Técnicas de minería de datos para el análisis de pruebas SABER

Diana Paola Ahumada Riaño

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2023

Técnicas de minería de datos para el análisis de pruebas SABER

Diana Paola Ahumada Riaño

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magíster en Estadística

Director:
Ph.D. Juan Carlos Correa Morales

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística.
Medellín, Colombia
2023

Un buen maestro puede crear esperanza, encender la imaginación e inspirar amor por el aprendizaje.
Brad Henry

"Recuerda mirar arriba, a las estrellas, y no abajo, a tus pies. Intenta encontrar el sentido a lo que ves y pregúntate qué es lo que hace que el Universo exista. Sé curioso. Por muy difícil que te parezca la vida, siempre hay algo que puedes hacer y en lo que puedes tener éxito. Lo único que cuenta es no rendirse."
Stephen Hawking

Agradecimientos

Quiero expresar mi profundo agradecimiento al Dr. Juan Carlos Correa, mi director de tesis, por la valiosa orientación que me brindó en el desarrollo de este trabajo, que cuenta con un importante componente social y que refleja su compromiso con la disminución de las brechas académicas que enfrentan los jóvenes colombianos.

También quisiera agradecer a la Universidad Nacional de Colombia, sede Medellín, por ser una parte integral de mi proceso profesional y académico, y aportarme las herramientas y el ambiente propicio para lograr mis metas.

Mi reconocimiento y gratitud a todos los docentes y administrativos que, con su gran sentido humano y científico, contribuyeron a mi formación académica en cada una de las clases y espacios de la Universidad.

Finalmente, no puedo dejar de agradecer a mi mamá, quien ha sido el ser más hermoso en mi vida, por su inquebrantable fe en mí y su apoyo incondicional. Gracias a ella, he sido capaz de enfrentar y superar los desafíos de este camino académico.

Resumen

En los últimos años, el concepto de 'calidad en educación' ha adquirido relevancia, en particular, se le ha dado bastante atención a los resultados de los exámenes estandarizados, como las pruebas Saber 11 y las pruebas PISA, y, así mismo, se han utilizado como herramientas para evaluar una parte de la calidad educativa. Por esta razón, es necesario realizar análisis estadísticos que permitan tener una mejor comprensión de los factores que influyen en los resultados de este tipo de pruebas, los cuales pueden estar relacionados con las condiciones socioeconómicas de los estudiantes, la infraestructura escolar, entre otros aspectos. Es así que, en función de obtener una visión más precisa y completa de la situación educativa en el departamento de Antioquía. En este trabajo se analizarán las variables socioeconómicas proporcionadas en los datos del Icfes con el propósito de ofrecer un panorama de las posibles causas de los resultados de las pruebas Saber 11 en cada una de las subregiones que la componen. El objetivo de esta investigación es, entonces, realizar un diagnóstico del nivel educativo en el departamento de Antioquia contrastando el desempeño en las pruebas Saber 11 de cada una de las subregiones del departamento, el puntaje global y la información socioeconómica de los estudiantes que presentaron la prueba durante el periodo 2017-2019. Para alcanzar este objetivo, se llevó a cabo una revisión de la literatura sobre los métodos y modelos que se utilizan para analizar pruebas estandarizadas en educación. Consecuentemente, se decidió emplear el proceso de clúster con K-Means para clasificar a los grupos de estudiantes según sus características socioeconómicas, utilizando los programas RStudio y Python. Así pues, se clasificó la población en 10 grupos mediante el proceso de clusterización –teniendo en cuenta las variables socioeconómicas presentadas en las bases de datos– con el fin de describir algunos de los comportamientos de las variables según la subregión. Finalmente, se llevó a cabo una prueba de homogeneidad para examinar las variables que podrían influir en los resultados de las pruebas Saber 11 en el departamento de Antioquia, considerando las características socioeconómicas de los estudiantes.

Palabras claves: pruebas Saber 11, pruebas estandarizadas, calidad educativa, características socioeconómicas, nivel educativo, clúster..

Abstract

Data mining techniques for the SABER evidence analysis

The concept of 'quality in education' has gained greater relevance in recent years, placing a significant importance on the results of standardized exams such as the Saber 11 and PISA tests, using them as tools to evaluate a part of the educational quality. For this reason, it is necessary to conduct statistical analyses that allow a better understanding of the factors that influence the results of these types of standardized tests, which may be related to socioeconomic conditions, school infrastructure, among other aspects. In this way, a more complete and accurate view of the educational situation in the department of Antioquia can be obtained. In this work only the socioeconomic variables provided by the Icfes data will be analyzed to give an insight into the possible causes of the Saber test results in each of the subregions of Antioquia. The objective of this work is, then, to diagnose the educational level in the department of Antioquia through the performance in the Saber 11 tests of each of the department's subregions, contrasting the overall score and the socioeconomic information of the students who took the test during the period 2017-2019. To achieve this objective, a literature review on the methods and models that can be used to analyze standardized tests in education was carried out. Consequently the K-Means clustering process was decided to be employed to classify groups of students according to their socioeconomic characteristics, using RStudio and Python programs. Thus the population was classified into 10 groups through the clustering process, taking into account the socioeconomic variables presented in the databases, in order to describe some behaviors of the variables according to the subregion. Finally, a homogeneity test was conducted to examine the variables that could influence the results of the Saber 11 tests in the department of Antioquia, considering the socioeconomic characteristics.

Keywords: Saber 11, standardized tests, educational quality, socioeconomic characteristics, educational level, cluster

Contenido

Lista de Figuras

Lista de Tablas

1 Introducción

El Ministerio de Educación Nacional (MEN) de Colombia define las pruebas Saber como evaluaciones externas estandarizadas aplicadas por el Instituto Colombiano para la Evaluación de la Educación (Icfes). Estas pruebas buscan evaluar el desempeño alcanzado por los estudiantes en competencias básicas definidas por el MEN al final de los ciclos de los niveles educativos de la educación básica y media. Las pruebas Saber 11 tienen como objetivo monitorear la calidad de la educación en los establecimientos educativos del país y ofrecer información estratégica para el establecimiento de políticas educativas. Por tanto, es importante analizar los factores que pueden afectar los resultados de las pruebas Saber 11 en función de identificar las causas de las brechas académicas en Colombia.

Según Sánchez (2020), las pruebas estandarizadas son una herramienta útil para hacer seguimiento a la educación que están recibiendo los estudiantes y evaluar su capacidad para aplicar los conocimientos adquiridos en situaciones propias del contexto. Además, se argumenta que estas pruebas pueden ayudar a las instituciones educativas a mejorar su calidad, puesto que proporcionan información objetiva sobre el rendimiento de los estudiantes y permiten identificar áreas que necesitan ser reforzadas.

Báez (2020), por su parte, destaca la importancia de repensar las comparaciones basadas en el desempeño de los estudiantes en pruebas estandarizadas para evaluar la calidad de la educación. El autor argumenta que estas comparaciones pueden ser problemáticas debido a la gran desigualdad e inequidad existente entre los colegios públicos y privados en Colombia. En otras palabras, considera que el resultado de estas pruebas puede ser desigual a causa de las diferencias socioeconómicas y de recursos entre estos dos tipos de instituciones educativas.

En esta misma línea, Borg y Gall (1984), reconocen que las pruebas estandarizadas son herramientas valiosas para medir el aprendizaje y el rendimiento de los estudiantes en diferentes áreas del conocimiento. Sin embargo, también advierten de sus limitaciones, como el posible sesgo cultural de las preguntas y ejemplos utilizados: estos pueden estar basados en la cultura dominante y no reflejar la diversidad cultural de los estudiantes, afectando así la validez de las pruebas y dando lugar a resultados inexactos.

El trabajo de Ríos-Cuesta (2023), titulado "Desempeño histórico en la Prueba Saber de matemática: la necesidad de revisar la política educativa del Chocó", analiza los resultados del departamento en el área de matemáticas durante 2016-2021 con el propósito de mostrar la necesidad de revisar la política educativa. Los investigadores encontraron una brecha entre los estudiantes urbanos y rurales, que para 2021 era de 7 puntos. Así mismo, consideran importante el intercambio de experiencias entre los profesores que enseñan matemáticas y proponen gestionar espacios de formación y apoyo pedagógico.

El objetivo de este trabajo es, entonces, realizar un diagnóstico del nivel educativo básico del departamento de Antioquia mediante el análisis de los resultados de las pruebas Saber 11 (2017-2019) y la información socioeconómica de los estudiantes que presentaron la prueba durante este periodo. Para tal fin, se utilizó un proceso de clusterización que clasificó a los estudiantes en 10 grupos con características socioeconómicas similares; luego, se identificaron las variables que afectan el rendimiento académico en el departamento de Antioquia, discriminando algunas variables importantes para cada una de las subregiones.

El trabajo está estructurado en seis capítulos. En el primero se encuentra la introducción, en el segundo se presentan los antecedentes, los referentes teóricos, seguidos de una descripción de la prueba Saber 11. En el capítulo 3 se describe la información y los datos que se utilizaron en el análisis para el departamento de Antioquia, antes del proceso de clusterización. El capítulo 4 presenta un análisis exploratorio de los datos de las pruebas Saber 11 para el departamento de Antioquia durante los años 2017 a 2019. De igual manera, en él se describe el proceso utilizado para clasificar los grupos mediante el uso del método de clúster K-Means y se elabora un análisis de algunas variables según la subregión correspondiente. En el capítulo 5 se presentan las pruebas de homogeneidad para algunas variables utilizadas y consideradas relevantes. Finalmente, en el capítulo 6 se exponen las conclusiones y recomendaciones del estudio.

2 Marco Teórico

En este capítulo, se presenta el marco teórico que respalda la investigación. Se comienza con la exposición del planteamiento del problema, cuyo objetivo es identificar la relación entre variables específicas de interés. A continuación, se revisan los antecedentes relacionados con el tema de estudio, en los que se examinan investigaciones previas y hallazgos relevantes en el campo. Posteriormente, se proporciona una introducción a las pruebas estandarizadas y el papel desempeñado por el Icfes en la evaluación de los estudiantes. Además, se presenta una introducción al concepto de minería de datos, destacando su capacidad para encontrar patrones, tendencias y relaciones ocultas en conjuntos de datos, y se aborda específicamente la técnica de Clúster K-Means, que permite la agrupación de observaciones similares en clústeres o grupos. Asimismo, se explora la técnica estadística *Propensity Score Matching*, utilizada para equilibrar las características de dos grupos diferentes y reducir el sesgo de selección en estudios observacionales. Finalmente, se describe la metodología GSK *(Generalized Structural Equation Modeling)*, la cual proporciona un enfoque integral para analizar variables categóricas.

2.1. Planteamiento del Problema

La preocupación por mejorar los indicadores de calidad de educación en Colombia ha llevado a que, durante los últimos años, se desarrollen políticas educativas que pretenden aumentar la cobertura de la educación pública y, también, la calidad de la educación. Por ejemplo, se han entregado colegios públicos en concesión a empresarios privados para que administren los recursos públicos y contraten libremente, bajo sus condiciones, al personal docente. No obstante, revistas y periódicos nacionales han presentado informes relacionados con los mejores colegios del país según las pruebas Saber 11, desconociendo el trabajo de instituciones públicas y realizando comparaciones que ignoran sus factores económicos y sociales.

En este sentido, comparar a los colegios o estudiantes únicamente basándose en el puntaje global no es la mejor manera de evaluar su desempeño, ya que esto desconoce el impacto del entorno social, económico y sociocultural en su desarrollo académico. Por esta razón, en esta investigación se plantea la necesidad de implementar un proceso de clusterización que permita agrupar a los estudiantes con características similares, de manera tal que la comparación sea más justa y se considere el contexto en el que se desenvuelven.

Al comparar colegios públicos y privados, es fundamental tener en cuenta todas las variables que pueden afectar los resultados académicos. Existen numerosos factores sociales que pueden influir en el rendimiento de los estudiantes: la desigualdad económica y social, el entorno familiar y social en el que crecen, los hábitos académicos fomentados en el hogar, las oportunidades laborales de los padres, el tiempo que dedican a sus hijos y las actividades extracurriculares en las que participan, así como la nutrición que reciben durante sus primeros años de vida.

Considerar estas variables nos permiten acceder a una visión más completa y precisa de la realidad educativa, evitando generalizaciones injustas y facilitando la identificación de aspectos que requieren atención y mejora. En otras palabras, al medir la incidencia de estas variables, se podrá realizar una comparación adecuada y justa entre los diferentes colegios, promoviendo un análisis más profundo y enriquecedor de los resultados académicos. Aristizabal, Rosero, Bedoya, y cols. (2016) apoyan esta idea, su trabajo establece que las brechas sociales se originan principalmente por las diferencias en recursos escolares, así como por las divergencias en el estatus socioeconómico y cultural del hogar, y, en último lugar, por las diferencias en características individuales. Autores como Chica Gómez, Galvis Gutiérrez, y Ramírez Hassan (2011) consideran que, a medida que el nivel de escolaridad de los padres aumenta, la probabilidad de estar en un nivel bajo disminuye. De igual forma, Manrique y Carreño (2014) concluyen que el nivel educativo de los padres debe ser alto para influir significativamente en el rendimiento educativo de sus hijos. Estas investigaciones sugieren que una vía potencial para mejorar la calidad educativa consiste en la inversión en la educación superior pública y en la mejora de las oportunidades educativas y laborales; igualmente, demuestran que es fundamental reconocer el papel que desempeñan los padres en la formación de sus hijos y promover su participación activa en este proceso.

En síntesis, fortalecer la educación superior pública implica brindar recursos adecuados para garantizar una educación de calidad, así como generar oportunidades de acceso equitativo a la educación superior para todos los estudiantes. Además, es necesario crear un entorno propicio que fomente la participación de los padres en la educación de sus hijos, promoviendo una colaboración efectiva entre la escuela y la familia. Asimismo, mejorar las oportunidades educativas y laborales implica establecer políticas que fomenten la igualdad de oportunidades y la movilidad social. Esto puede incluir la creación de programas de becas y subsidios para estudiantes de bajos recursos, la promoción de prácticas educativas inclusivas y la colaboración entre el sector educativo y el sector empresarial en el desarrollo de habilidades relevantes para el mercado laboral.

Las entidades encargadas de tomar acciones en el ámbito educativo, por otra parte, necesitan datos, resultados y análisis estadísticos que respalden sus decisiones. En consecuencia, es necesario que más profesionales se interesen por este campo de estudio y contribuyan a los

procesos de mejora. Desde hace varios años se han realizado estudios cuantitativos en educación. Trabajos como el de Ramoni Perazzi, Orlandoni Merli, Pérez Pulido, y Aguilar Galvis (2016) analizaron los resultados de las pruebas Saber para el ingreso a la educación superior en la Universidad de Santander en el año 2016; esta indagación tenía como finalidad identificar estudiantes con alto nivel de riesgo de deserción. Desde otra perspectiva, trabajos como el de Bernal, Bernal, y cols. (2016), denominado "Brechas de género en el rendimiento escolar", muestran algunas evidencias de que en Colombia las mujeres tienen menos acceso que los hombres a la educación superior. El trabajo realizado por Rodríguez (2016), denominado "Algunos factores que Influyen en los resultados de las pruebas Estandarizadas y censales", en el cual llevó a cabo una revisión documental de los factores que afectan los resultados de las Pruebas Saber 11 en Matemáticas de estudiantes de grado Undécimo en los últimos cinco años. En su análisis, destacó la importancia de la influencia de los padres en el rendimiento académico de los hijos, así como el impacto de las condiciones socioeconómicas externas en dicho rendimiento, enfatizando la necesidad de profundizar en este aspecto y considerar las diferencias entre las instituciones educativas en términos de recursos y nivel de formación. Además, resaltó que el nivel socioeconómico de la familia tiene una influencia constante en el rendimiento académico, y que es común que los niños de bajos recursos continúen siéndolo en la edad adulta. Por último, mencionó que los factores institucionales de la escuela también desempeñan un papel importante en el rendimiento académico. El Icfes –en su interés de aportar en el mejoramiento de la calidad de la educación– ha abierto el programa Icfes de Investigación. Mediante este, ha puesto a disposición de los investigadores su base de datos con el objetivo de que se realicen investigaciones rigurosas que aporten información confiable y, también, que ayuden a orientar la toma de decisiones en políticas públicas que permitan mejorar la calidad de la educación. De igual forma, está interesado en desarrollar un vínculo entre investigación, práctica educativa y política pública mediante el desarrollo de una agenda de investigación propia y de una estrategia de divulgación de sus resultados obtenidos. En esta línea, las investigaciones de Pérez-Pulido, Aguilar-Galvis, Orlandoni-Merli, Ramoni-Perazzi, y cols. (2016), Bernal y cols. (2016), han concluido que los mejores colegios del país son privados y que a mayor estrato social corresponde un mayor puntaje en la prueba Saber 11. Este tipo de estudios han ayudado a evidenciar la brecha social en el sistema educativo y han demostrado que no es válido comparar la calidad educativa por estratos sociales, pues existen varios factores del entorno familiar y social que pueden afectar el aprendizaje de un estudiante (como lo son factores económicos, oportunidades de aprendizaje, alimentación, problemáticas familiares y sociales, entre otros). Por esta razón, en este trabajo se pretende identificar los factores que pueden influir en los resultados en las pruebas Saber 11 en cada una de las subregiones del departamento de Antioquia y plantear una nueva estratificación de los resultados en las pruebas Saber 11 teniendo en cuenta las características sociales y económicas de los estudiantes.

2.2. Antecedentes

Al realizar la revisión bibliográfica, encontramos varias investigaciones, relativamente recientes, que analizan las pruebas Saber 11. Una de estas es "Calidad, cobertura y costos ocultos de la educación secundaria pública y privada en Colombia" de Guarín, Medina, y Posso (2018). Este trabajo combina técnicas de evaluación de impacto y un análisis de costo-beneficio para comparar la calidad de la educación media pública con la privada. En este se observa que el crecimiento de la matrícula pública está desplazando a la oferta privada debido a las dificultades económicas de las familias; igualmente, evidencia una gran variación en los resultados de las pruebas Saber 11 entre los colegios públicos. Adicionalmente, este trabajo indaga sobre las diferencias institucionales entre la educación pública y la privada con respecto a la cobertura, la calidad, los beneficiarios y la distribución del logro académico por ciudades y estrato.

Para identificar cuál ha sido el papel de los sectores públicos en la calidad de la educación secundaria, los investigadores estimaron la diferencia de los resultados en el logro empleando las bases de datos de los resultados de las pruebas de Estado. Asimismo, se sumó el efecto de la educación pública al promedio de la privada, obteniendo lo que se denomina el retorno de la educación pública. Este indicador se presenta en términos de la desviación estándar de todos los puntajes del país y se usa para construir el ranking de las ciudades con los colegios públicos que exhiben los retornos más altos, relacionados con los resultados en las pruebas Saber 11. Con base en esto, se encontró que los mejores resultados en las pruebas Icfes se encontraban en la ciudad de Bucaramanga.

El trabajo realizado por López, Virgüez, Silva, y Sarmiento (2017), denominado "Desigualdad de oportunidades en el sistema de educación pública en Bogotá, Colombia", considera que la educación es una herramienta útil para nivelar oportunidades entre individuos, así como para mejorar su bienestar e ingresos. Este realiza un análisis comparativo sobre la desigualdad de oportunidades en los resultados de la prueba Saber 11 de 2012 entre dos modelos de educación pública en Bogotá: tradicional y por concesión. Los autores utilizaron la técnica *propensity score matching* para escoger el grupo de estudiantes de colegios públicos tradicionales a comparar con los de colegios en concesión. Para este trabajo se usó la metodología no paramétrica de Checchi y Peragine (2010) y la de Ferreira, Gignoux, y Aran (2011) para estimar el límite inferior de la porción de la desigualdad de oportunidades. Los resultados indican que los estudiantes de colegios concesionados presentan un menor grado de desigualdad de oportunidades en el logro educativo que los estudiantes de los colegios públicos tradicionales; asimismo, estos obtienen mejores resultados en español y matemáticas.

Ahora bien, en la actualidad las técnicas de minería de datos aplicadas a los datos generados en los ambientes educativos están demostrando ser herramientas eficaces para predecir el rendimiento académico de los estudiantes. Chiok (2017) en el artículo "Predicción del rendimiento académico aplicando técnicas de minería de datos" utilizó los registros académicos de la Oficina de Estudios de la Universidad Nacional Agraria La Molina (UNALM). La muestra utilizada fue de 914 estudiantes matriculados en los ciclos 2013 II y 2014 I en el curso de Estadística General. Sobre esta se aplicó varias técnicas de minería de datos con la finalidad de predecir la aprobación del curso. El Software utilizado fue WEKA A (Waikato Environment for Knowledge Analysis), desarrollado por la Universidad de Waikato de Nueva Zelanda. Este es un programa de uso libre y está compuesto por un conjunto de algoritmos que implementan la mayoría de las técnicas de minería de datos. Después de aplicar varias técnicas de minería, los autores concluyeron que las TMD demostraban ser herramientas eficaces para obtener modelos que permitieran predecir el resultado de los estudiantes matriculados en el curso de Estadística General; de igual forma, que la técnica de la red naive de Bayes resultó ser la de mayor precisión, al obtener un 71 % de correcta clasificación.

Castro, Ortiz, y Lemus (2016), por su parte, realizaron un trabajo cuyo objetivo fue la construcción de un índice que permitiera medir el nivel socioeconómico de los estudiantes que presentaron las pruebas estandarizadas Saber 11 del 2012. Para este estudio utilizaron las bases de datos del Icfes y el cuestionario sociodemográfico compuesto por 58 preguntas, las cuales caracterizan el contexto familiar, social, económico y cultural del estudiante. Ellos realizaron un análisis de componentes principales (ACP), resumiendo la máxima información posible en un solo componente que se denomina *índice socioeconómico* y utilizando la metodología planteada por el Departamento Nacional De Planeación. Flórez, Espinosa, Sánchez, y Angulo (2008) en esta misma línea, realizaron un análisis del desempeño académico de los estudiantes en dichas pruebas con el índice socioeconómico, en el que se evidenció una correlación alta entre el puntaje obtenido y el índice construido.

Castro Aristizabal, Diaz Rosero, y Tobar Bedoya (2016), en "Causas de las diferencias en desempeño escolar entre los colegios públicos y privados: Colombia en las pruebas Saber 11", identificaron las posibles causas que dan origen a las brechas de desempeño escolar entre los colegios públicos y privados en las cinco principales ciudades de Colombia. Para ello, se empleó la información de los resultados de las pruebas Saber 11 del año 2014 y se aplicó la descomposición de Oaxaca Blinder, combinada con estimaciones de la Función de Producción Educativa para corregir el sesgo de selección. Así concluyeron que, en el contexto nacional, se encuentran diferencias significativas en todas las áreas evaluadas a favor de los colegios privados.

Los autores citados anteriormente se enfocan en analizar los factores que afectan el resultado de las pruebas Saber 11. Con base en esto, en el presente trabajo se pretende hacer una invitación a las directivas de universidades públicas a dar una mirada crítica a los requisitos de admisión y plantear prioridades que beneficien a los jóvenes con más dificultades socioeconómicas.

En este orden de ideas, la Universidad Nacional de Colombia (sede Medellín) se preocupa por el éxito académico y el bienestar de sus estudiantes, y ofrece una variedad de apoyos. Entre estos se encuentran becas y ayudas económicas para aquellos con necesidades financieras, servicios de orientación académica para la planificación de la carrera y la toma de decisiones, servicios de bienestar estudiantil que incluyen atención médica, deportes, recreación, cultura y arte, entre otros. Además, la universidad cuenta con programas de tutoría y mentoría que buscan mejorar el desempeño académico, y programas de intercambio y movilidad estudiantil que enriquecen la formación académica y personal. Asimismo, los estudiantes pueden aplicar al programa de monitorías académicas y administrativas; una oportunidad para adquirir experiencia en docencia, investigación y gestión universitaria, al tiempo que apoyan a sus compañeros en áreas específicas. La universidad ha establecido, de igual forma, convenios municipales enfocados en estudiantes de pregrado con vulnerabilidad socioeconómica; aunque la institución no puede dar cobertura al 100 % de la población con necesidades, logra un aporte significativo.

Algunas universidades públicas, por otra parte, utilizan criterios de selección basados en el desempeño académico del estudiante en la Prueba Saber 11 y el puntaje obtenido en el examen de ingreso, lo que deja por fuera a estudiantes que no cumplen con estos requisitos y no tienen los recursos para ingresar a una universidad privada. Para abordar esta situación, se podría idear una reforma al proceso de admisión que permita que los estudiantes con necesidades socioeconómicas cuenten con la oportunidad y el apoyo necesarios para acceder a la educación superior pública y de calidad que ofrecen las instituciones. De esta manera, se estaría fomentando una sociedad más equitativa y justa, en la que la educación sea un derecho para todos, independientemente de su origen socioeconómico. Según Samper (2021), en su libro *La inteligencia y el talento se desarrollan*, para mejorar la calidad de la educación básica y media en Colombia se requiere la implementación de la jornada única en la mayoría de los colegios públicos, al igual que la dotación de los recursos necesarios para crear mejores condiciones de estudio para los estudiantes de estratos bajos y reducir la brecha social. Además, el autor destaca que las competencias interpretativas, deductivas o argumentativas no son innatas, sino que se pueden adquirir con el tiempo y la práctica, especialmente durante la juventud.

Por esta razón, en caso de que estos procesos no se den en el colegio a edades tempranas, es importante que las universidades asuman un compromiso con los jóvenes del país y les brinden espacios para superar sus debilidades y tener la posibilidad de un mejor futuro. Un ejemplo destacable, como ya se mencionó, es el programa de tutorías de la Universidad Nacional (sede Medellín), que ofrece a los estudiantes la oportunidad de reforzar las temáticas en las que tienen debilidades y en las que deseen fortalecer según sus intereses. Es necesario que más universidades adopten este tipo de iniciativas para garantizar que todos los jóvenes tengan acceso a una educación de calidad, independientemente de sus recursos y antecedentes académicos.

2.3. Pruebas estandarizadas

La preocupación por medir la calidad de la educación ha llevado a algunos países a crear pruebas estandarizadas para verificar si los objetivos planeados se están cumpliendo. Dichas pruebas son instrumentos de evaluación que miden las fortalezas y debilidades particulares de los estudiantes. Con el objetivo de profundizar en este tema, se realizó una consulta sobre algunas pruebas estandarizadas internacionales, dentro de estas se encontraron: el Programa para la Evaluación Internacional de Alumnos (PISA), el Estudio de las Tendencias en Matemáticas y Ciencias (TIMMS), la Evaluación de Competencias de Adultos (PIAAC), y el Tercer Estudio Regional Comparativo y Explicativo (TERCE). Por otro lado, algunos países cuentan con sus propias pruebas a nivel nacional: PLANEA en México, SAT en EE.UU., Sentâ Shiken en Japón, Ser Bachiller (previamente ENES) en Ecuador, ENEM en Brasil, el Abitur en Alemania, Saber 11 en Colombia, etc. Adicional a esto, algunas universidades a nivel mundial tienen su propio examen de admisión.

Con todo, es importante revisar el concepto de *'calidad de la educación'* ya que esta determinada por múltiples factores, y sería un error simplificarla únicamente en función de los resultados de una prueba estandarizada, sin tener en cuenta el contexto particular de cada grupo de estudiantes. Si bien las pruebas estandarizadas pueden proporcionar información importante sobre el desempeño académico, es fundamental considerar otros aspectos, como las condiciones socioeconómicas, el entorno familiar, las características del sistema educativo y las oportunidades de aprendizaje disponibles para cada grupo de estudiantes. Según Toranzos (1996), el concepto de *calidad* incluye varias dimensiones. La primera es la de entender la calidad educativa como eficacia. Torrecilla (2008), al respecto afirma que una educación de calidad es aquella que logra que los estudiantes realmente aprendan lo que se supone que deben aprender según lo establecido en los programas curriculares. La segunda dimensión se refiere lo que se aprende en el sistema y a su relevancia en términos individuales y sociales. De esta forma, una educación de calidad posee contenidos que responden a lo que el individuo necesita para desarrollarse como persona intelectual, afectiva, moral, físicamente y para

desempeñarse adecuadamente en la sociedad. La tercera, por último, remite a la calidad de los procesos y medios que el sistema brinda a los estudiantes para el desarrollo de su experiencia educativa.

En síntesis, la calidad educativa es aquella que ofrece a los estudiantes un adecuado contexto físico para el aprendizaje, un cuerpo docente adecuadamente preparado para la tarea de enseñar, buenos materiales de estudio y de trabajo, y estrategias didácticas adecuadas, entre otros. En adición, Toranzos (1996) afirma que las preocupaciones actuales consisten en determinar quiénes aprenden en las escuelas, qué aprenden y en qué condiciones aprenden. También, invita a reflexionar sobre los sectores más desfavorecidos de la sociedad y argumenta que las escuelas no deben reducirse a ser un espacio de asistencia social al que se recurre únicamente en busca de alimentación o atención sanitaria.

ICFES

En nuestro país, el Instituto Colombiano para la Evaluación de la Educación (Icfes), entidad vinculada al Ministerio de Educación Nacional, tiene 50 años de existencia. Esta institución ofrece servicios de evaluación de la educación en todos sus niveles, adelanta investigaciones sobre los factores que inciden en la calidad de la educación y brinda información que contribuye al mejoramiento y la toma de decisiones en la calidad educativa. El actual examen Saber 11 es una evaluación estandarizada realizada semestralmente por el Icfes que tiene como objetivo servir de criterio para la entrada de estudiantes a las Instituciones de Educación Superior, monitorear la calidad de la formación que ofrecen los establecimientos de educación media y producir información para la estimación del valor agregado de la educación superior.

Desde 1966, esta prueba comienza a transformarse con el fin de brindar a todos los bachilleres del país la oportunidad de presentar sus exámenes de admisión en las diferentes universidades de Colombia. En 1968 nace el Icfes, Instituto Colombiano para el Fomento de la Educación Superior (50 Años del Icfes, s.f.). Este mismo año se creó la primera versión del examen Saber 11 con el propósito de facilitar los procesos de admisión en las instituciones de educación superior. Sin embargo, fue en 1980 cuando se estableció como requisito formal para ingresar a las universidades, y sus resultados comenzaron a considerarse como indicadores de la calidad educativa impartida en los colegios, según lo establecido en el Decreto 2343 de 1980. A partir de entonces, el examen Saber 11 se ha consolidado como una herramienta fundamental en el ámbito educativo para evaluar y monitorear el desempeño de los estudiantes en Colombia. En 1993 los exámenes conocidos como el programa 'Saber' comenzaron a ser reconocidos en el ámbito educativo. En esa fecha, se implementó una muestra maestra de los planteles educativos que permitió realizar mediciones periódicas en distintos momentos del año y en áreas específicas, como matemáticas, lenguaje, ciencias sociales y ciencias naturales, para los grados 3º, 5º, 7º y 9º. Esta iniciativa proporcionó una herramienta inestimable para evaluar

el rendimiento académico de los estudiantes en Colombia y obtener información relevante sobre su progreso en diferentes áreas del conocimiento.

Desde el año 2000, el enfoque de la prueba se centró en la evaluación por competencias en lugar de evaluar únicamente conocimientos y aptitudes. Esta transición estuvo en concordancia con los lineamientos curriculares y los estándares básicos de competencias establecidos por el Ministerio de Educación Nacional (MEN, 2006). El cambio en la orientación de la prueba permitió una evaluación enfocada en las habilidades y capacidades de los estudiantes, tomando en cuenta no solo sus conocimientos teóricos, sino, también, su capacidad para aplicar esos conocimientos en situaciones reales y resolver problemas de manera efectiva. Con el objetivo de consolidar un Sistema Nacional de Evaluación Estandarizada (SNEE) que consiga la alineación de todos los exámenes que lo conforman, la estructura del examen Saber 11 fue modificada a partir del segundo semestre de 2014 con el propósito de que sus resultados fueran comparables, en términos de la evaluación de competencias genéricas, con los de otras pruebas del SNEE, como las pruebas Saber 3°, 5° y 9° y el examen Saber Pro. Esta alineación implica que los exámenes deben estar articulados en torno a la evaluación de unas mismas competencias en diferentes grados de desarrollo. Esto permite pasar de un sistema con mediciones aisladas a uno que hace un seguimiento sistemático de los resultados de la educación a través de diferentes niveles (Icfes, 2013). En el 2001 los exámenes ECAES empezaron a desarrollarse en tres áreas profesionales (Medicina, Derecho e Ingeniería Mecánica). En los últimos años se han realizado algunos cambios en las pruebas y en las formas de evaluar a los estudiantes.

Así pues, el examen Saber 11 es una evaluación estandarizada que mide el desarrollo de las competencias de los estudiantes que están por finalizar la educación media. Este examen se diligencia en lápiz y papel, y consta de preguntas cerradas en las pruebas de matemáticas, sociales, ciencias y lectura. Debido a la existencia de dos calendarios académicos en Colombia, este examen tiene dos aplicaciones en el año. Por lo general, en el primer semestre, los estudiantes de colegios de calendario B toman el examen, mientras que en el segundo semestre lo toman los estudiantes que pertenecen a colegios de calendario A. Aunque este es el funcionamiento típico en la aplicación del examen, en algunos casos es posible encontrar estudiantes que provienen de calendario A tomando el examen en el primer semestre del año.

Por otro lado, conformar el SNEE (Sistema Nacional de Evaluación Estandarizada) requirió la alineación del examen Saber 11 y un cambio en la estructura de este. La Tabla **2-1** muestra cómo ha cambiado el examen en distintos periodos de tiempo. Uno de los cambios más significativos es la evaluación de competencias genéricas, el cual implicó reestructurar varios elementos: en primer lugar, con la creación de la subprueba de competencias ciudadanas; en segundo lugar, al diferenciar en la prueba de matemáticas una subprueba de razonamiento cuantitativo; y, finalmente, mediante la fusión de pruebas bajo el criterio de las competencias

genéricas que evalúan en común –lenguaje y filosofía se fusionaron en la prueba de lectura crítica; física, química y biología se fusionaron en la prueba de ciencias naturales; y las competencias ciudadanas se evalúan ahora a través de una prueba de sociales y ciudadanas (Icfes, 2013)–. A partir de 2014, se implementaron cambios significativos en el diseño y la aplicación del examen para adaptarse a las necesidades y demandas del sistema educativo colombiano.

Para finalizar Piñero, Sánchez, Bernal, y Jerez (2019), sobre la incidencia de las TIC en el mejoramiento de las pruebas Saber. Para esto tomaron como base una función de producción frontera estocástica aplicada en diferentes estudios para medir el impacto de distintos factores académicos y emplearon un modelo explicativo que permitió analizar los factores determinantes que inciden en los resultados obtenidos en las pruebas Saber 11. Los resultados demostraron que la condición socioeconómica de los estudiantes tiene un gran impacto sobre los resultados de las pruebas.

2000-1 a 2005-2	2006-1 a 2014-1	2014-2 en adelante
Lenguaje Filosofía	Lenguaje Filosofía	Lectura crítica
Matemáticas	Matemáticas	Matemáticas Razonamiento cuantitativo
Física Química Biología	Física Química Biología	Ciencias naturales
Historia Geografía	Ciencias sociales	Sociales y ciudadanas
Inglés Francés Alemán	Inglés Francés Alemán	Inglés

Tabla 2-1: Cambios en la estructura ICFES

2.4. Minería de datos

La minería de datos es una técnica cuyo propósito es extraer información valiosa y útil para la toma de decisiones, a partir del análisis y descubrimiento de patrones, tendencias y relaciones en grandes conjuntos de datos Berry y Linoff (2004). En términos generales, la minería de datos implica la aplicación de algoritmos y técnicas estadísticas y computacionales avanzadas para explorar, analizar y modelar grandes conjuntos de datos. Algunas de las técnicas utilizadas incluyen la clasificación, la regresión, el clustering y la asociación. De igual forma, involucra herramientas como el análisis estadístico, la inteligencia artificial, el aprendizaje automático y la visualización de datos. Esta técnica tiene aplicaciones en diversos campos, como la medicina, el marketing, la banca, la seguridad informática, entre otros.

Por ejemplo, en la medicina, se puede utilizar para analizar grandes cantidades de datos de pacientes y encontrar patrones que permitan mejorar el diagnóstico y el tratamiento de enfermedades Koh (2005). El método de clusterización K-Means, en particular, es una técnica de minería de datos útil para analizar bases de datos. La ventaja principal de este método de clusterización es su capacidad para identificar grupos homogéneos de individuos o casos en función de sus características. Este método generalmente es usado para la segmentación de mercado, es decir, para la identificación de grupos homogéneos de consumidores que comparten características similares en cuanto a sus necesidades, preferencias y comportamiento de compra. De esta manera, esta técnica resulta conveniente para las empresas que buscan diseñar estrategias de marketing más efectivas para cada segmento de mercado.

En el contexto de la prueba Saber 11, el método de K-Means puede ayudar a identificar grupos de estudiantes que comparten características similares, como nivel socioeconómico, educativo, cultural, entre otras. Al agrupar a los estudiantes de manera homogénea, los analistas pueden identificar patrones comunes en sus respuestas a las preguntas de la prueba, lo que ayuda a entender mejor los factores que influyen en el rendimiento académico de los estudiantes. Esto se debe a que el método de clusterización K-Means posee varias ventajas para analizar bases de datos con variables categóricas y variables respuesta numérica como la prueba Saber 11. En suma, el método de K-Means permite la identificación de grupos homogéneos de estudiantes, la identificación de patrones comunes en sus respuestas a las preguntas de la prueba y es útil para la segmentación de mercado en el contexto de la educación.

2.5. Cluster K MEANS

El método K-Means –propuesto por MacQueen en 1967– es un algoritmo de agrupamiento ampliamente utilizado en el campo del aprendizaje automático y la minería de datos. El objetivo del K-Means es dividir un conjunto de datos en k grupos distintos, en los que cada grupo está representado por su centroide. El algoritmo selecciona inicialmente k centroides y luego asigna cada objeto al grupo cuyo centroide sea el más cercano. Los centroides se actualizan iterativamente hasta alcanzar una convergencia. El método se basa en minimizar la suma de las distancias al cuadrado entre los objetos y sus centroides, maximizando la similitud dentro de los grupos y minimizando la similitud entre los grupos. A lo largo de los años, el método K-Means ha sido mejorado y adaptado a diferentes desafíos y aplicaciones en áreas como la segmentación de clientes, análisis de imágenes y procesamiento de señales. Es una herramienta poderosa en el análisis de datos y el aprendizaje automático, con una larga historia y amplia aplicabilidad en la exploración de conjuntos de datos complejos.

Tomando como referencia los apuntes de MacQueen (1967), la idea básica de la agrupación de K-Means consiste en definir los conglomerados de tal forma que se minimice la

variación total intraconglomerado (conocida como variación total dentro del conglomerado). Existen varios algoritmos K-Means. Kassambara (2017) plantea que el algoritmo estándar es el Hartigan-Wong (1979), que define la variación total intracluster como la suma de las distancias euclidianas al cuadrado entre los elementos y el correspondiente centroide:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{2-1}$$

- x_i diseña un punto de datos perteneciente al clúster C_k
- μ_k es el valor medio de los puntos asignados al conglomerado C_k

Cada observación (x_i) se asigna a un conglomerado determinado tal que la suma de cuadrados (SS) de la observación a sus centros de conglomerados asignados μ_k sea un mínimo.

Definimos la variación total dentro del conglomerado de la siguiente manera:

$$\text{tot.withinss} = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{2-2}$$

La suma cuadrática total dentro del clúster mide la compacidad es decir, la bondad de agrupación y se busca que sea lo más pequeña posible.

2.6. K-means algoritmo

La descripción del proceso del algoritmo K-Means se basó en el trabajo de Giordani y cols. (2020), estos especifican los pasos para poder aplicar K-Means como se detalla a continuación. El primer paso para aplicar el clustering K-Means consiste en indicar el número de clústeres (k) que se generaran en la solución final. De esta manera, el algoritmo comienza seleccionando aleatoriamente (k) objetos del conjunto de datos, los cuales servirán como centros iniciales de los conglomerados. Los objetos seleccionados también se conocen como medios de clúster o centroides.

A continuación, cada uno de los objetos restantes se asigna a su centroide más cercano en términos de la distancia euclidiana entre el objeto y la media del conglomerado. Este paso se denomina como paso de asignación de clústeres. Cabe aclarar que para utilizar la distancia de correlación, los datos se introducen como puntuaciones z.

Tras el paso de asignación, el algoritmo calcula el nuevo valor medio de cada clúster. Este paso se denomina actualización del centroide del clúster. Con los centros recalculados, se comprueba de nuevo cada observación para determinar si un objeto puede estar más cerca de un clúster diferente. Así pues, todos los objetos se reasignan de nuevo utilizando el clúster actualizado.

Los pasos de asignación de clústeres y actualización de centroides se repiten iterativamente hasta que la asignación de conglomerados deje de cambiar, es decir, una vez que los conglomerados formados en la iteración actual sean los mismos que los obtenidos en la iteración anterior Kassambara (2017).

Los pasos que sigue el algoritmo K-Means pueden resumirse como sigue:

- Especificar el número de conglomerados (K) que deben crearse (por el analista).
- Seleccionar aleatoriamente k objetos del conjunto de datos como centros o medias iniciales de los clústeres.
- Asignar cada observación a su centroide más cercano, basándose en la distancia euclídea entre el objeto y el centroide.
- Para cada uno de los k conglomerados, actualizar el centroide del conglomerado calculando los nuevos valores medios de todos los puntos de datos del conglomerado. El centroide de un k-ésimo conglomerado es un vector de longitud p que contiene las medias de todas las variables para las observaciones del k-ésimo conglomerado; p es el número de variables.
- Iterativamente minimizar el total dentro de suma de cuadrados. Es decir, iterar los pasos 3 y 4 hasta que las asignaciones de clúster dejen de cambiar o se alcance el número máximo de iteraciones. Por defecto, el software R utiliza 10 iteraciones como valor máximo por defecto.

2.7. Propensity Score Matching

El *Propensity Score Matching* (PSM) es una técnica estadística utilizada para reducir el sesgo de selección en estudios observacionales en los que los individuos no son asignados al azar en los grupos de tratamiento y control. Esta técnica se basa en la estimación de la probabilidad de que un individuo pertenezca a uno u otro grupo a partir de sus características observadas. Posteriormente, según esta probabilidad, iguala los grupos de tratamiento y control. En otras palabras, el PSM busca emparejar a los individuos en el grupo de tratamiento con individuos en el grupo de control que sean similares en términos de sus características

observadas. En adición, el PSM es una herramienta versátil para la investigación en diversos campos, como la salud, la educación y la economía, entre otros. Por ejemplo, en el caso de la salud, se podría utilizar el PSM para evaluar el impacto de un programa de prevención de enfermedades en una población de pacientes, en la que estos no fueron asignados al azar al grupo de tratamiento o control.

Para implementar el PSM es necesario seguir varios pasos: la selección de variables relevantes, la estimación del *propensity score* y la creación de grupos equilibrados. Cabe aclarar que el éxito del PSM depende, en gran medida, de la elección del método adecuado para estimar el *propensity score*. Como señala Austin (2011), el PSM es una técnica poderosa y flexible que puede ayudar a reducir el sesgo de selección en estudios observacionales, siempre y cuando se utilice un método adecuado para estimar el *propensity score*. Existen varios métodos para estimarlo, incluyendo la regresión logística, la regresión probit y la técnica de matching sin reemplazo. La elección del método dependerá, por lo tanto, del conjunto de datos y de la pregunta de investigación.

Igualmente, es importante tener en cuenta que, aunque el PSM puede mejorar la validez interna de los estudios observacionales, no puede compensar la falta de validez externa o generalización a otras poblaciones Guo (2010). Por lo tanto, es importante considerar cuidadosamente la selección de la población de estudio y la interpretación de los resultados. A pesar de estas limitaciones, el PSM sigue siendo uno de los métodos más populares para reducir el sesgo de selección en estudios observacionales. Como sugiere Ho (2007),el PSM puede ser utilizado en combinación con otros métodos, como la regresión discontinua o el instrumental variable, para fortalecer aún más la evidencia causal. Al utilizar múltiples métodos, los investigadores pueden abordar múltiples fuentes de sesgo y aumentar la confianza en la evidencia causal.

En conclusión, el *propensity score matching* es una técnica versátil que puede ayudar a reducir el sesgo de selección en estudios observacionales. Sin embargo, su eficacia depende de la elección del método adecuado para estimarlo. En este sentido, si bien el PSM puede mejorar la validez interna de los estudios observacionales, no puede compensar la falta de validez externa. Con todo, el PSM es uno de los métodos más efectivos al momento de reducir el sesgo de selección en estudios observacionales y, de igual forma, puede ser utilizado en combinación con otros métodos para fortalecer aún más la evidencia causal.

2.8. Metodología GSK

La metodología GSK (Grizzle-Starmer-Koch), propuesta por Grizzle, Starmer y Koch en 1969, es ampliamente utilizada para el análisis de datos categóricos Grizzle (1969). En particular, se destaca por su flexibilidad y capacidad para abordar una amplia gama de situaciones

en las que las variables son de naturaleza categórica.

Los avances en la implementación de la metodología GSK a nivel computacional, la hace ampliamente accesible para su aplicación en diversos contextos de investigación. Adicionalmente, su enfoque se centra en analizar las relaciones y asociaciones entre variables categóricas, permitiendo examinar la homogeneidad de las distribuciones marginales y evaluar la influencia de diferentes factores en dichas distribuciones. Por esta razón, la metodología GSK posibilita la realización de pruebas de homogeneidad para determinar si existen diferencias significativas entre las distribuciones marginales de las variables categóricas Agresti (1990). Estas pruebas brindan información valiosa sobre la dependencia o independencia de las variables en estudio. Por último, el método GSK consta de varias etapas, que incluyen la definición de la variable dependiente, la definición del modelo y la estimación y validación del mismo.

En relación con las pruebas de hipótesis, se empleó la Prueba para Homogeneidad de las Distribuciones Marginales siguiendo las pautas presentadas en las notas de clase del profesor Juan Carlos Correa. Además, se recurrió a los textos de Monroy, Rivera, y Dávila (2018) y Pardo (2020) para reforzar los conceptos previos sobre esta temática; igualmente, para obtener información adicional sobre la metodología GSK, se consultaron diversas fuentes, como el artículo de Zuluaga y Morales (2018). Asimismo, se utilizaron recursos adicionales disponibles de manera libre en el repositorio de GitHub[1] con el fin de ampliar la comprensión de los conceptos relacionados.

