Identificación de Perfiles de Rendimiento Académico mediante Clustering en la Pontificia Universidad Javeriana y la Universidad Nacional de Colombia: Un Análisis con Saber 11 y Saber Pro

Trabajo de grado

Diego Orlando Caballero Sarmiento

Tutor: Néstor Armando Nova Arévalo
CoTutora: Andrea González Sandoval

Pontificia Universidad
Javeriana Facultad de Ciencias
Pregrado en Ciencia de datos
Bogotá D.C
2026

Resumen

Este estudio analiza el desempeño académico de estudiantes de programas de Administración y áreas afines de la Pontificia Universidad Javeriana y la Universidad Nacional de Colombia, incluyendo sus sedes de Cali y Palmira respectivamente. La unidad de análisis corresponde a estudiantes que presentaron tanto la prueba Saber 11 entre 2014 y 2017 como la prueba Saber Pro entre 2018 y 2022, lo que permite evaluar la evolución del desempeño a lo largo de la trayectoria educativa de cada estudiante. La integración de ambas fuentes se realizó a través del archivo de llaves de vinculación del ICFES, siguiendo la metodología CRISP-DM, y resultó en una muestra final de aproximadamente 1.700 registros correspondientes al área de interés, depurada a partir de más de siete millones de registros iniciales.

Para identificar perfiles de desempeño, se compararon seis algoritmos de agrupamiento no supervisado construidos exclusivamente a partir de resultados académicos, con el fin de evitar sesgos: K-Means, Gaussian Mixture, Fuzzy C-Means, Clustering Aglomerativo, DBSCAN y Spectral Clustering. Aunque Spectral Clustering presentó las métricas internas más favorables coeficiente de Silhouette de 0.4123 e índice de Davies-Bouldin de 0.5706 fue descartado por su rendimiento excesivamente bajo en Calinski-Harabasz de 19.2 y su representación gráfica evidenció que concentraba prácticamente toda la muestra en un único clúster masivo con apenas unos pocos puntos aislados como grupos separados, lo que indica que optimizó las métricas sin producir una segmentación real y pedagógicamente útil. Se seleccionó K-Means con cuatro clústeres por obtener el mayor índice de Calinski-Harabasz (394.65) entre los algoritmos con particiones balanceadas, y por generar segmentos estables, coherentes e interpretables desde el punto de vista educativo, denominados Inicial, Intermedio, Avanzado y Superior. La estabilidad de esta solución fue verificada mediante un análisis de sensibilidad, al variar el número de clústeres entre k=3 y k=6, los valores de Silhouette fueron 0.1594, 0.1486, 0.1262 y 0.1196 respectivamente,

confirmando que k=4 ofrece el mejor equilibrio entre calidad métrica e interpretativa para los objetivos del estudio.

Los resultados evidencian que el desempeño en Saber 11 es un fuerte predictor del rendimiento en Saber Pro, así como una persistencia en las trayectorias académicas de los estudiantes. Se identificó además un descenso generalizado en los puntajes entre ambas pruebas en todos los perfiles. De manera complementaria, se evaluó la asociación entre la pertenencia a los clústeres y variables descriptivas mediante la V de Cramér, identificando algunas variables con mayor relación relativa que contribuyen a la caracterización de los perfiles encontrados. Estos resultados permiten identificar segmentos de estudiantes con características distintivas.

Tabla de contenido

1. Introducción 7
1.1 Planteamiento del problema 7
1.2 Justificación 8
1.3 Estado del arte 9
Objetivos 13
Objetivos específicos: 13
2. Marco Teórico 15
2.1 Educación Superior en Colombia 15
2.1.1 Contexto de investigación 16
2.2 Evaluación de la calidad educativa en Colombia 17
2.2.1 Pruebas estandarizadas en Colombia: Saber 11 y Saber Pro 17
2.2.2 Limitaciones y alcances de las pruebas estandarizadas 19
2.3 Fundamentos del Clustering 20
2.3.1 Aprendizaje No Supervisado 20
2.3.2 Análisis de Clusters 21
2.3.3 Fundamentos matemáticos del cluster 22
2.3.4 Métodos de Agrupamiento 23
2.3.5 Kmeans 24
2.3.6 Fuzzy C-Means 24
2.3.7 Modelo Gaussian Mixture 25
2.3.8 DBSCAN 26
2.3.9 Clustering Jerárquico 27
2.3.10 Spectral Clustering 28
2.4 Cramér's V 29
3. Metodología 31
4. Aplicación de la metodología 32
4.1 Entendimiento del negocio 32
4.1.1 Protección de datos personales y anonimización en el sistema educativo 34
4.2 Entendimiento de los datos 35
4.2.1 Datos de la prueba Saber 11 36
4.2.2 Datos de la prueba Saber Pro 37
4.2.3 Articulación entre las bases de datos 39
4.3 Preparación de los datos 39
4.3.1 Integración de las bases de datos Saber 11 y Saber Pro 39
4.3.2 Proceso de filtrado y vinculación 40
4.3.3 Consideraciones técnicas 43
4.3.4 Base de datos unida 44
4.4 Modelado 44
4.4.1 Algoritmos evaluados e hiperparámetros 45
4.5 Evaluación 47

4.6 Despliegue 51
5. Resultados preliminares 51
5.1 Análisis exploratorio 52
5.1.1 Resultados preliminares de la prueba Saber 11 52
5.1.2 Resultados preliminares de la prueba Saber Pro 55
6. Resultados Finales 59
6.1 Clústeres obtenidos únicamente para el área de administración y afines 59
6.2 Perfilamiento posterior al agrupamiento 72
6.3 Análisis de variables seleccionadas 73
6.3.1 Análisis de núcleo de pregrado de los estudiantes del clúster 74
6.3.2 Análisis del departamento donde está ubicado el colegio del estudiante 76
y donde presentó el examen Saber 11 76
6.3.3 Análisis del tipo de origen de institución universitaria 78
6.3.4 Análisis del nivel educativo del padre y la madre 80
6.3.5 Análisis complementario del perfil socioeconómico familiar por cluster 80
6.4 Discusión 85
6.5 Conclusiones 88
7. Anexos 90
Referencias 97

Índice de figuras

Figura 2. Fases de la metodología CRISP-DM. Fuente: Chapman et al. (2000) 33
Figura 3. Proceso de vinculación. 42
Figura 4. Flujo de la muestra desde los registros iniciales hasta la base analítica final. 43
Figura 5. Método de evaluación. 52
Figura 6. Distribución de puntaje Saber 11.. 54
Figura 7. Comparación de puntajes por módulo Saber 11 55
Figura 8. Cantidad de estudiantes por cada universidad en la muestra del estudio. 56
Figura 9. Comparación del puntaje promedio en Saber Pro. 58
Figura 10. Distribución por cada módulo en prueba Saber Pro. 59
Figura 11. Clústeres obtenidos con Spectral. 64
Figura 12. Clúster obtenido con K-means. 65
Figura 13. Centroides obtenidos con K-means. 67
Figura 14. Comparativa entre puntaje de cada cluster. 70
Figura 15. Matriz de transición de saber 11 a saber Pro. 71
Figura 16. Cantidad de estudiantes de cada carrera en cada clúster. 76
Figura 17. Cantidad de estudiantes de cada departamento en cada clúster. 78
Figura 18. Estudiantes por clúster de universidad privada o pública. 80
Figura 19. Nivel educativo del padre 82
Figura 20. Nivel educativo de la madre. 83
Figura 21. Perfil socioeconómico familiar por clúster. 85

