CARACTERISTICAS DE LOS ESTUDIANTES DEL DEPARTAMENTO DE NARIÑO SEGÚN EL NIVEL ALCANZADO EN MATEMÁTICAS EN LAS PRUEBAS SABER 11 EN EL PERIODO 2021 – B.

JORGE ANDRÉS SUÁREZ MUÑOZ

UNIVERSIDAD DE NARIÑO

FACULTAD DE CIENCIAS EXACTAS Y NATURALES

MAESTRÍA EN ESTADÍSTICA APLICADA

SAN JUAN DE PASTO

2023

CARACTERISTICAS DE LOS ESTUDIANTES DEL DEPARTAMENTO DE NARIÑO SEGÚN EL NIVEL ALCANZADO EN MATEMÁTICAS EN LAS PRUEBAS SABER 11 EN EL PERIODO 2021 – B.

Presentado por:

JORGE ANDRÉS SUÁREZ MUÑOZ

Trabajo presentado para optar al título de Magister en Estadística Aplicada

Asesor:

Mg. HERNÁN ABDON GARCÍA

UNIVERSIDAD DE NARIÑO

FACULTAD DE CIENCIAS EXACTAS Y NATURALES

MAESTRÍA EN ESTADÍSTICA APLICADA

SAN JUAN DE PASTO

2023

Las ideas y conclusiones aportadas en el siguiente trabajo son responsabilidad exclusiva del autor. **Artículo 1ro del Acuerdo No. 324 de octubre 11 de 1966** emanado del Honorable Consejo Superior de la Universidad de Nariño.

Nota de Aceptación

_____

_____

_____

_____

_____

_____

_____

_____

Presidente del Jurado

_____

Jurado

_____

Jurado

_____

Jurado

San Juan de Pasto, octubre de 2023.

AGRADECIMIENTOS

El autor del presente trabajo de grado expresa su más sincero agradecimiento a:

A Dios por permitir la posibilidad de cumplir este sueño, el cual nació hace mucho tiempo y solo ahora tuvo la fortuna de verse realizado. Gracias por la vida, la salud y las bendiciones recibidas, porque siempre extendio sus manos y auxilio, y sobre todo por el amor incondicional.

A mis padres y hermanas por el apoyo constante, comprensión y paciencia.

Al profesor Hernán Abdon Garcia, por su valiosa colaboración y dirección en el asesoramiento de este trabajo.

A la profesora Rocio Rosero por su paciencia, escucha y por brindar una voz amiga.

A Alfredo José Sacanambuy por sus consejos, paciencia y amistad ofrecida.

A todas las personas que de una u otra forma intervinieron en este proceso académico. Gracias y que Dios les bendiga.

DEDICATORIA

Dedico este trabajo inicialmente a Dios, a mi familia, y a todas aquellas personas que necesitan una voz de aliento y esperanza en sus vidas académicas. A veces parece que no hay mañana y que las ideas no fluyen, sin embargo, cuando dejamos todo en manos de Dios, las nubes se dispersan y de a poco el sol brilla de nuevo.

Quizá los resultados no sean inmediatos, pero es la paciencia y la constancia la que permite ver el fruto del esfuerzo.

JORGE ANDRES SUAREZ MUÑOZ.

RESUMEN ANALÍTICO DEL ESTUDIO

R.A.E.

Programa académico: Maestría en Estadística Aplicada.

Autor: Jorge Andrés Suárez Muñoz.

Asesor: Magister Hernán Abdón García.

Título: Características de los estudiantes del departamento de Nariño según el nivel alcanzado en matemáticas en las pruebas Saber 11 en el periodo 2021 – b.

Modalidad de trabajo: Investigación aplicada evaluativa.

Resumen

En esta investigación se determinan las relaciones existentes entre los puntajes de la prueba de matemáticas con los de lectura crítica, ciencias naturales, inglés, sociales y ciudadanas, de la prueba Saber 11 que presentaron los estudiantes de Nariño en el periodo 2021 – B, y se caracterizan los puntajes altos en matemáticas a partir de aspectos socioeconómicos, demográficos, familiares, institucionales y de rendimiento en la prueba. Este estudio fue de tipo descriptivo y correlacional, de enfoque cuantitativo y de diseño no experimental. Para la comprensión y procesamiento de la información se usó metodología CRISP-DM y minería de datos. Se utilizó la base de datos libre expuesta por el ICFES. El análisis de la información se realizó con WEKA 3.9.6 y RStudio. Los resultados descubiertos se sintonizan con los trabajos sobre capital cultural en donde a mejores condiciones socioeconómicas de los educandos mejores desempeños se obtienen en la prueba, mostrando también un número bajo de estudiantes nariñenses con desempeño avanzado en matemáticas.

Palabras clave: *análisis clúster, análisis factorial, pruebas Saber 11 y desempeño en matemáticas, metodología CRISP-DM, minería de datos, RStudio y WEKA.*

Abstract

In this research, the existing relationships between the scores of the mathematics test and those of critical reading, natural sciences, english, social and civic sciences, of the Saber 11 test taken by the students of Nariño in the period 2021 – B, are determined, and high scores in mathematics are characterized based on socioeconomic, demographic, family, institutional and test performance aspects. This research was descriptive and correlational, with a quantitative approach and a non-experimental design. Fort he understanding and transformation of the data, the CRISP-DM methodology and data mining were used. The free database exposed by the ICFES was used. The analisis of the information was carried out with WEKA 3.9.6 and RStudio. The results discovered are in tune with the works of cultural capital where the better socioeconomic conditions of the students, the better performance is obtained in the test, and also showed a low number of students from Nariño with advanced performance in mathematics.

Keywords: Cluster analysis, Saber 11 test and mathematics performance, CRISP-DM methodology, data mining, RStudio and WEKA.

CONTENIDO

Página.

LISTA DE TABLAS

Página.

LISTA DE FIGURAS

Página.

GLOSARIO

- **CNA:** Consejo Nacional de Acreditación.
- **COVID:** Enfermedad de coronavirus.
- **CRISP – DM:** CRoss Industry Standard Process for Data Mining.
- **DANE:** Departamento Administrativo Nacional de Estadística.
- **DBA:** Derechos Básicos de Aprendizaje.
- **DM:** Data Mining (Minería de datos).
- **EMD:** Educational Data Mining (Minería de datos educativa).
- **ICFES:** Anteriormente se reconocía como: Instituto Colombiano para el Fomento y Evaluación de la Educación Superior, actualmente se lo visualiza como Instituto Colombiano para la Evaluación de la Calidad de la Educación.
- **INSE:** Índice de nivel socioeconómico.
- **ISCE:** Índice Sintético de Calidad Educativa.
- **KDD:** Knowledge Discovery in Databases (Descubrimiento de conocimiento en bases de datos).
- **K – means:** Algoritmo de cluster de minería de datos, indica los centroides.
- **K - NN:** Algoritmo de minería de datos indica el vecino más cercano.
- **LLECE:** Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación.
- **MAR:** Missing at Random.
- **MCAR:** Missing Completely at Random.
- **MEN:** Ministerio de Educación Nacional.
- **NSE:** Nivel socioeconómico.
- **OCDE:** Organización para la Cooperación y Desarrollo Económicos.
- **PISA:** Programa para la Evaluación Internacional de Estudiantes.
- **TE:** Tamaño del Efecto.
- **TMD:** Técnicas de Minería de Datos.
- **TIMSS:** Estudio Internacional de Tendencias en Matemáticas y Ciencias.
- **UDENAR:** Universidad de Nariño.
- **WEKA:** Waikato Environment for Knowledge Analysis.

INTRODUCCIÓN

Desde hace más de una década el Gobierno colombiano viene desarrollando esfuerzos en procura de mejorar la calidad y cobertura educativa en el país, razón por la cual ha motivado a estudiantes, docentes, y demás personas afines con la educación a participar con el desarrollo de investigaciones, planes y propuestas encaminadas a brindar soluciones en torno a estos aspectos. En este sentido y con intención de evaluar el estado real de los procesos educativos efectuados en las aulas, el Gobierno nacional, promovió la participación de un grupo de estudiantes en pruebas internacionales como PISA y TIMSS, en donde, por desventura, se obtuvieron puntuaciones por debajo de la media establecida para cada prueba, hecho que conllevó a repensar la forma en la que se aborda la educación en Colombia y cómo ella se halla en sintonía con el cumplimiento de las exigencias académicas internacionales. Fue así como el Gobierno realizó una reestructuración de las pruebas de estado llevándolas desde una postura cognitiva a una de mayor reflexión, en donde se inspecciona el grado de dominio de las competencias que cada educando posee por área de estudio, tal como lo muestran Mullis et al. (2008), Fernandes Cristóvão (2010), ICFES (2020a), y, Sanabria James et al. (2020). Esta renovación implicó también que de las ocho áreas evaluadas en las pruebas Saber 11 para años previos al 2014, se eliminaran tres, dejando como referentes solamente a *matemáticas, lectura crítica, inglés, sociales y ciudadanas, y ciencias naturales* (MEN, s.f). Cabe señalar que el área de matemáticas ha jugado un rol importante en esta etapa, puesto que constituye uno de los ítems evaluados en donde el promedio de los educandos colombianos ha estado por debajo del valor medio esperado en cada prueba internacional en la que se ha intervenido.

Meditando en lo anterior, los investigadores en educación en Colombia han promovido reflexiones sobre los resultados obtenidos tras considerar cuestionamientos como, ¿qué hace que los resultados obtenidos en las pruebas internacionales para los estudiantes colombianos

no sean satisfactorios?, ¿cómo se abordan los procesos educativos en Colombia?, ¿qué reflejan las evaluaciones realizadas por el ICFES a través de sus pruebas Saber?, y para el caso particular de matemáticas, ¿qué competencias matemáticas se desarrollan en las aulas colombianas?, o, ¿cómo se suceden los procesos de alfabetización matemática en el país?

En aras de comenzar a brindar soluciones a estas inquietudes el MEN divulgó los estándares de competencias para las áreas de: matemáticas, lenguajes, ciencias, sociales y ciudadanas, en donde se especifica los Saberes que todo estudiante requiere conocer en su proceso de formación integral, y los que posteriormente fueron complementados con los DBA (MEN, 2006), (Ruta Maestra, 2017). Pese a ello, las dificultades exhibidas en esta introducción aún persisten y motivo por el cual el desarrollo de producciones que ahonden en estos tópicos fungen como importante, a razón de ello, investigadores como Bogoya (2006), Chica Gómez et al. (2010), Timarán Pereira et al. (2019), Collazos Valenzuela et al. (2021), Rodríguez Rosero et al. (2021), Peña Lozano & González Veloza (2022), entre otros; han esgrimido esfuerzos sobre evaluación en Colombia y la detección de factores que influyen en el rendimiento académico de los estudiantes en las pruebas Saber. Con base en las reflexiones expuestas en estos trabajos es de donde surge la presente investigación, ya que busca contribuir a los tópicos anteriormente propuestos sobre desempeño estudiantil en las pruebas y rendimiento en matemáticas y, esboza ideas sobre calidad educativa tomando como caso particular los resultados logrados por los estudiantes del departamento de Nariño en la prueba Saber 11 periodo 2021 – B.

El presente trabajo se halla dividido en cuatro capítulos, el primero muestra los aspectos generales de la investigación correspondientes al contexto, antecedentes, formulación del problema, objetivos, marco legal y marco teórico. El segundo capítulo aborda los aspectos metodológicos en donde se desarrollan las primeras cuatro fases del método CRISP – DM. El

tercer capítulo contiene el análisis y tratamiento de la información (quinta fase de CRISP – DM), y el último capítulo presenta los hallazgos a los que se llegó en esta investigación (última fase del método CRISP – DM), exhibiendo también ideas sobre futuras investigaciones respecto a los ámbitos aquí examinados. Al final del documento se describe la bibliografia recopilada en este trabajo.

El objetivo general acogido para este trabajo implicó determinar las relaciones entre los puntajes de la prueba de matemáticas con los de lectura crítica, ciencias naturales, inglés, sociales y ciudadanas, de la prueba Saber 11 que presentaron los estudiantes de Nariño en el periodo 2021 – B, y caracterizar los puntajes altos en matemáticas a partir de aspectos socioeconómicos, demográficos, familiares, institucionales y de rendimiento en las pruebas. Para su consecución se realizaron cinco objetivos específicos en donde se: (1) determinan las características de los estudiantes en lo referente a variables de tipo socioeconómicas, demográficas, familiares, institucionales y de rendimiento en las pruebas, (2) analiza la relación entre el nivel alcanzado en matemáticas y las variables de tipo socioeconómicas, demográficas, familiares, institucionales y de rendimiento en las pruebas, (3) determina la correlación existente entre los puntajes de las pruebas de lectura crítica, matemáticas, ciencias naturales, inglés, sociales y ciudadanas, (4) identifica las estructuras fundamentales o factores asociados a los puntajes dados en las pruebas de lectura crítica, matemáticas, ciencias naturales, inglés, sociales y ciudadanas, (5) analiza y compara modelos de agrupación que identifiquen grupos de estudiantes con características similares en torno al puntaje en matemáticas. Se utilizaron técnicas de minería de datos educativa para el tratamiento de la información como clustering para variables categóricas, análisis de factores y de correlación, y se realizó también análisis de tablas de frecuencia para elaborar cruces de variables cualitativas. Finalmente, se extiende invitación cordial al lector a detallar y reflexionar entorno a las ideas aquí expuestas.

1. ASPECTOS GENERALES DE LA INVESTIGACIÓN.

1.1 RESUMEN DEL CAPÍTULO.

En las siguientes líneas el lector encontrará datos asociados al Departamento de Nariño los cuales describen el contexto que recubre a los estudiantes en observación. Agregado a ello, se expone el planteamiento del problema bajo el análisis de sus antecedentes, la descripción del mismo y la formulación de la pregunta de investigación. En igual orden de ideas se presenta la justificación, los objetivos (general y específicos), el marco legal, y, el marco teórico el cual se dividió en los tópicos de rendimiento académico, consideraciones generales sobre la prueba Saber 11, competencias en matemáticas, minería de datos educativa y algoritmo clúster enfatizado en k – modas.

1.2 CONTEXTO.

La presente investigación fue realizada en el Departamento de Nariño, el cual, tal como señala Martínez (2019), se ubica al suroccidente del territorio colombiano, su capital es San Juan de Pasto, su extensión territorial es de 33.268 km^2, su población es aproximadamente de 1.765.906 habitantes de los cuales 879.565 son mujeres y 886.341 son hombres, hecho que representa una densidad poblacional de 53,08 habitante/km^2. El departamento de Nariño limita al norte con el departamento del Cauca, al oriente con el departamento del Putumayo, al sur con la República del Ecuador, y al occidente con el Océano Pacífico. Nariño cuenta con 64 municipios y 288 corregimientos, se caracteriza por tener una economía basada en la agricultura, ganadería, caza, silvicultura, y pesca como puntos más fuertes. En la parte contextual conviene decir también que la Universidad de Nariño, lugar donde se oferta la Maestría en Estadística Aplicada y la cual favoreció el desarrollo de esta investigación, se encuentra ubicada en mencionado departamento, siendo ella la institución de educación superior con mayor prestigio académico e investigativo de la región.

1.3 PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN.

1.3.1 *Antecedentes*.

En la inspección de referentes bibliográficos que aborden el desempeño de los estudiantes en las pruebas Saber 11 y el tratamiento de información mediante minería de datos, fueron relevantes los aportes realizados por autores como, Rodríguez Rosero et al. (2021), quien da cuenta de los factores que intervienen en el rendimiento académico de los discentes de educación media en el departamento de Nariño, situando como referentes las áreas de matemáticas, lectura crítica, inglés, ciencias naturales, ciudadanas y sociedades. Para lograr este objetivo acogen en su estudio la base de datos del ICFES del año 2018. Como conclusiones relevantes mencionados autores manifiestan que, para obtener un mejor desempeño en las pruebas Saber conviene que los estudiantes tengan facilidad de acceso a computadores y conexión a internet, de igual forma señalan al nivel educativo de los padres como un factor de gran influencia, al igual que pertenecer a un plantel educativo de carácter oficial y urbano. En este artículo se menciona también que la condición de ser varón influye en la obtención de mejores desempeños en matemáticas.

En el mismo orden de ideas se encontró el artículo de Peña Lozano & González Veloza (2022), quienes generan un modelo de predicción para los resultados obtenidos por los estudiantes en la prueba Saber 11 a partir de condiciones socioeconómicas. Para su estudio acogen la base de datos expuesta en el ICFES referida al año 2020, y con ella analizan la relación entre el puntaje de la prueba en matemáticas y las variables género, estrato, número de personas que conforman el hogar, educación de la madre, acceso a internet, disponibilidad de computadora en el hogar, horas de trabajo semanal del estudiante, naturaleza de la institución educativa y área de ubicación geográfica del colegio. Dicho análisis se aborda en consideración al modelo LightGBM ofrecido por el paquete PyCaret del software Python. Como

conclusión relevante exponen que, el acceso a internet influye en la obtención de mejores resultados en la prueba de matemáticas.

Otro referente importante fue el escrito por Timarán Pereira et al. (2019), en donde se usan árboles de decisión como técnica de minería de datos educativa para analizar factores vinculados al desempeño académico de los estudiantes de grado undécimo, quienes presentaron las pruebas Saber 11° entre los años 2015 y 2016. Se destaca de este texto la aplicación de la metodología CRISP – DM y el trabajo con la herramienta WEKA para el tratamiento de la información. Además, las conclusiones a las que llegan dichos autores permiten al lector meditar en los efectos sobre el agrupamiento de variables por dimensiones, las cuales facilitan la limpieza y tratamiento de los datos, y ubican como punto de reflexión que, en estos años, el porcentaje de estudiantes con desempeño bajo fue mayor al de nivel alto.

El texto de Junca Rodríguez (2019), analiza el desempeño en matemáticas de los educandos de bachillerato en las pruebas Saber 11 en el año 2009 – B, en Bogotá, a la luz de competencias argumentativas, comunicativas y de resolución de problemas. Para ello desarrolla un modelo multinivel jerárquico lineal mediante el cual establece tres niveles de estudio, el primero de ellos corresponde al de factores individuales, el segundo al de capital social (familia), y el tercero al factor de entorno escolar. El estudio concluye que el género del estudiante es un factor diferencial para el desempeño académico, siendo que los hombres presentan mejor rendimiento cuando trabajan de forma individual, mientras que las mujeres sobresalen cuando desenvuelven acciones en colectivo.

En Collazos Valenzuela et al. (2021), se analizan los determinantes del rendimiento académico de la educación media en la prueba Saber 11, referenciando para ello los resultados expuestos en la base de datos del ICFES concernientes a los periodos B de los años 2014 a

2019. En su componente metodológico los autores utilizaron un modelo de combinación de corte transversal, con el que lograron caracterizar a los estudiantes a través de posturas personales, familiares, socioeconómicas y del colegio de la población en estudio. Como conclusión exponen que, entre las pruebas evaluadas, la de inglés logró mejores resultados.

En Oviedo Carrascal & Jiménez Giraldo (2019) se cavila en el desempeño de los estudiantes de ingeniería en las pruebas de Estado Saber – Pro, en el departamento de Antioquía. El análisis de información se realizó con metodología CRISP- DM y bajo la instauración de variables académicas, económicas y sociodemográficas. Los autores utilizaron agrupación por clúster, selección de factores influyentes y predicción de desempeños con la base de datos de las pruebas Saber – Pro del año 2016. A manera de conclusión los autores exponen que el número de personas a cargo, método de enseñanza, hogar permanente, carácter académico de la institución y facilidades económicas, son los de mayor influencia en el desempeño dicha prueba. Se resalta de este artículo, entre otras cosas, la aplicación de la metodología abordada con CRISP – DM, la cual da pistas sobre cómo puede implementarse en el desarrollo del presente estudio. Agregado a ello, el artículo exhibe el tratamiento de los datos por medio de la herramienta “DQAnalyzer” para evaluar el perfil de las variables, “Frill” para detectar registros duplicados y WEKA versión 3.8.

En procura de identificar la deserción estudiantil de los estudiantes de Ingeniería de sistemas en una universidad privada en Colombia, Pérez Gutiérrez (2020) realiza un estudio entre diversas técnicas de minería de datos como los árboles de decisión, regresión logística y Naive Bayes. Este hecho resulta de interés en el presente estudio en tanto ubica referentes sobre la aplicación de las técnicas de minería de datos en el contexto educativo, además de exhibir herramientas como Jupyter Notebook para comprender datos, prepararlos y modelarlos;

Pandas para analizar los datos; Scikit-learn para conformar su estructura, y, Seaborn y Graphviz para visualizarlos.

El texto de Riquelme Santos et al. (2006), conceptualiza a la minería de datos (DM) y la diferencia de otras técnicas de análisis e interpretación de información. Muestra también la importancia que tiene la DM al ser utilizada en diferentes contextos y las posibilidades que ofrece a los investigadores para el tratamiento de datos. De este artículo se resalta la forma en la que los autores permiten al lector comprender términos propios de la minería de datos. Además, el artículo aborda la forma de preparar, seleccionar y limpiar datos dentro de la DM; define y clarifica los procesos de clasificación, regresión, clustering, generación de reglas, sumarización y análisis de secuencias. Por último, el artículo presenta las posibles aplicaciones de la DM y sus principales inconvenientes.

- ***Otras publicaciones relacionadas con rendimiento académico*.**

En la búsqueda de antecedentes bibliográficos también se encontraron investigaciones (nacionales e internacionales) que, si bien no mencionan directamente el desempeño académico de los estudiantes en las pruebas Saber 11, sí abordan temáticas sobre rendimiento académico de estudiantes y perspectivas docentes, formación y evaluación de competencias en el aula y lineamientos para la educación de calidad. Algunas de estas investigaciones analizaron el impacto que obtuvo la pandemia COVID – 19 en el desempeño de los educandos. Entre las investigaciones encontradas sobre rendimiento académico sobresalieron: Alkandari et al. (2021), Arias Serna (2009), Dicovskiy Riobóo & Pedroza Pacheco (2017), Haider & Al-Salman (2020), Lozano Díaz et al. (2020), Prada Núñez et al. (2021), Restrepo et al. (2020).

Entre las producciones que abordaron la formación y evaluación de competencias en el aula destacaron: Delors (1996), MEN (2006), Tobón (2013), Tobón Tobón et al. (2010). Entre

los trabajos que abordaron las perspectivas docentes frente al rendimiento académico de los estudiantes se analizaron los aportes de: Casali & Torres (2021), Cruz Guzmán & Benítez Granados (2020), Gamboa Unsihuay & Zuñiga Blanco (2021)**.** Y finalmente, los textos de CNA (1998) y MEN (2018), fungieron como referentes en el análisis de lineamientos sobre calidad educativa.

1.3.2 Descripción del problema.

Con intención de presentar al lector un texto esquemático sobre el problema que generó interés en esta investigación, se muestra en seguida la descripción del mismo abordando el conocimiento actual que se tiene sobre el problema y sus causas, sus posibles consensos o discrepancias, y los aspectos que aún no se logran conocer sobre este.

En Colombia, el análisis del rendimiento académico de los estudiantes en las pruebas Saber 11 es un aspecto que ha interesado a los investigadores en educación en las últimas décadas, puesto que sus resultados han permitido reflexionar sobre la calidad de los procesos pedagógicos y didácticos que se vienen desarrollando en las diferentes instituciones educativas, por ende que el desempeño de los estudiantes en estas pruebas haya sido laborado desde diferentes posturas, entre las que destacan los modelos econométricos de Chica Gómez et al. (2010), y, Rodríguez Rosero (2021); las técnicas de minería de datos educativa tratada por Timarán Pereira et al. (2019; 2020), Oviedo Carrascal & Jiménez Giraldo (2019); y otros modelos matemáticos susceptibles de ser adaptados a este fenómeno tal como los expuestos por Collazos Valenzuela et al. (2021), con los que se pretende determinar los factores de mayor influencia en la obtención de puntajes en la prueba Saber.

Sin embargo, en los resultados exhibidos por estos autores hay ciertos consensos como discrepancias, es decir, concuerdan al concluir que existe relación directa entre el desempeño

en la prueba y las variables: acceso a herramientas tecnológicas, nivel de formación de los padres, puntaje en la prueba de lectura, género, horas diarias dedicadas a lectura, tiempo invertido de los padres en la supervisión de las tareas escolares de sus hijos, por citar algunos casos, estableciendo así que un aumento en estas variables incrementa también los resultados de los estudiantes en las pruebas, pero no hay consenso en determinar un único método que analice o prediga con exactitud dichos resultados. Estos investigadores manifiestan que, debido a la presencia de factores sociales en la caracterización de los individuos, cada técnica de análisis y procesamiento de la información puede emanar aspectos que enriquezcan o complementen otros trabajos ya desarrollados, o sea, es como mirar un objeto desde diferentes ángulos en donde la unión de las consideraciones suscitas en los mismos aclaran el panorama del problema abordado.

Ahora, si bien es cierto que el análisis de los resultados de las pruebas Saber 11 ha implicado la proliferación de investigaciones que traten el rendimiento académico, sus factores determinantes y la caracterización de los educandos a través de los puntajes obtenidos, generando con ello un nuevo campo de estudio sobre estos temas (Junca Rodríguez, 2019), y que fruto de esto han surgido programas sociales como el "*Ser Pilo Paga"* y "*Generación E"*, los cuales buscan otorgar créditos condonables para que los estudiantes puedan acceder a la educación superior (ICETEX, 2022; Colombia aprende, 2022); es verdad también que los esfuerzos realizados por el Gobierno colombiano aún no son suficientes para lograr educación competitiva y de calidad en el contexto internacional, ya que como se detalla en Mullis et al. (2008), Fernandes Cristóvão (2010), MEN (2018), ICFES (2020b), los resultados de los educandos colombianos que han presentado las pruebas PISA y TIMSS, han posicionado al país por debajo de la media esperada en cada evaluación. En la prueba PISA, por ejemplo, los puntajes de los estudiantes han estado por debajo de la media de la prueba, siendo este evento más crítico en el área de matemáticas en donde sus puntajes promedio no han

alcanzado los 400 puntos, hecho que sí ha ocurrido en las otras áreas evaluadas y lo que posiciona a los resultados en matemáticas como los de menor puntuación promedio en el ámbito internacional. Los valores de los resultados de participación de Colombia en PISA se exhiben en detalle en MEN (2018). Agregado a ello se resalta que, los puntajes promedio de los estudiantes colombianos en matemáticas se han ubicado incluso por debajo del promedio latinoamericano y del OCDE, tal como se aprecia en ICFES (2020b).

Sanabria James et al. (2020), complementan las ideas precedentes al indicar que en la prueba PISA, Colombia ha ocupado los últimos lugares entre los países participantes (posición 62 de 65 totales), reflejando desempeños mínimos e insuficientes en las áreas de matemáticas y lenguaje, hecho que se agrava en los colegios públicos, ya que estos se sitúan por debajo del promedio exhibido por las instituciones de carácter privado. En específico, estos autores señalan que:

> El nivel de cada una de las competencias evaluadas es aún preocupante. Los resultados de las pruebas PISA para los años 2006 y 2009 registran aumentos en las tres áreas (matemáticas, lectura y ciencias). No obstante, el país se ubica en los rangos inferiores entre los países participantes. Para grado undécimo, el 44% de los estudiantes evaluados en pruebas nacionales en 2013 se ubicaban en el nivel bajo de la competencia matemática y en la prueba internacional PISA, Colombia ocupaba el puesto 62 en este componente de 65 países presentados. Para el periodo 2012 a 2015 más del 50% de los estudiantes de quinto grado se acumulan en los niveles de desempeño mínimo e insuficiente en las áreas de lenguaje y matemáticas. En noveno grado el porcentaje aumenta a más del 60% en matemáticas con una tendencia al 48 y 50% en lenguaje. Los resultados en las instituciones oficiales se registran por debajo de las instituciones privadas del país. (Sanabria James et al., 2020, p. 251).

En el caso de las pruebas TIMSS se sucede un evento similar, ya que, en la participación de las pruebas del año 2007, los estudiantes de grado cuarto y octavo no fueron ajenos a estos hechos situando al país por debajo del valor esperado en esta prueba. Al respecto Fernandes Cristóvão manifiesta:

> Tanto en matemáticas como en ciencias, en ambos grados, los estudiantes de los países asiáticos (Hong Kong, Singapur, Corea, Taipéi y Japón) tuvieron los promedios más altos. Inglaterra, Hungría y Rusia también lograron buenos resultados. Un número considerable de países evaluados, entre ellos Colombia, se ubicó por debajo del promedio TIMSS... El promedio global de los estudiantes colombianos de cuarto grado fue 355 puntos, el cual está muy por debajo de Hong Kong (607), Singapur (599), Taipéi (576) y Japón (568). En ese grado nuestro país superó solamente a Marruecos (341), El Salvador (330), Túnez (327), Kuwait (316), Qatar (296) y Yemen (224). Situación similar se observa en octavo, en donde el promedio global de Colombia fue 380, mientras que los de Taipéi, Corea y Singapur fueron, respectivamente, 598, 597 y 593. En ambos grados nuestro promedio fue significativamente inferior al promedio TIMSS… Naciones con nivel socioeconómico y de desarrollo humano similares a los de Colombia (Argelia, Irán, Ucrania y Turquía) lograron promedios significativamente más altos que los de nuestro país, aunque inferiores al promedio TIMSS. (Fernandes Cristóvão, 2010, p. 10).

Por lo anterior se expresa que los resultados de Colombia en TIMSS y PISA, han estado por debajo del promedio de los puntajes esperados, siendo el área de matemáticas la de menor estima, razón que incita a indagar sobre las posibles causas que motivan este hecho. Tal preocupación aumenta al considerar que en Colombia no se evidencian programas de capacitación que permitan tomar correctivos frente a este fenómeno, lo que se ha observado es

que, en los colegios existe preocupación por elaborar acciones de adiestramiento estudiantil en cuanto a la estructura de las pruebas, pero desde un punto de vista fragmentado, sin propuestas que transversalicen e integren la información. Al respecto, Sanabria James et al. (2020), exponen:

> En la búsqueda de resultados en pruebas nacionales e internacionales, el sector ha realizado múltiples intentos ajustados a distintas normas, pero sin transformar significativamente los resultados. Los esfuerzos se concentran en adiestrar a los estudiantes en la tipología de las evaluaciones, en los contenidos de las asignaturas o áreas y en el manejo de tipos de textos. Se hace un manejo fragmentado de la información con poca presencia de propuestas que transversalicen e integren la información, lo que dificulta al estudiante dar respuesta a posibles alternativas de solución que se le plantean en las evaluaciones. (Sanabria James et al. 2020, p. 250)

En concordancia con lo anterior, la preocupación por analizar el desempeño de los estudiantes en pruebas nacionales e internacionales crece al meditar en las consecuencias sufridas tras la afectación del COVID – 19, puesto que se ha detectado que el puntaje global de los estudiantes en las pruebas Saber 11 en los años 2020 y 2021 ha disminuido once puntos respecto a los resultados obtenidos en 2015 y 2016, hecho que conlleva a ahondar ideas sobre la manera en la que la pandemia influyó en el desarrollo de los procesos educativos en el aula, y ante lo cual se ha logrado concluir que se agudizó la desigualdad entre quienes gozan de recursos para acceder a la educación y quienes no los tienen en suficiencia, situación que puede promover el crecimiento de índices de deserción escolar y motivo por el cual el número de participantes en la prueba Saber 2020 fue menor respecto al del año anterior. En este punto, el informativo Portafolio sostiene:

> La última medición, en 2018, dio cuenta del bajo rendimiento de los alumnos colombianos, específicamente en lectura. "*Los estudiantes de Colombia tuvieron un rendimiento menor que la media de la OCDE en lectura, matemáticas y ciencias*", aclaró el certamen… Sumado a esto, este informe internacional resaltó la desigualdad entre adolescentes de bajos recursos y quienes tienen buenas condiciones de vida… De acuerdo con PISA, en el país los estudiantes con ventaja socioeconómica superaron a los estudiantes de bajos recursos en lectura por 86 puntos... Además, en el año 2020, el número de estudiantes que presentaron la prueba Saber 11 en colegios de calendario A se redujo un 5,4% frente al año inmediatamente anterior, pasando de 494.508 estudiantes en el 2019 a 467.896 en el 2020... De acuerdo con el ICFES, el promedio del puntaje global del calendario A en 2021 disminuyó 5 puntos respecto al 2014, pasando de 255 a 250 el año pasado. Ahora bien, el comparativo entre 2020 y 2021 tuvo como resultado una disminución del desempeño de 2 puntos. Respecto al calendario B, la institución sostuvo que se logró un incremento de 5 puntos entre 2020 y 2021; se pasó de 310 a 315, en cuanto al promedio del puntaje global. Sin embargo, este resultado ha disminuido 11 puntos en comparación con el año 2015. (Portafolio, 2022).

Sumado a lo anterior, Timarán Pereira et al. (2019), exponen que en los años 2015 y 2016 los resultados de la prueba Saber 11 reflejaron más estudiantes con desempeño académico bajo que con buen desempeño. Así entonces y considerando que la población mayoritaria en el 2015 y 2016 se hallaba en desempeño bajo y que los puntajes globales de la prueba Saber 2020 y 2021 han disminuido en 11 puntos en comparación al 2015, es plausible pensar que la calidad educativa en el país haya desmejorado, siendo el periodo pandémico y la virtualización de la educación un posible causante de ello.

De lo precedente emergen interrogantes como, ¿qué tan competentes están siendo los estudiantes actualmente en el desarrollo de las pruebas nacionales e internacionales? Si se sabe que los resultados del 2021 fueron menores que los del 2015 y que los desempeños en matemáticas han sido los de menor posicionamiento en las pruebas, entonces ¿cuál es el estado real de los procesos de alfabetización matemática en el país?, ¿en qué medida la instauración de las clases virtuales acaecidas por el COVID 19, mejoraron o empeoraron la asertividad del estudiante con los procesos de enseñanza, y en específico con las matemáticas?

Con intención de hacer frente a estos cuestionamientos, o al menos de sembrar la semilla en la solución de ellos, se elabora el presente trabajo de grado y con este, se inspecciona la relación entre rendimiento académico y pruebas Saber, haciendo énfasis en la búsqueda del porqué de los bajos desempeños de los estudiantes colombianos en el contexto internacional y en matemáticas en particular. La premisa que orienta esta investigación radica en determinar los factores que influyen en el desempeño de los estudiantes en las pruebas Saber, y en encontrar las características de aquellas personas que obtienen puntajes altos en matemáticas; sin embargo, y como un primer acercamiento a estos hechos, no se contempla a todo el colegiado colombiano, sino que se toma como caso de estudio a los educandos del departamento de Nariño, quienes en variadas ocasiones se han destacado en el ámbito nacional por sus buenos desempeños académicos en la prueba, llegando a posicionar a uno de sus planteles educativos (*Liceo de la Universidad de Nariño)* en el primer lugar de los colegios oficiales en las pruebas Saber (Caracol Radio, 2022).

Por ende, con este estudio se desea conocer ¿cómo fue el desempeño en matemáticas de los estudiantes del departamento de Nariño después de efectuar procesos desde la

virtualidad?, ¿qué relaciones se encuentran entre las áreas evaluadas?, y, ¿qué características se logran observar en las personas con puntajes altos en matemáticas en la prueba Saber 11?

El deseo de dar una primera aproximación a la solución de estos cuestionamientos motivó el desarrollo de la presente investigación, en donde se trabajó con RStudio y WEKA 3.9.6. Ante este evento surgieron dificultades para el estadístico en tanto RStudio precisa de un lenguaje de programación para su uso, lo que implica búsqueda constante de códigos para su adecuado manejo y análisis pertinente de las salidas computacionales. Agregado a esto, la base de datos proporcionada por el ICFES para el periodo 2021 – B presentó algunas limitaciones entre las que destacan el no mostrar con detalle cada pregunta realizada a los estudiantes para su posterior retroalimentación en las aulas de clase, lo que permitiría mejorar tanto a docentes y estudiantes sobre la forma de afrontar este tipo de preguntas. Otra limitación radicó en no exponer datos académicos sobre el profesorado y en no analizar el sentido de pertenencia de los docentes con la institución, lo que permitiría sospechar si los buenos o malos resultados de los colegiales están ligados con la actitud del docente hacia sus educandos en clase y con las estrategias metodológicas que él utiliza.

En igual sentido, otras dificultades refieren a que la base de datos requiere atravesar por un periodo de limpieza de información y a que son pocas las investigaciones realizadas con k – modas en el sector educativo (la mayoría usa k – means), motivo por el cual se dificulta encontrar producciones investigativas que sirvan de guía. Con todo lo anterior emerge la siguiente pregunta de investigación.

1.3.3 Pregunta de investigación.

¿Qué relación existe entre los puntajes de la prueba de matemáticas con los de lectura crítica, ciencias naturales, inglés, sociales y ciudadanas, de la prueba Saber 11 que

presentaron los estudiantes de Nariño en el periodo 2021 – B?, y, ¿qué características presentan los puntajes altos en matemáticas al situar como referentes aspectos socioeconómicos, demográficos, familiares, institucionales y de rendimiento en las pruebas?

1.4 JUSTIFICACIÓN DEL TRABAJO.

Desarrollar producciones investigativas en donde se acoja como tema de análisis el rendimiento escolar en las pruebas Saber 11, es un evento que ha generado interés en el contexto nacional puesto que sus resultados han servido de base para efectuar reflexiones sistemáticas de los procesos de enseñanza – aprendizaje que se vienen adelantando en el país, muestra de ello lo representa el trabajo elaborado por el Ministerio de Educación Nacional, quien con base en estos resultados y considerando también el desempeño de los estudiantes colombianos en pruebas internacionales, ha reorientado la evaluación Saber 11 desde una óptica cognitiva a otra más procedimental e interpretativa, teniendo en cuenta para ello los estándares básicos para la formación de competencias y los DBA que una persona debe desenvolver durante su paso por la estancia escolar, y que le permitan afrontar, desde posturas de ciudadanía crítica, las diversas eventualidades que posiblemente atravesaría en su diario vivir (MEN, 2006; Ruta Maestra, 2017).

Es así como el análisis de los resultados de las pruebas Saber 11 y la caracterización de los resultados de los estudiantes, han posibilitado la constante examinación de la calidad educativa que se viene desplegando en el país, sirviendo incluso de requisito para acceder a la educación superior, hecho que se evidencia en el artículo 14 de la Ley 30 de 1992 (Ley 30, 1992), y razón por la cual emerge la necesidad de comprender al educando como un ente participante de un mundo social, para quien su individualidad debe entenderse a partir de su propio lenguaje cultural y de los lineamientos propuestos por la comunidad académica global. El concepto de calidad educativa se ve ligado al de caracterización estudiantil y rendimiento

académico, siendo que este último se halla afectado por dimensiones sociales, económicas, políticas, culturales, entre otras; que subsisten en el aula y son pieza clave en los procesos de enseñanza - aprendizaje, y las que a decir de Tobón (2013) son poco tenidas en cuenta.

> Con frecuencia la evaluación de la calidad educativa se hace aplicando pruebas para determinar el logro cognitivo, y esto deja de lado las inteligencias múltiples, la actuación ante problemas reales y la ética. Además, pocas veces se evalúa la calidad de la educación considerando el ambiente socioeconómico, la alimentación de los estudiantes, la calidad de los materiales educativos, el entorno socio-familiar, la gestión de los directivos, la forma de trabajo de los docentes, las condiciones de los lugares de estudio, la influencia de los diferentes medios de comunicación, entre otros múltiples elementos interrelacionados. (Tobón, 2013, p. 17)

En el mismo orden de ideas, y reforzando el punto de vista de Tobón, Junca Rodríguez (2019) basado en D'Amore, menciona que, para comprender la caracterización del estudiante y su rendimiento académico en el contexto educativo, conviene considerar como factores influyentes al medio social que lo rodea, la interacción con sus docentes y otros que puedan estar presentes.

> Como plantea D'Amore (2006) para Brousseau el fenómeno de la enseñanza y el aprendizaje es un sistema didáctico y, como sistema conformado por tres elementos: maestro, estudiante, y Saber; inmersos en un medio externo "naturalmente existe un mundo externo, la sociedad en general, los padres, los matemáticos, etcétera" que forman parte del sistema y lo afectan. El concepto de medio ("milieu") es un concepto complejo que lleva consigo elementos explícitos e implícitos que están presentes y afectan de manera directa e indirecta las situaciones didácticas en la escuela.

> Para Bronfenbrenner (1979) existen diversos niveles de interacción que influyen en el desarrollo del individuo, como son las interacciones entre sus pares, la familia, la escuela, y la comunidad. (Junca Rodríguez, 2019, p. 17).

En sintonía con lo anterior, Rodríguez Espinar (1985) tras la lectura de los textos de Dyor y de Brookover, señala que la caracterización del estudiante y su rendimiento escolar se ve afectado por factores familiares, escolares, comunitarios y sociales, los cuales influyen de forma directa en el desempeño del educando.

> Así, Dyor (1972) opina que el ambiente (familiar, escolar y comunitario) afecta las actitudes de padres, profesores y compañeros; éstas influyen en la autopercepción del alumno que es la que definitivamente determina el rendimiento. Por su parte Brookover et. al. (1979) consideran que los diferentes inputs sociales de la escuela (características sociales del alumnado) influyen en los logros escolares tanto directamente como a través de la influencia mediacional de las características de la estructura social de la propia escuela y del clima social creado. (Rodríguez Espinar, 1985, p. 293).

En el análisis del medio externo y en la interacción del estudiante con sus pares, cobra sentido el desarrollo de la presente investigación, puesto que busca detectar la manera a través de la cual, la presencia de aspectos familiares, socioeconómicos, institucionales, demográficos y de rendimiento en la prueba, se relacionan con el desempeño y caracterización de los educandos en la prueba Saber 11, concentrando esfuerzos en los resultados obtenidos en matemáticas ya que, como se mencionó en la descripción del problema, los estudiantes colombianos no están siendo matemáticamente competentes en el ámbito nacional e internacional. Por efectos de delimitación del trabajo se toma como caso de estudio a los discentes del departamento de Nariño.

La realización del presente trabajo es importante además porque detecta características y analiza el rendimiento académico de estudiantes que atravesaron por periodos educativos de naturaleza virtual, enfatizando en los puntajes obtenidos en matemáticas con el fin de situar reflexiones pedagógicas que conlleven a mejorar el desempeño de los educandos en esa área de estudio ya que, como muestra Tobón (2013), existe, en términos generales, cierta tendencia a abordar el conocimiento matemático desde un enfoque de enseñanza tradicional, en donde se sitúa al docente como ente poseedor de conocimiento y a los discentes como recipientes sobre quienes verter dicho saber, de allí que Tobón llame la atención sobre la renuencia que tienen algunos docentes de matemáticas de cambiar su metodología, a una en donde el aprendiz desarrolle sus habilidades y no sólo cognitivas. En palabras de Tobón:

> Hace poco un supervisor educativo en México le envió al autor una comunicación en la cual está el siguiente párrafo: "(…) He venido trabajando con los docentes en mi zona la implementación de la reforma educativa en educación básica, pero los docentes de matemáticas se resisten de manera notable a trabajar por proyectos porque consideran que esta metodología no se aplica en este campo, ya que plantean que esto podría llevar a dejar de lado muchos contenidos del área por la dificultad de relacionarlos con un problema del contexto (…). Además, expresan que el plan de estudios viene por bloques y no se acomoda a los proyectos". (Tobón, 2013, p. 233).

En este sentido, el desarrollo del presente trabajo sirve de punto de apoyo para aquellas reflexiones educativas que promueven la renovación pedagógica y didáctica en las aulas de matemáticas, y las que propenden afinidad del aprendiz con esta área de estudio. Este trabajo también pretende servir de base para la instauración de la pedagógica multidireccional

propuesta por Moreno Olivos (2012) en donde el rendimiento matemático y la caracterización de los educandos se ve influido por factores socioeconómicos y culturales.

> La escuela del siglo XXI debe transitar de un modelo de pedagogía unidireccional centrado en la figura del profesor, cuya tarea principal ha sido la transmisión de conocimientos, hacia una pedagogía multidireccional y diferenciada que posibilite al alumno el desarrollo de una constelación de competencias tanto cognitivas como sociales, con las que haga frente de forma efectiva a los diversos problemas actuales (y futuros) caracterizados por ser abiertos, no estructurados y contradictorios, propios de la posmodernidad. (Moreno Olivos, 2012, p. 6).

Caracterizar a los estudiantes desde su desempeño en matemáticas se justifica también en tanto permite analizar la formación de competencias en el aula, y la discusión de cómo los procesos educativos facilitan que el conocimiento matemático pase del “Saber conocer” al “Saber hacer”, tal como se expone en los estándares de matemáticas (MEN, 2006), y lo que permite al estudiante poner en práctica sus conocimientos en la resolución de conflictos y situaciones problema. Aspecto que se torna interesante puesto que según Delors (1996), los docentes se preocupan más porque sus educandos memoricen fórmulas que por Saberlas aplicar en el contexto. En palabras de Delors:

> Más, en general, la enseñanza escolar se orienta esencialmente, por no decir que, de manera exclusiva, hacia el aprender a conocer y, en menor medida, el aprender a hacer. Las otras dos formas de aprendizaje dependen las más de las veces de circunstancias aleatorias, cuando no se las considera una mera prolongación, de alguna manera natural, de las dos primeras. (Delors, 1996, p. 96).