En el análisis de datos categóricos, las pruebas de homogeneidad desempeñan un papel fundamental para determinar si existen diferencias significativas entre las distribuciones marginales de dos o más variables categóricas. Estas pruebas permiten evaluar si las variables están relacionadas entre sí o si son independientes. Entre los métodos más comunes utilizados en este tipo de análisis se encuentran la prueba de Chi-cuadrado, la prueba exacta de Fisher y la prueba de la razón de verosimilitudes. La prueba de Chi-cuadrado compara las frecuencias observadas en cada categoría con las frecuencias esperadas bajo la hipótesis nula de independencia; si se encuentran diferencias significativas, se concluye que existe una asociación entre las variables categóricas. En casos de muestras pequeñas o que no cumplen las condiciones para aplicar la prueba de Chi-cuadrado, se recurre a la prueba exacta de Fisher, que calcula la probabilidad de obtener una distribución de frecuencias igual o más extrema que la observada bajo la hipótesis nula. Por otro lado, la prueba de la razón de verosimilitudes compara el ajuste de dos modelos: uno que asume la independencia entre las variables y otro que permite la dependencia. Se evalúa la razón de verosimilitudes entre ambos modelos y se compara con una distribución de referencia para determinar si hay suficiente evidencia para rechazar la hipótesis de independencia. La elección de la prueba

[1]Disponible en: https://github.com/JoaquinAmatRodrigo/Estadistica

de homogeneidad dependerá del tamaño de la muestra, las características de los datos y las suposiciones que se puedan hacer. Es importante interpretar los resultados en el contexto específico del estudio y considerar otras variables relevantes para obtener conclusiones más precisas.

3 Descripción base de datos

En este capítulo, se lleva a cabo una descripción de los datos de las pruebas Saber 11 proporcionados por el Icfes, los cuales se utilizaron como base para el presente trabajo. Se mencionan las variables que se encuentran en las bases de datos y se aborda específicamente el cálculo de la medida de riqueza de los estudiantes por parte del Icfes, utilizando la información disponible sobre la situación socioeconómica familiar, se expone el método utilizado por el Icfes para determinar esta medida, destacando su importancia en el análisis de los resultados académicos. Además, se presenta una clasificación de los datos recopilados para este estudio, organizados según los diferentes niveles socioeconómicos (NSE) propuestos por el Icfes. Esta clasificación permitirá una comprensión más profunda de las relaciones entre la situación socioeconómica y el desempeño académico de los estudiantes.

3.1. Descripción de los datos

Como se mencionó con anterioridad, el Icfes pone a disposición las bases de datos a los interesados en realizar actividades de consulta o investigación. Para esto el investigador solicita un código de acceso a los datos; después de ser autorizado puede ingresar y descargar la información que necesite para la investigación o consulta que desarrolla. Se resalta que las bases de datos descargables cumplen con la Política de Protección de Datos Personales (Ley 1581); en concordancia con estos lineamientos, no se expondrá nombre o número de documento del estudiante y se protegerá su anonimato. Los datos de las pruebas Saber 11 se encuentran en la página de Acceso a Bases de Datos y Diccionarios del Icfes[1].

En las pruebas Saber 11, los estudiantes evaluados deben presentar una serie de exámenes y responder un cuestionario socioeconómico, que se compone de preguntas cortas de selección múltiple y no se califica. Este permite obtener información relevante de los estudiantes y ayuda a dar una posible explicación de los resultados del examen. Además, indaga por características del núcleo familiar, características del hogar, entre otros. La información recopilada en los cuestionarios tiene propósitos académicos, de investigación y de política pública, esta información es de carácter confidencial y no afecta los resultados de los evaluados.

[1] https://www2.icfes.gov.co/acceso-a-bases-de-datos-y-diccionarios

Según el Diccionario de Variables Saber 11 periodo 20142-20182 (2018), las variables se encuentran agrupadas por ocho categorías o campos, estos son:

- **Información personal:** tipo de documento, nacionalidad, género, fecha de nacimiento, periodo, código consecutivo, tipo de inscripción, país de residencia, etnia, grupo étnico, limitaciones físicas y cognitivas.

- **Información contacto:** departamento de residencia, código DANE departamento, municipio de residencia, código DANE municipio, valor pensión del colegio, veces que ha presentado el examen, entre otros.

- **Información socioeconómica:** contiene el estrato socioeconómico de la vivienda, número de hermanos, material de los pisos, de la vivienda, personas en el hogar, cuartos en el hogar, educación del padre, educación de la madre, ocupación del padre, ocupación de la madre, trabajo del padre, trabajo de la madre, nivel del Sisbén, tiene internet, tiene servicio tv, tienen teléfono, tiene computador, tiene lavadora, tiene horno microondas, tiene horno, tiene automóvil, tiene DVD, tiene nevera, tiene celular, tiene motocicleta, tiene consola de videojuegos, cuántos libros tienen en casa, ingresos familiares mensuales, come leche y derivados; come carne, pescado y huevo; come cereal frutos y legumbre; situación económica respecto al año inmediatamente anterior, tiempo dedicado a leer por entretenimiento, tiempo dedicado a internet para actividades académicas, horas de trabajo a la semana, entre otros.

- **Módulo de antecedentes escolares:** inquiere por información como cantidad de años en preescolar, así como información de reprobación de años, número de años estudiando en el colegio actual, entre otros. Este cuestionario fue aplicado a una muestra aleatoria equivalente a cerca del 10 % de la población evaluada en cada semestre.

- **Módulo de expectativas:** pregunta por la probabilidad de que el próximo año ingrese a un programa de educación superior, el puntaje esperado en cada una de las áreas evaluadas, y el salario esperado si realiza estudios técnicos y profesionales.

- **Información del colegio:** código Icfes colegio, código DANE, colegio, nombre establecimiento, colegio género, colegio naturaleza, colegio calendario, colegio bilingüe, carácter del establecimiento, jornada, municipio del colegio y departamento.

- **Datos citación del examen:** preguntas como ¿se encuentra privado de la libertad?; solicita el municipio de presentación y departamento de presentación.

- **Resultados** puntaje en lectura crítica, matemáticas, ciencias naturales, sociales y ciudadanas, razonamiento cuantitativo, inglés; puesto, deciles y percentiles, puntaje global, nivel socioeconómico del evaluado, así como el nivel socioeconómico del establecimiento.

Mientras que en el Diccionario de Variables Saber 11 periodo 2019-1 - 2019-2 (2019), las variables se encuentran agrupadas por seis categorías/campos; en este diccionario desaparece la información relacionada con el Módulo de Antecedentes Escolares y Módulo de Expectativas. En él se presentan las siguientes categorías o campos:

- **Información personal**
- **Información de contacto**
- **Información socioeconómica**
- **Información del colegio**
- **Datos de citación del examen**
- **Resultados del examen**

Se evidencia que algunas variables no son incluidas dentro de estas categorías. Por esta razón, solo se toman las variables en común para el presente análisis.

En esta investigación se trabajó con las bases de datos del Icfes, cuya información corresponde a las pruebas Saber 11 desde 2017-2 hasta 2019-2. Para el departamento de Antioquia, solo se tuvieron en cuenta los estudiantes que tenían diligenciadas todas las variables seleccionadas para este estudio. Se debe tener presente que el Icfes presenta el diccionario en el que relaciona las variables que se tendrán en cuenta en cada examen. Durante el periodo comprendido entre el 2014-2 al 2018-2 comparten el mismo diccionario; mientras que para el 2019-1 y el 2019-2 hay una modificación en el diccionario, por esta razón solo se analizaron las variables en común desde el año 2017 hasta 2019.

A continuación, se enuncian algunas consideraciones respecto a la base de datos. Para el análisis de datos se agregó una nueva variable llamada "*edad*", la cual se calculó de acuerdo

con la fecha de presentación del examen, es decir, qué edad tenían al momento de presentar dicha prueba. Con la introducción de esta variable se encontraron varias inconsistencias: en la edad de los estudiantes existían estudiantes con edades menores a 10 años, e incluso se presentaron estudiantes con edades de cero y algunos con edades superiores a 80 años. Por esta razón, se tomó la decisión de trabajar solo con los datos de aplicantes que se encontraban en un rango de edad de 15 a 25 años. Otra de las consideraciones fue que solo se tomaron los resultados de estudiantes colombianos con tarjeta de identidad o cédula de ciudadanía. Se descartaron pruebas de estudiantes con discapacidad, puesto que esto requeriría de un análisis diferente. También se descartaron los estudiantes que pertenecían a grupos étnicos minoritarios; esto a raíz de considerar que no se encuentran en las mismas condiciones.

En el contexto colombiano, es importante considerar que las condiciones educativas para los grupos étnicos no son equiparables a las de otros grupos, lo cual tiene implicaciones en el análisis de las pruebas Saber. Según el estudio de Sanchez (2011), los estudiantes pertenecientes a una etnia en Colombia muestran un rendimiento académico inferior en comparación con sus pares no étnicos, específicamente en las áreas de matemáticas y lenguaje. Esta brecha se mantiene de manera persistente a nivel departamental, especialmente en aquellos departamentos con una alta proporción de población étnica. El análisis realizado a partir de los resultados de la prueba Saber 11 descompone esta brecha en factores relacionados con características observables, como el entorno familiar y el colegio, así como factores no observables. Estos hallazgos evidencian la necesidad de considerar la diversidad étnica y las condiciones particulares de los estudiantes para lograr comparaciones equitativas en el análisis de las pruebas Saber y promover estrategias que reduzcan la brecha en el rendimiento académico entre los grupos étnicos y no étnicos en Colombia Sanchez (2011).

En la variable *jornada escolar*, solo se contemplaron los estudiantes de jornada completa, mañana y única; se descartaron aquellos de la jornada tarde, noche y fines de semana por considerar que estos estudiantes no se encuentran en las mismas condiciones, lo que imposibilita una comparación equitativa con aquellos de jornada completa, mañana y única. Con esto me refiero a un contexto similar en términos de oportunidades y recursos disponibles para el estudio y el rendimiento académico. Los estudiantes de jornada tarde, noche y fines de semana a menudo tienen responsabilidades adicionales, como trabajos a tiempo parcial o responsabilidades familiares, que pueden limitar su disponibilidad de tiempo para dedicarse plenamente a los estudios. Estas condiciones particulares pueden influir en su capacidad para prepararse adecuadamente para las pruebas y, por lo tanto, podrían generar una disparidad en los resultados en comparación con los estudiantes de jornada completa.

Por otro lado, al considerar la equidad en el análisis de las pruebas Saber, se debe asegurar que los grupos de estudiantes que se comparan sean lo más similares posible –en términos de contexto y condiciones– con el fin de obtener resultados más precisos y justos. En con-

secuencia, al excluir a los estudiantes de jornada tarde, noche y fines de semana, se busca evitar posibles sesgos en los resultados que podrían surgir debido a las diferencias en las circunstancias y el tiempo disponible para el estudio. El total de estudiantes seleccionados para este estudio fueron 109195 y la cantidad de variables usadas fue 37.

A continuación, se describen las variables de la base de datos con la que se realizó el trabajo, La Tabla **3-1**, da cuenta de la información personal de los estudiantes que se utilizaron para este análisis.

Información Personal y Contacto

Información de Personal		
Campo	**Descripción del campo**	**Opción de respuesta**
estu-tipodocumento	Tipo de Documento	TI – Tarjeta de identidad, CC- Cédula de ciudadanía
estu-genero	Género	F-Femenino M-Hombre
edad	(Fecha Presentación - fecha nacimiento)	Edades de 15 a 25 años
año	Año de presentación del examen	2017 2018 2019
Periodo	Periodo de presentacion del examen	1 2
estu-mcpio-reside	Municipio de Residencia	Texto

Tabla 3-1: Información de Personal

A continuacion se presenta la Tabla **3-2** que contiene información socioeconómica relevante. Entre los datos se encuentra, la educación de los padres, el número de personas y cuartos en el hogar, las posibilidades de alimentación, los electrodomésticos disponibles, la dedicación a la lectura, entre otros. La información ofrece una aproximación de la situación socioeconómica de las familias, lo que permite comprender mejor los desafíos y las limitaciones que pueden enfrentar los estudiantes en su proceso educativo. Esta información es crucial para desarrollar políticas y programas que ayuden a mejorar el acceso a la educación y las oportunidades de aprendizaje para todos los estudiantes.

Información Socioeconómica

Campo	Descripción del campo	Opción de respuesta
fami-educacionpadre	Nivel educativo alcanzado por el padre	Ninguno Primaria incompleta Primaria completa Secundaria (Bachillerato) incompleta Secundaria (Bachillerato) completa Técnica o tecnológica incompleta Tecnica o tecnologia completa Educación profesional incompleta Educación profesional completa Posgrado No aplica No sabe
fami-educacionmadre	Nivel educativo alcanzado por la madre	Ninguno Primaria incompleta Primaria completa Secundaria (Bachillerato) incompleta Secundaria (Bachillerato) completa Técnica o tecnológica incompleta Técnica o tecnológica completa Educación profesional incompleta Educación profesional completa Postgrado No Aplica No sabe
fami-personashogar	¿Cuántas personas conforman el hogar donde vive actualmente, incluido usted?	1 a 2 3 a 4 5 a 6 7 a 8 9 o más
fami-cuartoshogar	¿En cuántos cuartos duermen las personas de su hogar?	Uno Dos Tres Cuatro Cinco Seis o más

Continuación de la Tabla

Campo	**Descripción del campo**	**Opción de respuesta**
fami-tienecomputador	Computador	No, Sí
fami-tienelavadora	Máquina lavadora de ropa	No, Sí
fami-tienemicroondas	Horno microondas u horno eléctrico o gas	No, Sí
fami-tieneautomovil	Automóvil particular	No, Sí
fami-tienemotocicleta	Motocicleta	No, Sí
fami-tiene consolavideojuegos	Consola para juegos electrónicos	No, Sí
fami-numlibros	¿Cuántos libros físicos o electrónicos hay en su hogar? excluyendo periódicos, revistas, directorios telefónicos y libros del colegio?	0 a 10 libros 11 a 25 libros 26 a 100 libros más de 100 libros
fami-comelechederivados	¿Cuantas veces por semana se comen leche o derivados (queso, yogur, etc?.)	1 o 2 veces por semana 3 a 5 veces por semana Nunca o rara vez comemos eso Todos o casi todos los días
fami-come-carnepescadohuevo	¿Cuantas veces por semana se comen carne, pescados o huevos?	1 o 2 veces por semana 3 a 5 veces por semana Nunca o rara vez comemos eso Todos o casi todos los días
fami-comecereal frutoslegumbre	¿Cuantas veces por semana se comen cereales(avena, granola), frutos secos (almendras, maní) o legumbres (frijoles, garbanzos, lentejas)	1 o 2 veces por semana 3 a 5 veces por semana Nunca o rara vez comemos eso Todos o casi todos los días
fami-situación economica	Con respecto al año inmediatamente anterior, la situación económica de su hogar es:	Igual Mejor Peor

Continuación de la Tabla

Campo	Descripción del campo	Opción de respuesta
estu-dedicacionlecturadiaria	¿Cuánto tiempo al día dedica a leer por entretenimiento?	No leo por entretenimiento 30 minutos o menos Entre 30 y 60 minutos Entre 1 y 2 horas Más de 2 horas
estu-dedicacioninternet	¿Cuánto tiempo al día dedica a navegar en internet? Excluye actividades académicas	No Navega Internet 30 minutos o menos Entre 30 y 60 minutos Entre 1 y 3 horas Más de 3 horas
estu-horassemanatrabaja	¿Cuántas horas trabajó usted durante la semana pasada?	0 Menos de 10 horas Entre 11 y 20 horas Entre 21 y 30 horas Más de 30 horas
estu-tiporemuneracion	¿Usted recibe algún tipo de remuneración por trabajar?	No Sí en efectivo Sí en especie Sí en efectivo y especie

Tabla 3-2: Información Socioeconómica

La Tabla **3-3** presenta los campos relacionados con la información del colegio que se utilizaron en el análisis. Es importante destacar que se seleccionaron colegios de jornada completa, mañana y única, con el objetivo de garantizar que los estudiantes estuvieran en condiciones similares y así evitar posibles sesgos en los resultados.

Información Colegio

Campo	**Descripción del campo**	**Opción de respuesta**
cole-nombre-establecimiento	Nombre del Establecimiento	Texto
cole-genero	Indica el género de la población del Establecimiento.	Femenino Masculino Mixto
cole-naturaleza	Indica la naturaleza del Establecimiento	No Oficial Oficial
cole-calendario	Calendario académico del Establecimiento	A B Otro
cole-area-ubicacion	Área de ubicación de la Sede	Rural Urbano
cole-jornada	Jornada de la Sede	Completa Mañana Unica
cole-mcpio-ubicacion	Nombre del municipio donde está ubicado	Texto

Tabla 3-3: Información del Colegio

En la Tabla **3-4**, se encuentran los campos relacionados con el puntaje en lectura crítica, puntaje global y el índice socioeconómico del evaluado propuesto por el ICFES.

Resultados

Campo	Descripción del campo	Opción de respuesta
punt-lectura-critica	Puntaje en lectura crítica	Numérica
percentil-lectura-critica	Percentil lectura crítica	Numérica
punt-global	Puntaje total obtenido	Numérica
percentil-global	Percentil global en que se encuentra el evaluado	Numérica
estu-inse-individual	Índice Socioeconómico del evaluado	Numérica

Tabla 3-4: Información de Resultados

Una vez agrupados y realizado el proceso de clusterización (el cual se describe en el capítulo 4), se agregó una nueva columna a los datos llamada *subregión* en la que se clasificaron los municipios según las subregiones de Antioquia.

3.2. Medida de la riqueza de los estudiantes

El Icfes ha mostrado gran interés en recolectar información acerca del entorno de los estudiantes, tales como antecedentes escolares, competencias socioemocionales y características socioeconómicas. Para el Icfes, estas variables, conocidas como *factores asociados al aprendizaje*, son relevantes en el estudio de la calidad de la educación, ya que tienen influencia sobre el logro educativo.

Autores como Murillo y Carrillo-Luna (2021), Cruz y cols. (2014), Marqués (2016), encuentran en sus investigaciones que el nivel socioeconómico tiene una alta incidencia sobre el desempeño académico. Por esta misma línea, investigaciones como la realizada por Villacís Mejía (2020) encuentra evidencias que vinculan "el IMC, la Talla/Edad y el nivel socioeconómico con el rendimiento académico, poniendo énfasis en las consecuencias atribuidas al retardo en el crecimiento y los bajos niveles socioeconómico en el crecimiento, desarrollo físico, cognitivo e intelectual en la primera infancia". Resulta coherente, entonces, que el Icfes considere indispensable contar con un indicador que logre consolidar las dimensiones que componen el nivel socioeconómico de cada estudiante. Por este motivo, incluye la base de datos el Índice de Nivel Socioeconómico (INSE) en el contexto de las pruebas Saber, pues para el Icfes:

> Dentro de la población evaluada en cada prueba, es probable que existan grupos de estudiantes que compartan características comunes que no necesariamente se logren identificar a partir del estrato socioeconómico, ya que dicha variable busca hacer una clasificación según inmueble

> residencial y no necesariamente apunta al acceso que tiene un estudiante de bienes y servicios. ICFES (2019)

Según el boletín 4 de Saber al Detalle, ICFES (2019), desde el 2009 hasta 2012 el INSE era calculado con base en la metodología de análisis multivariante. A partir de 2012 y hasta la fecha, en las pruebas Saber 359, Saber 11, Saber Pro y Saber TyT se emplea la metodología de teoría respuesta al ítem (TRI) para dicho calculo. Para el Icfes, una gran ventaja de emplear "TRI respecto al análisis multivariante es que es posible construir una escala histórica que mantiene la escala de medición entre periodos, mientras que con el análisis multivariante la escala puede variar ligeramente. En ese sentido, se garantiza comparabilidad entre periodos" ICFES (2019).

La TRI, entonces, modela en términos probabilísticos la relación que existe entre las respuestas que un individuo le da a un conjunto de ítems y una variable que no se puede observar directamente, como el INSE. Con base en esto, este análisis "permite (1) estudiar las relaciones que hay entre variables, (2) verificar que las variables seleccionadas para la medición apunten a medir una misma dimensión, (3) observar el signo de la relación y (4) la comunalidad que tienen, ICFES (2019)[2].

El Icfes, en particular, utiliza la TRI para comparar la probabilidad que tiene un estudiante de cierto nivel socioeconómico de contestar alguno de los ítems en el periodo actual (grupo referencia), respecto a un estudiante de un periodo previo (grupo focal) que pertenece al mismo nivel socioeconómico frente al mismo ítem. El puntaje del INSE, por su lado, se emplea como insumo para el cálculo del nivel socioeconómico categórico (NSE). El NSE, a su vez, se usa para caracterizar la población que presenta las pruebas Saber y se genera como reporte tanto para los establecimientos educativos como para los evaluados. El NSE se enmarca en una escala de NSE 1 hasta NSE 4, en la que el primer nivel hace referencia a estudiantes pertenecientes a niveles socioeconómicos bajos y se incrementa hasta el cuarto nivel, que corresponde a estudiantes pertenecientes a niveles socioeconómicos altos. Esta categorización le permite tener en cuenta no solo el nivel de ingresos, sino, también, la posesión de bienes, acceso a servicios y educación del núcleo familiar, lo cual brinda una perspectiva del hogar del estudiante.

En la Tabla **3-5**, se presentan los puntos de corte del puntaje del INSE para definir qué nivel del NSE caracteriza a cada estudiante.

Nivel socioeconómico	Puntos corte INSE
NSE 1	0-41.109
NSE 2	hasta 51.176
NSE 3	hasta 64.080
NSE 4	hasta 100

Tabla 3-5: Puntos de corte del INSE por NSE. Fuente: ICFES, 2019

[2]Ver: `https://www.icfes.gov.co/documents/39286/2231027/Edicion+4+-+boletin+saber+al+detalle+.pdf/f9a33ad6-7559-99a5-5f7f-16d2f9b16f76?version=1.4&t=1678150151066/`

El Icfes emplea una metodología de árboles de decisión con el propósito de caracterizar a los grupos que clasifican a los estudiantes en uno de los cuatro grupos NSE; así determina la frecuencia de cada respuesta en las diferentes categorías de los ítems asociadas a un nivel de NSE. Luego, ordena las frecuencias por categoría de respuesta en orden descendente y se establecen la disposición de cada uno de los niveles NSE. En la Tabla **3-6** realizada por el ICFES se presentan las categorías de respuesta de mayor frecuencia asociadas a cada NSE. Se toman las cinco primeras categorías más frecuentes por cada nivel NSE.

Tabla de frecuencias - Descriptores Socioeconómicos					
Variable de caracterización		**Niveles NSE categórico**			
		1	**2**	**3**	**4**
Internet	(No)	x	x		
Computador	(No)	x	x		
Horno microondas o a Gas	(No)	x			
Lavadora	(No)	x			
Educación de la madre:	Primaria Incompleta	x			
Lavadora	(Sí)		x		
Computador	(Sí)		x		x
Servicio de televisión	(Sí)		x	x	
Educación de la madre:	Secundaria completa			x	
Automóvil particular	(No)			x	
Horno microondas o a gas	(Sí)			x	
Automóvil particular	(Sí)			x	x
Educación de la madre:	Profesional completa				x
Internet	(Sí)				x
consola de videojuegos	(Sí)				x

Tabla 3-6: Caracterización por NSE. Fuente: ICFES, 2019

De los descriptores anteriores, el Icfes presenta un ejemplo de caracterización:

> Los estudiantes que presentan Saber 11 pertenecientes al NSE1 suelen no acceder servicio de internet ni tener computador, horno microondas ni lavadora. Típicamente la educación de la madre es primaria incompleta. Para los NSE4, es característico tener computador, automóvil, servicio de internet y consola de videojuegos. Las madres de los estudiantes pertenecientes a NSE4 tienen educación profesional completa, ICFES (2019).

Con la finalidad de observar la relación entre el promedio del puntaje global en los colegios oficiales y no oficiales y el nivel socioeconómico categórico (NSE), se presenta la siguiente grafica.

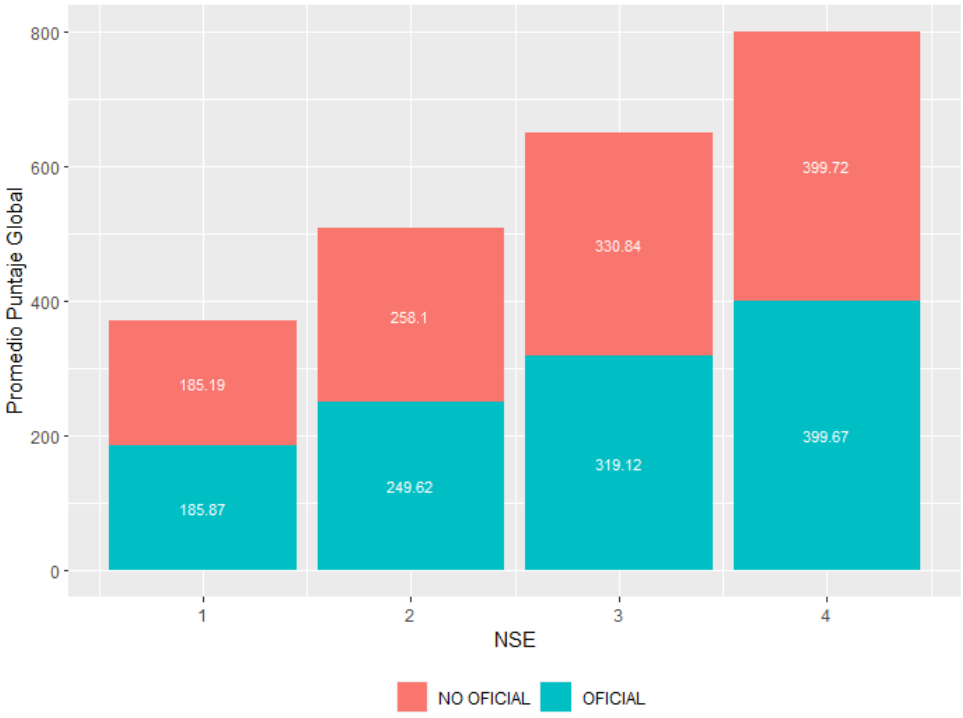

Figura 3-1: Promedio Puntaje Global en colegios oficiales y no oficiales - NSE (Elaboración propia).

En la Figura **3-1**, se puede observar que no existe una diferencia significativa entre el puntaje de las pruebas Saber en los colegios oficiales y no oficiales cuando se comparan en las mismas condiciones socioeconómicas y culturales. Es decir, un estudiante que presenta Saber 11 y se encuentra clasificado según sus condiciones socioeconómicas y culturales en NSE1, según la gráfica, el puntaje es mejor en los colegios oficiales que en los no oficiales, aunque no es muy grande. Para NSE2 y NSE3 se observan mejores puntajes en colegios no oficiales y en NSE4 –en el que, según el Icfes, se encuentran estudiantes pertenecientes a niveles socioeconómicos altos– se evidencia un mismo promedio en las pruebas Saber para colegios oficiales y no oficiales. Esto motiva a plantear que, aunque existen colegios privados (no oficiales) con excelentes condiciones de infraestructura y docentes altamente calificados, no se puede suponer que son mejores que los colegios oficiales, porque pueden influir las condiciones socioeconómicas y socioculturales en los resultados de las pruebas estandarizadas. En otras palabras, un estudiante de colegio privado con las mismas condiciones sociales, culturales, buena alimentación, padres con niveles superiores de estudio, condiciones de aprendizaje adecuadas, puede obtener resultados similares a los de un estudiante de una institución pública (esto se ve reflejado en la gráfica inicial). Por lo tanto, la comparación se debe realizar con estudiantes que se encuentren en las mismas condiciones socioeconómicas.

4 Análisis exploratorio de los datos

A continuación se presentan algunas variables de la base de datos, incluyendo la edad, los resultados en lectura crítica, puntaje global y variables categóricas. En el capítulo 5 de este trabajo, se llevarán a cabo las pruebas de homogeneidad de algunas de estas variables, con el fin de analizar posibles diferencias o asociaciones entre ellas.

Edad de los estudiantes

Se debe recordar que el rango de la edad para este trabajo fue de 15 a 25 años.

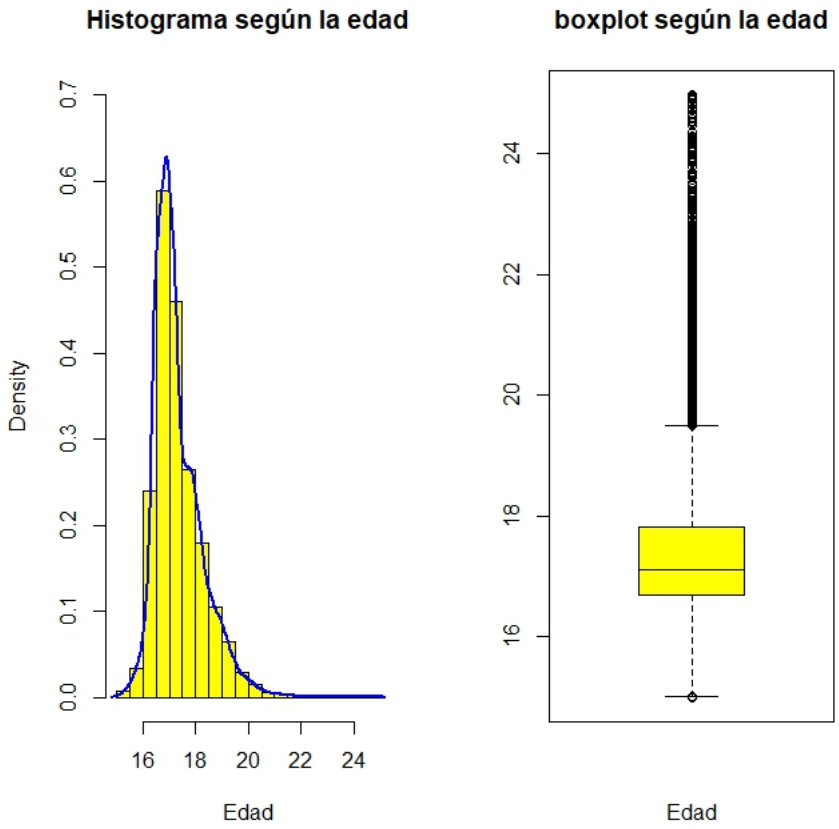

Figura 4-1: Descriptivo edad.(Elaboración propia)

En la Figura **4-1** se representa la distribución de los estudiantes según su edad. En esta, el promedio de la edad es de 17,34 años; la máxima edad que se ajusta a este conjunto es de 24,97 años y la desviación estándar es de 0,94 años. Se puede observar que los datos, aparentemente, presentan asimetría positiva, es

decir, que los mayores datos se alejan de la media. Para futuros análisis se podría disminuir el rango de la edad sabiendo que el 75 % de la población tiene menos de 17,82 años.

Puntaje en lectura crítica

A continuación, se presenta la distribución de los datos relacionados con el puntaje en lectura crítica.

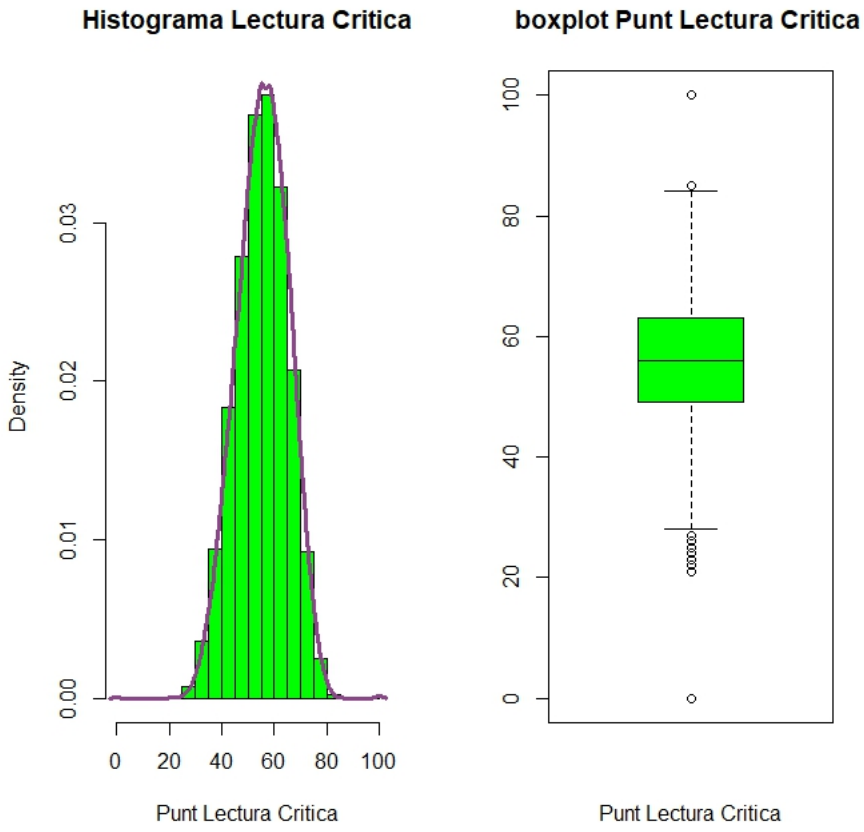

Figura 4-2: Descriptivo lectura crítica (Elaboración propia)

En la Figura **4-2** el promedio del puntaje en lectura crítica es de 55,63, el valor mínimo es cero y el máximo es 100 puntos, la desviación estándar es de 9,81. Los datos tienen un comportamiento aparentemente normal, el 75 % de los datos se encuentra con un puntaje menor de 63 puntos.

Puntaje global

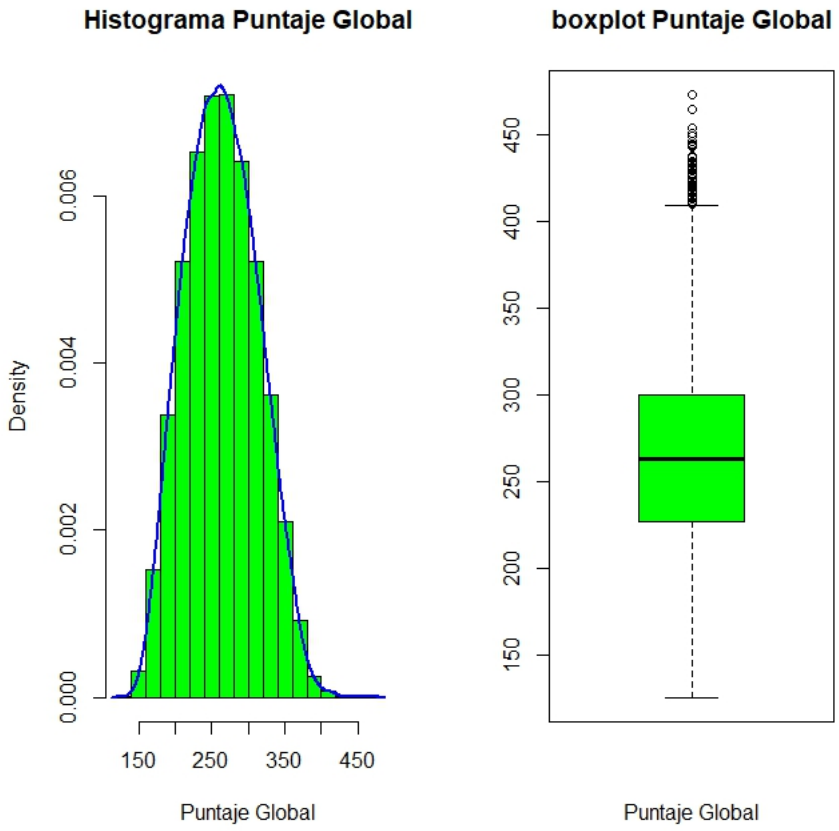

Figura 4-3: Descriptivo Global (Elaboración propia)

En la Figura **4-3** el promedio del puntaje global para Antioquia –con base en los datos usados (2017-2019)– es de 264,09 puntos y el 75 % de los datos tiene un puntaje menor a 300 puntos, presentando una desviación de 49,37, el puntaje mínimo es 126 y el máximo 473 puntos. El box-plot anterior nos muestra el puntaje global del Icfes e indica que los datos, en su mayoría, tienen una forma simétrica, además se observa que existen algunos puntajes por encima de 400 puntos.

Índice socioeconómico del ICFES

A continuación, se presentan las distribuciones para el INSE en sus 4 categorías para el puntaje global en los colegios oficiales y no oficiales.

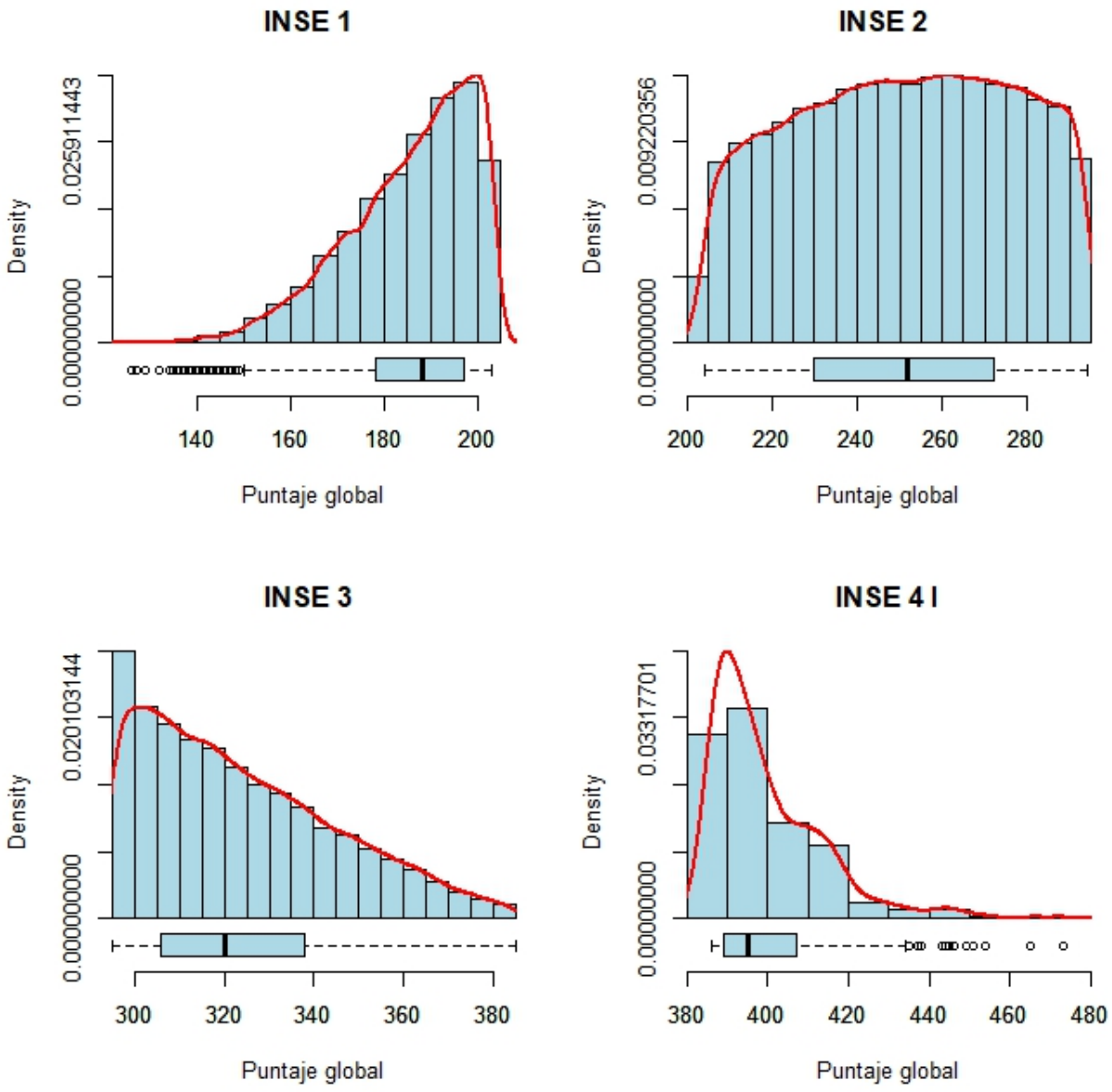

Figura 4-4: INSE (Elaboración propia)

En la Figura **4-4** se muestran los resultados del puntaje global para cada uno de los cuatro grupos INSE clasificados por el Icfes. El INSE 1 hace referencia a la población con nivel socioeconómico muy bajo. Este grupo presenta un promedio de 185,8 puntos, con un puntaje mínimo de 126 y un máximo de 203 puntos. El 75 % de los estudiantes en este grupo obtienen un puntaje menor a 197 puntos. Este resultado es preocupante si se considera el ingreso de estos estudiantes a la universidad, ya que no solo la parte económica será una barrera, sino que también se suma el componente académico.

El INSE 2 hace referencia a la población con nivel socioeconómico medio-bajo. Los resultados para el INSE 2 son mejores que los del INSE 1, con un promedio de 250,9 puntos, un puntaje mínimo de 204 y un máximo de 294 puntos. El 75 % de los estudiantes tienen un puntaje menor a 272 puntos. Sin embargo, los puntajes aún son muy bajos si se busca el ingreso a la universidad y la posibilidad de una beca académica.

A continuación, se presentan los resultados del grupo INSE 3, que se considera un grupo con nivel socioeconómico medio. El INSE 3 presenta un promedio de 324 puntos, con un puntaje mínimo de 295 y un máximo de 385 puntos. El 75 % de los estudiantes tienen un puntaje menor a 338 puntos. En este grupo se encuentra un porcentaje de estudiantes con posibilidades de ingreso a la educación pública, dependiendo de los criterios de admisión que cada universidad tenga.
El grupo INSE 4 se encuentra en una mejor condición socioeconómica, según el Icfes. Este grupo corresponde a estudiantes con condiciones socioeconómicas y culturales altas, según la clasificación del Icfes. Los estudiantes del grupo INSE 4 obtienen un promedio de 399,7 puntos en la prueba Saber, con un puntaje mínimo de 386 y un máximo de 473 puntos. Este grupo tiene la mayor posibilidad de ganar un apoyo tipo beca para continuar con sus estudios. Ahora es necesario reflexionar sobre la cantidad de estudiantes que se encuentran en cada grupo clasificado por el Icfes.

En la Figura **4-5** se presenta el porcentaje de estudiantes según el INSE propuesto por el Icfes para los colegios oficiales y no oficiales.

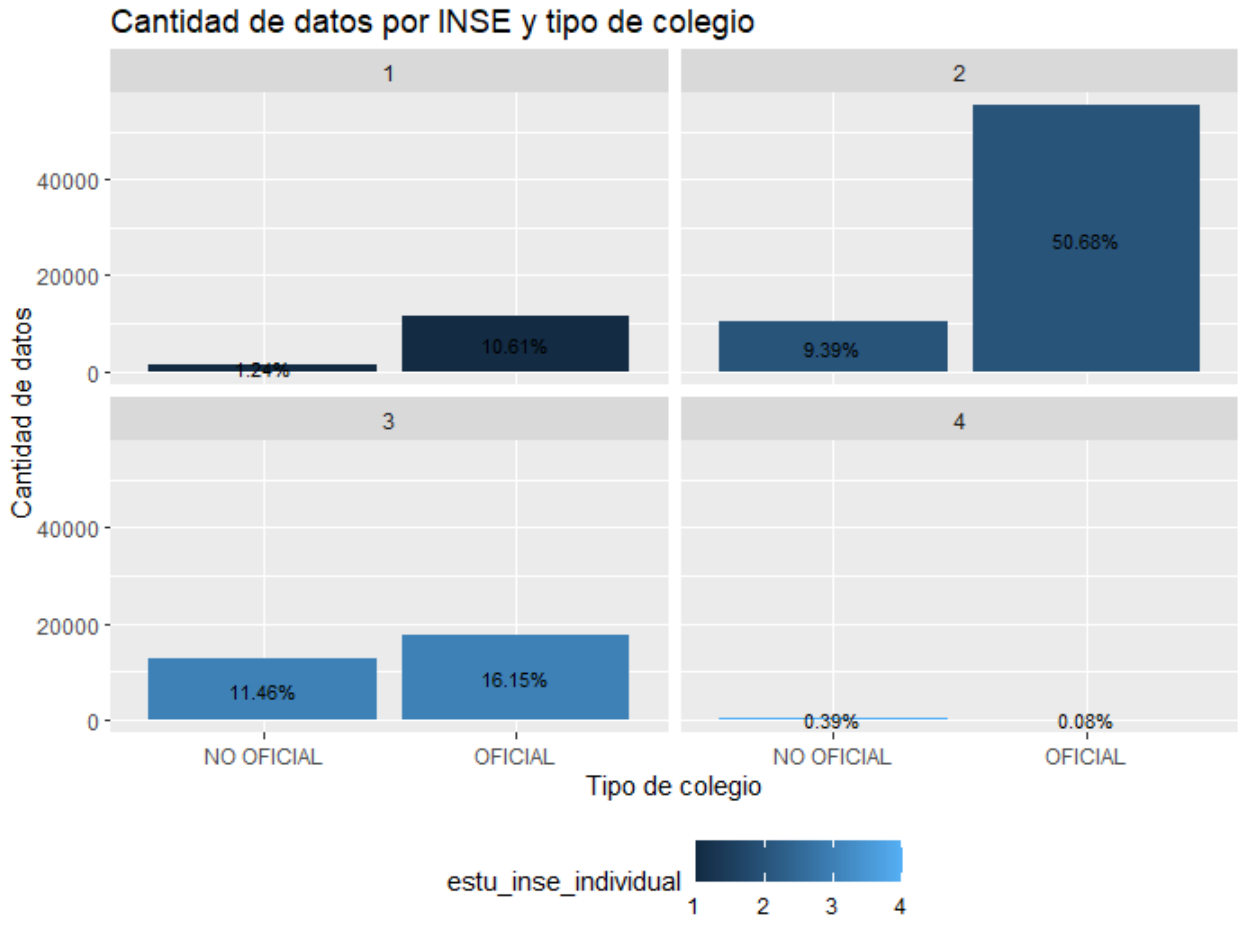

Figura 4-5: Porcentaje de estudiantes por INSE (Elaboración propia)

La Figura **4-5** refleja una realidad preocupante, relacionada con las grandes brechas sociales y académicas de Colombia, en este caso de Antioquia: el 71,92 % de los estudiantes se encuentra en nivel 1 y 2, el 27 % se encuentra en nivel 3 –considerado nivel socioeconómico medio– y solo el 0,47 % se encuentra en nivel 4, donde solo el 0,08 % de los estudiantes corresponden a colegios oficiales. Es decir, de 511 estudiantes en nivel 4, solo 87 estudiantes de colegios públicos se encuentran en este nivel; si se quiere ver a nivel general: solo 87 estudiantes de 109 195 que estudiaron en colegios públicos se encuentran clasificados en un nivel socioeconómico bueno, en el que se garantiza alimentación, posibilidad de computador, internet y transporte.

