

Rendimiento Académico, Nivel Socioeconómico y Brecha Digital: Un Análisis Big Data a las Pruebas Saber 11. (2008-2024)

Academic Performance, Socioeconomic Level and Digital Divide: A Big Data Analysis of the Saber 11 Tests. (2008-2024)

Juan José Vanegas Giraldo¹, Yan Carlos Ureña Villamizar², Camilo Zapata Molina³.

¹ Doctor en Ciencias de la Educación, Universidad Metropolitana de Ciencia y Tecnología, Ciudad de Panamá - Panamá. Orcid: <https://orcid.org/0000-0001-5120-786X>, Email: juanvanegas.est@umecit.edu.pa

² Doctor en Ciencias Gerenciales, Universidad Metropolitana de Ciencia y Tecnología, Ciudad de Panamá - Panamá. Orcid: <https://orcid.org/0000-0002-9970-159X> . Email: yanurena.doc@umecit.edu.pa

³ Magister en Educación, Tecnológico de Antioquia – Institución Universitaria, Antioquia - Colombia. Orcid: <https://orcid.org/0000-0002-9009-4498> . Email: camilo.zapata30@correo.tdea.edu.co

Recibido: Día 24, Mes 02, Año 2026.

Aprobado: Día 30, Mes 03, Año 2026.

Publicado: Día 22, Mes 07, Año 2026.

Citación: Vanegas, J.; Ureña, Y., & Zapata, C. (2026). Rendimiento Académico, Nivel Socioeconómico y Brecha Digital: Un Análisis Big Data a las Pruebas Saber 11. Revista Conocimiento, Investigación y Educación. CIE, 2 (23), 52- 61.

Resumen:

El artículo analiza la relación entre las condiciones socioeconómicas, el acceso a las Tecnologías de la Información y la Comunicación (TIC) y el rendimiento académico de los estudiantes colombianos en las pruebas Saber 11 durante el periodo 2008-2024. El estudio parte de la premisa de que la brecha digital constituye un factor determinante en la desigualdad educativa, especialmente tras la pandemia de COVID-19, que aceleró la transición hacia entornos de aprendizaje mediados por tecnologías digitales. Para ello, se empleó un enfoque cuantitativo, longitudinal y descriptivo-correlacional, utilizando una base de datos de más de nueve millones de registros depurados del ICFES. La metodología incluyó procesos de extracción, transformación y carga de datos (ETL), análisis estadístico, correlaciones de Pearson y técnicas de aprendizaje automático, particularmente Random Forest, MissForest y el algoritmo K-Means. Los resultados evidencian un crecimiento significativo del acceso a internet, que pasó del 22,2 % en 2008 al 73,4 % en 2024. Sin embargo, persisten diferencias importantes entre estudiantes de instituciones públicas y privadas, tanto en acceso como en intensidad de uso de las tecnologías. El análisis de conglomerados identificó tres perfiles de estudiantes. El grupo con mejor desempeño académico presentó mayores niveles socioeconómicos y acceso casi universal a internet y computadores. En contraste, el grupo de menor rendimiento evidenció una marcada exclusión digital y condiciones socioeconómicas desfavorables. Asimismo, las correlaciones mostraron asociaciones positivas entre acceso tecnológico, estrato socioeconómico y resultados en matemáticas, lectura crítica e inglés. Se concluye que la brecha digital trasciende el acceso a la tecnología y está estrechamente vinculada con factores socioeconómicos que condicionan el aprovechamiento educativo de las TIC y las oportunidades de aprendizaje.

Palabras Clave: Rendimiento académico, nivel socioeconómico, brecha digital, datos masivos, análisis de conglomerados.

Abstract:

This article analyzes the relationship between socioeconomic conditions, access to Information and Communication Technologies (ICTs), and the academic performance of Colombian students on the Saber 11 standardized tests during the period 2008–2024. The study is based on the premise that the digital divide is a determining factor in educational inequality, especially after the COVID-19 pandemic, which accelerated the transition to learning environments mediated by digital technologies. To this end, a quantitative, longitudinal, and descriptive-correlational approach was employed, using a database of more than nine million records cleaned from the Colombian Institute for the Evaluation of Education (ICFES). The methodology included data extraction, transformation, and loading (ETL) processes, statistical analysis, Pearson correlations, and machine learning techniques, particularly Random Forest, MissForest, and the K-Means algorithm. The results show a significant increase in internet access, rising from 22.2% in 2008 to 73.4% in 2024. However, significant differences persist between students in public and private institutions, both in access to and intensity of use of technology. Cluster analysis identified three student profiles. The group with the best academic performance exhibited higher socioeconomic levels and almost universal access to the internet and computers. In contrast, the lowest-performing group showed marked digital exclusion and unfavorable socioeconomic conditions. Furthermore, correlations showed positive associations between technological access, socioeconomic status, and performance in mathematics, critical reading, and English. It is concluded that the digital divide transcends access to technology and is closely linked to socioeconomic factors that influence the educational use of ICTs and learning opportunities.

Keywords: Academic achievement, socioeconomic status, digital divide, big data, cluster analysis.

1. INTRODUCCIÓN

La educación en la actualidad, precisa de una transformación adecuada de los medios y herramientas pedagógicas acorde a las potencialidades que ofrecen las Tecnologías de la Información y la Comunicación (TIC) para alinear la formación de ofertada por instituciones educativas a las características y necesidades de la era digital, no obstante a ello, en la actualidad se evidencian brechas importantes que se han agudizado a lo largo de los años, gestando situaciones estructurales de desigualdad que impiden dicha transformación (Miah, 2024).

Bajo la influencia de dichas situaciones estructurales de desigualdad, emerge la llamada brecha digital, haciendo referencia a la diferencia existente en el material y dispositivos tecnológicos a los que pueden acceder las personas en limitación a sus capacidades socioeconómicas, lo que deriva en ventajas educativas para aquellos sectores más favorecidos (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO], 2023). Siendo esta brecha profundizada resultado de la cuarentena dispuesta durante el periodo de pandemia causado por el COVID-19 que forzó a transitar hacia la educación mediada por entornos digitales, donde quienes contaban con tecnologías modernas aptas para mantener procesos formativos remotos e híbridos se vieron especialmente favorecidos en su rendimiento académico (García, 2025). Manifestando una problemática aguda que derivó en una formación y desarrollado de competencias limitados, manifestados en los resultados de aprendizaje (Organización para la Cooperación y el Desarrollo Económicos [OCDE], 2020).