Conviene recordar en esta instancia que las pruebas Saber 11 han sido diseñadas abordando dos aspectos, uno que permite revisar la alfabetización matemática evaluada en las pruebas PISA, y otro que conlleva a detectar el nivel de competencia logrado de acuerdo a los procesos educativos acaecidos en cada salón de clase (ICFES, 2019). Con base en ello el estudiante debe dar cuenta de cómo usa los conceptos matemáticos en la solución de situaciones problema, es decir deben “Saber hacer” con las matemáticas y no solo quedarse en el plano de la comprensión de conceptos. Al respecto Acebedo citado por Junca Rodríguez (2019) exponen lo siguiente:

> La evaluación de la competencia matemática está referida al Saber hacer en el contexto matemático escolar, es decir, a las formas de proceder asociadas al uso de los conceptos y estructuras matemáticas. La aproximación que se hace a la competencia matemática en la prueba tiene en cuenta las significaciones que el estudiante ha logrado construir y que pone en evidencia cuando se enfrenta a diferentes situaciones problema. En las pruebas es importante evaluar el significado de los conceptos matemáticos y la práctica significativa, relacionada esta última con la matematización que exige al estudiante simbolizar, formular, cuantificar, validar, esquematizar, representar, generalizar, entre otros, actividades que le permitirán desarrollar descripciones matemáticas, explicaciones o construcciones. (Acevedo y col. (2007), p. 22) en (Junca Rodríguez, 2019, p. 18).

Por otra parte, el desarrollo de esta investigación se justifica porque utiliza técnicas de minería de datos para el análisis de sus resultados, las cuales, como señala Pérez Gutiérrez, facilitan la evaluación de las necesidades de los educandos y ayudan a “predecir las tasas de deserción escolar, analizar y mejorar el rendimiento académico de los estudiantes” (Pérez Gutiérrez, 2020, p. 202). Agregado a ello, Timarán Pereira et al. (2019), presentan la necesidad

de abordar trabajos en donde se indaguen las interrelaciones acaecidas entre las variables de la prueba Saber y las cuales se potencian con el uso de minería de datos.

El trabajo con mineria de datos se muestra interesante puesto que facilita el análisis de los factores socioeconómicos presentes en las prueba Saber 11, en este punto, Oviedo Carrascal exhibe en su trabajo la manera en la que la mineria de datos ha sido usada para intepretar el desempeño en lectura de las pruebas PISA en España y con lo cual concluyó que "… variables como disponibilidad de computador, género y estado de inmigración son importantes para los resultados en matemáticas…", además situó al análisis de clúster como la técnica de minería datos más utilizada (Oviedo Carrascal & Jiménez Giraldo, 2019, p. 129). Trabajar con minería de datos es importante porque como señala Riquelme y comitiva, facilita "… la automatización para manejar grandes volúmenes de datos heterogéneos" (Riquelme Santos et al. 2006, p. 13), y porque, a pensar de Menacho Chiok (2017), las técnicas de minería de datos aplicadas al contexto educativo son eficaces para predecir el rendimiento académico estudiantil.

La realización de este trabajo se justifica también porque el análisis de sus resultados gira entorno de los referentes de evaluación de calidad propuestos en CNA (1998) y CNA (2006); puesto que promulgan la divulgación de conocimiento científico efectuado en las universidades, y razón por la cual la presente Maestría solicita a sus participantes la elaboración de este tipo de producciones. Además, elaborar estudios evaluativos que traten la caracterización del estudiantado en las pruebas Saber 11 y el análisis de los resultados obtenidos en ellas, se halla en sintonía con lo propuesto por Bogoya (2006) quien alude la importancia de realizar producciones a niveles de maestría y doctorado que traten sobre aspectos de calidad educativa. Al respecto Bogoya sostiene que:

> Colombia tiene cuatro grandes desafíos para su sistema de evaluación educativa: primero, brindar información robusta, confiable y oportuna acerca de los hallazgos que se revelan al realizar cada evaluación; segundo, fomentar la referenciación institucional entre las distintas instituciones que participan en una evaluación; tercero, convertir la evaluación en un campo de investigación propio de universidades en sus niveles de maestría y doctorado y de institutos especializados; y cuarto, crear y pulir nuevas estrategias y modelos de evaluación. Adicionalmente, es necesario desarrollar aplicaciones con tecnología de punta para ofrecer la presentación asincrónica y virtual de pruebas, es decir, en el momento que un usuario lo requiera y desde el sitio donde se encuentre. El logro de estos desafíos permitirá colocar al país a la vanguardia del tema en la región latinoamericana y en una posición apropiada para dialogar con solvencia conceptual con países del primer mundo sobre evaluación. (Bogoya, 2006, p. 23).

Finalmente, las ideas expuestas en este trabajo se justifican en tanto ellas se encuentran en sintonía con el interés propuesto por la Universidad de Nariño de abordar producciones investigativas que analicen problemáticas reales y las cuales produzcan impacto en la región, puesto que se orienta bajo los requerimientos establecidos por la Maestría en Estadística Aplicada de la UDENAR y funge como aspecto evaluativo del nivel académicamente competitivo en el que se encuentran los estudiantes del departamento de Nariño. Y porque además plantea nuevos problemas para futuros investigadores en los tópicos de rendimiento académico, formación de competencias para la ciudadanía, manejo de bases de datos por medio de clúster y análisis de factores, evaluación de calidad y uso de software RStudio y WEKA.

1.5 OBJETIVOS.

1.5.1 Objetivo general.

Determinar las relaciones entre los puntajes de la prueba de matemáticas con los de lectura crítica, ciencias naturales, inglés, sociales y ciudadanas, de la prueba Saber 11 que presentaron los estudiantes de Nariño en el periodo 2021 – B, y caracterizar los puntajes altos en matemáticas a partir de aspectos socioeconómicos, demográficos, familiares, institucionales y de rendimiento en las pruebas.

1.5.2 Objetivos específicos.

- Determinar las características de los estudiantes en lo referente a variables de tipo socioeconómicas, demográficas, familiares, institucionales y de rendimiento en las pruebas.
- Analizar la relación entre el nivel alcanzado en matemáticas y las variables de tipo socioeconómicas, demográficas, familiares, institucionales y de rendimiento en las pruebas.
- Determinar la correlación existente entre los puntajes de las pruebas de lectura crítica, matemáticas, ciencias naturales, inglés, sociales y ciudadanas.
- Identificar estructuras fundamentales o factores asociados a los puntajes dados en las pruebas de lectura crítica, matemáticas, ciencias naturales, inglés, sociales y ciudadanas.
- Analizar y comparar modelos de agrupación que identifiquen grupos de estudiantes con características similares en torno al puntaje en matemáticas.

1.6 MARCO LEGAL.

Los referentes legales tenidos en cuenta para el desarrollo de esta investigación fueron:

En primer lugar, se nombra a la Constitución Política de Colombia, ya que en su artículo 67 propone a la educación como un derecho para los colombianos la cual les permitirá formarse como ciudadanos libres, con deberes y valores democráticos. (Constitución Política de Colombia, 1991).

En segundo lugar, se menciona la Ley General de Educación (Ley 115, 1994), ya que propone los lineamientos que deben considerar las instituciones educativas para prestar sus servicios en los ámbitos formal y no formal. De esta Ley se destacan los artículos adjuntos.

- 8 literal d. Puesto que la presente investigación sirve de apoyo para las instituciones educativas, ya que sus resultados analizan fenómenos de educación de calidad.
- 29 y 32. En tanto los resultados de este trabajo dan cuenta del puntaje obtenido en la prueba Saber 11, válida para acceder a la educación superior.
- 77. Brinda autonomía a las instituciones educativas para organizar los contenidos temáticos escolares, hecho de interés en esta investigación porque contrasta los resultados de una prueba estandarizada (Saber 11) con la autonomía de cada plantel para armar sus estructuras curriculares.
- 80. Propone al ICFES como ente regulador de la calidad educativa brindada por los planteles educativos, declarando como obligación del Estado el elaborar acciones de refuerzo para las instituciones que así lo ameriten.
- 99. Garantiza el acceso a la educación superior a los 50 mejores puntajes obtenidos en el ámbito nacional en las pruebas Saber 11, así como a los dos puntajes más altos por departamento.

- 148. Establece las funciones de evaluación y control de los resultados de los planes y programas educativos, así como sostiene que es deber del MEN la evaluación permanente de la prestación de los servicios educativos.

En tercer lugar, emerge la Resolución 2343 propuesta por el MEN (Resolución 2343 de junio 5 de 1996), la que en sus diferentes artículos promueve consideraciones relevantes en torno a las prácticas educativas y evaluativas, artículos entre los que sobresalen los siguientes:

- 5. Acoge a la comunidad educativa en la estructuración del currículo.
- 7. Presenta la definición de currículo, hecho importante en este trabajo puesto que los resultados de las pruebas de Estado evalúan la manera en la que el diseño curricular de cada institución es acorde con las expectativas de calidad educativa propuestas en el país.
- 10. Estructura por grados escolares a la educación formal y por la cual se detalla que la educación media comprende a los grados décimo y undécimo, siendo estos los habilitados para presentar las pruebas Saber 11.
- 19. Habla sobre la evaluación del desempeño escolar, la cual exhorta a docentes, estudiantes, padres de familia y autoridades educativas, a concienciar sobre su rol y compromiso con los procesos educativos efectuados en las aulas, y a efectuar investigaciones que procuren mejoras en el rendimiento estudiantil.

Otro referente legal que se suscita como importante es la Ley 30 de 1992, la cual contiene los fundamentos de la Educación Superior y la que en su artículo 14 establece a la prueba Saber 11 como requisito de ingreso para los programas de pregrado. (Ley 30, 1992).

El Decreto 869 de 2010 en su artículo uno especifica la importancia que tiene el desarrollo de las pruebas de estado en el contexto nacional. En el artículo dos instaura al ICFES como director y coordinador del diseño, producción, aplicación, procesamiento, análisis de los resultados acaecidos en las pruebas de Estado. En el artículo seis le otorga la facultad de divulgar dichos resultados a la comunidad académica interesada. Este decreto sostiene también que la estructura esencial de la prueba debe estar vigente al menos doce años. (Decreto No. 869, 2010).

La Ley 1324 de 2009 establece los parámetros y criterios para establecer un sistema de evaluación confiriendo al ICFES la tarea de examinar la calidad educativa efectuada en el país. En su artículo 7 expone que se considera exámenes de Estado aquellos que evalúan la educación formal de quienes terminan la educación media, o quienes desean acreditar el dominio de sus competencias tras terminar dicho nivel escolar. También presenta como exámenes de Estado aquellos que evalúan la educación del pregrado. (Ley 1324, 2009).

1.7 MARCO TEÓRICO.

Los fundamentos teóricos que soportan el desarrollo de esta investigación se exponen considerando los tópicos de rendimiento académico, consideraciones generales sobre la prueba Saber 11 y las competencias en matemáticas, minería de datos educativa y algoritmo clúster enfatizado en k – modas.

1.7.1 Rendimiento académico.

Como han señalado varios autores, el rendimiento académico es en sí mismo un concepto de naturaleza multidimensional en donde y por la complejidad que este envuelve, no puede ser definido de manera única, de allí que sea posible estudiarlo como la calificación obtenida al final de un curso, o desde sus dimensiones socioeconómicas, políticas, familiares,

demográficas, las cuales acarrean consigo un sin número de subdimensiones favorecidas por la postura de un determinado modelo de investigación. Para comenzar con la conceptualización de este término se tiene en cuenta lo dicho por Solano Luengo (2015), quien lo presenta como el nivel de conocimientos que el alumno posee al concluir un curso y el cual se refleja en las valorativas finales. En palabras de mencionado autor:

> Nivel de conocimientos que el alumno demuestra tener en el campo, área o ámbito que es objeto de evaluación; es decir, el rendimiento académico es lo que el alumno demuestra saber en las áreas, materias o asignaturas en relación con los objetivos de aprendizaje y en comparación con sus compañeros de aula o grupo. Así pues, el rendimiento se define operativamente tomando como criterio las calificaciones que los alumnos obtienen. (Solano Luengo, 2015, p. 25).

Definición que, dicho sea de paso, concuerda con Jiménez (como se citó en Edel Navarro, 2003, p.3) quien lo precisa como: “nivel de conocimientos demostrado en un área ó materia comparado con la norma de edad y nivel académico”. Estas definiciones sitúan al rendimiento académico como un aspecto meramente cognitivo donde juegan un rol importante las destrezas y capacidades que cada individuo posee para hacer frente a las problemáticas planteadas en cada sesión de clase, así como la forma en la que el estudiante aborda las evaluaciones planteadas por el profesor y su actitud frente a los procesos de aprendizaje. Bajo esta definición, el rendimiento académico se ve afectado principalmente por el interés, voluntad y capacidad que cada individuo tiene con su proceso de formación, y por lo cual el éxito o fracaso escolar es causado únicamente por el grado de compromiso del discente. Al respecto, Covington (como se citó en Edel Navarro, 2003) manifiesta que existen tres tipos de estudiantes: los exitosos, los que aceptan el fracaso, y quienes evitan el fracaso. Los exitosos se caracterizan por tener alta motivación y confianza en sí mismos. Los que aceptan el fracaso

tienen autoestima baja y generalmente asumen conductas de personas académicamente derrotadas. Y quienes evitan el fracaso, son aquellas personas que sin esforzarse mucho procuran cumplir con los requisitos mínimos de cada asignatura.

Sin embargo, y según Álvaro Page et al. (1990), en el rendimiento académico toman parte también otro tipo de factores, diferentes al ámbito netamente cognitivo y que atañen a condiciones adjuntas al contexto educativo y de desarrollo de inteligencias múltiples. Estas condiciones son también responsables del desempeño de los estudiantes en las pruebas y deben considerarse en el análisis del rendimiento estudiantil. En palabras de dicho autor:

> Entre ellas figura la concepción del rendimiento basada en la voluntad (Kaczynska, M. 1965), según la cual el que un alumno rindiese o no, dependía de su buena o mala voluntad. Se olvidaban importantes factores individuales y sociales que inciden en el éxito o fracaso escolares, como son el nivel intelectual, las aptitudes, actitudes y ciertas condiciones de vida de los alumnos.
>
> Considerando el rendimiento desde el punto de vista de la capacidad se pueden hacer críticas similares. Según esta perspectiva mantenida por Muñoz Arroyo, A. (1977), si un niño no rinde es porque no tiene capacidad suficiente. Se suele esperar de un estudiante que tiene buena capacidad un alto nivel de rendimiento. Pero ésta es sólo una verdad a medias. Hay que tener en cuenta, de acuerdo con Secadas, F. (1952), que en el rendimiento influyen otros elementos como pueden ser la aplicación o esfuerzo del estudiante, así como condiciones temperamentales y situacionales del mismo. (Álvaro Page et al. 1990, p. 18).

Reforzando lo anterior, situar al rendimiento académico como la capacidad cognitiva del educando para resolver evaluaciones, deja de lado otros factores que influyen en esta etapa y

son determinantes, entre ellos el análisis del grado de objetividad de las pruebas, las cuales según Álvaro Page et al. (1990), pueden ser estandarizadas como las aplicadas por el ICFES, o no estandarizadas como las empleadas por cada profesor en sus clases y las cuales deben someterse a tratamientos estadísticos de revisión y reelaboración de elementos para lograr fiabilidad y validez. Elaborar pruebas objetivas para la medición del rendimiento académico es importante porque ellas mismas le permiten al docente adquirir un panorama del estado real del aprendizaje del alumno, evitando calificaciones subjetivas debidas a síntomas de cansancio al evaluar, posibles prejuicios sobre el discente, y la no comprensión de los resultados de una prueba debida a que los procesos presentados por el estudiante no tienen claridad, coherencencia y orden en su presentación. Al respecto, Lemus, L. (como se citó en Álvaro Page et al. (1990), señala que las pruebas estandarizadas deben caracterizarse por:

> 1. Estar compuestas de ítems o elementos seleccionados sobre la base de los objetivos específicos de la instrucción.
> 2. Los resultados de cada ítem en particular y los de toda la prueba en general deben ser analizados estadísticamente a efecto de determinar su grado de dificultad y de validez.
> 3. La prueba debe ir acompañada de instrucciones para su aplicación y calificación, y de normas para la interpretación de los resultados.
>
> Las pruebas objetivas, como medidas del rendimiento, ofrecen mayores ventajas; entre éstas destaca su grado de objetividad, debido a que las respuestas son cortas y precisas, sin la influencia subjetiva del profesor; estas pruebas poseen un alto grado de validez, debido a que cumplen específicamente los propósitos para los que fueron elaboradas; con estas pruebas, el profesor puede realizar una exploración mayor de los conocimientos del alumno y de una cantidad más amplia de materia en un tiempo relativamente breve. Lemus, L. (como se citó en Álvaro Page et al. 1990, p. 27).

Agregado a lo anterior, y en la búsqueda de factores que intervienen en el rendimiento académico del aprendiz, el ya citado Álvaro Page instaura como determinantes de desempeño a variables contextuales y de tipo personal. Las variables contextuales analizan el desempeño del alumno desde una óptica exterior a él mismo, se subdividen en sociofamiliares y en escolares. En las sociofamiliares se abordan tópicos de *clima educativo familiar, estructura familiar, origen social, medio sociocultural, y las características del hábitat o población de residencia*, y se parte del hecho de que la familia juega un rol importante en el desempeño del alumno, constituyendo un factor influyente el número de miembros que componen el núcleo familiar, así como la posición del educando en el mismo, el nivel educativo de los padres y su estatus social. Aquí se destaca que una mala relación del estudiante con sus progenitores puede conllevarle a visualizar a la escuela como un lugar desafiante posibilitando con ello el fracaso escolar. En este sentido Gilly citado en Álvaro Page et al. (1990) plantea lo siguiente:

> 1 En primer lugar, la mala calidad del clima educativo familiar es un factor de mala adaptación escolar dada la diferencia entre el sistema de valores que rige la vida familiar y el que rige la vida en la escuela. El paso de uno a otro puede ser realmente duro y desorientador para el niño.
>
> 2 La mala calidad del clima educativo familiar puede tener repercusiones sobre las condiciones materiales del trabajo escolar; es fácil comprender que la falta de tranquilidad, de paciencia y de autoridad paternas pueda perjudicar la buena marcha del trabajo que el niño tenga que hacer en casa.
>
> 3 La mala calidad del clima educativo puede engendrar problemas relacionales en la familia situando asi al niño en una atmósfera general de inseguridad afectiva de la que son responsables los padres. El fracaso escolar puede ser considerado por el niño

> como una especie de venganza de los agravios que haya recibido de su familia. Gilly (como se citó en Álvaro Page et al. 1990, p. 39).

Profundizando ideas en el componente sociofamiliar, entra en escena el análisis del estrato social, el cual es un factor clave para el desempeño escolar, puesto que según el medio y las condiciones en donde el estudiante se desenvuelva tendrá mayor facilidad o dificultad para acceder a la educación. En esta instancia el concepto de “clase social” es un indicativo de estratificación social, el cual según Poulantzas, N (como se citó en Álvaro Page et al. 1990) menciona que:

> Se ha convertido en una variable compleja que incluye factores como la ocupación, el nivel de ingresos, el prestigio social o la educación y permite diferenciar, en función de ellos, un cierto numero de posiciones (básicamente las clases "alta", "media" y "obrera"). Poulantzas, N (como se citó en Álvaro Page et al. 1990, p. 52).

En el mismo orden de ideas, Piñero & Rodríguez (como se citó en Edel Navarro, 2003) postulan que el análisis del estatus social de los individuos tiene efectos positivos sobre el desempeño estudiantil y motivo por el cual emerge como variable influyente en este proceso, a decir de estos autores:

> La riqueza del contexto del estudiante (medida como nivel socioeconómico) tiene efectos positivos sobre el rendimiento académico del mismo. Este resultado confirma que la riqueza sociocultural del contexto (correlacionada con el nivel socioeconómico, mas no limitada a él) incide positivamente sobre el desempeño escolar de los estudiantes. Ello recalca la importancia de la responsabilidad compartida entre la

familia, la comunidad y la escuela en el proceso educativo. Piñero & Rodríguez (como se citó en Edel Navarro, 2003, p. 5).

Las variables escolares por su parte, abordan a *la institución escolar* mediante características como ubicación, dotación, organización, dirección, gestión del centro educativo, carácter público o privado; asimismo acoge la figura del profesor a través de su metodología, edad, formación, experiencia profesional y forma de interrelacionarse con su discente, y como último examina al alumno desde la relación con los demás, con el medio y entorno escolar. En las variables de tipo personal se inspeccionan características propias del estudiante como lo son sus inteligencias y aptitudes, estilos cognitivos, sexo del aprendiz, personalidad en sus aspectos de extraversión, ansiedad, motivación y autoconcepto. Tratar este tipo de variables es importante porque, tal como señala Collazos Valenzuela et al. (2021) basada en Gallegos & Campos (2019), las características personales tienen una influencia mayor en el rendimiento estudiantil que las dimensiones sociales e institucionales del individuo.

En concordancia con lo anterior, Rodríguez Espinar (1985) también plantea que el concepto de rendimiento académico va más allá de la mera consideración de las calificaciones que los alumnos obtienen al final de un curso o etapa escolar, ya que en el mismo intervienen factores sociales, psicológicos, económicos y didácticos, que conllevan a inspeccionar las características del estudiantado a fin de precisar el porqué de un desempeño determinado. En este sentido mencionado autor propone al rendimiento académico como la determinación del éxito o fracaso escolar el cual se ve sujeto a dimensiones sociales, educativas y económicas. En lo social emerge como interesante el análisis del estatus social y bajo el cual plantea que, si a mejores rendimientos académicos se producen mejores índices de calidad de vida, entonces dicho rendimiento debe pensarse de la óptica del acceso a funciones productivas. La dimensión educativa por su parte, acoge los intereses institucionales bajo el desarrollo del interrogante

¿qué tan adecuados son los tratamientos educativos?, con lo cual se invita a que cada plantel establezca comparaciones de los logros obtenidos con los alcanzados por otras instituciones, a fin de realizar un análisis crítico introspectivo que le ayude a detectar falencias y fortalezas en el planteamiento y ejecución de sus procesos escolares. En la dimensión económica se habla de las inversiones en educación tanto en forma como en contenido, es decir, inversión en infraestructura física (baños, aulas, compra de materiales, etcétera), como en propuestas de capacitación al personal.

Complementando lo precedente, los trabajos de Schneider et al. (tal como se citó en Rodríguez Espinar, 1985), indican que el rendimiento académico va más allá de considerarlo como una nota al final de un curso o ciclo escolar, ya que esta postura subestima los efectos que genera la escuela sobre el educando y los que inmiscuyen las variadas relaciones interpersonales debidas a la interacción entre alumnos, docentes, padres de familia y comunidad escolar en general, y que toman partida en el análisis de los logros alcanzados en la formación y desarrollo profesional del individuo. De igual manera Glasman & Biniaminov (1981) en Rodríguez Espinar (1985), exponen que el rendimiento académico puede ser abordado desde posturas cognitivas y no cognitivas, siendo las primeras las de mayor elección por parte de investigadores en este ámbito. Es así como Edel Navarro (2003) señala al rendimiento desde una postura cognitiva, exhibiéndolo como:

> Un constructo susceptible de adoptar valores cuantitativos y cualitativos, a través de los cuales existe una aproximación a la evidencia y dimensión del perfil de habilidades, conocimientos, actitudes y valores desarrollados por el alumno en el proceso de enseñanza aprendizaje. Lo anterior en virtud de destacar que el rendimiento académico es una intrincada red de articulaciones cognitivas generadas por el hombre que sintetiza las variables de cantidad y cualidad como factores de medición y predicción de la

> experiencia educativa y que contrariamente de reducirlo como un indicador de desempeño escolar, se considera una constelación dinámica de atributos cuyos rasgos característicos distinguen los resultados de cualquier proceso de enseñanza aprendizaje. (Edel Navarro, 2003, p. 13).

Asimismo, Rodríguez Espinar (1985) plantea que el rendimiento académico puede ser entendido de mejor manera agrupando los factores cognitivos y no cognitivos del educando en un conjunto, y en otro se debe situar al componente temporal el cual liga los logros alcanzados por las personas en su interacción con la sociedad, en donde se precisa que las calificaciones obtenidas no son eternas, es decir se deben al lugar y tiempo en donde se produjeron con todas las variantes que pudieron estar allí establecidas, situación compartida por Edel Navarro (2003) para quien las calificaciones otorgadas en las pruebas, no deben verse como una etiqueta permanente de la persona, sino que son momentáneas y obedecen a condiciones por las cuales atraviesa el individuo al efectuar la prueba. Anexo a esto, Rodríguez Espinar elabora una clasificación de criterios sobre rendimiento académico, los cuales dan cuenta de: (a) factores de rendimiento, (b) tipología del diseño de investigación, (c) función atribuida a la institución escolar, (d) explicitación de relaciones entre variables, y, (e) modelos longitudinales causales en el rendimiento académico.

El criterio *factor de rendimiento* es el más utilizado para analizar el éxito o fracaso escolar y puede estudiarse en consideración a los subcriterios (1) psicológico, (2) sociológico y (3) carácter didáctico. El primero de ellos refiere a los atributos personales del alumno y el profesor. El segundo, acoge los elementos estructurales y estáticos del contexto como los componentes familiares, institucionales y sociales; y presenta a la escuela con un sistema social en donde existen interacciones entre padres de familia, profesores y alumnos, a partir de

variables como clase social, nivel de estudios, zona de ubicación. El tercero, trata sobre los métodos, recursos y organización del proceso de aprendizaje en el contexto educativo.

El criterio *tipología del diseño de investigación*, muestra que el rendimiento académico puede ser abordado desde estudios observacionales (ex – post – facto) o experimentales, siendo los primeros los de mayor acogida en la comunidad académica, asimismo, se plantea la aplicación de estudios transversales o longitudinales. Por su parte, el criterio *función atribuida a la institución escolar*, establece que la escuela puede ser estudiada bajo modelos de entrada y salida (input – output), que conlleven a determinar la calidad del proceso producto educativo, y las variables suscitadas en la interacción docente – alumno. El criterio *explicitación de relaciones entre variables*, propone la instauración de modelos (aditivos, mediacionales, y de interacción) para el trabajo con las variables. El modelo aditivo considera independientes a los factores que influyen en el rendimiento del alumno, pero con efecto aditivo, es decir, la suma de cada factor brinda la totalidad del modelo estudiado. Rodríguez Espinar et al. (1985), aclara que la regresión múltiple juega un papel importante en esta etapa ya que permite visualizar al rendimiento (R) bajo la participación de aptitudes, ambiente educativo e instrucciones, las cuales se aúnen en la función (f) así: $R = f(aptitudes + ambiente + instrucción)$.

El segundo modelo es el mediacional en donde se abordan los factores que influyen, aunque no en forma directa, en el rendimiento estudiantil, este tipo de modelos son conocidos como los interactivos de causalidad unidireccional y explican que el resultado académico viene dado por la interrelación profesor – alumno, y por las características particulares de cada uno de ellos, así como las condiciones particulares del contexto en donde se encuentran.

El tercer modelo mencionado por Rodríguez Espinar es el interactivo, aunque no exhibe mayores diferencias entre los mediacionales y estos últimos. El criterio *modelos longitudinales*

causales en el rendimiento académico, presenta tres puntos o formas de ser abordados. En primer lugar, sitúa el diseño del papel longitudinal 2W2V, en segunda instancia trata las relaciones causales – longitudinales en donde las variables medidas son susceptibles de contener error, y en tercer punto señala a los indicadores de la potencia con efectos recíprocos.

En resumen, el rendimiento académico es un concepto multidimensional, el cual puede ser abordado desde posturas cognitivas, no cognitivas, o a través de sus dimensiones personal – psicológica, demográfica, socioeconómica, temporal, escolar, sociofamiliar; siendo todas ellas influyentes en el resultado final reflejado por el aprendiz en sus procesos educativos. Anexo a ello, el rendimiento académico puede ser estudiado desde ópticas estadísticas al considerar el tratamiento de sus datos mediante modelos aditivos, mediacionales, interactivos, de entrada, o salida, o econométricos, por citar algunos ejemplos. En el análisis del rendimiento académico juega un rol importante la triada profesor, estudiantes y escuela, las cuales se interrelacionan de tal manera que se satisfagan los lineamientos propuestos por el Gobierno y le permitan al individuo formarse bajo posturas de pensamiento crítico y educación para la ciudadanía.

1.7.2 Consideraciones generales sobre la prueba Saber 11 y las competencias en matemáticas.

- **La prueba Saber.**

Tal como indica MEN (s.f), el ICFES como ente encargado de evaluar la calidad de la educación en el país, con el transcurrir del tiempo ha venido realizando cambios en el diseño y estructura de la prueba Saber 11 con el fin de satisfacer las exigencias planteadas por la comunidad educativa nacional e internacional. Estos cambios se han dado en tres momentos a través de los cuales las pruebas han presentado aspectos peculiares que no los hacen comparables entre periodos diferentes, es decir, las pruebas del periodo 1 (2000 – 1 a 2004 –

2) no pueden equiparar sus resultados con los del periodo 2 (2005 – 1 a 2014 – 1), ni con los del periodo 3 (2014 – 2 en adelante).

Enfatizando ideas en el tercer periodo cabe anotar que las pruebas genéricas de matemáticas, lectura crítica, ciencias naturales, sociales ciudadanas e inglés, se reestructuraron de tal manera que ya no se evalúan solamente los aspectos cognitivos sino también el grado del dominio de la competencia que tiene el educando sobre cada área de estudio. Para el periodo 3, y con intención de hacer estimaciones sobre los puntajes, el ICFES fijó como promedio para cada prueba el valor de 50 puntos con desviación estándar de 10 unidades, y como promedio del puntaje global a 250 con desviación estándar de 50. El valor máximo en cada prueba es de 100 y el máximo del puntaje global es 500, en ambos casos el valor mínimo es 0. El puntaje global (PG) se calcula con la expresión $PG = 5 * IG$, donde:

$$IG = \frac{3 * MATEMÁTICAS + 3 * LECTURA + 3 * CIENCIAS + 3 * SOCIALES + 1 * INGLES}{13}$$

Las expresiones: MATEMATICAS, LECTURA, CIENCIAS, SOCIALES, INGLES; representan los puntajes obtenidos en la prueba de su respectiva área.

Por otra parte, el ICFES clasifica a las instituciones educativas según categorías de rendimiento signadas como: **A+, A, B, C, D**. En donde A+ implica que el plantel educativo tiene "… más del 85% de sus estudiantes en el 33% más alto de la distribución de puntajes y los clasificados en D tienen menos del 40% de sus estudiantes en el mismo rango". (MEN, s.f, p. 8). Además, y con intención de brindar datos relevantes a los investigadores en este campo, el ICFES proporciona variables de estudio que pueden ser abordadas desde la información

personal, socioeconómica, académica y de citación, las cuales son conceptualizadas de la siguiente manera:

1. **Información personal:** Este módulo indaga por aspectos como el género del estudiante, pertenencia a una etnia, discapacidades, lugar de residencia, entre otros.
2. **Información Académica y de citación:** Este módulo indaga por aspectos cómo el colegio al que pertenece el estudiante, valor de la pensión que paga en su colegio (en caso de que lo haga), entre otros.
3. **Información socioeconómica:** Este módulo indaga por aspectos familiares como el nivel educativo de los padres, su ocupación, servicios con los que cuenta el hogar, entre otros. (MEN, s.f, p. 9).

Con intención de optimizar el tratamiento de la información personal, académica y socioeconómica anteriormente expuesta, emergen en esta instancia diversas investigaciones en donde se muestra la forma en la que algunos autores agruparon las variables brindadas por el ICFES y entre las cuales se destacan los aportes de Cuéllar Caicedo et al. (2016), quienes proponen estudiar el índice de nivel socioeconómico (INSE) en consideración a las dimensiones de potencial económico del hogar, capital cultural y dotación de la vivienda. Cuéllar Caicedo y compañía definen cada dimensión de la siguiente forma.

> El INSE se construye a partir de tres dimensiones: potencial económico del hogar, capital cultural y dotación de la vivienda. Para la dimensión potencial económico del hogar, se tienen en cuenta tres variables proxy relacionadas con los ingresos económicos del hogar, estas variables son: valor de la matrícula, ingreso familiar mensual y SISBEN, todas ellas ordinales.

> En cuanto a la dimensión capital cultural, se tienen en cuenta cuatro variables: educación de la madre y educación del padre, ocupación del padre y ocupación de la madre. Las variables de educación presentan categorías de respuesta como: primaria completa, primaria incompleta, secundaria incompleta, secundaria completa, técnico incompleto, técnico completo, universitario incompleto, universitario completo, posgrado, entre otros. Estas son consideradas variables ordinales y son de gran importancia si se tiene en cuenta que la educación de los padres es una variable que tiene una correlación alta con la educación de los hijos... En cuanto a la ocupación de los padres, entre las posibles categorías de respuesta se encuentran: estudiante, jubilado, hogar, empresario, trabajador independiente, entre otras. Las anteriores variables se consideran importantes debido a la relación que existe entre el capital cultural y el desempeño académico. (Coleman 1988, Gil Flores 2013).
>
> La última dimensión corresponde a dotación de la vivienda y es construida a partir de la información de la tenencia o no de ciertos elementos que podrían ser importantes o marcar la diferencia entre las condiciones socioeconómicas de vivienda de un estudiante u otro. (Cuéllar Caicedo et al. 2016, p. 97).

En el mismo orden de ideas, Timarán Pereira et al. (2019), clasifican las variables evaluadas en la prueba Saber 11 en consideración a aspectos socioecónomicos, académicos e institucionales. En los aspectos socioeconómicos contemplan las variables género, edad, estrato, nivel de sisben, ingreso familiar, educación de los padres, ocupación laboral de los progenitores, adquisición de automóvil, condiciones de la vivienda, facilidades de acceso a herramientas tecnológicas y condición de hacinamiento en casa. En los aspectos académicos evalúan el puntaje global obtenido por los estudiantes en la prueba Saber 11. Y en los aspectos institucionales analizan el tipo de institución educativa (oficial o privado), la jornada de estudio y la zona de ubicación geográfica del colegio.

En sentido similar, Rodríguez Rosero et al. (2021) categorizaron a las variables de las pruebas Saber 11 mediante los criterios de género, naturaleza del colegio, área de ubicación, estrato socioeconómico, escolaridad de los padres, facilidades de computador e internet en las viviendas, y número de horas que trabaja el estudiante entre semana. Y por último Blanco Villafañe (2015) clasifica a las variables mediante: datos sociodemográficos del estudiante y su familia, datos de la institución educativa, resultados de la prueba aplicada. En los datos sociodemográficos acoge a la edad, género, número de veces que el educando ha presentado el examen, estrato, condición laboral del estudiante, nivel educativo de los padres, ingreso familiar, nivel de Sisbén, número de habitaciones y de personas en el hogar. En los datos institucionales agrupa al tipo de calendario, género del colegio, naturaleza (oficial o privado), jornada, estado bilingüe, modalidad de egreso y valor de la pensión. Y en los resultados de la prueba analiza los puntajes obtenidos en lenguaje, matemáticas y ciencias sociales.

De lo anterior se considera que, la forma en la que se conceptualicen y agrupen las variables de la prueba Saber 11 permitirá fijar puntos de observación para el análisis e interpretación de los resultados obtenidos, de allí que agrupar las variables en torno a intereses sociodemográficos tal como se planteó en los objetivos del presente trabajo se consideró pertinente, puesto que rescató los principales aportes de los trabajos anteriormente citados sobre la prueba Saber y permitió visualizar de mejor manera el comportamiento de los atributos estudiados por el ICFES.

- **Consideraciones sobre las competencias en matemáticas.**

El concepto de competencias en educación define en sí mismo un amplio campo de estudio, y del cual autores como Delors (1996), Tobón Tobón et al. (2010), Tobón (2013), entre otros; han presentado variados aportes al respecto tanto en su conceptualización como en el

desarrollo y evaluación de estas en las aulas. Ahora bien, la intención en este apartado es presentar de manera somera a las competencias enfatizando en la manera en que el ICFES y el MEN han establecido el dominio de competencias en matemáticas, es así como, y con intención de brindar una aproximación inicial al lector sobre este concepto, que se expone el punto de vista de Delors, para quien las competencias representan:

> Una combinación de atributos con respecto al conocer y comprender (conocimiento teórico de un campo académico); el Saber cómo actuar (la aplicación práctica y operativa a base del conocimiento); y al Saber cómo ser (valores como parte integrante de la forma de percibir a los otros y vivir en un contexto). Este nuevo enfoque, además de no centrarse exclusivamente en los contenidos teóricos de un área del conocimiento, tiene una ventaja adicional que consiste en determinar las metas a lograrse en la formación de un profesional, es decir, <<el qué>> y dejar en libertad el <<cómo>>, primordial en el ambiente universitario de autonomías académicas (Delors, 1996, p. 25).

Desde esta postura, en el análisis de competencias es importante el desarrollo de las dimensiones del Saber, Hacer, Ser y Convivir del estudiante con su entorno, aspectos que no son ajenos al área de matemáticas puesto que se involucran en la noción de ser matemáticamente competente, y que llaman la atención en este trabajo ya que la prueba Saber 11 verifica en los colegiales el dominio de estas dimensiones. Es así como entonces el MEN, tras promover la redefinición de los fines de la educación matemática en el país y luego de considerar el interés de mejorar el desempeño de los estudiantes en las pruebas nacionales e internacionales, ha expuesto que una persona se considera matemáticamente competente cuando logra: *(1) formular y resolver problemas, (2) modelar procesos y fenómenos de la realidad, (3) comunicar, (4) razonar, y, (5) comparar y ejercitar procedimientos o algoritmos* (MEN, 2006).

De manera similar, LLECE (como se citó en ICFES, 2019, p. 11) propone que hablar de competencias en matemáticas implica estudiar:

> La capacidad de administrar nociones, representaciones y utilizar procedimientos matemáticos para comprender e interpretar el mundo real. Esto es, que el alumno tenga la posibilidad de matematizar el mundo real, lo que implica interpretar datos; establecer relaciones y conexiones; poner en juego conceptos matemáticos; analizar regularidades; establecer patrones de cambio; encontrar, elaborar, diseñar y/o construir modelos; argumentar; justificar; comunicar procedimientos y resultados… Llece (como se citó en ICFES, 2019, p. 11).

En apoyo a estas ideas, Roig & Llinares citados en ICFES (2019) exponen que:

> La competencia en matemática se vincula a una componente práctica relacionada con la capacidad que tiene una persona para hacer algo en particular, y también saber cuándo, y por qué utilizar determinados instrumentos. Se pueden considerar diferentes dimensiones del concepto de competencia matemática: comprensión conceptual de nociones matemáticas, desarrollo de destrezas procedimentales de carácter general, pensamiento estratégico. Roig & Llinares (como se citó en ICFES, 2019, p. 19).

Por último, el ICFES (2019) plantea que la competencia matemática da cuenta de:

> La relación entre el uso flexible y comprensivo del conocimiento matemático escolar y la diversidad de contextos, de la vida diaria, de la matemática misma y de otras ciencias. Este uso se evidencia, entre otros, en la capacidad del individuo para analizar, razonar y

> comunicar ideas efectivamente y para formular, resolver e interpretar problemas. (ICFES, 2019).

A este punto cabe señalar que una limitante de la base de datos de las pruebas Saber 11 en observación, radica en no contemplar con detalle la actuación del educando sobre cada ítem evaluado, es decir, no muestra sus aciertos o fallas en cada uno de los cinco procesos propuestos por (MEN, 2006) en la definición de ser matemáticamente competente y que toman rol primordial en la alfabetización matemática. Sin embargo, se conoce que la estructura de las preguntas contempla dos aspectos, por un lado, evalúa el ámbito cognitivo y por el otro el procedimental, siendo este último el que da cuenta del uso por parte de las personas del saber matemático en el contexto. En el texto del MEN, estos hechos se presentan de la siguiente manera:

> En el conocimiento matemático también se han distinguido dos tipos básicos: el conocimiento conceptual y el conocimiento procedimental. El primero está más cercano a la reflexión y se caracteriza por ser un conocimiento teórico, producido por la actividad cognitiva, muy rico en relaciones entre sus componentes y con otros conocimientos; tiene un carácter declarativo y se asocia con el saber qué y el saber por qué. Por su parte, el procedimental está más cercano a la acción y se relaciona con las técnicas y las estrategias para representar conceptos y para transformar dichas representaciones; con las habilidades y destrezas para elaborar, comparar y ejercitar algoritmos y para argumentar convincentemente. El conocimiento procedimental ayuda a la construcción y refinamiento del conocimiento conceptual y permite el uso eficaz, flexible y en contexto de los conceptos, proposiciones, teorías y modelos matemáticos; por tanto, está asociado con el saber cómo. (MEN, 2006, p. 50).

La prueba de matemáticas, según señala ICFES (2019) se compone aproximadamente de 50 preguntas, de las cuales el 34% dan cuenta de la competencia de interpretación y representación, el 43% de formulación y ejecución, y el 23% de argumentación. Los puntajes obtenidos pueden ser leídos con respecto a cuatro niveles de desempeño distinguidos entre los rangos de 0 a 35 puntos para el primer grupo, de 36 a 50 para el segundo, de 51 a 70 para el tercero y de 71 a 100 para el cuarto.

A manera de presentar consideraciones finales sobre las competencias se tiene que, en resumen, ellas mismas constituyen un campo de estudio amplio en donde se involucran aspectos sociodemográficos, actitudinales y aptitudinales. La formación de competencias en el aula no es una tarea fácil, requiere compromiso del estudiante y de su docente, puesto que tal como señala Delors (1996), el rol del profesor toma un papel relevante en esta etapa, ya que es el encargado de generar en el discente curiosidad e interés hacia el estudio. En igual forma se señala que un campo aún por explorar en las competencias es el de la evaluación y la forma en cómo estas son sometidas a juicio en las instancias educativas. No es fácil evaluar competencias porque ellas mismas requieren tiempo para su desarrollo, pero su medición ha permitido hacer frente a la nueva oleada de retos pedagógicos propuestos en el ámbito educativo. A decir de Moreno Olivos:

> Habrá que considerar que la evaluación de las competencias siempre será una aproximación al grado de dominio alcanzado en un momento determinado y de ninguna manera una medición exacta de su consecución por parte del alumnado. Además, como las competencias requieren tiempo para su desarrollo y maduración, lo más probable es que el dominio pleno de algunas de ellas en realidad se logre fuera del contexto de la escuela, en otro momento posterior y lejos de la mirada del profesor/evaluador, pues será en escenarios de la vida real –en situaciones inéditas o poco convencionales–

> cuando el alumno realmente pueda probar el dominio que posee de las competencias que la escuela intentó promover mediante el proceso formativo. No cabe duda que a los estudiantes del siglo XXI les ha tocado afrontar una educación cada vez más competitiva y desafiante. Por el bien de las nuevas generaciones, los educadores tenemos el compromiso ético de mejorar la enseñanza, el aprendizaje y la evaluación. A pesar de las múltiples críticas que el enfoque de competencias en educación ha recibido, algunas con razón y otras sin fundamento, la formación por competencias puede ser una posibilidad real de cambio y no mera retórica. (Moreno Olivos, 2012, p. 18).

Es así como la formación de competencias ha conllevado un reto pedagógico y didáctico en Colombia, tal interés evidenciado por el Gobierno Nacional ha dado paso al desarrollo de producciones investigativas en torno a estos temas, y sobre todo donde se conjuguen los conceptos de competencias, rendimiento académico y aspectos sociodemográficos del discente, no sin antes recordar que todos ellos se hallan estrechamente interrelacionados, y que el mal andamiaje de uno de ellos puede provocar el desánimo escolar, la pérdida del gusto hacia el estudio y la desilusión por parte del estudiante al no encajar en lugares que no valoren su ser cultural, intelectual y social. De allí que Bravo Salinas (2004) manifestara:

> Esto es lo que significa afirmar que no hay sujeto sin competencia, ni seres sin personalidad. Simplemente, hay seres humanos que no encontraron en la escuela y posiblemente desde la familia y la sociedad, los espacios culturales y formativos que le permitieran el despliegue de sus capacidades intelectuales, sus gustos y afectos, sentir la emoción del descubrimiento y el amor. (Bravo Salinas, 2004, p. 28).

1.7.3 Minería de datos.

Desde hace un tiempo en la comunidad de analistas de grandes bases de datos y tras la evolución del conocimiento tecnológico, se ha venido suscitando un campo de estudio en donde es posible escudriñar información que sirva de base para la construcción de información válida y útil para el investigador, dicho campo emergente se denomina *Minería de datos*, reconocida mayormente por sus siglas del inglés DM (Data Mining).

Para Beltrán Martínez (s.f, p. 5), la minería de datos es una disciplina que "… estudia métodos y algoritmos que permiten la extracción automática de información sintetizada que permite caracterizar las relaciones escondidas en la gran cantidad de datos", o visto de mejor manera como aquella que "… descubre relaciones, tendencias, desviaciones, comportamientos atípicos, patrones y trayectorias ocultas, con el propósito de soportar los procesos de toma de decisiones con mayor conocimiento." (p. 18). En el mismo sentido, Riquelme Santos et al. (2006, p. 15), la conciben "… como la construcción de un modelo que ajustado a unos datos proporciona un conocimiento" y en donde se hace hincapié en tres aspectos: (1) la escalabilidad del número de atributos y de instancias, (2) los algoritmos y arquitecturas, y (3) la automatización para manejar grandes volúmenes de información. Para Vargas Agurto (2014), la DM refiere al:

> Proceso que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores aparentemente caóticos que tienen una explicación que pueden descubrirse mediante diversas técnicas de esta herramienta. (Vargas Agurto, 2014, p. 12).