Dedicación a la lectura en Antioquia

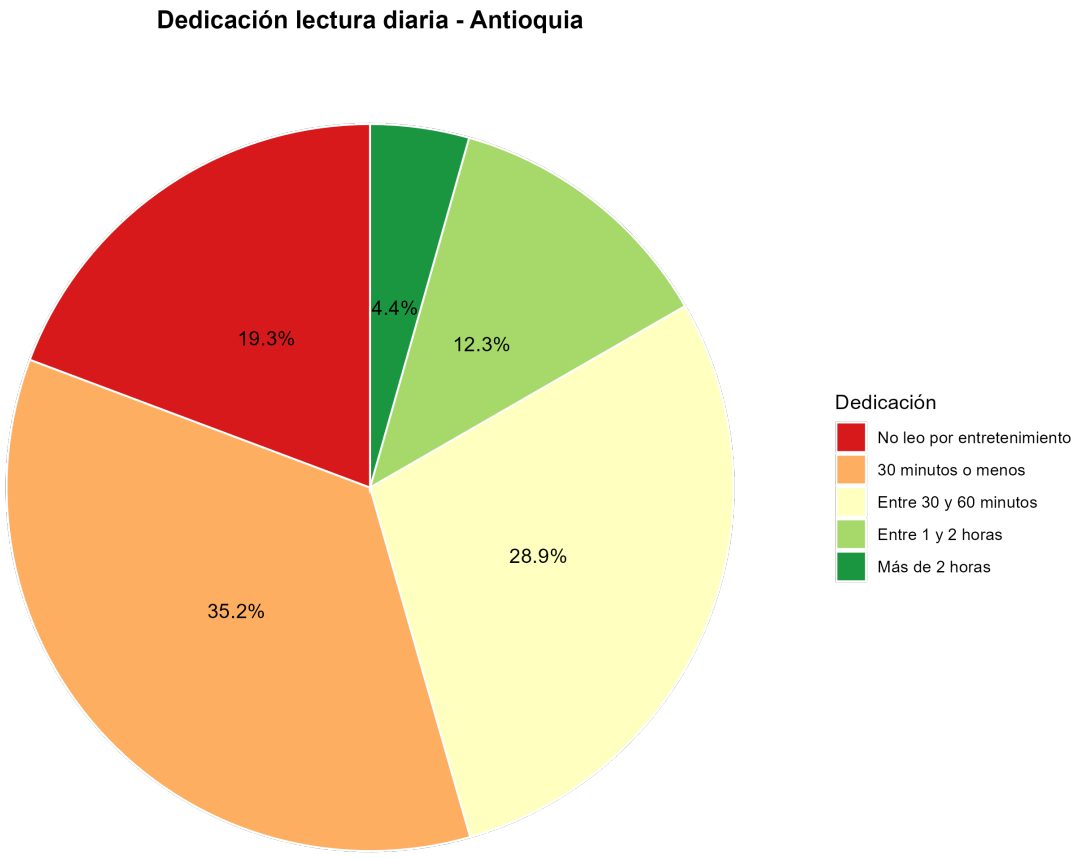

Figura 4-6: Dedicación lectura (Elaboración propia)

En la Figura **4-6** se puede evidenciar que el 35,18 % lee 30 minutos o menos, el 28,91 % lee entre 30 y 60 minutos, solo el 12,27 % lee entre 1 y 2 horas y, finalmente, el 4,38 % lee más de 2 horas. En este grupo de estudiantes, el 19,27 % manifiesta no leer por entretenimiento. Según estos datos, el 64,1 % de la población lee menos de 60 minutos al día.

En 2017, el DANE realizó una encuesta nacional de lectura (ENLEC 2017). Aunque no se registran datos para todo el departamento de Antioquia, se presentan los datos de la encuesta en Medellín. En cuanto a la lectura digital, el informe de la encuesta afirma que el 71,8 % de los encuestados manifiesta leer redes sociales (Facebook, Twitter, Instagram, WhatsApp, etc.), y solo el 19,1 % manifiesta leer libros digitales; en cuanto a la lectura impresa, el 48 % manifiesta leer periódicos y libros impresos. Estos datos demuestran que es necesario que desde los hogares y las escuelas se incentiven los hábitos de lectura crítica y de calidad.

Dedicación a la Lectura vs Puntaje Lectura en Antioquia

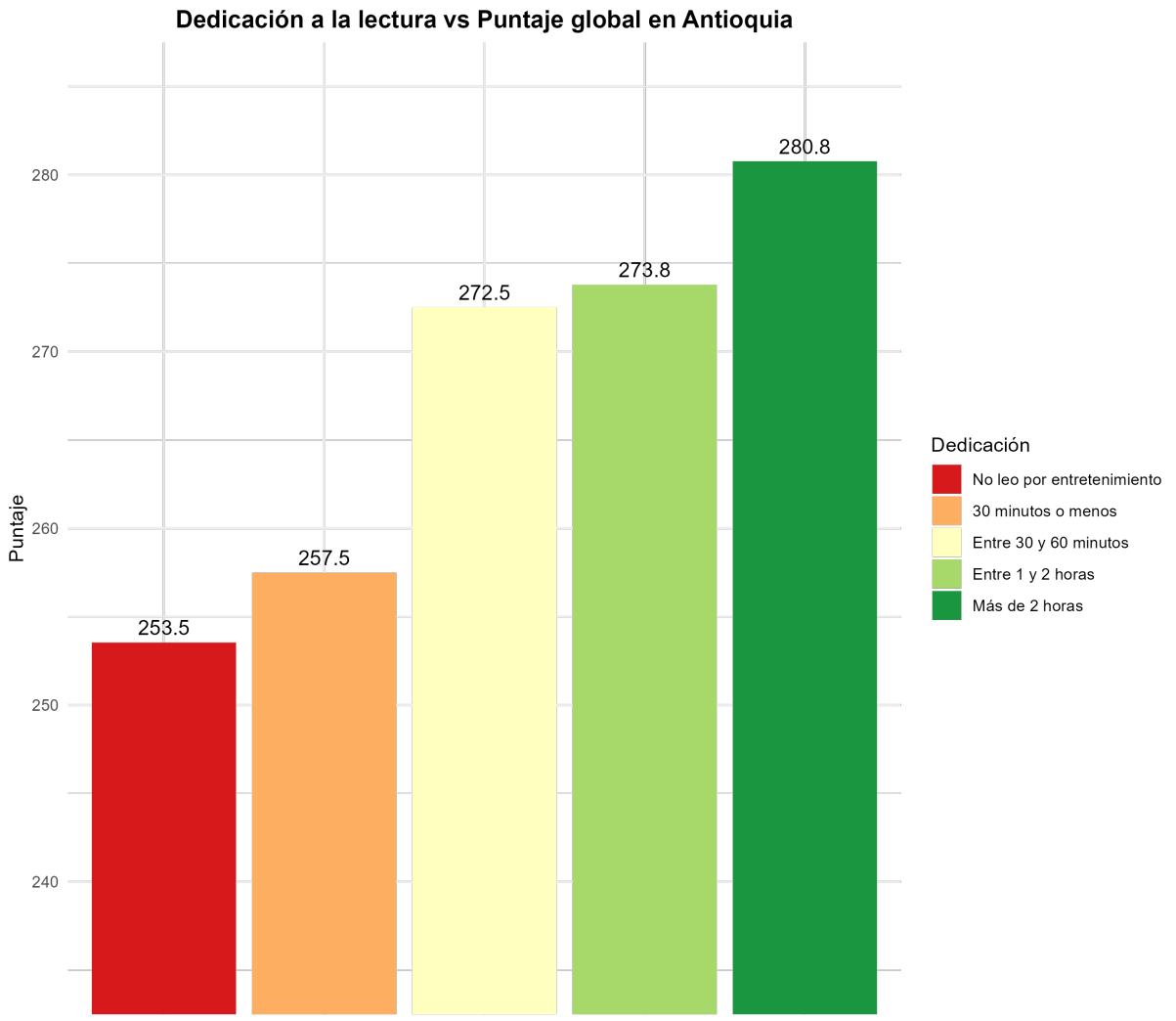

Figura 4-7: Dedicación a la lectura vs Puntaje Global

La Figura **4-7** muestra que los estudiantes del departamento de Antioquia que dedican más de dos horas diarias a la lectura obtienen los promedios más altos en el puntaje global. Esto puede indicar una relación entre el tiempo dedicado a la lectura y el rendimiento académico en el departamento de Antioquia.

Número de libros en la Familia vs Puntaje Lectura en Antioquia

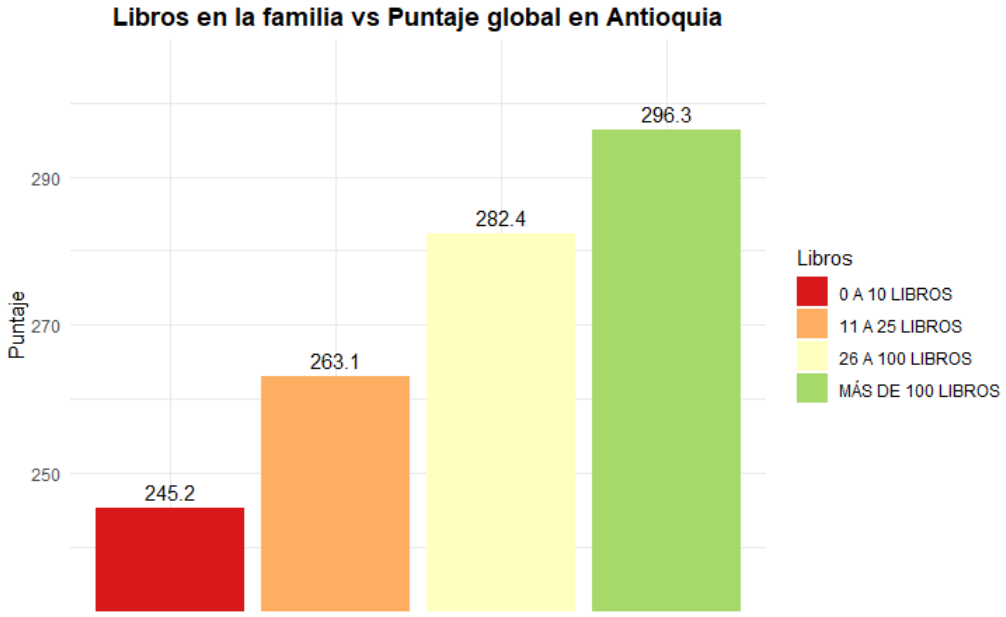

Figura 4-8: Número de libros en la familia vs Puntaje Global

La Figura **4-8** muestra una posible relación entre el número de libros en la familia y el puntaje global: a medida que aumenta el número de libros, el promedio de puntaje global también tiende a incrementarse. Los estudiantes que reportan tener más de 100 libros en su hogar obtienen el promedio más alto de puntaje global, mientras que aquellos que indican tener entre 0 y 10 libros registran el promedio más bajo, que es de 245,2. Estos resultados sugieren una posible influencia positiva de la disponibilidad de libros en el entorno familiar en el rendimiento académico de los estudiantes.

Familia tiene Moto vs Puntaje Global

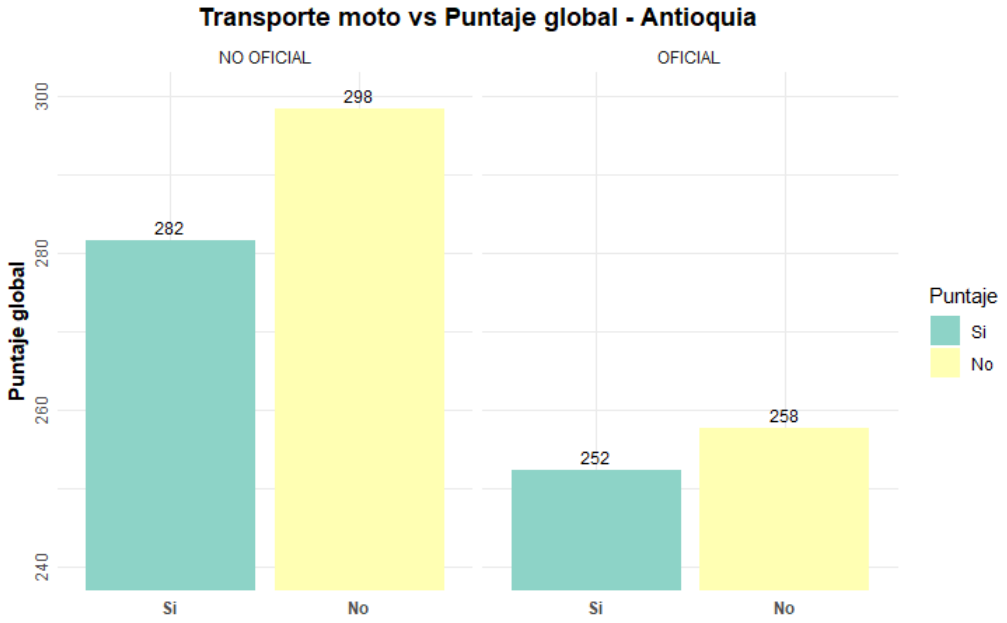

Figura 4-9: Familia tiene moto vs Puntaje Global en colegios oficiales y No oficiales (Elaboración propia)

En la Figura **4-9**se muestra el promedio en el puntaje global para estudiantes de colegios oficiales y no oficiales. Dentro de estos últimos, quienes que informaron no tener motocicleta en sus hogares obtuvieron mejores promedios en el puntaje global, con un promedio de 293 puntos, en comparación con los estudiantes que sí tienen motocicleta, que obtuvieron un promedio de 282 puntos. Mientras que, para los estudiantes de colegios oficiales que manifiestan tener motocicleta el promedio es de 252, los que no tienen motocicleta presentan un promedio de 258. Cabe señalar que el 63,1 % de la población estudiantil utilizada en este estudio su familia no cuenta con una motocicleta, mientras que el 36,87 % sí la posee.

Familia tiene carro

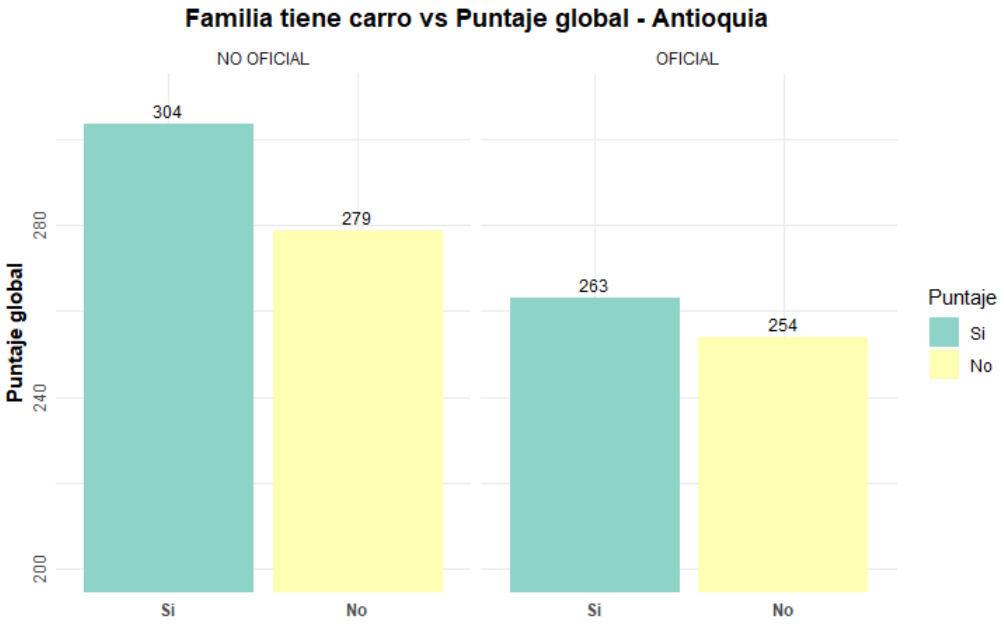

Figura 4-10: Familia tiene carro (Elaboración propia)

En la Figura **4-10**, se puede observar que los estudiantes que reportan no tener carro en su familia en colegios no oficiales obtienen menores promedios en el puntaje global, presentando un promedio de 279 para aquellos que no tienen carro y un promedio de 304 para aquellos que sí. Por otro lado, en colegios oficiales se reporta un promedio de 263 para los estudiantes que manifiestan tener carro y 254 para aquellos que no tienen carro. A partir de esto, se podría concluir que los estudiantes que pertenecen a familias con la posibilidad económica de tener carro obtienen mejores resultados y solo el 25.7 % manifiesta tiene carro; mientras que la mayoría manifiesta que no la tiene con un 74.3 %.
Se podría suponer que estudiantes que pertenecen a familias con posibilidades económicas de tener carro obtienen mejores resultados y solo el 25.7 % manifiesta tiene carro, siendo mayoría los que no tienen con un 74.3 %.

Estudiante dedicación en Internet

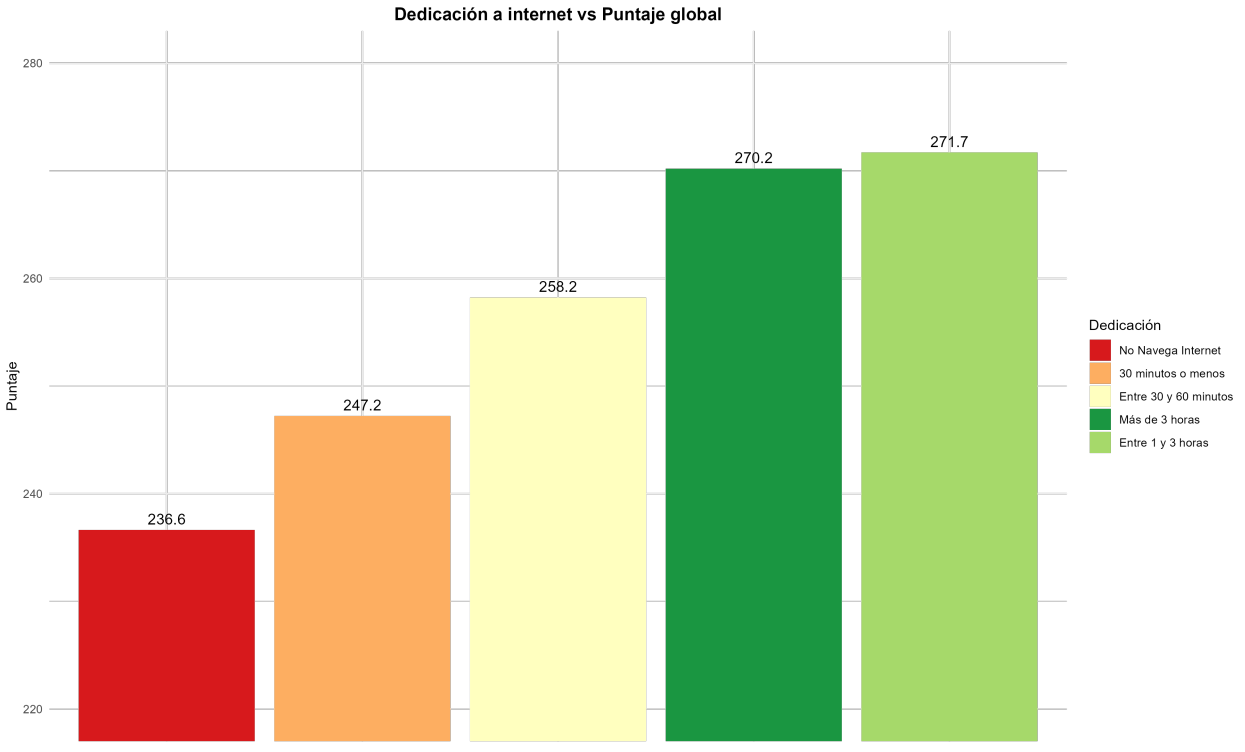

Figura 4-11: Estudiante dedicación en Internet (Elaboración propia)

La Figura **4-13**, observa que los estudiantes que navegan en Internet durante 1 a 3 horas al día obtienen mejores promedios en el puntaje global en comparación con aquellos que navegan por más de 3 horas. Aunque se podría argumentar que navegar en Internet podría influir en los resultados de la prueba Saber 11, se debe considerar que el tiempo que se dedica a esta actividad debe ser moderado para evitar distracciones y asegurarse de que se esté haciendo un uso efectivo del tiempo de estudio.

Familia tiene Computador

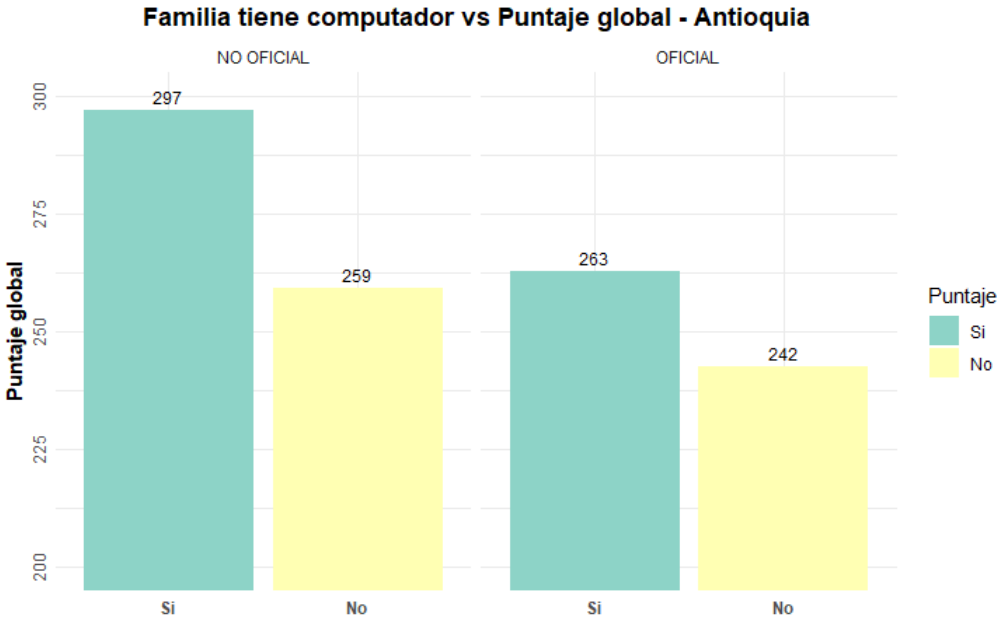

Figura 4-12: Familia tiene computador (Elaboración propia)

En la Figura **4-12**, se evidencia una diferencia en el promedio del puntaje global entre estudiantes de colegios no oficiales y colegios oficiales, dependiendo de si tienen acceso a un computador en sus hogares. Para aquellos estudiantes que tienen acceso a un computador, el promedio en el puntaje global es de 297 en colegios no oficiales y 263 en colegios oficiales. Por otro lado, para aquellos que no tienen acceso a un computador en sus hogares, el promedio en el puntaje global es de 259 en colegios no oficiales y 242 en colegios oficiales.
Es importante destacar que el 70,5 % de la población estudiantil manifiesta tener acceso a un computador en sus hogares. Estos hallazgos sugieren que el acceso a un computador puede tener una influencia en el rendimiento académico de los estudiantes. Asimismo, es necesario considerar estos resultados al diseñar políticas y estrategias educativas que promuevan la equidad y brinden oportunidades igualitarias para todos los estudiantes, independientemente de su acceso a la tecnología.

Educación de la Madre

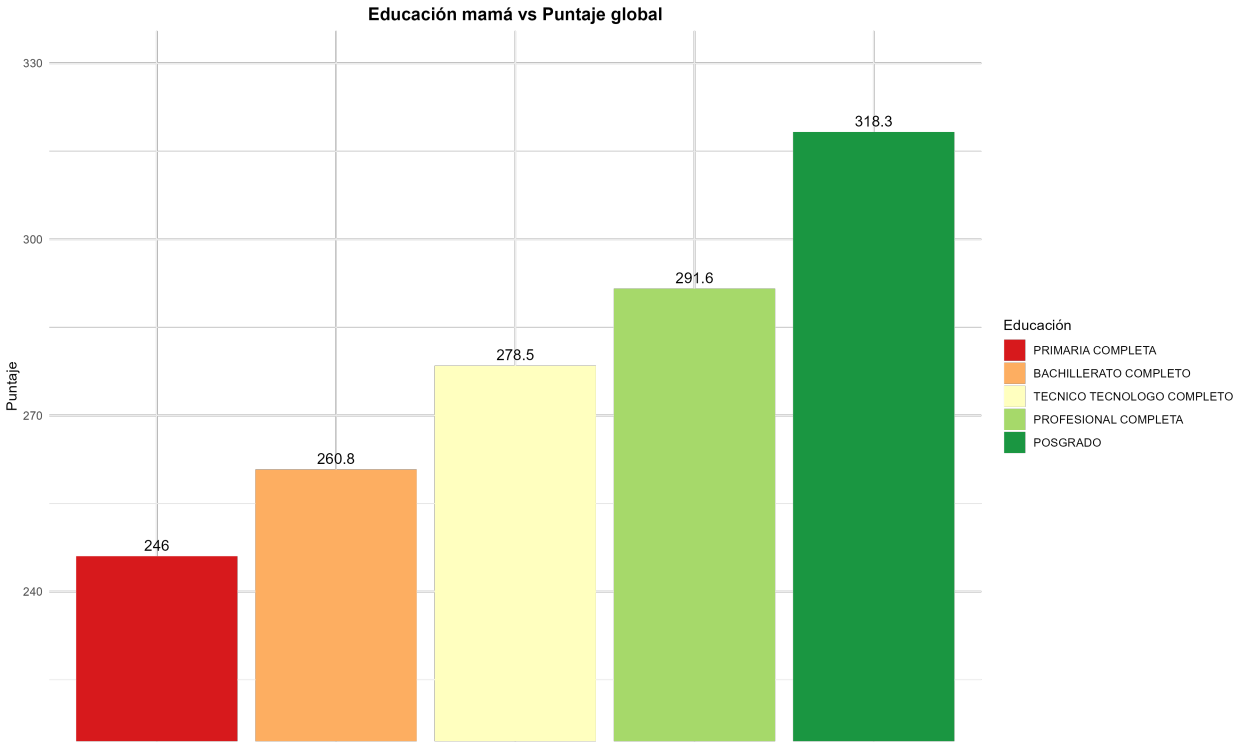

Figura 4-13: Educacción de la Madre (Elaboración propia)

En la Figura **4-13**, se puede evidenciar que a mayor nivel educativo de la madre mejor es el resultado en las pruebas Saber 11. Los estudiantes que tienen madres con posgrado presentan un promedio en el puntaje global de 318,3 puntos. Según los datos de 109195 solo 3531 madres tienen posgrado y de esas solo 843 tienen estudiando a sus hijos en colegios Oficiales.

Come leche, derivados vs Promedio en el Puntaje Global

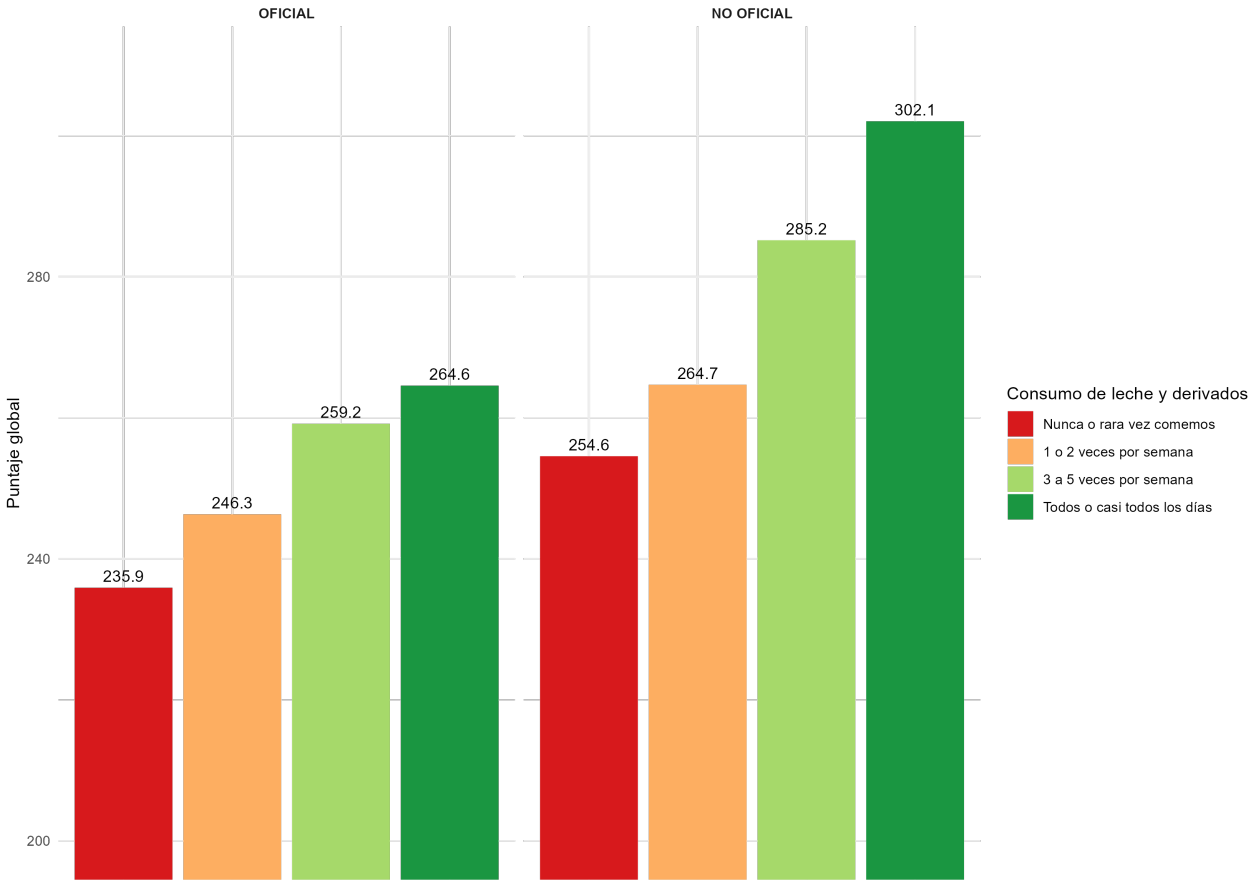

Figura 4-14: Estudiante Come leche, derivados (Elaboración propia)

En la Figura **4-14**, , se evidencia que el grupo de estudiantes que consumen leche y derivados lácteos diariamente obtiene el puntaje promedio más alto, tanto en colegios oficiales como no oficiales. En contraste, se observa que el grupo de estudiantes que rara vez o nunca consumen estos productos presenta los puntajes más bajos en la prueba Saber 11.

Estos hallazgos sugieren una posible relación entre el consumo de leche y derivados lácteos y el rendimiento académico de los estudiantes en las pruebas Saber 11. Sin embargo, es importante tener en cuenta que existen diversos factores que pueden influir en el desempeño académico, y el consumo de lácteos puede ser solo uno de ellos. Por otro lado, estos resultados resaltan la importancia de promover una alimentación balanceada y nutritiva, incluyendo el consumo regular de leche y derivados lácteos, como parte de las estrategias para mejorar el rendimiento académico de los estudiantes.

Come Carne, Pescado vs Promedio en el Puntaje Global

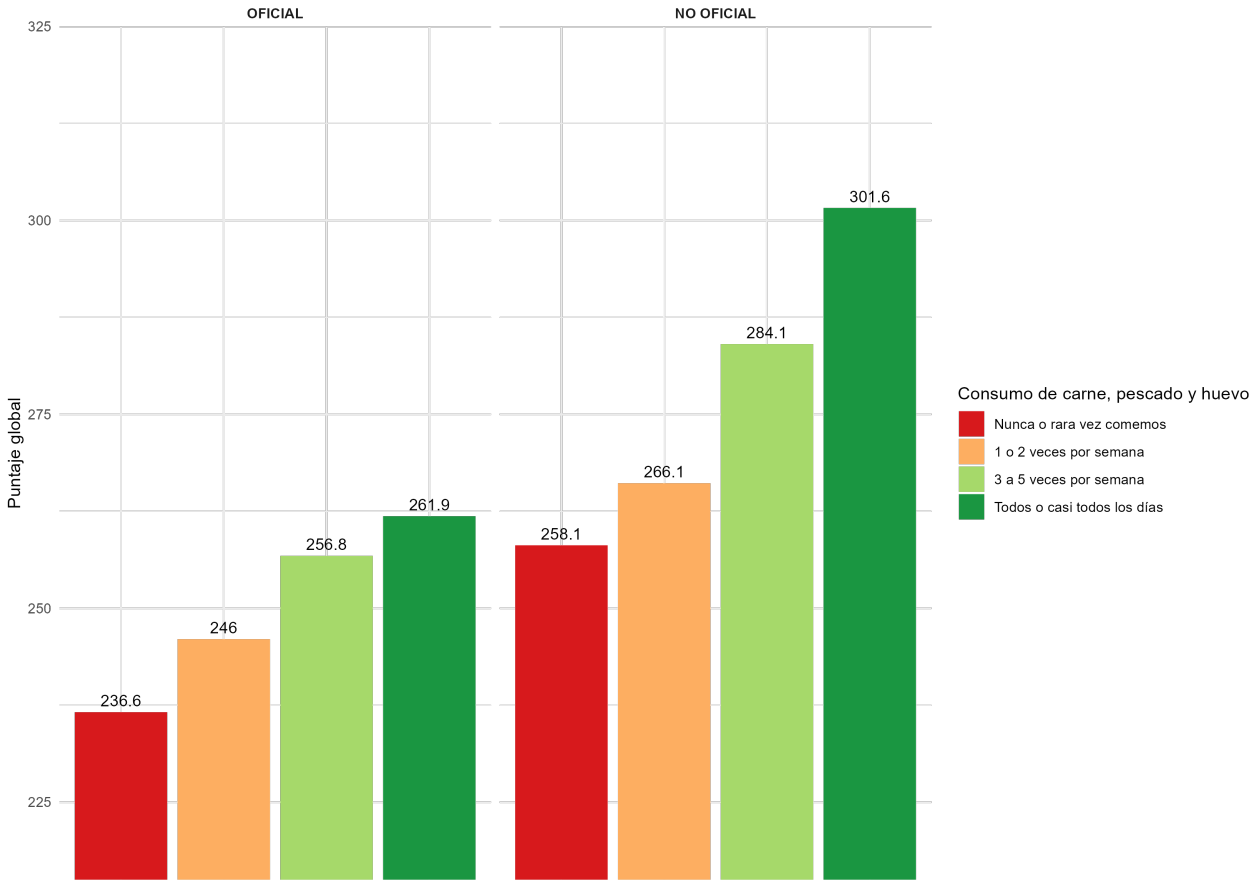

Figura 4-15: Estudiante come Carne, Pescado (Elaboración propia)

En la Figura **4-15**, se presentan los resultados relacionados con la frecuencia de consumo de carne, pescado y huevos, y se observa una clara asociación entre dicho consumo y el puntaje global obtenido en las pruebas Saber 11. Los estudiantes que indican consumir estos alimentos a diario presentan los puntajes más altos, mientras que aquellos que rara vez o nunca los consumen obtienen los puntajes más bajos. Cabe destacar que esta relación se aprecia tanto en colegios oficiales como no oficiales. Además, la figura revela que, en los colegios no oficiales, los estudiantes que reportan un consumo diario de proteínas alcanzan los mejores puntajes, con un promedio de 301,6, mientras que en los colegios oficiales el promedio es de 261,9. Esto sugiere que puede existir otras variables que pueden influir en el resultado del puntaje global.

Come Cereal, futos y legumbres vs Promedio puntaje Global

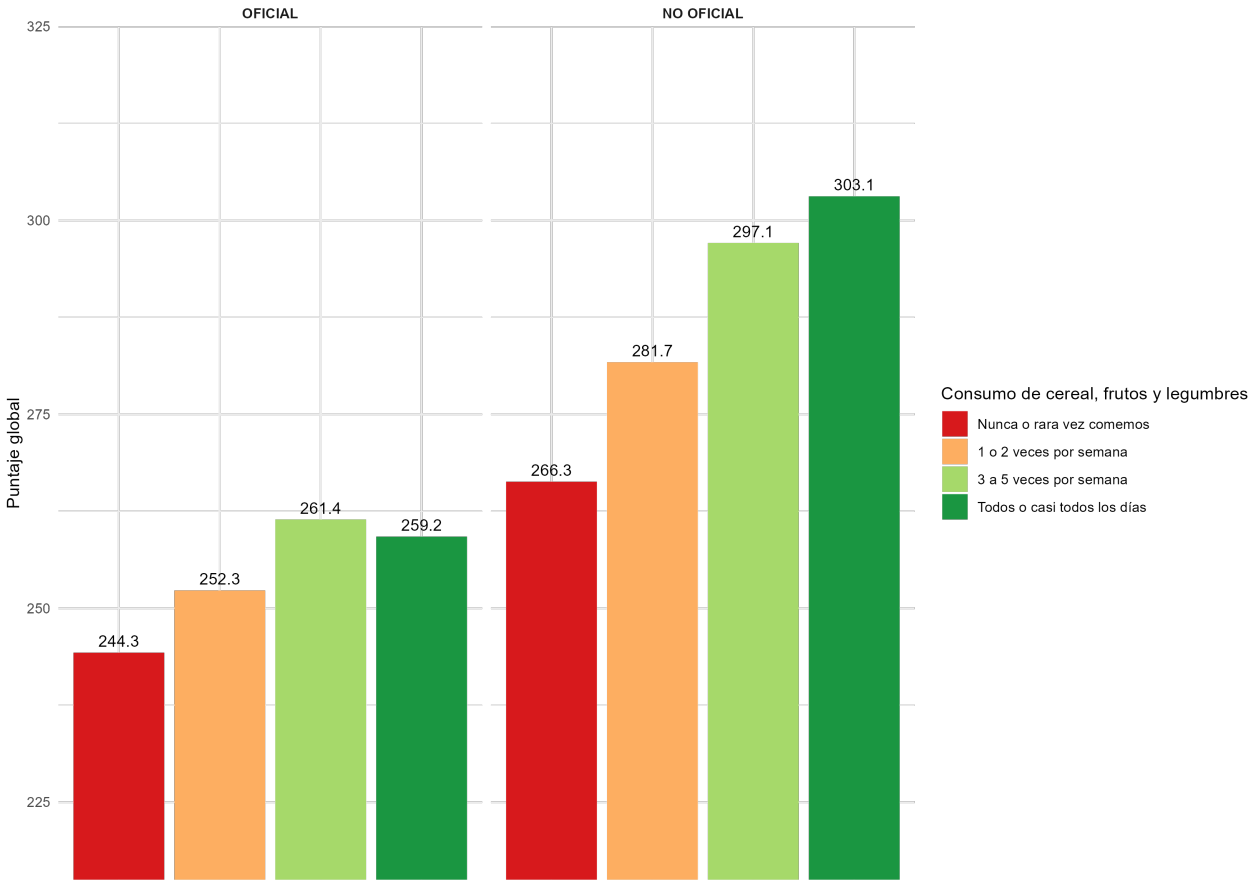

Figura 4-16: Estudiante Come cereal, futos y lejumbres (Elaboración propia)

En la Figura **4-16**, se puede observar que los estudiantes que consumen frutas, cereales y legumbres diariamente obtienen mejores resultados en promedio en la prueba Saber 11, mientras que aquellos que rara vez o nunca los consumen presentan puntajes más bajos. Además, a pesar de que algunos estudiantes de colegios oficiales manifiesten consumir alimentos como proteínas, verduras, cereales, lácteos y legumbres todos los días, obtienen menores resultados en la prueba Saber 11 que los colegios no oficiales, como se muestra en las tres graficas relacionadas con la alimentación. Es importante destacar que se trata de una tendencia y que pueden existir otros factores que influyan en los resultados de la prueba.

4.1. Cluster con K-MEANS

Se utilizó el algoritmo de K-Means para realizar la clasificación de los grupos de estudiantes. Se asignaron 10 clústeres y el algoritmo agrupó a los estudiantes teniendo en cuenta diversas variables previamente descritas. Para llevar a cabo este proceso de clasificación, se utilizó tanto RStudio como Python.
La elección de utilizar K-Means se debió a que es un algoritmo que busca aprender una estructura de agrupación en un conjunto de datos. En este sentido, el enfoque de K-Means comienza definiendo el número de grupos o clústeres que se cree existen en el conjunto de datos, representado por el valor de k. Por ejemplo, si se establece k=10, se identificarán diez clústeres, ya sea que representen una estructura de agrupación real o no.
Una vez definido el número de clústeres deseados, K-Means inicializa –generalmente de manera aleatoria– k centros o centroides en el conjunto de datos. Cada centroide puede no corresponder a un caso real de los datos, en cambio, tiene un valor aleatorio para cada variable presente en los datos. Cada uno de estos centroides representa un grupo, los casos se asignan al grupo cuyo centroide se encuentra más cercano a ellos. De forma iterativa, los centroides se mueven dentro del espacio de características con el objetivo de minimizar la variación de los datos dentro de cada grupo y maximizar la separación entre los diferentes clústeres. En cada iteración, los casos se asignan al grupo cuyo centroide se encuentra más cercano a ellos.

Los clústeres obtenidos mediante el algoritmo de K-Means tienden a ser esféricos en n dimensiones (siendo n el número de dimensiones del espacio de características). Esto implica que los conjuntos tienden a formar un círculo en dos dimensiones, una esfera en tres dimensiones y una hiperesfera de más de tres dimensiones. Además, los k grupos tienden a tener un diámetro similar, lo que indica cierta uniformidad en la distribución de los datos dentro de cada clúster Kassambara (2017). En resumen, el uso de K-Means en mi trabajo permitió clasificar a los estudiantes en grupos con base en diversas variables. De esta manera, el algoritmo proporcionó una forma efectiva de identificar estructuras de agrupación en los datos y permitió el análisis de más equitativo y características comunes entre los estudiantes.
En resumen, el uso de K-Means en mi trabajo permitió clasificar a los estudiantes en grupos con base en diversas variables. El algoritmo proporcionó una forma efectiva de identificar estructuras de agrupación en los datos y permitió el análisis de más equitativo y características comunes entre los estudiantes.

Flujo de trabajo para obtener cluster

En este trabajo se llevó a cabo un proceso de clusterización siguiendo los pasos propuestos por el autor Rhys (2020). Estos proporcionaron una guía metodológica para llevar a cabo la clusterización de manera sistemática y organizada. Este proceso se compone de los siguientes elementos:

- Carga de los datos: En la carga de los datos se utiliza la librería `readlx`, trabajada por Wickham, Bryan, y Müller (2021), para leer el archivo Excel de los datos utilizados para el presente trabajo. Para adecuar esta datos hay que formatear los nombres de las columnas para que se de uniformidad, para esto se utiliza la librería `janitor`, desarrollada por Firke, Bengtsson, Hill, y Wickham (2021). Se excluyen las variables 'estutipodocumento', 'estumcpioreside', 'periodo', 'estuinseindividual' debido que no agregan valor a la clusterizacion.

- Exploración de los datos: en esta etapa se utiliza la librería `skimr`, trabajada por Waring y Chang (2021) para identificar distribuciones, estadísticas descriptivas, valores faltantes, valores atípicos, frecuencia en los niveles de las variables categóricas.

- Preprocesamiento de los datos: el preprocesamiento utiliza la librería `tidymodels`, trabajada por Kuhn, Wickham, y RStudio (2021) usualmente empleada para normalizar las variables numéricas y convertir las variables categóricas a variables numéricas (como se muestra en la Fígura **4-1** mediante

el proceso One Hot Encoding). Este paso es necesario debido que el algoritmo K-Means solo funciona con variables numéricas.

Tabla 4-1: Proceso One Hot Encoding (Elaboración propia)

Edad	Punt_lectura	Percentil_lectura	Punt_Global	Percentil_global
1.295778	2.28063	1.41778	3.40098	1.524392
-0.94389	1.15939	0.69840	1.29448	0.516718
1.205074	0.03815	-0.52455	0.66658	-0.059094
0.921259	0.85360	0.374679	0.889386	0.1208468

- Librería para clustering: se escogen las librerías `mlr` trabajada por Bischl y cols. (2021) y la librería `tidyverse` trabajada por Wickham y Bryan (2021) de R debido que tiene una sintaxis sencilla y rápida para realizar la clusterización.

- Definir los datos a usar: se deben usar datos adecuados para la clusterización, es decir, con un preprocesamiento, con variables numéricas escaladas y variables categóricas convertidas a variables numéricas, debido que el algoritmo K-MEANS solo funciona con variables numérica. En este paso se realiza el proceso One Hot Encoding para convertir las variables categóricas a numéricas.

- Definir el algoritmo: La Librería `mlr` trabajada por Wickham y Bryan (2021), posee el algoritmo cluster k-Means para realizar el proceso de clusterizacion.

- Entrenar el algoritmo con los datos: En este paso se hace el entrenamiento, es decir, el algoritmo conocerá los datos que tiene y aplicará los pasos para agrupar en clúster.

- Guardar los datos de la clusterización: debido que el proceso de clusterización toma un tiempo, se debe guardar toda la información generada por el entrenamiento. Los datos a guardar consisten en el vector de los clústeres, es decir, un vector que indica el clúster que pertenece a cada registro.

- Agregar clústers a la data original: se realiza la adición de una columna que contiene el clúster de cada registro de la data, para realizar posteriores análisis.

- Visualización de los clusters

- Reducción de variables: para visualizar los clústeres se necesita reducir los datos de muchas variables a dos variables mediante una reducción de variables llamada Análisis Componentes Principales y así representar en un gráfico 2D los clústeres. Los datos se vuelven más escasos a medida que aumenta el número de dimensiones; se debe tener en cuenta el creciente espacio vacío con mayores dimensiones.

Cabe aclarar que lo primero que se hace antes de aplicar el algoritmo de reducción de dimensiones es centrar los datos restando la media de cada variable para cada caso. Esto coloca el origen en el centro de los datos, luego se encuentra el primer eje principal que es el eje que pasa por el origen y maximiza la varianza de los datos cuando se proyectan sobre él. Este nuevo eje principal es en realidad una combinación lineal de las variables predictoras. En un espacio de características bidimensional, el primer eje principal es el que maximiza la varianza y el segundo eje principal es ortogonal al primero. En esta situación, al trazar los componentes principales, simplemente computa como resultado una rotación de los datos.

- En la Tabla **4-2** se puede observar el resultado de aplicar el algoritmo para reducir las variables usando el algoritmo T-distributed Stochastic Neighbor Embedding. A diferencia de otros métodos de reducción de dimensionalidad, como PCA (Análisis de Componentes Principales), t-SNE se enfoca en preservar la estructura de vecindad local de los datos en el espacio de menor dimensión Maaten y Hinton (2008).

Tabla 4-2: Reducción Variables (Elaboración Propia)

V1	V2
-44.73264	-0.5006799
-45.66682	-3.6454339
-43.50692	-1.9392525
-44.08677	-1.6059699
-40.82912	-11.2227430
-45.8335	-4.1558437

- En la Tabla **4-3** se presenta la adición de los clústeres a las componentes (variables reducidas).

Tabla 4-3: Reducción de Variables con adición de clúster

V1	V2	Clúster
-44.73264	-0.5006799	8
-45.66682	-3.6454339	8
-43.50692	-1.9392525	8
-44.08677	-1.6059699	8
-40.82912	-11.2227430	8
-45.8335	-4.1558437	5

- Visualización de los clúster en componentes: Para la visualización se realiza con la librería `ggplot2` trabajada por Wickham (2016)

El siguiente clúster corresponde a los datos de las pruebas Saber 11 para el departamento de Antioquia para los años 2017-2019.

