

DESCUBRIMIENTO DE PATRONES

ASOCIADOS AL DESEMPEÑO ACADÉMICO
DE LOS ESTUDIANTES DE LOS DEPARTAMENTOS
DE CAUCA Y NARIÑO EN LAS
PRUEBAS **SABER 9°**



Silvio Ricardo Timarán Pereira
Fredy Alonso Vidal Alegría

editorial
Universidad de **Nariño**

DESCUBRIMIENTO DE **PATRONES**

ASOCIADOS AL DESEMPEÑO ACADÉMICO
DE LOS ESTUDIANTES DE LOS DEPARTAMENTOS
DE CAUCA Y NARIÑO EN LAS
PRUEBAS **SABER 9°**

DESCUBRIMIENTO DE **PATRONES**

ASOCIADOS AL DESEMPEÑO ACADÉMICO
DE LOS ESTUDIANTES DE LOS DEPARTAMENTOS
DE CAUCA Y NARIÑO EN LAS
PRUEBAS **SABER 9°**

Silvio Ricardo Timarán Pereira
Fredy Alonso Vidal Alegría

editorial
Universidad de **Nariño**

Timarán Pereira, Silvio Ricardo

Descubrimiento de patrones asociados al desempeño académico de los estudiantes de los departamentos de Cauca y Nariño en las Pruebas Saber 9° / Silvio Ricardo Timarán Pereira, Fredy Alonso Vidal Alegría. -- 1ª. ed. -- San Juan de Pasto : Editorial Universidad de Nariño, 2026.
241 p. : il., gráficas, tablas

Incluye bibliografía p. 222-233

ISBN: 978-628-7864-98-6 Impreso

ISBN: 978-628-7864-99-3 Digital

1. Calidad de la educación 2. Pruebas saber 3. Herramientas educativas 4. Didáctica y pedagogía 5. Rendimiento académico 6. Educación--Investigaciones 7. Sistemas educativos 8. Minería de datos 9. Rendimiento académico--Instituciones educativas--Departamentos del Cauca y Nariño I. Vidal Alegría, Fredy Alonso

371.26 T582 - SCDD-Ed. 23



Sección de Biblioteca
"Alberto Quijano Guerrero"

Descubrimiento de patrones asociados al desempeño académico de los estudiantes de los Departamentos de Cauca y Nariño en las Pruebas Saber 9°

© Editorial Universidad de Nariño

© Silvio Ricardo Timarán Pereira
Fredy Alonso Vidal Alegría

ISBN Impreso: 978-628-7864-98-6

ISBN Digital: 978-628-7864-99-3

DOI: <https://doi.org/10.22267/lib.udn.074>

Primera edición

Corrección de estilo: German Chaves Jurado

Diseño y diagramación: Angie Gabriela Ordoñez Chaves

Fecha de publicación: Junio 2026

San Juan de Pasto - Nariño - Colombia

Prohibida la reproducción total o parcial, por cualquier medio o cualquier propósito, sin la autorización escrita de sus Autores o de la Editorial Universidad de Nariño

AGRADECIMIENTOS

Expreso mi más profundo agradecimiento a la Universidad de Nariño por brindarme el apoyo necesario para llevar a cabo mi año sabático, tiempo fundamental en la elaboración de este libro. Asimismo, extiendo mi gratitud a la Institución Universitaria Colegio Mayor del Cauca por su invaluable colaboración durante mi estancia investigativa; su respaldo fue esencial para los resultados que hoy se plasman en estas páginas.

A mi esposa Amanda, por estar siempre presente con su apoyo incondicional en cada proyecto que emprendo y, especialmente, durante este periodo sabático. A mis hijos e hijas, quienes con su amor y alegría llenan de sentido mis esfuerzos. Finalmente, a los perritos y gatitos de casa, cuya compañía silenciosa y afectuosa hizo mucho más gratas las largas jornadas de escritura.

Silvio Ricardo Timarán Pereira

AGRADECIMIENTOS

A mi familia, por su apoyo incondicional a lo largo de mi vida.

A San, mi esposa, por su paciencia y por acompañarme con serenidad en cada etapa de mi camino. A Dani, cuya nobleza y creatividad son la motivación más profunda para seguir creciendo.

A Snoop, mi fiel amigo, por su silenciosa compañía durante las extensas jornadas de trabajo.

Expreso un sincero reconocimiento al profesor Ricardo Timarán Pereira, de la Universidad de Nariño, cuya colaboración, orientación y valioso aporte hicieron posible esta obra. Este libro es fruto del esfuerzo conjunto y del compromiso interinstitucional con el fortalecimiento de la educación y la investigación en nuestra región.

Fredy Alonso Vidal Alegría

CONTENIDO

INTRODUCCIÓN	12
--------------------	----

CAPÍTULO I:

PRUEBAS SABER 9º	18
------------------------	----

1.1 OBJETIVO PRUEBAS SABER 9º	18
-------------------------------------	----

1.2 NIVELES DE DESEMPEÑO EN LAS COMPETENCIAS	20
--	----

1.3 PRUEBA DE LENGUAJE	21
------------------------------	----

1.3.1 Competencia Comunicativa-Lectora (textual)	22
--	----

1.3.2 Competencia Comunicativa-Escritora (discursiva)	23
---	----

1.3.3 Componentes Evaluados en la Prueba de Lenguaje	24
--	----

1.4 PRUEBA DE MATEMÁTICAS	24
---------------------------------	----

1.4.1 Competencias Evaluadas en la Prueba de Matemáticas	25
--	----

1.4.2 Componentes Evaluados en la Prueba de Matemáticas	26
---	----

CAPÍTULO II:

ANÁLISIS EXPLORATORIO DE DATOS	28
--------------------------------------	----

2.1 CARACTERIZACIÓN DE LOS ESTUDIANTES DEL DEPARTAMENTO DEL CAUCA	28
--	----

2.1.1 Características Socioeconómicas de los Estudiantes Caucaños	29
--	----

2.1.2 Características Institucionales de los Colegios del Departamento del Cauca	31
---	----

2.1.3 Niveles de Desempeño Académico de los Estudiantes Caucaños en la Prueba de Lenguaje	33
--	----

2.1.4 Niveles de Desempeño Académico de los Estudiantes Caucaños en la Prueba de Matemáticas	40
---	----

2.2 CARACTERIZACIÓN DE LOS ESTUDIANTES DEL DEPARTAMENTO DE NARIÑO	44
--	----

2.2.1 Características Socioeconómicas de los Estudiantes Nariñenses	45
--	----

2.2.2 Características Institucionales de los Colegios del Departamento de Nariño	47
2.2.3 Niveles de Desempeño Académico de los Estudiantes Nariñenses en la Prueba de Lenguaje	54
2.2.4 Niveles de Desempeño Académico de los Estudiantes Nariñenses en la Prueba de Matemáticas	57
2.3 ANÁLISIS DEL DESEMPEÑO GENERAL EN LAS PRUEBAS SABER 9º DE LOS ESTUDIANTES DE LOS DEPARTAMENTOS DEL CAUCA Y NARIÑO	61

CAPÍTULO III:	
MATERIALES Y MÉTODOS	64
3.1 MATERIALES	64
3.2 MÉTODOS	74
3.3 METODOLOGÍA CRISP-DM (Cross-Industry Standard Process for Data Mining)	76
3.3.1 Análisis del Problema	77
3.3.2 Comprensión de los Datos	78
3.3.3 Preparación de los Datos	78
3.3.4 Modelado	96
3.3.5 Evaluación	100
3.3.6 Implementación	100

CAPÍTULO IV:	
RESULTADOS Y DISCUSIÓN	102
4.1 DEFINICIÓN DEL ALGORITMO DE ÁRBOLES DE DECISIÓN	104
4.2 DEFINICIÓN DEL MÉTODO DE PRUEBA DE LOS MODELOS	106
4.3 DEFINICIÓN DE HIPERPARÁMETROS	107
4.4 ÁRBOLES DE DECISIÓN PARA EL DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO EN LAS PRUEBAS SABER 9º DE LOS ESTUDIANTES DEL DEPARTAMENTO DEL CAUCA	110

4.4.1 Clasificación mediante Árboles de Decisión del Desempeño en la Competencia de Lenguaje de las Pruebas Saber 9º de los Estudiantes Caucanos	111
4.4.2 Clasificación mediante Árboles de Decisión del Desempeño en la Competencia de Matemáticas en las Pruebas Saber 9º de los Estudiantes Caucanos	121
4.4.3 Clasificación mediante Árboles de Decisión del Desempeño General en las Pruebas Saber 9º de los Estudiantes Caucanos	131
4.4.4 Discusión de Resultados de los Modelos Obtenidos de Desempeño Académico de los Estudiantes del Departamento del Cauca en las Pruebas Saber 9º	140
4.5 ÁRBOLES DE DECISIÓN PARA EL DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO EN LAS PRUEBAS SABER 9º DE LOS ESTUDIANTES DEL DEPARTAMENTO DE NARIÑO	142
4.5.1 Clasificación mediante Árboles de Decisión del Desempeño en la Competencia de Lenguaje en las Pruebas Saber 9º de los Estudiantes Nariñenses	143
4.5.2 Clasificación mediante Árboles de Decisión del Desempeño en la Competencia de Matemáticas de las Pruebas Saber 9º de los Estudiantes Nariñenses	153
4.5.3 Clasificación mediante Árboles de Decisión para el Desempeño General en las Pruebas Saber 9º de los Estudiantes Nariñenses	165
4.5.4 Discusión de Resultados de los Modelos Obtenidos de Desempeño Académico de los Estudiantes del Departamento de Nariño en las Pruebas Saber 9º	180
4.6 ÁRBOLES DE DECISIÓN PARA EL DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO EN LAS PRUEBAS SABER 9º DE LOS ESTUDIANTES DE LOS DEPARTAMENTOS DE CAUCA Y NARIÑO ...	183
4.6.1 Clasificación mediante Árboles de Decisión para el Desempeño en la Competencia de Lenguaje de los Estudiantes de los Departamentos de Cauca y Nariño	185
4.6.2 Clasificación mediante Árboles de Decisión para el Desempeño en la Competencia de Matemáticas de los Estudiantes de los Departamentos de Cauca y Nariño	193

4.6.3 Clasificación mediante Árboles de Decisión para el Desempeño General de los Estudiantes de los Departamentos de Cauca y Nariño en las pruebas Saber 9º	202
4.6.4 Discusión de Resultados de los Modelos Obtenidos de Desempeño Académico de los Estudiantes de los Departamento de Cauca y Nariño en las Pruebas Saber 9º	211
CONCLUSIONES Y TRABAJOS FUTUROS	214
REFERENCIAS	222

INTRODUCCIÓN

Los desafíos de la globalización, exigen, que los Estados preparen a sus ciudadanos, con habilidades y destrezas que les permitan enfrentar los nuevos paradigmas de la economía global. En este contexto, las evaluaciones en los distintos niveles educativos, se han convertido, en una herramienta fundamental, para medir los avances y retrocesos en los esfuerzos de cada país por mejorar continuamente (Torres, 2018).

El mejoramiento, tanto en la cantidad como en la calidad de la educación, es uno de los desafíos clave, que enfrenta Colombia, en su esfuerzo por impulsar el crecimiento económico y mejorar el bienestar de su población (OECD, 2016). Las evaluaciones nacionales e internacionales, revelan que, el sistema educativo colombiano, presenta bajos logros académicos en todos los niveles de estudio (Posada y Mendoza, 2014). Sin embargo, la calidad educativa, medida a través de pruebas de aprendizaje nacionales e internacionales, ha mostrado señales de progreso. Los resultados de las Pruebas PISA de 2006 y 2009, reflejan mejoras, en las tres áreas evaluadas: Matemáticas, Lectura y Ciencias. A pesar de estos avances, el país sigue ubicado en los rangos inferiores entre las naciones participantes y por debajo de otros países latinoamericanos, como Chile, México y Argentina (OECD, 2013).

El Ministerio de Educación Nacional (MEN) y el Instituto Colombiano para la Evaluación de la Educación (ICFES), han puesto en marcha, un conjunto de estrategias, para mejorar la calidad y pertinencia del sistema educativo. Estas iniciativas están enfocadas, en fortalecer las competencias básicas y ciudadanas, optimizar el sistema nacional de evaluación, reforzar el aseguramiento de la calidad y promover el desarrollo profesional de directivos y docentes.

Las Pruebas Saber, representan una herramienta clave, para evaluar la educación en los establecimientos públicos y privados del país. Su aplicación tiene como propósito, monitorear el rendimiento académico, entendido como el desarrollo de competencias básicas de los estudiantes, y hacer seguimiento a la calidad educativa en las instituciones. Asimismo, permiten valorar los avances en determinados periodos, medir el impacto de los programas y definir acciones de mejoramiento específicas.

Actualmente, el ICFES, diseña y aplica las Pruebas Saber 3°, Saber 5°, Saber 9° y Saber 11°, dirigidas a la Educación Básica y Media, así como la Prueba Saber Pro, destinada a evaluar la Educación Superior (ICFES, 2014a).

Para fortalecer el sistema de evaluación de los estudiantes de Educación Básica Secundaria, se implementan, las Pruebas Saber 9°, las cuales, se han aplicado de manera censal hasta el año 2017. Estas pruebas permiten, medir el desempeño académico de los estudiantes de 9° grado, a través de la evaluación periódica de competencias básicas, proporcionando información sobre la calidad educativa en instituciones oficiales, privadas, urbanas y rurales.

Las pruebas, valoran las competencias adquiridas por los estudiantes, durante la Básica Secundaria, facilitando una transición adecuada hacia la Educación Media (Grados 10 y 11) en concordancia con los estándares básicos de competencias establecidos por el MEN. Estos estándares sirven, como referentes comunes, para determinar, en qué medida los estudiantes y el sistema educativo en su conjunto, cumplen con las expectativas de calidad educativa (ICFES, 2017a).

La mayoría de los estudios realizados en Colombia para evaluar el desempeño académico en las Pruebas Saber, como los de Gaviria y Barrientos, 2001; Correa, 2004; Barrientos, 2008; Chica *et al.*, 2010; ICFES, 2011; Gómez, 2014; Torres *et al.*, 2014; Hernández, 2015; Gutiérrez, 2015; ICFES, 2017b y Agudelo *et al.*, 2020, se han basado en técnicas estadísticas tradicionales. Sin embargo, estos estudios, no han considerado diversas variables sociodemográficas, económicas, académicas e institucionales, que afectan el rendimiento de los estudiantes y la calidad

educativa en las instituciones de Educación Secundaria, dejando valiosa información sin aprovechar.

La aplicación de enfoques más avanzados, como la minería de datos (Timarán *et al.*, 2021a; Timarán *et al.*, 2021b), permitiría revelar, patrones ocultos en los datos y proporcionar una comprensión más profunda, de los factores que inciden en el desempeño académico, contribuyendo así, al mejoramiento del sistema educativo.

Según Frawley *et al.*, (1992) la minería de datos, es “un proceso de detección de patrones válidos, novedosos, potencialmente útiles y fácilmente entendibles por el usuario, a partir de los datos”. 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 (por su sigla en inglés *Educational Data Mining*) (Timarán *et al.*, 2021a).

La Minería de Datos Educativos, permite analizar grandes volúmenes de información para identificar tendencias, correlaciones y anomalías, facilitando la toma de decisiones informadas en el diseño de entornos de aprendizaje, la asignación de recursos y la evaluación de estrategias educativas. Sus técnicas pueden emplearse, para explicar o predecir diversos fenómenos dentro del ámbito educativo (Timarán *et al.*, 2021a; Timarán *et al.*, 2021b).

Por ejemplo, mediante la Minería de Datos, es posible prever, con un alto grado de precisión, la probabilidad de deserción estudiantil (Valero, 2009; Valero *et al.*, 2010; Timarán *et al.*, 2017; Díaz-Martínez *et al.*, 2021). Además, las instituciones educativas, pueden utilizar estos métodos, para realizar análisis detallados sobre las características de sus estudiantes y sus sistemas de evaluación, detectando prácticas exitosas o, por el contrario, identificando fraudes e inconsistencias (Valero *et al.*, 2010; Cedillo *et al.*, 2024).

Este libro de investigación, describe el proceso de aplicación de técnicas de Minería de Datos Educativa, con el propósito, de identificar patrones relacionados con el desempeño académico de los estudiantes

en competencias de Lenguaje, Matemáticas y Desempeño General. El estudio se centra, en alumnos de instituciones educativas de los departamentos de Nariño y Cauca, que, cursando el Grado Noveno, presentaron las Pruebas Saber 9°.

Las fuentes de datos iniciales, provienen, de los resultados de las Pruebas Saber 9°, almacenados en el repositorio de datos abiertos del ICFES, conocido como Data Icfes (ICFES, 2024a). Los registros más recientes y disponibles de los estudiantes colombianos que realizaron estas pruebas, corresponden al año 2017.

Con base en esta información y siguiendo las diferentes fases de la metodología CRISP-DM (Villena, 2016), ampliamente utilizada en entornos académicos e industriales para el desarrollo de proyectos de Minería de Datos, se seleccionaron los datos socioeconómicos, académicos e institucionales de los estudiantes de los departamentos de Cauca y Nariño, que presentaron las Pruebas Saber 9° en el año 2017.

Se realizó un análisis exploratorio de datos, para evaluar su calidad, seguido de un proceso de limpieza y transformación, para obtener conjuntos de datos minables por departamento, además de uno unificado que abarcara ambos. Se optó, por la técnica de clasificación, mediante árboles de decisión, debido a su facilidad de interpretación y comprensión de patrones (Frawley *et al.*, 1992).

Se evaluaron varios modelos construidos con distintos algoritmos de clasificación, mediante árboles de decisión, seleccionando el generado por el algoritmo J48, basado en el C4.5 propuesto por Quinlan, (1993) por ofrecer los mejores resultados. Para evaluar la exactitud de los modelos, los conjuntos de datos de cada departamento y el unificado, fueron divididos en dos grupos: uno de entrenamiento y otro de prueba. Se asignó el 80% del total de instancias, al conjunto de entrenamiento y el 20% al de prueba.

En cada competencia evaluada, se clasificó como buen desempeño, a aquellos estudiantes, cuyo puntaje superaba la media nacional, mientras

que se consideró bajo desempeño, a quienes obtuvieron puntajes por debajo de dicho promedio.

En esta investigación, fue de gran interés, conocer especialmente, los factores asociados al desempeño académico de los estudiantes de los departamentos del Cauca y Nariño, que están por debajo de la media nacional en las Competencias de Lenguaje, Matemáticas y Desempeño General de las Pruebas Saber 9º, por las siguientes razones:

- Permite identificar los problemas específicos, que afectan a estos estudiantes, lo que facilita la creación de estrategias de intervención más efectivas.
- Los datos obtenidos pueden informar, la creación de políticas educativas, que aborden las necesidades de los estudiantes con bajo desempeño, promoviendo una educación más equitativa.
- Conocer los factores asociados al bajo desempeño, permite a las instituciones educativas, asignar recursos de manera más eficiente, asegurando que se dirijan a los estudiantes que más lo necesitan.
- Los estudiantes con bajo desempeño, son más propensos a abandonar sus estudios. Identificar sus perfiles y factores asociados, permite implementar medidas preventivas para reducir la deserción escolar.
- Al entender los factores que contribuyen al bajo desempeño, se pueden desarrollar programas de apoyo y tutoría, que ayuden a mejorar el rendimiento académico de estos estudiantes.

Este libro documenta y difunde, el conocimiento descubierto, convirtiéndolo en una fuente adicional de consulta sobre el desempeño académico de los estudiantes de Educación Básica Secundaria en las Competencias de Lenguaje, Matemáticas y Desempeño General, evaluadas en las Pruebas Saber 9º del año 2017. Además, esta información puede integrarse, al conocimiento existente, para fortalecer los procesos

de toma de decisiones de los organismos gubernamentales y las instituciones educativas de los departamentos de Cauca y Nariño, así como del MEN, contribuyendo a la implementación de estrategias y planes, orientados a mejorar la calidad de la educación secundaria en la región.

La presente investigación, proyectó dar respuesta a la pregunta: ¿Cuáles son los factores socioeconómicos, académicos e institucionales, asociados al desempeño académico de los estudiantes de los departamentos de Cauca y Nariño que presentaron las Pruebas Saber 9º?

Este libro se estructura en cinco capítulos. En el Capítulo I, se presenta, una conceptualización sobre las Pruebas Saber 9º, abordando su propósito y las competencias que evalúan. En el Capítulo II, se lleva a cabo, un análisis exploratorio de datos, con el objetivo de comprender en mayor profundidad, los resultados obtenidos por los estudiantes de Grado 9º, en los departamentos de Cauca y Nariño, en las Pruebas Saber 9º. El Capítulo III, describe, los conjuntos de datos utilizados en esta investigación y detalla las fases de la metodología CRISP-DM, que orientó el proceso de Minería de Datos Educativa aplicado. En el Capítulo IV, se presentan los modelos de desempeño académico, obtenidos mediante la técnica de clasificación con árboles de decisión, considerando los resultados en Lenguaje, Matemáticas y Desempeño General en las Pruebas Saber 9º, de los estudiantes, de los departamentos de Cauca y Nariño. Además, se analizan los hallazgos y se discuten sus implicaciones. Finalmente, en el Capítulo V, se exponen las conclusiones del estudio y se plantean posibles líneas de investigación futuras, orientadas a continuar el análisis del desempeño académico, con enfoques de Minería de Datos Educativa.

CAPÍTULO I PRUEBAS SABER 9º

1.1 OBJETIVO PRUEBAS SABER 9º

Uno de los principales objetivos de la política educativa en Colombia, es asegurar, que todos los estudiantes, sin distinción de su origen o condiciones socioeconómicas y culturales, accedan a una educación de excelencia en las instituciones escolares. Esta formación, debe fortalecer el desarrollo de competencias esenciales para su vida, promoviendo la convivencia, la productividad en diversos entornos y la capacidad de aprendizaje continuo, a lo largo de su trayectoria personal y profesional.

La Ley General de Educación (Ley 115 de 1994) establece el Sistema Nacional de Evaluación de la Calidad de la Educación, cuya gestión recae en el Ministerio de Educación Nacional (MEN) y el Instituto Colombiano para la Evaluación de la Educación (ICFES). Su propósito, es realizar evaluaciones periódicas para medir la calidad educativa que reciben niños, niñas y jóvenes en el país (ICFES, 2009).

Por otro lado, la Ley 1324 de 2009, otorga al ICFES (2022a), la responsabilidad de evaluar la formación impartida en el sistema educativo, mediante exámenes externos estandarizados. Estos exámenes siguen criterios de estandarización, lo que implica, que tanto su aplicación como el procesamiento de los resultados, se realizan bajo condiciones uniformes. Asimismo, la normativa establece, que el MEN, es la entidad encargada de definir los aspectos que deben ser evaluados en dichas pruebas (ICFES, 2014b).

Las Pruebas Saber, administradas por el ICFES, son evaluaciones externas estandarizadas, diseñadas para medir el desempeño de los estudiantes en relación con las competencias básicas establecidas por el

Ministerio de Educación Nacional (MEN). Estas pruebas valoran, los logros adquiridos por los estudiantes, al concluir diferentes etapas del sistema educativo: Saber 3º y 5º en Educación Primaria, Saber 9º al finalizar la Educación Secundaria Básica y Saber 11º al término de la Educación Media.

Estos procesos de evaluación, buscan supervisar el desempeño del sistema educativo y ofrecer datos fundamentales para la toma de decisiones estratégicas, con el fin de impulsar mejoras continuas. Tanto el Ministerio de Educación Nacional como las instituciones educativas, utilizan esta información para fortalecer la calidad del aprendizaje y la enseñanza (ICFES, 2014b).

Las Pruebas Saber 3º, 5º y 9º, han sido un recurso clave, en el diseño de políticas públicas destinadas a mejorar la calidad educativa en Colombia. Junto con otras evaluaciones, sus resultados han proporcionado información periódica, precisa y comparable, sobre la calidad de la Educación Básica Primaria y Secundaria en el país.

Desde 2009 hasta 2017, Colombia presentó informes, sobre los resultados de estas pruebas, comenzando con los estudiantes de Grado 5 y 9 y posteriormente, incorporando a los de Grado 3. Esta información ha sido fundamental, para orientar la política educativa, utilizando los datos obtenidos para analizar los niveles de competencia alcanzados en áreas como Matemáticas, Lectura, Ciencias Naturales, Competencias Ciudadanas y cuestionarios sobre factores asociados (ICFES, 2022b).

Las Pruebas Saber 9º, tienen como propósito, mejorar la calidad educativa en Colombia, mediante evaluaciones periódicas, que analizan las competencias básicas de los estudiantes y los factores que influyen en sus logros. Los resultados obtenidos permiten, que los establecimientos educativos, las secretarías de educación, el Ministerio de Educación Nacional (MEN) y la sociedad en general, identifiquen fortalezas y áreas de oportunidad, facilitando la formulación de estrategias de mejoramiento en cada ámbito.

Además, la regularidad de estas pruebas, permite evaluar los avances a lo largo del tiempo y medir el impacto de programas y acciones, destinadas a fortalecer la educación (ICFES, 2017a).

Las Pruebas Saber 9°, se aplican periódicamente, para evaluar las competencias básicas de los estudiantes de Grado 9°, permitiendo medir su desempeño y la calidad educativa de instituciones oficiales, privadas, urbanas y rurales. Estas pruebas analizan, las habilidades adquiridas entre los Grados 8° y 9°, conforme a los estándares básicos de competencia establecidos por el MEN. Dichos estándares proporcionan, un marco común, para determinar, en qué medida los estudiantes y el sistema educativo en su conjunto cumplen con las expectativas de calidad.

Las evaluaciones Saber 9°, se centran en las áreas de Matemáticas y Lenguaje, aunque no abarcan la totalidad de los contenidos ni estándares definidos para cada disciplina y algunos logros, solo pueden observarse en el contexto de las actividades escolares. Estas pruebas constituyen, un indicador clave del avance en el proceso de formación en Educación Básica. A través de ellas, es posible establecer los niveles de desempeño, identificar fortalezas y debilidades y desarrollar estrategias concretas, para mejorar la calidad educativa (ICFES, 2017a).

1.2 NIVELES DE DESEMPEÑO EN LAS COMPETENCIAS

Al evaluar los Niveles de Desempeño, es fundamental comprender, que estos describen cualitativamente las competencias de los estudiantes en cada prueba. Según Agudelo et al., (2020) esta información permite:

- Determinar la capacidad del estudiante, para resolver preguntas o problemas de diversos grados de complejidad.
- Analizar los niveles de aprendizaje alcanzados, identificando diferencias entre los estudiantes. Esto facilita, la formulación de metas específicas y/o diferenciadas para el mejoramiento educativo.

- Comparar los resultados en distintos niveles de agregación, como el porcentaje de estudiantes dentro de un establecimiento educativo o municipio, que presentan un nivel de desempeño mayor, menor o similar al evaluado.

Los Niveles de Desempeño, son una herramienta clave, para identificar lo que un estudiante es capaz de lograr con base en los resultados de la evaluación.

En la tabla 1, se describen las capacidades que debe tener un estudiante de acuerdo con el nivel que obtenga en la competencia evaluada

Tabla 1. Nivel de Desempeño

Nivel	Capacidades
Avanzado	Muestra un desempeño sobresaliente en las competencias esperadas para el área y grado evaluado
Satisfactorio	Muestra un desempeño adecuado en las competencias exigibles para el área y grado evaluado. Este es el nivel esperado que todos, o la gran mayoría de los estudiantes, deberían alcanzar
Mínimo	Muestra un desempeño mínimo en las competencias exigibles para el área y grado evaluado
Insuficiente	No supera las preguntas de menor complejidad de la prueba

Fuente: Agudelo *et al.*, 2020

1.3 PRUEBA DE LENGUAJE

De acuerdo con la Guía de Orientación de las Pruebas Saber 9º (ICFES, 2017a), el examen de Lenguaje, evalúa dos competencias fundamentales: la Comunicativa-Lectora, enfocada en la Comprensión Textual y la Comunicativa-Escritora, centrada en la construcción discursiva. Estas habilidades deben reflejarse, en la estructura y organización del texto, su significado y sentido, así como en los contextos y propósitos de la comunicación. A su vez, la evaluación se desarrolla en tres niveles de competencia: Literal, Inferencial y Crítico.

1.3.1 Competencia Comunicativa-Lectora (*textual*)

Esta competencia evalúa, la capacidad de comprensión, uso y análisis de la información presente en diversos tipos de textos, estableciendo una interacción dinámica, entre el contenido y el lector. Se examina, cómo los estudiantes leen e interpretan distintos textos, esperando que puedan comprender, tanto la información explícita como la implícita, conectar sus contenidos con sus conocimientos previos sobre un tema específico y desarrollar inferencias, formular conclusiones y adoptar posiciones argumentadas. En términos generales, la prueba lectora, mide la habilidad de los estudiantes, para interpretar el significado y la estructura de diversos textos, promoviendo una reflexión sobre qué dice el texto, cómo lo dice, con qué propósito, por qué, cuándo y quién lo expresa (ICFES, 2017a).

Siguiendo las orientaciones establecidas en los estándares básicos de competencias, las preguntas diseñadas para evaluar esta habilidad, se estructuran en torno a textos seleccionados, con base en los siguientes criterios:

- La pertinencia del tema según la edad y el nivel educativo de los estudiantes.
- El uso del vocabulario.
- La complejidad de la sintaxis.
- Los conocimientos previos relacionados con el grado cursado.
- La sofisticación del estilo.
- La estructura del texto y su nivel de complejidad.
- La extensión del contenido.

Asimismo, se consideran distintos tipos de textos, teniendo en cuenta la diversidad en la organización y estructuración de la información:

- Textos continuos, compuestos por oraciones y párrafos, ya sea en prosa o en verso.
- Textos discontinuos, como listas, formularios, gráficos o diagramas.
- Textos mixtos, como historietas o cómics.

1.3.2 Competencia Comunicativa-Escritora (discursiva)

Esta competencia evalúa, cómo el estudiante selecciona y emplea, tanto los mecanismos que garantizan el uso coherente de la lengua como las estrategias discursivas necesarias para construir significado y dar unidad al escrito. Esto implica, la capacidad de adoptar una postura frente a la lectura, utilizando diversas estrategias de pensamiento y generando nuevos significados (ICFES, 2017a).

Se espera, que el estudiante aplique sus conocimientos sobre el uso y funcionamiento de la lengua en situaciones comunicativas específicas, respondiendo eficazmente a las necesidades del intercambio y a las demandas particulares del contexto. Para ello, debe seguir procedimientos sistemáticos, que aporten coherencia y cohesión al texto (ICFES, 2017a).

Según los estándares básicos de competencias, esta habilidad está relacionada, con la producción de textos escritos, que cumplan los siguientes requisitos:

- Satisfacer las necesidades comunicativas, ya sea narrar, informar, exponer, solicitar o argumentar sobre un tema específico.
- Seguir procedimientos sistemáticos para su elaboración.
- Aplicar los conocimientos del autor, sobre los temas tratados, así como el funcionamiento de la lengua en distintas situaciones comunicativas.

La prueba evalúa, tanto el proceso de escritura como la escritura en sí misma. Esto significa, que los estudiantes no tendrán que redactar textos, sino que las preguntas explorarán qué tipos de textos elegirían para cumplir un propósito comunicativo específico, cómo los organizarían para expresar un mensaje o una idea y qué aspectos considerarían para utilizar palabras y frases de manera adecuada, garantizando que el texto tenga sentido (ICFES, 2017a).

1.3.3 Componentes Evaluados en la Prueba de Lenguaje

Para evaluar las competencias textual y discursiva, se consideran tres componentes transversales: Sintáctico, Semántico y Pragmático (ICFES, 2017a).

- **Componente Semántico:** se refiere al significado del texto y su sentido. Este aspecto indaga, sobre qué se dice en el texto.
- **Componente Sintáctico:** está relacionado con la organización del texto, garantizando su coherencia y cohesión. Este componente examina, cómo se expresa la información.
- **Componente Pragmático:** aborda el propósito del mensaje en función del contexto comunicativo, es decir, para qué se transmite la información (ICFES, 2017a).

1.4 PRUEBA DE MATEMÁTICAS

En esta prueba se adopta, una perspectiva integradora, basada en los lineamientos curriculares y los estándares básicos de competencias, los cuales constituyen el fundamento pedagógico de la disciplina y orientan su desarrollo en el ejercicio educativo. Estos estándares reconocen, que las Matemáticas van más allá de un sistema teórico, pues representan una herramienta práctica fundamental para afrontar y comprender diversas situaciones. Por esta razón, la educación en esta área, debe otorgar gran importancia a la formación de conceptos, pero, sobre todo, al desarrollo de las destrezas necesarias para la resolución de problemas en distintos contextos y para la comunicación efectiva a través del lenguaje matemático. Se privilegian como escenarios de aprendizaje, las situaciones problemáticas relacionadas con la vida cotidiana, otras disciplinas científicas y la propia naturaleza de las Matemáticas. La evaluación, en este sentido, busca medir la capacidad del individuo, para aplicar conceptos y estructuras matemáticas en un contexto específico, así como las estrategias empleadas en su interpretación (ICFES, 2017a).

En la Prueba de Matemáticas, se busca evaluar las significaciones que el estudiante ha construido y que pone en juego, al enfrentarse a diversas situaciones problema. En ella, se analiza tanto el significado de los

conceptos matemáticos como su aplicación, particularmente en el proceso de matematización, el cual exige al estudiante, simbolizar, formular, cuantificar, validar, representar y generalizar, entre otras habilidades. Estas actividades le permiten, realizar descripciones matemáticas, ofrecer explicaciones y seleccionar posibles construcciones (ICFES, 2017a).

1.4.1 Competencias Evaluadas en la Prueba de Matemáticas

La Prueba Saber en el área de Matemáticas, evalúa tres competencias fundamentales: Comunicación, Razonamiento y Resolución de Problemas, las cuales, los estudiantes deben demostrar en tres contextos del conocimiento matemático. El primero, está relacionado con los números, las operaciones y sus transformaciones; el segundo, abarca los problemas propios de la Geometría y la medición y el tercero, se centra en los fundamentos de la Estadística (ICFES, 2017a).

Para la construcción de la prueba, las competencias matemáticas se organizan en tres grandes categorías: Razonamiento y Argumentación, Comunicación, Representación y Modelación y Planteamiento y Resolución de Problemas (ICFES, 2017a):

- **La Competencia de Razonamiento y Argumentación:** se centra, en la habilidad del estudiante para analizar y justificar los caminos seguidos en la resolución de problemas. Implica la capacidad de explicar el cómo y el porqué de los procedimientos utilizados, así como de fundamentar estrategias y métodos empleados en situaciones problemáticas. Además, abarca la formulación de hipótesis, la generación de opiniones e ideas, la exploración de ejemplos y contraejemplos y la validación de argumentos. También incluye, la identificación de patrones y su expresión matemática, la generalización de propiedades y relaciones, el planteamiento de preguntas y el reconocimiento de distintos tipos de razonamiento. Finalmente, permite evaluar cadenas de argumentos con criterio y precisión (ICFES, 2017a).

- **La Competencia de Comunicación, Representación y Modelación:** Esta competencia evalúa la facultad del estudiante para transmitir ideas y decodificar conceptos matemáticos mediante múltiples registros, ya sean gráficos, simbólicos o algebraicos. Según el ICFES (2017a), no solo implica el dominio de fórmulas y variables, sino también la habilidad de transitar entre el lenguaje formal y el natural. En última instancia, esta capacidad se manifiesta cuando el evaluado logra transformar estructuras matemáticas complejas en explicaciones claras y manejables en sus propias palabras.
- **La Competencia de Planteamiento y Resolución de Problemas:** se refiere, a la capacidad del estudiante, para formular problemas a partir de diversas situaciones, tanto dentro como fuera del ámbito matemático. También implica el desarrollo y aplicación de estrategias para la solución de problemas, así como la justificación de los métodos e instrumentos utilizados. Además, abarca la evaluación de la pertinencia de un cálculo exacto o aproximado, la valoración de la razonabilidad de una respuesta obtenida y la verificación e interpretación de los resultados en función del problema original. Finalmente, esta competencia permite, generalizar soluciones y estrategias, para abordar nuevas situaciones problemáticas de manera eficaz (ICFES, 2017a).

1.4.2 Componentes Evaluados en la Prueba de Matemáticas

Para estructurar la evaluación de la Prueba de Matemáticas, las competencias se han agrupado en tres componentes principales: Numérico-Variacional, Geométrico-Métrico y Aleatorio. Cada uno, abarca aspectos fundamentales de las Matemáticas, incluyendo operaciones y relaciones entre cantidades, medición de figuras, experimentación y análisis de probabilidades. Esta clasificación, no busca dividir las Matemáticas en áreas aisladas, sino proporcionar un esquema de organización que refleje la totalidad de los ejes matemáticos establecidos en los estándares básicos de competencias, permitiendo así, una comprensión integral de la disciplina (ICFES, 2017a):

- El Componente Numérico-Variacional: abarca los aspectos fundamentales de los números y la numeración, incluyendo su significado y la estructura del sistema numérico. También considera, las operaciones Matemáticas, sus propiedades, efectos y las relaciones entre ellas. Además, implica el reconocimiento de regularidades y patrones, para identificar cambios constantes en una serie de valores y analizar su comportamiento. Este componente comprende, la identificación de variables y la descripción de fenómenos de cambio y dependencia, así como los conceptos y procedimientos relacionados con la variación directa, la proporcionalidad y la variación lineal, tanto en contextos aritméticos como geométricos. Asimismo, integra el uso del lenguaje simbólico (álgebraico), articulado a través de convenciones y esquemas representativos, para el manejo de variables, la variación inversa y el concepto de función (ICFES, 2017a).
- El Componente Geométrico-Métrico: se enfoca en la construcción y manipulación de representaciones de objetos en el espacio, así como en el análisis de sus relaciones y transformaciones. Incluye la comprensión del espacio, el estudio abstracto de figuras y formas en el plano y en el espacio y la identificación de patrones y regularidades. También abarca el razonamiento geométrico, la resolución de problemas de medición, la estimación y descripción de magnitudes y la transformación de figuras en distintos planos. Asimismo, contempla la selección adecuada de unidades de medida, patrones e instrumentos, el uso correcto de las unidades y la aplicación de conceptos fundamentales, como perímetro, área y volumen (ICFES, 2017a).
- El Componente Aleatorio: se refiere, a la capacidad de representar, leer e interpretar datos en distintos contextos. Incluye, el análisis cualitativo de regularidades y tendencias, así como la formulación de inferencias y argumentos, con base en medidas de tendencia central y dispersión. Además, abarca el reconocimiento, la descripción y el estudio de eventos aleatorios, para comprender mejor su comportamiento y aplicación en diversas situaciones (ICFES, 2017a).

ANÁLISIS EXPLORATORIO DE DATOS

CAPÍTULO II

El Análisis Exploratorio de Datos (EDA), (por sus siglas en inglés: *Exploratory Data Analysis*), es un conjunto de técnicas estadísticas, diseñadas para examinar, describir y resumir la naturaleza de los datos, facilitando la comprensión de las relaciones entre las variables de interés. Su objetivo principal es, permitir un análisis preliminar de los datos antes de formular hipótesis, ayudando a detectar errores evidentes, identificar patrones, reconocer valores atípicos o eventos anómalos y descubrir conexiones relevantes entre variables.

La evaluación del desempeño académico en instituciones de Educación Secundaria, es fundamental, ya que proporciona evidencia sólida sobre la calidad educativa en distintos municipios, subregiones y departamentos. Con el propósito de profundizar en los resultados obtenidos por los estudiantes de Grado 9º de los departamentos de Cauca y Nariño en las Pruebas Saber 9º, se llevó a cabo un EDA, sobre los conjuntos de datos extraídos de las bases de datos del ICFES, permitiendo una mejor interpretación y análisis de los resultados.

2.1 CARACTERIZACIÓN DE LOS ESTUDIANTES DEL DEPARTAMENTO DEL CAUCA

En el año 2017, 16,392 estudiantes del departamento del Cauca, participaron en las Pruebas Saber 9º. Para el análisis de los resultados, se consideraron 19 variables, distribuidas en tres categorías: 9 variables asociadas a factores socioeconómicos, 6 variables relacionadas con factores institucionales y 4 variables enfocadas en aspectos académicos.

2.1.1 Características Socioeconómicas de los Estudiantes Caucanos

En la tabla 2, se presentan las características socioeconómicas de estos estudiantes.

Tabla 2. Características Socioeconómicas de los Estudiantes Caucanos que Presentaron la Prueba Saber 9º

Variables socioeconómicas		N	%
Sexo	Masculino	8,578	52.3%
	Femenino	7,814	47.7%
Edad	13 años o menos	512	3.1%
	14 años	5,589	34.1%
	15 años	5,616	34.3%
	16 años o mas	4,675	28.5%
Educación del Padre	Primaria	7,690	46.9%
	Bachillerato	5,851	35.7%
	Técnico o Tecnólogo	1,232	7.5%
	Universitario – Pregrado	751	4.6%
	Universitario – Posgrado	868	5.3%
Educación de la Madre	Primaria	7,135	43.5%
	Bachillerato	5,917	36.1%
	Técnico o Tecnólogo	1,644	10.0%
	Universitario – Pregrado	816	5.0%
	Universitario – Posgrado	880	5.4%
Hacinamiento	Hacinamiento crítico	115	0.7%
	Hacinamiento medio	901	5.5%
	Sin hacinamiento	15,376	93.8%
TIC	Malo	3,435	21.0%
	Regular	8,243	50.2%
	Bueno	4,714	28.8%

Variables socioeconómicas		N	%
Electrodomésticos	Malo	7,296	44.5%
	Regular	5,126	31.3%
	Bueno	3,970	24.2%
Reprobado	No	10,024	61,2%
	Si	6,368	38.8%
Cuartil INSE	Q1	3,886	23.7%
	Q2	4,076	24.9%
	Q3	4,228	25.8%
	Q4	4,202	25.6%

Fuente. Esta Investigación

El análisis de los datos de la tabla 2, revela que, en 2017, un total de 16,392 estudiantes de Grado 9º del departamento del Cauca, presentaron las Pruebas Saber 9º. Del total de participantes, el 52.3% fueron hombres y el 47.7% mujeres, con una diferencia de 4.6% a favor de los hombres. La mayoría de los estudiantes, tenían entre 14 y 15 años, representando el 68.4% del total.

En cuanto al nivel educativo de los padres, la proporción de quienes alcanzaron Educación Primaria o Secundaria, fue ligeramente superior entre los padres (82.6%) que entre las madres (79.6%). Sin embargo, en Educación Superior (tecnológica, pregrado y posgrado), el porcentaje de madres con estudios avanzados (20.5%) superó al de los padres (17.4%).

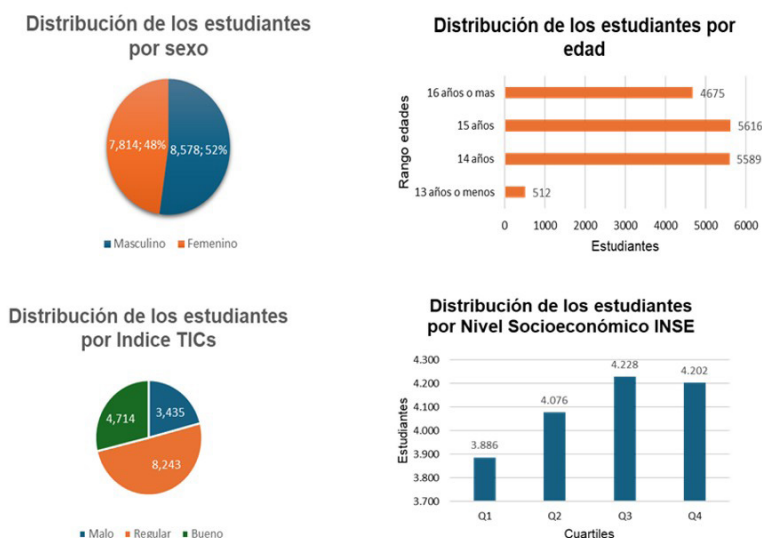
Respecto a las condiciones de vivienda, solo el 6.2% de los estudiantes vivía en situación de hacinamiento. En términos de acceso a tecnologías de la información y comunicación (TIC), el 28.8% reportó un nivel bueno, mientras que en el acceso a electrodomésticos (estufa, microondas, nevera), el 24.2% manifestó tener un nivel bueno.

Además, el 38.8% de los estudiantes que presentaron la prueba, reprobaron al menos un grado. Finalmente, en cuanto al Índice de Nivel Socioeconómico (INSE), el 48.6% de los estudiantes se encontraba en

condiciones favorables (cuartiles 1 y 2), mientras que el 50.4% presentaba condiciones menos favorables (cuartiles 3 y 4).

En la figura 1, se pueden observar algunas características socioeconómicas de los estudiantes del departamento del Cauca, que presentaron las Pruebas Saber 9º, discriminadas por sexo, rango de edades, índice TICs y por índice INSE

Figura 1. Caracterización de los Estudiantes Caucaños que Presentaron las Pruebas Saber 9º.



Fuente. Esta Investigación

2.1.2 Características Institucionales de los Colegios del Departamento del Cauca

En la tabla 3, se presentan las características institucionales de los colegios del departamento del Cauca, a los cuales pertenecían los estudiantes que presentaron las Pruebas Saber 9º.

Tabla 3. Características Institucionales de los Colegios de los Estudiantes Caucanos que Presentaron las Pruebas Saber 9º

Variables Institucionales		N	%
Tipo de Colegio	Femenino	636	3.9%
	Masculino	211	1.3%
	Mixto	15,545	94.8%
Régimen	Oficial	15,025	91.7%
	No oficial	1,367	8.3%
Carácter	Académico	8,139	49.7%
	Técnico	6,431	39.2%
	Técnico/Académico	1,822	11.1%
Área ubicación	Urbano	9,201	56.1%
	Rural	7,191	43.9%
Jornada	Mañana	11,847	72.3%
	Tarde	1,713	10.5%
	Nocturna	9	0.1%
	Completa	2,418	14.8%
	Única	405	2.5%
Subregión	Norte	5,409	33.0%
	Popayán	3,854	23.5%
	Centro	2,771	16.9%
	Oriente	1,044	6.4%
	Occidente	777	4.7%
	Sur	1,511	9.2%
	Macizo	822	5.0%
	Bota Caucana	204	1.2%

Fuente. Esta Investigación

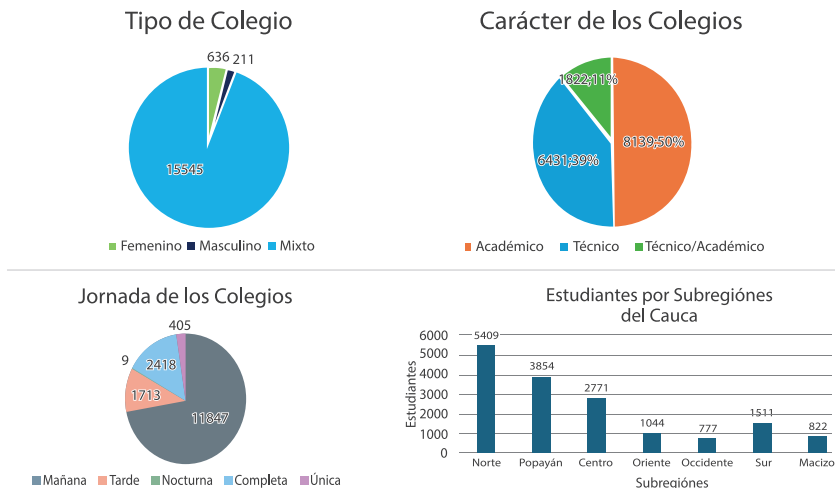
Según los datos de la tabla 3, la mayoría de los estudiantes Caucanos que presentaron las Pruebas Saber 9º en 2017, provenían de colegios oficiales y mixtos, de orientación académica, ubicados en el casco urbano

de los distintos municipios del departamento. La jornada escolar predominante, fue la diurna (mañana).