En esta etapa es importante reconocer y diferenciar términos que son de uso común en el trabajo con DM y que corresponden a los conceptos de datos, información y conocimiento. Es así como el Equipo editorial Etecé (2020) plantea que:

> Un dato es la representación de una variable que puede ser cuantitativa o cualitativa que **indica un valor que se le asigna a las cosas y se representa a través de una secuencia de símbolos, números o letras.**
>
> Los datos describen hechos empíricos. Para examinarlos deben ser organizados o tabulados, ya que un dato por sí mismo no puede demostrar demasiado. (Etecé, 2020).

En este sentido los datos pueden ser de tipo numérico (entero o real), texto (carácter o cadena), o, lógico (booleano), y se diferencian de *información* en el sentido que un dato es un carácter en bruto, el cual puede tener o no contexto y no ha sufrido proceso alguno de transformación, mientras que la *información*, y a decir de Vargas Agurto (2014):

> Puede definirse como un mensaje significativo que se transmite de la fuente a los usuarios, es la expresión material del conocimiento con fines de uso…La información para que sea utilizable debe tener tres características básicas: completa, confiable y oportuna. Una información completa debe contar con los elementos necesarios para que pueda ser analizada y procesada; confiable, debe provenir de una fuente veraz y creíble; oportuna, debe llegar a tiempo para su empleo. (Vargas Agurto, 2014, p. 7).

De otra parte, se considera conocimiento al proceso que conlleva el análisis, interpretación y comprensión de la información, razón por la cual se ubica en un estadio superior a los dos anteriores. *Conocimiento* es lo que se deduce de la información y genera

utilidad para el investigador, de allí que los datos al ser procesados faciliten la existencia de información la cual promueve conocimiento.

Convenir en estas diferencias es importante porque la minería de datos forma parte de un proceso de mayor envergadura conocido como KDD (Knowledge Discovery in Databases) y en el que según Vargas Agurto (2014), es importante determinar las fuentes de información útiles para crear un almacén de datos (Data Warehouse) que facilite los procesos de selección de datos objetivos, su limpieza, transformación, integración y reducción, así como la detección de valores ruido o anómalos, que posteriormente contribuyan al descubrimiento de conocimiento para el investigador. La DM es un proceso dentro del KDD, la cual facilita la evaluación e interpretación de los patrones extraídos y que conducen a la obtención de conocimiento relevante. La DM puede ser directa (cuando se sabe claramente lo que se busca), o indirecta (cuando no se sabe lo que se busca).

Agregado a lo anterior UNAYTA (2019) y Núñez Cárdenas (s.f), establecen que una forma de abordar acciones con DM implica, definir el problema e identificar los datos necesarios que servirán de base para elaborar un proceso de modelamiento, en el cual suele ser importante ejecutar un proceso de entrenamiento y prueba que facilite la validación e interpretación del conocimiento obtenido. Es importante considerar que, al trabajar en minería de datos suele ser conveniente realizar una imputación de los mismos.

- **Imputación de datos.**

Medina & Galván (2007) proponen en su texto que el proceso de imputar datos se puede realizar mediante diferentes maneras como, por ejemplo, analizar datos completos (*listwise o case deletion LD*), analizar datos disponibles (*pairwise o available case AC*),

reponderar, imputar de forma simple, imputar por regresión, usar el método de máxima verosimilitud y la imputación múltiple, entre otros.

El método *listwise* tiene como particularidad que omite las observaciones faltantes puesto que asume que ellas siguen un patrón definido, hecho por el cual trabaja solamente con las filas que tienen registros completos, provocando así la aparición de sesgos en los coeficientes de correlación y asociación, así como dificultades en el trabajo con muestras probabilísticas. El método *pairwise* asume un patrón MCAR (Missing Completely at Random) en los datos omitidos. En este punto conviene tener en cuenta que seguir un patrón MCAR quiere decir que cualquier observación tiene la misma probabilidad de ser considerada como dato perdido, por tanto, la ausencia de una observación no depende del comportamiento de las variables en estudio, o sea, los valores faltantes en Y no dependen ni de la variable Y ni de X. Así entonces, y según Araneda (2021), la probabilidad de ausencia de los datos es igual para todos los casos estudiados con lo que se establece que los datos perdidos tienen un comportamiento aleatorio. Desde la teoría, el patrón MCAR se visualiza como el estado ideal, ya que muestra que no existe sesgo en la pérdida de información.

Continuando con las ideas de Medina & Galván (2007), el método de *reponderación* busca compensar la falta de respuesta, ponderando los registros que permanecen en la muestra a través de modelos de probabilidad con información completa y datos exógenos, buscando con ello obtener estimadores robustos.

La imputación simple puede abordarse mediante los métodos de medias no condicionadas, medias condicionadas, variables ficticias, distribución no condicionada, regresión, máxima verosimilitud; de los cuales grosso modo se detalla que las medias no condicionadas asumen que los datos faltantes siguen un patrón MCAR, su uso provoca

intervalos de confianza más estrechos pero no es tan aconsejable puesto que al usar la media como reemplazo de los valores faltantes se generan problemas de subestimar la varianza y afectaciones en la correlación de las variables y en la distribución de los datos.

Imputar por medias condicionadas implica formar grupos a partir de variables correlacionadas con la variable en estudio, necesita que los datos perdidos sigan un patrón MCAR y provoca menos sesgo que el caso anterior. La imputación por variables ficticias implica crear un indicador que ayude a identificar los valores faltantes, sin embargo, este método genera inconsistencias en la capacidad para explicar a los estimadores. La imputación mediante distribución no condicionada favorece el uso del método *hot – deck*, en donde se llenan los registros vacíos (receptores) con información de campos completos (donantes), los datos faltantes se llenan a partir de una sección aleatoria para no introducir sesgos. El algoritmo *hot – deck* presenta variantes como el algoritmo secuencial y el método aleatorio.

Otro de los métodos de imputación simple es el de regresión, se usa cuando hay patrones y variables correlacionadas, acá se eliminan las observaciones con datos incompletos y se ajusta una ecuación de regresión para la respuesta. Una de sus desventajas es que no es útil para análisis de varianza o de correlación. El método de máxima verosimilitud estima cada parámetro del modelo con la función de máxima verosimilitud con el fin de predecir los valores omitidos, este proceso se repite hasta lograr convergencia en los parámetros.

La imputación múltiple utiliza métodos de simulación para el tratamiento de los datos, asume un patrón de datos faltantes aleatorio de tipo MAR (Missing at Random), lo que indica que la ausencia de datos no es aleatoria. Para Araneda (2021), el método MAR indica que la ausencia de una observación no se relaciona con los valores de su propia variable, sino que depende de los valores de las otras variables en observación, así entonces, dadas las variables

Y, X y Z, la información ausente en Y se halla relacionada con los valores obtenidos para X y Z, de allí que el método MAR requiera correlación alta entre la variable a imputar y el vector de variables modelo. Un ejemplo del método MAR acontece al preguntar la edad a un grupo de personas, ya que se esperaría que los hombres manifiesten su edad abiertamente mientras que las mujeres eviten esa pregunta frente a un grupo de hombres, así entonces, los valores faltantes de la edad de las mujeres son condicionados por la presencia de hombres.

Medina & Galvan (2007) concluyen en su texto que no hay un método de imputación que se pueda disponer como válido para todos los casos, la selección del mismo depende de los datos que se trabajen, sin embargo, si la elección se da en consideración a métodos estadísticos entonces cualquier forma de imputación analizada puede resultar válida, si en cambio se recurre a la teoría es probable que se recomiende el uso de la imputación múltiple o de la máxima verosimilitud, si se tiene en cuenta la forma de distribución entonces sobresale el método hot-deck y para el caso de estimaciones de índices de pobreza se usan criterios de sensibilidad.

Por otra parte, y ahondando ideas sobre minería de datos, Beltrán Martínez (s.f), señala que, al conformar el banco de base de datos en el KDD, pueden encontrarse dos grupos de investigadores, el primero refiere a los *picapedreros o granjeros*, quienes son los encargados de realizar informes periódicos de los datos en observación, reflexionan sobre la evolución de los parámetros y supervisan los valores anómalos. En el otro grupo se hallan los *exploradores*, quienes buscan nuevos patrones significativos mediante las técnicas de minerías de datos. De igual manera Beltrán Martínez indica que en la selección, limpieza y transformación de los datos, los histogramas y los boxplot juegan un rol particular puesto que permiten identificar datos anómalos, outliers o valores faltantes, los cuales deberán tratarse en consideración a parámetros previamente establecidos por el investigador y de acuerdo al sustento estadístico

que los involucra. Es así como los investigadores en minería de datos señalan que una forma plausible para el acceso a los datos se obtiene al hacer uso de la metodología CRISP – DM (Cross-Industry Standard Process for Data Mining).

- **Metodología CRISP – DM.**

Wirth & Hipp (2000) señalan en su texto que una de las dificultades que subyace en la minería de datos implica que esta no tiene un método estándar que le permita actuar sobre los datos, suscitando con ello posibles actuaciones erróneas sobre el tratamiento de la información por parte del investigador sobre todo cuando no se es perito en este contexto, por ende que para dichos autores el uso de la metodología CRISP – DM surja como importante ya que plantea una estructura automatizada para el tratamiento de datos a través de fases, tareas genéricas, tareas especializadas e instancias de procesos. Agregado a ello, la metodología CRISP – DM diferencia entre *modelo de referencia* y *guía de usuario*, siendo que la tarea del modelo consiste en brindar una descripción general de fases, tareas, resultados y orienta sobre el qué hacer en un proyecto de DM. La guía del usuario brinda consejos y sugerencias a profundidad para cada tarea y fases de la DM, y otorga indicaciones del cómo hacer el proyecto.

El *modelo de referencia genérico*, útil para la planificación, documentación y comunicación, se fundamenta en seis etapas las cuales dan cuenta del entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación. En la fase *entendimiento del negocio* se analizan los objetivos y requisitos del proyecto desde su contexto original para, con base en ellos, diseñar un plan preliminar que facilite el *entendimiento de los datos*, fase en el cual se analiza la calidad de ellos y se elaboran hipótesis sobre información oculta. En la *preparación de los datos*, se elaboran tareas de elección y construcción de atributos, limpieza y transformación de datos. En el *modelado* se

asocia la técnica de DM más pertinente para la consecución del objetivo planteado y la cual se validará en la etapa de *evaluación*. El despliegue del modelo se ejecuta en la última fase e implica la generación de conocimiento o redacción de conclusiones las cuales dependen del usuario y no tanto del analista de datos.

- **Minería de datos educativa.**

Un campo de aplicación de la DM se da en el ámbito educativo, hecho que da origen a la minería de datos educativa o más conocida por sus siglas en inglés como EDM (*Educational Data Mining*). En vista que DM y EDM son similares (difieren solo en el contexto), en adelante se entenderá que al referirse a la DM en realidad se está tratando con la EDM puesto que el contexto abordado en el presente trabajo es meramente educativo. Cabe tener en cuenta también que la aplicación de la EDM acarrea consigo numerosas ventajas, las cuales se detallan a lo amplio de este texto con la exposición de algoritmos válidos para el tratamiento de la información, pero asimismo, y en voz de Usman & Atumoshi (2017), trae algunas desventajas como el no poder establecer un modelo único que permita predecir con exactitud del 100% los resultados académicos de un aprendiz, esto debido a la dimensionalidad de los datos, a las técnicas asociadas con el aprendizaje probabilístico y a la dificultad de mejorar la asignación docente por alumno en las instituciones escolares.

Dentro de las ventajas, Oviedo Carrascal & Jiménez Giraldo (2019) señalan que la EDM ha cobrado fuerza en las investigaciones abordadas en el contexto educativo, sirviendo de herramienta útil en la comprensión y análisis de los resultados de la prueba PISA, por ejemplo, ya que facilita la detección de la interrelación entre factores socioeconómicos y el desempeño con el área de lectura. La EDM contiene en sus haberes técnicas que facilitan el estudio de la información entre las que se destacan "… árboles de decisión, redes bayesianas, regresión, correlación y análisis por clústeres…, siendo el clustering la técnica más usada". (p. 129), y

razón por la cual Riquelme Santos et al. (2006), explicitan en su texto que las tareas más comunes de la DM involucran los procesos de clasificación, regresión, clustering, generación de reglas, resumen y análisis de secuencias, describiendo cada acto de la siguiente manera:

- Clasificación: clasifica un dato dentro de una de las clases categóricas predefinidas. Responde a preguntas tales como, ¿Cuál es el riesgo de conceder un crédito a este cliente? ¿Dado este nuevo paciente qué estado de la enfermedad indican sus análisis?
- Regresión: el propósito de este modelo es hacer corresponder un dato con un valor real de una variable. Responde a cuestiones como ¿Cuál es la previsión de ventas para el mes que viene? ¿De qué depende?
- Clustering: se refiere a la agrupación de registros, observaciones, o casos en clases de objetos similares. Un clúster es una colección de registros que son similares entre sí, y distintos a los registros de otro clúster. ¿Cuántos tipos de clientes vienen a mi negocio? ¿Qué perfiles de necesidades se dan en un cierto grupo de pacientes?
- Generación de reglas: aquí se extraen o generan reglas de los datos. Estas reglas hacen referencia al descubrimiento de relaciones de asociación y dependencias funcionales entre los diferentes atributos. ¿Cuánto debe valer este indicador en sangre para que un paciente se considere grave? ¿Si un cliente de un hipermercado compra pañales también compra cerveza?
- Resumen o sumarización: estos modelos proporcionan una descripción compacta de un subconjunto de datos. ¿Cuáles son las principales características de mis clientes?
- Análisis de secuencias: se modelan patrones secuenciales, como análisis de series temporales, secuencias de genes, etc. El objetivo es modelar los estados del

> proceso, o extraer e informar de la desviación y tendencias en el tiempo. ¿El consumo de energía eléctrica de este mes es similar al del año pasado? Dados los niveles de contaminación atmosférica de la última semana cuál es la previsión para las próximas 24 horas. (Riquelme Santos et al. 2006, p. 13).

Anexo a lo anterior, las técnicas utilizadas de DM pueden ser supervisadas o no supervisadas. Las técnicas supervisadas conocidas también como *predictivas* permiten la clasificación (clases disjuntas estimadas por funciones), categorización (clases no disjuntas estimadas por correspondencias), y regresión de elementos (Oviedo Carrascal & Jiménez Giraldo, 2019); en esta etapa los datos pueden recibir tratamientos de entrenamiento y de prueba. Beltrán Martínez (s. f) destaca que como técnicas predictivas se encuentran: K – NN (vecino más cercano), aprendizaje perceptrón, métodos multicapa ANN, funciones radiales básicas, árboles de decisión, clasificación de Bayes, métodos de división central, reglas (CN2), pseudo – relacionales, Pick – and – Mix, relacional: ILP, IFLP, SCIL.

Las técnicas no supervisadas también llamadas *descriptivas* o *de descubrimiento*, permiten visualizar regularidades entre datos y detectar rasgos característicos de ellos mediante asociaciones, clustering y selección de factores. Acá los datos pueden ser analizados desde posturas exploratorias o de segmentación. Como técnicas exploratorias aparecen los estudios correlacionales, las dependencias, detección datos anómalos y análisis de dispersión. En las técnicas de segmentación o clustering aparecen: K – means, K – modas, redes neuronales de Kohonen, EM (estimación de medias) y la autoclasificación. Cabe decir que es probable que el lector en la revisión de textos encuentre técnicas compartidas por estos dos tipos de aprendizajes, ejemplo de ello lo constituyen las redes neuronales, sin embargo, debe diferenciar el fin con la que van a ser utilizadas, ya sea de manera predictiva o de descubrimiento.

En sintonia con lo anterior, y con intención de exponer ideas sobre cómo y cuándo utilizar un algoritmo de DM en un contexto específico, Blanco Villafañe (2015) ha elaborado un cuadro comparativo a través del cual brinda sugerencias del uso de tales técnicas según el entorno educativo en donde se desenvuelva el trabajo, la población considerada y el objetivo a investigar; de esta forma dicho autor sugiere que si se desea por ejemplo, descubrir patrones pedagógicamente interesantes en relación a docentes, estudiantes y datos expuestos en la web, lo más conveniente es utilizar los algoritmos k – means para datos numéricos, k - modas para categóricos, árboles de decisión o reglas de asociación. En cambio, si se pretende estudiar las correlaciones de diversos patrones y su influencia en estudiantes universitarios, se sugiere usar redes neuronales, árboles de decisión o análisis discriminante.

En el trabajo de Blanco Villafañe se destaca también que la acción de categorizar estudiantes según el desempeño en matemáticas, puede ser guiada por los algoritmos de k – means, k - modas, árboles de decisión, reglas de asociación, redes neuronales o Bayes. De allí que, con ánimo de delimitar el campo de acción de la minería de datos, se ahonda esfuerzos en las técnicas de segmentación (clustering) y la correlación a través del análisis de factores para evaluar relación entre variables.

1.7.4 Análisis factorial.

Hair et al. (1999) presentan al análisis factorial como un conjunto de "… métodos estadísticos multivariantes cuyo propósito principal es definir la estructura subyacente en una matriz de datos" (p. 80). El análisis de factores pretende tambien resumir y reducir datos con mínima perdida de información y puede abordarse desde perspectivas exploratorias o confirmatorias. Las posturas exploratorias exponen al análisis factorial como "… útil para la búsqueda de una estructura entre una serie de variables o como un método de reducción de

datos” (p. 81), hecho que conlleva a trabajar con lo que proporcionan los datos sin aplicar restricciones a priori. Por otro lado, las posturas confirmatorias permiten al investigador probar hipótesis preconcebidas y con las cuales se evalúe el grado de ajuste de los datos a una estructura esperada.

El desarrollo de la presente investigación se orienta en consideración al análisis factorial exploratorio, por lo cual en seguida se amplían ideas respecto a este hecho.

Hair y comitiva señalan que el análisis factorial exploratorio se subdivide en tipo **R** y **Q**. Se llama **R** cuando se buscan dimensiones latentes no fácilmente observables a un grupo de variables. Se denomina **Q** cuando se estudia la matriz de correlaciones de individuos. El método **R** es el de uso más frecuente mientras que el **Q** suele ser reemplazado por el análisis clúster. Sin embargo, conviene decir que, la diferencia entre análisis tipo **Q** y clúster radica en que el primero se basa en las intercorrelaciones entre encuestados mientras que, en clúster, se forman grupos a partir de la distancia entre las puntuaciones de los individuos sobre las variables en estudio.

Agregado a lo anterior, Hair et al. (1999) señalan que para realizar un análisis de factores se debe tener en cuenta la matriz de correlación, la selección de variables y el tamaño de la muestra. Para la matriz de correlación se requiere que los datos sean numéricos. Al seleccionar variables se sugiere laborar con al menos cinco de ellas. En cuanto al tamaño muestral se recomienda acoger más de 50 observaciones, o, “… por lo menos un número de observaciones cinco veces mayor que el número de variables analizadas” (p.88), ya que para muestras menores el análisis pierde su significancia.

Como anotación especial se relata que la matriz de correlación debe reflejar que las variables se encuentren relacionadas, caso contrario el análisis factorial pierde importancia ya que la multicolinealidad entre las variables es uno de los requisitos a cumplir para el desarrollo del estudio factorial. En esta etapa surge como importe el análisis del *contraste de esfericidad de Bartlett,* el cual determina si las variables en observación presentan o no correlación bajo el sistema de hipótesis:

$$\begin{cases} H_0\colon\ La\ matriz\ de\ correlación\ es\ la\ identidad. \\ H_1\colon La\ matriz\ de\ correlación\ no\ es\ la\ identidad. \end{cases}$$

Si se verifica que la matriz es la identidad, entonces se infiere que los datos no están correlacionados, por tanto, en la aplicación del test de Bartlett se requiere el rechazo de la hipótesis nula (H_0). Agregado a lo anterior se resalta que, en el análisis de factores entran en juego tres tipos de varianza conocidas como: *(1) común, (2) específica (única),* y *(3) del error*. La varianza común es la varianza compartida por todas las variables en estudio y da origen al estudio de las *comunalidades* las cuales son estimaciones de la varianza común. La varianza específica es la que se relaciona con una sola variable en específico. La varianza del error es debida a la poca fiabilidad en la recolección de los datos. Así entonces, la relación entre la varianza total y las demás variabilidades puede ser modelada bajo la expresión

Varianza total = Variabilidad común + Variabilidad específica (Comunalidad) + varianza error.

Por otro lado, en el estudio del análisis de factores cabe diferenciar dos tipos de análisis por los cuales puede optar el investigador, los que conciernen a los métodos de componentes principales y de factor común, siendo que el primero se usa cuando se requiere explicar la

máxima varianza de las variables originales, mientras que la segunda se usa para identificar dimensiones latentes. Al respecto Hair et al. (1999) sostienen que:

> La selección de un modelo u otro se basa en dos criterios: (1) los objetivos del análisis factorial y (2) el grado de conocimiento anterior acerca de la varianza en las variables. El análisis de componentes principales es apropiado cuando el interés primordial se centra en la predicción o el mínimo número de factores necesarios para justificar la porción máxima de la varianza representada en la serie de variables original, y cuando el conocimiento previo sugiere que la varianza específica y de error representa una proporción relativamente pequeña de la varianza total. Por el contrario, cuando el objetivo principal es identificar las dimensiones latentes o las construcciones representadas en las variables originales y el investigador tiene poco conocimiento acerca de la varianza específica y de error y por tanto quiere eliminar esta varianza, lo más apropiado es utilizar el modelo factorial común. (Hair et al. 1999, p. 91).

Sin embargo, y a decir del mencionado Hair y su grupo, una debilidad del análisis factorial común, radica en que este método, a diferencia del de componentes principales, no exhibe soluciones únicas; además, las comunalidades presentadas en este análisis factorial no siempre pueden ser válidas.

Reforzando lo anterior, Bolaños (2020) plantea que el estudio de la matriz de correlación, la selección de variables y del tamaño de la muestra, se puede condensar en el desarrollo de 5 pasos, en donde se debe: (1) verificar que la matriz de datos sea factorizable, (2) extraer los factores, (3) determinar el número correcto de factores, (4) rotar los factores, (5) interpretar los resultados. Para el desarrollo del primer paso, la ejecución del test de Bartlett y la prueba de Kaiser-Meyer- Olkin (KMO) son relevantes. En la extracción de factores, Bolaños

propone el uso de métodos como análisis de componentes principales, mínimos cuadrados no ponderados, mínimos cuadrados generalizados, máxima verosimilitud, factorización de ejes principales, alfa y factorización imagen, los cuales describe así:

- **Análisis de componentes principales.** Método para la extracción de factores utilizada para formar combinaciones lineales no correlacionadas de las variables observadas. El primer componente tiene la varianza máxima. Las componentes sucesivas explican progresivamente proporciones menores de la varianza y no están correlacionadas unas con otras. El análisis principal de las componentes se utiliza para obtener la solución factorial inicial. No se puede utilizar cuando una matriz de correlaciones es singular.
- **Método de mínimos cuadrados no ponderados.** Método de extracción de factores que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlación observada y reproducida, ignorando las diagonales.
- **Método de mínimos cuadrados generalizados.** Método de extracción de factores que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlación observada y reproducida. Las correlaciones se ponderan por el inverso de su exclusividad, de manera que las variables que tengan un valor alto de exclusividad reciban una ponderación menor que aquéllas que tengan un valor bajo de exclusividad.
- **Método de máxima verosimilitud.** Método de extracción factorial que proporciona las estimaciones de los parámetros que con mayor probabilidad ha producido la matriz de correlaciones observada, si la muestra procede de una distribución normal multivariada. Las correlaciones se ponderan por el inverso de la exclusividad de las variables, y se emplea un algoritmo iterativo.
- **Factorización de ejes principales.** Método para la extracción de factores que parte de la matriz de correlaciones original con los cuadrados de los coeficientes de

correlación múltiple insertados en la diagonal principal como estimaciones iniciales de las comunalidades. Las cargas factoriales resultantes se utilizan para estimar de nuevo las comunalidades que reemplazan a las estimaciones previas de comunalidad en la diagonal. Las iteraciones continúan hasta que el cambio en las comunalidades, de una iteración a la siguiente, satisfaga el criterio de convergencia para la extracción.

- **Alfa.** Método de extracción factorial que considera a las variables incluidas en el análisis como una muestra del universo de las variables posibles. Este método maximiza el Alfa de Cronbach para los factores.
- **Factorización imagen.** Método para la extracción de factores, desarrollado por Guttman y basado en la teoría de las imágenes. La parte común de una variable, llamada la imagen parcial, se define como su regresión lineal sobre las restantes variables, en lugar de ser una función de los factores hipotéticos. (Bolaños, 2020).

- **Número de factores a extraer.**

Hair et al. (1999), señalan que para determinar el número de factores a extraer se pueden utilizar los criterios de raíz latente, a priori, porcentaje de la varianza, contraste de caída, los que grosso modo implican lo siguiente.

- **Raíz latente:** Es la técnica más utilizada y establece seleccionar variables con *raíces latentes o autovalores* superiores a la unidad.
- **A priori:** Posibilita que el analista defina con anterioridad el número de factores a extraer, aspecto útil cuando se requiere repetir un experimento.
- **Porcentaje de la varianza:** Busca obtener un porcentaje acumulado específico de la varianza total extraída. Hair y su grupo sugieren que para las ciencias naturales conviene

que los factores den cuente al menos del 95% de la varianza y en ciencias sociales sea por lo menos del 60%.

- **Contraste de caída:** Permite identificar el número de factores óptimos a extraer antes que la varianza única absorba a la varianza común.

Bolaños (2020) complementa lo anterior proponiendo el uso de los métodos: *Kaiser, análisis scree plot y análisis paralelo*, los cuales concibe de la siguiente manera:

- **Kaiser Criterion (Guttman, 1954):** esta regla sugiere que se deben retener todos los factores que tengan un eigenvalue de 1.0 o mayor; con el razonamiento de que un factor no debe explicar menos que la varianza equivalente que hubiera explicado una sola de las variables incluidas en el análisis. La regla sin embargo no es estricta y debe analizarse en conjunto con otros criterios.
- **Análisis del Scree Plot (Cattell, 1966):** este método complementa al anterior y se basa también el análisis de la magnitud de los eigenvalues pero a partir de la tendencia que se observa en el Scree Plot. Se procuran seleccionar un grupo reducido de factores que tengan eigenvalues significativamente superiores a los demás, para lo cual se identifica el punto de inflexión en la curva del scree plot (también referido como el codo por su semejanza con un brazo) a partir del cual la curva se transforma a una línea “plana” o relativamente recta.
- **Análisis paralelo (Horn, 1965):** Esta regla suele complementar las anteriores cuando el número de variables iniciales y factores resultantes es elevado. El procedimiento es basado en el principio de que los factores a extraer deben dar cuenta de más varianza que la que es esperada de manera aleatoria. El procedimiento reordena las observaciones de manera aleatoria entre cada variable y los eigenvalues son recalculados a partir de esta nueva base de datos aleatoriamente ordenada. Los

factores con eigenvalues mayores a los valores aleatorios son retenidos para interpretación. (Bolaños, 2020).

Una vez extraídos los factores se debe analizar *las cargas factoriales* las cuales dan cuenta de las correlaciones entre cada variable y el factor. Así entonces, en cada factor se retienen las variables con cargas factoriales más altas y si ellas no fueren fácilmente leíbles se sugiere elaborar una rotación de factores. Hair et al. (1999), señalan que las cargas de factores superiores a ± 0,30 se ubican en nivel mínimo, las cargas de ± 0,40 cobran mayor importancia y las cargas de ± 0,50, o, mayores son significativas.

- **Rotación de la matriz.**

Se sugiere rotar la matriz de cargas original cuando se desea visualizar, desde otro ángulo, las cargas de las variables sobre los ejes factoriales, entendiendo que se busca con ello "…redistribuir la varianza de los primeros factores a los últimos para lograr un patrón de factores más simple y teóricamente más significativo" (Hair et al. 1999, p. 95). Dicha rotación puede ser *ortogonal u oblicua*. Las rotaciones ortogonales, las cuales son de uso más frecuente, giran los ejes conservando un ángulo de 90 grados. En el trabajo con rotaciones ortogonales emergen los métodos *varimax, quartimax, y equimax.* Como métodos de rotaciones oblicuas se encuentran la *oblimin y promax*. Bolaños (2020), teoriza a estas rotaciones de la siguiente manera:

- **Varimax:** Método de rotación ortogonal que minimiza el número de variables que tienen saturaciones altas en cada factor. Simplifica la interpretación de los factores.
- **Criterio oblimin directo:** Método para la rotación oblicua (no ortogonal). El método necesita un valor delta que servirá para ajustar los ejes en función de las

saturaciones buscando una mejor aproximación, pero considerando que la varianza se distribuirá entre todos los factores.

- **Método quartimax:** Método de rotación que minimiza el número de factores necesarios para explicar cada variable.
- **Método equimax:** Método de rotación que es combinación del método varimax, que simplifica los factores, y el método quartimax, que simplifica las variables. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable.
- **Rotación promax:** Rotación oblicua que permite que los factores estén correlacionados. Esta rotación se puede calcular más rápidamente que una rotación oblimin directa, por lo que es útil para conjuntos de datos grandes. (Bolaños, 2020).

Además, y a decir de Peña (2002), el método varimax permite "…maximizar la varianza de los coeficientes que definen los efectos de cada factor sobre las variables observadas" (p. 379) y se postula como la rotación más usualmente utilizada. El método quartimax, como indica Hair et al. (1999) "…no ha demostrado gran capacidad para generar estructuras más simples" (p. 98) por esa razón en este trabajo se hace uso de varimax. Asimismo, Hair y su grupo muestran que una de las desventajas de las rotaciones oblicuas es que "… permiten la existencia de factores correlacionados en lugar de mantener la independencia" (p. 98) entre ellos, hecho que requiere de mayor experticia para el investigador en el análisis de las cargas factoriales.

1.7.5 Análisis clúster.

Hair et al. (1999), conceptualizan al análisis clúster "… como un grupo de técnicas multivariantes cuyo principal propósito es agrupar objetos basándose en las características que poseen" (p. 492), con la particularidad que en los conglomerados resultantes debe existir alta

homogeneidad entre sus observaciones (dentro de cada clúster) y alta heterogeneidad entre los grupos formados. El concepto de valor teórico se define por el conjunto de variables que determinan el carácter de los objetos, siendo dicho valor definido por el investigador, puesto que "… El objetivo del análisis clúster es la comparación de objetos basándose en el valor teórico, no en la estimación del valor teórico en sí misma". En esta instancia cabe diferenciar entre análisis clúster y factorial, puesto que por su estructura similar es posible que el lector los pueda considerar igual, sin embargo, el primero se encarga con mayor detalle del agrupamiento de objetos (individuos), mientras que el segundo lo hace aludiendo a las variables.

Una de las ventajas de trabajar con análisis clúster radica en que facilita la clasificación de individuos mediante conglomerados, hecho acogido por variadas disciplinas de estudio y con las que han logrado hacer taxonomías de individuos. De igual forma permite elaborar una simplificación de los datos y una identificación de patrones; empero, entre sus mayores desventajas se halla que puede clasificarse como descriptivo, no inferencial, de múltiples soluciones, que pretende formar grupos aun teniendo evidencia clara de una auténtica estructura de los datos y que es una técnica sensible a valores atípicos. Hair y su grupo señalan:

> Sin embargo, junto a los beneficios del análisis *clúster* existen algunos inconvenientes. El análisis *clúster* puede caracterizarse como descriptivo, ateórico y no inferencial. El análisis *clúster* no tiene bases estadísticas sobre las cuales deducir inferencias estadísticas para una población a partir de una muestra, y se utiliza fundamentalmente como una técnica exploratoria. Las soluciones no son únicas, en la medida en que la pertenencia al conglomerado para cualquier número de soluciones depende de muchos

> elementos del procedimiento… el análisis *clúster* siempre creará conglomerados a pesar de una <<auténtica>> estructura en los datos. (Hair et al. 1999, p. 493).

Estos autores señalan también que en el análisis clúster se debe tener en cuenta que, si los datos son métricos, pueden ser tratados desde un enfoque de proximidad o de patrones. Como medidores de proximidad se encuentran las distancias euclídeas, city – block y Mahalanobis. Con respecto a la verificación de patrones toman relevancia los coeficientes de correlación. Por otro lado, si los datos no son métricos, se utilizan los coeficientes emparejados como medidas de asociación de similitud. Asimismo, se debe considerar que los algoritmos dados en los clústeres pueden ser *jerárquicos, no jerárquicos, mixtos.* Los métodos jerárquicos se clasifican en *aglomerativos* y *divisivos,* y para la conformación de sus conglomerados se pueden usar algoritmos como los métodos de encadenamiento simple, encadenamiento completo, encadenamiento medio, Ward y de centroide. Todos ellos pueden ser representados por dendrogramas.

El método del encadenamiento simple se basa en la distancia mínima, es decir, detecta la pareja de objetos separados por la distancia más corta para proceder a agruparlos (vecino más cercano). El encadenamiento completo realiza un proceso similar al simple, con la salvedad que ahora la distancia debe ser máxima (vecino más lejano). El encadenamiento medio tiene en cuenta la distancia media de todas las observaciones de un clúster con todos los individuos de otro. El método de Ward considera la distancia entre dos grupos como la sumatoria elevada al cuadrado entre dos conglomerados. En el método del centroide se trabaja con la distancia euclidiana como herramienta de medición, se ve menos afectado por valores atípicos, pero su desventaja es que el centro de gravedad se recalcula cuando ingresa una nueva observación al clúster.

En los métodos no jerárquicos no se construyen dendrogramas y parten del hecho que el número de agrupaciones a realizar ya es conocido con anterioridad. La selección de puntos semillas como centro de los conglomerados es necesaria en este tipo de métodos. En esta etapa surgen como importante los algoritmos: K – means (datos numéricos), K – modas (datos categóricos) y K – prototipos (datos mixtos); en donde la asignación de individuos se realiza en consideración a los métodos de umbral secuencial, umbral paralelo, optimización para k - means, y obtención de modas en k – modas.

Santamaría Ruiz (2006) propone la clasificación de clúster en 5 grupos que se caracterizan por ser: (1) bien separados, (2) basados en el centro, (3) contiguos, (4) basados en densidad, (5) de propiedad o conceptual. Cada clúster lo asume de la siguiente manera:

- Clústeres bien separados. Esta definición idealista parte del hecho que todos los objetos de un grupo deben ser suficientemente similares.
- Clústeres basados en el centro. Un clúster es un conjunto de objetos en el que un objeto está más cerca al centro del clúster, que al centro de otro clúster.
- Clústeres contiguos. Un clúster es un conjunto de puntos, donde un punto en el clúster está más próximo a otro punto o puntos del clúster, que a cualquier otro punto que no pertenezca al clúster.
- Clústeres basados en densidad. Este tipo de agrupamiento, se basa en el hecho de tener grupos en regiones de alta densidad, separados por regiones de baja densidad.
- Clúster de propiedad o Conceptual. Son clústeres que tienen propiedad compartida o representan un concepto particular, es decir, hay puntos en común entre dos grupos. (Santamaría Ruiz, 2006, p. 4).

- **Datos relevantes sobre k – means (clústeres cuantitativos).**

Ahondando en lo anterior y esgrimiendo ideas sobre k – means, cabe decir que Ochoa et al. (2017), lo presentan como técnica basada en el clúster particional y útil para el tratamiento de variables cuantitativas, en donde se "… utiliza el promedio para representar los centros de los clústeres" (p.5). Agregado a ello, Peña (2002), expone que en la labor con k – means conviene aplicar 4 etapas en donde se deben: (1) seleccionar los centros de los grupos iniciales, (2) calcular las distancias euclídeas, (3) definir un criterio de optimización, y, (4) terminar el proceso. En palabras de Peña.

> El algoritmo de k-medias (que con nuestra notación deberá ser de G−medias) requiere las cuatro etapas siguientes:
>
> (1) Seleccionar G puntos como centros de los grupos iniciales. Esto puede hacerse:
>
> a) asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos formados;
>
> b) tomando como centros los G puntos más alejados entre sí
>
> c) construyendo los grupos con información a priori, o bien seleccionando los centros a priori.
>
> (2) Calcular las distancias euclídeas de cada elemento al centro de los G grupos, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.
>
> (3) Definir un criterio de optimalidad y comprobar si reasignando uno a uno cada elemento de un grupo a otro, mejora el criterio.
>
> (4) Si no es posible mejorar el criterio de optimalidad, terminar el proceso. (Peña, 2002, p. 228).

Complementando lo precedente, Zhexue Huang (1998) expone que dado un conjunto de objetos numéricos X y un valor entero K, el k – means busca particionar los X objetos en K grupos en donde se minimice la suma de errores al cuadrado dentro de cada conglomerado, hecho que implica minimizar la función de costo P definida por:

$$Minimizar\ P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}\ d(X_i, Q_l)$$

$$Sujeto\ a \sum_{l=1}^{k} w_{i,l} = 1, \quad 1 \leq i \leq n$$

$$w_{i,l} \in \{0,1\}, \qquad 1 \leq i \leq n, \quad 1 \leq l \leq k \quad (Traducción\ de\ \text{Zhexue Huang, 1998, p. 287}).$$

Donde W es una partición de la matriz Q = {Q_1, Q_2, … Q_k} de un conjunto de objetos de igual dominio.

Por otro lado, Pastrán Ramírez & Roa Peña (2015) refuerzan estas ideas al exponer que, en la conformación de conglomerados con datos númericos emergen como importante el análisis de las medidas de similiaridad (similitud) y disimilaridad (distancias) entre individuos, en donde juegan papel protagónico el coeficiente de correlación para las similitudes y las distancias euclidiana y de mahalanobis. En las similaridades se cumple la simetría y la no negatividad. En las distancias se cumplen la no negatividad, la simetría y la desigualdad triangular.

- **Datos relevantes sobre k – modas (clústeres cualitativos).**

El trabajo realizado por Zhexue Huan (1998) ha servido como guía para la agrupación de datos cualitativos en donde el algoritmo k – means no es efectivo. Es por ello que mencionado autor propone la agrupación de variables categóricas por medio del método denominado k – modas, y la agrupación de datos mixtos por medio del algoritmo k – prototipos.

En esta sección se aborda el método k – modas, puesto que uno de los intereses que persigue el desarrollo de la presente investigación es conformar clústeres con variables cualitativas.

Antes de presentar el algoritmo K – modas, conviene acudir a la conceptualización de variable categórica expuesta por Pastrán Ramírez & Roa Peña (2015) quienes la conciben como un conjunto de observaciones $\{C_1, C_2, \ldots, C_k\}$ donde cada C_i es una categoría o modalidad de la variable X. De igual forma mencionados autores, en conjunto con Rendón et al. (2015) y basados en los aportes de Zhexue Huang (1998), argumentan que en el estudio del k – modas conviene comprender los términos de medidas de disimilaridad (distancia), moda y frecuencia, puesto que ellos mismos juegan un rol importante en esta etapa. Es así entonces que, en aras de brindar una definición para las medidas de disimilaridad, Rendón y su grupo exponen que:

> Sean X, Y dos objetos categóricos descritos por m atributos categóricos. La medida de disimilaridad entre X y Y se define por el total de las no coincidencias de los atributos categóricos de los objetos. El número más pequeño de las diferencias significa que los objetos son similares (Zhexue., 1998).
>
> Formalmente:
>
> $$d(X,Y) = \sum_{j=1}^{m} \delta(X_j, Y_i)$$
>
> Donde
>
> $$\delta(X_j, Y_i) = \begin{cases} 0 \ (x_j = \ y_i) \\ 1 \ (x_j \neq \ y_i) \end{cases}$$
>
> $d(X,Y)$ da igual importancia a cada categoría del atributo. Si se toma en cuenta las frecuencias de las categorías en el conjunto de datos, se define la medida de disimilaridad como:

$$d_{x^2}(X,Y) = \sum_{j=1}^{m} \frac{n_{xj} + n_{yj}}{n_{xj}\, n_{yj}}\, \delta(X_j, Y_i)$$

(Rendón et al. 2015, p. 933).

En esta instancia cabe destacar también el aporte de Clavijo M. & Granada D., (2016) quienes para comprender de mejor manera la expresión $\delta(X_j, Y_i)$ explicitan que dados dos individuos x_j, y_i, el resultado 0 indica una semejanza máxima, es decir, son muy parecidos, mientras que un valor de 1 se sucede cuando hay diferencias entre ellos.

Con intención de clarificar la función de disimilaridad expuesta, se propone el análisis del siguiente ejemplo el cual toma como referencia lo expuesto por The Academician (2020).

Ejemplo: Sean los vectores X = (a, b, c); Y = (a, b, c); Z = (a, c, b); W = (m, n, p), analice las disimilaridades entre X y Y, X y Z, X y W.

La distancia d(X, Y) es 0, porque al analizar componente a componente los elementos de X y de Y, se tiene que contienen los mismos elementos, es decir $x_j = y_i$. Por tanto, la suma de sus distancias es 0.

La distancia d(X, Z) = 0 + 1 + 1 = 2. Ya que la primera componente en ambos vectores es “a”, y por ser igual tiene distancia cero. La segunda componente ya no es igual porque en X es “b”, y en *Z* es “c”, por lo que su distancia es 1. Cabe recordar que la disimilaridad indica que elementos diferentes tienen distancia 1, hecho que también se evidencia en la tercera componente. Para la distancia d(X, W) se observa que todos los elementos de X son distintos a los de W, luego se infiere que la suma de sus distancias sea 3.

El método de k – modas propuesto por Zhexue Huang (1998) es muy similar al algoritmo k – means, con la diferencia que el foco de observación no es la media sino la moda de un conjunto de datos, motivo que conlleva a trabajar con medidas de disimilitud entre atributos y el uso de frecuencias de los datos. En el k – modas se busca minimizar la función de costo dada por

$$P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,l}\,\delta\left(x_{i,j},\, q_{l,j}\right)$$

Donde $w_{i,l} \in W\ y\ Q_l = \left[q_{l,1}, q_{l,2}, \ldots, q_{l,m}\right] \in Q$. (Traducción de Zhexue Huang, 1998, p. 290).

Para minimizar la función de costos, el k – modas se fundamenta en 3 pasos.

Paso 1: Seleccionar k modas iniciales, una para cada grupo.

Paso 2: Asignar cada objeto a la moda más cercana utilizando la distancia δ. Actualizar la moda del grupo después de cada asignación.

Paso 3: Después que todos los objetos han sido asignados a un grupo, volver a examinar la disimilaridad de los objetos con las modas actuales. Si un objeto es encontrado tal que su moda más cercana corresponde a otro grupo, asignar el objeto a su nueva moda y actualizar la moda de ambos grupos.

Paso 4: Repetir el paso 3 hasta que no existan objetos cambiados de grupo. (Rendón et al. 2015, p. 934).

Para la elección del k – modo inicial se pueden seleccionar los primeros k registros distintos del conjunto de datos, o, elegir las motas a partir del método de frecuencias. En palabras de Zhexue:

Like the k-means algorithm the k-modes algorithm also produces locally optimal solutions that are dependent on the initial modes and the order of objects in the data set. In our current implementation of the k-modes algorithm we include two initial mode selection methods. The first method selects the first k distinct records from the data set as the initial *k* modes. The second method is implemented with the following steps.

1. Calculate the frequencies of all categories for all attributes and store them in a category array in descending order of frequency as shown in figure 1. Here, $c_{i,j}$ denotes category i of attribute j and $f(c_{i,j}) \geq f(c_{i+1,\ j})$ where $f(c_{i,j})$ is the frequency of category $c_{i,j}$

2. Assign the most frequent categories equally to the initial k modes. For example in figure 1, assume k=3. We assign $Q_1=[q_{1,1}=c_{1,1}, q_{1,2}=c_{2,2}, q_{1,3}=c_{3,3}, q_{1,4}=c_{1,4}]$, $Q_2=[q_{2,1}=c_{2,1}, q_{2,2}=c_{1,2}, q_{2,3}=c_{4,3}, q_{2,4}=c_{2,4}]$ and $Q_3=[q_{3,1}=c_{3,1}, q_{3,2}=c_{2,2}, q_{3,3}=c_{1,3}, q_{3,4}=c_{3,4}]$.

3. Start with Q_1. Select the record most similar to Q_1 and replace Q_1 with the record as the first initial mode. Then select the record most similar to Q_2 and replace Q_2 with the record as the second initial mode. Continue this process until Q_k is replaced. In these selections $Q_l \neq Q_t$ for $l \neq t$. (Zhexue Huang, 1998, p. 290).

En cuanto al primer método, en la selección de modas cabe considerar que el concepto de moda presentado en de Zhexue y explicado por Pastrán Ramírez & Roa Peña (2015), lo exhibe como un vector que minimiza la disimilaridad, hecho que se evidencia en la siguiente definición.

Sea $X = \{X_1, X_2,..., X_n\}$ un conjunto de n individuos caracterizados por variables categóricas de tipo $(A_1, A_2, ..., A_m)$. Una moda de $X = \{X_1, X_2,..., X_n\}$ es un vector $Q = [q_1, q_2,..., q_m]$ que minimiza

$$d(X, Q) = \sum_{i=1}^{n} \delta(X_i, Q)$$

Aquí Q no es necesariamente un elemento de X. (Pastrán Ramírez & Roa Peña, 2015, p. 30).

