioeconómicas de dichos estudiantes. Por género la mayoría son mujeres, con un 55.7%, por edad el mayor porcentaje se encuentran debajo de 18 años con un 59.7%. En un alto valor el estrato social bajo es el que más prevalece con un 92.3%. Más de la mitad de las familias de los estudiantes (53,2%) tienen ingresos inferiores a un salario mínimo mensual, el 44,4% entre uno y cinco salarios mínimos, únicamente el 1,4% superan los cinco salarios mínimos. El 1,2% vive en hacinamiento crítico y 15,4% en hacinamiento medio y el 60,4% vive en malas condiciones de vivienda. Solamente el 22,6% cuenta con TIC en condiciones regulares, el 74,4% restante en malas condiciones. La mayoría de estudiantes pertenece a colegios públicos (80,6%) y en jornada de la mañana (81,1%). Más de la mitad de las madres y los padres solamente cuentan con estudios de primaria, 57,5% y 59,3% respectivamente, mientras que el 33,2% y 31,4% cuentan con estudios de secundaria y únicamente un 6,5% y 4.6% cuentan con estudios de educación superior.

Tabla 1. ***Características Socioeconómicas de estudiantes de secundaria del departamento de Nariño que presentaron las pruebas SABER 11º en el período 2015-2016.***

Variable Socioeconómica		N	%
Género	Femenino	18.564	55,7%
	Masculino	14.755	44,3%
Grupos de edad	Menor que 18 años	19.969	59,7%
	Entre 18 y 22 años	12.108	36,2%
	Mayor que 22 años	1.361	4,1%
Estrato social	Bajo	30.862	92,3%
	Medio	2.272	6,8%
	Alto	304	0,9%
Ingreso familiar	Menos de 1 SM	17.889	53,5%
	Entre 1 y menos de 5 SM	14.832	44,4%
	Entre 5 y menos de 10 SM	355	1,1%
	10 o más SM	87	0,3%
	Sin dato	275	0,8%
Tipo de Colegio	Privado	3.826	11,4%
	Público	29.612	88,6%
Jornada	Completa u Ordinaria	215	0,6%
	Única	149	0,4%
	Mañana	27.114	81,1%
	Tarde	3.629	10,9%
	Noche	2.331	7,0%
Condición en la que vive	Hacinamiento crítico	391	1,2%
	Hacinamiento medio	5.160	15,4%
	Sin hacinamiento	27.887	83,4%
Condición de la vivienda	Mala	20.197	60,4%
	Regular	4.987	14,9%
	Buena	8.254	24,7%

Variable Socioeconómica		N	%
Condición de las TIC	Mala	25.887	77,4%
	Regular	7.551	22,6%
	Buena	0	0,0%
Nivel Educativo de Madre	Ninguna	662	2,0%
	Primaria	19.217	57,5%
	Secundaria	11.109	33,2%
	Superior	2.175	6,5%
	Sin dato	275	0,8%
Nivel Educativo de Padre	Ninguna	1.303	3,9%
	Primaria	19.827	59,3%
	Secundaria	10.484	31,4%
	Superior	1.549	4,6%
	Sin dato	275	0,8%
Total		**33.438**	**100.0%**

Fuente: elaboración propia

La Tabla 2 presenta los diferentes municipios a los cuales pertenecían las instituciones de Educación Secundaria del departamento de Nariño que presentaron las pruebas SABER 11º en el periodo de estudio. Casi una tercera parte (30,1%) pertenecen a la capital del departamento, Pasto, seguida por los municipios de Tumaco con 13,7%, e Ipiales con 8,3%, y luego con porcentajes por encima del 2%, Cumbal, Túquerres, La Unión y Samaniego.

Tabla 2. ***Municipios de las Instituciones de Educación Secundaria del departamento de Nariño que participaron en las pruebas SABER 11º en el período 2015-2016.***

Municipio	N	%
Pasto	10.087	30,20%
Tumaco	4.565	13,70%
Ipiales	2.773	8,30%
Cumbal	866	2,60%
Túquerres	765	2,30%
La Unión	734	2,20%
Samaniego	687	2,10%

Municipio	N	%
Sandoná	533	1,60%
Barbacoas	521	1,60%
Guachucal	453	1,40%
La Cruz	417	1,20%
San Lorenzo	410	1,20%
Buesaco	393	1,20%
Taminango	392	1,20%
El Charco	383	1,10%
Pupiales	373	1,10%
El Tambo	353	1,10%
Córdoba	345	1,00%
El Tablón	340	1,00%
Olaya Herrera	313	0,90%
Nariño	305	0,90%
San Pablo	304	0,90%
Chachagüí	270	0,80%
Albán (San José)	260	0,80%
Ricaurte	254	0,80%
Guaitarilla	247	0,70%
Los Andes (Sotomayor)	246	0,70%
Linares	240	0,70%
Mallama (Piedrancha)	234	0,70%
Potosí	230	0,70%
Yacuanquer	229	0,70%
Consacá	224	0,70%
Ancuyá	212	0,60%
Imués	209	0,60%
La Florida	208	0,60%
Tangua	207	0,60%
Colón (Génova)	191	0,60%
Leiva	173	0,50%
Francisco Pizarro	169	0,50%
Cuaspud (Carlosama)	162	0,50%

Municipio	N	%
Belén	160	0,50%
Funes	157	0,50%
Iles	155	0,50%
Roberto Payán (San José)	154	0,50%
Arboleda (Berruecos)	153	0,50%
Gualmatán	153	0,50%
Aldana	151	0,50%
Magüí (Payán)	150	0,40%
Puerres	149	0,40%
Policarpa	147	0,40%
El Rosario	144	0,40%
Santacruz (Guachavés)	142	0,40%
San Bernardo	140	0,40%
Contadero	138	0,40%
El Peñol	137	0,40%
San Pedro de Cartago	132	0,40%
Cumbitara	117	0,30%
Ospina	117	0,30%
Mosquera	114	0,30%
La Tola	111	0,30%
La Llanada	105	0,30%
Sapuyes	93	0,30%
Providencia	73	0,20%
Santa Barbara (Iscuandé)	69	0,20%
Total	**33.438**	**100.0%**

Fuente: elaboración propia

2.2 CORRELACIÓN ENTRE LAS PRUEBAS DEL SABER 11º

A través del coeficiente de correlación de Pearson se establece como se asocian linealmente las cinco competencias y su resultado global de las pruebas SABER 11º presentadas por los estudiantes de Educación Secundaria del departamento de Nariño en el período 2015-2016. Los resultados se presentan en la Tabla 3.

Tabla 3. ***Matriz de correlaciones de las Competencias de las pruebas SABER 11º***

Pruebas	Ciencias Naturales	Inglés	Lectura Crítica	Matemáticas	Ciudadanas y Sociales	Global
Ciencias Naturales	1	0,670**	0,756**	0,810**	0,789**	0,920**
Inglés		1	0,629**	0,648**	0,631**	0,747**
Lectura Crítica			1	0,733**	0,772**	0,885**
Matemáticas				1	0,766**	0,917**
Ciudadanas					1	0,911**
Global						1

** = La correlación es significativa al nivel 0,01 (bilateral).
Fuente: elaboración propia

De acuerdo a la tabla anterior, todas las correlaciones resultan altamente significativas ($p\ valor < 0.01$) por el tamaño del número de datos. Siguiendo la clasificación de Cohen (1998), para la interpretación del coeficiente de Pearson, se observó que Ciencias Naturales presenta unas correlaciones altas ($0,5 < r < 0,8$) con inglés, Lectura Crítica, Competencias Ciudadanas y muy alta ($0,8 < r < 1$) con Matemáticas y con el puntaje Global. Inglés correlaciones altas Lectura Crítica, Matemáticas, Competencias Ciudadanas y puntaje Global. Lectura Crítica presenta correlaciones altas con Matemáticas y Competencias Ciudadanas y muy altas con el puntaje Global. Matemáticas correlaciones altas Competencias Ciudadanas y muy altas con el puntaje Global; y Competencias Ciudadanas correlación muy alta con el puntaje Global.

2.3 EFECTO LOCALIDAD DE LAS PRUEBAS DEL SABER 11º EN EL CONTEXTO DEPARTAMENTAL

Para establecer el "efecto localidad" en el contexto departamental, se elaboró un ranking según el municipio donde se ubican las diferentes Instituciones de Educación Secundaria del departamento de Nariño que participaron en las pruebas SABER 11º, durante el período 2015 a 2016.

Para la elaboración del ranking se utilizó el estadístico d de Cohen (1998) que permite calcular el Tamaño del Efecto Localidad a partir de las diferencias estandarizadas entre los promedios en cada uno de las instituciones frente a la localidad que alcanzó el máximo puntaje promedio en cada una de las cinco competencias genéricas y que se toma como referencia, así:

$$d_i = \frac{\overline{X}_O - \overline{X}_i}{S_{io}} \quad (1)$$

Donde:

$$S_{io} = \sqrt{\frac{(n_o - 1){S_o}^2 + (n_i - 1){S_i}^2}{n_o + n_i - 2}}$$

La interpretación de cada símbolo se presenta en la Tabla 4.

Tabla 4. ***Símbolos de la fórmula d de Cohen***

Variable	Descripción
d_i	Tamaño del efecto de la *institución i* con relación al *referente*
$\overline{X}_o$	Media de la institución de referencia (de mayor puntaje)
$\overline{X}_i$	Media de la institución *i*.
${S_o}^2$	Varianza de la institución de referencia.
${S_i}^2$	Varianza de la institución de *i*.
N_o	Número de estudiantes que presentaron las pruebas de la institución de referencia.
N_i	Número de estudiantes que presentaron las pruebas de la institución *i*.

Fuente: elaboración propia

Cohen (1998), propone la siguiente escala para interpretar las diferencias en unidades estándar, obtenidas con el estadístico de la expresión (1): [0.0, 0.2] trivial o irrelevante, [0.2, 0.5] pequeña, [0.5, 0.8] moderada y [0.8, ∞] grande.

2.3.1 Efecto Localidad en Lectura Crítica

Según el ranking que presenta la Tabla 5 en Lectura Crítica la localidad de Gualmatán junto con Colón (Génova) y Belén ocupan las primeras posiciones, con diferencias irrelevantes de acuerdo a Cohen. A diferencias pequeñas están los municipios que ocupan los lugares 4 a 18 de la tabla. A diferencias moderadas están los municipios que ocupan los lugares 19 a 44. Las localidades en los lugares 45 a 64 de la tabla se encuentran a distancias grandes en esta competencia.

Tabla 5. ***Ranking por Municipio de las Instituciones de Educación Secundaria del departamento de Nariño en Lectura Crítica, pruebas SABER 11º - 2015-2016.***

No.	Municipio	N	Media	DE	d Cohen
1	Gualmatán	153	56,60	7,8	-
2	Colón (Génova)	191	56,56	11,0	0,00
3	Belén	160	55,9	10,5	0,08
4	Pasto	10.087	54,5	9,3	0,23
5	Pupiales	373	54,5	8,4	0,26
6	Puerres	149	54,3	7,9	0,29
7	La Cruz	417	53,9	10,4	0,27
8	San Pablo	304	53,8	10,6	0,29
9	Sandoná	533	53,7	9,6	0,31
10	Ipiales	2.773	53,6	9,0	0,34
11	San Bernardo	140	53,5	7,9	0,40
12	Guachucal	453	53,3	8,5	0,40
13	Aldana	151	53,1	7,7	0,45
14	Contadero	138	52,8	8,0	0,48
15	Guaitarilla	247	52,8	8,3	0,47
16	Potosí	230	52,8	7,3	0,51
17	Túquerres	765	52,8	8,5	0,46
18	El Tambo	353	52,5	8,8	0,48
19	Buesaco	393	52,5	7,6	0,54
20	Policarpa	147	52,3	8,6	0,52
21	Providencia	73	52,2	9,0	0,54

No.	Municipio	N	Media	DE	d Cohen
22	Ospina	117	52,0	7,8	0,59
23	El Tablón	340	51,9	8,3	0,57
24	Consacá	224	51,8	7,7	0,62
25	Funes	157	51,8	8,2	0,60
26	Los Andes (Sotomayor)	246	51,7	9,5	0,55
27	Ancuya	212	51,7	8,7	0,59
28	Cumbal	866	51,6	8,4	0,60
29	San Lorenzo	410	51,6	8,3	0,61
30	El Peñol	137	51,5	7,5	0,67
31	Sapuyes	93	51,5	9,4	0,61
32	La Unión	734	51,4	9,3	0,58
33	Iles	155	50,9	7,3	0,76
34	Albán (San José)	260	50,9	9,0	0,67
35	San Pedro de Cartago	132	50,9	6,7	0,79
36	Samaniego	687	50,7	9,0	0,67
37	Nariño	305	50,6	9,9	0,64
38	Taminango	392	50,6	8,9	0,69
39	Chachagüí	270	50,5	7,8	0,78
40	Tangua	207	50,3	8,2	0,78
41	La Florida	208	50,3	8,9	0,74
42	Santacruz (Guachavés)	142	50,2	8,3	0,79
43	Cumbitara	117	50,2	8,6	0,79
44	Imués	209	50,2	8,2	0,80
45	Arboleda (Berruecos)	153	50,1	7,9	0,83
46	Yacuanquer	229	50,1	7,6	0,85
47	Córdoba	345	49,7	8,2	0,85
48	Cuaspud (Carlosama)	162	49,6	8,3	0,87
49	El Rosario	144	49,0	8,3	0,94
50	Linares	240	48,4	8,9	0,96
51	La Llanada	105	48,3	8,6	1,02
52	Mallama (Piedrancha)	234	48,2	8,5	1,02
53	Ricaurte	254	46,8	9,4	1,11
54	Leiva	173	45,8	8,2	1,34
55	Francisco Pizarro	169	44,9	8,9	1,40

No.	Municipio	N	Media	DE	d Cohen
56	Mosquera	114	44,6	7,5	1,56
57	Barbacoas	521	44,3	7,3	1,66
58	Tumaco	4.565	43,7	8,6	1,50
59	El Charco	383	43,3	7,7	1,73
60	Olaya Herrera	313	42,9	7,5	1,81
61	Roberto Payán (San José)	154	42,6	7,0	1,89
62	Magüí (Payán)	150	42,1	6,0	2,09
63	La Tola	111	41,3	6,8	2,07
64	Santa Bárbara (Iscuandé)	69	40,9	8,4	1,97

Fuente: elaboración propia

2.3.2 Efecto Localidad en Matemáticas

El municipio de Colón (Génova), ocupa la primera casilla en la prueba de Matemáticas, seguido de las localidades de Belén, La Cruz, Gualmatán y San Pablo en ese orden a distancias pequeñas, como se aprecia en la Tabla 6. A distancias moderadas están los municipios ubicados en las casillas 6 a 20 y con distancias grandes los que están en los lugares 21 a 64.

Tabla 6. ***Ranking por Municipio de las Instituciones de Educación Secundaria del departamento de Nariño en Matemáticas, pruebas SABER 11º - 2015-2016.***

No.	Municipio	N	Media	DE	d Cohen
1	Colón (Génova)	191	64,57	17,7	-
2	Belén	160	59,52	14,6	0,31
3	La Cruz	417	59,13	14,4	0,35
4	Gualmatán	153	57,54	9,5	0,48
5	San Pablo	304	57,21	13,7	0,48
6	Sandoná	533	56,49	12,5	0,57
7	Puerres	149	56,26	11,5	0,54
8	Pasto	10.087	55,61	12,4	0,72
9	Pupiales	373	55,00	10,9	0,70
10	Consacá	224	54,79	9,8	0,70
11	Aldana	151	54,70	9,9	0,67

No.	Municipio	N	Media	DE	d Cohen
12	Ipiales	2.773	54,66	11,6	0,82
13	Túquerres	765	54,59	11,4	0,77
14	San Bernardo	140	54,34	10,0	0,68
15	Guaitarilla	247	54,29	11,2	0,71
16	Ancuya	212	54,24	11,5	0,70
17	Guachucal	453	53,81	10,9	0,81
18	Ospina	117	53,56	9,5	0,73
19	Potosí	230	53,50	10,3	0,78
20	El Tambo	353	53,46	11,2	0,80
21	Contadero	138	52,31	10,5	0,81
22	La Unión	734	52,31	12,3	0,90
23	Buesaco	393	52,28	10,4	0,93
24	Albán (San José)	260	52,27	11,3	0,85
25	Imués	209	52,18	11,0	0,85
26	Sapuyes	93	52,13	11,7	0,78
27	Iles	155	52,00	10,0	0,85
28	Taminango	392	51,98	11,4	0,91
29	Cumbal	866	51,91	10,7	1,03
30	Policarpa	147	51,88	10,7	0,84
31	El Peñol	137	51,58	9,1	0,88
32	Chachagüí	270	51,56	10,8	0,92
33	Los Andes (Sotomayor)	246	51,52	11,5	0,90
34	Providencia	73	51,21	11,6	0,82
35	Funes	157	51,11	10,5	0,90
36	Yacuanquer	229	50,99	10,3	0,96
37	Samaniego	687	50,95	12,1	1,01
38	San Lorenzo	410	50,95	10,7	1,02
39	Córdoba	345	50,93	10,6	1,00
40	La Florida	208	50,88	10,1	0,96
41	San Pedro de Cartago	132	50,73	10,7	0,91
42	El Tablón	340	50,44	11,0	1,02
43	Nariño	305	50,38	11,8	0,99
44	Linares	240	50,02	12,0	0,98

No.	Municipio	N	Media	DE	d Cohen
45	Santacruz (Guachavés)	142	49,44	9,7	1,02
46	El Rosario	144	49,26	10,0	1,03
47	Tangua	207	49,04	10,1	1,09
48	Arboleda (Berruecos)	153	48,93	9,9	1,06
49	Cuaspud (Carlosama)	162	48,92	9,2	1,08
50	La Llanada	105	48,61	11,9	1,00
51	Cumbitara	117	48,03	9,7	1,09
52	Leiva	173	45,77	10,4	1,28
53	Mallama (Piedrancha)	234	45,60	10,7	1,33
54	Ricaurte	254	45,39	12,0	1,30
55	Mosquera	114	43,75	9,0	1,38
56	Francisco Pizarro	169	41,82	10,1	1,55
57	Barbacoas	521	41,41	8,9	1,94
58	Tumaco	4.565	40,93	10,3	2,22
59	El Charco	383	40,11	8,7	1,97
60	La Tola	111	39,47	9,1	1,66
61	Olaya Herrera	313	39,27	8,3	1,99
62	Magüí (Payán)	150	38,25	7,7	1,85
63	Roberto Payán (San José)	154	37,80	7,1	1,91
64	Santa Bárbara (Iscuandé)	69	37,58	8,0	1,71

Fuente: elaboración propia

2.3.3 Efecto Localidad en Ciencias Naturales

La Tabla 7 presenta el ranking por municipio de las Instituciones de Educación Secundaria en Ciencias Naturales en las pruebas SABER 11º en el período 2015-2016, en dicha competencia los municipios de Colón (Génova) y Belén, presentan el mejor rendimiento en esta competencia con diferencias irrelevantes según Cohen. A diferencias pequeñas (*d Cohen* > 0,2), los municipios que se observan en la tabla ocupando las posiciones 3 a 14, a diferencias moderadas (*d de Cohen* > 0,5), municipios en las posiciones 15 a 38 y a diferencias grandes (*d de Cohen* > 0,5), los municipios que en la tabla ocupan las posiciones 39 a 64.

Tabla 7. ***Ranking por Municipio de las Instituciones de Educación Secundaria del departamento de Nariño en Ciencias Naturales, pruebas SABER 11º - 2015-2016.***

No.	Municipio	N	Media	DE	d Cohen
1	Colón (Génova)	191	60,25	11,7	-
2	Belén	160	58,16	11,1	0,18
3	Gualmatán	153	57,76	9,1	0,23
4	Puerres	149	57,28	8,2	0,29
5	Guachucal	453	56,48	9,3	0,37
6	Pupiales	373	56,44	8,5	0,39
7	La Cruz	417	56,42	10,9	0,34
8	Pasto	10.087	56,24	10,2	0,39
9	San Pablo	304	56,24	10,3	0,37
10	Consacá	224	55,65	8,6	0,45
11	San Bernardo	140	55,54	8,2	0,45
12	Sandoná	533	55,41	10,4	0,45
13	Aldana	151	55,32	7,1	0,50
14	Contadero	138	55,14	8,1	0,49
15	Potosí	230	55,04	8,0	0,53
16	Ipiales	2.773	54,81	9,1	0,59
17	Buesaco	393	54,80	7,9	0,58
18	Túquerres	765	54,78	9,0	0,57
19	El Tambo	353	54,72	9,8	0,53
20	Guaitarilla	247	54,62	9,0	0,55
21	Policarpa	147	54,38	8,7	0,56
22	El Peñol	137	54,28	7,1	0,59
23	Ospina	117	54,28	7,8	0,57
24	Ancuya	212	53,75	8,6	0,64
25	Sapuyes	93	53,71	9,3	0,60
26	La Unión	734	53,15	9,7	0,70
27	Cumbal	866	53,01	8,2	0,81
28	San Lorenzo	410	52,89	8,1	0,78
29	El Tablón	340	52,84	8,5	0,76
30	Cuaspud (Carlosama)	162	52,51	8,1	0,76
31	Imués	209	52,44	8,0	0,79

No.	Municipio	N	Media	DE	d Cohen
32	La Florida	208	52,40	8,1	0,79
33	Albán (San José)	260	52,33	9,2	0,77
34	Taminango	392	52,30	9,3	0,78
35	Chachagüí	270	52,26	7,9	0,83
36	Yacuanquer	229	52,25	9,1	0,77
37	Los Andes (Sotomayor)	246	52,13	9,4	0,78
38	Funes	157	51,97	9,0	0,78
39	Iles	155	51,88	7,4	0,84
40	Providencia	73	51,75	8,3	0,78
41	Nariño	305	51,66	10,2	0,80
42	San Pedro de Cartago	132	51,58	8,0	0,84
43	Arboleda (Berruecos)	153	51,42	8,5	0,85
44	Santacruz (Guachavés)	142	51,35	8,3	0,86
45	Tangua	207	51,29	8,1	0,90
46	El Rosario	144	51,15	8,6	0,87
47	Córdoba	345	51,11	8,0	0,97
48	Samaniego	687	50,66	9,6	0,95
49	Cumbitara	117	50,19	6,9	0,99
50	Linares	240	49,37	10,7	0,97
51	La Llanada	105	49,12	9,2	1,02
52	Mallama (Piedrancha)	234	48,50	9,6	1,11
53	Ricaurte	254	48,08	9,5	1,16
54	Leiva	173	47,78	8,0	1,23
55	Mosquera	114	46,21	7,2	1,37
56	Barbacoas	521	44,18	7,2	1,86
57	Tumaco	4.565	42,78	8,7	1,98
58	Francisco Pizarro	169	42,59	8,1	1,74
59	El Charco	383	42,33	7,6	1,95
60	La Tola	111	41,67	7,5	1,79
61	Roberto Payán (San José)	154	41,61	6,2	1,93
62	Olaya Herrera	313	41,44	7,4	2,03
63	Magüí (Payán)	150	41,29	6,4	1,95
64	Santa Bárbara (Iscuandé)	69	40,93	6,6	1,82

Fuente: elaboración propia

2.3.4 Efecto Localidad en Inglés

Según la Tabla 8, durante el periodo estudiado, igualmente el municipio de Colón (Génova) ocupa la primera posición en el desempeño promedio de instituciones de Educación Secundaria en inglés. A diferencias pequeñas se sitúan los municipios que se sitúan en las posiciones 2 a 15. En las posiciones 16 a 37 se ubican los municipios a diferencias moderadas y con distancias grandes los que ocupan los lugares 38 a 64.

Tabla 8. ***Ranking por Municipio de las Instituciones de Educación Secundaria del departamento de Nariño en inglés, pruebas SABER 11º - 2015-2016.***

No.	Municipio	N	Media	DE	d Cohen
1	Colón (Génova)	191	58,78	13,8	-
2	Belen	160	56,01	12,7	0,21
3	La Cruz	417	55,56	13,6	0,24
4	Pupiales	373	55,39	8,9	0,31
5	Pasto	10.087	54,99	11,2	0,34
6	Aldana	151	54,19	9,1	0,38
7	Gualmatán	153	53,95	9,3	0,40
8	Sapuyes	93	53,95	11,6	0,37
9	Buesaco	393	53,41	8,7	0,51
10	San Pablo	304	53,35	10,9	0,45
11	Ipiales	2.773	53,26	9,9	0,54
12	San Bernardo	140	53,19	8,3	0,47
13	Potosí	230	52,97	8,4	0,52
14	Policarpa	147	52,89	10,6	0,47
15	Contadero	138	52,71	9,2	0,50
16	Funes	157	52,66	9,7	0,51
17	Sandoná	533	52,63	9,4	0,57
18	Ospina	117	52,46	7,6	0,53
19	Guaitarilla	247	51,98	9,7	0,58
20	Túquerres	765	51,80	9,5	0,66
21	Cuaspud (Carlosama)	162	51,78	8,1	0,61
22	Guachucal	453	51,43	8,0	0,73
23	Consacá	224	51,34	8,8	0,65
24	Puerres	149	51,15	8,5	0,65
25	Providencia	73	51,14	10,0	0,59
26	Cumbal	866	51,11	8,3	0,81

No.	Municipio	N	Media	DE	d Cohen
27	Cumbitara	117	50,87	7,7	0,67
28	Imués	209	50,36	8,2	0,75
29	La Unión	734	50,34	10,0	0,77
30	El Tablón	340	50,32	9,0	0,77
31	San Lorenzo	410	50,25	8,1	0,83
32	Chachagüí	270	50,21	8,0	0,79
33	La Florida	208	50,19	7,9	0,77
34	Ancuya	212	50,09	9,1	0,75
35	Arboleda (Berruecos)	153	50,04	8,3	0,75
36	Los Andes (Sotomayor)	246	49,83	9,6	0,77
37	Albán (San José)	260	49,64	9,4	0,80
38	Nariño	305	49,28	10,3	0,81
39	Samaniego	687	49,26	10,1	0,87
40	La Llanada	105	49,25	8,6	0,78
41	El Peñol	137	49,12	7,7	0,83
42	Taminango	392	49,07	8,5	0,92
43	Córdoba	345	49,06	7,5	0,95
44	El Tambo	353	48,99	9,1	0,89
45	Tangua	207	48,98	7,5	0,89
46	Iles	155	48,68	7,4	0,89
47	El Rosario	144	48,26	6,7	0,93
48	San Pedro de Cartago	132	48,17	8,1	0,90
49	Santacruz (Guachavés)	142	48,15	7,2	0,93
50	Mallama (Piedrancha)	234	47,70	8,4	0,99
51	Ricaurte	254	47,64	9,8	0,96
52	Yacuanquer	229	47,53	8,0	1,02
53	Linares	240	46,99	7,7	1,09
54	Mosquera	114	46,66	7,8	1,02
55	La Tola	111	46,05	7,2	1,08
56	Leiva	173	45,51	7,4	1,18
57	Barbacoas	521	45,25	6,8	1,47
58	El Charco	383	43,98	6,6	1,54
59	Francisco Pizarro	169	43,95	6,1	1,36
60	Tumaco	4.565	43,75	7,7	1,87
61	Olaya Herrera	313	43,38	7,2	1,51
62	Roberto Payán (San José)	154	43,19	6,4	1,40
63	Magüí (Payán)	150	42,93	6,1	1,43
64	Santa Bárbara (Iscuandé)	69	41,36	6,1	1,42

Fuente: elaboración propia

2.3.5 Efecto Localidad en Sociales y Competencias Ciudadanas

Como se observa en la Tabla 9, en la prueba de Competencias Ciudadanas y Sociales, las instituciones de Educación Secundaria del municipio de Colón (Génova) nuevamente presentan el mejor desempeño, conjuntamente con las instituciones de Belén y Gualmatán, con diferencias irrelevantes. A distancias moderadas de las anteriores están los municipios que ocupan las posiciones 4 a 19 en la tabla. A distancias moderadas están los municipios en los lugares 20 a 47 y distancias grandes los que se ubican en las casillas 48 a 64.