Índice de tablas

Tabla 1. Primer paso del método Cramer's V 30
Tabla 2. Hiperparametros usados por cada algoritmo. 46
Tabla 3. Comparación de métodos. Nota. todo el conjunto de datos estandarizado. 6
Tabla 4. Distribución de los clústeres 65
Tabla 5. Variables seleccionadas por valor de Cramer 's V. 73
Tabla 6. Proporción por núcleo de pregrado según clúster 75

1. Introducción

1.1 Planteamiento del problema

El análisis del desempeño académico de los estudiantes universitarios constituye un desafío relevante dentro del sistema de educación superior en Colombia. Si bien existen pruebas estandarizadas como Saber 11 y Saber Pro, aplicadas por el ICFES para evaluar competencias en distintos momentos de la trayectoria académica, todavía está en proceso conocer sobre los factores académicos, sociodemográficos e institucionales que inciden de manera específica en los resultados de los estudiantes de la Pontificia Universidad Javeriana y Universidad Nacional de Colombia en la prueba Saber Pro.

La falta de estudios sistemáticos y profundos sobre esta población limita la comprensión de patrones de desempeño, desigualdades internas o variables explicativas de los resultados. En consecuencia, las universidades disponen de evidencia general sobre su contribución a la formación, pero con menor información focalizada que les permita orientar con mayor precisión sus procesos de formación y estrategias institucionales de mejora.

Un elemento adicional es que, si bien existen en Colombia investigaciones que vinculan los resultados de Saber 11 y Saber Pro a nivel individual, son menos frecuentes los trabajos que adoptan un enfoque exploratorio orientado a identificar perfiles internos de trayectoria académica dentro de una misma área de formación, y aún menos los que combinan esta caracterización con un análisis posterior de asociación con variables sociodemográficas, socioeconómicas e institucionales. Esta perspectiva permite comprender de forma más detallada cómo las condiciones de ingreso, el contexto socioeconómico, las trayectorias académicas y las características institucionales se reflejan en los desempeños de egreso, complementando la evidencia ya disponible con una mirada más centrada en la heterogeneidad interna de la población estudiantil en la Universidad Javeriana y Universidad Nacional de Colombia.

En este trabajo se aborda esta limitación mediante el uso de dos bases de datos complementarias, Saber 11 y Saber Pro, que permiten caracterizar de manera integral la trayectoria de los estudiantes de la Pontificia Universidad Javeriana y la Universidad Nacional de Colombia. Con más de 50 variables académicas, sociodemográficas e institucionales disponibles, se busca identificar patrones de desempeño y asociaciones entre estas variables y los perfiles de rendimiento académico. El análisis se restringe a los módulos genéricos del examen Saber Pro Lectura Crítica, Razonamiento Cuantitativo, Competencias Ciudadanas, Comunicación Escrita e Inglés con el fin de garantizar una medición homogénea y comparable entre todos los estudiantes independientemente de su programa académico, dado que los módulos específicos varían según el campo de formación y no permiten una comparación directa entre áreas.

En este contexto, el estudio busca responder las siguientes preguntas: ¿qué patrones de desempeño académico se pueden identificar en los estudiantes de Administración y afines de las 2 universidades analizadas a partir de sus resultados en Saber 11 y Saber Pro?, ¿Qué tan consistentes son los perfiles de desempeño académico al comparar los resultados de los estudiantes en Saber 11 y Saber Pro?, y ¿qué asociación presentan las variables sociodemográficas, socioeconómicas e institucionales con los perfiles de desempeño identificados?

1.2 Justificación

Comprender los factores asociados al rendimiento académico en la prueba Saber Pro es fundamental para que la Pontificia Universidad Javeriana y Universidad Nacional de Colombia puedan diseñar políticas internas que fortalezcan la calidad educativa y promuevan la equidad en la formación de sus estudiantes. Sin embargo, este análisis no puede limitarse únicamente a la medición de puntajes finales, sino que requiere un enfoque integral que tenga en cuenta la trayectoria del estudiante, sus condiciones de ingreso y su contexto socioeconómico.

Este trabajo representa una contribución con un enfoque distinto, ya que integra información de las pruebas Saber 11 y Saber Pro, permitiendo trazar un recorrido completo desde el momento en que los estudiantes finalizan la educación media hasta el cierre de su formación universitaria. Esta perspectiva longitudinal ofrece una oportunidad única para identificar variables explicativas de los resultados que trascienden los puntajes, incorporando factores académicos, sociodemográficos, socioeconómicos e institucionales.

La Ciencia de Datos constituye un enfoque idóneo para este propósito, pues posibilita el análisis de grandes volúmenes de información y la detección de patrones no visibles con métodos tradicionales. A través de técnicas estadísticas y métodos de agrupamiento no supervisado, este trabajo identifica patrones de desempeño y segmenta grupos de estudiantes mediante clustering, complementa este análisis con medidas de asociación entre los perfiles obtenidos y variables sociodemográficas, socioeconómicas e institucionales.

En conjunto, esta investigación genera evidencia aplicable a las 2 universidades analizadas, cuyas características institucionales y perfiles estudiantiles particulares limitan la generalización directa de los hallazgos a la del sistema educativo colombiano general. No obstante, la metodología de Saber 11 y Saber Pro mediante clustering y análisis de variables externas puede servir como referente para estudios similares en otras instituciones del país o incluso podrían aportar elementos para la comprensión más general del sistema de educación superior en Colombia, como los perfiles de ingreso, la diversidad regional y otras características asociadas al contexto de los estudiantes.

1.3 Estado del arte

El estudio del rendimiento académico en pruebas estandarizadas ha sido abordado desde diferentes perspectivas en Colombia y en el ámbito internacional. A continuación, se presentan algunos trabajos relevantes y sus aportes en relación con esta investigación:

Gallego Vega (2021), en su estudio sobre estudiantes de la Universidad Nacional, identifica brechas significativas de desempeño en Saber Pro según facultades y condiciones sociodemográficas; su principal fortaleza radica en el análisis detallado a nivel institucional, combinando datos académicos y contextuales, aunque se limita a una sola universidad, lo que restringe la posibilidad de generalización; no obstante, su contribución es valiosa al ofrecer evidencia empírica sobre la incidencia de factores sociodemográficos en el rendimiento en educación superior, constituyendo un referente metodológico útil y aplicable al análisis de la población de estudio en este proyecto [1].

Chaves-Manzano y Ordoñez-López (2020), critican la medición de la calidad educativa basada exclusivamente en pruebas estandarizadas, señalando que estas tienden a invisibilizar aspectos pedagógicos e institucionales; su principal fortaleza es ofrecer un marco teórico alternativo que cuestiona el rol dominante de los exámenes en la evaluación, aunque no desarrollan propuestas metodológicas concretas para superar dichas limitaciones; su contribución radica en destacar la necesidad de complementar las pruebas con el análisis de factores institucionales, lo cual respalda la inclusión de estas variables en el análisis de este trabajo [2].

OCDE (2020), a través del Informe PISA, hace un estudio comparativo entre países cómo las desigualdades socioeconómicas inciden en el rendimiento en pruebas estandarizadas; destaca por el uso de bases de datos masivas y metodologías de minería de datos que permiten identificar patrones globales, aunque su escala internacional dificulta la aplicación directa de sus conclusiones a contextos locales; su principal contribución es demostrar el potencial de la Ciencia de Datos para detectar brechas y orientar políticas educativas, lo que inspira el uso de técnicas avanzadas de análisis en el estudio de la educación superior colombiana [3].