De aquí que, y a decir de Zhexue (1998), la moda no es única. Para aclarar esta situación, sea A = {[a, b], [a c], [c, b], [b, c]}. En A se puede observar que, de las primeras componentes, la moda refiere al elemento "a" (por tener mayor número de observaciones), pero, en la segunda componente los elementos “b” y “c” están en igual frecuencia (2 veces cada uno) hecho que permite exponer que la moda puede ser [a, b], o, [a, c].

Otro ejemplo en donde se construye la disimilaridad a partir de modas preestablecidas y donde se analiza los clústeres obtenidos mediante k – modas, es el expuesto por The Academician (2020), donde relaciona las modas (a, c, b) y (x, y, z), con las ternas (a, b, c), (a, c, b), (a, b, c), (b, a, c), (a, b, c), (x, y, z), (x, y, z), (x, z, y), (a, z, y), (x, y, c).

El hecho de tener como modas preestablecidas a (a, c, b) y (x, y, z), motiva a agrupar en un primer conglomerado a las ternas (a, b, c), (a, c, b), (a, b, c), (b, a, c), (a, b, c), (a, z, y), y, en otro clúster a (x, y, z), (x, y, z), (x, z, y), (x, y, c); ya que al comparar las distancias entre cada terna con las modas iniciales se verifica que, las ternas del clúster 1 presentar menor distancia respecto a la moda (a, c, b), mientras que las ternas del segundo conglomerado presentan menor distancia con la moda (x, y, z). Con intención de clarificar el cálculo de las distancias, tómese la primera terna (a, b, c) la cual se va asociar con las modas iniciales,

entonces, la distancia entre (a, b, c) y la moda inicial (a, c, b) es 2, porque solamente diferente en dos elementos (los de la segunda y la tercera componente), mientras que la distancia entre (a, b, c) y la moda (x, y, z) es 3, ya que todos sus elementos son diferentes. Así entonces, la terna (a, b, c) se atribuye al grupo de la moda (a, c, b) por presentar menor distancia. Este proceso se realiza con cada terna a fin de obtener los dos conglomerados descritos al comienzo de este párrafo.

Luego se observa que, en el primer conglomerado destaca como nueva moda la terna (a, b, c) ya que es la de mayor frecuencia, mientras en el segundo grupo continua como moda la terna (x, y, z). Lo que hace el k – modas ahora es analizar la disimilaridad de las ternas dadas al comienzo con estas nuevas modas, a fin de observar si las ternas cambian o no de clúster. Se debe iterar hasta no encontrar elementos que cambien de clúster y una vez se llegue a ello el proceso habrá terminado.

Por otro lado, en cuanto a las frecuencias, Pastrán Ramírez & Roa Peña (2015) convienen en conceptualizarlas bajo el cociente $\frac{n_{c_{k,j}}}{n}$, donde $n_{c_{k,j}}$ es la cantidad de veces que $c_{k,j}$ pertenece a las p-uplas asociadas a los n elementos de X; además, recalcan la importancia de usar el teorema que identifica modas mediante las frecuencias y que minimiza la función distancia. A decir de estos autores se tiene que:

Se define la frecuencia de la k – ésima categoría de cada variable A_j como:

$$\mathrm{fr}\left(A_j = n_{c_{k,j}} \middle| \mathrm{X}\right) = \mathrm{fr}\left(A_j = n_{c_{k,j}}\right) = \frac{n_{c_{k,j}}}{n}$$

Una manera rápida de identificar modas en un conjunto es la que nos ofrece el siguiente teorema.

Teorema. La función D(X,Q) es minimizada si y sólo si

$$\mathrm{fr}(A_j = q_j|\mathrm{X}) \geq \mathrm{fr}(A_j = C_{k,j}|\mathrm{X})$$

para $q_j \neq C_{k,j}$ para todo j = 1, ..., m. (Pastrán Ramírez & Roa Peña, 2015, p. 31).

Para comprender de mejor manera el concepto de frecuencias abordado por Pastrán Ramírez y compañía, se expone el contiguo ejemplo.

Ejemplo. Dadas las variables categóricas A_1, A_2, A_3, las cuales refieren en su orden al **sexo** (masculino = 1, femenino = 2), **edad** (joven = 1, adulto = 2, anciano = 3), **estado civil** (soltero = 1, casado = 2, viudo = 3), de un grupo de 9 individuos (X_1, X_2, …, X_9), cuyo valor en las variables A_1, A_2, A_3, corresponden respectivamente a las ternas (1, 2, 2) para X_1, (2, 2, 2) para X_2, (2, 1, 1) para X_3, (2, 3, 3) para X_4, (2, 3, 2) para X_5, (1, 3, 2) para X_6, (1, 2, 1) para X_7, (2, 2, 1) para X_8 y (1, 3, 3) para X_9.

La definición de frecuencia indica que para la categoría sexo existen 4 hombres y 5 mujeres, ya que en las primeras componentes de cada terna se identifican cuatro números 1, los cuales dan cuenta de los hombres presentes en el estudio, y cinco números 2, que en A_1 se asocian con personas del sexo femenino. De manera similar se identifica para A_2, una persona joven, cuatro adultos y cuatro ancianos. En A_3 se visualizan 3 personas solteras, 4 casadas y 2 viudas. Así entonces, el concepto de frecuencia abordado por Pastrán Ramírez & Roa Peña, en esencia refiere a la cantidad de veces que aparece un suceso.

- **Datos relevantes sobre k – prototipos (clústeres mixtos).**

Al trabajar con clústeres es poco frecuente encontrar softwares de libre acceso que agrupen variables categóricas mediante k – modas, puesto que la mayoría de programas que

tratan este hecho lo hacen, por lo general, recurriendo al método de clústeres para variables mixtas fundamentado en el k – prototipos.

A decir de Zhexue Huang (1998), el algoritmo k – prototipos define una medida de disimilitud combinada, en donde se acopla al k – means y al k – modas con el fin de agrupar atributos numéricos y categóricos. El proceso de agrupación en el k – prototipos es semejante al implementado en el k – means, con la diferencia que para actualizar los atributos categóricos utiliza el k – modas. Una de las ventajas de usar el k – prototipos es que permite trabajar con todo el conjunto de datos, evento que garantiza al menos un agrupamiento localmente óptimo, a diferencia de los algoritmos PAM (Partitioning Around Medoids), o CLARA (Clustering Large Applications) del k – means en donde se agrupa mediante muestras. En palabras de Zhexue:

> The major differences between CLARA and the k-prototypes algorithm are as follows: (1) CLARA clusters a large data set based on samples, whereas k-prototypes directly works on the whole data set. (2) CLARA optimises its clustering result at the sample level. A good clustering based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased. The k-prototypes algorithm optimises the cost function on the whole data set. It guarantees at least a locally optimal clustering. (3) The efficiency of CLARA depends on the sample size. The larger and more complex the whole data set is, the larger the sample is required. CLARA will no longer be efficient when the sample size exceeds a certain range, say thousands of objects. The k-prototypes algorithm has no such limitations (Zhexue Huang, 1998, p. 286).

Por otro lado, en el k – prototipos la disimilaridad entre dos objetos X, Y, puede medirse mediante la expresión

$$d(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta\,(x_j, y_j)$$

Donde el primer término representa el cuadrado de la distancia euclidiana para datos numéricos, y el segundo término es la medida de disimilitud de coincidencia simple de los valores categóricos. Finalmente se expone que la función de costo a minimizar en el k – prototipos es

$$P(W,Q) = \sum_{l=1}^{k}\left(\sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{p}\left(x_{i,j} - q_{l,J}\right)^2 + \gamma \sum_{i=1}^{n} w_{i,l} \sum_{j=p+1}^{m} \delta\left(x_{i,j}, q_{l,j}\right)\right)$$

- **Viabilidad e importancia de un clúster.**

En el trabajo con clúster aparecen diferentes formas de determinar el grado de cohesión y separación de los clústeres, entre ellas destacan el coeficiente silueta y la viabilidad e importancia de la agrupación realizada. En seguida se describe brevemente cada uno de ellos.

Coeficiente silueta. Es una medida de clustering que utiliza promedios de proximidades y es útil cuando se busca clústeres compactos y separados. En esencia, el coeficiente silueta es un número entre -1 y 1, en donde se debe tener en cuenta que, a mayor medida de la silueta mejor resulta la calidad del clúster. En aras de interpretar de mejor manera dicho coeficiente, Velásquez H. (2021) expone que valores próximos a -1 indican mal agrupamiento para cada clúster, siluetas cercanas a 0 son indicio de clúster traslapados, y coeficientes cercanos a 1 muestran clúster bien separados o altamente densos.

Para calcular el coeficiente silueta se tiene en cuenta la expresión.

$$s = \frac{b - a}{\max(a, b)}$$

En donde *“a”* es el promedio de la distancia entre una muestra y los puntos pertenecientes a una misma clase, y *“b”* representa la distancia entre la muestra y los puntos del clúster más cercano.

Viabilidad. En los clústeres la viabilidad es una medida de cohesión y separación de conglomerados, la cual establece que la agrupación efectuada no es fuerte cuando presenta valores entre -1 y 0.2. Es aceptable si el índice de agrupación se halla entre 0.2 y 0.5. Y se considera buena cuando el coeficiente de agrupación es de 0.5 a 1.

Importancia. La importancia de un clúster se considera una medición de cohesión, la cual establece que la agrupación efectuada no es fuerte cuando presenta valores entre 0 y 0.2. Es aceptable si el índice de agrupación se halla entre 0.2 y 0.6. Y se dice buena cuando el coeficiente de agrupación oscila entre 0.6 a 1.

Precisión. Zhexue Hang (1998) señala que la precisión como medida de validación de clúster en k – modas se define bajo la expresión

$$r = \frac{\sum_{i=1}^{cantidad\ cluster} a_i}{total\ datos}$$

En donde el error de agrupación se define como $r = 1 - e$, siendo buenas agrupaciones cuando $r > 0.87$.

Gráfica: Por último, la representación gráfica de los clústeres funge un rol importante en el análisis de la cohesión y separación de los conglomerados, aunque para ello se requiere cierta perspicacia y conocimiento de los datos por parte del investigador.

1.7.6 Conformación de clústeres en WEKA.

Pese a la variedad de instructivos sobre el uso de la plataforma WEKA, pocos de ellos trabajan a profundidad la conformación de clústeres. Por lo general, los textos encontrados exponen brevemente dos de los ocho algoritmos que WEKA dispone para formar grupos (EM y K-means). Este hecho motivó a que, en adelante, con base en las ayudas que ofrece el software, se desarrolle una explicación más amplia sobre el trabajo con clústeres en WEKA, no sin antes advertir que, a decir de Echeverria Castillo et al. (2020), la mayoría de esfuerzos en la programación de este software se han centrado para comprender de mejor manera los resultados de los algoritmos supervisados y no tanto de los no supervisados, acto que ha influido en que WEKA carezca de índices de validación de agrupamientos, los cuales permitirían detectar el mejor algoritmo para agrupar un conjunto de datos dado. Así entonces, este software no brinda al usuario medidas de viabilidad e importancia sobre los clústeres conformados, tampoco otorga el coeficiente silueta, pero si ofrece una visualización del comportamiento de las instancias en cada conglomerado lo que implica que el usuario predisponga de cierta habilidad en el reconocimiento de patrones y separación de elementos. Por otro lado, una de las ventajas de usar WEKA radica en su agilidad para laborar con variables cuantitativas y cualitativas utilizando para ello agrupación de variables mixtas. Otro punto a favor se presenta en que este software construye una representación gráfica de los clústeres formados, lo cual posibilita la inspección visual de posibles clústeres traslapados. Conviene señalar que la explicación detallada en seguida concierne a la versión WEKA 3.9.6, la que ofrece al usuario interactuar desde las pestañas *Explorer, Experimenter, KnowledgeFlow, Workbench,* y *Simple CLI*; donde para trabajar con clústeres se selecciona el modo *Explorer*.

En la ventana de *preprocesamiento* se requiere cargar la base de datos, caso contrario la plataforma no permitirá activar la pestaña *Clúster*. En esta instancia se sugiere elaborar una

exploración de los datos a fin de determinar si el programa los leyó bien o conviene activar algún filtro en particular. Posterior a ello, en la parte superior se debe activar la pestaña *Clúster*, la que según ULA.VE (s.f), se compone de 4 elementos que permiten al usuario: (1) *seleccionar y configurar el algoritmo,* (2) *evaluar el resultado del clúster,* (3) *visualizar los resultados,* (4) *ver los clústeres en texto*.

En la opción *Choose* de la pestaña clúster se elige el algoritmo con el cual se va agrupar, acá el analista puede escoger los métodos de *Canopy, Cobweb, EM, FarthestFirst, FilteredClusterer, HierarchicalClusterer, MakeDensityBasedClusterer, SimpleKMeans*. Un resumen de estos métodos ofrecidos en la misma interfaz de WEKA advierte lo siguiente:

- ***Canopy.*** Requiere solo una pasada sobre los datos. Puede ejecutarse en modo por lotes o incremental. Por lo general, los resultados no son tan buenos cuando se ejecutan de forma incremental ya que no se conoce de antemano el mínimo/máximo de cada atributo numérico. Tiene una heurística basada en desviaciones estándar de atributos que se puede usar en modo por lotes para establecer la distancia T2. La distancia T2 determina el número de agrupaciones que se forman.
- ***Cobweb.*** Clase que implementa los algoritmos de agrupamiento Cobweb y Classit. La aplicación de los operadores de nodos (fusión, división, etc.) en términos de ordenación y prioridad difiere entre los documentos originales de Cobweb y Classit. Este algoritmo siempre compara el mejor host, agrega una nueva hoja, fusiona los dos mejores hosts y divide el mejor host al considerar dónde colocar una nueva instancia.
- ***EM.*** (maximización de expectativas). EM asigna una distribución de probabilidad a cada instancia que indica la probabilidad de que pertenezca a cada uno de los clústeres. EM puede decidir cuántos clústeres crear mediante validación cruzada, o puede especificar a priori cuántos clústeres generar. En la validación, el número de

conglomerados se obtiene considerando el conjunto de entrenamiento que se divide aleatoriamente en 10 pliegues. La EM se realiza 10 veces y el logaritmo de verosimilitud se promedia sobre los 10 resultados. Si el logaritmo de verosimilitud ha aumentado, el número de conglomerados aumenta en 1 y el programa continúa en el paso 2.
El número de pliegues se fija en 10, siempre que el número de instancias en el conjunto de entrenamiento no sea inferior a 10. Si este es el caso, el número de pliegues se establece igual al número de instancias. Los valores que faltan se reemplazan globalmente con un remplazo de valores perdidos.

- ***FarthestFirst.*** Funciona como un agrupador aproximado simple y rápido. Modela a partir de SimpleKMeans y podría ser un inicializador útil para ello.
- ***HierarchicalClusterer:*** Clase de agrupamiento jerárquico. Implementa una serie de métodos clásicos de agrupación en clústeres jerárquicos.
- ***MakeDensityBasedClusterer:*** Clase para envolver un Clúster para que devuelva una distribución y densidad. Se ajusta a distribuciones normales y distribuciones discretas dentro de cada grupo producido por el agrupador envuelto.
- ***SimpleKMeans:*** Datos de clúster utilizando el algoritmo de k medias. Puede utilizar la distancia euclidiana (predeterminada) o la distancia de Manhattan. Si se utiliza la distancia de Manhattan, los centroides se calculan como la mediana de los componentes en lugar de la media. (Fuente: traducción realizada en esta investigación con base en las ayudas de WEKA versión 3.9.6).

Cada uno de los anteriores algoritmos acarrean consigo diferentes opciones las cuales pueden potenciar su implementación y acoplarse a las necesidades del analista de datos. Una vez elegido el algoritmo, el usuario puede observar un listado de los elementos que componen a cada clúster conformado y una representación gráfica de los grupos obtenidos.

2 ASPECTOS METODOLÓGICOS.

2.1 RESUMEN DEL CAPÍTULO.

En este capítulo se abordan las etapas de la metodología CRISP – DM referentes a: (1) comprender el negocio, (2) comprender los datos, (3) preparar los datos y (4) modelar. A lo largo de estas fases el lector encontrará la modalidad de trabajo de grado, el enfoque, alcance y diseño, la población y su unidad de análisis, las técnicas de análisis de la información, el cronograma de trabajo, presupuesto y la matriz instrumental.

2.2 COMPRENSIÓN DEL NEGOCIO.

El desarrollo de esta etapa permitió conocer, profundizar y apropiar el contexto que recubrió a la prueba Saber 11 – 2021 B, la cual acogió las particularidades de las evaluaciones realizadas en el periodo 3 (2014 – 2 en adelante) propuesto en MEN (s.f) y que, como dato adicional planteó el escenario del análisis educacional en tiempos de pandemia.

La prueba Saber 11 – 2021 B, evaluó 5 pruebas genéricas correspondientes a matemáticas, lectura crítica, ciencias naturales, sociales ciudadanas e inglés. Con base en ellas calculó el puntaje total el cual tenía un valor máximo de 500 puntos. Las pruebas genéricas se desenvolvieron en el marco del desarrollo de competencias propuestas por Delors (1996), en donde se inspeccionó en el desarrollo de las dimensiones del Saber, Hacer y Ser de cada estudiante, y su habilidad para convivir en armonía con sus semejantes. Agregado a ello, la prueba Saber 11 recogió información sobre las condiciones sociodemográficas de cada educando a partir de aspectos personales, socioeconómicos, académicos y de citación, hecho que favoreció que la base de datos tuviera variables cuantitativas y cualitativas, las que se analizaron a través de las técnicas de minerías de datos.

Así entonces, para comprender de mejor manera los resultados de la prueba Saber 11, se hizo necesario conceptualizar aspectos de rendimiento académico y minería de datos, lo que facilitaría en una etapa posterior el tratamiento metodológico y procedimental de los datos. La fundamentación teórica sobre rendimiento académico, prueba Saber 11, minería de datos y clustering, se encuentra detallada en el marco teórico (capítulo 1.7 de esta investigación).

2.2.1 Modalidad de trabajo de grado.

El presente trabajo acoge a la Investigación aplicada evaluativa como guía para el desarrollo metodológico. Es investigación aplicada puesto que tal como se exhibe en el artículo tercero del Acuerdo 014 del 11 de octubre (2022) emanado por el Comité Curricular y de Investigaciones del Programa de Maestría en Estadística Aplicada, esta busca dar soluciones a problemas planteados que generen impacto en la comunidad donde se desenvuelven, acogen los resultados de la investigación básica para el procesamiento de su información y obliga la construcción de un marco teórico para el fundamento de las ideas. Así entonces, en este trabajo se buscó dar solución a un problema en específico el cual aludió al desempeño en matemáticas de los estudiantes del departamento de Nariño (ver literal 1.3), y se logró que las conclusiones conseguidas dieran pistas sobre los factores relacionados con los desempeños altos en matemáticas. Además, para el análisis y discusión de los resultados fue necesario construir un marco teórico (ver literal 1.7) el cual se desagregó en los tópicos de rendimiento académico, consideraciones generales sobre la prueba Saber, minería de datos y clustering. Con dicho marco teórico se evidenció la necesidad de utilizar conocimientos emergentes en la investigación básica asociados a las técnicas de minería de datos.

Anexo a lo anterior cabe señalar que la construcción del marco teórico permitió precisar que esta investigación giró entorno de intereses evaluativos, puesto que valoró el desempeño en matemáticas de los estudiantes de Nariño, determinó los factores con mayor relación en la

caracterización de puntajes altos en dicha área y analizó el rendimiento académico de los nariñenses tras afrontar un periodo pandémico, el cual los obligó a abandonar la escuela en su naturaleza presencial e introducirse en el mundo del aprendizaje virtual.

2.2.2 Enfoque.

Cuantitativo, ya que utiliza métodos estadísticos y de minería de datos para el tratamiento de la base de datos de la prueba Saber 11 periodo 2021 – B.

2.2.3 Alcance de la investigación.

La presente investigación es descriptiva y correlacional ya que busca determinar las relaciones entre los puntajes de la prueba de matemáticas con los puntajes de las otras áreas evaluadas. Busca además caracterizar los desempeños altos en matemáticas acudiendo para ello al análisis clúster, análisis de factores y la correlación entre variables.

Para Hernández Sampieri et al. (2014), la intención de los estudios descriptivos es medir información pero sin indicar correlación existente entre sus variables, buscando "especificar las propiedades, las características y los perfiles de personas, grupos, comunidades, procesos, objetos o cualquier otro fenómeno que se someta a un análisis." (Hernández Sampieri et al. 2014, p. 92).

Agregado a ello, la presente investigación tiene alcance correlacional en tanto estudia la relación existente entre los puntajes de las pruebas de matemáticas, inglés, ciencias naturales, lectura crítica, y sociales ciudadanas. El estudio correlacional permite identificar el grado de asociación existente entre las variables la cual puede ser positiva, negativa, no presentar relación fuerte o ser nula. La diferencia entre estos dos tipos de alcances mencionados, a decir de Hernández Sampieri et al. (2014), es que los estudios correlacionales detallan con precisión

el grado de vinculación existente entre las variables, mientras que los descriptivos miden con precisión las variables individuales.

2.2.4 Diseño de la investigación.

El presente trabajo se considera *no experimental de corte transeccional o transversal*, puesto que como muestra el mencionado Hernández Sampieri y su equipo, la investigación no experimental es aquella en donde las variables se estudian sin sufrir manipulación alguna por parte del observador, es decir los datos son analizados tal y como se presentan y no son sometidos a ningún tipo de experimentación. Es de tipo transeccional porque los datos expuestos en la base de datos obedecen a un único momento de realización de la prueba concerniente al periodo 2021 – B.

2.2.5 Población.

La presente investigación realiza un estudio censal hecho que implica no elaborar procesos de muestreo.

2.2.6 Unidad de análisis.

Estudiantes del departamento de Nariño que presentaron la prueba Saber 11 en el periodo 2021 – B.

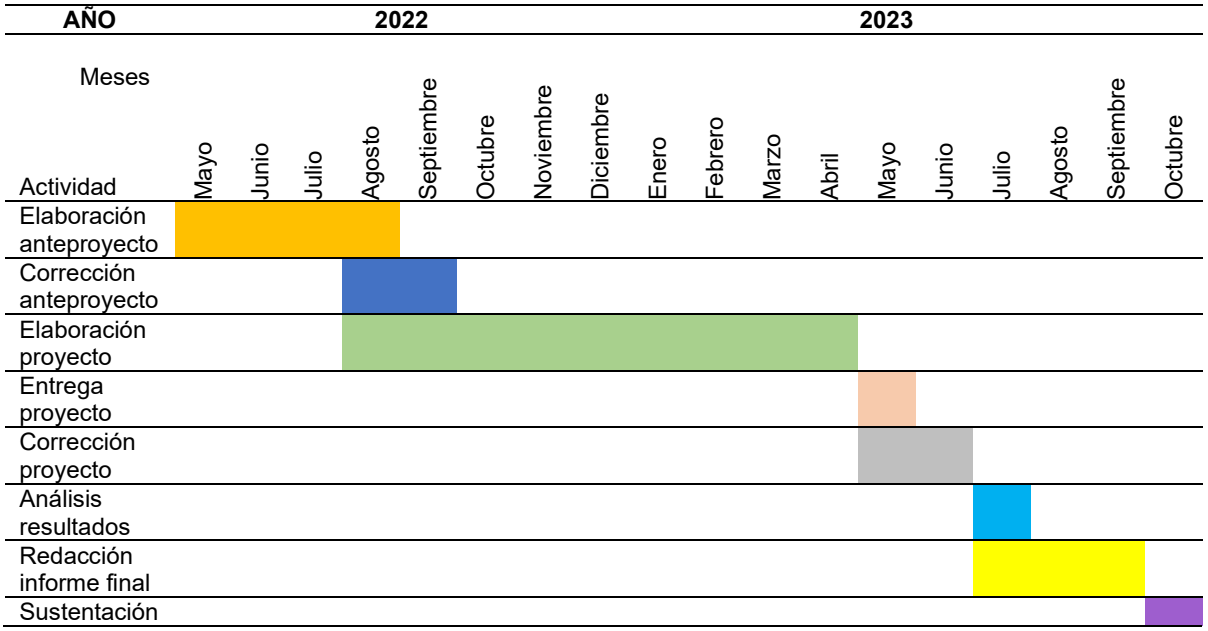
2.2.7 Cronograma de actividades.

Las actividades proyectadas para el presente trabajo se realizaron teniendo en cuenta las fechas y actividades descritas en la tabla 1, de donde se destaca que en el año 2022 se elaboró el anteproyecto en los meses de mayo a agosto, la corrección del mismo se hizo en agosto y septiembre. De forma simultánea se fue realizando el proyecto durante los meses de agosto de 2022 a abril de 2023. En el mes de mayo se hace entrega del proyecto al comité

curricular y se comienza con la corrección del proyecto. El análisis de resultados y la redacción del informe final se comienza en julio. Se prevé que la sustentación del trabajo se realiza en octubre de 2023.

Tabla 1.

Cronograma de trabajo.



Fuente: Esta investigación.

2.2.8 Presupuesto de inversión.

Los rubros planteados para este proyecto se describen en la tabla 2, en donde se explicita el valor a considerar para pago de inscripción de tesis, matrícula de egresado, servicios públicos, seguridad social, entre otros.

Tabla 2.

Presupuesto de inversión.

RUBROS	DESCRIPCIÓN	VALOR UNITARIO	CANTIDAD	TOTAL
INSCRIPCIÓN	Inscripción del proyecto de trabajo de grado.	1.500.000	1	1.500.000
MATRICULA	Pago matrícula de egresado	133.000	1	133.000
	Recibo de energía	40.000	17	680.000

RUBROS	DESCRIPCIÓN	VALOR UNITARIO	CANTIDAD	TOTAL
SERVICIOS PÚBLICOS	Recibo de agua potable, acueducto, alcantarillado.	42.000	17	714.000
	Internet, telefonía, televisión.	98.000	17	1.666.000
SEGURIDAD SOCIAL	Aporte al sistema de salud.	125.000	17	2.125.000
	Aporte al sistema de pensión.	160.000	17	2.720.000
HONORARIOS	Pago mensual de honorarios del investigador. Referencia del salario mínimo legal vigente.	1.000.000	17	17.000.000
PAPELERIA	Lapiceros.	800	3	2.400
	Resma de papel tamaño carta.	11.000	1	11.000
	Impresión tesis.	80.000	1	80.000
	Fotocopiado de tesis.	15.000	4	60.000
	Encuadernación trabajo.	25.000	2	50.000
VARIOS	Gastos varios y/o imprevistos.	500.000	1	500.000
			TOTAL	**27.241.400**

Fuente: Esta investigación.

2.3 COMPRENSIÓN DE LOS DATOS.

El desarrollo de esta etapa permitió identificar la información expuesta en la base de datos abiertos del ICFES concerniente a la prueba Saber 11 – 2021 B, la que en principio contenía 532.980 filas y 82 variables que daban cuenta de la información personal, académica, de citación a las pruebas y de la situación socioeconómica de los estudiantes de los 32 departamentos de Colombia. En dichas variables se hallaban aspectos como documento de identificación, nacionalidad y género, entre otras. Los nombres de las variables se caracterizaban porque a cada atributo antecedía el prefijo ESTU para reconocer los valores de información personal, COLE para tratar los aspectos académicos y de citación, y FAMI para identificar los datos socioeconómicos, por ende que se hallaran nombres de variables como ESTU_GÉNERO que refería al género del discente, COLE_GENERO que indicaba si la institución era de tipo masculino, femenino o mixto, y FAMI_CUARTOSHOGAR la cual indicaba el número de cuartos existente en el domicilio del educando. Además, los resultados de las pruebas genéricas fueron entregados mediante puntajes, desempeños y percentiles, siendo los puntajes y los percentiles variables cuantitativas y los desempeños variables categóricas.

Los desempeños según ICFES (2020a) se clasificaron en *bajo o insuficiente, básico o mínimo, satisfactorio, y, avanzado*. Cada uno obedecía a un rango de puntajes obtenido por los estudiantes en la prueba y daba cuenta de las habilidades que por materia manejaba el aprendiz. La tabla 3 esquematiza estos hechos.

Tabla 3.

*Clasificación de n*iveles de desempeño.

Taxonomía de los niveles de desempeño		
Nivel	**Puntaje**	**Descripción**
1	0 a 35	Bajo o insuficiente
2	36 a 50	Básico o mínimo
3	51 a 70	Satisfactorio
4	71 a 100	Avanzado

Fuente: Esta investigación con base en los datos de ICFES (2020a).

2.4 PREPARACIÓN DE LOS DATOS.

Es esta etapa se realizó un depuramiento de la base de datos original, el cual comenzó con un filtrado por departamento a fin de seleccionar a los discentes oriundos de Nariño, fue así como surgió el archivo *datos_nariño_2021b* que se caracterizó por tener 16.615 registros de individuos. Posteriormente y con intención de ganar información, se eliminaron las variables esgrimidas en la tabla 4, ya que se determinó que no aportaban información relevante para este estudio.

Tabla 4.

Variables eliminadas de la prueba Saber 11.

VARIABLE ELIMINADA	**DESCRIPCIÓN**
DOCUMENTO DE IDENTIDAD	Número de documento del estudiante
ESTU_NACIONALIDAD	País de residencia del estudiante
PERIODO	Refiere a la etapa 2021 – B
ESTU_CONSECUTIVO	Código asignado al estudiante por parte del ICFES
ESTU_ESTUDIANTE	Clasifica a todos los individuos como estudiante.
ESTU_DEPTO_RESIDE	Departamento de residencia del estudiante, en este caso correspondió a Nariño.
ESTU_COD_RESIDE_MCPIO.	Código postal de cada municipio.

VARIABLE ELIMINADA	DESCRIPCIÓN
ESTU_COD_RESIDE_DEPTO	Código postal del departamento, tiene un valor de 52 para todos los estudiantes.
FAMI_TIENESERVICIOTV	Indica si la familia cuenta o no con internet.
FAMI_TIENELAVADORA	Indica si en casa tienen o no con lavadora.
FAMI_TIENEHORNOMICROOGAS.	Indica si la familia posee o no con microondas o gas.
FAMI_TIENEAUTOMOVIL.	Indica si en casa cuentan o no con automóvil.
FAMI_TIENEMOTOCICLETA.	Indica si la familia tiene o no con motocicleta.
FAMI_TIENECONSOLAVIDEOJUEGOS.	Indica si en casa hay o no consola de videojuegos.
FAMI_COMELECHEDERIVADOS.	Indica la frecuencia en el consumo de alimentos derivados de la leche
FAMI_COMECARNEPESCADOHUEVO.	Indica la frecuencia en el consumo de alimentos derivados de la carne, pescado o huevo.
FAMI_COMECEREALFRUTOSLEGUMBRE	Indica la frecuencia en el consumo de alimentos derivados de cereales, frutas o legumbres.
FAMI_SITUACIONECONOMICA	Establece el estado socioeconómico de la familia (igual, mejor o peor) pero no da punto de comparación con otro periodo de tiempo.
ESTU_TIPOREMUNERACION	Indica si el estudiante recibe dinero por alguna actividad efectuada.
COLE_CODIGO_ICFES	Código del colegio ante el ICFES
COLE_COD_DANE_ESTABLECIMIENTO	Código del colegio ante el DANE
COLE_COD_DANE_SEDE	Código de la sede del colegio ante el DANE
COLE_NOMBRE_SEDE	Nombre de la sede del colegio.
COLE_SEDE_PRINCIPAL	Indica si la sede es principal o no.
COLE_COD_MCPIO_UBICACION	Código postal del colegio ante el ICFES según el municipio.
COLE_COD_DEPTO_UBICACION	Código postal del colegio ante el ICFES según el departamento.
ESTU_PRIVADO_LIBERTAD	Indica si el estudiante es o no privado de libertad
ESTU_COD_DEPTO_PRESENTACION	Código postal del departamento donde el estudiante presenta la prueba
ESTU_COD_MCPIO_PRESENTACION	Código postal del municipio donde el estudiante presenta la prueba
ESTU_ESTADOINVESTIGACION	Indica el estado de publicaciones, las cuales se hallan con sus subniveles de publicar, validez oficina jurídica.
PERCENTIL_ESPECIAL_GLOBAL	Indica el percentil global en la cual se ubicó el estudiante en la prueba. Se eliminaron también los percentiles de cada área y se dejó solo los puntajes.

Fuente: Esta investigación.

Con lo anterior se logró que *datos_nariño_2021b*, contara con 16.615 individuos y 41 variables de las cuales 35 eran categóricas y 6 numéricas. Consecuente a esto se procedió a crear la matriz instrumental, la cual acopló las variables expuestas en la base de datos del ICFES con los aspectos socioeconómicos, familiares, institucionales, demográficos y

académicos (rendimiento en las pruebas), los cuales se trataron en el literal 1.7.1 del presente texto.

2.4.1 Matriz instrumental.

La matriz instrumental presenta en su primera columna las dimensiones en las que fueron acopladas las variables a investigar, en la segunda expone la definición de estas dimensiones, luego muestra las variables que componen a cada dimensión y finalmente esboza una breve descripción de los mismos. Con intención de brindar claridad para la lectura de la tabla 5 se tiene por ejemplo que, los aspectos socioeconómicos se definen como el conjunto de factores sociales y económicos presentes en el contexto del estudiante, son estudiados mediante las variables fami_estratovivienda, fami_educacionpadre, fami_educacionmadre, fami_trabajolaborpadre. La primera de estas variables describe el estrato del estudiante, la segunda explica el máximo nivel educativo del padre, la tercera expone el máximo nivel educativo de la madre y así sucesivamente.

Tabla 5.

Matriz instrumental.

DIMENSION	DEFINICIÓN	VARIABLE	DESCRIPCIÓN
Aspectos socioeconómicos	Conjunto de factores sociales y económicos presentes en el contexto del estudiante	FAMI_ESTRATOVIVIENDA	Estrato de la vivienda
		FAMI_EDUCACIONPADRE	Máximo nivel educativo del padre
		FAMI_EDUCACIONMADRE	Máximo nivel educativo de la madre
		FAMI_TRABAJOLABORPADRE	Dedicación, oficio o trabajo del padre
		FAMI_TRABAJOLABORMADRE	Dedicación, o trabajo de la madre
		ESTU_NSE_INDIVIDUAL	Nivel socioeconómico del estudiante.
Aspectos familiares	Refiere a las condiciones o comodidades que dispone el estudiante en su núcleo familiar.	FAMI_PERSONASHOGAR	Número de personas en el hogar.
		FAMI_CUARTOSHOGAR	Número de cuartos en la vivienda.
		FAMI_TIENEINTERNET	Disponibilidad de internet en la vivienda.
		FAMI_TIENECOMPUTADOR	Disponibilidad de computador en la vivienda.
		FAMI_NUMLIBROS	Disponibilidad de libros en la vivienda.
Aspectos institucionales	Describe característica asociadas a la institución educativa del estudiante	COLE_NOMBRE_ESTABLECIMIENTO	Nombre del establecimiento educativo
		COLE_GENERO	Género del colegio.
		COLE_NATURALEZA	Tipo de colegio.
		COLE_CALENDARIO	Calendario académico del colegio.
		COLE_BILINGUE	Institución bilingüe
		COLE_CARACTER	Modalidad del colegio
		COLE_AREA_UBICACION	Ubicación geográfica del colegio

DIMENSION	DEFINICIÓN	VARIABLE	DESCRIPCIÓN
		COLE_JORNADA	Jornada en la que se trabaja en el colegio.
		COLE_MCPIO_UBICACION	Municipio de ubicación del colegio
Aspectos demográficos	Información del estudiante sobre atributos personales como: edad, sexo, origen étnico, entre otros.	ESTU_GENERO	Género del estudiante
		EDAD	Rango de edad del estudiante al presentar la prueba
		ESTU_TIENEETNIA	El estudiante pertenece a algún grupo étnico
		ESTU_MCPIO_RESIDE	Municipio de residencia del estudiante
		ESTU_DEDICACIONLECTURADIARIA	Tiempo diario que dedica el estudiante a actividades de lectura
		ESTU_DEDICACIONINTERNET	Tiempo diario que dedica el estudiante a navegar en internet
		ESTU_HORASSEMANATRABAJA	Tiempo en horas que el estudiante trabaja a la semana
		ESTU_GENERACION-E	Clasificación según el programa generación Excelencia.
Rendimiento en las pruebas	Mide el nivel del rendimiento a partir de los puntajes y desempeños obtenidos en las pruebas.	PUNT_LECTURA_CRITICA	Puntaje obtenido en la prueba
		PUNT_MATEMATICAS	Puntaje obtenido en la prueba
		PUNT_C_NATURALES	Puntaje obtenido en la prueba
		PUNT_SOCIALES_CIUDADANAS	Puntaje obtenido en la prueba
		PUNT_INGLES	Puntaje obtenido en la prueba
		PUNT_GLOBAL	Puntaje global obtenido en la prueba
		DESEMP_LECTURA_CRITICA	Desempeño obtenido en la prueba
		DESEMP_PUNT_MATEMATICAS	Desempeño obtenido en la prueba
		DESEMP_PUNT_C_NATURALES	Desempeño obtenido en la prueba
		DESEMP_PUNT_ SOCIALES_CIUDADANAS	Desempeño obtenido en la prueba
		DESEMP_PUNT_INGLES	Desempeño obtenido en la prueba
		DESEMP_PUNT_GLOBAL	Desempeño obtenido en la prueba

Fuente: Esta investigación con sustento en la base de datos libres publicados por el ICFES.

Algunos de las variables expuestas en la matriz anterior fueron codificados teniendo en cuenta detalles como ubicar una línea al piso a fin que no queden espacios entre los nombres, o dar palabras claves a cada variable a fin de acortar la longitud de las mismas. La tabla 6 muestra la codificación realizada.

Tabla 6.

Valores de los indicadores de las variables Saber 11

VARIABLE	CATEGORIA		CODIFICACIÓN
Estrato de la vivienda	Sin estrato	=	Sin_estrato
	Estrato 1	=	1
	Estrato 2	=	2
	Estrato 3	=	3
	Estrato 4	=	4
	Estrato 5	=	5
	Estrato 6	=	6
	Ninguno	=	Ninguno
	Primaria incompleta	=	Prim_incomp
	Primaria completa	=	Prim_comp
	Secundaria incompleta	=	Secund_incomp
	Secundaria completa	=	Secund_comp

VARIABLE	CATEGORIA		CODIFICACIÓN
Escolaridad del padre	Técnica o tecnológica incompleta	=	Tec_incomp
	Técnica o tecnológica completa	=	Tec_comp
	Educación profesional incompleta	=	Prof_incomp
	Educación profesional completa	=	Prof_comp
	Postgrado	=	Postg
	No aplica	=	No_aplica
	No sabe	=	N_S
Escolaridad de la madre	Ninguno	=	Ninguno
	Primaria incompleta	=	Prim_incomp
	Primaria completa	=	Prim_comp
	Secundaria incompleta	=	Secund_incomp
	Secundaria completa	=	Secund_comp
	Técnica o tecnológica incompleta	=	Tec_incomp
	Técnica o tecnológica completa	=	Tec_comp
	Educación profesional incompleta	=	Prof_incomp
	Educación profesional completa	=	Prof_comp
	Postgrado	=	Postg
	No aplica	=	No_aplica
	No sabe	=	N_S
Ocupación laboral del padre	No aplica	=	No_aplica
	No sabe	=	N_S
	Pensionado	=	Pensionado
	Dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial.	=	Gran_empresario_directivo
	Dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo, tienda, papelería, etc.)	=	Microempresario
	Operario de máquinas o conduce vehículos (taxista, chofer)	=	Conductor_vehiculos
	Trabaja en el hogar, no trabaja o estudia.	=	Trabaja_en_casa, no_trabaja_estudia
	Agricultor, pesquero o jornalero	=	Agricultor_pesquero_jornalero
	Trabaja por cuenta propia (por ejemplo, plomero, electricista)	=	Oficios_varios_independiente
	Trabaja como profesional (por ejemplo, médico, abogado, ingeniero)	=	Trabajador_profesional
	Trabaja como personal de limpieza, mantenimiento, seguridad o construcción	=	Empleado_servicios_generales
	Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente)	=	Auxiliar_administrativo
	Es vendedor o trabaja en atención al público.	=	Ventas
Ocupación laboral de la madre	No aplica	=	No_aplica
	No sabe	=	N_S
	Pensionado	=	Pensionado
	Dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial.	=	Gran_empresario_directivo
	Dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo, tienda, papelería, etc.)	=	Microempresario
	Operario de máquinas o conduce vehículos (taxista, chofer)	=	Conductor_vehiculos
	Trabaja en el hogar, no trabaja o estudia.	=	Trabaja_en_casa, no_trabaja_estudia
	Agricultor, pesquero o jornalero	=	Agricultor_pesquero_jornalero
	Trabaja por cuenta propia (por ejemplo, plomero, electricista)	=	Oficios_varios_independiente

VARIABLE	CATEGORIA		CODIFICACIÓN
	Trabaja como profesional (por ejemplo, médico, abogado, ingeniero)	=	Trabajador_profesional
	Trabaja como personal de limpieza, mantenimiento, seguridad o construcción	=	Empleado_servicios_generales
	Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente)	=	Auxiliar_administrativo
	Es vendedor o trabaja en atención al público.	=	Ventas
Número de personas en el hogar	1 a 2	=	1_a_2
	3 a 4	=	3_a_4
	5 a 6	=	5_a_6
	7 a 8	=	7_a_8
	9 o más	=	9_o_mas
Número de cuartos en el hogar	Seis o más	=	Seis_o_mas
Internet en casa	No responde	=	N_R
Computador en casa	No responde	=	N_R
Número de libros en casa	0 a 10		0_a_10
	11 a 25		11_a_25
	26 a 100		26_a_100
	Más de 100		Mas_de_100
Naturaleza del colegio	No oficial	=	No_oficial
Colegio bilingüe	No responde	=	N_R
Modalidad del colegio	No responde	=	N_R
Género del estudiante	Femenino	=	F
	Masculino	=	M
Edad			X<18
			18<=X<=22
			X>22
Estudiante con etnia	No responde	=	N_R
Tiempo promedio de lectura al día	30 minutos o menos	=	30_minutos_o_menos
	Entre 30 y 60 minutos	=	30_y_60 minutos
	Entre 1 y 2 horas	=	1_y_2_horas
	Más de 2 horas	=	Mas_de_2_horas
	No leo por entretenimiento	=	No_lee_por_entret
	No responde	=	N_R
Dedicación diaria a internet	30 minutos o menos	=	30_minutos_o_menos
	Entre 30 y 60 minutos	=	30_y_60 minutos
	Entre 1 y 3 horas	=	1_y_3_horas
	Más de 3 horas	=	Mas_de_3_horas
	No Navega Internet	=	No_navega
	No responde	=	N_R
Número de horas que trabaja el estudiante	Menos de 10 horas	=	[1, 10]
	Entre 11 y 20 horas	=	[11, 20]
	Entre 21 y 30 horas	=	[21, 30]
	Más de 30 horas	=	Mas_30_horas
	No responde	=	N_R
Generación E	Excelencia departamental	=	Excelencia_departamental
	Excelencia nacional	=	Excelencia_nacional

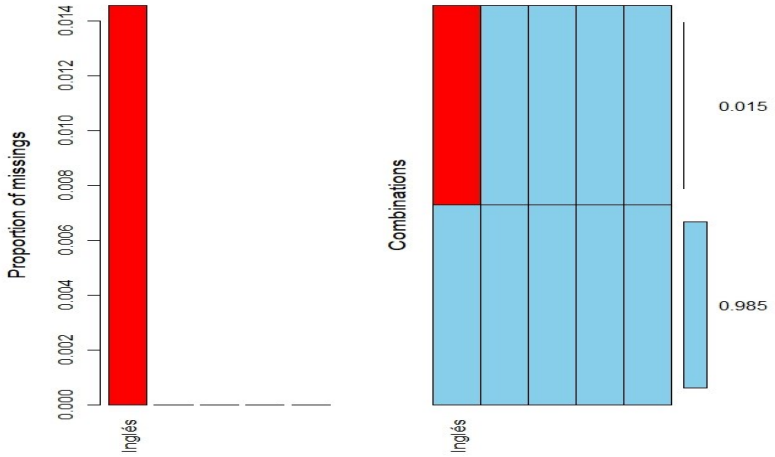
Fuente: Esta investigación basada en la base de datos libres publicada por el ICFES.

2.4.2 Imputación y limpieza.

Creada la matriz instrumental, se procedió a realizar un proceso de limpieza e imputación de datos, en donde se corrigieron problemas de caracteres que dificultaban la labor en RStudio como tildes, espacios en blanco y el símbolo “-” del desempeño en inglés, el cual se cambió por la palabra “menos” ya que el software lo asumía como un operador y no como un elemento distintivo de la prueba. De igual forma se realizó un proceso de imputación de valores faltantes de las observaciones numéricas, para lo cual se utilizó el comando *mice* con método “pmm” debido a su eficacia y simplicidad en los cálculos. Cabe mencionar que en la prueba de inglés se hallaron 242 datos faltantes equivalentes al 1.46% del total de registros de estudiantes. El comando *aggr* de RStudio permitió contemplar de manera gráfica esta situación (ver figura 1), notando que, en la gráfica izquierda, solo el área de inglés presentaba valores perdidos, y a la derecha, que estos se daban cuando las combinaciones de los demás valores se hallaban completas. En igual orden de ideas, el comando *vis_dat()* de la librería visdat, reconfirmó visualmente que eran muy pocos los datos ausentes en comparación con el conjunto de datos en estudio, hecho que se aprecia de mejor manera en la figura 2

Figura 1.