De forma más explícita, la preexistente problemática socioeconómica condicionó el acceso a las TIC a ciertos sectores de la sociedad y, por tanto, el desarrollo de competencias básicas requeridas para su uso efectivo, en consecuencia, la transición forzosa a la educación remota mediada por entornos digitales desencadenó resultados de

aprendizaje indicados dispares, siendo así un predictor actual del rendimiento académico (Sirin, 2005; Zabala et al., 2023). Paralizando así la movilidad social y el progreso de las naciones (Escobar, 2024).

En el contexto colombiano, dicha transición forzosa evidenció la magnitud de la brecha digital, informa la Universidad Nacional de Colombia (UNAL, 2021), que aproximadamente cinco de ocho millones de estudiantes cuyo proceso formativo se desarrollaba en instituciones educativas públicas, no contaban con computador ni acceso a internet, lo que, según Melo-Becerra et al. (2021), aumentó la deserción, repitencia de año escolar y resultó determinante en los resultados dispares de las pruebas Saber 11.

Dichas pruebas, son ejecutadas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), cuyos resultados históricos son compilados en sus bases de datos públicas que pueden ser consultadas y analizadas; en particular, se encuentra información relativa al nivel de competencia en cinco áreas de formación: lectura crítica, matemática, sociales y ciudadanas, ciencias naturales e inglés, así como de aspectos colaterales que pueden influir en el desempeño de alguna de estas competencias, como lo es la presencia de herramientas tecnológicas y estrato socioeconómico.

Con este contexto como base, el presente artículo se propone analizar la correlación de tres de las áreas mencionadas: Matemáticas, inglés y lectura crítica con los otros datos recopilados por el ICFES en las pruebas Saber 11 que dan cuenta de las condiciones socioeconómicas, tecnológicas y los entornos educativos, tomando las bases de datos que abarcan desde el año 2008 hasta el 2024. Cabe resaltar que, debido al masivo volumen de los datos implicados, se utilizarán técnicas de manipulación Big Data, en particular, se ejecutará un análisis mediante Machine Learning (ML) que hacen parte de algoritmos de Inteligencia Artificial (IA) no generativa. De esta manera, se busca identificar patrones, tendencias y relaciones subyacentes que no son fácilmente observables

mediante métodos estadísticos tradicionales. Asimismo, el estudio pretende establecer perfiles de estudiantes según sus características socioeconómicas y tecnológicas, permitiendo comprender con mayor profundidad los factores asociados al desempeño académico. Los hallazgos obtenidos servirán como insumo para la formulación de estrategias educativas y políticas públicas orientadas a reducir las brechas de aprendizaje y promover una mayor equidad en el sistema educativo colombiano.

2. REVISIÓN DE LITERATURA

Inteligencia Artificial

La IA, es entendida como una tecnología avanzada que permite mayor autonomía a gran cantidad de procesos, ofreciendo grandes oportunidades respecto a la personalización y eficiencia en el procesamiento de datos, creación de contenidos y retroalimentación inmediata a diversa cantidad de tareas, optimizando la educación misma al aportar nuevas experiencias de aprendizaje y de gestión académica, promoviendo un aprendizaje significativo.

Para encontrar una definición clara, es preciso explorar diversas perspectivas, al respecto Perezchica-Vega et al. (2024), se centran en definir la IA generativa en la esfera educativa como aquella tecnología que posibilita automatizar la creación de contenidos, interacción con el estudiante y procesamiento de gran cantidad de información, dicha capacidad da paso a la personalización y aumento tanto de eficiencia como eficacia de los procesos formativos, no obstante, también está plagada de desafíos éticos y de responsabilidad académica que se han de tener en constante consideración, requiriendo constante monitoría y reflexión crítica para garantizar que no comprometa de manera negativa principios fundamentales como la equidad, transparencia y privacidad de quienes la utilizan.

Desde una perspectiva afín, Villegas-José y Delgado-García (2024), conciben la IA en el ámbito educativo como una herramienta que permite personalizar procesos formativos y optimizar la manera en que se gestionan las instituciones que prestan servicios educativos, de este modo, favorece interacción entre estudiante-docente, facilitando el análisis de tendencias en el aprendizaje, dicho enfoque, contribuye a la adaptación de contenidos formativos a las necesidades individuales de cada estudiante, centrando todo el proceso en este, ajustándolo a su ritmo y necesidades.

Por su parte, Hernández (2024), afirma que la IA generativa es una tecnología contemporánea de reciente aparición que posibilita crear contenido mediante algoritmos, facilitando el procesamiento de datos y su interpretación desde diversas fuentes para apoyar las actividades educativas, asimismo, se resalta la reciente adaptabilidad de las interfaces conversacionales como ChatGPT que han presentado virtudes y utilidades que bajo su correcta aplicación, complementan los procesos formativos de manera efectiva.

En consideración con las perspectivas expuestas, se resalta la perspectiva de Villegas-José y Delgado-García (2024), pues aporta una perspectiva que converge con los demás autores expuestos y desataca la

importancia de la IA en entornos educativos, en especial, para personalizar el proceso formativo de los estudiantes, optimizar la gestión académica, así como crear entornos dinámicos de aprendizaje centrados en el estudiante. De este modo, se plasman conceptualmente características que son de interés para el proceso investigativo y que resalta el potencial transformador de la IA para la educación superior, lo que justifica su inclusión para el diseño de un modelo pedagógico.

Clusterización

La clusterización es una técnica clave para el análisis de gran cantidad de datos, permitiendo identificar patrones que derivan en la creación de soluciones personalizadas a gran cantidad de problemas complejos que se ven implicados por gran cantidad de variables. Dicha técnica es fundamental para la investigación educativa en la actualidad, pues permite comprender de manera adecuada los diversos factores que inciden en el proceso de aprendizaje de los estudiantes, facilitando en última instancia, el diseño de un modelo de aprendizaje que tenga en consideración las dinámicas de la industria 4.0 y la IA, en particular, la generativa.

Al respecto, Tan Choon Keong et al. (2025), destacan que esta técnica es utilizada para dar cuenta de temáticas, grupos y tendencias dentro un conjunto de datos, es comúnmente aplicada en ciencias de la computación y de datos, no obstante, se presenta como poderosa e innovadora en la educación, en especial, bajo las particularidades de la integración y transformación tecnológica actuales, pues se recaban datos más fácilmente y se observan dinámicas desde una perspectiva más estructurada, posibilitando la optimización de los procesos formativos.

Procesos de Aprendizaje

El proceso de aprendizaje bajo una mirada educativo-tecnológica, puede entenderse a rasgos generales como aquella construcción de conocimientos que emerge tras la articulación entre pedagogía, tecnología y contextos. Bajo dicha mirada, existen diversos enfoques metodológicos que son adaptables a las necesidades de un modelo de educación 4.0 basado en la IA que resaltan por generar en los estudiantes una comprensión profunda de los temas tratados y una motivación sostenida, a continuación, se describen algunos de estos.