Demarchi Sánchez (2023), en su estudio en el Tecnológico de Antioquia, emplea un enfoque cualitativo para identificar factores académicos, sociales y económicos que inciden en el

desempeño en Saber Pro; su aporte subraya la importancia de considerar factores externos al ámbito académico y respalda la inclusión de variables sociodemográficas e institucionales en el análisis, sirviendo como base para contrastar con el enfoque cuantitativo propuesto en este trabajo [4].

Nafuri et al. (2022), Aplican técnicas de clustering no supervisado como K-Means o DBSCAN para clasificar el rendimiento académico de más de 117.000 estudiantes en Malasia, utilizando variables demográficas, académicas y de comportamiento. Los modelos fueron evaluados mediante índices de Davies-Bouldin, Silhouette y Calinski-Harabasz, los mismos empleados en este estudio. Los resultados muestran que K-Means supera consistentemente a los demás algoritmos, destacándose por su estabilidad y desempeño en datos de alta dimensionalidad, lo que respalda su uso en este trabajo [16].

Balcázar y Ñopo (2016), en su estudio "Broken Gears", analizan el valor agregado de la educación superior en el rendimiento académico de los futuros docentes colombianos, comparando los puntajes obtenidos por los mismos individuos en Saber 11 (2002-2007) y Saber Pro (2011-II) mediante z-scores y modelos que controlan por sesgo de selección; su fortaleza radica en el uso de una metodología longitudinal individualizada y en el reconocimiento explícito de las diferencias entre estudiantes de pedagogía y sus pares de otras carreras, aunque sus conclusiones se limitan al ámbito de la formación docente; su principal contribución para este trabajo es demostrar la viabilidad metodológica de vincular ambas pruebas a nivel individual para evaluar trayectorias académicas, así como evidenciar que el valor agregado no es uniforme entre programas, lo que refuerza la decisión metodológica de restringir el presente análisis a un área específica de formación [29].

Gómez Soler, Bernal Nisperuza y Herrera Idárraga (2020), profesoras del Departamento de Economía de la Pontificia Universidad Javeriana, analizan la relación entre las actividades de

preparación para exámenes y el desempeño en la prueba Saber 11 en Colombia, utilizando técnicas econométricas sobre los resultados del examen de Estado para estimar el efecto neto de la preparación previa sobre los puntajes obtenidos; su fortaleza radica en el uso de una metodología cuantitativa rigurosa aplicada a datos nacionales y en encontrar que, en promedio, las actividades de preparación están asociadas a un incremento de aproximadamente 0.06 desviaciones estándar en los puntajes, aunque el estudio se limita a la prueba Saber 11 y no incorpora seguimiento longitudinal hasta Saber Pro; su contribución a este trabajo está en evidenciar que las condiciones previas a la presentación del examen como la preparación, la trayectoria escolar y el contexto influyen significativamente en el desempeño estandarizado, lo que respalda la decisión metodológica de incorporar los puntajes de Saber 11 como insumo para caracterizar trayectorias académicas en educación superior [30].

Burgos Flórez, Ortiz Benavides, Rúales Suárez y Bastidas García (2019), calculan el valor agregado generado por la Universidad de Nariño mediante el cruce de resultados Saber 11 y Saber Pro entre 2010 y 2014; su fortaleza es la aplicación institucional concreta del marco metodológico de valor agregado a una universidad regional específica, su contribución a este trabajo es evidenciar la utilidad de los estudios institucionales focalizados como referencia para que cada universidad comprenda su propia contribución a la formación de sus estudiantes, lo cual respalda el enfoque institucional adoptado en esta investigación [31].

Estos estudios evidencian avances importantes en el análisis longitudinal del desempeño académico en Colombia, particularmente en torno al concepto de valor agregado, que ha sido abordado desde enfoques econométricos, multinivel y cuasi-experimentales. La integración de las pruebas Saber 11 y Saber Pro a nivel individual no es, por tanto, una práctica metodológica única en el país. Sin embargo, los trabajos existentes se han concentrado mayoritariamente en estimar la contribución institucional al rendimiento o en comparar instituciones mediante promedios y

modelos paramétricos, mientras que el uso de técnicas de aprendizaje no supervisado para identificar perfiles de desempeño a partir de la trayectoria individual ha sido poco explorado. Asimismo, es en este espacio donde se ubica el aporte diferencial del presente trabajo: a diferencia de los estudios de valor agregado que estiman la contribución institucional, esta investigación adopta un enfoque exploratorio y descriptivo orientado a la identificación de perfiles homogéneos de trayectoria académica mediante clustering no supervisado en el área de formación específica de Administración y afines, complementado con una validación externa que permite caracterizar contextualmente cada grupo. La contribución no reside, por tanto, de manera principal en vincular ambas pruebas, sino en la perspectiva analítica puesta y en el tipo de evidencia que ofrece para comprender la población estudiantil estudiada.

Objetivos

Objetivo general: Identificar perfiles de desempeño académico en estudiantes de programas de Administración y áreas afines de la Pontificia Universidad Javeriana y la Universidad Nacional de Colombia, con sus sedes en Cali y Palmira respectivamente, mediante técnicas de clustering, integrando desempeño en Saber 11, Saber Pro y variables sociodemográficas para caracterizar grupos homogéneos y comparar patrones de rendimiento.

Objetivos específicos:

- Construir, depurar y documentar la base de datos, seleccionando las variables académicas, sociodemográficas e institucionales más relevantes para el análisis de la trayectoria educativa de los estudiantes de ambas universidades.
- Modelar el desempeño académico de la población estudiantil mediante modelos de aprendizaje automático (clustering), con el fin de identificar patrones de desempeño.
- Analizar la composición de cada clúster identificado, describiendo sus

características predominantes.

- Evaluar las diferencias de los clusters, características institucionales, sociodemográficas o económicas, identificando patrones únicos o comunes a cada uno.

- Comunicar los hallazgos a través de visualizaciones y reportes analíticos que presenten los perfiles de desempeño identificados y las asociaciones encontradas con variables sociodemográficas, socioeconómicas e institucionales.

2. Marco Teórico

Para comprender a profundidad los patrones y factores que inciden en el rendimiento académico de los estudiantes de la Universidad Nacional de Colombia y Universidad Javeriana en las pruebas Saber Pro, es esencial establecer con claridad los conceptos fundamentales que orientan esta investigación. A continuación, se presentan las definiciones y enfoques teóricos más relevantes que permitirán contextualizar y analizar el fenómeno estudiado.

2.1 Educación Superior en Colombia

La educación superior en Colombia ha sido moldeada por una evolución normativa y estructural que comenzó a consolidarse desde finales del siglo XX. La creación del ICFES en 1968 y la implementación del examen de Estado en 1980 (antecedente directo de la actual Prueba Saber 11) marcaron los primeros pasos hacia la regulación y estandarización del sistema. Con la Constitución de 1991 y la promulgación de la Ley 30 de 1992, se establecieron principios clave como la autonomía universitaria, la coexistencia entre instituciones públicas y privadas, y la creación de mecanismos de evaluación como el Sistema Nacional de Acreditación y el Registro Calificado [5].