Tabla 9. ***Ranking por Municipio de las Instituciones de Educación Secundaria del departamento de Nariño en Competencias Sociales y Ciudadanas, pruebas SABER 11º - 2015-2016.***

No.	Municipio	N	Media	DE	d Cohen
1	Colón (Génova)	191	57,23	12,3	-
2	Belén	160	56,26	11,0	0,08
3	Gualmatán	153	56,03	8,8	0,11
4	San Bernardo	140	54,92	9,3	0,21
5	Pasto	10.087	54,39	10,6	0,27
6	San Pablo	304	54,29	12,0	0,24
7	Puerres	149	54,19	9,4	0,27
8	La Cruz	417	53,98	13,4	0,25
9	Pupiales	373	53,45	9,9	0,35
10	Sandoná	533	53,17	11,4	0,35
11	Guaitarilla	247	53,08	9,9	0,38
12	Ipiales	2.773	52,94	10,0	0,42
13	Guachucal	453	52,50	9,9	0,44
14	Túquerres	765	52,47	9,4	0,47
15	Buesaco	393	52,40	8,8	0,48
16	Ospina	117	52,33	9,0	0,44
17	Policarpa	147	52,32	10,1	0,43
18	Potosí	230	51,74	8,9	0,52
19	Contadero	138	51,61	10,0	0,49
20	Aldana	151	51,56	8,7	0,52
21	Consacá	224	51,38	9,1	0,55
22	Sapuyes	93	51,35	9,5	0,51
23	El Tablón	340	51,29	9,8	0,55
24	El Tambo	353	51,25	10,5	0,54
25	La Unión	734	51,25	11,3	0,52

No.	Municipio	N	Media	DE	d Cohen
26	Ancuya	212	50,97	10,4	0,55
27	Samaniego	687	50,75	10,9	0,58
28	Funes	157	50,61	10,3	0,58
29	Providencia	73	50,59	11,0	0,56
30	Cumbal	866	50,58	9,7	0,65
31	Tangua	207	50,43	9,4	0,63
32	Los Andes (Sotomayor)	246	50,33	10,5	0,61
33	Santacruz (Guachavés)	142	50,32	9,5	0,62
34	Albán (San José)	260	50,24	10,3	0,62
35	San Lorenzo	410	50,14	9,8	0,67
36	Taminango	392	50,04	10,2	0,66
37	Iles	155	50,01	8,8	0,66
38	Yacuanquer	229	49,98	9,5	0,67
39	La Florida	208	49,92	8,7	0,69
40	Chachagüí	270	49,69	9,3	0,71
41	Imués	209	49,67	9,3	0,70
42	El Peñol	137	49,65	9,3	0,68
43	Nariño	305	49,61	11,1	0,66
44	El Rosario	144	49,46	9,9	0,69
45	Arboleda (Berruecos)	153	49,39	9,5	0,70
46	Cumbitara	117	49,15	9,8	0,71
47	Cuaspud (Carlosama)	162	48,98	8,1	0,78
48	Córdoba	345	48,44	9,2	0,84
49	San Pedro de Cartago	132	48,26	8,8	0,82
50	Linares	240	47,89	10,7	0,82
51	La Llanada	105	47,63	10,3	0,83
52	Mallama (Piedrancha)	234	46,59	10,5	0,94
53	Ricaurte	254	45,88	10,9	0,99
54	Leiva	173	45,28	9,7	1,07
55	Mosquera	114	44,71	8,5	1,14
56	Francisco Pizarro	169	42,69	9,5	1,32
57	Barbacoas	521	42,52	8,5	1,52
58	Olaya Herrera	313	41,30	8,8	1,55
59	Tumaco	4.565	41,21	9,9	1,59
60	La Tola	111	41,00	7,7	1,50
61	El Charco	383	40,89	8,8	1,62
62	Magüí (Payán)	150	39,71	7,7	1,67
63	Roberto Payán (San José)	154	39,47	8,7	1,64
64	Santa Bárbara (Iscuandé)	69	38,43	8,3	1,65

Fuente: elaboración propia

Cabe destacar el alto rendimiento de las instituciones de Educación Secundaria de los municipios de Colón (Génova), Belén y Gualmatán en todas las competencias de las Pruebas SABER 11º e igualmente el bajo rendimiento de las instituciones que pertenecen a los municipios de la zona costera del departamento de Nariño.

2.4 EFECTO DE LAS SUBREGIONES EN LAS PRUEBAS SABER 11º

2.4.1 Efecto de la Subregión en Lectura Crítica

Las subregiones Centro, Río Mayo y Obando, como se observa en la Tabla 10, se destacan por su mejor rendimiento en la prueba de inglés, presentando diferencias no relevantes ($d \leq 0.20$). La Sabana, Occidente, Guambuyaco, Juanambú, Los Abades y La Cordillera están a distancias pequeñas ($d \leq 0.50$) con las primeras. A diferencias moderadas Piedemonte Costero ($d \leq 0.80$) y a distancias grandes Pacífico Sur, Telembí y Sanquianga ($d \leq 0.80$) .

Tabla 10. ***Ranking por Subregión las Instituciones de Educación Secundaria del departamento de Nariño en Lectura Crítica, pruebas SABER 11º - 2015-2016.***

Subregión	N	Media	DE	d Cohen
Centro	11306	54,07	9,3	-
Río Mayo	1812	53,51	9,9	0,06
Obando	6105	52,92	8,7	0,13
Sin subregión	247	52,81	8,3	0,13
La Sabana	1184	52,13	8,5	0,21
Occidente	1209	51,97	9,2	0,23
Guambuyaco	841	51,60	8,9	0,27
Juanambú	1822	51,52	8,4	0,28
Los Abades	902	50,71	8,9	0,36
La Cordillera	973	49,74	8,9	0,47
Piedemonte costero	488	47,49	9,0	0,71
Pacífico Sur	4734	43,73	8,6	1,13
Telembí	825	43,59	7,1	1,14
Sanquianga	990	42,91	7,6	1,22

Fuente: elaboración propia

2.4.2 Efecto de la Subregión en Matemáticas

En Matemáticas, las instituciones educativas de las subregiones Río Mayo, Centro, Occidente y La Sabana presentan los altos rendimientos con diferencias no relevantes (d ≤ 0.20). Las instituciones de Obando, Guambuyaco, Juanambú, Los Abades y La Cordillera, presentan diferencias pequeñas (d ≤ 0.50) con las anteriores. Piedemonte Costero, Pacífico Sur, Telembí y Sanquianga a distancias grandes (d ≤ 0.80) (ver Tabla 11).

Tabla 11. ***Ranking por Subregión las Instituciones de Educación Secundaria del departamento de Nariño en Matemáticas, pruebas SABER 11º - 2015-2016.***

Subregión	N	Media	DE	d Cohen
Río Mayo	1812	56,43	14,1	-
Centro	11306	55,07	12,3	0,11
Occidente	1209	54,49	12,0	0,15
Sin Subregión	247	54,29	11,2	0,16
La Sabana	1184	53,87	11,2	0,20
Obando	6105	53,72	11,2	0,23
Guambuyaco	841	51,98	11,1	0,34
Juanambú	1822	51,60	11,3	0,38
Los Abades	902	50,74	11,7	0,43
La Cordillera	973	49,98	11,0	0,49
Piedemonte Costero	488	45,49	11,4	0,81
Pacífico Sur	4734	40,97	10,2	1,35
Telembí	825	40,16	8,6	1,29
Sanquianga	990	40,02	8,7	1,32

Fuente: elaboración propia

2.4.3 Efecto de la Subregión en Ciencias Naturales

Como se observa en la Tabla 12, en Ciencias Naturales, durante el período 2015 a 2016, la subregión Centro del departamento de Nariño alcanzó el mejor desempeño, sin embargo, Río

Mayo, Obando, La Sabana y Occidente presentan diferencias no relevantes (d ≤ 0.20) con dicha subregión. Guambuyaco, Juanambú, La Cordillera y Los Abades presentan diferencias pequeñas (d ≤ 0.50) con el Centro. A diferencias moderadas Piedemonte Costero (d ≤ 0.80) y a distancias grandes Telembí, Pacífico Sur y Sanquianga (d ≤ 0.20).

Tabla 12. ***Ranking por Subregión las Instituciones de Educación Secundaria del departamento de Nariño en Ciencias Naturales, pruebas SABER 11º - 2015-2016.***

Subregión	N	Media	DE	d Cohen
Centro	11306	55,78	10,1	-
Río Mayo	1812	55,62	10,4	0,02
Sin Subregión	247	54,62	9,0	0,12
Obando	6105	54,53	8,8	0,13
La Sabana	1184	54,23	8,8	0,16
Occidente	1209	53,97	10,1	0,18
Guambuyaco	841	53,19	9,4	0,26
Juanambú	1822	53,19	8,8	0,26
La Cordillera	973	51,39	8,8	0,44
Los Abades	902	50,86	9,3	0,49
Piedemonte Costero	488	48,28	9,6	0,74
Telembí	825	43,17	7,0	1,27
Pacífico Sur	4734	42,77	8,7	1,34
Sanquianga	990	42,33	7,6	1,36

Fuente: elaboración propia

2.4.4 Efecto de la Subregión en inglés

Según la Tabla 13, durante el período 2015 a 2016, en la prueba de inglés, las subregiones Centro, Río Mayo, Obando, alcanzan el mejor desempeño, presentando, entre ellas, diferencias no relevantes (d ≤ 0.20). La Sabana, Occidente, Juanambú, Guambuyaco, La Cordillera y Los Abades presentan diferencias pequeñas (d ≤ 0.50) con las subregiones iniciales. A diferencias moderadas Piedemonte Costero (d ≤ 0.80) y a distancias grandes Telembí, Sanquianga y Pacífico Sur (d ≤ 0.80).

Tabla 13. ***Ranking por Subregión las Instituciones de Educación Secundaria del departamento de Nariño en inglés, pruebas SABER 11º - 2015-2016.***

Subregión	N	Media	DE	d Cohen
Centro	11306	54,37	11,1	-
Río Mayo	1812	53,55	11,8	0,07
Obando	6105	52,51	9,2	0,18
Sin Subregión	247	51,98	9,7	0,22
La Sabana	1184	51,78	9,4	0,24
Occidente	1209	50,83	9,2	0,32
Juanambú	1822	50,80	9,2	0,33
Guambuyaco	841	49,29	8,9	0,46
Los Abades	902	49,24	9,7	0,47
La Cordillera	973	49,11	8,6	0,48
Piedemonte costero	488	47,67	9,1	0,61
Telembí	825	44,44	6,7	0,91
Sanquianga	990	44,15	7,1	0,94
Pacífico Sur	4734	43,76	7,6	1,04

Fuente: elaboración propia

2.4.5 Efecto de la Subregión en Sociales y Ciudadanas

Según la Tabla 14, en las pruebas de Ciudadanas y Sociales, las instituciones educativas que más se destacan, pertenecen a las subregiones Centro, Río Mayo, Obando y La Sabana que presentan los mejores rendimientos con diferencias no relevantes ($d \leq 0.20$). Las instituciones del Occidente, Juanambú, Los Abades, Guambuyaco y La Cordillera, están a distancia pequeñas de las anteriores ($d \leq 0.50$). A diferencias moderadas Piedemonte Costero ($d \leq 0.80$) y a distancias grandes Telembí, Sanquianga y Pacífico Sur ($d \leq 0.80$).

Tabla 14. ***Ranking por Subregión las Instituciones de Educación Secundaria del departamento de Nariño en Sociales y Ciudadanas, pruebas SABER 11º - 2015-2016.***

Subregión	N	Media	DE	d Cohen
Centro	11306	53,91	10,6	-
Río Mayo	1812	53,61	11,7	0,03
Sin Subregión	247	53,08	9,9	0,08
Obando	6105	52,11	9,8	0,17
La Sabana	1184	51,87	9,4	0,19
Occidente	1209	51,40	10,8	0,24
Juanambú	1822	50,87	10,2	0,29
Los Abades	902	50,67	10,7	0,30
Guambuyaco	841	50,27	10,3	0,34
La Cordillera	973	49,35	10,2	0,43
Piedemonte costero	488	46,22	10,7	0,73
Telembí	825	41,44	8,5	1,19
Sanquianga	990	41,30	8,7	1,21
Pacífico Sur	4734	41,27	9,9	1,22

Fuente: elaboración propia

2.5 EFECTO DE LAS VARIABLES SOCIECONÓMICAS EN LAS PRUEBAS SABER 11º

A continuación, se establece el efecto de las diferentes variables socioeconómicas consideradas en este estudio, sobre los puntajes obtenidos en las Pruebas SABER 11º, a partir de las diferencias estandarizadas de los promedios obtenidos en las pruebas, en los diferentes grupos que conforman dichas variables.

2.5.1 Efecto en Lectura Crítica

En Lectura Crítica de las pruebas SABER 11º - 2015-2016, los hombres y mujeres presentan similar desempeño pues sus diferencias son no relevantes, como se observa en la Tabla 15. Los más jóvenes, menores de 18 años, tienen mejor desempeño que los de mayor edad, alcanzando con éstos diferencias moderadas o grandes. Los estudiantes de estratos sociales más altos tienen mejor

desempeño que los de estratos más bajos mostrando diferencias no relevantes con los estratos altos y el medio, pero diferencias grandes con el bajo.

De manera consistente con lo anterior se observa que a mayor ingreso familiar mejor desempeño en esta competencia y de igual manera estudiantes que pertenecen a instituciones de carácter privado presentaron mejor desempeño que los de las públicas. Los estudiantes de jornada completa u ordinaria presentaron un desempeño mucho más alto que los de otro tipo de jornada de estudio, siendo los de jornada de la noche los que presentaron más bajo desempeño. Las condiciones en que viven los estudiantes también afectan el desempeño en Lectura crítica, quienes tienen mejores condiciones de vida presentan mejores resultados.

También mejores condiciones de TIC implican mejores resultados en esta prueba. Finalmente, estudiantes cuyos padres tiene niveles de educación superior rinden más en esta competencia que aquellos estudiantes cuyos padres presentan niveles educativos más bajos.

Tabla 15. ***Variables Sociodemográficas y desempeño académico en Lectura Crítica en las Instituciones de Educación Secundaria del departamento de Nariño de las pruebas SABER 11º - 2015-2016.***

Variable Socioeconómica		N	Media	DT	d Cohen
Género	Femenino	18.564	50,88	9,83	0,05
	Masculino	14.755	51,41	9,61	-
Edad	Menor de 18 años	19.969	53,9	9,2	-
	Entre 18 y 22 años	12.108	47,8	8,8	0,67
	Mayor de 22 años	1.361	40,3	8,2	1,48
Estrato social	Bajo	30.862	50,5	9,6	0,85
	Medio	2.272	58,2	9,1	0,05
	Alto	130	58,7	10,2	-
Ingreso familiar	Menos de 1 SM	17.889	49,9	9,0	1,22
	Entre 1 y menos de 5 SM	14.832	52,3	10,2	0,84
	Entre 5 y menos de 10 SM	355	60,6	10,0	0,04
	10 o más SM	87	60,9	9,5	-
	Sin dato	275	50,3	8,7	1,19
Tipo de Colegio	Privado	3.826	54,6	9,8	-
	Público	29.612	50,7	9,6	0,41

Variable Socioeconómica		N	Media	DT	d Cohen
Jornada	Completa u Ordinaria	215	61,7	9,6	-
	Única	149	53,2	9,1	0,91
	Mañana	27.114	51,6	9,6	1,06
	Tarde	3.629	52,2	8,9	1,07
	Noche	2.331	42,3	8,2	2,32
Condición en la que vive	Hacinamiento crítico	391	48,8	8,8	0,26
	Hacinamiento medio	5.160	49,7	9,1	0,17
	Sin hacinamiento	27.887	51,4	9,8	-
Condición de la vivienda	Mala	20.197	50,8	9,2	0,43
	Regular	4.987	45,9	9,6	0,93
	Buena	8.254	54,9	9,6	-
Condición de las TIC	Mala	25.887	49,9	9,4	0,56
	Regular	7.551	55,2	9,7	-
Educación de Madre	Ninguna	662	44,8	9,2	1,12
	Primaria	19.217	50,4	9,6	0,54
	Secundaria	11.109	51,9	9,6	0,38
	Superior	2.175	55,5	9,8	-
	Sin dato	275	50,3	8,7	0,54
Educación de Padre	Ninguna	1.303	47,6	9,8	0,80
	Primaria	19.827	50,4	9,7	0,52
	Secundaria	10.484	52,3	9,5	0,33
	Superior	1.549	55,4	9,8	-
	Sin dato	275	50,3	8,7	0,53

Fuente: elaboración propia

2.5.2 Efecto en Matemáticas

De la Tabla 16 se deduce que, en la prueba de Matemáticas, las diferencias por sexo son pequeñas ($d \leq 0.5$), con mejor desempeño en los hombres. Los menores de 18 años se desempeñan mejor que los de mayor edad alcanzando diferencias moderadas y grandes en esta competencia. Se observa también en la Tabla 16 que a mayor estrato social más alto desempeño mostrando diferencias no relevantes entre los estratos alto y medio y diferencias grandes con el bajo. De la misma manera los estudiantes que pertenecen a familias con ingresos más altos y a instituciones de carácter privado se desempeñan mejor en Matemáticas. Como se ha observado en todas las competencias, los estudiantes con jornada completa u ordinaria presentaron un desempeño mucho más alto

que los de otras jornadas, presentando diferencias grandes según el estadístico d de Cohen.

Mejores condiciones en que vive el estudiante, de la vivienda y de la TIC implica mejor rendimiento en la prueba de Matemáticas. Finalmente, los niveles educativos de la madre y del padre del estudiante tienen un efecto importante en los resultados de esta prueba, si los padres tienen niveles de educación más altos entonces los estudiantes alcanzan mejores desempeños.

Tabla 16. ***Variables Sociodemográficas y desempeño académico en Matemáticas en las Instituciones de Educación Secundaria del departamento de Nariño de las pruebas SABER 11º - 2015-2016.***

Variable Socioeconómica		N	Media	DT	d Cohen
Género	Femenino	18.564	49,77	12,46	0,28
	Masculino	14.755	53,36	12,97	-
Edad	Menor de 18 años	19.969	55,4	12,2	-
	Entre 18 y 22 años	12.108	46,2	11,0	0,78
	Mayor de 22 años	1.361	36,7	8,9	1,56
Estrato social	Bajo	30.862	50,6	12,5	0,97
	Medio	2.272	60,9	13,1	0,14
	Alto	130	62,8	16,2	-
Ingreso familiar	Menos de 1 SM	17.889	49,9	11,7	1,44
	Entre 1 y menos de 5 SM	14.832	52,8	13,6	1,04
	Entre 5 y menos de 10 SM	355	64,9	14,9	0,13
	10 o más SM	87	66,9	16,7	-
	Sin dato	275	47,7	10,6	1,56
Tipo de Colegio	Privado	3.826	55,7	13,5	-
	Público	29.612	50,8	12,6	0,39
Jornada	Completa u Ordinaria	215	67,8	15,0	-
	Única	149	51,7	11,1	1,19
	Mañana	27.114	52,1	12,6	1,25
	Tarde	3.629	52,7	11,4	1,30
	Noche	2.331	39,1	9,5	2,84

Variable Socioeconómica		N	Media	DT	d Cohen
Condición en la que vive	Hacinamiento crítico	391	49,1	11,8	0,20
	Hacinamiento medio	5.160	49,7	11,7	0,16
	Sin hacinamiento	27.887	51,7	13,0	-
Condición de la vivienda	Mala	20.197	51,2	12,0	0,41
	Regular	4.987	44,0	12,2	0,96
	Buena	8.254	56,2	12,9	-
Condición de las TIC	Mala	25.887	49,8	12,3	0,55
	Regular	7.551	56,6	13,0	-
Educación de Madre	Ninguna	662	42,6	11,4	1,14
	Primaria	19.217	50,5	12,8	0,49
	Secundaria	11.109	52,3	12,5	0,36
	Superior	2.175	56,8	12,8	-
	Sin dato	275	47,7	10,6	0,73
Educación de Padre	Ninguna	1.303	46,0	12,1	0,86
	Primaria	19.827	50,6	12,8	0,49
	Secundaria	10.484	52,7	12,4	0,34
	Superior	1.549	56,9	13,2	-
	Sin dato	275	47,7	10,6	0,72

Fuente: elaboración propia

2.5.3 Efecto en Ciencias Naturales

De la Tabla 17 se concluye que en la prueba de Ciencias Naturales de las pruebas SABER 11º - 2015-2016, los hombres presentan mejor desempeño que las mujeres con diferencias pequeñas ($d \leq 0.5$). Por edad se desempeñan mejor los menores de 18 años alcanzando diferencias moderadas o grandes con los de mayor edad. El estrato social más alto tiene mejor desempeño mostrando diferencias no relevantes con el estrato medio, pero diferencias grandes con el bajo. Igualmente se observa que los estudiantes de familias con ingresos más altos se desempeñan mejor en esta competencia. Las instituciones de carácter privado presentaron mejor desempeño que las públicas. De la misma manera los estudiantes que asisten a jornada completa u ordinaria presentan un desempeño mucho más alto que los de otro tipo de jornada de estudio.

Las condiciones en que viven los estudiantes también afectan el desempeño en esta competencia, quienes tienen mejores condiciones de vida presentan mejores resultados. Mejores condiciones de TIC implican mejores resultados en esta prueba. Finalmente, los niveles educativos de la madre y del padre tienen un efecto importante en los resultados de la prueba de Ciencias Naturales, estudiantes cuyos padres tiene niveles de educación superior rinden más en esta competencia.

Tabla 17. ***Variables Sociodemográficas y desempeño académico en Ciencias Naturales en las Instituciones de Educación Secundaria del departamento de Nariño de las pruebas SABER 11º - 2015-2016.***

Variable Socioeconómica		N	Media	DT	d Cohen
Género	Femenino	18.564	51,29	10,34	0,22
	Masculino	14.755	53,55	10,73	-
Edad	Menor de 18 años	19.969	55,5	9,9	-
	Entre 18 y 22 años	12.108	48,4	9,4	0,73
	Mayor de 22 años	1.361	39,2	7,8	1,66
Estrato social	Bajo	30.862	51,7	10,3	1,05
	Medio	2.272	60,4	10,6	0,18
	Alto	130	62,4	13,4	-
Ingreso familiar	Menos de 1 SM	17.889	51,1	9,6	1,52
	Entre 1 y menos de 5 SM	14.832	53,4	11,3	1,11
	Entre 5 y menos de 10 SM	355	63,9	12,4	0,15
	10 o más SM	87	65,8	12,8	-
	Sin dato	275	50,2	8,6	1,60
Tipo de Colegio	Privado	3.826	56,0	11,5	-
	Público	29.612	51,8	10,3	0,40
Jornada	Completa u Ordinaria	215	69,0	12,0	-
	Única	149	53,8	9,1	1,39
	Mañana	27.114	52,9	10,3	1,56
	Tarde	3.629	53,4	9,3	1,64
	Noche	2.331	41,7	8,3	3,16
Condición en la que vive	Hacinamiento crítico	391	50,2	9,0	0,22
	Hacinamiento medio	5.160	51,0	9,7	0,15
	Sin hacinamiento	27.887	52,6	10,7	-

Variable Socioeconómica		N	Media	DT	d Cohen
Condición de la vivienda	Mala	20.197	52,2	9,7	0,43
	Regular	4.987	45,8	10,6	1,01
	Buena	8.254	56,5	10,6	-
Condición de las TIC	Mala	25.887	51,0	10,1	0,56
	Regular	7.551	56,8	10,8	-
Educación de Madre	Ninguna	662	44,9	10,0	1,12
	Primaria	19.217	51,6	10,5	0,48
	Secundaria	11.109	53,1	10,4	0,35
	Superior	2.175	56,7	10,7	-
	Sin dato	275	50,2	8,6	0,62
Educación de Padre	Ninguna	1.303	48,3	10,6	0,79
	Primaria	19.827	51,7	10,5	0,49
	Secundaria	10.484	53,3	10,2	0,35
	Superior	1.549	56,9	11,2	-
	Sin dato	275	50,2	8,6	0,62

Fuente: elaboración propia

2.5.4 Efecto en Inglés

Como se observa en la Tabla 18 en la prueba de inglés, los hombres y mujeres presentan diferencias no relevantes ($d \leq 0.2$) . Por edad, los menores de 18 años se desempeñan mejor, alcanzando diferencias moderadas o grandes con los estudiantes de mayor edad. Estudiantes de estrato social más alto tienen mejor desempeño presentando diferencias pequeñas con el estrato medio, pero diferencias grandes con el bajo. De la misma manera se observa que a mayor ingreso familiar mejor desempeño en la prueba de inglés. Y de manera consistente con lo anterior, los estudiantes de colegios privados presentan mejor desempeño en esta prueba que aquellos que pertenecen a colegios públicos. La jornada de estudio marca una diferencia importante en el desempeño de inglés, los que estudian en jornada completa presentan un más alto rendimiento que los de otras jornadas, siendo los de jornada de la noche los de más bajo rendimiento. Las condiciones en que viven los estudiantes también afectan el desempeño en esta competencia, estudiantes con mejores condiciones de vida presentan mejores resultados. De igual, manera mejores condiciones de TIC

implica mejores resultados en esta competencia. Finalmente, los niveles educativos de la madre y del padre tienen un efecto importante en los resultados de la prueba de inglés, estudiantes cuyos padres tiene niveles de educación superior rinden más en la prueba de inglés.

Tabla 18. ***Variables Sociodemográficas y desempeño académico en inglés en las Instituciones de Educación Secundaria del departamento de Nariño de las pruebas SABER 11º - 2015-2016.***

Variable Socioeconómica		N	Media	DT	d Cohen
Género	Femenino	18.564	50,63	10,40	0,08
	Masculino	14.755	51,47	10,60	-
Edad	Menor de 18 años	19.969	53,6	10,7	-
	Entre 18 y 22 años	12.108	47,7	8,7	0,60
	Mayor de 22 años	1.361	41,8	7,4	1,12
Estrato social	Bajo	30.862	50,2	9,9	1,46
	Medio	2.272	60,7	12,3	0,32
	Alto	130	64,7	15,5	-
Ingreso familiar	Menos de 1 SM	17.889	49,4	9,1	1,94
	Entre 1 y menos de 5 SM	14.832	52,5	11,4	1,27
	Entre 5 y menos de 10 SM	355	64,9	13,5	0,16
	10 o más SM	87	67,1	14,1	-
	Sin dato	275	50,4	8,7	1,63
Tipo de Colegio	Privado	3.826	55,8	12,4	-
	Público	29.612	50,4	10,1	0,53
Jornada	Completa u Ordinaria	215	67,4	13,6	-
	Única	149	52,3	9,6	1,25
	Mañana	27.114	51,4	10,4	1,54
	Tarde	3.629	52,0	9,9	1,52
	Noche	2.331	43,3	7,5	2,94
Condición en la que vive	Hacinamiento crítico	391	48,7	8,8	0,25
	Hacinamiento medio	5.160	49,2	9,1	0,20
	Sin hacinamiento	27.887	51,3	10,7	-

Variable Socioeconómica		N	Media	DT	d Cohen
Condición de la vivienda	Mala	20.197	50,4	9,6	0,49
	Regular	4.987	46,3	9,3	0,83
	Buena	8.254	55,4	11,7	-
Condición de las TIC	Mala	25.887	49,5	9,6	0,65
	Regular	7.551	56,1	11,8	-
Educación de Madre	Ninguna	662	45,6	8,5	0,90
	Primaria	19.217	50,3	10,4	0,49
	Secundaria	11.109	51,6	10,3	0,36
	Superior	2.175	55,4	11,4	-
	Sin dato	275	50,4	8,7	0,45
Educación de Padre	Ninguna	1.303	47,6	9,4	0,75
	Primaria	19.827	50,4	10,4	0,50
	Secundaria	10.484	51,9	10,3	0,36
	Superior	1.549	55,6	11,8	-
	Sin dato	275	50,4	8,7	0,46

Fuente: elaboración propia

2.5.5 Efecto en Sociales y Ciudadanas

Según la Tabla 19 en la prueba de Sociales y Ciudadanas de las pruebas SABER 11º - 2015-2016, los hombres y mujeres presentan desempeños con diferencias no relevantes ($d \leq 0.2$). Por edad se desempeñan mejor los menores de 18 años con diferencias moderadas o grandes con los de mayor edad. El estrato social más alto tiene mejor desempeño mostrando diferencias no relevantes con el estrato medio, pero diferencias grandes con el bajo. De igual manera se observa que los estudiantes de familias con ingresos más altos se desempeñan mejor en esta competencia y los que pertenecen a instituciones de carácter privado presentaron mejor desempeño que las públicas. Los estudiantes que tienen jornada completa u ordinaria presentaron un más alto desempeño en esta competencia que los de otro tipo de jornada de estudio, observando diferencias grandes con éstas. Las condiciones en que viven los estudiantes, de su vivienda y de las TIC afecta también el desempeño en esta prueba, observando que quienes tienen mejores condiciones presentan mejores resultados. Finalmente, los niveles

educativos de la madre y del padre de los estudiantes son factores que tienen un efecto importante en los resultados de esta competencia, padres con niveles educativos más altos implica mejor desempeño en la prueba de Competencias Sociales y Ciudadanas.