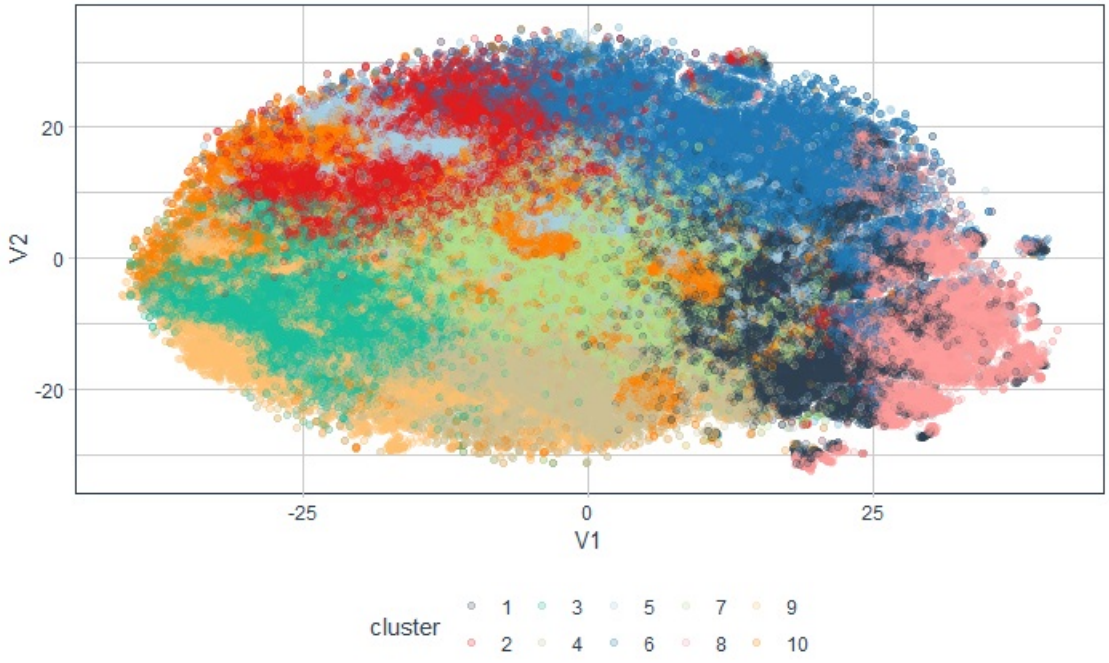

Figura 4-17: Clúster de los datos(Elaboración Propia)

Se seleccionaron 10 clústeres, los cuales están representados en la Figura **4-17** . En estos se presentan agrupados los estudiantes según características cercanas en las variables usadas. Se pueda observar la representación por colores para cada uno de los clústeres. Se utilizaron las librerías tidyverse, tidytext; luego se creó una función para clasificar las variables en los clústeres y luego se seleccionaron las variables categóricas y, finalmente, se imprime el clúster que se desea observar.

Cuando se miran las características del clúster 1 y 10 para la base de datos se encuentra lo siguiente: el clúster 10 se caracteriza por una población de estudiantes en su mayoría provenientes de áreas urbanas; asisten a colegios mixtos en jornada completa, principalmente en colegios no oficiales; dedican entre 1 y 3 horas diarias al uso de internet y entre 30 y 60 minutos a la lectura. Los estudiantes no trabajan y tienen una alimentación regular que incluye carne, pescado, huevos, cereales, frutas, leche y derivados todos los días. La mayoría de ellos vive en hogares con tres cuartos; sus padres son profesionales y tienen entre 26 y 100 libros en el hogar; en su casa residen entre tres y cuatro personas. Además, ellos cuentan con automóvil, computadora, consola de videojuegos, horno microondas y lavadora, pero no tienen moto. Este clúster se destaca por obtener los mejores puntajes globales y en lectura crítica.

En contraste, si miramos el clúster 1 se encuentra que: el clúster 1 se distingue por tener una población mayoritariamente compuesta por estudiantes urbanos que asisten a colegios mixtos en la jornada de la mañana. Estos estudiantes estudian en colegios oficiales y dedican 30 minutos o menos a la lectura. Su consumo de carne, pescado, huevo, frutos y leche se da de 3 a 5 veces por semana. En cuanto a la formación de sus padres, la mayoría tiene estudios de bachillerato. La cantidad de libros en sus hogares oscila entre 0 y 10. No cuentan con automóvil ni consola de videojuegos, y en sus hogares residen de 3 a 4 personas. Sin

embargo, poseen motocicleta. Este clúster se caracteriza por obtener los puntajes más bajos en la prueba Saber 11.

4.2. Descripción de los datos incluyendo el Clúster

A continuación se presenta una descripción de los datos para algunas variables en la que se incluye el clúster y las subregiones.

Promedio Puntaje Puntaje Global por Clúster

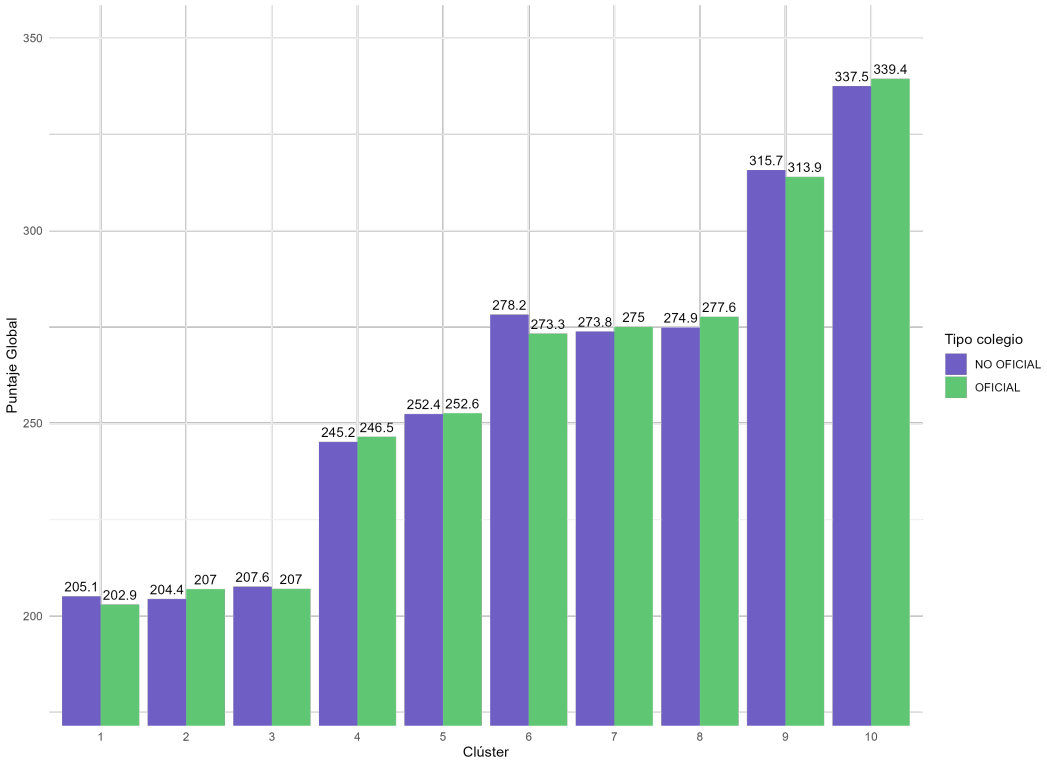

Figura 4-18: Clúster vs Promedio puntaje global (Elaboración Propia)

En la Figura **4-18** se muestra clúster para Antioquia y el promedio global en cada clúster para colegios oficiales y no oficiales. Se puede evidenciar que los clústeres con menor promedio en el puntaje global en las pruebas Saber 11 son el clúster 1, 2 y 3. En los clústeres 2, 4, 5, 7, 8, 10, el promedio en el puntaje global es mejor en los colegios oficiales, lo que nos lleva a reflexionar sobre las condiciones socioeconómicas y culturales de los estudiantes. Si observamos el clúster con mejores promedios, este corresponde al clúster 10 y el mejor promedio lo tienen los colegios oficiales en ese clúster. Para verificar estos posibles resultados se realizará una prueba de hipótesis, mediante una prueba de medias. Con el fin de evaluar los datos con mayor cuidado se presenta la prueba de hipótesis con ajuste Bonferroni:

Prueba de Hipótesis con ajuste Bonferroni

Hipótesis nula (H_o): No hay diferencia significativa en el puntaje global entre los colegios oficiales y no oficiales.
Hipótesis alternativa (H_1): Existe una diferencia significativa en el puntaje global entre los colegios oficiales y no oficiales.

Prueba de medias con ajuste Bonferroni

Clúster	Conclusión	Valor_p
Clúster 1	Rechazar la hipótesis nula	0.00008
Clúster 2	No se puede rechazar la hipótesis nula	0.00882
Clúster 3	No se puede rechazar la hipótesis nula	0.63423
Clúster 4	No se puede rechazar la hipótesis nula	0.01652
Clúster 5	No se puede rechazar la hipótesis nula	0.81774
Clúster 6	Rechazar la hipótesis nula	0
Clúster 7	Rechazar la hipótesis nula	0.00254
Clúster 8	Rechazar la hipótesis nula	0.00495
Clúster 9	Rechazar la hipótesis nula	0.00073
Clúster 10	No se puede rechazar la hipótesis nula	0.01767

Tabla 4-4: Prueba de hipótesis con ajuste de Bonferroni

De acuerdo con los resultados mostrados en la Tabla **4-11** para cada clúster, podemos interpretar lo siguiente: en los clústeres 2, 3, 4, 5 y 10, no se puede rechazar la hipótesis nula. Esto sugiere que no hay evidencia estadística suficiente para afirmar que existe una diferencia significativa en el puntaje global entre los colegios oficiales y no oficiales en estos clústeres. En contraste, en los clústeres 1, 6, 7, 8 y 9, se rechaza la hipótesis nula, lo que indica que hay diferencias significativas en el puntaje global entre los colegios oficiales y no oficiales.

Se encontraron diferencias significativas en el puntaje global entre los colegios oficiales y no oficiales en los clústeres 1, 6, 7, 8 y 9, mientras que no se encontraron diferencias significativas en los clústeres 2, 3, 4, 5 y 10. Estas conclusiones se basan en la comparación de las medias de los dos grupos y se ajustan con el método de Bonferroni para controlar el nivel de significancia global. Con esto se pude concluir que las condiciones culturales y sociales tienen un impacto significativo en el rendimiento académico de los estudiantes. De acuerdo con un estudio realizado por Diamond (2021), se ha demostrado que los factores culturales y sociales –como la educación familiar, el acceso a recursos y el ambiente socioeconómico– pueden afectar la motivación, la actitud y las habilidades de los estudiantes. Por lo tanto, es importante tener en cuenta estos factores a la hora de evaluar y mejorar el rendimiento académico.

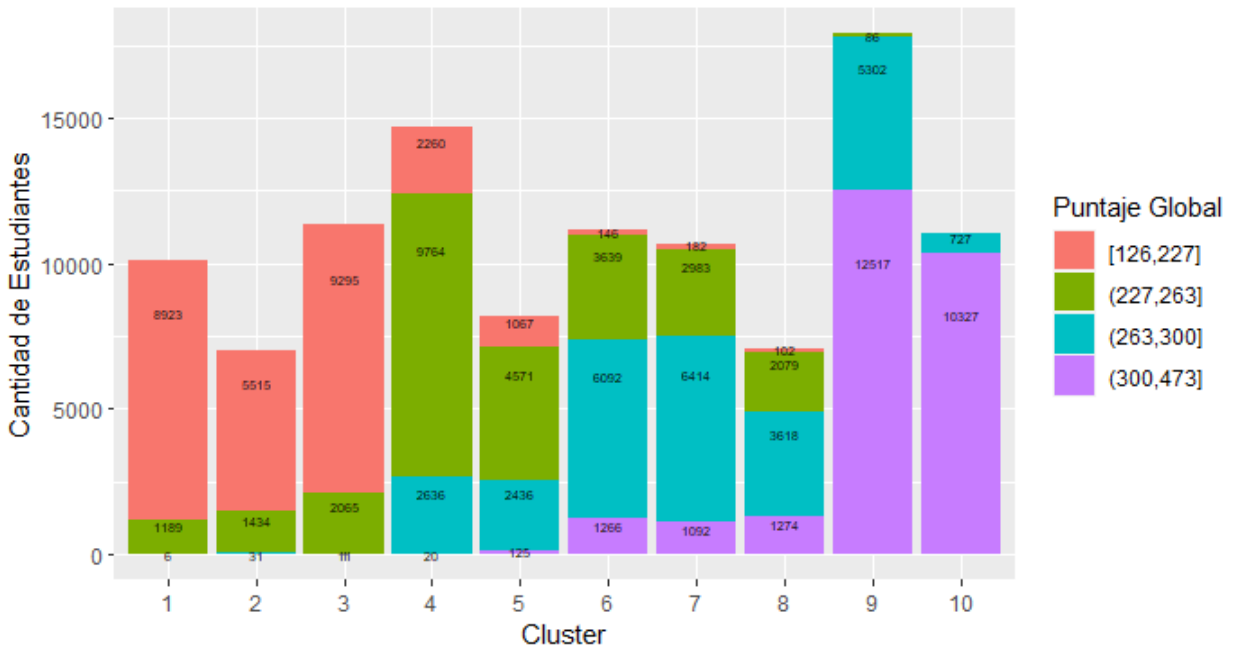

Figura 4-19: Población por cluster en Puntaje global (Elaboración propia)

En la Figura **4-19** se puede observar la organización por cuartiles para el puntaje global distribuido en cada uno de los clústeres, los puntajes más altos se encuentran en el clúster 9 y 10 y los menores en los clústeres 1, 2 y 3. En el capítulo 5 se utilizarán los cuartiles del puntaje global para poder realizar las pruebas de homogeneidad.

Tabla 4-5: Clúster Vs Puntaje Global

Clúster	**(126, 227]**	**(227, 263]**	**(263, 300]**	**(300, 473]**
1	88,2 %	11,8 %	0,1 %	0,0 %
2	79,0 %	20,5 %	0,4 %	0,0 %
3	81,7 %	18,2 %	0,1 %	0,0 %
4	15,4 %	66,5 %	18,0 %	0,1 %
5	13,0 %	55,8 %	29,7 %	1,5 %
6	1,3 %	32,7 %	54,7 %	11,4 %
7	1,7 %	28,0 %	60,1 %	10,2 %
8	1,4 %	29,4 %	51,2 %	18,0 %
9	0,0 %	0,5 %	29,6 %	69,9 %
10	0,0 %	0,0 %	6,6 %	93,4 %

En la Tabla **4-5** se muestra el porcentaje de la población en cada clúster según los cuartiles del puntaje global. En el clúster 1, se observa que el 88,2 % de los estudiantes obtiene puntajes en el rango de 126 a 227 en el puntaje global, y no hay presencia de estudiantes en el cuarto cuartil. Además, se evidencia que los estudiantes ubicados en los últimos clústeres tienden a obtener puntajes más altos y se ubican en los cuartiles 3 y 4. Esto se aprecia en el clúster 10, donde el 93,4 % de la población obtiene puntajes entre 300 y 473 puntos.

Promedio Puntaje en Lectura Crítica en cada Clúster

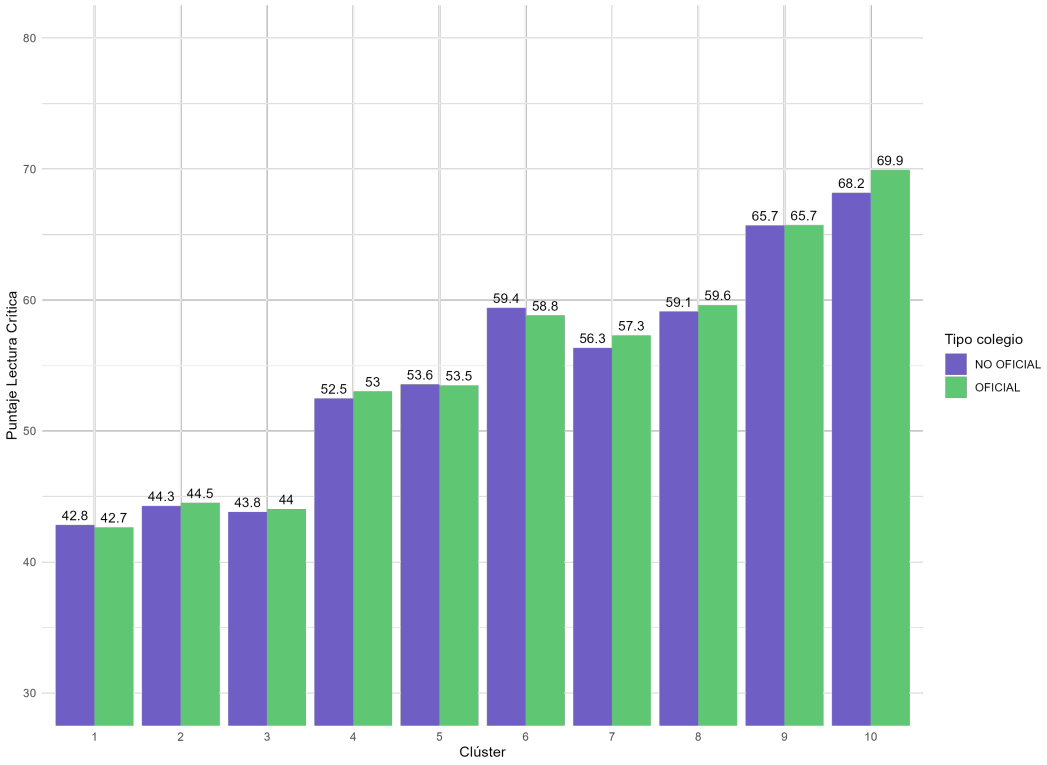

Figura 4-20: Cluster vs Promedio puntaje Lectura Crítica (Elaboración propia)

En la Figura **4-20** se puede observar que los mayores promedios en el puntaje de lectura crítica corresponden al clúster 9 y 10, con puntajes de 69,9 para los colegios oficiales y 68,2 para los colegios no oficiales. Los puntajes más bajos corresponden al clúster 1, 2 y 3; el primer clúster tiene un promedio con 42,7 para los colegios oficiales y 42,8 para los no oficiales. Los colegios oficiales presentan mayor puntaje en los clústeres 2, 3, 4, 7, 8, 10; los colegios no oficiales presentan mejor puntaje en el clúster 1, 5, 6 y en el clúster 9 tienen el mismo puntaje los colegios oficiales con los no oficiales.

Prueba de hipótesis con ajuste Bonferroni

Con el propósito de aplicar una prueba de hipótesis, se parte de las siguientes premisas:
Hipótesis nula (H0): No hay diferencia significativa en el puntaje de Lectura critica entre los colegios oficiales y no oficiales.
Hipótesis alternativa (H1): Existe una diferencia significativa en el puntaje de Lectura critica entre los colegios oficiales y no oficiales.

En los Clusters 1, 2, 3, 5, 8 y 9, no se encontró evidencia suficiente para afirmar que existe una diferencia significativa en el puntaje de Lectura crítica entre colegios oficiales y no oficiales.
En los Clústeres 1, 2, 3, 5, 8 y 9, no se encontró evidencia suficiente para afirmar que existe una diferencia significativa en el puntaje de Lectura crítica entre colegios oficiales y no oficiales. En los Clústeres 4, 6, 7 y 10, se encontró evidencia suficiente para afirmar que existe una diferencia significativa en el puntaje de

Lectura crítica entre colegios oficiales y no oficiales. Estos resultados sugieren que la diferencia en el puntaje de Lectura crítica entre colegios oficiales y no oficiales puede variar dependiendo del clúster en el que se encuentren.

Clasificación del Clúster según Puntaje en lectura crítica

A continuación, se presenta la clasificación de los estudiantes según el cuartil en el que están ubicados los estudiantes en las pruebas de lectura crítica.

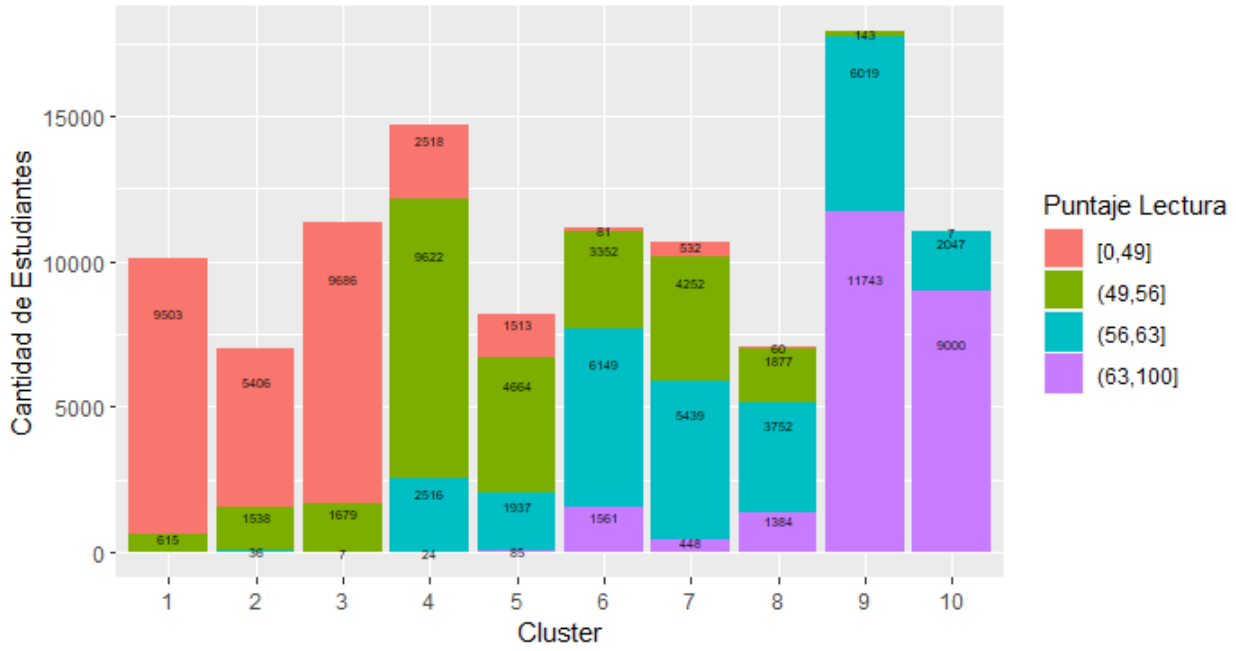

Figura 4-21: Población por clúster en lectura crítica (Elaboración propia)

En la Figura **4-21**, se puede observar la organización por cuartiles para el puntaje en lectura crítica distribuido en cada uno de los clústeres, se puede observar que los puntajes más altos se encuentran en el clúster 9 y 10 y los puntajes más bajos en el clúster 1, 2 y 3.

Tabla 4-6: Clúster Vs Cuartiles del Puntaje Lectura Crítica

Clúster	(0, 49]	(49, 56]	(56, 63]	(63, 100]
1	93,9 %	6,1 %	0,0 %	0,0 %
2	77,4 %	22,0 %	0,5 %	0,0 %
3	85,2 %	14,8 %	0,1 %	0,0 %
4	17,2 %	65,5 %	17,1 %	0,2 %
5	18,5 %	56,9 %	23,6 %	1,0 %
6	0,7 %	30,1 %	55,2 %	14,0 %
7	5,0 %	39,8 %	51,0 %	4,2 %
8	0,8 %	26,5 %	53,0 %	19,6 %
9	0,0 %	0,8 %	33,6 %	65,6 %
10	0,0 %	0,1 %	18,5 %	81,4 %

En la Tabla 4.2, se puede observar que en el clúster 1 el 93.9 % de los estudiantes obtiene puntajes entre 0 y 49 puntos, en este no hay representación de estudiantes en los cuartiles 3 y 4. En contraste, en el clúster 10, el 81.4 % de los estudiantes obtiene puntajes en lectura crítica entre 63 y 100 puntos y no tiene estudiantes ubicados en el cuartil 1. A continuación, se presenta el porcentaje de estudiantes por Clúster en cada subregión para los colegios Oficiales y No Oficiales.

Clúster Colegios Oficiales

La Tabla **4-7** presenta el porcentaje de estudiantes de colegios oficiales, desglosado por clúster y subregión.

Tabla 4-7: Clúster Colegios Oficiales

Subregión	1	2	3	4	5	6	7	8	9	10
Bajo Cauca	6,0 %	8,9 %	22,1 %	11,9 %	6,4 %	23,3 %	1,2 %	5,6 %	14,5 %	0,1 %
Magdalena medio	17,1 %	8,9 %	19,4 %	13,6 %	14,2 %	11,3 %	3,1 %	3,6 %	8,9 %	0,1 %
Nordeste	12,5 %	12,5 %	24,7 %	10,0 %	12,8 %	13,6 %	2,4 %	4,2 %	6,9 %	0,3 %
Norte	8,3 %	11,0 %	19,4 %	10,5 %	13,5 %	14,8 %	3,9 %	6,2 %	11,2 %	1,2 %
Occidente	7,7 %	12,9 %	24,6 %	8,3 %	13,9 %	17,5 %	2,3 %	5,0 %	7,2 %	0,6 %
Oriente	6,6 %	6,9 %	12,0 %	11,8 %	14,8 %	14,7 %	5,9 %	7,4 %	18,4 %	1,5 %
Suroeste	7,7 %	11,6 %	19,0 %	9,9 %	14,4 %	17,5 %	2,8 %	5,6 %	10,7 %	1,0 %
Urabá	9,8 %	13,0 %	33,0 %	11,5 %	7,0 %	16,7 %	0,7 %	3,3 %	5,1 %	0,1 %
Valle de Aburra	10,8 %	4,8 %	5,9 %	20,4 %	6,5 %	9,1 %	6,7 %	9,1 %	25,3 %	1,3 %

Se observa que en la subregión Bajo Cauca, que el mayor porcentaje de población se encuentra en el clúster 6 con un 23,3 % y el clúster 3 con un 22,1 %, mientras que el clúster 10 tiene la menor población con apenas un 0,1 %. En la subregión del Magdalena Medio, el clúster 3 es el más poblado con un 19,4 %, mientras que el clúster 10 tiene la menor presencia con un 0,1 %. En el Nordeste, el clúster 3 cuenta con la mayor población con un 24,7 %, mientras que el clúster 10 tiene la menor presencia con un 0,3 %. En la subregión Norte, el clúster 3 representa el mayor porcentaje de la población con un 19,4 %, mientras que el clúster 10 tiene la menor presencia con un 1,2 %. En el Occidente, el clúster 3 cuenta con el mayor porcentaje de población

con un 24,6 %, mientras que el clúster 8 tiene la menor presencia con un 0,6 %. En el Oriente, el clúster 9 es el más poblado con un 18,4 %, mientras que el clúster 10 tiene la menor presencia con un 1,5 %. En el Suroeste, el clúster 3 representa el mayor porcentaje de población con un 19 %, mientras que el clúster 10 tiene la menor presencia con un 1 %. En Urabá, el clúster 3 tiene el mayor porcentaje de población con un 33 %, mientras que el clúster 10 tiene la menor presencia con un 0,1 %. En el Valle de Aburrá, el clúster 9 es el más poblado con un 25,3 %, mientras que el clúster 10 tiene la menor presencia con un 1,3 %.

Clúster Colegios No oficiales

Ahora si realizamos el mismo ejercicio para colegios no oficiales, se puede observar en la Tabla **4-8** que en el Bajo Cauca, el mayor porcentaje se encuentra en el clúster 10 con un 26,9 %, mientras que los clústeres 5 y 8 tienen un porcentaje menor, cada uno con un 2,2 %. En el Magdalena Medio, el clúster 7 tiene el mayor porcentaje con un 37,2 %, y no hay población ubicada en los clústeres 2 y 3.

En el Nordeste, los clústeres 1 y 7 tienen el mayor porcentaje, ambos con un 26,5 %, mientras que los clústeres 5, 6, 9 y 10 tienen el menor porcentaje, cada uno con un 2,9 %. En la subregión del Norte, el clúster 7 tiene el mayor porcentaje con un 27,8 %, mientras que el clúster 3 tiene el menor porcentaje con un 2,1 %. En la subregión Occidente, los clústeres 7 y 10 tienen el mayor porcentaje, con un 41,2 % y un 35,3 %, respectivamente, y no hay población en los clústeres 3 y 8. En la subregión Oriente, el clúster 10 tiene el mayor porcentaje con un 59,8 %, mientras que el clúster 2 tiene el menor porcentaje con un 0,3 %. En el Suroeste, el clúster 10 tiene el mayor porcentaje con un 30,3 %, mientras que los clústeres 2 y 3 tienen un 3 % cada uno. En la subregión de Urabá, el clúster 10 tiene el mayor porcentaje con un 19,1 %, y el clúster 8 tiene el menor porcentaje con un 3 %. Finalmente, en el Valle de Aburra, el clúster 10 tiene el mayor porcentaje de la población con un 41,4 %, y el clúster 3 tiene el menor porcentaje con un 1 %. En resumen, se puede evidenciar que el clúster con mayor población en cada una de las subregiones para colegios no oficiales es el clúster 10, mientras que el clúster 3 presenta un porcentaje bajo en todas las subregiones.

Tabla 4-8: Clúster Colegios No Oficiales

Subregión	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Bajo Cauca	9,7 %	3,7 %	9,0 %	9,7 %	2,2 %	11,2 %	9,0 %	2,2 %	16,4 %	26,9 %
Magdalena medio	11,6 %	0,0 %	0,0 %	4,7 %	2,3 %	7,0 %	37,2 %	4,7 %	9,3 %	23,3 %
Nordeste	26,5 %	5,9 %	11,8 %	11,8 %	2,9 %	2,9 %	26,5 %	5,9 %	2,9 %	2,9 %
Norte	12,0 %	2,4 %	2,1 %	7,9 %	8,2 %	9,3 %	27,8 %	2,7 %	7,9 %	19,6 %
Occidente	2,9 %	2,9 %	0,0 %	2,9 %	5,9 %	5,9 %	41,2 %	0,0 %	2,9 %	35,3 %
Oriente	3,4 %	0,3 %	0,6 %	1,8 %	2,7 %	2,0 %	24,2 %	1,5 %	3,8 %	59,8 %
Suroeste	6,1 %	3,0 %	3,0 %	6,1 %	6,1 %	9,1 %	21,2 %	6,1 %	9,1 %	30,3 %
Urabá	8,5 %	4,3 %	11,6 %	7,5 %	6,2 %	12,7 %	16,9 %	3,0 %	10,1 %	19,1 %
Valle de Aburra	8,0 %	3,0 %	1,0 %	4,4 %	2,1 %	2,0 %	26,6 %	3,3 %	8,1 %	41,4 %

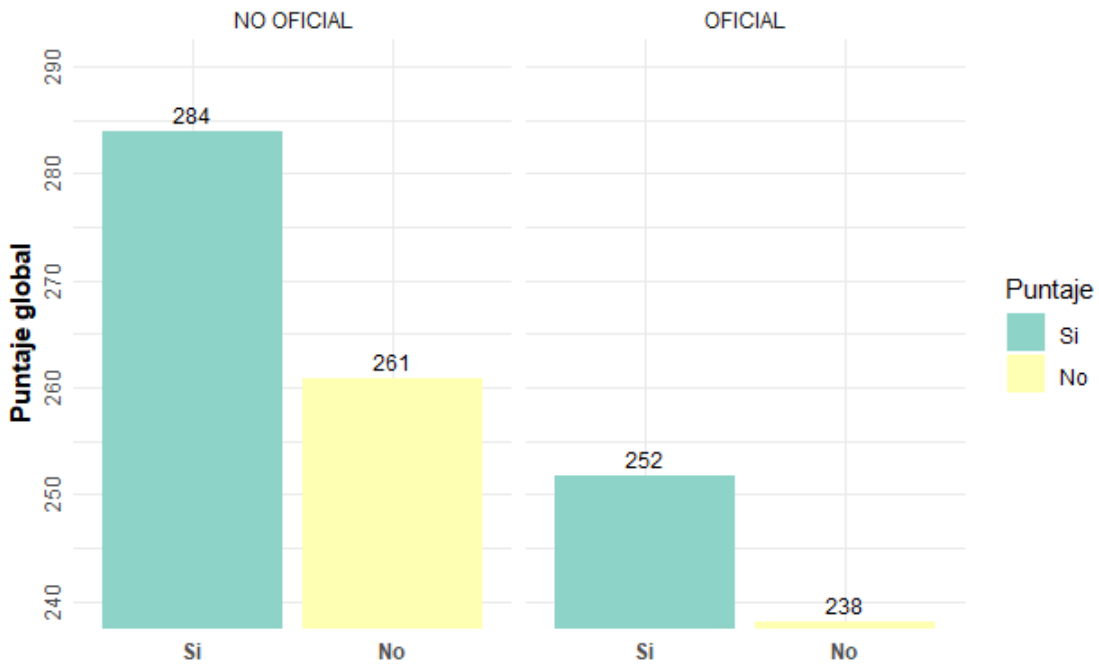

Figura 4-22: Familia tiene computador. (Elaboración propia)

En la Figura **4-22**, se puede apreciar que los estudiantes de colegios no oficiales que tienen acceso a un computador obtienen un promedio más alto en el puntaje global, alcanzando los 284 puntos. En oposición, los estudiantes de colegios oficiales que manifiestan tener computador obtienen un promedio de 252 puntos.

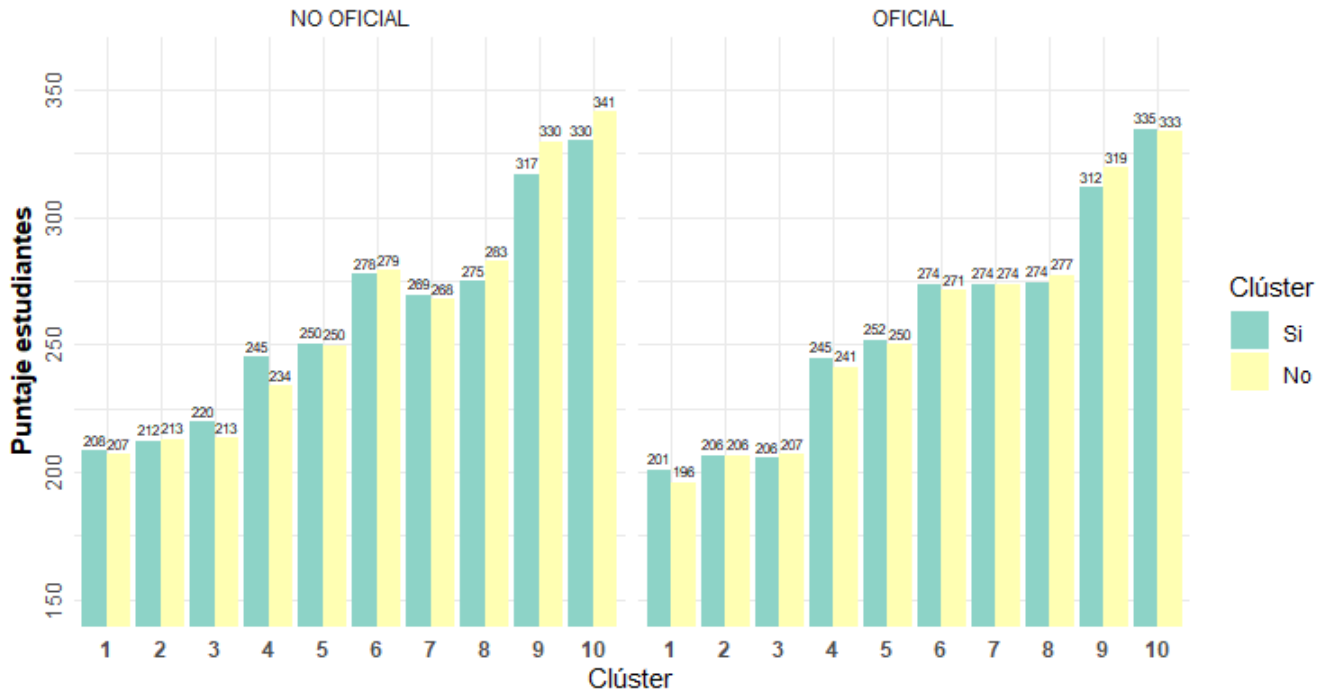

Figura 4-23: Familia tiene computador por clúster (Elaboración propia)

En la Figura **4-23**, se puede observar que los estudiantes se agrupan en 10 clústeres distintos utilizando el proceso de clusterización. Al analizar los resultados, se evidencia que los clústeres 1, 2 y 3, tanto en colegios oficiales como no oficiales, muestran los puntajes más bajos en promedio. Sin embargo, llama la atención que en el clúster 10, conformado por estudiantes de colegios no oficiales que indican no tener acceso a un computador, se obtiene el promedio más alto de puntaje. Esto puede sugerir que el acceso a un computador no es necesariamente un factor determinante en el rendimiento académico, y que existen otros aspectos que pueden estar influyendo en dicho resultado. Es importante considerar que la clusterización nos permite comparar a los estudiantes de manera más equitativa, agrupándolos con base en características comunes para poder realizar una comparación más justa entre colegios oficiales y no oficiales. Estos hallazgos nos invitan a reflexionar sobre la importancia de tener en cuenta variables adicionales que puedan impactar el rendimiento académico de los estudiantes, más allá del acceso a recursos tecnológicos. Aspectos como el entorno socioeconómico, el apoyo familiar y otros factores socioculturales pueden desempeñar un papel significativo en el desempeño estudiantil.

Cantidad de Libros vs puntaje global

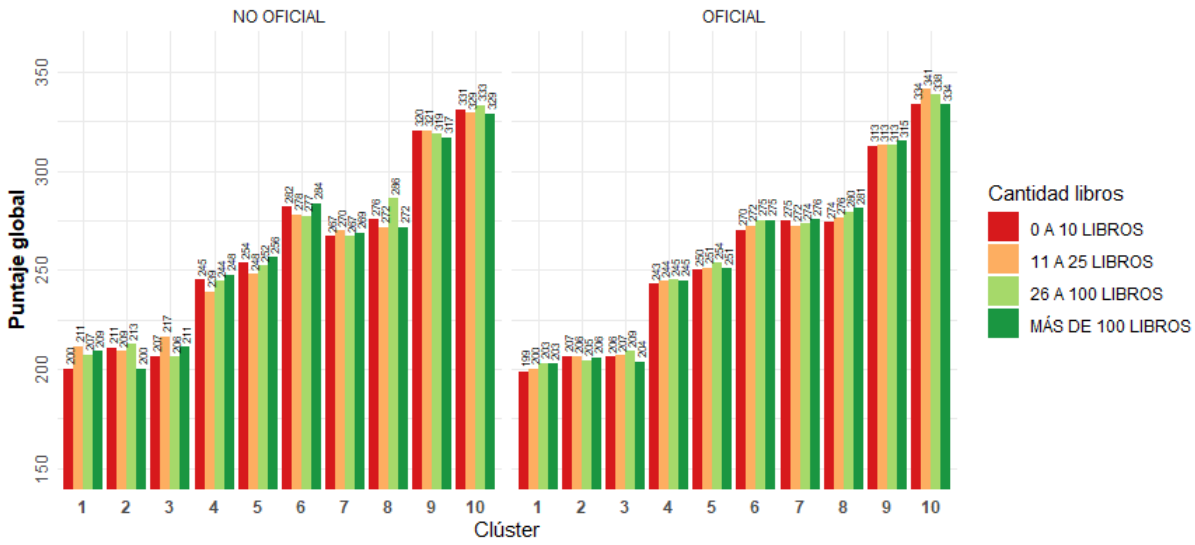

Figura 4-24: Cantidad de libros en la familia por clúster (Elaboración propia)

En la Figura **4-24**, se pueden observar los mejores promedios en los clústeres 9 y 10, tanto para colegios oficiales como no oficiales. Es interesante notar que, en el clúster 10 los estudiantes que tienen entre 11 y 25 libros en su hogar obtienen un promedio superior en comparación con aquellos que afirman tener más de 100 libros, pero esto solo aplica para los colegios oficiales. Por otro lado, en los colegios no oficiales, el puntaje más alto se encuentra en el clúster 10, específicamente entre los estudiantes que tienen entre 26 y 100 libros en su hogar.

Dedicación a la lectura vs puntaje global

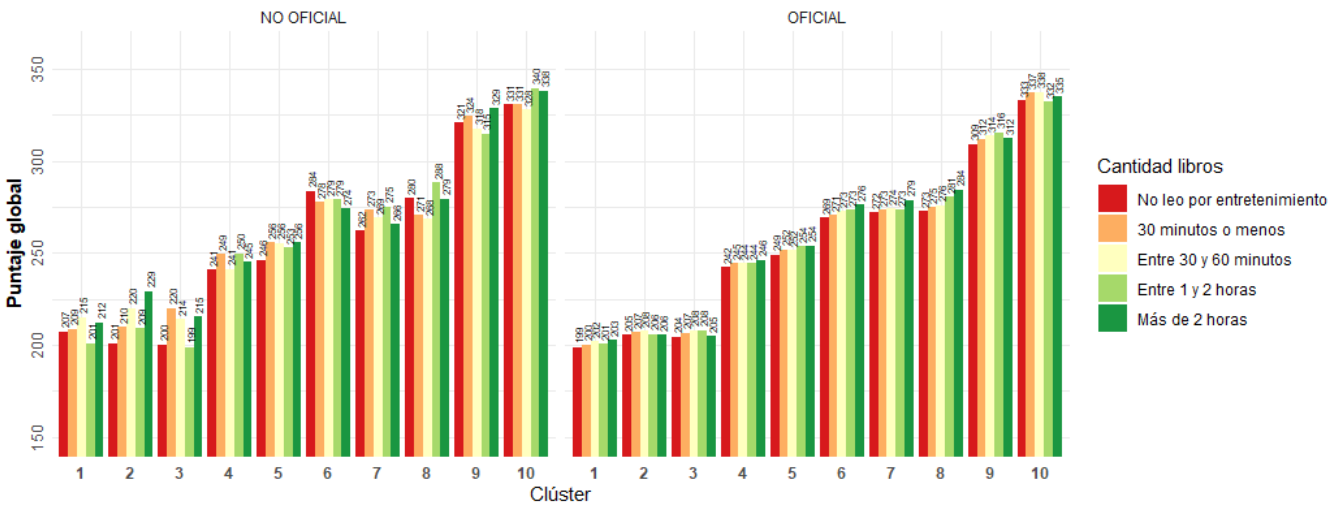

Figura 4-25: Dedicación Lectura por clúster (Elaboración propia)

En la Figura **4-25**, se puede observar que, en los colegios no oficiales, los puntajes más altos se obtienen en el clúster 10 entre los estudiantes que manifiestan dedicar entre 1 y 2 horas a la lectura. En cambio, para los colegios oficiales, el puntaje más alto se encuentra entre los estudiantes que leen de 30 a 60 minutos. Por otro lado, los puntajes más bajos se observan en los clústeres 1, 2 y 3, tanto en los colegios oficiales como en los no oficiales.

Educación de la Madre vs puntaje global

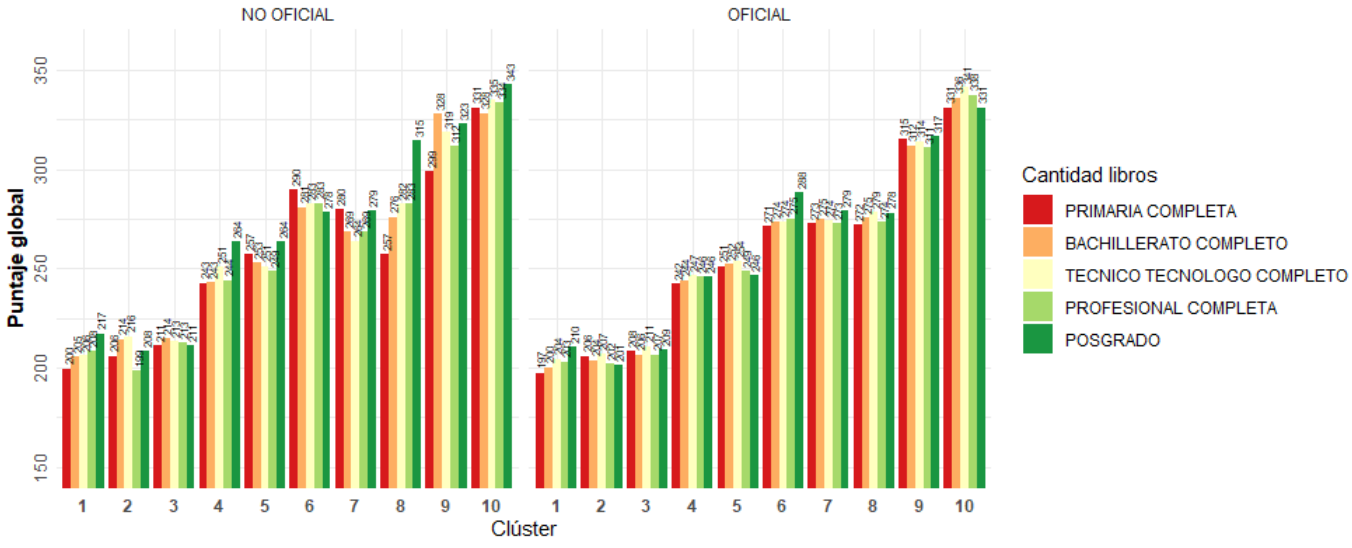

Figura 4-26: Educación de la Madre clúster (Elaboración propia)

En la Figura **4-26**, se puede observar que en el clúster 10, los estudiantes de colegios no oficiales cuyas madres tienen posgrado obtienen el puntaje más alto, mientras que, en los colegios oficiales del mismo clúster, el promedio del puntaje global lo alcanzan los estudiantes cuyas madres tienen educación técnica o tecnológica completa. Se considera importante esta variable, porque el papel de la madre en la educación de sus hijos es fundamental y tiene un impacto significativo en el desarrollo académico y personal de los niños. En este sentido, las madres desempeñan múltiples roles en el proceso educativo, tanto dentro como fuera del hogar.

En primer lugar, las madres suelen ser las principales cuidadoras de sus hijos, brindándoles amor, atención y apoyo emocional. Desde temprana edad, las madres juegan un papel crucial en estimular el aprendizaje de sus hijos, fomentando su curiosidad, exploración y desarrollo cognitivo. A través de actividades como leerles cuentos, enseñarles canciones, resolver problemas y participar en juegos interactivos, las madres contribuyen al desarrollo de habilidades lingüísticas, cognitivas y sociales de sus hijos.

Además, las madres son responsables de establecer rutinas y hábitos de estudio en el hogar. Ellas supervisan las tareas escolares, ayudan a sus hijos a organizar su tiempo y espacio de estudio, y los motivan a alcanzar sus metas académicas. La participación de las madres en la educación implica mantener una comunicación constante con los maestros, asistir a reuniones escolares y estar al tanto del progreso académico de sus hijos. Asimismo, las madres desempeñan un papel importante en la transmisión de valores, principios y normas educativas en el hogar. A través de su ejemplo y enseñanzas, transmiten la importancia de la responsabilidad, el respeto, la disciplina y la perseverancia en el proceso de aprendizaje. Las madres también fomentan el desarrollo de habilidades socioemocionales en sus hijos, como la empatía, la resiliencia y la resolución de conflictos, lo cual contribuye a su éxito académico y desarrollo integral.