El Sistema Nacional de Acreditación garantiza que las instituciones de educación superior cumplan con altos estándares de calidad mediante procesos de evaluación y reconocimiento público. El Registro Calificado, por su parte, autoriza la oferta de programas académicos tras verificar condiciones mínimas de calidad en aspectos como currículo, infraestructura y planta docente [6].

2.1.1 Contexto de investigación

El sistema de educación superior en Colombia integra instituciones públicas y privadas bajo la supervisión del Ministerio de Educación Nacional, mediante mecanismos como el registro calificado y la acreditación para garantizar la calidad. Aunque comparten fines académicos y de investigación, difieren en su estructura: las públicas se financian principalmente con recursos estatales y promueven políticas de acceso, mientras las privadas dependen de matrículas y recursos propios, lo que refleja avances y brechas persistentes [6].

El sistema enfrenta tres tensiones principales: expandir cobertura manteniendo calidad educativa, garantizar sostenibilidad financiera, y asegurar pertinencia regional de la oferta académica. Aunque instituciones como el ICETEX han logrado ampliar el acceso estudiantil, persisten desafíos significativos en equidad y calidad según organismos internacionales como la OCDE y el Banco Mundial [6].

La Universidad Nacional representa el modelo público con misión de excelencia académica con acceso equitativo y servicio a la nación a través de múltiples sedes regionales [7]. La Pontificia Universidad Javeriana ejemplifica el modelo privado con enfoque en excelencia académica, por su alta calidad humana, ética, profesional y por su responsabilidad social [8]. Ambas instituciones ilustran vocaciones distintas que se reflejan en su composición estudiantil, énfasis investigativos y estrategias de relacionamiento con el entorno.

La inclusión de la Pontificia Universidad Javeriana Cali permite una visión más amplia del desempeño institucional, al compartir la identidad y el prestigio académico de la Javeriana pero en otra región. Su enfoque se orienta a la formación integral de profesionales con sensibilidad social, lo que la convierte en un referente importante dentro del ámbito educativo privado en la región [9]. Por su parte, la Universidad Nacional de Colombia sede Palmira representa una extensión del modelo público de excelencia académica y equidad en el acceso en el contexto regional [10].

La selección de estas universidades responde a un criterio metodológico orientado a garantizar comparabilidad académica y diversidad institucional. Estas instituciones permiten contrastar contextos educativos distintos al incluir tanto universidades privadas como públicas de alta calidad académica y reconocimiento nacional, así como sedes ubicadas en diferentes regiones del país, lo que introduce variabilidad socioeconómica, cultural y educativa relevante para el análisis.

2.2 Evaluación de la calidad educativa en Colombia

La evaluación educativa en Colombia cumple una función crucial como mecanismo regulador y de mejora continua. Más allá de calificar estudiantes, se ha consolidado como una herramienta esencial para verificar condiciones de calidad, gestión institucional y cumplimiento de normas. La evaluación en sus modalidades interna (autoevaluación, evaluación institucional) y externa (como las pruebas estandarizadas) permite diagnosticar fortalezas y debilidades, orientar estrategias pedagógicas, garantizar rendición de cuentas y promover la mejora. En este sentido, diversos autores coinciden en que la calidad educativa solo puede perseguirse con sistemas de evaluación sólidos, equitativos y sistemáticos, que involucren múltiples actores en procesos transparentes y formativos, no meramente sancionatorios [11].

2.2.1 Pruebas estandarizadas en Colombia: Saber 11 y Saber Pro

La prueba Saber 11 es un examen estandarizado que evalúa a los estudiantes al finalizar la educación media y constituye un requisito obligatorio para acceder a la educación superior. Su objetivo es valorar competencias fundamentales en áreas clave: Lectura Crítica, Matemáticas, Ciencias Naturales, Sociales y Ciudadanas, e Inglés. Para este estudio, se utilizaron los resultados de Saber 11 correspondientes al periodo 2014-2017, es decir, cinco años previos a los datos de Saber Pro (2018-2022), con el fin de aproximar el historial académico de los estudiantes, considerando que la duración típica de los programas de educación superior es de cinco años. Es importante destacar que la prueba solo cambió de 268 a las 278 preguntas, distribuidas en dos

sesiones: 131 preguntas en la primera y 147 en la segunda, aplicadas el mismo día con una duración total de 9 horas. Cada una de estas áreas está diseñada para medir habilidades como la comprensión, el razonamiento y la aplicación de conocimientos en contextos cotidianos. Esta prueba no solo permite a las instituciones de educación superior establecer criterios de admisión, sino que también ofrece un panorama general del estado de la educación media en el país, constituyéndose en un insumo fundamental para las políticas públicas y la toma de decisiones en materia educativa [12].

Por su parte, la prueba Saber Pro es un requisito de grado para todas las universidades que está orientado hacia los estudiantes próximos a culminar sus estudios universitarios, mientras que la Universidad Javeriana se rige por la norma estándar del ICFES que permite presentar el examen Saber Pro al completar el 75% de los créditos, la Universidad Nacional de Colombia exige, bajo su propio estatuto interno, que el estudiante haya aprobado al menos el 80% de su plan de estudios para autorizar la inscripción. Esta prueba busca identificar competencias genéricas que deben caracterizar a un profesional, independientemente de su programa académico. Dentro de cada uno de los módulos, además del puntaje, se reporta un nivel de desempeño que puede alcanzar hasta los 300 puntos. En el año 2020, debido a la pandemia, sólo se aplicaron los módulos genéricos, mientras que a partir de 2021 se retomó también la aplicación de los módulos específicos. Los módulos genéricos incluyen Lectura Crítica (35 preguntas), Razonamiento Cuantitativo (35 preguntas), Competencias Ciudadanas (35 preguntas), Comunicación Escrita (1 pregunta) e Inglés (55 preguntas), además de un cuestionario socioeconómico que no se califica. Dentro de estos, la evaluación se organiza en porcentajes específicos: en Lectura Crítica, un 33 % de las preguntas corresponden a interpretación y representación, otro 33 % a formulación y ejecución, y el 34 % restante a argumentación. En Competencias Ciudadanas, la prueba se divide en conocimientos (30 %), argumentación (20 %), multiperspectivismo (30 %) y pensamiento sistémico (20 %). Finalmente, el módulo de Inglés se compone de siete partes: las secciones 1, 2, 3 y 6 representan

cada una el 11 % del total; la sección 4, el 18 %; la sección 5, el 16 %; y la sección 7, el 22 % [12]. En este estudio no se tendrá en cuenta la puntuación de los módulos específicos de la prueba, debido a que el análisis se centra exclusivamente en los componentes genéricos. Esto permite garantizar una medición homogénea y comparable entre estudiantes de distintas carreras, dado que los módulos específicos varían según el área de formación y no son equivalentes entre sí.

2.2.2 Limitaciones y alcances de las pruebas estandarizadas

Las pruebas estandarizadas como Saber 11 y Saber Pro, aunque útiles para el diagnóstico educativo y la rendición de cuentas, presentan importantes limitaciones conceptuales y metodológicas. Estas evaluaciones imponen modelos educativos iguales que no consideran las condiciones socioculturales reales de la población colombiana, especialmente las limitaciones de recursos, infraestructura y capacitación docente que caracterizan contextos más vulnerables. Además, estas pruebas pueden actuar como mecanismos de diferenciación que tienden a reflejar desigualdades sociales y económicas, lo que en ocasiones impacta de manera más evidente a estudiantes. Al no considerar completamente las condiciones iniciales de los evaluados, sus resultados pueden ofrecer sólo una visión parcial [13].