Tabla 19. ***Variables Sociodemográficas y desempeño académico en Sociales y Ciudadanas en las Instituciones de Educación Secundaria del departamento de Nariño de las pruebas SABER 11º - 2015-2016.***

Variable Socioeconómica		N	Media	DT	d Cohen
Género	Femenino	18.564	49,64	11,02	0,14
	Masculino	14.755	51,22	11,53	-
Edad	Menor de 18 años	19.969	53,6	10,6	-
	Entre 18 y 22 años	12.108	46,3	10,3	0,69
	Mayor de 22 años	1.361	38,5	9,4	1,43
Estrato social	Bajo	30.862	49,7	11,0	0,93
	Medio	2.272	59,0	10,5	0,09
	Alto	130	59,9	14,1	-
Ingreso familiar	Menos de 1 SM	17.889	49,0	10,5	1,37
	Entre 1 y menos de 5 SM	14.832	51,7	11,8	0,99
	Entre 5 y menos de 10 SM	355	62,5	11,8	0,07
	10 o más SM	87	63,3	10,8	-
	Sin dato	275	48,1	9,7	1,52
Tipo de Colegio	Privado	3.826	54,7	11,7	-
	Público	29.612	49,8	11,1	0,44
Jornada	Completa u Ordinaria	215	65,7	12,1	-
	Única	149	51,1	9,7	1,31
	Mañana	27.114	50,9	11,1	1,34
	Tarde	3.629	51,7	10,0	1,38
	Noche	2.331	40,0	9,5	2,63
Condición en la que vive	Hacinamiento crítico	391	48,0	10,3	0,23
	Hacinamiento medio	5.160	48,9	10,6	0,16
	Sin hacinamiento	27.887	50,6	11,4	-
Condición de la vivienda	Mala	20.197	50,1	10,6	0,43
	Regular	4.987	44,1	11,2	0,95
	Buena	8.254	54,7	11,0	-

Variable Socioeconómica		N	Media	DT	d Cohen
Condición de las TIC	Mala	25.887	48,9	10,9	0,57
	Regular	7.551	55,1	11,1	-
Educación de Madre	Ninguna	662	43,3	10,8	1,09
	Primaria	19.217	49,6	11,3	0,50
	Secundaria	11.109	51,1	11,0	0,37
	Superior	2.175	55,2	11,0	-
	Sin dato	275	48,1	9,7	0,65
Educación de Padre	Ninguna	1.303	46,1	11,5	0,83
	Primaria	19.827	49,6	11,3	0,51
	Secundaria	10.484	51,5	11,0	0,36
	Superior	1.549	55,4	11,0	-
	Sin dato	275	48,1	9,7	0,67

Fuente: elaboración propia

Capítulo III

MATERIALES Y MÉTODOS

La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Como fuentes de información se utilizaron los datos que se encontraban disponibles, al momento de la investigación, en las bases de datos del Icfes de los resultados de los estudiantes que presentaron las pruebas Saber 11º. Los datos más actualizados eran de los años 2015 y 2016. Para el descubrimiento de patrones asociados al desempeño académico en las pruebas Saber 11º, se construyó un modelo de clasificación basado en árboles de decisión, utilizando el algoritmo J48 de la herramienta WEKA (Witten, Frank & Hall, 2011). Se escogió este modelo porque según la experiencia de algunos autores (Han, Kamber & Pei, 2012), (Sattler & Dunemann, 2001), (Timarán & Millán, 2006), para este tipo de proyectos, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender los resultados obtenidos. Además, la importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación. Por otra parte, se escogió WEKA por ser una herramienta de minería de datos de software libre, distribuida bajo licencia GPL, que contiene una colección de algoritmos para rea-

lizar análisis de datos y modelado predictivo, tiene herramientas para la visualización de estos datos y provee una interfaz gráfica que unifica las herramientas para acceder fácilmente a sus funcionalidades (Calleja, 2010), (García, 2016).

Para el descubrimiento de patrones, se aplicó la metodología CRISP-DM (Chapman et al., 2000), (Villena, 2016). En cuanto a las metodologías para desarrollar análisis de minería de datos y en un intento de normalización del proceso, de forma similar a como se hace en ingeniería para normalizar el proceso de desarrollo software, surgieron a finales de los 90 dos metodologías principales: CRISP-DM (Chapman et al., 2000), (Villena, 2016) y SEMMA (Sample, Explore, Modify, Model, and Assess) (Azevedo & Santos, 2008). Las dos especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase. Azevedo y Santos (2008) comparan ambas implementaciones y llegan a la conclusión de que, aunque se puede establecer un paralelismo claro entre ellas, CRISP-DM es más completo porque tiene en cuenta la aplicación al entorno de negocio de los resultados, y por ello es la que se adoptó popularmente. En encuestas realizadas en KDNuggets en 2002, 2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada, cuatro veces más que SEMMA. La metodología CRISP-DM para proyectos de minería de datos no es la "más actual" o "la mejor", pero es muy útil para comprender esta tecnología o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características (Azevedo & Santos, 2008).

CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos (Hernández et al, 2005) y contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

En la fase de análisis del problema se identificó con exactitud la problemática que se solucionaría utilizando la minería de datos, esto permitió recolectar la información necesaria para interpretar con asertividad los resultados encontrados (Villena, 2016). En

la fase de análisis de los datos se realizó la recolección inicial de datos, para establecer un primer contacto con el problema, familiarizarse con ellos, identificando su calidad y establecer las relaciones más evidentes que permitieron definir las primeras hipótesis. En la fase de preparación se seleccionó los datos a los cuales se les aplicaría una determinada técnica de modelado, limpieza, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato (Villena, 2016). En la fase de modelado se seleccionaron las técnicas de minería de datos más apropiadas para el proyecto. En la fase de evaluación se verificó si el modelo se ajusta a las necesidades establecidas en el proyecto. Se evaluaron los patrones encontrados con el fin de determinar su validez, remover los redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario. Finalmente, en la fase de implementación, se trató de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión del MEN, Icfes y de las instituciones gubernamentales y educativas que velan por la calidad de la educación en Colombia y difundir informes sobre el conocimiento extraído (Villena, 2016).

3.1 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 actual examen Saber 11º es una evaluación estandarizada realizada semestralmente por el Icfes, que tiene como objetivos: servir de criterio para la entrada de estudiantes a las Instituciones de Educación Superior, monitorear la calidad de la formación que ofrecen los establecimientos de educación media y producir información para la estimación del valor agregado de la educación superior (Icfes, 2016).

El examen Saber 11º es una evaluación estandarizada del desarrollo de las competencias de los estudiantes que están por finalizar la educación media. Este examen se diligencia en su totalidad

a lápiz y papel y consta de preguntas cerradas en las pruebas de matemáticas, sociales, ciencias y lectura. Debido a la existencia de dos calendarios académicos en Colombia, este examen tiene dos aplicaciones en el año. Por lo general, en el primer semestre del año, los estudiantes en colegios de calendario B toman el examen, mientras que en el segundo semestre lo toman los estudiantes que pertenecen a colegios de calendario A.

Con el objetivo de consolidar un Sistema Nacional de Evaluación Estandarizada (SNEE) que consiga la alineación de todos los exámenes que lo conforman, la estructura del examen Saber 11º fue modificada a partir del segundo semestre de 2014 para que sus resultados fueran comparables en términos de las competencias evaluadas, con otras pruebas del SNEE como las pruebas Saber 3, 5 y 9, y el examen Saber Pro (Icfes, 2013). Esto llevó a una nueva estructura del examen para evaluar pruebas genéricas: matemáticas, lectura crítica, ciencias naturales, sociales y ciudadanas e inglés. Este examen produce resultados a nivel individual de estudiantes que están próximos a culminar la educación media. Los resultados contienen puntajes del examinando en cada una de las cinco pruebas genéricas en una escala fijada en la segunda aplicación del año 2014 con promedio 50 y desviación estándar 10 (fijar la media y desviación estándar permite establecer una línea de base y tener un punto de referencia para las estimaciones) y un puntaje global, construido a partir de un promedio ponderado de los puntajes en las cinco pruebas genéricas bajo la siguiente fórmula:

$$PG = 5 \times IG$$

Donde:

$$IG = \frac{3 \times Matemáticas + 3\ x\ Lectura + 3\ x\ Ciencias + 3 \times Sociales + Inglés}{13}$$

Matemáticas: puntaje en la prueba de matemáticas.

Lectura: puntaje en la prueba de lectura.

Ciencias: puntaje en la prueba de ciencias naturales.

Sociales: puntaje en la prueba de sociales y ciudadanas.

Inglés: puntaje en la prueba de inglés.

En esta fase, descubrir factores asociados al desempeño académico de los estudiantes nariñenses que encontrándose finalizando el grado undécimo de educación media, presentaron las pruebas Saber 11º, se convirtió en un problema a resolver con minería de datos.

3.2 COMPRENSIÓN DE LOS DATOS

En esta fase, se identificó, recopiló y familiarizó con la información socioeconómica, académica e institucional, disponible en las bases de datos del Icfes, correspondiente a los resultados de los estudiantes de educación media que presentaron las pruebas Saber 11º en los años 2015 y 2016.

Debido a que los archivos planos del Icfes tienen diversas estructuras, inicialmente se construyeron cuatro repositorios con ayuda del SGBD PostgreSQL, denominados *sb*11_20151, *sb*11_20152, *sb*11_20161 y *sb*11_20162, correspondientes, respectivamente, a la información de los estudiantes que presentaron las pruebas Saber 11º en los periodos A y B en los años 2015 y 2016- 2. Las características de estos repositorios se muestran en la Tabla 20.

Tabla 20. ***Características repositorios pruebas Saber 11º - 2015 y 2016***

Repositorio	No. Atributos	No. Registros
*sb*11_20151	128	108258
*sb*11_20152	86	573128
*sb*11_20161	83	74127
*sb*11_20162	81	605192
Total		**1361495**

Fuente: elaboración propia

Analizando los atributos de estos repositorios, se dan algunos casos en los que la información registrada de un año ya no se reporta en otro año y viceversa. También existen casos en los que se solicita la misma información en todos los años, pero con un formato o un nombre de atributo diferente. Con base al diccionario de datos disponible en el FTP del Icfes, se seleccionaron los atributos cuya información se encuentra presente tanto en

el 2015 como en el 2016. Este análisis se muestra en la Tabla 21. Como resultado de este análisis, se construyó un repositorio inicial compuesto por 1.361.495 registros y 69 atributos al cual se lo denominó *sb*11_*inicial*, donde se integraron los repositorios de cada año. De este repositorio se eliminaron los registros de los estudiantes que no son del departamento de Nariño y aquellos que presentaron más de una vez las pruebas de estado Saber 11º. El resultado fue el repositorio *sb*11_*inicial_narino* con 33.438 estudiantes nariñenses que presentaron las pruebas Saber 11º entre los años 2015 y 2016 con 69 atributos. Este repositorio sirvió de base para las subsiguientes fases.

Tabla 21. ***Análisis de atributos de los repositorios pruebas Saber 11º - 2015-2016.***

#	Atributo	15	16	Descripción	Valores
1	ESTU_TIPODOCUMENTO	✕	✕	Tipo de documento	CC - Cédula de ciudadanía CE - Cédula extranjera CR - Certificado Registraduría NIP - Número de identificación personal NUI - Número único de identificación PC - Pasaporte colombiano PE - Pasaporte extranjero PV - Por verificar RC - Registro civil de nacimiento TI - Tarjeta de identidad
2	ESTU_NACIONALIDAD	✕	✕	Nacionalidad	
3	ESTU_GENERO	✕	✕	Género	F - Femenino M - Masculino
4	ESTU_FECHANACIMIENTO	✕	✕	Fecha de nacimiento	
5	PERIODO	✕	✕	Periodo de presentación del examen	20151 20152 20161 20162
6	ESTU_CONSECUTIVO	✕	✕	Código consecutivo identificador del inscrito	

#	Atributo	15	16	Descripción	Valores
7	ESTU_ESTUDIANTE	X	X	Indica si el inscrito realizó la inscripción por medio de un colegio (estudiante) o fue de manera particular (individual)	ESTUDIANTE INDIVIDUAL
8	ESTU_PAIS_RESIDE	X	X	País donde reside actualmente	
9	ESTU_TIENEETNIA	X	X	¿Pertenece usted a un grupo étnico minoritario?	No Si
10	ESTU_ETNIA	X	X	¿Cuál es el grupo étnico minoritario al que pertenece?	Arhuaco Cancuamo Comunidad afrodescendiente Comunidades Rom (Gitanas) Cubeo Emberá Guambiano Huitoto Inga Páez Palenquero Pasto Pijao Raizal Sikuani Tucano Wayúu Zenú Otro grupo étnico minoritario Ninguno
11	ESTU_DEPTO_RESIDE	X	X	Departamento de residencia	
12	ESTU_COD_RESIDE_DEPTO	X	X	Código Dane del departamento de residencia	Numérico [99999 extranjero]
13	ESTU_MCPIO_RESIDE	X	X	Municipio de Residencia	
14	ESTU_COD_RESIDE_MCPIO	X	X	Código Dane del municipio de residencia	
15	ESTU_AREARESIDE	X	X	Área de residencia	Área Rural Cabecera Municipal

#	Atributo	15	16	Descripción	Valores
16	ESTU_VALORPENSION-COLEGIO	X	X	Valor mensual de la pensión que paga actualmente	No paga Pensión Menos de 87.000 Entre 87.000 y menos de 120.000 Entre 120.000 y menos de 150.000 Entre 150.000 y menos de 250.000 250.000 o más
17	ESTU_VECESPRESENTO-EXAMEN	X	X	Número de veces que ha presentado el examen SB11 el ESTUDIANTE	Ninguna vez Una vez Dos veces Tres veces o más
18	FAMI_ESTRATOVIVIENDA	X	X	Estrato socioeconómico de la vivienda según recibo de energía eléctrica	Estrato 1 Estrato 2 Estrato 3 Estrato 4 Estrato 5 Estrato 6 Sin Estrato
19	FAMI_NUMHERMANOS	X	X	Numero de hermanas y hermanos en total	Ninguno Uno Dos Tres Cuatro Cinco Seis Siete Ocho Nueve Más de nueve
20	FAMI_PISOSHOGAR	X	X	Material de los pisos que predomina en la vivienda	Cemento, gravilla, ladrillo. Madera burda, tabla, tablón. Madera pulida, baldosa, tableta, mármol, alfombra. Tierra, arena.
21	FAMI_PERSONASHOGAR	X	X	Número de personas que conforman el hogar donde vive actualmente	Una Dos Tres Cuatro Cinco Seis Siete Ocho Nueve Diez Once Doce o más

#	Atributo	15	16	Descripción	Valores
22	FAMI_CUARTOSHOGAR	✕	✕	Número de habitaciones del hogar	Una Dos Tres Cuatro Cinco Seis Siete Ocho Nueve Diez o más
23	FAMI_EDUCACIONPADRE	✕	✕	Nivel educativo más alto alcanzado por el padre	Ninguno Primaria incompleta Primaria completa Secundaria (Bachillerato) incompleta Secundaria (Bachillerato) completa Técnica o tecnológica incompleta Técnica o tecnológica completa Educación profesional incompleta Educación profesional completa Postgrado No sabe
24	FAMI_EDUCACIONMADRE	✕	✕	Nivel educativo más alto alcanzado por la madre	Ninguno Primaria incompleta Primaria completa Secundaria (Bachillerato) incompleta Secundaria (Bachillerato) completa Técnica o tecnológica incompleta Técnica o tecnológica completa Educación profesional incompleta Educación profesional completa Postgrado No sabe

#	Atributo	15	16	Descripción	Valores
25	FAMI_OCUPACIONPADRE	X	X	Ocupación u oficio del padre	Empleado con cargo como director o gerente general Empleado de nivel auxiliar o administrativo Empleado de nivel directivo Empleado de nivel técnico o profesional Empleado obrero u operario Empresario Hogar Otra actividad u ocupación Pensionado Pequeño empresario Profesional independiente Trabajador por cuenta propia
26	FAMI_OCUPACIONMADRE	X	X	Ocupación u oficio del padre	Empleado con cargo como director o gerente general Empleado de nivel auxiliar o administrativo Empleado de nivel directivo Empleado de nivel técnico o profesional Empleado obrero u operario Empresario Hogar Otra actividad u ocupación Pensionado Pequeño empresario Profesional independiente Trabajador por cuenta propia
27	FAMI_NIVELSISBEN	X	X	Puntaje de SISBEN en el que está clasificada su familia	Nivel 1 Nivel 2 Nivel 3 Está clasificada en otro nivel del SISBEN No está clasificada por el SISBEN
28	FAMI_TIENEINTERNET	X	X	Cuenta con servicio o conexión a internet	No Si
29	FAMI_TIENESERVICIOTV	X	X	Cuenta con servicio de televisión por cable	No Si
30	FAMI_TELEFONO	X	X	Cuenta con Teléfono fijo	No Si
31	FAMI_TIENECOMPUTADOR	X	X	Cuenta con computador	No Si

#	Atributo	15	16	Descripción	Valores
32	FAMI_TIENELAVADORA	X	X	Cuenta con lavadora	No Si
33	FAMI_TIENEMICROONDAS	X	X	Cuenta con microondas	No Si
34	FAMI_TIENEHORNO	X	X	Cuenta con horno eléctrico o a gas	No Si
35	FAMI_TIENEAUTOMOVIL	X	X	Cuenta con automóvil particular	No Si
36	FAMI_TIENEDVD	X	X	Cuenta con DVD	No Si
37	FAMI_NUMLIBROS	X	X	Número de libros físicos o electrónicos hay en el hogar	A 10 LIBROS 11 A 25 LIBROS 26 A 100 LIBROS MÁS DE 100 LIBROS
38	FAMI_INGRESOFMILIARMENSUAL	X	X	Total, de ingresos mensuales del hogar, en términos de salarios mínimos (SMMLV)	Menos de 1 SMLV Entre 1 y menos de 2 SMLV Entre 2 y menos de 3 SMLV Entre 3 y menos de 5 SMLV Entre 5 y menos de 7 SMLV Entre 7 y menos de 10 SMLV 10 o más SMLV
39	ESTU_TRABAJAACTUALMENTE	X	X	Trabaja actualmente	No Si, 20 horas o más a la semana Si, menos de 20 horas a la semana
40	ESTU_RECIBESALARIO	X	X	Recibe algún salario por trabajar	No Si
41	COLE_CODIGO_ Icfes	X	X	Código Icfes de la sede-jornada	
42	COLE_COD_DANE_ESTABLECIMIENTO	X	X	Código Dane del Establecimiento Educativo	
43	COLE_NOMBRE_ESTABLECIMIENTO	X	X	Nombre del Establecimiento Educativo	
44	COLE_GENERO	X	X	Indica el género de la población del Establecimiento.	FEMENINO MASCULINO MIXTO

#	Atributo	15	16	Descripción	Valores
45	COLE_NATURALEZA	X	X	Indica la naturaleza del Establecimiento	NO OFICIAL OFICIAL
46	COLE_CALENDARIO	X	X	Calendario académico del Establecimiento	A B OTRO (Esta opción aplica para la población INDIVIDUAL)
47	COLE_BILINGUE	X	X	Indica si el Establecimiento es bilingüe o no	No Si
48	COLE_CARACTER	X	X	Indica el carácter del Establecimiento	ACADÉMICO TÉCNICO TÉCNICO/ACADEMICO NO APLICA
49	COLE_COD_DANE_SEDE	X	X	Código Dane de la Sede	
50	COLE_NOMBRE_SEDE	X	X	Nombre de la Sede	
51	COLE_SEDE_PRINCIPAL	X	X	Es la sede principal del Establecimiento Educativo	N - No S - Si
52	COLE_AREA_UBICACION	X	X	Área de ubicación de la Sede	Rural Urbana
53	COLE_JORNADA	X	X	Jornada de la Sede	COMPLETA MAÑANA NOCHE SABATINA TARDE UNICA
54	COLE_COD_MCPIO_UBICACION	X	X	Código Dane del municipio donde está ubicada la Sede	
55	COLE_MCPIO_UBICACION	X	X	Nombre del municipio donde está ubicada la Sede	
56	COLE_COD_DEPTO_UBICACION	X	X	Código Dane del departamento donde está ubicada la Sede	

#	Atributo	15	16	Descripción	Valores
57	COLE_DEPTO_UBICACION	X	X	Nombre del departamento donde está ubicada la Sede	
58	ESTU_PRIVADO_LIBERTAD	X	X	Privado de la libertad	N - No S - Si
59	ESTU_COD_MCPIO_PRESENTACION	X	X	Código Dane del municipio presentación del examen	
60	ESTU_MCPIO_PRESENTACION	X	X	Nombre municipio presentación del examen	
61	ESTU_DEPTO_PRESENTACION	X	X	Nombre departamento presentación del examen	
62	ESTU_COD_DEPTO_PRESENTACION	X	X	Código Dane del departamento presentación del examen	
63	PUNT_LECTURA_CRITICA	X	X	Puntaje en Lectura Crítica	Rango [0, 100]
64	PUNT_MATEMATICAS	X	X	Puntaje en Matemáticas	Rango [0, 100]
65	PUNT_C_NATURALES	X	X	Puntaje en Ciencias Naturales	Rango [0, 100]
66	PUNT_SOCIALES_CIUDADANAS	X	X	Puntaje Sociales y Ciudadanas	Rango [0, 100]
67	PUNT_INGLES	X	X	Puntaje Sociales y Ciudadanas	Rango [0, 100]
68	DESEMP_INGLES	X	X	Desempeño en Inglés	A– A1 A2 B+ B1
69	PUNT_GLOBAL	X	X	Puntaje total obtenido	Rango [0, 500]

Fuente: Icfes

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 69 atributos del conjunto de datos *Sb*11*_inicial_narino* se clasificaron en cuatro dimensiones: sociodemográfica, económica, académica e institucional. En la Tabla 22 se muestra esta clasificación.

Tabla 22. ***Clasificación de atributos en dimensiones***

Dimensión	Atributo
Sociodemográfica	ESTU_TIPODOCUMENTO, ESTU_NACIONALIDAD, ESTU_GENERO, ESTU_FECHANACIMIENTO, ESTU_CONSECUTIVO, ESTU_ESTUDIANTE, ESTU_PAIS_RESIDE, ESTU_TIENEETNIA, ESTU_ETNIA, ESTU_DEPTO_RESIDE, ESTU_COD_RESIDE_DEPTO, ESTU_MCPIO_RESIDE, ESTU_COD_RESIDE_MCPIO, ESTU_AREARESIDE, ESTU_VECESPRESENTOEXAMEN, FAMI_NUMHERMANOS, FAMI_PERSONASHOGAR, FAMI_CUARTOSHOGAR, FAMI_TIENEINTERNET, FAMI_TIENESERVICIOTV, FAMI_TELEFONO, FAMI_TIENECOMPUTADOR, FAMI_TIENELAVADORA, FAMI_TIENEMICROONDAS, FAMI_TIENEHORNO, FAMI_TIENEAUTOMOVIL, FAMI_TIENEDVD, FAMI_NUMLIBROS, ESTU_PRIVADO_LIBERTAD, ESTU_COD_MCPIO_PRESENTACION, ESTU_MCPIO_PRESENTACION, ESTU_DEPTO_PRESENTACION, ESTU_COD_DEPTO_PRESENTACION
Económica	FAMI_ESTRATOVIVIENDA , FAMI_PISOSHOGAR, FAMI_OCUPACIONPADRE, ,FAMI_OCUPACIONMADRE, FAMI_NIVELSISBEN, FAMI_INGRESOFMILIARMENSUAL, ESTU_TRABAJAACTUALMENTE, ESTU_RECIBESALARIO
Académica	PERIODO, FAMI_EDUCACIONPADRE, FAMI_EDUCACIONMADRE, PUNT_LECTURA_CRITICA, PUNT_MATEMATICAS, PUNT_C_NATURALES, PUNT_SOCIALES_CIUDADANAS, PUNT_INGLES, DESEMP_INGLES, PUNT_GLOBAL
Institucional	COLE_CODIGO_ Icfes, COLE_COD_DANE_ESTABLECIMIENTO, COLE_NOMBRE_ESTABLECIMIENTO, COLE_GENERO, COLE_NATURALEZA, COLE_CALENDARIO, COLE_BILINGUE, COLE_CARACTER, COLE_COD_DANE_SEDE, COLE_NOMBRE_SEDE, COLE_SEDE_PRINCIPAL, COLE_AREA_UBICACION, COLE_JORNADA, COLE_COD_MCPIO_UBICACION, COLE_MCPIO_UBICACION, COLE_COD_DEPTO_UBICACION, COLE_DEPTO_UBICACION

Fuente: esta investigación

3.3 PREPARACIÓN DE LOS DATOS

En esta fase se realizó inicialmente un análisis de la calidad de los datos del repositorio *Sb*11*_inicial_narino*, con el fin de conocer por cada atributo el número de valores distintos, el número de valores nulos, el valor máximo, valor mínimo, moda, media y un histograma para determinar cuáles técnicas de limpieza de datos se debían aplicar.

Los 69 atributos del repositorio base, considerados por el Icfes como los más importantes para capturar la información de las pruebas Saber 11º comunes en 2015 y 2016, 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, se realizaron transformaciones o cambios de formato a los valores de los atributos que se consideraron necesarios, se eliminaron los atributos reemplazados, así como los registros de estudiantes que presentaron más de una vez las pruebas Saber 11º. 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. Como resultado de esta fase se obtuvo un repositorio de datos limpio y transformado, con 33.438 y 23 atributos, listo para aplicarle las técnicas de minería de datos y al cual se le denominó *Sb*11*_final_narino*. En la Tabla 23 se muestra el diccionario de datos de este repositorio.

Tabla 23. ***Diccionario de datos Sb11_final_narino***

No.	Atributo	Descripción	Valores
Socioeconómicos			
1	estu_genero	Sexo del estudiante	M, F
2	estu_edad_intervalo	Rango de edad del estudiante en el momento de presentar la prueba	<18 Entre 18 y 22 >22

No.	Atributo	Descripción	Valores
3	fami_estrato	Estrato socioeconómico del estudiante	BAJO, MEDIO, ALTO
4	fami_nivel_sisben	Nivel de clasificación en el SISBEN al que pertenece el estudiante	NIVELES 1, 2, 3, OTRO NIVEL, NO ESTÁ EN SISBEN
5	fami_ingreso_fmiliar_mensual	Ingresos mensuales familiares en salarios mínimos	Hasta 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más 10 salarios
6	fami_educa_madre	Máximo nivel educativo de la madre	PRIMARIA, SECUNDARIA, TÉCNICO, TECNOLÓGICO, PROFESIONAL, POSTGRADO, NINGUNO
7	fami_educa_padre	Máximo nivel educativo del padre	PRIMARIA, SECUNDARIA, TÉCNICO, TECNOLÓGICO, 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
10	eco_condicion_electrodomesticos	Condición relacionado con los electrodomésticos en hogar del estudiante	BUENA, MALA, REGULAR
11	econ_condicion_vivienda	Condición de la vivienda del estudiante	BUENA, MALA, REGULAR
12	eco_condicion_tic	Condición de uso de TIC en el hogar del estudiante	BUENA, REGULAR, MALA
13	eco_condicion_vive	Condición de vida del estudiante	SIN HACINAMIENTO, HACINAMIENTO MEDIO, HACINAMIENTO CRÍTICO

No.	Atributo	Descripción	Valores
Académicos			
14	punt_global_cuali	Puntaje cualitativo global obtenido por el estudiante en las pruebas Saber 11º	POR ENCIMA DE LA MEDIA NACIONAL POR DEBAJO DE LA MEDIA NACIONAL
15	punt_lectura_critica_cuali	Puntaje cualitativo en lectura crítica obtenido por el estudiante en las pruebas Saber 11º	POR ENCIMA DE LA MEDIA NACIONAL POR DEBAJO DE LA MEDIA NACIONAL
16	punt_matematicas_cuali	Puntaje cualitativo en matemáticas obtenido por el estudiante en las pruebas Saber 11º	POR ENCIMA DE LA MEDIA NACIONAL POR DEBAJO DE LA MEDIA NACIONAL
17	punt_c_:naturales_cuali	Puntaje cualitativo en ciencias naturales obtenido por el estudiante en las pruebas Saber 11º	POR ENCIMA DE LA MEDIA NACIONAL POR DEBAJO DE LA MEDIA NACIONAL
18	punt_ingles_cuali	Puntaje cualitativo en ingles obtenido por el estudiante en las pruebas Saber 11º	POR ENCIMA DE LA MEDIA NACIONAL POR DEBAJO DE LA MEDIA NACIONAL
19	punt_sociales_ciudadanas_cuali	Puntaje cualitativo en sociales y ciudadanas obtenido por el estudiante en las pruebas Saber 11º	POR ENCIMA DE LA MEDIA NACIONAL POR DEBAJO DE LA MEDIA NACIONAL
Institucionales			
20	Tipo_cole	Tipo de institución educativa	PÚBLICA, PRIVADA
21	Cole_jornada	Jornada de estudio del estudiante	MAÑANA, TARDE, NOCHE, ÚNICA, SABATINA-DOMINICAL
22	cole_subregion	Subregión geográfica de Nariño donde se encuentra la institución educativa	Ver TABLA 25.

Fuente: elaboración propia

El proceso de limpieza y transformación de este repositorio se realizó de la siguiente manera. 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; tampoco se tomaron en cuenta los atributos con valores constantes y aquellos que servían de identificadores de cada estudiante.

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 *Sb*11*_final_narino*, 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 24 se detalla este proceso para el atributo *estu_edad* que fue reemplazado por el nuevo atributo *estu_edad_intervalo*.

Tabla 24. ***Valores discretizados del atributo estu_intervalo***

Valor	No. Estudiantes
Edad hasta 18	19.969
Edad entre 18 y 22	12.108
Edad mayor 22	1.361
Total	**33.438**

Fuente: esta investigación

En esta investigación, se consideró conveniente crear un nuevo atributo denominado *cole_subregion* que hace referencia a la subregión geográfica del departamento de Nariño donde se encuentra la institución educativa, teniendo en cuenta los municipios del departamento. Esta clasificación se describe en la Tabla 25.