Es importante destacar que el papel de la madre en la educación no se limita solo a las responsabilidades dentro del hogar. Muchas madres también enfrentan desafíos económicos y sociales: es una tendencia en crecimiento que la madre fuera del hogar para brindar sustento a sus familias. A pesar de estas circunstancias, muchas se esfuerzan por involucrarse activamente en la educación de sus hijos, buscando recursos y oportunidades para su desarrollo educativo.

En el artículo “Las madres en la educación, una voz siempre presente, pero ¿reconocida?” de Arenas (2002) destaca el papel crucial de las madres como primeras educadoras de sus hijos y cuestiona si su labor es verdaderamente valorada en la sociedad. La autora resalta la participación activa de las madres en el proceso educativo, su influencia en el desarrollo académico y personal de los niños, y su compromiso emocional. Sin embargo, también señala los desafíos que enfrentan las madres, como la falta de reconocimiento y apoyo institucional, barreras de comunicación con las escuelas y la limitación de tiempo debido a responsabilidades laborales y domésticas. El artículo invita a reflexionar sobre la importancia de reconocer y valorar el papel de las madres en la educación de sus hijos.
El papel de la madre en los hogares colombianos es invaluable. Su influencia se extiende desde el cuidado y la estimulación temprana hasta el apoyo en el proceso de aprendizaje, la transmisión de valores y el fomento de habilidades socioemocionales. Reconocer y valorar el papel de las madres en la educación es esencial para promover un desarrollo integral y equitativo de los niños en Colombia.

4.3. Subregiones

Antioquia es un departamento ubicado en la región Andina de Colombia. Se divide en varias subregiones, cada una con su propia cultura, tradición y paisaje único. Las subregiones de Antioquia se diferencian por sus características geográficas y geomorfológicas, presentan algunos modelos de desarrollo que se han ido estableciendo con condiciones sociales, culturales económicas y ambientales diversas. Las situaciones de estos territorios han conllevado a su reconfiguración: actualmente estas son conformadas por 125 municipios que se agrupan en nueve subregiones (Bajo Cauca, Magdalena Medio, Nordeste, Norte, Occidente, Oriente, Suroeste, Urabá y Valle de Aburrá).

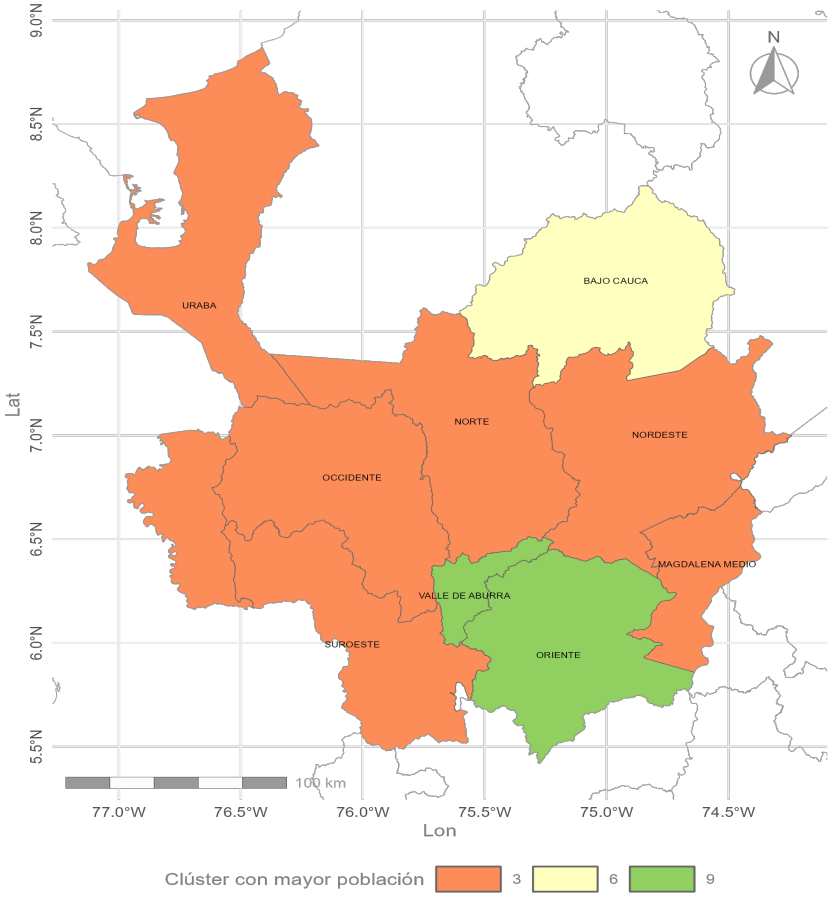

Figura 4-27: Cluster vs Población en las subregiones. (Elaboración propia)

En la Figura **4-27** se muestran los clústeres con mayor presencia en las subregiones, el clúster 3 destaca como el más poblado en la mayoría de las subregiones, mientras que el clúster 10 solo se encuentra con mayor presencia en la subregión Valle de Aburra y Oriente. Esto puede indicar que, en estas últimas dos subregiones, los estudiantes –en su mayoría– cuentan con mayores posibilidades de alimentación, conectividad de internet, padres con mayor nivel de escolaridad, entre otros.

A continuación, en la Figura **4-28** se presenta el promedio del puntaje global de las pruebas Saber 11 para cada una de las subregiones durante los años 2017 al 2019.

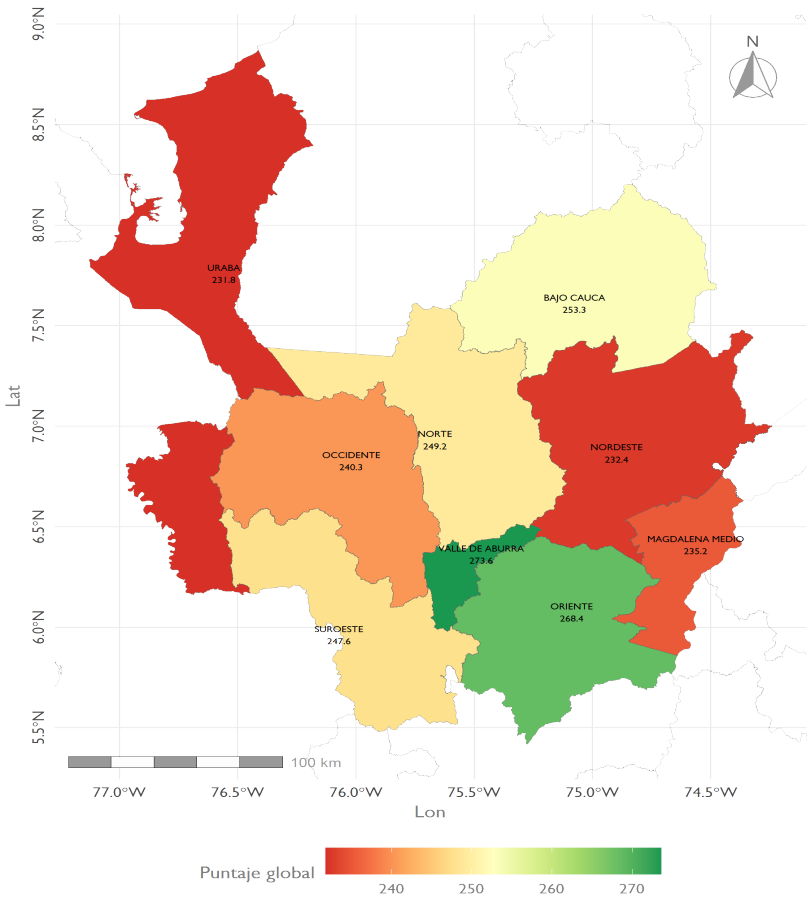

Figura 4-28: Promedio Puntaje Global en cada Subregión. (Elaboración propia)

Como se puede observar, las subregiones con menores promedios en las pruebas Saber 11 para los años 2017-2019 son Urabá, Nordeste y Magdalena Medio. Por otro lado, las subregiones con mejores promedios en la prueba Saber 11 son Oriente y Valle de Aburra.
El estudio titulado "Modelo estadístico para estimar la influencia de la lectura crítica en las competencias evaluadas en las pruebas Saber 11" Padilla-Escorcia, González-Tinoco, y Fernández-Díaz (2022), tiene como objetivo principal la verificación de la correlación entre la competencia de lectura crítica y las ciencias básicas evaluadas en las pruebas Saber 11 del 2018. La investigación se desarrolló con la participación de estudiantes de educación básica secundaria pertenecientes a 177 instituciones educativas ubicadas en la ciudad de Barranquilla. Los resultados obtenidos en el estudio mostraron la existencia de una correlación significativa entre la competencia de lectura crítica y las matemáticas, evidenciada por un coeficiente superior al 90 %. Esto respalda la suposición de que un mejor desempeño en lectura crítica se traduce en mejores resultados en el ámbito de las matemáticas. Esto va de la mano con las variables usadas en este trabajo.

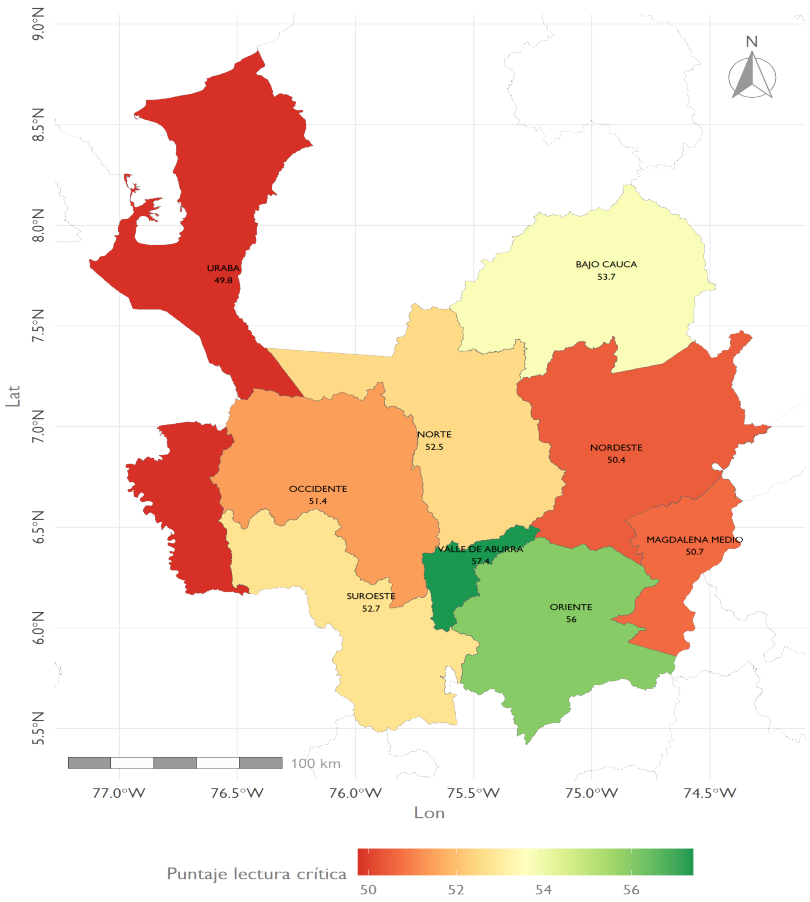

Figura 4-29: Promedio Puntaje Lectura crítica en cada subregión (Elaboración propia)

En la Figura **4-29**, se puede apreciar que las subregiones de Urabá, Nordeste y Magdalena Medio exhiben un promedio más bajo en el puntaje de lectura crítica. Por otro lado, las subregiones de Oriente y Valle de Aburrá muestran un promedio más alto en dicho puntaje. Estos resultados coinciden con las mismas subregiones presentadas en la Figura **4-28** para el puntaje global, lo cual indica una posible relación entre el puntaje de lectura crítica y el puntaje global.

Establecimientos educativos Oficiales

Subregión	Centro Educativo	Institución Educativa
Valle de Aburrá	16	338
Bajo Cauca	21	47
Magdalena Medio	5	24
Nordeste	7	34
Norte	20	47
Occidente	9	52
Oriente	19	94
Suroeste	19	70
Urabá	30	127

Tabla 4-9: Número de Establecimientos educativos Oficiales, Fuente: Directorio Único de Establecimientos Educativos -DUE-, Corte 01/12/2019

En la Tabla **4-9** se presenta el número de instituciones y centros educativos para cada una de las subregiones del Departamento de Antioquia hasta el 2019, la información no incluye la cantidad de sedes educativas adscritas a los establecimientos educativos de cada subregión. Es importante tener en cuenta que una institución educativa se refiere a un establecimiento que brinda servicios educativos desde el nivel de preescolar hasta el noveno grado, abarcando los nueve grados obligatorios. Además, algunas instituciones educativas también pueden ofrecer los dos grados de educación media y el preescolar completo. Estas instituciones suelen ser administradas por el gobierno a nivel departamental, distrital o municipal.

Por otro lado, un centro educativo es un establecimiento que no proporciona todos los nueve grados de educación básica y preescolar de manera completa. En consecuencia, estos centros educativos deben asociarse con otros para ofrecer el ciclo completo de educación básica. Los centros educativos pueden ser de gestión pública o privada y ofrecer servicios educativos especializados en diferentes áreas o disciplinas, como artes, deportes, música, idiomas, entre otros.

Establecimientos educativos No Oficiales

Subregión	Centro Educativo	Institución Educativa
Valle de Aburrá	475	231
Bajo Cauca	11	10
Magdalena Medio	7	1
Nordeste	1	9
Norte	8	16
Occidente	2	12
Oriente	39	31
Suroeste	6	10
Urabá	18	35

Tabla 4-10: Número de Establecimientos educativos no oficiales, Fuente: Directorio Único de Establecimientos Educativos -DUE-, Corte 01/12/2019

En la Tabla **4-10** se puede apreciar una diferencia en la cantidad de centros educativos en comparación con las instituciones educativas, lo cual indica que son escasos los centros educativos que brindan una educación completa. Al observar detenidamente, se nota que la mayor concentración de centros e instituciones educativos se encuentra en el Valle de Aburrá, la subregión que obtiene los mejores resultados en las pruebas Saber 11. Por otro lado, las subregiones de Nordeste y Magdalena Medio presentan una menor cantidad de centros e instituciones educativos en comparación con otras zonas.

A continuación, se dará una mirada de las pruebas Saber 11 en cada una de las subregiones teniendo como referencia el proceso de clusterización en el que se clasifica la población en 10 grupos socioeconómicos.

4.3.1. Subregión Bajo Cauca

Se compone de los municipios de Cáceres, Caucasia, El Bagre, Nechí, Tarazá y Zaragoza. La subregión Bajo Cauca del departamento de Antioquia, Colombia, es una zona que ha enfrentado diversos desafíos en materia de desarrollo social y económico. En lo que sigue, se abordarán algunos aspectos relevantes sobre educación, alimentación, economía, conectividad, internet, conflicto armado e inversión en educación entre los años 2017 y 2019.

En cuanto a la educación, el Bajo Cauca presenta una tasa de analfabetismo del 11,8 %, una de las más altas del departamento (Gobernación de Antioquia, 2018). Además, el acceso a la educación superior es limitado, con solo un 4,4 % de la población mayor de 15 años matriculada en programas de este nivel (Departamento Nacional de Planeación, 2019). Sin embargo, se han implementado estrategias para mejorar la calidad educativa, como el Programa de Alimentación Escolar, que ha beneficiado a más de 23 000 estudiantes en la subregión (Gobernación de Antioquia, 2019).

En cuanto a la alimentación, el Bajo Cauca enfrenta altos índices de pobreza y desnutrición. Según el DANE (2018), el 38,4 % de la población vive en situación de pobreza, mientras que el 17,5 % se encuentra en situación de pobreza extrema. Además, el 12,5 % de los niños menores de cinco años presenta desnutrición crónica (Gobernación de Antioquia, 2019). Para hacer frente a esta problemática, se han desarrollado programas de seguridad alimentaria como –el ya mencionado– Programa de Alimentación Escolar y el Programa de Apoyo Nutricional a la Primera Infancia (Gobernación de Antioquia, 2019).

En cuanto a la economía, el Bajo Cauca es una zona con gran potencial para la minería, la agricultura y la ganadería. Sin embargo, el conflicto armado ha afectado seriamente el desarrollo de estas actividades, especialmente en zonas rurales (Alcaldía de Caucasia, 2019). La inversión en infraestructura y servicios básicos ha sido limitada, lo que ha generado una brecha significativa en comparación con otras regiones del país Alcaldía de Caucasia (2019).

Respecto a la conectividad y el acceso a internet, el Bajo Cauca enfrenta un importante rezago en este ámbito. Según el MinTIC (2019), solo el 10 % de la población de la región tiene acceso a internet fijo y el 15 % tiene acceso a internet móvil. Esto representa una barrera para el desarrollo de nuevas actividades económicas y para la inclusión digital de la población.

Finalmente, en relación con el conflicto armado, el Bajo Cauca ha sido una de las zonas más afectadas por la violencia en Colombia. Según el Centro Nacional de Memoria Histórica (2018), entre 1985 y 2013 se registraron más de 2.700 víctimas de violaciones a los derechos humanos en la subregión.

Puntaje global en cada clúster para colegios oficiales y no oficiales

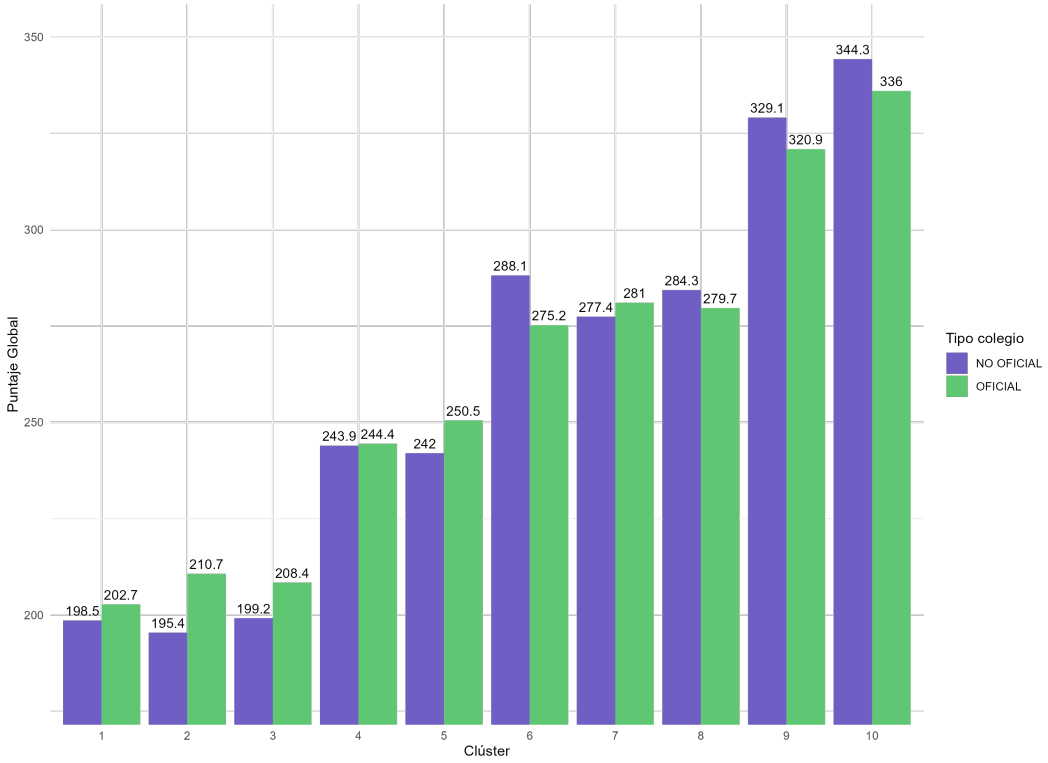

Figura 4-30: Clúster vs Puntaje Global.(Elaboración propia)

En la Figura **4-30** se puede observar que los mejores promedios se encuentran en el clúster 8, 9 y 10, tanto para colegios oficiales como no oficiales. El mejor promedio en el puntaje global se encuentra en el clúster 10 entre los colegios no oficiales.

Prueba de medias con ajuste Bonferroni

Clúster	Conclusión	pvalor
Cluster 1	No se puede rechazar la hipótesis nula	0.53679
Cluster 2	No se puede rechazar la hipótesis nula	0.37503
Cluster 3	No se puede rechazar la hipótesis nula	0.20081
Cluster 4	No se puede rechazar la hipótesis nula	0.92126
Cluster 5	No se puede rechazar la hipótesis nula	0.71484
Cluster 6	No se puede rechazar la hipótesis nula	0.02057
Cluster 7	No se puede rechazar la hipótesis nula	0.58709
Cluster 8	No se puede rechazar la hipótesis nula	0.37422
Cluster 9	No se puede rechazar la hipótesis nula	0.13972
Cluster 10	No se puede rechazar la hipótesis nula	0.76122

Tabla 4-11: Prueba de hipótesis con ajuste de Bonferroni

Basados en los resultados presentados en la Tabla **4-11**, podemos concluir que en los 10 clusters analizados no se encontró evidencia suficiente para rechazar la hipótesis nula. Esto significa que no hay suficiente información estadística para afirmar que existe una diferencia significativa en el puntaje global entre los colegios oficiales y no oficiales.

Puntaje en Lectura crítica para cada clúster en colegios Oficiales y No oficiales.

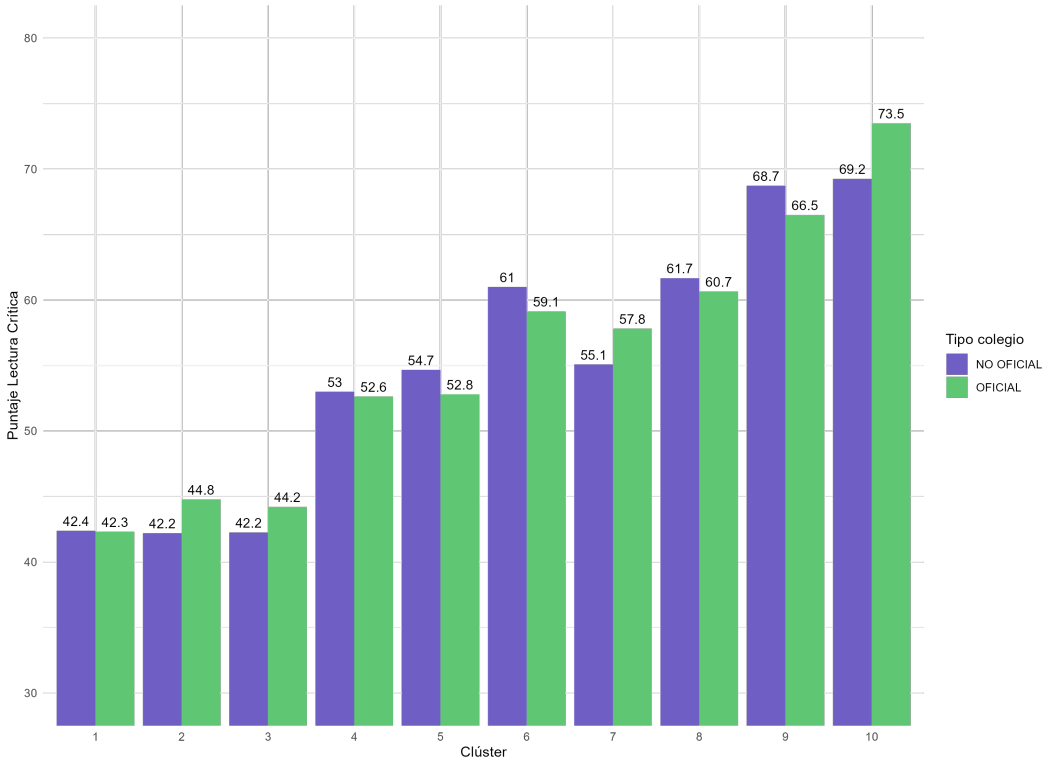

Figura 4-31: Cluster vs Puntaje Lectura Crítica(Elaboración propia)

En la Figura **4-31** se presenta el promedio en la prueba de lectura crítica en cada clúster en colegios oficiales y no oficiales. Se puede observar que los mejores puntajes en lectura crítica para el Bajo Cauca se encuentran en el clúster 9 y 10. Los puntajes más bajos corresponden a los grupos de estudiantes ubicados en los clústeres 1, 2 y 3; el mejor promedio en lectura crítica es 73,5 y corresponde a colegios oficiales.

Prueba de medias con ajuste Bonferroni para la prueba de Lectura crítica

Clúster	Conclusión	Valor de p
1	No se puede rechazar H0	0.96295
2	No se puede rechazar H0	0.35883
3	No se puede rechazar H0	0.29445
4	No se puede rechazar H0	0.72474
5	No se puede rechazar H0	0.42117
6	No se puede rechazar H0	0.10883
7	No se puede rechazar H0	0.02920
8	No se puede rechazar H0	0.83131
9	No se puede rechazar H0	0.00651
10	No se puede rechazar H0	0.31723

Tabla 4-12: Prueba de hipótesis con ajuste de Bonferroni lectura Crítica

En la Tabla **4-12** se puede concluir que en los 10 clústeres analizados no se encontró evidencia suficiente para rechazar la hipótesis nula. Esto significa que no hay suficiente información estadística para afirmar que existe una diferencia significativa en el puntaje de lectura crítica entre los colegios oficiales y no oficiales.

Porcentaje de Estudiantes vs dedicación Lectura Diaria

En la Figura **4-32**, se puede observar que en la subregión Bajo Cauca la mayor parte de la población lee 30 minutos o menos y el 32 % lee entre 30 y 60 minutos. Solo el 6 % de la población lee más de 2 horas.
Se puede observar en la **4-32**, que los estudiantes que dedican más de 2 horas a la lectura obtienen los mejores promedios en el puntaje de lectura crítica que los que dedican menos tiempo.

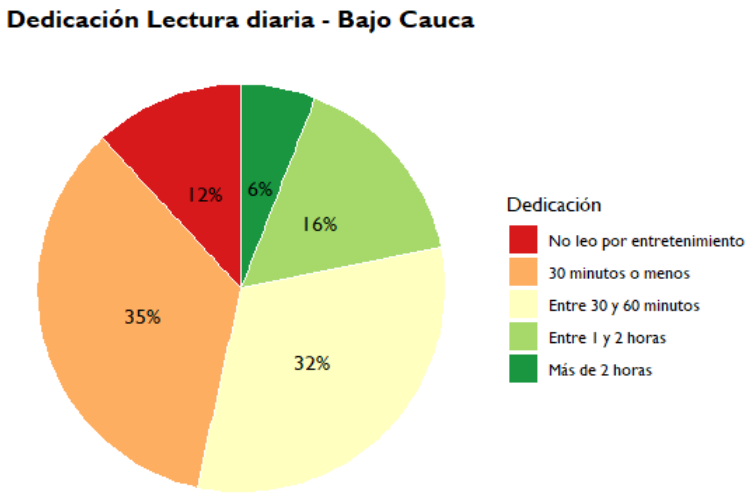

Figura 4-32: Dedicación a Lectura diaria (Elaboración Propia)

Se puede observar en la Figura **4-32**, que los estudiantes que dedican más de 2 horas a la lectura obtienen los mejores promedios en el puntaje de lectura crítica que los que dedican menos tiempo.

Dedicacion a la lectura vs Puntaje global

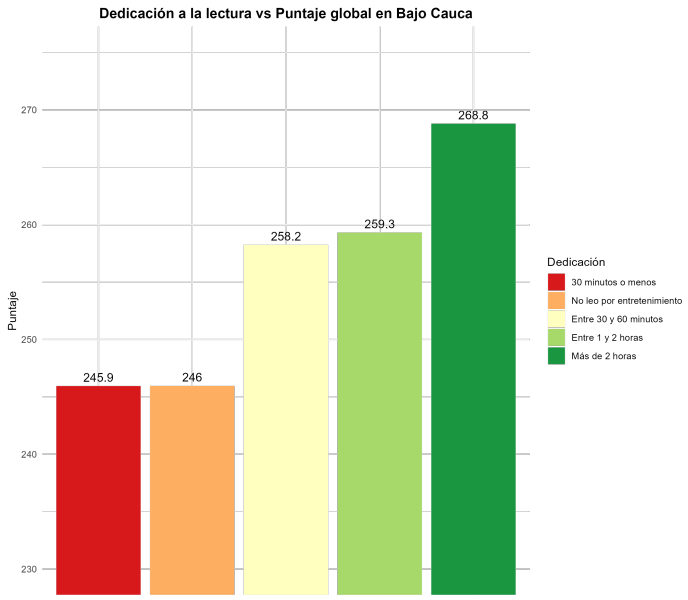

Figura 4-33: Dedicación a Lectura vs Puntaje Global (Elaboración Propia)

En la Figura **4-33**, indica que a mayor tiempo de dedicación a la lectura se obtienen mejores promedios en el puntaje global. Los promedios de los puntajes para las categorías 30 minutos o menos y no leo por entretenimiento son muy similares.

Dedicacion a la lectura vs Puntaje global para cada clúster en Colegios oficiales y No oficiales

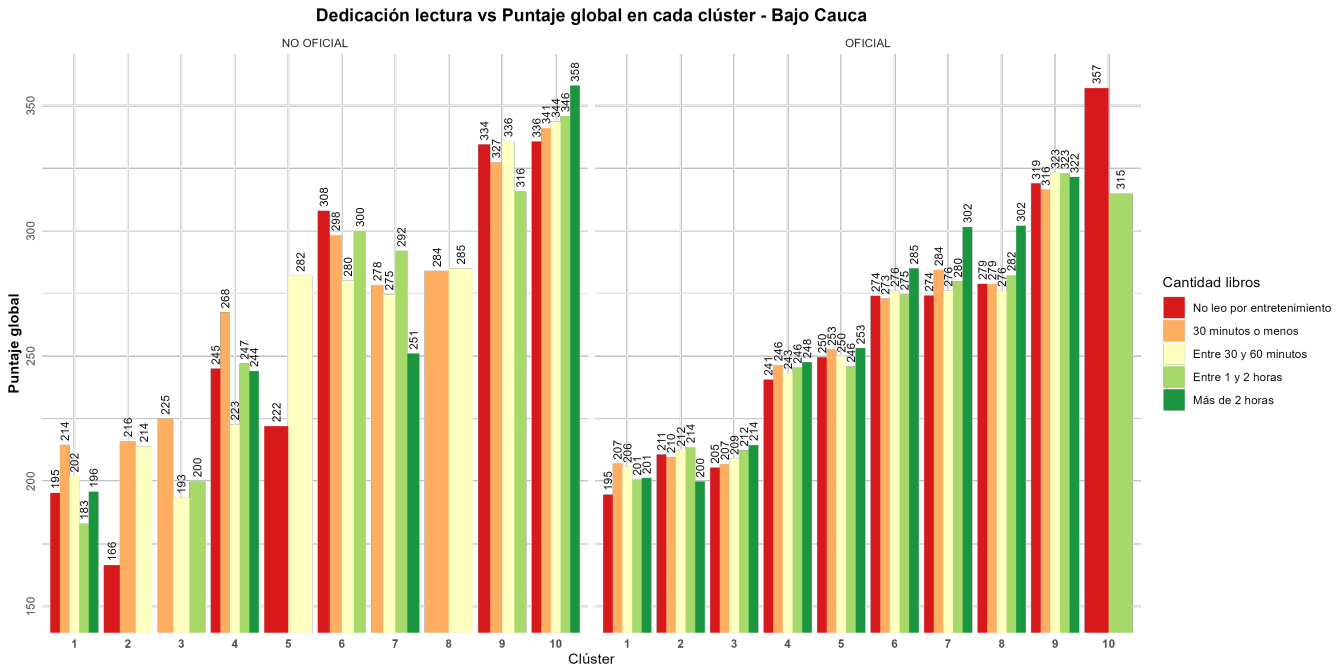

Figura 4-34: Dedicación a Lectura vs Puntaje Global para cada clúster en Colegios oficiales y No oficiales (Elaboración Propia)

En la Figura **4-34**, se puede observar que en los colegios Oficiales en el clúster 10 solo seleccionaron don ítems: no leo por entretenimiento y leo entre 1 y 2 horas; el primero posee el mayor promedio en el puntaje global. Es importante que las instituciones y familias evalúan las posibilidades de lectura para que esta sea no sea vista como una obligación. En los colegios no oficiales el mejor puntaje en el promedio global lo obtienen estudiantes que manifiestan leer más de dos horas, pero en los demás clústeres el que resalta con los mejores puntajes es leo 30 min o menos.

Familia tiene computador vs Puntaje Global

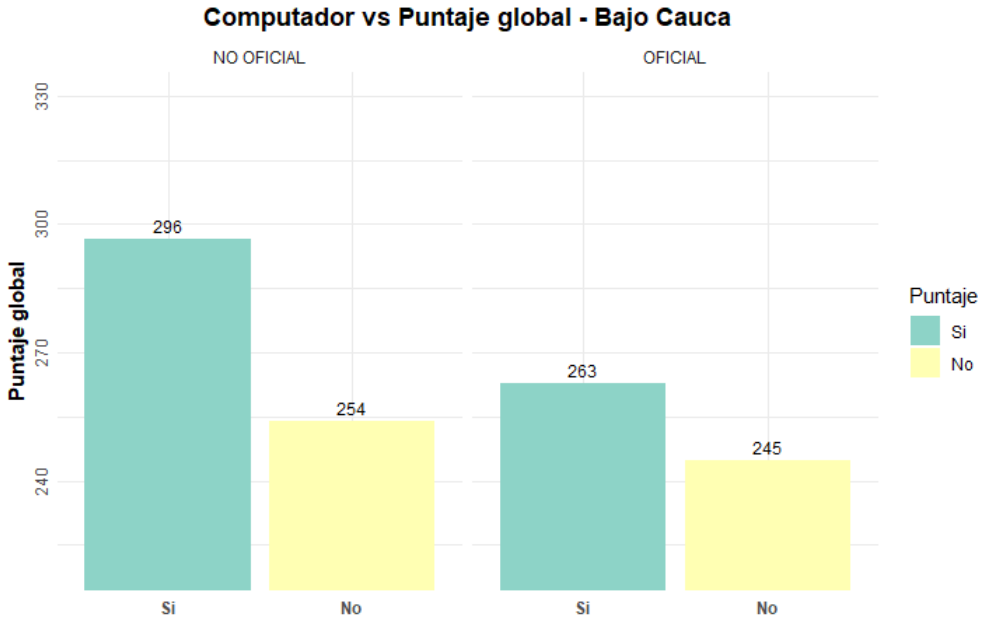

Figura 4-35: Familia tiene computador vs puntaje global(Elaboración Propia)

En la Figura **4-35**, muestra la relación entre el puntaje global y la disponibilidad de computadoras en las familias para estudiantes de colegios oficiales y no oficiales. Se observa que los estudiantes que tienen acceso a computadoras en sus casas y que estudiaron en colegios oficiales, tienen un promedio más alto, mientras que aquellos que no tienen acceso a los computadores y que estudiaron en colegios oficiales obtienen un promedio más bajo. Es importante destacar que el 57,1 % de los estudiantes de colegios oficiales en la subregión manifiestan no tener computadoras, mientras que solo el 1,21 % de los estudiantes de colegios no oficiales reportan no tener computadoras.

Familia come carne, pescado y huevo

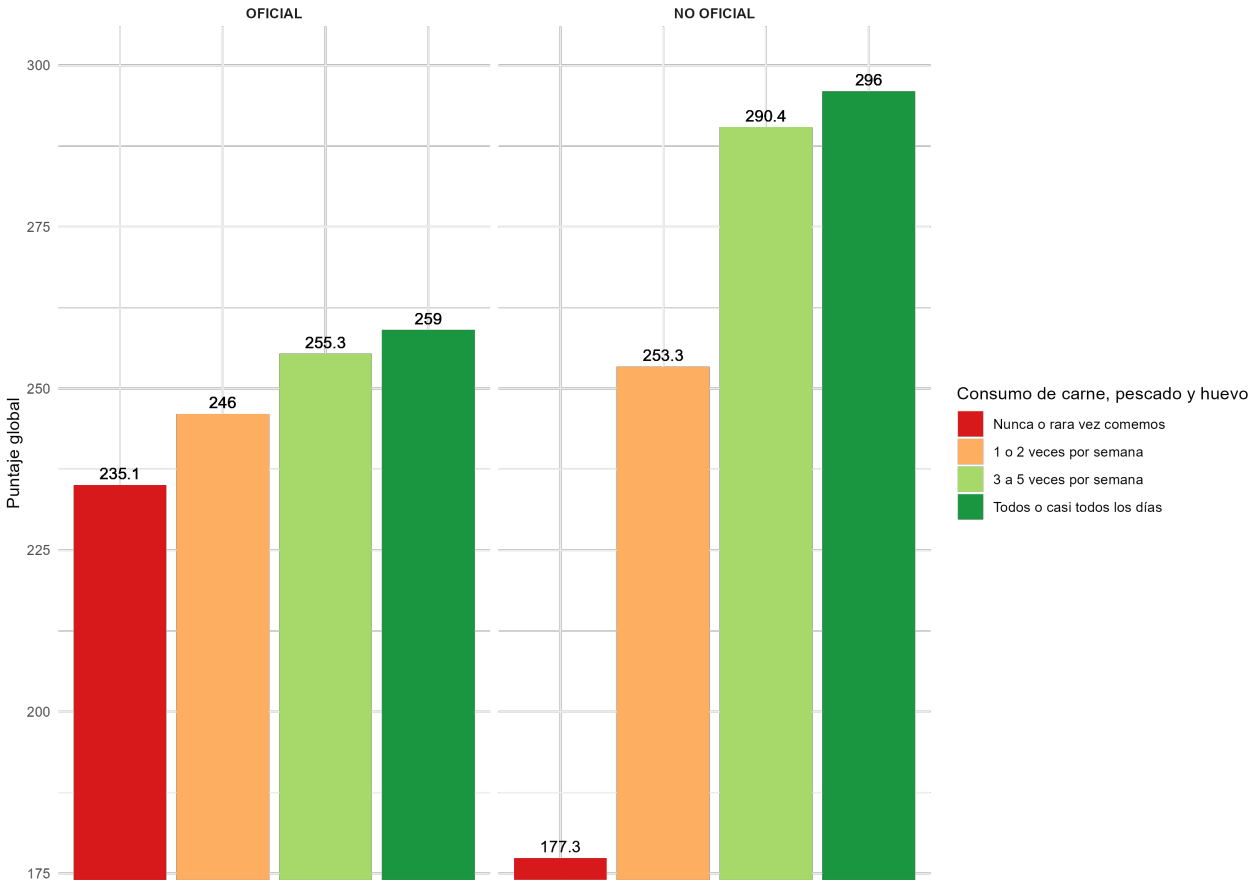

Figura 4-36: Familia consume carne Bajo Cauca(Elaboración propia)

En la Figura **4-36** se puede observar que los mejores promedios se encuentran en los estudiantes que reportan comer todos los días carne, pescado y huevos y los puntajes más bajos se encuentran en los que reportan nunca o rara vez. En lo relacionado a la distribución entre instituciones oficiales y no oficiales, se evidencia un mayor promedio en los colegios no oficiales.

Familia tiene moto vs Puntaje Global

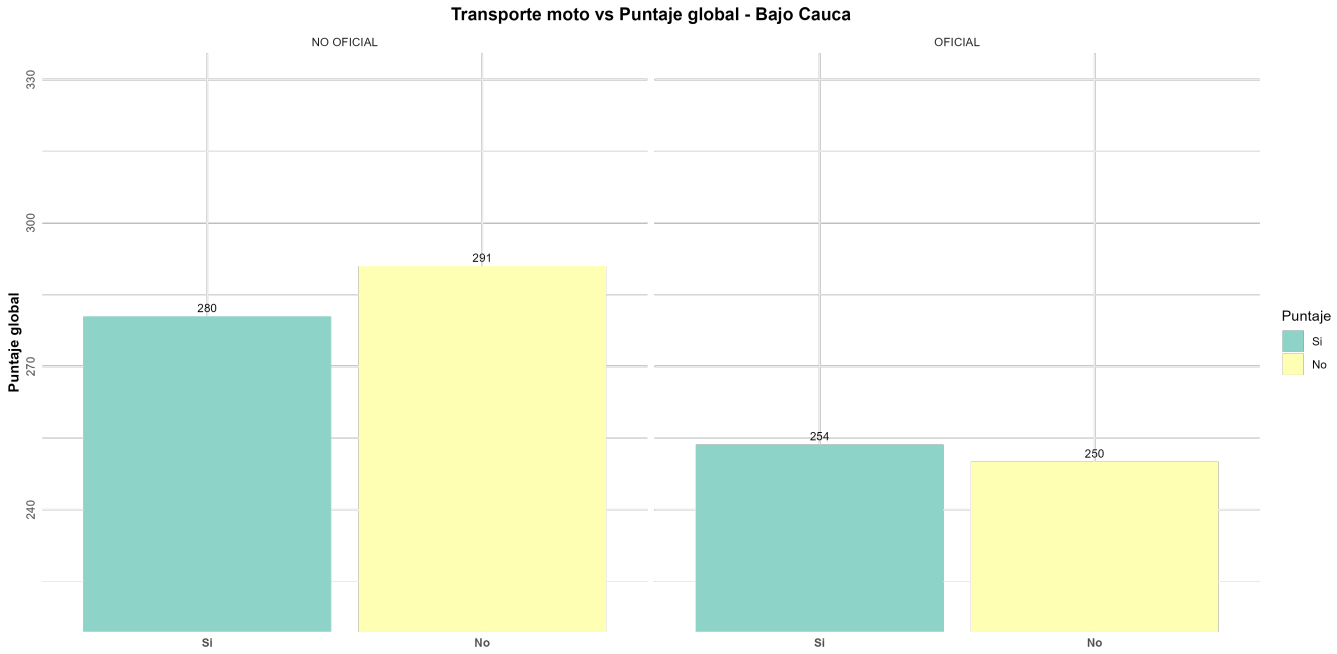

Figura 4-37: Familia tiene Moto Bajo Cauca(Elaboración propia)

Como se evidencia en la Figura **4-37**, en esta subregión los estudiantes que manifiestan tener moto en la familia tienen mejores promedios en el puntaje global. Esto puede ser porque la moto es un medio importante de transporte y no es un distractor sino un facilitador. Los mejores colegios públicos en esta región son: I. E. La Misericordia, I. E. Liceo Caucasia, I. E. Divino Nino y las tres instituciones ubicadas en Caucasia.

4.3.2. Subregión Magdalena Medio

La subregión de Magdalena Medio se compone de 6 municipios: Puerto Nare, Puerto Triunfo, Puerto Berrío, Yondó, Maceo y Caracolí. La subregión del Magdalena Medio, ubicada en el departamento de Antioquia, se destaca por su riqueza natural y su potencial turístico y agroindustrial.

Según un informe de la Fundación para la Conservación y el Desarrollo Sostenible (FOLU),[1] esta región es clave para la conservación de los bosques y la biodiversidad debido a sus áreas protegidas, como el Parque Natural Regional Serranía de Las Quinchas y la Reserva Natural Otún-Quimbaya. Además, el Magdalena Medio es reconocido por su producción de cacao, café y frutas tropicales, lo que lo convierte en un área estratégica para el desarrollo de la agroindustria y la exportación de productos de alta calidad. Sin embargo, la región también enfrenta importantes desafíos en términos de desarrollo sostenible, como la deforestación, la contaminación de los ríos y la falta de acceso a servicios básicos como agua potable y energía eléctrica. Es por ello que, se necesitan esfuerzos conjuntos de los diferentes actores del territorio para garantizar un futuro sostenible para la región.

En cuanto al acceso a internet, según el Índice de Conectividad Municipal del 2020 –elaborado por el Ministerio de Tecnologías de la información y las Comunicaciones–, la subregión de Magdalena Medio presenta una baja conectividad en comparación con otros municipios de Antioquia. Por ejemplo, el municipio de Puerto Nare cuenta con una penetración de internet del 34,6 %, mientras que el municipio de Puerto Berrío tiene una penetración del 37,8 %. Estas cifras muestran la necesidad de invertir en infraestructura y tecnología orientada a mejorar el acceso a internet de la subregión.

En cuanto a la educación, la subregión de Magdalena Medio cuenta con una oferta de instituciones educativas en todos sus municipios, desde preescolar hasta educación media y técnica. Adicionalmente, según el informe del Observatorio Antioquia Presente 2021[2], se han realizado esfuerzos significativos para mejorar la calidad de la educación en la subregión. Entre las estrategias implementadas se encuentran la formación de docentes, la implementación de tecnologías educativas y el fortalecimiento de la educación en valores.

No obstante, en este mismo informe se detalla la persistencia de brechas en materia de calidad educativa y acceso a la educación superior. Por ejemplo, en el municipio de Puerto Nare, solo el 13 % de la población mayor de 25 años cuenta con educación superior. Por lo tanto, se deben promover estrategias que fomenten la formación y capacitación de la población para reducir estas brechas. En este sentido, se requieren esfuerzos conjuntos de los diferentes actores del territorio para garantizar una educación de calidad y acceso a la educación superior para toda la población de la subregión de Magdalena Medio.

Por último, respecto a la economía, la subregión de Magdalena Medio se enfoca principalmente en la agroindustria y el comercio. El Plan de Desarrollo 2020-2023 del municipio de Puerto Triunfo busca fortalecer el sector agrícola y pecuario, fomentar la innovación y el emprendimiento. Asimismo, el turismo representa una oportunidad de crecimiento económico para la región, especialmente en municipios como Puerto Berrío y Puerto Nare, que cuentan con atractivos naturales como el rio Magdalena y la Serranía de Las Quinchas.

En conclusión, la subregión de Magdalena Medio presenta una gran riqueza natural y potencial económico, pero también enfrenta desafíos importantes en términos de desarrollo sostenible y acceso a servicios básicos.