A modo de inicio, se hace mención a Guo et al. (2020), quienes exponen el Aprendizaje Basado en Proyectos (ABP) como aquella metodología inmersiva que integra a los estudiantes en proyectos que buscan dar respuesta a necesidades o problemáticas reales, dicho enfoque, permite el desarrollo de conocimiento situado, productos prácticos y un continuo aprendizaje significativo ligado a contextos reales.

Brecha Digital

Las iniciativas normativas internacionales las lidera actualmente la UNESCO (2021), que ha plasmado un marco ético fundamental alrededor de la gobernanza de la IA buscando mantener un uso centrado en la

promoción y prevención de los derechos humanos, la inclusión y el desarrollo sostenible, destacando la potencialidad que tiene la IA para contribuir al fomento de una educación de calidad que evite riesgos como la discriminación algorítmica y el uso indebido de los datos personales.

Esto se alinea con lo plasmado por UN (2015), en la agenda 2030, donde se incentiva a los estados miembros a integrar tecnologías emergentes en la generalidad de los procesos educativos, buscando la disminución del conjunto de brechas existentes que reproducen la situación de desigualdad en las sociedades humanas y asegurar la calidad educativa. En esta medida, la UNESCO (2021), también expone la importancia de la transparencia, explicabilidad y responsabilidad que se ha de tener al utilizar la IA, siendo estos principios esenciales que generan confianza en los sistemas educativos y los habilita a automatizar procesos y mantener una adecuada protección de los derechos de estudiante y docentes.

3. METODOLOGIA

El presente artículo, en su desarrollo metodológico, se adscribió al paradigma empírico-analítico, cuyo fundamento es la observación, medición y análisis estadístico del objeto de estudio, en esta medida, se entiende que los fenómenos ocurren en un mundo cuantificable y medible (Ortiz, 2015). Asumiendo así que el rendimiento académico es factible a ser descrito y correlacionado mediante procedimientos replicables apoyados de métodos propios de la Big Data y la estadística inferencial sin la necesidad de recurrir a interpretaciones subjetivas sujetas a sesgos del investigador.

En congruencia, se adoptó el enfoque cuantitativo, el cual, en su estructura permite el análisis de datos numéricos para identificar regularidades, distribuciones y relaciones estadísticas, así como patrones o tendencias en la muestra (Hernández-Sampieri y Mendoza, 2018). También se aclara que el presente estudio se enmarcó en el tipo longitudinal, pues se utilizó la base de datos ICFES, para el análisis de las variaciones que se produjeron entre los años 2008 a 2024 en las variables de interés (Arias, 2012).

Asimismo, es de diseño de descriptivo correlacional, pues siguiendo a Arias (2012), se buscó dar cuenta del grado de relación no causal entre los resultados dados de las pruebas Saber 11 en matemáticas, inglés y lectura crítica, así como las características socioeconómicas, tecnológicas y los entornos educativos. Respecto al procedimiento, la información fue sometida a un proceso de Extracción, Transformación y Carga (ETL), que se alineó con los principios de la ingeniería de datos y el modelado dimensional, derivando en la construcción de una base de datos normalizada y ordenada (Kimball y Ross, 2013).

El proceso mencionado, incluyó la lectura, estandarización y unificación de los archivos descargados desde el histórico de resultados de las pruebas Saber 11, así como la codificación de variables cualitativas en representaciones numéricas a partir de la biblioteca Pandas de Python (McKinney, 2010). De allí, se utilizaron pruebas de ML, en específico, la técnica de Random Forest con un enfoque MissForest, que demuestra especial precisión en datos mixtos, siendo esto concordante con los

principios estadísticos que orientan el manejo de información perdida (Stekhoven y Bühlmann, 2012; Little y Rubin, 2019).

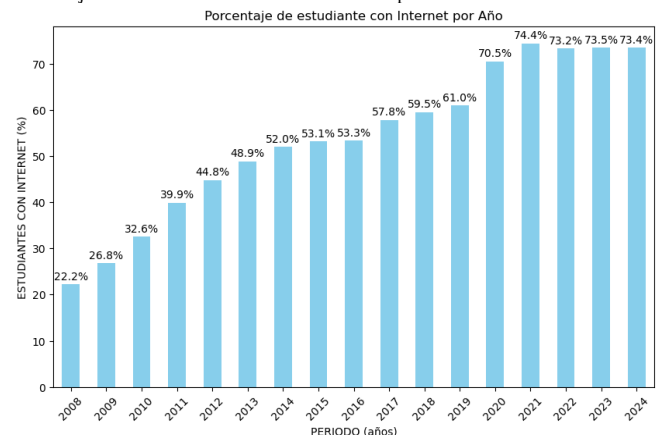
En última instancia, se procedió a analizar los datos inicialmente mediante estadística descriptiva y la posterior estadística inferencial, abarcando el análisis correlacional Pearson, el método del codo y la aplicación del algoritmo de clústers K-Means para identificar tendencias y patrones en los datos (MacQueen, 1967), acatando criterios técnicos para asegurar el rigor del estudio al determinar el número óptimo de conglomerados (Kaufman y Rousseeuw, 2005).

La población objeto, se constituyó por la totalidad de registros de los resultados de pruebas Saber 11 aplicados por el ICFES en Colombia entre el 2008 y 2024, lo que correspondió a un universo de 11.073.003 registros, representando dicho conjunto la base empírica del estudio. Posteriormente, se depuró, limpió y validó los datos, reduciendo los registros a un total de 9.474.998 que constituyeron la muestra del estudio, dicho proceso de depuración, no se basó en un muestreo probabilístico, sino en criterios de consistencia, completitud y calidad de la información recabada.

4. RESULTADOS

A continuación, se presentan los hallazgos encontrados a través de los procesos descritos anteriormente:

Figura 1
Porcentaje de Estudiantes con Acceso a Internet por Año.