Tabla 25. ***Subregiones de Nariño***

Subregión	Municipios
Centro	Pasto, Chachagüí, La Florida, Nariño, Tangua, Yacuanquer.
Guambuyaco	El Peñol, El Tambo, La Llanada, Los Andes, Sotomayor.
Juanambú	Arboleda, Buesaco, La Unión, San Pedro de Cartago, San Lorenzo.
La Cordillera	Cumbitara, El Rosario, Leiva, Policarpa, Taminango.
La Sabana	Guaitarilla, Imués, Ospina, Sapuyes, Túquerres.
Los Abades	Providencia, Samaniego, Santacruz.
Obando	Aldana, Contadero, Córdoba, Cuaspud, Cumbal, Funes, Iles, Ipiales, Guachucal, Gualmatán, Potosí, Pupiales, Puerres.
Occidente	Ancuya, Consacá, Linares, Sandoná.
Pacífico Sur	Francisco Pizarro, Tumaco.
Piedemonte Costero	Mallama, Ricaurte.
Río Mayo	Albán, Belén, Colón, El Tablón de Gómez, La Cruz, San Bernardo, San Pablo.
Sanquianga	El Charco, La Tola, Mosquera, Olaya Herrera, Santa Bárbara.
Telembí	Barbacoas, Magüí Payán, Roberto Payán.

Fuente: Instituto Geográfico Agustín Codazzi IGAC - Diccionario Geográfico Gobernación del Departamento de Nariño.

Se creó un índice para determinar la condición de la vivienda, teniendo en cuenta el material de los pisos del atributo *fami_piso_hogar*. A este índice se le denominó *eco_condición_vivienda*. En la Tabla 26 se muestran estos valores y el material de los pisos.

Tabla 26. ***Valores del atributo eco_condición_vivienda***

Material de los Pisos de la Vivienda	Condición Vivienda
Tierra, Arena	MALA
Cemento, Gravilla, Ladrillo	MALA
Madera Burda, Tabla - Tablón	REGULAR
Madera Pulida, Baldosa, Tableta, Mármol, Alfombra	BUENA
Madera Pulida, Mármol, Alfombra - Tapete de Pared a Pared	BUENA

Fuente: esta investigación.

De igual manera se creó un índice para determinar la condición de electrodomésticos con que cuenta el hogar del estudiante. Este índice se denominó *eco_condición_electrodomesticos.* 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 es 4 la condición de electrodomésticos es BUENA, si el índice está entre 2 y 3 la condición de electrodomésticos es MEDIA y finalmente, si el índice está entre 0 y 1 de electrodomésticos es MALA. En la Tabla 27 se muestran las variables que intervienen en el cálculo del índice *eco_condición_electrodomesticos*.

Tabla 27. ***Cálculo del índice de eco_condición_electrodomésticos***

Electrodomésticos	Si	No
fami_tienelavadora	1	0
fami_tienemicroondas	1	0
fami_tienehorno	1	0
fami_tienedvd	1	0

Fuente: esta investigación.

Se creó el índice *eco_condición_tic* que tiene en cuenta el número de servicios con que cuenta la vivienda. Se procedió de igual manera para los valores de índice *eco_condición_tic* para los valores BUENA, MEDIA y MALA. En la Tabla 28 se muestran los atributos que intervienen en el cálculo del índice *eco_condición_tic*.

Tabla 28. ***Cálculo del índice de eco_condición_tic***

Servicios	Si	No
fami_tieneinternet	1	0
fami_tienetv	1	0
fami_telefono	1	0
fami_tienecomputador	1	0

Fuente: esta investigación.

Se creó el índice *eco_condición_vive*. Para asignar los valores a este índice, 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_personashogar* entre *fami_cuartoshogar*. 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.

3.4 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 socioeconómicos, académicos e institucionales asociados al rendimiento académico en las pruebas Saber 11º que presentaron los estudiantes nariñenses de educación media en los años 2015 y 2016, 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). Las descriptivas se aplican a problemas en los que hay que describir los valores de los datos existentes (Hernández et al., 2005).

Como tarea predictiva se escogió clasificación con árboles de decisión (Han, Kamber & Pei, 2012). En esta tarea se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes de educación media, los factores socioeconómicos, académicos e institucionales asociados a un probable buen o mal desempeño académico en las pruebas Saber 11º.

Como tarea descriptiva se escogió clustering. Con clustering se pretende agrupar los estudiantes teniendo en cuenta las similitu-

des en los resultados de las pruebas Saber 11º. Los resultados del modelado se describen en el capítulo siguiente.

3.5 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 y Discusión.

3.6 IMPLEMENTACIÓN

En esta fase, a través de la difusión de los informes de esta investigación, el conocimiento descubierto se incorporará al existente y se podrá integrar a los procesos de toma de decisiones del MEN, Icfes y de las instituciones educativas nariñenses que velan por la calidad de la educación media y superior en Colombia y en el departamento de Nariño. Una vez estas instituciones intervengan los factores asociados al desempeño académico en las Pruebas Saber 11º, será posible analizar los resultados y determinar sus efectos.

Capítulo IV

RESULTADOS

4.1 DESCUBRIMIENTO DE PATRONES PREDICTIVOS ASOCIADOS A LAS PRUEBAS SABER 11º

4.1.1 Factores asociados al desempeño académico global en las Pruebas Saber 11º

Se seleccionó la tarea de clasificación con árboles de decisión como la técnica predictiva de minería de datos más adecuada para descubrir patrones asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media en las pruebas Saber 11º. 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 & Pei, 2012), (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).

Con esta técnica se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes de grado 11 del departamento de Nariño, cuales son los factores socioeconómicos, académicos e institucionales asociados al buen (por encima de la media) o mal (por debajo de la media) desempeño académico en las pruebas Saber 11º, teniendo en cuenta el puntaje global obtenido por el estudiante en las pruebas Saber 11º y los puntajes en Lectura Crítica, Matemáticas, Ciencias Naturales, Inglés y en Sociales y Ciudadanas , como atributo clase.

Para la construcción de los modelos de clasificación con árboles de decisión se utilizó la herramienta WEKA ver 3.9.4 (Witten et al, 2011) (Hall et al., 2011) y su algoritmo J48, el cual implementa al algoritmo C.45, (Quinlan, 1993). El algoritmo J48 se basa en la utilización del criterio de ganancia de información (*information gain*). 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 tuvo en cuenta para la poda fue el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25% y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños (García & Álvarez, 2010). Otro parámetro utilizado para variar el tamaño del árbol fue a través del factor M que especifica el mínimo número de instancias o registros por nodo del árbol (Witten et al, 2011).

Antes de construir los modelos se definió el procedimiento para probar la calidad de estos y su validez, teniendo en cuenta que, para entrenar y probar un modelo de clasificación, se divide los datos en dos conjuntos: entrenamiento y prueba (Witten et al, 2011). 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).

Por esta razón, se utilizó la validación cruzada como método de evaluación del clasificador. Para este caso particular se utilizó el método de evaluación validación cruzada con n pliegues (n – *fold cross validation)*. Este método consiste en dividir el conjunto de entrenamiento en n subconjuntos disjuntos de similar tamaño llamados pliegues (*folds*) de forma aleatoria. El número de subconjuntos se puede introducir en el campo Folds. Posteriormente se realizan *n* iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes $n - 1$ (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último, se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones. Otra ventaja de la validación cruzada es que la varianza de los n errores de muestra parciales, permite estimar la variabilidad del método de aprendizaje con respecto al con-

junto de datos. Para esta investigación, se utilizó 10 particiones (10 – *fold cross validation)* teniendo en cuenta lo recomendado por Hernández et al. (2005).

Por otra parte, se evaluó o estimó el coste del clasificador para el repositorio *sb_final_narino* a través de la matriz de confusión. La matriz de confusión representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila $i, i = 1 \cdots n$ constituyen el número de instancias que realmente pertenecen a la clase i. Similarmente la sumatoria de los ejemplos o registros en cada columna $j, j = 1 \cdots n$ son las instancias que ha predicho el algoritmo al valor j de la clase. Los valores en la diagonal son los aciertos y se los conoce, en el caso de que el atributo clase tenga dos valores, como verdaderos positivos *(VP)* y verdaderos negativos *(VN)* y el resto son los errores de clasificación conocidos como falsos positivos *(FP)* y falsos negativos *(FN)* (ejemplos que pertenecían a la clase i de la fila i y fueron clasificados incorrectamente en otra) (Fernández, 2009).

Debido a que el clasificador pudiera estar reconociendo solamente los verdaderos positivos o los verdaderos negativos, entonces se evaluó la manera como el clasificador reconoce dichas clases, utilizando las métricas conocidas como sensibilidad y especificidad. La sensibilidad mide la proporción de verdaderos positivos del clasificador, mientras que la especificidad mide la proporción de verdaderos negativos. Para calcular dichas métricas se utilizaron las siguientes fórmulas:

$$Sensibilidad = \frac{VP}{VP + FN}$$

$$Especificidad = \frac{VN}{VN + FP}$$

Donde *VP* son los verdaderos positivos, *VN* los verdaderos negativos, *FP* los falsos positivos y *FN* los falsos negativos.

Teniendo en cuenta los parámetros de evaluación anteriores, se procedió a construir los diferentes árboles de decisión con el

algoritmo *J*48. Se escogieron como clase los puntajes obtenidos por los estudiantes en cada una de las pruebas Saber 11º, las cuales fueron discretizados en los valores *"por encima de la media nacional"* , y *"por debajo de la media nacional"* . En la Tabla 29 se muestra la distribución de los estudiantes de *sb*11*_final_narino* con respecto a cada uno de los atributos clase que se escogieron para las pruebas Saber 11º.

Tabla 29. *Clases de Sb11_final_narino*

Clase	Bajo la media	%	Sobre la media	%	Media nacional
Punt_global	16.758	50,1	16.680	49,9	257,82/500
Punt_lectura_critica	17.920	53,6	15.518	46,4	51,76/100
Punt_matemáticas	17.474	52,3	15.964	47,7	51,35/100
Punt_c_naturales	17.364	51,9	16.074	48,1	52,07/100
Punt_sociales_ciudadanas	17.056	51,0	16.382	49,0	50,93/100
Punt_ingles	20.485	61,3	12.953	38,7	52,04/100

Fuente: esta investigación.

4.1.2 Factores asociados al desempeño académico en lectura crítica en las pruebas Saber 11º

Se escogió como clase el puntaje en lectura crítica de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores "por encima de la media nacional ", y "por debajo de la media nacional", siendo la media nacional 52 sobre 100. Con el fin de obtener diferentes modelos de árboles y escoger el de mejores resultados, se establecieron 2 porcentajes de prepoda del árbol para el factor *M* igual a 1% y 2% del total de registros del repositorio de datos, y 2 porcentajes para el factor confianza *C* igual a 25% y 50% y se construyeron los diferentes modelos combinando estos factores. Se escogió el árbol construido con los parámetros $M = 33C$ (1%) y $C = 25\%$ por los mejores resultados obtenidos y por la facilidad de análisis de los patrones. Una vez construidos 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 1% y una confianza del 65%. En la Figura 1 se muestra la precisión del árbol

y su matriz de confusión. El árbol construido con los parámetros $M = 33C$ y $C = 25\%$ se muestra en la Figura 2.

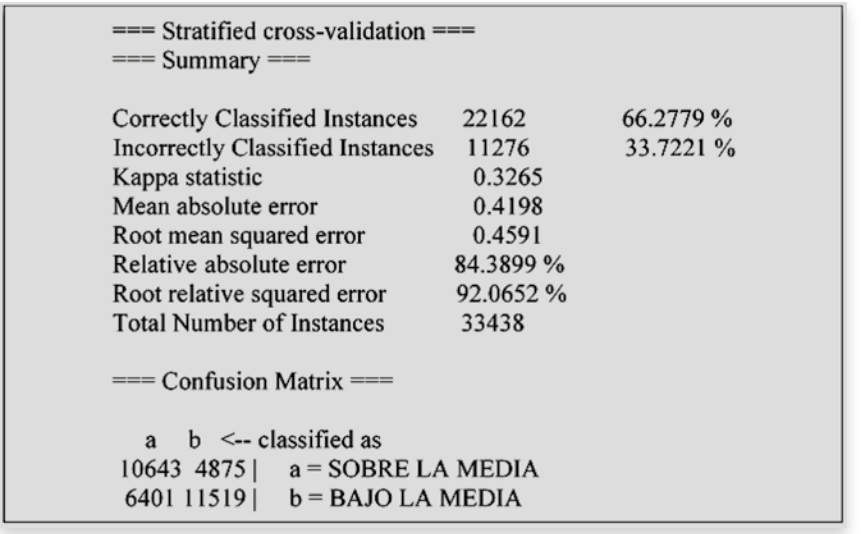


Figura 1. *Precisión y matriz de confusión del árbol de lectura crítica.*

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos *sb11_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las pruebas Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje en lectura crítica (*puntaje_lectura_critica_cuali*) como clase, se puede observar que este clasifica correctamente a 22.162 instancias, que corresponde a un porcentaje de precisión del 66% y 11.276 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 34% (ver Figura 1).

Teniendo en cuenta la matriz de confusión (Figura 1), del total de 33.438 estudiantes evaluados, el modelo clasifica a 17.044 estudiantes con desempeño académico sobre la media, correspondiente a un 51% del total de estudiantes y a 16.394 estudiantes con un desempeño académico bajo la media, que corresponde al 49%. Del 51% de estudiantes que están sobre la media, el modelo clasifica correctamente a 10.643 estudiantes, que corresponde a un

porcentaje de 62% (casos correctos) y clasifica incorrectamente a 6.401 estudiantes, correspondiente al 38% (falsos casos). Del 49% de estudiantes que están bajo la media, el modelo clasifica correctamente a 11.519 estudiantes, que corresponde a un porcentaje de 70% (casos correctos) e incorrectamente a 4.875 estudiantes, correspondiente al 30% (falsos casos).

```
J48 pruned tree
------------------
estu_edad_intervalo = Entre 18 y 22 años
|  subregion = PACIFICO SUR: BAJO LA MEDIA (2333.0/243.0)
|  subregion = OBANDO
|  |  fami_educa_madre = Primaria incompleta: BAJO LA MEDIA (675.0/256.0)
|  |  fami_educa_madre = Primaria completa: BAJO LA MEDIA (572.0/218.0)
|  subregion = LA CORDILLERA: BAJO LA MEDIA (404.0/109.0)
|  subregion = CENTRO
|  |  fami_nivel_sisben = Nivel 1
|  |  |  cole_jornada = Mañana
|  |  |  |  fami_educa_madre = Primaria incompleta: BAJO LA MEDIA (538.0/202.0)
|  |  |  |  fami_educa_madre = Primaria completa: BAJO LA MEDIA (363.0/132.0)
|  |  |  cole_jornada = Tarde: BAJO LA MEDIA (400.0/177.0)
|  |  fami_nivel_sisben = Nivel 2: BAJO LA MEDIA (558.0/265.0)
|  |  fami_nivel_sisben = No está clasificada: SOBRE LA MEDIA (397.0/144.0)
|  subregion = LOS ABADES: BAJO LA MEDIA (368.0/120.0)
|  subregion = TELEMBI: BAJO LA MEDIA (544.0/51.0)
|  subregion = JUANAMBU: BAJO LA MEDIA (722.0/272.0)
|  subregion = SANQUIANGA: BAJO LA MEDIA (636.0/55.0)
|  subregion = OCCIDENTE: BAJO LA MEDIA (412.0/155.0)
estu_edad_intervalo = Menor que 18 años
|  subregion = GUAMBUYACO: SOBRE LA MEDIA (489.0/204.0)
|  subregion = PACIFICO SUR: BAJO LA MEDIA (1660.0/557.0)
|  subregion = OBANDO: SOBRE LA MEDIA (4212.0/1688.0)
|  subregion = CENTRO: SOBRE LA MEDIA (8020.0/2771.0)
|  subregion = RIO MAYO: SOBRE LA MEDIA (1172.0/462.0)
|  subregion = OCCIDENTE: SOBRE LA MEDIA (724.0/288.0)
estu_edad_intervalo = Mayor que 22 años: BAJO LA MEDIA (1361.0/117.0)

Number of Leaves  :  54
Size of the tree :  61
```

Figura 2. *Árbol textual de la prueba lectura crítica.*

Con respecto al total de casos correctos que están sobre la media según la Tabla 29, en el puntaje de lectura crítica (15.518), la exactitud del modelo es del 68,6%, esto significa que la sensibilidad del modelo es de 0,69. De igual manera, con respecto al total de casos correctos que están bajo la media (17.920), la exactitud del modelo es del 64,3%, es decir la especificidad del modelo es 0,64.

Se escogieron los patrones más representativos del mejor árbol obtenido (ver Figura 2), teniendo en cuenta un mínimo soporte del 1% y una confianza mínima de 60%, tanto los que se ubican

por encima de la media, como aquellos que se sitúan por debajo de ella. Entre los patrones más importantes están:

Regla 1. Si el estudiante tiene entre 18 y 22 años y es de la subregión de Pacífico Sur del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 7% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 89,6% de los 2333 estudiantes que se clasifican así, están correctamente clasificados y el 11,7% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 2. Si el estudiante tiene entre 18 y 22 años, es de la subregión de Obando del departamento de Nariño y la educación de la madre es primaria incompleta, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 62,1% de los 675 estudiantes que se clasifican así, están correctamente clasificados y el 2,3% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 3. Si el estudiante tiene entre 18 y 22 años, es de la subregión de Obando del departamento de Nariño y la educación de la madre es primaria completa, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 1,7% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 61,9% de los 572 estudiantes que se clasifican así, están correctamente clasificados y el 2% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 4. Si el estudiante tiene entre 18 y 22 años y es de la subregión de La Cordillera del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 1,2% del total de estudiantes de Nariño que presentaron las pruebas Saber

11º se clasifican de esta manera. El 73% de los 404 estudiantes que se clasifican así, están correctamente clasificados y el 1,6% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 5. Si el estudiante tiene entre 18 y 22 años, es de la subregión Centro del departamento de Nariño, es de nivel sisben 1, estudia en la jornada de la mañana y la educación de la madre es primaria incompleta entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 1,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 62,5% de los 538 estudiantes que se clasifican así, están correctamente clasificados y el 1,9% de los 17.920 estudiantes que están bajo la media (ver tabla 29), cumplen este patrón.

Regla 6. Si el estudiante tiene entre 18 y 22 años, es de la subregión Centro del departamento de Nariño, es de nivel sisben 1, estudia en la jornada de la mañana y la educación de la madre es primaria completa entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 1,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 63,6% de los 363 estudiantes que se clasifican así, están correctamente clasificados y el 1,3% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 7. Si el estudiante tiene entre 18 y 22 años, es de la subregión Centro del departamento de Nariño y no está clasificado en el sisben, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté sobre la media nacional. El 1,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 63,7% de los 397 estudiantes que se clasifican así, están correctamente clasificados y el 1,6% de los 15.518 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 8. Si el estudiante tiene entre 18 y 22 años y es de la subregión Los Abades del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 1,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 67,4% de los 368 estudiantes que se clasifican así, están correctamente clasificados y el 1,4% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 9. Si el estudiante tiene entre 18 y 22 años y es de la subregión Telembí del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 1,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 90,6% de los 544 estudiantes que se clasifican así, están correctamente clasificados y el 2,8% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 10. Si el estudiante tiene entre 18 y 22 años y es de la subregión Juanambú del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 2,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 62,63% de los 722 estudiantes que se clasifican así, están correctamente clasificados y el 2,5% de los 17.920 estudiantes que están bajo la media (ver tabla 29), cumplen este patrón.

Regla 11. Si el estudiante tiene entre 18 y 22 años y es de la subregión Sanquianga del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 1,9% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 91,3% de los 636 estudiantes que se clasifican así, están correctamente clasificados y el 3,2%

de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 12. Si el estudiante tiene entre 18 y 22 años y es de la subregión Occidente del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 1,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 62,4% de los 412 estudiantes que se clasifican así, están correctamente clasificados y el 1,4% de los 17.920 estudiantes que están bajo la media (ver tabla 29), cumplen este patrón.

Regla 13. Si el estudiante es menor que 18 años y es de la subregión Pacífico Sur del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 5% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 66,4% de los 1.660 estudiantes que se clasifican así, están correctamente clasificados y el 6,2% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 14. Si el estudiante es menor que 18 años y es de la subregión Obando del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté sobre la media nacional. El 12,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60% de los 4212 estudiantes que se clasifican así, están correctamente clasificados y el 16,3% de los 15.518 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 15. Si el estudiante es menor que 18 años y es de la subregión Centro del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté sobre la media nacional. El 24% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se

clasifican de esta manera. El 65,4% de los 8020 estudiantes que se clasifican así, están correctamente clasificados y el 33,8% de los 15.518 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 16. Si el estudiante es menor que 18 años y es de la subregión Juanambú del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté sobre la media nacional. El 24% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 65,4% de los 8.020 estudiantes que se clasifican así, están correctamente clasificados y el 31,5% de los 15.518 estudiantes que están sobre la media (ver tabla 29), cumplen este patrón.

Regla 17. Si el estudiante es menor que 18 años y es de la subregión Río Mayo del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté sobre la media nacional. El 3,5% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,6% de los 1.172 estudiantes que se clasifican así, están correctamente clasificados y el 4,6% de los 15.518 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 18. Si el estudiante es menor que 18 años y es de la subregión occidente del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté sobre la media nacional. El 2,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,2% de los 724 estudiantes que se clasifican así, están correctamente clasificados y el 2,8% de los 15.518 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 19. Si el estudiante es menor que 18 años y es de la subregión Occidente del departamento de Nariño, entonces su desempeño académico en la prueba de lectura crítica del Saber

11º es posible que esté sobre la media nacional. El 2,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,2% de los 724 estudiantes que se clasifican así, están correctamente clasificados y el 2,8% de los 15.518 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 20. Si el estudiante es mayor que 18 años, entonces su desempeño académico en la prueba de lectura crítica del Saber 11º es posible que esté bajo la media nacional. El 4,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 91,4% de los 1361 estudiantes que se clasifican así, están correctamente clasificados y el 6,9% de los 17.920 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

4.1.3 Factores asociados al desempeño académico en matemáticas en las pruebas Saber 11º

Se escogió como clase el puntaje en matemáticas de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores "por encima de la media nacional "y "por debajo de la media nacional", siendo la media nacional 51 sobre 100.

Con el fin de obtener diferentes modelos de árboles por competencia y reglas de clasificación generalizadas hasta reglas más detalladas, se establecieron 2 porcentajes de prepoda del árbol para el factor *M* igual a 1% y 2% del total de registros del repositorio de datos, y 2 porcentajes para el factor confianza *C* igual a 25% y 50%. Se construyeron los diferentes modelos de árboles combinando estos factores. Se escogió el árbol construido con los parámetros M=330 (1%) y C=25% por los mejores resultados obtenidos y por la facilidad de análisis de los patrones. Una vez construidos 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 1% y una confianza del 65%. En la Figura 3 se muestra la precisión del árbol y su matriz de confusión. El árbol construido con los parámetros M=330 y C=25% se muestra en la Figura 4.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       23065               68.9784 %
Incorrectly Classified Instances     10373               31.0216 %
Kappa statistic                          0.3812
Mean absolute error                      0.4049
Root mean squared error                  0.45
Relative absolute error                 81.1445 %
Root relative squared error             90.0949 %
Total Number of Instances            33438

=== Confusion Matrix ===

    a     b   <-- classified as
 11643  4321 |     a = SOBRE LA MEDIA
  6052 11422 |     b = BAJO LA MEDIA
```

Figura 3. *Precisión y matriz de confusión del árbol de matemáticas.*

```
J48 pruned tree
------------------

estu_edad_intervalo = Entre 18 y 22 años
|   estu_genero = F: BAJO LA MEDIA (6205.82/1434.99)
|   estu_genero = M
|   |   eco_condicion_tic = MALA: BAJO LA MEDIA (4997.82/1711.57)
estu_edad_intervalo = Menor que 18 años
|   fami_nivel_sisben = Nivel 1
|   |   subregion = GUAMBUYACO: SOBRE LA MEDIA (434.0/158.0)
|   |   subregion = PACIFICO SUR: BAJO LA MEDIA (1594.0/475.0)
|   |   subregion = OBANDO: SOBRE LA MEDIA (3101.0/1147.0)
|   |   subregion = CENTRO
|   |   |   cole_jornada = Mañana: SOBRE LA MEDIA (3148.0/1213.0)
|   |   |   cole_jornada = Tarde: SOBRE LA MEDIA (818.0/298.0)
|   |   subregion = LOS ABADES: SOBRE LA MEDIA (456.0/179.0)
|   |   subregion = JUANAMBU
|   |   |   estu_genero = M: SOBRE LA MEDIA (431.42/137.42)
|   |   subregion = RIO MAYO: SOBRE LA MEDIA (1088.0/358.0)
|   |   subregion = OCCIDENTE: SOBRE LA MEDIA (690.0/199.0)
|   |   subregion = LA SABANA: SOBRE LA MEDIA (654.0/251.0)
|   fami_nivel_sisben = Nivel 2: SOBRE LA MEDIA (1941.0/626.0)
|   fami_nivel_sisben = No esta clasificada: SOBRE LA MEDIA (3110.0/704.0)
estu_edad_intervalo = Mayor que 22 años: BAJO LA MEDIA (1361.0/89.0)

Number of Leaves  :     28

Size of the tree :      35
```

Figura 4. *Árbol textual de la prueba matemáticas.*

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos *sb*11*_final_narino* , en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje en matemáticas (*punt_matematicas_cuali*) como clase, se puede observar que este clasifica correctamente a 23065 instancias, que corresponde a un porcentaje de precisión del 69% y 10.373 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 31% (ver Figura 3).

Teniendo en cuenta la matriz de confusión (Figura 3), del total de 33.438 estudiantes evaluados, el modelo clasifica a 17.695 estudiantes con desempeño académico sobre la media, correspondiente a un 53% del total de estudiantes y a 15.743 estudiantes con un desempeño académico bajo la media, que corresponde al 47%. Del 53% de estudiantes que están sobre la media, el modelo clasifica correctamente a 11.643 estudiantes, que corresponde a un porcentaje de 66% (casos correctos) y clasifica incorrectamente a 6.052 estudiantes, que corresponde a un porcentaje de 34% (falsos casos). Del 47% de estudiantes que están bajo la media, el modelo clasifica correctamente a 11.422 estudiantes, que corresponde a un porcentaje de 73% (casos correctos) e incorrectamente a 4.321 estudiantes, correspondiente al 27% (falsos casos).

Con respecto al total de casos correctos que están sobre la media según la Tabla 29, en el puntaje de matemáticas (15.964), la exactitud del modelo es del 72,9%, esto significa que la sensibilidad del modelo es de 0,73. De igual manera, con respecto al total de casos correctos que están bajo la media (17.920), la exactitud del modelo es del 65,4%, es decir la especificidad del modelo es de 0,65.

Se escogieron los patrones más representativos del mejor árbol obtenido (ver Figura 4), teniendo en cuenta un mínimo soporte del 1% y una confianza mínima de 60%, tanto los que se ubican por encima de la media, como aquellos que se sitúan por debajo de ella. Entre los patrones más importantes están:

Regla 1. Si el estudiante tiene entre 18 y 22 años y es de género femenino, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté bajo la media nacional. El 18,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 76,9% de los 6205 estudiantes que se clasifican así, están correctamente clasificados y el 27,3% de los 17.474 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 2. Si el estudiante tiene entre 18 y 22 años, es de género masculino y su condición TIC es mala, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté bajo la media nacional. El 14,9% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 65,8% de los 4997 estudiantes que se clasifican así, están correctamente clasificados y el 18,8% de los 17.474 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 3. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1 y es de la subregión Guambuyaco del departamento de Nariño, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 1,3% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 63,6% de los 434 estudiantes que se clasifican así, están correctamente clasificados y el 1,7% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 4. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1 y es de la subregión Pacífico Sur del departamento de Nariño, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté bajo la media nacional. El 4,8% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 70,82% de los 1594 estudiantes que se clasifican así, están correctamente clasificados y el 6,4% de los 17.474 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 5. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1 y es de la subregión Obando del departamento de Nariño, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 9,3% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 63% de los 3101 estudiantes que se clasifican así, están correctamente clasificados y el 12,2% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 6. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1, es de la subregión Centro del departamento de Nariño y su jornada de estudio es en la mañana, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 9,4% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 61,5% de los 3148 estudiantes que se clasifican así, están correctamente clasificados y el 12,1% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 7. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1, es de la subregión Centro del departamento de Nariño y su jornada de estudio es en la tarde, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 2,4% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 63,6% de los 818 estudiantes que se clasifican así, están correctamente clasificados y el 3,3% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 8. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1 y es de la subregión Los Abades del departamento de Nariño, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 1,4% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,7% de los 456 estudiantes que se clasifican así, están correctamente

clasificados y el 1,7% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 9. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1, es de la subregión Juanambú del departamento de Nariño y es de género masculino, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 1,3% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 68,2% de los 431 estudiantes que se clasifican así, están correctamente clasificados y el 1,8% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 10. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1 y es de la subregión Río Mayo del departamento de Nariño, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 3,3% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 67,1% de los 1088 estudiantes que se clasifican así, están correctamente clasificados y el 4,6% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 11. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1 y es de la subregión Occidente del departamento de Nariño, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 2,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 71,2% de los 690 estudiantes que se clasifican así, están correctamente clasificados y el 3,1% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 12. Si el estudiante es menor que 18 años, su familia es de nivel sisben 1 y es de la subregión La Sabana del departamento de Nariño, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 61,6% de los 654 estudiantes que se clasifican así, están correctamente

clasificados y el 2,5% de los 15.964 estudiantes que están sobre la media (ver tabla 29), cumplen este patrón.

Regla 13. Si el estudiante es menor que 18 años y su familia es de nivel sisben 2, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 5,8% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 67,7% de los 1941 estudiantes que se clasifican así, están correctamente clasificados y el 8,2% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 14. Si el estudiante es menor que 18 años y su familia no está clasificada en el sisben, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté sobre la media nacional. El 9,3% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 77,4% de los 3110 estudiantes que se clasifican así, están correctamente clasificados y el 15,1% de los 15.964 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 15. Si el estudiante es mayor que 22 años, entonces su desempeño académico en la prueba de matemáticas del Saber 11º es posible que esté bajo la media nacional. El 4,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 93,5% de los 1361 estudiantes que se clasifican así, están correctamente clasificados y el 2,3% de los 17.474 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

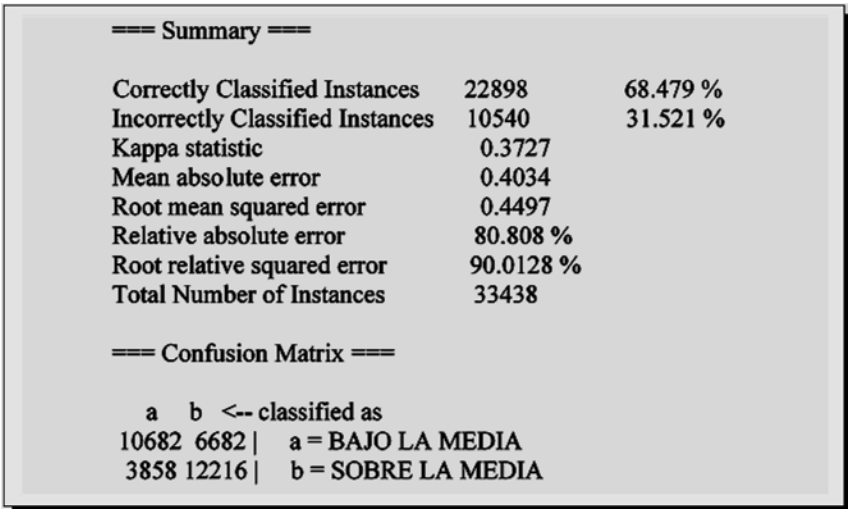
4.1.4 Factores asociados al desempeño académico en ciencias naturales en las pruebas Saber 11º

Se escogió como clase el puntaje en ciencias naturales de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores "*por encima de la media nacional*" y "*por debajo de la media nacional*", siendo la media nacional 52 sobre 100.

Con el fin de obtener diferentes modelos de árboles por competencia y reglas de clasificación generalizadas hasta reglas más detalladas, se establecieron 2 porcentajes de prepoda del árbol para el factor *M* igual a 1% y 2% del total de registros del repositorio de

datos, y 2 porcentajes para el factor confianza *C* igual a 25% y 50%. Se construyeron los diferentes modelos de árboles combinando estos factores. Se escogió el árbol construido con los parámetros $M = 33C$ (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 1% y una confianza del 60%. En la Figura 5 se muestra la precisión del árbol y su matriz de confusión. El árbol construido con los parámetros $M = 33C$ y $C = 25\%$ se muestra en la Figura 6.

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos *sb11_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje en ciencias naturales (*punt_c_naturales_cuali*) como clase, se puede observar que este clasifica correctamente a 22898 instancias, que corresponde a un porcentaje de precisión del 68,5% y 10.540 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 31,5% (ver Figura 5).