El enfoque estandarizado presenta también algunas limitaciones pedagógicas que vale la pena reconocer. Al centrarse en competencias medibles y comparables, estas pruebas pueden no capturar en su totalidad habilidades transversales como el pensamiento crítico, la creatividad o el trabajo colaborativo, que forman parte fundamental de la formación integral. En Colombia, además de las pruebas Saber 11 y Saber Pro, se han implementado evaluaciones en grados intermedios como Saber 3°, 5° y 9°, aunque estas han enfrentado dificultades de continuidad metodológica en algunos periodos, lo que ha limitado la consolidación de una línea histórica estable de seguimiento del desempeño. Estos elementos no invalidan la utilidad de las pruebas estandarizadas como herramientas de diagnóstico y política educativa, sino conducen a complementarlas con otros

instrumentos y enfoques que permitan una comprensión más integral de los procesos formativos y del contexto socioeconómico en el que se desarrollan los estudiantes. [13].

2.3 Fundamentos del Clustering

En esta sección se presentan los fundamentos teóricos y conceptuales que sustentan el desarrollo del presente estudio. Se exponen los conceptos fundamentales necesarios para comprender las técnicas de análisis de clusters empleadas en esta investigación. Se presentan los principios del aprendizaje no supervisado, la formulación matemática de los métodos de agrupamiento, así como las métricas utilizadas para evaluar la calidad de los resultados. Estos elementos constituyen la base metodológica para el análisis del desempeño académico de los estudiantes [14].

2.3.1 Aprendizaje No Supervisado

El aprendizaje automático (Machine Learning) es una rama de la inteligencia artificial orientada al desarrollo de métodos capaces de identificar patrones y construir modelos a partir de datos. Según la disponibilidad de una variable objetivo, estos métodos se clasifican en aprendizaje supervisado, no supervisado y por refuerzo [14].

A diferencia del aprendizaje supervisado, el aprendizaje no supervisado analiza datos sin etiquetas previas, con el objetivo de descubrir estructuras ocultas, relaciones internas y patrones naturales. Para ello, los algoritmos identifican similitudes entre observaciones mediante medidas de distancia o criterios de semejanza, permitiendo agrupar individuos con características similares o detectar comportamientos diferenciados dentro de una población [14].

En el ámbito educativo, este enfoque facilita el análisis del desempeño académico desde una perspectiva exploratoria, identificando perfiles de estudiantes a partir de los resultados en pruebas estandarizadas. Entre sus técnicas más representativas se encuentra el análisis de clusters, que será abordado como metodología central del presente estudio.

2.3.2 Análisis de Clusters

El análisis de clusters, también conocido como análisis de agrupamiento, constituye una de las técnicas principales del aprendizaje no supervisado y tiene como objetivo identificar grupos homogéneos dentro de un conjunto de datos. Este método busca organizar las observaciones en clusters de tal manera que los individuos pertenecientes a un mismo grupo presenten características similares entre sí, mientras que aquellos ubicados en diferentes grupos muestran diferencias [15].

La formación de clusters se basa en la comparación de las observaciones a partir de sus variables académicas, utilizando medidas de distancia o similitud que permiten cuantificar qué tan cercanos o diferentes son los individuos dentro del espacio de datos. De esta forma, el análisis de agrupamiento permite descubrir estructuras latentes que no son evidentes mediante análisis descriptivos tradicionales [15].

Existen diversos enfoques para realizar el análisis de clusters, entre los que destacan los métodos particionales, los métodos jerárquicos, los algoritmos basados en densidad y los modelos probabilísticos. Cada uno de estos enfoques presenta diferentes supuestos y estrategias para la formación de grupos, dependiendo de la naturaleza de los datos y del objetivo del análisis [15].

En educación, el análisis de clusters identifica perfiles de estudiantes con desempeños similares mediante variables de pruebas estandarizadas. Facilita el reconocimiento de patrones de rendimiento y segmentaciones sin clasificaciones previas, convirtiéndose en una herramienta adecuada para estudiar la relación entre los resultados de las pruebas Saber 11 y Saber Pro, objetivo de esta investigación.

Un paso previo al clúster es la estandarización de las variables de entrada. Dado que los métodos de agrupamiento basados en distancias como K-Means dependen de la magnitud absoluta de las variables, aquellas con valores numéricamente mayores tienden a dominar el cálculo de las distancias entre observaciones, lo que distorsiona la formación de los grupos. Para evitar este sesgo, se aplica una transformación que homogeniza la escala de todas las variables, asegurando que cada

una contribuya en igualdad de condiciones al análisis [15].

La técnica más utilizada para este propósito es la estandarización Z-score, implementada en bibliotecas como StandardScaler de scikit-learn [28]. Esta transformación se aplica de forma independiente a cada variable y consiste en restar la media y dividir por la desviación estándar:

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}$$

Donde:

- x_{ik} es el valor original de la variable k para la observación i
- μ_k es la media de la variable $k$calculada sobre todas las observaciones
- σ_k es la desviación estándar de la variable k
- z_{ik} es el valor estandarizado resultante

Cada variable presenta media cero y desviación estándar uno, independientemente de su escala original. Es importante destacar que la estandarización se aplica columna por columna, es decir, cada variable se transforma usando su propia media y desviación, sin mezclar información entre variables. Esto preserva las relaciones internas entre observaciones mientras elimina las diferencias de escala [15].

2.3.3 Fundamentos matemáticos del clúster

El análisis de clústeres se fundamenta en la representación matemática de las observaciones dentro de un espacio multidimensional definido por sus variables características. Sea un conjunto de datos compuesto por n observaciones, donde cada individuo puede representarse mediante un vector $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, siendo p el número de variables consideradas, cada observación corresponde a un punto dentro de un espacio de características p-dimensional. Bajo esta representación, el conjunto completo de datos puede expresarse mediante una matriz $X \in \mathbb{R}^{nxp}$. La

similitud entre observaciones se evalúa comúnmente mediante medidas de distancia, siendo la distancia euclidiana una de las más utilizadas, definida como [15]:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

2.3.4 Métodos de Agrupamiento

En temas educativos, el clustering es útil cuando se busca descubrir estructuras latentes en los datos sin imponer categorías previas, especialmente en poblaciones donde el desempeño no sigue divisiones naturales evidentes. Su valor radica en permitir una segmentación exploratoria de estudiantes con características similares de rendimiento, facilitando la identificación de perfiles diferenciados que pueden llegar a orientar decisiones pedagógicas e institucionales sin requerir etiquetas o clasificaciones predefinidas. Estudios previos han demostrado la utilidad de estas técnicas como K-means, DBSCAN o Fuzzy C-Means en educación, aplicando análisis de clustering para clasificar el rendimiento académico de estudiantes universitarios, evidenciando que los métodos no supervisados permiten detectar agrupaciones significativas a partir de variables de desempeño [16]. De manera similar en Colombia, se han empleado aproximaciones de segmentación para identificar brechas de rendimiento en la Universidad Nacional de Colombia a partir de resultados en pruebas estandarizadas [1]. No obstante, estos métodos no deben interpretarse como evidencia causal ni como una clasificación definitiva de capacidades, sino como una aproximación exploratoria a estructuras presentes en los datos, cuyos resultados deben leerse siempre en función del contexto institucional y socioeconómico de la población analizada.