Nota: Elaboración Propia (2026)

El análisis de las bases de datos revela una reconfiguración fundamental en el acceso al capital tecnológico por parte de la población estudiantil de 2008 a 2024. Se parte de un escenario de exclusión digital, donde apenas un 22.2% del estudiantado disponía de conectividad, para transitar hacia un nuevo paradigma evidenciado en el máximo histórico del 74.4% alcanzado en 2021. Este sufrió un incremento de 235,14% con diferencia de 52.2 puntos porcentuales de aumento. La primera fase, comprendida entre 2008 y 2014, se caracteriza por una expansión acelerada del acceso, con un crecimiento promedio anual de 4.97 puntos,

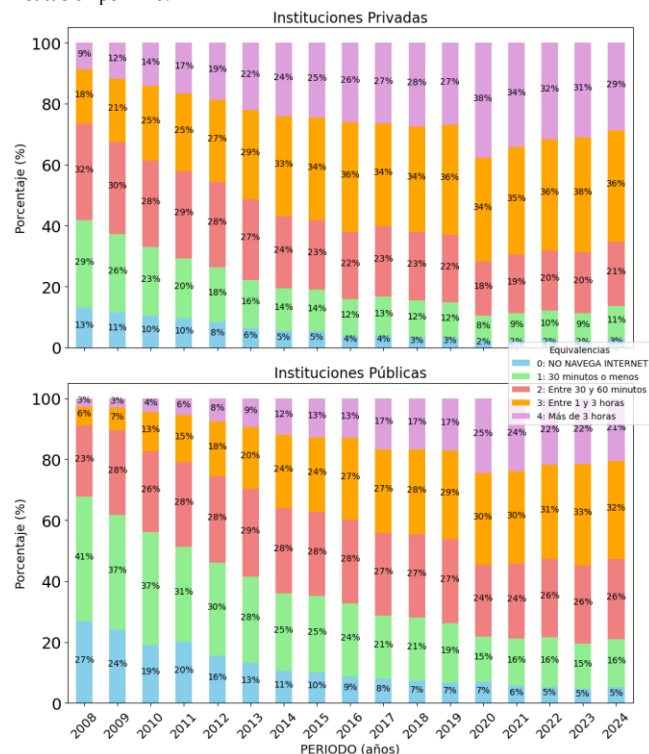
lo cual sugiere el efecto de las políticas iniciales de masificación de la infraestructura y la progresiva digitalización de los recursos pedagógicos.

Posteriormente, entre 2015 y 2019, la dinámica de crecimiento se modera, evidenciando una transición hacia una fase de maduración en la tasa de adopción tecnológica. Sin embargo, este patrón es alterado por un evento disruptivo como lo fue la pandemia de COVID-19. El salto de 9.5 puntos porcentuales registrado entre 2019 y 2020 constituye el mayor incremento interanual de estudiantes con acceso a internet y debe ser interpretado como la consecuencia directa de la digitalización forzosa del sistema educativo. Dicho cambio de paradigma transformó el acceso a Internet de un recurso complementario a una necesidad urgente para garantizar la continuidad pedagógica, permeando así la conectividad en segmentos poblacionales que hasta entonces permanecían al margen.

Por último, en el periodo de 2022 a 2024 se observa una fase de estabilización, con valores que se estabilizan en torno al 73%. Esta estabilización sugiere que el crecimiento que hubo antes ha alcanzado un techo o umbral, dando paso a un escenario donde la brecha digital se estanca. La persistencia de casi un cuarto de la población estudiantil sin conectividad ya no responde a una tendencia de crecimiento, sino a barreras estructurales de índole socioeconómica y geográfica que hace difíciles cerrarla sin los requerimientos necesarios. Ante esta exclusión digital se presenta un desafío para la equidad educativa, pues condiciona diferencialmente las trayectorias de aprendizaje y amenaza con perpetuar ciclos de desigualdad en el sistema.

Figura 2

Porcentaje de Tiempo Dedicado a Internet por Estudiantes según tipo de Institución por Año.



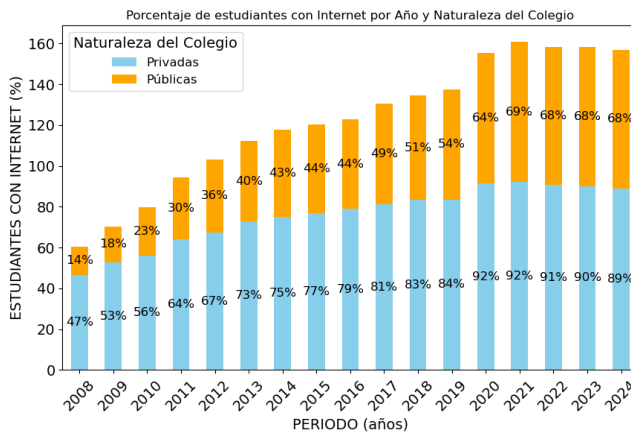
Nota. Adaptado de Porcentaje de estudiantes según dedicación a internet, por tipo de institución (públicas y privadas) en el periodo de tiempo entre el año 2008 y el año 2023, por J. Vanegas-Giraldo, 2025, Impacto de la IA en la Brecha Educativa en Colombia.

En Figura 2, el análisis del período 2008-2024 revela una profunda transformación digital tanto en instituciones públicas como privadas; la tendencia más destacada es el drástico aumento en las personas que usan internet; el porcentaje de personas que no navega disminuyó del 13% al 3% en el sector privado y, de manera aún más significativa, del 27% al 5% en el público, lo que indicaría que se redujo aún más la brecha del acceso a internet en las instituciones privadas y públicas. Asimismo, se produjo una clara intensificación en los hábitos de uso, con un desplazamiento masivo de las sesiones cortas (menos de una hora) hacia periodos de conexión más prolongados (más de una hora), indicando una mayor integración de internet en la vida diaria de los estudiantes.

Al comparar ambos sectores, se observa que las instituciones privadas mostraron una adopción más temprana y un uso más intensivo. Para 2024, el 65% de los estudiantes dedicaban más de una hora diaria a navegar, con un 29% superando las tres horas. Aunque las instituciones públicas siguieron una trayectoria de crecimiento similar y lograron cerrar la brecha de acceso, su intensidad de uso fue menor, alcanzando un 32% de usuarios con más de una hora de conexión y un 21% superando las tres horas. A pesar de que el 2020 actuó como catalizador, acelerando la adopción de tecnología y patrones de uso más prolongados en ambos grupos y consolidando la dependencia de las actividades en línea en los dos tipos de instituciones. Esto demuestra que, si bien el acceso se ha universalizado, el dominio y manejo de estas herramientas es desigual, debido a una diferencia sostenida en la intensidad de uso, donde el sector privado lidera. Por tanto, el desafío ya no reside únicamente en conectar, sino en trabajar en la capacitación y el desarrollo de habilidades, una brecha de competencias que se vuelve aún más crítica con la irrupción de tecnologías como la IA, que exigen un mayor dominio conceptual y práctico.

Figura 3

Porcentaje de Estudiantes con Internet por Año y Tipo de Colegio.