```
=== Summary ===

Correctly Classified Instances       22898               68.479 %
Incorrectly Classified Instances     10540               31.521 %
Kappa statistic                          0.3727
Mean absolute error                      0.4034
Root mean squared error                  0.4497
Relative absolute error                 80.808 %
Root relative squared error             90.0128 %
Total Number of Instances            33438

=== Confusion Matrix ===

    a     b   <-- classified as
 10682  6682 |   a = BAJO LA MEDIA
  3858 12216 |   b = SOBRE LA MEDIA
```

Figura 5. *Precisión y matriz de confusión del árbol de Ciencias Naturales.*

Teniendo en cuenta la matriz de confusión (Figura 5), del total de 33.438 estudiantes evaluados, el modelo clasifica a 18.898 estudiantes con desempeño académico sobre la media, correspondiente a un 57% del total de estudiantes y a 14.540 estudiantes con un desempeño académico bajo la media, que corresponde al 43%. Del 57% de estudiantes que están sobre la media, el modelo clasifica correctamente a 12.216 estudiantes, que corresponde a un porcentaje de 65% (casos correctos) y clasifica incorrectamente a 6.682 estudiantes, que corresponde a un porcentaje de 35% (falsos casos). Del 43% de estudiantes que están bajo la media, el modelo clasifica correctamente a 10.682 estudiantes, que corresponde a un porcentaje de 73% (casos correctos) e incorrectamente a 3.858 estudiantes, correspondiente al 27% (falsos casos).

Con respecto al total de casos correctos que están sobre la media según la tabla 29, en el puntaje de ciencias naturales (16.074), la exactitud del modelo es del 76%, esto significa que la especificidad del modelo es de 0,76. De igual manera, con respecto al total de casos correctos que están bajo la media (17.324), la exactitud del modelo es del 61,5%, es decir la sensibilidad del modelo es de 0,62.

```
J48 pruned tree
------------------

estu_edad_intervalo = Entre 18 y 22 años
|   subregion = PACIFICO SUR: BAJO LA MEDIA (2333.0/146.0)
|   subregion = LA CORDILLERA: BAJO LA MEDIA (404.0/124.0)
|   subregion = CENTRO
|   |   estu_genero = F: BAJO LA MEDIA (1545.28/590.9)
|   subregion = LOS ABADES: BAJO LA MEDIA (368.0/107.0)
|   subregion = TELEMBI: BAJO LA MEDIA (544.0/31.0)
|   subregion = JUANAMBU: BAJO LA MEDIA (722.0/287.0)
|   subregion = SANQUIANGA: BAJO LA MEDIA (636.0/31.0)
estu_edad_intervalo = Menor que 18 años
|   subregion = GUAMBUYACO: SOBRE LA MEDIA (489.0/182.0)
|   subregion = PACIFICO SUR: BAJO LA MEDIA (1660.0/436.0)
|   subregion = OBANDO: SOBRE LA MEDIA (4212.0/1504.0)
|   subregion = CENTRO
|   |   cole_jornada = Mañana: SOBRE LA MEDIA (6282.0/1941.0)
|   |   cole_jornada = Tarde: SOBRE LA MEDIA (1431.0/512.0)
|   subregion = JUANAMBU: SOBRE LA MEDIA (1076.0/426.0)
|   subregion = RIO MAYO: SOBRE LA MEDIA (1172.0/421.0)
|   subregion = OCCIDENTE: SOBRE LA MEDIA (724.0/245.0)
|   subregion = LA SABANA: SOBRE LA MEDIA (717.0/254.0)
estu_edad_intervalo = Mayor que 22 años: BAJO LA MEDIA (1361.0/70.0)

Number of Leaves  :     53

Size of the tree :     61
```

Figura 6. *Árbol textual de la prueba Ciencias Naturales*

Se escogieron los patrones más representativos del mejor árbol obtenido (ver Figura 6), teniendo en cuenta un mínimo soporte del 1% y una confianza mínima de 60%, tanto los que se ubican por encima de la media, como aquellos que se sitúan por debajo de ella. Los patrones más importantes se observan en la Figura 6.

Regla 1. Si el estudiante tiene entre 18 y 22 años y es de la subregión Pacífica Sur del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté bajo la media nacional. El 7% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 93,7% de los 2333 estudiantes que se clasifican así, están correctamente clasificados y el 12,6% de los 17.364 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 2. Si el estudiante tiene entre 18 y 22 años y es de la subregión La Cordillera del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté bajo la media nacional. El 1,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 69,3% de los 404 estudiantes que se clasifican así, están correctamente clasificados y el 1,6% de los 17.364 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 3. Si el estudiante tiene entre 18 y 22 años, es de la subregión Centro del departamento de Nariño y es de género femenino, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté bajo la media nacional. El 4,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 61,8% de los 1545 estudiantes que se clasifican así, están correctamente clasificados y el 5,5% de los 17.364 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 4. Si el estudiante tiene entre 18 y 22 años y es de la subregión Telembí del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté bajo la media nacional. El 1,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º

se clasifican de esta manera. El 94,3% de los 544 estudiantes que se clasifican así, están correctamente clasificados y el 3% de los 17.364 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 5. Si el estudiante tiene entre 18 y 22 años y es de la subregión Juanambú del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté bajo la media nacional. El 2,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,2% de los 722 estudiantes que se clasifican así, están correctamente clasificados y el 2,5% de los 17.364 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 6. Si el estudiante tiene entre 18 y 22 años y es de la subregión Sanquianga del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté bajo la media nacional. El 1,9% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 95,1% de los 636 estudiantes que se clasifican así, están correctamente clasificados y el 3,5% de los 17.364 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 7. Si el estudiante es menor que 18 años y es de la subregión Guambuyaco del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté sobre la media nacional. El 1,5% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 62,8% de los 489 estudiantes que se clasifican así, están correctamente clasificados y el 1,9% de los 16.074 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 8. Si el estudiante es menor que 18 años y es de la subregión Pacífico Sur del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté bajo la media nacional. El 5% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 73,7% de los 1660 estudiantes

que se clasifican así, están correctamente clasificados y el 7% de los 17.364 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 9. Si el estudiante es menor que 18 años y es de la subregión Obando del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté sobre la media nacional. El 12,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 64,3% de los 4212 estudiantes que se clasifican así, están correctamente clasificados y el 16,8% de los 16.074 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 10. Si el estudiante es menor que 18 años, es de la subregión Centro del departamento de Nariño y su jornada de estudio es en la mañana, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté sobre la media nacional. El 18,8% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 69,1% de los 6.282 estudiantes que se clasifican así, están correctamente clasificados y el 27% de los 16.074 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 11. Si el estudiante es menor que 18 años, es de la subregión Centro del departamento de Nariño y su jornada de estudio es en la tarde, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté sobre la media nacional. El 4,3% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 64,2% de los 1.431 estudiantes que se clasifican así, están correctamente clasificados y el 5,7% de los 16.074 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 12. Si el estudiante es menor que 18 años y es de la subregión Juanambú del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté sobre la media nacional. El 3,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,4% de los 1.076 estudiantes que se clasifican así, están correctamente clasificados y el 4% de

los 16.074 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 13. Si el estudiante es menor que 18 años y es de la subregión Río Mayo del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté sobre la media nacional. El 3,5% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 64,1% de los 1.172 estudiantes que se clasifican así, están correctamente clasificados y el 4,7% de los 16.074 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 14. Si el estudiante es menor que 18 años y es de la subregión Occidente del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté sobre la media nacional. El 2,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 66,2% de los 724 estudiantes que se clasifican así, están correctamente clasificados y el 3% de los 16.074 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 15. Si el estudiante es menor que 18 años y es de la subregión La Sabana del departamento de Nariño, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté sobre la media nacional. El 2,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 64,6% de los 717 estudiantes que se clasifican así, están correctamente clasificados y el 2,9% de los 16.074 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 16. Si el estudiante es mayor que 22 años, entonces su desempeño académico en la prueba de ciencias naturales del Saber 11º es posible que esté bajo la media nacional. El 4,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 94,9% de los 1361 estudiantes que se clasifican así, están correctamente clasificados y el 7,4% de los 17.364 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

4.1.5 Factores asociados al desempeño académico en inglés en las pruebas Saber 11º

Se escogió como clase el puntaje en inglés de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores "por encima de la media nacional "y "por debajo de la media nacional", siendo la media nacional 52 sobre 100.

Con el fin de obtener diferentes modelos de árboles por competencia y reglas de clasificación generalizadas hasta reglas más detalladas, se establecieron 2 porcentajes de prepoda del árbol para el factor M igual a 1% y 2% del total de registros del repositorio de datos, y 2 porcentajes para el factor confianza C igual a 25% y 50%. Se construyeron los diferentes modelos de árboles combinando estos factores. Se escogió el árbol construido con los parámetros $M = 33C$ (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 1% y una confianza del 60%. En la Figura 7 se muestra la precisión del árbol y su matriz de confusión. El árbol construido con los parámetros $M = 33C$ (1%) y $C = 25\%$, se muestra en la Figura 8.

```
=== Summary ===

Correctly Classified Instances       22589               67.5549 %
Incorrectly Classified Instances     10849               32.4451 %
Kappa statistic                          0.2802
Mean absolute error                      0.4127
Root mean squared error                  0.455
Relative absolute error                 86.9516 %
Root relative squared error             93.3965 %
Total Number of Instances            33438

=== Confusion Matrix ===

   a    b   <-- classified as
 5754  7199 |    a = SOBRE LA MEDIA
 3650 16835 |    b = BAJO LA MEDIA
```

Figura 7. *Precisión y matriz de confusión del árbol de inglés.*

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos *sb*11*_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje en inglés (*punt_ingles_cuali*) como clase, se puede observar que este clasifica correctamente a 22589 instancias, que corresponde a un porcentaje de precisión del 67,5% y 10.849 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 32,5% (ver Figura 7).

Teniendo en cuenta la matriz de confusión (Figura 7), del total de 33.438 estudiantes evaluados, el modelo clasifica a 9.404 estudiantes con desempeño académico sobre la media, correspondiente a un 28% del total de estudiantes y a 24.034 estudiantes con un desempeño académico bajo la media, que corresponde al 72%. Del 28% de estudiantes que están sobre la media, el modelo clasifica correctamente a 5.754 estudiantes, que corresponde a un porcentaje de 61% (casos correctos) y clasifica incorrectamente a 3.650 estudiantes, que corresponde a un porcentaje de 39% (falsos casos). Del 72% de estudiantes que están bajo la media, el modelo clasifica correctamente a 16.835 estudiantes, que corresponde a un porcentaje de 70% (casos correctos) e incorrectamente a 7.199 estudiantes, correspondiente al 30% (falsos casos).

Con respecto al total de casos correctos que están sobre la media según la Tabla 29, en el puntaje de inglés (12.953), la exactitud del modelo es del 44,4%, esto significa que la sensibilidad del modelo es de 0,44. De igual manera, con respecto al total de casos correctos que están bajo la media (20.485), la exactitud del modelo es del 82,2%, es decir la especificidad del modelo es de 0,82.

Se escogieron los patrones más representativos del mejor árbol obtenido (ver Figura 8), teniendo en cuenta un mínimo soporte del 1% y una confianza mínima de 60%, tanto los que se ubican por encima de la media, como aquellos que se sitúan por debajo de ella. Entre los patrones más importantes están:

```
J48 pruned tree
------------------

estu_edad_intervalo = Entre 18 y 22 años: BAJO LA MEDIA (12108.0/3087.0)
estu_edad_intervalo = Menor que 18 años
|  fami_estrato = BAJO
|  |  eco_condicion_tic = MALA
|  |  |  subregion = GUAMBUYACO: BAJO LA MEDIA (472.0/188.0)
|  |  |  subregion = PACIFICO SUR: BAJO LA MEDIA (1256.0/210.0)
|  |  |  subregion = LA CORDILLERA: BAJO LA MEDIA (524.0/197.0)
|  |  |  subregion = CENTRO
|  |  |  |  fami_educa_madre = Secundaria incompleta: BAJO LA MEDIA (470.0/183.0)
|  |  |  |  fami_educa_madre = Primaria completa: BAJO LA MEDIA (860.0/343.0)
|  |  |  subregion = LOS ABADES: BAJO LA MEDIA (464.0/165.0)
|  |  eco_condicion_tic = REGULAR
|  |  |  subregion = PACIFICO SUR: BAJO LA MEDIA (393.0/116.0)
|  |  |  subregion = CENTRO: SOBRE LA MEDIA (2578.0/1018.0)
|  fami_estrato = MEDIO
|  |  fami_ingreso_fmiliar_mensual = Entre 1 y menos de 2 SM: SOBRE LA MEDIA (500.0/165.0)
|  |  fami_ingreso_fmiliar_mensual = Entre 2 y menos de 3 SM: SOBRE LA MEDIA (594.0/137.0)
|  |  fami_ingreso_fmiliar_mensual = Entre 3 y menos de 5 SM: SOBRE LA MEDIA (473.0/74.0)
estu_edad_intervalo = Mayor que 22 años: BAJO LA MEDIA (1361.0/100.0)

Number of Leaves  :  56

Size of the tree :  66
```

Figura 8. *Árbol textual de la prueba inglés.*

Regla 1. Si la edad del estudiante está entre 18 y 22 años, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 36,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 74,5% de los 12.108 estudiantes que se clasifican así, están correctamente clasificados y el 44% de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 2. Si el estudiante es menor que 18 años, de estrato bajo, con una condición TIC mala y de la subregión Guambuyaco del departamento de Nariño, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 1,4% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,2% de los 472 estudiantes que se clasifican así, están correctamente clasificados y el 1,4% de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 3. Si el estudiante es menor que 18 años, de estrato bajo, con una condición TIC mala y de la subregión Pacífico Sur del departamento de Nariño, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 3,8% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 83,3% de los 1256 estudiantes que se clasifican así, están correctamente clasificados y el 5,1% de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 4. Si el estudiante es menor que 18 años, de estrato bajo, con una condición TIC mala y de la subregión La Cordillera del departamento de Nariño, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 1,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 62,4% de los 524 estudiantes que se clasifican así, están correctamente clasificados y el 1,6% de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 5. Si el estudiante es menor que 18 años, de estrato bajo, con una condición TIC mala, de la subregión Centro del departamento de Nariño y la educación de la madre es secundaria incompleta, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 1,4% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 61,1% de los 470 estudiantes que se clasifican así, están correctamente clasificados y el 1,4% de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 6. Si el estudiante es menor que 18 años, de estrato bajo, con una condición TIC mala, de la subregión Centro del departamento de Nariño y la educación de la madre es primaria completa, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 2,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,1% de los 860 estudiantes que se clasifican así, están correctamente clasificados y el 2,5%

de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 7. Si el estudiante es menor que 18 años, de estrato bajo, con una condición TIC mala, y es de la subregión Los Abades del departamento de Nariño, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 1,4% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 64,4% de los 464 estudiantes que se clasifican así, están correctamente clasificados y el 1,5% de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 8. Si el estudiante es menor que 18 años, de estrato bajo, con una condición TIC regular, y es de la subregión Pacífico Sur del departamento de Nariño, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 1,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 70,5% de los 393 estudiantes que se clasifican así, están correctamente clasificados y el 1,4% de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 9. Si el estudiante es menor que 18 años, de estrato bajo, con una condición TIC regular, y es de la subregión Centro del departamento de Nariño, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté sobre la media nacional. El 7,7% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,5% de los 393 estudiantes que se clasifican así, están correctamente clasificados y el 12% de los 12.953 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 10. Si el estudiante es menor que 18 años, de estrato medio y los ingresos familiares están entre 1 y menos de 2 salarios mínimos mensuales vigentes, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté sobre la media nacional. El 1,5% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 67% de los 500 estudiantes que se clasifican así, están correcta-

mente clasificados y el 2,6% de los 12.953 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 11. Si el estudiante es menor que 18 años, de estrato medio y los ingresos familiares están entre 2 y menos de 3 salarios mínimos mensuales vigentes, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté sobre la media nacional. El 1,8% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 76,9% de los 594 estudiantes que se clasifican así, están correctamente clasificados y el 3,5% de los 12.953 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

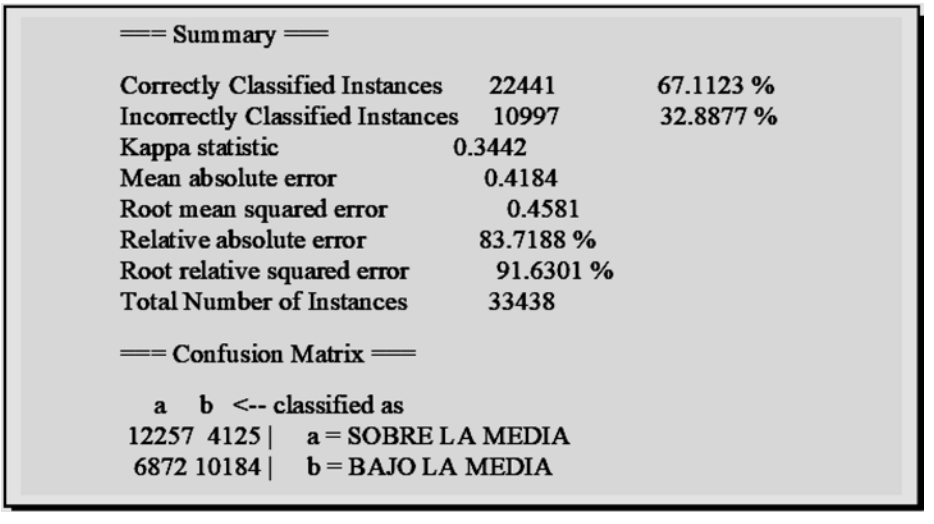
Regla 12. Si el estudiante es menor que 18 años, de estrato medio y los ingresos familiares están entre 3 y menos de 5 salarios mínimos mensuales vigentes, entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté sobre la media nacional. El 1,4% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 84,4% de los 473 estudiantes que se clasifican así, están correctamente clasificados y el 3,1% de los 12.953 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 13. Si el estudiante es mayor que 18 años entonces su desempeño académico en la prueba de inglés del Saber 11º es posible que esté bajo la media nacional. El 4,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 92,7% de los 1361 estudiantes que se clasifican así, están correctamente clasificados y el 6,2% de los 20.485 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

4.1.6 Factores asociados al desempeño académico en competencias ciudadanas en las pruebas Saber 11º

Se escogió como clase el puntaje en sociales y ciudadanas de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores “por encima de la media nacional “y “por debajo de la media nacional”, siendo la media nacional 51 sobre 100.

Con el fin de obtener diferentes modelos de árboles por competencia y reglas de clasificación generalizadas hasta reglas más detalladas, se establecieron 2 porcentajes de prepoda del árbol para el factor M igual a 1% y 2% del total de registros del repositorio de datos, y 2 porcentajes para el factor confianza C igual a 25% y 50%. Se construyeron los diferentes modelos de árboles combinando estos factores. Se escogió el árbol construido con los parámetros M=330 (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 1% y una confianza del 60%. En la Figura 9 se muestra la precisión del árbol y su matriz de confusión. El árbol construido con los parámetros M=330 y C=25% se muestra en la Figura 10.



```
=== Summary ===

Correctly Classified Instances       22441               67.1123 %
Incorrectly Classified Instances     10997               32.8877 %
Kappa statistic                          0.3442
Mean absolute error                      0.4184
Root mean squared error                  0.4581
Relative absolute error                 83.7188 %
Root relative squared error             91.6301 %
Total Number of Instances            33438

=== Confusion Matrix ===

    a     b   <-- classified as
 12257  4125 |    a = SOBRE LA MEDIA
  6872 10184 |    b = BAJO LA MEDIA
```

Figura 9. *Precisión y matriz de confusión del árbol de competencias ciudadanas.*

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos *sb*11*_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje

en sociales y ciudadanas (*punt_sociales_ciudadanas_cuali*) como clase, se puede observar que este clasifica correctamente a 22.441 instancias, que corresponde a un porcentaje de precisión del 67,1% y 10.997 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 32,9% (ver Figura 9).

Teniendo en cuenta la matriz de confusión (Figura 9), del total de 33.438 estudiantes evaluados, el modelo clasifica a 19.129 estudiantes con desempeño académico sobre la media, correspondiente a un 57% del total de estudiantes y a 14.309 estudiantes con un desempeño académico bajo la media, que corresponde al 43%. Del 57% de estudiantes que están sobre la media, el modelo clasifica correctamente a 12.257 estudiantes, que corresponde a un porcentaje de 64% (casos correctos) y clasifica incorrectamente a 6.872 estudiantes, que corresponde a un porcentaje de 36% (falsos casos). Del 72% de estudiantes que están bajo la media, el modelo clasifica correctamente a 10.184 estudiantes, que corresponde a un porcentaje de 71% (casos correctos) e incorrectamente a 4.125 estudiantes, correspondiente al 29% (falsos casos).

```
J48 pruned tree
------------------

estu_edad_intervalo = Entre 18 y 22 años
|   subregion = PACIFICO SUR: BAJO LA MEDIA (2333.0/269.0)
|   subregion = OBANDO
|   |   fami_educa_madre = Primaria completa: BAJO LA MEDIA (572.0/227.0)
|   subregion = LA CORDILLERA: BAJO LA MEDIA (404.0/134.0)
|   subregion = CENTRO
|   |   fami_nivel_sisben = No esta clasificada: SOBRE LA MEDIA (397.0/146.0)
|   subregion = LOS ABADES: BAJO LA MEDIA (368.0/145.0)
|   subregion = TELEMBI: BAJO LA MEDIA (544.0/54.0)
|   subregion = JUANAMBU: BAJO LA MEDIA (722.0/281.0)
|   subregion = SANQUIANGA: BAJO LA MEDIA (636.0/65.0)
|   subregion = OCCIDENTE: BAJO LA MEDIA (412.0/159.0)
estu_edad_intervalo = Menor que 18 años
|   subregion = PACIFICO SUR: BAJO LA MEDIA (1660.0/527.0)
|   subregion = OBANDO: SOBRE LA MEDIA (4212.0/1584.0)
|   subregion = CENTRO
|   |   cole_jornada = Mañana: SOBRE LA MEDIA (6282.0/1932.0)
|   |   cole_jornada = Tarde: SOBRE LA MEDIA (1431.0/479.0)
|   subregion = LOS ABADES: SOBRE LA MEDIA (490.0/169.0)
|   subregion = RIO MAYO: SOBRE LA MEDIA (1172.0/412.0)
|   subregion = OCCIDENTE: SOBRE LA MEDIA (724.0/255.0)
|   subregion = LA SABANA: SOBRE LA MEDIA (717.0/279.0)
estu_edad_intervalo = Mayor que 22 años: BAJO LA MEDIA (1361.0/142.0)