Por lo tanto, en esta parte se describen los principales métodos considerados en la investigación, presentando su formulación matemática y el procedimiento mediante el cual identifican estructuras dentro de los datos.

2.3.5 Kmeans

El algoritmo K-means es un método de agrupamiento particional cuyo objetivo consiste en dividir un conjunto de n observaciones en K clusters previamente definidos. Cada grupo se representa mediante un centroide, el cual corresponde al promedio de las observaciones pertenecientes al clúster. El algoritmo busca formar grupos compactos y bien separados, minimizando la variabilidad interna de cada clúster [15].

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

Donde:

C_k = conjunto de puntos del clúster k

$|C_k|$ = número de observaciones del clúster

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_i - \mu_k|^2$$

El algoritmo K-means busca minimizar la suma de las distancias cuadráticas entre cada observación y el centroide del clúster al que pertenece, conocida como suma de cuadrados intra-cluster (Within-Cluster Sum of Squares) [15].

2.3.6 Fuzzy C-Means

El algoritmo Fuzzy C-Means es un método de agrupamiento difuso cuyo objetivo consiste en dividir un conjunto de n observaciones en K clusters permitiendo que cada observación pertenezca simultáneamente a varios grupos con distintos grados de pertenencia. A diferencia de K-means, donde cada dato pertenece exclusivamente a un único cluster, Fuzzy C-Means asigna valores de pertenencia continuos entre 0 y 1, reflejando la incertidumbre o transición entre grupos [17].

Cada cluster se representa mediante un centroide calculado como un promedio ponderado de las observaciones según su grado de pertenencia, de la siguiente forma [17]:

$$\mu_k = \frac{\sum_{i=1}^{n} u_{ik}^{m} \cdot x_i}{\sum_{i=1}^{n} u_{ik}^{m}}$$

Donde:

u_{ik} = grado de pertenencia de la observación *i* al clúster *k*

m = parámetro de difusividad (m > 1), que controla el nivel de solapamiento entre clústeres

x_i = observación i

μ_k = centroide del clúster k

El algoritmo Fuzzy C-Means busca minimizar la siguiente función objetivo:

$$J_m = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^{m} \cdot |x_i - \mu_k|^2$$

Mientras K-means asigna cada observación a un único clúster mediante una partición rígida, Fuzzy C-Means introduce una partición difusa en la cual las observaciones pueden pertenecer parcialmente a múltiples clústeres, proporcionando una representación más flexible de la estructura de los datos [17].

2.3.7 Modelo Gaussian Mixture

El modelo de mezclas gaussianas (Gaussian Mixture Model, GMM) es un método de agrupamiento probabilístico que asume que los datos fueron generados por una combinación de K distribuciones gaussianas, cada una representando un clúster. A diferencia de K-Means, que asigna cada observación a un único clúster de forma rígida, GMM estima la probabilidad de que cada observación pertenezca a cada uno de los K componentes, lo que permite una asignación suave y flexible de los datos. La densidad de probabilidad del modelo se define como una suma ponderada de K distribuciones gaussianas:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Donde π_k es el peso del componente k, que representa la proporción de datos asignada a ese clúster y cumple que la suma de todos los pesos es igual a 1; μ_k es el vector de medias del componente k; y Σ_k es la matriz de covarianza del componente k, que determina la forma y orientación del clúster en el espacio de datos.

Los parámetros del modelo se estiman mediante el algoritmo de Esperanza-Maximización (EM), que alterna entre dos pasos. En el paso E (Esperanza) se calcula la probabilidad de pertenencia de cada observación a cada componente, conocida como responsabilidad [27]:

$$r_{ik} = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}$$

En el paso M (Maximización) se actualizan los parámetros del modelo utilizando las responsabilidades calculadas en el paso anterior, maximizando la verosimilitud de los datos. Este proceso se repite hasta que los parámetros convergen. La principal ventaja de GMM frente a K-Means es su capacidad para capturar clústeres con formas elípticas y diferentes tamaños, al modelar explícitamente la covarianza de cada componente [27].

2.3.8 DBSCAN

El algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un método de agrupamiento basado en densidad cuyo objetivo consiste en identificar regiones de alta concentración de datos (clústeres) separadas por regiones de baja densidad. A diferencia de métodos como K-means o Fuzzy C-Means, DBSCAN no requiere especificar previamente el número de clústeres y es capaz de identificar ruido (outliers) de manera explícita [17].

El algoritmo se basa en dos parámetros fundamentales:

- ε(epsilon): radio máximo de vecindad para considerar que dos puntos son cercanos
- MinPts: número mínimo de puntos necesarios para formar una región densa

Los puntos en DBSCAN se clasifican en tres categorías: los puntos núcleo, que son aquellos que

tienen al menos MinPts puntos dentro de su vecindad de radio ε; los puntos frontera, que no cumplen la condición de núcleo, pero se encuentran dentro de la vecindad de un punto núcleo; y los puntos ruido, que no pertenecen a ningún clúster al no cumplir ninguna de las condiciones anteriores. La vecindad de un punto se define como [17]:

$$N_\varepsilon(x_i) = \{x_j \in D : |x_j - x_i| \leq \varepsilon\}$$

Un punto x_i es considerado núcleo si:

$$|N_\varepsilon(x_i)| \geq \text{MinPts}$$

El algoritmo construye clústeres conectando puntos núcleo que son densamente alcanzables entre sí. Dos puntos pertenecen al mismo clúster si existe una cadena de puntos núcleo tal que cada uno está dentro de la vecindad ε del siguiente [17].

2.3.9 Clustering Jerárquico

El algoritmo de **c**lustering jerárquico es un método de agrupamiento cuyo objetivo consiste en construir una estructura jerárquica de clústeres, representada comúnmente mediante un dendrograma, que muestra cómo las observaciones se van agrupando de manera progresiva. A diferencia de anteriores métodos este enfoque no requiere fijar previamente el número de clústeres, sino que permite obtener distintas particiones al cortar el dendrograma en diferentes niveles.

Este enfoque aglomerativo (bottom-up) permite construir los clústeres de forma progresiva a partir de las similitudes locales entre observaciones, facilitando la interpretación y evitando la complejidad, además de el alto costo computacional asociado al enfoque divisivo (top-down) [15].

La distancia entre dos observaciones suele calcularse mediante la distancia euclidiana:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Para definir la distancia entre clústeres, se utilizan distintos criterios de enlace: Los métodos de

enlace jerárquico se diferencian por cómo miden la distancia entre clústeres: el enlace simple utiliza la distancia mínima entre los puntos de ambos clústeres; el enlace completo emplea la distancia máxima entre sus puntos; el enlace promedio calcula el promedio de todas las distancias entre puntos de un clúster y otro; y el enlace de Ward busca minimizar el aumento de la varianza intra-cluster al fusionar dos grupos [15].

2.3.10 Spectral Clustering

El Spectral Clustering es un método de agrupamiento basado en grafos cuyo objetivo consiste en identificar clusters a partir de la estructura de similitud entre los datos, transforma los datos a un nuevo espacio utilizando los autovalores y autovectores de una matriz de similitud, lo que permite capturar estructuras complejas y no lineales [18].