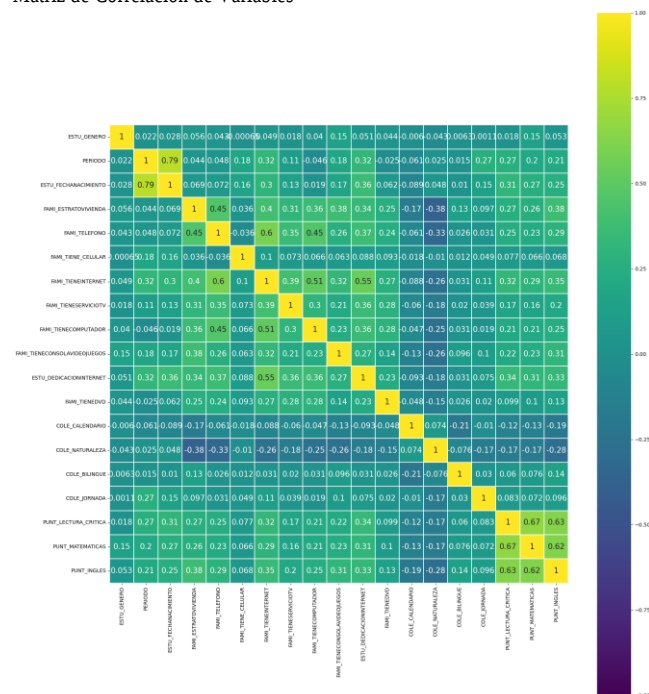


Nota. Adaptado de Porcentaje de estudiantes con acceso a internet por año (2008-2023) y naturaleza del colegio (públicas y privadas), por J. Vanegas-Giraldo, 2025, Impacto de la IA en la Brecha Educativa en Colombia.

En la figura 3, los resultados evidencian un crecimiento sostenido en el acceso a internet por parte de los estudiantes, tanto en instituciones privadas como públicas, durante el periodo 2008 a 2024. En las instituciones privadas, el porcentaje de estudiantes con acceso pasó de un 47% en 2008 a un 89% en 2024, mostrando un incremento constante, sin embargo, desde 2022 a 2024 bajó en un 1% cada por año, a pesar de que 2020 y 2021 mantuvieron sus valores más altos del 92%. Por otra parte, las instituciones públicas, aunque también presentan una tendencia creciente, partieron de niveles significativamente más bajos (14% en 2008) y alcanzaron el 64% en 2020, incrementando hasta 68% del 2021 a 2024, lo que sugiere analizar los 3 últimos años.

Para las siguientes gráficas, se mostrarán los resultados de cálculo de los clústeres explicados en la metodología donde primeramente se aplicará el método del codo, debido a que esta nos indica de mejor manera cuantos clústeres aplican para este set de datos.

Figura 4
Matriz de Correlación de Variables

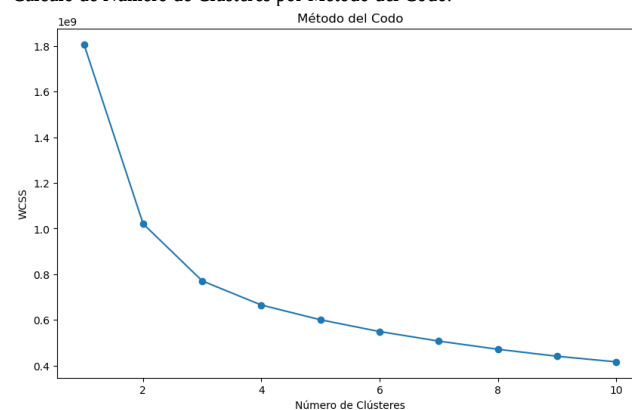


Nota: Elaboración Propia (2026)

En la Figura 4, la matriz de correlaciones revela asociaciones significativas entre los puntajes en lectura crítica, matemáticas e inglés, con valores moderados y positivos, lo que evidencia una tendencia a que el buen desempeño en una competencia se relacione con un buen rendimiento en las demás. Entre las variables contextuales, el estrato socioeconómico, el acceso a internet y a computador muestran correlaciones positivas con los puntajes, sugiriendo que mejores

condiciones materiales se asocian con mayores logros académicos. En contraste, variables como tener celular, dedicar tiempo a videojuegos, el calendario escolar o la jornada presentan correlaciones muy bajas, lo que indica una influencia mínima en el rendimiento. Por lo tanto, el capital cultural y tecnológico del hogar surge como un factor más relevante que ciertos aspectos demográficos o escolares.

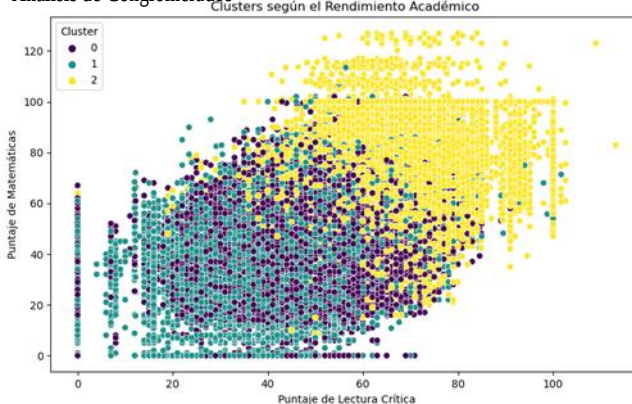
Figura 5
Cálculo de Número de Clústeres por Método del Codo.



Nota: Elaboración Propia (2026)

La Figura 5 presenta la curva obtenida a partir del método del codo, aplicada a los puntajes de las pruebas ICFES. En el eje horizontal se representa el número de clústeres considerados y en el eje vertical el valor del Within-Cluster Sum of Squares (WCSS), que mide la variabilidad interna de cada agrupación. Por lo tanto, se observa una disminución pronunciada del WCSS entre uno y tres clústeres, lo que indica que en este rango se logra una mejora sustancial en la homogeneidad interna de los grupos formados. A partir de tres clústeres, la pendiente de la curva se reduce notablemente, evidenciando que la ganancia en términos de reducción de variabilidad es secundaria al incrementar el número de grupos. Este punto indica que tres es el número óptimo de clústeres para segmentar los datos de manera eficiente, ya que añadir más grupos no aporta una mejora sustancial en la varianza explicada, razón por la cual se tomaron 3 clústeres en el estudio.