Number of Leaves  :  46

Size of the tree :  52
```

Figura 10. *Árbol textual de la prueba competencias ciudadanas.*

Con respecto al total de casos correctos que están sobre la media según la Tabla 29, en el puntaje de competencias ciudadanas (16.382), la exactitud del modelo es del 74,8%, esto significa que la sensibilidad del modelo es de 0,75. De igual manera, con respecto al total de casos correctos que están bajo la media (17.056), la exactitud del modelo es del 59,7%, es decir la especificidad del modelo es de 0,60.

Se escogieron los patrones más representativos del mejor árbol obtenido (ver Figura 10), teniendo en cuenta un mínimo soporte del 1% y una confianza mínima de 60%, tanto los que se ubican por encima de la media, como aquellos que se sitúan por debajo de ella. Entre los patrones más importantes están:

Regla 1. Si la edad del estudiante está entre 18 y 22 años y es de la subregión Pacífico Sur del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 7% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 88,5% de los 2.333 estudiantes que se clasifican así, están correctamente clasificados y el 12,1% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 2. Si la edad del estudiante está entre 18 y 22 años, es de la subregión Obando del departamento de Nariño y la educación de la madre es de primaria completa, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 1,7% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,3% de los 572 estudiantes que se clasifican así, están correctamente clasificados y el 2% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 3. Si la edad del estudiante está entre 18 y 22 años y es de la subregión La Cordillera del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 1,2% del total de estudiantes de Nariño que presentaron las

pruebas Saber 11º se clasifican de esta manera. El 66,8% de los 404 estudiantes que se clasifican así, están correctamente clasificados y el 1,6% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 4. Si la edad del estudiante está entre 18 y 22 años, es de la subregión Centro del departamento de Nariño y no está clasificado en el sisben, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté sobre la media nacional. El 1,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 63,2% de los 397 estudiantes que se clasifican así, están correctamente clasificados y el 1,5% de los 16.382 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 5. Si la edad del estudiante está entre 18 y 22 años y es de la subregión Los Abades del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 1,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 60,6% de los 368 estudiantes que se clasifican así, están correctamente clasificados y el 1,3% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 6. Si la edad del estudiante está entre 18 y 22 años y es de la subregión Telembí del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 1,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 90,1% de los 544 estudiantes que se clasifican así, están correctamente clasificados y el 2,9% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 7. Si la edad del estudiante está entre 18 y 22 años y es de la subregión Juanambú del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 2,2% del total de estudiantes de Nariño que presentaron las pruebas

Saber 11º se clasifican de esta manera. El 61,1% de los 722 estudiantes que se clasifican así, están correctamente clasificados y el 2,6% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 8. Si la edad del estudiante está entre 18 y 22 años y es de la subregión Sanquianga del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 1,9% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 89,8% de los 636 estudiantes que se clasifican así, están correctamente clasificados y el 3,3% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 9. Si la edad del estudiante está entre 18 y 22 años y es de la subregión Occidente del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 1,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 61,4% de los 412 estudiantes que se clasifican así, están correctamente clasificados y el 1,5% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 10. Si el estudiante es menor que 18 años y es de la subregión Pacífico Sur del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 5% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 68,3% de los 1.660 estudiantes que se clasifican así, están correctamente clasificados y el 6,6% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

Regla 11. Si el estudiante es menor que 18 años y es de la subregión Obando del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté sobre la media nacional. El 12,6% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º

se clasifican de esta manera. El 62,4% de los 4.212 estudiantes que se clasifican así, están correctamente clasificados y el 16% de los 16.382 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 12. Si el estudiante es menor que 18 años, es de la subregión Centro del departamento de Nariño y la jornada de estudio es en la mañana, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté sobre la media nacional. El 18,8% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 69,2% de los 6.282 estudiantes que se clasifican así, están correctamente clasificados y el 26,6% de los 16.382 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 13. Si el estudiante es menor que 18 años, es de la subregión Centro del departamento de Nariño y la jornada de estudio es en la tarde, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté sobre la media nacional. El 4,3% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 66,5% de los 1.431 estudiantes que se clasifican así, están correctamente clasificados y el 5,8% de los 16.382 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 14. Si el estudiante es menor que 18 años y es de la subregión Los Abades del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté sobre la media nacional. El 1,5% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 65,5% de los 490 estudiantes que se clasifican así, están correctamente clasificados y el 2% de los 16.382 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 15. Si el estudiante es menor que 18 años y es de la subregión Río Mayo del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté sobre la media nacional. El 3,5% del total de estudiantes de Nariño que presentaron las pruebas

Saber 11º se clasifican de esta manera. El 64,8% de los 1.172 estudiantes que se clasifican así, están correctamente clasificados y el 4,6% de los 16.382 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 16. Si el estudiante es menor que 18 años y es de la subregión Occidente del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté sobre la media nacional. El 2,2% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 64,8% de los 724 estudiantes que se clasifican así, están correctamente clasificados y el 2,9% de los 16.382 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 17. Si el estudiante es menor que 18 años y es de la subregión La Sabana del departamento de Nariño, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté sobre la media nacional. El 2,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 61,1% de los 717 estudiantes que se clasifican así, están correctamente clasificados y el 2,7% de los 16.382 estudiantes que están sobre la media (ver Tabla 29), cumplen este patrón.

Regla 18. Si el estudiante es mayor que 22 años, entonces su desempeño académico en la prueba de competencias ciudadanas del Saber 11º es posible que esté bajo la media nacional. El 4,1% del total de estudiantes de Nariño que presentaron las pruebas Saber 11º se clasifican de esta manera. El 89,6% de los 1361 estudiantes que se clasifican así, están correctamente clasificados y el 7,1% de los 17.056 estudiantes que están bajo la media (ver Tabla 29), cumplen este patrón.

4.2 DESCUBRIMIENTO DE PATRONES DESCRIPTIVOS ASOCIADOS A LAS PRUEBAS SABER 11º

Se seleccionó la tarea de agrupamiento o clustering particional con el algoritmo K-means, como la técnica descriptiva de aprendizaje no supervisado más adecuada para descubrir grupos

de estudiantes con desempeños similares en las pruebas Saber 11º, por ser uno de los métodos más utilizados y populares de agrupamiento (Hernández, Ramírez & Ferri, 2005). En la tarea de *clustering* se trata de encontrar grupos similares entre un conjunto de datos basado en el concepto de distancia (Han, Kamber & Pei, 2012). Los *clusters* tienen una alta homogeneidad interna (dentro del *cluster*) y una alta heterogeneidad externa (entre *cluster*). Por homogeneidad se entiende que los registros en un cluster o segmento están próximos unos a otros, donde la proximidad se expresa por medio de una medida, dependiendo de la distancia de los registros al centro del segmento. Por heterogeneidad se entiende que los registros en diferentes segmentos no son similares de acuerdo a una medida de similaridad (Pérez & Santín, 2006).

El algoritmo K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia euclidiana. K-means necesita como dato de entrada el número de grupos en los que se va a segmentar la población. A partir de este número *k* de clusters, el algoritmo coloca primero *k* puntos aleatorios (centroides). Luego asigna a cualquiera de esos puntos todas las muestras con las distancias más pequeñas. A continuación, el punto se desplaza a la media de las muestras más cercanas. Esto generará una nueva asignación de muestras, ya que algunas muestras están ahora más cerca de otro centroide. Este proceso se repite de forma iterativa y los grupos se van ajustando hasta que la asignación no cambia más moviendo los puntos. Este resultado final representa el ajuste que maximiza la distancia entre los distintos grupos y minimiza la distancia intragrupo.

Se utilizó la herramienta *WEKA* ver 3.9.4 (Witten et al, 2011) (Hall et al., 2011) con el algoritmo *K-means*, en el cual se configura el número de grupos (*NumClusters*) a formar y la semilla (*seed*), que se utiliza en la generación de un número aleatorio, el cual es usado para hacer la asignación inicial de instancias a los grupos. Como se conoce el atributo clase de los diferentes conjuntos de datos de las pruebas Saber 11º, que son los puntajes obtenidos

por los estudiantes en cada una de las pruebas Saber 11º, las cuales fueron discretizados en los valores "*por encima de la media nacional*" , y "*por debajo de la media nacional*", para evaluar los resultados del agrupamiento, se utilizó la opción de evaluación por clases (*Clases to clusters evaluation*), con el fin de que el número de clústeres sea el mismo de los valores del atributo clase, que en este caso son 2. Como K-means es sensible a los valores iniciales de asignación de ejemplos a los grupos, se tomaron 5 valores de semilla: 2, 5, 8, 10, 12 con el fin de escoger el mejor resultado en cada prueba. Con el mejor resultado se interpretaron las características de cada centroide de los clústeres que se formaron y de acuerdo a estas se obtuvieron patrones descriptivos por cada prueba del Saber 11º. También, a través de gráficos en tres dimensiones (*X*, *Y*, *color*) que permite la herramienta WEKA la visualización de los clústeres, se analizaron las características de estos, tomando como parámetro en el eje *X*, el número de clústeres, en el eje *Y*, el número de instancia dado por WEKA y en la dimensión color, la distribución por clústeres por los diferentes atributos, como se muestra en la Figura 11.

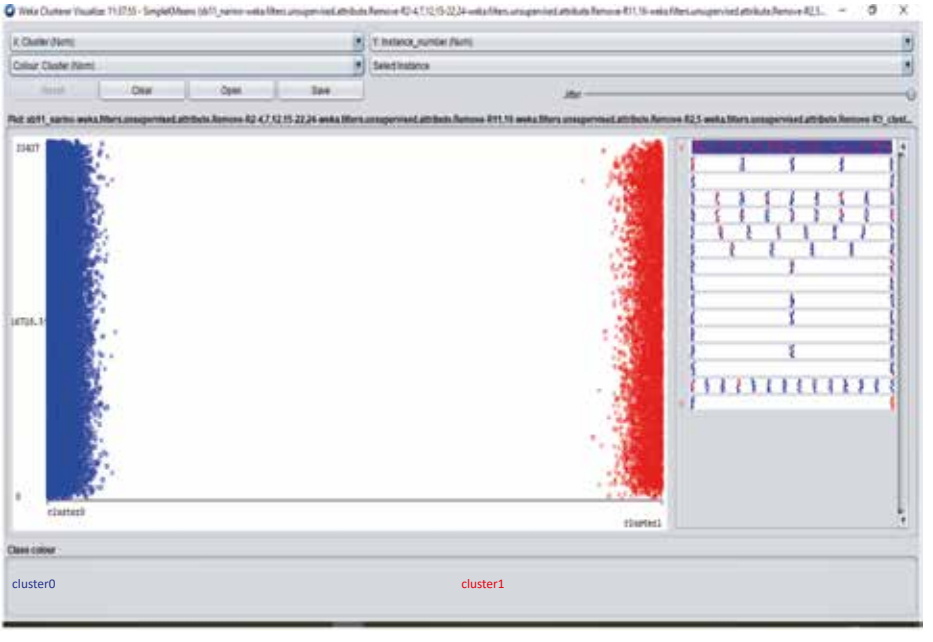


Figura 11. *Visualización de la distribución de los clústeres.*

4.2.1 Factores descriptivos asociados al desempeño académico al Puntaje Global en las pruebas Saber 11º

Teniendo en cuenta los parámetros de evaluación anteriores, se procedió a construir los diversos clústeres tomando como clase el puntaje global de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores "*por encima de la media nacional*" y "*por debajo de la media nacional*". Se generaron varios modelos variando la semilla (*sed*) del valor inicial del centroide con los valores 2, 5, 8, 10 y 12. El mejor resultado se obtuvo con el parámetro semilla 8. La exactitud del modelo obtenido y la descripción de los clústeres producidos con la herramienta WEKA se muestran en Figura 12 y Figura 13, respectivamente.

Analizando los resultados obtenidos con el algoritmo K-means con el conjunto de datos *sb*11*_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron la prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje global (*puntaje_global_cuali*) como clase, se puede observar que el modelo agrupa correctamente a 19.979 estudiantes, que corresponde a un porcentaje de precisión del 60% y 13.459 estudiantes incorrectamente agrupados, correspondiente a un porcentaje del 40% (ver Figura 12).

```
=== Model and evaluation on training set ===

Clustered Instances

0     26753 ( 80%)
1      6685 ( 20%)

Class attribute: punt_global_media
Classes to Clusters:

    0     1  <-- assigned to cluster
11766  4992 | DEBAJO MEDIA
14987  1693 | SOBRE MEDIA

Cluster 0 <-- SOBRE MEDIA
Cluster 1 <-- DEBAJO MEDIA

Incorrectly clustered instances :        13459.0    40.2506 %
```

Figura 12. *Exactitud del modelo con puntaje global de las pruebas Saber 11º.*

Final cluster centroids:

Attribute	Full Data (33438.0)	Cluster# 0 (26753.0)	1 (6685.0)
cole_jornada	Mañana	Mañana	Mañana
estu_genero	F	F	F
fami_educa_madre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_educa_padre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_ingreso_fmiliar_mensual	Menos de 1 SM	Menos de 1 SM	Entre 1 y menos de 2 SM
fami_nivel_sisben	Nivel 1	Nivel 1	Nivel 1
estu_edad_intervalo	Menor que 18 años	Menor que 18 años	Entre 18 y 22 años
tipo_cole	PUBLICO	PUBLICO	PUBLICO
eco_condicion_vive	SIN HACINAMIENTO	SIN HACINAMIENTO	SIN HACINAMIENTO
eco_condicion_vivienda	MALA	MALA	MALA
eco_condicion_tic	MALA	MALA	MALA
fami_estrato	BAJO	BAJO	BAJO
subregion	CENTRO	CENTRO	PACIFICO SUR

Figura 13. *Características de los clústeres del modelo con puntaje global de las pruebas Saber 11º.*

Según la Figura 13, en el clúster 0 el modelo clasifica a 26.753 estudiantes sobre la media y que corresponde al 80% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º y en el clúster 1 el modelo clasifica a 6.685 estudiantes bajo la media y que corresponde al 20% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º. Del 80% de los estudiantes que el modelo asignó sobre la media en el clúster 0, están correctamente asignados 14.987 estudiantes, que corresponde a un porcentaje de 56%. Del 20% de los estudiantes que el modelo asignó bajo la media en el clúster 1, están correctamente asignados 4.992 estudiantes, que corresponde a un porcentaje de 75%.

Con respecto al total de casos correctos que están sobre la media según la Tabla 29, en el puntaje global (16.680), la exactitud del modelo en el clúster 0 es del 90%. De igual manera, con respecto al total de casos correctos que están bajo la media (16.758), la exactitud del modelo en el clúster 1 es del 30%.

Interpretando las características de cada centroide de los clústeres que se muestran en la Figura 12, se pueden obtener los siguientes patrones descriptivos:

Clúster 0. El 80% de todos los estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media en el

puntaje global, estudian en la jornada de la mañana, son de género femenino, la educación de la madre y del padre es primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, menores que 18 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Centro del departamento de Nariño. El 56% de estos estudiantes están correctamente agrupados.

Clúster 1. El 20% de todos los estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media en el puntaje global, estudian en la jornada de la mañana, son de género femenino, la educación de la madre y del padre es primaria incompleta, los ingresos familiares están entre 1 y menos de 2 salarios mínimos (SMMLV), son de nivel sisben 1, su edad está entre 18 y 22 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión pacífico Sur del departamento de Nariño. El 75% de estos estudiantes están correctamente agrupados.

Las características que hacen diferentes a los dos clústeres son:

Clúster 0. Los estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media en el puntaje global, tienen unos ingresos familiares menores que 1 salario mínimo (SMMLV), menores que 18 años y de la subregión Centro del departamento de Nariño.

Clúster 1. Los estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media en el puntaje global, tienen unos ingresos familiares entre 1 y menos de 2 salarios mínimos (SMMLV), su edad está entre 18 y 22 años y de la subregión Pacífico Sur del departamento de Nariño.

Para el caso de los ingresos familiares, en la Figura 14 se puede observar la distribución de los estudiantes en cada clúster. Las instancias de color azul corresponden a los estudiantes de ingresos

menores que 1 SMMLV y las instancias de color rojo corresponden a los estudiantes de ingresos entre 1 y menos de 2 SMMLV. De acuerdo a la Figura 14, los primeros son mayoría en el clúster 0 (sobre la media) y los segundos en el clúster 1 (bajo la media).

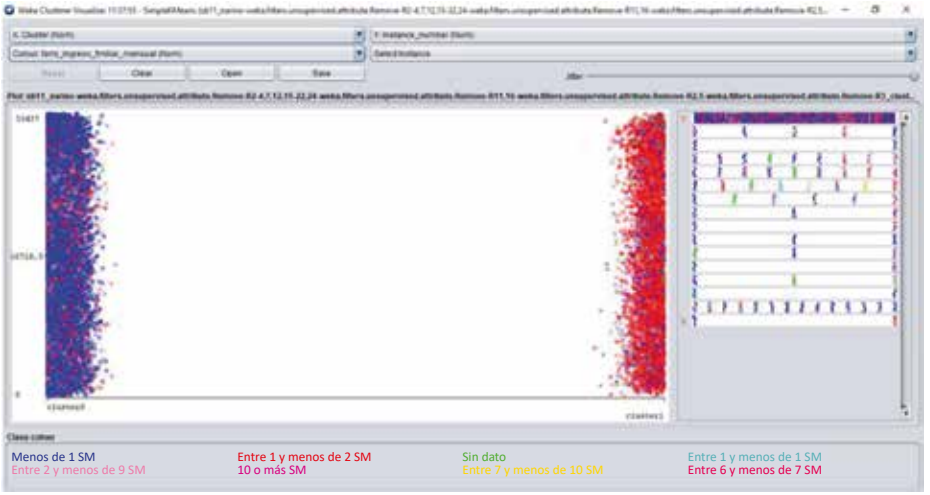


Figura 14. *Características de los clústeres según ingresos mensuales con respecto al puntaje global.*

4.2.2 Factores descriptivos asociados al desempeño académico en Lectura Crítica en las pruebas Saber 11º

Teniendo en cuenta los parámetros de evaluación descritos al inicio de este capítulo, se procedió a construir los diferentes clústeres tomando como clase el puntaje en la prueba de lectura crítica de cada estudiante obtenido en el Saber 11º, el cual fue discretizado en los valores "*por encima de la media nacional*" y "*por debajo de la media nacional*". Se generaron varios modelos variando la semilla (*sed*) del valor inicial del centroide con los valores 2,5,8,10 y 12. El mejor resultado se obtuvo con el parámetro semilla=8. La exactitud del modelo obtenido y la descripción de los clústeres obtenidos con la herramienta WEKA se muestran en Figura 15 y Figura 16, respectivamente.

Analizando los resultados obtenidos con el algoritmo K-means con el conjunto de datos *sb*11*_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, aca-

démicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo lectura crítica (*punt_lectura_critica_cuali*) como clase, se puede observar que el modelo agrupa correctamente a 20.508 estudiantes, que corresponde a un porcentaje de precisión del 61% y 12.930 estudiantes incorrectamente agrupados, correspondiente a un porcentaje del 39% (ver Figura 15).

```
=== Model and evaluation on training set ===

Clustered Instances

0     21480 ( 64%)
1     11958 ( 36%)

Class attribute: punt_lectura_critica_cuali
Classes to Clusters:

   0    1 <-- assigned to cluster
 8245  7273 | SOBRE LA MEDIA
13235  4685 | BAJO LA MEDIA

Cluster 0 <-- BAJO LA MEDIA
Cluster 1 <-- SOBRE LA MEDIA

Incorrectly clustered instances :         12930.0    38.6686 %
```

Figura 15. *Exactitud del modelo en lectura crítica de las pruebas Saber 11º.*

Final cluster centroids:

Attribute	Full Data (33438.0)	Cluster# 0 (26753.0)	1 (6685.0)
cole_jornada	Mañana	Mañana	Mañana
estu_genero	F	F	F
fami_educa_madre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_educa_padre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_ingreso_fmiliar_mensual	Menos de 1 SM	Menos de 1 SM	Entre 1 y menos de 2 SM
fami_nivel_sisben	Nivel 1	Nivel 1	Nivel 1
estu_edad_intervalo	Menor que 18 años	Menor que 18 años	Entre 18 y 22 años
tipo_cole	PUBLICO	PUBLICO	PUBLICO
eco_condicion_vive	SIN HACINAMIENTO	SIN HACINAMIENTO	SIN HACINAMIENTO
eco_condicion_vivienda	MALA	MALA	MALA
eco_condicion_tic	MALA	MALA	MALA
fami_estrato	BAJO	BAJO	BAJO
subregion	CENTRO	CENTRO	PACIFICO SUR

Figura 16. *Características de los clústeres del modelo con puntaje de lectura crítica pruebas Saber 11º.*

Según la Figura 15, en el clúster 0 el modelo clasifica a 21.480 estudiantes bajo la media y que corresponde al 64% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º y en el clúster 1 el modelo clasifica a 11958 estudiantes sobre la media y que corresponde al 36% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º. Del 64% de los estudiantes que el modelo asignó bajo la media en el clúster 0, están correctamente asignados 13.235 estudiantes, que corresponde a un porcentaje de 62%. Del 36% de los estudiantes que el modelo asignó sobre la media en el clúster 1, están correctamente asignados 7.273 estudiantes, que corresponde a un porcentaje de 61%.

Con respecto al total de casos correctos que están bajo la media según la Tabla 29. *Clases de Sb11_final_narino*, en lectura crítica (17.920), la exactitud del modelo en el clúster 0 es del 74%. De igual manera, con respecto al total de casos correctos que están sobre la media (15.518), la exactitud del modelo en el clúster 1 es del 47%.

Interpretando las características de cada centroide de los clústeres que se muestran en la Figura 16, se pueden obtener los siguientes patrones descriptivos:

Clúster 0. El 64% de todos los estudiantes del departamento de Nariño que presentaron la prueba de lectura crítica en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media estudian en la jornada de la mañana, son de género femenino, la educación de la madre y del padre es primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, su edad está entre 18 y 22 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Obando del departamento de Nariño. El 62% de estos estudiantes están correctamente agrupados.

Clúster 1. El 36% de todos los estudiantes del departamento de Nariño que presentaron la prueba de lectura crítica en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media estudian en la jornada de la mañana, son de género femeni-

no, la educación de la madre y del padre es secundaria completa, los ingresos familiares están entre 1 y menos de 2 salarios mínimos (SMMLV), son de nivel sisben 1, menores que 18 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición buena, sus condiciones TIC son malas, son de estrato bajo y de la subregión Centro del departamento de Nariño. El 61% de estos estudiantes están correctamente agrupados.

Las características que hacen diferentes a los dos clústeres son:

Clúster 0. Los estudiantes del departamento de Nariño que presentaron la prueba de lectura crítica en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media tienen unos ingresos familiares menores que 1 salario mínimo (SMMLV), su edad está entre 18 y 22 años, de una condición de vivienda mala y de la subregión Obando del departamento de Nariño.

Clúster 1. Los estudiantes del departamento de Nariño que presentaron la prueba de lectura crítica en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media tienen unos ingresos familiares entre 1 y menos de 2 salarios mínimos (SMMLV), menores que 18 años, de una condición vivienda buena y de la subregión Centro del departamento de Nariño.

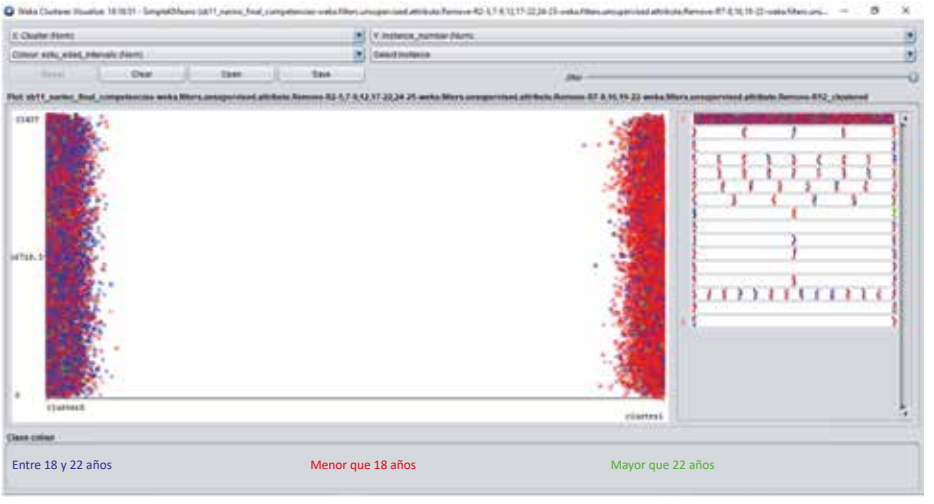


Figura 17. *Características de los clústeres según la edad de los estudiantes con respecto al puntaje en lectura crítica.*

Para el caso de la edad de los estudiantes, en la Figura 17 se puede observar la distribución de cada uno de ellos en cada clúster. Las instancias de color azul corresponden a los estudiantes cuya edad está entre 18 y 22 años y las instancias de color rojo corresponden a los estudiantes menores que 18 años. De acuerdo a la Figura 17, los primeros son mayoría en el clúster 0 (bajo la media) y los segundos en el clúster 1 (sobre la media).

4.2.3 Factores descriptivos asociados al desempeño académico en Matemáticas en las pruebas Saber 11º

Teniendo en cuenta los parámetros de evaluación descritos al inicio de este capítulo, se procedió a construir los diferentes clústeres tomando como clase el puntaje en matemáticas de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores "*por encima de la media nacional*", y "*por debajo de la media nacional*". Se generaron varios modelos variando la semilla (*sed*) del valor inicial del centroide con los valores 2, 5, 8, 10 y 12. El mejor resultado se obtuvo con el parámetro semilla=10. La exactitud del modelo obtenido y la descripción de los clústeres obtenidos con la herramienta Weka se muestran en Figura 18 y Figura 19, respectivamente.

Analizando los resultados obtenidos con el algoritmo K-means con el conjunto de datos *sb*11*_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje de matemáticas (*punt_matematicas_cuali*) como clase, se puede observar que el modelo agrupa correctamente a 21.288 estudiantes, que corresponde a un porcentaje de precisión del 64% y 12.150 estudiantes incorrectamente agrupados, correspondiente a un porcentaje del 36% (ver Figura 18).

Según la Figura 18, en el clúster 0 el modelo clasifica a 24.330 estudiantes sobre la media y que corresponde al 73% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º y en el clúster 1 el modelo clasifica a 9.108 estudiantes bajo la media y que corresponde al 27% del total de

estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º. Del 73% de los estudiantes que el modelo asignó sobre la media en el clúster 0, están correctamente asignados 14.072 estudiantes, que corresponde a un porcentaje de 58%. Del 27% de los estudiantes que el modelo asignó bajo la media en el clúster 1, están correctamente asignados 7.216 estudiantes, que corresponde a un porcentaje de 79%.

```
=== Model and evaluation on training set ===

Clustered Instances

0     24330 ( 73%)
1      9108 ( 27%)

Class attribute: punt_matematicas_cuali
Classes to Clusters:

    0    1  <-- assigned to cluster
14072  1892 | SOBRE LA MEDIA
10258  7216 | BAJO LA MEDIA