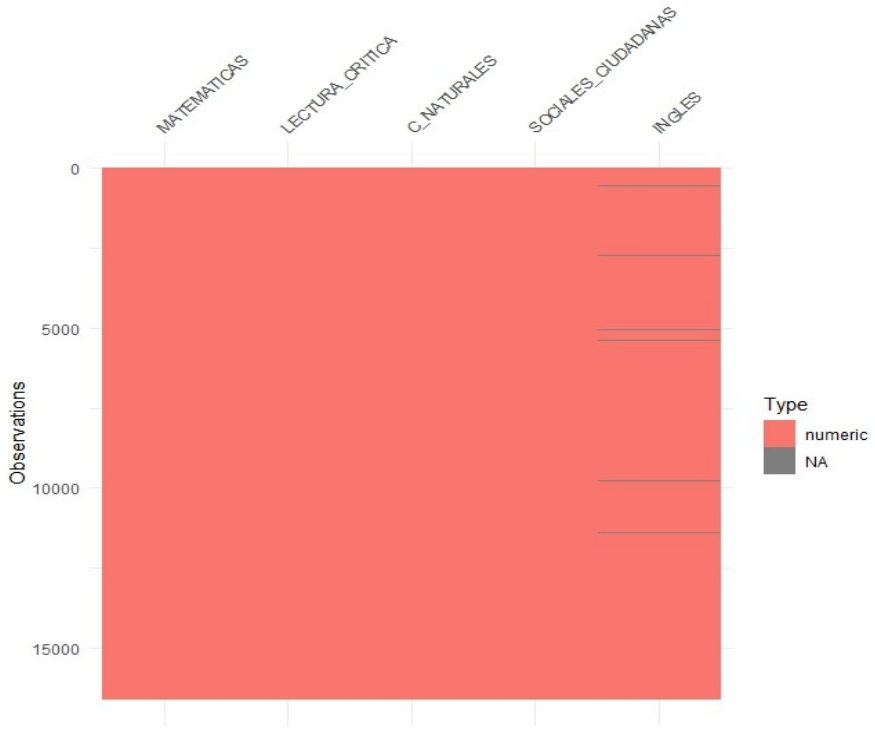
Identificación valores perdidos con comando aggr.



Fuente: esta investigación.

Figura 2.

Identificación valores perdidos con comando vis_dat.



Fuente: esta investigación.

Como se mencionó, el porcentaje de datos perdidos fue bajo, sin embargo, se indagó mediante el test de Baylor si la ausencia de ellos seguía un patrón MCAR, por esta razón se utilizó el comando *LittleMCAR(x)* de la librería *BaylorEdPsych* en consideración a la hipótesis:

$$\begin{cases} H_0: Los\ datos\ faltantes\ siguen\ un\ patrón\ MCAR \\ H_1: Los\ datos\ faltantes\ no\ siguen\ un\ patrón\ \ MCAR \end{cases}$$

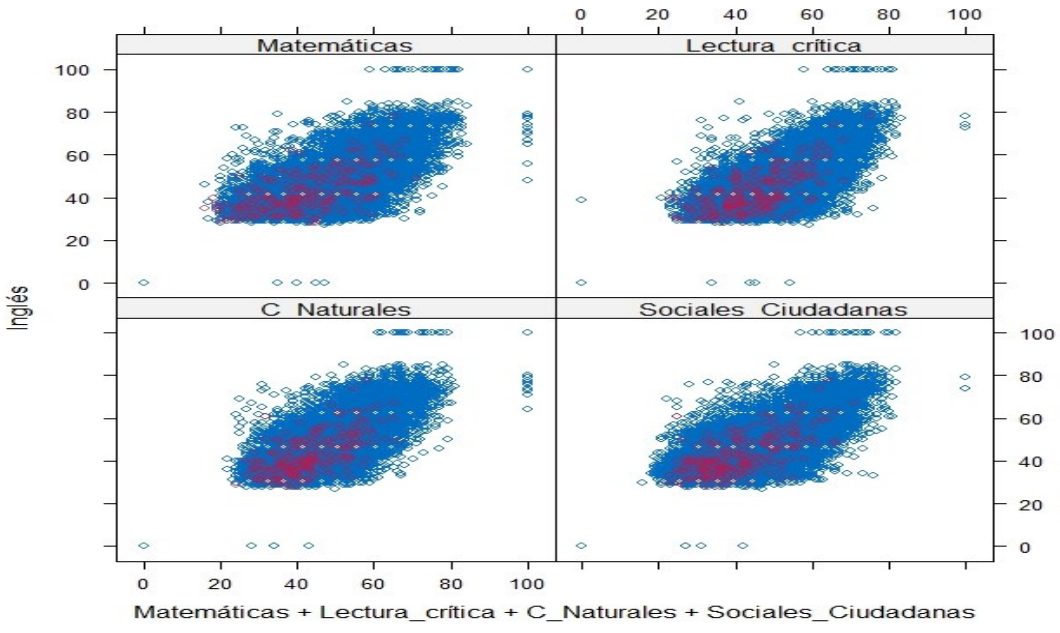
Tras la aplicación del test de Baylor se obtuvo un p – valor de 0, con lo cual se rechazó H_0 y se planteó que la ausencia de valores no tenía un patrón MCAR, pero sí uno MAR, ya que como exhibe la figura 1, los valores faltantes de inglés se sucedían cuando las combinaciones de los demás valores se hallaban completas, es decir, dependían de las demás pruebas y razón por la cual, para la imputación de datos se usaron técnicas de regresión múltiple en

donde el comando *mice* jugó un papel importante, ya que permitió visualizar el acople de cada uno de sus métodos en la imputación de valores perdidos para el área de inglés, determinando así que el método "*pmm*" era el que mejor ajuste brindaba en la nube de puntos, es decir, la nube de color rojo que se observa en la figura 3 se adecuó de mejor manera a cada una de las nubes azules, las cuales daban cuenta del comportamiento de los valores presentes en las demás pruebas evaluadas.

El método "*pmm*" realiza internamente una regresión para estimar los valores faltantes ubicando los valores de inglés en función de matemáticas, lectura crítica, ciencias naturales, y sociales ciudadanas. Otro método que emergió como posible candidato de imputación fue *midas touch*, sin embargo, por elección del investigador se prefirió *pmm*. Los métodos de *regresión lasso* y *booststrap*, entre otros, no exhibieron buenos ajustes visuales.

Figura 3.

Nube de puntos valores imputados.

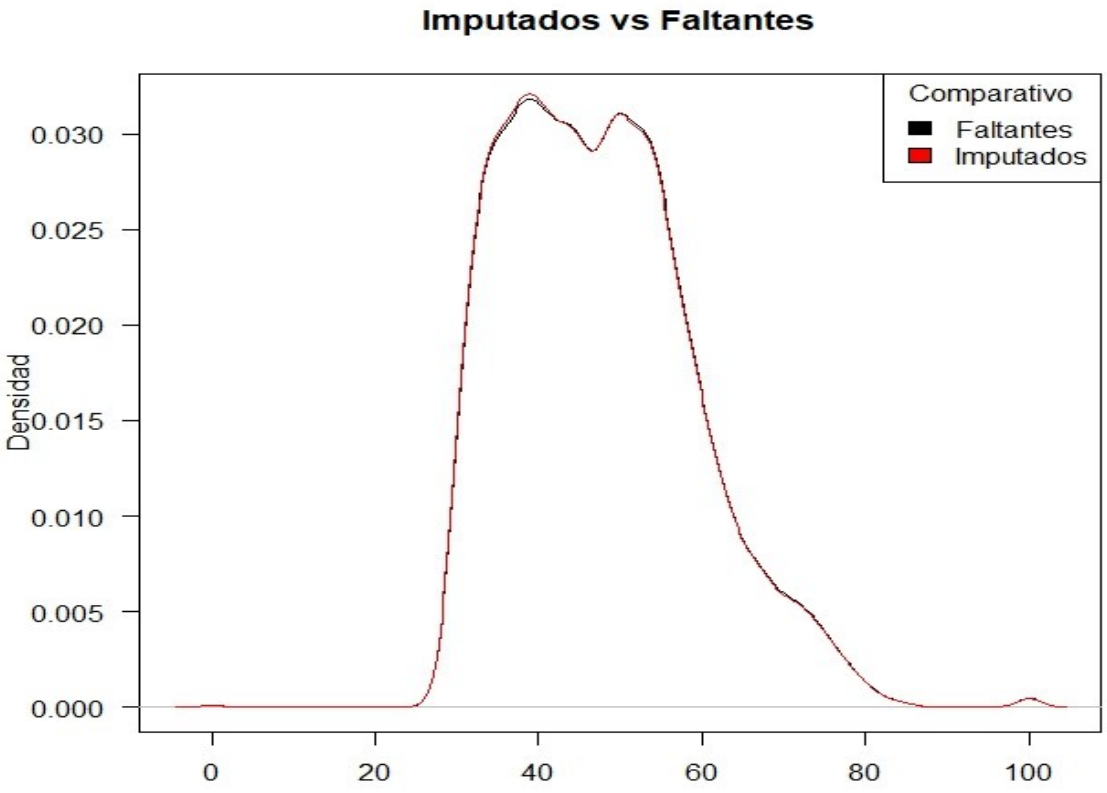


Fuente: esta investigación.

Seguido a lo anterior, se procedió a comparar los gráficos de densidad de los datos imputados en contra de los no imputados, con el fin de precisar si la imputación conservaba la distribución original de los datos o por el contrario la distorsionaba totalmente, fue así como en la figura 4 se evidenció que, la curva de color rojo (valores imputados) no se distanciaba mucho de la curva color negro (sin imputar) hecho que favoreció concluir que la imputación efectuada brindaba resultados acordes al comportamiento esperado de los individuos. Por otro lado, los valores ausentes de las variables cualitativas se suplieron con la moda de cada categoría.

Figura 4.

Comparativo de curvas de datos faltantes versus datos imputados.



Fuente: esta investigación.

Realizado este proceso se elaboró un primer análisis clúster a fin de brindar un panorama general del estado real de los datos, con el que se determinó la importancia de trabajar con 3 agrupaciones, sin embargo se precisó que variables categóricas como estrato, número de personas en el hogar, educación padre, educación madre, número de libros en casa, tiempo de lectura por diversión, género del colegio, naturaleza del colegio, calendario del colegio, colegio bilingüe, jornada del colegio y generación del estudiante; presentaban valores atípicos, los cuales no influían en el modelo del clúster, es decir, al realizar una inspección previa sobre cómo iban a quedar los clústeres se observó que las agrupaciones conseguidas sin los datos atípicos, tenían los mismos elementos que los conglomerados que si consideraban estos valores extraños.

En una etapa posterior se procedió a analizar la estructura de los datos, de donde cabe decir que las variables categóricas se clasificaron en diferentes niveles, por ejemplo, la variable género del estudiante contó con 2 niveles, “F” que indicaba *Femenino* y “M” que traducía *Masculino*, la variable edad se categorizó con 3 niveles correspondientes a edades menores a 18 años (x<18), entre 18 y 22 (18<=x<=22), y mayores de 22 años (x > 22). La descripción en detalle de cada variable se expone en la tabla 7.

Tabla 7.

Estructura de los datos del departamento de Nariño.

N°	VARIABLE	TIPO	NIVELES	DETALLE NIVELES
1	Género del estudiante	Factor	2	“Femenino”, “Masculino”.
2	Edad	Factor	3	“X<18”, “18<=X<=22”, “X>22”
3	Estudiante con etnia	Factor	2	“No”, “Si”
4	Municipio de residencia	Factor	64	"Albán", "Aldana", …
5	Estrato de la vivienda	Factor	3	“1”, “2”, “3”
6	Número de personas en el hogar	Factor	4	"1_a_2", "3_a_4", "5_a_6", "7_a_8"
7	Número de cuartos en el hogar	Factor	6	“Uno”, “Dos”, “Tres”, “Cuatro”, “Cinco”, “Seis o más”
8	Nivel educativo del padre	FACTOR	9	"Ninguno", "Primaria incompleta", "Primaria completa", "Secundaria incompleta", "Secundaria completa", "Técnico incompleto",

N°	VARIABLE	TIPO	NIVELES	DETALLE NIVELES
				"Técnico completo", "Profesional incompleta", "Profesional completa".
9	Nivel educativo de la madre	FACTOR	9	"Ninguno", "Primaria incompleta", "Primaria completa", "Secundaria incompleta", "Secundaria completa", "Técnico incompleto", "Técnico completo", "Profesional incompleta", "Profesional completa".
10	Ocupación laboral del padre	FACTOR	13	"Agricultor, pesquero, jornalero", "Auxiliar administrativo", "Conductor de vehículos", "Empleado en servicios generales", "Gran empresario o directivo", "Microempresario", "No sabe, "No aplica", "Oficios varios o independiente", "Pensionado", "Trabaja en casa, no trabaja o estudia", "Trabajador profesional", "Ventas".
11	Ocupación laboral de la madre	FACTOR	13	"Agricultor, pesquero, jornalero", "Auxiliar administrativo", "Conductor de vehículos", "Empleado en servicios generales", "Gran empresario o directivo", "Microempresario", "No sabe, "No aplica", "Oficios varios o independiente", "Pensionado", "Trabaja en casa, no trabaja o estudia", "Trabajador profesional", "Ventas".
12	Internet en casa	Factor	2	"No", "Si"
13	Computador en casa	Factor	2	"No", "Si"
14	Número de libros en casa	Factor	3	"0 a 10", "11 a 25", "26 a 100"
15	Tiempo promedio de lectura al día	Factor	2	“30 minutos o menos”, “30 a 60 minutos”
16	Dedicación diaria a internet	Factor	5	"30 minutos o menos", "30 a 60 minutos”, "1 a 3 horas", "Más de 3 horas", "No navega"
17	Número de horas que trabaja a la semana	Factor	5	"0", "[1, 10]", "[11, 20]", "[21, 30]", "más de 30 horas"
18	Nombre del colegio.	Factor	402	"Bachillerato de adultos UNAD", …
19	Género del colegio	Factor	2	"Mixto", “Femenino”
20	Naturaleza del colegio	Factor	2	"Oficial", “No oficial”
21	Calendario académico	Factor	3	"A", “B”, “Otro”.
22	Nivel bilingüe del colegio	Factor	2	"No", “Si”
23	Modalidad de egreso	Factor	4	"Académico", "No aplica", "Técnico", "Técnico/académico"
24	Zona geográfica	Factor	2	"Rural", "Urbano"
25	Jornada de estudio	Factor	6	"Mañana", “tarde”, “noche”, “sabatina”, “única”, “completa”
26	Municipio de ubicación del colegio	Factor	88	"Albán", "Aldana", …
27	Desempeño en lectura crítica	Factor	4	"1", "2", "3", "4"
28	Desempeño en matemáticas	Factor	4	"1", "2", "3", "4"
29	Desempeño en ciencias naturales	Factor	4	"1", "2", "3", "4"
30	Desempeño en sociales y ciudadanas	Factor	4	"1", "2", "3", "4"
31	Desempeño en inglés	Factor	3	"A – ", "A1", "A2"
32	Puntaje del decil global	Factor	4	"4", "5", "6", "7"
33	Nivel socioeconómico del estudiante	Factor	4	"1", "2", "3", "4"
34	Nivel socioeconómico de la institución	Factor	1	"2"
35	Generación de la excelencia	Factor	1	"Gratuidad"
36	Puntaje en lectura crítica	Numérica		
37	Puntaje en matemáticas	Numérica		

N°	VARIABLE	TIPO	NIVELES	DETALLE NIVELES
38	Puntaje en ciencias naturales	Numérica		
39	Puntaje en sociales y ciudadanas	Numérica		
40	Puntaje en inglés	Numérica		
41	Puntaje global	Numérica		

Fuente: esta investigación.

Lo anterior favoreció dividir el archivo *datos_nariño_2021b* en dos documentos, uno denominado *Nariño_imputados_*2021b, en donde se integró a las 35 variables categóricas y cuya utilidad se vio al desarrollar los objetivos específicos 1, 2, y 5 del presente trabajo, y otro denominado *solo_puntajes*, el cual contenía información sobre los resultados de los estudiantes en las pruebas (variables numéricas) y que fue usado para el desarrollo de los objetivos 3 y 4.

2.5 MODELADO.

En esta investigación se consideró importante trabajar con técnicas de minería de datos como clustering para variables mixtas y categóricas, análisis de correlación entre variables, análisis de factores y cruces de variables. Para las variables numéricas se usó análisis de correlación y de factores, y para las variables categóricas se utilizó análisis de clúster y cruces de variables expuestas en tablas de frecuencia. La descripción de dichas técnicas se encuentra en el marco teórico literal 1.73 a 1.7.5 del presente texto.

La modelación de los datos mediante análisis clúster permitió verificar qué tan cohesionados y separados se hallan entre sí los conglomerados formados, hecho que facilitó vislumbrar de mejor manera las características de los estudiantes que presentaban desempeños bajos, básicos, satisfactorios y avanzados en la prueba Saber 11. El análisis factorial permitió encontrar estructuras o ejes fundamentales que recubrían a las variables numéricas, la correlación entre ellas y su relación con el rendimiento en la prueba Saber 11.

3 ANÁLISIS DE RESULTADOS.

3.1 RESUMEN DEL CAPÍTULO.

En este capítulo el lector encontrará el desarrollo de los objetivos específicos propuestos para esta investigación y el cumplimiento de la etapa 5 de la metodología CRISP – DM denominada *evaluación del modelo*. Los resultados se presentan mediante tablas, gráficos o un texto que describe las relaciones halladas. Al final del capítulo se expone la discusión de resultados.

3.2 DESARROLLO DEL OBJETIVO ESPECÍFICO 1.

Conviene recordar que el primer objetivo busca *determinar las características de los estudiantes en lo referente a variables de tipo socioeconómicas, demográficas, familiares, institucionales y de rendimiento en las pruebas,* razón por la cual se elaboró un análisis por variable en donde se expusieron los hallazgos más relevantes de cada una de ellas. Se comienza entonces con el estudio del conjunto de las variables socioeconómicas.

3.2.1 Análisis descriptivo de los aspectos socioeconómicos.

En esta etapa se identificaron 6 variables los cuales aludían al estrato de la vivienda del estudiante, el nivel educativo de los padres, la ocupación laboral de los progenitores y la clasificación del nivel socioeconómico individual (NSE) del educando.

En la categoría referente al estrato se encontró que el 65.48% de los estudiantes, correspondientes a 10879 individuos pertenecían al estrato 1. El 24.47% (4066 educandos) se halló en estrato 2 y el 10.05% (1670 personas) eran de estrato 3, con lo que se resalta que el 89.95% de los colegiales se hallaban en estrato 1 o 2, y que más de la mitad de los estudiantes del departamento tenían como característica principal estar adscritos al estrato 1.

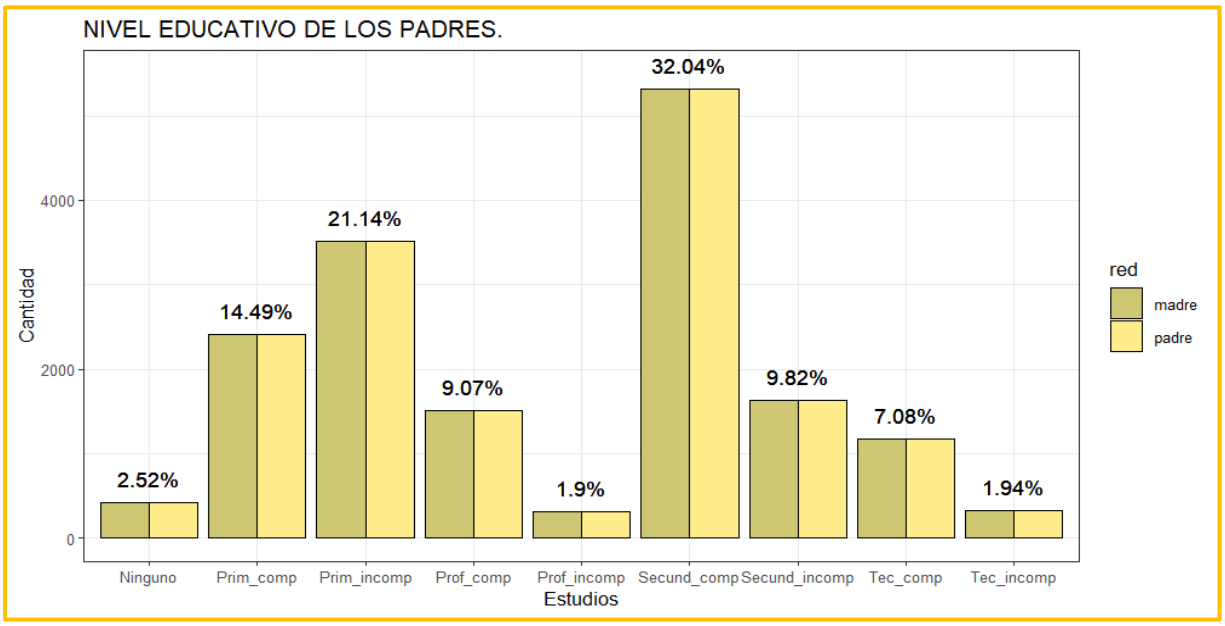
Entre los datos más relevantes al inspeccionar el nivel educativo del padre establecieron que el 41.86% de los papás contaban con educación secundaria (completa o incompleta), 35.63% tenían educación primaria (terminada o sin terminar), y 19.99% afirmaron tener educación postsecundaria (con título o sin título) de donde el 16.15% habían culminados sus estudios técnicos o profesionales. Se observó también que, con el 32.04%, el evento más común fue encontrar padres con secundaria completa y el menos común (1.9%) era hallar padres que no habían culminado sus estudios universitarios.

De igual manera ocurrió con el nivel educativo de la madre, donde el comportamiento de los datos fue el mismo que el observado en los padres, así entonces se distinguió también para ellas que el 41.86% contaban con educación secundaria (completa o incompleta), seguidas por el 35.63% de quienes tenían educación primaria (terminada o sin terminar), y en tercer lugar se ubicaban el 19.99% de las madres con educación postsecundaria incluyendo con título o sin título. Se observó también que el evento más común fue encontrar madres con secundaria completa y el menos común era hallar madres sin culminar estudios universitarios.

La figura 5 expone gráficamente el comportamiento porcentual del nivel educativo de los padres.

Figura 5.

Nivel educativo de los padres.



Fuente: esta investigación.

Por otro lado, en la ocupación laboral del padre se destacó que el 32.81% de ellos se dedicaban a la agricultura, pesca o trabajo al jornal, siendo el evento más representativo. En segundo lugar, aunque con menor porcentaje de representación (9.88%) lo ocuparon los padres que trabajaban como independientes, seguidos por el 8.11% de quienes laboraban en casa, no trabajan o estudiaban. Se encontró también que solo el 1.1% eran pensionados.

Entre las ocupaciones más comunes desempeñadas por las madres se encontró, con el 45.95%, a quienes ejercían el rol de amas de casa, o quienes no trabajaban o estudiaban. El 10.51% se dedicaban a labores de agricultura, pesca o trabajo al jornal. Se observó también que el 9.05% eran microempresarias, 7.33% se dedicaban a las ventas, 6.05% eran trabajadoras profesionales y un 6.01% eran empleadas de servicios generales. El empleo menos común visualizado en este grupo fue el de conductoras de vehículos (0.23%). La tabla 8

presenta los porcentajes y frecuencias encontradas para la ocupación laboral del padre y la madre.

Tabla 8.

Ocupación laboral de los padres.

	TRABAJO DEL PADRE		TRABAJO DE LA MADRE	
DESCRIPCIÓN	**CANTIDAD**	**PORCENTAJE (%)**	**CANTIDAD**	**PORCENTAJE (%)**
No aplica	1105	6.65%	597	3.59%
Agricultor, pesquero, jornalero	5451	32.81%	1747	10.51%
Microempresario	1139	6.86%	1503	9.05%
Conductor de vehículos	1296	7.80%	38	0.23%
Oficios varios o independiente.	1641	9.88%	460	2.77%
Trabajador profesional	919	5.53%	1006	6.05%
Trabaja en casa, no trabaja, o estudia	1347	8.11%	7634	45.95%
Auxiliar administrativo	550	3.31%	913	5.50%
Empleado de servicios generales	859	5.17%	999	6.01%
No sabe	1110	6.68%	291	1.75%
Ventas	829	4.99%	1218	7.33%
Pensionado	182	1.10%	49	0.29%
Gran empresario o directivo	187	1.13%	160	0.96%
TOTAL	**16615**	**100%**	**16615**	**100%**

Fuente: esta investigación.

En cuanto al NSE individual del estudiante se encontró que el 43.35% de ellos se clasificaban en nivel 2, seguidos por el 31.02% que se posicionaron en nivel 1, el 21.96% se categorizaron en nivel 3 y el 3.67% se ubicaron en nivel 4.

- **Conclusiones de las características socioeconómicas.**

El análisis de las características socioeconómicas permitió establecer que los estudiantes del departamento de Nariño en el periodo 2021 – B, se caracterizaron por:

- Contar con el 89.95% de los estudiantes clasificados en estrato 1 o 2, siendo el estrato 1 el evento más frecuente.
- Estar adscritos a núcleos familiares donde el nivel educativo del padre y la madre presentaban, como rasgos más destacados, contar con estudios secundarios culminados como mayor referente, seguidos por quienes tenían primaria incompleta y primaria completa.
- Contar con padres mayormente dedicados al trabajo agrícola, a la pesca o al jornal, seguidos por quienes laboraban como independientes.
- Contar con madres que en su gran mayoría ejercían labores del hogar (amas de casa), no tenían trabajo o se hallaban estudiando, o por madres dedicadas a ejercer labores de agricultura, pesca o que trabajo al jornal.
- Clasificarse en NSE 2 (evento más común), seguido por NSE 1.

3.2.2 Análisis descriptivo de los aspectos familiares.

Las características familiares integran 5 variables concernientes al número de personas presentes en el hogar, cantidad de cuartos en la vivienda, presencia de internet y computador en las casas, y, cantidad de libros existentes en cada residencia.

En cuanto al número de personas en el hogar se detectó que el 56.35% de las familias de los estudiantes se integraban por 3 o 4 personas, 28.47% por 5 a 6 individuos, 7.88% por 7 a 8 sujetos, y el 7.29% daba cuenta de 1 a 2 personas. Por otro lado, referente al número de libros se halló que el 57.11% de los estudiantes tenían entre 0 a 10 libros en casa, de 11 a 25 el 28.80%, y el 14.10% contaban con 26 a 100 textos.

En el mismo orden de ideas se detectó que el 63.73% de los aprendices contaban con acceso a internet mientras que para el 36.27% este no era un evento favorable. El 59.29%

manifestó no tener computador, acto contrario al 40.71% que afirmó contar con dicho artefacto. Por último, la variable número de cuartos en el hogar reflejó que el 38.47% de los educandos contaban con tres alcobas en su vivienda, dos el 33.48%, cuatro el 14.52%, una el 5.45%, cinco el 5.13%, y tan solo el 2.94% contaba con seis o más cuartos en sus viviendas.

- **Conclusiones de las características familiares.**

El análisis de los aspectos familiares permitió establecer que los estudiantes del departamento de Nariño en el periodo 2021 – B, se caracterizaron porque:

- En la mayoría de las viviendas había dos o tres habitaciones.
- Los núcleos familiares se integraban por 3 a 4 personas en la mayoría de los casos, seguidos por las familias con 5 a 6 individuos.
- Más de la mitad de los educandos tenían en casa un máximo de 10 libros (57.11%), contaban con cobertura de internet (63.73%), pero no disponían de computador (59.29%).

3.2.3 Análisis descriptivo de los aspectos demográficos.

Las características demográficas aunaron 8 variables las cuales tratan sobre el género del discente, su edad, la pertenencia a grupos étnicos, el municipio de residencia, el tiempo dedicado a realizar actividades de lectura diaria y a navegar en internet. Contempló además las horas que el estudiante trabajaba entre semana y si fue clasificado como generación de la excelencia o estuvo en otro nivel.

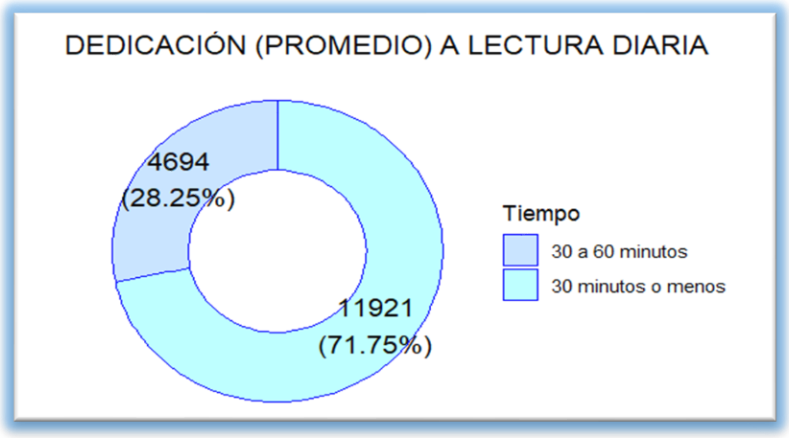
En cuanto al género se obtuvo que, el 43.65% de los estudiantes eran hombres y el 56.35% eran mujeres, notando así mayor presencia del sector femenino. En lo referente a la edad se halló que el 63.73% eran menores de edad, es decir tenían menos de 18 años, el

33.67% tenían entre 18 y 22 años y tan solo el 2.6% superaban los 22 años. De la pertenencia a grupos étnicos se evidenció que el 74.58% no lo hacía mientras que el 25.42% sí.

Agregado a lo anterior, se evidenció que el 71.75% de los estudiantes dedicaban a la lectura por diversión, un tiempo promedio de 30 minutos o menos, y el 28.25% hacía lo propio en un lapso entre 30 a 60 minutos. La figura 6 destaca el tiempo promedio de los estudiantes al realizar actividades de lectura por diversión.

Figura 6.

Tiempo promedio de lectura diaria.



Fuente: Esta investigación.

Del tiempo promedio diario de navegación en internet se obtuvo que el 34.04% lo hacía entre 30 a 60 minutos, el 25.71% pasaba entre 1 a 3 horas, el 19.28% empleaba 30 minutos o menos, se halló también que el 14.66% invertía más de tres horas en internet y el 6.31% no navegaba en la web.

Otro aspecto característico que logró evidenciarse fue que el 54.58% de los estudiantes no ejercían actividad laboral alguna, a diferencia del 27.74% que dedicaban entre 1 a 10 horas

semanales. El 10.66% trabajaba entre 11 a 20 horas, 3.65% entre 21 a 30 horas, y el 3.37% lo hacía en más de 30 horas. En cuanto a la variable generación E se notó que la mayoría de los estudiantes pertenecía a la categoría denominada *gratuidad*.

Del municipio de residencia del estudiante se logró precisar que el 27.88% de discentes habitaban en Pasto, 10.83% en Tumaco y 9.76% en Ipiales, suceso que indicó que estas tres ciudades aglomeraban al 48.47% del total poblacional, evento considerable ya que representó un valor cercano a la mitad de colegiales del departamento. Se observó también que en los demás municipios habitaban menos del 2.78% de personas, ya que, en Cumbal residían el 2.77%, en Túquerres el 2.71% y en la Unión el 2.59% como datos más relevantes. El municipio donde habitaban menos educandos fue Mosquera con el 0.17% lo que equivalía a 29 personas.

- **Conclusiones de las características demográficas.**

El análisis de la variable demográfica permitió establecer las siguientes características.

- Las mujeres conformaban el grupo mayoritario de la población, enfrentando al 56.35% de ellas contra el 43.65% de hombres, aspecto que invita a reflexionar en torno a la relación entre género y desempeño en matemáticas, suceso que se inspeccionará en el cruce de variables.
- El 63.73% de los colegiales tenía menos de 18 años al presentar la prueba, el 33.67% oscilaba entre los 18 a 22 años y solo el 2.6% eran mayores a 22 años.
- El evento más común fue hallar estudiantes que no pertenecían a grupos étnicos
- Más de la mitad de discentes no trabajaban entre semana, el 27.74% laboraban entre 1 a 10 horas y el 10.66% lo hacía entre 11 a 20 horas.
- Más del 95% de educandos hacían parte de la categoría denominada *gratuidad*.
- El 71.75% de los educandos no leía más de 30 minutos al día.

- El tiempo promedio de navegación en internet más común oscilaba entre 30 a 60 minutos seguido por quienes empleaban de 1 a 3 horas.
- Los municipios con mayor número de estudiantes fueron Pasto, Tumaco e Ipiales, y el de menor número fue Mosquera.

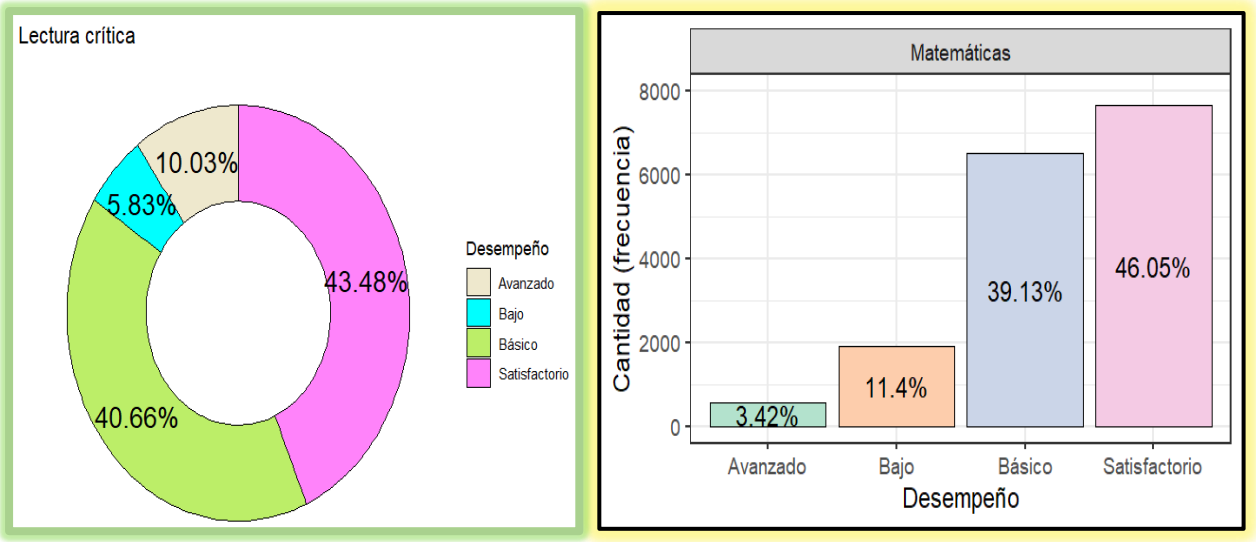
3.2.4 Análisis descriptivo del rendimiento en las pruebas.

En el rendimiento en las pruebas se analizaron los desempeños alcanzados por los estudiantes en lectura crítica, matemáticas, ciencias naturales, sociales ciudadanas, inglés y el decil global.

En el desempeño en lectura crítica se evidenció que el 5.83% se ubicó en nivel *bajo*, *básico* el 40.66%, 43.48% en *satisfactorio* y 10.03% en *avanzado*. En la prueba de matemáticas se halló que el 11.40% de las personas obtuvieron desempeño *bajo*, *básico* el 39.13%, *satisfactorio* el 46.05% y 3.42% en *avanzado*. Por ende, tanto en lenguaje como en matemáticas el evento más común fue hallar discentes en nivel satisfactorio, seguidos por quienes se encontraban en nivel básico. Se evidenció también que no era común distinguir estudiantes en nivel avanzado, sobre todo en matemáticas en donde el porcentaje de representación fue menor al 3.5%. La figura 7 presenta de forma gráfica los porcentajes obtenidos por los estudiantes en cada nivel de desempeño en las pruebas de *lectura crítica* y *matemáticas*.

Figura 7.

Desempeños en lectura crítica y matemáticas.



Fuente: esta investigación.

Para ciencias naturales, el 22.95% se ubicó en desempeño *bajo*, 53.08% en *básico*, *satisfactorio* el 22.25%, y 1.72% en *avanzado*. En sociales y ciudadanas, el 33.96% obtuvo desempeño *bajo*, *básico* el 42.53%, *satisfactorio* el 21.60%, y 1.91% en *avanzado*. En inglés, el 58.97% de los estudiantes obtuvo un puntaje de “A –”, siendo este el evento más común, seguido por el 28.60% del nivel “A1” y 12.43% de “A2”. Finalmente, con referencia al decil global, se observó que el 38.39% de los educandos se clasificaron en decil cinco, el 30.29% en decil seis, 19.27% en decil cuatro y 12.05% en decil siete.

- **Conclusiones del rendimiento en las pruebas.**

En esta etapa, los estudiantes se caracterizaron por lo siguiente.

- El desempeño *satisfactorio* fue el más sobresaliente para las pruebas de lectura crítica y matemáticas, seguido de cerca por quienes se clasificaron en *básico*.

- En las pruebas de ciencias naturales y sociales ciudadanas se encontró que, el nivel *básico* fue el puntaje más representativo obtenido por los estudiantes, seguido por quienes se clasificaron en desempeño *bajo*.
- Más de la mitad de los discentes obtuvo nivel “A –” en inglés.
- El porcentaje de individuos clasificados en nivel *avanzado* es escaso, siendo *lectura crítica* la de mayor ponderación (5.83%), seguido por *matemáticas* (3.42%), *sociales ciudadanas* (1.91%) y ciencias *naturales* (1.72%).
- Los resultados de las pruebas favorecieron que el 68.68% de los evaluados se ubicasen en decil 5 o en 6 como eventos más representativos, y en donde el 38.39% daba cuenta de los estudiantes del decil 5 y el 30.29% de los del decil 6.

3.2.5 Análisis descriptivo de los aspectos institucionales.

Los aspectos institucionales se integraron por variables como el género del colegio, su naturaleza (oficial o privado), tipo de calendario, estado bilingüe o no, jornada de trabajo, carácter o modalidad de egreso, y área de ubicación (rural o urbano).

El análisis de los datos permitió determinar que el 96.15% de las instituciones educativas eran mixtas y el 3.85% de corte femenino. Públicas el 88.5% y privadas el 11.5%. 73.46% se ubicaban en el casco urbano mientras que 26.54% en zona rural. 96.06% no eran bilingües mientras que el 3.94% sí lo eran. De igual forma se determinó que el 54.75% tenían modalidad académica, 30.7% técnica/académica, educación técnica el 13.40% y para el 1.15% no aplicó a este ítem. Se evidenció también que el 99.58% de las instituciones tenían calendario “A”, 0.38% tipo “B”, y 0.04% otro calendario.

Agregado a lo anterior se halló que, el 74.32% de los estudiantes abordaban sus encuentros de clase en la jornada de la mañana, 11.27% en jornada única, en la tarde el

8.23%, sabatina el 3.69%, el 2.47% en la noche y 0.03% en jornada completa. Del NSE institucional se relata que el 66.80% de los planteles se catalogó en nivel 2, 18.94% en nivel 1, 13.11% en 3 y tan solo el 1.15% se categorizó en nivel 1.

- **Conclusiones de los aspectos institucionales.**

Los aspectos institucionales permitieron establecer que los estudiantes se caracterizaban porque en su mayoría abordaban acciones educativas en instituciones de género mixto, oficial, no bilingüe, pertenecientes a la zona urbana, con calendario "A" y jornada de la mañana. Además, la modalidad de egreso más frecuente era académico o técnico/académico. Se halló también que el nivel socioeconómico más común de las instituciones educativas fue el NSE 2 seguido por el NSE 1. Estos dos niveles acogieron al 85.74% de las escuelas.

3.3 DESARROLLO DEL OBJETIVO ESPECÍFICO 2.

Analizar la relación entre el nivel alcanzado en matemáticas y las variables de tipo socioeconómicas, demográficas, familiares, institucionales y de rendimiento en las pruebas.

Para dar cumplimento a este objetivo se realizó un cruce entre cada variable y el desempeño conseguido por los estudiantes en la prueba de matemáticas. Los resultados fueron registrados en tablas de contingencia en donde cada variable se describió por filas mientras que en las columnas se ubicó al desempeño en matemáticas, hecho que motivó a realizar análisis de frecuencias mediante perfil fila. Además, se estudió el grado de asociación entre variables categóricas a partir de una prueba Chi cuadrado, la cual al 5% de significancia y un p-valor de $2.2 * e^{-16}$, denotó que las variables se encontraban relacionadas. Se calculó también el *coeficiente de contingencia, la V de Cramer* y la *Tau* de Kendall, a fin de establecer

el grado de asociación efectuado en cada cruce. Los hallazgos más importantes por variable destacan lo siguiente.

3.3.1 Cruce de los aspectos socioeconómicos con el desempeño en matemáticas.

De la relación existente entre estrato y desempeño en la prueba cabe decir que, de los estudiantes adscritos al estrato 1, el 44.2% obtuvieron desempeños en nivel *satisfactorio*, 40.59% en nivel *básico*, 12.73% en *bajo* y 2.49% en *avanzado*. En el estrato 2 se incrementó la representación de estudiantes en nivel *satisfactorio* (50.5%) y aumentó también el porcentaje de estudiantes con desempeño *avanzado*, pasando del 2.5% del estrato 1 al 4.9% en estrato dos. En el estrato 3 destacaron, con el 47.5%, los estudiantes ubicados en nivel satisfactorio y se notó que el porcentaje de discentes en nivel avanzado subió de 4.9% del estrato dos a 5.9%. De este cruce se concluyó que la relación entre estrato y desempeño en la prueba indicó que, a mayor estrato de los estudiantes se evidenciaron mejores desempeños en matemáticas. Sin embargo, fueron pocas las personas que lograron clasificarse en nivel avanzado. Por otro lado, las medidas de asociación lograron establecer una relación débil entre las variables, ya que el coeficiente de contingencia obtuvo un valor de 0.11, la V de Cramer de 0.08 y la prueba de Kendall de 0.08.

Al cruzar el nivel educativo del padre con el desempeño en matemáticas se halló que, conforme mejoraba el grado de estudios del padre, los resultados obtenidos en las pruebas también mejoraban, ya que los hijos de padres sin ningún estudio obtenían con mayor frecuencia desempeños en nivel bajo o básico, mientras que los estudiantes cuyos progenitores contaban con estudios profesionales (completos o incompletos) y técnicos (terminados o sin terminar), adquirían mayor representación porcentual en los niveles *satisfactorios y avanzados*. Se observó también que los hijos de padres con educación primaria (completa o incompleta), y secundaria (culminada o sin culminar) acaparaban mayor representación en los desempeños

básicos y satisfactorio. El coeficiente de contingencia asociado al nivel educativo de los padres, obtuvo un valor de 0.20, la V de Cramer de 0.12 y la prueba de Kendall de 0.08, indicando asociación débil entre las variables.

Con relación al cruce entre el nivel educativo de la madre y el desempeño de los educandos en matemáticas, se obtuvo un suceso igual al anterior ya que los resultados fueron idénticos tanto en las tablas de frecuencia como en las medidas de asociación. Es decir, los valores expuestos en la tabla 9 fueron los obtenidos en el cruce del desempeño en matemáticas de los estudiantes con el nivel educativo tanto del padre como de la madre.

Tabla 9.

Cruce del nivel educativo de los padres y el desempeño en matemáticas.

Estudios de los padres. (Padre y madre).	**Nivel de desempeño en matemáticas**								**TOTAL**
	BAJO		**BÁSICO**		**SATISFACTORIO**		**AVANZADO**		
	Frecuencia	**%**	**Frecuencia**	**%**	**Frecuencia**	**%**	**Frecuencia**	**%**	**Frecuencia**
Ninguno	122	29.2%	191	45.7%	101	24.2%	4	1.0%	**418**
Primaria incompleta	422	12.0%	1651	47.0%	1383	39.4%	56	1.6%	**3512**
Primaria completa	269	11.2%	1042	43.3%	1055	43.8%	41	1.7%	**2407**
Secundaria incompleta	213	13.1%	680	41.7%	710	43.5%	29	1.8%	**1632**
Secundaria completa	590	11.1%	1957	36.8%	2552	47.9%	225	4.2%	**5324**
Técnico incompleto	40	12.4%	100	31.0%	173	53.6%	10	3.1%	**323**
Técnico completo	70	6.0%	336	28.6%	704	59.9%	66	5.6%	**1176**
Profesional incompleta	36	11.4%	104	32.9%	152	48.1%	24	7.6%	**316**
Profesional completa	132	8.8%	440	29.2%	822	54.5%	113	7.5%	**1507**
TOTAL	**1894**	**11.4%**	**6501**	**39.1%**	**7652**	**46.1%**	**568**	**3.4%**	**16615**

Fuente: esta investigación.