Figura 6
Análisis de Conglomerados

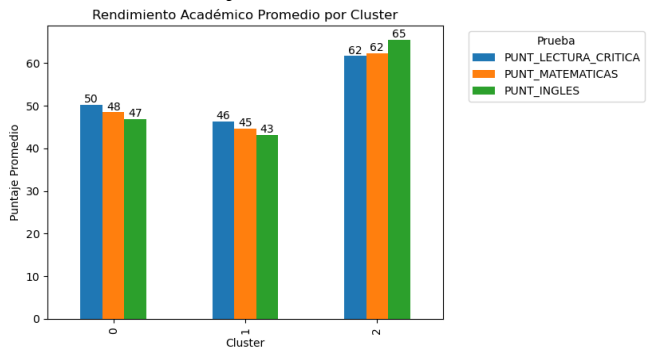


En la Figura 6 se realizó un análisis de conglomerados (clustering) para clasificar a los participantes en función de su rendimiento académico, utilizando como variables los puntajes obtenidos en las pruebas de Matemáticas y Lectura Crítica. El análisis identificó tres grupos distintos, denominados Cluster 0, Cluster 1 y Cluster 2, cuya distribución se detalla a continuación:

Los clústeres 0 (púrpura) y 1 (verde azulado) agrupan a los individuos con rendimientos bajos y medios. Ambos clústeres muestran una superposición considerable y se localizan en la misma zona del gráfico. Los puntajes de Lectura Crítica para estos dos grupos se encuentran predominantemente en el rango de 10 a 70, mientras que los puntajes de Matemáticas se sitúan, en su gran mayoría, por debajo de 80. De otro modo, se encuentran valores de Lectura Crítica y Matemática en valores de 0.

El Clúster 2 (amarillo) corresponde al grupo de alto rendimiento. Este conglomerado se distingue claramente de los otros dos, ocupando la región superior derecha del gráfico; los estudiantes de este grupo presentan de forma consistente puntajes de Lectura Crítica superiores a 50 y, mayoritariamente, superiores a 60. De manera similar, sus puntajes en Matemáticas son consistentemente altos, ubicándose casi en su totalidad por encima de 60 y extendiéndose hasta los valores máximos de la escala de evaluación.

Figura 7
Rendimiento Académico por Clúster



Nota: Elaboración Propia (2026)

El clúster 2 representa el grupo de mayor rendimiento: sus promedios son sustancialmente superiores en las tres pruebas (62–65) respecto a los demás clústeres, lo que confirma que este clúster corresponde a los estudiantes de mejor desempeño global.

Por otro lado, el clúster 0 y el clúster 1 muestran rendimientos similares entre sí, pero el clúster 0 tiene promedios ligeramente superiores al clúster 1 en las tres pruebas (diferencias de: 4 puntos en lectura, 3 en matemáticas, 4 en inglés). Esto sugiere dos grupos de desempeño medio-bajo (por debajo de 50) para el clúster 0 y para el clúster 1 desempeño bajo en los 3 puntajes de la prueba.

Finalmente, el clúster 2 (altas puntuaciones en las tres áreas) sugiere un perfil de rendimiento transversal (competencias consolidadas tanto en lectura como en matemáticas e inglés). En cambio, clústeres 0 y

1 muestran perfiles paralelos de rendimiento moderado-bajo en las tres pruebas.

Tabla 1
Análisis de Variable por Clúster

Variable	Clúster 0	Clúster 1	Clúster 2
Estrato	2.128675	1.439336	3.162231
Tiene teléfono	0.770081	0.143875	0.914526
Tiene celular	0.973211	0.939355	0.99209
Tiene internet	0.896358	0.051431	0.961395
Tiene servicio de TV	0.852589	0.439636	0.928406
Tiene computador	0.912793	0.164567	0.989938
Tiene consola de videojuegos	0.095921	0.016923	0.584889
Dedicación a internet	2.65313	1.339358	3.079003
Tiene DVD	0.901781	0.652392	0.95364
Calendario del Colegio	1.015576	1.014811	0.890758
Naturaleza del colegio	0.763634	0.881587	0.280541
Colegio bilingüe	0.004848	0.012019	0.049917
Jornada del colegio	1.503671	1.383864	2.284541
Puntaje de lectura crítica	50.25648	46.28518	61.626122

Puntaje de matemáticas	48.42988	44.56028	62.260311
Puntaje de inglés	46.77292	43.13290	65.453226
Población por Clúster	3,660,718	4,254,95	1,559,329

El clúster 2 está compuesto por la población con el mayor nivel socioeconómico, con un estrato promedio de 3 A su vez, cuentan con el mayor acceso a medios y herramientas de las TIC: un 99% posee celular, un 98% tiene computador y un 96% reporta tener acceso a internet. Este perfil se correlaciona con el desempeño académico más alto, alcanzando puntajes promedio de 61.62 en lectura crítica, 62.26 en matemáticas y 65.45 en inglés. Este grupo está conformado por 1,559,329 estudiantes.

El clúster 0 representa a la población con un nivel socioeconómico intermedio, reflejado en un estrato promedio de 2.12. Cuentan con un notable acceso a la tecnología, donde un 97% tiene celular, un 91% dispone de computador y un 89% tiene conexión a internet. Su rendimiento académico es moderado, con puntajes promedio de 50.25 en lectura crítica, 48.42 en matemáticas y 46.77 en inglés. Este es el segundo clúster más grande, con 3,660,718 estudiantes.

Por último, el clúster 1 es el más numeroso, con 4,254,951 estudiantes, agrupando a la población con el menor nivel socioeconómico, con un estrato promedio de 1.43. Este grupo enfrenta una profunda brecha digital; a pesar de que un 93% tiene celular, solo un 16% posee computador y apenas un 5.1% cuenta con acceso a internet. Esta falta de acceso a herramientas TIC se refleja en los resultados académicos más bajos de los tres grupos, con puntajes promedio de 46.28 en lectura crítica, 44.56 en matemáticas y 43.13 en inglés.

5. DISCUSIÓN

El análisis de resultados efectuado permitió confirmar que el acceso a las TIC, las condiciones socioeconómicas y el entorno educativo ejercen un impacto significativo y multidimensional sobre el rendimiento mostrado por la población estudiantil en las pruebas de estado Saber 11 entre los años 2008 y 2024, evidenciando el cumplimiento del objetivo planteado, estos datos permiten discutir su implicación según estudios de naturaleza similar para encontrar similitudes y discordancias en los datos.

En primera instancia, se revela una tendencia creciente en el acceso a Internet desde 2008 con estabilización desde 2021, así como el aumento en acceso a las TIC por parte de la totalidad de los estudiantes, estos datos coinciden con lo expresado en informes oficiales del ICFES, donde se subrayan los avances en infraestructura y políticas públicas, como el programa “Computadoras para educar” que sustenta el incremento sostenido en la proporción de estudiantes con acceso a internet y herramientas tecnológicas aunque no ha logrado subsanar la disparidad evidenciada entre instituciones privadas y públicas. Esto indica una diferencia en el acceso a tecnologías digitales que afecta en gran medida a

los estudiantes menos favorecidos, exacerbando las desigualdades existentes y limitando el potencial educativo de las TIC (Escobar, 2024; Vanegas-Giraldo 2025).