Cluster 0 <-- SOBRE LA MEDIA
Cluster 1 <-- BAJO LA MEDIA

Incorrectly clustered instances :        12150.0    36.3359 %
```

Figura 18. *Exactitud del modelo con puntaje en matemáticas de las pruebas Saber 11º.*

Final cluster centroids:

Attribute	Full Data (33438.0)	Cluster# 0 (24330.0)	1 (9108.0)
cole_jornada	Mañana	Mañana	Mañana
estu_genero	F	M	F
fami_educa_madre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_educa_padre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_ingreso_fmiliar_mensual	Menos de 1 SM	Menos de 1 SM	Menos de 1 SM
fami_nivel_sisben	Nivel 1	Nivel 1	Nivel 1
estu_edad_intervalo	Menor que 18 años	Menor que 18 años	Entre 18 y 22 años
tipo_cole	PUBLICO	PUBLICO	PUBLICO
eco_condicion_vive	SIN HACINAMIENTO	SIN HACINAMIENTO	SIN HACINAMIENTO
eco_condicion_vivienda	MALA	MALA	MALA
eco_condicion_tic	MALA	MALA	MALA
fami_estrato	BAJO	BAJO	BAJO
subregion	CENTRO	CENTRO	PACIFICO SUR

Figura 19. *Características de los clústeres del modelo con puntaje en matemáticas pruebas Saber 11º.*

Con respecto al total de casos correctos que están sobre la media según la Tabla 29, en el puntaje de matemáticas (15.964), la exactitud del modelo en el clúster 0 es del 88%. De igual manera, con respecto al total de casos correctos que están bajo la media (17.474), la exactitud del modelo en el clúster 1 es del 41%.

Interpretando las características de cada centroide de los clústeres que se muestran en la Figura 19, se pueden obtener los siguientes patrones descriptivos:

Clúster 0. El 73% de todos los estudiantes del departamento de Nariño que presentaron la prueba de matemáticas en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media estudian en la jornada de la mañana, son de género masculino, la educación de la madre y del padre es primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, menores que 18 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Centro del departamento de Nariño. El 58% de estos estudiantes están correctamente agrupados.

Clúster 1. El 27% de todos los estudiantes del departamento de Nariño que presentaron la prueba de matemáticas en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media estudian en la jornada de la mañana, son de género femenino, la educación de la madre y del padre es primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, su edad esta entre 18 y 22 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Pacífico Sur del departamento de Nariño. El 79% de estos estudiantes están correctamente agrupados.

Las características que hacen diferentes a los dos clústeres son:

Clúster 0. Los estudiantes del departamento de Nariño que presentaron la prueba de matemáticas en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media

son de sexo masculino, menores que 18 años y de la subregión Centro del departamento de Nariño.

Clúster 1. Los estudiantes del departamento de Nariño que presentaron la prueba de matemáticas en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media son de sexo femenino, su edad está entre 18 y 22 años y de la subregión Pacífico Sur del departamento de Nariño.

Para el caso del sexo de los estudiantes, en la Figura 20 se puede observar la distribución de cada uno de ellos en cada clúster. Las instancias de color azul corresponden a los estudiantes de sexo femenino y las instancias de color rojo corresponden a los estudiantes de sexo masculino. De acuerdo a la Figura 17, los primeros son mayoría en el clúster 1 (bajo la media) y los segundos en el clúster 0 (sobre la media).

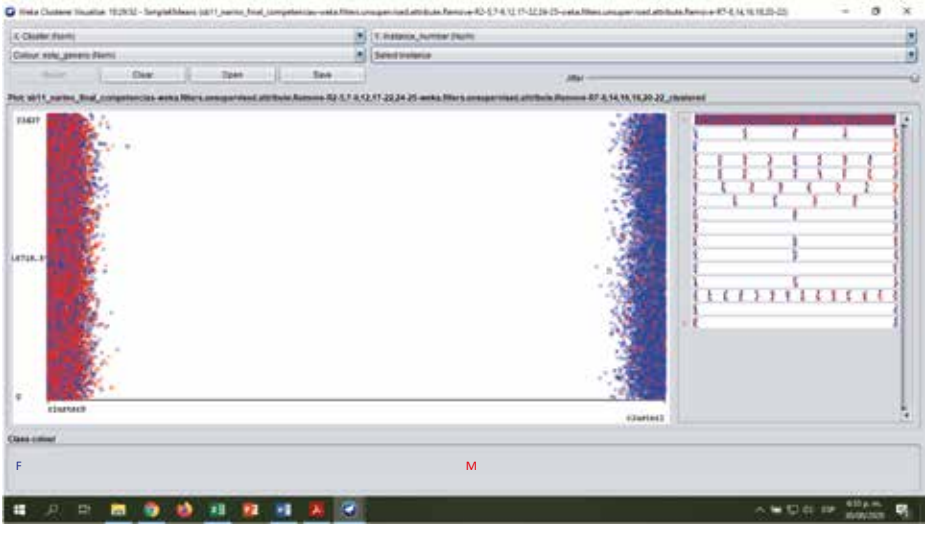


Figura 20. *Características de los clústeres según el sexo de los estudiantes con respecto al puntaje en matemáticas.*

4.2.4 Factores descriptivos asociados al desempeño académico en Ciencias Naturales en las pruebas Saber 11º

Teniendo en cuenta los parámetros de evaluación descritos al inicio de este capítulo, se procedió a construir los diferentes

clústeres tomando como clase el puntaje en ciencias naturales de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores "por encima de la media nacional", y "por debajo de la media nacional". Se generaron varios modelos variando la semilla (*sed*) del valor inicial del centroide con los valores 2, 5, 8, 10 y 12. El mejor resultado se obtuvo con el parámetro semilla=10. La exactitud del modelo obtenido y la descripción de los clústeres obtenidos con la herramienta WEKA se muestran en Figura 21 y Figura 22, respectivamente.

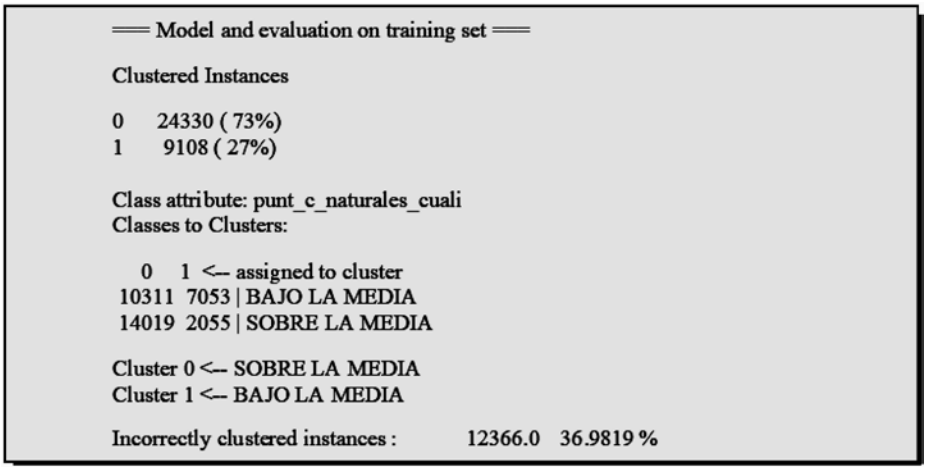


Figura 21. *Exactitud del modelo con puntaje en ciencias naturales de las pruebas Saber 11º.*

Analizando los resultados obtenidos con el algoritmo K-means con el conjunto de datos *sb*11*_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje de ciencias naturales (*punt_c_naturales_cuali*) como clase, se puede observar que el modelo agrupa correctamente a 21.072 estudiantes, que corresponde a un porcentaje de precisión del 63% y 12.366 estudiantes incorrectamente agrupados, correspondiente a un porcentaje del 37% (ver Figura 21).

Según la Figura 21, en el clúster 0 el modelo clasifica a 24.330 estudiantes sobre la media y que corresponde al 73% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º y en el clúster 1 el modelo clasifica a 9.108 estudiantes bajo la media y que corresponde al 27% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º. Del 73% de los estudiantes que el modelo asignó sobre la media en el clúster 0, están correctamente asignados 14.019 estudiantes, que corresponde a un porcentaje de 58%. Del 27% de los estudiantes que el modelo asignó bajo la media en el clúster 1, están correctamente asignados 7.053 estudiantes, que corresponde a un porcentaje de 77%.

Con respecto al total de casos correctos que están sobre la media según la Tabla 29, en el puntaje de ciencias naturales (16.074), la exactitud del modelo en el clúster 0 es del 87%. De igual manera, con respecto al total de casos correctos que están bajo la media (17.364), la exactitud del modelo en el clúster 1 es del 41%.

Final cluster centroids:

Attribute	Cluster# Full Data (33438.0)	0 (24330.0)	1 (9108.0)
cole_jornada	Mañana	Mañana	Mañana
estu_genero	F	M	F
fami_educa_madre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_educa_padre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_ingreso_fmiliar_mensual	Menos de 1 SM	Menos de 1 SM	Menos de 1 SM
fami_nivel_sisben	Nivel 1	Nivel 1	Nivel 1
estu_edad_intervalo	Menor que 18 años	Menor que 18 años	Entre 18 y 22 años
tipo_cole	PUBLICO	PUBLICO	PUBLICO
eco_condicion_vive	SIN HACINAMIENTO	SIN HACINAMIENTO	SIN HACINAMIENTO
eco_condicion_vivienda	MALA	MALA	MALA
eco_condicion_tic	MALA	MALA	MALA
fami_estrato	BAJO	BAJO	BAJO
subregion	CENTRO	CENTRO	PACIFICO SUR

Figura 22. *Características de los clústeres del modelo con puntaje en ciencias naturales pruebas Saber.*

Interpretando las características de cada centroide de los clústeres que se muestran en la Figura 22, se pueden obtener los siguientes patrones descriptivos:

Clúster 0. El 73% de todos los estudiantes del departamento de Nariño que presentaron la prueba de ciencias naturales en

el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media estudian en la jornada de la mañana, son de género masculino, la educación de la madre y del padre es primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, menores que 18 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Centro del departamento de Nariño. El 58% de estos estudiantes están correctamente agrupados.

Clúster 1. El 27% de todos los estudiantes del departamento de Nariño que presentaron la prueba de ciencias naturales en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media estudian en la jornada de la mañana, son de género femenino, la educación de la madre y del padre es primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, su edad esta entre 18 y 22 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Pacífico Sur del departamento de Nariño. El 77% de estos estudiantes están correctamente agrupados.

Las características que hacen diferentes a los dos clústeres son:

Clúster 0. Los estudiantes del departamento de Nariño que presentaron la prueba de ciencias naturales en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media son de sexo masculino, menores que 18 años y de la subregión Centro del departamento de Nariño.

Clúster 1. Los estudiantes del departamento de Nariño que presentaron la prueba de ciencias naturales en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media son de sexo femenino, su edad está entre 18 y 22 años y de la subregión Pacífico Sur del departamento de Nariño.

Para el caso del tipo de colegio a los que asisten los estudiantes, en la Figura 23 se puede observar la distribución de cada uno de ellos en cada clúster. Las instancias de color azul corresponden a

los estudiantes de colegios públicos y las instancias de color rojo corresponden a los estudiantes de colegios privados. De acuerdo a la Figura 23, la mayoría de estudiantes del departamento de Nariño tanto del clúster 0 (sobre la media) como del clúster 1 (bajo la media) asisten a colegios públicos.

Figura 23. *Características de los clústeres según el tipo de colegio de los estudiantes con respecto al puntaje en ciencias naturales.*

4.2.5 Factores descriptivos asociados al desempeño académico en Inglés en las pruebas Saber 11º

Teniendo en cuenta los parámetros de evaluación descritos al inicio de este capítulo, se procedió a construir los diferentes clústeres tomando como clase el puntaje en la prueba de inglés de cada estudiante obtenido en el Saber 11º, el cual fue discretizado en los valores "*por encima de la media nacional*" y "*por debajo de la media nacional*". Se generaron varios modelos variando la semilla (*sed*) del valor inicial del centroide con los valores 2, 5, 8, 10 y 12. El mejor resultado se obtuvo con el parámetro semilla=8. La exactitud del modelo obtenido y la descripción de los clústeres obtenidos con la herramienta Weka se muestran en Figura 24 y Figura 25, respectivamente.

Analizando los resultados obtenidos con el algoritmo K-means con el conjunto de datos *sb*11*_final_narino*, en el cual se alma-

cenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo de puntaje en inglés (*punt_ingles_cuali*) como clase, se puede observar que el modelo agrupa correctamente a 21.083 estudiantes, que corresponde a un porcentaje de precisión del 63% y 12.355 estudiantes incorrectamente agrupados, correspondiente a un porcentaje del 37% (ver Figura 24).

```
=== Model and evaluation on training set ===

Clustered Instances

0     21480 (64%)
1     11958 (36%)

Class attribute: punt_ingles_cuali
Classes to Clusters:

   0    1 <-- assigned to cluster
 6675  6278 | SOBRE LA MEDIA
14805  5680 | BAJO LA MEDIA

Cluster 0 <-- BAJO LA MEDIA
Cluster 1 <-- SOBRE LA MEDIA

Incorrectly clustered instances :         12355.0   36.949 %
```

Figura 24. *Exactitud del modelo en inglés de las pruebas Saber 11º.*

Según la Figura 24, en el clúster 0 el modelo clasifica a 21.480 estudiantes bajo la media y que corresponde al 64% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º y en el clúster 1 el modelo clasifica a 11958 estudiantes sobre la media y que corresponde al 36% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º. Del 64% de los estudiantes que el modelo asignó bajo la media en el clúster 0, están correctamente asignados 14.805 estudiantes, que corresponde a un porcentaje de 69%. Del 36% de los estudiantes que el modelo asignó sobre la media en el clúster 1, están correctamente asignados 6.278 estudiantes, que corresponde a un porcentaje de 53%.

Final cluster centroids:

Attribute	Full Data (33438.0)	Cluster# 0 (21480.0)	1 (11958.0)
cole_jornada	Mañana	Mañana	Mañana
estu_genero	F	F	F
fami_educa_madre	Primaria incompleta	Primaria incompleta	Secundaria completa
fami_educa_padre	Primaria incompleta	Primaria incompleta	Secundaria completa
fami_ingreso_fmiliar_mensual	Menos de 1 SM	Menos de 1 SM	Entre 1 y menos de 2 SM
fami_nivel_sisben	Nivel 1	Nivel 1	Nivel 1
estu_edad_intervalo	Menor que 18 años	Entre 18 y 22 años	Menor que 18 años
tipo_cole	PUBLICO	PUBLICO	PUBLICO
eco_condicion_vive	SIN HACINAMIENTO	SIN HACINAMIENTO	SIN HACINAMIENTO
eco_condicion_vivienda	MALA	MALA	BUENA
eco_condicion_tic	MALA	MALA	MALA
fami_estrato	BAJO	BAJO	BAJO
subregion	CENTRO	OBANDO	CENTRO

Figura 25. *Características de los clústeres del modelo con puntaje en inglés de las pruebas Saber 11º.*

Con respecto al total de casos correctos que están bajo la media según la Tabla 29, en inglés (20.485), la exactitud del modelo en el clúster 0 es del 72%. De igual manera, con respecto al total de casos correctos que están sobre la media (12.953), la exactitud del modelo en el clúster 1 es del 48%.

Interpretando las características de cada centroide de los clústeres que se muestran en la Figura 25, se pueden obtener los siguientes patrones descriptivos:

Clúster 0. El 64% de todos los estudiantes del departamento de Nariño que presentaron la prueba de inglés en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media estudian en la jornada de la mañana, son de género femenino, la educación de la madre y del padre son primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, su edad esta entre 18 y 22 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Obando del departamento de Nariño. El 69% de estos estudiantes están correctamente agrupados.

Clúster 1. El 36% de todos los estudiantes del departamento de Nariño que presentaron la prueba de inglés en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media estudian en la jornada de la mañana, son de género femenino, la educación de la madre y del padre son secundaria completa, los

ingresos familiares están entre 1 y menos de 2 salarios mínimos (SMMLV), son de nivel sisben 1, menores que 18 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición buena, sus condiciones TIC son malas, son de estrato bajo y de la subregión Centro del departamento de Nariño. El 53% de estos estudiantes están correctamente agrupados.

Las características que hacen diferentes a los dos clústeres son:

Clúster 0. Los estudiantes del departamento de Nariño que presentaron la prueba de inglés en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media tienen unos padres con educación primaria incompleta, ingresos familiares menores que 1 salario mínimo (SMMLV), su edad esta entre 18 y 22 años, de una condición de vivienda mala y de la subregión Obando del departamento de Nariño.

Clúster 1. Los estudiantes del departamento de Nariño que presentaron la prueba de inglés en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media tienen unos padres con educación secundaria completa, ingresos familiares entre 1 y menos de 2 salarios mínimos (SMMLV), menores que 18 años, de una condición vivienda buena y de la subregión Centro del departamento de Nariño.

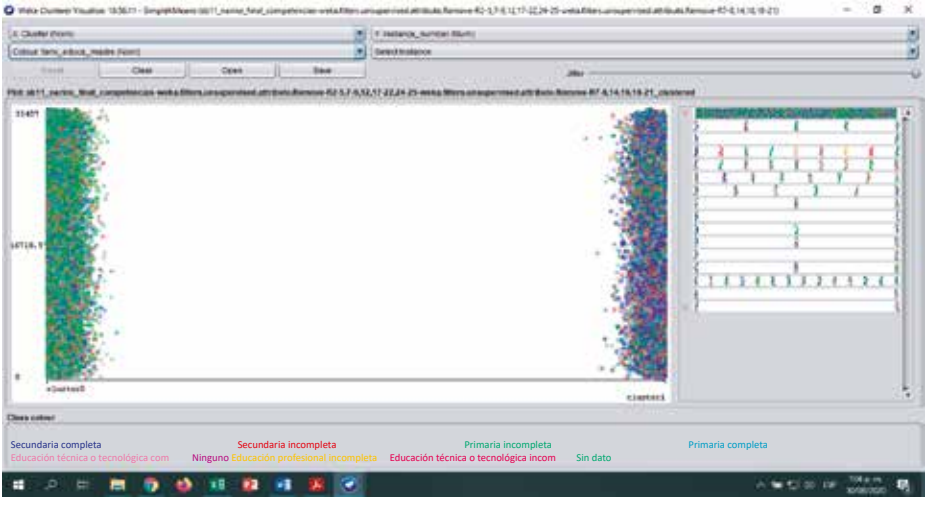


Figura 26. *Características de los clústeres según la educación de los padres con respecto al puntaje en inglés.*

Para el caso de la educación de los padres de los estudiantes, en la Figura 26 se puede observar la distribución de cada uno de ellos en cada clúster. Las instancias de color verde corresponden a los estudiantes cuyos padres tienen el nivel educativo de primaria incompleta y las instancias de color rojo corresponden a los estudiantes cuyos padres tienen el nivel educativo de secundaria completa. De acuerdo a la Figura 26, los primeros son mayoría en el clúster 0 (bajo la media) y los segundos en el clúster 1 (sobre la media).

4.2.6 Factores descriptivos asociados al desempeño académico en Competencias Ciudadanas en las pruebas Saber 11º

Teniendo en cuenta los parámetros de evaluación descritos al inicio de este capítulo, se procedió a construir los diferentes clústeres tomando como clase el puntaje en sociales y ciudadanas de cada estudiante obtenido en las pruebas Saber 11º, el cual fue discretizado en los valores "*por encima de la media nacional*" y "*por debajo de la media nacional*". Se generaron varios modelos variando la semilla (*sed*) del valor inicial del centroide con los valores 2,5,8,10 y 12. El mejor resultado se obtuvo con el parámetro semilla=10. La exactitud del modelo obtenido y la descripción de los clústeres obtenidos con la herramienta WEKA se muestran en Figura 27 y Figura 28, respectivamente.

Analizando los resultados obtenidos con el algoritmo K-means con el conjunto de datos *sb*11*_final_narino*, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 33.438 estudiantes nariñenses, quienes presentaron las prueba Saber 11º en los periodos 2015 y 2016 donde se escogió el atributo puntaje en sociales y ciudadanas (*punt_sociales_ciudadanas_cuali*) como clase, se puede observar que el modelo agrupa correctamente a 20.698 estudiantes, que corresponde a un porcentaje de precisión del 62% y 12.740 estudiantes incorrectamente agrupados, correspondiente a un porcentaje del 38% (ver Figura 27).

Según la Figura 27, en el clúster 0 el modelo clasifica a 24.330 estudiantes sobre la media y que corresponde al 73% del total de estudiantes del departamento de Nariño que presentaron las

pruebas Saber 11º y en el clúster 1 el modelo clasifica a 9.108 estudiantes bajo la media y que corresponde al 27% del total de estudiantes del departamento de Nariño que presentaron las pruebas Saber 11º. Del 73% de los estudiantes que el modelo asignó sobre la media en el clúster 0, están correctamente asignados 13.986 estudiantes, que corresponde a un porcentaje de 57%. Del 27% de los estudiantes que el modelo asignó bajo la media en el clúster 1, están correctamente asignados 6.712 estudiantes, que corresponde a un porcentaje de 74%.

```
=== Model and evaluation on training set ===

Clustered Instances

0      24330 ( 73%)
1       9108 ( 27%)


Class attribute: punt_sociales_ciudadanas_cuali
Classes to Clusters:

   0    1  <-- assigned to cluster
13986  2396 | SOBRE LA MEDIA
10344  6712 | BAJO LA MEDIA