Por otro lado, al compararse el desempeño de los estudiantes en matemáticas con la ocupación laboral del padre se encontró que, a mejor estatus laboral del padre, mejores desempeños obtenían los colegiales. Así entonces, el 43.1% de los estudiantes cuyos padres eran agricultores se clasificaron en *básico*, seguido por el 42.8% del nivel satisfactorio, 12.1%

del nivel *bajo* y 2% de *avanzado*. El 51.8% de los educandos cuyos padres eran auxiliares administrativos se clasificaron en *satisfactorio* y el 12% en nivel *bajo*. El 54.2% de los hijos de padres conductores de vehículos se ubicaron en *satisfactorio*, 36.7% en *básico,* 5.4% en *bajo*, y 3.7% en *avanzado*. El 44.4% de los hijos de padres empleados en servicios generales destacaron en *satisfactorio*, 42.1% en *básico* y 4% en *avanzado*. En *satisfactorio* se ubicaron también, como nivel más destacado, el 49.2% de los hijos de los empresarios, 52.2% de los microempresarios, 50.7% de los independientes, 57.1% de los pensionados, 54.3% de los profesionales, y 43.7% de los de ventas. Se observó además que el 43.4% de los estudiantes con padres dedicados a labores del hogar se clasificaron en nivel *básico* como dato más relevante, y que los hijos de padres profesionales tenían la representación más alta (11.9%) de todas las ocupaciones laborales en nivel *avanzado*.

El coeficiente de contingencia para este cruce obtuvo un valor de 0.21, la V de Cramer de 0.13 y la prueba de Kendall de 0.03, indicando asociación débil entre las variables.

Al contrastar la ocupación laboral de la madre con el desempeño en matemáticas de los estudiantes se encontró que, en nivel *básico* se encontraban, con mayor frecuencia, los hijos de madres dedicadas a la agricultura (46.4%). El nivel *satisfactorio* lo ocuparon el 54.4% de los hijos de las madres auxiliares administrativas, 44.7% de las choferes, 53.2% de las microempresarias, 45.9% de las independientes y 55.9% de las profesionales. En las ocupaciones de servicios generales, empresarias, pensionadas, amas de casa y mujeres dedicadas a las ventas, pese a que los estudiantes se clasificaron en nivel *satisfactorio*, no se vio una marcada diferencia con quienes se ubicaron en nivel *básico*, ya que, para la variable servicios generales, el 46.1% de los colegiales estaba en *satisfactorio* mientras que el 41% en *básico*; en la variable empresarias, el 36.9% se posicionó en *satisfactorio* mientras el 32.5% en *básico*; las madres amas de casa dejaron entrever que el 45.4% de sus hijos se hallaba en

satisfactorio mientras el 40.2% en básico; en ventas, el 46% en satisfactorio y el 40.3% en básico. Las madres pensionadas dejaron entrever que el 34.7% de sus hijos se hallaba en *básico* y en igual porcentaje se encontraban los de *satisfactorio*. En nivel *avanzado* tomaron prioridad el 11.1% de los hijos de madres profesionales, de igual manera sobresalieron en este nivel, con el 5% de representación, los hijos de mujeres empresarias y el 4.6% de los hijos de las auxiliares administrativas.

El coeficiente de contingencia para este cruce obtuvo un valor de 0.18, la V de Cramer de 0.11 y la prueba de Kendall de 0.04, indicando asociación débil entre las variables.

El cruce entre el nivel socioeconómico individual del estudiante y su desempeño en matemáticas mostró que, en nivel satisfactorio sobresalieron el 44.9% de los estudiantes con NSE2, el 57.4% con NSE3 y el 59.7% con NSE4. Se vio también que, en nivel *avanzado*, la representación porcentual más alta (16.7%) correspondió a quienes tenían NSE4 y el mayor porcentaje de individuos en nivel *bajo* correspondió a quienes tenían NSE1 (13.8%). Las medidas de asociación establecieron que, el coeficiente de contingencia obtuvo un valor de 0.24, la V de Cramer de 0.14 y la prueba de Kendall de 0.18, indicando asociación débil entre las variables.

- **Conclusiones del presente cruce.**

Tras el análisis de los resultados se verificó la bibliografía referente a *capital cultural* y demás artículos relacionados con rendimiento académico en donde se expone que, a mejores condiciones sociales, labores y económicas, crece la posibilidad de obtener mejores resultados en la prueba. Sin embargo, son pocos los estudiantes nariñenses que consiguieron desempeño avanzado en matemáticas.

Se observó también que la mayoría de estudiantes en todos los estratos, se clasificaban en desempeño *básico o* en *satisfactorio*, siendo el nivel *satisfactorio* el de mayor prevalencia. En el estrato 3 se halló la mayor representación porcentual de educandos con desempeño *avanzado* en matemáticas.

Al cruzar el nivel educativo tanto del padre como de la madre se halló que, los hijos de progenitores con mayor escolaridad presentaron mejores desempeños en la prueba, y los hijos de padres profesionales lograron clasificarse con mayor facilidad en nivel *avanzado*.

En cuanto a la ocupación laboral del padre se encontró que, los hijos cuyos padres ejercían la agricultura, o, no trabajaban, o, se dedicaban a labores del hogar, se clasificaron mayormente en nivel *básico*, mientras que en los demás trabajos sobresalió el nivel *satisfactorio*. Además, los hijos de padres con empleos profesionales se clasificaron mayormente en nivel *avanzado*. Suceso similar ocurrió al examinar la ocupación laboral de la madre, en donde los hijos de madres dedicadas a la agricultura se posicionaron con mayor frecuencia en nivel *básico*, mientras que para las demás profesiones se observó que los colegiales se clasificaban, generalmente, en nivel *satisfactorio*. Las madres pensionadas dejaron entrever que el porcentaje de sus hijos catalogado en *básico* fue igual al de s*atisfactorio*.

En el estudio del NSE se halló que, las personas con nivel socioeconómico 1 se posicionaron mayormente en desempeño *básico* en la prueba, mientras que para los demás NSE sobresalió el desempeño *satisfactorio*, notando también que en NSE4 se encontró la mayor representación de estudiantes con desempeño *avanzado*. Por último, las medidas de asociación establecieron una relación débil entre las variables estudiadas y el desempeño en matemáticas, siendo la variable NSE la de mayor asociación.

3.3.2 Cruce de los aspectos familiares con el desempeño en matemáticas.

Al compararse el número de personas en el hogar con el desempeño logrado en la prueba de matemáticas se advirtió que, en nivel básico sobresalieron las familias compuestas por 1 a 2 sujetos, ya que aquí el 41.3% de los educandos obtuvieron desempeños en nivel *básico,* seguidos por el 40.3% de los de nivel *satisfactorio.* De igual forma ocurrió con las familias integradas por 7 a 8 individuos, en donde el porcentaje de colegiales en nivel básico fue de 44.1% y el de *satisfactorio* fue el 35.6%. En nivel *satisfactorio* destacaron, con el 49.2%, las familias integradas por 3 a 4 miembros, y con el 44.2% las de cinco a seis sujetos. Se notó también que a medida que las familias superaban los 4 integrantes, la relación de ser clasificado en nivel *básico* también aumentaba. En nivel *avanzado* aparecieron el 4% de las familias con uno a dos miembros y el 3.8% de las compuestas por tres a cuatro sujetos. El coeficiente de contingencia obtuvo un valor de 0.11, la V de Cramer de 0.06 y la prueba de Kendall de -0.05, indicando asociación débil entre las variables.

Por otro lado, la relación entre número de cuartos en el hogar y desempeño en la prueba reflejó que, el 41.5% de los estudiantes se hallaban en nivel *bajo* y el 41.1% en nivel *satisfactorio.* En nivel *bajo* sobresalieron también el 42.3% de educandos cuyas casas tenían 5 habitaciones y con igual porcentaje se encontró a los estudiantes cuyos hogares contaban con 6 cuartos. Las personas en nivel *satisfactorio* se caracterizaron porque el 49.4% tenían dos habitaciones, 46.8% tenían 3 y el 42.5% contaba con 4 alcobas. Los educandos del nivel *avanzando* mostraron como datos más relevantes que en sus casas, el 3.2% tenía dos habitaciones y el 3.2% solo una. En cuanto a las medidas de asociación, el coeficiente de contingencia obtuvo un valor de 0.1, la V de Cramer de 0.06 y la prueba de Kendall de 0.013, indicando asociación débil entre las variables.

La relación entre tener *internet en casa* y desempeño en matemáticas identificó que, el 51.4% de los estudiantes que sí tenían internet se clasificaron en *satisfactorio*, 35.9% se categorizaron en *básico,* 8.2% en *bajo* y 4.5% en *avanzado*. De los colegiales que manifestaron no contar con internet en sus hogares se notó que el 44.8% se clasificaron en *básico*, 36.6% en *satisfactorio,* 16.9% en *bajo* y 1.6% en *avanzado*. De lo anterior, se concluyó que las personas con internet obtuvieron mejores resultados. Con referencia a las medidas de asociación, el coeficiente de contingencia tuvo un valor de 0.19, la V de Cramer de 0.19 y la prueba de Kendall de 0.18, indicando asociación débil entre las variables.

El cruce *computador en casa* y desempeño en matemáticas permitió establecer que el 54.8% de los educandos que sí tenían computador se clasificaron en *satisfactorio*, 31.7% en *básico*, 7.4% en *bajo* y 6.1% en *avanzado*. Por otro lado, el 44.2% de las personas que manifestaron no tener computador en casa obtuvieron desempeños en nivel *básico*, 40.1% en *satisfactorio*, 14.2% en *bajo* y 1.6% en *avanzado*, razón por la que se determinó que quienes tenían computador en sus viviendas exhibieron mejores desempeños en la prueba. El coeficiente de contingencia obtuvo un valor de 0.21, la V de Cramer de 0.21 y la prueba de Kendall de 0.2, siendo esta asociación más fuerte que la inmediatamente anterior.

Finalmente, al inspeccionar en la relación entre cantidad de libros existentes en el hogar y desempeño en la prueba se encontró que, el 43% de los educandos con menos de 10 textos se clasificaron en *básico*, 41.3% en nivel *satisfactorio*, 13.4% en nivel *bajo* y 2.3% en *avanzado.* Para las personas con 11 a 25 libros en casa, el 50.2% obtuvieron desempeños *satisfactorios*, 36.6% en *básico,* 9.6% en *bajo* y 3.7% en *avanzado*. Los discentes con más de 26 textos en su hogar, mostraron que el 56.9% se posicionaron en *satisfactorio*, 28.8% en *básico*, 7.3% en *avanzado* y 7% en *bajo*; con lo que se notó que, a mayor número de libros en casa, mejores resultados se obtuvieron en la prueba. En cuanto a las medidas de asociación, el coeficiente de

contingencia tuvo un valor de 0.16, la V de Cramer de 0.12 y la prueba de Kendall de 0.14, indicando asociación débil entre las variables.

- **Conclusiones del presente cruce.**

Al comparar las variables de los aspectos familiares con el desempeño en matemáticas se determinó que, los estudiantes cuyas familias se integraban por 1 a 2 personas, o, por 7 a 8 individuos, obtenían con mayor frecuencia puntajes en nivel *básico*, mientras que los colegiales cuyos hogares conformados por 3 a 4 personas, o, 5 a 6 sujetos, presentaban mayormente desempeños *satisfactorios*. El mayor porcentaje de estudiantes con nivel avanzado se vio en las familias de 1 a 2 personas.

Del número de habitaciones en casa se detectó que los educandos con uno, cinco o seis cuartos, se clasificaron con mayor frecuencia en nivel *bajo*, mientras que quienes contaban con dos, tres o cuatro habitaciones, se clasificaron mayormente en nivel *satisfactorio*. Por otro lado, más de la mitad de colegiales que contaban con internet en sus viviendas lograron clasificarse en nivel *satisfactorio*, mientras que quienes no tenían internet en casa se clasificaban con mayor frecuencia en desempeño *básico*. Además, los estudiantes con internet en casa obtuvieron mayor representación en nivel avanzado frente aquellos que no disponían de ello.

Se observó también que más de la mitad de estudiantes con computador en casa se categorizaron en desempeño *satisfactorio*, mientras que quienes no contaban con dicho artefacto se clasificaron mayormente en nivel *básico*. Se notó también que, los estudiantes con computador en casa obtuvieron mayor representación en nivel avanzado frente aquellos que no contaban con este artefacto. Con referencia a la cantidad de libros existentes en el hogar del estudiante y su desempeño en matemáticas se encontró que, las personas con menos de 10

textos se clasificaron con mayor frecuencia en nivel *básico*, mientras que quienes tenían de más de 11 libros se clasificaron, generalmente, en nivel *satisfactorio*, notando así que los colegiales con mayor número de libros en casa, presentaban, en términos generales, mejores desempeños en la prueba.

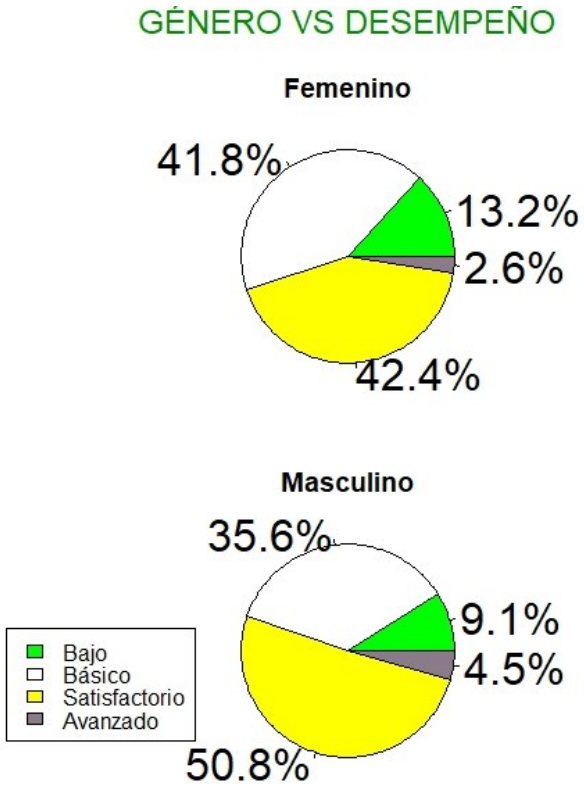
Las medidas de asociación permitieron establecer que, de los aspectos familiares analizados, las variables internet y computador en casa, presentaron mejor asociación con el desempeño en matemáticas.

3.3.3 Cruce de los aspectos demográficos con el desempeño en matemáticas.

Al cruzar el género del estudiante con el desempeño en la prueba se halló que los hombres presentaban mejores resultados que las mujeres, ya que el 50.8% de ellos se clasificó en nivel *satisfactorio*, 35.6% en *básico*, 9.1% en *bajo* y 4.5% en *avanzado*; mientras que para las mujeres se observó que el 42.4% se clasificaron en *satisfactorio*, 41.8% en *básico*, 13.2% en *bajo*, y 2.6% en *avanzado*. Así entonces, los hombres tenían mayor representación en los desempeños altos mientras que las mujeres lo hacían en *básico o bajo*. Las medidas de asociación reflejaron que, el coeficiente de contingencia obtuvo un valor de 0.11, la V de Cramer de 0.11 y la prueba de Kendall de 0.10, indicando asociación débil entre las variables. La figura 8 presenta la clasificación de los estudiantes en la prueba de matemáticas, según su género.

Figura 8.

Cruce entre género y desempeño en matemáticas.



Fuente: esta investigación.

Al cruzar la variable edad con el desempeño en la prueba se halló que el 54.2% de las personas menores a 18 años consiguieron nivel *satisfactorio* en la prueba, 34.4% nivel *básico*, 6.4% *bajo* y 4.9% *avanzado*. Para los individuos cuya edad oscilaba entre 18 y 22 años, el 47.6% se clasificó en *básico*, 33.7% en *satisfactorio*, 17.9% en *bajo* y 0.9% en *avanzado*. Y el 49.3% de quienes tenían más de 22 años obtuvieron desempeño *bajo*, 44.4% *básico*, 6.2% *satisfactorio* y no se tuvieron registros de este grupo de edad en nivel *avanzado*. Las medias de asociación destacaron que el coeficiente de contingencia tuvo un valor de 0.31, la V de Cramer de 0.23 y la prueba de Kendall de 0.17, indicando asociación no tan fuerte entre las variables.

Al contrastar la pertenencia a grupos étnicos con desempeño en matemáticas se determinó que, el 51.6% de los estudiantes sin etnia obtuvieron desempeños *satisfactorios*, 37.5% nivel *básico*, 6.9% nivel *bajo* y 4% *avanzado*; mientras que el 44% de los colegiales adscritos a grupos étnicos se clasificaron en nivel *básico*, 29.7% *satisfactorio*, 24.6% *bajo* y 1.7% en *avanzado*, con lo que se observó que el no tener etnia favoreció la obtención de mejores desempeños. Las medidas de asociación para este cruce denotaron que el coeficiente de contingencia obtuvo un valor de 0.27, la V de Cramer de 0.28 y la prueba de Kendall fue -0.24, indicando asociación no tan fuerte entre las variables. Se resalta que el test de Kendall señaló una relación inversa, es decir a medida que se aumenta la pertenencia a grupos étnicos, se disminuye el desempeño en matemáticas.

La relación entre *tiempo promedio de lectura diaria y desempeño en matemáticas* dejó entrever que el 44% de quienes leían menos de 30 minutos al día se clasificaron en *satisfactorio*, 40.7% en *básico*, 12.5% en *bajo* y 2.9% en *avanzado*; mientras que el 51.3% de quienes leían entre 30 a 60 minutos se clasificaron en nivel *satisfactorio*, 35.2% en *básico*, 8.7% en *bajo* y 4.8% en *avanzado*. El coeficiente de contingencia tuvo un valor de 0.09, la V de Cramer de 0.09 y la prueba de Kendall de 0.09, indicando asociación débil entre las variables.

Sobre el tiempo promedio diario que pasaban los estudiantes en internet y su relación con el desempeño en matemáticas se encontró que, quienes no navegaban en la red o pasaban menos de 30 minutos en internet, fueron clasificados mayoritariamente en nivel *básico* obteniendo representaciones del 45.8% para los que no navegaban y del 43.8% para los menos de 30 minutos. Se observó también que las posibilidades de clasificarse en nivel *satisfactorio* aumentaban conforme incrementaba el tiempo de navegación representando esto el 45.6% para quienes navegaban de 30 a 60 minutos, 51.6% para los de una a tres horas y 52.8% para los tiempos mayores a tres horas. El porcentaje con mayor representación en nivel

avanzado se observó para quienes invertían más de tres horas en la red y correspondió al 5.4% de educandos. Por otro lado, el coeficiente de contingencia tuvo un valor de 0.17, la V de Cramer de 0.1 y la prueba de Kendall de -0.04, indicando asociación débil entre las variables.

Del número de horas que trabajaban los estudiantes entre semana y su desempeño en la prueba se observó que el 49.8% de quienes no laboraban se clasificaron en *satisfactorio*, 36.2% en básico y el 4.6% en *avanzado*. El 43.6% de quienes laboraban entre 1 a 10 horas se clasificaron en nivel *básico,* seguidos por el 40.1% de *satisfactorio* y el 2.2% de *avanzado*. Para quienes trabajan entre 11 a 20 horas, el nivel avanzado fue alcanzado por el 2%, 45.2% logró nivel *satisfactorio*, 39.4% *básico* y 13.4% *bajo*. Estos hechos indicaron que, a más horas de trabajo entre semana, menos estudiantes se clasificaban en *avanzado*. Igual suceso ocurrió con quienes laboran más de 30 horas, en donde el 0.4% alcanzó nivel *avanzado*, 46.4% se clasificó en *básico*, 38.6% en *satisfactorio* y 14.6% en *bajo*. Las medidas de asociación indicaron que, el coeficiente de contingencia obtuvo un valor de 0.13, la V de Cramer de 0.08 y la prueba de Kendall fue 0.09, indicando asociación débil entre las variables.

Finalmente, la relación entre municipio de residencia y nivel de desempeño en matemáticas determinó que, en términos generales, los estudiantes de la zona costera obtuvieron desempeños en nivel *bajo o básico*, mientras que los colegiales de Pasto e Ipiales, destacaban en nivel *satisfactorio o* avanzado. Se halló también que municipios cercanos a Pasto como San Bernardo, Colón, Samaniego, Sandoná, y Tangua, resaltaban en desempeño *avanzado* con representaciones del 8.3%, 8.1%, 6.6%, 5.6%, 5%, respectivamente. Pasto clasificó al 5.7% de sus estudiantes en nivel *avanzado* y al 57.7% en *satisfactorio*. Ipiales posicionó al 4.9% de sus discentes en *avanzado* y al 56.1% en *satisfactorio*. Tumaco, la tercera cabecera municipal del departamento, categorizó al 0.7% de sus educandos en *avanzado* y al

18.6% en *satisfactorio*. En nivel *satisfactorio* sobresalieron también municipios como, San Bernardo con el 70% de representación, Gualmatán con el 67% y Consacá con el 61.1%.

Se precisó también que quienes vivían en los municipios de Arboleda, Barbacoas, Cuaspud, El Charco, El Peñol, El Rosario, Francisco Pizarro, La Tola, Magüi Payán, Mosquera, Olaya Herrera, Roberto Payan, Santa Barbará, Santacruz y Sapuyes; no lograron puntajes mayores 70 en la prueba por lo que su representación en nivel *avanzado* fue nula. En el mismo aspecto, los residentes de Santa Barbará tampoco tuvieron puntuaciones mayores a 50, por lo que su desempeño se relegó a estar en nivel *bajo o básico*.

Las medidas de asociación mostraron un coeficiente de contingencia con valor de 0.42, la V de Cramer con 0.27 y la prueba de Kendall con -0.03, indicando asociación moderada entre las variables.

- **Conclusiones del presente cruce.**

El presente cruce permitió entrever que el desempeño en matemáticas de los hombres fue superior al de las mujeres, notando así que el género del estudiante favoreció la clasificación de personas según su rendimiento en la prueba. Se notó también que los estudiantes con edades menores o iguales a 18 años, sin etnia, que disponían más de 30 minutos en internet, no trabajaban entre semana y residían en Pasto, Ipiales, o sectores aledaños, se clasificaban mayoritariamente en nivel *avanzado,* o, *satisfactorio*. Por otra parte, las personas que residían en zonas costeras o cercanas, adscritos a grupos étnicos, mayores de 18 años y que trabajan entre semana, generalmente se clasificaban, en nivel *bajo*, o, *básico*. Se observó también que entre más tiempo dedicaban los estudiantes a realizar actividades de lectura por diversión, mejores resultados obtenían en la prueba. Por otro lado, las medidas de

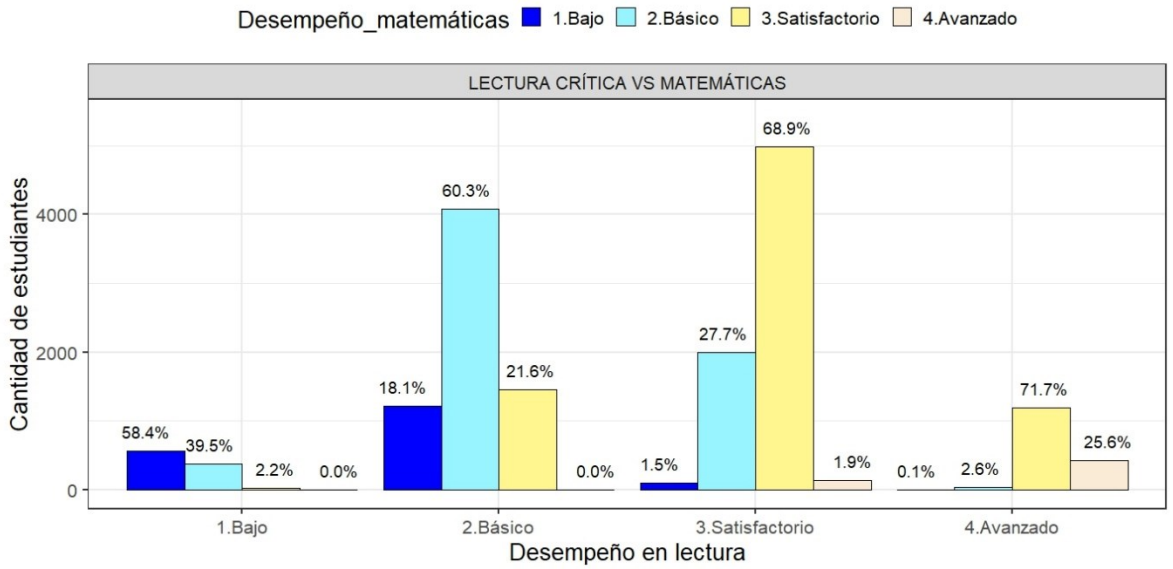
asociación reflejaron que las variables género, etnia y municipio de residencia, eran las de mayor asociación con el desempeño en la prueba.

3.3.4 Cruce del rendimiento en las pruebas con el desempeño en matemáticas.

Al relacionar el desempeño en lectura crítica con el de matemáticas se encontró que, el 58.4% de quienes se calificaron en nivel *bajo* en lectura, también obtuvieron nivel *bajo* en matemáticas. 60.3% de los *básicos* en lectura, fueron *básicos* en matemáticas. 68.9% de los *satisfactorios* en lectura, obtuvieron *satisfactorio* en matemáticas, y, 71.7% de los *satisfactorios* y 25.6% de los *avanzados* en lectura, se clasificaron como *avanzados* en matemáticas. Este aspecto motivó a pensar en que, altos desempeños en lectura se asociaban con altos desempeños en matemáticas, y bajos puntajes en lectura se relacionaban con bajos puntajes en matemáticas. Las medidas de asociación indicaron relación positiva entre las variables ya que, el coeficiente de contingencia obtuvo un valor de 0.6, la V de Cramer de 0.44 y la prueba de Kendall de 0.59. La figura 9 expone de manera visual los resultados del presente cruce.

Figura 9.

Cruce entre desempeños de lectura crítica y matemáticas.



Fuente: esta investigación.

Con respecto al cruce entre ciencias naturales y matemáticas se encontró un evento similar al anterior, en donde el 53.6% de los educandos clasificados en nivel *bajo* en ciencias obtuvieron desempeño *bajo* en matemáticas. 48.6% del nivel *básico* en ciencias se clasificaron en *básico* en matemáticas. 86.2% de los *satisfactorios* en ciencias fueron también *satisfactorios* en matemáticas, y 74.7% de los *avanzados* en ciencias también registraron nivel *avanzado* en matemáticas. Como dato peculiar se evidenció que no existían personas clasificadas en nivel *avanzado* en ciencias que obtuvieran desempeños *bajos* en matemáticas, ni tampoco personas categorizadas en nivel *avanzado* en matemáticas que obtuvieran puntajes *bajos* en ciencias. Las medidas de asociación indicaron relación positiva entre las variables ya que, el coeficiente de contingencia fue de 0.66, la V de Cramer de 0.50 y la prueba de Kendall de 0.62.

De manera similar a lo expuesto hasta el momento se encontró que, en el cruce entre los desempeños de sociales ciudadanas y matemáticas, 55.7% de los estudiantes con nivel *bajo* en sociales, obtuvo *básico* en matemáticas. 53.1% de los de nivel *básico* en sociales, se clasificaron en *satisfactorio* en matemáticas. 81.5% de los *satisfactorios* en sociales, se categorizaron en *satisfactorio* en matemáticas. Y el 58.2% de los *avanzados* en sociales, también obtuvieron nivel *avanzado* en matemáticas. Además, no existieron colegiales clasificados en nivel *bajo* en matemáticas con desempeño *avanzado* en sociales ciudadanas.

Las medidas de asociación indicaron relación positiva entre las variables ya que, el coeficiente de contingencia obtuvo un valor de 0.61, la V de Cramer de 0.44 y la prueba de Kendall de 0.57. La tabla 10 expone los valores obtenidos en el presente cruce.

Tabla 10.

Desempeño en sociales ciudadanas versus desempeño en matemáticas.

Nivel de desempeño en sociales y ciudadanas	Nivel de desempeño en matemáticas								TOTAL
	Bajo		Básico		Satisfactorio		Avanzado		
	Frecuencia	%	Frecuencia	%	Frecuencia	%	Frecuencia	%	Frecuencia
Bajo	1648	29.2%	3145	55.7%	846	15%	3	0.1%	**5642**
Básico	235	3.3%	3048	43.1%	3752	53.1%	32	0.5%	**7067**
Satisfactorio	11	0.3%	306	8.5%	2923	81.5%	348	9.7%	**3588**
Avanzado	0	0%	2	0.6%	131	41.2%	185	58.2%	**318**
TOTAL	**1894**	**11.4%**	**6501**	**39.12%**	**7652**	**46.1%**	**568**	**3.42%**	**16615**

Fuente: Esta investigación.

En el cruce dado entre las pruebas de inglés y matemáticas se evidenció que, 48.9% de los colegiales con nivel A – en inglés, obtuvieron *básico* en matemáticas, 64.9% del nivel A1 en inglés obtuvo *satisfactorio* en matemáticas, y 78.8% del nivel A2 en inglés se registraron como *satisfactorios* en matemáticas. Se notó además que el 9% de los estudiantes con nivel A2 en inglés, se clasificaron en nivel *avanzado* en matemáticas. Las medidas de asociación indicaron relación moderada entre las variables ya que, el coeficiente de contingencia tuvo un valor de 0.40, la V de Cramer de 0.31 y la prueba de Kendall de 0.38.

Finalmente, al cruzar el decil global y la prueba de matemáticas se encontró que el 48.9% y el 50.4% de los educandos del decil 4, se clasificaron, respectivamente, en nivel *bajo y básico* en matemáticas. 64.5% del decil 5 obtuvieron *básico* en matemáticas, 26.2% lograron *satisfactorio*, y 4.2% *avanzado*. Además, 84.3% de los discentes del decil 6 lograron nivel *satisfactorio* en matemáticas, 15.3% *básico* y 0.3% *avanzado*. Y el 85.6% de los colegiales del decil 7 obtuvieron desempeño *satisfactorio* en matemáticas, 14.2% en *avanzado* y 0.2% en *básico*. Además, no se hallaron estudiantes que estando en decil 7 obtuvieran calificaciones menores a 35 puntos, ni tampoco personas que estando en decil 4 fueran clasificadas en desempeño *avanzado* en matemáticas. El coeficiente de contingencia para este caso fue de 0.66, la V de Cramer fue de 0.51 y la prueba de Kendall arrojó un valor de 0.67; indicando asociación positiva entre las variables.

- **Conclusiones del presente cruce.**

Los cruces efectuados en esta instancia permitieron observar que las pruebas de lectura crítica, sociales y ciudadanas, y ciencias naturales, se asociaban positivamente con el desempeño en matemáticas. Quizá un punto en discordia se situó con la prueba de inglés, la cual pese a mostrar asociación con matemáticas, no presentó relación tan fuerte como en los anteriores casos, hecho que denotó menores valores en sus coeficientes de asociación, sin embargo, se pudo concluir que, en términos generales, a puntajes bajos en matemáticas se relacionaban puntajes bajos en las demás áreas evaluadas, y a desempeños altos en matemáticas se verificaban desempeños altos en las demás pruebas. Cabe considerar también que, hasta el momento, los aspectos académicos son los que mejor relación han presentado con el desempeño en matemáticas hecho que se vislumbró en los valores descritos en los coeficientes de asociación.

Por otro lado, este cruce permitió identificar las siguientes características por nivel de desempeño.

- ***Nivel bajo.*** Los estudiantes con desempeño *bajo* en matemáticas obtuvieron, con mayor frecuencia, nivel *bajo* en lectura crítica y en ciencias naturales. En inglés resaltó la categoría "A −". No se hallaron personas que simultáneamente cuenten con desempeño bajo en matemáticas y *avanzado* en ciencias naturales o en sociales ciudadanas.
- ***Nivel básico.*** Los discentes con desempeño básico en matemáticas obtuvieron, con mayor frecuencia, nivel *básico* en lectura crítica y en ciencias naturales, y desempeño *bajo* en sociales ciudadanas. En inglés resaltó la categoría "A −". El puntaje global se posicionó en decil 5.

- ***Nivel satisfactorio***. Los colegiales con desempeño *satisfactorio* en matemáticas se categorizaron, con mayor frecuencia, en nivel *satisfactorio* en lectura, ciencias naturales y sociales ciudadanas. En inglés resaltó la categoría “A1”. Además, los resultados de este nivel conllevaron a ubicar los puntajes globales en los deciles 6 y 7.
- ***Nivel avanzado.*** Las personas categorizadas en desempeño *avanzado en matemáticas* puntuaron en nivel *avanzado* en lectura, ciencias naturales y sociales ciudadanas. En inglés resaltó la categoría “A2”. En este grupo se evidenció ausencia de personas que simultáneamente cuenten con desempeño *avanzado* en matemáticas y *bajo en lectura crítica* y *ciencias naturales*. No se hallaron personas dentro del decil 7 que obtuvieran desempeño *bajo* en matemáticas.

3.3.5 *Cruce de los aspectos institucionales con el desempeño en matemáticas*.

Al comparar el desempeño en matemáticas con las variables *género del colegio, estado bilingüe de la institución y calendario académico*, no se advirtieron asociaciones sobresalientes ya que más del 96% de las instituciones se caracterizaban por ser de naturaleza mixta, no bilingüe y adscritas al calendario “A” en los diferentes niveles de desempeño. De igual forma ocurrió con la variable *jornada del colegio* en donde más del 70% de la población estudiaba en jornada de la mañana. La variable *naturaleza del colegio (oficial o privado)* reflejó que, en todos los niveles de desempeño, más del 80% de educandos pertenecían a instituciones públicas. De manera no tan diciente se logró observar que los estudiantes del sector privado se lograron clasificar con mayor frecuencia en nivel *satisfactorio y avanzado* en matemáticas.

La ubicación geográfica del plantel educativo mostró que los estudiantes de colegiales rurales obtuvieron desempeños *bajos y básicos* en matemáticas como mayor referente, mientras que los discentes de la zona urbana se clasificaron con mayor frecuencia en nivel *satisfactorio y avanzado*. En el atributo *carácter del colegio* no se observó una relación que

permita catalogar a los estudiantes en uno u otro desempeño, ya que, la modalidad *académica* resaltó en todos los niveles de desempeño.

3.4 DESARROLLO DEL OBJETIVO ESPECÍFICO 3.

Determinar la correlación existente entre los puntajes de las pruebas de lectura crítica, matemáticas, ciencias naturales, inglés, sociales y ciudadanas.

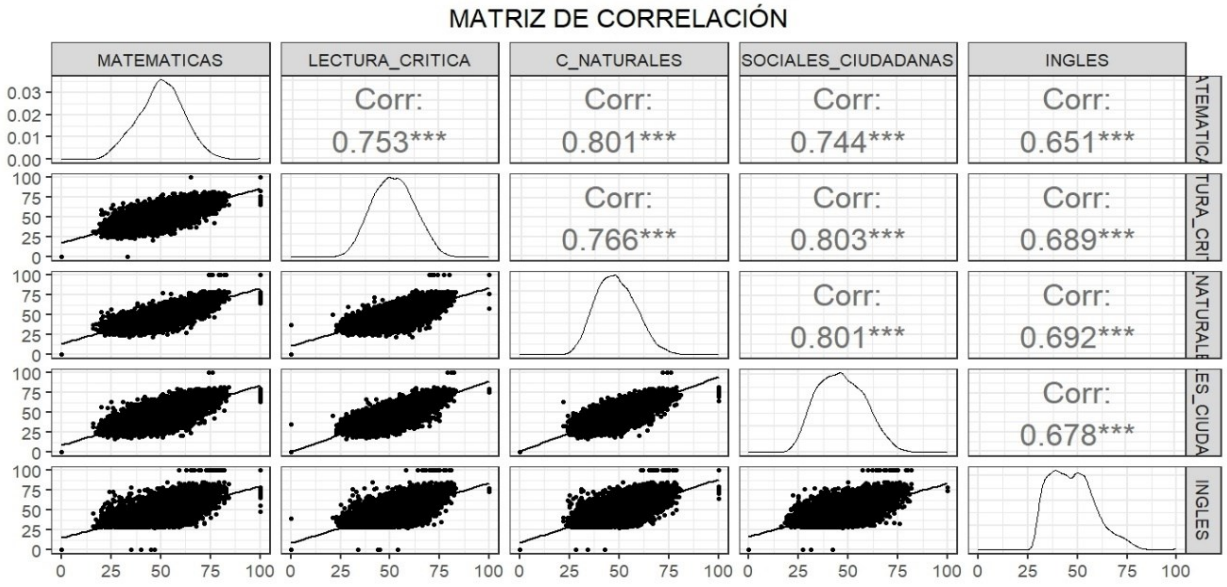
Para dar solución a este objetivo se creó la matriz de correlación, la que en su parte superior describe las correlaciones existentes entre las pruebas, en la diagonal muestra los gráficos de densidad y en la parte baja esboza el ajuste lineal (smooth) que presentaron los datos. Con esta matriz se determinó que la correlación entre matemáticas y cada una de las pruebas fue, 0.801 con ciencias naturales, 0.753 con lectura crítica, 0.744 con sociales ciudadanas y 0.651 con inglés; hecho que mostró que, en términos generales, a mejores desempeños de los estudiantes en las pruebas de ciencias naturales, lectura crítica, sociales e inglés, mejores resultados se obtenían en matemáticas. Además, según lo visto en Saravia (2015), la correlación existente entre matemáticas e inglés se consideró positiva moderada, y fue positiva alta para matemáticas y las demás pruebas evaluadas.

Anexo a lo anterior, las gráficas de densidad mostraron comportamiento casi simétrico para las pruebas de matemáticas, lectura crítica y ciencias naturales. En Sociales existió una ligera tendencia hacia la izquierda y en inglés se notó la presencia de dos crestas las cuales rompieron la simetría de la gráfica. Por otro lado, los smooth indicaron que los datos, en términos generales, se ajustaban a un comportamiento lineal, ya que en cada recuadro se dibujó una recta que atravesó la nube de puntos. Se precisó también que la variabilidad entre los puntos y la recta no era grande, dicho de otra forma, la nube de puntos se hallaba cercana

al smooth lineal hecho que reforzó la consideración del ajuste lineal en los datos. La matriz de correlación generada por RStudio se describe en la figura 10.

Figura 10.

Matriz de correlación de los puntajes de la prueba Saber 11.

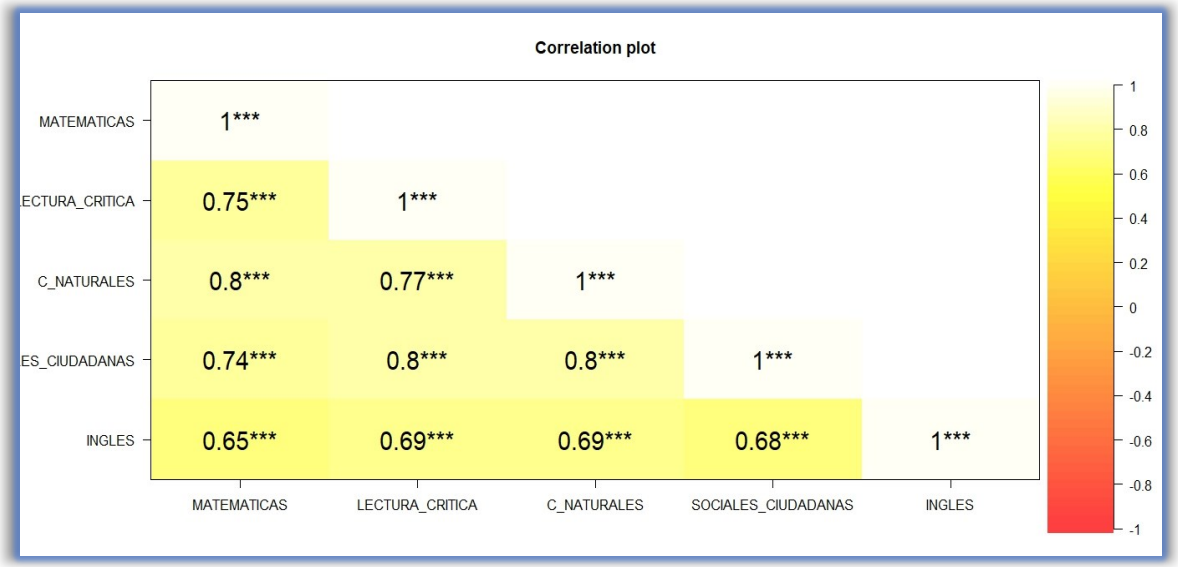


Fuente: esta investigación.

Con intención de reforzar el anterior análisis, en la figura 11 se presenta el gráfico de calor correlacional en donde se precisa que la intensidad y transparencia de los colores de las correlaciones obtenidas son próximas a la de la unidad, indicando relación lineal positiva.

Figura 11.

Gráfico de calor de correlaciones.



Fuente: esta investigación.

En resumen, considerando los resultados anteriores es plausible pensar que, en términos generales, desempeños altos en matemáticas se relacionaban con desempeños altos en las demás pruebas evaluadas, o, al contrario, a puntajes bajos en matemáticas se correspondían puntajes bajos en las demás pruebas, tal como se vio en el cruce de variables.

3.5 DESARROLLO DEL OBJETIVO ESPECÍFICO 4.

Identificar estructuras fundamentales o factores asociados a los puntajes dados en las pruebas de lectura crítica, matemáticas, ciencias naturales, inglés, sociales y ciudadanas.

A continuación se esgrimen los resultados generados tras la aplicación del análisis factorial exploratorio, y se labora desde las concepciones teóricas expuestas en el literal 1.7.4 del presente documento y los aportes de Bolaños (2020) y Vallejo Medina (2020), quienes

plantean que para desarrollar este análisis se requiere, (1) *verificar que la matriz de datos sea factorizable,* (2) *extraer los factores,* (3) *determinar el número correcto de factores,* (4) *rotar los factores*, (5) i*nterpretar los resultados*.

3.5.1 Paso 1. Verificar que la matriz sea factorizable.

Con el literal 3.4 de la presente investigación se concluyó que los puntajes de las pruebas de matemáticas, lectura crítica, ciencias naturales, sociales ciudadanas e inglés se encontraban correlacionadas. No se tuvo en cuenta los valores del puntaje global puesto refería una combinación de las cinco anteriores. Las figuras 10 y 11 exhibieron este hecho desde una postura gráfica, por lo que se procede ahora a determinar si la matriz de datos que contiene a dichos puntajes es o no factorizable, por ende, se recurre a la prueba de esfericidad de Bartlett y la de Kaiser Meyer Olkin (KMO).

La prueba de Bartlett se estudió en consideración al siguiente sistema de hipótesis.

$$\begin{cases} H_0\colon\ La\ matriz\ de\ correlación\ es\ la\ identidad. \\ H_1\colon La\ matriz\ de\ correlación\ no\ es\ la\ identidad. \end{cases}$$

La salida computacional mostró que, a un nivel de significancia del 5%, el p – valor obtenido fue de $1.743267\ *e^{-80}$, hecho que sugirió rechazar la hipótesis nula H_0.

Por otro lado, la prueba KMO indicó qué tan adecuados se encontraban los datos para el análisis factorial, estudiando la proporción de varianza entre variables que podrían presentar comunalidad. Bolaños (2020) esgrime como valores referencia para la interpretación del KMO los expuestos en la tabla 11, ante los cuales se analizó que la salida computacional de RStudio

arrojó un valor de ***0.9***, hecho que lo ubicó en rango *maravilloso* y que reforzó la intención de continuar con el desarrollo del análisis factorial.

Tabla 11.

Valores para interpretar el KMO.

TABLA DE VALORES PARA KMO	
RANGO	**DESCRIPCIÓN**
0.00 a 0.49	Inaceptable
0.50 a 0.59	Miserable
0.60 a 0.69	Mediocre
0.70 a 0.79	Medio
0.80 a 0.89	Meritorio
0.90 a 1.0	Maravilloso

Fuente: Adaptado de Bolaños (2020).

3.5.2 Paso 2. Elección del método para extraer factores.

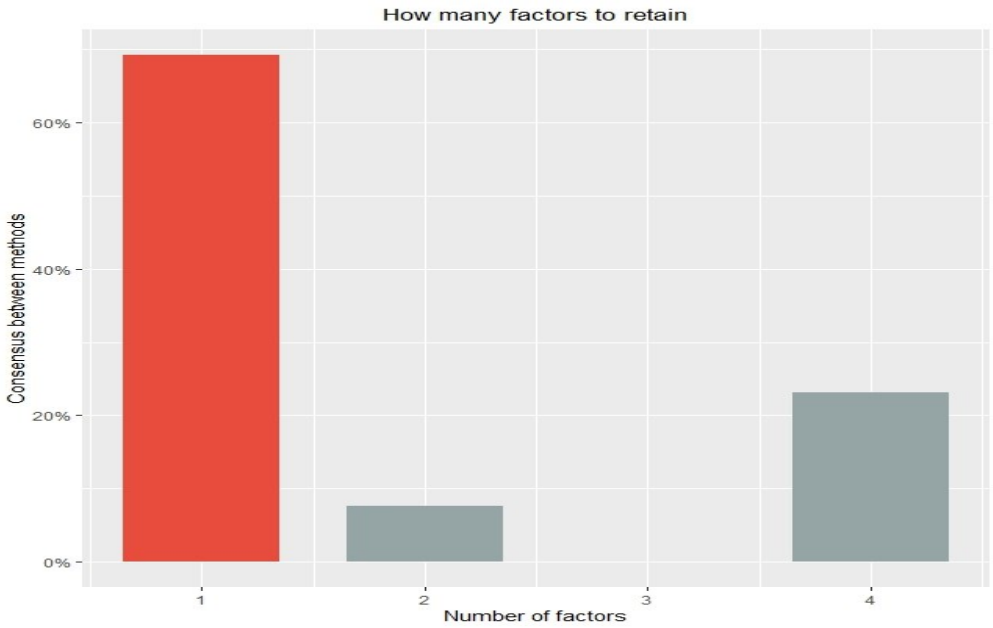
En RStudio se encuentran disponibles los métodos de mínimo residuo, máxima verosimilitud, factorización de ejes principales, factorización alfa, mínimos cuadrados y rango mínimo. De ellos se eligió el método de ejes principales ya que no requiere como supuesto la normalidad multivariada.