[1] Ver:https://folucolombia.org/wp-content/uploads/2022/03/Subregiones-FOLU-Antioquia.pdf

[2] Ver:https://www.undp.org/es/colombia/publications/antioquia-retos-y-desaf%C3%ADos-para-el-desarrollo-sostenible

Puntaje Global por clúster

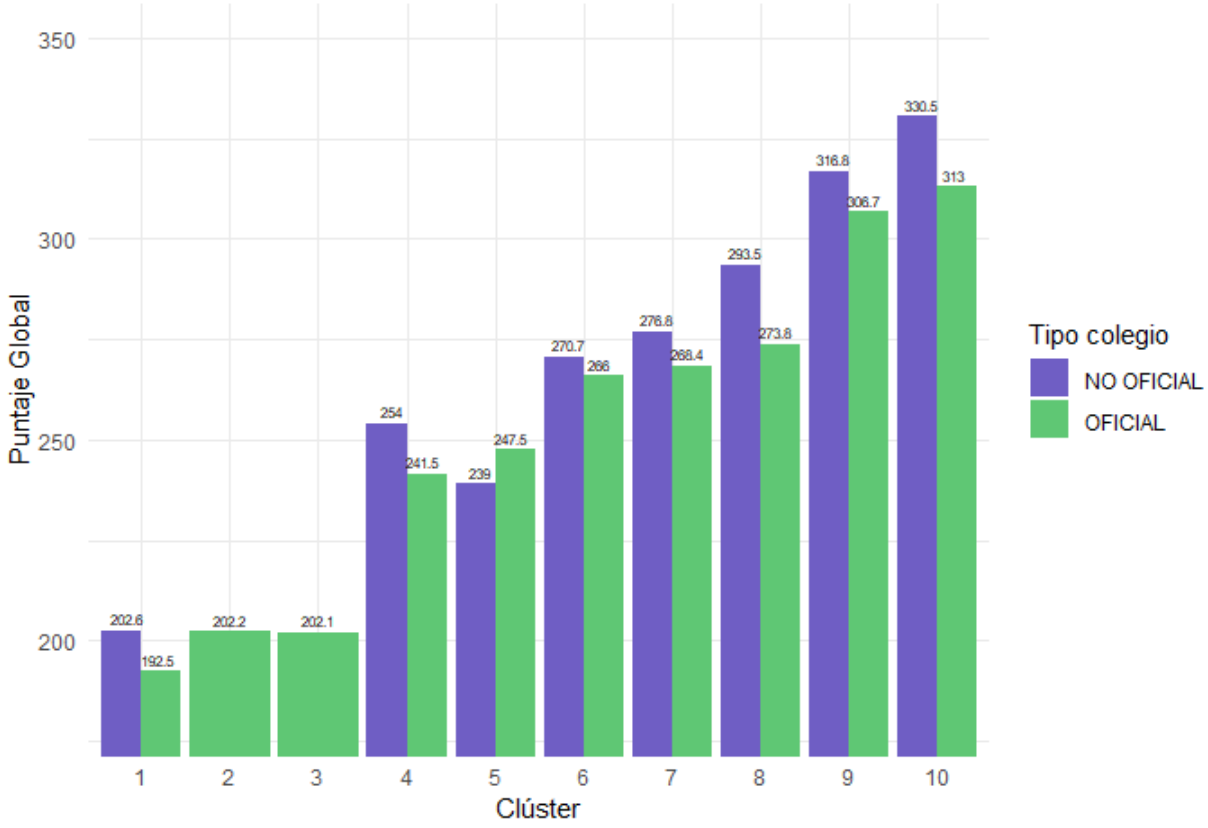

Figura 4-38: Clúster vs Puntaje Global (Elaboración propia)

En la Figura **4-38**, se observa que, en la subregión de Magdalena Medio, en el análisis de los resultados de las pruebas Saber 11 en los 10 clústeres, no se encuentra población de colegios no oficiales en los clústeres 2 y 3. Además, se destaca que los mejores promedios se encuentran en los clústeres 9 y 10, tanto para los colegios oficiales como para los no oficiales; el colegio no oficial ubicado en el clúster 10 es el que presenta el mejor promedio en el puntaje global.

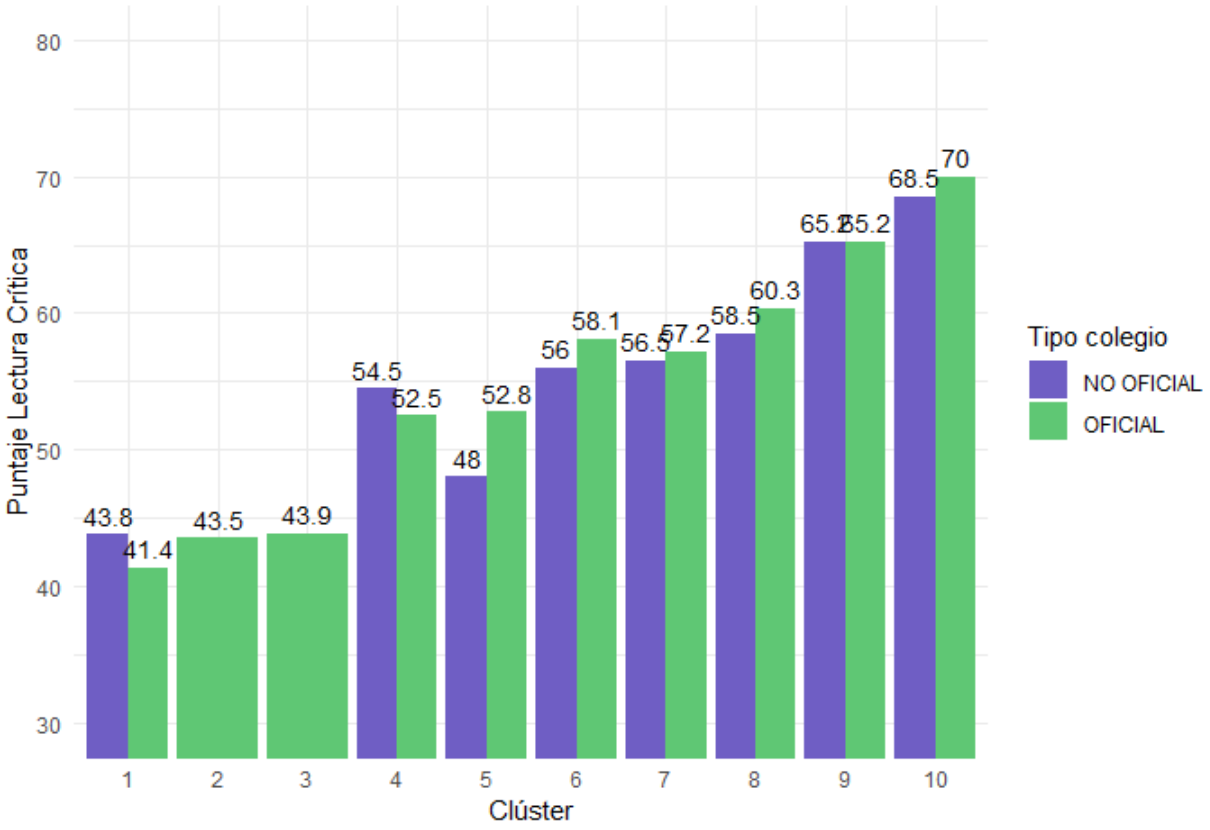

Figura 4-39: Clúster vs Puntaje Lectura crítica (Elaboración propia)

En la Figura **4-39**, se presenta el promedio de la prueba de lectura crítica para cada clúster en colegios oficiales y no oficiales en la subregión de Magdalena Medio. Se observa que los clústeres 9 y 10 presentan los puntajes más altos, mientras que los clústeres 1, 2 y 3 muestran los puntajes más bajos. Es importante resaltar que no hay estudiantes de colegios no oficiales en los clústeres 2 y 3.

Familia come carne , pescado y huevos.

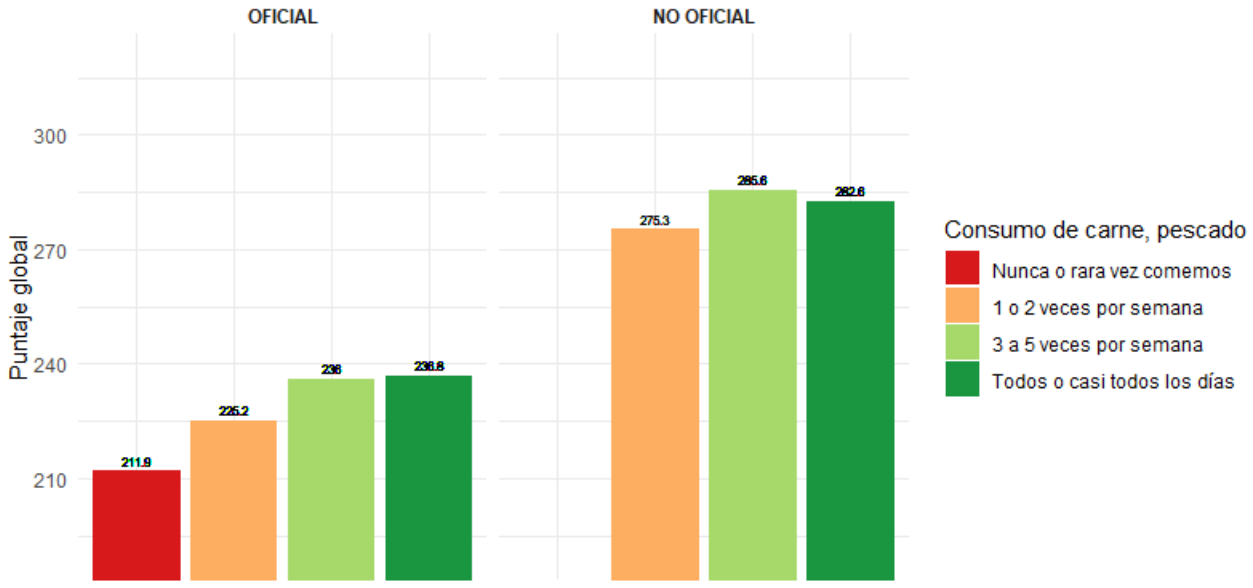

Figura 4-40: Familia come carne , pescado y huevos. (Elaboración propia)

En la Figura **4-40**, se aprecia que los colegios no oficiales presentan mejores resultados que los colegios oficiales en la subregión de Magdalena Medio. Es interesante observar que los estudiantes de colegios no oficiales que indican consumir carne, pescado y huevos de 3 a 5 veces por semana obtienen los mejores resultados, mientras que, en los colegios oficiales, a pesar de manifestar un consumo de carne superior a 3 veces por semana, no se alcanzan resultados elevados. Esto sugiere que otros factores pueden influir en el desempeño académico, y no solo el consumo de carne, pescado y huevos.

Hábitos de lectura

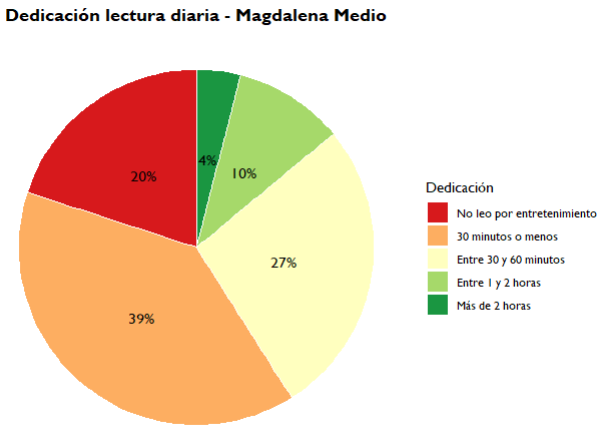

Figura 4-41: Hábitos de lectura Magdalena Medio (Elaboración propia)

En la Figura **4-41**, se puede apreciar que el 66 % de la población tiene una lectura inferior a 60 minutos. Por otro lado, solo el 4 % de la población dedica más de dos horas a la lectura. Además, un 20 % de los encuestados indica que no lee por entretenimiento. Estos datos reflejan patrones interesantes en los hábitos de lectura de la subregión Magdalena Medio.

Dedicación a la lectura vs Puntaje Global

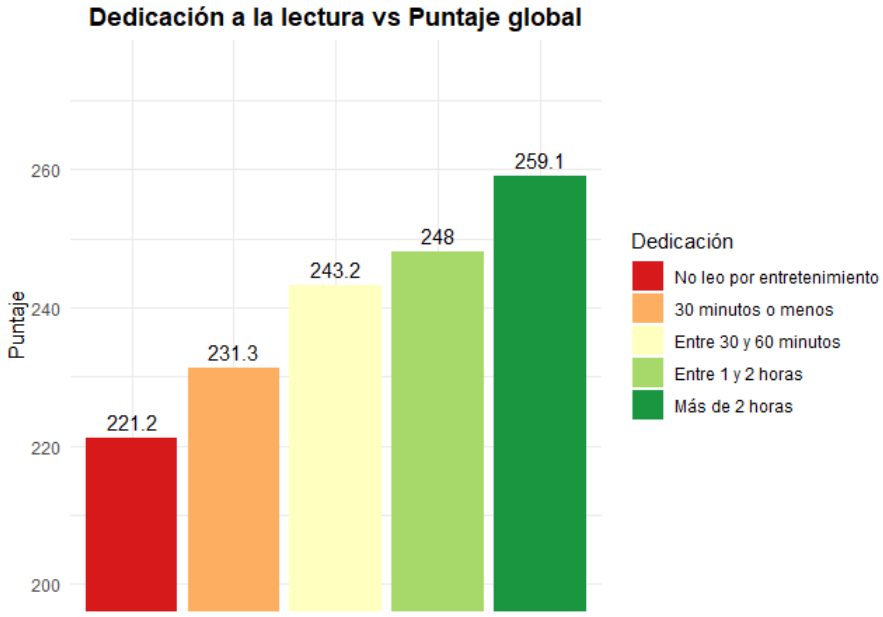

Figura 4-42: Dedicacción lectura vs Puntaje Global magdalena Medio (Elaboración propia)

En la Figura **4-42**, se puede observar que, a mayor tiempo de dedicación, se obtiene un mejor promedio en el puntaje global en la subregión de Magdalena Medio. Aquellos estudiantes que manifiestan no leer por entretenimiento obtienen un puntaje de 221,2, lo cual indica que consideran la lectura como una obligación académica y no aprovechan al máximo la lectura crítica. Los mejores puntajes son obtenidos por aquellos que afirman leer más de 2 horas, con un promedio de 259,1 en el puntaje global.

Dedicación al internet vs Puntaje Global

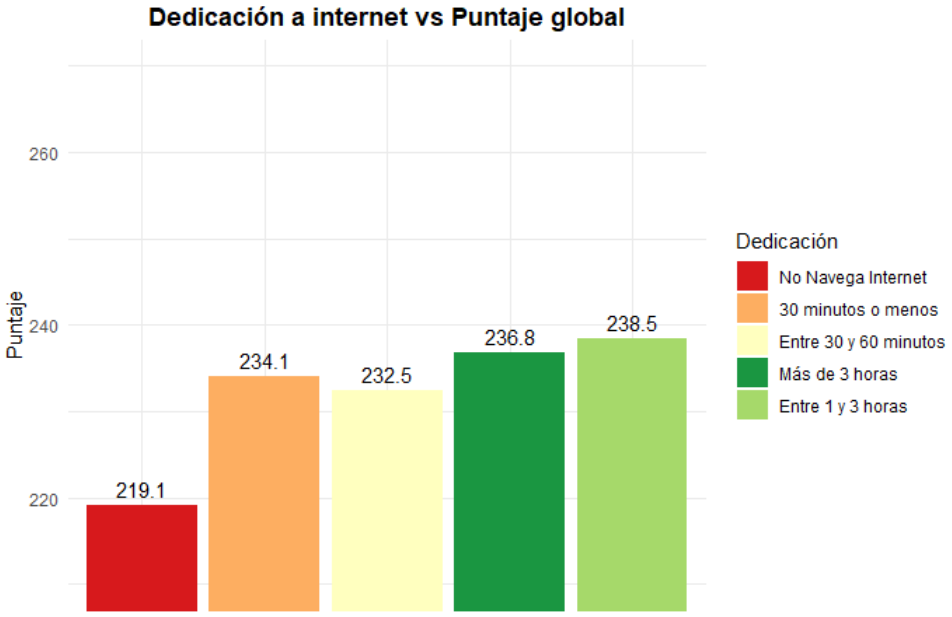

Figura 4-43: Dedicación al internet vs Puntaje Global (Elaboración propia)

Al analizar los resultados relacionados con el puntaje global de la prueba Saber 11 y el tiempo dedicado a Internet, se observa en la Figura **4-43** que los promedios están muy cercanos para aquellos estudiantes que navegan en Internet durante 30 minutos o más de tres horas. Sin embargo, es interesante destacar que el promedio es ligeramente mayor para aquellos que indican navegar entre 1 y 3 horas. Esto sugiere que puede existir una relación entre el tiempo dedicado a Internet y el rendimiento académico. Es importante considerar otros factores que puedan influir en los resultados, como la calidad del contenido al que acceden los estudiantes durante su tiempo en línea y como lo utilizan para fines educativos o recreativos.

Educación de la Madre vs Puntaje Global

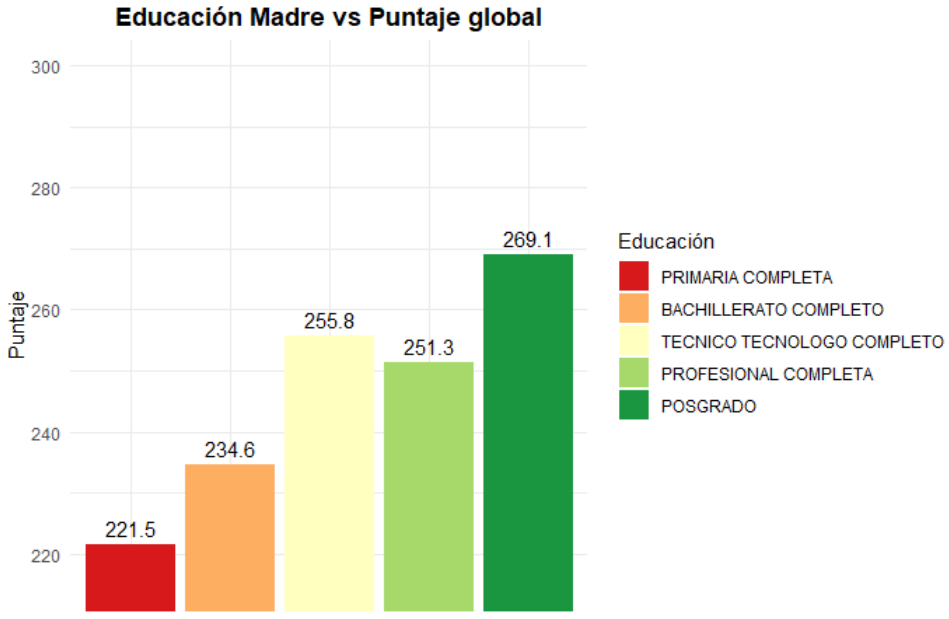

Figura 4-44: Clúster vs ¿Tiene computador? (Elaboración propia)

En la Figura **4-44**, se puede apreciar que los puntajes más altos se observan en los estudiantes cuyas madres tienen un mayor nivel de escolaridad. Los estudiantes cuyas madres poseen posgrados muestran un promedio de 269 puntos, mientras que aquellos cuyas madres solo han completado la educación primaria presentan un promedio de 221,5 puntos. Estos resultados sugieren una relación entre el nivel educativo de las madres y el rendimiento académico de sus hijos.

Mejores Colegios de la Subregión

Las mejores tres instituciones públicas en la subregión son la I. E. R. La Sierra en Puerto Nare, I. E. R. Puerto Perales en Puerto Triunfo, I. E. R. Doradal en Puerto Triunfo.

4.3.3. Subregión Nordeste

La subregión nordeste del departamento de Antioquia está compuesta por 10 municipios: Amalfi, Anorí, Cisneros, Remedios, San Roque, Santo Domingo, Segovia, Vegachí, Yalí y Yolombó. Esta región es conocida por su importante papel en la producción de oro, plata y cobre en Colombia; la actividad minera es una de las principales fuentes de ingresos.

De acuerdo con el Anuario Estadístico de Antioquia[3] 2019, la población de la subregión nordeste es de 181 752 habitantes, lo que representa el 2,9 % de la población total del departamento. La tasa de natalidad es del 17,4 %, la tasa de mortalidad es del 4,4 % y la tasa de fecundidad es del 2,1 %. Es importante destacar que la subregión nordeste forma parte de la iniciativa "FOLU Antioquia" (Food, Land Use and Ecosystem Services), cuyo objetivo es implementar prácticas sostenibles en la producción de alimentos y el uso del suelo para contribuir a la mitigación del cambio climático. Se han identificado oportunidades para la restauración de tierras y la implementación de sistemas agroforestales en la región.

Sin embargo, la subregión nordeste presenta una brecha digital significativa en cuanto al acceso a internet y a computadores, con cifras del 29,6 % y 34,5 %, respectivamente (Anuario Estadístico de Antioquia, 2019). Esto puede dificultar el acceso a información y educación en línea para los habitantes de la zona y afectar negativamente su desarrollo económico y social. Asimismo, el acceso a la educación superior en la subregión es limitado, lo que obliga a muchos estudiantes a trasladarse a otras ciudades para continuar con sus estudios.

[3] Ver:https://www.antioquiadatos.gov.co/index.php/biblioteca-estadistica/anuario-estadistico-de-antioquia/anuario-estadistico-de-antioquia-2019/

Puntaje Global por Clúster en colegios oficiales y No oficiales

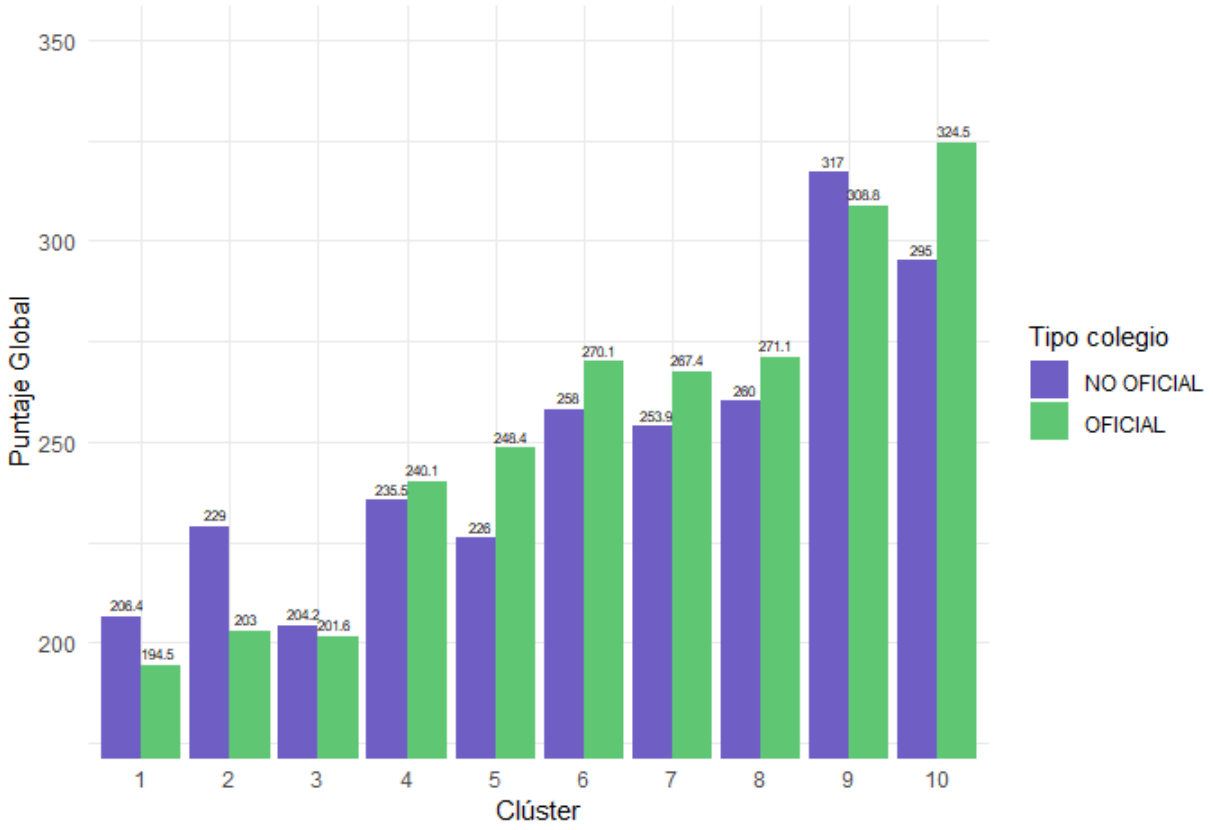

Figura 4-45: Clúster vs Puntaje Global Nordeste (Elaboración propia)

En la Figura **4-45** se puede observar que los puntajes más altos se encuentran en el clúster 9 y 10; dentro de estos, el puntaje mayor corresponde a los colegios oficiales. En el clúster 4, 5, 6, 7, 8 y 10 los promedios son mayores que en los colegios no oficiales y los colegios no oficiales obtienen mejores puntajes en el clúster 1, 2, 3 y 9.

Prueba de medias con ajuste Bonferroni

Para verificar las hipótesis, se realiza una prueba de medias con ajuste Bonferroni.
Hipótesis nula (H0): No hay diferencia significativa en el puntaje global entre los colegios oficiales y no oficiales.
Hipótesis alternativa. (H1): Existe una diferencia significativa en el puntaje global entre los colegios oficiales y no oficiales.

En los clústeres 1, 2, 3, 4 y 7 no se encontró evidencia suficiente para rechazar la hipótesis nula, lo que indica que no hay una diferencia significativa en el puntaje entre los colegios oficiales y no oficiales en estos clústeres. Los clústeres 8 presenta evidencia estadística para rechazar la hipótesis nula, lo que sugiere que existe una diferencia significativa en el puntaje entre los colegios oficiales y no oficiales en este clúster. En

los clústeres 5, 6, 9 y 10 no se pudieron realizar pruebas estadísticas debido a la falta de observaciones suficientes en alguno de los grupos. Es importante tener en cuenta las limitaciones en los clústeres con falta de observaciones, ya que no se pueden hacer afirmaciones concluyentes sobre las diferencias significativas en el puntaje entre los colegios oficiales y no oficiales en dichos clústeres.

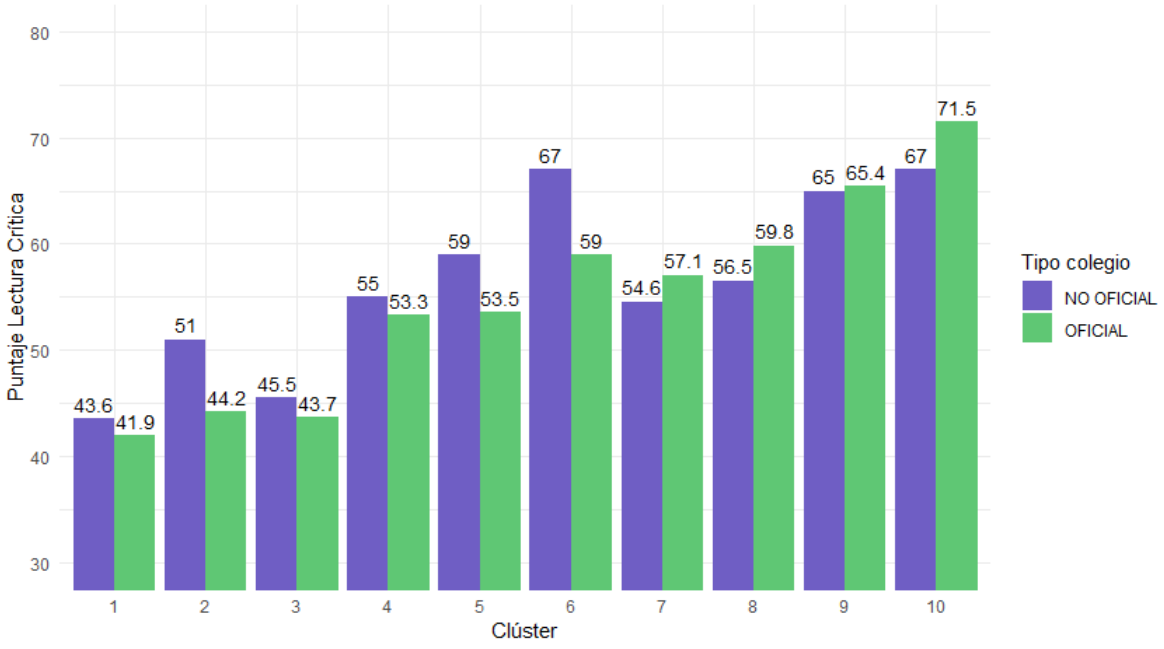

Figura 4-46: Clúster vs Puntaje Lectura crítica Nordeste(Elaboración propia)

En la Figura **4-46** Se puede observar que el mayor promedio en lectura crítica se encuentra en el clúster 10 con un promedio de 71,5 puntos y corresponde a los colegios oficiales, los colegios no oficiales obtienen mejores resultados en los clústeres 1, 2, 3, 4, 5 y 6, mientras que los colegios no oficiales obtienen mejores resultados en el clúster 7, 8, 9 y 10 para la subregión Nordeste.

Familia Come carne, pescado y huevos vs Puntaje Global

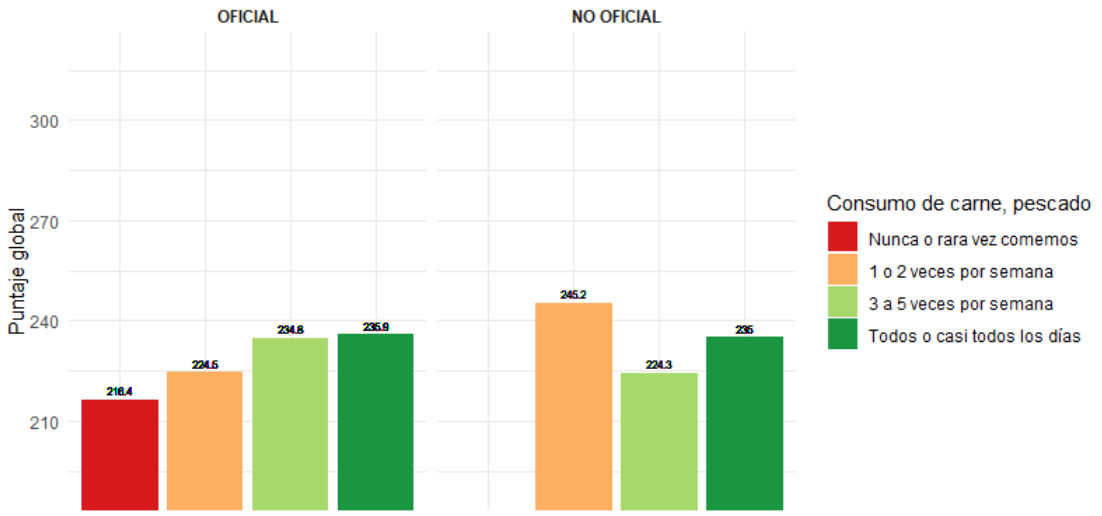

Figura 4-47: Familia come carne, pescado y huevos Nordeste (Elaboración propia)

En la Figura **4-47** se puede observar que los estudiantes en colegios no oficiales manifiestan comer carne mínimo una vez por semana y el mejor promedio se encuentra en los estudiantes que reportan comer 1 o 2 veces por semana. Para los colegios oficiales el mejor promedio se encuentra en los estudiantes que manifiestan comer carne todos los días.

Los mejores colegios públicos de la región

I. E. R. Pedro Pablo Castrillon I. E. Técnico Industrial Tomas Carrasquilla,I. E. R. Botero, Las tres Instituciones Ubicadas En Santo Domingo

4.3.4. Subregion Norte

La subregión Norte del departamento de Antioquia está conformada por 17 municipios, entre los que se encuentran Angostura, Belmira, Briceño, Campamento, Carolina del Príncipe, Donmatías, Entrerríos, Gómez Plata, Guadalupe, Ituango, San Andrés de Cuerquía, San José de la Montana, San Pedro de los Milagros, Santa Rosa de Osos, Toledo, Valdivia y Yarumal. Esta región es conocida por ser una de las zonas más afectadas por el conflicto armado en Colombia y, por ende, presenta desafíos importantes en materia de salud, educación, economía y acceso a internet.

En cuanto a la salud, según el Anuario Estadístico de Antioquia 2019, la subregión Norte presenta una tasa de mortalidad infantil de 14,1 por cada mil nacidos vivos y una tasa de mortalidad materna de 63,5 por cada cien mil nacidos vivos. Estas cifras son superiores a las del departamento y del país en general, lo que muestra una situación de vulnerabilidad en la región en cuanto al acceso a servicios de salud de calidad. Además, el conflicto armado ha afectado el acceso a la atención medica en la zona, especialmente en las áreas rurales. Según datos de la Secretaría de Educación de Antioquia, la tasa de cobertura en educación secundaria es del 62 %, lo que indica que una parte importante de la población no tiene acceso a este nivel educativo. Además, el acceso a la educación superior es limitado en la región y muchos estudiantes tienen que trasladarse a otras ciudades para continuar con sus estudios.

En cuanto a la economía, la subregión Norte es conocida por ser una zona productora de café y cacao, entre otros productos agrícolas. De acuerdo con los datos de la iniciativa FOLU Antioquia, esta región tiene una alta potencialidad para el desarrollo de prácticas sostenibles en la producción agrícola y forestal, lo que podría contribuir al fortalecimiento de la economía local y a la conservación del medioambiente.

Sin embargo, la situación de conflicto armado en la región ha afectado negativamente el desarrollo económico y social de la zona, así como la seguridad de las comunidades locales. Según datos del Observatorio de Derechos Humanos y DIH de la Defensoría del Pueblo, en el año 2021 se registraron 84 casos de violencia en la subregión Norte, lo que muestra la necesidad de fortalecer la presencia del Estado y de implementar estrategias de construcción de paz en la zona.

Sin embargo, la situación de conflicto armado en la región ha afectado negativamente el desarrollo económico y social de la zona, así como la seguridad de las comunidades locales. Según datos del Observatorio de Derechos Humanos y DIH de la Defensoría del Pueblo, en el año 2021 se registraron 84 casos de violencia en la subregión Norte, lo que muestra la necesidad de fortalecer la presencia del Estado y de implementar estrategias de construcción de paz en la zona.

Por otro lado, el acceso a internet en la subregión Norte es limitado, lo que dificulta el acceso a información y a la educación en línea para los habitantes de la zona. Según el Anuario Estadístico de Antioquia 2019, el acceso a internet en la región es del 26,9 %, mientras que el acceso a computador es del 32,2 %. Esta brecha digital afecta negativamente el desarrollo económico y social de la región, así como la capacidad de las comunidades locales para acceder a servicios de salud y educación en línea.

Puntaje Global en cada clúster

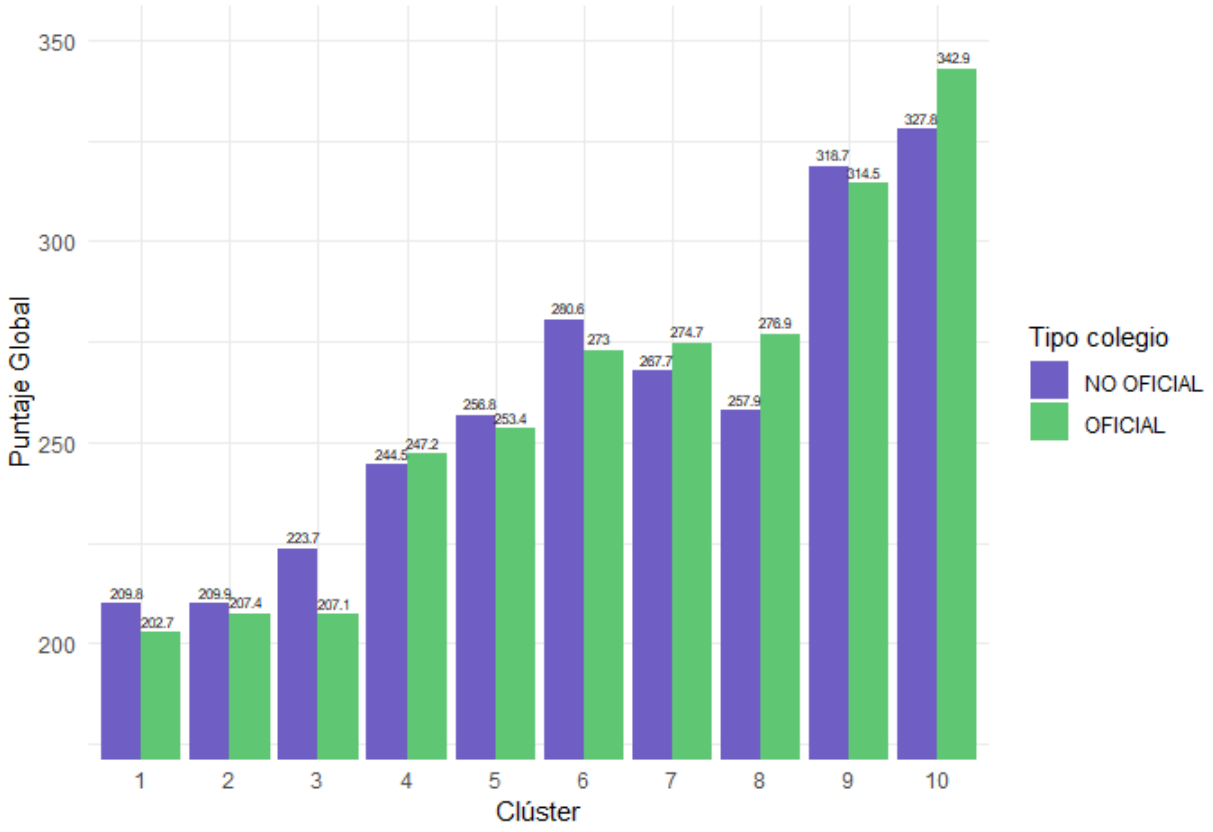

Figura 4-48: Puntaje Global en cada clúster para colegios Oficiales y No oficiales

En la Figura **4-48** se evidencia una diferencia en los resultados obtenidos entre los colegios oficiales y los no oficiales en los diferentes clústeres. Los colegios oficiales muestran un mejor desempeño en los clústeres 4, 7, 8 y 10, mientras que los colegios no oficiales lideran en los clústeres 1, 2, 3, 5, 6 y 9. Es importante destacar que los puntajes más bajos se registran en los clústeres 1, 2 y 3. Esta información revela la posible existencia de diferencias en el rendimiento académico entre los colegios oficiales y no oficiales en la subregión en cada clúster.

Prueba de medias con ajuste Bonferroni

Se realiza la prueba de medias con ajuste Bonferroni para verificar:

- Hipótesis nula (H0): No hay diferencia significativa en el puntaje global entre los colegios oficiales y no oficiales.

- Hipótesis alternativa (H1): Existe una diferencia significativa en el puntaje global entre los colegios oficiales y no oficiales.

Al examinar los resultados de los diferentes clústeres en relación con la hipótesis planteada, se observó que en los clústeres 1, 2, 3, 4, 5, 6, 7, 8 y 9 no se encontraron evidencias suficientes para rechazar la hipótesis nula. Esto indica que no existe una diferencia significativa en el puntaje de interés entre los colegios oficiales y no oficiales en estos clústeres. Sin embargo, en el clúster 10 se encontraron pruebas concluyentes para rechazar la hipótesis nula, lo que indica que existe una diferencia considerable en el puntaje entre los dos tipos de colegios en este clúster específico. Estos hallazgos resaltan la importancia de considerar las características y particularidades de cada clúster al analizar el impacto de la condición del colegio en los resultados de la prueba.

Puntaje Lectura Crítica en cada clúster

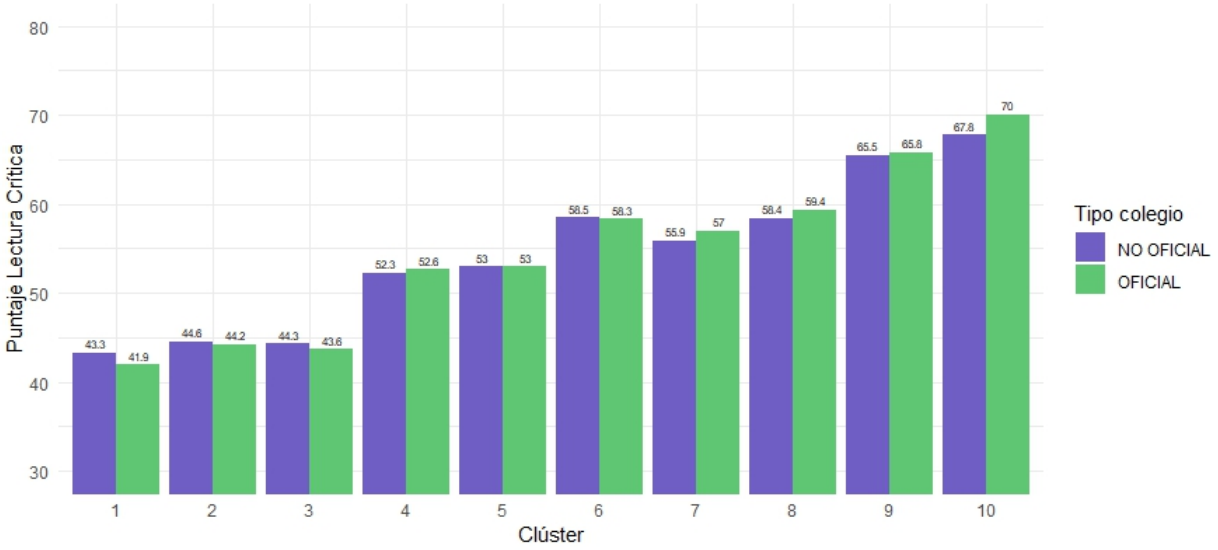

Figura 4-49: Puntaje Lectura Crítica en cada clúster

En la Figura **4-49**, se puede observar que el promedio es mayor en los colegios no oficiales en los cluster 1,2,3,5, 6 y 9 mientras que los colegios oficiales presentan mayor promedio en los clúster 4,7,8,10. el puntaje más alto fue de 70 puntos en el clúster 10 para colegios oficiales.

Mejores Colegios de la subregión

Los mejores colegios públicos de la región son: I. E. Escuela Normal Superior Pedro Justo Berrio en Santa Rosa De Osos , I. E. San Luis en Yarumal, I. E. Gómez Plata en Gómez Plata.

4.3.5. Subregión Occidente

Se compone por 19 municipios: Abriaquí, Anzá, Armenia, Buriticá, Caicedo, Cañasgordas, Dabeiba, Ebéjico, Frontino, Giraldo, Heliconia, Liborina, Olaya, Peque, Sabanalarga, San Jerónimo, Santa Fe de Antioquia, Sopetran y Uramita. La subregión Occidente de Antioquia es una zona importante del departamento de Antioquia, Colombia, que comprende 19 municipios. En términos de educación, se destaca una cobertura de educación primaria del 96,2 % y una cobertura de educación secundaria del 58,1 %. Además, hay una presencia significativa de instituciones de educación superior en la región FOLU Antioquia(2019).

En cuanto a la economía, la agricultura, la ganadería y la minería son los sectores principales, en los que se destaca el cultivo de café, plátano y maíz, así como la extracción de oro en algunos municipios. También hay una creciente actividad turística en la región, gracias a su patrimonio histórico y cultural y sus paisajes naturales. En relación con el acceso a internet, se ha registrado una baja conectividad en la subregión, con una tasa de penetración del 26,5 % en 2018, principalmente en áreas rurales de la región que carecen de infraestructura de telecomunicaciones. Por otro lado, la subregión ha sido afectada históricamente por el conflicto armado en Colombia, con la presencia de grupos armados ilegales y la ocurrencia de violencia y desplazamiento forzado de población en algunos municipios.

Puntaje Global en cada clúster

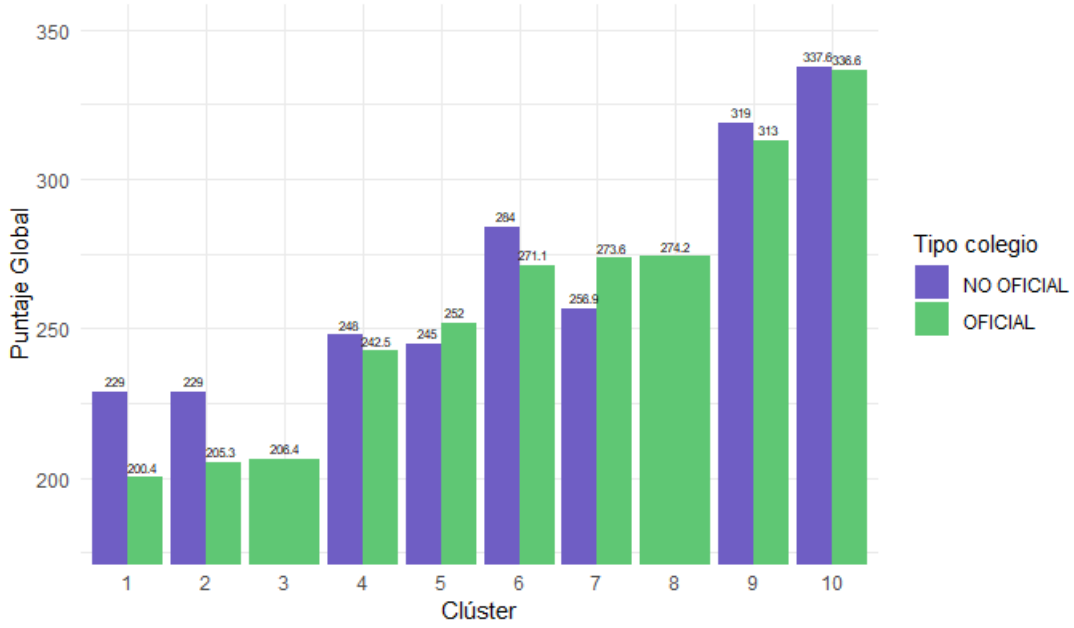

Figura 4-50: Puntaje global en cada clúster (Elaboración propia)

En la Figura **4-50**, se puede apreciar que los colegios no oficiales no tienen representación en el clúster 3 y 8. Por otro lado, los mejores puntajes se encuentran en el clúster 10, donde los colegios no oficiales obtienen el puntaje más alto. Se observa que los colegios no oficiales obtienen mejores promedios en los clústeres 1, 2, 4, 6, 9 y 10, mientras que los colegios no oficiales se encuentran con mejor promedio en los clústeres 5 y 7. Estos resultados sugieren que existe una variabilidad en el rendimiento académico entre los colegios oficiales y no oficiales en diferentes clústeres de la subregión.