En cuanto a la distribución geográfica, la mayor concentración de estudiantes correspondió a las subregiones del norte del Cauca, que incluyen municipios, como Buenos Aires, Caldono, Caloto, Corinto, Guachené, Jambaló, Miranda, Padilla, Puerto Tejada, Santander de Quilichao, Suárez, Toribio y Villa Rica. Le siguieron en cantidad de estudiantes, la capital Popayán y la subregión del Centro, conformada por Cajibío, El Tambo, Morales, Piendamó, Puracé, Silvia y Timbío.

En la figura 2, se puede observar algunas características de los colegios del departamento del Cauca, cuyos estudiantes presentaron las Pruebas Saber 9º.

Figura 2. Características de los Colegios del Departamento del Cauca con Número de Estudiantes que Presentaron las Pruebas Saber 9º.



Fuente: Esta investigación

Como se puede observar en la figura 2, la gran mayoría de estudiantes caucanos que presentaron las pruebas Saber 9º en el año 2017, pertenecen a colegios de tipo mixto, de carácter académico, de jornada única y son de las subregiones del Norte, Popayán y Centro del departamento.

La tabla 4, presenta los municipios del departamento del Cauca, a los que pertenecían las instituciones de Educación Media, cuyos estudiantes participaron en las Pruebas Saber 9º, durante el periodo de estudio de esta investigación.

Tabla 4. Número de Estudiantes de los Municipios del Departamento del Cauca que Presentaron las Pruebas Saber 9º

Subregión	Municipio	N	%
Norte	Santander de Quilichao	1,286	7.8%
	Puerto Tejada	649	4.0%
	Caldono	614	3.7%
	Caloto	443	2.7%
	Toribio	414	2.5%
	Miranda	370	2.3%
	Corinto	300	1.8%
	Buenos Aires	261	1.6%
	Villa rica	260	1.6%
	Suarez	232	1.4%
	Guachené	227	1.4%
	Jambaló	217	1.3%
	Padilla	136	0.8%
	Popayán	3,854	23.5%
Centro	Piendamó	539	3.3%
	El Tambo	533	3.3%
	Cajibío	467	2.8%
	Timbío	390	2.4%
	Silvia	353	2.2%
	Morales	311	1.9%
	Puracé	178	1.1%
Oriente	Páez	382	2.3%
	Inza	380	2.3%

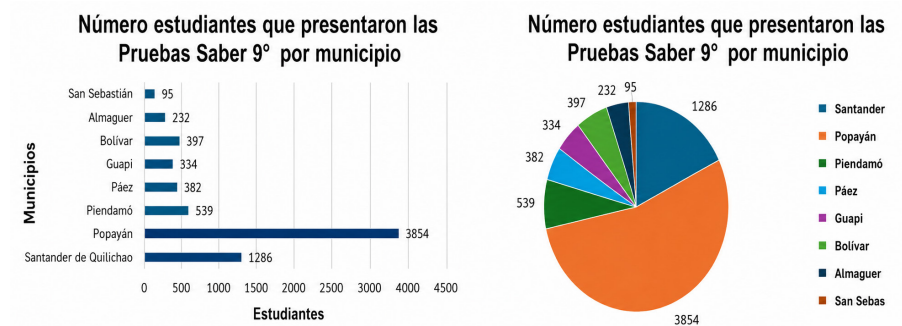
Subregión	Municipio	N	%
Occidente	Totoró	282	1.7%
	Guapi	334	2.0%
	Timbiquí	244	1.5%
	López de Micay	199	1.2%
	Bolívar	397	2.4%
	Patía	375	2.3%
	Balboa	224	1.4%
	Argelia	200	1.2%
	Mercaderes	172	1.0%
	Sucre	80	0.5%
	Florencia	63	0.4%
Macizo	Almaguer	232	1.4%
	La Vega	201	1.2%
	Sotará	143	0.9%
	Rosas	134	0,8%
	La Sierra	112	0.7%
	San Sebastián	95	0.6%
	Santa Rosa	61	0.4%
	Piamonte	48	0.3%
Bota Caucana			
Totales		16,392	100.0%

Fuente. Esta Investigación

El análisis de los resultados, muestra, que los municipios con mayor número de estudiantes que presentaron la prueba, organizados por subregión, fueron: Santander de Quilichao en la subregión Norte, Popayán en la subregión de Popayán, Piendamó en la subregión Centro, Páez en la subregión Oriente, Guapi en la subregión Occidente, Bolívar en la subregión Sur, Almaguer en la subregión Macizo y San Sebastián en la subregión Bota Caucana.

Asimismo, en la figura 3 se muestra los tres municipios con mayor porcentaje de estudiantes respecto al total de 16,392 participantes.

Figura 3. Municipios con Mayor Número de Estudiantes que Presentaron las Pruebas Saber 9º, en cada Subregión del Departamento del Cauca.



Fuente. Esta Investigación

Según se observa en la figura 3, la participación en las pruebas Saber 9º se concentró principalmente en tres municipios, los cuales agrupan una porción significativa de los 16,392 participantes totales. Popayán lidera la muestra con el 23.5%, seguido por Santander de Quilichao con un 7.8% y Puerto Tejada con el 4.0%. Estas cifras subrayan la relevancia demográfica de estas localidades en el censo educativo del departamento del Cauca.

2.1.3 Niveles de Desempeño Académico de los Estudiantes Caucaños en la Prueba de Lenguaje

La tabla 5, muestra los niveles de desempeño académico en la Prueba de Lenguaje de los estudiantes del departamento del Cauca, que presentaron el examen Saber 9º en el año 2017.

Tabla 5. Niveles de Desempeño Académico de los Estudiantes Caucanos en la Prueba de Lenguaje

Variables Académicas		N	%
Nivel Desempeño	Avanzado	507	3.1%
	Satisfactorio	5,151	31.4%
	Mínimo	8,015	48.9%
	Insuficiente	2,719	16.6%
Clase Desempeño	Sobre la media nacional	5,775	35.2%
	Bajo la media nacional	10,617	64.8%

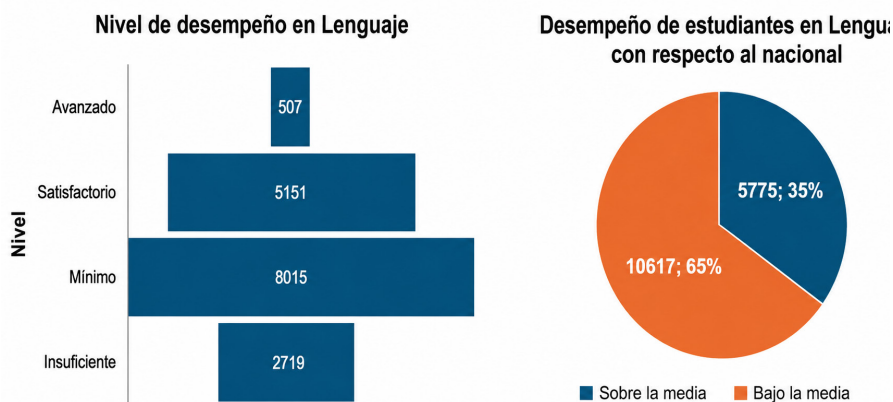
Fuente. Esta Investigación

Según el análisis de los resultados obtenidos por los estudiantes de Grado 9º del departamento del Cauca, en la competencia de Lenguaje, reflejados en la tabla 5, se observa, que la mayoría (48.9%) alcanzó un nivel de desempeño mínimo. Esto indica, que lograron responder correctamente las preguntas de menor complejidad, demostrando un conocimiento básico en esta área. Por otro lado, un 31.4% de los estudiantes obtuvo un nivel de desempeño satisfactorio, evidenciando, el dominio de las competencias esperadas para su nivel educativo y una adecuada comprensión y aplicación de los conocimientos evaluados. En contraste, apenas un 3.1% alcanzó un nivel de desempeño avanzado, lo que significa, que estos estudiantes sobresalieron en las competencias de Lenguaje para Grado 9º, superando las expectativas establecidas. Finalmente, un 16.6% de los evaluados, presentó un nivel de desempeño insuficiente, lo que indica, que no lograron responder correctamente las preguntas más básicas de la prueba, evidenciando dificultades para alcanzar las competencias fundamentales en esta área.

Comparando el puntaje obtenido por los estudiantes caucanos en la competencia de lenguaje de las Pruebas Saber 9º, con relación a la media nacional, se observa que el 64.8% obtuvo un puntaje inferior al promedio nacional, mientras que únicamente el 35.2% logró superar la media en esta área.

La figura 4, muestra la cantidad de estudiantes del departamento del Cauca, que presentaron las Pruebas Saber 9º, categorizados según su nivel y tipo de desempeño académico en la competencia de Lenguaje.

Figura 4. Número de Estudiantes Caucanos que Presentaron la Prueba de Lenguaje



Fuente. Esta Investigación

De acuerdo con la figura 4, el desempeño en la Prueba de Lenguaje de los estudiantes del Cauca se concentra mayoritariamente en los niveles mínimo y satisfactorio. Esta tendencia se refleja en una brecha significativa respecto al promedio del país: el 65% de la población evaluada se sitúa por debajo de la media nacional, mientras que solo el 35% restante logra superar dicho estándar, lo que evidencia retos importantes en el fortalecimiento de las competencias comunicativas en el departamento.

La tabla 6 muestra, la correspondencia entre los niveles de desempeño y los puntajes obtenidos por los estudiantes del departamento del Cauca, en la Prueba de Lenguaje del Saber 9º. Asimismo, la figura 5, presenta un histograma que ilustra, la distribución de los puntajes en Lenguaje, categorizados por intervalos y el número de estudiantes en cada uno de ellos. Este gráfico incluye, métricas clave, como el puntaje mínimo y máximo, la media y la desviación estándar, proporcionando una visión más detallada del rendimiento académico en esta competencia.

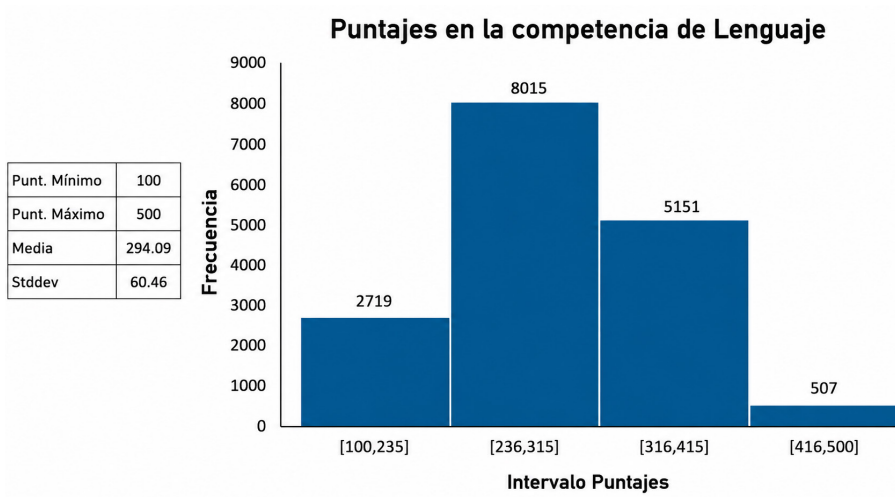
Tabla 6. Equivalencia entre Nivel de Desempeño y Puntaje Obtenido por los Estudiantes Caucanos en la Prueba de Lenguaje

Nivel Desempeño	Intervalo Puntaje
Avanzado	[416,500]
Satisfactorio	[316,415]
Mínimo	[236,315]
Insuficiente	[100,235]

Fuente. Esta Investigación

A partir de los niveles de desempeño definidos en la tabla 6 para la Prueba de Lenguaje, la clasificación de los estudiantes caucanos se establece según los siguientes rangos de puntaje: el nivel Insuficiente comprende resultados entre 100 y 235; el nivel Mínimo se sitúa entre 236 y 315; el nivel Satisfactorio abarca de 316 a 415; y, finalmente, el nivel Avanzado corresponde a puntajes entre 416 y 500.

Figura 5. Histograma de Puntajes de los Estudiantes Caucanos en la Competencia de Lenguaje



Fuente. Esta Investigación

Como se evidencia en el histograma de la figura 5, la mayor parte de la población estudiantil en el Cauca alcanzó puntajes en la Prueba de Lenguaje que oscilan entre los 236 y 415 puntos. Esta concentración de resultados sitúa a la mayoría de los evaluados en los niveles de desempeño Mínimo y Satisfactorio, reflejando una tendencia hacia el rango medio de la escala de evaluación.

2.1.4 Niveles de Desempeño Académico de los Estudiantes Caucanos en la Prueba de Matemáticas

La tabla 7, muestra el desempeño académico de los estudiantes del departamento del Cauca, en la Prueba de Matemáticas del Saber 9º, aplicada en 2017.

Tabla 7. Niveles de Desempeño Académico de los Estudiantes Caucanos en la Prueba de Matemáticas

Variables Académicas		N	%
Nivel Desempeño	Avanzado	542	3.3%
	Satisfactorio	2,287	14.0%
	Mínimo	8,929	54.5%
	Insuficiente	4,634	28.3%
Clase Desempeño	Sobre la media nacional	5,705	34.8%
	Bajo la media nacional	10,687	65.2%

Fuente. Esta Investigación

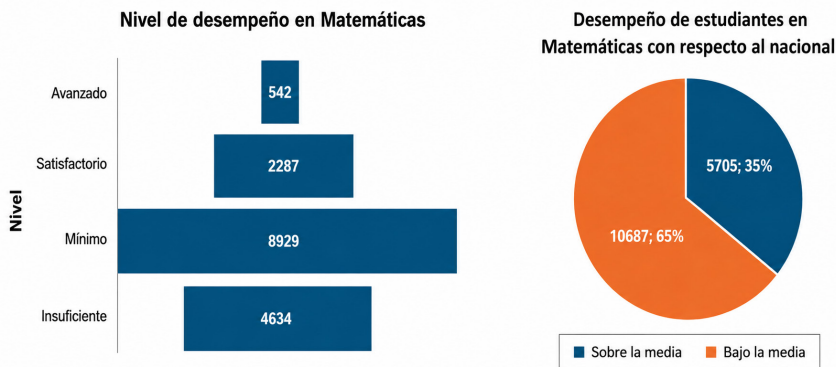
Según el análisis de los resultados, obtenidos por los estudiantes de Grado 9º del departamento del Cauca en la competencia de Matemáticas, reflejados en la tabla 7, se observa, que la mayoría (54.5%) alcanzó un nivel de desempeño mínimo, lo que indica, que lograron responder correctamente las preguntas de menor complejidad, demostrando un conocimiento básico en esta área. Por otro lado, un 14% de los estudiantes, obtuvo un nivel de desempeño satisfactorio, evidenciando el dominio de las competencias esperadas para su nivel educativo y una adecuada comprensión y aplicación de los conocimientos evaluados. En contraste,

apenas un 3.3% alcanzó un nivel de desempeño avanzado, lo que significa, que estos estudiantes sobresalieron en las competencias Matemáticas para Grado 9º, superando las expectativas establecidas. Finalmente, un 28.3% de los evaluados, presentó un nivel de desempeño insuficiente, lo que indica, que no lograron responder correctamente las preguntas más básicas de la prueba, evidenciando dificultades, para alcanzar las competencias fundamentales en esta área.

Comparando el puntaje obtenido por los estudiantes caucanos en la competencia de Matemáticas de las Pruebas Saber 9º, con relación a la media nacional, se observa que el 65.2% obtuvo un puntaje por debajo del promedio nacional, mientras que solo el 34.8% logró superar la media en esta área.

En la figura 6 se puede observar, el número de estudiantes del departamento del Cauca que presentaron las Pruebas Saber 9º, discriminado por nivel y clase de desempeño académico en la competencia de Matemáticas.

Figura 6. Número de estudiantes caucanos que presentaron la prueba de matemáticas



Fuente. Esta investigación

De acuerdo con la figura 6, el desempeño en la Prueba de Matemáticas de los estudiantes del Cauca se concentra mayoritariamente en los niveles Mínimo e Insuficiente. Esta tendencia se refleja en una brecha

significativa respecto al promedio del país: el 65% de la población evaluada se sitúa por debajo de la media nacional, mientras que solo el 35% restante logra superar dicho estándar, lo que evidencia retos importantes en el fortalecimiento de la competencia en Matemáticas en el departamento del Cauca.

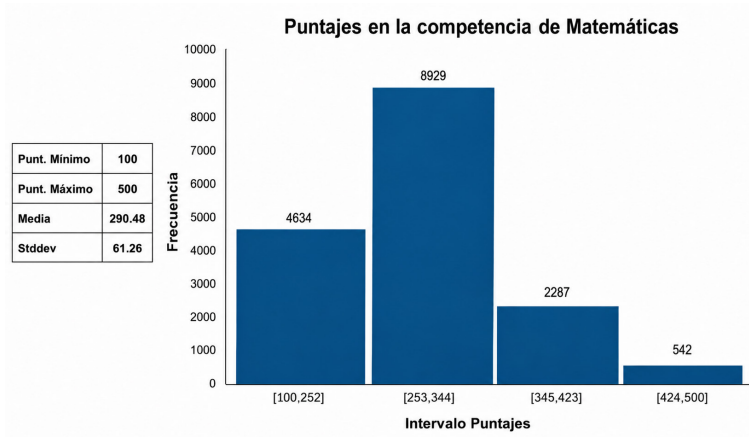
La tabla 8 muestra, la relación entre los niveles de desempeño y los puntajes obtenidos por los estudiantes del departamento del Cauca en la Prueba de Matemáticas del Saber 9º. Asimismo, la figura 7, presenta un histograma, que ilustra la distribución de los puntajes en Matemáticas, clasificados por intervalos. Este gráfico incorpora métricas clave, como el puntaje mínimo y máximo, la media y la desviación estándar, ofreciendo un análisis detallado del rendimiento académico en esta área.

Tabla 8. Equivalencia Entre Nivel de Desempeño y Puntaje Obtenido en la Prueba de Matemáticas

Nivel desempeño	Intervalo puntaje
Avanzado	[424,500]
Satisfactorio	[345,423]
Mínimo	[253,344]
Insuficiente	[100,252]

Fuente. Esta Investigación

Figura 7. Histograma de Puntajes de los Estudiantes Caucanos en la Competencia de Matemáticas



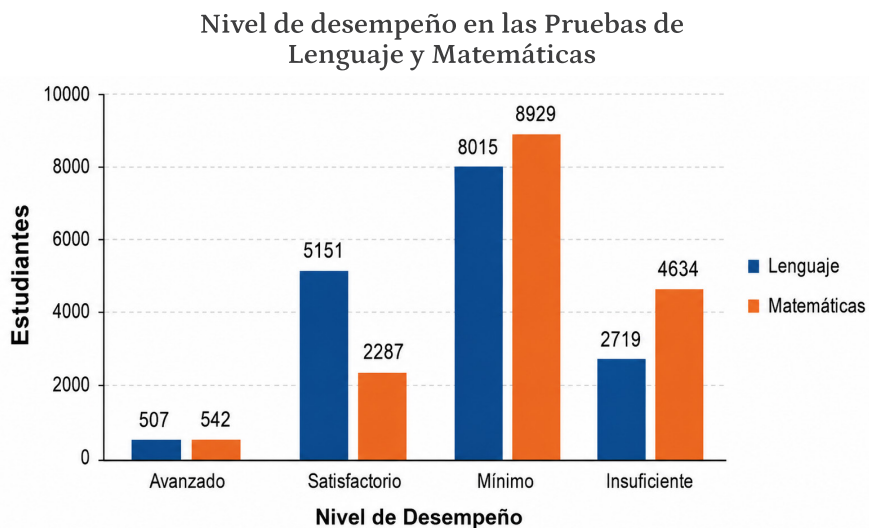
Fuente. Esta Investigación

A partir de los niveles de desempeño definidos en la tabla 8 para la Prueba de Matemáticas, la clasificación de los estudiantes caucanos se establece según los siguientes rangos de puntaje: el nivel Insuficiente comprende resultados entre 100 y 252; el nivel Mínimo se sitúa entre 253 y 344; el nivel Satisfactorio abarca de 345 a 423; y, finalmente, el nivel Avanzado corresponde a puntajes entre 424 y 500.

Como se evidencia en el histograma de la figura 7, la mayor parte de la población estudiantil en el Cauca alcanzó puntajes en la Prueba de Matemáticas que oscilan entre los 100 y 344 puntos. Esta concentración de resultados sitúa a la mayoría de los evaluados en los niveles de desempeño Mínimo e Insuficiente, reflejando una tendencia hacia el rango bajo de la escala de evaluación.

En la figura 8 se presenta, un comparativo entre los niveles de desempeño en las Pruebas de Lenguaje y Matemáticas que presentaron los estudiantes de 9º Grado del departamento del Cauca en las Pruebas Saber 9º.

Figura 8. Desempeño de los Estudiantes Caucanos en las Pruebas de Lenguaje y Matemáticas



Fuente. Esta investigación

El análisis de los resultados en las Pruebas de Lenguaje y Matemáticas, que se muestran en la figura 8, revela que el número de estudiantes con desempeño Insuficiente es mayor en Matemáticas que en Lenguaje. En cuanto a los niveles Mínimo y Avanzado, la diferencia entre ambas pruebas no es significativa. Sin embargo, en el nivel Satisfactorio, la cantidad de estudiantes que alcanzaron esta categoría, es superior en Lenguaje que en Matemáticas. En términos generales, los estudiantes de Grado 9º del departamento del Cauca, obtuvieron mejores resultados en la Prueba de Lenguaje que en la de Matemáticas en las Pruebas Saber 9º aplicadas en 2017.

2.2 CARACTERIZACIÓN DE LOS ESTUDIANTES DEL DEPARTAMENTO DE NARIÑO

En 2017, un total de 18,971 estudiantes del departamento de Nariño, presentaron las Pruebas Saber 9º. Para el análisis de los resultados, se consideraron 19 variables, distribuidas en tres categorías: 9 relacionadas

con factores socioeconómicos, 6 con factores institucionales y 4 con factores académicos.

2.2.1 Características Socioeconómicas de los Estudiantes Nariñenses

La tabla 9 describe las características socioeconómicas de los estudiantes del departamento de Nariño que presentaron las pruebas Saber 9° en 2017. Por su parte, la figura 9 ilustra algunas de estas variables, organizando los datos según sexo, rango de edades, acceso a TIC y el Índice de Nivel Socioeconómico (INSE).

Tabla 9. Características Socioeconómicas de los Estudiantes Nariñenses que Presentaron las Pruebas Saber 9°

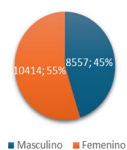
Variable Socioeconómica		N	%
Sexo	Masculino	8,557	45.1%
	Femenino	10,414	54.9%
Edad	13 años o menos	494	2.6%
	14 años	4,469	23.6%
	15 años	7,393	39.0%
	16 años o mas	6,615	34.9%
Educación del Padre	Primaria	9,764	51.5%
	Bachillerato	5,788	30.5%
	Técnico o Tecnólogo	1,234	6.5%
	Universitario – Pregrado	959	5.1%
	Universitario – Posgrado	1,226	6.5%
Educación de la Madre	Primaria	8,994	47.4%
	Bachillerato	5,945	31.3%
	Técnico o Tecnólogo	1,249	6.6%
	Universitario – Pregrado	1,622	8.5%
	Universitario – Posgrado	1,161	6.1%

Variable Socioeconómica		N	%
Hacinamiento	Hacinamiento crítico	68	0.4%
	Hacinamiento medio	1,155	6.1%
	Sin hacinamiento	17,748	93.6%
TIC	Malo	3,656	19.3%
	Regular	9,596	50.6%
	Bueno	5,719	30.1%
Electrodomésticos	Malo	7,302	38.5%
	Regular	6,356	33.5%
	Bueno	5,313	28.0%
Reprobado	No	12,005	63.3%
	Si	6,966	36.7%
Cuartil INSE	Q1	4,950	26.1%
	Q2	4,769	25.1%
	Q3	4,551	24.0%
	Q4	4,701	24.8%

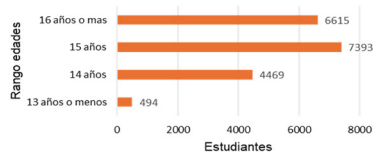
Fuente. Esta Investigación

Figura 9. Caracterización de los Estudiantes Nariñenses que Presentaron las Pruebas Saber 9º

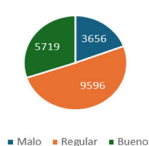
Distribución de los estudiantes por sexo



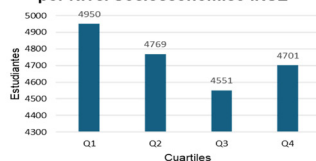
Distribución de los estudiantes por edad



Distribución de los estudiantes por Índice TICs



Distribución de los estudiantes por Nivel Socioeconómico INSE



Fuente. Esta Investigación

El análisis de los datos de la tabla 9 y figura 9, muestra que, en 2017, del total de 18,971 estudiantes de Grado 9º del departamento de Nariño que presentaron las Pruebas Saber 9º, la mayoría fueron mujeres, representando el 54.9% frente al 45.1% de hombres, lo que indica, una diferencia del 9.8%.

En cuanto a la edad, el grupo predominante, tiene entre 14 y 15 años, lo que equivale al 62.6% del total de estudiantes. El 37.4% restante se divide entre aquellos de 16 años o más (34.9%) y los que tienen 13 años o menos (2.6%).

Respecto al nivel educativo de los padres, el porcentaje de aquellos con estudios entre primaria y secundaria, es ligeramente superior al de las madres (82% frente a 78.7%). Sin embargo, en Educación Superior (tecnológica, pregrado y posgrado), el porcentaje de madres es mayor, con un 21.2% frente al 18.1% de los padres.

Además, solo el 6.5% de los estudiantes vive en condiciones de hacinamiento. En términos de acceso tecnológico, el 30.1% considera tener un buen nivel de TIC (internet y televisión), mientras que el 28%, reporta un nivel adecuado de electrodomésticos (estufa, microondas y nevera).

Por otro lado, el 36.7% de los estudiantes que presentaron la prueba, reprobaron algún grado. Finalmente, en cuanto al Índice de Nivel Socioeconómico (INSE), el 51.2% se encuentra en los cuartiles 1 y 2, indicando condiciones favorables, mientras que el 48.8% pertenece a los cuartiles 3 y 4, reflejando condiciones menos favorables.

2.2.2 Características Institucionales de los Colegios del Departamento de Nariño

En la tabla 10 se presentan, las características institucionales de los colegios del departamento de Nariño, a los cuales pertenecen los estudiantes que presentaron las Pruebas Saber 9º.

De acuerdo con los datos de la tabla 10, la mayoría de los estudiantes nariñenses que presentaron las Pruebas Saber 9º, en 2017, provenían de

instituciones mixtas (95.7%) y oficiales (92.6%), con un enfoque académico (55.5%). Además, el 69.3% de estos colegios, estaban ubicados en zonas urbanas de diversos municipios del departamento de Nariño, con jornadas escolares diurnas (84.2%).

Tabla 10. Características Institucionales de los Colegios de los Estudiantes Nariñenses que Presentaron las Pruebas Saber 9º

Variables institucionales		N	%
Tipo de Colegio	Femenino	811	4.3%
	Mixto	18,160	95.7%
Régimen	Oficial	17,571	92.6%
	No oficial	1,400	7.4%
Carácter	Académico	10,535	55.5%
	Técnico	2,602	13.7%
	Técnico/Académico	5,834	30.8%
Área ubicación	Urbano	13,150	69.3%
	Rural	5,821	30.7%
Jornada	Mañana	15,965	84.2%
	Tarde	2,116	11.2%
	Sabatina	12	0,1%
	Completa	119	0.6%
		759	4.0%
Subregión	Rio Mayo	1,011	5.3%
	Pasto	5,106	26.9%
	Obando	3,633	19.2%
	Occidente	617	3.3%
	La Cordillera	726	3.8%
	Pacífico Sur	2,584	13.6%
	Los Abades	440	2.3%
	Telembí	627	3.3%

Variables institucionales		N	%
Subregión	La Sabana	931	4.9%
	Guambuyaco	456	2.4%
	Centro	686	3.6%
	Sanquianga	816	4.3%
	Juanambú	1,048	5.5%
	Piedemonte Costero	290	1.5%

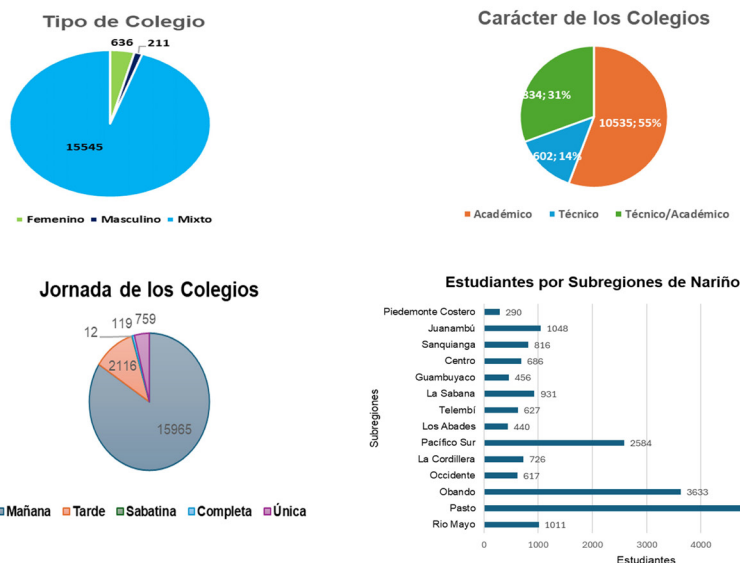
Fuente. Esta Investigación

En cuanto a la distribución geográfica, la subregión de Pasto, capital del departamento, concentró el mayor porcentaje de estudiantes (26.9%), seguida por la subregión de Obando (19.2%), que abarca los municipios de Aldana, Contadero, Córdoba, Cuaspud, Cumbal, Funes, Iles, Ipialles, Guachucal, Gualmatán, Potosí, Pupiales y Puerres. Por su parte, la subregión del Pacífico Sur, conformada por los municipios de Tumaco y Francisco Pizarro, representó el 13.6% del total de estudiantes.

En la figura 10, se puede observar ciertas características de los colegios del departamento de Nariño, cuyos estudiantes presentaron las Pruebas Saber 9º.

Como se puede observar en la figura 10, la gran mayoría de estudiantes nariñenses que presentaron las pruebas Saber 9º en el año 2017, pertenecen a colegios de tipo mixto, de carácter académico, de jornada de la mañana y son de las subregiones de Pasto, Obando y Pacífico Sur del departamento.

Figura 10. Características de los Colegios del Departamento de Nariño con Número de Estudiantes que Presentaron las Pruebas Saber 9º.



Fuente. Esta Investigación

La tabla 11 muestra, la distribución de estudiantes por municipio en el departamento de Nariño, pertenecientes a instituciones de Educación Básica Secundaria, cuyos alumnos presentaron las Pruebas Saber 9º, durante el periodo analizado en esta investigación.

Tabla 11. Número de Estudiantes de los Municipios del Departamento de Nariño que Presentaron las Pruebas Saber 9º

Subregión	Municipio	N	%
Rio Mayo	La Cruz	236	1.2%
	El Tablón de Gómez	185	1.0%
	San Pablo	177	0.9%
	San Bernardo	123	0.6%
	Alban	119	0.6%
	Colon	99	0.5%
	Belén	72	0.4%

Subregión	Municipio	N	%
Pasto	San Juan de Pasto	5,106	26.9%
Obando	Ipiales	1,607	8.5%
	Cumbal	497	2.6%
	Guachucal	258	1.4%
	Pupiales	253	1.3%
	Córdoba	187	1.0%
	Iles	135	0.7%
	Potosí	133	0.7%
	Aldana	125	0.7%
	Funes	102	0.5%
	Puerres	101	0.5%
	Cuaspud	90	0.5%
	Gualmatán	85	0.4%
	Contadero	60	0.3%
Occidente	Sandoná	237	1.2%
	Consacá	165	0.9%
	Linares	115	0.6%
	Ancuya	100	0.5%
La Cordillera	Taminango	259	1.4%
	Leiva	139	0.7%
	Policarpa	127	0.7%
	Cumbitara	104	0.5%
	El Rosario	97	0.5%
Pacífico Sur	San Andrés de Tumaco	2,506	13.2%
	Francisco Pizarro	78	0.4%
Los Abades	Samaniego	265	1.4%
	Santacruz	99	0.5%
	Providencia	76	0.4%

Subregión	Municipio	N	%
Telembí	Barbacoas	416	2.2%
	Roberto Payan	125	0.7%
	Magüi	86	0.5%
La Sabana	Túquerres	513	2.7%
	Guaitarilla	149	0.8%
	Imués	91	0.5%
	Ospina	91	0.5%
	Sapuyes	87	0.5%
Guambuyaco	El Tambo	161	0.8%
	Los Andes	140	0.7%
	El Peñol	90	0.5%
	La Llanada	65	0.3%
Centro	Chachagüí	212	1.15%
	Yacuanquer	179	0,9%
	Tangua	126	0,7%
	La Florida	102	0,5%
	Nariño	67	0,4%
Sanquianga	Olaya Herrera	303	1,6%
	El Charco	207	1,1%
	La Tola	114	0,6%
	Santa Barbara	101	0,5%
	Mosquera	91	0,5%
Juanambú	La Unión	358	1,9%
	San Lorenzo	272	1,4%
	Buesaco	216	1,1%
	Arboleda	117	0,6%
	San Pedro de Cartago	85	0,4%
Piedemonte Costero	Ricaurte	190	1,0%
	Mallama	100	0,5%
Totales		18,971	100.0%

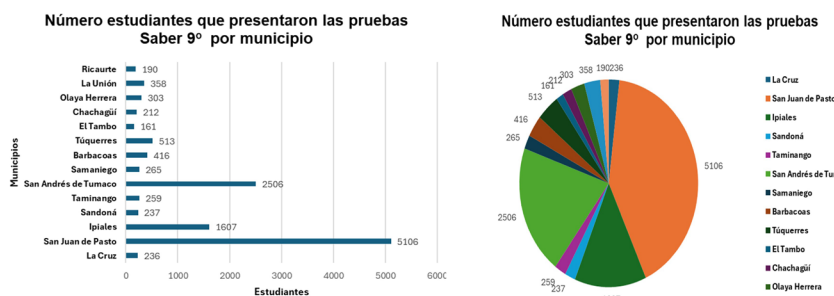
Fuente. Esta Investigación

El análisis de los resultados de la tabla 11 indica, que los municipios del departamento de Nariño con mayor número de estudiantes que presentaron las Pruebas Saber 9º, clasificados por subregión, son: La Cruz (Río Mayo), San Juan de Pasto (Pasto), Ipiales (Obando), Sandoná (Occidente), Taminango (La Cordillera), San Andrés de Tumaco (Pacífico Sur), Samaniego (Los Abades), Barbacoas (Telembí), Túquerres (La Sabana), El Tambo (Guambuyaco), Chachagüí (Centro), Olaya Herrera (Sanquianga), La Unión (Juanambú) y Ricaurte (Pedemonte Costero).

Los tres municipios con mayor participación de estudiantes en las Pruebas Saber 9º, en relación con el total de 18,971 estudiantes evaluados en Nariño, fueron: San Juan de Pasto (26.9%), San Andrés de Tumaco (13.2%) e Ipiales (8.5%). Esta distribución se presenta de manera visual en la figura 11.

Según se observa en la figura 11, la participación en las pruebas Saber 9º se concentró principalmente en tres municipios, los cuales agrupan una porción significativa de los 18,971 participantes totales. San Juan de Pasto lidera la muestra con el 27%, seguido por San Andrés de Tumaco con un 13% e Ipiales con el 8.4%. Estas cifras subrayan la relevancia demográfica de estas localidades en el censo educativo del departamento de Nariño.

Figura 11. Municipios con Mayor Número de Estudiantes que Presentaron las Pruebas Saber 9º, en cada Subregión del Departamento de Nariño.



Fuente. Esta Investigación

2.2.3 Niveles de Desempeño Académico de los Estudiantes Nariñenses en la Prueba de Lenguaje

La tabla 12 muestra, el desempeño académico de los estudiantes del departamento de Nariño, en la Prueba de Lenguaje del Saber 9º, aplicada en 2017.

Tabla 12. Niveles de Desempeño Académico de los Estudiantes Nariñenses en la Prueba de Lenguaje

Variables Académicas		N	%
Nivel desempeño	Avanzado	894	4.7%
	Satisfactorio	7,484	39.4%
	Mínimo	8,210	43.3%
	Insuficiente	2,383	12.6%
Clase desempeño	Sobre la media nacional	8,505	44.8%
	Bajo la media nacional	10,466	55.2%

Fuente. Esta Investigación

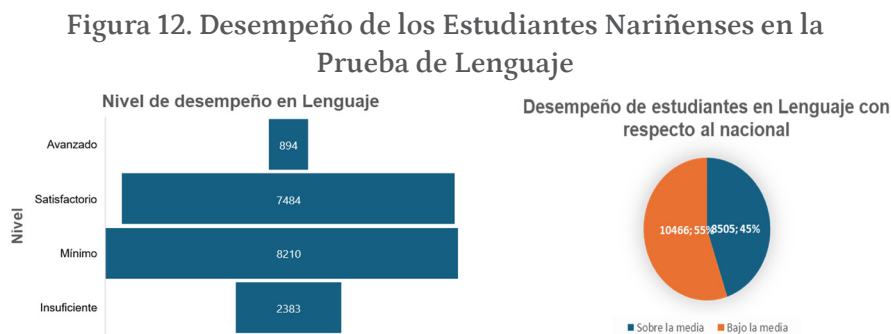
Analizando los resultados obtenidos por los estudiantes nariñenses de Grado 9º en la Competencia de Lenguaje que se muestra en la tabla 12, se puede afirmar, que la mayoría obtuvo un nivel de desempeño mínimo (43.3%), lo que significa que estos estudiantes superaron las preguntas de menor complejidad de la Prueba de Lenguaje, demostrando un entendimiento básico de esta competencia. Un menor porcentaje de estudiantes, obtuvieron un nivel de desempeño satisfactorio (39.4%), lo que significa, que demostraron las competencias esperadas para su nivel de formación, mostrando una comprensión y aplicación adecuada, de los conocimientos y habilidades evaluados. Únicamente, el 4.7% obtuvo un nivel de desempeño avanzado en la Competencia de Lenguaje, lo que significa, que este grupo de estudiantes, demostraron un desempeño sobresaliente en las competencias esperadas para el área de lenguaje del Grado 9º, superando las expectativas generales. Un 12.6% de los estudiantes evaluados, tienen un nivel de desempeño en Lenguaje de insuficiente, lo que significa, que este grupo de estudiantes, no lograron responder correctamente las

preguntas más sencillas de la prueba, indicando, que no han alcanzado las competencias básicas esperadas en esta área.

Comparando el puntaje obtenido por los estudiantes nariñenses en la Competencia de Lenguaje de las Pruebas Saber 9º, el 55.2% obtuvo un puntaje menor que el puntaje promedio nacional y el 44.8% está sobre la media nacional en esta competencia.

En la figura 12 se puede observar, el desempeño en la Prueba de Lenguaje de los estudiantes del departamento de Nariño, que presentaron las Pruebas Saber 9º, de acuerdo con el nivel y clase de desempeño.

De acuerdo con la figura 12, el desempeño en la Prueba de Lenguaje de los estudiantes del departamento de Nariño se concentra mayoritariamente en los niveles mínimo y satisfactorio. Esta tendencia se refleja en una brecha significativa respecto al promedio del país: el 55% de la población evaluada se sitúa por debajo de la media nacional, mientras que solo el 45% restante logra superar dicho estándar, lo que evidencia retos importantes en el fortalecimiento de las competencias comunicativas en el departamento.



Fuente. Esta Investigación

En la tabla 13 se presenta, la equivalencia del nivel de desempeño con relación al puntaje obtenido en la Prueba de Lenguaje, que presentaron los estudiantes del departamento de Nariño, en el Saber 9º. Además,

en la figura 13, se muestra el histograma de los puntajes obtenidos en lenguaje discriminado por intervalos de puntaje y acompañado con métricas como el puntaje mínimo, máximo, media y desviación estándar.

Tabla 13. Equivalencia entre Nivel de desempeño y puntaje obtenido en la Prueba de Lenguaje

Nivel desempeño	Intervalo puntaje
Avanzado	[416,500]
Satisfactorio	[316,415]
Mínimo	[236,315]
Insuficiente	[100,235]

Fuente. Esta investigación

Figura 13. Histograma de Puntajes en la Competencia de Lenguaje



Fuente. Esta Investigación

A partir de los niveles de desempeño definidos en la tabla 13 para la Prueba de Lenguaje, la clasificación de los estudiantes nariñenses se establece según los siguientes rangos de puntaje: el nivel Insuficiente comprende resultados entre 100 y 235; el nivel Mínimo se sitúa entre 236 y 315; el nivel Satisfactorio abarca de 316 a 415; y, finalmente, el nivel Avanzado corresponde a puntajes entre 416 y 500.

Como se evidencia en el histograma de la figura 13, la mayor parte de la población estudiantil en Nariño alcanzó puntajes en la Prueba de Lenguaje que oscilan entre los 236 y 415 puntos. Esta concentración

de resultados sitúa a la mayoría de los evaluados en los niveles de desempeño Mínimo y Satisfactorio, reflejando una tendencia hacia el rango medio de la escala de evaluación.

2.2.4 Niveles de Desempeño Académico de los Estudiantes Nariñenses en la Prueba de Matemáticas

La tabla 14 muestra, el desempeño académico de los estudiantes del departamento de Nariño, en la Prueba de Matemáticas del Saber 9º, aplicada en 2017.

Tabla 14. Niveles de Desempeño Académico de los Estudiantes Nariñenses en la Prueba de Matemáticas

Variables Académicas		N	%
Nivel desempeño	Avanzado	937	4.9%
	Satisfactorio	4,022	21.2%
	Mínimo	9,923	52.3%
	Insuficiente	4,089	21.6%
Clase desempeño	Sobre la media nacional	8,730	46.0%
	Bajo la media nacional	10,241	54.0%

Fuente. Esta investigación

Al analizar los resultados de los estudiantes nariñenses de Grado 9º en la Competencia de Matemáticas, según lo mostrado en la tabla 14, se observa que la mayoría obtuvo un nivel de desempeño Mínimo (52.3%). Esto indica, que lograron responder correctamente las preguntas de menor complejidad en la Prueba de Matemáticas, demostrando un entendimiento básico de esta competencia.

Por otro lado, un porcentaje menor de estudiantes (21.2%) alcanzó un nivel de desempeño Satisfactorio, lo que evidencia, que han desarrollado las competencias esperadas para su nivel de formación, mostrando una comprensión adecuada y una correcta aplicación de los conocimientos y habilidades evaluados.

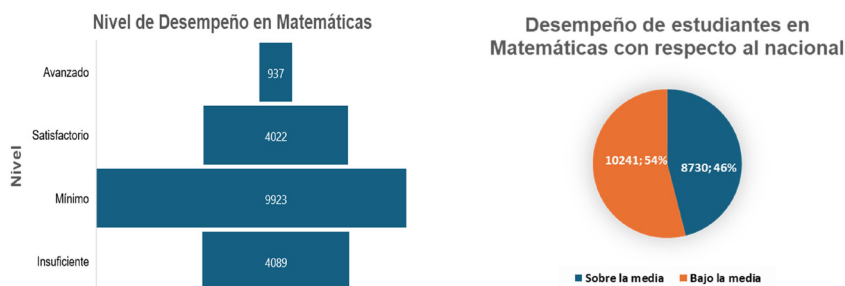
En contraste, solo el 4.9% logró un nivel de desempeño Avanzado en la Competencia de Matemáticas, lo que demuestra, que este grupo de estudiantes tuvo un rendimiento sobresaliente, en los aspectos evaluados para el área de Matemáticas en Grado 9º, superando las expectativas generales.

Finalmente, un 21.6% de los estudiantes evaluados presentó un desempeño Insuficiente en Matemáticas. Esto indica, que no lograron responder correctamente las preguntas más sencillas de la prueba, lo que evidencia, que aún no han alcanzado las competencias básicas esperadas en esta área.

Al comparar los resultados obtenidos por los estudiantes nariñenses en la Competencia de Matemáticas de las Pruebas Saber 9º, con la media nacional, se observa que el 54.0% obtuvo un puntaje inferior al promedio nacional, mientras que el 46.0% superó esta media.

La figura 14, muestra el número de estudiantes del departamento de Nariño que presentaron la Prueba de Matemáticas, según el nivel y clase de desempeño.

Figura 14. Número de Estudiantes Nariñenses que Presentaron la Prueba de Matemáticas Según Nivel y Clase de Desempeño



Fuente. Esta Investigación

De acuerdo con la figura 14, el desempeño en la Prueba de Matemáticas de los estudiantes de Nariño se concentra mayoritariamente en el nivel Mínimo (52.3%) seguido del nivel Insuficiente (21.5%) y Satisfactorio (21.2%). Únicamente el 4.9% se encuentra en el nivel Avanzado. Esta

tendencia se refleja en una brecha significativa respecto al promedio del país: el 54% de la población evaluada se sitúa por debajo de la media nacional, mientras que el 46% restante logra superar dicho estándar, lo que evidencia retos importantes en el fortalecimiento de la competencia en Matemáticas en el departamento de Nariño.

La tabla 15 muestra, la equivalencia entre los niveles de desempeño y los puntajes obtenidos en la Prueba de Matemáticas, por los estudiantes del departamento de Nariño, en el Saber 9º.

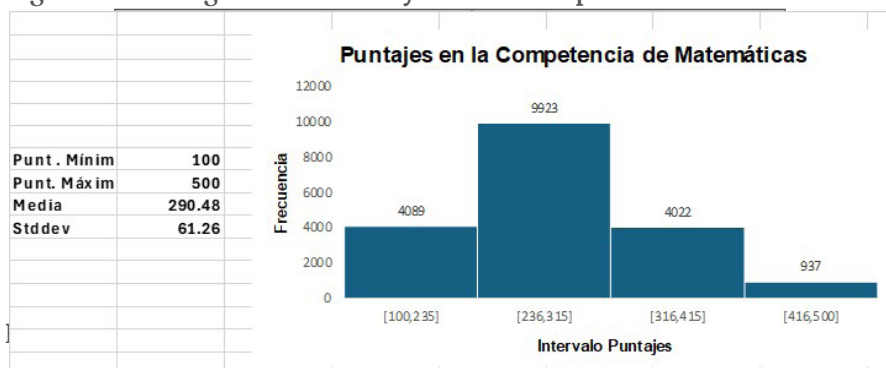
Asimismo, la figura 15, presenta un histograma, que distribuye los puntajes en Matemáticas por intervalos, acompañado de métricas clave, como el puntaje mínimo, máximo, promedio y la desviación estándar.