3.5.3 Paso 3. Determinar el número correcto de factores.

Al usar el comando *n_factors* en RStudio, el software plantea que existe un factor que emerge como relevante debido a la alta correlación existente entre los puntajes. La figura 12 esquematiza este aspecto.

Figura 12.

Número de factores a retener



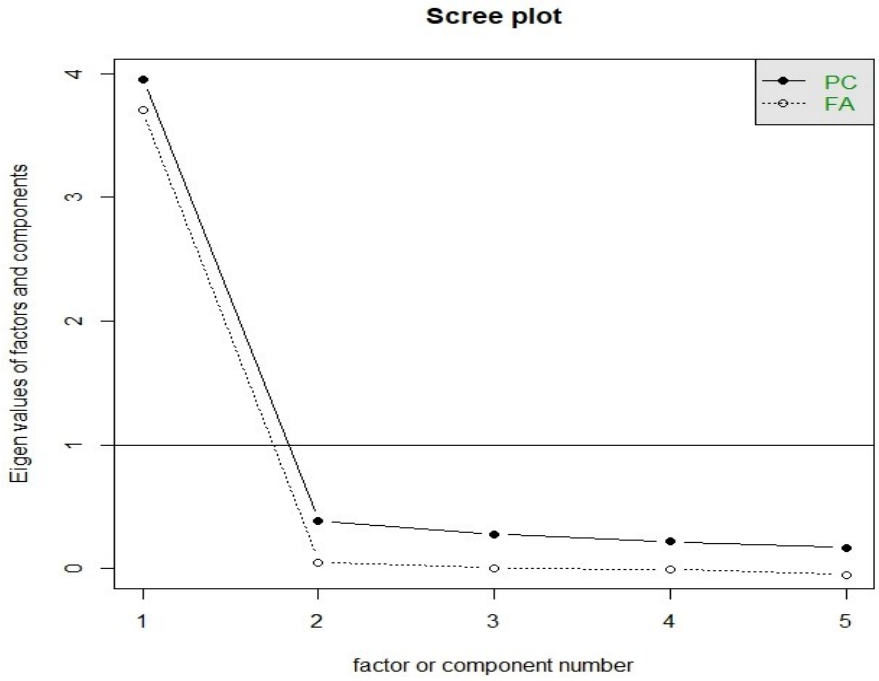
Fuente: esta investigación.

De igual forma, Bolaños (2020) expone que para elegir el número de factores conviene considerar los criterios *Kaiser, análisis Scree Plot y análisis paralelo.* En Kaiser se retienen todos los factores con valor propio mayor a la unidad. En Scree Plot se selecciona un grupo reducido de factores con valores propios superiores a los demás analizando para ello el punto de inflexión de la curva o “codo”. El análisis paralelo contempla la extracción de factores a través de sus varianzas más altas y a partir de allí elige los valores propios.

La gráfica del Scree Plot asociada a los puntajes de la prueba Saber 11 mostró que los valores se acoplaban de mejor manera en el primer factor, puesto que fue el único con valor propio mayor a la unidad (ver figura 13).

Figura 13.

Scree plot de los puntajes de los estudiantes en la prueba Saber.



Fuente: esta investigación.

El análisis paralelo también sugirió que el primer factor tenía valor propio superior a la unidad, por esta razón se consideró que trabajar los datos con dos factores sería suficiente. Cabe decir que conviene dejar como mínimo dos factores porque de lo contrario el proceso de rotación “varimax” no podría realizarse.

De lo precedente se determinó que, el primer factor, presentaba un valor propio de 3.95 y explicaba el 79.08% de la varianza total. El valor propio del segundo factor era de 0.38 y explicaba el 7.65% de la varianza total. Además, con la participación del segundo factor se logró acumular el 86.73% de la varianza total. Los resultados del modelo factorial permitieron entrever también que el factor 1 presentaba las cargas factoriales más altas. Agregado a ello, las comunalidades de todos los puntajes eran bastante altas si se retenía el primer factor, hecho debido a la alta correlación existente entre los puntajes en estudio y con lo que fue

posible establecer que a puntajes altos en matemáticas correspondían puntajes altos en las otras cuatro pruebas evaluadas. La tabla 12 presenta las cargas factoriales y las comunalidades obtenidas en esta instancia.

Tabla 12.

Cargas factoriales y comunalidades.

PUNTAJE	CARGAS FACTORIALES		COMUNALIDADES	
	Factor 1	Factor 2	Factor 1	Factor 2
Lectura crítica	-0.903115	0.065363	0.815618	0.819890
Matemáticas	-0.888838	0.207950	0.790033	0.833277
Ciencias naturales	-0.915191	0.122815	0.837574	0.852658
Sociales y ciudadanas	-0.907654	0.113554	0.823836	0.836730
Inglés	-0.828788	-0.554221	0.686889	0.994050
Valor propio	3.953950	0.382655		
% de varianza explicado	0.790790	0.076531		
% de varianza acumulado	0.790790	0.867321		

Fuente: esta investigación.

3.5.4 Paso 4. Rotar la matriz.

Con intención de aclarar un poco más el análisis precedente se realizó una rotación *varimax*, la cual permitió establecer que los puntajes de *lectura crítica, matemáticas, ciencias naturales y sociales ciudadanas*, se encontraban altamente correlacionados en un factor; sin embargo, los resultados en inglés se distanciaron del comportamiento de las demás pruebas siendo elementos distintivos de otro factor. La tabla 13 muestra las cargas factoriales obtenidas con la rotación varimax.

Tabla 13.

Cargas factoriales con rotación varimax.

	ROTACIÓN VARIMAX	
PUNTAJE	**FACTOR 1**	**FACTOR 2**
Lectura crítica	0.774159	0.469646
Matemáticas	0.845082	0.345127
Ciencias naturales	0.817271	0.429797
Sociales y ciudadanas	0.805764	0.432984
Inglés	0.354798	0.931756
Valor propio	2.756556	1.580049
% de varianza explicado	0.551311	0.316010

Fuente: esta investigación.

3.5.5 Paso 5. Interpretar los resultados.

La rotación varimax permitió concluir que, pese a que en un principio se notó alta correlación positiva entre los resultados obtenidos por los estudiantes en las diferentes pruebas, los puntajes de lectura crítica, matemáticas, ciencias naturales y sociales ciudadanas, encontraron puntos en común entre sí, lo que favoreció establecer un factor que los acoja y que se denominó *áreas básicas*. Por otro lado, el área de inglés jugó un rol diferente en esta etapa ya que, a pesar de hallarse correlacionada con matemáticas, la relación entre estas dos era moderada, aspecto que sugirió no agruparla en el factor de áreas básicas sino distinguirla en un factor alterno el cual se lo nombró *inglés*.

A manera de consideraciones finales se relata que, con el desenvolvimiento del presente objetivo se logró detectar que, pese a encontrar correlación entre los puntajes de matemáticas, ciencias naturales, lectura crítica, inglés, sociales y ciudadanas, se observó que las pruebas realizadas por los estudiantes se podían agrupar en dos factores, uno denominado áreas básicas y otro inglés, aspecto que en el interés de determinar relaciones entre los puntajes se tornó relevante para esta investigación.

3.6 DESARROLLO DEL OBJETIVO ESPECÍFICO 5.

Analizar y comparar modelos de agrupación que identifiquen grupos de estudiantes con características similares en torno al puntaje en matemáticas.

La participación del software WEKA 3.9.6 emergió como importante en esta etapa ya que sus algoritmos permitieron elaborar conglomerados con variables cualitativas, sin necesidad de convertir manualmente los atributos categóricos en variables dummy o de hacer codificaciones numéricas de ellos. Fue así que, en primera medida, se ingresó en WEKA el

conjunto de datos *Nariño_imputados_2021b* el cual se componía netamente por variables categóricas, tal como se explicitó en el literal 2.4.2 del presente texto.

Con intención de ganar información y obtener clústeres mejor conformados, en la fase de *preprocesamiento* de WEKA se omitieron las variables de los aspectos institucionales, ya que en el cruce de variables se determinó que ellos no advirtieron relaciones sobresalientes con el desempeño de los estudiantes en matemáticas. Seguido a ello se aplicó el atributo *numeric to nominal* de los filtros no supervisados que ofrece el programa con el objetivo que WEKA reconociera como categóricas a todas las instancias.

Posteriormente, en la sección *Cluster* se activó la opción *Simple K–Means*, allí se eligió el algoritmo *k-means ++,* y como función de distancia la euclídea. Cabe decir que a pesar que los datos eran categóricos, el software los leía como variables mixtas. Entre los algoritmos a disposición se hallaban *k-means ++, Canopy,* y, *Farthest first.* Como distancias estaban *Chebishev, Euclidean, Filtered, Manhattan, Minkowski*, sin embargo, las distancias *Chebishev, Filtered y Minkowski* no se hallaban activas para datos cualitativos.

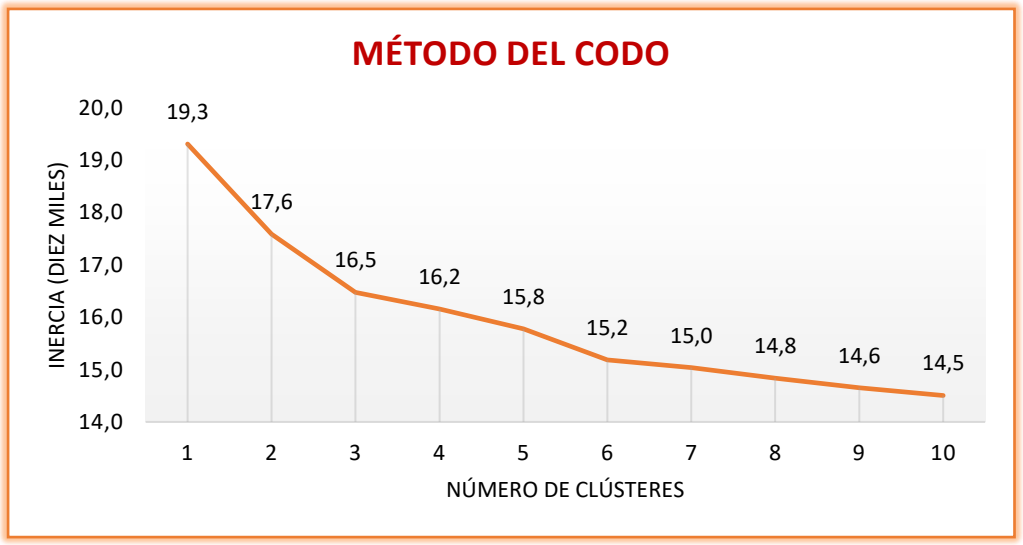
Se eligió k-means ++ y distancia euclídea ya que, tras probar con las demás distancias y algoritmos estos fueron los que mostraron clústeres mejor diferenciados, y los que elaboraron una distribución más homogénea de las instancias clasificadas en cada clúster. Las otras funciones y algoritmos agrupaban a más del 50% de instancias en un clúster, lo que provocaba poca diferenciación entre los elementos de cada conglomerado.

Seguido a ello, en la opción *numClusters* se fueron digitando números del 1 al 10 con el fin de contrastar la suma de cuadrados del error de las diez primeras agrupaciones. Este hecho permitió construir la gráfica *Inercia vs Número de Clústeres*, la cual fue elaborada en Excel y

estudiada en consideración al *método del codo*. Los valores expuestos en el eje de la *inercia* se midieron en diez miles siendo que el dato 19,3 por ejemplo, indicaba un error cuadrado real de 193.000. Se observó además que la curva que describía el comportamiento de la inercia presentaba un punto de inflexión o cambio de concavidad a partir del tercer clúster, es decir, a partir del tercer clúster los datos se estabilizaban o suavizaban, razón por la que se tuvo a bien elaborar la agrupación de los datos mediante tres conglomerados.

Figura 14.

Método del codo para elegir el número de clústeres.



Fuente: esta investigación.

Al contar con tres conglomerados WEKA dispuso en el clúster 0 a 5229 individuos los cuales aludían al 31% del total de datos, en el clúster 1 clasificó al 31% y en el clúster dos ubicó al 38%. La tabla 14 muestra la clasificación realizada por WEKA.

Tabla 14.

Clasificación en WEKA del número de individuos y porcentajes por clúster.

CLUSTER	NÚMERO DE INDIVIDUOS	PORCENTAJE (%)
Clúster 0	5229	31
Clúster 1	5080	31
Clúster 2	6306	38

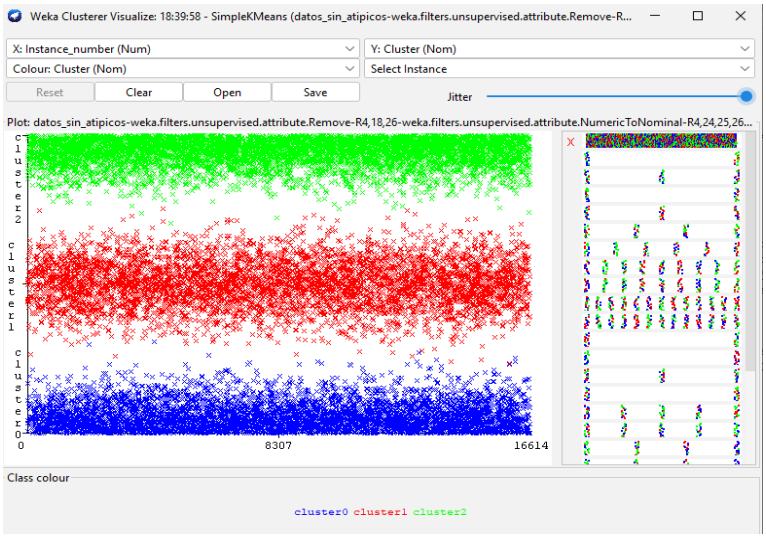
Fuente: esta investigación.

Como anotación importante se menciona que los clústeres elaborados por WEKA no presentaron separación ni cohesión perfecta, ya que había atributos compartidos entre los grupos formados. Esto destaca que una de las dificultades de WEKA radica en no brindar medidas de viabilidad e importancia para el análisis de la separación y cohesión de los grupos, ni tampoco el valor del coeficiente silueta. Lo único que muestra es el porcentaje de instancias clasificadas exhibidas en la tabla 14, el número de iteraciones efectuadas (5 en este caso) y la suma de cuadrado del error para tres clústeres la cual fue de 164.622. Evento que motivó a tomar decisiones respecto a la representación gráfica sobre los clústeres conformados y ante lo cual se precisó que, al trabajar con tres clústeres, los datos tendían a estar más separados que si se ponían cinco grupos, por citar un caso.

La figura 15 muestra que los elementos del clúster 0 (color azul) presentaban más cercanía a los del clúster 1 que a los del 2, o sea, los datos del clúster 0 distaban más de los del grupo 2 que de los del 1. El clúster 1 compartía información con ambos grupos ya que había cruces de color rojo en el clúster 2 y en el 1, hecho que validó la idea inicial sobre la no tan fuerte separación entre grupos.

Figura 15.

Agrupación en WEKA de los datos mediante 3 clústeres.

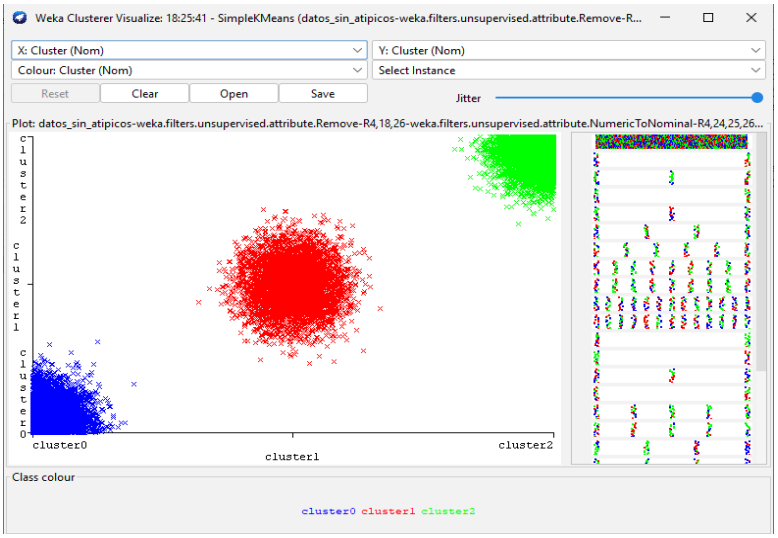


Fuente: esta investigación.

Agregado a la figura 15, es probable que el lector esté acostumbrado a la representación circular de los clústeres, razón por la cual se muestra en la figura 16 dicha gráfica, no sin antes advertir que esta imagen es engañosa porque pareciera separar de buena manera los grupos, hecho que como se dijo anteriormente no es del todo verdadero.

Figura 16.

Representación de clústeres en WEKA por 3 grupos.

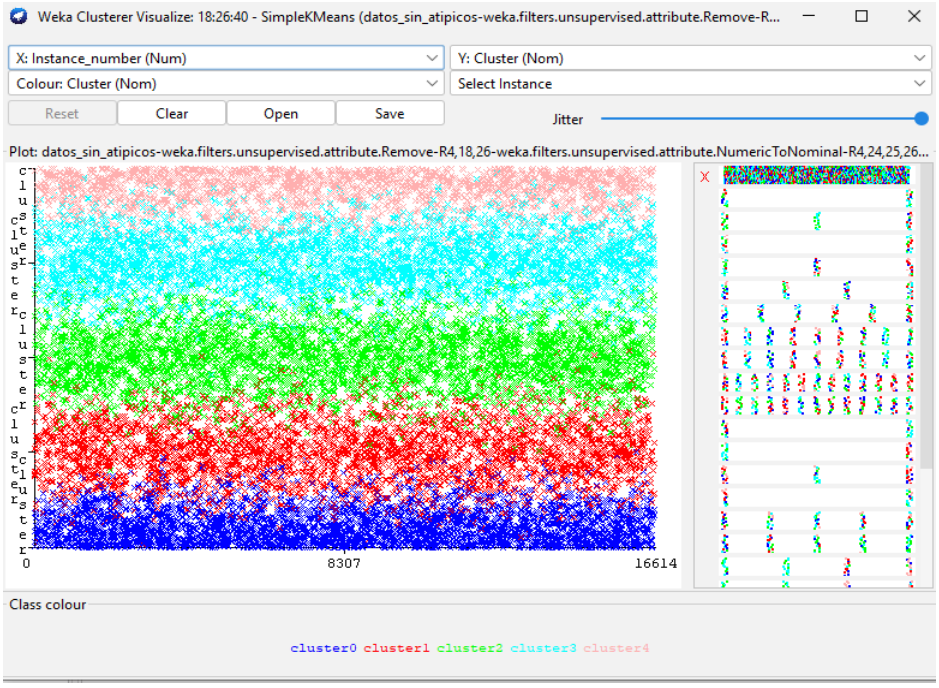


Fuente: esta investigación.

Con intención de mostrar al lector la representación elaborada por WEKA al agrupar los datos mediante 5 conglomerados, se expone la figura 17 en donde se detalla que la separación de instancias es más confusa haciendo que los atributos tiendan a mezclarse entre los diferentes grupos. El color azul se confunde con el rojo (clúster 1) y con el verde (clúster 2). Hecho similar ocurre con los elementos de los demás grupos.

Figura 17.

Agrupación en WEKA de los datos mediante 5 clústeres.



Fuente: esta investigación.

Ahora bien, pese a que WEKA no brindó medidas de viabilidad y separación de los clústeres, sí logró establecer una característica general para el total de estudiantes y también diferenció las características particulares de cada uno de los 3 conglomerados formados.

Como característica para el grupo completo WEKA estableció que los estudiantes nariñenses se caracterizaban por ser mayoritariamente mujeres, menores de edad, sin etnia, con estrato 1, con núcleo familiar integrado por 3 a 4 personas y quienes habitaban en tres cuartos. Tanto el padre como la madre contaban con educación secundaria completa. El padre se dedicaba a labores de agricultura, pesca o trabajo al jornal, y la madre a trabajar en casa, no trabajar o estudiar. Tenían internet, pero no computador. En sus casas existían hasta 10 libros. Dedicaban menos de 30 minutos a realizar actividades de lectura y de 30 a 60 minutos a navegar en internet. No trabajaban entre semana. Los desempeños en las pruebas más característicos fueron, *satisfactorio* en matemáticas y lectura crítica, *básico* en sociales ciudadanas y ciencias naturales, y "A −" en inglés. Su puntaje promedio del decil global fue 5.0 y se clasificaron en NSE 2.

Anexo a lo anterior, las características otorgadas por WEKA para cada uno de los tres clústeres conformados fueron las descritas a continuación.

Los individuos del ***clúster 0*** se caracterizaron, en términos generales, por tener calificación de 7 en el puntaje decil global, hecho que indicó aprobación del 70% de la prueba. Su NSE fue de 2.0. En sociales ciudadanas, lectura crítica, ciencias naturales y matemáticas tuvieron desempeño *satisfactorio*. Contaban con padres con estudios secundarios culminados. Los rasgos propios del ***clúster 0*** y que lo diferenciaron de los demás, lo constituyeron que el grupo mayoritario lo conformaban hombres con desempeños sobresalientes en sociales ciudadanas y ciencias naturales, tenían la mejor calificación en puntaje decil global (7) y se clasificaban en NSE 2. Los demás atributos eran compartidos con alguno de los otros dos grupos restantes.

El centro del ***clúster 1*** mostró que sus estudiantes se caracterizaban, en términos generales, por clasificarse en nivel 6 en el puntaje decil global, lo que indicó aprobación del 60% de la prueba. Su NSE tuvo calificación de 3.0. Obtuvieron desempeño *básico* en sociales ciudadanas y ciencias naturales, y *satisfactorio* en matemáticas y lectura crítica. Tanto el padre como la madre habían culminado sus estudios de secundaria. El aspecto diferenciador de los educandos de este clúster radicó en la tenencia de computadores en casa y en la categorización en nivel básico en ciencias naturales y sociales ciudadanas, hecho que lo asemejaba a la condición presentada para el contexto global (clúster con los datos completos). Además, los estudiantes de este grupo puntuaron en A1 en inglés, y su NSE 3 era el más alto entre los tres clústeres conformados. Los demás atributos encontraron algo en común con uno de los otros dos conglomerados.

El centro del ***clúster 2*** mostró como condiciones particulares que los estudiantes tenían, en términos generales, dos cuartos en el hogar y tanto el padre como la madre no habían terminado la educación primaria, además no disponían de internet en casa y los desempeños en las pruebas fueron los más bajos encontrados, ya que en lectura y matemáticas lograron un nivel *básico*, mientras que en sociales ciudadanas y ciencias naturales se posicionaron en *bajo*. Anexo a lo anterior, los discentes obtuvieron nivel 5 en su puntaje decil global o sea aprobaron el 50% de la prueba. Su nivel socioeconómico tuvo calificación de 1.0. Los demás atributos encontraban algo en común con uno de los otros dos conglomerados.

La tabla 15 condensa las ideas precedentes exhibiendo las características encontradas tanto en su aspecto global, o sea acogiendo a los 16.615 registros, y los datos más relevantes de cada conglomerado.

Tabla 15.

Características de cada clúster encontradas en WEKA.

VARIABLE	CENTROS DE LOS CLÚSTERES			
	DATOS COMPLETOS (16.615)	**CLÚSTER 0 (5.229)**	**CLÚSTER 1 (5.080)**	**CLÚSTER 2 (6.306)**
Género del estudiante	F	M	F	F
Edad	X < 18	X < 18	X < 18	X < 18
Adscrito a etnia	No	No	No	No
Estrato de la vivienda	1	1	1	1
Número de personas en el hogar	3_a_4	3_a_4	3_a_4	3_a_4
Número de cuartos en el hogar	Tres	Tres	Tres	Dos
Escolaridad del padre	Secundaria completa	Secundaria completa	Secundaria completa	Primaria incompleta
Escolaridad de la madre	Secundaria completa	Secundaria completa	Secundaria completa	Primaria incompleta
Ocupación laboral del padre	Agricultor, pesquero o jornalero	Agricultor, pesquero o jornalero	Agricultor, pesquero o jornalero	Agricultor, pesquero o jornalero
Ocupación laboral de la madre	Trabaja en casa, no trabaja o estudia	Trabaja en casa, no trabaja o estudia	Trabaja en casa, no trabaja o estudia	Trabaja en casa, no trabaja o estudia
Internet en casa	Si	Si	Si	No
Computador en casa	No	No	Si	No
Número de libros en casa	0 a 10	0 a 10	0 a 10	0 a 10
Tiempo promedio de lectura al día	30 minutos o menos	30 minutos o menos	30 minutos o menos	30 minutos o menos
Dedicación diaria a internet	30 a 60 minutos	30 a 60 minutos	30 a 60 minutos	30 a 60 minutos
Número de horas que trabaja a la semana	0	0	0	0
Desempeño en lectura crítica	3	3	3	2
Desempeño en matemáticas	3	3	3	2
Desempeño en ciencias naturales	2	3	2	1
Desempeño en sociales y ciudadanas	2	3	2	1
Desempeño en inglés	A –	A –	A1	A –
Decil global	5	7	6	5
NSE individual	2	2	3	1

Fuente: esta investigación.

- **Clústeres conformados en RStudio con k – modas.**

Al igual que WEKA, RStudio también brindó la posibilidad de elaborar clústeres para variables categóricas con la ventaja que, para la agrupación de las variables cualitativas implementa el algoritmo k – modas. Sin embargo, las opciones presentadas en RStudio para agrupar variables cualitativas, son más limitadas que las de WEKA en el sentido que solo

presenta visualización de clústeres para valores discretos, pero no para atributos expresados mediante texto. Al trabajar con el algoritmo k – modas, RStudio tampoco expuso medidas viabilidad, importancia, ni el coeficiente silueta para valores categóricos dados en caracteres.

Para utilizar el k – modas en RStudio se debe instalar los paquetes *klaR y MASS*. Se usa el comando *kmodes* con 3 clústeres ya que la aplicación del método del codo sugirió la conformación de tres conglomerados. Cabe decir que al correr el comando *kmodes* se obtuvieron diferentes agrupaciones en cada ejecución lo que comprobó la teoría de Hair et al. (1999) al exponer la multiplicidad de soluciones en el trabajo con clúster, por lo que fue necesario precisar una semilla con los números 1234.

El comando *kmodes* otorgó 6 componentes los cuales se denominan: *"clúster", "size", "modes", "withindiff", "iterations", "weighted"*. En la opción *"clúster"* representa mediante valores enteros el número del grupo al que pertenece cada observación. "Size" define el tamaño de cada conglomerado, en este caso el primer clúster contiene 6062 observaciones, 6058 el segundo y 4495 el tercero. *"Modes"* presenta las modas de cada agrupación efectuada. *"Withindiff"* da la distancia de coincidencia simple dentro del clúster para cada grupo, en este caso los valores obtenidos fueron 69128, 58329, 46155 respectivamente, y señala que para este proceso se realizaron 3 iteraciones. "W*eighted"* muestra si se realizaron o no distancias ponderadas, hecho que para este caso resultó en falso. La tabla 16 presenta los clústeres formados en RStudio mediante el algoritmo kmodes.

Tabla 16.

Centros de los clústeres en RStudio mediante kmodes.

VARIABLE	CENTROS DE LOS CLÚSTERES		
	CLUSTER 0 (5959)	CLUSTER 1 (5204)	CLUSTER 2 (5452)
Género del estudiante	M	F	F
Edad	X < 18	X < 18	X < 18

VARIABLE	CENTROS DE LOS CLÚSTERES		
	CLUSTER 0 (5959)	CLUSTER 1 (5204)	CLUSTER 2 (5452)
Adscrito a etnia	No	No	No
Estrato de la vivienda	1	1	1
Número de personas en el hogar	3 a 4	3 a 4	3 a 4
Número de cuartos en el hogar	Tres	Tres	Tres
Nivel educativo del padre	Secundaria completa	Primaria incompleta	Primaria incompleta
Nivel educativo de la madre	Secundaria completa	Primaria incompleta	Secundaria completa
Ocupación laboral del padre	Agricultor, pesquero, jornalero	Agricultor, pesquero, jornalero	Agricultor, pesquero, jornalero
Ocupación laboral de la madre	Trabaja en casa, no trabaja o estudia	Trabaja en casa, no trabaja o estudia	Trabaja en casa, no trabaja o estudia
Internet en casa	Si	Si	No
Computador en casa	Si	No	No
Número de libros en casa	11 a 25	0 a 10	0 a 10
Tiempo promedio de lectura al día	30 minutos o menos	30 minutos o menos	30 minutos o menos
Dedicación diaria a internet	1 a 3 horas	30 a 60 minutos	30 minutos o menos
Número de horas que trabaja el estudiante a la semana	0	0	0
Desempeño en lectura crítica	3	2	2
Desempeño en matemáticas	3	2	2
Desempeño en ciencias naturales	2	2	1
Desempeño en sociales y ciudadanas	2	2	1
Desempeño en inglés	A1	A –	A –
Decil global	6	5	4
NSE individual	2	2	1

Fuente: esta investigación.

La salida computacional exhibida por RStudio permitió observar resultados similares a las agrupaciones elaboradas en WEKA. No fueron idénticos, pero coincidieron en variables como el género, etnia, estrato, número de personas en el hogar y otros registros, lo que permitió validar las ideas expuestas en WEKA.

Finalmente, en aras de buscar una medida de validación para los clústeres generados se recurrió a utilizar la agrupación por variables mixtas en RStudio, para ello fue necesario instalar las librerías clustMixType, cluster, ggpubr, y el comando *validation_kproto*. Esto permitió determinar que la medida del coeficiente silueta era 0.182 lo que indicó indicios de clústeres traslapados.

Entre otros datos importantes que arrojó la salida computacional de RStudio, se encontró al valor estimado de lambda el cual fue 1.041472, útil para conformar grupos de forma directa con el comando *kproto* y que define un valor ponderado entre las distancias de datos categóricos y numéricos. El tamaño de cada uno de los 3 clústeres fue de 4200, 5266, 7149 observaciones respectivamente. Los errores acogidos en cada conglomerado fueron de 46335.82, 50064.89, 73510.98. Con todo ello, las agrupaciones logradas por el k – prototipos en RStudio se presentan en la tabla 17.

Tabla 17.

Centros de los clústeres en RStudio mediante K-prototipos.

VARIABLE	CENTROS DE LOS CLUSTERES		
	CLUSTER 0 (5959)	**CLUSTER 1 (5204)**	**CLUSTER 2 (5452)**
Género del estudiante	M	F	F
Edad	X < 18	X < 18	18<=X<=22
Adscrito a etnia	No	No	No
Estrato de la vivienda	2	1	1
Número de personas en el hogar	3 a 4	3 a 4	3 a 4
Número de cuartos en el hogar	Tres	Dos	Tres
Nivel educativo del padre	Secundaria completa	Primaria incompleta	Primaria incompleta
Nivel educativo de la madre	Secundaria completa	Secundaria completa	Secundaria completa
Ocupación laboral del padre	Agricultor, pesquero, jornalero	Agricultor, pesquero, jornalero	Agricultor, pesquero, jornalero
Ocupación laboral de la madre	Trabaja en casa, no trabaja, o estudia	Trabaja en casa, no trabaja, o estudia	Trabaja en casa, no trabaja, o estudia
Internet en casa	Si	Si	No
Computador en casa	Si	No	No
Número de libros en casa	11 a 25	0 a 10	0 a 10
Tiempo promedio de lectura al día	30 minutos o menos	30 minutos o menos	30 minutos o menos
Dedicación diaria a internet	1 a 3 horas	30 a 60 minutos	30 minutos o menos
Número de horas que trabaja el estudiante a la semana	0	0	0
Desempeño en lectura crítica	3	3	2
Desempeño en matemáticas	3	3	2
Desempeño en ciencias naturales	3	2	1
Desempeño en sociales y ciudadanas	3	2	1
Desempeño en inglés	A1	A 1	A –
Decil global	7	6	5
NSE individual	3	2	2

Fuente: esta investigación.

Así entonces, y de acuerdo a los clústeres conformados en WEKA y RStudio, se logró identificar que los estudiantes del departamento de Nariño se dividían en tres grupos, en donde

si bien es cierto existían variables compartidas, también se destacan algunas particularidades en cada uno de ellos y que su análisis e interpretación se encuentran acordes a los trabajos referentes a capital cultural, donde se establece que, a mejores condiciones socioeconómicas del estudiante, mayor posibilidad presenta de obtener mejores resultados en la prueba.

En el primer grupo se encontraban los hombres menores de edad, sin etnia, estratificados en nivel 2, cuyos padres tenían educación secundaria completa y que se distinguían por tener internet y computador en casa. Contaban con 11 a 25 libros en sus hogares y el tiempo promedio de navegación en la web era de 1 a 3 horas. Los desempeños en las materias los ubicó en rango satisfactorio, con puntaje decil global nivel 7 (el más alto de los 3 grupos). Su desempeño en inglés fue A1. En este grupo se hallaban las personas con mejores desempeños en la prueba y mejores condiciones socioeconómicas.

En el segundo grupo se posicionaron las mujeres con menos de 18 años, sin etnia y estrato 1, cuyos padres tenían educación secundaria completa. Las características propias de este clúster, mostraron que eran personas con internet, pero sin computador en casa. Tenían de 0 a 10 libros en sus viviendas. Navegaban en internet de 30 a 60 minutos. Se clasificaron en desempeño básico en ciencias naturales y sociales ciudadanas, acto que conllevó a bajar el puntaje en decil global, el cual fue de 6 en este caso.

El tercer grupo fue integrado por mujeres entres 18 a 22 años, cuyo padre no había culminado los estudios de primaria mientras que la madre contaba con educación secundaria completa. No tenían internet ni computador. Presentaron el menor desempeño en las pruebas siendo básico en lectura y matemáticas, y bajo en ciencias naturales y sociales ciudadanas, lo que los clasificó en decil 5. Su NSE fue 2 (el menor de todos los tres grupos).

3.6.1 Características del desempeño en matemáticas.

La agrupación por clúster realizada hasta el momento, ha permitido esbozar tres conglomerados en los cuales se distinguen las características principales de los estudiantes del departamento de Nariño y de donde se resalta que, en promedio, no fue frecuente encontrar personas con desempeños en nivel *avanzado* en la prueba Saber 11 pero sí en *bajo, básico y satisfactorio*. Se logró detectar también que, la mayoría de colegiales contaban con padres de familia dedicados a la agricultura, la pesca o trabajo al jornal, pero, eran los hijos de padres con ocupación profesional los que obtuvieron mejores puntajes en matemáticas. Agregado a ello se evidenció que el nivel educativo del padre y la madre se relacionaba con el desempeño del estudiante en la prueba, y que los resultados de ciencias naturales, matemáticas, lectura crítica y sociales ciudadanas presentaban una alta correlación positiva entre sí, notando que puntajes altos en una de ellas implicaban puntajes altos en las otras, o, al contrario, a puntajes bajos en una de ellas se correspondían puntajes bajos en las demás.

En adelante, y con intención de suscitar reflexiones tomando como rejilla analítica el desempeño en matemáticas, se usó una de las ventajas que ofrece el software WEKA la cual permite realizar conglomerados especificando como base un atributo en específico, por ello que en la opción *Classes to clusters evaluation* se eligió el atributo *desempeño en matemáticas*. En la opción *Choose* se activó nuevamente como función de distancia la euclidiana y como método de inició al algoritmo *k – means* ++. En la opción *numClusters* se escribió 4 conviniendo que refiere a los cuatro niveles de desempeño (*bajo, básico, superior, avanzado*).

Cabe anotar que WEKA no clasificó a plenitud todos los datos, ya que el porcentaje de instancias incorrectamente clasificadas fue del 59,91% lo que equivale a 9954 individuos. Desde otra perspectiva, en desempeño *bajo* se clasificaron 4809 personas en total, de las cuales 1020 pertenecían realmente a este nivel y 3789 no. En básico se encontraron 5059

registros, de donde 2806 eran correctamente clasificados y 2253 no lo eran. De los 3167 estudiantes del desempeño *satisfactorio*, 2355 se clasificaron adecuadamente y 812 no. Finalmente, de los 3580 registros del nivel *avanzado*, 293 pertenecían realmente a este nivel mientras que 3287 no lo hacían. La tabla 18 expone la matriz de confusión asociada a estos hechos.

Tabla 18.

Clasificación por desempeños en matemáticas según WEKA.

DESEMPEÑO	BAJO	BÁSICO	SATISFACTORIO	AVANZADO
BAJO	1020	701	50	123
BÁSICO	2604	2806	720	371
SATISFACTORIO	1161	1343	2355	2793
AVANZADO	24	209	42	293
TOTAL	**4809**	**5059**	**3167**	**3580**
% DEL TOTAL	**29**	**30**	**19**	**22**

Fuente: esta investigación.

Complementando lo anterior, en la tabla 19 se esgrime la clasificación de las variables por nivel desempeño elaborada por WEKA, la cual se explicitará más adelante.

Tabla 19.

Características de los estudiantes por nivel de desempeño en matemáticas.

	CARACTERISTICAS POR NIVEL DE DESEMPEÑO EN MATEMÁTICAS			
VARIABLE	**BAJO**	**BÁSICO**	**SATISFACTORIO**	**AVANZADO**
Género del estudiante	F	F	M	M
Edad	18 <= X < 22	X<18	X<18	X<18
Adscrito a etnia	No	No	No	No
Estrato de la vivienda	1	1	1	1
Número de personas en el hogar	3 a 4	3 a 4	3 a 4	3 a 4
Número de cuartos en el hogar	Dos	Tres	Tres	Tres
Nivel educativo del padre	Primaria incompleta	Secundaria completa	Secundaria completa	Secundaria completa
Nivel educativo de la madre	Primaria incompleta	Secundaria completa	Secundaria completa	Secundaria completa
Ocupación laboral del padre	Agricultor, pesquero, jornalero	Agricultor, pesquero, jornalero	Agricultor, pesquero, jornalero	Agricultor, pesquero, jornalero
Ocupación laboral de la madre	Trabaja en casa, no trabaja, estudia	Trabaja en casa, no trabaja, estudia	Trabaja en casa, no trabaja, estudia	Trabaja en casa, no trabaja, estudia
Internet en casa	No	Si	Si	Si

VARIABLE	CARACTERISTICAS POR NIVEL DE DESEMPEÑO EN MATEMÁTICAS			
	BAJO	**BÁSICO**	**SATISFACTORIO**	**AVANZADO**
Computador en casa	No	Si	Si	No
Número de libros en casa	0 a 10	0 a 10	0 a 10	0 a 10
Tiempo promedio de lectura al día	30 minutos o menos	30 minutos o menos	30 minutos o menos	30 minutos o menos
Dedicación diaria a internet	30 a 60 minutos	30 a 60 minutos	1 a 3 horas	30 a 60 minutos
Número de horas que trabaja a la semana	[1, 10]	0	0	0
Desempeño en lectura crítica	2	2	3	3
Desempeño en ciencias naturales	2	2	2	3
Desempeño en sociales y ciudadanas	1	2	2	3
Desempeño en inglés	A -	A -	A1	A1
Decil global	5	5	6	7
NSE individual	1	2	2	2

Fuente: esta investigación.

Con base en lo anterior es posible establecer que las características particulares por cada nivel de desempeño fueron.

- **Desempeño *bajo*.**

Las mujeres entre 18 a 22 años integraron el grupo mayoritario. Tanto el padre como la madre no habían terminado sus estudios de primaria. En promedio tenían dos habitaciones en el hogar. Los estudiantes no tenían computador ni internet en casa y trabajaban hasta 10 horas entre semana. Obtuvieron desempeño *básico* en lectura crítica y ciencias naturales, y *bajo* en sociales ciudadanas. En inglés alcanzaron el nivel “A –”. Su puntaje en decil global los catalogó en decil 5 y su NSE fue de 1.

- **Desempeño *básico*.**

Se distinguieron las mujeres menores a 18 años, cuyas viviendas contaban con 3 habitaciones. La escolaridad del padre y la madre mostraba educación secundaria completa. Contaban con internet y computador en casa. A diferencia de los educandos del desempeño *bajo*, acá los estudiantes no trabajaban entre semana. Los puntajes de las pruebas los

clasificaron en nivel *básico* en ciencias naturales, sociales ciudadanas y lectura crítica. Se observó incremento en el NSE el cual pasó de 1 a 2, con lo que se mejoró las condiciones socioeconómicas del núcleo familiar del estudiante. Las demás variables se comportaron de igual forma que los del *bajo*.

- **Desempeño *satisfactorio*.**

A diferencia de los casos anteriores, este grupo se conformó mayoritariamente por hombres cuyos padres cuentan con educación secundaria completa. Como novedad, el tiempo promedio de uso de internet era de 1 a 3 horas. El desempeño de los estudiantes en lectura crítica se halló en *satisfactorio* y la clasificación del decil global fue de 6, lo que significa que aprobaron el 60% de la prueba. Se vio mejoras en el nivel de inglés el cual pasó de "A –" a "A1". Las demás variables de este nivel se comportaron de igual forma que las del *básico*.

- **Desempeño *avanzado*.**

Sobresalieron los hombres menores de edad como grupo mayoritario de este nivel, quienes no tenían computador en casa y dedicaban un tiempo promedio de 30 a 60 minutos a navegar en internet. En las pruebas se clasificaron en nivel *satisfactorio* y en inglés se posicionaron en nivel A1. Su decil global se ubicó en nivel 7. Las demás variables de este nivel se comportaron de igual forma que las del *satisfactorio*.

3.7 DISCUSIÓN DE RESULTADOS.

Con intención de contrastar los referentes teóricos con los hallazgos encontrados en esta investigación, se comienza elaborando reflexiones sobre el desempeño en matemáticas que obtuvieron los educandos nariñenses en la prueba Saber 11 y se finaliza con el análisis del aporte de la minería de datos educativa para este estudio. Así entonces, se observó que en los resultados de los discentes nariñenses jugaron un rol importante tanto las habilidades

cognitivas del aprendiz como los aspectos socioeconómicos, demográficos, familiares, institucionales y de rendimiento en las pruebas, tal como lo exponían Álvaro Page et al. (1990), y Rodríguez Espinar (1985); aspectos que, en cierta manera, condicionaron la obtención de resultados en uno de los cuatro niveles de desempeño establecidos por el ICFES (2020a). En este sentido, la concepción sobre desempeño académico de Solano Luengo (2015) se quedó corta, puesto que a pesar que la prueba Saber 11 pretende medir con un puntaje total el rendimiento estudiantil, en esta investigación se determinó que tal puntaje se vio influenciado no solo por lo que la persona demuestra saber respecto a un área de estudio, sino que también se relaciona con el contexto personal del aprendiz.

En este punto de la discusión, el aporte de Collazos Valenzuela et al. (2019) cobra importancia puesto que en su texto presenta la forma en la que diversos autores relacionaron al rendimiento académico con los antecedentes familiares, las condiciones económicas, el número de horas de estudio y los conocimientos previos de cada aprendiz, obteniendo asociación positiva entre estas variables y el desempeño en las pruebas. De igual forma aparece el aporte de Piñero & Rodríguez (como se citó en Edel Navarro, 2003), para quien el contexto del educando "… incide positivamente sobre el desempeño escolar de los estudiantes. Ello recalca la importancia de la responsabilidad compartida entre la familia, la comunidad y la escuela en el proceso educativo". Piñero & Rodríguez (como se citó en Edel Navarro, 2003, p.5). Las conclusiones de estos autores conllevaron a establecer que el desarrollo de esta investigación encontró sintonía con las producciones elaboradas sobre capital cultural, donde a mejores condiciones socioeconómicas del aprendiz, mejores resultados se esperan en la prueba. De las producciones de los citados Collazos Valenzuela, y Piñero & Rodríguez, se determinó que las variables que ellos destacaron como importantes, también encontraron asociación con el desempeño en matemáticas, aunque en este trabajo, tales asociaciones no gozaron de tanta preminencia como en las conclusiones exhibidas por ellos, puesto que para el

contexto del educando nariñense, las correlaciones más fuertes entre las variables se distinguieron al examinar el desempeño en las pruebas.

De igual manera, el trabajo de Rodríguez Rosero et al. (2021) cobró importancia ya que sostenía que los hombres, con computadores en casa, conexión a internet y con padres con alto nivel educativo, obtenían mejores desempeños en la prueba que las mujeres con iguales características; sin embargo, un punto en discordia entre este estudio y el trabajo de Rodríguez Rosero y su grupo consistió en que ellos exponían que, el factor de mayor influencia en el rendimiento en las pruebas aludía el contar con estudiantes adscritos a instituciones oficiales y urbanas, ante lo cual cabe decir que, a pesar que el cruce de variables realizado en este trabajo no exhibió resultados diferenciadores para los aspectos institucionales, sí permitió observar, aunque mediante una leve relación, que los estudiantes del sector urbano de instituciones privadas obtuvieron mejores desempeños que los de colegios públicos.