El algoritmo comienza construyendo una matriz de similitud W, donde cada elemento mide qué tan parecidos son dos puntos. La forma común de definirla es mediante el kernel gaussiano [18]:

$$W_{ij} = \exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right)$$

A partir de esta matriz, se construye la matriz de grado D, que es diagonal y cuyos elementos son:

$$D_{ii} = \sum_j W_{ij}$$

Luego se define la matriz Laplaciana del grafo, siendo una de las formas más utilizadas la Laplaciana no normalizada [18]:

$$L = D - W$$

El paso clave del método consiste en calcular los k autovectores asociados a los menores autovalores de la matriz Laplaciana. Estos autovectores se utilizan para proyectar los datos en un nuevo espacio de menor dimensión, donde la estructura de los clústeres es más evidente. Finalmente, sobre esta nueva representación, se aplica un algoritmo de clustering clásico

(generalmente K-means) para obtener la partición final [18].

2.4 Cramér's V

El Cramér's V es una medida de asociación entre dos variables categóricas, basada en el estadístico de la prueba Chi-cuadrado. El primer paso consiste en construir una tabla de contingencia que cruce los clústeres con la variable categórica de interés. En esta tabla, las columnas representan los k clústers identificados en el proceso previo, mientras que las filas corresponden a las r categorías de la variable externa. Cada celda n_{ij} contiene la frecuencia observada de observaciones que pertenecen simultáneamente a la categoría i de la variable externa y al clúster j. Esta tabla de doble entrada permite visualizar la distribución conjunta de ambas variables y constituye la base para los cálculos posteriores de independencia [19].

	Cluster 1	Cluster 2	...	Cluster k
Categoría 1	n_{11}	n_{12}	...	n_{1k}
Categoría 2	n_{21}	n_{22}	...	n_{2k}
...	...	...	...	...
Categoría r	n_{r1}	n_{r2}	...	n_{rk}

Tabla 1. Primer paso del método Cramer's V

A partir de la tabla de contingencia, se calcula el estadístico Chi-cuadrado ($\chi 2$) para evaluar si existe una asociación significativa entre los clústers y la variable categórica. La fórmula empleada es [19]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Donde n_{ij} representa la frecuencia observada en cada celda y E_{ij} es la frecuencia esperada bajo el supuesto de independencia entre las variables. La frecuencia esperada se calcula como $E_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$, siendo $n_{i\cdot}$ el total de la fila i, $n_{\cdot j}$ el total de la columna j, y n el número total de observaciones.

Un valor elevado de χ^2 indica que las diferencias entre lo observado y lo esperado son grandes, sugiriendo que las variables no son independientes.

Dado que el estadístico Chi-cuadrado es sensible al tamaño de la muestra y al número de categorías, se utiliza la V de Cramér (V) como medida estandarizada de la fuerza de la asociación. Este coeficiente se calcula mediante la siguiente expresión [19]:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, k-1)}}$$

Este coeficiente ajusta el Chi-cuadrado por las dimensiones de la tabla usando el mínimo entre r−1 y k−1 en el denominador lo que garantiza que el valor resultante siempre esté entre 0 y 1, donde r es el número de filas, k el número de columnas y n el total de observaciones [19].

La V de Cramér permite interpretar la magnitud de la asociación entre los clústeres y la variable categórica en un rango de 0 a 1: un valor de 0 indica ausencia de relación, mientras que un valor de 1 señala una asociación perfecta [19]. Para su interpretación práctica, valores entre 0 y 0.1 corresponden a una asociación débil, entre 0.1 y 0.3 a una moderada, y superiores a 0.3 a una fuerte.

3. Metodología

Para estructurar de manera sistemática el proceso analítico, este proyecto adopta el modelo metodológico CRISP-DM (Cross Industry Standard Process for Data Mining). Este enfoque es uno de los estándares en proyectos de minería y análisis de datos, que permite organizar las actividades en seis fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. La metodología CRISP-DM se adapta adecuadamente a este estudio, ya que integra tanto la comprensión del contexto educativo como el manejo de grandes volúmenes de información y la aplicación de técnicas analíticas. A continuación, se presenta el desarrollo de cada una de las fases en el marco de la investigación [20].

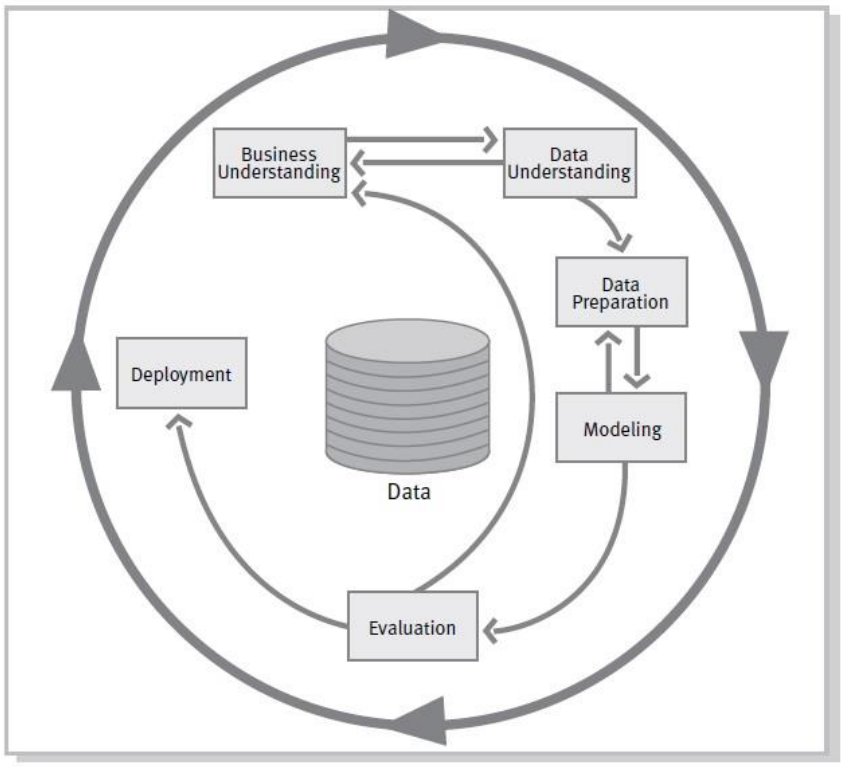


Figura 2. Fases de la metodología CRISP-DM. Fuente: Chapman et al. (2000)

4. Aplicación de la metodología

4.1 Entendimiento del negocio

El presente proyecto busca identificar patrones de desempeño académico en estudiantes de programas de Administración y áreas afines de la Pontificia Universidad Javeriana y la Universidad Nacional de Colombia, a partir de las pruebas de estado. Estas pruebas, diseñadas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), constituyen un instrumento estandarizado para evaluar competencias genéricas en la educación superior y un referente para medir la calidad educativa. El análisis se restringe a los módulos genéricos del examen, lo que garantiza una medición homogénea y comparable entre todos los estudiantes independientemente de su programa académico.

La selección del área de Administración y áreas afines responde a varios criterios. En primer lugar, las cuatro sedes analizadas cuentan con programas consolidados y de larga trayectoria en este campo, lo que garantiza comparabilidad institucional y disponibilidad de una muestra

suficiente en cada sede. En segundo lugar, Administración y áreas afines constituye un caso relevante de análisis dentro del sistema de educación superior colombiano: es una de las áreas con mayor número de estudiantes matriculados a nivel nacional, agrupa programas de Administración de Empresas, Economía y Contaduría Pública que comparten un núcleo común de competencias en razonamiento cuantitativo, lectura crítica e inglés, y forma profesionales con alta inserción en sectores estratégicos de la economía, lo que la convierte en un campo de interés tanto académico como institucional. En tercer lugar, restringir el análisis a un área específica permite trabajar con una muestra homogénea en términos de formación, evitando que las diferencias estructurales entre disciplinas dominen sobre las diferencias de trayectoria académica que son el objeto de estudio, y facilitando una interpretación más precisa de los perfiles identificados.