Tabla 15. Equivalencia entre Nivel de Desempeño y Puntaje Obtenido en la Prueba de Matemáticas

Nivel Desempeño	Intervalo Puntaje
Avanzado	[424,500]
Satisfactorio	[345,423]
Mínimo	[253,344]
Insuficiente	[100,252]

Fuente. Esta Investigación

Figura 15. Histograma de Puntajes en la Competencia de Matemáticas



Fuente. Esta Investigación

A partir de los niveles de desempeño definidos en la tabla 15 para la Prueba de Matemáticas, la clasificación de los estudiantes nariñenses se establece según los siguientes rangos de puntaje: el nivel Insuficiente comprende resultados entre 100 y 252; el nivel Mínimo se sitúa entre 253 y 344; el nivel Satisfactorio abarca de 345 a 423; y, finalmente, el nivel Avanzado corresponde a puntajes entre 424 y 500.

Como se evidencia en el histograma de la figura 15, la mayor parte de la población estudiantil en el departamento de Nariño alcanzó puntajes en la Prueba de Matemáticas que oscilan entre los 100 y 315 puntos. Esta concentración de resultados sitúa a la mayoría de los evaluados en los niveles de desempeño Mínimo e Insuficiente (74%), reflejando una tendencia hacia el rango bajo de la escala de evaluación.

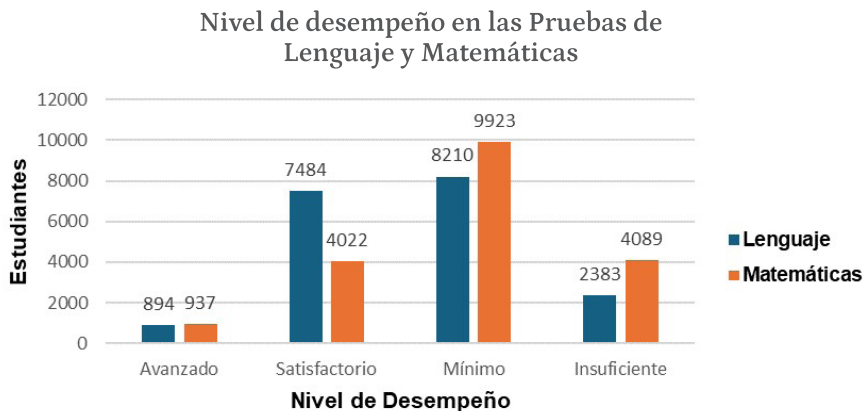
La figura 16 muestra, una comparación entre los niveles de desempeño alcanzados por los estudiantes de 9º Grado del departamento de Nariño en las Pruebas Saber 9º, tanto en Lenguaje como en Matemáticas.

El análisis de los resultados obtenidos en las pruebas de Lenguaje y Matemáticas, representados en la figura 16, revela, que el número de estudiantes con desempeño insuficiente, es mayor en Matemáticas que en Lenguaje.

En cuanto al nivel de desempeño avanzado, la diferencia entre ambas pruebas, es mínima. Sin embargo, en el nivel satisfactorio, el número de estudiantes que alcanzaron esta categoría, es mayor en Lenguaje que en Matemáticas. Por el contrario, en el nivel mínimo ocurre lo opuesto, siendo más elevado, el número de estudiantes con esta calificación en Matemáticas que en Lenguaje.

En términos generales, los estudiantes de Grado 9º del departamento de Nariño, obtuvieron mejores resultados en la Prueba de Lenguaje que en la de Matemáticas en las Pruebas Saber 9º realizadas en 2017.

Figura 16. Desempeño de los Estudiantes Nariñenses en las Pruebas de Lenguaje y Matemáticas



Fuente. Esta Investigación

El análisis de los resultados en las Pruebas de Lenguaje y Matemáticas, que se muestran en la figura 16, revela que el número de estudiantes con desempeños Insuficiente y Mínimo es mayor en Matemáticas que en Lenguaje. En cuanto al nivel Satisfactorio, el número de estudiantes con un este desempeño es mayor en Lenguaje que en Matemáticas. El nivel Avanzado, la diferencia entre ambas pruebas no es significativa. En términos generales, los estudiantes de Grado 9º del departamento de Nariño obtuvieron mejores resultados en la Prueba de Lenguaje que en la de Matemáticas en las Pruebas Saber 9º aplicadas en 2017.

2.3 ANÁLISIS DEL DESEMPEÑO GENERAL EN LAS PRUEBAS SABER 9º DE LOS ESTUDIANTES DE LOS DEPARTAMENTOS DEL CAUCA Y NARIÑO

A partir del promedio de los puntajes obtenidos en las Pruebas de Lenguaje y matemáticas del Saber 9º, por los estudiantes de los departamentos de Cauca y Nariño, se determinó su calificación global, en las Pruebas Saber 9º. Para evaluar el desempeño general, se comparó, el puntaje total de cada estudiante, con la media nacional. Para aquellos, cuyo puntaje era igual o superior a la media nacional, se les asignó la categoría de ‘Sobre la

media', mientras que quienes obtuvieron un puntaje inferior, fueron clasificados como 'Bajo la media'.

En la tabla 16 se muestra, el número de estudiantes de los departamentos del Cauca y Nariño, cuyo desempeño general en las Pruebas Saber 9º, esta "Sobre la media" o "Bajo la media".

Analizando la información de la tabla 16, en el año 2017, un total de 35,363 estudiantes de los departamentos de Cauca y Nariño, presentaron las Pruebas Saber 9º. De ellos, el 46.4% pertenecía al Cauca y el 53.6% a Nariño. En cuanto al desempeño académico general, el 39% de los estudiantes de ambos departamentos, obtuvo resultados por encima de la media nacional, mientras que el 61% quedó por debajo de ella. Dentro del grupo que superó la media nacional, el 38.9% eran del Cauca y el 61.1% de Nariño. Por otro lado, entre los estudiantes con puntajes por debajo de la media, el 51.1% correspondía al Cauca y el 48.9% a Nariño.

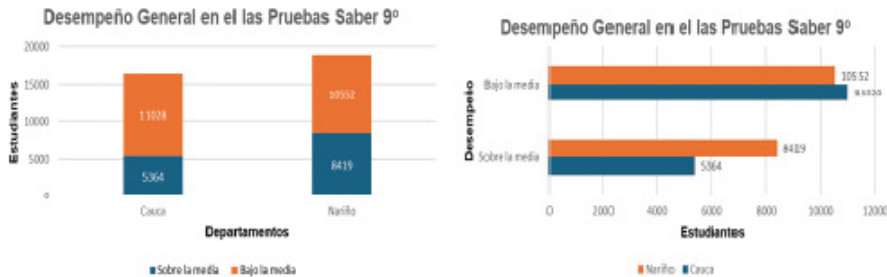
Tabla 16. Desempeño General en las Pruebas Saber 9º de los Estudiantes del Cauca y Nariño

Desempeño General			
Departamento	Sobre la Media	Bajo la Media	Totales
Cauca	5,364 (38.9%)	11,028 (51.1%)	16,392 (46.4%)
Nariño	8,419 (61.1%)	10,552 (48.9%)	18,971 (53.6%)
Totales	13,783 (39%)	21,580 (61%)	35,363 (100%)

Fuente. Esta Investigación

En la figura 17 se muestra, la relación de desempeño general de los estudiantes de los departamentos del Cauca y Nariño en las Pruebas Saber 9º.

Figura 17. Desempeño General de los Estudiantes Caucaños y Nariñenses en las Pruebas Saber 9º



Fuente. Esta Investigación

Como se muestra en la tabla 16 y la figura 17, aunque el número de estudiantes de Nariño que presentaron las Pruebas Saber 9º fue superior al de Cauca, el porcentaje de aquellos con un Desempeño General por debajo de la media nacional (48.9%) resultó menor en comparación con el de Cauca (51.1%). Esto indica, que, en términos de Desempeño General en las Pruebas Saber 9º del año 2017, los estudiantes de Nariño, tuvieron mejores resultados que los de Cauca.

CAPÍTULO III MATERIALES Y MÉTODOS

Los estudios descriptivos, se centran, en analizar la distribución de variables, sin abordar hipótesis de causalidad u otras interpretaciones. Por esta razón, esta investigación se enmarcó, en el enfoque cuantitativo y corresponde a un estudio descriptivo con un diseño no experimental.

3.1 MATERIALES

Para esta investigación, se emplearon las últimas actualizaciones disponibles en el momento de su realización, obtenidas del repositorio de datos abiertos del ICFES, conocido como Data Icfes (ICFES, 2024a). La información correspondía a los resultados de las Pruebas Saber 9º, del año 2017, en las que participaron 589,876 estudiantes colombianos de Grado 9º de Educación Básica Secundaria. De este total, el 2.8% (16,392 estudiantes) pertenecía al departamento del Cauca, mientras que el 3.2% (18,971 estudiantes), correspondía al departamento de Nariño.

Los datos socioeconómicos, institucionales y académicos de los estudiantes que presentaron las Pruebas Saber 9º, almacenados en el repositorio Data Icfes, fueron importados a una base de datos, diseñada con el Sistema Gestor de Bases de Datos (SGBD) PostgreSQL, denominada *Saber9db*. La tabla principal, *sbr9_colombia*, contenía 589,876 registros, distribuidos en 56 atributos. El diccionario de datos correspondiente a este conjunto de datos, se presenta en la tabla 17.

A partir del conjunto de datos *sbr9_colombia*, se seleccionaron inicialmente, los registros correspondientes a los estudiantes de Cauca y Nariño, formando el conjunto *sbr9_cauca_narino*, objeto de esta investigación, con 35,363 registros y 56 atributos. Posteriormente, la información se organizó en dos subconjuntos: uno con los datos de los estudiantes del Cauca y otro con los de Nariño. Estos fueron almacenados en las tablas *sbr9_cauca*, con 16,392 registros y 56 atributos y *sbr9_narino*, con 18,971 registros y 56 atributos.

Tabla 17. Diccionario de Datos del Conjunto Sbr9_Colombia

N	Atributo	Descripción	Valores	
			Información Personal	
1	ESTU_TIPODOCUMENTO	Tipo de Documento	CC – Cédula de ciudadanía CCB- Certificado de cabildo CE – Cédula extranjera CR – Certificado registraduría NES-Número establecido por secretaria de educación NUJP – Número único de identificación personal NUI – Número único de identificación PC – Pasaporte colombiano PE – Pasaporte extranjero RC – Registro civil de nacimiento TI – Tarjeta de identidad	
2	ESTU_GENERO	Género	Hombre Mujer	
3	ESTU_EDAD	Edad del Estudiante	13 años o menos 14 años 15 años 16 años o mas	
4	ESTU_GRADO	Grado del Estudiante	9o	
5	ESTU_FECHANACIMIENTO	Fecha de Nacimiento	DD/MM/AAAA	
6	PERIODO	Periodo de Presentación	2017	
7	ESTU_CONSECUTIVO	Id público del Estudiante	Texto	
8	ESTU_ESTUDIANTE	Indica si realizó la inscripción por medio de un colegio (estudiante) o fue individual	ESTUDIANTE	

N	Atributo	Descripción	Valores
Información Socioeconómica			
9	FAMI_EDUCACIONPADRE	Nivel Educativo más alto alcanzado por el Padre	Primaria Bachillerato Técnico o Tecnólogo Universitario - Pregrado Universitario – Posgrado NA
10	FAMI_EDUCACIONMADRE	Nivel Educativo más alto alcanzado por la Madre	Primaria Bachillerato Técnico o Tecnólogo Universitario - Pregrado Universitario – Posgrado NA
11	FAMI_PERSONASHOGAR	Número de personas que conforman el hogar donde vive actualmente, incluido el estudiante	2,3,4,5,6
11	FAMI_CUARTOSHOGAR	Número de cuartos donde duermen las personas del hogar	1,2,3,4,5,6, NA
13	FAMI_TIENEINTERNET	Si el hogar cuenta con servicio o conexión a internet	No, Sí, NA
14	FAMI_TIENESERVICIOTV	Si el hogar cuenta con servicio por cable de televisión	No, Sí, NA
15	FAMI_TIENECOMPUTADOR	Si en el hogar hay o no computador	No, Sí, NA
16	FAMI_TIENELAVADORA	Si en el hogar hay o no lavadora	No, Sí, NA

N	Atributo	Descripción	Valores
17	FAMI_TIENEHORNOMICROOGAS	Si en el hogar hay o no microondas, horno eléctrico o a gas	No, Sí, NA
18	FAMI_TIENEAUTOMOVIL	Si en el hogar hay o no automóvil particular	No, Sí, NA
19	FAMI_TIENECONSOLAVIDEOJUEGOS	Si en el hogar hay o no consola de video juegos (PlayStation, Xbox, Nintendo, etc.)	No, Sí, NA
20	FAMI_TRABAJOLABORPADRE	Trabajo que realizó el padre durante la mayor parte del último año	Trabaja en el hogar; no trabaja o estudia Es dueño de un negocio pequeño (tiene pocos empleados o no tiene; por ejemplo, tienda, papelería, salón de belleza, lavandería, etc.) Es dueño de un negocio grande (empresa o institución) o tiene un cargo de nivel directivo o gerencial Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente) Es vendedor; vende por catálogo o trabaja en atención al público Es agricultor, pescador o jornalero Trabaja por cuenta propia (por ejemplo, plomero, electricista) Es operario de máquinas o conduce vehículos (taxista, chofer) Realiza labores de limpieza, mantenimiento, seguridad o construcción Trabaja como profesional (por ejemplo, médico, abogado) Es pensionado NA

N	Atributo	Descripción	Valores
21	FAMI_TRABAJO LABORMADRE	Trabajo que realizó la madre durante la mayor parte del último año	Trabaja en el hogar; no trabaja o estudia Es dueño de un negocio pequeño (tiene pocos empleados o no tiene; por ejemplo, tienda, papelería, salón de belleza, lavandería, etc.) Es dueño de un negocio grande (empresa o institución) o tiene un cargo de nivel directivo o gerencial Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente) Es vendedor, vende por catálogo o trabaja en atención al público Es agricultor, pescador o jornalero Trabaja por cuenta propia (por ejemplo, plomero, electricista) Es operario de máquinas o conduce vehículos (taxista, chofer) Realiza labores de limpieza, mantenimiento, seguridad o construcción Trabaja como profesional (por ejemplo, médico, abogado) Es pensionado NA
22	ESTU_DEDICACION INTERNET	Tiempo al día dedicado a navegar en internet	No navega en internet 30 minutos o menos Entre 30 y 60 minutos Entre 1 y 3 horas Más de 3 horas NA

N	Atributo	Descripción	Valores
23	ESTU_COMODESPLAZACOLEGIO	Modo de desplazarse al colegio	Caminando o en bicicleta En bus o transporte público En transporte escolar o particular Otro (burro o caballo, canoa, lancha, etc.) NA
24	ESTU_TIEMPODESPLAZACOLEGIO	Tiempo de desplazamiento de la casa al colegio	Menos de 15 minutos Entre 15 y 30 minutos Entre 30 minutos y una hora Más de 1 hora NA
25	ESTU_ANOSPREESCOLAR	Años cursado en Preescolar	Ninguno Un año Dos años Tres años No me acuerdo NA
26	ESTU_ESTECOLEGIO_ANOPASADO	Estudió el año pasado en el mismo colegio	No, Sí, NA
27	ESTU_GRADOCOMENZO_ESTECOLEGIO	Desde que grado estudia en el colegio actual	Primero Segundo Tercero Cuarto Quinto NA
28	ESTU_REPROBOGRADO	Si ha repetido algún grado	No, Sí, NA

N	Atributo	Descripción	Valores
29	ESTU_TARDEACLASE	Número de veces que ha llegado tarde el último mes	Nunca Una o dos veces por semana Una o dos veces por mes Todos los días NA
30	ESTU_NOFUEALCOLEGIO	Número de días que no ha ido al colegio en el último mes	Nunca Uno o dos días Tres a cinco días Más de cinco días NA
31	ESTU_REMUNERA_OTRASACTIVIDADES	Si ha recibido remuneración por actividades diferentes a las escolares	No, Sí, NA
32	ESTU_SERETIROCOLEGIO	Tiempo de retiro o suspensión de estudios	No Sí, menos de un año Sí, un año Sí, dos años o más NA
33	FAMI_NUMLIBROS	Numero de libros que hay en el hogar	0 a 10 libros 11 a 25 libros 26 a 100 libros Más de 100 libros NA
Información del Colegio			
34	COLE_COD_DANE_ESTABLECIMIENTO	Código DANE del establecimiento	Texto

N	Atributo	Descripción	Valores
35	COLE_NOMBRE_ESTABLECIMIENTO	Nombre del Establecimiento	Texto
36	COLE_GENERO	Indica el género de la población del establecimiento	FEMENINO MASCULINO MIXTO
37	COLE_NATURALEZA	Indica la naturaleza del establecimiento	NO OFICIAL OFICIAL
38	COLE_BILINGUE	Indica si el establecimiento es bilingüe o no	N, S
39	COLE_CHARACTER	Indica el carácter del establecimiento	ACADÉMICO TÉCNICO TÉCNICO/ACADÉMICO NO APLICA
40	COLE_COD_DANE_SEDE	Código DANE de la sede	Texto
41	COLE_NOMBRE_SEDE	Nombre de la Sede	Texto
42	COLE_AREA_UBICACION	Área de ubicación de la sede	RURAL URBANO
43	COLE_JORNADA	Jornada de la sede	COMPLETA MAÑANA NOCHE SABATINA TARDE UNICA
44	COLE_COD_MCPIO_UBICACION	Código DANE del municipio donde está ubicada la sede	Número

N	Atributo	Descripción	Valores
45	COLE_MCPIO_UBICACION	Nombre del municipio donde está ubicada la sede	Texto
46	COLE_COD_DEPTO_UBICACION	Código DANE del departamento de la sede	Número
47	COLE_DEPTO_UBICACION	Nombre del departamento donde está ubicada la sede	Texto
Datos Citación del Examen			
48	ESTU_COD_MCPIO_PRESENTACION	Código DANE del municipio presentación del examen	Número
49	ESTU_MCPIO_PRESENTACION	Municipio de presentación del examen	Texto
50	ESTU_COD_DEPTO_PRESENTACION	Código DANE del departamento del municipio de presentación	Número
Resultados			
51	PUNT LENGUAJE	Puntaje del estudiante en lenguaje	Rango [100, 500]
52	NIVEL_DESEMP LENGUAJE	Nivel de desempeño en lenguaje	Insuficiente Mínimo Satisfactorio Avanzado
53	PUNT_MATEMATICAS	Puntaje del estudiante en Matemáticas	Rango [100, 500]

N	Atributo	Descripción	Valores
54	NIVEL_DESEMP_MATEMATICAS	Nivel de desempeño en Matemáticas	Insuficiente Mínimo Satisfactorio Avanzado
55	ESTU_INSE_INDIVIDUAL	Índice socioeconómico a nivel de estudiante	Número
56	ESTU_NSE_ESTABLECIMIENTO	Nivel Socioeconómico del Establecimiento	1, 2, 3, 4, 5

Fuente: elaboración propia con datos de ICFES, (2024b).

3.2 MÉTODOS

La Minería de Datos, comprende un conjunto de técnicas y herramientas utilizadas en el complejo proceso de extraer y presentar conocimiento implícito, previamente desconocido, pero potencialmente útil y comprensible para el ser humano. Aplicada a grandes volúmenes de información, permite automatizar la predicción de tendencias y comportamientos, así como identificar modelos previamente desconocidos (Frawley *et al.*, 1992).

Para identificar patrones relacionados con el desempeño académico en las Pruebas Saber 9º, mediante técnicas de Minería de Datos Educativa, se evaluaron tres metodologías y se seleccionó la más adecuada, para el propósito de esta investigación. Las metodologías consideradas fueron SEMMA (Azevedo y Santos, 2008), CATALYST (P3TQ) (Pyle, 2003) y CRISP-DM (Chapman *et al.*, 2000; Shearer, 2000). Todas ellas, comparten una estructura, basada en fases interrelacionadas, lo que transforma el proceso de Minería de Datos, en un procedimiento iterativo e interactivo.

La metodología SEMMA (acrónimo en inglés de *Sample, Explore, Modify, Model y Assess*), propuesta por el SAS Institute en 2012, es definida por Moiné y Haedo, (2015) como un proceso para muestrear, explorar, modificar, modelar y evaluar, grandes volúmenes de datos, con el propósito de identificar patrones previamente desconocidos, que puedan representar una ventaja comercial. Este enfoque se centra, en aspectos técnicos y excluye actividades relacionadas con el análisis y comprensión del problema en estudio. Por lo general, su aplicación está restringida, al uso del software SAS y al modelado predictivo. No obstante, el SAS Institute, enfatiza, que SEMMA no constituye una metodología de Minería de Datos, sino más bien, una estructura lógica, que organiza las herramientas funcionales de SAS *Enterprise Miner*.

La metodología Catalyst (Pyle, 2003), también conocida como P3TQ (*Product, Place, Price, Time, Quantity*), está diseñada, para abordar proyectos de Minería de Datos, de manera estructurada y adaptable. Su enfoque se basa, en la identificación y el análisis de patrones dentro de cinco

dimensiones clave, lo que la hace especialmente útil en ámbitos como el comercio, la logística y la gestión de inventarios.

P3TQ, busca optimizar procesos y mejorar la toma de decisiones basada en datos, proporcionando un marco sólido, para la estructuración de proyectos. Para ello, plantea la formulación de dos modelos: el Modelo de Negocio y el Modelo de Explotación de Información. El primero, guía la identificación de problemas u oportunidades y define los requerimientos organizacionales, mientras que el segundo, establece los pasos para construir y ejecutar modelos de Minería de Datos, a partir del análisis previo (Moiné *et al.*, 2011).

La metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), desarrollada en el año 2000 por SPSS, NCR y Daimler Chrysler, es ampliamente reconocida, como la referencia principal en proyectos de Minería de Datos (Hernández *et al.*, 2005). Este modelo de proceso estándar, de carácter abierto, describe los enfoques más utilizados por expertos en el área y se distingue por su flexibilidad y capacidad de adaptación, a distintos sectores y tipos de proyectos.

CRISP-DM, sigue una estructura jerárquica, comenzando con fases generales y desglosándose en tareas más específicas, hasta llegar a instancias concretas del proceso. Para facilitar su aplicación, se acompaña de dos documentos clave: el Modelo de Referencia y la Guía del Usuario. El primero, proporciona una visión general de las fases, tareas y resultados esperados en cualquier proyecto de Minería de Datos, mientras que el segundo, ofrece una orientación práctica con consejos y listas de verificación para cada etapa del proceso.

Azevedo y Santos, (2008), Moiné *et al.*, (2011) y Moiné y Haedo, (2015) compararon las tres metodologías y concluyeron, que, aunque presentan similitudes, CRISP-DM, es la más completa, debido a su enfoque en la aplicación de los resultados al entorno empresarial, lo que llevó a su amplia adopción. Según encuestas realizadas en el sitio web KDNuggets en 2002, 2004, 2007 y 2014, CRISP-DM, fue la metodología más utilizada, con una frecuencia cuatro veces mayor que SEMMA. Aunque CRISP-DM no es necesariamente la más reciente ni la mejor metodología, su

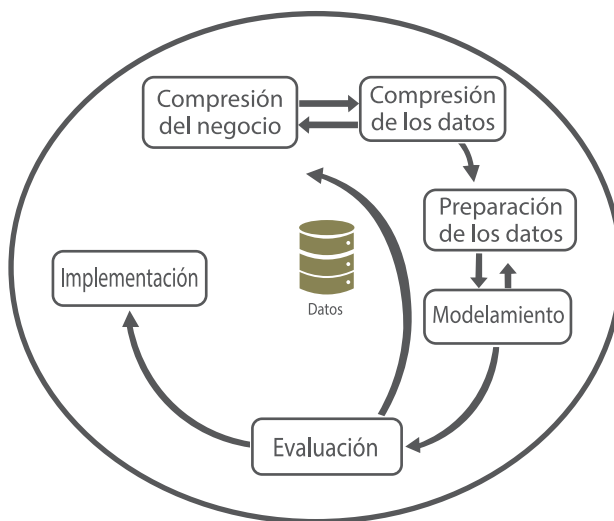
estructura resulta, altamente útil, para comprender la Minería de Datos y generar ideas, para diseñar o mejorar métodos de trabajo, en proyectos con características similares (Azevedo & Santos, 2008).

Debido a estas razones, se escogió la metodología CRISP-DM, para descubrir patrones asociados al desempeño académico, con técnicas de Minería de Datos Educativa, de los estudiantes que estaban cursando el Grado 9º de Educación Media, de los departamentos del Cauca y Nariño, que presentaron las Pruebas Saber 9º, en el año 2017.

3.3 METODOLOGÍA CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM, organiza el proceso de Minería de Datos, en seis fases: Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelamiento, Evaluación e Implementación (Chapman *et al.*, 2000; Villena, 2016) como se muestra en la figura 18.

Figura 18. Fases de la Metodología CRISP-DM



Fuente. Mancilla-Vela et al., 2020

Como se muestra en la figura 18, el modelo ofrece, una representación integral del ciclo de vida de un proyecto de Minería de Datos. Su naturaleza dinámica e iterativa, permite que los procesos no se ejecuten de manera estricta, facilitando la transición flexible entre ellos en ambas direcciones. La progresión dentro del modelo, depende tanto de los resultados obtenidos en cada fase como de la planificación de la siguiente tarea a realizar. Cada fase se organiza en tareas generales, las cuales se desglosan en tareas específicas, que detallan las acciones necesarias, para abordar situaciones concretas (Larose y Larose, 2014).

3.3.1 Comprensión del Negocio

En esta fase, se identificó con precisión, la problemática a resolver, mediante Minería de Datos. Se llevaron a cabo actividades, que permitieron una comprensión profunda y completa sobre las Pruebas Saber 9º. Dado que el propósito de estas pruebas, es evaluar el rendimiento académico de los estudiantes de Noveno Grado en Colombia, para contribuir a la mejora de la calidad educativa, se planteó desarrollar un modelo de Minería de Datos, capaz de identificar patrones asociados al desempeño académico de los estudiantes de Cauca y Nariño, que participaron en las Pruebas Saber 9º, en 2017.

Este enfoque facilitó la recolección de los datos pertinentes, para obtener resultados adecuados. Para detectar patrones de desempeño en los estudiantes de Educación Media de estos departamentos, se optó, por un proceso supervisado, en el que cada estudiante fue etiquetado según su puntaje en las Competencias de Lenguaje, Matemáticas y Desempeño General. Aquellos con puntajes superiores a la media nacional, fueron clasificados con un desempeño bueno, mientras que los que obtuvieron puntajes inferiores, fueron considerados con un desempeño bajo.

Este criterio sirvió, como base para el desarrollo de un modelo predictivo, que permite identificar perfiles académicos en las Pruebas Saber 9º, clasificando a los estudiantes, en función de si se encuentran por encima o por debajo de la media nacional. La investigación se enfocó principalmente, en el análisis de patrones de los estudiantes con desempeño bajo,

ya que conocer esta información, puede ayudar a las secretarías de educación, instituciones educativas y docentes de Cauca y Nariño, a tomar decisiones estratégicas, para fortalecer el proceso de enseñanza-aprendizaje en las áreas evaluadas por Saber 9º, contribuyendo así, a mejorar la calidad de la Educación Secundaria en la región.

3.3.2 Comprensión de los Datos

Durante esta fase, se llevó a cabo la recolección inicial de datos, con el propósito de establecer un primer acercamiento al problema, familiarizarse con la información, evaluar su calidad y detectar relaciones fundamentales, que permitieran formular las primeras hipótesis. Como resultado de esta etapa, se desarrolló el análisis exploratorio de datos, descrito en el Capítulo anterior y se construyeron repositorios de datos, que sirven como materiales de apoyo para la investigación, los cuales están referenciados en la Sección 3.1 de este Capítulo.

3.3.3 Preparación de los Datos

Para garantizar un conjunto de datos correctamente estructurado y listo para su uso en los modelos de Minería de Datos, en esta fase, se llevó a cabo un análisis inicial de la calidad de los datos en los conjuntos *sbr9_cauca* y *sbr9_narino*. Este análisis permitió identificar, para cada atributo, el número de valores distintos, la cantidad y porcentaje de valores nulos, así como medidas estadísticas clave, como la moda, la media, el valor máximo y mínimo y la desviación estándar. Con esta información, se determinaron las técnicas de limpieza y transformación de datos más adecuadas para su procesamiento. Los resultados del análisis de calidad de los atributos de tipo texto en el conjunto *sbr9_cauca* se presentan en la tabla 18, mientras que los correspondientes a los atributos numéricos, se muestran en la tabla 19.

Tabla 18. Calidad de Datos del Conjunto sbr9_cauca con 16,392 Registros y 52 Atributos Tipo Texto

N	Atributo	Valores no Nulos	Valores Nulos	% Valores Nulos	No. Valores Distintos	Valor Único
1	estu_tipodocumento	14,414	1978	12.07	9	
2	estu_genero	16,074	318	1.94	2	
3	estu_edad	16,013	379	2.31	4	
3	estu_grado	16,392	0	0.00	1	9
5	estu_fechanacimiento	15,594	798	4.87	2,075	
6	periodo	16,392	0	0.00	1	2,017
7	estu_consecutivo	16,392	0	0.00	16,392	
8	estu_estudiante	16,392	0	0.00	1	ESTUDIANTE
9	fami_educacionpadre	16,392	0	0.00	7	
10	fami_educacionmadre	16,392	0	0.00	7	
11	fami_personashogar	15,990	402	2.45	5	
12	fami_cuartoshogar	16,392	0	0.00	7	
13	fami_tieneinternet	16,392	0	0.00	3	
14	fami_tieneserviciotv	16,392	0	0.00	3	
15	fami_tienecomputador	16,392	0	0.00	3	
16	fami_tienelavadora	16,392	0	0.00	3	
17	fami_tienehornomicroogas	16,392	0	0.00	3	
18	fami_tieneautomovil	16,392	0	0.00	3	
19	fami_tieneconsolavideojuegos	16,392	0	0.00	3	

N	Atributo	Valores no Nulos	Valores Nulos	% Valores Nulos	No. Valores Distintos	Valor Único
20	fami_trabajolabpadre	16,392	0	0.00	13	
21	fami_trabajolabmadre	16,392	0	0.00	13	
22	estu_dedicacioninternet	16,392	0	0.00	6	
23	estu_comodesplazacolegio	16,392	0	0.00	5	
24	estu_tiempodesplazacolegio	16,392	0	0.00	5	
25	estu_anospreescolar	16,392	0	0.00	6	
26	estu_estecolegio_anopasado	16,392	0	0.00	3	
27	estu_gradocomenzo_estecolegio	16,392	0	0.00	1	NA
28	estu_reprobogrado	16,392	0	0.00	5	
29	estu_tardeaclase	16,392	0	0.00	5	
30	estu_nofuealcolegio	16,392	0	0.00	5	
31	estu_remunera_otrasactividades	16,392	0	0.00	3	
32	estu_seretirocolegio	16,392	0	0.00	5	
33	fami_numlibros	16,392	0	0.00	5	
34	cole_cod_dane_establecimiento	16,392	0	0.00	458	
35	cole_nombre_establecimiento	16,392	0	0.00	455	
36	cole_genero	16,392	0	0.00	3	
37	cole_naturaleza	16,392	0	0.00	2	
38	cole_bilingue	12,831	3561	21.72	2	
38	cole_caracter	16,250	142	0.87	4	

N	Atributo	Valores no Nulos	Valores Nulos	% Valores Nulos	No. Valores Distintos	Valor Único
40	cole_cod_dane_sede	16,392	0	0.00	497	
41	cole_nombre_sede	16,392	0	0.00	483	
42	cole_area_ubicacion	16,392	0	0.00	2	
43	cole_jornada	16,392	0	0.00	5	
44	cole_cod_mcpio_ubicacion	16,392	0	0.00	42	
45	cole_mcpio_ubicacion	16,392	0	0.00	42	
46	cole_cod_depto_ubicacion	16,392	0	0.00	1	19
47	cole_depto_ubicacion	16,392	0	0.00	1	CAUCA
48	estu_cod_mcpio_presentacion	16,392	0	0.00	42	
49	estu_mcpio_presentacion	16,392	0	0.00	42	
50	estu_cod_depto_presentacion	16,392	0	0.00	1	19
51	nivel_desemp_lenguaje	16,392	0	0.00	4	
52	nivel_desemp_matematicas	16,392	0	0.00	4	

Fuente. Esta Investigación

Tabla 19. Calidad de Datos del Conjunto sbr9_cauca con 16,392 Registros y 4 Atributos Tipo Numérico

N	Atributo	Vr. no Nulos	Vr. Nulos	% Vr. Nulos	Vr. Distin- tos	Min	Max	Media	σ DE	Moda	Mdn.
1	punt_lenguaje	16,392	0	0.00	375	100	500	294.09	60.45	283	290
2	punt_matematicas	16,392	0	0.00	374	100	500	290.48	61.26	287	283
3	estu_inse_individual	16,392	0	0.00	2,634	0	74.68	41.31	9.72	27.93	40.07
4	estu_nse_establecimiento	16,392	0	0.00	5	1	5	1.80	0.78	2	2

Fuente. Esta Investigación

Tabla 20. Calidad de Datos del Conjunto sbr9_narino con 18,971 Registros y 52 Atributos Tipo Texto

N	Atributo	Valores no Nulos	Valores Nulos	% Valores Nulos	No. Valores Distintos	Valor Único
1	estu_tipodocumento	17,167	1804	9.51	9	
2	estu_genero	18,486	485	2.56	2	
3	estu_edad	18,403	568	2.99	4	
3	estu_grado	18,971	0	0.00	1	9
5	estu_fechanacimiento	18,188	783	4.13	2,298	
6	periodo	18,971	0	0.00	1	2,017
7	estu_consecutivo	18,971	0	0.00	18,971	
8	estu_estudiante	18,971	0	0.00	1	ESTUDIANTE

N	Atributo	Valores no Nulos	Valores Nulos	% Valores Nulos	No. Valores Distintos	Valor Unico
9	fami_educacionpadre	18,971	0	0.00	7	
10	fami_educacionmadre	18,971	0	0.00	7	
11	fami_personashogar	18,387	584	3.08	5	
12	fami_cuartoshogar	18,971	0	0.00	7	
13	fami_tieneinternet	18,971	0	0.00	3	
14	fami_tieneserviciotv	18,971	0	0.00	3	
15	fami_tienecomputador	18,971	0	0.00	3	
16	fami_tienelavadora	18,971	0	0.00	3	
17	fami_tienehormomicroogas	18,971	0	0.00	3	
18	fami_tieneautomovil	18,971	0	0.00	3	
19	fami_tieneconsolavideojuegos	18,971	0	0.00	3	
20	fami_trabajolaborpadre	18,971	0	0.00	13	
21	fami_trabajolabormadre	18,971	0	0.00	13	
22	estu_dedicacioninternet	18,971	0	0.00	6	
23	estu_comodesplazacollegio	18,971	0	0.00	5	
24	estu_tiemposdesplazacollegio	18,971	0	0.00	5	
25	estu_anospreescolar	18,971	0	0.00	6	
26	estu_estecolegio_anopasado	18,971	0	0.00	3	
27	estu_gradocomenzo_estecolegio	18,971	0	0.00	1	NA

N	Atributo	Valores no Nulos	Valores Nulos	% Valores Nulos	No. Valores Distintos	Valor Unico
28	estu_reprobogrado	18,971	0	0.00	5	
29	estu_tardeaclase	18,971	0	0.00	5	
30	estu_nofuealcolegio	18,971	0	0.00	5	
31	estu_remunera_otrasactividades	18,971	0	0.00	3	
32	estu_seretirocolegio	18,971	0	0.00	5	
33	fami_numlibros	18,971	0	0.00	5	
34	cole_cod_dane_establecimiento	18,971	0	0.00	406	
35	cole_nombre_establecimiento	18,971	0	0.00	387	
36	cole_genero	18,971	0	0.00	2	
37	cole_naturaleza	18,971	0	0.00	2	
38	cole_bilingue	16,828	2,143	11.30	2	
38	cole_caracter	18,861	110	0.58	4	
40	cole_cod_dane_sede	18,971	0	0.00	421	
41	cole_nombre_sede	18,971	0	0.00	411	
42	cole_area_ubicacion	18,971	0	0.00	2	
43	cole_jornada	18,971	0	0.00	5	
44	cole_cod_mcpio_ubicacion	18,971	0	0.00	64	
45	cole_mcpio_ubicacion	18,971	0	0.00	64	
46	cole_cod_depto_ubicacion	18,971	0	0.00	1	52

N	Atributo	Valores no Nulos	Valores Nulos	% Valores Nulos	No. Valores Distintos	Valor Unico
47	cole_depto_ubicacion	18,971	0	0.00	1	NARIÑO
48	estu_cod_mcpio_presentacion	18,971	0	0.00	64	
49	estu_mcpio_presentacion	18,971	0	0.00	64	
50	estu_cod_depto_presentacion	18,971	0	0.00	1	52
51	nivel_desemp_lenguaje	18,963	8	0.04	4	
52	nivel_desemp_matematicas	18,936	35	0.18	4	

Fuente. Esta Investigación

Tabla 21. Calidad de Datos del Conjunto Sbr9_Narino con 18,971 Registros y 4 Atributos Tipo Numérico

N	Atributo	Vr. no Nulos	Vr. Nulos	% Vr. Nulos	Vr. Distintos	Min	Max	Media	σ DE	Moda	Mdn.
1	punt_lenguaje	18,963	8	0.04	378	100	500	308.13	62.64	288	306
2	punt_matematicas	18,936	35	0.18	380	100	500	305.55	65.59	500	300
3	estu_inse_individual	18,971	0	0.00	2,703	0	74.68	41.63	10.47	27.93	40.66
4	estu_nse_establecimiento	18,971	0	0.00	5	1	5	2.09	0.80	2	2

Fuente. Esta Investigación

Tras analizar la calidad de los datos del conjunto *sbr9_cauca* (tablas 18 y 19), se concluye que esta es satisfactoria. La mayoría de los atributos carece de valores nulos; en aquellos casos donde existen, su proporción es inferior al 10%, lo que permite su tratamiento mediante medidas de tendencia central (media, moda o mediana) o de posición (máximo y mínimo). Se exceptúan los atributos *estu_tipodocumento* (12.1%) y *cole_bilingue* (21.7%), cuyos valores faltantes requieren un tratamiento diferenciado o, en su defecto, la eliminación de dichas variables.

El análisis de la calidad de datos del conjunto *sbr9_narino* de los atributos de tipo texto, se muestra en la tabla 20 y de tipo numérico se muestra en la tabla 21.

La calidad de los datos del conjunto *sbr9_narino* es buena, tal como se observa en las tablas 20 y 21. La mayoría de los atributos no presenta valores nulos y los que sí, lo hacen en una proporción menor al 10%, permitiendo su reemplazo por medidas de tendencia central o de posición. No obstante, el atributo *cole_bilingue* (11.3%) presentan porcentajes superiores, por lo cual deben ser tratados de forma especial o descartados del análisis.

Basándose en el análisis previo de la calidad de los datos en los conjuntos *sbr9_cauca* y *sbr9_narino* y considerando las técnicas de Minería de Datos a emplear, se llevó a cabo el proceso de limpieza y transformación de datos. En primer lugar, se eliminaron los atributos con un alto porcentaje de valores nulos, así como aquellos, con una gran cantidad de valores distintos que dificultaban la identificación de patrones generales. También se descartaron valores constantes, atributos con un único valor, claves primarias y variables irrelevantes, que no aportaban al objetivo del estudio. La tabla 22 presenta la lista de atributos eliminados en los conjuntos *sbr9_cauca* y *sbr9_narino*.

Tabla 22. Lista de Atributos Inicialmente Eliminados de los Conjuntos de Datos Sbr9_Cauca y Sbr9_Narino

N	Atributo	Causa Eliminación
1	estu_tipodocumento	No relevante para el estudio
2	estu_grado	Valor único: grado 9
3	estu_fechanacimiento	Redundante, se cuenta con el rango de edades
4	periodo	Valor único: periodo 2017
5	estu_consecutivo	Llave primaria.
6	estu_estudiante	Valor único: ESTUDIANTE
7	fami_tieneautomovil	No relevante
8	fami_trabajolaborpadre	No relevante para el estudio. Mas del 16% tiene el valor de NA
9	fami_trabajolabormadre	No relevante para el estudio. La mayoría es de ocupación hogar
10	estu_dedicacioninternet	No relevante para el estudio
11	estu_comodesplazacolegio	No relevante para el estudio
12	estu_tiempodesplazacolegio	No relevante para el estudio
13	estu_anospreescolar	No relevante para el estudio
14	estu_estecolegio_anopasado	No relevante para el estudio
15	estu_gradocomenzo_estecolegio	Valor único: NA. No relevante para el estudio
16	estu_tardeaclase	No relevante para el estudio
17	estu_nofuealcolegio	No relevante para el estudio
18	estu_remunera_otrasactividades	No relevante para el estudio
19	estu_seretirocolegio	No relevante para el estudio
20	fami_numlibros	No relevante para el estudio
21	cole_cod_dane_establecimiento	Alto número de valores distintos. No permite generalizar
22	cole_nombre_establecimiento	Alto número de valores distintos. No permite generalizar
23	cole_bilingue	No relevante para el estudio. La gran mayoría de colegios no son bilingues
24	cole_cod_dane_sede	Alto número de valores distintos. No permite generalizar

N	Atributo	Causa Eliminación
25	cole_nombre_sede	Alto número de valores distintos. No permite generalizar
26	cole_cod_depto_ubicacion	Único valor: 19 o 52
27	cole_depto_ubicacion	Único valor: CAUCA o NARIÑO
28	estu_cod_mcpio_presentacion	No relevante para el estudio: Interesa la ubicación del colegio
29	estu_mcpio_presentacion	No relevante para el estudio. Interesa la ubicación del colegio
30	estu_cod_depto_presentacion	Único valor 19 o 52. No relevante para el estudio
31	estu_nse_establecimiento	Se tomo el INSE individual del estudiante

Fuente. Esta Investigación

Siguiendo las técnicas de imputación de datos descritas por Quitiaquez, (2024), se sustituyeron los valores nulos y aquellos considerados nulos camuflados (como NA y No Aplica) en los atributos de los conjuntos *sbr9_cauca* y *sbr9_narino*, cuyo porcentaje de estos valores, fuera inferior al 10%. La imputación se realizó, utilizando la moda de la institución educativa del estudiante, en los atributos categóricos y la media de la misma institución en los atributos numéricos.

Además, se llevó a cabo un proceso de estandarización, eliminando las tildes de los atributos que las contenían y reemplazando el carácter 'ñ' por 'n', asegurando así, una correcta interpretación por la herramienta de Minería de Datos, empleada para el descubrimiento de patrones. En cuanto a los valores nulos correspondientes a la edad del estudiante, estos fueron calculados, con base en su fecha de nacimiento y la fecha de aplicación de las Pruebas Saber 9º, en 2017 (13 de septiembre de 2017).

Un gran número de valores distintos dentro de una variable, puede dificultar la identificación de patrones en la Minería de Datos (Hernández *et al.*, 2005). Para solucionar este problema, se aplica un proceso de discretización, que consiste, en reemplazar la variable original por una nueva, con un número reducido de valores distintos, lo que facilita la identificación de patrones con un alcance más general. Posteriormente, la variable original es eliminada.

Siguiendo este criterio, se incorporó el atributo *cole_region_ubicación* en los conjuntos de datos *sbr9_cauca* y *sbr9_narino*, el cual agrupa, los municipios de ambos departamentos en subregiones geográficas. La tabla 23 presenta, las subregiones del departamento del Cauca, mientras que la tabla 24, muestra las correspondientes al departamento de Nariño.

Tabla 23. Subregiones del Departamento del Cauca

Subregión	Municipios
POPAYAN	Popayán.
CENTRO	Cajibío, El Tambo, Morales, Piendamó, Tunía, Puracé, Silvia, Timbío.
NORTE	Buenos Aires, Caldono, Caloto, Corinto, Guachené, Jambaló, Miranda, Padilla, Puerto Tejada, Santander de Quilichao, Toribio, Villa Rica.
SUR	Argelia, Balboa, Bolívar, Florencia, Mercaderes, Patía, Sucre.
OCCIDENTE	Guapi, López de Micay, Timbiquí.
ORIENTE	Inza, Páez, Totoró
MACIZO	Almaguer, La Vega, La Sierra, Rosas, Sotará.
BOTA CAUCANA	Piamonte, San Sebastián, Santa Rosa.

Fuente. Fuente. Wikipedia, s.f. (a)

Tabla 24. Subregiones del Departamento de Nariño

Subregión	Municipios
PASTO	Pasto
CENTRO	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.

Subregión	Municipios
PACÍFICO SUR	Francisco Pizarro, San Andrés de Tumaco.
PIEDEMONTES 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.
TELEMBI	Barbacoas, Magüí Payán, Roberto Payán.

Fuente. Fuente. Wikipedia, s.f. (b)

Se diseñó un índice denominado *fami_electrodomesticos*, para evaluar la condición de los electrodomésticos en el hogar del estudiante, basado en el número de dispositivos presentes en la vivienda. Este índice se calcula, sumando los valores de presencia (1) o ausencia (0) de cada electrodoméstico. Según el resultado, la condición de electrodomésticos se clasifica en tres niveles: buena, si el índice es 2; media, si es 1 y mala, si es 0. La tabla 25 presenta, las variables utilizadas en el cálculo de este índice (Timarán *et al.*, 2021a).

Tabla 25. Cálculo del Índice *fami_electrodomesticos*

Índice	Lavadora	Horno	Suma
BUENO	1	1	2
REGULAR	1	0	1
REGULAR	0	1	1
MALO	0	0	0

Fuente. Esta Investigación.

Se desarrolló el índice *fami_tics*, el cual evalúa la cantidad de servicios de tecnologías de la información y comunicaciones presentes en la vivienda. Su cálculo siguió el mismo enfoque empleado en el índice *fami_electrodomesticos*. Según los valores obtenidos, si el índice se encuentra entre 3 y 4, la condición TIC es buena; entre 1 y 2, es regular y si es 0, se considera mala. La tabla 26 presenta, los atributos utilizados en este cálculo (Timarán *et al.*, 2021b).

Asimismo, se creó el índice *fami_hacinamiento*, basado en el cálculo del nivel de hacinamiento, el cual mide, la relación entre el número de personas que habitan una vivienda y el número de habitaciones disponibles (Spicker *et al.*, 2009). De acuerdo con los valores generalmente aceptados, hasta 2.4 se considera sin hacinamiento; de 2.5 a 4.9, hacinamiento medio y de 5.0 o más, hacinamiento crítico. Aplicando estos conceptos, el índice *fami_hacinamiento* se obtuvo, dividiendo los valores de los atributos *fami_personashogar* entre *fami_cuartoshogar*, reemplazando ambas variables por el nuevo índice. La asignación de valores siguió la escala propuesta por Spicker *et al.*, (2009).