En concordancia con lo anterior, los resultados de esta investigación también se hallaron en sintonía con lo expuesto en Peña Lozano & González Veloza (2022), quienes relacionaron al puntaje en matemáticas con las variables: género, estrato, número de personas en el hogar, educación de la madre, acceso a internet, disponibilidad de computadora en casa y horas de trabajo semanal del estudiante; empero, a diferencia de Peña Lozano y comitiva, la variable acceso a internet, la cual mencionados autores la situaron como el factor de mayor relación con el desempeño en matemáticas, no fue la de mayor preponderación en este trabajo ya que las variables con mayor relación fueron las del rendimiento en las pruebas.

Ahondando un poco más en la variable género, en este trabajó también se encontró que los hombres obtenían mejores desempeños en matemáticas que las mujeres; sin embargo, los datos ofrecidos por el ICFES no permitieron entrever la conclusión exhibida por Junca

Rodríguez (2019), quien sostiene que los hombres presentan mejor rendimiento cuando trabajan de forma individual, mientras que las mujeres sobresalen cuando desenvuelven acciones en colectivo, por tal motivo que se invita a futuros investigadores en estos tópicos a inspeccionar más al detalle en este evento.

El hecho de enfatizar que en la presente investigación sobresalieron las relaciones entre los puntajes obtenidos por los colegiales en las pruebas frente a las variables de tipo socioeconómico, demográfico, institucional y familiar expuestas por los autores anteriormente nombrados, conllevó a profundizar aún más en el estudio de Collazos Valenzuela et al. (2021), puesto que también validaron como importante estas relaciones, sin embargo, los resultados obtenidos en este trabajo se movieron en horizontes distintos a los fijados por Collazos Valenzuela y comitiva, ya que estos autores postularon al área de inglés como la de mejor desempeño en la prueba Saber, hecho que contrastó con los resultados de esta investigación ya que, para el caso de los colegiales nariñenses, el área de inglés fue la de menor puntuación puesto que la mayoría de discentes se categorizó en nivel “A –“. Además, inglés obtuvo la menor relación con matemáticas, ya que las asociaciones y las correlaciones estudiadas entre ellas no fueron tan fuertes como las observadas entre matemáticas y las demás pruebas evaluadas.

Al considerar el estudio de Timarán Pereira et al. (2019), se encontró que ellos argüían que en los años 2015 y 2016, el porcentaje de colegiales con desempeño bajo en la aplicación de las pruebas Saber 11 era mayor al signado como alto, evento ante lo cual cabe decir que, en esta investigación, salvo el área de lectura crítica en donde el 10.03% de los educandos obtuvieron nivel *avanzado* mientras que el 5.83% *bajo*, en las pruebas de matemáticas, ciencias, sociales e inglés, también se vislumbró mayor presencia de estudiantes en nivel *bajo* frente al *avanzado*. Empero, se resalta el hecho que, en lectura y matemáticas, con el 43.48% y

46.05% respectivamente, predominó la categorización de educandos en nivel *satisfactorio*, mientras que en ciencias naturales y sociales sobresalía el nivel *básico* y en inglés el "A –". Estos hechos conllevaron a pensar que, en las aulas de Nariño, los esfuerzos por mejorar el desempeño en matemáticas poco a poco van dando frutos, puesto que el haber clasificado a la mayoría de educandos en nivel *satisfactorio* se concibió como un evento favorable en la intención de matematizar a los estudiantes. Pero, ¿por qué son pocos los estudiantes con desempeño *avanzado* en matemáticas?

La respuesta a esta pregunta implicaría abordar un estudio explicativo en donde se analicen más a fondo los factores de influencia en el desempeño en matemáticas y ante lo cual se invita a futuros investigadores a ahondar más en detalle en estos temas, sin embargo, a manera de una primera reflexión puede meditarse en qué tan alineadas se encuentran las pruebas Saber 11 con los documentos guías para la educación básica y media expuestos en MEN (2006), Ruta Maestra (2017), MEN (2018), por citar algunos textos; y qué tan sintonizadas se hallan con las evaluaciones internacionales. Al respecto López (2013), enfatizando en la alienación de las pruebas externas con la Saber noveno, establece que, en términos generales, no hay una fuerte alineación entre los que se evalúa internacionalmente con lo abordado en el país, siendo esto un factor que motiva los bajos desempeños en el ámbito externo, y tal vez sea una de las razones por las que en el texto de Fernandes Cristóvão (2010), se encontró que países con nivel socioeconómico y de desarrollo humano similares al Colombia, hayan logrado promedios significativamente más altos en el contexto internacional.

Así entonces, una posible razón por la que se hallen pocos estudiantes en nivel *avanzado* sea por la falta de alineación entre las pruebas, en donde el componente cognitivo evaluado en esta fase sea muy profundo, o trate temas que por diversas razones los docentes no alcanzan a trabajar con los estudiantes en las aulas, o porque quizá la estructura de la

prueba no favorece la solución de problemáticas desde saberes ancestrales en donde juegue rol importante el conocimiento etnomatemático, hecho que posiblemente potenciaría los resultados de los discentes nariñenses en matemáticas ya que en el departamento se encuentran diversos grupos étnicos en donde la educación formal acopla los intereses culturales.

Agregado a lo anterior, el bajo número de personas con desempeño *avanzado* en matemáticas, plantea la posibilidad que en el proceso de alfabetización matemática de los aprendices, y retomando las ideas de Delors (1996), Tobón (2013), Tobón Tobón et al. (2010), se priorice las competencias del Saber sobre el Hacer, es decir, es probable que en los recintos de clase se dé prelación a la aprehensión y conceptualización de los objetos matemáticos, y poco a la resolución de ejercicios enfatizada en el contraste de diversos procedimientos para llegar a un resultado, donde se considere las ventajas o desventajas que cada forma de solución pueda traer consigo. En este sentido entonces, las palabras de Delors (1996) y de Moreno Olivos (2012) cobran validez, al manifestar que "… la enseñanza escolar se orienta esencialmente, por no decir que, de manera exclusiva, hacia el aprender a conocer y, en menor medida, el aprender a hacer". (Delors, 1996, p. 96).

De allí que reflexionar en la forma cómo se trabaja y evalúa en las aulas colombianas, le facilitaría al ICFES instaurar jornadas de capacitación para que los docentes aprendan a formular y resolver preguntas al estilo de las evaluadas en la prueba Saber, estableciendo así lineamientos más claros sobre la medición de competencias en matemáticas para los estudiantes. En este punto se medita en que quizá la autonomía que brinda el Estado colombiano a las instituciones educativas en la estructuración curricular de sus planes educativos (Resolución 2343 de junio 5 de 1996), no ha sido ampliamente aprovechada por algunos colegios ya que en la evaluación de los aprendizajes se ha favorecido la medición del

componente cognitivo, dejando relegado el componente procedimental y las dimensiones socioeconómicas del estudiante, de allí que Tobón haya manifestado que "… con frecuencia la evaluación de la calidad educativa se hace aplicando pruebas para determinar el logro cognitivo, y esto deja de lado las inteligencias múltiples, la actuación ante problemas reales y la ética". (Tobón, 2013, p. 17). Así entonces, emerge aquí un nuevo campo de estudio, en donde existen muchas cosas por develar pero que por desventura, la base de datos con la que se trabajó en esta investigación, no permitió elaborar mayores conclusiones sobre la forma en la que los docentes efectúan su trabajo en las aulas, por lo que se invita a los lectores a profundizar en estos aspectos en futuras investigaciones.

Como conclusión sobre las ideas expuestas para el rendimiento académico se expone que, a pesar que la prueba Saber 11 atravesó por un proceso de reestructuración el cual le permitió brindar nuevas pautas para la evaluación de competencias en matemáticas a través de los procesos de formular y resolver problemas, modelar procesos y fenómenos de la realidad, comunicar ideas en matemáticas, razonar, y, comparar y ejercitar procedimientos o algoritmos (MEN, 2006), es probable que una tarea pendiente para el ICFES sea la de fomentar jornadas de capacitación para docentes en redacción de preguntas basadas en competencias, e instauración de clases para el ejercicio del pensamiento crítico y vivencias ciudadanas, con las cuales los profesores puedan efectuar acciones de retroalimentación escolar que permitan mejorar el desempeño estudiantil en el área de matemáticas. Además, y según mostró López (2013), conviene revisar con mayor vehemencia los procesos de alineación entre los ítems evaluados en la prueba Saber, los lineamientos dados para la educación básica y media, y lo inspeccionado en los test internacionales, a fin de conseguir una educación de calidad fundamentada en el ejercicio de las competencias.

Como dato adicional se resalta que sería importante que en la revisión de la alineación de las pruebas se implementen también la evaluación de saberes ancestrales, hecho que aviva al componente etnomatemático. Además, sería interesante que el ICFES contemple integrar en el puntaje total una ponderación especial de los aspectos socioeconómicos, institucionales, demográficos y familiares, de cada región, o como mencionaba Rodríguez Espinar (1985), también se acojan los subcriterios psicológicos, sociológicos y de carácter didáctico en la valoración de resultados.

En otro ángulo, la implementación de las fases de la metodología CRISP-DM en esta investigación favoreció la estructuración de un plan de trabajo el cual permitió comprender y organizar la información disponible, en este sentido, las ideas de Riquelme Santos et al. (2006), Oviedo Carrascal & Jiménez Giraldo (2019), y, Pérez Gutiérrez (2020), encontraron sintonía en el presente estudio al exhibir a CRISP-DM como una metodología funcional y complementaria al trabajo con minería de datos.

En cuanto a la labor con DM, este trabajo encontró eco con las concepciones de Beltrán Martínez (s.f), y Vargas Agurto (2014), al visualizar a la minería de datos como el proceso que facilitó descubrir las relaciones existentes entre el desempeño en matemáticas y las variables que integraron la prueba Saber 11. En este punto se dio prioridad al uso de técnicas exploratorias y de segmentación, las que según Blanco Villafañe (2015), son útiles en la categorización del desempeño estudiantil. Esta investigación se sintonizó también con Medina & Galván (2007) en lo concerniente a los procesos de imputación de datos, ya que permitió analizar la aleatoriedad de los datos faltantes y con base en ello elegir el modelo más adecuado de imputación.

Finalmente, al trabajar con clúster se observó lo dicho por Hair et al. (1999), quienes sostenían que el *clúster* siempre crearía grupos así se cuente con una estructura definida para los datos, este aspecto se notó al contrastar los clústeres formados en RStudio y en WEKA, ya que los conglomerados formados, aunque parecidos, no eran exactamente iguales.

4 CONCLUSIONES.

El trabajo hasta aquí desarrollado ha facilitado caracterizar a los estudiantes nariñenses según sus desempeños en la prueba Saber 11 periodo 2021 B, y ha evidenciado las relaciones existentes entre los puntajes obtenidos en matemáticas y las demás áreas evaluadas por el ICFES. Este evento permitió elaborar un acercamiento al desempeño de los estudiantes en la prueba Saber tomando como referencia particular el área de matemáticas. A continuación, se destacan los hallazgos más relevantes identificados en la población estudiada, los cuales se encuentran discriminados en, características generales de los estudiantes, características de los puntajes altos en matemáticas, y, problemas abiertos para próximas investigaciones.

4.1 CARACTERÍSTICAS GENERALES DE LOS ESTUDIANTES.

Con referencia al rendimiento en las pruebas, los resultados de la prueba Saber 11 exhibieron que la mayoría de estudiantes nariñenses obtuvieron desempeños *satisfactorios* en lectura crítica y matemáticas, y *básicos* en ciencias naturales y sociales ciudadanas, mostrando con ello mejores resultados en lectura y matemáticas que en naturales y sociales. Además, no fue común encontrar personas clasificadas en desempeño *avanzado* en las pruebas ya que, a excepción de lectura crítica, menos del 6% de los individuos obtenían más de 70 puntos. En inglés se encontró que más del 55% de los educandos se hallaban en nivel “A –”. Estos aspectos favorecieron que el puntaje global más distintivo de los estudiantes fuese el decil 5.

El cruce de variables realizado entre el rendimiento en las pruebas y el desempeño en matemáticas mostró una asociación fuerte, de hecho la mayor entre las relaciones observadas, acontecimiento que fue validado al estudiar la correlación entre los puntajes de matemáticas, lectura crítica, sociales ciudadanas y ciencias naturales, y que conllevó a concluir que se encontraban altamente correlacionadas positivamente, indicando que, en términos generales,

puntajes altos en una de ellas implicaban puntajes altos en las otras, o, al contrario, a puntajes bajos en una de ellas se correspondían puntajes bajos en las demás.

La prueba de inglés se movió en horizontes diferentes, ya que mostró una correlación moderada entre ella y el desempeño en matemáticas, hecho que favoreció que en el análisis factorial se posicionara en un eje distinto y con lo que se estableció que la relación entre inglés y desempeño en matemática no se clasificara como fuerte sino como moderada. Agregado a ello, cabe recordar que, en la obtención del puntaje global, la prueba de inglés era la que menor valor ponderado tenía (ver literal 1.7.2 del presente texto), denotando con eso que para el ICFES las pruebas de lectura crítica, matemáticas, sociales ciudadanas y ciencias naturales eran de mayor relevancia.

En los aspectos socioeconómicos se observó que, en cuanto a la escolaridad, los padres con estudios secundarios completos constituyeron el grupo mayoritario y tan solo el 19.99% de ellos contaban con estudios técnicos o profesionales (terminados o sin terminar), con lo que se resalta que no fue frecuente encontrar estudiantes cuyos padres tuvieran alta preparación académica. Se observó también que la ocupación laboral más común para los padres aludía a personas agricultoras, pesqueros o jornaleros, mientras que las madres oficiaban como amas de casa, no trabajaban o se hallaban estudiando. Adjunto a esto se vio que la mayoría de colegiales se clasificaron en estrato 1 y NSE 1.

Los aspectos familiares permitieron establecer que más de la mitad de la población contaba con hogares integrados por 3 a 4 personas y con viviendas de dos o tres habitaciones, siendo estos los eventos más frecuentes, además era común encontrar que, en casa, los estudiantes tuvieran internet, computador y hasta 10 libros para realizar actividades de lectura por entretenimiento.

De los aspectos demográficos se concluyó que más de la mitad de los estudiantes que presentaron la prueba eran mujeres y que el 97.4% no superaban los 22 años de edad, con lo que se tenía una población joven que, por lo general, no trabajaba entre semana ni tampoco pertenecía a grupos étnicos. Se halló también que, en general, los estudiantes dedicaban menos de 30 minutos a leer por diversión e invertían más tiempo en internet que en actividades de lectura. En esta etapa se halló también que los municipios de Pasto, Tumaco e Ipiales albergaban casi que a la mitad de los discentes del departamento.

En cuanto a los aspectos institucionales se encontró que más del 96% de los planteles educativos eran mixtos, no bilingües y de calendario A. El 88.5% eran públicos y el 73.46% se ubicaban en zona urbana. Se halló que la modalidad de egreso más común era ser académico y que en el 74.32% de los casos, los estudiantes estaban adscritos a la jornada de la mañana. La clasificación socioeconómica de las instituciones las ubicaba, mayoritariamente, en nivel 2. Como aspecto a detallar en esta instancia se tuvo que, el cruce de variables efectuado entre los aspectos institucionales y el desempeño en matemáticas, no mostró cruces de gran importancia.

En resumen, los estudiantes del departamento de Nariño presentaban como características generales que eran mayoritariamente mujeres, menores de edad, sin etnia, con estrato 1, para quienes su núcleo familiar se integraba por 3 a 4 personas, las cuales habitaban en dos o tres habitaciones. Tanto el padre como la madre contaban con secundaria completa, y el padre se dedicaba a labores de agricultura, pesca o trabajo al jornal, y la madre a trabajar en casa, no trabajar o estudiar. Tenían internet, computador y hasta 10 libros en casa. Dedicaban menos de 30 minutos a realizar actividades de lectura y de 30 a 60 minutos a navegar en internet. No trabajaban entre semana, y los desempeños en las pruebas más característicos

fueron: *satisfactorio* en matemáticas y lectura crítica, *básico* en sociales ciudadanas y ciencias naturales, “A –” en inglés. Su puntaje promedio del decil global fue 5.0 y tenían NSE 2.

Agregado a lo anterior, y gracias al análisis clúster, se notó también los educandos nariñenses podían clasificarse en tres subgrupos, los cuales aunque no perfectamente cohesionados ni separados puesto que el coeficiente silueta obtenido en el k-prototipos fue bajo (0.18), sí logró encontrar rasgos referentes de cada conglomerado como por ejemplo que, en el primer grupo se hallaban los hombres menores de edad sin etnia y con estratificación nivel 2, cuyos padres contaban con educación secundaria completa y que se distinguían por tener internet y computador en casa, contaban con 11 a 25 libros y el tiempo de navegación en la web era de 1 a 3 horas. Los desempeños en las materias los ubicaba en rango satisfactorio, su puntaje decil global era nivel 7 correspondiente al más alto de los 3 grupos, y su desempeño en inglés fue A1. Cabe decir que en este grupo se hallaban las personas con mejores desempeños en la prueba y mejores condiciones socioeconómicas.

En el segundo grupo se posicionaron las mujeres menores a 18 años, sin etnia y estrato 1, cuyos padres tenían educación secundaria completa. Las características propias de este clúster, mostraron que, en promedio, eran personas con internet, pero sin computador en casa, además, el número de libros en los hogares era menor (0 a 10) y navegaban en internet de 30 a 60 minutos. En el desempeño académico disminuyó en ciencias naturales y sociales ciudadanas ya que se clasificaron en desempeño básico acto que conllevó a bajar el puntaje en decil global, el cual fue de 6 en este caso.

El tercer grupo fue integrado por mujeres entres 18 a 22 años, cuyo padre no había culminado los estudios de primaria mientras que la madre contaba con educación secundaria completa. No tenían internet ni computador. Presentaron el menor desempeño en las pruebas

siendo básico en lectura y matemáticas, y bajo en ciencias naturales y sociales ciudadanas, lo que los clasificó en decil 5 y su NSE fue 2, el menor de todos los tres grupos.

Por último, los cruces de variables permitieron establecer características de los estudiantes según su desempeño en matemáticas, siendo los de nivel bajo, básico y satisfactorio los descritos a continuación, mientras que los puntajes altos (avanzados) se describen en el literal 4.2

- ***Desempeño bajo*.** Refiere a estudiantes mujeres cuya edad superaba los 22 años. Más de la mitad de los individuos de este grupo pertenecían a alguna etnia. Tenían el mayor porcentaje de padres sin estudios en comparación con los otros niveles y con padres que tenían educación primaria incompleta. Sobresalía el padre que trabajaba en casa, no trabajaba o estudiaba, y la madre que laboraba como agricultora, pesquera o jornalera. Su NSE más común fue 1. Las familias se integraban por 5 a 8 personas y más de la mitad de estudiantes no tenían internet en casa ni computador. Por lo general, invertían hasta 1 hora de su tiempo en internet o por el contrario no navegaban en la red. Más de la mitad de educandos en este nivel trabajan más de 1 hora a la semana. Y residían en municipios como Tumaco, Barbacoas, El Charco y la zona costera.

Presentaban nivel *bajo o básico* en lectura crítica y *nivel bajo* en ciencias naturales y sociales ciudadanas En inglés resaltó la categoría “A –”. En este grupo se evidenció ausencia de personas que simultáneamente cuenten con desempeño bajo en matemáticas y *avanzado* en ciencias naturales o en sociales ciudadanas. Además, los resultados de este nivel ubicaron el puntaje global en el decil 4, resaltando que ningún estudiante del decil 4 logró puntuaciones altas en matemáticas.

- ***Desempeño básico*.** Refiere a estudiantes mujeres cuya edad oscilaba entre los 18 y 22 años. No pertenecían a grupos étnicos. Navegaban en internet un tiempo promedio de 30 minutos a 3 horas. Contaban con el mayor porcentaje de padres con educación primaria (completa o incompleta), y con padres que ejercían labores de oficios varios o independientes, mientras que las madres se dedicaban a la agricultura, pesca o trabajo al jornal. Su NSE más común era nivel 1. Más de la mitad de educandos no contaban con computador en sus viviendas, pero sí con internet. Residían en municipios como Tumaco y zonas aledañas, en Santa Barbará y Cumbal.

Además, los discentes con desempeño básico en matemáticas tendían a clasificarse en nivel *básico* en lectura crítica y nivel *bajo o básico* en ciencias naturales y sociales ciudadanas. En inglés resaltó la categoría “A –”. En este grupo no se evidenciaron personas que simultáneamente cuenten con desempeño *básico* en matemáticas y *avanzado* en ciencias naturales o en sociales ciudadanas. Además, los resultados de este nivel ubicaron el puntaje global en el decil 5.

- ***Desempeño satisfactorio*.** Fue integrado en su mayoría por hombres menores de 18 años y sin etnia, cuyos padres trabajaban como independientes o conductor de vehículos y las madres eran microempresarias. Su NSE más común los posicionó en nivel 2. Más de la mitad de educandos no contaban con computador en sus viviendas, pero sí con internet. El tiempo promedio de uso de internet era de 30 minutos a 3 horas. Residían en Pasto, Ipiales, San Bernardo, Gualmatán y Consacá.

Por lo general, los educandos con desempeño *satisfactorio* en matemáticas tendían a posicionarse en nivel *satisfactorio* en lectura y *básico o satisfactorio* en ciencias naturales y

sociales ciudadanas. En inglés resaltó la categoría “A1”. Y los resultados de este nivel ubicaron el puntaje global en el decil 6.

4.2 CARACTERÍSTICAS DE LOS PUNTAJES ALTOS EN MATEMÁTICAS.

Los estudiantes con puntajes avanzados en matemáticas presentaron como aspectos relevantes los descritos en seguida.

- ***Desempeño avanzado*.** Refería a hombres menores de edad, sin etnia y cuyos padres tenían educación técnica (completa o incompleta), profesional (completa o incompleta), o, secundaria (completa o incompleta). Tanto el padre como la madre se desempeñaban como trabajadores profesionales. Su NSE los clasificó en nivel 3. Mas del 73% de colegiales tenían internet y computador en casa, y más del 60% contaban con 11 a 100 textos en el hogar. Su tiempo promedio de navegación en internet superaba las 3 horas. Residían en lugares como Pasto, Ipiales, San Bernardo, Colón, Samaniego, Sandoná y Tangua.

Con generalidad, las personas categorizadas en desempeño *avanzado en matemáticas* puntuaron en nivel *avanzado* en lectura y en desempeño *satisfactorio o avanzado* en ciencias naturales y sociales ciudadanas. En inglés resaltó la categoría “A2”. En este grupo se evidenció ausencia de personas que simultáneamente cuenten con desempeño *avanzado* en matemáticas y *bajo en lectura crítica* y *ciencias naturales*. Además, los resultados de este nivel ubicaron el puntaje global en el decil 7, resaltando que ningún estudiante del decil 7 obtuvo calificaciones menores a 35 puntos en matemáticas.

4.3 PROBLEMAS ABIERTOS PARA PRÓXIMAS INVESTIGACIONES.

Los resultados aquí analizados tuvieron un factor adicional ya que tras la pandemia COVID 19, los educandos se vieron forzados a desarrollar sesiones de clase virtual en donde

se pusieron en juego múltiples factores, los que en muchas ocasiones afectaron la comunicación entre los fenómenos escolares, los métodos de enseñanza-aprendizaje y los resultados esperados, hecho que pudo poner en contravía la asertividad de colegiales con las clases y sus docentes. El área de matemáticas juega un rol importante en esta etapa ya que ella misma se recubre de objetos abstractos y los que en ocasiones no son de fácil manipulación por parte de los educandos, hecho que suscita la aparición de algunas ideas las cuales se plantean como posibles futuras investigaciones. Por un lado, se invita a profundizar en trabajos que analicen las estrategias pedagógicas y didácticas que se llevaron a cabo en municipios que destacaron por tener puntajes altos en matemáticas, con el fin de hacer un comparativo con los municipios que obtuvieron puntajes bajos para estructurar planes de aula que fortalezcan las competencias de los estudiantes y mejoren la enseñanza y aprendizaje de las matemáticas. Se invita también a hacer un contraste entre las ventajas o desventajas que pudo traer consigo la educación en tiempos de pandemia y el rendimiento escolar, detectar el grado de compromiso de los docentes, padres de familia, personal de la institución y demás actores educativos frente al aprendizaje de los educandos desde la virtualidad.

Se invita también a analizar la alineación existente entre las preguntas expuestas en la prueba Saber 11 y lo establecido en los estándares y lineamientos curriculares para la educación básica y media, bajo cuestionamientos como ¿lo que evalúa el ICFES en la prueba Saber se halla en sintonía con lo exigido por el MEN?, o, ¿qué tan sintonizados estuvieron los planes de estudio de los docentes frente a los requerimientos de la educación en tiempos de pandemia? Otro aspecto por investigar que cobra interés acoge la inquietud sobre ¿en qué manera las clases de matemáticas favorecen el desarrollo de competencias ciudadanas en los estudiantes?, y ¿cómo estas clases se orientan al ejercicio de la ciudadanía crítica?

Se invita al lector a comparar los datos aquí obtenidos con el uso de otros softwares distintos a RStudio y WEKA, a fin de establecer qué programa entrega mejor tratamiento en los datos. Los dos programas con los que se trabajó en esta investigación son de naturaleza libre es decir no requieren licencia, quizá un software con licencia pueda brindar los coeficientes silueta, de viabilidad e importancia en la agrupación de clústeres categóricos, hecho que WEKA no lo dejó entrever en sus resultados. O puede también el lector trabajar en la adaptación o creación de un código en RStudio que permita visualizar con el comando K-modes, las medidas de viabilidad e importancia para los clústeres categóricos, es decir cuyos datos estén en texto y no discretizados, asimismo, puede elaborar códigos que faciliten la representación visual de los clústeres elaborados con el comando k-modes.

Por último, se invita a los lectores a ampliar las ideas expuestas en este estudio, referenciando también los resultados de los educandos que presentan la prueba Saber 11 en toda Colombia, bajo un estudio por bloques o regiones. El análisis puede enriquecerse con otras técnicas de minería de datos como árboles de decisión, o aquellas que además de clasificar permitan también realizar modelos predictivos.

REFERENCIAS

ACUERDO 014 del 11 de Octubre. (2022). *Modalidades de Trabajos de Grado para Optar por el Título de Maestría en Estadística Aplicada.* (UDENAR, Ed.)

Alkandari, A., Law, J., Alhashmi, H., Alshammari, O., & Bhandari5, P. (15 de Enero de 2021). Staying (Mentally) Healthy: The Impact of COVID-19 on Personal and Professional Lives. *Techniques and Innovations in Gastrointestinal Endoscopy*, 199-206. doi:https://doi.org/10.1016/j.tige.2021.01.003

Álvaro Page, M., Bueno Monreal, M. J., Calleja Sopeña, J. Á., Cerdán Victoria, J., Echeverría Cubias, M. J., García López, C., . . . Trillo Marco, C. (1990). *Hacia un modelo causal del rendimiento académico.* Madrid: CIDE.

Araneda, P. (08 de Agosto de 2021). *Datos faltantes.* Obtenido de https://rpubs.com/paraneda/missingdata

Arias Serna, D. (30 de Mayo de 2009). *¿Número de estudiantes por aula incide en calidad?* Obtenido de https://www.cronicadelquindio.com/noticias/cultura-2/nmero-de-estudiantes-por-aula-incide-en-calidad

Beltrán Martínez, B. (s.f). *Minería de datos.* México: Benemérita Universidad Autónoma de Puebla. Recuperado el 28 de Junio de 2022, de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://bbeltran.cs.buap.mx/NotasMD.pdf

Blanco Villafañe, V. P. (2015). *Análisis del Desempeño Académico del Examen de Estado para el Ingreso a la Educación Superior Aplicando Minería de Datos.* Bogotá: Universidad Nacional.

Bogoya, D. (2006). *Evaluación educativa en Colombia. Seminario internacional de evaluación.* Recuperado el 16 de Febrero de 2021, de http://www.catedras-bogota.unal.edu.co/web/ancizar/ancizar/2009I/2007II/documentos/s9_bogoyapres.pdf

Bolaños, L. (10 de Marzo de 2020). *Anális factorial exploratorio.* Obtenido de https://rpubs.com/luis_bolanos/FA

Bravo Salinas, N. (2004). Acerca de las competencias desde un enfoque pedagógico. *Areté, 4*(1), 25-43. Recuperado el 8 de Enero de 2021, de https://arete.ibero.edu.co/article/view/559

Caracol Radio. (22 de Febrero de 2022). *El Liceo de la Udenar el mejor colegio público del país.* Obtenido de https://caracol.com.co/emisora/2022/02/22/pasto/1645550824_035361.html

Casali, A., & Torres, D. (2021). Impacto del COVID-19 en docentes universitarios argentinos: cambio de prácticas, dificultades y aumento del estrés. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*(28), 423-431. doi:https://doi.org/10.24215/18509959.28.e53

Chica Gómez, S. M., Galvis Gutiérrez, D. M., & Ramírez Hassan, A. (2010). Determinantes del rendimiento académico en Colombia. Pruebas ICFES - Saber 11o, 2009*. *Revista Universidad EAFIT, 46*(160), 48-72. Obtenido de http://hdl.handle.net/10784/16801

Clavijo M., J. A., & Granada D., H. A. (2016). Una técnica de clasificación con variables categóricas. *Ciencia en Desarrollo, 7*(1), 15-20. doi:doi.org/10.19053/01217488.4226

CNA. (1998). *Sistema Nacional de Acreditación en Colombia.* Obtenido de https://www.mineducacion.gov.co/CNA/1741/article-186365.html

CNA. (2006). *Lineamientos para la acreditación de programas.* Bogotá: MEN.

Collazos Valenzuela, A. C., Quintero Medina, M. V., & Trujillo Caicedo, K. N. (2021). Determinantes del rendimiento académico de la prueba saber 11 en Colombia duante el periodo 2014-2019. *Panoráma, 15*(29), 103-126. doi:https://doi.org/10.15765/pnrm.v15i29.1723

Colombia aprende. (2022). *Generación E. Programa.* Obtenido de https://especiales.colombiaaprende.edu.co/generacione/programa.html

Constitución Política de Colombia. (20 de Julio de 1991). Obtenido de http://wsp.presidencia.gov.co/Normativa/Documents/Constitucion-Politica-Colombia.pdf

Cruz Guzmán, O. d., & Benítez Granados, J. (2020). Las crisis también pueden promover el aprendizaje, impacto del Covid-19 en prácticas docentes. *Revista Latinoamericana de Estudios Educativos, L*, 291-302. doi:https://doi.org/10.48102/rlee.2020.50.ESPECIAL.114

Cuéllar Caicedo, E. J., Guerrero, S., & López, D. (2016). Propuesta para la construcción de un índice socioeconómico para los estudiantes que presentan las pruebas Saber Pro. *Comunicaciones en Estadística, 9*(1), 93-106. doi:https://doi.org/10.15332/s2027-3355.2016.0001.05

Decreto No. 869. (17 de Marzo de 2010). *Por el cual se reglamenta el Examen de Estado de la Educación Media, ICFES - SABER 11°.* Obtenido de https://www.mineducacion.gov.co/1621/articles-221588_archivo_pdf_decreto_869.pdf

Delors, J. (1996). *La educación encierra un tesoro. Informe a la UNESCO de la Comisión Internacional sobre la educación para el siglo XXI.* Madrid, España: Santillana Ediciones UNESCO.

Dicovskiy Riobóo, L. M., & Pedroza Pacheco, M. E. (2017). Métodos univariados y multivariados para analizar el rendimiento académico de la carrera de Ingeniería Agroindustrial en la UNI región norte, Estelí, Nicaragua. *Revista Cientifica de FAREM-Esteli, 6*(22), 3-17. Recuperado el 10 de Mayo de 2022, de https://www.lamjol.info/index.php/FAREM/article/view/4513/4234

Echeverria Castillo, Y., Laborí de la Nuez, B., & Soriano Sifonte, R. (2020). Extensión de índices de validación de grupo en la herramienta WEKA para la evaluación de algoritmos de agrupamiento. *Serie Científica de la Universidad de las Ciencias Informáticas, 13*(9), 179-187.

Edel Navarro, R. (2003). El rendimiento académico: concepto, investigación y desarrollo. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 1*(2), 0. Recuperado el 13 de Junio de 2022, de https://www.redalyc.org/comocitar.oa?id=55110208

Equipo editorial, Etecé. (1 de Octubre de 2020). *Dato*. Recuperado el 7 de Julio de 2022, de Concepto.de.: https://concepto.de/dato/

Fernandes Cristóvão, M. I. (2010). *Resultados de Colombia en TIMSS 2007. Resumen ejecutivo.* Bogotá: ICFES.

Gamboa Unsihuay, J. E., & Zuñiga Blanco, A. (2021). Modelos de minería de datos aplicados al rendimiento académico universitario: Educación virtual durante pandemia COVID-19. *Tierra Nuestra, 15*(1), 18-28. doi:http://dx.doi.org/10.21704/rtn.v15i1.1812

Haider, A. S., & Al-Salman, S. (2020). Dataset of Jordanian university students' psychological health impacted by using e-learning tools during COVID-19. *Data in Brief, 32*(106104). doi:https://doi.org/10.1016/j.dib.2020.106104

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1999). *Análisis multivariante* (Quinta ed.). (A. Otero, Ed., E. Prentice, & D. Cano, Trads.) Madrid: PRENTICE HALL.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (Sexta ed.). México D. F: Mc Graw Hill Education.

ICETEX. (04 de Junio de 2022). *Ser Pilo Paga 1*. Obtenido de https://web.icetex.gov.co/es/-/ser-pilo-paga-1#:~:text=Es%20el%20otorgamiento%20de%20cr%C3%A9ditos,del%20programa%20Ser%20Pilo%20Paga.

ICFES. (2019). *Marco de referencia de la prueba de matemáticas Saber 11.°.* Bogotá: Dirección de Evaluación, Icfes.

ICFES. (2020a). *Establecimiento de estándares de desempeño: descripción de niveles y puntos de corte.* Obtenido de https://www.icfes.gov.co/documents/39286/443287/Niveles+de+desempe%C3%B1o.pdf

ICFES. (2020b). *Informe Nacional de Resultados para Colombia - PISA 2018.* Obtenido de https://www.studocu.com/co/document/corporacion-universitaria-u-de-colombia/sistemas-informacion/informe-nacional-de-resultados-pisa-2018/16434748

Junca Rodríguez, G. A. (2019). Desempeño académico en las pruebas Saber 11. *Panoráma Económico, 27*(1), 8-38. Recuperado el 01 de Junio de 2022, de https://dialnet.unirioja.es/servlet/articulo?codigo=7456221

Ley 115. (8 de Febrero de 1994). *Ley General de Educación y Desarrollos Reglamentarios.* Obtenido de https://www.mineducacion.gov.co/1621/articles-85906_archivo_pdf.pdf

Ley 1324. (13 de Julio de 2009). *LEY 1324 DE 2009.* Obtenido de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=36838

Ley 30. (28 de Diciembre de 1992). *Fundamentos de la Educación Superior.* Obtenido de http://www.legal.unal.edu.co/rlunal/home/doc.jsp?d_i=34632

Lopez, A. A. (2013). Alineación entre las evaluaciones externas y los estándares académicos: el caso de la prueba Saber. *RELIEVE. Revista Electrónica de Investigación, 19*(2), 1-16.

Lozano Díaz, A., Fernández Prados, J. S., Figueredo Canosa, V., & Martínez Martínez, A. M. (2020). Impactos del Confinamiento por el COVID-19 entre Universitarios: Satisfacción Vital, Resiliencia y Capital Social Online. *International Journal of Sociology of Education, Special Issue: COVID-19*

Crisis and Socioeducative Inequalities and Strategies to Overcome them, 9(Extra 1), 79-104. doi:10.17583/rise.2020.5925

Martínez, A. (21 de Febrero de 2019). *Toda Colombia.* Obtenido de https://www.todacolombia.com/departamentos-de-colombia/narino/index.html

Medina , F., & Galván, M. (Julio de 2007). *Imputación de datos: teoría y práctica.* Obtenido de https://repositorio.cepal.org/handle/11362/4755

MEN. (2006). *Estándares Básicos de Competencias en Lenguaje, Matemáticas, Ciencias y Ciudadanas.* Bogotá: Ministerio de Educación Nacional.

MEN. (8 de Agosto de 2018). *La calidad: esencia de la educación en las aulas de clase.* Obtenido de https://www.mineducacion.gov.co/portal/salaprensa/Noticias/373629:La-calidad-esencia-de-la-educacion-en-las-aulas-de-clase

MEN. (2018). *Resumen Ejecutivo PISA.* Bogotá: ICFES.

MEN. (s.f). *Documentación del examen Saber 11.* Bogotá: Mineducación.

Menacho Chiok, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos, 78*(1), 26-33. doi:http://dx.doi.org/10.21704/ac.v78i1.811

Moreno Olivos, T. (2012). La evaluación de competencias en educación. *Revista electronica Sinéctica*(39), 1-20. Recuperado el 15 de Marzo de 2021, de https://www.redalyc.org/pdf/998/99826889010.pdf

Mullis, I., Martin, M., & Foy, P. (2008). *TIMSS 2007 International Mathematics Report.* Chestnut Hill, Massachusetts: TIMSS & PIRLS International Study Center.

Núñez Cárdenas, F. d. (s.f). *El proceso de minería de datos*. Obtenido de https://www.uaeh.edu.mx/scige/boletin/huejutla/n1/m2.html#refe

Ochoa, L. L., Rosas Paredes, K., & Esquicha Tejada, J. (2017). Estudio Comparativo de Técnicas no Supervisadas de Minería de Datos para Segmentación de Alumnos. *Global Partnerships for Development and Engineering Education: Proceedings of the 15th LACCEI International Multi-Conference for Engineering, Education and Technology* (págs. 1-10). Boca Raton, Florida: LACCEI. doi:http://dx.doi.org/10.18687/LACCEI2017.1.1.115

Oviedo Carrascal, A. I., & Jiménez Giraldo, J. (2019). Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO. *Revista Politécnica, 15*(29), 128-140. doi:https://doi.org/10.33571/rpolitec.v15n29a10

Pastrán Ramírez , L. F., & Roa Peña, N. J. (Febrero de 2015). *Clasificación mediante K - modas para el caso de variables categóricas.* Obtenido de https://1library.co/document/myjnd2pq-clasificacion-mediante-k-modas-caso-variables-categoricas.html

Peña Lozano, Y., & González Veloza, J. J. (2022). Modelo de predicción de los resultados de la prueba icfes saber 11 en el área de matemáticas a partir de variables socioeconómicas. *Studies in Engineering and Exact Sciences, 3*(1), 52-68. doi:10.54021/seesv3n1-006

Peña, D. (2002). *Análisis de datos multivariantes.* McGraw-Hill.

Pérez Gutiérrez, B. R. (2020). Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico. *Revista UIS Ingenierías, 19*(1), 193-204. doi:https://doi.org/10.18273/revuin.v19n1-2020018

Portafolio. (06 de 02 de 2022). *Saber y Pisa medirán golpe de la pandemia en el aprendizaje.* Recuperado el 19 de Mayo de 2022, de https://www.portafolio.co/tendencias/educacion-pruebas-saber-y-pisa-mediran-golpe-de-la-pandemia-en-el-aprendizaje-561408

Prada Núñez, R., Gamboa Suárez, A. A., & Hernández Suárez, C. A. (2021). Efectos depresivos del aislamiento preventivo obligatorio asociados a la pandemia del Covid-19 en docentes y estudiantes de una universidad pública en Colombia. *Psicogente, 24*(45), 1-20. doi:https://doi.org/10.17081/psico.24.45.4156

Rendón, E., Zepeda, R., Barrueta, E., & Itzel - María, A. (2015). El algoritmo de agrupamiento K-Modas: Un caso de estudio. *Revista de Tecnología e Innovación, 2*(5), 929-941.

Resolución 2343 de junio 5 de 1996. Obtenido de http://bibliotecadigital.usb.edu.co:8080/bitstream/10819/1079/1/Ministerio_de_Educacion_Resolucion_2343_junio_5_de_1996.pdf

Restrepo, J. E., Sánchez, O. A., & Castañeda Quirama, T. (2020). Estrés académico en estudiantes universitarios. *Revista Psicoespacios, 14*(24), 23-47. doi:https://doi.org/10.25057/21452776.1331

Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10*(29), 11-18. Recuperado el 20 de Mayo de 2022, de https://idus.us.es/handle/11441/43290

Rodríguez Espinar, S. (1985). Modelos de investigación sobre el rendimiento académico. Problemática y tendencias. *Revista Investigación Educativa, 3*(6), 284-303. Recuperado el 14 de Junio de 2022, de http://hdl.handle.net/10201/97141

Rodríguez Rosero, D. D., Ordoñez Ortega, R. E., & Hidalgo Villota, M. E. (2021). Determinantes del rendimiento académico de la educación media en el Departamento de Nariño, Colombia. *Lecturas de Economía*(94), 87-126. doi:https://doi.org/10.17533/udea.le.n94a341834

Ruta Maestra. (1 de Febrero de 2017). *DBA Derechos Básicos de Aprendizaje.* Obtenido de https://rutamaestra.santillana.com.co/dba-derechos-basicos-de-aprendizaje/

Sanabria James, L. A., Pérez Almagro, M. C., & Riascos Hinestroza, L. E. (2020). Pruebas de evaluación Saber y PISA en la Educación Obligatoria de Colombia. *Educatio siglo XXI, 38*(3), 231-254. doi:https://doi.org/10.6018/educatio.452891

Santamaría Ruiz, W. (Enero de 2006). *Técnicas de Minería de Datos Aplicadas en la Detección de Fraude:Estado del Arte.* Recuperado el 6 de Julio de 2022, de https://www.researchgate.net/publication/240724702_Tecnicas_de_Mineria_de_Datos_Aplicadas_en_la_Deteccion_de_FraudeEstado_del_Arte

Saravia, J. C. (10 de Marzo de 2015). *¡Pero qué linda relación tienen! La correlación de Pearson.* Obtenido de https://statssos.online/2015/03/10/pero-que-linda-relacion-tienen-la-correlacion-de-pearson/

Solano Luengo, L. O. (2015). *Rendimiento académico de los estudiantes de secundaria obligatoria y su relación con las aptitudes mentales y las actitudes ante el estudio.* España: Universidad Nacional de Educación a Distancia (UNED). Recuperado el 6 de Junio de 2022, de http://hdl.handle.net/11162/161183

The Academician. (1 de Mayo de 2020). *47. K Modes clustering using simple example with Python implementation.* Recuperado el 2 de Febrero de 2023, de https://youtu.be/x7pdjCxzo1A

Timarán Pereira, R., Caicedo Zambrano, J., & Hidalgo Troya, A. (Febrero de 2019). Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. *Revista de Investigación, Desarrollo e Innovación, 9*(2), 363-378. doi:https://doi.org/10.19053/20278306.v9.n2.2019.9184

Timarán Pereira, R., Hidalgo Troya, A., & Caicedo Zambrano, J. (2020). Factores asociados al desempeño académico en Lectura Crítica en las pruebas Saber 11o con árboles de decisión. *Investigación e Innovación en Ingenerías, 8*(3), 29-37. doi:https://doi.org/10.17081/ invinno.8.3.4701

Tobón Tobón, S., Pimienta Prieto, J. H., & García Fraile, J. A. (2010). *Secuencias didácticas: Aprendizaje y evaluación de competencias* (Primera ed.). México: PEARSON.

Tobón, S. (2013). *Formación integral y competencias. Pensamiento complejo, currículo, didáctica y evaluación* (Cuarta ed.). Bogotá, Colombía : ECOE.

ULA.VE. (s.f). *Capítulo 1. Técnicas de análisis de datos en WEKA.* Obtenido de http://webdelprofesor.ula.ve/economia/angelz/archivos/analisis_datos_con_weka.pdf

UNAYTA. (28 de Enero de 2019). *Data mining y Big data: Qué es y cómo aplicarlo en mi negocio*. Obtenido de https://unayta.es/data-mining-big-data/

Usman, M., & Atumoshi, A. Y. (2017). Educational Data Mining And Its Applications. *International Journal of Innovative Research and Advanced Studies (IJIRAS), 4*(1), 221-226. Recuperado el 7 de Julio de 2022, de https://www.ijiras.com/2017/Vol_4-Issue_1/paper_45.pdf

Vallejo Medina, P. (5 de Mayo de 2020). *Cómo hacer un Análisis Factorial Exploratorio en R. [Chupitos de R].* Obtenido de https://youtu.be/V0KOVwoU9gk

Vargas Agurto, W. (2014). *Mineria de datos y extracción del conocimiento.* Recuperado el 30 de Junio de 2022, de Académia.edu: https://www.academia.edu/9407662/Miner%C3%ADa_de_Datos

Vargas Cordero, Z. R. (2009). La investigación aplicada: una forma de conocer las realidades con evidencia científica. *Educación, 33*(1), 155-165.

Velásquez H., J. D. (13 de Octubre de 2021). *Coeficiente de la silueta — 4:24 min.* Obtenido de Cursos de analítica y machine learning: https://jdvelasq.github.io/courses/notebooks/sklearn_unsupervised_03_clustering/1-03_metodo_de_la_silueta.html

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. 1*, págs. 29-39. ResearchGate. Recuperado el 9 de Julio de 2022, de https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining

Zhexue Huang, J. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data mining and knowledge discovery, 2*(3), 283-304. doi:https://doi.org/10.1023/A:1009769707641