Puntaje Lectura crítica en cada clúster

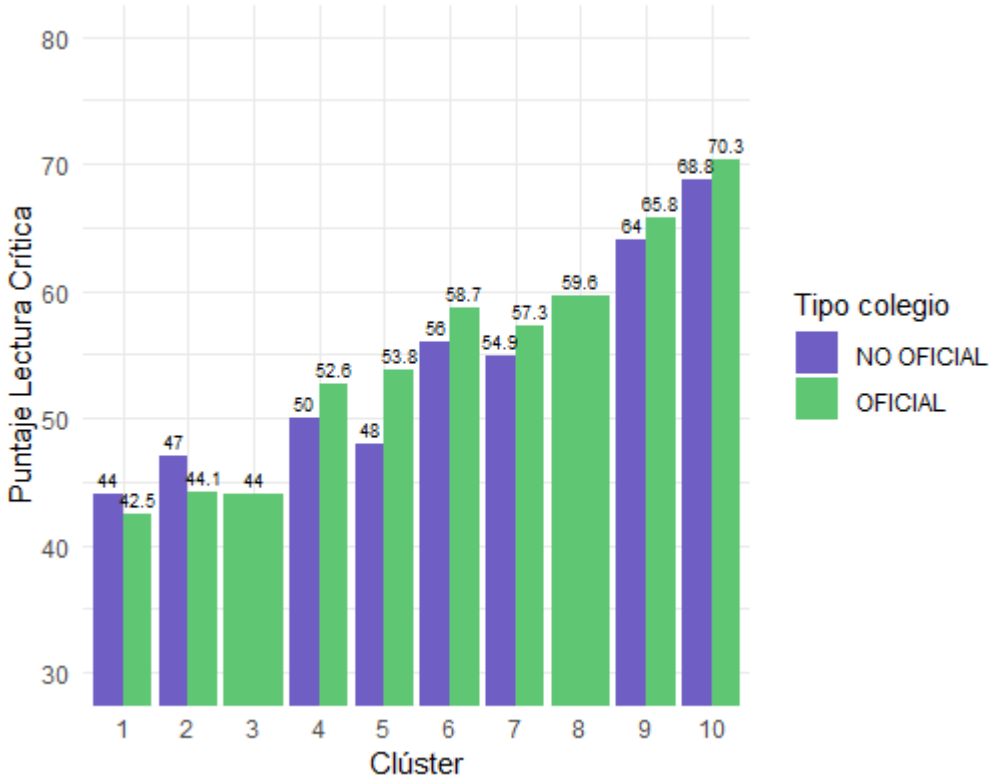

Figura 4-51: Lectura por cada clúster. (Elaboración propia)

En la Figura **4-51**, se aprecia una representación de colegios oficiales en los clústeres 3 y 8. Es importante destacar que el clúster 10, muestra el promedio más alto en la prueba de lectura crítica, alcanzando los 70,3 puntos en los colegios oficiales. Por otro lado, se identifica que el puntaje más bajo en la prueba se encuentra en el clúster 1 y pertenece a los colegios no oficiales, registrando solamente 44 puntos.

Prueba de medias con ajunte Bonferroni

Al realizar la prueba de medias con ajuste Bonferroni para (Ho) no hay diferencia significativa en el puntaje de lectura critica entre los colegios oficiales y no oficiales. Se encontró que en los clústeres 1, 2, 3, 4, 8 y 9, no disponen de suficientes observaciones para realizar la prueba. Por otro lado, en los clústeres 5, 6 y 10, se obtienen resultados de prueba en los que no se puede rechazar la hipótesis nula, con valores de p de 0.09942, 0.22034 y 0.27184 respectivamente. Sin embargo, en el clúster 7 se obtiene un valor de p igual a 0.02269, lo que indica que se puede rechazar la hipótesis nula en ese caso específico.

Familia Come carne Pescado y huevos

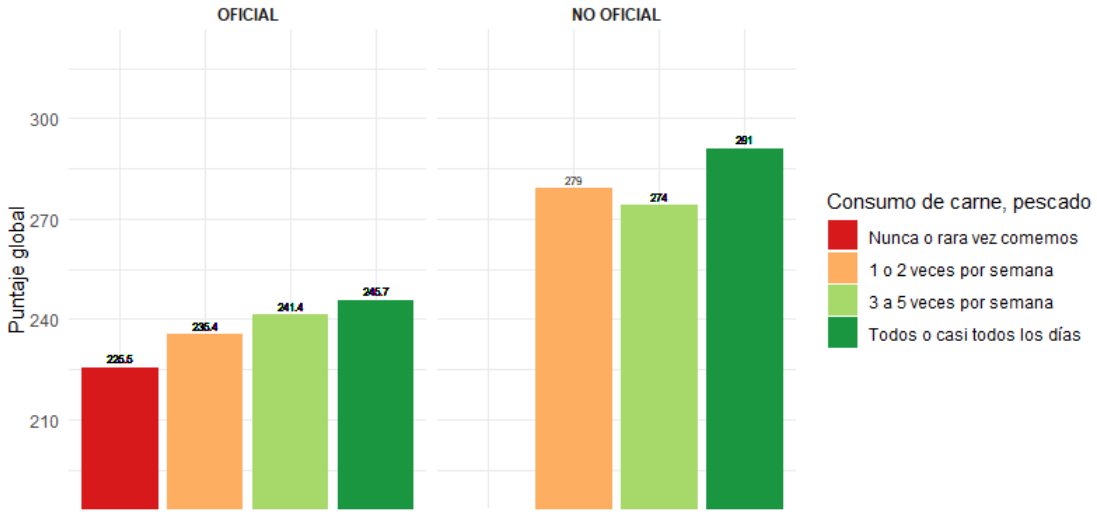

Figura 4-52: Familia come carne, percado y huevos . (Elaboración propia)

El la Figura **4-52**, se puede observar que los estudiantes de colegios no oficiales para esta subregión manifiestan consumir proteina minimo una o dos veces por semana. Al analizar la situación en los colegios oficiales, se observa que, a pesar de que algunos estudiantes de colegios oficiales afirman consumir carne diariamente, no logran obtener los mismos resultados en el puntaje global que los estudiantes de colegios no oficiales.

Mejores colegios de la subregión

Los mejores colegios públicos de la región son: I. E. Anza en Anza, I. E. Rosa Mesa De Mejia en Armenia, I. E. San Luis Gonzaga en Santafé De Antioquia.

4.3.6. Subregión Oriente

La subregión del Oriente Antioqueno es una de las 9 subregiones del departamento de Antioquia, Colombia, que está conformada por los 23 municipios de Abejorral, Alejandría, Argelia, Cocorná, Concepción, El Carmen de Viboral, El Peñol, El Retiro, El Santuario, Granada, Guarne, Guatapé, La Ceja, La Unión, Marinilla, Nariño, Rionegro, San Carlos, San Francisco, San Luis, San Rafael, San Vicente Ferrer y Sonsón. Según el Anuario Estadístico de Antioquia de 2019, la población de la subregión es de 1 079 302 habitantes y su tasa de analfabetismo es del 3,2 %.

En cuanto a la educación, la cobertura de la primaria en la subregión del Oriente Antioqueno es del 95,8 %, mientras que la cobertura de educación secundaria es del 64,7 %, según "El informe sobre el estado y evolución del desarrollo sostenible en Antioquia" (FOLU Antioquia, 2019). Además, la subregión cuenta con una variedad de instituciones de educación superior, incluyendo la Universidad Nacional de Colombia sede Medellín, la Universidad Católica de Oriente y la Universidad de Antioquia, entre otras.

EEn términos económicos, la subregión se destaca por su producción agroindustrial, en particular el cultivo de café, caña de azúcar, aguacate, cítricos, banano y flores, así como la producción de leche y carne. Según el Anuario Estadístico de Antioquia de 2019, el sector agropecuario y forestal es el principal generador de empleo en la subregión, seguido por el sector comercio y servicios.

En cuanto al acceso a internet, la subregión del Oriente Antioqueno presenta una tasa de penetración del 36,9 %, según el informe de FOLU Antioquia (2019). Sin embargo, existen diferencias significativas entre las áreas urbanas y rurales, con una menor penetración en las zonas rurales debido a la falta de infraestructura de telecomunicaciones.

Desafortunadamente, la subregión del Oriente Antioqueno ha sido afectada por el conflicto armado en Colombia, con la presencia de grupos armados ilegales y la ocurrencia de violencia y desplazamiento forzado de población en algunos municipios. De acuerdo con el informe de FOLU Antioquia, el municipio de San Carlos ha sido uno de los más afectados por la violencia en la subregión.

En síntesis, la subregión del Oriente Antioqueno es una región importante en términos de educación, economía y acceso a internet en el departamento de Antioquia, aunque aún enfrenta desafíos relacionados con el conflicto armado y la desigualdad entre áreas urbanas y rurales. A pesar de ello, la región ha experimentado un crecimiento significativo en los últimos años y continúa siendo un importante motor de desarrollo en el departamento.

Puntaje global en cada clúster

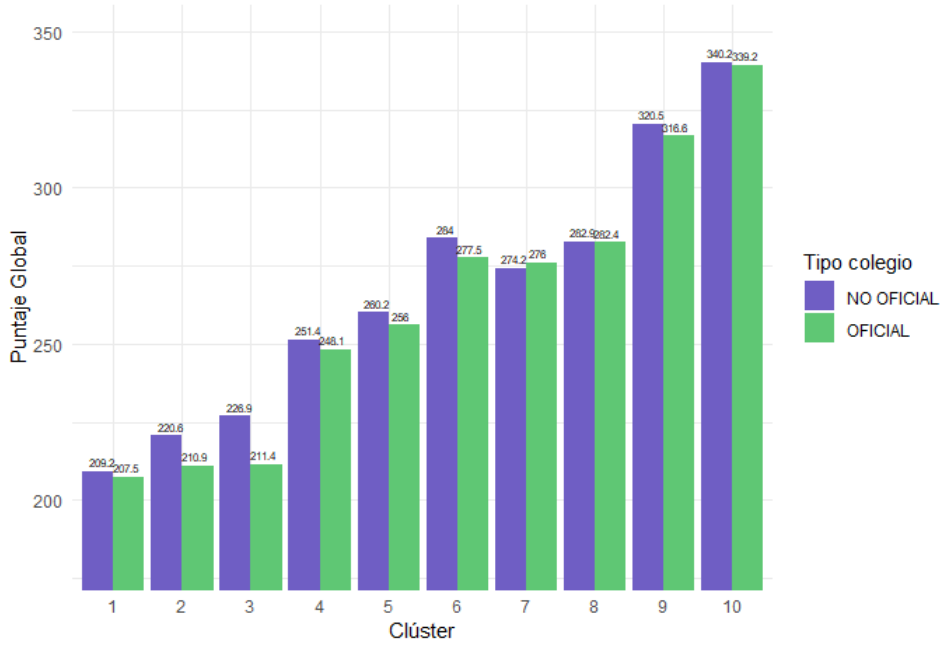

Figura 4-53: Clúster vs Puntaje Global Oriente (Elaboración propia)

En la Figura **4-53**, se puede observar que en el clúster 10 se encuentran los mejores promedios tanto para los colegios oficiales como para los no oficiales con puntajes de 340,2 y 339,2. Los menores puntajes se encuentran en el clúster 1, 2 y 3, en los que los colegios oficiales presentan un menor puntaje en estos clústeres.

Puntaje Lectura Crítica en cada clúster

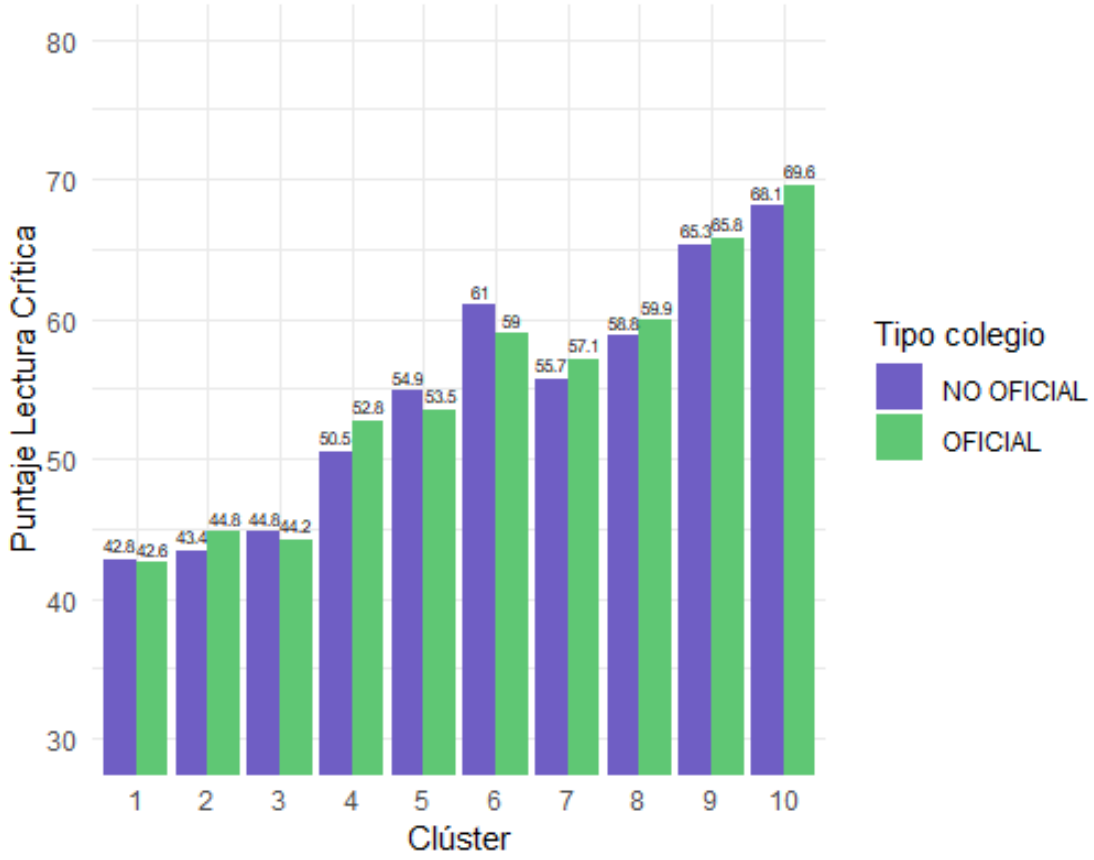

Figura 4-54: Clúster vs Puntaje Lectura crítica Oriente (Elaboración propia)

En la Figura **4-54** se puede observar que los mejores puntajes se encuentran en el clúster 10, es decir que los colegios oficiales en este clúster son los que tienen mejor promedio. Los clúster 1, 2, 3 son los que presentan menor promedio en el puntaje de lectura crítica. Los mejores promedios corresponden al clúster 8, 6, 5 y 2. El promedio en el puntaje más alto se logró en los colegios oficiales.

Familia come carne

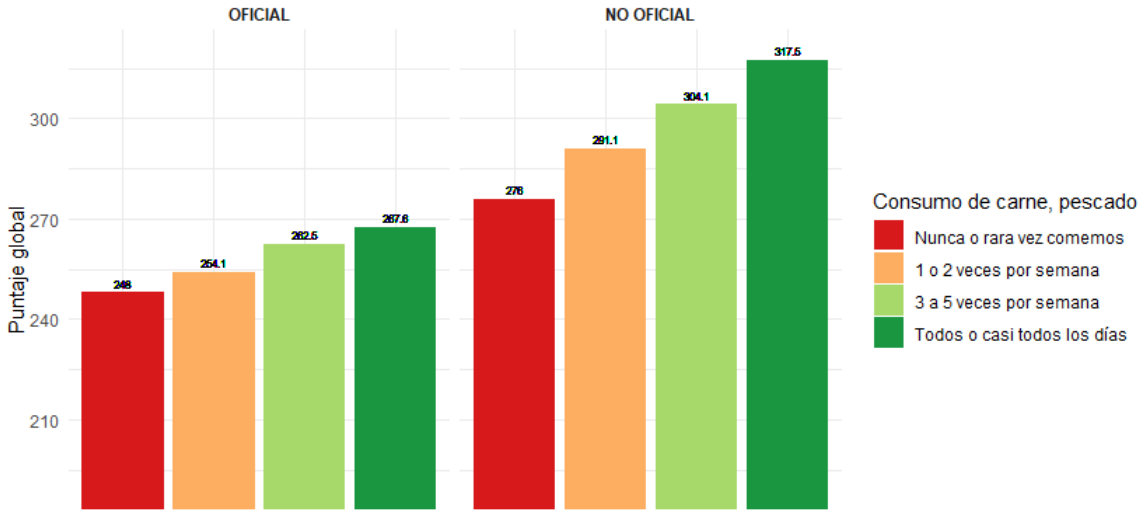

Figura 4-55: Familia come carne (Elaboración propia)

La figura **4-55** se observa que, aunque los estudiantes de colegios oficiales manifiesten comer carne todos los días, no tienen los mismos resultados que los de colegios no oficiales.

Mejores colegios de la subregión

Los mejores colegios públicos en la subregión son: I. E. R. San Miguel en Sonsón, I. E. Antonio Donado Camacho en Rionegro, I. E. Fray Julio Tobón B en El Carmen De Viboral.

4.3.7. Subregión Suroeste

La subregión Suroeste es una de las nueve subregiones del departamento de Antioquia, Colombia. Está conformada por 23 municipios, Amagá, Andes, Angelópolis, Betania, Betulia, Caramanta, Ciudad Bolívar, Concordia, Fredonia, Hispania, Jardín, Jericó, La Pintada, Montebello, Pueblorrico, Salgar, Santa Bárbara, Támesis, Tarso, Titiribí, Urrao, Valparaíso y Venecia. Esta subregión se caracteriza por su rica cultura, su belleza natural y su gente amable y trabajadora.

En cuanto a su economía, la subregión Suroeste se destaca por la producción de café, banano, aguacate, plátano, entre otros cultivos. Además, la subregión cuenta con una importante actividad ganadera y minera, especialmente en la explotación de oro. Según el Anuario Estadístico de Antioquia de 2019, el sector agropecuario es el principal generador de empleo en la subregión, seguido por el sector comercio y servicios.

En cuanto a su economía, la subregión Suroeste se destaca por la producción de café, banano, aguacate, plátano, entre otros cultivos. Además, la subregión cuenta con una importante actividad ganadera y minera, especialmente, la explotación de oro. Según el Anuario Estadístico de Antioquia de 2019, el sector agropecuario es el principal generador de empleo en la subregión, seguido por el sector comercio y servicios.

En términos de educación, la subregión cuenta con una amplia oferta educativa, desde la educación básica hasta la educación superior. El informe sobre el estado y evolución del desarrollo sostenible en Antioquia realizado (FOLU Antioquia, 2019) asevera que la cobertura de educación primaria en la subregión es del 98,6 %, mientras que la cobertura de educación secundaria es del 66,2 %. Además, la subregión cuenta con importantes instituciones de educación superior como la Universidad de Antioquia y la Universidad Católica de Manizales, entre otras.

En cuanto al acceso a internet, la subregión Suroeste presenta una tasa de penetración del 29,5 %, según el informe de FOLU Antioquia. Sin embargo, se han realizado importantes inversiones en infraestructura de telecomunicaciones para mejorar el acceso a internet en la región.
Desafortunadamente, la subregión Suroeste también ha sido afectada por el conflicto armado en Colombia, especialmente en la década de los años 90 y principios de los 2000. No obstante, gracias a los esfuerzos de las comunidades y del gobierno, la región ha logrado recuperarse y hoy en día es considerada como una de las zonas más seguras y tranquilas del departamento de Antioquia.

La subregión Suroeste es una región importante en términos de educación, economía y acceso a internet en el departamento de Antioquia, aunque aún enfrenta desafíos relacionados con el conflicto armado y la desigualdad en el acceso a servicios públicos. A pesar de ello, la región se destaca por su rica cultura, su belleza natural y continúa siendo un importante motor de desarrollo en el departamento.

Puntaje global en cada Clúster

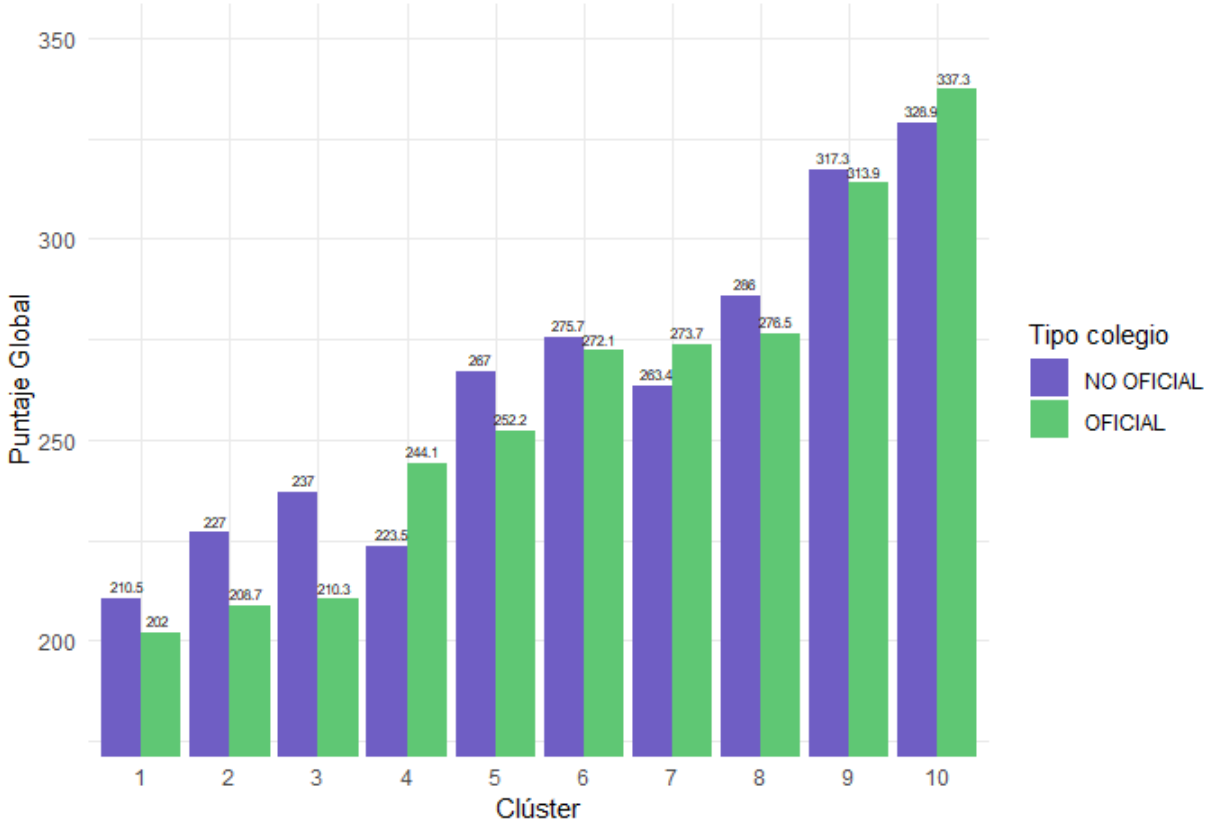

Figura 4-56: Clúster vs Puntaje Global Suroeste (Elaboración propia)

En la Figura **4-56** se evidencia que el mayor promedio en el puntaje global se logró en los colegios oficiales en el clúster 10, mientras que en los clúster 1, 2 y 3, los colegios no oficiales obtienen mejor promedio que los colegios oficiales.

Puntaje en Lectura Crítica en cada Clúster

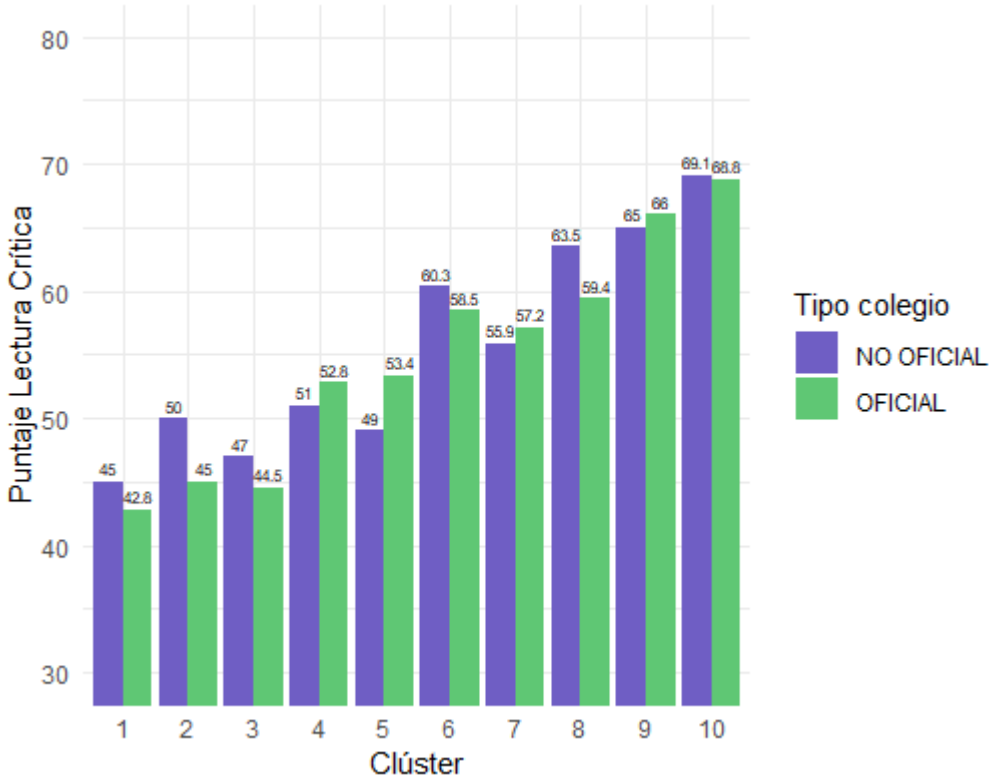

Figura 4-57: Clúster vs Puntaje Lectura crítica Suroeste (Elaboración propia)

En la figura **4-57**, se puede observar que en el clúster 1, 2, 3, 6, 8 y 10, los colegios no oficiales obtienen mejor promedio en el puntaje de lectura crítica. Los promedios más bajos se encuentran en el clúster 1, 3, 2, 4 y 5.

Familia come Carne, pescado y huevos vs puntaje global

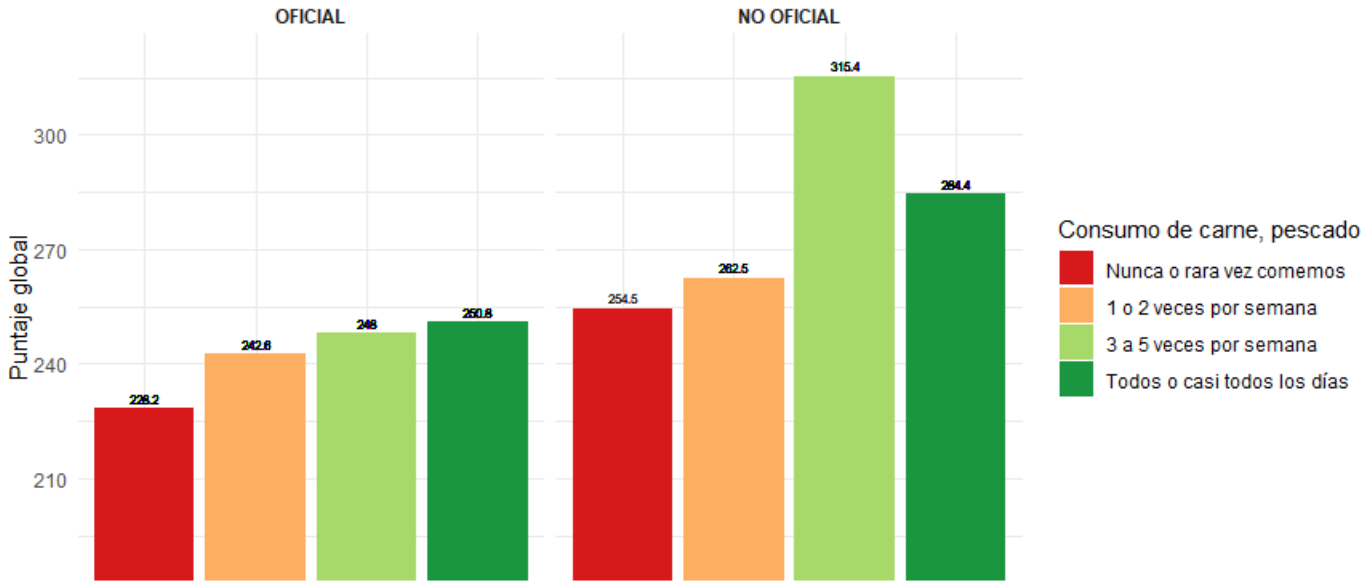

Figura 4-58: Familia come carne vs puntaje global)

En la Figura **4-58**, se observan puntajes muy cercanos en todas las categorías de consumo de carne, pescado y huevos en los colegios oficiales. Sin embargo, en los colegios no oficiales se puede apreciar que los mejores promedios se encuentran en los estudiantes que manifiestan consumir carne, pescado y huevos de 3 a 5 veces por semana. Estos estudiantes obtienen un rendimiento académico destacado en comparación con aquellos que consumen menos frecuentemente estos alimentos.

Mejores colegios de la subregión

Los mejores colegios públicos son: I. E. San Antonio, Jardín, I. E. Escuela Normal Superior Sagrada Familia, Urrao, I. E. Escuela Normal Superior Amaga, Amaga.

4.3.8. Subregión Urabá

La subregión de Urabá es una de las nueve subregiones que conforman el departamento de Antioquia, Colombia. Está ubicada en el extremo noroeste y se compone de 11 municipios (Apartadó, Arboletes, Carepa, Chigorodó, Murindo, Mutatá, Necoclí, San Juan de Urabá, San Pedro de Urabá, Turbo y Vigía del Fuerte).

En cuanto a la economía de la subregión, se destaca por ser una región con gran potencial agroindustrial, con cultivos de banano, plátano, cacao, palma africana y frutas tropicales. Además, la actividad pesquera y la producción de ganado bovino son importantes para la economía local. Según el Anuario Estadístico de Antioquia de 2019, el sector agropecuario representa el 27,3 % del PIB de la subregión.

En materia de educación, la subregión cuenta con una amplia oferta educativa, desde la educación básica hasta la educación superior. Según .[El] informe sobre el estado y evolución del desarrollo sostenible en Antioquia" realizado por FOLU Antioquia en 2021, la cobertura de educación primaria en la subregión es del 95 %, mientras que la cobertura de educación secundaria es del 63 %. Además, la subregión cuenta con importantes instituciones de educación superior como la Universidad de Antioquia y la Universidad de Córdoba, entre otras.

Respecto a la salud, la subregión cuenta con una buena oferta de servicios de salud, con hospitales y clínicas en las principales ciudades de la región. Según el Anuario Estadístico de Antioquia de 2019, la subregión cuenta con una tasa de mortalidad infantil de 10,3 por cada mil nacidos vivos, lo que la ubica por debajo del promedio nacional.

La infraestructura vial de la subregión cuenta con una amplia red de carreteras que la conectan con el resto del departamento y el país. Sin embargo, según el informe de FOLU Antioquia (2021), existen aún importantes desafíos en términos de mantenimiento y construcción de nuevas vías, especialmente en las zonas rurales. En cuanto al acceso a internet, la subregión de Urabá presenta una tasa de penetración del 32,2 % (FOLU Antioquia, 2021). Sin embargo, aún existen importantes brechas en el acceso a internet en las zonas rurales y en las comunidades más vulnerables.

La subregión de Urabá es una región con gran potencial agroindustrial y con una amplia oferta educativa y de servicios de salud. Con todo, aún persisten importantes desafíos en materia de infraestructura vial y acceso a internet, especialmente en las zonas rurales. A pesar de ello, la región se destaca por su belleza natural y continúa siendo un importante motor de desarrollo en el departamento de Antioquia.

Puntaje Global por Clúster

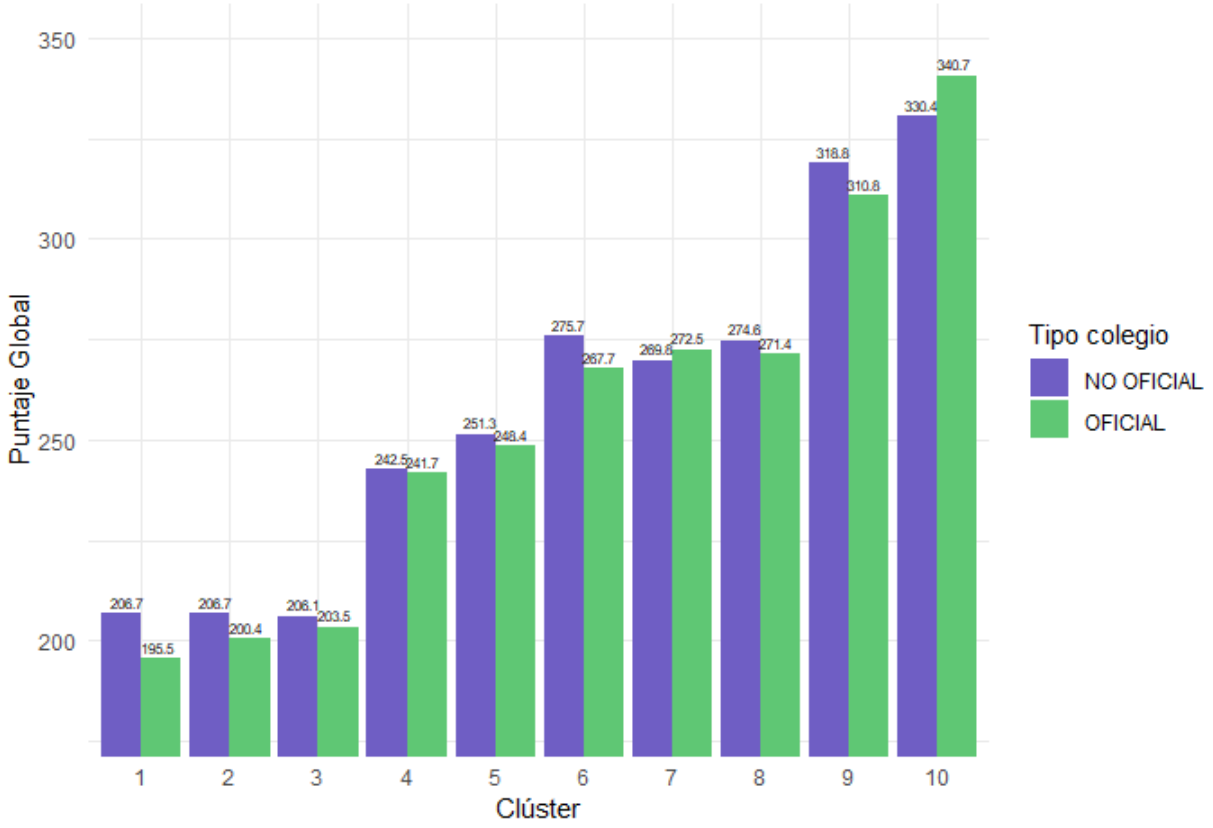

Figura 4-59: Clúster vs Puntaje Global Urabá (Elaboración propia)

En la Figura **4-59** Se evidencia que el mejor promedio en el puntaje global fue logrado por los colegios oficiales con 340.7 puntos y se encuentra en el clúster 10, el menor puntaje para colegios oficiales se encuentra en el clúster 1 con 195.5 puntos.

Puntaje Lectura por Clúster

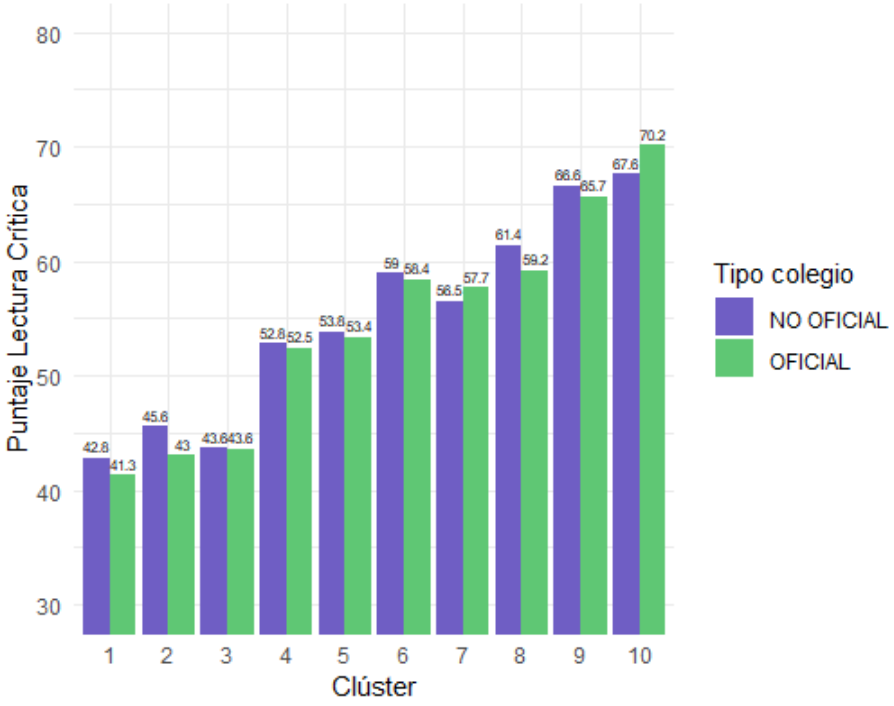

Figura 4-60: Clúster vs Puntaje Lectura crítica Uraba (Elaboración propia)

En la Figura **4-60** se observa que el mejor promedio en lectura crítica se encuentra en el clúster 10 y corresponde al promedio de colegios oficiales. Los menores promedios se encuentran en el clúster 1, 2 y 3.

Familia consume Proteina

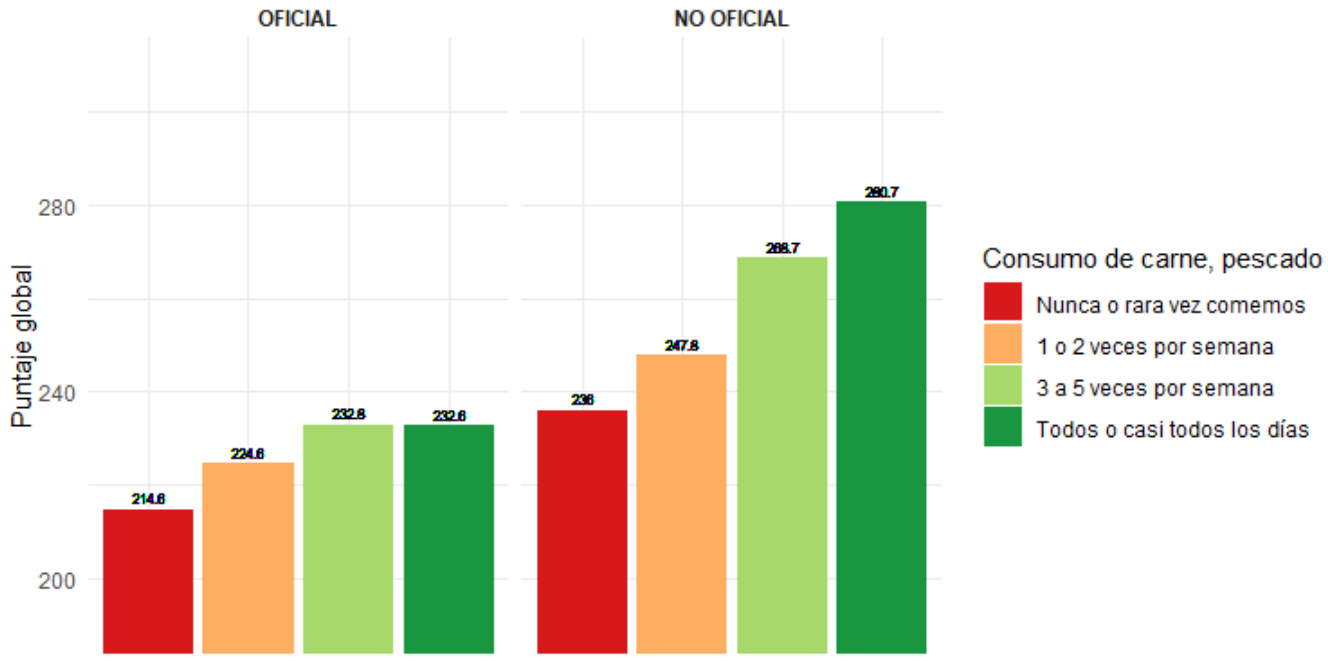

Figura 4-61: Familia come proteina Urabá (Elaboración propia)

En la Figura **4-61** se muestra que los estudiantes de colegios no oficiales que consumen carne, pescado y huevo a diario obtienen un promedio de puntaje de 280,7, mientras que en los colegios oficiales el promedio es de 232,6. Estos resultados sugieren que el consumo regular de carne, pescado y huevo puede tener un impacto positivo en el rendimiento académico de los estudiantes de colegios no oficiales. Sin embargo, es importante tener en cuenta que estos datos se basan en promedios y que existen otros factores que pueden influir en el rendimiento académico de los estudiantes.

Mejores colegios de la subregión

Los mejoress colegios públicos del la subregión son: I. E. Los Andes , Chigorodó I. E. Luis Carlos Galan Sarmiento,Carepa I.E. Cadena Las Playas , Apartadó

4.3.9. Subregión Valle de Aburrá

El Valle de Aburrá es una subregión del departamento de Antioquia, en Colombia, que se compone de 10 municipios: Medellín, Barbosa, Bello, Caldas, Copacabana, Envigado, Girardota, Itagüí, La Estrella y Sabaneta. Esta subregión es la más densamente poblada de Antioquia y una de las más importantes en términos económicos y culturales de todo el país.

En cuanto a la economía, el Valle de Aburra es una región altamente industrializada y comercial, con un importante sector manufacturero que incluye textiles, alimentos, productos químicos y maquinaria. Además, la subregión es el centro financiero de Antioquia y cuenta con una amplia oferta de servicios, incluyendo turismo, tecnología y salud. Según el Anuario Estadístico de Antioquia de 2019, el sector servicios representa el 73,5 % del PIB de la subregión.

Respecto a la educación, la subregión cuenta con una amplia oferta educativa, desde la educación básica hasta la educación superior. Según la FOLU Antioquia (2021), la cobertura de educación primaria en la subregión es del 99 %, mientras que la cobertura de educación secundaria es del 82 %. Además, la subregión cuenta con importantes instituciones de educación superior como la Universidad de Antioquia, la Universidad Nacional y la Universidad EAFIT, entre otras.

Por otro lado, la subregión cuenta con una buena oferta de servicios de salud, con hospitales y clínicas en las principales ciudades de la región. Según el Anuario Estadístico de Antioquia de 2019, el Valle de Aburrá cuenta con una tasa de mortalidad infantil de 10,6 por cada mil nacidos vivos, lo que la ubica por debajo del promedio nacional.

En acceso a internet, la subregión del Valle de Aburra presenta una tasa de penetración del 56,7 %, según el informe de FOLU Antioquia (2021). Si bien esta tasa es alta en comparación con otras subregiones de Antioquia, aún existen importantes brechas en el acceso a internet en las zonas rurales y en las comunidades más vulnerables.

Puntaje Global por Clúster

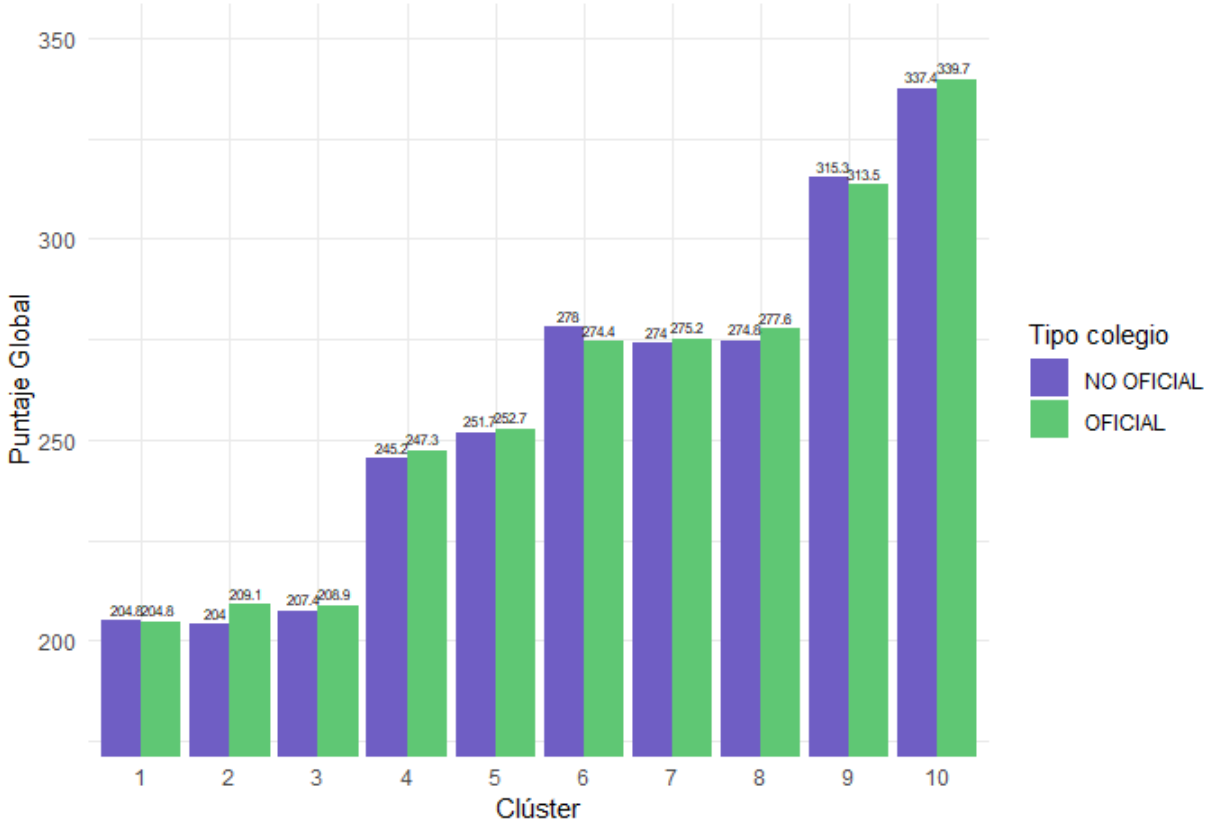

Figura 4-62: Cluster vs Puntaje Global Valle de Aburrá (Elaboración propia)

En la Figura **4-62** se observa que los menores puntajes se encuentran en el clúster 1, 2 y 3, y los mejores puntajes se encuentran en el clúster 9 y 10 tanto para los colegios oficiales como no oficiales.

Prueba de medias con ajuste Bonferroni

- Hipótesis nula (H0): No hay diferencia significativa en el puntaje GLOBAL entre los colegios oficiales y no oficiales.

- Hipótesis alternativa (H1): Existe una diferencia significativa en el puntaje GLOBAL entre los colegios oficiales y no oficiales.

La evaluación de los resultados por clúster encontró lo siguiente:

- Clúster 1: no se puede rechazar la hipótesis nula, lo que indica que no hay una diferencia significativa en el puntaje global entre colegios oficiales y no oficiales en este clúster.

- Clúster 2: se rechaza la hipótesis nula, lo que sugiere que existe una diferencia significativa en el puntaje global entre ambos tipos de colegios en este clúster.
- Clústeres 3, 5, 7, 8, y 10: no se puede rechazar la hipótesis nula, lo que indica que no hay una diferencia significativa en el puntaje global entre colegios oficiales y no oficiales en estos clústeres.
- Clústeres 4, 6, y 9: se rechaza la hipótesis nula, lo que sugiere que existe una diferencia significativa en el puntaje global entre ambos tipos de colegios en estos clústeres.

Los resultados muestran que la diferencia en el puntaje global entre colegios oficiales y no oficiales varía dependiendo del clúster en la subregión del Valle de Aburra. Es importante destacar que en algunos clústeres se encontró una diferencia significativa, mientras que en otros no se encontraron diferencias. Esto sugiere que factores adicionales, como las características socioeconómicas o culturales de cada clúster, pueden influir en los resultados académicos de los estudiantes.