Tabla 26. Cálculo del Índice *fami_tics*

Índice	Internet	Servicio TV	Computador	Videojuegos	Suma
BUENO	1	1	1	1	4
BUENO	0	1	1	1	3
BUENO	1	1	1	0	3
BUENO	1	1	0	1	3
BUENO	1	0	1	1	3
REGULAR	1	1	0	0	2
REGULAR	1	0	1	0	2
REGULAR	1	0	0	1	2
REGULAR	0	1	1	0	2
REGULAR	0	1	0	1	2
REGULAR	0	0	1	1	2
REGULAR	1	0	0	0	1
REGULAR	0	1	0	0	1
REGULAR	0	0	1	0	1
REGULAR	0	0	0	1	1
MALO	0	0	0	0	0

Fuente. Esta Investigación.

Se creó el atributo *punt_general*, para representar el puntaje global obtenido por los estudiantes de Cauca y Nariño, en las Pruebas Saber 9º. Este se calcula, como la media entera de la suma de los puntajes obtenidos en las competencias de Lenguaje (*punt_lenguaje*) y Matemáticas (*punt_matematicas*).

Adicionalmente, se establecieron los atributos *clase_desemp_lenguaje*, *clase_desemp_matematicas* y *clase_desemp_general*, para clasificar el desempeño de los estudiantes en cada competencia, identificando si se encuentran por encima o por debajo de la media nacional. Este procedimiento, también permitió suavizar los valores atípicos en los puntajes. La tabla 27 presenta, la cantidad de estudiantes del conjunto *sbr9_cauca* que se encuentran sobre o bajo la media nacional, mientras que la tabla 28, muestra la misma información para el conjunto *sbr9_narino*.

Tabla 27. Desempeño de los Estudiantes del Departamento del Cauca en las Pruebas Saber 9º

Clase	Bajo la Media	Sobre la Media	Media Nacional/500
<i>clase_desemp_lenguaje</i>	10,617	5,775	314.84
<i>clase_desemp_matematicas</i>	10,687	5,705	307.07
<i>clase_desemp_global</i>	11028	5364	310.91

Fuente. Esta Investigación.

Según los datos de la tabla 27, el desempeño de los estudiantes en el Cauca muestra que, en Lenguaje, un 64.8 % no alcanza la media nacional, frente al 35.2 % que sí lo hace. En Matemáticas, la tendencia es similar: el 65.2% se ubica bajo el promedio nacional y el 34.8 % por encima de este. Esta brecha se acentúa en el Desempeño General de las Pruebas Saber 9º, donde el 67.3 % de los estudiantes se posiciona por debajo del promedio nacional, dejando solo un 32.7 % en el nivel superior.

Tabla 28. Desempeño de los Estudiantes del Departamento de Nariño en las Pruebas Saber 9º

Clase	Bajo la Media	Sobre la Media	Media Nacional/500
clase_desemp_lenguaje	10,466	8,505	314.84
clase_desemp_matematicas	10,241	8,730	307.07
clase_desemp_global	10,552	8,419	310.91

Fuente. Esta investigación.

El análisis de la tabla 28 revela que el 55.1 % de los estudiantes nariñenses se sitúa por debajo de la media nacional en la prueba de Lenguaje, mientras que el 44.9% restante se encuentra por encima. En Matemáticas, la tendencia es similar: el 53.9 % se ubica bajo el promedio nacional y el 46.1 % por encima de este. Finalmente, en cuanto al Desempeño General de las Pruebas Saber 9º, los resultados mantienen la consistencia, con un 55.6 % de los evaluados por debajo de la media nacional y un 44.4 % que logra posicionarse sobre ella.

Se creó el atributo *estu_cuartil_inse* para clasificar a los estudiantes, según su índice socioeconómico (*estu_inse_individual*) (ICFES, 2019), siendo Q1 el cuartil más alto y Q4 el más bajo. La asignación de cuartiles se realizó conforme a los siguientes criterios:

- Si *estu_inse_individual* ≤ 34.36 entonces *cuartil_inse* = Q4
- Si *estu_inse_individual* > 34.36 y *estu_inse_individual* ≤ 40.37 entonces *cuartil_inse* = Q3
- Si *estu_inse_individual* > 40.37 y *estu_inse_individual* ≤ 48.03 entonces *cuartil_inse* = Q2
- Si *estu_inse_individual* > 48.03 entonces *cuartil_inse* = Q1

Posteriormente, se eliminaron los atributos reemplazados por los nuevos. Como resultado de esta fase, se generaron los conjuntos de datos definitivos *sbr9_cauca_final* con 16,392 registros y 20 atributos, y *sbr9_narino_final* con 18,971 registros y 18 atributos. El diccionario de

datos final de ambos conjuntos, que servirá como base para la etapa de modelado, se encuentra en la tabla 29.

Tabla 29. Diccionario de Datos de los Conjuntos de Datos *sbr9_cauca* y *sbr9_narino*

N	Atributo	Descripción	Valores
Socioeconómicos			
1	ESTU_GENERO	Género	Hombre Mujer
2	ESTU_EDAD	Edad del Estudiante	13 años o menos 14 años 15 años 16 años o mas
3	FAMI_EDUCACIONPADRE	Nivel Educativo más alto alcanzado por el Padre	Primaria Bachillerato Técnico o Tecnólogo Universitario - Pregrado Universitario – Posgrado NA
4	FAMI_EDUCACIONMADRE	Nivel Educativo más alto alcanzado por la Madre	Primaria Bachillerato Técnico o Tecnólogo Universitario - Pregrado Universitario – Posgrado NA
5	FAMI_HACINAMIENTO	Índice de hacinamiento	Bueno, Regular, Malo. Reemplaza a los atributos: fami_personas hogar y fami_cuartos-hogar.
6	FAMI_TICS	Índice Tics	Bueno, Regular, Malo.
7	FAMI_ELECTRODOMMESTICOS	Índices electrodomésticos	Bueno, Regular, Malo. Reemplaza atributos: fami_tienelavadora, fami_tienehornomicrogas.
8	ESTU_REPROBOGRADO	Si ha repetido algún grado	N, S
9	ESTU_CUARTIL_INSE	Cuartil del INSE del estudiante	Q1, Q2, Q3, Q4 Reemplaza atributo estu_inse_individual

N	Atributo	Descripción	Valores
Institucionales			
10	COLE_GENERO	Indica el género de la población del establecimiento	FEMENINO MASCULINO MIXTO
11	COLE_NATURALEZA	Indica la naturaleza del establecimiento	NO OFICIAL OFICIAL
12	COLE_CHARACTER	Indica el carácter del establecimiento	ACADÉMICO TÉCNICO TÉCNICO/ACADÉMICO
13	COLE_AREA_UBICACION	Área de ubicación de la sede	RURAL URBANO
14	COLE_JORNADA	Jornada de la sede	COMPLETA MAÑANA NOCHE SABATINA TARDE UNICA
15	COLE_REGION_UBICACION	Subregiones de los departamentos o del Cauca o de Nariño	8 subregiones del Cauca 14 subregiones de Nariño
Académicos			
16	CLASE_DESEMP LENGUAJE	Clase de desempeño académico en lenguaje	Sobre la Media y Bajo la Media. Reemplaza atributo punt_lenguaje
17	CLASE_DESEMP MATEMATICAS	Clase de desempeño académico en matemáticas	Sobre la Media y Bajo la Media. Reemplaza atributo punt_matematicas
18	CLASE_DESEMP_GENERAL	Clase de desempeño académico general	Sobre la Media y Bajo la Media. Reemplaza atributo punt_general

Fuente. Esta Investigación.

3.3.4 Modelado

En esta fase, se seleccionó el modelo más apropiado, que permita descubrir patrones asociados al desempeño académico de los estudiantes de los departamentos del Cauca y Nariño, que presentaron las Pruebas Saber 9º. Según Frawley *et al.* (1992) dichos patrones, deben ser fácilmente comprensibles e interpretables por los usuarios.

El modelo de clasificación mediante árboles de decisión, resultó ser el más adecuado para esta investigación. Su elección, se basó en diversas razones: a) emplea una estructura arbórea para representar posibles resultados, lo que lo convierte en un modelo sencillo e interpretable al examinar su configuración, facilitando así su aplicabilidad; b) al ser un modelo de caja blanca, permitió a los investigadores comprender e interpretar los patrones descubiertos (Han & Kamber, 2001; Azevedo & Santos, 2008; Hernández *et al.*, 2005; Timarán *et al.*, 2017), a diferencia de otras técnicas, como las redes neuronales artificiales, que operan como modelos de caja negra y dificultan la interpretación de los resultados por parte de los usuarios; c) se buscó un modelo, que posibilitara la identificación de patrones en el desempeño académico de los estudiantes, en las Pruebas de Lenguaje, Matemáticas y Desempeño General del Saber 9º, clasificándolos según estuvieran “sobre la media nacional” o “bajo la media nacional”, atributos asignados como clase o *target*, lo que el modelo de clasificación logró eficazmente; d) el modelo incorpora en los patrones, únicamente los atributos relevantes para la toma de decisiones, omitiendo aquellos que no contribuyen a la precisión del árbol (Han *et al.*, 2011; Sattler & Dunemann, 2001; Timarán *et al.*, 2021b); e) los árboles de decisión, pueden manejar tanto datos numéricos como categóricos y requieren poca preparación previa, a diferencia de otras técnicas, que exigen la normalización de datos y trabajan con un único tipo de información; f) la validación del modelo, puede realizarse fácilmente, mediante las métricas obtenidas en la matriz de confusión, lo que permite determinar su fiabilidad; g) los árboles de decisión, son robustos y muestran un buen desempeño con grandes volúmenes de datos.

La clasificación mediante árboles de decisión, se basa en clases disjuntas, lo que significa, que cada árbol conduce a una única hoja, asignando una sola clase a la predicción (Hernández & Lorente, 2009). En estas estructuras, las hojas, representan las etiquetas de clase, mientras que las ramas, simbolizan las conjunciones de características que llevan a dichas etiquetas. Este enfoque permite, obtener resultados a través de un proceso supervisado.

El aprendizaje de árboles de decisión, sigue una estrategia de divide y vencerás, realizando una búsqueda codiciosa, para determinar los puntos óptimos de división dentro del árbol. Este proceso de segmentación, se ejecuta de manera descendente y recursiva, hasta que la mayoría o la totalidad de los registros han sido clasificados con etiquetas de clase específicas (Mor *et al.*, 2005).

La clasificación de datos consta de dos etapas. En la primera, se construye un modelo, a partir de un conjunto de datos denominado, Conjunto de Entrenamiento, en el cual, cada registro pertenece a una clase predefinida, determinada por un atributo de clase o variable dependiente. Los demás atributos se conocen, como Atributos Predictores o Variables Independientes. En la segunda etapa, el modelo se evalúa utilizando un conjunto de datos distinto, denominado Conjunto De Prueba.

La exactitud del modelo se determina, según el porcentaje de ejemplos, en el conjunto de prueba que son clasificados correctamente (Timarán *et al.*, 2021a).

Uno de los principales desafíos de los árboles de decisión, es su susceptibilidad al sobreajuste (*overfitting*), un fenómeno que ocurre, cuando el modelo se ajusta excesivamente a los datos de entrenamiento, dificultando su capacidad de generalización a nuevos datos de prueba. Para mitigar este problema, se emplean procesos de poda, que consisten, en la eliminación de ramas que contienen características de poca relevancia.

Existen dos tipos de poda: la Poda Previa (*prepoda*) y la Poda Posterior (*postpoda*). La Prepoda, detiene el crecimiento del árbol, cuando la cantidad de datos es insuficiente, evitando la generación de divisiones

innecesarias. Por otro lado, la Postpoda, elimina los subárboles, con datos inadecuados, una vez que el árbol ha sido construido, mejorando así su capacidad de generalización (Han *et al.*, 2011).

Para comparar y evaluar los árboles de decisión, es fundamental, analizar su Complejidad y el Proceso de Poda. La Complejidad, se refiere a la cantidad de nodos, ramas y niveles presentes en la estructura del árbol, mientras que la Poda consiste, en eliminar nodos y ramas innecesarios o redundantes, para optimizar su funcionamiento.

Un árbol de decisión complejo, puede alcanzar una alta precisión, pero corre el riesgo de sobreajustar los datos, es decir, captar demasiado ruido y perder capacidad de generalización ante nuevos datos. Por otro lado, un árbol de decisión demasiado simple, podría presentar una baja precisión y no ajustarse adecuadamente a los datos, lo que conlleva, la pérdida de patrones significativos e información valiosa.

El Proceso de Poda, ayuda a encontrar un equilibrio entre complejidad y precisión, ya que permite reducir el tamaño del árbol, sin comprometer su desempeño (Hernández *et al.*, 2005; Han *et al.*, 2011).

La poda de árboles de decisión, se lleva a cabo, mediante la definición de ciertos hiperparámetros. Estos son variables de configuración externa, utilizadas para gestionar el entrenamiento de modelos en Minería de Datos y aprendizaje automático (Sánchez-Jiménez *et al.*, 2023). A diferencia de los parámetros, que se generan automáticamente durante el proceso de aprendizaje, los hiperparámetros, se establecen de manera manual antes de entrenar un modelo.

El ajuste de hiperparámetros, permite optimizar el rendimiento del modelo para obtener resultados óptimos. Se trata de una práctica experimental, en la que, a través de múltiples iteraciones, se prueban diferentes valores hasta identificar los más adecuados (Belcic & Stryker, 2024). Este ajuste puede realizarse de manera manual o automática. Aunque el ajuste manual puede ser lento y laborioso, brinda una mejor comprensión

del impacto de los hiperparámetros en el modelo. Sin embargo, en la mayoría de los casos, se emplean algoritmos de aprendizaje de hiperparámetros, ampliamente utilizados.

El proceso de ajuste de hiperparámetros, es iterativo y requiere probar diversas combinaciones de parámetros y valores. Generalmente, se inicia definiendo una métrica principal, como la precisión y luego se ajusta para optimizarla. Es recomendable, utilizar técnicas de validación cruzada, para evitar, que el modelo se enfoque únicamente en una porción específica de los datos.

Según Mor *et al.*, (2005) otra manera de comparar y evaluar los árboles de decisión, es mediante Técnicas de Validación y Validación Cruzada. La Validación, consiste en evaluar el rendimiento del árbol de decisión, utilizando un conjunto de datos independiente que no fue empleado en su construcción. Por otro lado, la Validación Cruzada, implica dividir los datos en varios subconjuntos y evaluar el desempeño del árbol en cada uno, mientras el resto de los datos, se utilizan para su entrenamiento.

Ambas técnicas, ayudan a prevenir el sobreajuste y el mal ajuste, proporcionando una estimación más precisa del rendimiento del árbol frente a nuevos datos. La comparación entre árboles de decisión, puede basarse en sus puntuaciones de validación y validación cruzada, seleccionando aquel con la mayor puntuación media en distintos subconjuntos de datos.

Una tercera estrategia para comparar y evaluar los árboles de decisión, es mediante el Análisis de su Sensibilidad y Especificidad. La Sensibilidad, representa la proporción de verdaderos positivos, es decir, los casos correctamente clasificados como positivos, mientras que la Especificidad, mide la proporción de verdaderos negativos, o los casos correctamente identificados como negativos.

Estas métricas, son particularmente útiles en problemas de clasificación binaria, en los que el árbol de decisión, debe distinguir entre dos posibles resultados, como sí o no, comprar o vender, o enfermo o saludable. Un mayor valor de Sensibilidad y Especificidad, indica un mejor

desempeño del árbol de decisión; sin embargo, ambas medidas no son independientes y suele existir un equilibrio entre ellas.

Para comparar y evaluar árboles de decisión en función de estos criterios, se puede utilizar, la curva de la Característica Operativa del Receptor (ROC), que grafica la Sensibilidad frente a la Especificidad, para distintos umbrales o puntos de corte (Mor *et al.*, 2005).

Los hallazgos obtenidos durante la fase de modelado, serán expuestos en el próximo capítulo.

3.3.5 Evaluación

Durante la fase de Evaluación, se verificó, si el modelo cumplía con los requisitos establecidos en el proyecto. En este proceso, se examinaron los patrones descubiertos para determinar su validez, eliminar aquellos redundantes o irrelevantes y traducir los patrones útiles en términos comprensibles para el usuario.

Para interpretar los resultados, se utilizó la Matriz de Confusión, una tabla que muestra la cantidad de clasificaciones realizadas para cada categoría. En esta tabla, la diagonal representa, las clasificaciones correctas, permitiendo evaluar la precisión del modelo.

El análisis y la interpretación de los patrones descubiertos, se detallan en el capítulo de Resultados y Discusión.

3.3.6 Implementación

En esta etapa, se elaboró el informe final de la investigación, con el propósito de documentar y presentar los resultados de manera accesible para el usuario, fomentando así, el incremento del conocimiento. A través de su difusión, los hallazgos obtenidos, se incorporarán al conocimiento existente y podrán integrarse en los procesos de toma de decisiones del MEN, el ICFES y las instituciones educativas responsables de garantizar la calidad de la educación media en los departamentos del Cauca y Nariño.

Además, la información contenida en el informe, permitirá identificar tendencias en el desempeño académico de los estudiantes, facilitando el desarrollo de estrategias pedagógicas basadas en evidencia. La implementación de estas estrategias, podrá contribuir al diseño de políticas educativas más efectivas, adaptadas a las necesidades específicas de cada contexto regional.

Una vez, estas entidades intervengan en los factores asociados al desempeño académico en las Pruebas Saber 9°, será posible analizar los resultados y evaluar sus efectos, lo que podría derivar en nuevas investigaciones, orientadas a la optimización de los modelos de evaluación educativa y la mejora continua del aprendizaje en la educación media (Villena, 2016).

CAPÍTULO IV RESULTADOS Y DISCUSIÓN

Para identificar patrones en el desempeño académico de los estudiantes que cursaban el Grado 9º, en los departamentos del Cauca y Nariño y presentaron las Pruebas Saber 9º, en 2017, se seleccionó la técnica de Minería de Datos, basada en clasificación mediante árboles de decisión, por considerarse la más adecuada.

Para la construcción de los distintos árboles de decisión, se establecieron como atributos clase o target, los desempeños en Lenguaje, Matemáticas y el Desempeño General, con base en los conjuntos de datos *sbr9_cauca_final* y *sbr9_narino_final*, descritos en las tablas 27, 28 y 29. Estos atributos permitieron, identificar los factores (*variables independientes*) asociados a los estudiantes, cuyo desempeño en las competencias mencionadas, era bueno (*sobre la media*) o bajo (*bajo la media*).

Además, con el propósito de seleccionar el algoritmo de árboles de decisión más adecuado, para clasificar a los estudiantes de los departamentos del Cauca y Nariño, así como el mejor método para evaluar el modelo seleccionado, se unificaron los conjuntos de datos *sbr9_cauca_final* y *sbr9_narino_final* en un único conjunto denominado *sbr9_cauca_narino_final*, que contiene 35,363 registros y 17 atributos.

Se empleó la herramienta WEKA (*Waikato Environment for Knowledge Analysis*) (Witten *et al.*, 2011) para la construcción de distintos modelos de clasificación, mediante árboles de decisión y la interpretación de los patrones descubiertos. Asimismo, se utilizó esta herramienta, para definir el procedimiento de evaluación del modelo, verificando su calidad y validez.

Para entrenar y probar un modelo de clasificación, los datos se dividen en dos conjuntos: Entrenamiento y Prueba. Esta metodología permite, evaluar el desempeño del modelo en datos no utilizados durante la fase de entrenamiento, garantizando su capacidad de generalización (Witten *et al.*, 2011).

Las siguientes opciones en la herramienta WEKA, permiten evaluar la calidad y validez de un clasificador:

- Uso del conjunto de entrenamiento (*Use training set*): el modelo se entrena y se prueba con el mismo conjunto de datos. Aunque esta técnica puede generar resultados óptimos en ese conjunto específico, su precisión, puede verse afectada al aplicarse a nuevos datos.
- Uso de un conjunto de prueba (*Supplied test set*): se entrena el modelo, con un conjunto de datos y se evalúa con otro independiente. Sin embargo, existe el riesgo de que el conjunto de prueba, no refleje las características del conjunto de entrenamiento, lo que puede afectar la validez de los resultados.
- División por porcentaje (*Percentage Split*): se selecciona aleatoriamente una parte de los datos para el entrenamiento y otra para la prueba. A diferencia del método anterior, ambos conjuntos pertenecen al mismo universo de datos, minimizando el riesgo de discrepancia.
- Validación cruzada (*Cross validation*): esta técnica reduce la influencia de la partición de datos en los resultados del experimento. Consiste en dividir los datos en varios subconjuntos y alternar el entrenamiento y la prueba en cada iteración para obtener una evaluación más precisa (Hernández *et al.*, 2005)

4.1 DEFINICIÓN DEL ALGORITMO DE ÁRBOLES DE DECISIÓN

Para determinar el algoritmo más adecuado, para clasificar a los estudiantes de los conjuntos de datos *sbr9_cauca_final* y *sbr9_narino_final*, según los atributos de clase establecidos por su desempeño en Lenguaje, Matemáticas y el Desempeño General, se evaluaron diversos algoritmos de árboles de decisión, mediante la herramienta WEKA. Tras comparar los resultados, se seleccionó el algoritmo, con las mejores métricas. Para garantizar una evaluación rigurosa, se empleó la Técnica de Validación Cruzada, como método de análisis de los algoritmos.

Como se mencionó anteriormente, para llevar a cabo esta selección, se integraron ambos conjuntos de datos en uno solo, denominado *sbr9_cauca_narino_final*, el cual contiene 35,363 registros y 17 atributos. Posteriormente, se probaron en WEKA, utilizando la opción de validación cruzada (*cross-validation*) con 10 particiones (*folds*) y los parámetros de poda predeterminados. Según Hernández *et al.*, (2005), esta configuración es la más recomendada.

Para la construcción de los modelos con los diferentes algoritmos, se consideró como atributo clase o *target* exclusivamente, el desempeño general. Los modelos resultantes, fueron evaluados, con base en diversas métricas, incluyendo Exactitud, Precisión, Sensibilidad, *F-measure*, Área Bajo la Curva PRC, la Curva ROC y el Tiempo de Ejecución.

La Exactitud (*accuracy*), representa la proporción de predicciones correctas realizadas por el modelo, en relación con el total de predicciones efectuadas. Es una métrica útil, cuando las clases están equilibradas, es decir, cuando la distribución de instancias, es similar en cada categoría.

La Precisión (*precision*), mide la fiabilidad del modelo en la identificación de casos positivos. Se calcula, como la proporción de verdaderos positivos con respecto al total de predicciones positivas realizadas. En escenarios donde las clases están desbalanceadas, se recomienda emplear Precisión o Sensibilidad como métricas de evaluación.

La Sensibilidad (*recall*), indica la capacidad del modelo, para detectar correctamente las clases positivas. Se obtiene, al calcular la proporción de verdaderos positivos identificados por el algoritmo en comparación con el total de casos positivos reales.

La F-measure o F1-score, combina Precisión y Sensibilidad, en una única métrica, a través de una media armónica. Es particularmente útil, cuando la distribución de clases, es desigual, ya que permite encontrar un equilibrio entre ambas métricas.

El PRC área (*Precision-Recall Curve Area*), cuantifica el área bajo la curva de *precision-recall*, que muestra la relación entre la *precisión* y el *recall*, a lo largo de distintos umbrales de clasificación. Esta métrica, es especialmente relevante en problemas con clases desbalanceadas, donde *precision* y *recall*, ofrecen información más significativa que la exactitud. Un valor elevado de PRC área, indica que el modelo logra un equilibrio adecuado entre ambas métricas, reduciendo la cantidad de falsos positivos.

La curva ROC (*Receiver Operating Characteristic*), es una representación gráfica, que muestra la relación, entre la tasa de verdaderos positivos y la tasa de falsos positivos, conforme se modifica el umbral de clasificación. En el gráfico, una línea diagonal divide el espacio ROC: los puntos por encima de esta línea, representan modelos con un buen desempeño de clasificación, mientras que los puntos por debajo reflejan un rendimiento deficiente.

En la tabla 30 se detallan los algoritmos evaluados y sus respectivas métricas. De acuerdo con estos resultados, el algoritmo J48 (una implementación del C4.5 (Quinlan, 1993)) obtuvo el mejor desempeño. Este método emplea el criterio de razón de ganancia (*gain ratio*), lo cual previene el sesgo hacia variables con rangos de valores amplios durante la selección de los nodos del modelo. Gracias a su alto rendimiento, particularmente en el área bajo la curva ROC (AUC), el J48 demuestra una notable capacidad para discriminar entre estudiantes con desempeño superior e inferior al promedio. Por tales motivos, se seleccionó este algoritmo para desarrollar los modelos de clasificación orientados

a detectar patrones de desempeño académico de los estudiantes de los departamentos de Cauca y Nariño en las Pruebas Saber 9^º.

Tabla 30. Evaluación de Algoritmos de Árboles de Decisión

No	Algoritmo	Exactitud	Precisión	Recall	F1	PRC	ROC	Time (sgs)
1	DecisionStump	0.652	0.640	0.653	0.633	0.600	0.597	0.01
2	ID3	0.618	0.64	0.637	0.643	0.606	0.625	0.27
3	J48	0.699	0.693	0.699	0.693	0.702	0.726	0.12
4	HoeffdingTree	0.698	0.693	0.699	0.686	0.713	0.721	0.28
5	Random Forest	0.675	0.670	0.675	0.671	0.714	0.713	4.51
6	RandomTree	0.645	0.638	0.645	0.640	0.615	0.636	0.16
7	REPTree	0.694	0.688	0.695	0.687	0.720	0.736	0.15
8	AdaBoostM1	0.664	0.655	0.664	0.640	0.704	0.703	0.17
9	Bagging	0.697	0.691	0.697	0.692	0.752	0.749	1.64

Fuente. Esta Investigación.

4.2 DEFINICIÓN DEL MÉTODO DE PRUEBA DE LOS MODELOS

Según Witten *et al.*, (2011) y Timaran *et al.*, (2021b) para probar la calidad de un modelo de clasificación y su validez, se divide los datos en dos conjuntos: Entrenamiento y Prueba.

Por esta razón, con el objetivo de establecer la mejor forma de particionar el conjunto de datos *sbr9_cauca_narino_final*, compuesto por 35,363 instancias, se aplicaron los filtros no supervisados de instancia *Randomize* y *Resample* de la herramienta WEKA. Estos filtros, permitieron dividir aleatoriamente y sin reemplazo, el conjunto, en seis subconjuntos: dos con una distribución 70-30 (70% para entrenamiento y 30% para prueba), dos con 80-20 (80% para entrenamiento y 20% para prueba) y dos con 90-10 (90% para entrenamiento y 10% para prueba).

Adicionalmente, se realizó una evaluación, utilizando el conjunto completo de datos, mediante la opción de validación cruzada (*cross-validation*). Para la construcción de los diferentes modelos de clasificación,

se empleó el algoritmo J48, utilizando los hiperparámetros predeterminados. Los resultados obtenidos en esta evaluación, se presentan en la Tabla 31.

Tabla 31. Evaluación de la Calidad del Modelo de Árboles de Decisión Mediante Particiones

Modelo	Conjunto de Entrenamiento	Conjunto de Prueba	Precisión	Recall	ROC
Tree_sbr9_cauca_nariño_final_70_30	24,754	10,609	0.73	0.81	0.73
Tree_sbr9_cauca_nariño_final_80_20	28,290	7,073	0.74	0.81	0.74
Tree_sbr9_cauca_nariño_final_90_10	21,655	2,407	0.74	0.80	0.74
Tree_sbr9_cauca_nariño_final-cross-v	35363		0.73	0.81	0.73

Fuente. Esta Investigación.

A partir de los resultados presentados en la tabla 31, se determinó que el método más adecuado para evaluar los modelos de clasificación (diseñados para identificar patrones de desempeño en Lenguaje, Matemáticas y Desempeño General en Cauca y Nariño) es el uso de un conjunto de prueba externo (Supplied Test Set). Se optó por una partición de 80 % para entrenamiento y 20 % para validación, debido a que esta configuración maximizó las métricas de desempeño del modelo, garantizando un equilibrio óptimo entre el aprendizaje y la capacidad de generalización.

4.3 DEFINICIÓN DE HIPERPARÁMETROS

El algoritmo J48, aplica un proceso de poda, al árbol de clasificación, una vez que ha sido generado (Hernández & Lorente, 2009). Entre los hiperparámetros considerados para la poda, el más relevante, es el factor de confianza *C* (*confidence factor*), el cual influye tanto en el tamaño del árbol como en su capacidad de predicción. Este valor oscila entre 0 y 1. A medida que la probabilidad disminuye, se requiere una diferencia

más significativa en los errores de predicción antes y después de la poda, para evitar eliminar nodos. Por defecto, este factor se establece en 0.25. Sin embargo, cuando su valor se reduce, el algoritmo, permite más operaciones de poda, dando lugar a árboles progresivamente más pequeños (García, 2016).

Otro hiperparámetro que afecta el tamaño del árbol, es el factor M , que determina el número mínimo de instancias o registros por nodo, necesarios para subdividir el árbol (Witten *et al.*, 2011).

Para determinar el modelo óptimo basado en árboles de decisión, se ajustaron los hiperparámetros de poda, utilizando la técnica de búsqueda por cuadrícula. Este método consiste, en definir una lista de hiperparámetros, junto con una métrica de rendimiento, permitiendo que el algoritmo pruebe todas las combinaciones posibles, para encontrar la configuración más adecuada.

El proceso se realizó por separado, para los conjuntos de datos *sbr9_cauca_final* y *sbr9_narino_final*, considerando, que los hiperparámetros de poda del algoritmo J48, pueden variar en cada caso. En particular, se evaluaron distintas combinaciones del factor de confianza C (*confidence factor*), con valores de 0.25, 0.15 y 0.05, junto con el factor M , definido como 1% y 0.5% del total de instancias de cada conjunto de datos.

Para la validación, se empleó la técnica de *cross-validation* con 10 *folds* y se analizaron métricas clave, como precisión, *recall* y curva ROC, enfocándose en aquellos modelos, cuyo desempeño general, se encontraba por debajo del promedio. Además, se evaluaron la cantidad de hojas y la profundidad del árbol generado.

Los resultados de este análisis, se presentan en la tabla 32, para el conjunto de datos *sbr9_cauca_final* (16,392 instancias) y en la tabla 33, los resultados para el conjunto *sbr9_narino_final* (18,971 instancias).

Tabla 32. Ajuste de Hiperparámetros para el Conjunto de Datos Sbr9_Cauca_Final

Modelo	C	M%	Precisión	Recall	ROC	Hojas	Tamaño
Tree_CO25M164	0.25	1	0.73	0.94	0.67	27	37
Tree_CO25M82	0.25	0.5	0.74	0.92	0.70	36	48
Tree_CO15M164	0.15	1	0.72	0.96	0.60	6	9
Tree_CO15M82	0.15	0.5	0.72	0.96	0.60	6	9
Tree_CO05M164	0.05	1	0.72	0.96	0.60	6	9
Tree_CO05M82	0.05	0.5	0.72	0.96	0.60	6	9

Fuente. Esta Investigación.

Según los resultados de la tabla 32, el modelo de árboles de decisión con mejor desempeño fue *Tree_CO25M82*, configurado con los hiperparámetros $C = 0.25$ y $M = 0.5\%$, destacándose especialmente, en las métricas de precisión y curva ROC. Estos hiperparámetros servirán, como referencia, para la construcción de distintos modelos de árboles de decisión, cuyo objetivo es, identificar patrones relacionados con el desempeño académico en las Competencias de Lenguaje, Matemáticas y Desempeño General de los estudiantes del departamento del Cauca, que presentaron las Pruebas Saber 9º, en el año 2017.

Dado que estos parámetros actuarán como referencia, podría ser necesario, ajustarlos en determinadas competencias, si el árbol resultante, presenta un nivel excesivo de poda.

Tabla 33. Ajuste de Hiperparámetros para el Conjunto de Datos Sbr9_Narino_Final

Modelo	C	M%	Precisión	Recall	ROC	Hojas	Tamaño
Tree_CO25M190	0.25	1	0.72	0.72	0.75	46	65
Tree_CO25M95	0.25	0.5	0.72	0.73	0.76	36	48
Tree_CO15M190	0.15	1	0.71	0.72	0.75	44	62
Tree_CO15M95	0.15	0.5	0.72	0.73	0.75	57	81
Tree_CO05M190	0.05	1	0.72	0.72	0.74	43	60
Tree_CO05M95	0.05	0.5	0.72	0.72	0.75	55	78

Fuente. Esta investigación.

De acuerdo con los resultados de la tabla 33, el modelo de árboles de decisión *Tree_CO25M95* (configurado con los hiperparámetros $C = 0.25$ y $M = 0.5$ %) alcanzó el mejor desempeño, destacando especialmente en las métricas de recall y curva ROC. Estos parámetros servirán como referencia para desarrollar los modelos orientados a identificar patrones en el desempeño académico de los estudiantes de Nariño en las Pruebas Saber 9º. El análisis se centrará en Lenguaje, Matemáticas y Desempeño General para comprender los factores incidentes en los resultados. Cabe precisar que, si bien estos hiperparámetros son la referencia inicial, se contempla un ajuste fino (fine-tuning) en caso de detectar un sobre-podamiento del árbol, con el fin de garantizar que el modelo final capture la complejidad real de los datos.

4.4 ÁRBOLES DE DECISIÓN PARA EL DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO EN LAS PRUEBAS SABER 9º DE LOS ESTUDIANTES DEL DEPARTAMENTO DEL CAUCA

Para desarrollar modelos de árboles de decisión, capaces de identificar patrones en el desempeño académico de los estudiantes de Noveno Grado del departamento del Cauca, quienes presentaron las Pruebas Saber 9º, en 2017, se utilizó el conjunto de datos *sbr9_cauca_final*. Este conjunto, está compuesto por 16,392 instancias y 18 atributos. A partir de él, se generaron tres subconjuntos de datos: *sbr9_cauca_lenguaje*, *sbr9_cauca_matemáticas* y *sbr9_cauca_general*, donde el atributo de clasificación (*target*), es el desempeño en cada competencia, categorizado en “sobre la media” y “bajo la media”.

En la tabla 34 (igual que en la tabla 27) se presenta, la distribución observada del desempeño en las Competencias de Lenguaje, Matemáticas y Desempeño General. Para evaluar la calidad de los modelos generados, cada conjunto de datos se dividió, en dos subconjuntos: 80% de las instancias se destinaron al entrenamiento del modelo y el 20% restante a la fase de prueba.

Tabla 34. Distribución del Desempeño Observado de los Estudiantes Caucanos en las Competencias de Lenguaje, Matemáticas y General

Conjunto	Desempeño				
	Bajo la Media	%	Sobre la Media	%	Total
sbr9_cauca_lenguaje	10,617	64.8	5,775	35.2	16,392
sbr9_cauca_matematicas	10,687	65.2	5,705	34.8	16,392
sbr9_cauca_general	11,028	67.3	5,364	32.7	16,392

Fuente. Esta Investigación.

Los resultados consolidados en la tabla 34 revelan un panorama crítico para el departamento del Cauca. De los 16,392 estudiantes evaluados en la prueba de Lenguaje, el 64.8 % (10,617 alumnos) no alcanzó la media nacional, frente a un 35.2 % que logró superarla. Esta tendencia se mantiene en Matemáticas, donde el 65.2 % se sitúa por debajo del promedio nacional y solo el 34.8 % (5,705 alumnos) se posiciona por encima de este. Esta brecha se acentúa drásticamente en el Desempeño General de las Pruebas Saber 9º, donde el rezago alcanza al 67.3 % de los evaluados, dejando apenas a un 32.7 % en el nivel superior.

4.4.1 Clasificación mediante Árboles de Decisión del Desempeño en la Competencia de Lenguaje de las Pruebas Saber 9º de los Estudiantes Caucanos

Para construir un modelo de clasificación, basado en árboles de decisión con el algoritmo J48, capaz de identificar patrones en el desempeño académico de los estudiantes del departamento del Cauca, en la Competencia de Lenguaje de las Pruebas Saber 9º, se utilizó el conjunto de datos *sbr9_cauca_lenguaje*. Como atributo de clasificación (target), se estableció el desempeño en Lenguaje, representado por el atributo *clase_desemp_lenguaje*.

El modelo, fue probado, mediante una partición del conjunto de datos, en dos subconjuntos:

- Conjunto de entrenamiento, con el 80% de las instancias (13,113 registros), de acuerdo con los parámetros definidos en la tabla 31.
- Conjunto de prueba, con el 20% de las instancias (3,279 registros) destinadas a la validación.

Para optimizar el rendimiento del modelo, el árbol de decisión, fue podado, utilizando los hiperparámetros de referencia definidos en la tabla 32: factor de confianza $C = 0.25$ y mínimo número de instancias por nodo $M = 0.5\%$.

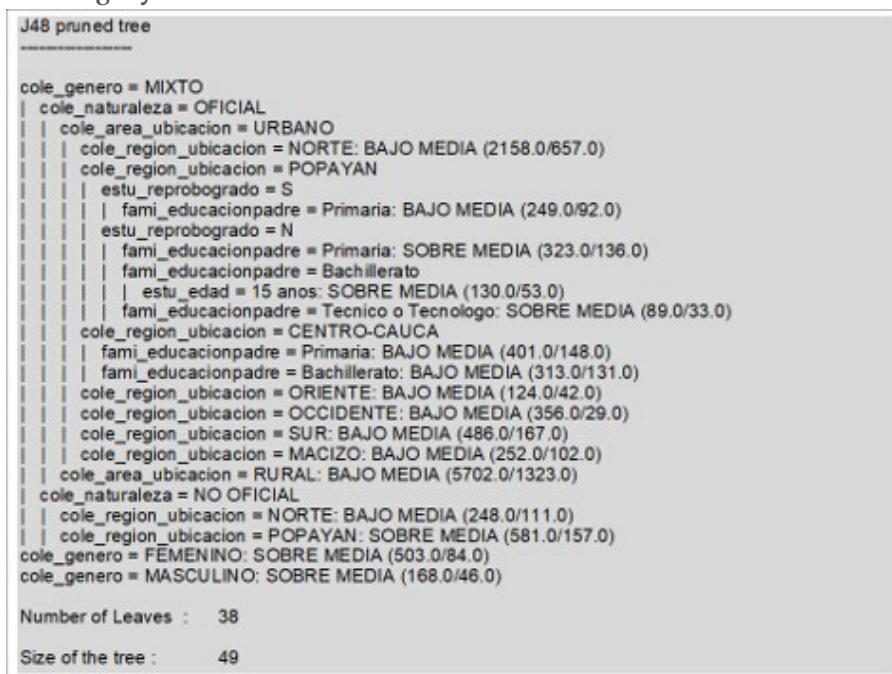
La validez de las reglas o patrones obtenidos, se determinó, analizando cada rama del árbol, desde el nodo raíz hasta el nodo hoja, mediante dos métricas clave:

- Soporte mínimo: porcentaje mínimo de instancias, que llegan a un nodo hoja, con respecto al total de instancias del conjunto de datos.
- Confianza mínima: porcentaje de instancias correctamente clasificadas, en cada nodo, con respecto al total de instancias que llegan al nodo hoja.

Estas métricas, fueron adaptadas de la técnica de Minería de Datos Reglas de Asociación (Agrawal *et al.*, 1993) y solo se conservaron las ramas, que superaran un soporte mínimo de 0.5% y una confianza mínima de 60%.

La Figura 19 ilustra el modelo de árbol de decisión resultante para la competencia de Lenguaje.

Figura 19. Árbol de Decisión Podado para la Competencia de Lenguaje de los Estudiantes Caucanos en las Pruebas Saber 9º.



Fuente. Esta Investigación.

Con el fin de interpretar los patrones identificados en el modelo de la figura 19, se seleccionaron las reglas más representativas del árbol de decisión resultante. Dichas reglas fueron filtradas bajo un criterio de soporte mínimo del 0.5 % y una confianza superior al 60 %, calculados sobre el total de las 13,113 instancias que conforman el conjunto de entrenamiento.

Entre los patrones más relevantes, organizados descendientemente según la métrica de soporte, se destacan los siguientes:

Regla 1. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en el área rural de un municipio del departamento del Cauca, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9º, es probable que este bajo la media nacional. El 43.5% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 76.8% de

los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 51.6% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 2. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en el área urbana de un municipio de la subregión Norte del departamento del Cauca, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 16.5% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 69.6% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 17.7% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 3. Si el estudiante, es de un colegio mixto, de carácter no oficial y ubicado en el municipio de Popayán, capital del departamento del Cauca, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este sobre la media nacional. El 4.4% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 73% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 9.2% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 4. Si el estudiante, es de un colegio femenino, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este sobre la media nacional. El 3.8% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 83.3% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 9.1% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 5. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana de un municipio de la subregión Sur del departamento del Cauca, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este bajo la media nacional. El 3.7% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 65.6% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 3.8% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 6. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana de un municipio de la subregión Centro del departamento del Cauca y la máxima educación del padre, es primaria, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este bajo la media nacional. El 3.1% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 63.1% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 3.0% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 7. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana de un municipio de la subregión Occidente del departamento del Cauca, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este bajo la media nacional. El 2.7% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 91.9% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 3.9% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 8. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana del municipio de Popayán, capital del departamento del Cauca, no reprobó ningún grado y la máxima educación del padre, es primaria, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este sobre la media nacional. El 2.5% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 60 % de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 4% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 9. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana de un municipio de la subregión Centro del departamento del Cauca y la máxima educación del padre, es bachillerato, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este bajo la media nacional. El 2.4% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 60.0% de los estudiantes que se clasifican en esta categoría, han sido

correctamente identificados y el 2.1% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 10. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana del municipio de Popayán, capital del departamento del Cauca, reprobó algún grado y la máxima educación del padre, es primaria, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este bajo la media nacional. El 1.9% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 63.1% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 1.9% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 11. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana de un municipio de la subregión Macizo del departamento del Cauca, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este bajo la media nacional. El 1.9% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 60% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 1.8% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 12. Si el estudiante, es de un colegio masculino, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este sobre la media nacional. El 1.3% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 72.6% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 2.6% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 13. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana del municipio de Popayán, capital del departamento del Cauca, no reprobó ningún grado, la máxima educación del padre, es bachillerato y tiene 15 años, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este sobre la media nacional. El 1% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 60 % de los estudiantes que se clasifican

en esta categoría, han sido correctamente identificados y el 1.7% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 14. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana de un municipio de la subregión Oriente del departamento del Cauca, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este bajo la media nacional. El 0.9% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 66.1% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 1.0% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 15. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana del municipio de Popayán, capital del departamento del Cauca, no reprobó ningún grado y la máxima educación del padre, es técnico o tecnólogo, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable que este sobre la media nacional. El 0.7% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 62.9 % de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 1.2% de los estudiantes observados sobre la media, cumplen este patrón.

Por otra parte, la figura 20 presenta la evaluación del modelo para la prueba de Lenguaje, fundamentada en las métricas derivadas de la Matriz de Confusión.

Figura 20. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Lenguaje de los Estudiantes Caucanos en las Pruebas Saber 9°.

=== Summary ===									
Correctly Classified Instances	2330							71.0582 %	
Incorrectly Classified Instances	949							28.9418 %	
Kappa statistic								0.2821	
Mean absolute error								0.3936	
Root mean squared error								0.4441	
Relative absolute error								86.3383 %	
Root relative squared error								93.1108 %	
Total Number of Instances	3279								
=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,910	0,660	0,719	0,910	0,803	0,312	0,695	0,769	BAJO MEDIA
	0,340	0,090	0,670	0,340	0,451	0,312	0,695	0,566	SOBRE MEDIA
Weighted Avg.	0,711	0,461	0,702	0,711	0,680	0,312	0,695	0,698	
=== Confusion Matrix ===									
a	b	<-- classified as							
1940	192	a = BAJO MEDIA							
757	390	b = SOBRE MEDIA							

Fuente. Esta Investigación.

Como se muestra en la figura 20, el modelo de árbol de decisión, obtuvo una exactitud del 71.1%, lo que significa, que clasifica correctamente el 71.1% de las instancias del conjunto de prueba, mientras que el 28.9% de las instancias fueron clasificadas incorrectamente.

El error cuadrático medio (RMSE, por sus siglas en inglés: *Root Mean Squared Error*) del modelo, es de 0.44, lo que indica, que el ajuste es adecuado, ya que los valores observados (reales) y los predichos, están relativamente próximos entre sí.

Asimismo, según la Matriz de Confusión, el modelo de clasificación evaluado sobre el conjunto de prueba *sbr9_cauca_lenguaje*, compuesto por 3,279 estudiantes, presentó los siguientes resultados:

- 59.2% de los estudiantes con desempeño en Lenguaje “bajo la media”, fueron correctamente clasificados como verdaderos negativos (VN), lo que equivale a 1,940 estudiantes.

- 11.9% de los estudiantes con desempeño “sobre la media”, fueron correctamente identificados como verdaderos positivos (VP), correspondiente a 390 estudiantes.
- 5.9% de los estudiantes con desempeño “bajo la media”, fueron clasificados erróneamente como “sobre la media” (falsos positivos, FP), lo que representa 192 estudiantes.
- 23.1% de los estudiantes con desempeño “sobre la media”, fueron incorrectamente categorizados como “bajo la media” (falsos negativos, FN), equivalente a 757 estudiantes.

En cuanto a la exactitud del modelo de árboles de decisión por clase, se observa, que, para el caso de los estudiantes caucanos con desempeño bajo la media (negativos) en Lenguaje, el modelo alcanza una precisión de predicción de 0.719. Esto significa, que el 71.9% de los casos clasificados como “bajo la media”, fueron correctamente identificados.

La especificidad o tasa de verdaderos negativos (*TP Rate*), junto con el *recall*, es de 0.91, lo que indica, que el modelo, logra detectar correctamente, al 91% de los estudiantes caucanos, que realmente tienen un desempeño bajo la media.

Por otro lado, la tasa de falsos negativos (*FP Rate*) es de 0.66, lo que implica, que el 66% de los estudiantes con rendimiento sobre la media, fueron erróneamente clasificados como bajo la media.

El modelo también exhibe una *F-Measure* (*F1-Score*) de 0.803, lo que representa un equilibrio sólido entre precisión y *recall* en la clasificación de estudiantes con desempeño bajo la media. Este resultado indica, que el modelo, no solo logra identificar correctamente a estos estudiantes, sino que también, minimiza los errores en las predicciones negativas.

Finalmente, el área bajo la curva PRC (*PRC Área*) es de 0.769, lo que refleja un desempeño razonablemente bueno, en la clasificación de los estudiantes con rendimiento bajo la media. Cuanto mayor sea esta área, mejor será la capacidad del modelo para equilibrar precisión y *recall*.

Con un 76.9%, el modelo demuestra, una sólida habilidad, para gestionar conjuntos de datos desbalanceados, como ocurre en el caso del *sbr9_cauca_lenguaje*.

En el caso de los estudiantes caucanos con desempeño sobre la media en Lenguaje, el modelo de árboles de decisión, presenta una precisión de predicción de 0.67, lo que significa que el 67% de los casos clasificados como “sobre la media”, fueron correctamente identificados.

Sin embargo, la sensibilidad o tasa de verdaderos positivos (*TP Rate*) y *recall* del modelo es de 0.34, lo que indica, que solo logra reconocer correctamente al 34% de los estudiantes, que realmente tienen un desempeño sobre la media.

Por otro lado, la tasa de falsos positivos (*FP Rate*), es de 0.090, lo que implica, que el 9% de los estudiantes con desempeño bajo la media, fueron erróneamente clasificados como sobre la media.