Esto es de importancia, pues según García (2025), la pandemia actuó como catalizadora de transformación tecnológica en el contexto educativo, dicha rápida transición aumentó la brecha en los resultados de aprendizaje, afectando de manera desproporcional a los estudiantes según su contexto socioeconómico. Esta brecha digital se evidenció con fuerza en el análisis de clúster realizado, donde el nivel socioeconómico y el acceso a herramientas de las TIC se mostraron como determinantes para el resultado de las pruebas Saber 11, dichos datos respaldan lo establecido por Barrios et al. (2021), quienes informaron que la presencia de computadores e internet en el hogar se correlaciona con el desempeño en los resultados de las pruebas de estado, en especial, con el puntaje obtenido en las pruebas de inglés, cabe resaltar que estos autores indican que es crucial el acceso al internet. Mientras que Ariza et al. (2021), encontraron que dichas condiciones influyen en el desempeño de lectura crítica y matemáticas, indicando un impacto diferencial de las TIC según la familiaridad y presencia de habilidades digitales previas de los estudiantes. Por su parte, Rodríguez y Gamboa (2023), documentaron datos similares, aunque centrados únicamente en los puntajes de matemática.

En concordancia, Alvarez (2024), usó los datos administrativos del ICFES 2015-2019, y encontró una relación condicional positiva entre la presencia de TIC en el hogar y un mejor resultado en pruebas estandarizadas, conversamente, la relación se dató como negativa si solo había posibilidad de acceso en la escuela, resaltando la importancia de la condición socioeconómica de la familia. Asimismo, Ramos (2025), encontró efectos significativos entre el tipo de institución, el estrato y los resultados en dichas pruebas, indicando que aquellos provenientes de un nivel socioeconómico alto y colegio privado tienden a tener mayor rendimiento, aspecto que se evidenció también en el análisis de Clúster desarrollado. Esto es también confirmado por Ballesteros-Alfonso y Gómez-Velasco (2022), quien encontró que mejores condiciones económicas y tecnológicas se correlacionan significativamente en todas las áreas evaluadas por las pruebas Saber 11.

6. CONCLUSIONES

Con lo expuesto anteriormente, se permite concluir que el acceso a las Tecnologías de la Información y la Comunicación (TIC), las condiciones socioeconómicas y los entornos educativos son factores interdependientes que determinan de manera significativa los resultados de los estudiantes en las pruebas Saber 11 entre 2008 y 2024. A pesar de la notable expansión en la conectividad a internet durante este período, que pasó de un 22.2% a un 73.4%, el análisis revela que la brecha digital no se ha cerrado, sino que se ha transformado; debido a que este estudio demuestra que la disponibilidad de infraestructura tecnológica es insuficiente si no va acompañada de condiciones socioeconómicas favorables que permitan el aprovechamiento de esta.

Los hallazgos del análisis de clústeres son contundentes al respecto: el grupo de estudiantes con mayor rendimiento (Clúster 2) no

solo presenta el más alto nivel socioeconómico (estrato promedio de 3), sino también un acceso casi universal a computador (98%) e internet (96%). En contraposición, el clúster más numeroso y de más bajo rendimiento (Clúster 1) evidencia una profunda exclusión digital en sentido de aprovechamiento de este como está ocurriendo con el aprendizaje y uso de la IA, con solo un 16% de acceso a computador y un 5% a internet, lo que se correlaciona directamente con sus bajos puntajes. Por lo tanto, este análisis longitudinal confirma que el capital tecnológico y el capital cultural del hogar son predictores del éxito académico que otros factores escolares.

Igualmente, la persistencia de una brecha de casi un 25% de estudiantes sin conectividad, sumada a las discrepancias en la intensidad de uso entre instituciones públicas y privadas, subraya que las desigualdades estructurales preexistentes se replican y amplifican en el mundo digital. El nuevo desafío para las políticas educativas no puede quedar solo en el acceso, tiene que centrarse también en la equidad, apuntando a cerrar las brechas de segundo y tercer nivel, relacionadas con las competencias para un uso adecuado de las TIC y los beneficios tangibles que de ello se derivan.

Si bien el presente estudio ofrece un análisis exhaustivo y longitudinal de los factores que inciden en los resultados de las pruebas Saber 11, es fundamental reconocer ciertas limitaciones inherentes a su diseño y alcance para futuras investigaciones académicas.

En primer lugar, la naturaleza correlacional del estudio, aunque permite identificar asociaciones significativas entre las variables, no autoriza el establecimiento de relaciones de causalidad directa. Así, factores no medidos, como la calidad pedagógica docente, el capital cultural familiar o la motivación intrínseca del estudiante, podrían estar influyendo en los resultados académicos y su interacción con el acceso a las TIC.

En segundo lugar, el análisis se basa en datos de autoinforme proporcionados por los estudiantes en los cuestionarios del ICFES, especialmente en lo que respecta a la posesión de dispositivos tecnológicos y los hábitos de uso de internet. Estos datos pueden estar sujetos a sesgos de deseabilidad social, a imprecisiones en la memoria o marcar aleatoriamente por falta de tiempo, lo que podría afectar la precisión de las mediciones sobre la brecha digital.

Para finalizar, aunque se utiliza un macroconjunto de datos que abarca un período de 17 años, las variables recopiladas por el ICFES han podido variar en su formulación y alcance a lo largo del tiempo, a pesar de que se usará el diccionario propio de las variables del ICFES. Si bien se realizó un riguroso proceso de extracción, transformación y carga (ETL) para normalizar y unificar los datos, es posible que existan matices en la definición de las variables entre diferentes cohortes que no se capturan completamente en el análisis.

REFERENCIAS BIBLIOGRÁFICAS

Álvarez, O. (2024). Essay on education: evidence from a developing country [Tesis de Doctorado, Universitat de Barcelona]. https://www.tdx.cat/bitstream/handle/10803/693349/OAV_PhD_THESIS.pdf?sequence=1&isAllowed=y

Arias, F. (2012). El proyecto de investigación: Introducción a la metodología científica (6ª ed.). Editorial Episteme. <https://abacoenred.org/wp-content/uploads/2019/02/El-proyecto-de-investigaci%C3%B3n-F.G.-Arias-2012-pdf-1.pdf>

Ariza, J. F., Saldarriaga, J. P., Reinoso, K. Y., y Tafur, C. D. (2021). Tecnologías de la información y la comunicación y desempeño académico en la educación media en Colombia. *Lecturas de Economía*, (94), 47-86. <https://doi.org/10.17533/udea.le.n94a338690>