Puntaje Lectura Crítica por Clúster

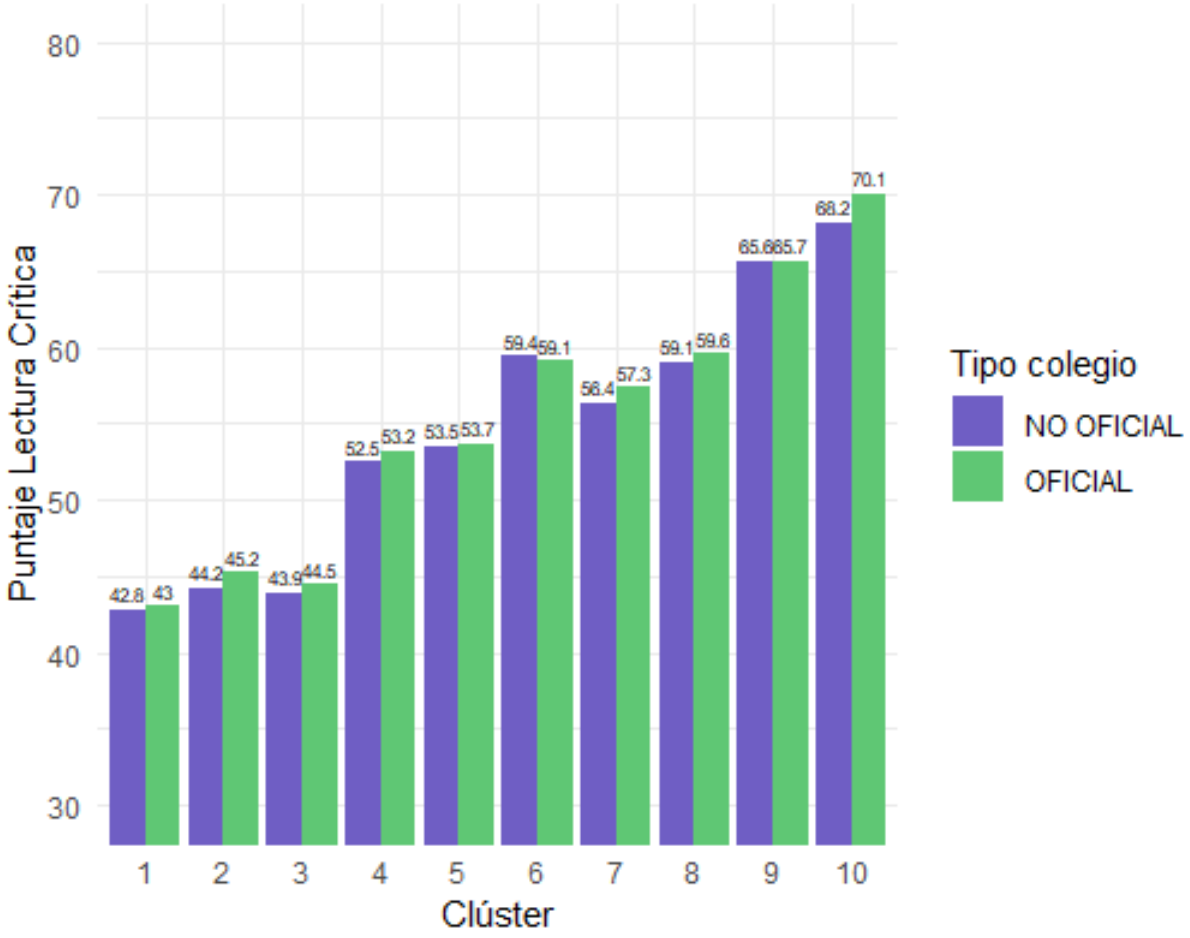

Figura 4-63: Cluster vs Puntaje Lectura crítica Valle de Aburrá (Elaboración propia)

En la Figura **4-63** Se observa que el mejor promedio en el puntaje de lectura crítica se encuentra en el clúster 10 tanto para colegios oficiales como no oficiales, en el clúster 10 el mejor promedio es 70,1 y corresponde a los colegios oficiales.

Prueba de medias con ajuste Bonferroni

Al interpretar los resultados obtenidos para el puntaje de Lectura Crítica en los colegios oficiales y no oficiales, se planteó una hipótesis nula que afirmaba que no existía una diferencia significativa entre ambos tipos de colegios y una alternativa que sostenía lo contrario.
Al analizar los resultados por clúster, se encontraron las siguientes conclusiones:

- Clústeres 1, 3, 5, 6, 8, y 9: no se puede rechazar la hipótesis nula, lo que indica que no se encontró una diferencia significativa en el puntaje de lectura crítica entre los colegios oficiales y no oficiales en estos clústeres.
- Clústeres 2, 4, 7, y 10: se rechaza la hipótesis nula, lo que sugiere que sí existe una diferencia significativa en el puntaje de lectura crítica entre ambos tipos de colegios en estos clústeres.

Los resultados indican que la presencia de una diferencia significativa en el puntaje de lectura crítica entre los colegios oficiales y no oficiales varía según el clúster en el que se encuentren. Es importante tener en cuenta que en algunos clústeres se encontró una diferencia significativa, mientras que en otros no se observaron diferencias. Estos hallazgos pueden señalar la influencia de factores adicionales –como el entorno socioeconómico o cultural de cada clúster– en el desempeño académico de los estudiantes en lectura crítica.

Consumo de carne, pescado y huevos en la familia

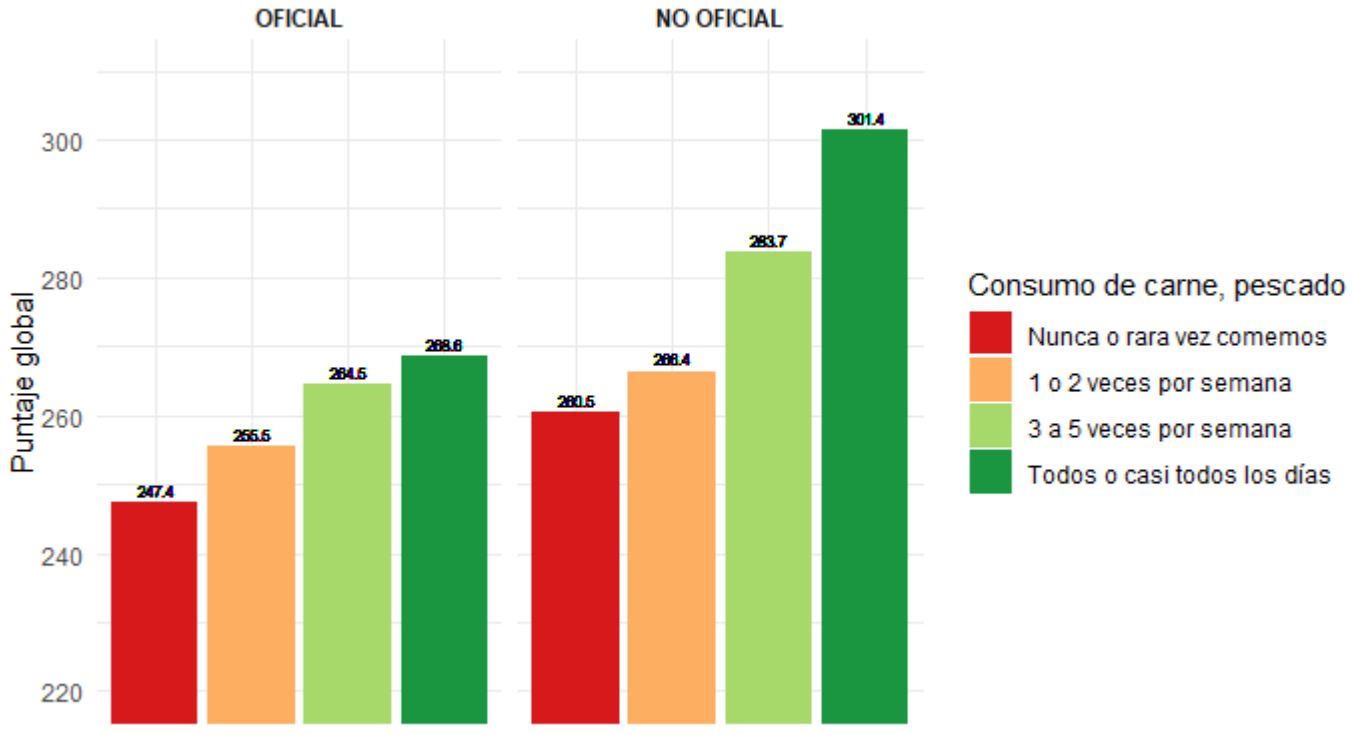

Figura 4-64: Familia tiene moto Valle de Aburrá (Elaboración propia)

En la Figura **4-64** se puede observar que los estudiantes que consumen proteína todos los días tienen los promedios más altos, con valores de 268,6 para los colegios oficiales y 301,4 para los colegios no oficiales. Por otro lado, los estudiantes que rara vez o nunca consumen proteína muestran los promedios más bajos. Esta información sugiere una posible relación entre el consumo de proteína y el rendimiento académico de los estudiantes. Es importante tener en cuenta que estos resultados se basan en promedios y pueden variar en función de otros factores relacionados con la alimentación y el entorno educativo.

Mejores colegios de la subregión

Los mejores Colegios públicos son: I. E. Escuela Normal Superior María Auxiliadora , Copacabana, I.E. Colegio Loyola Para La Ciencia Y La Innovación, Medellín, Ie Nueva Generación , Bello .

5 Pruebas de Homogeneidad

En este capítulo, se abordan las pruebas de homogeneidad, las cuales son de vital importancia en el estudio de variables relevantes. Para llevar a cabo estas pruebas, se requirió realizar una transformación de la variable de interés, en este caso, el puntaje global, en términos de cuartiles. Es importante tener en cuenta que, al aplicar pruebas de homogeneidad con variables categóricas, la variable de respuesta también debe ser categórica, de modo que pueda ser comparada y analizada de manera adecuada.

El proceso de trabajar el puntaje global en cuartiles nos permite clasificar los datos en diferentes categorías, lo cual facilita la comparación de las distribuciones de frecuencia entre los grupos considerados. Esto nos permite evaluar si existen diferencias significativas en las proporciones de cada cuartil entre los grupos en estudio.

Al utilizar pruebas de homogeneidad, se busca determinar si las diferencias observadas en las frecuencias de las categorías son estadísticamente significativas o si pueden ser atribuidas al azar. Para ello, se establecen hipótesis nulas y alternativas, y se utiliza un estadístico de prueba, como el chi-cuadrado, para comparar las frecuencias observadas con las esperadas bajo la hipótesis nula. Si el valor del estadístico de prueba supera un umbral crítico, se rechaza la hipótesis nula, indicando la existencia de una diferencia significativa en las distribuciones de las categorías.

Las pruebas de homogeneidad son una herramienta valiosa en el análisis estadístico de variables categóricas. A través de ellas, podemos examinar y comparar las frecuencias de diferentes categorías en distintos grupos, lo que nos brinda información relevante para comprender y tomar decisiones basadas en los patrones de comportamiento o características de interés en las distintas subpoblaciones.

5.1. Prueba de homogeneidad educación de la Madre

Hipótesis para homegeneidad marginal,

$$H_0 : \pi_{(1+)} = \pi_{(+1)}; \pi_{(2+)} = \pi_{(+2)}; \pi_{(3+)} = \pi_{(+3)}; \pi_{(4+)} = \pi_{(+4)},$$

o equivalentemente

$$H_0 : \pi_{(1+)} - \pi_{(+1)} = 0; \pi_{(2+)} - \pi_{(+2)} = 0; \pi_{(3+)} - \pi_{(+3)} = 0; \pi_{(4+)} - \pi_{(+4)} = 0.$$

es decir:

$$H_0 : \pi_{(puntajesinestudios)} = \pi_{(puntajeconbachillerato)} = \pi_{(puntajeprofesional)} = \pi_{(puntajeconposgrado)}$$

Lo anterior mostramos se puede expresar en forma

$$H_0 : A_\pi = 0$$

Notar que

$$\pi_1 + = \pi + 1$$

implica

$$\pi_{(1+)} = \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21} - \pi_{31} - \pi_{41}$$

$$\boldsymbol{A}_{4\times 16} = \begin{pmatrix} 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 1 & 1 & 1 & 0 \end{pmatrix} \tag{5-1}$$

$$\boldsymbol{\pi}_{16\times 1} = \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{14} \\ \pi_{21} \\ \vdots \\ \pi_{44} \end{pmatrix}$$

La matriz A (5-1) da la hipótesis correcta pero es singular ya que la suma de las dos primeras filas es igual al negativo de las últimas dos, entonces consideramos A^* que es igual a A pero con una fila menos. Por lo tanto

$$X^2 = \hat{\pi} A^{*T} \left(A^* \sum_{\hat{\pi}} A^{*T} \right)^{-1} A^* \hat{\pi} = f^{*T} \sum_{f^*} f^* \tag{5-2}$$

Para esto, se tomaron solo los datos de estudios completos en cada categoria, es decir el nivel de estudio terminado.

Tabla 5-1: Educación de la Madre

Clúster	**(126, 227]**	**(227, 263]**	**(263, 300]**	**(300, 473]**
Ninguno	757	374	218	73
Bachiller_completo	7649	8365	8190	6257
Profesional_completo	1515	2192	3130	5888
Posgrado	161	293	635	2442

Para realizar la prueba de homogeneidad en relación con la variable respuesta –que en este caso es el puntaje global– fue necesario transformar dicha variable en términos de cuartiles. En la Tabla **5-1**, se presentan los datos de formación de la madre que se utilizaron para llevar a cabo la prueba de homogeneidad. Estos datos están agrupados según los cuartiles del puntaje global,Los resultados obtenidos de la prueba son los mostrados en la Tabla **5-2**.

Tabla 5-2: Resultados de la Prueba

Prueba	**Resultados**
Chicuadrada calculada:	28661.3435530454
Grados de libertad:	9
Valor-p:	0

Utilizando la función de GSK para probar homogeneidad, se obtiene un valor p de 0 y por lo tanto se concluye que el nivel educativo de la madre influye en el puntaje global de los estudiantes.

5.2. Prueba de homogeneidad para lectura

A continuación se presenta la prueba de homogeneidad.

$$H_0 : \pi_{(1+)} = \pi_{(+1)}; \pi_{(2+)} = \pi_{(+2)}; \pi_{(3+)} = \pi_{(+3)}; \pi_{(4+)} = \pi_{(+4)}$$

o equivalente

$$H_0 : \pi_{(1+)} - \pi_{(+1)} = 0; \pi_{(2+)} - \pi_{(+2)} = 0; \pi_{(3+)} - \pi_{(+3)} = 0; \pi_{(4+)} - \pi_{(+4)} = 0$$

Es decir:

$$H_o : \pi_{Puntaje\ Lectura\ no\ lee} = \pi_{Puntaje\ Lectura\ lee\ 30a60min}$$

$$= \pi_{Puntaje\ Lectura\ lee\ entre1hry2hrs} = \pi_{Puntaje\ Lectura\ lee\ más\ de\ 2hrs}$$

Lo anterior se puede expresar en forma

$$H_0 : A_\pi = 0$$

$$\pi_1 + = \pi + 1$$

implica $\pi_{1+} = \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21} - \pi_{31} - \pi_{41}$

A continuación se presenta la Tabla **5-3** relacionada con dedicación a la lectura.

Tabla 5-3: Dedicación a la lectura

Clúster	**(0, 49]**	**(49, 56]**	**(56, 63]**	**(63, 100]**
Entre 1 y 2 horas	2714	3043	3672	3966
Entre 30 y 60 minutos	6696	7715	8708	8451
Más de 2 horas	732	926	1443	1679
No. Leo	7195	5760	4824	3259

Tabla 5-4: Resultados de la Prueba

En la Tabla **5-4** se puede evidenciar que el **Prueba**	**Resultados**
Chicuadrada calculada:	14603.5456234867
Grados de libertad:	9
Valor-p:	0

Utilizando la función de GSK para probar homogeneidad se obtiene un valor p de 0, cómo se muestra en **5-4** y por lo tanto se concluye que el total de horas dedicadas a lectura influye en el nivel de lectura.

5.3. Prueba de homogeneidad Come leche y derivados

$$H_0 : \pi_{(1+)} = \pi_{(+1)}; \pi_{(2+)} = \pi_{(+2)}; \pi_{(3+)} = \pi_{(+3)}; \pi_{(4+)} = \pi_{(+4)}$$

o equivalente

$$H_0 : \pi_{(1+)} - \pi_{(+1)} = 0; \pi_{(2+)} - \pi_{(+2)} = 0; \pi_{(3+)} - \pi_{(+3)} = 0; \pi_{(4+)} - \pi_{(+4)} = 0$$

Es decir:

$$H_o : \pi_{Puntaje\ no\ come\ lecheyderivados} = \pi_{Puntaje\ come\ 1a2\ vecesporsemana}$$

$$= \pi_{Puntaje\ Come\ 3a5\ vecesporsemana} = \pi_{Puntaje\ Come\ Todos\ o\ casi\ todos\ los\ días}$$

Lo anterior se puede expresar en forma $H_o : A_\pi = 0$. Notar que $\pi_{1+} = \pi + 1$ implica $\pi_{1+} = \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21} - \pi_{31} - \pi_{41}$

Tabla 5-5: Come leche y derivados

Clúster	**(126, 227]**	**(227, 263]**	**(263, 300]**	**(300, 473]**
Nunca o rara vez comemos	2664	1649	1030	560
1 o 2 veces por semana	10115	8686	6686	3770
3 a 5 veces por semana	6503	7646	7547	6403
Todos o casi todos los días	8702	9829	12010	15889

En la Tabla **5-5**, se muestran los datos utilizados para realizar la prueba, clasificados por cuartiles en función de los grupos que indicaron su consumo de leche y sus derivados, así como la frecuencia de consumo.

Tabla 5-6: Resultados de la Prueba

Prueba	**Resultados**
Chicuadrada calculada:	21369.2109487678
Grados de libertad:	9
Valor-p:	0

El la Tabla **5-6**, se muestran los resultados de la prueba, se puede observar que utilizando la función de GSK para probar homogeneidad se obtiene un valor p de 0 y por lo tanto se concluye que comer leche y derivados influye en el puntaje global de los estudiantes.

5.4. Prueba de homogeneidad Come carne pescado y huevos

$$H_0 : \pi_{(1+)} = \pi_{(+1)}; \pi_{(2+)} = \pi_{(+2)}; \pi_{(3+)} = \pi_{(+3)}; \pi_{(4+)} = \pi_{(+4)}$$

o equivalente

$$H_0 : \pi_{(1+)} - \pi_{(+1)} = 0; \pi_{(2+)} - \pi_{(+2)} = 0; \pi_{(3+)} - \pi_{(+3)} = 0; \pi_{(4+)} - \pi_{(+4)} = 0$$

Es decir:

$$H_o : \pi_{Puntaje\ no\ come\ carne\ pescado\ y\ huevos} = \pi_{Puntaje\ come\ 1a2\ vecesporsemana}$$

$$= \pi_{Puntaje\ Come\ 3a5\ vecesporsemana} = \pi_{Puntaje\ Come\ Todos\ o\ casi\ todos\ los\ días}$$

Lo anterior se puede expresar en forma
$H_o : A_\pi = 0$. Notar que $\pi_{1+} = \pi + 1$ implica $\pi_{1+} = \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21} - \pi_{31} - \pi_{41}$

Tabla 5-7: Come carne pescado y huevos

Clúster	**(126, 227]**	**(227, 263]**	**(263, 300]**	**(300, 473]**
Nunca o rara vez comemos	1656	911	678	412
1 o 2 veces por semana	7652	6427	5083	2915
3 a 5 veces por semana	8269	9033	8598	7086
Todos o casi todos los días	9912	11439	12914	16209

En la Tabla **5-7**, se muestran los datos utilizados para realizar la prueba, clasificados por cuartiles en función de los grupos que indicaron su consumo de carne, pescado y huevos así como la frecuencia de consumo.

Tabla 5-8: Resultados de la Prueba

Prueba	**Resultados**
Chicuadrada calculada:	31541.8556299037
Grados de libertad:	9
Valor-p:	0

El la Tabla **5-8**, se muestran los resultados de la prueba, se puede observar que Utilizando la función de GSK para probar homogeneidad se obtiene un valor p de 0 y por lo tanto se concluye que comer carne, pescado y huevo influye en el puntaje global de los estudiantes.

5.5. Prueba homogeneidad Come cereal frutos legumbres

$$H_0 : \pi_{(1+)} = \pi_{(+1)}; \pi_{(2+)} = \pi_{(+2)}; \pi_{(3+)} = \pi_{(+3)}; \pi_{(4+)} = \pi_{(+4)}$$

o equivalente

$$H_0 : \pi_{(1+)} - \pi_{(+1)} = 0; \pi_{(2+)} - \pi_{(+2)} = 0; \pi_{(3+)} - \pi_{(+3)} = 0; \pi_{(4+)} - \pi_{(+4)} = 0$$

Es decir:

$$H_o : \pi_{Puntaje\ no\ come\ frutos\ lejumbres} = \pi_{Puntaje\ come\ 1a2\ veces\ por\ semana}$$

$$= \pi_{Puntaje\ Come\ 3a5\ vecesporsemana} = \pi_{Puntaje\ Come\ Todos\ o\ casi\ todos\ los\ días}$$

Lo anterior se puede expresar en forma $H_o : A_\pi = 0$. Notar que $\pi_{1+} = \pi + 1$ implica $\pi_{1+} = \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21} - \pi_{31} - \pi_{41}$

Tabla 5-9: Come Cereales, frutos, legumbres

Clúster	**(126, 227]**	**(227, 263]**	**(263, 300]**	**(300, 473]**
Nunca o rara vez comemos	4021	3109	2409	1434
1 o 2 veces por semana	11239	10786	9711	7434
3 a 5 veces por semana	7459	8914	9530	10113
Todos o casi todos los días	4770	5001	5623	7641

En la Tabla **5-9**, se muestran los datos utilizados para realizar la prueba, clasificados por cuartiles en función de los grupos que indicaron su consumo de cereales, frutos y legunbres así como la frecuencia de consumo.

Tabla 5-10: Resultados de la Prueba

Prueba	**Resultados**
Chicuadrada calculada:	10462.7158160804
Grados de libertad:	9
Valor-p:	0

El la Tabla **5-10**, se muestran los resultados de la prueba, se puede observar que utilizando la función de GSK para probar homogeneidad se obtiene un valor p de 0 y por lo tanto se concluye que comer cereal, frutos y lejumbres, influye en el puntaje global de los estudiantes.

5.6. Prueba Familia tiene Moto

A continuación se plantean las pruebas de hipótesis relacionadas con los estudiantes que manifiestan tener motocicleta en su familia y los resultados en las pruebas SABER, ubicados por cuartiles.

Ho: Las proporciones de los resultados de la prueba en los cuatro niveles son iguales tanto para las personas con moto como para las personas sin moto.

Ha: Las proporciones de los resultados de la prueba en los cuatro niveles son diferentes para las personas con moto y para las personas sin moto.

$$H_0 : \pi_{(1+)} = \pi_{(+1)}; \pi_{(2+)} = \pi_{(+2)}$$

o equivalente

$$H_0 : \pi_{(1+)} - \pi_{(+1)} = 0; \pi_{(2+)} - \pi_{(+2)} = 0$$

Es decir:

$$H_o : \pi_{Puntaje\ no\ familia\ tiene\ moto} = \pi_{Puntaje\ tiene\ moto}$$

Lo anterior se puede expresar en forma $H_o : A_\pi = 0$. Notar que $\pi_{1+} = \pi + 1$ implica $\pi_{1+} = \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21}$

Tabla 5-11: Familia tiene Moto

Clúster	**(126, 227]**	**(227, 263]**	**(263, 300]**	**(300, 473]**
No	15737	17033	17472	18697
Si	11752	10777	9801	7925

En la Tabla **5-11**, se muestran los datos utilizados para realizar la prueba, clasificados por cuartiles en función de los grupos que indicaron tener o no motocicleta.

Tabla 5-12: Según el test Pearson's

Prueba	**Resultados**
Chi-squared test X-squared:	1037.9
df:	3
p-value:	2.2e-16

Según la Tabla **5-12**, se puede concluir que se rechaza la hipótesis nula y se establece una asociación significativa entre las dos variables. Existe una relación significativa entre poseer una motocicleta o no y los resultados de la prueba, los cuales se agrupan en cuartiles.

Con el objetivo de validar los hallazgos, se calculará el coeficiente de contingencia (V de Cramer), el cual es una medida de asociación entre dos variables categóricas. El coeficiente de contingencia es una medida más robusta de la asociación entre las variables, brindando una perspectiva más sólida.

Tabla 5-13: Coeficiente de contingencia Familia tiene moto

	X^2	df	$P(X^2)$
Likelihood Ratio	1048.2	3	2.2e-16
Pearson	1037.9	3	0

Al observar los resultados en la Tabla **5-13** se sugiere que existe una asociación significativa entre dos variables categóricas que se están comparando y que la hipótesis nula de que no hay asociación entre las variables es poco probable. Para verificar lo encontrado se aplica el coeficiente de cramer, (tambien conocido como V de Cramer).

Tabla 5-14: Según el test Pearson's

Prueba	**Resultados**
Phi-Coefficient:	NA
Contingency Coeff:	0.097
Cramer's V :	0.097

Se puede observar en la Tabla **5-14** que Un coeficiente de contingencia y una V de Cramer de 0.097, puede sugerir que la asociación entre las dos variables categóricas es débil. Esto significa que tener una moto o no y los resultados de la prueba agrupados en cuartiles tiene una relación debil o poco significativa. Por esta razón en el capítulo anterior se presenta el comportamiento de esta variable en cada una de las regiones.

5.7. Prueba Familia tiene Automóvil

$$H_0 : \pi_{(1+)} = \pi_{(+1)}; \pi_{(2+)} = \pi_{(+2)}$$

o equivalente

$$H_0 : \pi_{(1+)} - \pi_{(+1)} = 0; \pi_{(2+)} - \pi_{(+2)} = 0$$

Es decir:

$$H_o : \pi_{Puntaje\ familia\ no\ tiene\ automovil} = \pi_{Puntaje\ familia\ tiene\ automovil}$$

Lo anterior se puede expresar en forma $H_o : A_\pi = 0$. Notar que $\pi_{1+} = \pi + 1$ implica $\pi_{1+} = \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21}$

Tabla 5-15: Familia tiene Automovil

Clúster	**(126, 227]**	**(227, 263]**	**(263, 300]**	**(300, 473]**
No	23050	22545	20292	15254
Si	4439	5265	6981	11368

En la Tabla **5-15**, se muestran los datos utilizados para realizar la prueba, clasificados por cuartiles en función de los grupos que indicaron tener o no automovil.

Tabla 5-16: Según el test Pearson's

Prueba	**Resultados**
Chi-squared test X-squared:	6011.9
df:	3
p-value:	2.2e-16

Cómo se observa en Tabla **5-16**, según el test como X-squared es muy alto y p-value muy pequeño esto puede indicar una fuerte asociacion entre dos variables y una evidencia estadísticamente significativa para rechazar la hipótesis nula de independencia.

Tabla 5-17: Coeficiente de contingencia Familia tiene Automovil

	X^2	df	$P(X^2)$
Likelihood Ratio	5779.5	3	0
Pearson	6011.9	3	0

Según los resultados mostrados en la Tabla **5-17** estos resultados sugieren que existe una asociación significativa entre la variable (Familia tiene Automóvil) y los resultados del puntaje global. El coeficiente de contingencia se utiliza para medir la fuerza de esta asociación. Para verificar lo encontrado se aplica el coeficiente de cramer, (tambien conocido como V de Cramer).

Tabla 5-18: Según el test Pearson's

Prueba	**Resultados**
Phi-Coefficient:	NA
Contingency Coeff:	0.228
Cramer's V :	0.235

Según los resultados de la prueba presentados en la Tabla **5-18**, se puede observar una asociación moderada entre las variables analizadas, aunque no necesariamente una asociación fuerte.

5.8. Prueba Familia tiene Computador

$$H_0 : \pi_{(1+)} = \pi_{(+1)}; \pi_{(2+)} = \pi_{(+2)}$$

o equivalente

$$H_0 : \pi_{(1+)} - \pi_{(+1)} = 0; \pi_{(2+)} - \pi_{(+2)} = 0$$

Es decir:

$$H_o : \pi_{Puntaje\ no\ familia\ tiene\ computador} = \pi_{Puntaje\ tiene\ computador}$$

Lo anterior se puede expresar en forma $H_o : A_\pi = 0$. Notar que $\pi_{1+} = \pi + 1$ implica $\pi_{1+} = \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21}$

Tabla 5-19: Familia tiene Computador

Clúster	**(126, 227]**	**(227, 263]**	**(263, 300]**	**(300, 473]**
No	12371	9665	6724	3469
Si	15118	18145	20549	23153

En la Tabla **5-19**, se muestran los datos utilizados para realizar la prueba, clasificados por cuartiles en función de los grupos que indicaron tener o no computador.

Tabla 5-20: Según el test Pearson's

Prueba	**Resultados**
Chi-squared test X-squared:	7323.7
df:	3
p-value:	2.2e-16

Según los datos mostrados en la En la Tabla **5-20**, los resultados obtenidos a partir del test de chi-cuadrado utilizando el método de Pearson indican una asociación altamente significativa entre las variables analizadas. Esto implica que existe una relación estadísticamente significativa entre las categorías evaluadas. El valor p extremadamente pequeño sugiere que las diferencias observadas entre los datos esperados y los datos observados no pueden ser atribuidas al azar. Estos hallazgos respaldan la hipótesis de que existe una relación sustancial entre las variables en estudio.

Tabla 5-21: Coeficiente de contingencia Familia tiene Computador

	X^2	df	$P(X^2)$
Likelihood Ratio	7670.1	3	0
Pearson	7323.7	3	0

Los resultados mostrados en la Tabla **5-21** presentan el coeficiente de contingencia para la variable (Familia tiene Computador) en relación con los resultados del puntaje global.El valor p es igual a 0, lo que indica una asociación significativa entre las variables. Estos resultados sugieren que existe una relación entre la presencia de un computador en la familia y los puntajes globales obtenidos.

Tabla 5-22: Según el test Pearson's

Prueba	**Resultados**
Phi-Coefficient:	NA
Contingency Coeff:	0.251
Cramer's V :	0.259

Según los resultados de presentados en la Tabla **5-22** estos, sugieren una fuerte asociación estadísticamente significativa entre las variables categóricas en la tabla de contingencia, con una relación moderada según los coeficientes de contingencia y Cramer's V.

6 Conclusiones y recomendaciones

6.1. Conclusiones

Uno de los hallazgos principales de este trabajo resulta a partir de del proceso de clusterización –el cual nos permite definir alrededor de 10 grupos socioeconómicos para las instituciones educativas–, considerando el estrato promedio de los estudiantes, es que el 90 % de los estudiantes del estrato más bajo, no supera el puntaje del percentil 20 % del estrato más alto. Así mismo, si nos fijamos por el estrato socioeconómico de la vivienda, nos encontramos con que los mejores estudiantes en los estratos bajos no pueden competir sino con los estudiantes más bajos de los estratos altos. Algo que podemos cuestionar, puesto que un estudiante de un estrato alto recibe educación significativa por parte de su entorno social, lo cual no es posible para los estudiantes de estratos bajos, dadas las condiciones tan limitadas de su entorno.

En el trabajo descriptivo, después de realizar el proceso de clusterización, se evidencia que si la comparación fuera más equitativa se podría resaltar el logro de varias instituciones públicas así como el verdadero nivel socioeconómico del estudiante, que es necesario para visualizar las grandes brechas sociales y económicas que enfrentan. Por esta razón, en la base de datos se trabajó bajo algunas condiciones para realizar este proceso de comparación de la manera más justa posible; lo común es comparar público y privado sin tener presente que estudiantes de jornada noche y sábados tienen otras condiciones y manejan tiempos de estudio más cortos, pues en su mayoría trabaja.

Asimismo, se encontró que al agrupar estudiantes bajo las mismas condiciones (K-Means), cuando se realizan comparaciones entre instituciones públicas y privadas, las diferencias son mínimas. En estratos propuestos por el clúster como bajos, los colegios públicos tienen mejores resultados que los colegios privados en algunos clúster para puntaje global y lectura crítica. Cabe destacar que variables como la educación de la madre, la alimentación, tener internet, computador, dedicación a la lectura, muestran gran influencia en el resultado de las pruebas Saber 11.

Según lo anterior, un estudiante de un estrato bajo que ocupe un percentil superior en las pruebas Saber dentro de su estrato, tiene condiciones que hacen difícil que logre pasar a una universidad pública (que, en general, realiza un examen de admisión donde solo se considera el resultado global de este examen, y solo considera la condición económica para determinar el costo de la matrícula). Por esta razón, la misma universidad se vuelve una fortaleza inexpugnable para estudiantes de estratos bajos con estos exámenes que desconocen sus realidades. Por ello, si un estudiante con estas condiciones logra entrar a la universidad, se encuentra con un sistema pedagógico que está en su contra, ya que esta mezcla indiscriminadamente estudiante que provienen de estratos altos, medios y bajos, y enfrenta a docentes que solo se interesan por los estudiantes capaces de sobrevivir en sus cursos.

En este sentido, la universidad pública realizaría una labor social importante si, en lugar de escoger a los mejores de los mejores, escogieran a los mejores de su estrato, privilegiando el potencial. Esto a razón de que, si un estudiante de un estrato bajo termina su carrera, tendrá una movilidad social mucho mayor que uno de estrato alto, adicionalmente su impacto en su entorno podría ser mayor al volverse un referente.

El gobierno nacional da gran importancia a la inversión en educación básica, lo cual es reconocible; pero, considero que el camino más rápido para disminuir las brechas sociales es brindando la oportunidad para que jóvenes y adultos continúen con los procesos de educación: ya se ha visto que la educación de los padres influye positivamente en el resultado de pruebas estandarizadas de los estudiantes, entonces, educar a los padres ayudaría a que ellos puedan influir en la educación de sus hijos y se posibilite una mejor calidad de vida que garantice cubrir las necesidades básicas y que el tiempo que se determina para disminuir las brechas sea menor.

Cabe resaltar que, según los datos encontrados, se evidencia que tener moto está relacionado con los niveles bajos socioeconómicos y que, en promedio, los resultados son más bajos en las pruebas Saber 11 de las familias que tienen moto. Solo en dos regiones se observa que tener moto estaría relacionado con conseguir un buen puntaje en las pruebas Saber 11. De igual forma, tener internet, computador, libros en casa, buenos hábitos de lectura, una buena alimentación, puede estar influyendo en los resultados de las pruebas estandarizadas. Este tipo de afirmaciones son respaldadas por investigaciones como Nuñez, Zambrano, Alarcón, Monar, y Cisneros (2017), en su estudio vinculado con desarrollo del cerebro relacionados con la buena alimentación.

6.2. Recomendaciones

Se recomienda realizar un análisis sobre el conflicto armado en Antioquia y como este puede haber influido en el resultado de las pruebas Saber. También, se recomienda hacer un análisis con las bases de datos de los estudiantes de la Universidad Nacional que permita visualizar el nivel socioeconómico real de los estudiantes que pertenecen a la institución; y un análisis sobre el impacto que tienen en el rendimiento académicos las asesorías de estudiantes monitores y maestros.

Sería muy valioso que la Universidad Nacional se uniera a la campaña de invitar estudiantes de colegios públicos a tomar clases con estudiantes universitarios. Para esto se requiere de un enlace con los colegios y que ellos sean los que seleccionen a sus estudiantes, así como el compromiso social de los docentes y en general de la comunidad educativa.

Una recomendación para mejorar los resultados en las pruebas Saber es promover el trabajo colaborativo entre los colegios que se encuentran en situaciones similares. Sería beneficioso establecer un espacio de intercambio entre los equipos docentes, donde se compartan estrategias exitosas y se discuta sobre las practicas pedagógicas que están generando buenos resultados. Este trabajo conjunto podría involucrar a docentes, padres y estudiantes, creando una comunidad educativa comprometida con la mejora académica.

Por otro lado, es clave evitar el error común de comparar estudiantes y colegios de diferentes estratos socioeconómicos, ya que esto no tiene en cuenta el contexto social, cultural y económico, así como las posibilidades académicas adicionales disponibles en cada entorno. En cambio, se debe priorizar la comparación equitativa entre colegios y estudiantes que se encuentren en condiciones similares, para generar un análisis más justo y constructivo. Dentro del trabajo colaborativo entre colegios, se debe fomentar la participación activa de cada institución, permitiendo que, desde su realidad y posibilidades, puedan proponer estrategias de mejora efectivas. Por último, cada colegio tiene sus particularidades y desafíos, por lo que es importante valorar y aprovechar el conocimiento y la experiencia de cada comunidad educativa. Al permitir que cada institución aporte soluciones desde su contexto, se promoverá un enfoque más inclusivo y adaptado a las necesidades específicas de cada colegio, lo que contribuirá a una mejora integral de los resultados en las pruebas Saber 11.

Referencias

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley.

Alcaldía de Caucasia. (2019). *Plan de desarrollo municipal caucasia 2020.* Descargado de `https://www.caucasia-antioquia.gov.co/inicio/index.shtml`

Arenas, G. (2002). Las madres en la educación, una voz siempre presente pero,¿ reconocida. *El harén pedagógico: perspectiva de género en la organización escolar*, 103–118.

Aristizabal, G. C., Rosero, M. D., Bedoya, J. T., y cols. (2016). *Causas de las diferencias en desempeño escolar entre los colegios públicos y privados: Colombia en las pruebas saber11 2014* (Inf. Téc.).

Austin, P. C. (2011). *An introduction to propensity score methods for reducing the effects of confounding in observational studies* (Vol. 1).

Bernal, L. K. A.-G., Bernal, G., y cols. (2016). *Brechas de género en el rendimiento escolar a lo largo de la distribución de puntajes: evidencia pruebas saber 11°* (Inf. Téc.). Universidad Javeriana-Bogotá.

Berry, M. J., y Linoff, G. S. (2004). *Data mining techniques: For marketing, sales, and customer relationship management*. John Wiley & Sons.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... [...] (2021). mlr: Machine learning in r [Manual de software informático]. Descargado de `https://mlr.mlr-org.com/` (R package version 2.18.0)

Borg, W. R., y Gall, M. D. (1984). Educational research: An introduction. *British Journal of Educational Studies*, *32*(3), 274–274.

Báez, L. G. (2020). Confiabilidad de las pruebas estandarizadas que se aplican en colombia para medir y evaluar la calidad de la educación. *Revista Espacios*, *41*(35), 1–15.

Castro, L. M. C., Ortiz, F., y Lemus, D. F. (2016). Construcción de un índice socioeconómico familiar para los estudiantes que presentan la prueba saber 11. *Comunicaciones en Estadística*, *9*(1), 79–92.

Castro Aristizabal, G., Diaz Rosero, M., y Tobar Bedoya, J. (2016). *Causas de las diferencias en desempeño escolar entre los colegios públicos y privados: Colombia en las pruebas saber11 2014* (Inf. Téc.). Faculty of Economics and Management, Pontificia Universidad Javeriana Cali.

Checchi, D., y Peragine, V. (2010). Inequality of opportunity in italy. *The Journal of Economic Inequality*, *8*(4), 429–450.

Chica Gómez, S., Galvis Gutiérrez, D., y Ramírez Hassan, A. (2011). Determinantes del rendimiento académico en colombia: Pruebas icfes saber 11, 2009 (academic performance determinants in colombia: Icfes saber 11, 2009 exam). *Center for Research in*

Economics and Finance (CIEF), Working Papers(11-5).

Chiok, C. H. M. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. En *Anales científicos* (Vol. 78, pp. 26–33).

Cruz, Z., Medina, J., Vázquez, J., Espinosa, E., Antonio, A., y Carmona, G. (2014). Influencia del nivel socioeconómico en el rendimiento académico de los alumnos del programa educativo de ingeniería industrial en la universidad politécnica de altamira. *N. Carmona, Y. Santamaría y L. Almanza (coords.), Ciencias Administrativas y Sociales. Handbook TV*, 24–38.

Diamond, J. (2021). The influence of cultural and social conditions on student academic performance. *Journal of Education*, *55*(3), 123-134.

Ferreira, F. H., Gignoux, J., y Aran, M. (2011). Measuring inequality of opportunity with imperfect data: the case of turkey. *The Journal of Economic Inequality*, *9*(4), 651–680.

Firke, S., Bengtsson, J., Hill, S., y Wickham, H. (2021). janitor: Simple tools for examining and cleaning dirty data [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=janitor` (R package version 2.1.0)

Flórez, C., Espinosa, F., Sánchez, L., y Angulo, R. (2008). Diseño del índice sisben en su tercera versión. *Bogotá, Colombia. https://www. sisben. gov. co/Portals/0/Documentos/Documentos Tecnicos/02. Resumen Ejecutivo Sisb*.

Giordani, P., Ferraro, M. B., Martella, F., Giordani, P., Ferraro, M. B., y Martella, F. (2020). *Introduction to clustering*. Springer.

Grizzle, S.-C. F. y. K. G. G., J. E. (1969). Analysis of categorical data by linear models biometrics. , *25*(Nov), 489–504.

Guarín, A., Medina, C., y Posso, C. (2018). Calidad, cobertura y costos ocultos de la educación secundaria pública y privada en colombia. *Revista desarrollo y sociedad*(81), 61–114.

Guo, . F.-M. W., S. (2010). *Propensity score analysis: Statistical methods and applications* (Vol. 1).

Ho, I.-K. K. G. . S. E. A., D. E. (2007). *Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference* (Vol. 3).

ICFES. (2019). Instituto colombiano para la evaluación de la educación - icfes. (2021). boletín saber al detalle (edición 4).

Kassambara, A. (2017). *Practical guide to cluster analysis in r* (Vol. 1). Createspace.

Koh, T. G., Hian. (2005). Data mining applications in healthcare. *Journal of healthcare information management*(19), 64–72.

Kuhn, M., Wickham, H., y RStudio. (2021). tidymodels: Easily install and load the 'tidymodels' packages [Manual de software informático]. Descargado de `https://www.tidymodels.org/` (R package version 0.1.4)

López, Á., Virgüez, A., Silva, C., y Sarmiento, J. (2017). Desigualdad de oportunidades en el sistema de educación pública en bogotá, colombia. *Lecturas de Economía*(87), 165–190.

Maaten, L. v. d., y Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En L. M. Le Cam y J. Neyman (Eds.), *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Berkeley, CA: University of California Press.

Manrique, I. J. A., y Carreño, C. A. A. (2014). Influencia de los padres en el rendimiento académico de los hijos: una aproximación econométrica en el contexto de la educación media colombiana. *Educación y Desarrollo Social*, *8*(2), 184–199.

Marqués, I. (2016). Apuntes sobre el informe coleman. sobre la difícil convivencia de los principios igualitarios en un mundo desigual. *International Journal of Sociology of Education*, *5*(2), 107–126.

Monroy, L. G. D., Rivera, M. A. M., y Dávila, L. R. L. (2018). *Análisis estadístico de datos categóricos*. Universidad Nacional de Colombia.

Murillo, F. J., y Carrillo-Luna, S. (2021). Segregación escolar por nivel socioeconómico en colombia y sus departamentos. *Magis, Revista Internacional de Investigación en Educación*, *14*, 1–23.

Nuñez, R. B., Zambrano, M. Q., Alarcón, M. S., Monar, L. V., y Cisneros, J. C. (2017). Alimentación saludable como factor influyente en el rendimiento escolar de los estudiantes de instituciones educativas en ecuador. *FACSALUD-UNEMI*, *1*(1), 34–39.

Padilla-Escorcia, I. A., González-Tinoco, N. E., y Fernández-Díaz, O. R. (2022). Modelo estadístico para estimar la influencia de la lectura crítica en las competencias evaluadas en las pruebas saber 11°. *Trilogía Ciencia Tecnología Sociedad*, *14*(26).

Pardo, C. E. (2020). Estadística descriptiva multivariada.

Pérez-Pulido, M. O., Aguilar-Galvis, F., Orlandoni-Merli, G., Ramoni-Perazzi, J., y cols. (2016). Análisis estadístico de los resultados de las pruebas de estado para el ingreso a la educación superior en la universidad de santander, colombia-statistical analysis of the results of state tests for admission to higher education at the university of santander, colombia. *Revista científica*, *4*(27), 328–339.

Piñero, J. C. M., Sánchez, M. C. C., Bernal, I. A. M., y Jerez, S. A. R. (2019). Incidencia de las tic en el mejoramiento de las pruebas saber 11: un análisis a partir del modelo tpack. *Encuentro Internacional de Educación en Ingeniería*.

Ramoni Perazzi, J., Orlandoni Merli, G., Pérez Pulido, M. O., y Aguilar Galvis, F. (2016). Análisis estadistico de los resultados de las pruebad de estado para el ingreso a la educación superior en la universidad de santander, colombia. *Revista Científica*.

Rhys, H. (2020). *Machine learning with r, the tidyverse, and mlr* (Vol. 1). Shelter Island.

Rodríguez, D. F. M. (2016). Algunos factores que influyen en los resultados de las pruebas estandarizadas y censales. *Boletín Redipe*, *5*(3), 136–145.

Ríos-Cuesta, W. (2023). Desempeño histórico en la prueba saber de matemáticas: la necesidad de revisar la política educativa del chocó. *Encuentros*, *21*(01), 30–39.

Samper, J. D. Z. (2021). *La inteligencia y el talento se desarrollan* (Vol. 1). Magisterio.

Sanchez, A. (2011). Etnia y rendimiento académico en colombia. , *14*(Dic), 189–227.

Sánchez, G. D. D. (2020). La evaluación desde las pruebas estandarizadas en la educación en latinoamérica. *Revista En-Contexto*, *8*(13), 107–133.

Toranzos, L. (1996). Evaluación y calidad. *Revista iberoamericana de educación*, *10*.

Torrecilla, F. J. M. (2008). Los modelos multinivel como herramienta para la investigación educativa. *Magis. Revista Internacional de Investigación en Educación*, *1*(1), 45–62.

Villacís Mejía, J. E. (2020). Estado nutricional antropométrico, nivel socioeconómico y rendimiento académico en niños escolares de 6 a 12 años las islas galápagos, ecuador 2019.

Waring, D., y Chang, E. (2021). skimr: Compact and flexible summaries of data [Manual de software informático]. Descargado de `https://cran.r-project.org/package=skimr` (R package version 2.1.3)

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis [Manual de software informático]. Descargado de `https://cran.r-project.org/web/packages/ggplot2/index.html` (R package version 3.3.0)

Wickham, H., y Bryan, J. (2021). tidyverse: Easily install and load the 'tidyverse' [Manual de software informático]. Descargado de `https://tidyverse.org/` (R package version 1.3.1)

Wickham, H., Bryan, J., y Müller, K. (2021). readxl: Read excel files [Manual de software informático]. Descargado de `https://readxl.tidyverse.org/` (R package version 1.3.1)

Zuluaga, K. J. H., y Morales, J. C. C. (2018). Regresión logística bivariable para tablas de contingencia usando metodología gsk. *Comunicaciones en Estadística*, *11*(2), 153–170.