El *F-Measure* (*F1-Score*), tiene un valor de 0.451, lo que sugiere un desequilibrio significativo entre precisión y *recall*. Esto indica, que el modelo no está funcionando de manera óptima, ya que su baja sensibilidad, le impide identificar correctamente a suficientes estudiantes con desempeño sobre la media.

Por último, el área bajo la curva PRC (*PRC Área*) es de 0.566, lo que refleja una capacidad moderada para diferenciar entre los estudiantes con desempeño sobre la media y aquellos con desempeño bajo la media, en la competencia de Lenguaje de las Pruebas Saber 9º.

El área bajo la curva ROC (*ROC Área*) del modelo es de 0.7, un valor superior a 0.5, lo que indica, que el modelo tiene una capacidad aceptable, para diferenciar entre los estudiantes caucanos con desempeño bajo la media y aquellos con desempeño sobre la media. En otras palabras, el modelo, logra detectar más verdaderos negativos (VN) y verdaderos positivos (VP) en comparación con los falsos negativos (FN) y falsos positivos (FP), lo que refleja, un rendimiento razonable, en la clasificación de los estudiantes según su desempeño en Lenguaje.

En general, el análisis de las métricas, indica, que el modelo tiene un mejor desempeño en la clasificación de los estudiantes con desempeño bajo la media, en la competencia de Lenguaje. En particular, la métrica PRC, es más alta que la ROC, lo que sugiere, que el modelo funciona de manera más efectiva, cuando se evalúan conjuntamente precisión y *recall*.

Esta investigación se centró en los estudiantes con desempeño bajo la media, ya que es en este grupo, donde resulta fundamental, identificar los factores que influyen en el rendimiento académico en distintas competencias. Comprender los perfiles de los estudiantes caucanos, cuyo desempeño en Lenguaje en las Pruebas Saber 9º, no ha sido óptimo, permitirá a las directivas de las instituciones de Educación Básica Secundaria, a la Secretaría de Educación del Departamento del Cauca y al Ministerio de Educación Nacional (MEN) tomar decisiones basadas en datos, que contribuyan a mejorar la calidad educativa en el nivel básico secundario, con especial énfasis en el desarrollo de habilidades en la Competencia de Lenguaje.

4.4.2 Clasificación mediante Árboles de Decisión del Desempeño en la Competencia de Matemáticas en las Pruebas Saber 9º de los Estudiantes Caucaños

Para construir un modelo de clasificación, basado en árboles de decisión con el algoritmo J48, enfocado en identificar patrones de desempeño académico en la Competencia de Matemáticas de las Pruebas Saber 9º, que presentaron los estudiantes del departamento del Cauca, se utilizó el conjunto de datos *sbr9_cauca_matematicas*. Se escogió como atributo clase o *target*, el desempeño en Matemáticas representado por el atributo *clase_desemp_matematicas*.

El modelo se probó, dividiendo el conjunto de datos en dos subconjuntos:

- Conjunto de entrenamiento, que contiene el 80% de las instancias (13,113 registros), de acuerdo con los parámetros previamente establecidos.

- Conjunto de prueba, compuesto por el 20% de las instancias (3,279 registros) para evaluar el rendimiento del modelo.

Para optimizar su desempeño, se aplicó un Proceso de Poda con los hiperparámetros factor de confianza $C = 0.30$ y mínimo número de instancias por nodo $M = 0.5\%$. Este ajuste se realizó, tras comprobar que un C de 0.30 ofrecía mejores resultados en comparación con el valor de referencia $C = 0.25$.

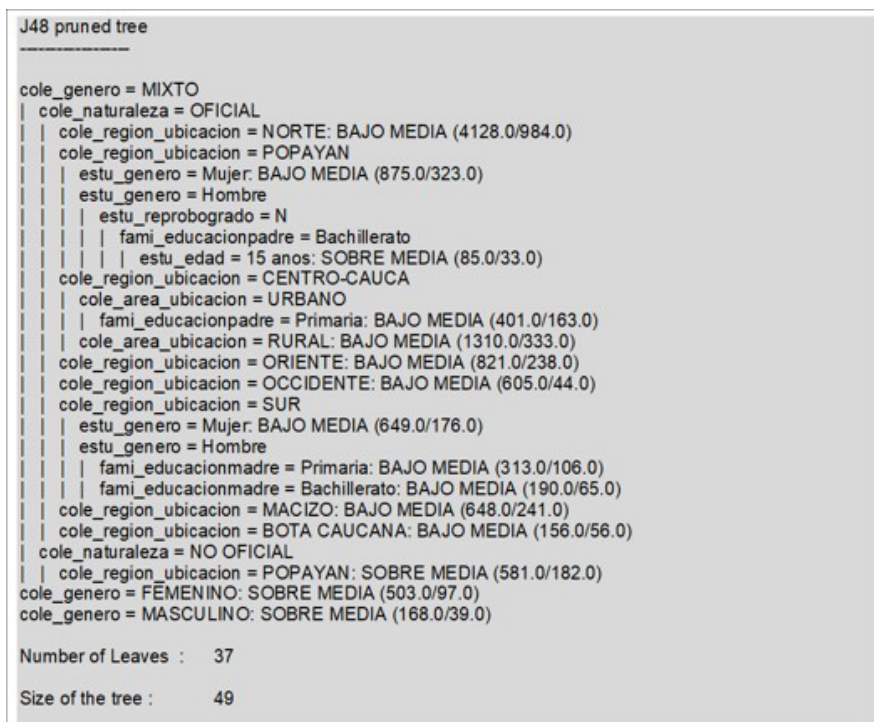
Para evaluar la validez de los patrones generados por el modelo de árboles de decisión, se analizaron las ramas del árbol desde el nodo raíz hasta el nodo hoja, utilizando dos métricas clave:

- Soporte mínimo: Representa el porcentaje mínimo de instancias, que llegan a un nodo hoja, en relación con el total de instancias en el conjunto de datos.
- Confianza mínima: Indica el porcentaje mínimo de instancias correctamente clasificadas en cada nodo, con respecto al total de instancias que llegan al nodo hoja.

Estas métricas fueron adaptadas de la técnica de Minería de Datos, conocida como Reglas de Asociación (Agrawal *et al.*, 1993). Para garantizar la calidad del modelo, solo se conservaron aquellas ramas, que superaran un soporte mínimo del 0.5% y una confianza mínima del 60%.

La Figura 19 presenta el modelo de árbol de decisión resultante para la Competencia de Matemáticas.

Figura 21. Árbol de Decisión Podado para la Competencia de Matemáticas de los Estudiantes Caucanos en las Pruebas Saber 9º.



Fuente. Esta Investigación.

Con el fin de interpretar los patrones identificados en el modelo de la figura 21, se seleccionaron las reglas más representativas del árbol de decisión resultante. Dichas reglas fueron filtradas bajo un criterio de soporte mínimo del 0.5 % y una confianza superior al 60 %, calculados sobre el total de las 13,113 instancias que conforman el conjunto de entrenamiento

A continuación, se presentan los patrones más relevantes, organizados descendientemente según la métrica de soporte:

Regla 1. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en un municipio de la subregión Norte del departamento del Cauca, entonces su desempeño académico en la Prueba de Matemáticas

del Saber 9°, es probable, que este bajo la media nacional. El 31.5% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 76.2% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 48.1% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 2. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en el área rural de un municipio de la subregión Centro del departamento del Cauca, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 10% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 74.6% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 15.3% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 3. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el municipio de Popayán, capital del departamento del Cauca y es mujer, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 6.7% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 63.1% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 10.2% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 4. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en un municipio de la subregión Oriente del departamento del Cauca, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 6.3% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 71% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 9.6% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 5. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en un municipio de la subregión Sur del departamento del Cauca y es mujer, entonces su desempeño académico en la Prueba de

Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 4.9% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 72.9% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 7.6% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 6. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en un municipio de la subregión Macizo del departamento del Cauca, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 4.9% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 62.8% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 7.6% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 7. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en un municipio de la subregión Occidente del departamento del Cauca, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 4.6% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 92.7% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 7.1% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 8. Si el estudiante, es de un colegio mixto, de carácter no oficial y ubicado en el municipio de Popayán, capital del departamento del Cauca, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 4.4% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 68.7% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 12.8% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 9. Si el estudiante, es de un colegio femenino, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 3.8% del total de estudiantes

caucanos evaluados, se clasifican de esta manera. El 80.7 % de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 11.1% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 10. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el área urbana de un municipio de la subregión Centro del departamento del Cauca y la máxima educación del padre es primaria, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 3.1% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 60% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 4.7% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 11. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en un municipio de la subregión Sur del departamento del Cauca, es hombre y la máxima educación de la madre es primaria, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 2.4% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 66.1% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 3.6% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 12. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en un municipio de la subregión Sur del departamento del Cauca, es hombre y la máxima educación de la madre es bachillerato, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 1.4% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 65.8% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 2.6% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 13. Si el estudiante, es de un colegio masculino, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 1.3% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 76.8 % de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 3.7% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 14. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en un municipio de la subregión Bota Caucana del departamento del Cauca, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 1.2% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 64.1% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 1.8% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 15. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el municipio de Popayán, capital del departamento del Cauca, es hombre, no reprobó grado, la educación máxima del padre es bachillerato y tiene 15 años, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 0.6% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 61.2% de los estudiantes que se clasifican en esta categoría, han sido correctamente identificados y el 1.9% de los estudiantes observados sobre la media, cumplen este patrón.

Por otra parte, en la figura 22 se presenta la evaluación del modelo para la prueba de Matemáticas, fundamentada en las métricas derivadas de la Matriz de Confusión.

Figura 22. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Matemáticas de los Estudiantes Caucanos en las Pruebas Saber 9º.

```

=== Summary ===
Correctly Classified Instances    2290      69.8384 %
Incorrectly Classified Instances  989      30.1616 %
Kappa statistic                  0.2427
Mean absolute error              0.3988
Root mean squared error          0.4486
Relative absolute error          87.4707 %
Root relative squared error      93.5997 %
Total Number of Instances       3279

=== Detailed Accuracy By Class ===
      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
      0.931   0.721   0.699    0.931   0.799    0.287  0.691    0.769   BAJO MEDIA
      0.279   0.069   0.693    0.279   0.398    0.287  0.691    0.563   SOBRE MEDIA
Weighted Avg. 0.698   0.488   0.697    0.698   0.656    0.287  0.691    0.696

=== Confusion Matrix ===
  a  b  <- classified as
1963 145 | a = BAJO MEDIA
 844 327 | b = SOBRE MEDIA

```

Fuente. Esta Investigación.

Como se muestra en la figura 22, el modelo de árbol de decisión, alcanzó una exactitud del 69.8%, lo que significa, que clasifica correctamente la mayoría de las instancias del conjunto de prueba. Sin embargo, el 30.2% de las instancias, fueron clasificadas incorrectamente.

El Error Cuadrático Medio (RMSE) del modelo es de 0.45, lo que indica un buen ajuste, ya que la diferencia entre los datos observados (reales) y los valores predichos, no es significativa.

Según la Matriz de Confusión, el modelo de clasificación evaluado sobre el conjunto de prueba *sbr9_cauca_matemáticas*, compuesto por 3,279 estudiantes, obtuvo los siguientes resultados:

- 59.9% de los estudiantes con desempeño en Matemáticas “bajo la media”, fueron correctamente clasificados como verdaderos negativos (VN), lo que equivale a 1,963 estudiantes.

- 10% de los estudiantes con desempeño “sobre la media”, fueron correctamente identificados como verdaderos positivos (VP), correspondiente a 327 estudiantes.
- 4.4% de los estudiantes con rendimiento “bajo la media”, fueron clasificados erróneamente como “sobre la media” (falsos positivos, FP), lo que representa 145 estudiantes.
- 25.7% de los estudiantes con desempeño “sobre la media”, fueron incorrectamente categorizados como “bajo la media” (falsos negativos, FN), equivalente a 844 estudiantes.

En cuanto a la exactitud del modelo de árboles de decisión por clase, se observa, que, para los estudiantes caucanos con desempeño bajo, la media en Matemáticas, el modelo presenta una precisión de predicción de 0.699. Esto significa, que el 69.9% de los casos clasificados como “bajo la media”, fueron correctamente identificados.

La especificidad o tasa de verdaderos negativos (*TP Rate*), junto con el *recall*, es de 0.931, lo que indica, que el modelo logra detectar correctamente al 93.1% de los estudiantes que realmente tienen un desempeño bajo la media.

Por otro lado, la tasa de falsos negativos (*FP Rate*) es de 0.721, lo que implica, que el 72.1% de los estudiantes con rendimiento sobre la media, fueron erróneamente clasificados como bajo la media.

El modelo exhibe una *F-Measure* (*F1-Score*) de 0.799, lo que representa un equilibrio sólido entre precisión y *recall* en la clasificación de estudiantes con desempeño bajo la media. Esto indica, que el modelo no solo logra identificar correctamente a estos estudiantes, sino que también minimiza los errores en las predicciones negativas.

Por último, el área bajo la curva PRC (*PRC Área*), es de 0.769, lo que refleja un desempeño razonablemente bueno, en la clasificación de los estudiantes con rendimiento bajo la media. La curva PRC, es especialmente útil, cuando se trabaja con datos desbalanceados, como en el caso

del *sbr9_cauca_matemáticas*, ya que se enfoca, en la capacidad del modelo para clasificar correctamente los ejemplos negativos. Un valor de 0.769, sugiere que el modelo es bastante efectivo en esta tarea.

En el caso de los estudiantes caucanos con desempeño sobre la media en Matemáticas, el modelo de árboles de decisión, presenta una precisión de predicción de 0.693, lo que significa, que clasifica correctamente el 69.3% de los casos predichos como “sobre la media”.

Sin embargo, la sensibilidad o tasa de verdaderos positivos (*TP Rate*) y *recall*, es de 0.279, lo que indica, que el modelo solo logra identificar correctamente al 27.9% de los estudiantes que realmente tienen un desempeño sobre la media.

Por otro lado, la tasa de falsos positivos (*FP Rate*) es de 0.069, lo que significa que el 6.9% de los estudiantes con desempeño en Matemáticas bajo la media, fueron clasificados erróneamente como sobre la media.

El *F-Measure* (*F1-Score*), es de 0.398, lo que sugiere, que el modelo no está funcionando de manera óptima. Este valor refleja, un desequilibrio importante entre precisión y *recall*, lo que indica, que el modelo no identifica correctamente a un número suficiente de estudiantes con rendimiento sobre la media debido al bajo *recall*.

Finalmente, el área bajo la curva PRC (*PRC Área*), es de 0.563, lo que indica, que el modelo tiene una capacidad de discriminación moderada, para identificar correctamente, a los estudiantes sobre la media en la Competencia de Lenguaje de las Pruebas Saber 9º.

El área bajo la curva ROC (*ROC Área*) del modelo, es de 0.70, un valor superior a 0.5, lo que indica, que tiene una capacidad aceptable para distinguir entre los estudiantes caucanos con desempeño bajo la media y aquellos con desempeño sobre la media en la Prueba de Matemáticas del Saber 9º.

En otras palabras, el modelo logra detectar más verdaderos negativos (VN) y verdaderos positivos (VP) que falsos negativos (FN) y falsos positivos (FP), lo que refleja, un rendimiento adecuado en la clasificación de los estudiantes según su desempeño.

En general, el análisis de las métricas refleja, un mejor desempeño del modelo en la clasificación de los estudiantes con desempeño bajo la media en la Prueba de Matemáticas del Saber 9º. Para esta categoría, la métrica PRC, es superior a la ROC, lo que indica, que el modelo funciona más eficientemente cuando se evalúan conjuntamente precisión y *recall*.

Esta investigación se enfocó, en identificar los factores asociados al bajo desempeño en las distintas competencias evaluadas en las Pruebas Saber 9º, de los estudiantes caucanos. En este caso interesa el desempeño en Matemáticas. Comprender los perfiles de los estudiantes caucanos, cuyo desempeño en esta competencia es inferior a la media nacional, permitirá a las directivas de instituciones de Educación Básica Secundaria, a la Secretaría de Educación del Departamento del Cauca y al Ministerio de Educación Nacional (MEN), tomar decisiones fundamentadas en datos. Estas decisiones contribuirán, al mejoramiento de la calidad educativa en el nivel básico secundario, con especial énfasis en el fortalecimiento de las habilidades Matemáticas de los estudiantes caucanos.

4.4.3 Clasificación mediante Árboles de Decisión del Desempeño General en las Pruebas Saber 9º de los Estudiantes Caucaños.

Para desarrollar un modelo de clasificación, basado en árboles de decisión, utilizando el algoritmo J48, con el objetivo de identificar patrones en el desempeño general de los estudiantes del departamento del Cauca en las Pruebas Saber 9º, se trabajó con el conjunto de datos *sbr9_cauca_general*. Como variable objetivo o target, se seleccionó el Desempeño General, representado por el atributo *clase_desemp_general*.

Para evaluar la Efectividad del Modelo, el conjunto de datos se dividió en dos partes: un conjunto de entrenamiento, que representó el 80% del total de instancias (13,113), siguiendo los parámetros previamente establecidos y un conjunto de prueba, conformado por el 20% restante (3,279 instancias).

Además, se aplicó una poda al árbol de decisión, utilizando los siguientes hiperparámetros: un factor de confianza de $C=0.25$ y un mínimo número de instancias por nodo de $M=0.5\%$.

El Soporte Mínimo, se define, como el porcentaje mínimo de instancias que llegan a un nodo hoja en relación con el total de instancias del conjunto de datos, mientras que la Confianza Mínima, representa el porcentaje de instancias correctamente clasificadas en cada nodo, con respecto al total de instancias que llegan a ese nodo hoja. Estas métricas fueron adaptadas de la técnica de Minería de Datos Reglas de Asociación (Agrawal *et al.*, 1993).

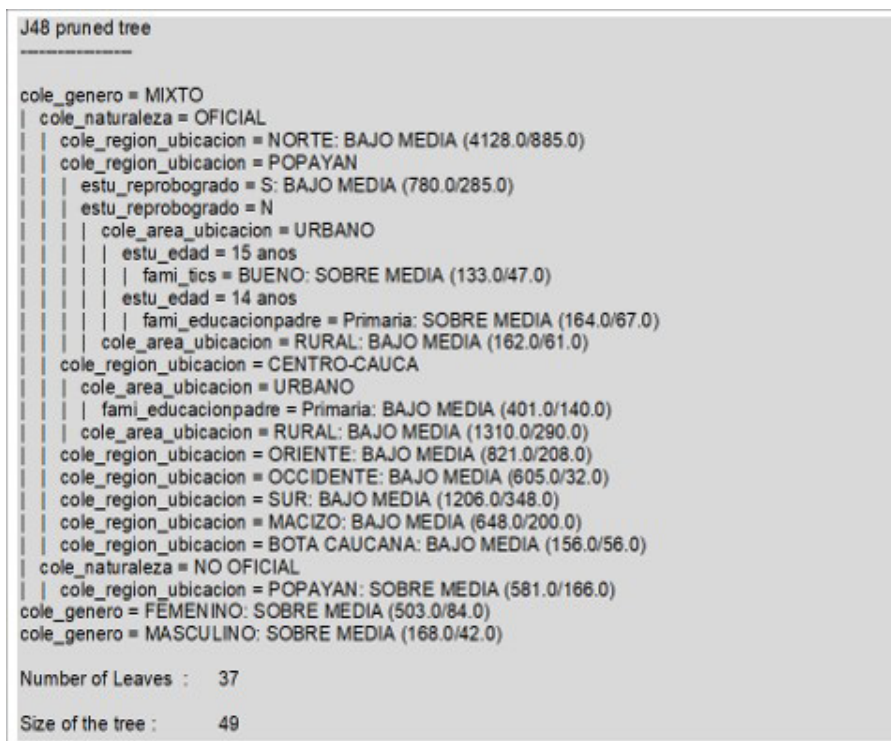
Para garantizar la calidad del modelo, se conservaron únicamente, las ramas que superaron un soporte mínimo de 0.5% y una confianza mínima de 60% .

En la figura 23, se muestra el modelo de árbol de decisión resultante para el Desempeño General en las pruebas Saber 9°.

Con el fin de interpretar los patrones identificados en el modelo de la figura 23, se seleccionaron las reglas más representativas del árbol de decisión resultante. Dichas reglas fueron filtradas bajo un criterio de soporte mínimo del 0.5% y una confianza superior al 60% , calculados sobre el total de las 13,113 instancias que conforman el conjunto de entrenamiento. Entre los patrones más importantes ordenados por soporte están:

Regla 1. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el uno de los municipios de la zona Norte del departamento del Cauca, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 31.5% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 78.6% de los estudiantes que se clasifican así, están correctamente clasificados y el 46.7% de los estudiantes observados bajo la media, cumplen este patrón.

Figura 23. Árbol de Decisión Podado del Desempeño General de los Estudiantes Caucanos en las Pruebas Saber 9°.



Fuente. Esta Investigación.

Regla 2. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en uno de los municipios de la región Centro del departamento del Cauca y está en el área rural, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 10% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 77.9% de los estudiantes que se clasifican así, están correctamente clasificados y el 14.8% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 3. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en uno de los municipios de la región Sur del departamento del Cauca, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 9.2% del total de estudiantes

caucanos evaluados, se clasifican de esta manera. El 71.1% de los estudiantes que se clasifican así, están correctamente clasificados y el 13.7% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 4. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en uno de los municipios de la región Oriente del departamento del Cauca, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 6.3% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 74.7% de los estudiantes que se clasifican así, están correctamente clasificados y el 9.3% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 5. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el municipio de Popayán, capital del departamento del Cauca y reprobó algún grado, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 5.9% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 63.5% de los estudiantes que se clasifican así, están correctamente clasificados y el 8.8% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 6. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en uno de los municipios de la región Macizo del departamento del Cauca, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 4.9% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 69.1% de los estudiantes que se clasifican así, están correctamente clasificados y el 7.3% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 7. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en uno de los municipios de la región Occidente del departamento del Cauca, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 4.6% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 94.7% de los estudiantes que se clasifican así, están correctamente clasificados y el 6.9% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 8. Si el estudiante, es de un colegio mixto, de carácter no oficial, ubicado en el municipio de Popayán, capital del departamento del Cauca, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 4.4% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 71.4% de los estudiantes que se clasifican así, están correctamente clasificados y el 3.9% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 9. Si el estudiante, es de un colegio femenino, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 3.8% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 83.3% de los estudiantes que se clasifican así, están correctamente clasificados y el 2.0% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 10. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en uno de los municipios de la región Centro del departamento del Cauca, en el área urbana y la máxima educación del padre, es Primaria, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 3.1% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 65.1% de los estudiantes que se clasifican así, están correctamente clasificados y el 4.5% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 11. Si el estudiante, es de un colegio masculino, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 1.3% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 75% de los estudiantes que se clasifican así, están correctamente clasificados y el 1.0% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 12. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el municipio de Popayán, capital del departamento del Cauca, no reprobó ningún grado, está en el área urbana, tiene 14 años y el máximo nivel educativo del padre es Primaria, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media

nacional. El 1.3% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 60% de los estudiantes que se clasifican así, están correctamente clasificados y el 3.8% de los estudiantes observados sobre la media, cumplen este patrón.

Regla 13. Si el estudiante, es de un colegio mixto, de carácter oficial y ubicado en uno de los municipios de la región Bota Caucana del departamento del Cauca, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 1.2% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 64.1% de los estudiantes que se clasifican así, están correctamente clasificados y el 1.8% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 14. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el municipio de Popayán, capital del departamento del Cauca, no reprobó ningún grado y está en el área rural, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 1.2% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 62.3% de los estudiantes que se clasifican así, están correctamente clasificados y el 1.8% de los estudiantes observados bajo la media, cumplen este patrón.

Regla 15. Si el estudiante, es de un colegio mixto, de carácter oficial, ubicado en el municipio de Popayán, capital del departamento del Cauca, no reprobó ningún grado, está en el área urbana, tiene 15 años y el índice TICs de su hogar, es bueno, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 1% del total de estudiantes caucanos evaluados, se clasifican de esta manera. El 64.7% de los estudiantes que se clasifican así, están correctamente clasificados y el 3.1% de los estudiantes observados sobre la media, cumplen este patrón.

Asimismo, la figura 24 presenta la evaluación del modelo para el Desempeño General en las pruebas Saber 9°, fundamentada en las métricas derivadas de la matriz de confusión.

Figura 24. Métricas de Evaluación del Modelo de Árbol de Decisión para el Desempeño General de los Estudiantes Caucanos en las Pruebas Saber 9º.

=== Summary ===									
Correctly Classified Instances	2374					72.4001 %			
Incorrectly Classified Instances	905					27.5999 %			
Kappa statistic					0.2748				
Mean absolute error					0.3749				
Root mean squared error					0.4339				
Relative absolute error					84.9865 %				
Root relative squared error					92.2535 %				
Total Number of Instances	3279								
=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.930	0.693	0.731	0.930	0.819	0.313	0.715	0.804	BAJO MEDIA
	0.307	0.070	0.683	0.307	0.423	0.313	0.715	0.568	SOBRE MEDIA
Weighted Avg.	0.724	0.488	0.715	0.724	0.688	0.313	0.715	0.727	
=== Confusion Matrix ===									
a	b	<-- classified as							
2042	154	a = BAJO MEDIA							
751	332	b = SOBRE MEDIA							

Fuente. Esta Investigación.

Como se muestra en la figura 24, el modelo de árbol de decisión obtenido, presenta una exactitud del 72.4%, lo que significa, que clasifica correctamente, el 72.4% de las instancias del conjunto de prueba, mientras que un 27.6% de ellas, son clasificadas incorrectamente.

El Error Cuadrático Medio (RMSE, por sus siglas en inglés: *Root Mean Squared Error*) del modelo, es de 0.44, lo que indica un buen ajuste. En otras palabras, los valores predichos por el modelo, están relativamente cercanos a los datos observados (reales), lo que sugiere una adecuada representación del comportamiento de las instancias en el conjunto de prueba.

Igualmente, de acuerdo con la Matriz de Confusión de la figura 24, del total de 3,279 estudiantes (instancias) que conforman el conjunto de prueba del *sbr9_cauca_general*, el modelo predice correctamente, al 62.3% de estudiantes, con Desempeño General “bajo la media”, (verdaderos negativos VN), que corresponde a 2042 estudiantes y al 10.1% de estudiantes con desempeño académico “sobre la media” (verdaderos

positivos VP), que corresponde a 332 estudiantes. Además, el modelo predice incorrectamente, al 4.7% de estudiantes que están “bajo la media” como “sobre la media” (falsos positivos FP), correspondiente a 154 estudiantes y al 22.9% de estudiantes que están “sobre la media” como “bajo la media” (falsos negativos FN), correspondiente a 751 estudiantes.

En cuanto a la exactitud del modelo de árboles de decisión por clase, se puede observar, que para el caso de los estudiantes caucanos que están bajo la media (negativos) en el Desempeño General de las Pruebas Saber 9º, el modelo tiene una precisión de predicción de 0.731, lo que significa, que, del total de casos predichos que están bajo la media, el 73.1% están correctos.

La Especificidad o tasa de verdaderos negativos (*TP Rate*) y *Recall* del modelo, es de 0.93, lo cual indica, que, el modelo es capaz de identificar al 93% de los estudiantes caucanos, que realmente están bajo la media.

Por otra parte, la tasa de falsos negativos (*FP Rate*) del modelo, es de 0.693, lo que significa, que el 69.3% de estudiantes que estaban sobre la media, fueron predichos como bajo la media.

La *F-Measure* o *F1-Score*, es de 0.819, que es el promedio armónico entre la precisión y el *recall* de los que están bajo la media, lo que significa, que el modelo tiene un rendimiento sólido y balanceado, entre precisión y *recall*. El modelo identifica correctamente, los estudiantes que están bajo la media (instancias negativas (*recall*)) y no comete demasiados errores en las predicciones negativas (*precisión*).

Finalmente, el Área Bajo la Curva (PRC) (*PRC Área*) es de 0.804. este valor indica, que el modelo tiene una muy buena capacidad para identificar correctamente a los estudiantes caucanos que están bajo la media en el Desempeño General. Con un 80.4%, el modelo tiene una capacidad más sólida, para equilibrar correctamente las predicciones “bajo la media”. Un PRC más alto, significa que el modelo es efectivo en minimizar los falsos positivos y falsos negativos. Útil cuando se trata de conjuntos de datos desbalanceados, como es el caso del conjunto de datos *sbr9_cauca_general*.

Para el caso de los estudiantes caucanos que están sobre la media (positivos) en el Desempeño General en las Pruebas Saber 9^o, el modelo tiene una precisión de predicción de 0.683, lo que significa que, del total de casos predichos que están sobre la media, el 68.3% están correctos. La Sensibilidad o tasa de verdaderos positivos (*TP Rate*) y *Recall* del modelo, es de 0.307, lo cual indica, que, el modelo es capaz de identificar al 30.7% de los estudiantes caucanos, cuyo Desempeño General, está realmente, sobre la media. Por otra parte, la tasa de falsos positivos (*FP Rate*) del modelo, es de 0.070, lo que significa, que el 7% de estudiantes que estaban sobre la media, fueron predichos como bajo la media. La *F-Measure* o *F1-Score*, es de 0.423, lo que significa, que el modelo no está funcionando particularmente bien. Este valor indica, que hay un desequilibrio importante entre *precisión* y *recall*. El modelo está fallando, en identificar correctamente, suficientes estudiantes que estén sobre la media (instancias positivas), debido al bajo *recall*. Finalmente, el Área Bajo la Curva (PRC) (*PRC Área*), es de 0.568. Este valor indica, que el modelo tiene una capacidad de discriminación moderada, para identificar correctamente, a los estudiantes caucanos que están sobre la media en el Desempeño General. Por lo tanto, un PRC de 0.568, sugiere, que el modelo no es muy efectivo para clasificar correctamente a los estudiantes cuyo Desempeño General, esta sobre la media.

El Área Bajo la Curva (ROC) (*ROC Área*) del modelo, es de 0.72 (mayor que 0.5), lo que indica, que, el modelo tiene una aceptable probabilidad de que sea capaz de distinguir los estudiantes caucanos que están etiquetados como “bajo la media”, de los estudiantes caucanos que están etiquetados como “sobre la media”, ya que, es capaz de detectar más verdaderos negativos (VN) y verdaderos positivos (VP) que falsos negativos (FN) y falsos positivos (FP).

En la combinación de estas métricas, se aprecia un mejor desempeño del modelo, en la clasificación de los estudiantes etiquetados como “bajo la media” que, como se expresó anteriormente, es para el tipo de estudiantes que interesa para esta investigación. Además, para esta clase, la métrica PRC, es más alto que el ROC, lo que muestra, que el modelo se comporta mejor identificando correctamente a los estudiantes caucanos que están bajo la media en el Desempeño General.

El hecho de conocer los perfiles de los estudiantes caucanos, cuyo Desempeño General en las Pruebas Saber 9º, no es el mejor, ayudará a las directivas de las instituciones caucanas de Educación Básica Secundaria, de la Secretaria de Educación del departamento del Cauca y del MEN, a la toma de decisiones, basadas en datos para el mejoramiento de la calidad de la Educación Básica Secundaria y específicamente, en mejorar las habilidades de los estudiantes caucanos en las competencias evaluadas por las Pruebas Saber 9º, en el año 2017.

4.4.4 Discusión de Resultados de los Modelos Obtenidos de Desempeño Académico de los Estudiantes del Departamento del Cauca en las Pruebas Saber 9º

El análisis de los patrones de desempeño, de los estudiantes caucanos, en las Competencias de Lenguaje, Matemáticas y Desempeño General en las Pruebas Saber 9º, del año 2017, revela varios factores clave, que influyen en los resultados académicos.

El principal factor común asociado al rendimiento, es el Género del Colegio. Se observa que, los estudiantes de colegios mixtos, donde conviven alumnos de ambos sexos con iguales oportunidades y normas, tienden a obtener un desempeño inferior en las Pruebas Saber 9º. Por el contrario, los colegios de educación diferenciada por género—femeninos o masculinos—se asocian con mejores resultados académicos. Según Velandia, (2012), Cárcamo y Mola, (2012), Gordillo, (2015) y Correa, (2004) esto puede explicarse, por una mayor homogeneidad en el desempeño dentro de los colegios de un solo género, el enfoque académico más riguroso o incluso el proceso de selección de estudiantes con altos potenciales. Páez, (2014) también sostiene, que hombres y mujeres, aprenden y se desarrollan a ritmos distintos, lo que podría influir en estos resultados. Además, la experiencia demuestra, que los hombres y las mujeres tienen ritmos diferentes de desarrollo y maduran a edades dispares.

El segundo factor determinante, es la Naturaleza del Colegio, ya sea oficial o privado. Se asocia, un desempeño inferior a los estudiantes caucanos que asisten a colegios oficiales, mientras que aquellos que estudian

en colegios privados, tienden a obtener mejores resultados. Giraldo y Fajardo, (2018) explican, que en Colombia, existen diferencias significativas en financiación, gestión y calidad entre la educación pública y privada, lo que influye en los resultados académicos. Con la educación privada, a menudo, se asocia a mejores resultados académicos, de mejor calidad, con mejor infraestructura, pero a un costo mayor. El asistir a un colegio privado, está relacionado con estudiantes con mayores recursos económicos. Investigaciones como las de Seibold, (2000), Montero *et al.*, (2004) y Garbanzo, (2007) han demostrado, una fuerte relación entre el nivel socioeconómico del estudiante y su desempeño escolar. Por otra parte, López y González, (2016) indican, que existe una evidencia, en la relación entre la pertenencia a una clase social y el desempeño académico, es decir, que los estudiantes con un estatus socioeconómico más alto, demuestran mayores logros académicos.

Otro factor relevante, es la Ubicación del Colegio, ya sea en zona rural o urbana. Se observa, que los estudiantes caucanos de colegios rurales, tienden a obtener un desempeño inferior, en comparación con aquellos que estudian en colegios urbanos. Beltrán, (2023) señala, que el modelo educativo en Colombia, ha sido diseñado, desde una perspectiva urbana, dejando de lado las particularidades de la sociedad rural. Aguilar y Monge, (2000) resalta, la brecha educativa, entre las escuelas rurales y urbanas, mencionando problemas, como la falta de infraestructura y recursos, la desvinculación con programas de desarrollo agrícola y la exclusión de la perspectiva rural en los planes educativos. Además, Marcillo y Acero, (2016) destaca, la ausencia de políticas para mejorar la enseñanza en aulas multigrado. Las aulas multigrado, se encuentran en las zonas rurales del país y se caracteriza, por reunir estudiantes con edades distintas y de diferentes grados de escolaridad en una sola aula, en la cual, el proceso de enseñanza-aprendizaje, es asumido por un solo docente, lo que afecta sustancialmente la calidad educativa en zonas rurales.

Por último, la subregión del Cauca, en la que se encuentra el colegio, también influye en el desempeño escolar. Se observa, que los estudiantes de municipios ubicados en las zonas Norte y Centro del departamento, tienden a obtener resultados más bajos en las Pruebas Saber 9°, mientras

que los estudiantes de Popayán, la capital del departamento, muestran mejores desempeños. Según el Sistema Estadístico Nacional (Dane, 2020a), la pobreza multidimensional, es más prevalente en el Norte y Centro del Cauca, lo que impacta negativamente, el acceso a educación de calidad. Investigaciones como las de Serpa y Reyes, (2005), Buitrago, (2008), Ramírez *et al.*, (2011), Andrade *et al.*, (2015) y Mosquera, (2021) han demostrado, que la pobreza limita el acceso a recursos educativos fundamentales, afectando el rendimiento académico y las oportunidades de movilidad social. En contraste, las ciudades capitales suelen contar con mejores instalaciones, bibliotecas y acceso a tecnología, lo que explica los mejores resultados en Popayán.

Este análisis confirma que factores estructurales y socioeconómicos tienen un impacto significativo en el desempeño académico de los estudiantes caucanos, en las pruebas Saber 9º.

4.5 ÁRBOLES DE DECISIÓN PARA EL DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO EN LAS PRUEBAS SABER 9º DE LOS ESTUDIANTES DEL DEPARTAMENTO DE NARIÑO

Para construir los modelos de árboles de decisión, que permitan descubrir patrones asociados al desempeño académico en las Competencias de Lenguaje, Matemáticas y Desempeño General, de los estudiantes del departamento de Nariño, que estaban cursando el Grado 9º y presentaron las Pruebas Saber 9º, en el año 2017, se tomó el conjunto de datos *sbr9_narino_final*, compuesto por 18,971 instancias y 18 atributos y se formaron tres conjuntos de datos: *sbr9_narino_lenguaje*, *sbr9_narino_matemáticas* y *sbr9_narino_general*, tomando como atributo clase o ‘target’, el desempeño en la respectiva competencia, cuyos valores están discretizados como ‘sobre la media’ y ‘bajo la media’.

En la tabla 35 (igual que en la tabla 28), se muestra, la distribución observada o real del desempeño en las Competencias de Lenguaje, Matemáticas y Desempeño General. Además, para evaluar la calidad de los respectivos modelos, cada uno de estos conjuntos, se particionó en dos subconjuntos con el 80% de instancias para entrenamiento y 20% para prueba.

Tabla 35. Distribución del Desempeño Observado de los Estudiantes Nariñenses en las Competencias de Lenguaje, Matemáticas y Desempeño General

Conjunto	Desempeño				
	Bajo la Media	%	Sobre la Media	%	Total
sbr9_narino_lenguaje	10,466	55.1	8,505	44.9	18,971
sbr9_narino_matematicas	10,241	53.9	8,730	46.1	18,971
sbr9_narino_general	10,552	55.6	8419	44.4	18,971

Fuente. Esta Investigación

De acuerdo con los datos de la tabla 35, el desempeño académico en el departamento de Nariño presenta una distribución más equilibrada en comparación con el desempeño en el departamento del Cauca, aunque persiste una tendencia mayoritaria por debajo del estándar nacional. En la prueba de Lenguaje, de un universo de 18,971 evaluados, el 55.1 % (10,466 alumnos) se sitúa bajo la media, mientras que el 44.9 % restante logró superarla. En Matemáticas, los resultados son similares: un 53.9 % se encuentra por debajo del promedio nacional frente a un 46.1 % que se posiciona por encima. Finalmente, en el Desempeño General de las Pruebas Saber 9º, el 55.6 % de los estudiantes (10,552 registros) se ubica por debajo de la media, dejando a un 44.4 % en el nivel superior.

4.5.1 Clasificación mediante Árboles de Decisión del Desempeño en la Competencia de Lenguaje en las Pruebas Saber 9º de los Estudiantes Nariñenses

Para construir el modelo de clasificación por árboles de decisión, con el algoritmo J48, que permita descubrir patrones de desempeño académico, en la Competencia de Lenguaje en las Pruebas Saber 9º, que presentaron los estudiantes del departamento de Nariño, se utilizó el conjunto de datos *sbr9_narino_lenguaje*. Se escogió como clase o target, el desempeño en Lenguaje, representado por el atributo *clase_desemp_lenguaje*. Se probó el modelo, particionando este conjunto de datos, en dos conjuntos: el conjunto de entrenamiento con el 80% (15,176) del total

de instancias (de acuerdo con los parámetros definidos en la tabla 31) y el conjunto de prueba con el 20% (3,795) del total de instancias para el conjunto de prueba. Se podó el árbol de decisión, con los hiperparámetros: factor de confianza $C=0.15$ (se modificó este parámetro por mejores resultados) y mínimo número de instancias por nodo $M=0.5\%$, según los parámetros de referencia establecidos en la tabla 33.

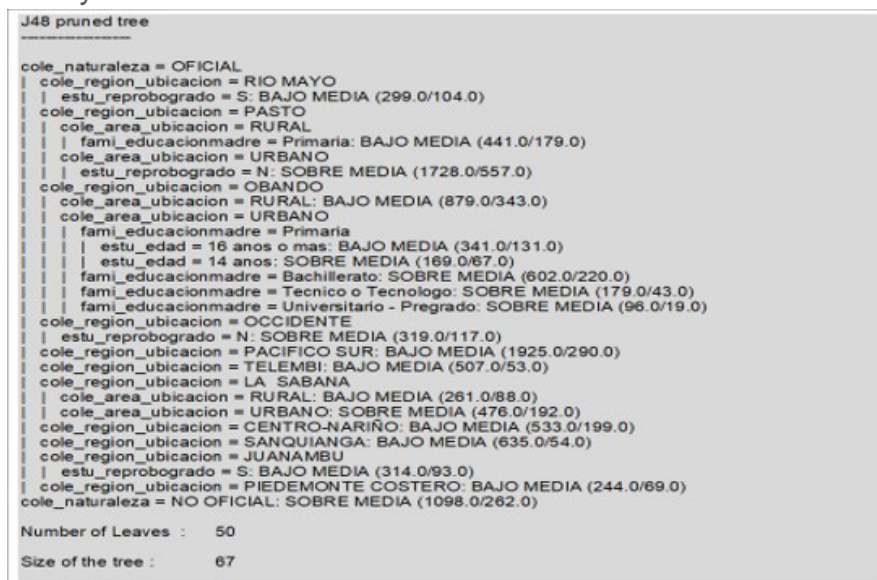
Para determinar la validez de las reglas o patrones, resultantes de interpretar cada rama del árbol desde el nodo raíz hasta el nodo hoja, se utilizaron las métricas de *soporte mínimo* (i.e. porcentaje mínimo del número de instancias que llegan a un nodo hoja, con respecto al número total de instancias del conjunto de datos) y *confianza mínima* (i.e. el porcentaje mínimo de instancias correctamente clasificadas en cada nodo, con respecto al número total de instancias que llegan al nodo hoja), adaptadas de la técnica de Minería de Datos Reglas de Asociación (Agrawal *et al.*, 1993) y se dejaron solo las ramas que superen el soporte mínimo de 0.5% y una confianza mínima de 60%.

En la figura 25, se aprecia el modelo de árbol de decisión resultante para la Competencia de Lenguaje.

Con el fin de interpretar los patrones identificados en el modelo de la figura 25, se seleccionaron las reglas más representativas del árbol de decisión resultante. Dichas reglas fueron filtradas bajo un criterio de soporte mínimo del 0.5 % y una confianza superior al 60 %, calculados sobre el total de las 15,176 instancias que conforman el conjunto de entrenamiento. Entre los patrones más representativos ordenados por soporte están:

Regla 1. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión Pacífico Sur, en el departamento de Nariño, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 12.7% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 84.9% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 23% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Figura 25. Árbol de Decisión Podado para la Competencia de Lenguaje de los Estudiantes Nariñenses en las Pruebas Saber 9º.



Fuente. Esta Investigación.

Regla 2. Si el estudiante, es de un colegio de carácter oficial, ubicado en el municipio de Pasto, capital del departamento de Nariño, en el área urbana y no reprobó ningún grado, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9º, es probable, que este sobre la media nacional. El 11.4% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 67.8% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 25.3% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 3. Si el estudiante, es de un colegio de carácter no oficial, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9º, es probable, que este sobre la media nacional. El 7.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 76.1% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 16.1% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 4. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Obando del departamento de Nariño, en el área rural, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 5.8% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 10.5% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 5. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Sanquianga del departamento de Nariño, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 4.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 91.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 7.6% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 6. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Obando del departamento de Nariño, en el área Urbana, el nivel educativo máximo de la madre, es Bachillerato, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 4.0% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 63.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 8.8% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 7. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión del Centro del departamento de Nariño, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 3.5% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 62.7% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 6.4% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 8. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión del Telembí del departamento de Nariño, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 3.3% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 89.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 6.1% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 9. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de La Sabana del departamento de Nariño y en el área Urbana, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 3.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 7% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 10. Si el estudiante, es de un colegio de carácter oficial, ubicado en el municipio de Pasto, capital del departamento de Nariño, en el área Rural y el máximo nivel educativo de la madre es Primaria, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 2.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 5.3% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 11. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Obando del departamento de Nariño, en el área Urbana, el nivel educativo máximo de la madre es Primaria y tiene 16 años o más, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 2.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 4.1% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 12. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Juanambú del departamento de Nariño y reprobó algún grado, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 2.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 70.4% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.8% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 13. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Occidente del departamento de Nariño y no reprobó ningún grado, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 2.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 63.3% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 4.7% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 14. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Rio Mayo del departamento de Nariño y reprobó algún grado, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 65.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.6% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 15. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de La Sabana del departamento de Nariño y en el área Rural, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 1.7% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 66.3% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.1% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 16. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Piedemonte Costero del departamento de Nariño, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 1.6% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 71.7% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2.9% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 17. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Obando del departamento de Nariño, en el área Urbana, el nivel educativo máximo de la madre es Técnico o Tecnólogo, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 1.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 76% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2.6% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 18. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Obando del departamento de Nariño, en el área urbana, el nivel educativo máximo de la madre, es Primaria y tiene 14 años, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 1.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60.4% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2.5% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 19. Si el estudiante, es de un colegio de carácter oficial, ubicado en un municipio de la subregión de Obando del departamento de Nariño, en el área urbana, el nivel educativo máximo de la madre es Universitario -Pregrado, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 0.6% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 80.2% de los estudiantes clasificados en esta categoría, están

correctamente clasificados y el 1.4% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

De igual manera, la figura 26 presenta la evaluación del modelo para la prueba de Lenguaje, fundamentada en las métricas derivadas de la matriz de confusión.

Figura 26. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Lenguaje de los Estudiantes Nariñenses en las Pruebas Saber 9º

=== Summary ===									
Correctly Classified Instances	2559						67.4308 %		
Incorrectly Classified Instances	1236						32.5692 %		
Kappa statistic			0.338						
Mean absolute error			0.4098						
Root mean squared error			0.4552						
Relative absolute error			82.9162 %						
Root relative squared error			91.6207 %						
Total Number of Instances	3795								
=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,719	0,383	0,703	0,719	0,711	0,338	0,730	0,764	BAJO MEDIA
	0,617	0,281	0,636	0,617	0,627	0,338	0,730	0,641	SOBRE MEDIA
Weighted Avg.	0,674	0,337	0,673	0,674	0,674	0,338	0,730	0,710	
=== Confusion Matrix ===									
a	b	<- classified as							
1521	593	a = BAJO MEDIA							
643	1038	b = SOBRE MEDIA							

Fuente. Esta Investigación.

Como se observa en la figura 26, el modelo de árbol de decisión resultante, tiene una exactitud del 67.4, es decir, que el modelo clasifica correctamente al 67.4% de las instancias del conjunto de prueba e incorrectamente a un 32.6% de las instancias probadas. La Raíz del Error Cuadrático Medio (RMSE) (por sus siglas en ingles *Root Mean Squared Error*) del modelo, es de 0.46, lo que indica, que el modelo esta ajustado, es decir, que los datos observados (reales) de los predichos están cerca.

Igualmente, de acuerdo con la Matriz de Confusión, de la figura 26, del total de 3,795 estudiantes (instancias) que conforman el conjunto de prueba del *sbr9_narino_lenguaje*, el modelo predice correctamente, al

40.1% de estudiantes con desempeño en Lenguaje “bajo la media” (verdaderos negativos VN), que corresponde a 1521 estudiantes y al 27.4% de estudiantes con desempeño académico “sobre la media” (verdaderos positivos VP), que corresponde a 1038 estudiantes. Además, el modelo predice incorrectamente, al 15.6% de estudiantes que están “bajo la media” como “sobre la media” (falsos positivos FP), correspondiente a 593 estudiantes y al 16.9% de estudiantes que están “sobre la media” como “bajo la media” (falsos negativos FN), correspondiente a 643 estudiantes.