Ballesteros-Alfonso, A. L., y Gómez-Velasco, N. Y. (2022). Desigualdad de resultados pruebas Saber-11 antes y durante la pandemia covid-19 (2014-2021). *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud*, 20(3), 46-68. <https://doi.org/10.11600/rllcsnj.20.3.5189>

Barrios Aguirre, F., Forero, D. A., Castellanos Saavedra, M. P., y Mora Malagón, S. Y. (2021). The impact of computer and internet at home on academic results of the saber 11 national exam in colombia. *SAGE Open*, 11(3), 21582440211040810. <https://doi.org/10.1177/21582440211040810>

Escobar, W. F. (2024). Análisis a la Brecha Digital en la Educación Colombiana: Un Estudio Documental [Tesis de Maestría, Corporación Universitaria Minuto de Dios]. <https://repository.uniminuto.edu/server/api/core/bitstreams/b94d92af-de3e-4349-9720-198b747c4b88/content>

García Sagrado, R. (2025). Desigualdad educativa y brecha digital: un análisis post-aceleración tecnológica. *Revista Investigación & Praxis En CS Sociales*, 4(1), 26–37. <https://doi.org/10.24044/ripcs.v4i1.4011>

Guo, P., Saab, N., Post, L. S., y Admiraal, W. (2020). A review of project-based learning in higher education: Student outcomes and measures. *International Journal of Educational Research*, 102, Artículo 101586. <https://doi.org/10.1016/j.ijer.2020.101586>

Hernández González, M., Ramos Quiroz, J. M., Chávez Maciel, F. J., y Trejo Cázares, M. C. (2024). Ventajas y riesgos de la Inteligencia Artificial Generativa desde la percepción de los estudiantes de educación superior en México. *European Public & Social Innovation Review*, 9, 1–18.

Hernández-Sampieri, R. y Mendoza, C. (2018). Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta. McGrawHillEducation.

Kaufman, L., y Rousseeuw, P. J. (2005). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley. <http://dx.doi.org/10.1002/9780470316801>

Kimball, R., y Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3ª ed.). Wiley. <https://ia801609.us.archive.org/14/items/the-data-warehouse-toolkit-kimball/The%20Data%20Warehouse%20Toolkit%20-%20Kimball.pdf>

Little, R., y Rubin, D. (2019). *Statistical Analysis with Missing Data* (3ª ed.). Wiley. <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119482260>

McKinney, W. (2010). Data structures for statistical computing in Python. *Scipy*, 445(1), 51-56. <https://doi.org/10.25080/Majora-92bf1922-00a>

McQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Computer and Chemistry*, 4, 257-272. <https://scispace.com/pdf/some-methods-for-classification-and-analysis-of-multivariate-4pswti19oz.pdf>

Melo-Becerra, L. A., Ramos-Forero, J. E., Arenas, J. L. R., y Zárate-Solano, H. M. (2021). Efecto de la pandemia sobre el sistema educativo: El caso de Colombia. *Borradores de Economía*, (1179), 1-56. <https://doi.org/10.32468/be.1179>

Miah, M. (2024). Digital inequality: the digital divide and educational outcomes. *ACETJCER*, 17(1). https://openurl.ebsco.com/EPDB%3Agcd%3A11%3A423099/detailv2?sid=ebsco%3Aplink%3Acrawler&id=ebsco%3Agcd%3A175598103&link_origin=www.google.com

Naciones Unidas. (2015). *Agenda 2030 para el desarrollo sostenible*. Naciones Unidas. (Citado como UN, 2015).

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2023). *Global education monitoring report, 2023: technology in education: a tool on whose terms?* UNESCO. <https://doi.org/10.54676/UZQV8501>

Organización para la Cooperación y el Desarrollo Económicos. (2020). *Are Students Ready to Thrive in an Interconnected World?* OECD. <https://www.oecd.org/pisa/publications/pisa-2018-results-volume-vi-d5f68679-en.htm>

Ortiz, O. (2015). Enfoques y métodos de investigación en las ciencias sociales. Ediciones de la U. https://www.researchgate.net/publication/315842152_Enfoques_y_metodos_de_investigacion_en_las_ciencias_humanas_y_sociales

Perezchica-Vega, J. E., Sepúlveda-Rodríguez, J. A., y Román-Méndez, A. D. (2024). Inteligencia artificial generativa en la educación superior: Usos y opiniones de los profesores. *European Public & Social Innovation Review*, 9, 1-20.

Ramos Sarmiento, H. M. (2025). Análisis del rendimiento académico en pruebas Saber 11° mediante un diseño factorial: influencia del estrato y naturaleza del colegio [Trabajo de Grado, Fundación Universitaria Los Libertadores]. <https://repository.libertadores.edu.co/items/4fd74460-c270-42a6-9cb5-e4d84443faac>

Rodríguez, J. L., y Gamboa, L. F. (2023). Access to ICTs during middle education: Who may benefit in the case of Colombia (Documentos de Trabajo Saber Investigar No. 4). Instituto Colombiano para la Evaluación de la Educación (ICFES). <https://www.icfes.gov.co/web/guest/saber-investigar>

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3), 417-453. <https://doi.org/10.3102/00346543075003417>

Stekhoven, D. J., y Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118. <https://doi.org/10.1093/bioinformatics/btr597>

Tan Choon Keong, y Gao JiMei. (2025). Analyzing Flipped Classroom Themes Trends in Computer Science Education (2007–2023) Using CiteSpace. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 46(1), 15–27. <https://doi.org/10.37934/araset.46.1.1527> (Citado como Tan Choon Keong et al., 2025).

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO.

Universidad Nacional de Colombia. (12 de abril de 2021). El daño irreparable de la pandemia en la educación colombiana. *Periódico UNAL*. <https://periodico.unal.edu.co/articulos/el-dano-irreparable-de-la-pandemia-en-la-educacion-colombiana>

Vanegas-Giraldo, J. (2025). Impacto de la IA en la Brecha Educativa en Colombia. *Revista Interamericana de Investigación, Educación y Pedagogía*, 18(2), 11-30. <https://doi.org/10.15332/25005421.10376>

Villegas-José, V., y Delgado-García, M. (2024). Inteligencia artificial: revolución educativa innovadora en la Educación Superior. *Pixel-Bit. Revista de Medios y Educación*, 71, 159–177.

Zabala, O. J., Archila, B. F. L., y Hernández, J. D. (2023). Documento de trabajo N°12: El impacto de las condiciones socioeconómicas en el rendimiento académico (ICFES SABER 11) de los estudiantes de Norte de Santander. *Observatorio Socioeconómico Regional de la Frontera*. https://www.unipamplona.edu.co/unipamplona/portalIG/home_72/recursos/01_general/18122014/pag_documentos.jsp