Cluster 0 <-- SOBRE LA MEDIA
Cluster 1 <-- BAJO LA MEDIA

Incorrectly clustered instances :         12740.0    38.1004 %
```

Figura 27. *Exactitud del modelo con puntaje en sociales y ciudadanas de las pruebas Saber 11º.*

Final cluster centroids:

Attribute	Full Data (33438.0)	Cluster# 0 (24330.0)	1 (9108.0)
cole_jornada	Mañana	Mañana	Mañana
estu_genero	F	M	F
fami_educa_madre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_educa_padre	Primaria incompleta	Primaria incompleta	Primaria incompleta
fami_ingreso_fmiliar_mensual	Menos de 1 SM	Menos de 1 SM	Menos de 1 SM
fami_nivel_sisben	Nivel 1	Nivel 1	Nivel 1
estu_edad_intervalo	Menor que 18 años	Menor que 18 años	Entre 18 y 22 años
tipo_cole	PUBLICO	PUBLICO	PUBLICO
eco_condicion_vive	SIN HACINAMIENTO	SIN HACINAMIENTO	SIN HACINAMIENTO
eco_condicion_vivienda	MALA	MALA	MALA
eco_condicion_tic	MALA	MALA	MALA
fami_estrato	BAJO	BAJO	BAJO
subregion	CENTRO	CENTRO	PACIFICO SUR

Figura 28. *Características de los clústeres del modelo con puntaje en sociales y ciudadanas de las pruebas Saber 11º.*

Con respecto al total de casos correctos que están sobre la media según la Tabla 29, en el puntaje de sociales y ciudadanas (16.382), la exactitud del modelo en el clúster 0 es del 85%. De igual manera, con respecto al total de casos correctos que están bajo la media (17.056), la exactitud del modelo en el clúster 1 es del 39%.

Interpretando las características de cada centroide de los clústeres que se muestran en la Figura 28, se pueden obtener los siguientes patrones descriptivos:

Clúster 0. El 73% de todos los estudiantes del departamento de Nariño que presentaron la prueba de sociales y ciudadanas en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media estudian en la jornada de la mañana, son de género masculino, la educación de la madre y del padre son primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, menores que 18 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Centro del departamento de Nariño. El 57% de estos estudiantes están correctamente agrupados.

Clúster 1. El 27% de todos los estudiantes del departamento de Nariño que presentaron la prueba de matemáticas en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media estudian en la jornada de la mañana, son de género femenino, la educación de la madre y del padre son primaria incompleta, los ingresos familiares son menores que 1 salario mínimo (SMMLV), son de nivel sisben 1, su edad esta entre 18 y 22 años, son de colegio público, viven sin hacinamiento, en una vivienda de condición mala, sus condiciones TIC son malas, son de estrato bajo y de la subregión Pacífico Sur del departamento de Nariño. El 74% de estos estudiantes están correctamente agrupados.

Las características que hacen diferentes a los dos clústeres son:

Clúster 0. Los estudiantes del departamento de Nariño que presentaron la prueba de sociales y ciudadanas en el Saber 11º

entre los años 2015 y 2016 y que se agrupan en el clúster sobre la media son de sexo masculino, menores que 18 años y de la subregión Centro del departamento de Nariño.

Clúster 1. Los estudiantes del departamento de Nariño que presentaron la prueba de matemáticas en el Saber 11º entre los años 2015 y 2016 y que se agrupan en el clúster bajo la media son de sexo femenino, su edad esta entre 18 y 22 años y de la subregión Pacífico Sur del departamento de Nariño.

Para el caso de la jornada del colegio en la que asisten los estudiantes, en la Figura 29 se puede observar la distribución de cada uno de ellos en cada clúster. Las instancias de color azul corresponden a los estudiantes cuya jornada de estudios es en la mañana, las instancias de color rojo corresponden a los estudiantes de jornada de la tarde, las instancias de color verde corresponden a los estudiantes de jornada de la noche y las instancias de color amarillo a los estudiantes de jornada única. De acuerdo a la Figura 29, la mayoría de estudiantes del departamento de Nariño tanto del clúster 0 (sobre la media) como del clúster 1 (bajo la media) asisten a sus colegios en la jornada de la mañana.

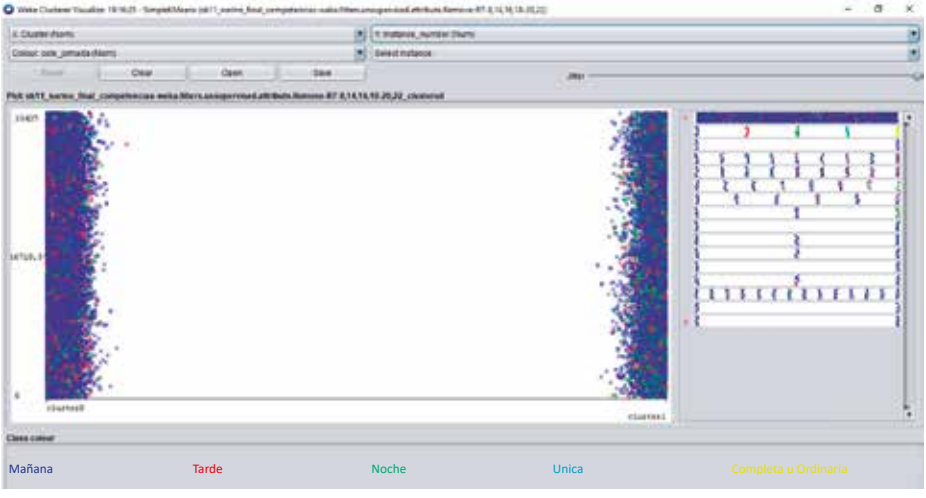


Figura 29. *Características de los clústeres según la jornada escolar con respecto al puntaje en sociales y ciudadanas.*

Capítulo V

DISCUSIÓN DE RESULTADOS Y CONCLUSIONES

5.1 DISCUSIÓN DE RESULTADOS DE LOS PATRONES PREDICTIVOS ENCONTRADOS

Para la discusión de resultados se realizó un resumen sobre los patrones que se destacan por el mayor porcentaje de estudiantes clasificados correctamente por el modelo, en relación con el total de estudiantes que se encuentran bajo la media o sobre la media, dependiendo de que el patrón se ubique por encima o por debajo de la media nacional en cada competencia que se analice.

5.1.1 Desempeño en Lectura Crítica

Durante el período 2015-2016, presentaron las pruebas Saber 11º un total de 33.438 estudiantes de Educación Secundaria del departamento de Nariño; entre quienes, según la edad, el 59,7% son menores de 18 años, el 36,2% tienen edades entre 18 y 22 años y apenas el 4,1% tiene más de 22 años (ver Tabla 30); por lo cual, los mayores de 18 años corresponden al 40,3%.

Tabla 30. ***Patrones destacados en lectura crítica por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Regla	Edad	Subregión	Bajo la media	Sobre la media	%
1	18-22	Pacífico Sur	X		11,7
14	<18	Obando		X	16,3
15	<18	Centro		X	33,8
16	<18	Juanambú		X	31,5

Fuente: elaboración propia

En los factores asociados al desempeño académico en lectura crítica y que forman parte de los patrones descubiertos están la edad de los estudiantes y la subregión donde se encuentra la institución educativa. En la tabla 30 se nota que los menores que 18 años tienen mejor desempeño académico que los que están entre 18 y 22 años. Según Isorna et al. (2017), esto se debe a lo que se conoce como el efecto de la edad relativa RAE (de inglés *Relative Age Efect*) que es la diferencia de edad entre los integrantes de un curso y sus consecuencias. En el campo educativo se requiere el agrupamiento por edad de los estudiantes, cuyo objetivo es el de asegurar un proceso de formación estándar a todos los estudiantes. Normalmente siempre existirán diferencias de edad y por lo tanto potencialmente de maduración y experiencia entre los estudiantes de un curso o grupo que incide muchas veces en los logros académicos. Diversos estudios han mostrado que las diferencias en edad relativa de los niños que comparten la misma clase en los periodos escolares suponen unas diferencias de desarrollo que podrían perjudicar la igualdad de condiciones y las posibilidades de éxito entre los sujetos de un mismo grupo de corte (Cobley, Abraham, & Baker, 2008). Tejedor (2003) señala que la edad es un factor explicativo del rendimiento académico, mencionando que dentro de un mismo curso aquellos alumnos que son más jóvenes obtienen un mejor promedio. Coincidiendo con estudios previos de Richardson (1994) y posteriores como el de Valli Jayanthi et al. (2014).

Por otra parte, la subregión es otro factor asociado al desempeño académico en lectura crítica. De acuerdo a la tabla 30, las subre-

giones de Obando, Centro y Juanambú tienen mejor desempeño académico que la subregión Pacífico Sur. Esto se puede explicar por el nivel de desarrollo de las subregiones y de la categoría de los colegios. Si se observa la Tabla 25 estas tres primeras subregiones están compuestas por municipios situados en la zona andina del departamento de Nariño y la última compuesta por municipios de la costa pacífica nariñense.

Según un estudio realizado por Viloria (2007) y publicado por el Banco de la República, la zona de la costa pacífica nariñense (Tumaco, Barbacoas, Francisco Pizarro entre otros) es la más pobre del departamento y donde se ubican colegios clasificados por el Icfes en su mayoría como categoría baja. En cambio, en la zona andina (entre ellos Pasto, Ipiales, Túquerres, La Unión entre otros) están los municipios que concentran la tercera parte de la economía del departamento, que es el agropecuario. Aquí se encuentran colegios clasificados en las categorías baja, media y alta.

Con respecto a la categoría de los colegios, en un estudio realizado por Willms y Somers (2001) determinaron que los colegios con mejores resultados académicos eran aquellos que se caracterizaban por un conjunto de ventajas académicas: maestros que estaban satisfechos con sus salarios y altos niveles de recursos escolares (una biblioteca grande, más materiales instructivos, profesores bien entrenados y una baja proporción de alumnos por maestro). Por otra parte, Piñeros y Rodríguez (1998), estimaron que una adecuada dotación de insumos para las escuelas secundarias (asociados con la infraestructura) tenía un efecto positivo sobre el rendimiento académico de los estudiantes. Iregui, Melo y Ramos (2007) también encontraron que los insumos (la existencia de laboratorios, biblioteca y canchas deportivas) tenían un impacto positivo y estadísticamente significativo en el logro académico de escolares colombianos

Estos resultados coinciden con el efecto localidad donde se realizó un ranking de los municipios del departamento de Nariño utilizando el estadístico d de Cohen, con los resultados en la prueba de lectura crítica y que se puede apreciar en la Tabla 7.

Finalmente, Caso y Hernández (2007), afirman que la comprensión lectora de un estudiante es una habilidad que se correlaciona con el rendimiento académico de este y que en este estudio se corrobora con la alta correlación que existe en las pruebas Saber 11º entre lectura crítica y el puntaje global de la prueba Saber 11º en la matriz de correlaciones de la tabla 3. Esto quiere decir que si en la lectura crítica su desempeño esta sobre la media, seguramente el desempeño en el puntaje general de la prueba Saber 11º estará sobre la media. De igual manera para el desempeño bajo la media.

5.1.2 Desempeño en Matemáticas

Según la Tabla 31, el 36,2% de la población estudiada tiene edad entre 18 y 22 años, de la cual, las mujeres son las que más obtienen puntajes por debajo de la media nacional; asimismo, en este grupo etario, están los hombres con unas condiciones malas de las TIC. Si se tiene en cuenta que, el 77,4% de estudiantes tienen malas condiciones en relación con las TIC, se podría afirmar que toda política educativa orientada al fortalecimiento del desempeño académico en el área de matemáticas, debería estar dirigida a toda la población, sin discriminación de género.

Tabla 31. ***Patrones destacados en matemáticas por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Regla	Edad	Subregión	Sexo	Sisben	TIC	Bajo	Sobre	%
1	18-22		F			X		27,3
2	18-22		M		Mala	X		18,8
5	<18	Obando		1			X	12,2
6	<18	Centro		1			X	12,1
13	<18			2			X	8,2
14	<18			No			X	15,1

Fuente: elaboración propia

Para los menores de 18 años, población estudiantil que constituye el 59,7% de los estudiantes objeto de esta investigación, los alumnos que más aportan a la suma de puntajes sobre la media nacional, son los que habitan en las regiones de Obando y Centro,

y también se suman los alumnos con Sisben categoría 2 y los no clasificados en Sisben. Por la estructura de los patrones de la tabla anterior, se puede afirmar que, las cantidades de estudiantes de las otras regiones no son destacables para cambiar el promedio sobre la media ni bajo la media. Vale anotar que, según la Tabla 13, las regiones Centro y Obando, están entre las regiones mejor clasificadas en el ranking de regiones según el desempeño en el área de matemáticas.

La edad del estudiante es un factor asociado al rendimiento académico de los estudiantes en la prueba de matemáticas en el Saber 11º; hombres y mujeres con edades entre 18 y 22 años, aportan significativamente al promedio por debajo de la media nacional, lo cual puede ser mejorado, teniendo en cuenta lo que indica Murillo (2013), en el sentido que los jóvenes, poseen un gran potencial de imaginación y talento creativo, que debe aprovecharse para el trabajo matemático en la Educación Básica y Media.

En el grupo de edades entre 18 y 22 años, que constituye el 59,7% de los estudiantes evaluados, el sexo es una variable asociada a los promedios por debajo de la media nacional, las mujeres constituyen el 27,3% y los hombres el 18,8% en este grupo de edades. Esto deja ver que, en este grupo de edades, los hombres tienen mejor desempeño académico que las mujeres en la evaluación de matemáticas en Saber 11º; un tanto contrario a los hallazgos de Gómez y Soares (2013), quienes expresan que, según el sexo, no hay diferencias significativas en el desempeño académico; que no se puede afirmar de modo definitivo que exista una relación directa entre el rendimiento académico y el sexo. Sin embargo, de acuerdo a los resultados obtenidos los hombres si tienen un ligero mejor desempeño académico que las mujeres en la prueba de matemáticas en el Saber 11º.

El índice de condición TIC de los estudiantes que mide la posibilidad que tienen de utilizar internet, computador, telefonía en su casa, es otro factor asociado al desempeño académico de los estudiantes que presentaron la prueba de matemáticas en el Saber 11º; específicamente, según los hallazgos, los hombres con

edades entre 18 y 22 años, en el área de matemáticas arrastran al promedio bajo la media nacional. En este contexto, las investigaciones realizadas por Botello y Guerrero (2014) quienes investigan el impacto de las tecnologías de la información y comunicación en el desempeño académico de los estudiantes de América Latina, tomando como fuente la prueba PISA del 2012, encontraron que, la tenencia de tecnologías y el uso de éstas en el aprendizaje escolar mediante actividades de contenido digital, inciden de manera positiva en el desempeño académico de los alumnos.

5.1.3 Desempeño en Ciencias Naturales

En la Tabla 32 se observa que, la población estudiantil que más aporta a la suma de puntajes bajo le media nacional en Ciencias Naturales, son los de la región Pacífico Sur cuya edad está entre 18 y 22 años, y también, los que tienen más de 22 años de todas las regiones. Por su parte, los que más aportan al número de estudiantes sobre la media nacional, son los menores de 18 años de las regiones de Obando y Centro, esto es coincidente con el buen desempeño de estas regiones en el área de matemáticas, tal como se evidencia en la Tabla 13, que presenta el ranking por regiones según el desempeño en esa área. De modo que, en general, para mejorar el desempeño en ciencias naturales, es necesario implementar estrategias para fortalecer el desempeño académico en esta área en la región Pacífico Sur.

Tabla 32. ***Patrones destacados en ciencias naturales por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Regla	Edad	Región	Bajo la media	Sobre la media	%
1	18-22	Pacífico Sur	X		12,6
8	<18	Pacífico Sur	X		7
9	<18	Obando		X	16,8
10	<18	Centro		X	27
16	>22		X		7,4

Fuente: elaboración propia

Los estudiantes con edades menores de 22 años que residen en la región Pacífico Sur, constituyen el 19,6% de la población cuyos puntajes están bajo la media nacional.

En general, los factores asociados al desempeño académico en ciencias naturales son similares a los obtenidos en la prueba de lectura crítica y por lo tanto su discusión es igual.

5.1.4 Desempeño en Inglés

Las reglas de los patrones 1 y 13 de la Tabla 33, indican que, los estudiantes que más contribuyen al promedio por debajo de la media nacional, son los alumnos con edades entre 18 y 22 años, y también los mayores de 22 años; pero entre ellos, sobresalen los alumnos con edad entre 18 y 22 años. Por su parte, los que más aportan al número de estudiantes cuyos puntajes están sobre la media nacional, son los menores de 22 años, de la región Centro, de estrato bajo y con regulares condiciones TIC. Este hecho sugiere que, una política orientada a mejorar el desempeño académico en inglés, tendría que estar dirigida a toda la población estudiantil, pero con predominio de la población con edades entre 18 y 22 años.

Tabla 33. ***Patrones destacados en inglés por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Regla	Edad	Estrato	Región	TIC	Bajo	Sobre	%
1	18-22				X		44
9	<18	Bajo	Centro	Regular		X	12
13	>22				X		6,2

Fuente: elaboración propia

Por otra parte, según la Tabla 8, 26 municipios de los 64 del departamento de Nariño, presentan diferencias significativas en el desempeño en la prueba en inglés, en relación con los mejores puntajes del departamento, lo cual, considerando el nivel socioeconómico del Departamento, hay coincidencia con lo indicado por Garbanzo (2007), Seibold (2000) y Montero, Villalobos y Valverde

(2004), en cuanto que existe asociación significativa entre el nivel socioeconómico del estudiante y el desempeño académico.

Es interesante la información de la regla 9, en el sentido que, a pesar de que los estudiantes pertenezcan a estrato bajo, se puede mejorar el nivel de desempeño en inglés fortaleciendo las condiciones TIC. Hecho que es coherente con lo que indica Botello y Guerrero (2014) que el uso de tecnologías TIC en los procesos de aprendizaje escolar incrementan el puntaje promedio en cada una de las áreas de estudio entre un 5% y un 6%.

5.1.5 Desempeño en sociales y competencias ciudadanas

Según las reglas 1 y 18 de la Tabla 34, los estudiantes que más aportan el promedio bajo la media nacional, son los de edad entre 18 y 22 años, de la región Pacífico Sur, jornada de la mañana; y también los mayores de 22 años de todas las regiones. Por su parte, los alumnos que más aportan a mantener un promedio sobre la media nacional, son los menores de 18 años, de las regiones de Obando y Centro, destacándose los de la región Centro. Es claro que, una política orientada a mejorar el desempeño en competencias ciudadanas, tendría que dirigirse a toda la población estudiantil.

Tabla 34. ***Patrones destacados en competencias ciudadanas por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Regla	Edad	Región	Jornada	Bajo	Sobre	%
1	18-22	Pacífico Sur	Mañana	X		12,1
11	<18	Obando			X	16
12	<18	Centro			X	26,6
18	>22			X		7,1

Fuente: elaboración propia

La jornada académica es un factor asociado al rendimiento académico de los estudiantes en la prueba de competencias ciudadanas en las pruebas Saber 11º; concretamente aparece visible para el caso de estudiantes entre 18 y 22 años de la región Pacífico Sur, donde los alumnos de la jornada de la mañana aportan de

manera importante al promedio bajo la media en competencias ciudadanas. En este contexto, el estudio de Chica, Galvis y Ramírez (2010) utilizando un Modelo Logit Ordenado Generalizado, encontraron que los estudiantes de jornada completa obtienen mejores puntajes en relación con los estudiantes de otras jornadas; de modo similar, Ridao y Gil (2010)], en el estudio "La jornada escolar y el rendimiento de los alumnos", encontraron que, se obtienen mejores calificaciones en las instituciones educativas con jornada completa en comparación con los de jornada continua.

5.2 DISCUSIÓN DE RESULTADOS DE PATRONES DESCRIPTIVOS

5.2.1 Desempeño en Lectura Crítica

La Tabla 35 indica que la población estudiantil que más aporta a la cantidad de estudiantes con puntajes bajo la media nacional en Lectura Crítica, son los que tienen edad entre 18 y 22 años, con ingresos menores a 1 smmlv, con malas condiciones de vivienda y que residen en la Región Obando; por su parte, los que más aportan al incremento de estudiantes con puntajes sobre la media nacional en Lectura Crítica, son los que tienen menos de 18 años, con ingresos entre 1 y dos smmlv, con buenas condiciones de vivienda y residen en la Región Centro.

Tabla 35. ***Patrones descriptivos destacados en lectura crítica por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Clúster	Edad	Ingresos (smmlv)	Cond. Vivienda	Región
Bajo la media	18-22	<1	Mala	Obando
Sobre la media	<18	Entre 1 y 2	Buena	Centro

Fuente: elaboración propia

5.2.2 Desempeño en Matemáticas

En la Tabla 36 se observa que, la población estudiantil que más aporta a la cantidad de estudiantes con puntajes sobre la media nacional en Matemáticas, son los hombres menores de 18 años

que viven en la Región Centro; por su parte, los que más aportan al incremento de estudiantes con puntajes bajo la media nacional en Matemáticas, son las mujeres cuya edad está entre 18 y 22 años que viven en la Región Pacífico Sur.

Tabla 36. ***Patrones descriptivos destacados en matemáticas por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Clúster	Edad	Género	Región
Sobre la media	<18	M	Centro
Bajo la media	18-22	F	Pacífico Sur

Fuente: elaboración propia

5.2.3 Desempeño en Ciencias Naturales

En Tabla 37 se evidencia que, la población estudiantil que más aporta a la cantidad de estudiantes con puntajes sobre la media nacional en Ciencias Naturales, son los hombres menores de 18 años que viven en la Región Centro; por su parte, los que más aportan al incremento de estudiantes con puntajes bajo la media nacional en Ciencias Naturales, son las mujeres cuya edad está entre 18 y 22 años que viven en la Región Pacífico Sur.

Tabla 37. ***Patrones descriptivos destacados en ciencias naturales por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Clúster	Edad	Género	Región
Sobre la media	<18	M	Centro
Bajo la media	18-22	F	Pacífico Sur

Fuente: elaboración propia

Es interesante observar que los clústeres de matemáticas y de Ciencias Naturales son iguales; en cierto modo, esto se debe a que, el desempeño en Ciencias Naturales está altamente relacionado con el desempeño en Matemáticas, tal como se indicó en la tabla 3.

5.2.4 Desempeño en Inglés

En la Tabla 38 se indica que la población estudiantil que más aporta a la cantidad de estudiantes con puntajes bajo la media nacional en inglés, son los de edades entre 18 y 22 años, con ingresos menores a 1 smmlv, sus padres tienen primaria incompleta, su vivienda tiene malas condiciones y residen en la Región Obando. Por su parte, la población estudiantil que más aporta a la cantidad de estudiantes con puntajes sobre la media nacional en inglés, son los menores de 18 años, con ingresos entre 1 y 2 smmlv, sus padres tienen secundaria completa, su vivienda tiene buenas condiciones y residen en la Región Centro del departamento de Nariño.

Tabla 38. ***Patrones descriptivos destacados en inglés por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Clúster	Edad	Ingresos (smmlv)	Niv. Educ. Padr.	Cond. Viv.	Región
Bajo la media	18-22	<1	P. Incompl.	Mala	Obando
Sobre la media	<18	Entre 1 y 2	Sec. Compl.	Buena	Centro

Fuente: elaboración propia

5.2.5 Desempeño en sociales y competencias ciudadanas

La Tabla 39 muestra que la población estudiantil que más aporta a la cantidad de estudiantes con puntajes sobre la media nacional en Competencias Ciudadanas, son los hombres menores de 18 años que viven en la Región Centro; por su parte, los que más aportan al incremento de estudiantes con puntajes bajo la media nacional en Competencias Ciudadanas, son las mujeres cuya edad está entre 18 y 22 años que viven en la Región Pacífico Sur.

Tabla 39. ***Patrones descriptivos destacados en sociales y competencias ciudadanas por su aporte al incremento de puntajes por encima o por debajo de la media nacional en esta área***

Clúster	Edad	Sexo	Región
Sobre la media	<18	M	Centro
Bajo la media	18-22	F	Pacífico Sur

Fuente: elaboración propia

De las tablas anteriores se tiene que los estudiantes que más aportan al incremento de puntajes bajo la media, son las mujeres de la región Pacífico Sur; y esto ocurre en las áreas de Matemáticas, Ciencias Naturales y Competencias Ciudadanas. Este resultado podría indicar que es necesario continuar fortaleciendo las políticas de inclusión educativa, tanto en lo relativo a las regiones como al caso del género.

5.3 CONCLUSIONES Y TRABAJOS FUTUROS

El objetivo de esta investigación fue "Descubrimiento de factores asociados al desempeño académico en las pruebas Saber 11º de los estudiantes de las instituciones educativas del departamento de Nariño con técnicas de minería de datos" , el cual se logró en su totalidad. Las fuentes de datos fueron las bases de datos del Icfes de los años 2015 y 2016, de las cuales se seleccionaron los datos socioeconómicos, académicos e institucionales de los estudiantes nariñenses. La metodología utilizada para cumplir este objetivo fue CRISP-DM, y la técnica de minería de datos aplicada para el descubrimiento de patrones de desempeño académico, fue clasificación, basada en árboles de decisión.

Los resultados obtenidos con el modelo de clasificación con árboles de decisión para descubrir factores asociados al desempeño académico de los estudiantes nariñenses que encontrándose finalizando el grado undécimo de educación media, presentaron las pruebas Saber 11º entre los años 2015 y 2016, indican que este es capaz de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos que se encuentran almacenados en las bases de datos del Icfes.

Considerando el buen desempeño académico en las pruebas Saber 11º como aquellos puntajes globales por encima de la media y un bajo desempeño en estas pruebas como aquellos puntajes globales por debajo de la media, es mayor el porcentaje de estudiantes nariñenses que tienen un desempeño académico bajo comparado con el porcentaje de estudiantes que tienen un buen desempeño.

Por otra parte, entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos asociados

al buen desempeño académico en las pruebas Saber 11º están: el estrato socioeconómico medio o alto, la jornada de estudio en la mañana o completa, el índice tic regular y la edad menor que 18 años.

De igual manera, entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos asociados a un bajo desempeño académico en las pruebas Saber 11º están: el estrato socioeconómico bajo, el índice tic bajo y el nivel Sisben 1.

Entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos tanto en la prueba de Lectura Crítica como el de Matemáticas, se destacan el estrato socioeconómico, la jornada de estudio, el índice tic, la edad y el sexo de los estudiantes como factores importantes asociados al buen o bajo desempeño académico de los estudiantes en estas pruebas.

Por otra parte, entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos asociados al buen desempeño académico en Ciencias Naturales de las pruebas Saber 11º están: el estrato socioeconómico medio o alto, la jornada de estudio en la mañana o completa y el índice tic regular. De igual manera, entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos asociados a un bajo desempeño académico en Ciencias Naturales de las pruebas Saber 11º están: el estrato socioeconómico bajo, el índice tic bajo y la jornada de la tarde.

Igualmente, entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos en la prueba de inglés, se destacan el estrato socioeconómico, la jornada de estudio, el índice tic y la edad como factores importantes asociados al buen o bajo desempeño académico de los estudiantes en esta prueba.

Se plantea como trabajos futuros complementar este estudio utilizando otras técnicas predictivas y algoritmos, con el fin de comparar los resultados obtenidos con árboles de decisión con

el algoritmo J48. Por otra parte, aplicar otras tares de minería de datos que permitan relacionar cuales atributos se presentan juntos asociados al desempeño académico en las pruebas Saber 11º y cómo se agrupan los individuos de acuerdo a su rendimiento en dichas pruebas.

Además, sería recomendable realizar estudios sobre la relación entre el rendimiento académico de los estudiantes en las pruebas Saber 11º, el desempeño académico en las Instituciones de Educación Superior en su formación profesional y las pruebas Saber Pro que presentan los estudiantes próximos a terminar una carrera profesional en Colombia.

REFERENCIAS

Ahmad, F., Ismail, N. & Aziz, A. (2015). The prediction of student's academic perfomance using classification data mining techniques. Applied Mathematical Sciences, 9 (129), 6415-6426.

Azevedo, A. & Santos, M. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. In: Proceedings of IADIS European Conference on Data Mining. Amsterdam, Netherlands, pp. 182-185. ISBN: 978-972-8924-63-8.

Badr, G., Algobail, A., Almutairi, H., y Almutery, M. (2016). Predicting students' performance in university courses: A case study and tool in KSU Mathematics Department. Symposium on Data Mining Applications. Procedia Computer Science, 82, pp. 80-89.

Barrientos, J., 2008. Calidad de la educación pública y logro académico en Medellín 2004-2006: Una aproximación por regresión intercuartil. Revista Lecturas de Economía, Núm. 68, pp. 121-144. ISSN: 0120-2596. Universidad de Antioquia. Medellín, Colombia.

Blanco, V., 2015. Análisis del Desempeño Académico del Examen de Estado para el Ingreso a la Educación Superior Aplicando Minería de Datos. Trabajo de investigación presentado como requisito para optar al título de Magister en Ingeniería Sistemas y Computación. Universidad Nacional de Colombia, Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial. Valledupar, Colombia.

Botello, H. & Guerrero, A. (2014). La influencia de las TIC en el desempeño académico de los estudiantes en América Latina: Evidencia de la prueba PISA. Memorias Virtual Educa. Lima, Perú. http://hdl.handle.net/20.500.12579/4050.

Calleja, A. (2010). Minería de Datos con Weka para la Predicción del Precio de Automóviles de Segunda Mano. Proyecto de fin de carrera para obtener el título en Ingeniería Informatica. Universidad Politécnica de Valencia. Disponible en: https://riunet.upv.es/bitstream/handle/10251/10097/PFC_DSIC-80_Agust%C3%ADnCalleja.pdf.

Caso, J. & Hernández, L. (2007). Variables que inciden en el rendimiento académico de adolescentes mexicanos. Revista Latinoamericana de Psicología. Vol. 39, No. 3, pp. 487-501.

Cobley, S., Abraham, C. & Baker, J. (2008). Relative age effects on physical education attainment and school sport representation. Physical Education & Sport Pedagogy, 13 (3), pp. 267-276.

Corsi, L., García, M., Archila, M. & Niño, J. (2012). Factores asociados a desempeños destacados y no destacados en las pruebas saber 11º (2009-2). Tesis de Maestría en Educación. Pontifica Universidad Javeriana. Bogotá D.C. (Colombia).

Correa, J. J. (2004). Determinantes del Rendimiento Educativo de los Estudiantes de Secundaria en Cali: Un análisis multinivel. En: Revista Sociedad y Economía. No. 6, pp. 81-105.

Costa, E., Fonseca, B., Almeida, M., Ferreira, F. & Rego, J. (2016). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 73, pp. 247-256.

Chapman, P., Clinton, J., Kerber, R, Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark), Daimler Chrysler AG (Germany), SPSS Inc. (USA), and OHRA Verzekeringen en Bank Groep B.V. (The Netherlands).

Chica, S., Galvis, D. y Ramírez, A. (2010). Determinantes del rendimiento académico en Colombia: pruebas Icfes Saber 11º. Revista Universidad EAFIT, Vol. 46, Núm. 160. ISSN: 0120-341X. Medellín, Colombia.

Fernández, H. (2005). Como interpretar la evaluación pruebas Saber. Subdirección de Estándares y Evaluación. Ministerio de Educación Nacional. Bogotá, Colombia.

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.

García, J. (2016). Comenzando con Weka: Filtrado y selección de subconjuntos de atributos basada en su relevancia descriptiva para

la clase. Technical report. Universidad de Malaga. Disponible en: https://www.researchgate.net/publication/308141950.

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

Gaviria, A. y Barrientos, J. (2001). Calidad de la educación y rendimiento académico en Bogotá. Revista Coyuntura Social, Núm. 24. ISSN: 0121-2532. Bogotá D.C., Colombia.

Gómez, J. (2014). Análisis de las competencias en matemáticas y lenguaje de los bachilleres colombianos. Trabajo de grado, Facultad de Ciencias Administrativas y Económicas Economía y Negocios Internacionales. Universidad ICESI. Cali, Colombia. Disponible en: https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/77946/1/gomez_analisis_competencias_2014.pdf.

Hamsa, H., Indiradevi, S. & Kizhakkethottam, J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. Procedia Technology, 25, pp. 326-332.

Han, J., Kamber, M. (2001). Data Mining: Concepts and Techniques. San Francisco, USA: Morgan Kaufmann Publishers; 2001. 550 p.

Hernández, O. (2015). Determinantes del Rendimiento Académico en la Educación Media de Cundinamarca. Trabajo de grado, Facultad de Economía, Escuela Colombiana de Ingeniería Julio Garavito. Bogotá D.C., Colombia. Disponible en: http://repositorio.escuelaing.edu.co/bitstream/001/349/1/AA-Econom%C3%ADa-1077087614.pdf.

Hernández, J., Ramírez, M. y Ferri, C. (2005). Introducción a la Minería de Datos. Madrid (España): Pearson Prentice Hall. ISBN: 84-205-4091-9.

Icfes (2014). Alineación del examen SABER 11º Lineamientos generales 2014 - 2 Sistema Nacional de Evaluación Estandarizada de la Educación, Instituto Colombiano para la Evaluación de la Educación (Icfes). ISBN: 978-958-11-0630-1. Bogotá, Colombia.

Icfes (2016). Sistema Nacional de Evaluación Estandarizada de la Educación: Lineamientos generales para la presentación del examen de Estado Saber 11º. Instituto Colombiano para la Evaluación de la Educación (Icfes). ISBN: 978-958-11-0680-6. Bogotá, Colombia.

Icfes (2018). Guía de orientación Saber 11º. 2ª Edición. Instituto Colombiano para la Evaluación de la Educación, Bogotá D.C., Colombia: Icfes, 2018, pp. 19-31.

Iregui, A., Melo, L. & Ramos, J. (2007). Análisis de eficiencia de la educación en Colombia. Revista de Economía del Rosario, 10 (1): 21-41.

Isorna, M., Rial, A., Felpeto, M. & Rodríguez, L. (2017). Evaluación del Impacto del Efecto Relativo de la Edad en el Rendimiento Escolar, Bullying, Autoestima, Diagnóstico de TDAH y Consumo de Tabaco en el Paso de Educación Primaria a Secundaria. Revista Iberoamericana de Diagnóstico y Evaluación Psicológica. RIDEP, Nº 44. Vol. 2, pp. 92-104.

Khobragade, L. & Mahadik, P. (2015). Students academic failure prediction using data mining. International Journal of Advanced Research in Computer and Communication Engineering, 4 (11), 290-298.

López, H. y González, M. (2016). Factores Asociados a la Competencia de Inglés como Lengua Extranjera de los Estudiantes Universitarios Colombianos. Tesis M.S., Fac. Educación, Univ. La Sabana, Chía, Colombia.

MEN (2006). Estándares Básicos de Competencias en Lenguaje, Matemáticas, Ciencias y Ciudadanas: Guía sobre lo que los estudiantes deben saber y saber hacer con lo que aprenden. Ministerio de Educación Nacional. ISBN: 958-691-290-6. Bogotá, Colombia.

Montero, E. y 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.

Morales, M. (2019). Factores Determinantes en el Rendimiento de los estudiantes de la región Pacífico-colombiana en las Pruebas Saber 11º. Trabajo de grado Facultad de Ciencias Administrativas y Económicas, Universidad Icesi. Cali (Colombia). Disponible en: https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/84736/1/TG02508.pdf

Muñoz, A. (2017). Determinantes de las diferencias en los resultados de las pruebas académicas de estado. El caso de la educación media oficial en Bogotá. Trabajo de grado, Escuela Colombiana de Ingeniería Julio Garavito, Facultad de Economía. Bogotá D.C., Colombia. Disponible en: https://repositorio.escuelaing.

edu.co/bitstream/001/700/1/Mu%C3%B1oz%20Sanchez%2C%20Andr%C3%A9s%20Mauricio-2017.pdf

Murillo, E. (2013). Factores que inciden en el Rendimiento Académico en el área de Matemáticas de los estudiantes de noveno grado en los Centros de Educación Básica de la ciudad de Tela, Atlántida Tesis de Maestría. Universidad Pedagógica Nacional Francisco Morazán. San Pedro Sula, Honduras.

Osmanbegović, E., y Suljić, M. (2012). Data mining approach for predicting student performance. Economic Review-Journal of Economics and Business, 10(1), 2-12.

Pérez, C. & Santín, D. (2007). Data Mining: Soluciones con Enterprise Miner. México: Editorial Alfaomega. ISBN: 84-7897-695-7.

Piñeros, L. & Rodríguez. (1998). Los Insumos Escolares en la Educación Secundaria y su Efecto Sobre el Rendimiento Académico de los Estudiantes: Un estudio en Colombia. Banco Mundial, LCSHD, Paper, Series No. 36.

Posada, J. & Mendoza, F. (2014) Determinantes del logro académico de los estudiantes de grado 11 en el periodo 2008-2010. Una perspectiva de género y región, Bogotá, Colombia: Estudios sobre calidad de la educación en Colombia, Icfes, Ministerio de Educación Nacional, 2014.

Ridao,I. y Gil, J. (2002). La jornada escolar y el rendimiento de los alumnos. En revista de Educación, No. 327, pp. 141-156.

Richardson, J.T.E. (1994). Mature Students in Higher Education: Academic Performance and Intellectual Ability. Higher Education: 28(3), pp. 373-386.

Rodríguez, F., Benavides, H. & Riascos, A. (2019). Predicción del desempeño académico usando técnicas de aprendizaje de máquinas. Disponible en: https://www.icfes.gov.co/documents/20143/234129/Prediccion+desempeno+academico+usando+un+enfoque+de+mineria+de+datos.pdf.

Sánchez, A., & Otero, A. (2012). Educación y Reproducción de la desigualdad en Colombia. RE Reportes del Emisor, Investigación e Información Económica. Bogotá D.C. (Colombia), No. 154, pp. 1-4. ISSN: 01240625.

Sattler, K, Dunemann, O., 2001. SQL Database Primitives for Decision Tree Classifiers. In: Paques H, Liu L, Grossman D, editors. The 10th ACM International Conference on Information and Knowledge Management. Atlanta, USA: ACM New York, pp. 379-86.

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, 23. Disponible en: http://www.rieoei.org/rie23a07.htm.

Tejedor, J. (2003). Poder explicativo de algunos determinantes del rendimiento en los estudios universitarios. Revista Española de Pedagogía, No. 224, pp. 5-32.

Timarán, R. y Millán (2006). New algebraic operators and SQL primitives for mining classification rules. En Computational Intelligence (pp. 61-65). Disponible en: http://www.actapress.com/PaperInfo.aspx?PaperID=29048&reason=500.

Timarán, R., Calderón, A. y Jiménez, J. (2013). Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. Revista Ventana Informática, No. 28 (ene.-jun.). Manizales, Colombia: Facultad de Ciencias e Ingeniería, Universidad de Manizales, pp. 31-47. ISSN: 0123-9678.

Timarán, S.R., Hernández, I., Caicedo, J., Hidalgo, A. & Alvarado, J (2016). Descubrimiento de Patrones de Desempeño Académico con árboles de decisión en las competencias genéricas de la formación profesional. Ediciones Universidad Cooperativa de Colombia. Bogotá, Colombia: Ediciones Universidad Cooperativa de Colombia, 2016. https://dx.doi.org/10.16925/9789587600490.

Timarán, R., Jiménez, J. y Calderón, A. (2017). Detección de patrones de deserción estudiantil con minería de datos. Editorial Universitaria, Universidad de Nariño. Pasto, Colombia. ISBN: 978-958-8958-38-5.

Timarán, R., Caicedo, J. & Hidalgo, A. (2019). Árboles de decisiones para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11º. Revista investigación, desarrollo e innovación, No. 9 (2), pp. 363-378. Doi: 10.19053/20278306.v9. n2.2019.9184.

Valero, S. (2009). Aplicación de técnicas de minería de datos para predecir deserción. Puebla, México: Universidad Tecnológica de Izúcar

de Matamoros. Disponible en: http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf.

Valero, S., Salvador, A. y 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. Puebla, México: Universidad Tecnológica de Izúcar de Matamoros. Disponible en: www.utim.edu.mx/~svalero/docs/e1.pdf.

Valli Jayanthi, S., Balakrishnan, S., Lim Siok Ching, A., Aaqilah Abdul Latiff, N. & Nasirudeen, A. M. A. (2014). Factors Contributing to Academic Performance of Students in a Tertiary Institution in Singapore. http://doi.org/10.12691/education-2-9-8, American Journal of Educational Research: 2 (9), pp. 752-758.

Villena, J. (2016). CRISP-DM: La metodología para poner orden en los proyectos de Data Science. Disponible en: https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science.

Viloria, J. (2007). Economía del Departamento de Nariño: Ruralidad y aislamiento geográfico. Documentos de trabajo sobre economía regional. Banco de la República. No. 87. ISSN: 1692-3715.

Willms, J. y Somers, M. (2001). Resultados Escolares en América Latina. Informe preparado para la Unesco. Por el Canadian Research Institute for Social Policy y la Universidad de New Brunswick en colaboración con el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación.

Witten, I., Frank, E. and Hall, M., 2011. Data Mining: Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufmann ISBN: 978-0-12-374856-0.

En este libro se presenta los resultados obtenidos al aplicar técnicas de minería de datos educativa con el fin de detectar factores asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media, que presentaron las pruebas Saber 11° en los años 2015 y 2016. La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Siguiendo la metodología CRISP-DM, se seleccionó, de las bases de datos del ICFES, la información socioeconómica, académica e institucional de estos estudiantes. Se construyó, limpió y transformó un repositorio de datos y utilizando la herramienta de minería de datos WEKA, se generaron árboles de decisión que permitieron identificar patrones asociados al buen o mal desempeño académico de los estudiantes en las pruebas Saber 11°. Los patrones descubiertos ayudarán en los procesos de toma de decisiones del MEN, ICFES y de las instituciones educativas que velan por la calidad de la educación en Colombia.

Universidad de Nariño

GRIAS

ISBN: 978-958-5123-81-6