En cuanto a la exactitud del modelo de árboles de decisión por clase, en la figura 26, se puede observar, que para el caso de los estudiantes nariñenses que están bajo la media (negativos) en el desempeño en Lenguaje en las Pruebas Saber 9º, el modelo tiene una *precisión* de Predicción de 0.703, lo que significa, que, del total de casos predichos que están bajo la media, el 70.3% están correctos. La Especificidad o tasa de verdaderos negativos (*TP Rate*) y *Recall* del modelo, es de 0.719, lo cual indica, que, el modelo es capaz de identificar al 71.9% de los estudiantes nariñenses que realmente están bajo la media. Indica, que el modelo tiene un buen nivel de Especificidad o capacidad para identificar correctamente, el 72% de las instancias negativas reales. En términos simples, de todas las instancias que realmente pertenecen a la clase negativa, el modelo ha logrado reconocer correctamente el 72% de ellas. Por otra parte, la tasa de falsos negativos (*FP Rate*) del modelo, es de 0.383, lo que significa, que el 38.3% de estudiantes que estaban sobre la media fueron predichos como bajo la media. La *F-Measure* o *F1-Score* es de 0.711, que es el promedio armónico entre la precisión y el *recall* de los que están bajo la media. Este valor sugiere, que el modelo tiene un buen rendimiento al equilibrar *precisión* y *recall*, pero no de manera óptima. El modelo, tiene un rendimiento razonable y aceptable entre *precisión* y *recall*. El modelo identifica de manera razonable, los estudiantes que están bajo la media (instancias negativas (*recall*)) y comete errores, de forma aceptable, en las predicciones negativas (*precisión*). Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.764. Este valor indica, que el modelo tiene una buena capacidad para identificar correctamente a los estudiantes nariñenses, que están bajo la media en el desempeño en Lenguaje. Con un 76.4%, el modelo tiene una capacidad más sólida para equilibrar correctamente las predicciones

“bajo la media”. Útil cuando se trata de conjuntos de datos desbalanceados, como es el caso del conjunto de datos *sbr9_narino_lenguaje*.

Para el caso de los estudiantes nariñenses que están sobre la media (positivos) en el desempeño en Lenguaje en las Pruebas Saber 9º, el modelo tiene una precisión de Predicción de 0.636, lo que significa, que, del total de casos predichos que están sobre la media, el 63,6% están correctos. La Sensibilidad o tasa de verdaderos positivos (*TP Rate*) y *Recall* del modelo, es de 0.617, lo cual indica que, el modelo es capaz de identificar al 61.7% de los estudiantes nariñenses que realmente están sobre la media. Por otra parte, la tasa de falsos positivos (*FP Rate*) del modelo es de 0.281, lo que significa, que el 28.1% de estudiantes que estaban sobre la media, fueron predichos como bajo la media. La *F-Measure* o *F1-Score* es de 0.617, lo que indica, que el modelo, muestra un desempeño moderado. No es excelente, pero tampoco es pobre, lo que indica, que hay un equilibrio aceptable entre *precisión* y *recall*, aunque ambos podrían mejorar. Sugiere, que el modelo identifica correctamente algunas instancias positivas (*recall* razonable) y mantiene un nivel aceptable de *precisión*, pero aún hay margen significativo de mejora. Finalmente, el Área Bajo la Curva (PRC) (*PRC Área*), es de 0.641. Este valor indica, que el modelo tiene una capacidad de discriminación aceptable, para identificar correctamente a los estudiantes nariñenses, que están sobre la media en la Competencia de Lenguaje.

El Área Bajo la Curva ROC (*ROC Área*) del modelo, es de 0.73 (mayor que 0.5) lo que indica, que, el modelo tiene una probabilidad aceptable, de que sea capaz de distinguir los estudiantes nariñenses que están etiquetados como “bajo la media”, de los estudiantes nariñenses que están etiquetados como “sobre la media”, ya que, es capaz de detectar más verdaderos negativos (VN) y verdaderos positivos (VP) que de falsos negativos (FN) y falsos positivos (FP).

En la combinación de estas métricas, se aprecia un mejor desempeño del modelo, en la clasificación de los estudiantes nariñenses etiquetados como “bajo la media” en la Competencia de Lenguaje. Para esta clase, la métrica PRC, es más alto que el ROC, lo que muestra, que el modelo se

comporta mejor cuando las métricas de *precisión* y *recall* se evalúan juntas. Esta investigación se centró, en los estudiantes con desempeño bajo la media. El hecho de conocer los perfiles de los estudiantes nariñenses, cuyo desempeño en la Competencia de Lenguaje de las Pruebas Saber 9º, no es el mejor, ayudará a las directivas de las instituciones nariñenses de Educación Básica Secundaria, de la Secretaria de Educación del departamento de Nariño y del MEN, a la toma de decisiones basadas en datos para el mejoramiento de la calidad de la Educación Básica Secundaria y específicamente, en mejorar las habilidades de los estudiantes nariñenses, en la Competencia de Lenguaje.

4.5.2 Clasificación mediante Árboles de Decisión del Desempeño en la Competencia de Matemáticas de las Pruebas Saber 9º de los Estudiantes Nariñenses

Para desarrollar el modelo de clasificación basado en árboles de decisión utilizando el algoritmo J48, con el propósito de identificar patrones de desempeño académico en la Competencia de Matemáticas de las Pruebas Saber 9º, presentadas por los estudiantes del departamento de Nariño, se empleó el conjunto de datos *sbr9_narino_matemáticas*. Como variable objetivo (target), se seleccionó el desempeño en Matemáticas, representado por el atributo *clase_desemp_matemáticas*.

Para evaluar el modelo, el conjunto de datos, se dividió en dos partes: un conjunto de entrenamiento, que abarcó el 80% de las instancias (15,176), siguiendo los parámetros definidos en la tabla 31 y un conjunto de prueba, conformado por el 20% restante (3,795 instancias). Además, se aplicó un proceso de poda al árbol de decisión, utilizando los hiperparámetros establecidos: un factor de confianza $C = 0.25$ y un mínimo de instancias por nodo $M = 0.5\%$, según los valores de referencia indicados en la tabla 33.

Para validar las reglas o patrones derivados de la interpretación de cada rama del árbol, desde el nodo raíz hasta el nodo hoja, se emplearon las métricas de Soporte Mínimo y Confianza Mínima. El Soporte Mínimo, se define, como el porcentaje de instancias que llegan a un nodo hoja

con respecto al total del conjunto de datos, mientras que la Confianza Mínima, corresponde al porcentaje de instancias correctamente clasificadas en cada nodo en relación con el total de instancias que llegan al nodo hoja. Estas métricas fueron adaptadas de la Técnica de Minería de Datos Reglas de Asociación (Agrawal *et al.*, 1993).

Para garantizar la relevancia de las reglas extraídas, solo se conservaron, aquellas ramas que alcanzaran un soporte mínimo de 0.5% y una confianza mínima de 60%.

En la figura 27, se muestra, el modelo de árbol de decisión resultante, para la Competencia de Matemáticas.

Figura 27. Árbol de Decisión Podado para la Competencia de Matemáticas de los Estudiantes Nariñenses en las Pruebas Saber 9^o.

```

J48 pruned tree
-----
cole_region_ubicacion = RIO MAYO
| estu_reprobogrado = S: BAJO MEDIA (299.0/121.0)
| estu_reprobogrado = N
| | cole_caracter = ACADEMICO: SOBRE MEDIA (183.0/65.0)
| | cole_caracter = TECNICO/ACADEMICO: SOBRE MEDIA (258.0/93.0)
cole_region_ubicacion = PASTO
| cuartil_inse = Q2
| | estu_reprobogrado = S
| | | estu_genero = Hombre: SOBRE MEDIA (201.0/78.0)
| | | estu_genero = Mujer: BAJO MEDIA (189.0/73.0)
| | | estu_reprobogrado = N: SOBRE MEDIA (642.0/224.0)
| | cuartil_inse = Q4: BAJO MEDIA (468.0/179.0)
| | cuartil_inse = Q3
| | | estu_edad = 16 anos o mas: BAJO MEDIA (217.0/86.0)
| | | cuartil_inse = Q1: SOBRE MEDIA (2017.0/569.0)
cole_region_ubicacion = OBANDO
| fami_educacionmadre = Primaria
| | estu_edad = 16 anos o mas: BAJO MEDIA (551.0/198.0)
| | fami_educacionmadre = Bachillerato
| | | estu_genero = Hombre: SOBRE MEDIA (406.0/145.0)
| | | fami_educacionmadre = Universitario - Posgrado: SOBRE MEDIA (138.0/37.0)
| | | fami_educacionmadre = Tecnico o Tecnologo: SOBRE MEDIA (229.0/55.0)
| | | fami_educacionmadre = Universitario - Pregrado: SOBRE MEDIA (138.0/37.0)
cole_region_ubicacion = OCCIDENTE
| estu_reprobogrado = N: SOBRE MEDIA (319.0/123.0)
cole_region_ubicacion = LA CORDILLERA
| fami_educacionmadre = Primaria
| | estu_genero = Mujer: BAJO MEDIA (231.0/90.0)
cole_region_ubicacion = PACIFICO SUR
| cole_naturalaleza = OFICIAL: BAJO MEDIA (1925.0/244.0)
| cole_naturalaleza = NO OFICIAL: SOBRE MEDIA (132.0/46.0)
cole_region_ubicacion = LOS ABADES
| estu_genero = Hombre: SOBRE MEDIA (178.0/56.0)
cole_region_ubicacion = TELEMBI: BAJO MEDIA (507.0/43.0)
cole_region_ubicacion = LA SABANA
| cole_area_ubicacion = URBANO: SOBRE MEDIA (476.0/169.0)
cole_region_ubicacion = GUAMBUYACO
| estu_genero = Mujer: BAJO MEDIA (190.0/72.0)
cole_region_ubicacion = CENTRO-NARIÑO
| cole_caracter = ACADEMICO: BAJO MEDIA (363.0/126.0)
| cole_caracter = TECNICO/ACADEMICO: SOBRE MEDIA (114.0/44.0)
cole_region_ubicacion = SANQUIANGA: BAJO MEDIA (635.0/67.0)
cole_region_ubicacion = JUANAMBU
| estu_reprobogrado = S: BAJO MEDIA (314.0/110.0)

Number of Leaves :    58
Size of the tree :    82

```

Fuente. Esta Investigación.

Para la interpretación de los patrones, se seleccionaron las reglas más representativas del modelo de árbol de decisión obtenido (figura 27), asegurando que cumplieran con un mínimo soporte del 0.5% y una confianza mínima del 60%, en relación con el total de instancias del conjunto de entrenamiento (15,176 instancias).

A continuación, se presentan los patrones más relevantes, ordenados según su valor de soporte:

Regla 1. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño y su índice de nivel socioeconómico INSE está en el cuartil Q1, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 13.3% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 71.8% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 29% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 2. Si el estudiante es de un colegio ubicado en un municipio de la subregión Pacífico Sur del departamento de Nariño y la naturaleza del colegio es oficial, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 12.7% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 87.3% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 23.4% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 3. Si el estudiante, es de un colegio ubicado en un municipio de la subregión Sanquianga del departamento de Nariño, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 4.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 89.4% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 7.7% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 4. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño, su índice de nivel socioeconómico INSE está en el cuartil Q2 y no reprobó ningún grado, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 4.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 65.1% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 9.2% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 5. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando del departamento de Nariño, el máximo nivel educativo de la madre, es Primaria y tiene 16 años o más, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 3.6% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 64.1% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 6.7% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 6. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Telembí del departamento de Nariño, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 3.3% del total de estudiantes nariñenses evaluados se clasifican de esta manera. El 91.5% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 6.2% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 7. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de La Sabana del departamento de Nariño y está en el área urbana del municipio, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 3.1% del total de estudiantes nariñenses evaluados se clasifican de esta manera. El 64.5% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 6.8% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 8. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño y su índice de nivel socioeconómico INSE está en el cuartil Q4, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 3.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61.8% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 5.7% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 9. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando del departamento de Nariño, el máximo nivel educativo de la madre es Bachillerato y es hombre, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 2.7% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 64.3% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 5.8% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 10. Si el estudiante, es de un colegio ubicado en un municipio de la subregión Centro del departamento de Nariño y es un colegio de carácter académico, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 2.4% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 65.3% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 4.4% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 11. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Juanambú del departamento de Nariño y reprobó algún grado, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 2.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 65% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 3.8% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 12. Si el estudiante, es de un colegio ubicado en el municipio de la subregión de Occidente del departamento de Nariño y no reprobó ningún grado, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 2.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61.4% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 4.6% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 13. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Rio Mayo del departamento de Nariño y reprobó algún grado entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 3.6% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 14. Si el estudiante, es de un colegio ubicado en un municipio de la subregión Rio Mayo del departamento de Nariño, no reprobó ningún grado y es un colegio de carácter técnico/académico, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 1.7% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 64% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 3.7% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 15. Si el estudiante, es de un colegio ubicado en el municipio de la subregión de Obando del departamento de Nariño y el máximo nivel educativo de la madre es técnico o tecnólogo, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 1.5% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 76% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 3.3% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 16. Si el estudiante, es de un colegio ubicado en el municipio de la subregión de La Cordillera del departamento de Nariño, el máximo nivel educativo de la madre es primaria y es mujer, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 1.5% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2.8% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 17. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño, su índice de nivel socioeconómico INSE está en el cuartil Q3 y tiene 16 años o más, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable que este bajo la media nacional. El 1.4% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60.4% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2.6% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 18. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Guambuyaco del departamento de Nariño y es una mujer, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 1.3% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 62.1% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2.3% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 19. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño, su índice de nivel socioeconómico INSE está en el cuartil Q2, reprobó algún grado y es hombre, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 1.3% del total de estudiantes nariñenses evaluados se clasifican de esta manera. El 61.2% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2.9% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 20. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Los Abades del departamento de Nariño y es un hombre, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 1.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 68.5% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2.6% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 21. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Rio Mayo del departamento de Nariño, no reprobó ningún grado y es un colegio de carácter académico, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 1.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 64.5% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2.6% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 22. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño, su índice de nivel socioeconómico INSE está en el cuartil Q2, reprobó algún grado y es mujer, entonces su desempeño académico en la prueba de matemáticas del Saber 9° es probable que este bajo la media nacional. El 1.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61.4% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2.3% de los estudiantes observados que están bajo la media cumplen este patrón.

Regla 23. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando del departamento de Nariño y el máximo nivel educativo de la madre es universitario con postgrado, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 0.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 73.2% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 24. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando del departamento de Nariño y el máximo nivel educativo de la madre es universitario pregrado, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 0.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 73.2% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 2% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 25. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Pacífico Sur del departamento de Nariño y la naturaleza del colegio es no oficial, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 0.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 65.2% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 1.9% de los estudiantes observados que están sobre la media cumplen este patrón.

Regla 26. Si el estudiante, es de un colegio ubicado en un municipio de la subregión Centro del departamento de Nariño y es un colegio de carácter técnico/académico, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 0.8% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61.4% de los estudiantes que se clasifican en esta categoría, están correctamente clasificados y el 1.6% de los estudiantes observados que están sobre la media cumplen este patrón.

Por otra parte, en la figura 28 se presenta la evaluación del modelo para la prueba de Matemáticas, fundamentada en las métricas derivadas de la matriz de confusión.

Figura 28. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Matemáticas de los Estudiantes Nariñenses en las Pruebas Saber 9º

=== Summary ===									
Correctly Classified Instances	2529							66.6403 %	
Incorrectly Classified Instances	1266							33.3597 %	
Kappa statistic								0.3303	
Mean absolute error								0.4086	
Root mean squared error								0.4551	
Relative absolute error								82.1734 %	
Root relative squared error								91.2195 %	
Total Number of Instances	3795								
=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.682	0.352	0.689	0.682	0.686	0.330	0.731	0.755	BAJO MEDIA
	0.648	0.318	0.641	0.648	0.645	0.330	0.731	0.663	SOBRE MEDIA
Weighted Avg.	0.666	0.336	0.667	0.666	0.667	0.330	0.731	0.712	
=== Confusion Matrix ===									
a	b	<-- classified as							
1381	643	a = BAJO MEDIA							
623	1148	b = SOBRE MEDIA							

Fuente. Esta Investigación.

Como se muestra en la figura 28, el modelo de árbol de decisión alcanzó una exactitud del 66.6%, lo que significa, que clasifica correctamente el 66.6% de las instancias del conjunto de prueba (tanto positivas como negativas), mientras que el 33.4% restante corresponde a clasificaciones incorrectas.

Además, la Raíz del Error Cuadrático Medio (RMSE, por sus siglas en inglés: *Root Mean Squared Error*) del modelo, es de 0.46, lo que indica un buen ajuste. En otras palabras, los valores predichos por el modelo, se encuentran relativamente cerca de los datos observados, lo que refleja, una adecuada capacidad de generalización.

Según la Matriz de Confusión, el conjunto de prueba del modelo *sbr9_narino_matemáticas*, está conformado por 3,795 estudiantes (instancias). De ellos, el modelo predice correctamente, el desempeño en Matemáticas, de 36.4% de los estudiantes que obtuvieron resultados *bajo la media* (1,381 estudiantes, verdaderos negativos - VN) y de 30.3% de los estudiantes con desempeño *sobre la media* (1,148 estudiantes, verdaderos positivos - VP).

Sin embargo, el modelo también presenta errores en la clasificación: 16.9% de los estudiantes con desempeño *bajo la media* (643 estudiantes) fueron clasificados incorrectamente como *sobre la media* (falsos positivos - FP), mientras que 16.4% de los estudiantes con desempeño *sobre la media* (623 estudiantes) fueron categorizados erróneamente como *bajo la media* (falsos negativos - FN).

En cuanto a la exactitud del modelo de árboles de decisión por clase, se observa, que, para los estudiantes nariñenses con desempeño *bajo la media* en Matemáticas en las Pruebas Saber 9º (instancias negativas), el modelo alcanza una precisión de Predicción de 0.689, lo que significa, que el 68.9% de los casos predichos como “bajo la media” son correctos.

La Especificidad del modelo, también conocida como Tasa de Verdaderos Negativos (*TP Rate*) o *recall*, es de 0.682, lo que indica, que el modelo identifica correctamente el 68.2% de los estudiantes que realmente tienen un desempeño bajo la media. En términos sencillos, esto refleja una buena capacidad del modelo, para reconocer correctamente las instancias negativas.

Por otra parte, la Tasa de Falsos Negativos (*FP Rate*), es de 0.352, lo que significa que el 35.2% de los estudiantes con desempeño “sobre la media” fueron clasificados erróneamente como “bajo la media”.

La *F-Measure* o *F1-Score* del modelo, es de 0.686, lo que representa el promedio armónico entre precisión y *recall*. Este valor indica, que el modelo mantiene un equilibrio moderado entre identificar correctamente a los estudiantes con bajo desempeño y evitar clasificaciones incorrectas.

Finalmente, el Área Bajo la Curva PRC (PRC Área), es de 0.755, lo que indica una buena capacidad del modelo, para diferenciar correctamente a los estudiantes con desempeño *bajo la media*. Con un 75.5%, el modelo demuestra un sólido rendimiento en la clasificación de estudiantes en conjuntos de datos desbalanceados, como el conjunto *sbr9_narino_matematicas*.

Para los estudiantes nariñenses, con desempeño sobre la media en Matemáticas en las Pruebas Saber 9^o (instancias positivas), el modelo presenta una precisión de Predicción de 0.641, lo que indica, que el 64.1% de los casos clasificados como “sobre la media” son correctos.

La Sensibilidad, también conocida como Tasa de Verdaderos Positivos (*TP Rate*) o *recall*, es de 0.648, lo que significa, que el modelo identifica correctamente al 64.8% de los estudiantes que realmente tienen un desempeño sobre la media.

Por otro lado, la Tasa de Falsos Positivos (*FP Rate*), es de 0.318, lo que implica, que el 31.8% de los estudiantes con desempeño “bajo la media”, fueron clasificados erróneamente como “sobre la media”.

La *F-Measure* o *F1-Score* es de 0.645, lo que refleja, un desempeño moderado del modelo. Si bien, muestra un equilibrio aceptable entre precisión y *recall*, todavía hay margen de mejora. En términos prácticos, el modelo reconoce correctamente, algunas instancias positivas, pero aún podría optimizar su capacidad para reducir errores de clasificación.

Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.633, lo que indica, que el modelo posee una capacidad de discriminación aceptable, para identificar correctamente a los estudiantes con desempeño sobre la media en Matemáticas. Un PRC de 0.633, sugiere, que el modelo es moderadamente efectivo, especialmente en contextos de conjuntos de datos desbalanceados como *sbr9_narino_matemáticas*.

El Área Bajo la Curva ROC (*ROC Área*) del modelo, es de 0.73, un valor superior a 0.5, lo que indica, una capacidad aceptable de diferenciación entre los estudiantes nariñenses con desempeño bajo la media y aquellos con desempeño sobre la media.

Este resultado sugiere, que el modelo detecta más verdaderos negativos (VN) y verdaderos positivos (VP) que falsos negativos (FN) y falsos positivos (FP), lo que refleja un desempeño razonable en la clasificación de las instancias y una adecuada capacidad para discriminar correctamente entre ambas categorías de estudiantes.

La combinación de estas métricas, revela, que el modelo tiene un mejor desempeño en la clasificación de los estudiantes etiquetados como *bajo la media*, lo cual se evidencia en el valor de PRC más alto que el ROC. Esto indica, que el modelo es eficaz en la identificación precisa de los estudiantes nariñenses, con menor rendimiento en Matemáticas.

Como se mencionó anteriormente, esta investigación se centra en este grupo de estudiantes. Conocer sus perfiles, permitirá a las directivas de las instituciones de Educación Básica Secundaria en Nariño, a la Secretaría de Educación del departamento y al Ministerio de Educación Nacional (MEN) tomar decisiones, basadas en datos que contribuyan a mejorar la calidad educativa y, específicamente, a fortalecer las habilidades Matemáticas de los estudiantes nariñenses.

4.5.3 Clasificación mediante Árboles de Decisión para el Desempeño General en las Pruebas Saber 9º de los Estudiantes Nariñenses

Para desarrollar el modelo de clasificación, basado en árboles de decisión, con el algoritmo J48, cuyo objetivo es identificar patrones de desempeño general en las Pruebas Saber 9º, de los estudiantes del departamento de Nariño, se trabajó con el conjunto de datos *sbr9_nariño_general*. Se estableció como variable objetivo (*target*), el Desempeño General, representado por el atributo *clase_desemp_general*.

Para evaluar el modelo, se dividió el conjunto, de datos en dos partes: un 80% destinado al entrenamiento (15,176 instancias, según los parámetros de la tabla 31) y un 20% reservado para la prueba (3,795 instancias). Además, se realizó una poda del árbol de decisión, utilizando los hiperparámetros de referencia: un factor de confianza de $C=0.25$ y un mínimo de instancias por nodo de $M=0.5\%$, conforme a lo establecido en la tabla 33.

Para validar las reglas o patrones obtenidos al interpretar cada rama del árbol, desde el nodo raíz hasta el nodo hoja, se aplicaron las métricas de Soporte Mínimo y Confianza Mínima. El Soporte Mínimo, representa el porcentaje mínimo de instancias que llegan a un nodo hoja en

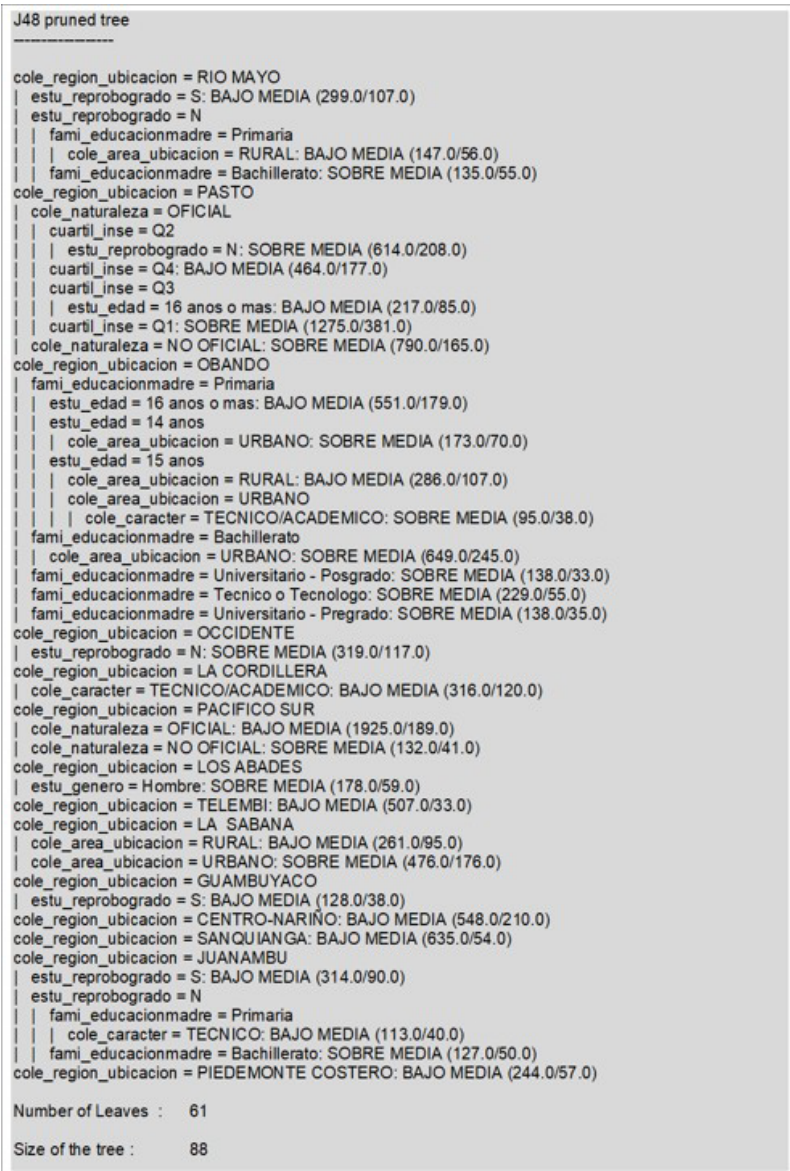
relación con el total del conjunto de datos, mientras que la Confianza Mínima, indica el porcentaje de instancias correctamente clasificadas en cada nodo, respecto al total que llega al nodo hoja.

Estas métricas, fueron adaptadas de la Técnica de Minería de Datos conocida como Reglas de Asociación (Agrawal *et al.*, 1993). Como criterio de selección, solo se conservaron las ramas que superaron un soporte mínimo del 0.5% y una confianza mínima del 60%.

En la figura 29, se muestra, el árbol de decisión obtenido para el Desempeño General en las Pruebas Saber 9º de los estudiantes del departamento de Nariño.

Para efectos de la interpretación de los patrones de Desempeño General de los estudiantes del departamento de Nariño en las Pruebas Saber 9º, se escogieron las reglas más representativas del modelo de árbol de decisión obtenido (figura 29) que cumplan un mínimo soporte del 0.5% y una confianza mínima del 60%, con relación al total de instancias del conjunto de entrenamiento (15,176). Entre los patrones más importantes ordenados por soporte están:

Figura 29. Árbol de Decisión Podado para el Desempeño General en las Pruebas Saber 9º de Los Estudiantes Nariñenses en las Pruebas Saber 9º.



Fuente. Esta Investigación.

Regla 1. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Pacífico Sur, en el departamento de Nariño y asiste a un colegio oficial, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 12.7% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 90.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 22.8% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 2. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño, asiste a un colegio oficial y pertenece al cuartil Q1, según su índice de nivel socioeconómico INSE, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 8.4% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 70.1% de los estudiantes clasificados en esta categoría, se encuentran correctamente clasificados y el 19% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 3. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño y asiste a un colegio no oficial, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 5.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 79.1% de los estudiantes clasificados en esta categoría, se encuentran correctamente clasificados y el 11.7% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 4. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando, en el departamento de Nariño, su madre tiene como máximo nivel educativo el Bachillerato y asiste a un colegio situado en la zona urbana, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 4.3% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 62.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 9.6% de los estudiantes observados con resultados por encima de la media, cumplen con este patrón.

Regla 5. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Sanquianga, en el departamento de Nariño, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 4.2% nariñenses evaluados, se clasifican de esta manera. El 91.5% de los estudiantes del total de estudiantes clasificados en esta categoría, están correctamente clasificados y el 7.5% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 6. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño, asiste a un colegio oficial, pertenece al cuartil Q2, según su índice de nivel socioeconómico INSE y no ha reprobado ningún grado, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 4% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 66.1% de los estudiantes clasificados en esta categoría, se encuentran correctamente clasificados y el 9.1% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 7. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando, en el departamento de Nariño, su madre posee como máximo nivel educativo la Primaria y tiene 16 años o más, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 3.6% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 67.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 6.5% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 8. Si el estudiante, es de un colegio ubicado en un municipio de la subregión Centro, en el departamento de Nariño, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 3.6% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61.7% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 6.5% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 9. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Telembí, en el departamento de Nariño, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 3.3% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 93.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 6% de los estudiantes observados con resultados bajo la media, cumplen con este patrón.

Regla 10. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de La Sabana, en el departamento de Nariño y asiste a un colegio en el área urbana, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 3.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 63% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 7.1% de los estudiantes observados con resultados superiores a la media, cumplen con este patrón.

Regla 11. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño y pertenece al cuartil Q4, según su índice de nivel socioeconómico INSE, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 3.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61.9% de los estudiantes clasificados en esta categoría, se encuentran correctamente clasificados y el 5.5% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 12. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Juanambú, en el departamento de Nariño y ha reprobado algún grado, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 2.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 71.3% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.7% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 13. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Occidente, en el departamento de Nariño y no ha reprobado ningún grado, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 2.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 63.3% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 4.7% de los estudiantes observados con resultados por encima de la media, cumplen con este patrón.

Regla 14. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de La Cordillera, en el departamento de Nariño y asiste a un colegio de carácter Técnico/Académico, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 2.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 62% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.7% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 15. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Rio Mayo, en el departamento de Nariño y ha reprobado algún grado, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 64.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.5% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 16. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando, en el departamento de Nariño, cuya madre tiene como máximo nivel educativo la Primaria, tiene 15 años y asiste a un colegio situado en la zona rural, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 1.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 62.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.4% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 17. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de La Sabana, en el departamento de Nariño y asiste a un colegio en el área rural, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 1.7% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 63.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados, y el 3.1% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 18. Si el estudiante, es de un colegio ubicado en un municipio de la subregión del Piedemonte Costero, en el departamento de Nariño, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 1.6% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 76.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2.9% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 19. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando, en el departamento de Nariño y su madre posee como máximo nivel educativo un título de Técnico o Tecnólogo, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 1.5% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 76% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.4% de los estudiantes observados con resultados por encima de la media, cumplen con este patrón.

Regla 20. Si el estudiante, es de un colegio ubicado en el municipio de Pasto, capital del departamento de Nariño, pertenece al cuartil Q3 según su índice de nivel socioeconómico INSE y tiene 16 años o más, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 1.4% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60.8% de los estudiantes clasificados en esta categoría, se encuentran correctamente clasificados y el 2.6% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 21. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Los Abades, en el departamento de Nariño y es un hombre, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 1.2% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 66.9% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2.6% de los estudiantes observados con resultados superiores a la media, cumplen con este patrón.

Regla 22. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando, en el departamento de Nariño, cuya madre tiene como máximo nivel educativo la Primaria, tiene 14 años y asiste a un colegio situado en la zona urbana, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 1.1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2.6% de los estudiantes observados con resultados superiores a la media, cumplen con este patrón.

Regla 23. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Rio Mayo, en el departamento de Nariño, no ha reprobado ningún grado, su madre tiene como máximo nivel educativo la Primaria y asiste a un colegio situado en la zona rural, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 1% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 61.9% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.7% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 24. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando, en el departamento de Nariño y su madre posee como máximo nivel educativo un título universitario con postgrado, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 0.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 76.1% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2.1% de

los estudiantes observados con resultados por encima de la media, cumplen con este patrón.

Regla 25. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando, en el departamento de Nariño y su madre tiene como máximo nivel educativo un título universitario, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 0.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 74.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2.1% de los estudiantes observados con resultados por encima de la media, cumplen con este patrón.

Regla 26. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Pacífico Sur, en el departamento de Nariño y asiste a un colegio no oficial, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 0.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 68.9% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2% de los estudiantes observados con resultados superiores a la media, cumplen con este patrón.

Regla 27. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Río Mayo, en el departamento de Nariño, no ha reprobado ningún grado y su madre tiene como máximo nivel educativo el Bachillerato, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 0.9% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, se encuentran correctamente clasificados y el 2% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 28. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Guambuyaco, en el departamento de Nariño y reprobó algún grado, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 0.8% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 70.3% de los

estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.5% de los estudiantes observados con resultados inferiores a la media, cumplen con este patrón.

Regla 29. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Juanambú, en el departamento de Nariño, no ha reprobado ningún grado y su madre tiene como máximo nivel educativo el Bachillerato, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 0.8% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.9% de los estudiantes observados con resultados sobre la media, cumplen con este patrón.

Regla 30. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Juanambú, en el departamento de Nariño, no ha reprobado ningún grado, su madre tiene como máximo nivel educativo la Primaria y asiste a un colegio de carácter Técnico, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 0.7% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 64.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.3% de los estudiantes observados con resultados bajo la media, cumplen con este patrón.

Regla 31. Si el estudiante, es de un colegio ubicado en un municipio de la subregión de Obando, en el departamento de Nariño, cuya madre tiene como máximo nivel educativo la Primaria, tiene 15 años, asiste a un colegio situado en la zona urbana de carácter Técnico/Académico, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 0.6% del total de estudiantes nariñenses evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.4% de los estudiantes observados con resultados superiores a la media, cumplen con este patrón.

Por otra parte, en la figura 30 se presenta la evaluación del modelo obtenido para el Desempeño General de los estudiantes nariñenses en las Pruebas Saber 9º, fundamentada en las métricas derivadas de la matriz de confusión.

Como se aprecia en la figura 30, el modelo de árbol de decisión obtenido presenta una exactitud del 69.1%. Esto significa que clasifica correctamente el 69.1% de las instancias del conjunto de prueba, tanto positivas como negativas, mientras que el 30.9% de las instancias son clasificadas incorrectamente.

La raíz del error cuadrático medio (*Root Mean Squared Error*, RMSE) del modelo es de 0.44, lo que indica que, en promedio, la magnitud de los errores en las predicciones es de aproximadamente 0.44. En otras palabras, las estimaciones del modelo se encuentran, en promedio, a 0.44 de distancia de los valores reales u observados.

Figura 30. Métricas de Evaluación del Modelo de Árbol de Decisión para el Desempeño General de los Estudiantes Nariñenses en las Pruebas Saber 9º.

=== Summary ===										
Correctly Classified Instances	2624								69.1436 %	
Incorrectly Classified Instances	1171								30.8564 %	
Kappa statistic								0.3741		
Mean absolute error								0.3907		
Root mean squared error								0.4444		
Relative absolute error								79.1119 %		
Root relative squared error								89.4108 %		
Total Number of Instances	3795									
=== Detailed Accuracy By Class ===										
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
	0,731	0,358	0,718	0,731	0,724	0,374	0,757	0,791	BAJO MEDIA	
	0,642	0,269	0,658	0,642	0,650	0,374	0,757	0,677	SOBRE MEDIA	
Weighted Avg.	0,691	0,318	0,691	0,691	0,691	0,374	0,757	0,741		
=== Confusion Matrix ===										
a	b	<-- classified as								
1537	566	a = BAJO MEDIA								
605	1087	b = SOBRE MEDIA								

Fuente. Esta Investigación.

Según la Matriz de Confusión, el conjunto de prueba *sbr9_narino_general* está compuesto por 3,795 instancias (estudiantes). De ellos, el modelo, predice correctamente el desempeño académico, de un 41% de los estudiantes con rendimiento “bajo la media” (*verdaderos negativos*, VN), lo que equivale a 1,537 estudiantes. Asimismo, acierta en un 29% de los casos al identificar a estudiantes con desempeño “sobre la media” (*verdaderos positivos*, VP), que representan 1,087 estudiantes.

El modelo comete errores en un 15% de los casos, clasificando erróneamente a 566 estudiantes con desempeño “bajo la media” como si estuvieran “sobre la media” (*falsos positivos*, FP). De manera similar, un 16% de los estudiantes que realmente tienen un desempeño “sobre la media” son clasificados incorrectamente como si estuvieran “bajo la media” (*falsos negativos*, FN), lo que equivale a 605 estudiantes.

En la figura 30, se observa la exactitud del modelo de árboles de decisión, según cada clase. Para el caso de los estudiantes nariñenses con Desempeño General “bajo la media” en las Pruebas Saber 9º (*clase negativa*), el modelo alcanza una precisión de Predicción de 0.718, lo que significa, que el 71.8% de las instancias clasificadas como “bajo la media” son correctas.

La especificidad, también conocida como tasa de verdaderos negativos (*True Positive Rate*, TP Rate) y *recall*, es de 0.731. Esto indica que el modelo identifica correctamente el 73.1% de los estudiantes que realmente tienen un desempeño bajo la media, mostrando una buena capacidad para reconocer instancias negativas con precisión. En términos simples, de todas las instancias que efectivamente pertenecen a esta categoría, el modelo logra clasificar correctamente el 73.1%.

Por otro lado, la tasa de falsos positivos (*False Positive Rate*, FP Rate), es de 0.358, lo que significa, que el 35.8% de los estudiantes con desempeño “sobre la media” fueron incorrectamente clasificados como “bajo la media”.

El *F-Measure* o *F1-Score* es de 0.724, lo que representa, un equilibrio entre la precisión y el *recall* en la clasificación de estudiantes con desempeño bajo la media. Esto sugiere, que el modelo de clasificación, logra identificar correctamente un 72.4% de los casos sin comprometer demasiado la exactitud ni generar clasificaciones erróneas en exceso.

Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.791, lo que indica, una alta capacidad del modelo para detectar estudiantes con desempeño bajo la media en las Pruebas Saber 9º. La curva PRC, es especialmente útil, en conjuntos de datos desbalanceados, como es el caso de *sbr9_narino_general*, ya que mide, la Efectividad del modelo, al clasificar correctamente los ejemplos negativos. Un valor de 0.791, sugiere, que el modelo es bastante confiable en esta tarea.

Para los estudiantes nariñenses, con un Desempeño General “sobre la media” en las Pruebas Saber 9º (*clase positiva*), el modelo alcanza una precisión de Predicción de 0.658. Esto significa, que el 65.8% de las instancias clasificadas como “sobre la media” son correctas.

La Sensibilidad, también conocida como Tasa de Verdaderos Positivos (*True Positive Rate*, TP Rate) y *recall*, es de 0.642, lo que indica, que el modelo identifica correctamente al 64.2% de los estudiantes que realmente tienen buen desempeño.

Por otro lado, la Tasa de Falsos Positivos (*False Positive Rate*, FP Rate), es de 0.269, lo que implica, que el 26.9% de los estudiantes que el modelo clasificó como “sobre la media” en realidad tienen un desempeño “bajo la media”. En términos prácticos, aproximadamente 1 de cada 4 clasificaciones positivas realizadas por el modelo, es incorrecta.

El *F-Measure* o *F1-Score*, es de 0.65, lo que refleja, un equilibrio entre la precisión y el *recall*. Es decir, el modelo, logra clasificar correctamente, al 65% de los estudiantes con buen desempeño académico, considerando tanto los falsos positivos como los falsos negativos. Este resultado sugiere, un rendimiento aceptable, aunque con margen de mejora.

Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.677, lo que indica, que el modelo tiene una capacidad de discriminación moderada, para identificar correctamente, a los estudiantes con Desempeño General “sobre la media” en las Pruebas Saber 9º. La curva PRC, es particularmente útil, en conjuntos de datos desbalanceados, como *sbr9_narino_general*, ya que permite evaluar la efectividad del modelo, al clasificar correctamente los ejemplos positivos. Un valor de 0.677, sugiere que el modelo tiene una eficiencia razonable en esta tarea.

El Área Bajo la Curva ROC (*ROC Área*) del modelo, es de 0.757, indica, que el modelo tiene un buen nivel de discriminación, siendo capaz de diferenciar correctamente, las dos clases en el 75.7% de los casos. El modelo, es capaz de distinguir los estudiantes nariñenses que están etiquetados como “bajo la media” (bajo desempeño), de los estudiantes nariñenses que están etiquetados como “sobre la media” (buen desempeño), ya que, es capaz de detectar más verdaderos negativos (VN) y verdaderos positivos (VP) que de falsos negativos (FN) y falsos positivos (FP).

Al analizar estas métricas en conjunto, se observa, que el modelo tiene un mejor desempeño, en la clasificación de los estudiantes nariñenses, con Desempeño General “bajo la media” en las pruebas Saber 9º. Este resultado se refuerza con el hecho, de que, en esta clase, el Área Bajo la Curva PRC, es mayor que el Área Bajo la Curva ROC, lo que indica, que el modelo es particularmente eficaz en términos de precisión y *recall*, para identificar a los estudiantes con desempeño inferior al promedio.

Conocer los perfiles de los estudiantes nariñenses, que presentan un bajo rendimiento en las Pruebas Saber 9º, permitirá a las directivas de las instituciones de Educación Básica Secundaria, a la Secretaría de Educación del departamento de Nariño y al Ministerio de Educación Nacional (MEN), tomar decisiones fundamentadas en datos. Estas acciones contribuirán, a mejorar la calidad educativa en el nivel básico secundario, con un enfoque especial, en el fortalecimiento de las habilidades de los estudiantes en las Competencias de Lenguaje y Matemáticas.

4.5.4 Discusión de Resultados de los Modelos Obtenidos de Desempeño Académico de los Estudiantes del Departamento de Nariño en las Pruebas Saber 9º

El análisis de los patrones de desempeño de los estudiantes nariñenses en las Competencias de Lenguaje, Matemáticas y Desempeño General en las Pruebas Saber 9º, del año 2017, revela, que los factores más determinantes en los resultados académicos, son la subregión del municipio, en la que se encuentra el colegio y la naturaleza de la institución educativa. Estos elementos influyen significativamente, en el nivel de aprendizaje de los estudiantes, evidenciando disparidades en el acceso a recursos y oportunidades educativas dentro del departamento de Nariño.

El desempeño académico en las Pruebas Saber 9º, varía significativamente según la subregión del municipio, en el que se ubica el colegio dentro del departamento de Nariño. Un rendimiento más bajo, se observa en los estudiantes de instituciones ubicadas en municipios pertenecientes a las subregiones del Pacífico, como Pacífico Sur, Piedemonte Costero, Sanquianga y Telembí. En contraste, los mejores resultados, se asocian con estudiantes de colegios ubicados en la capital del departamento, Pasto, así como, en menor medida, en las subregiones de Obando y La Sabana. Estos patrones reflejan diferencias en el acceso a infraestructura educativa, recursos y oportunidades académicas entre las distintas subregiones.

De acuerdo con el Departamento Administrativo Nacional de Estadística (DANE, 2020b) y Santacruz (2024), las subregiones del Pacífico nariñense, tales como Tumaco, Barbacoas y Roberto Payán, enfrentan desafíos críticos asociados a la pobreza multidimensional y a las necesidades básicas insatisfechas (NBI). Estos altos índices de precariedad son el resultado de factores estructurales como el aislamiento geográfico, la carencia de infraestructura adecuada y el acceso limitado a servicios esenciales. Asimismo, la persistente inseguridad y la vulnerabilidad ambiental en estas zonas actúan como barreras adicionales que obstaculizan el desarrollo socioeconómico y, por extensión, el fortalecimiento de los procesos educativos en el territorio.

Por otro lado, investigaciones como las de Serpa y Reyes, (2005), Buitrago, (2008), Ramírez *et al.*, (2011), Andrade *et al.*, (2015) y Mosquera, (2021) destacan, cómo la pobreza en Colombia, está estrechamente vinculada al bajo desempeño escolar. Esto repercute negativamente en el acceso a una educación de calidad y limita las oportunidades de movilidad social. La falta de recursos esenciales para el aprendizaje, como materiales escolares, acceso a internet, alimentación adecuada y un entorno familiar seguro y estimulante, agrava esta situación.

En Colombia, la educación en los colegios de las ciudades capitales de departamento, suele ser de mejor calidad y contar con más recursos que en los municipios no capitales. Las capitales, destacan, por disponer de mejores infraestructuras educativas, como escuelas, bibliotecas, laboratorios y acceso a tecnología, lo que favorece un entorno más propicio para el aprendizaje. Este es el caso del departamento de Nariño, donde el rendimiento académico, sobresale en su capital Pasto y en la exprovincia de Obando. Esta última incluye, municipios como Ipiales, Túquerres, Iles y Funes entre otros, los cuales, presentan mejores indicadores en comparación con las zonas más rurales y aisladas del departamento. Estas áreas se benefician, de un mayor acceso a servicios educativos, de salud y empleo, lo que contribuye a reducir las carencias asociadas a la pobreza multidimensional (DANE, 2020b).

El desempeño en las Pruebas Saber 9º, muestra diferencias marcadas según la naturaleza del colegio. En el caso de los estudiantes nariñenses, un bajo rendimiento, se asocia con aquellos que asisten a colegios oficiales, mientras que los estudiantes de colegios privados tienden a obtener mejores resultados. Según Giraldo y Vergara, (2018) esto puede explicarse, por las diferencias significativas entre la educación pública y privada en Colombia, en términos de financiación, gestión y calidad. La educación privada, aunque implica un mayor costo, suele ofrecer mejor infraestructura, mayores recursos y mejores resultados académicos. Además, asistir a un colegio privado, está estrechamente relacionado con estudiantes que tienen mayores recursos económicos.

Estos hallazgos coinciden, con estudios previos, como los de Seibold, (2000), Montero *et al.*, (2007) y Garbanzo, (2007), que destacan una relación significativa entre el nivel socioeconómico de los estudiantes y su desempeño académico. Por otro lado, López y González, (2016) indican, que existe evidencia de la influencia de la clase social en los logros académicos, siendo los estudiantes, con un estatus socioeconómico más alto, los que demuestran un mejor desempeño.

La educación de la madre, es un factor determinante en el desempeño académico de los estudiantes nariñenses en las Pruebas Saber 9º. Se observa, que un bajo rendimiento, está asociado con estudiantes, cuyas madres tienen un nivel educativo de primaria o bachillerato, mientras que un desempeño destacado, suele corresponder a aquellos, cuyas madres poseen títulos de educación superior, ya sean de pregrado o postgrado. En Colombia, aproximadamente el 40% de los menores, viven únicamente con su madre (Profamilia, 2015), lo que refuerza, la influencia directa que ellas tienen en el desarrollo académico de sus hijos.

Según Espejel y Jiménez, (2019) el nivel educativo de la madre, tiene un impacto positivo en el rendimiento escolar, especialmente, cuando supera la educación básica. Estudios como los de Mullis *et al.*, (2003), Moreira, (2007), Matute *et al.*, (2009), Nieto y Ramos, (2011) y García y Skrita, (2019) han demostrado, que una madre con mayor nivel educativo, sirve como un modelo a seguir, ofreciendo orientación y apoyo efectivos en el ámbito académico. Además, las madres con niveles educativos más altos, tienden a tener mejores oportunidades económicas, lo que les permite, proporcionar recursos esenciales, como materiales escolares, acceso a tecnología y un entorno adecuado para el aprendizaje. Por otro lado, también suelen estar más comprometidas e involucradas en la vida escolar de sus hijos, lo que se traduce en mayor motivación y respaldo emocional.

Otro de los factores clave en el desempeño académico de los estudiantes nariñenses en las Pruebas Saber 9º, es la repetición de grados a lo largo de su trayectoria escolar. Se ha observado, que aquellos alumnos que han reprobado algún curso, suelen obtener resultados más bajos en

estas evaluaciones, mientras que quienes han avanzado sin interrupciones académicas, tienden a lograr un mejor rendimiento. Diversos estudios, como los de Canales y Solís, (2009), Bayona y López, (2018), Pinzón, (2018) y Díaz y Ruiz, (2018) han demostrado que repetir un año, no siempre contribuye a mejorar el aprendizaje; por el contrario, en muchos casos, puede generar efectos negativos a largo plazo. Estos hallazgos sugieren, que la continuidad en el proceso educativo y el aprendizaje sin interrupciones, favorecen un mejor desempeño académico. En este sentido, la repetición de grados escolares, puede representar una barrera significativa para el desarrollo académico de los estudiantes.

4.6 ÁRBOLES DE DECISIÓN PARA EL DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO EN LAS PRUEBAS SABER 9º DE LOS ESTUDIANTES DE LOS DEPARTAMENTOS DE CAUCA Y NARIÑO

Para identificar patrones comunes en el desempeño académico de los estudiantes de Grado 9º de las instituciones educativas de los departamentos del Cauca y Nariño, que presentaron las Pruebas Saber 9º, en 2017, mediante árboles de decisión, se unificaron los conjuntos de datos *sbr9_cauca_final* y *sbr9_narino_final* en uno solo, denominado *sbr9_cauca_narino_final*. Con el objetivo de mejorar la capacidad de generalización del modelo, se eliminó el atributo correspondiente a la subregión del departamento, donde se ubica cada institución (*cole_region_ubicacion*).

Como resultado, se obtuvo un conjunto de datos, compuesto por 35,363 registros y 17 atributos. A partir de este conjunto, se generaron tres subconjuntos específicos: *sbr9_cauca_narino_lenguaje*, *sbr9_cauca_narino_matemáticas* y *sbr9_cauca_narino_general*. En cada caso, el atributo de clase o target, corresponde al desempeño en la competencia respectiva, con valores discretizados en “sobre la media” y “bajo la media”.

En la tabla 36, se presenta la distribución observada del desempeño en las Competencias de Lenguaje, Matemáticas y en el Desempeño General de los estudiantes de los departamentos de Cauca y Nariño en las pruebas Saber 9º.

Para evaluar la calidad de los modelos generados, cada conjunto fue particionado en dos subconjuntos: el 80% de las instancias se destinó al entrenamiento y el 20% restante a la prueba.

Tabla 36. Distribución del Desempeño Observado de los Estudiantes del Cauca y Nariño en las Competencias de Lenguaje, Matemáticas y Desempeño General

Conjunto	Desempeño				
	Bajo la Media	%	Sobre la Media	%	Total
sbr9_cauca_narino_lenguaje	21,083	59.6	14,280	40.4	35,363
sbr9_cauca_narino_matematicas	20,928	59.1	14,435	40.9	35,363
sbr9_cauca_narino_general	21,580	61	13,783	39	35,363

Fuente. Esta Investigación

De acuerdo con los datos presentados en la tabla 36, el desempeño académico en los departamentos de Cauca y Nariño revela una tendencia crítica frente al estándar nacional. En la prueba de Lenguaje, de un universo de 35,363 evaluados, el 59.6 % (21,083 estudiantes) se sitúa por debajo de la media nacional, mientras que solo el 40.4 % logra superarla. Un comportamiento similar se observa en Matemáticas, donde el 59.1 % de los estudiantes no alcanza el promedio país, frente a un 40.9 % que se posiciona por encima. Esta brecha se acentúa en los resultados del Desempeño General de las Pruebas Saber 9^o; en este indicador, el 61 % (21,580 alumnos) se ubica por debajo de la media nacional, lo que deja apenas a un 39 % (13,783 alumnos) en el nivel superior. Estas cifras evidencian que más de seis de cada diez estudiantes en la región enfrentan rezagos significativos en sus competencias básicas comparados con el resto del país.

4.6.1 Clasificación Mediante Árboles de Decisión para el Desempeño en la Competencia de Lenguaje de los Estudiantes de los Departamentos de Cauca y Nariño

Para construir un modelo de clasificación, basado en árboles de decisión, con el algoritmo J48, con el objetivo de identificar patrones en el desempeño académico en la Competencia de Lenguaje de las Pruebas Saber 9º, presentadas por estudiantes de los departamentos de Cauca y Nariño, se utilizó el conjunto de datos *sbr9_cauca_narino_lenguaje*.

Se estableció como variable objetivo (*target*), el desempeño en Lenguaje, representado por el atributo *clase_desemp_lenguaje*. Para evaluar el modelo, se dividió el conjunto de datos en dos partes: un 80% destinado al entrenamiento (28,290 instancias, según los parámetros de la tabla 31) y un 20% reservado para la prueba (7,073 instancias).

Además, se realizó una poda del árbol de decisión, utilizando los hiperparámetros de referencia: un factor de confianza de $C=0.25$ y un mínimo de instancias por nodo de $M=0.5\%$, conforme a lo establecido en la tabla 33.

Para validar las reglas o patrones obtenidos al interpretar cada rama del árbol, desde el nodo raíz hasta el nodo hoja, se aplicaron las métricas de Soporte y Confianza Mínimos. El Soporte Mínimo, representa el porcentaje mínimo de instancias, que llegan a un nodo hoja en relación con el total del conjunto de datos, mientras que la Confianza Mínima indica, el porcentaje de instancias correctamente clasificadas en cada nodo, respecto al total que llega al nodo hoja.

Estas métricas fueron adaptadas, de la Técnica de Minería de Datos, conocida como Reglas de Asociación (Agrawal *et al.*, 1993). Como criterio de selección, solo se conservaron las ramas que superaron un soporte mínimo del 0.5% y una confianza mínima del 60%.

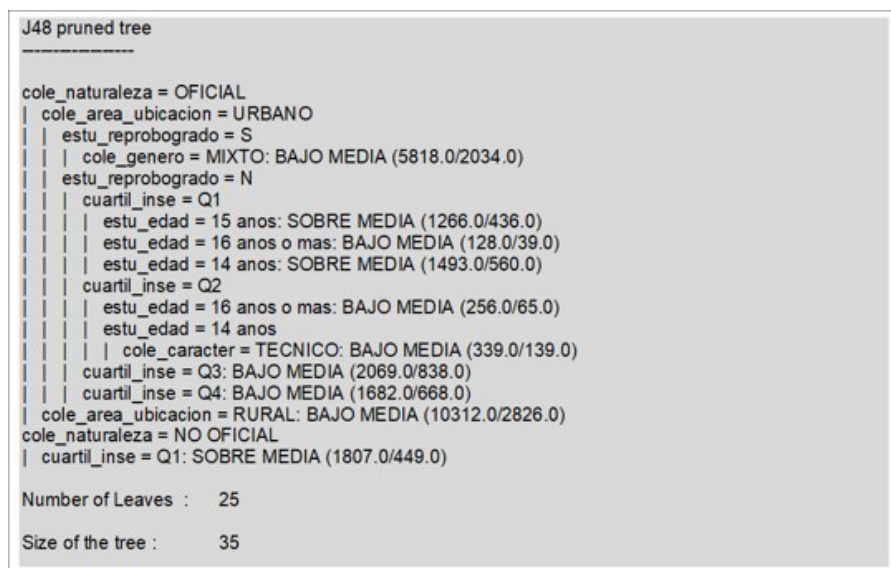
La Figura 31 ilustra el árbol de decisión generado para analizar la Competencia de Lenguaje de los estudiantes de Cauca y Nariño en las pruebas

Saber 9º. Con el fin de interpretar los patrones identificados, se seleccionaron las reglas más representativas del modelo, bajo el criterio de cumplir con un soporte mínimo del 0.5 % y una confianza superior al 60 %. Estos umbrales se aplicaron sobre el total de instancias del conjunto de entrenamiento, el cual comprende un registro de 28,290 datos, garantizando así la significancia estadística y la robustez de las reglas obtenidas.

Entre los patrones más relevantes, ordenados por soporte, se encuentran:

Regla 1. Si el estudiante, asiste a un colegio oficial, ubicado en el área rural, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9º, es probable, que este bajo la media nacional. El 36.5 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 72.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 61.1% de los estudiantes observados con resultados por debajo de la media, cumplen este patrón.

Figura 31. Árbol de Decisión Podado para el Desempeño en la Competencia de Lenguaje de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º



Fuente. Esta Investigación.

Regla 2. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, ha reprobó algún grado y el colegio es mixto, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 20.6% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 65% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 34.5% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 3. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado y se encuentra en el cuartil Q3 del índice socioeconómico, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 7.3 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 12.3% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 4. Si el estudiante, asiste a un colegio no oficial y se encuentra en el cuartil Q1 del índice socioeconómico, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 6.4 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 75.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 15.8% de los estudiantes observados con resultados por encima de la media, cumplen este patrón.

Regla 5. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado y se encuentra en el cuartil Q4 del índice socioeconómico, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 5.9 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 60.3% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 10% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 6. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, se encuentra en el cuartil Q1 del índice socioeconómico y tiene 14 años, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 5.3 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 62.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 13.1% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 7. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, se encuentra en el cuartil Q1 del índice socioeconómico y tiene 15 años, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este sobre la media nacional. El 4.5 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 65.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 11.1% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 8. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, se encuentra en el cuartil Q2 del índice socioeconómico, tiene 14 años y el colegio es de carácter técnico, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 1.2 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 2% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 9. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, se encuentra en el cuartil Q2 del índice socioeconómico y tiene 16 años o más, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9°, es probable, que este bajo la media nacional. El 0.9 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El

74.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.5% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 10. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, se encuentra en el cuartil Q1 del índice socioeconómico y tiene 16 años o más, entonces su desempeño académico en la Prueba de Lenguaje del Saber 9º, es probable, que este bajo la media nacional. El 0.5 % del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 69.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 0.8% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Igualmente, en la figura 32 se presenta la evaluación del modelo obtenido para la Competencia de Lenguaje de los estudiantes caucanos y nariñenses en las Pruebas Saber 9º, fundamentada en las métricas derivadas de la matriz de confusión.

Como se muestra en la figura 32, el modelo de árbol de decisión obtenido, tiene una exactitud del 66.6%. Esto significa, que clasifica correctamente el 66.6% de las instancias del conjunto de prueba, mientras que el 33.4% son clasificadas incorrectamente.

Por otro lado, la Raíz del Error Cuadrático Medio (*Root Mean Squared Error*, RMSE) del modelo, es de 0.47, lo que indica, que, en promedio, la diferencia entre las predicciones y los valores reales u observados, es moderada. Este nivel de error, es aceptable, dentro del contexto del análisis, reflejando una distancia razonable entre los valores predichos y los reales.

De acuerdo con la Matriz de Confusión, del total de 7,073 estudiantes (instancias) que conforman el conjunto de prueba del *sbr9_cauca_nariño_lenguaje*, el modelo predice correctamente, al 50.2% de estudiantes con desempeño en Lenguaje “bajo la media” (verdaderos negativos VN), que corresponde a 3,553 estudiantes y al 16.3% de estudiantes con desempeño académico “sobre la media” (verdaderos positivos VP), que

corresponde a 1156 estudiantes. Además, el modelo predice incorrectamente, al 9.1% de estudiantes que están “bajo la media” como “sobre la media” (falsos positivos FP), correspondiente a 645 estudiantes y al 24.3% de estudiantes que están “sobre la media” como “bajo la media” (falsos negativos FN), correspondiente a 1719 estudiantes.

Figura 32. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia De Lenguaje de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º

=== Summary ===									
Correctly Classified Instances	4709					66.5771 %			
Incorrectly Classified Instances	2364					33.4229 %			
Kappa statistic			0.264						
Mean absolute error			0.4364						
Root mean squared error			0.4673						
Relative absolute error			90.5536 %						
Root relative squared error			95.1438 %						
Total Number of Instances	7073								
=== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,846	0,598	0,674	0,846	0,750	0,280	0,664	0,698	BAJO MEDIA
	0,402	0,154	0,642	0,402	0,494	0,280	0,664	0,567	SOBRE MEDIA
Weighted Avg.	0,666	0,417	0,661	0,666	0,646	0,280	0,664	0,645	
=== Confusion Matrix ===									
a	b	<-- classified as							
3553	645	a = BAJO MEDIA							
1719	1156	b = SOBRE MEDIA							

Fuente. Esta Investigación.

En cuanto a la exactitud del modelo de árboles de decisión por clase, en la figura 32, se observa, que el modelo de árboles de decisión presenta, una precisión de Predicción de 0.674 para los estudiantes de los departamentos del Cauca y Nariño, cuyo desempeño en Lenguaje en las Pruebas Saber 9º, está “bajo la media” (*clase negativa*). Esto significa, que, de los casos clasificados en esta categoría, el 67.4% son correctos. En otras palabras, cuando el modelo predice que un estudiante tiene un desempeño bajo la media, hay una probabilidad del 67.4% de que esa clasificación sea acertada.

La Especificidad, también conocida como Tasa de Verdaderos Negativos (*True Positive Rate*, TP Rate) y *recall*, es de 0.846, lo que indica, que el modelo identifica correctamente al 84.6% de los estudiantes que realmente, presentan un desempeño bajo la media. Este resultado refleja, que el modelo tiene un buen nivel de Especificidad, para capturar instancias negativas reales. En términos simples, de todos los estudiantes que efectivamente pertenecen a esta categoría, el modelo reconoce correctamente al 84.6%.

Por otro lado, la Tasa de Falsos Positivos (*False Positive Rate*, FP Rate), es de 0.598, lo que significa, que el 59.8% de los estudiantes con desempeño “bajo la media” fueron incorrectamente clasificados como “sobre la media”.

El *F-Measure* o *F1-Score* es de 0.75, lo que indica, un equilibrio entre precisión y *recall*, aunque no de manera óptima. El modelo identifica de manera razonable, los estudiantes que están bajo la media (instancias negativas (*recall*)) y comete errores, de forma aceptable, en las predicciones negativas (precisión). Este resultado refleja, que el modelo tiene un compromiso entre evitar falsos negativos (precisión) y capturar los verdaderos negativos (*recall*).

Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.698, lo que indica, que el modelo tiene un mejor desempeño en la clasificación de los estudiantes con desempeño bajo la media, cuando se priorizan la Precisión y el *recall* en esta clase. Con un 69.8%, el modelo demuestra una capacidad sólida, para equilibrar correctamente estas predicciones.

Para el caso de los estudiantes de los departamentos del Cauca y Nariño, que están sobre la media (positivos) en el desempeño en Lenguaje en las Pruebas Saber 9º, el modelo tiene una *precisión* de Predicción de 0.642, lo que significa, que, del total de casos predichos que están sobre la media, el 64,2% están correctos. Cuando el modelo clasifica a un estudiante con desempeño sobre la media, hay un 64.2% de probabilidad de que esa Predicción sea correcta.

La Sensibilidad o Tasa de Verdaderos Positivos (*TP Rate*) y *recall* del modelo, es de 0.402, lo cual indica, que, el modelo es capaz de identificar correctamente al 40.2% de los estudiantes de los departamentos del Cauca y Nariño que realmente están sobre la media.

Por otra parte, la Tasa de Falsos Positivos (*FP Rate*) del modelo, es de 0.154, lo que significa, que el 15.4% de los estudiantes observados que estaban bajo la media, fueron predichos incorrectamente como sobre la media.

La *F-Measure* o *F1-Score*, es de 0.494, lo que indica, que, el modelo muestra un desempeño moderado. No es excelente, pero tampoco es pobre, lo que indica, que, hay un equilibrio aceptable entre *precisión* y *recall*, aunque ambos podrían mejorar. Sugiere que el modelo identifica correctamente algunas instancias positivas (*recall* razonable) y mantiene un nivel aceptable de precisión, pero aún hay margen significativo de mejora.

Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.567. Este valor indica, que el modelo tiene una capacidad moderada para identificar correctamente los casos positivos (“sobre la media”) mientras mantiene un nivel aceptable de precisión.

El Área Bajo la Curva ROC (*ROC Área*) del modelo, es de 0.664, un valor superior a 0.5, lo que indica, que el modelo tiene una capacidad moderada para diferenciar entre los estudiantes de los departamentos del Cauca y Nariño con desempeño “bajo la media” y aquellos con desempeño “sobre la media”.

Este resultado refleja, que el modelo logra detectar más verdaderos negativos (*VN*) y verdaderos positivos (*VP*) que falsos negativos (*FN*) y falsos positivos (*FP*), lo que representa un rendimiento superior al de un modelo aleatorio ($AUC = 0.5$). Sin embargo, aún existe margen de mejora, en términos de discriminación entre las clases.

La combinación de estas métricas, refleja un mejor desempeño del modelo en la clasificación de los estudiantes con rendimiento “bajo la media”. En esta categoría específica, la métrica PRC es superior a la ROC,

lo que indica, que el modelo funciona mejor cuando se consideran conjuntamente las métricas de precisión y *recall*.

Esta investigación, ha puesto el foco en los estudiantes con desempeño bajo la media. Identificar sus perfiles en los departamentos del Cauca y Nariño, permitirá a las directivas de las Instituciones Educativas, a las Secretarías de Educación y al Ministerio de Educación Nacional (MEN) tomar decisiones fundamentadas en datos. Esto contribuirá, a mejorar la calidad de la Educación Básica Secundaria y, en particular, a reforzar las habilidades de Lenguaje de los estudiantes de estos departamentos.

4.6.2 Clasificación mediante Árboles de Decisión para el Desempeño en la Competencia de Matemáticas de los Estudiantes de los Departamentos de Cauca y Nariño

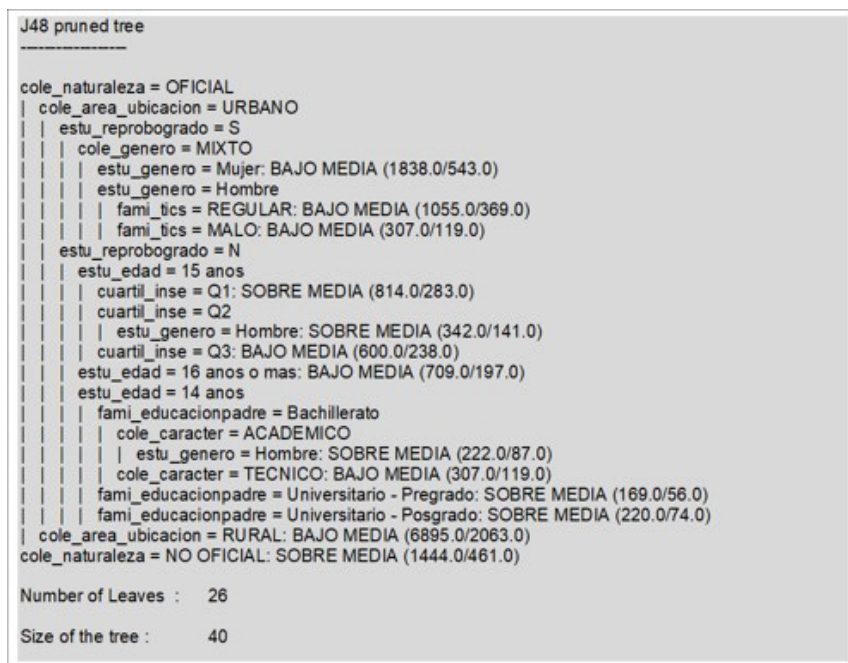
Para construir el modelo de clasificación por árboles de decisión, con el algoritmo J48, que permita descubrir patrones de desempeño académico en la Competencia de Matemáticas de las Pruebas Saber 9º, que presentaron los estudiantes de los departamentos de Cauca y Nariño, se utilizó el conjunto de datos *sbr9_cauca_narino_matematicas*. Se escogió como clase o target el desempeño en Matemáticas representado por el atributo *clase_desemp_matematicas*. Se probó el modelo, particionando este conjunto de datos, en dos conjuntos: el conjunto de entrenamiento con el 80% (28,290) del total de instancias (de acuerdo con los parámetros definidos en la tabla 31) y el conjunto de prueba con el 20% (7,073) del total de instancias para el conjunto de prueba. Se pudo el árbol de decisión con los hiperparámetros: factor de confianza $C=0.25$ y mínimo número de instancias por nodo $M=0.5\%$, según los parámetros de referencia establecidos en la tabla 33.

Para determinar la validez de las reglas o patrones, resultantes de interpretar cada rama del árbol desde el nodo raíz hasta el nodo hoja, se utilizaron las métricas de *soporte mínimo* (i.e. porcentaje mínimo del número de instancias, que llegan a un nodo hoja con respecto al número total de instancias del conjunto de datos) y *confianza mínima* (i.e. el porcentaje mínimo de instancias correctamente clasificadas en cada nodo

con respecto al número total de instancias que llegan al nodo hoja), adaptadas de la Técnica de Minería de Datos Reglas de Asociación (Agrawal *et al.*,1993) y se dejaron solo las ramas que superen el soporte mínimo de 0.5% y una confianza mínima de 60%.

La Figura 33 presenta el árbol de decisión generado para analizar la competencia de Matemáticas en los estudiantes de Cauca y Nariño.

Figura 33. Árbol de Decisión Podado Para el Desempeño en la Competencia de Matemáticas de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9^o



Fuente. Esta Investigación.

Con el fin de interpretar los patrones identificados, se seleccionaron las reglas más representativas del modelo, bajo el criterio de cumplir con un soporte mínimo del 0.5 % y una confianza superior al 60 %. Estos umbrales se aplicaron sobre el total de instancias del conjunto de entrenamiento, el cual comprende un registro de 28,290 datos, garantizando así la significancia estadística y la robustez de las reglas obtenidas.

Entre los patrones más importantes ordenados por soporte están:

Regla 1. Si el estudiante, asiste a un colegio oficial, ubicado en el área rural, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 24.4% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 70.1% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 41.1% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 2. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, ha reprobó algún grado, el colegio es mixto y es mujer, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 6.5% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 70.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 11% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 3. Si el estudiante, asiste a un colegio no oficial, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 5.1% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 68.1% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 12.5% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 4. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, ha reprobó algún grado, el colegio es mixto, es hombre y el índice TICs familiar es regular, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 3.7% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 65% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 6.3% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 5. Si el estudiante asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, tiene 15 años y pertenece al cuartil Q1 del índice socioeconómico, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 2.9% del total de estudiantes de los departamentos del Cauca y Nariño evaluados se clasifican de esta manera. El 65.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 7.1% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 6. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado y tiene 16 años o más entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 2.5% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 72.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 4.2% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 7. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, tiene 15 años y pertenece al cuartil Q3 del índice socioeconómico, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 2.1% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 60.3% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.6% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 8. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, tiene 15 años, pertenece al cuartil Q2 del índice socioeconómico y es hombre, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 1.2% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, están correcta-

mente clasificados y el 3% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 9. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, ha reprobó algún grado, el colegio es mixto, es hombre y el índice TICs familiar es malo, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9° es probable, que este bajo la media nacional. El 1.1% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 61.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.8% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 10. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, tiene 14 años, el nivel educativo máximo del padre es bachillerato y el colegio es de carácter técnico, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este bajo la media nacional. El 1.1% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 61.2% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.8% de los estudiantes observados con resultados por debajo de la media, cumplen este patrón.

Regla 11. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, tiene 14 años y el nivel educativo máximo del padre es postgrado, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 0.8% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 66.4% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.9% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 12. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, tiene 14 años, el nivel educativo máximo del padre es bachillerato, el colegio es de carácter académico

y es hombre, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 0.8% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 60.8% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.9% de los estudiantes observados con resultados por encima de la media, cumplen este patrón.

Regla 13. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobó ningún grado, tiene 14 años y el nivel educativo máximo del padre es universitario, entonces su desempeño académico en la Prueba de Matemáticas del Saber 9°, es probable, que este sobre la media nacional. El 0.6% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 66.9% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.8% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Por otra parte, en la figura 34 se presenta la evaluación del modelo obtenido para la Competencia de Matemáticas de los estudiantes caucanos y nariñenses en las Pruebas Saber 9°, fundamentada en las métricas derivadas de la matriz de confusión.

Figura 34. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Matemáticas de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º

```

=== Summary ===
Correctly Classified Instances      4578      64.725 %
Incorrectly Classified Instances    2495      35.275 %
Kappa statistic                    0.2232
Mean absolute error                 0.4464
Root mean squared error             0.4733
Relative absolute error             92.3425 %
Root relative squared error         96.2225 %
Total Number of Instances          7073

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,841    0,632    0,657    0,841    0,738     0,240    0,647    0,684    BAJO MEDIA
      0,368    0,159    0,617    0,368    0,461     0,240    0,647    0,547    SOBRE MEDIA
Weighted Avg. 0,647    0,438    0,641    0,647    0,624     0,240    0,647    0,628

=== Confusion Matrix ===

  a   b  ← classified as
3510 662 |  a = BAJO MEDIA
1833 1068 | b = SOBRE MEDIA

```

Fuente. Esta Investigación.

Como se observa en la figura 34, el modelo de árbol de decisión resultante, tiene una exactitud del 64.7, es decir, que el modelo clasifica correctamente al 64.7% de las instancias del conjunto de prueba e incorrectamente a un 35.3% de las instancias probadas. La Raíz del Error Cuadrático Medio RMSE (por su sigla en inglés *Root Mean Squared Error*) del modelo, es de 0.47, lo que indica, que el nivel promedio de error entre las predicciones del modelo y los valores reales u observados, es aceptable. La distancia entre los valores predichos y reales es moderada.

De acuerdo con la Matriz de Confusión, del total de 7,073 estudiantes (instancias) que conforman el conjunto de prueba del *sbr9_cauca_nariño_matematicas*, el modelo predice correctamente al 49.6% de estudiantes con desempeño en Matemáticas “bajo la media” (verdaderos negativos VN), que corresponde a 3,510 estudiantes y al 15.1% de estudiantes con desempeño académico “sobre la media” (verdaderos positivos VP), que corresponde a 1068 estudiantes. Además, el modelo predice incorrectamente, al 9.4% de estudiantes que están “bajo la media” como “sobre la media” (falsos positivos FP), correspondiente a 662 estudiantes y al

25.9% de estudiantes que están “sobre la media” como “bajo la media” (falsos negativos FN), correspondiente a 1833 estudiantes.

En cuanto a la exactitud del modelo de árboles de decisión por clase, en la figura 34, se puede observar, que para el caso de los estudiantes de los departamentos del Cauca y Nariño, que están bajo la media (negativos) en el desempeño en Matemáticas en las Pruebas Saber 9º, el modelo tiene una *precisión* de Predicción de 0.657, lo que significa, que, del total de casos predichos que están bajo la media, el 65.7% están correctos, es decir, cuando el modelo clasifica a un estudiante con desempeño bajo la media, hay un 65.7% de probabilidad de que esa Predicción sea correcta.

La Especificidad o Tasa de Verdaderos Negativos (*TP Rate*) y *recall* del modelo, es de 0.841, lo cual indica, que, el modelo es capaz de identificar correctamente al 84.1% de los estudiantes de los departamentos del Cauca y Nariño, que realmente están bajo la media. Indica, que el modelo tiene un buen nivel de Especificidad o capacidad para identificar correctamente el 84.1% de las instancias negativas reales. En términos sencillos, de todas las instancias que realmente pertenecen a la clase negativa, el modelo, ha logrado reconocer correctamente el 84.1% de ellas.

Por otra parte, la Tasa de Falsos Negativos (*FP Rate*) del modelo, es de 0.632, lo que significa, que el 63.2% de los estudiantes observados que estaban sobre la media, fueron predichos como bajo la media.

La *F-Measure* o *F1-Score*, es de 0.74, que es el promedio armónico entre la *precisión* y el *recall* de los que están bajo la media. Indica, un buen equilibrio entre *precisión* y *recall*. Significa, que el modelo tiene un rendimiento razonablemente bueno, en identificar correctamente a los estudiantes bajo la media, minimizando tanto los falsos positivos como los falsos negativos. El modelo identifica de manera razonable, los estudiantes que están bajo la media (instancias negativas (*recall*)) y comete errores, de forma aceptable, en las predicciones negativas (*precisión*). Es útil esta métrica, cuando las clases están desbalanceadas como es el caso de esta investigación.

Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.684. Este valor indica, que el modelo tiene un rendimiento razonablemente bueno, en la clasificación de los estudiantes bajo la media. Cuanto mayor sea el Área Bajo la Curva, mejor será el rendimiento del modelo en términos de precisión y *recall*. Con un 68.4%, el modelo tiene una capacidad más sólida, para equilibrar correctamente las predicciones “bajo la media”. Útil, cuando se trata de conjuntos de datos desbalanceados como es el caso del conjunto de datos *sbr9_cauca_narino_matematicas*.

Para el caso de los estudiantes de los departamentos del Cauca y Nariño que están sobre la media (positivos) en el desempeño en Matemáticas en las Pruebas Saber 9º, el modelo tiene una *precisión* de Predicción de 0.617, lo que significa, que, del total de casos predichos que están sobre la media, el 61,7% están correctos. Cuando el modelo clasifica a un estudiante con desempeño sobre la media, hay un 61.7% de probabilidad, de que esa Predicción sea correcta.

La Sensibilidad o Tasa de Verdaderos Positivos (*TP Rate*) y *Recall* del modelo, es de 0.368, lo cual indica, que, el modelo es capaz de identificar correctamente al 36.8% de los estudiantes de los departamentos del Cauca y Nariño, que realmente están sobre la media.

Por otro lado, la Tasa de Falsos Positivos (*FP Rate*) del modelo, es de 0.159, lo que significa, que el 15.9% de los estudiantes observados que estaban bajo la media, fueron predichos incorrectamente, como sobre la media.

La *F-Measure* o *F1-Score*, es de 0.461, lo que indica el modelo, muestra un desempeño moderado. No es excelente, pero tampoco es pobre, lo que indica, que hay un equilibrio aceptable entre *precisión* y *recall*, aunque ambos podrían mejorar. Sugiere, que el modelo identifica correctamente algunas instancias positivas (*recall* razonable) y mantiene un nivel aceptable de precisión, pero aún hay margen significativo de mejora.

Finalmente, el área bajo la curva PRC (*PRC Área*) es de 0.547. Este valor indica que el modelo tiene una capacidad moderada para identificar correctamente los casos positivos (“sobre la media”) mientras mantiene un nivel aceptable de precisión.

El Área Bajo la Curva ROC (*ROC Área*) del modelo, es de 0.647 (mayor que 0.5), lo que indica, que, el modelo tiene una capacidad razonable para distinguir los estudiantes de los departamentos del Cauca y Nariño, que están etiquetados como “bajo la media”, de los estudiantes que están etiquetados como “sobre la media”, ya que, es capaz de detectar más verdaderos negativos (VN) y verdaderos positivos (VP) que de falsos negativos (FN) y falsos positivos (FP). Esto es mejor que un modelo al azar ($AUC = 0.5$), pero aún deja espacio para mejorar en términos de discriminación entre las clases.

En la combinación de estas métricas, se aprecia un mejor desempeño del modelo en la clasificación de los estudiantes etiquetados como “bajo la media” en el desempeño en Matemáticas. Para esta clase, la métrica PRC, es más alto que el ROC, lo que muestra, que el modelo se comporta mejor, cuando las métricas de *precisión* y *recall* se evalúan juntas.

Esta investigación se centró, en los estudiantes con desempeño bajo la media. Identificar los perfiles de los estudiantes de los departamentos del Cauca y Nariño, cuyo desempeño en la Competencia de Matemáticas en las Pruebas Saber 9º, resulta bajo, proporcionará a las directivas de las Instituciones Educativas de estos departamentos, a las Secretarías de Educación y al MEN, información valiosa para la toma de decisiones fundamentadas en datos. Esto contribuirá al mejoramiento de la calidad de la Educación Básica Secundaria y, en particular, al fortalecimiento de las habilidades en Matemáticas de los estudiantes del Cauca y Nariño.

4.6.3 Clasificación mediante Árboles de Decisión para el Desempeño General de los Estudiantes de los Departamentos de Cauca y Nariño en las Pruebas Saber 9º

Para construir el modelo de clasificación por árboles de decisión, con el algoritmo J48, que permita descubrir patrones de Desempeño General, en las Pruebas Saber 9º, que presentaron los estudiantes de los departamentos de Cauca y Nariño, se utilizó el conjunto de datos *sbr9_cauca_narino_general*. Se escogió como clase o target, el Desempeño General, representado por el atributo *clase_desemp_general*. Se probó

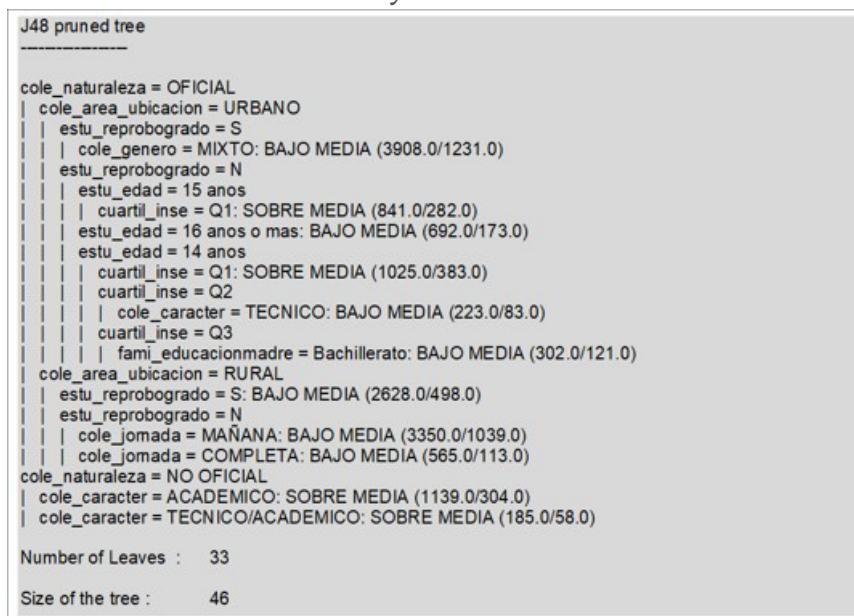
el modelo, particionando este conjunto de datos en dos conjuntos: el conjunto de entrenamiento con el 80% (28,290) del total de instancias (de acuerdo con los parámetros definidos en la tabla 31 y el conjunto de prueba con el 20% (7,073) del total de instancias para el conjunto de prueba. Se podó el árbol de decisión con los hiperparámetros: factor de confianza $C=0.25$ y mínimo número de instancias por nodo $M=0.5\%$, según los parámetros de referencia establecidos en la tabla 33.

Para determinar la validez de las reglas o patrones, resultantes de interpretar cada rama del árbol desde el nodo raíz hasta el nodo hoja, se utilizaron las métricas de *soporte mínimo* (i.e. porcentaje mínimo del número de instancias que llegan a un nodo hoja con respecto al número total de instancias del conjunto de datos) y *confianza mínima* (i.e. el porcentaje mínimo de instancias correctamente clasificadas en cada nodo con respecto al número total de instancias que llegan al nodo hoja), adaptadas de la Técnica de Minería de Datos Reglas de Asociación (Agrawal et al.,1993) y se dejaron solo las ramas que superen el soporte mínimo de 0.5% y una confianza mínima de 60%.

En la figura 35, se aprecia el modelo de árbol de decisión resultante, para el Desempeño General de los estudiantes de estos dos departamentos.

Con el fin de interpretar los patrones identificados de la figura 35, se seleccionaron las reglas más representativas del modelo, bajo el criterio de cumplir con un soporte mínimo del 0.5 % y una confianza mínima del 60 %. Estos umbrales se aplicaron sobre el total de instancias del conjunto de entrenamiento, el cual comprende un registro de 28,290 datos, garantizando así la significancia estadística y la robustez de las reglas obtenidas.

Figura 35. Árbol de Decisión Podado para el Desempeño General de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º



Fuente. Esta Investigación.

Entre los patrones más importantes ordenados por soporte están:

Regla 1. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, ha reprobado algún grado y el colegio es mixto, entonces su Desempeño General en las Pruebas Saber 9º, es probable, que este bajo la media nacional. El 13.8% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 68.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 22.6% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 2. Si el estudiante, asiste a un colegio oficial, ubicado en el área rural, no ha reprobado ningún grado y asiste en la jornada de la mañana, entonces su Desempeño General en las Pruebas Saber 9º, es probable, que este bajo la media nacional. El 11.8% del total de estudiantes de

los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 69% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 19.4% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 3. Si el estudiante, asiste a un colegio oficial, ubicado en el área rural y ha reprobado algún grado, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 9.3% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 81.1% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 15.2% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 4. Si el estudiante, asiste a un colegio no oficial de carácter académico entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 4% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 73.3% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 10.4% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 5. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobado ningún grado, tiene 14 años y está en el cuartil Q1 del índice socioeconómico, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 3.6% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 62.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 9.3% de los estudiantes observados con resultados sobre la media, cumplen este patrón.

Regla 6. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobado ningún grado, tiene 15 años y está en el cuartil Q1 del índice socioeconómico, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 3% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se

clasifican de esta manera. El 66.5% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 7.6% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Regla 7. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobado ningún grado y tiene 16 años o más, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 2.4% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 75% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 4% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 8. Si el estudiante, asiste a un colegio oficial, ubicado en el área rural, no ha reprobado ningún grado y estudia en un colegio de jornada completa, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 2% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 80% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 3.3% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 9. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobado ningún grado, tiene 14 años, está en el cuartil Q3 del índice socioeconómico y el nivel educativo máximo de la madre, es bachillerato, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este bajo la media nacional. El 1.1% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 60% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.7% de los estudiantes observados con resultados bajo la media, cumplen este patrón.

Regla 10. Si el estudiante, asiste a un colegio oficial, ubicado en el área urbana, no ha reprobado ningún grado, tiene 14 años, está en el cuartil Q2 del índice socioeconómico y el carácter del colegio es técnico, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este

bajo la media nacional. El 0.8% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 64.4% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.3% de los estudiantes observados con resultados inferiores a la media, cumplen este patrón.

Regla 11. Si el estudiante, asiste a un colegio no oficial de carácter técnico/académico, entonces su Desempeño General en las Pruebas Saber 9°, es probable, que este sobre la media nacional. El 0.7% del total de estudiantes de los departamentos del Cauca y Nariño evaluados, se clasifican de esta manera. El 68.6% de los estudiantes clasificados en esta categoría, están correctamente clasificados y el 1.7% de los estudiantes observados con resultados superiores a la media, cumplen este patrón.

Por otra parte, en la figura 36 se presenta la evaluación del modelo obtenido para el Desempeño General de los estudiantes caucanos y nariñenses en las Pruebas Saber 9°, fundamentada en las métricas derivadas de la matriz de confusión.

Figura 36. Métricas de Evaluación del Modelo de Árbol de Decisión para el Desempeño General de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9°

```

=== Summary ===

Correctly Classified Instances   4755      67.2275 %
Incorrectly Classified Instances 2318      32.7725 %
Kappa statistic                 0.2609
Mean absolute error             0.4244
Root mean squared error         0.4626
Relative absolute error         89.0835 %
Root relative squared error     94.6539 %
Total Number of Instances       7073

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.863    0.621    0.681    0.863    0.761     0.280    0.682    0.733    BAJO MEDIA
      0.379    0.137    0.643    0.379    0.477     0.280    0.682    0.570    SOBRE MEDIA
Weighted Avg. 0.672    0.430    0.666    0.672    0.649     0.280    0.682    0.669

=== Confusion Matrix ===

  a  b  <-- classified as
3697 588 |  a = BAJO MEDIA
1730 1058 | b = SOBRE MEDIA

```

Fuente. Esta Investigación.

Como se observa en la figura 36, el modelo de árbol de decisión resultante, tiene una exactitud del 67.2, es decir, que el modelo clasifica correctamente, al 67.2% de las instancias del conjunto de prueba e incorrectamente, a un 32.8% de las instancias probadas. La Raíz del Error Cuadrático Medio RMSE (por su sigla en inglés *Root Mean Squared Error*) del modelo, es de 0.46, lo que indica, que el nivel promedio de error entre las Predicciones del modelo y los valores reales u observados, es aceptable. La distancia entre los valores predichos y reales, es moderada.

De acuerdo con la Matriz de Confusión, del total de 7,073 estudiantes (instancias) que conforman el conjunto de prueba del *sbr9_cauca_nariño_general*, el modelo predice correctamente al 52.3% de estudiantes con Desempeño General “bajo la media” (verdaderos negativos VN), que corresponde a 3,697 estudiantes y al 15% de estudiantes con desempeño académico “sobre la media” (verdaderos positivos VP), que corresponde a 1058 estudiantes. Además, el modelo predice incorrectamente, al 8.3% de estudiantes que están “bajo la media” como “sobre la media” (falsos positivos FP), correspondiente a 588 estudiantes y al 24.5% de estudiantes que están “sobre la media” como “bajo la media” (falsos negativos FN), correspondiente a 1730 estudiantes.

En cuanto a la Exactitud del modelo de árboles de decisión por clase, se puede observar, que para el caso de los estudiantes de los departamentos del Cauca y Nariño que están bajo la media (negativos) en el Desempeño General de las Pruebas Saber 9º, el modelo tiene una precisión de Predicción de 0.681, lo que significa, que, del total de casos predichos que están bajo la media, el 68.1% están correctos, es decir, cuando el modelo clasifica a un estudiante con desempeño bajo la media, hay un 68.1% de probabilidad, de que esa Predicción sea correcta.

La Especificidad o Tasa de Verdaderos Negativos (*TP Rate*) y *Recall* del modelo, es de 0.863, lo cual indica, que, el modelo es capaz de identificar correctamente al 86.3% de los estudiantes de los departamentos del Cauca y Nariño, que verdaderamente tienen su Desempeño General bajo la media. Indica, que el modelo, tiene un buen nivel de Especificidad o capacidad para identificar correctamente el 86.3% de las instancias

negativas reales. En términos sencillos, de todas las instancias que realmente pertenecen a la clase negativa, el modelo, ha logrado reconocer correctamente el 86.3% de ellas.

Por otra parte, la Tasa de Falsos Negativos (*FP Rate*) del modelo, es de 0.621, lo que significa, que el 62.1% de los estudiantes observados que estaban sobre la media, fueron predichos como bajo la media.

La *F-Measure* o *F1-Score*, es de 0.761, que es el promedio armónico entre la Precisión y el *recall* de los que están bajo la media. Indica, un buen equilibrio entre *precisión* y *recall*. Significa, que el modelo tiene un rendimiento razonablemente bueno, en identificar correctamente a los estudiantes bajo la media, minimizando tanto los falsos positivos como los falsos negativos. El modelo identifica de manera razonable, los estudiantes que están bajo la media (instancias negativas (*recall*)) y comete errores, de forma aceptable, en las predicciones negativas (*precisión*). Es útil esta métrica, cuando las clases están desbalanceadas, como es el caso de esta investigación.

Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.733. Este valor indica, que el modelo tiene un rendimiento razonablemente bueno, en la clasificación de los estudiantes bajo la media. Cuanto mayor sea el área bajo la curva, mejor será el rendimiento del modelo en términos de precisión y *recall*. Con un 73.3%, el modelo tiene una capacidad más sólida para equilibrar correctamente las predicciones “bajo la media”. Útil, cuando se trata de conjuntos de datos desbalanceados, como es el caso del conjunto de datos *sbr9_cauca_narino_general*.

Para el caso de los estudiantes de los departamentos del Cauca y Nariño, que están sobre la media (positivos) en el Desempeño General de las Pruebas Saber 9º, el modelo tiene una *precisión* de Predicción de 0.643, lo que significa, que, del total de casos predichos que están sobre la media, el 64,3% están correctos. Cuando el modelo clasifica a un estudiante con Desempeño General sobre la media, hay un 64.3% de probabilidad, de que esa Predicción sea correcta.

La Sensibilidad o Tasa de Verdaderos Positivos (*TP Rate*) y *recall* del modelo, es de 0.379. Indica que, el modelo es capaz de identificar correctamente al 37.9% de los estudiantes de los departamentos del Cauca y Nariño, que realmente están sobre la media en el Desempeño General de las Pruebas Saber 9^o.

Por otra parte, la Tasa de Falsos Positivos (*FP Rate*) del modelo, es de 0.137, lo que significa, que el 13.7% de los estudiantes observados que estaban bajo la media, fueron predichos incorrectamente como sobre la media.

La *F-Measure* o *F1-Score*, es de 0.477, lo que indica, que el modelo muestra un desempeño moderado. No es excelente, pero tampoco es pobre, lo que indica, que hay un equilibrio aceptable entre precisión y *recall*, aunque ambos podrían mejorar. Sugiere, que el modelo identifica correctamente algunas instancias positivas (*recall* razonable) y mantiene un nivel aceptable de precisión, pero aún hay margen significativo de mejora.

Finalmente, el Área Bajo la Curva PRC (*PRC Área*), es de 0.57. Este valor indica, que el modelo tiene una capacidad moderada para identificar correctamente los casos positivos (“sobre la media”) mientras mantiene un nivel aceptable de precisión.

El Área Bajo la Curva ROC (*ROC Área*) del modelo, es de 0.682 (mayor que 0.5) lo que indica, que, el modelo tiene una capacidad moderada para distinguir los estudiantes de los departamentos del Cauca y Nariño, que están etiquetados como “bajo la media”, de los estudiantes que están etiquetados como “sobre la media”, ya que, es capaz de detectar más verdaderos negativos (VN) y verdaderos positivos (VP) que de falsos negativos (FN) y falsos positivos (FP). Esto es mejor que un modelo al azar ($AUC = 0.5$), pero aún deja espacio para mejorar en términos de discriminación entre las clases.

En la combinación de estas métricas, se aprecia un mejor desempeño del modelo, en la clasificación de los estudiantes etiquetados como “bajo la media” en el desempeño en Matemáticas. Para esta clase, la métrica PRC, es más alto que el ROC, lo que muestra, que el modelo se comporta mejor cuando las métricas de *precisión* y *recall* se evalúan juntas.

Esta investigación se centró, en los estudiantes con desempeño bajo la media. Identificar los perfiles de los estudiantes de los departamentos del Cauca y Nariño, cuyo Desempeño General en las Pruebas Saber 9º, resulta bajo, proporcionará a las directivas de las Instituciones Educativas de estos departamentos, a las Secretarías de Educación y al MEN, información valiosa para la toma de decisiones fundamentadas en datos. Esto contribuirá, al mejoramiento de la calidad de la Educación Básica Secundaria y, en particular, al fortalecimiento de las habilidades en Lenguaje y Matemáticas de los estudiantes del Cauca y Nariño, competencias que fueron evaluadas en las Pruebas Saber 9º, en el año 2017.

4.6.4 Discusión de Resultados de los Modelos Obtenidos de Desempeño Académico de los Estudiantes de los Departamento de Cauca y Nariño en las Pruebas Saber 9º

Al unificar los dos conjuntos de datos de los estudiantes de los departamentos del Cauca y Nariño (*sbr9_cauca_final* y *sbr9_narino_final*) en un solo conjunto de datos (*sbr9_cauca_narino_final*), con el fin de obtener patrones comunes de desempeño en las Competencias de Lenguaje, Matemáticas y Desempeño General de las Pruebas Saber 9º, el análisis de estos patrones revela, que los factores más determinantes en los resultados académicos, son la naturaleza de la institución educativa y el área de ubicación del colegio.

El desempeño en las Pruebas Saber 9º, muestra diferencias marcadas, según la naturaleza del colegio. En el caso de los estudiantes de los departamentos del Cauca y Nariño, un bajo rendimiento, se asocia con aquellos que asisten a colegios oficiales, mientras que los estudiantes de colegios privados, tienden a obtener mejores resultados. Según Giraldo y Fajardo, (2018) esto puede explicarse, por las diferencias significativas entre la educación pública y privada en Colombia, en términos de financiación, gestión y calidad. La educación privada, aunque implica un mayor costo, suele ofrecer mejor infraestructura, mayores recursos y mejores resultados académicos. Además, asistir a un colegio privado está estrechamente relacionado con estudiantes que tienen mayores recursos económicos. Estos hallazgos coinciden con estudios previos, como los de

Seibold, (2000), Montero et al., (2007) y Garbanzo, (2007) que destacan, una relación significativa entre el nivel socioeconómico de los estudiantes y su desempeño académico. Por otro lado, López y González, (2016) indican, que existe evidencia de la influencia de la clase social en los logros académicos, siendo los estudiantes con un estatus socioeconómico más alto, los que demuestran un mejor desempeño. Por otro lado, investigaciones como las de Serpa y Reyes, (2005), Buitrago, (2008), Ramírez et al., (2011), Andrade et al., (2015) y Mosquera, (2021) destacan, cómo la pobreza en Colombia, está estrechamente vinculada al bajo desempeño escolar. Esto repercute negativamente, en el acceso a una educación de calidad y limita las oportunidades de movilidad social. La falta de recursos esenciales para el aprendizaje, como materiales escolares, acceso a internet, alimentación adecuada y un entorno familiar seguro y estimulante, agrava esta situación.

Otro factor relevante, es la ubicación del colegio, ya sea, en una zona rural o urbana. Se observa que los estudiantes de los departamentos del Cauca y Nariño que asisten a colegios rurales, tienden a obtener un desempeño inferior, en comparación con aquellos que estudian en colegios urbanos. Las posibles causas de esta diferencia, según Beltrán, (2023) radican, en que el modelo educativo en Colombia, ha sido diseñado, desde una perspectiva urbana, dejando de lado las particularidades de la sociedad rural. Según Aguilar y Monge, (2000) la brecha educativa entre las escuelas rurales y urbanas, se debe, a problemas como la falta de infraestructura y recursos, la desvinculación de programas de desarrollo agrícola y la exclusión de la perspectiva rural, en los planes educativos. Además, Marcillo y Acero, (2016) destaca la ausencia de políticas, para mejorar la enseñanza en aulas multigrado. Estas aulas, comunes en las zonas rurales del país, reúnen a estudiantes de distintas edades y grados de escolaridad, en una sola clase, donde el proceso de enseñanza-aprendizaje, es asumido por un solo docente, lo que afecta sustancialmente la calidad educativa en las zonas rurales.

Además de las causas mencionadas, hay varios otros factores que pueden contribuir al desempeño inferior de los estudiantes en colegios rurales:

- La necesidad de trabajar para apoyar a sus familias. Esto limita su desarrollo personal y perpetúa la pobreza y la desigualdad (Callysaya & Nuni, 2024).
- La falta de atención de los padres, el desinterés de los estudiantes y el bajo presupuesto destinado a la educación en comunidades rurales, son factores que contribuyen al fracaso escolar (Aguilar *et al.*, 2019).
- La distancia a los centros educativos, las malas condiciones de los caminos y la escasez de materiales de aprendizaje, desmotivan a los estudiantes (Callysaya & Nuni, 2024).
- La rotación frecuente de maestros y la falta de capacitación adecuada, afectan la continuidad y calidad del proceso educativo en los colegios rurales (Callysaya & Nuni, 2024).

Otro de los factores comunes en el desempeño académico de los estudiantes de los departamentos del Cauca y Nariño, en las Pruebas Saber 9º, es la repetición de grados a lo largo de su trayectoria escolar. Se ha observado, que aquellos alumnos que han reprobado algún curso, suelen obtener resultados más bajos en estas evaluaciones, mientras que quienes han avanzado sin interrupciones académicas, tienden a lograr un mejor rendimiento. Diversos estudios, como los de Canales y Solís, (2009), Bayona y López, (2018), Pinzón, (2018) y Díaz y Ruiz, (2018) han demostrado, que repetir un año, no siempre contribuye a mejorar el aprendizaje; por el contrario, en muchos casos, puede generar efectos negativos a largo plazo. Estos hallazgos sugieren, que la continuidad en el proceso educativo y el aprendizaje sin interrupciones, favorecen un mejor desempeño académico. En este sentido, la repetición de grados escolares, puede representar una barrera significativa para el desarrollo académico de los estudiantes.

CONCLUSIONES Y TRABAJOS FUTUROS

Con el objetivo de descubrir patrones asociados al desempeño académico de los estudiantes de las Instituciones de Educación Secundaria de los departamentos de Cauca y Nariño, en las Pruebas Saber 9º, se utilizaron Técnicas de Minería de Datos Educativa. La fuente de datos, fue la información actualizada y disponible en el repositorio de datos abiertos del ICFES, denominado Data Icfes.

Para las Pruebas Saber 9º, la información más reciente, correspondía al año 2017, en el cual, se evaluaron las Competencias de Lenguaje y Matemáticas y el Desempeño General (el promedio de los puntajes de las dos competencias).

Del repositorio Data Icfes, se seleccionaron los datos socioeconómicos, académicos e institucionales de los estudiantes caucanos y nariñenses. La metodología utilizada para cumplir este objetivo, fue CRISP-DM

Se seleccionó la técnica de clasificación, mediante árboles de decisión, por ser especialmente, adecuada para interpretar fácilmente los patrones de desempeño académico.

Se evaluaron varios modelos, contruidos con diferentes algoritmos de clasificación por árboles de decisión y se seleccionó el modelo generado por algoritmo J48, por sus mejores resultados.

Para evaluar la exactitud de los diferentes modelos de clasificación, se particionaron los conjuntos de datos de cada departamento y el unificado de los dos, por cada competencia, en un conjunto de entrenamiento con el 80% del total de instancias y un conjunto de prueba con el 20% del total de instancias.

En cada competencia evaluada, se consideró como buen desempeño, aquellos estudiantes, cuyo puntaje en la respectiva competencia, estaba por encima de la media nacional y como bajo desempeño, aquellos cuyo puntaje estaba por debajo de la media nacional.

De acuerdo con el análisis exploratorio de datos, se observó que, en común entre los estudiantes de los dos departamentos que presentaron las Pruebas Saber 9º, que el porcentaje de estudiantes con puntajes por debajo de la media, es mayor que el de aquellos con puntajes por encima de la media.

El análisis de resultados de esta investigación, se centró particularmente, en conocer los factores asociados, al desempeño académico de los estudiantes de los departamentos del Cauca y Nariño, que están por debajo de la media nacional en las Competencias de Lenguaje, Matemáticas y Desempeño General de las Pruebas Saber 9º.

De acuerdo con el análisis de los patrones de desempeño académico en las Pruebas Saber 9º de los estudiantes del departamento del Cauca, se puede decir en general, que los estudiantes que asisten a colegios mixtos, tienden a tener un desempeño inferior en comparación con los colegios de un solo género.

Los estudiantes de colegios privados, tienen mejores resultados académicos que los de colegios oficiales. Los estudiantes de colegios rurales tienen un desempeño inferior, en comparación con los de colegios urbanos.

Los estudiantes de municipios de las subregiones del Norte y Centro del Cauca, tienen resultados más bajos, en comparación con el resto de los municipios del Cauca.

Estos resultados subrayan la importancia de considerar factores estructurales y socioeconómicos en la educación y pueden servir, como base, para mejorar la calidad educativa y la equidad en el acceso a oportunidades académicas en el Cauca.

Se espera que estos resultados impacten en políticas públicas y gestión de recurso como las siguientes:

1. Políticas Educativas:

- Los resultados, pueden guiar la formulación de políticas educativas más inclusivas y equitativas, enfocadas en mejorar la calidad de la educación en colegios mixtos, oficiales y rurales.
- Es necesario desarrollar programas específicos, para apoyar a los estudiantes en las zonas Norte y Centro del Cauca.

2. Inversión en Infraestructura:

- Incrementar la inversión en infraestructura y recursos educativos en colegios oficiales y rurales, puede ayudar a cerrar la brecha de desempeño.
- Mejorar las instalaciones y el acceso a tecnología en las zonas más desfavorecidas.

3. Programas de Apoyo Socioeconómico:

- Implementar programas de apoyo socioeconómico, para estudiantes de bajos recursos, puede mejorar su rendimiento académico.
- Facilitar el acceso a becas y ayudas financieras para estudiantes de colegios oficiales y rurales.

4. Investigación y Monitoreo:

- Continuar investigando y monitoreando el impacto de estos factores en el desempeño académico para ajustar las políticas y programas según sea necesario.
- Fomentar la colaboración entre instituciones educativas, gobiernos y organizaciones no gubernamentales para abordar estos desafíos.

De igual manera, teniendo en cuenta el análisis de los patrones de desempeño académico en las Pruebas Saber 9º, de los estudiantes del departamento de Nariño, se puede decir en general, que los estudiantes de las subregiones del Pacífico, como Pacífico Sur, Piedemonte Costero, Sanquianga y Telembí, tienen un rendimiento académico inferior en comparación con los de Pasto y las subregiones de Obando y La Sabana.

Los estudiantes de colegios privados, tienen mejores resultados académicos, que los de colegios oficiales.

Un bajo rendimiento está asociado con estudiantes, cuyas madres tienen un nivel educativo de primaria o bachillerato, mientras que un desempeño destacado, se observa en aquellos, cuyas madres poseen títulos de Educación Superior.

El nivel educativo de la madre, tiene un impacto positivo en el rendimiento escolar, proporcionando orientación y apoyo efectivos.

Los estudiantes que han reprobado algún curso, suelen obtener resultados más bajos en las Pruebas Saber 9º. La continuidad en el proceso educativo sin interrupciones favorece un mejor desempeño académico.

Estos resultados, subrayan la importancia de considerar factores estructurales y socioeconómicos en la educación y pueden servir como base, para mejorar la calidad educativa y la equidad en el acceso a oportunidades académicas en Nariño.

Se espera que estos resultados, impacten en políticas públicas y gestión de recurso como las siguientes:

1. Políticas Educativas:

- Los resultados, pueden guiar la formulación de políticas educativas más inclusivas y equitativas, enfocadas en mejorar la calidad de la educación, en las subregiones del Pacífico y en colegios oficiales.
- Es necesario desarrollar programas específicos, para apoyar a los estudiantes en las subregiones más desfavorecidas.

2. Inversión en Infraestructura:

- Incrementar la inversión en infraestructura y recursos educativos en las subregiones del Pacífico y en colegios oficiales, puede ayudar a cerrar la brecha de desempeño.
- Mejorar las instalaciones y el acceso a tecnología en las zonas más desfavorecidas.

3. Programas de Apoyo Socioeconómico:

- Implementar programas de apoyo socioeconómico para estudiantes de bajos recursos, puede mejorar su rendimiento académico.
- Facilitar el acceso a becas y ayudas financieras para estudiantes de colegios oficiales y en subregiones desfavorecidas.

4. Apoyo a la Educación de las Madres:

- Promover programas de educación para madres, puede tener un impacto positivo en el rendimiento académico de sus hijos.
- Ofrecer oportunidades de formación y desarrollo para madres con niveles educativos bajos.

5. Investigación y Monitoreo:

- Continuar investigando y monitoreando el impacto de estos factores en el desempeño académico, para ajustar las políticas y programas según sea necesario.
- Fomentar la colaboración entre instituciones educativas, gobiernos y organizaciones no gubernamentales, para abordar estos desafíos.

Finalmente, al buscar patrones de desempeño académico comunes en las Pruebas Saber 9º, de los estudiantes de los departamentos de Cauca y Nariño, el análisis de los resultados obtenidos, muestra que los estudiantes de colegios privados, tienen mejores resultados académicos que los de colegios oficiales.

Las diferencias en financiación, gestión y calidad entre la educación pública y privada, influyen significativamente en los resultados.

Los estudiantes de colegios rurales, tienden a obtener un desempeño inferior en comparación con los de colegios urbanos.

La falta de infraestructura, recursos y políticas adecuadas para las aulas multigrado en zonas rurales, afecta la calidad educativa.

Los estudiantes que han reprobado algún curso, suelen obtener resultados más bajos en las Pruebas Saber 9º.

La continuidad en el proceso educativo sin interrupciones favorece un mejor desempeño académico.

Estos resultados subrayan la importancia de considerar factores estructurales y socioeconómicos en la educación y pueden servir como base, para mejorar la calidad educativa y la equidad en el acceso a oportunidades académicas en los departamentos del Cauca y Nariño.

Se espera, que estos resultados impacten en políticas públicas y gestión de recurso, como las siguientes:

1. Políticas Educativas:

- Los resultados, pueden guiar la formulación de políticas educativas más inclusivas y equitativas, enfocadas en mejorar la calidad de la educación, en colegios oficiales y rurales.
- Es necesario desarrollar programas específicos para apoyar a los estudiantes en las zonas rurales y en colegios oficiales.

2. Inversión en Infraestructura:

- Incrementar la inversión en infraestructura y recursos educativos en colegios oficiales y rurales, puede ayudar a cerrar la brecha de desempeño.
- Mejorar las instalaciones y el acceso a tecnología en las zonas más desfavorecidas.

3. Programas de Apoyo Socioeconómico:

- Implementar programas de apoyo socioeconómico, para estudiantes de bajos recursos, puede mejorar su rendimiento académico.
- Facilitar el acceso a becas y ayudas financieras para estudiantes de colegios oficiales y rurales.

4. Apoyo a la Educación de las Madres:

- Promover programas de educación, para madres puede tener un impacto positivo en el rendimiento académico de sus hijos.
- Ofrecer oportunidades de formación y desarrollo para madres, con niveles educativos bajos.

5. Investigación y Monitoreo:

- Continuar investigando y monitoreando el impacto de estos factores, en el desempeño académico, para ajustar las políticas y programas según sea necesario.
- Fomentar la colaboración entre instituciones educativas, gobiernos y organizaciones no gubernamentales, para abordar estos desafíos.

Esta investigación, proporciona un diagnóstico valioso sobre el desempeño académico de los estudiantes de Noveno Grado, en los departamentos de Cauca y Nariño en las Pruebas Saber 9º. Además, establece una base sólida, para implementar mejoras estratégicas en el sistema educativo de estas regiones. Al identificar los factores que influyen en el desempeño académico en la Educación Básica Secundaria, las autoridades educativas, podrán diseñar intervenciones más efectivas, reducir las desigualdades educativas y garantizar, que los estudiantes de ambos departamentos tengan mayores oportunidades de éxito en su trayectoria educativa.

Para trabajos futuros, se sugiere, ampliar este estudio, utilizando diferentes técnicas predictivas y algoritmos, con el fin de comparar los resultados obtenidos, mediante árboles de decisión, con el algoritmo J48. Además, sería beneficioso, aplicar técnicas descriptivas de Minería de Datos Educativa, para identificar, qué atributos se asocian al desempeño académico en las Pruebas Saber 9º y cómo se agrupan los estudiantes según su rendimiento en estas pruebas.

También sería recomendable investigar, la relación entre los factores que influyen en el desempeño académico de los estudiantes en las Pruebas Saber 5º, Saber 9º y Saber 11º, con el objetivo de evaluar, el valor agregado de la Educación Secundaria, en la formación de los bachilleres colombianos, especialmente en los departamentos de Cauca y Nariño.

REFERENCIAS

- Agrawal, R., Imielinski, T. & Swami, A., (1993). Mining association rules between sets of items in large databases. En P. Buneman, S. Jajodia, (Eds.). *Proceedings of the ACM SIGMOD Conference* (pp. 207–216). ACM Press.
- Agudelo, Y., Aramendi, R., Cañón, R. & Galvis, C., (2020). *Guía para la interpretación y uso de los resultados históricos de las pruebas Saber*. Ministerio de Educación Nacional.
- Aguilar, M. & Monge, M. (2000). *Hacia una pedagogía rural para las escuelas rurales costarricenses*. Editorial Universidad de Costa Rica. <https://catalogosiidca.csuca.org/Record/CR.UNA01000217131/Details>
- Aguilar, K., Ek, G., Alamilla, P. & Rodríguez, J., (2019). Desigualdades estructurales en el vínculo entre escuela y comunidad rural: Tres casos de abandono escolar. *Perspectiva Educacional. Formación de Profesores*, 58(2), 98–120.
- Andrade, M., Barón, D., Cepeda, S. & Murillo, C., (2015). *Influencia de la pobreza en el rendimiento académico de los niños del municipio de Arauca*. [Trabajo de grado, Universidad Cooperativa de Colombia]. <https://repository.ucc.edu.co/bitstreams/56e8b07a-7010-4b4f-9e36-ff-288d648eb8/download>
- Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. En H. Weghorn & A. P. Abraham (Eds.), *Proceedings of the IADIS European Conference on Data Mining 2008* (pp. 182–185). IADIS.

- Barrientos, J., (2008). Calidad de la educación pública y logro académico en Medellín 2004–2006: Una aproximación por regresión intercuartil. *Lecturas de Economía*, (68), 121–144.
- Bayona, H. & López, A., (2018). *La reprobación y deserción en Colombia entre 1980 y 2015*. Secretaría de Educación Municipal. [http://www.semlorica.gov.co/images/noticias/2019/01/Anexo%205.%20Lectura%20para%20el%20tutor%20\(1\).pdf](http://www.semlorica.gov.co/images/noticias/2019/01/Anexo%205.%20Lectura%20para%20el%20tutor%20(1).pdf)
- Belcic, I. & Stryker, C., (2024). *¿Qué es el ajuste de hiperparámetros?*. Boletín The Think. <https://www.ibm.com/es-es/think/topics/hyperparameter-tuning>.
- Beltrán, L., (2023). Retos para la enseñanza de la lectura en el contexto rural de Colombia. *Ciencia Latina Revista Científica Multidisciplinar*, 7(5), 4893-4915. https://doi.org/10.37811/cl_rcm.v7i5.8084.
- Buitrago, C., (2008). *Educación y pobreza en Colombia: Un enfoque institucional*. Universidad de los Andes. <https://repositorio.uniandes.edu.co/bitstreams/Oc694a18-782c-47dd-a82b-8ce142f4d4cc/download>
- Callisaya, Y.M. & Nuni, C., (2024). *La crisis de la deserción escolar en las comunidades rurales*. <https://lapalabradelbeni.com/post/la-crisis-de-la-desercion-escolar-en-las-comunidades-rurales>.
- Canales, D. & Solís, R., (2009). Impacto de la Reprobación en Primaria y Secundaria sobre la Evaluación Pisa [conferencia]. *X Congreso Nacional de Investigación Educativa, área 10: Interrelaciones Educación-Sociedad*. Veracruz, México. https://www.comie.org.mx/congreso/memoriaelectronica/v10/pdf/area_tematica_10/ponencias/1113-F.pdf
- Cárcamo, C. & Mola, J., (2012). Diferencias por sexo en el desempeño académico en Colombia: Un análisis regional. *Economía & Región*, 6(1), 133–169.

- Cedillo, J., Beltrán, H., Saltos, M. & Soriano, F., (2024). Explorando la minería de datos en la gestión educativa superior: Desafíos y oportunidades en la era digital. *Revista Reincisol*, 3(5), 1367–1385.
- Correa, J. J., (2004). Determinantes del rendimiento educativo de los estudiantes de secundaria en Cali: Un análisis multinivel. *Sociedad y Economía*, (6), 81–105. https://sociedadyeconomia.univalle.edu.co/index.php/sociedad_y_economia/article/view/3866
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R., (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. CRISP-DM Consortium. <https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf>
- Chica, S., Galvis, D. & Ramírez, A., (2010). Determinantes del rendimiento académico en Colombia: Pruebas ICFES Saber 11°. *Revista Universidad EAFIT*, 46(160), 48–72.
- DANE, (2020a). *La información del DANE en la toma de decisiones de los departamentos: Cauca*. Sistema Estadístico Nacional. <https://www.dane.gov.co/files/investigaciones/planes-desarrollo-territorial/200313-Info-Dane-Cauca.pdf>.
- DANE, (2020b). *Pobreza multidimensional región Pacífica (sin incluir Valle del Cauca): Departamento de énfasis Nariño*. Boletín técnico. https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2019/Boletin_Region_bt_pobreza_multidimensional_19_pacifica.pdf.
- Díaz, D. & Ruiz, A., (2018). Reprobación escolar en el nivel medio superior y su relación con el autoconcepto en la adolescencia. *Revista Latinoamericana de Estudios Educativos*, 48(2), 125–142.
- Díaz-Martínez, M., Ahumada-Cervantes, M. & Melo-Morin, J., (2021). Árboles de decisión como metodología para determinar el rendimiento académico en educación superior. *Revista Lasallista de Investigación*, 18(2), 94–104.

- Espejel, M. & Jiménez, M. (2019). Nivel educativo y ocupación de los padres: Su influencia en el rendimiento académico de estudiantes universitarios. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, 10(19), 1-20.
- Frawley, W.J., Piatetsky-Shapiro, G. & Matheus, C.J., (1992). Knowledge discovery in databases: *An overview*. *AI Magazine*, 13(3), 57–70 <https://doi.org/10.1609/aimag.v13i3.1011>.
- Garbanzo, G. (2007). Factores asociados al rendimiento académico en estudiantes universitarios: Una reflexión desde la calidad de la educación superior pública. *Educación*, 31(1), 43–63. www.redalyc.org/pdf/440/44031103.pdf.
- García, J. (2016). *Comenzando con Weka: Filtrado y selección de subconjuntos de atributos basada en su relevancia descriptiva para la clase* (Informe técnico). Universidad de Málaga. <https://www.researchgate.net/publication/308141950>.
- García, J., & Skrita, A. (2019). Predicting academic performance based on students' family environment: Evidence for Colombia using classification trees. *Psychology, Society, & Education*, 11, 299–311.
- Gaviria, A., & Barrientos, J. (2001). Calidad de la educación y rendimiento académico en Bogotá. *Coyuntura Social*, (24), 211-227.
- Giraldo, C. & Fajardo, R. (2018). La educación pública frente a la educación privada: Una comparación de beneficios. *Pensamiento Udecino*, 2(1), 6-13.
- Gómez, A., (2023). *Construcción de la Narrativa sobre la Educación Diferenciada de una Institución Educativa*. [Tesis de Maestría, Universidad de la Sabana]. <https://intellectum.unisabana.edu.co/server/api/core/bitstreams/c73d349b-092d-4f63-b906-501e524c0fa0/content>.

- Gómez, J., (2014). *Análisis de las competencias en matemáticas y lenguaje de los bachilleres colombianos* [Trabajo de grado, Universidad ICESI]. <https://repository.icesi.edu.co/items/5f9b8cea-4048-7785-e053-2cc003c84dc5>.
- Gordillo, E. (2015). Historia de la educación mixta y su difusión en la educación formal occidental. *Historia de la Educación Latinoamericana*, 17(25), 107–124.
- Gutiérrez, Y. (2015). *Relación entre la estructura familiar y el rendimiento académico en el área de matemáticas, en estudiantes de tercero y quinto de primaria*. Instituto Latinoamericano de Altos Estudios. <https://libroselectronicos.ilae.edu.co/index.php/ilae/catalog/book/43>
- Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers. <https://archive.org/details/the-morgan-kaufmann-series-in-data-management-systems-jiawei-han-micheline-kambe>
- Hernández, O. (2015). *Determinantes del rendimiento académico en la educación media de Cundinamarca* [Trabajo de grado, Escuela Colombiana de Ingeniería Julio Garavito]. <http://repositorio.escuelaing.edu.co/bitstream/001/349/1/AA-Econom%C3%ADa-1077087614.pdf>.
- Hernández, E. & Lorente, R. (2009). *Minería de datos aplicada a la detección de Cáncer de Mama*. Universidad Carlos III de Madrid. https://www.researchgate.net/publication/265891193_Minera_de_datos_aplicada_a_la_deteccion_de_Cancer_de_Mama
- Hernández, J., Ramírez, M. & Ferri, C., (2005). *Introducción a la Minería de Datos*. Pearson Prentice Hall.

- ICFES. (2024a). *Data Icfes: Repositorio de Datos Abiertos del Icfes: Conjunto de datos Saber 3,5 y 9*. Instituto Colombiano para la Evaluación de la Educación -Icfes. <https://www.icfes.gov.co/investigaciones/data-icfes/>
- ICFES. (2024b). *Diccionario de variables Saber 3º, 5º y 9º*. Data Icfes: Repositorio de Datos Abiertos del Icfes: Documentación. Instituto Colombiano para la Evaluación de la Educación -Icfes. <https://www.icfes.gov.co/investigaciones/data-icfes/>
- ICFES. (2022a). *Marco de referencia y especificaciones trayectorias escolares. Pruebas Saber 3º, 5º, 7º y 9º-2022*. Instituto Colombiano para la Evaluación de la Educación (ICFES).
- ICFES. (2022b). *Informe nacional Saber 3º, 5º y 9º: Aplicación 2021*. Instituto Colombiano para la Evaluación de la Educación (ICFES).
- ICFES. (2019). *¿Como se Construye el Índice de Nivel Socioeconómico (INSE) en el Contexto de las Pruebas Saber?* (Boletín Saber al Detalle, Edición 04). Instituto Colombiano para la Evaluación de la Educación (ICFES).
- ICFES. (2017a). *Saber 9º: Guía de Orientación*. Instituto Colombiano para la Evaluación de la Educación (ICFES).
- ICFES. (2017b). *Experiencia acumulada en el levantamiento de datos de factores asociados 2012-2015, Saber 3º, 5º y 9º*. Instituto Colombiano para la Evaluación de la Educación (ICFES).
- ICFES. (2014a). *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).
- ICFES. (2014b). *Pruebas Saber 3º, 5º y 9º. Lineamientos para las aplicaciones muestral y censal 2014*. Instituto Colombiano para la Evaluación de la Educación (ICFES).

- ICFES. (2011). *Informe técnico Saber 5o. y 9o 2009*. Instituto Colombiano para la Evaluación de la Educación (ICFES).
- ICFES. (2009). *Lineamientos Generales Saber 2009 grados 5º y 9º*. Instituto Colombiano para la Evaluación de la Educación (ICFES).
- Larose, D. & Larose, C. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons. <https://www.doi.org/10.1002/9781118874059>.
- López, H. & González, M. (2016). *Factores asociados a la competencia de inglés como lengua extranjera de los estudiantes universitarios colombianos* [Tesis de maestría, Universidad de La Sabana]. <https://hdl.handle.net/10818/28345>.
- Mancilla-Vela, G. Leal-Gatica, P., Sánchez-Ortiz, A. & Vidal-Silva, C., (2020). Factores asociados al éxito de los estudiantes en modalidad de aprendizaje en línea: Un análisis en minería de datos. *Formación Universitaria*, 13(6), 23-36. <http://dx.doi.org/10.4067/S0718-50062020000600023>.
- Marcillo, F. & Acero, G., (2016). *Necesidades de formación de educadores en servicio de escuelas rurales multigrado a partir del análisis de la gestión académica y la planeación curricular en el marco de la política pública educativa* [Tesis de maestría, Universidad Nacional de Colombia]. Repositorio Institucional. <https://repositorio.unal.edu.co/handle/unal/58011>.
- Matute, E., Sanz, A., Gumá, E., Rosselli, M., & Ardila, A. (2009). Influencia del nivel educativo de los padres, el tipo de escuela y el sexo en el desarrollo de la atención y la memoria. *Revista Latinoamericana de Psicología*, 41(2), 257–276.
- Moine, J. M., Haedo, A. & Gordillo, S. (2011). *Estudio comparativo de metodologías para minería de datos*. Red de Universidades con Carreras en Informática. <http://sedici.unlp.edu.ar/handle/10915/20034>

- Moiné, J. & Haedo, A. (2015). *Una herramienta para la evaluación y comparación de metodologías de minería de datos*. Universidad Nacional del Noroeste de la Provincia de Buenos Aires. https://sedici.unlp.edu.ar/bitstream/handle/10915/50428/Documento_completo.pdf?sequence=1
- Montero, E., Villalobos, J. & Valverde, A. (2007). Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico en la Universidad de Costa Rica: Un análisis multinivel. *RELIEVE*, 13(2), 215–234. www.redalyc.org/pdf/916/91613205.pdf.
- Mor, E., Sanguesa, R. & Molina, L. (2005). *Data mining*. Universidad Oberta de Cataluña UOC.
- Moreira, T. (2007). Perfil sociodemográfico y académico de estudiantes en deserción del sistema educativ. *Actualidades en Psicología*, 21(108), 145-165
- Mosquera, D. (2021). Desigualdades socioeconómicas y rendimiento escolar. *Revista Universitaria de Informática RUNIN*, 9(12), 59–65.
- Mullis, R. L., Rathge, R., & Mullis, A. K. (2003). Predictors of academic performance during early adolescence: A contextual view. *International Journal of Behavioral Development*, 27(6), 541–548.
- Nieto, S., & Ramos, R. (2011). *La sobreeducación de los padres afecta al rendimiento académico de sus hijos?* Research Institute of Applied Economics, Universidad de Barcelona.
- OECD. (2013). *Pisa 2012 Results: What Makes Schools Successful? Resources, Policies and Practices* (Volume IV). OECD Publishing.
- OECD. (2016). *Education in Colombia. Organization for Economic Co-operation and Development*. OECD Publishing.

- Páez, R., (2014). Coeducación y educación diferenciada: Una opción por la igualdad de género en la escuela pública mixta. *Educación y Ciudad*, (27), 127–138.
- Pinzón, D. (2018). *Reprobación y desempeño académico: Evidencia de la implementación de la promoción automática en Colombia*. Universidad de los Andes. <http://hdl.handle.net/1992/7862>.
- Pinzón Hernández, D. (2018). *Reprobación y desempeño académico: evidencia de la implementación de la promoción automática en Colombia*. Universidad de los Andes. <http://hdl.handle.net/1992/7862>
- 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*. ICFES & Ministerio de Educación Nacional.
- Pyle, D. (2003). *Business modeling and data mining*. Morgan Kaufmann. <https://doi.org/10.1016/B978-1-55860-653-1.X5000-3>.
- Profamilia. (2015). *Resumen ejecutivo encuesta nacional de demografía y salud – ENDS Colombia 2015*. Profamilia & Ministerio de Salud y Protección Social. <https://profamilia.org.co/wp-content/uploads/2019/06/Resumen-Ejecutivo-Encuesta-Nacional-De-Demografia-Y-Salud-ends-2015.pdf>.
- Quinlan, J. R. (1993). *Programs for machine learning*. Morgan Kaufmann Editors.
- Quitiaquez, A. (2024). Revisión teórica de la imputación de datos con una aplicación en R. *Revista Sigma*, 20(2), 35–42. <https://revistas.ude-nar.edu.co/index.php/rsigma>.
- Ramírez, M., Quiñones, D., Erasmo, R., Salazar, L., Aníbal, R. (2011). Pobreza y rendimiento escolar: Estudio de caso de jóvenes de alto rendimiento. *Educere*, 15(52), 663–672.

- Sánchez-Jiménez, E., Hernández, Y., Ortiz-Hernández, J., Martínez-Rebollar, A. & Estrada-Esquivel, H. (2023). Configuración de hiperparámetros mediante algoritmos de optimización: Aplicación en la predicción de enfermedades cardiovasculares. *Research in Computing Science*, 152(8), 141–155.
- Santacruz, A. (2024). *Contexto Subregional Pacífico Sur*. Centro de Estudios de Desarrollo Regional CEDRE.
- Sattler, K. & Dunemann, O. (2001). SQL database primitives for decision tree classifiers. En H. Paques, L. Liu, & D. Grossman (Eds.), *Proceedings of the 10th ACM International Conference on Information and Knowledge Management* (pp. 379–386). ACM
- 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), 215–231.
- Serpa, M. & Reyes, N. (2005). *Influencia de la pobreza en el rendimiento académico de los niños y niñas de los grados tercero del Colegio Nueva Vida del barrio La Nevada de Valledupar, Cesar* [Tesis de grado, Universidad Nacional Abierta y a Distancia]. <https://repository.unad.edu.co/bitstream/handle/10596/20470/mcserpac.pdf>
- Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *The Journal of Data Warehousing*, 5(4), 13-22.
- Spicker, P., Leguizamón, S. & Gordon, D. (2009). *Pobreza: Un glosario internacional* (Primera edición). Consejo Latinoamericano de Ciencias Sociales. <https://www.crop.org/publications/clacso/pobreza-un-glosario-internacional.aspx>
- Timarán, S. R., Caicedo, J. & Hidalgo, A. (2021a). *Minería de datos educativa para el descubrimiento de factores asociados al desempeño académico en las pruebas Saber 11°*. Editorial Universidad de Nariño.

- Timarán, S. R., Caicedo, J. & Hidalgo, A. (2021b). *Aplicación de la minería de datos en la detección de patrones de desempeño académico en las pruebas Saber Pro*. Editorial Universidad de Nariño.
- Timarán, R., Jiménez, J. & Calderón, A. (2017). *Detección de patrones de deserción estudiantil con minería de datos*. Editorial Universitaria, Universidad de Nariño.
- Torres, J., Pachajoa, L. y Pantoja, R. (2014). Resultados de las Pruebas Saber en el grado quinto del área de las ciencias naturales en tres instituciones educativas oficiales del municipio de Pasto. *Revista Fedumar Pedagogía y Educación*, 1(1), 55-69.
- Torres, C. (2018). *Relación entre el currículo de educación básica secundaria y media de la Institución Educativa Divino Niño y los componentes que evalúa la prueba Saber 11* [Tesis de maestría, Corporación Universitaria Minuto de Dios – UNIMINUTO].
- Valero, S. (2009). *Aplicación de técnicas de minería de datos para predecir deserción*. Universidad Tecnológica de Izúcar de Matamoros. <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*. Universidad Tecnológica de Izúcar de Matamoros. www.utim.edu.mx/~svalero/docs/e1.pdf.
- Velandia, L. (2012). *Impacto del programa de educación diferenciada en el Colegio “El Minuto de Dios”* [Tesis de maestría, Universidad Libre de Colombia & Universidad de Las Tunas].
- Villena J. (2016). CRISP-DM: La metodología para poner orden en los proyectos de Data Science. <https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science>.

Wikipedia. (s.f.a). *Subregiones de Cauca (Colombia)*. Wikipedia la Enciclopedia libre. [https://es.wikipedia.org/wiki/Subregiones_de_Cauca_\(Colombia\)](https://es.wikipedia.org/wiki/Subregiones_de_Cauca_(Colombia)).

Wikipedia. (s.f.b). *Subregiones de Nariño*. Wikipedia la Enciclopedia libre. https://es.wikipedia.org/wiki/Subregiones_de_Nariño.

Witten, I., Frank, E. & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (Third Edition). Morgan Kaufmann.

ÍNDICE DE TABLAS

Tabla 1. Nivel de Desempeño	21
Tabla 2. Características Socioeconómicas de los Estudiantes Caucanos que Presentaron la Prueba Saber 9º	29
Tabla 3. Características Institucionales de los Colegios de los Estudiantes Caucanos que Presentaron las Pruebas Saber 9º	32
Tabla 4. Número de Estudiantes de los Municipios del Departamento del Cauca que Presentaron las Pruebas Saber 9º	34
Tabla 5. Niveles de Desempeño Académico de los Estudiantes Caucanos en la Prueba de Lenguaje	37
Tabla 6. Equivalencia entre Nivel de Desempeño y Puntaje Obtenido por los Estudiantes Caucanos en la Prueba de Lenguaje	39
Tabla 7. Niveles de Desempeño Académico de los Estudiantes Caucanos en la Prueba de Matemáticas	40
Tabla 8. Equivalencia entre Nivel de Desempeño y Puntaje Obtenido en la Prueba de Matemáticas	42
Tabla 9. Características Socioeconómicas de los Estudiantes Nariñenses que Presentaron las Pruebas Saber 9º	45
Tabla 10. Características Institucionales de los Colegios de los Estudiantes Nariñenses que Presentaron las Pruebas Saber 9º	48
Tabla 11. Número de Estudiantes de los Municipios del Departamento de Nariño que Presentaron las Pruebas Saber 9º	50
Tabla 12. Niveles De Desempeño Académico de los Estudiantes Nariñenses en la Prueba de Lenguaje	54
Tabla 13. Equivalencia Entre Nivel de Desempeño y Puntaje Obtenido en la Prueba de Lenguaje	56
Tabla 14. Niveles de Desempeño Académico de los Estudiantes Nariñenses en la Prueba De Matemáticas	57

Tabla 15. Equivalencia Entre Nivel de Desempeño y Puntaje Obtenido en la Prueba de Matemáticas	59
Tabla 16. Desempeño General en las Pruebas Saber 9º de los Estudiantes del Cauca y Nariño	62
Tabla 17. Diccionario de Datos del Conjunto Sbr9_Colombia	65
Tabla 18. Calidad de Datos del Conjunto Sbr9_Cauca con 16,392 Registros y 52 Atributos Tipo Texto	79
Tabla 19. Calidad de Datos del Conjunto Sbr9_Cauca con 16,392 Registros y 4 Atributos Tipo Numérico	82
Tabla 20. Calidad de Datos del Conjunto Sbr9_Narino con 18,971 Registros y 52 Atributos Tipo Texto	82
Tabla 21. Calidad de Datos del Conjunto Sbr9_Narino con 18,971 Registros y 4 Atributos Tipo Numérico	85
Tabla 22. Lista de Atributos Inicialmente Eliminados de los Conjuntos de Datos	87
Tabla 23. Subregiones del Departamento del Cauca	89
Tabla 24. Subregiones del Departamento de Nariño	89
Tabla 25. Cálculo del Índice Fami_Lectrodomesticos	90
Tabla 26. Cálculo del Índice Fami_Tics	91
Tabla 27. Desempeño de los Estudiantes del Departamento del Cauca ...	92
Tabla 28. Desempeño de los Estudiantes del Departamento de Nariño	93
Tabla 29. Diccionario de Datos de los Conjuntos de Datos Sbr9_Cauca y Sbr9_Narino	94
Tabla 30. Evaluación de Algoritmos de Árboles de Decisión	106
Tabla 31. Evaluación de la Calidad del Modelo de Árboles de Decisión Mediante Particiones	107
Tabla 32. Ajuste de Hiperparámetros para el Conjunto de Datos Sbr9_Cauca_Final	109

Tabla 33. Ajuste de Hiperparámetros para el Conjunto de Datos Sbr9_ Narino_Final 109

Tabla 34. Distribución del Desempeño Observado de los Estudiantes Caucanos en las Competencias de Lenguaje, Matemáticas y General 111

Tabla 35. Distribución del Desempeño Observado de los Estudiantes Nariñenses en las Competencias de Lenguaje, Matemáticas y General 143

Tabla 36. Distribución del Desempeño Observado de los Estudiantes del Cauca y Nariño en las Competencias de Lenguaje, Matemáticas y General 184

ÍNDICE DE FIGURAS

Figura 1. Caracterización de los Estudiantes Caucanos que Presentaron las Pruebas Saber 9º	31
Figura 2. Características de los Colegios del Departamento del Cauca con Número de Estudiantes que Presentaron las Pruebas Saber 9º	33
Figura 3. Municipios con Mayor Número de Estudiantes que Presentaron las Pruebas Saber 9º en Cada Subregión del Departamento del Cauca	36
Figura 4. Número de Estudiantes Caucanos que Presentaron la Prueba de Lenguaje	38
Figura 5. Histograma de Puntajes de los Estudiantes Caucanos en la Competencia de Lenguaje	39
Figura 6. Número de Estudiantes Caucanos que Presentaron la Prueba de Matemáticas	41
Figura 7. Histograma de Puntajes de los Estudiantes Caucanos en la Competencia de Matemáticas	43
Figura 8. Desempeño de los Estudiantes Caucanos en las Pruebas de Lenguaje y Matemáticas	44
Figura 9. Caracterización de los Estudiantes Nariñenses que Presentaron las Pruebas Saber 9º	46
Figura 10. Características de los Colegios del Departamento de Nariño con Número de Estudiantes que Presentaron las Pruebas Saber 9º	50
Figura 11. Municipios con Mayor Número de Estudiantes que Presentaron las Pruebas Saber 9º en cada Subregión del Departamento de Nariño	53
Figura 12. Desempeño de los Estudiantes Nariñenses en la Prueba de Lenguaje	55
Figura 13. Histograma de Puntajes en la Competencia de Lenguaje	56

Figura 14. Número de Estudiantes Nariñenses que Presentaron la Prueba de Matemáticas Según Nivel y Clase de Desempeño	58
Figura 15. Histograma de Puntajes en la Competencia de Matemáticas ..	59
Figura 16. Desempeño de los Estudiantes Nariñenses en las Pruebas de Lenguaje y Matemáticas	61
Figura 17. Desempeño General de los Estudiantes Caucanos y Nariñenses en las Pruebas Saber 9º	63
Figura 18. Fases de la Metodología CRISP-DM	76
Figura 19. Árbol de Decisión Podado para la Competencia de Lenguaje de los Estudiantes Caucanos en las Pruebas Saber 9º	113
Figura 20. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Lenguaje De los Estudiantes Caucanos en las Pruebas Saber 9º	118
Figura 21. Árbol de Decisión Podado para la Competencia de Matemáticas de los Estudiantes Caucanos en las Pruebas Saber 9º	123
Figura 22. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Matemáticas de los Estudiantes Caucanos en las Pruebas Saber 9º	128
Figura 23. Árbol de Decisión Podado del Desempeño General de los Estudiantes Caucanos en las Pruebas Saber 9º	133
Figura 24. Métricas de Evaluación del Modelo de Árbol de Decisión para el Desempeño General de los Estudiantes Caucanos en las Pruebas Saber 9º	137
Figura 25. Árbol de Decisión Podado para la Competencia de Lenguaje de los Estudiantes Nariñenses en las Pruebas Saber 9º	145
Figura 26. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Lenguaje de los Estudiantes Nariñenses en las Pruebas Saber 9º	150
Figura 27. Árbol de Decisión Podado para la Competencia de Matemáticas de los Estudiantes Nariñenses en las Pruebas Saber 9º	154

Figura 28. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Matemáticas de los Estudiantes Nariñenses en las Pruebas Saber 9º	162
Figura 29. Árbol de Decisión Podado para el Desempeño General en las Pruebas Saber 9º de los Estudiantes Nariñenses en las Pruebas Saber 9º	167
Figura 30. Métricas de Evaluación del Modelo de Árbol de Decisión para el Desempeño General de los Estudiantes Nariñenses en las Pruebas Saber 9º	176
Figura 31. Árbol de Decisión Podado para el Desempeño en la Competencia de Lenguaje de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º	186
Figura 32. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Lenguaje de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º	190
Figura 33. Árbol de Decisión Podado para el Desempeño en la Competencia de Matemáticas de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º	194
Figura 34. Métricas de Evaluación del Modelo de Árbol de Decisión para la Competencia de Matemáticas de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º	199
Figura 35. Árbol de Decisión Podado para el Desempeño General de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º	204
Figura 36. Métricas de Evaluación del Modelo de Árbol de Decisión para el Desempeño General de los Estudiantes de Cauca y Nariño en las Pruebas Saber 9º	207

AUTORES

Silvio Ricardo Timarán Pereira

Doctor en Ingeniería con énfasis en Ciencias de la Computación, Magíster en Ingeniería, Especialista en Multimedia Educativa e Ingeniero de Sistemas y Computación por formación. Su sólida trayectoria académica lo ha llevado a ser Profesor Titular del Departamento de Sistemas en la Universidad de Nariño, donde además dirige el Grupo de Investigación Aplicada en Sistemas (GRIAS).

Como investigador activo y autor de diversas publicaciones científicas y ponencias internacionales y nacionales, ha centrado su trabajo en la intersección de la Ciencia de Datos y el Machine Learning. En esta obra, el autor aplica su experticia en minería de datos para analizar un factor crítico en el desarrollo regional: el desempeño académico en las pruebas Saber 9°. A través de un análisis profundo de los departamentos de Cauca y Nariño, el Dr. Timarán ofrece una herramienta analítica fundamental para entender los patrones educativos y fortalecer la toma de decisiones basada en evidencia tecnológica en el territorio.

Fredy Alonso Vidal Alegría

Magíster en Educación, Especialista en Administración de la Información y Bases de Datos, e Ingeniero de Sistemas. Su destacada trayectoria en la gestión académica y técnica lo ha llevado a desempeñarse actualmente como Decano de la Facultad de Ingeniería de la Institución Universitaria Colegio Mayor del Cauca.

Con una sólida experiencia en el diseño y gestión de Sistemas de Información, sus áreas de investigación se centran en las Bases de Datos y la Inteligencia de Negocios, temas sobre los cuales ha sido ponente en diversos escenarios académicos nacionales e internacionales. En esta obra, el profesor Vidal aporta su visión estratégica y técnica para desentrañar los factores que inciden en la calidad educativa del suroccidente colombiano, consolidando un análisis riguroso sobre el desempeño académico en el departamento del Cauca.

èditorial

Universidad de **Nariño**

Año de publicación: 2026
San Juan de Pasto - Nariño - Colombia

El Ministerio de Educación Nacional (MEN) y el ICFES han consolidado el sistema de pruebas Saber 9° como una herramienta vital para evaluar y fortalecer las competencias básicas en Colombia. Estas evaluaciones no solo miden el rendimiento académico, sino que ofrecen una ventana hacia los factores socioeconómicos e institucionales que inciden en el aprendizaje.

Este libro de investigación presenta un análisis profundo y riguroso sobre el desempeño académico en las Competencias de Lenguaje, Matemáticas y Desempeño General en las pruebas Saber 9° que presentaron los estudiantes de grado noveno en los departamentos de Cauca y Nariño. A diferencia de los informes tradicionales, esta obra utiliza técnicas de Minería de Datos Educativa (EDM) bajo el estándar de la metodología CRISP-DM, la guía de referencia más sólida en proyectos de ciencia de datos a nivel mundial.

A través de un proceso exhaustivo que abarca desde la exploración de datos del ICFES hasta la limpieza y construcción de modelos predictivos, los autores logran transformar datos brutos en conocimiento accionable. El libro detalla el descubrimiento de patrones asociados a las realidades específicas de los estudiantes del suroccidente colombiano, interpretando los hallazgos desde una perspectiva técnica y pedagógica.

Más que un reporte estadístico, esta obra busca socializar el conocimiento descubierto para convertirlo en un insumo clave para la toma de decisiones. Es una fuente de consulta indispensable para: Organismos gubernamentales y Secretarías de Educación; Directivos docentes e instituciones educativas e Investigadores en Ciencia de Datos y Pedagogía.

Este libro es una apuesta por la calidad educativa basada en la evidencia, proporcionando las bases para planes de mejoramiento que respondan a la realidad territorial de Cauca y Nariño.

ISBN: 978-628-7864-99-3



9 786287 864993



Universidad de Nariño
FUNDADA EN 1994



editorial
Universidad de Nariño