No obstante, esto implica reconocer ciertas limitaciones interpretativas. Al excluir otras áreas del conocimiento como ingenierías, ciencias de la salud, ciencias sociales o humanidades, los perfiles identificados no son directamente extrapolables a estudiantes de campos disciplinares distintos, dado que cada área presenta dinámicas de formación, criterios de admisión, exigencias curriculares y perfiles de ingreso particulares. Asimismo, ciertos módulos del examen Saber Pro como razonamiento cuantitativo o lectura crítica pueden tener un peso interpretativo diferente según el campo de formación, lo que condiciona la lectura de los resultados al perfil específico de los programas analizados. Por tanto, los hallazgos de este estudio deben entenderse como una caracterización única de la población de Administración y áreas afines en las cuatro sedes analizadas, no como un retrato general del sistema de educación superior colombiano. Esta delimitación, lejos de ser una debilidad, constituye una decisión metodológica orientada a obtener mayor profundidad analítica dentro de un campo disciplinar concreto, y abre la puerta a que estudios futuros repliquen este enfoque en otras áreas del conocimiento para construir una visión comparativa más amplia.

Tradicionalmente, los análisis de Saber Pro se han enfocado en el puntaje global o en desempeños por competencias en un único momento. En contraste, este estudio adopta un enfoque longitudinal que permite caracterizar la trayectoria individual de los estudiantes desde la educación media hasta el egreso universitario, integrando las bases de datos oficiales de Saber 11 y Saber Pro mediante el archivo de llaves de vinculación del ICFES. Este enfoque no busca establecer relaciones causales sino identificar asociaciones entre las condiciones académicas, sociodemográficas e institucionales al momento de ingreso y los perfiles de desempeño observados.

4.1.1 Protección de datos personales y anonimización en el sistema educativo

En Colombia, el tratamiento de información personal está regulado por un marco jurídico que garantiza el derecho fundamental a la intimidad y a la protección de datos. La Ley 1581 de 2012, establece el régimen general de protección de datos personales, aplicable a todas las entidades públicas y privadas que recolectan, almacenen o utilicen información de carácter personal. Esta normativa desarrolla el artículo 15 de la Constitución Política, que consagra el derecho de todas las personas a conocer, actualizar y rectificar la información que sobre ellas se haya recogido en bases de datos o archivos, y establece principios rectores como la legalidad, finalidad, libertad, veracidad, transparencia, acceso y circulación restringida, seguridad y confidencialidad [22].

En este contexto, el Instituto Colombiano para la Evaluación de la Educación (ICFES), como entidad encargada de las pruebas Saber 11 y Saber Pro, tiene la obligación de proteger la identidad y la información sensible de los estudiantes. Por ello, no suministra números de identificación personal en las bases de datos públicas, sino que aplica procesos de anonimización y encriptación que impiden la identificación directa de los individuos [23].

Como parte de estas medidas, el ICFES genera llaves que permite relacionar de forma segura y anónima los registros de las pruebas Saber 11 y Saber Pro, evitando conocer la identidad de cada estudiante. Este archivo es un instrumento técnico que facilita el seguimiento longitudinal de

trayectorias educativas, respetando al mismo tiempo las disposiciones legales sobre privacidad y tratamiento de datos [23].

4.2 Entendimiento de los datos

Para el desarrollo de esta investigación se utilizaron dos fuentes de información oficiales dadas por el Instituto Colombiano para la Evaluación de la Educación (ICFES): la base de datos de la prueba Saber Pro y la base de datos de la prueba Saber 11, con un enfoque particular en los estudiantes de programas de Administración y áreas afines de la Pontificia Universidad Javeriana en sus sedes de Bogotá y Cali, y de la Universidad Nacional de Colombia en sus sedes de Bogotá y Palmira. El rango de datos definido para Saber 11 comprende el período 2014-2 a 2017, y para Saber Pro el período 2018 a 2022. Saber 11 se estableció desde el segundo semestre de 2014 debido a que los registros anteriores corresponden a una estructura de prueba distinta, con módulos y escalas de calificación diferentes a los vigentes en el período de estudio, lo que impide la comparación directa. El límite superior de 2017 para Saber 11 responde a que, dado que los estudiantes deben completar por lo menos el 75% de su plan de estudios antes de presentar Saber Pro, se requiere un mínimo de cuatro años entre ambas pruebas, lo que sitúa la presentación de Saber Pro entre 2018 y 2022, período que coincide con la disponibilidad completa de datos al momento del análisis. Esto garantiza que cada estudiante vinculado cuente con registros comparables en ambas pruebas dentro del marco temporal del estudio.

Para el análisis se utilizó Python en el entorno Google Colaboratory. Python fue seleccionado por su ecosistema de librerías especializadas para el manejo y análisis de datos, como pandas y scikit-learn [28], mientras que Colaboratory permite procesar grandes volúmenes de información en la nube sin requerir hardware especializado, garantizando la reproducibilidad del análisis. Los detalles del entorno de ejecución, versiones de librerías e infraestructura utilizada se presentan en el Anexo 1.

4.2.1 Datos de la prueba Saber 11

La base de Saber 11 corresponde al examen presentado por los estudiantes al finalizar la educación media. Contiene información sobre características escolares, familiares, sociodemográficas y de desempeño académico en las distintas áreas evaluadas, los detalles de todas las variables de la prueba Saber 11 se presentan en el anexo 7. Algunas variables destacadas son:

- Información del colegio: área de ubicación (urbana/rural), carácter (oficial o no oficial), jornada, bilingüismo, calendario académico, naturaleza jurídica y códigos DANE de sede y establecimiento.
- Información del estudiante: género, fecha de nacimiento, nacionalidad, municipio, departamento de residencia y presentación del examen.
- Condiciones familiares y del hogar: número de cuartos en el hogar, nivel educativo de madre y padre, número de personas en el hogar, tenencia de bienes (automóvil, computador, internet, lavadora) y estrato socioeconómico.
- Resultados académicos: puntajes individuales en áreas específicas (lectura crítica, matemáticas, inglés, sociales y ciudadanas, ciencias naturales) y el puntaje global.

En total, esta base contiene 51 variables, lo que permite caracterizar el contexto escolar y familiar de cada estudiante en el momento de finalizar su educación media.

Durante la etapa de limpieza, se realizó una revisión exhaustiva de la estructura de las bases y del nivel de completitud de las variables. En este proceso se identificó que los registros correspondientes al año 2014-1 y anteriores presentan valores nulos en todos los puntajes de la prueba Saber 11. Esto se debe a que, en dichos periodos, el examen utilizaba una estructura y categorización distinta, lo que impide la comparación directa con los años posteriores.

Por esta razón, se decidió excluir los datos anteriores a 2014-2 del análisis, concentrando el estudio en los años en los que el formato de evaluación es homogéneo, además de eliminar años

que no encajan con los tiempos del examen saber Pro debido a que desde el año de presentación de la saber 11 deben pasar por lo menos 4 años para la presentación de la saber Pro. Esta decisión implica la eliminación aproximada del 43% del total de registros, conservando así una base más consistente y comparable para los análisis.

Variables con datos faltantes moderados (1-10%)*:* La variable de bilingüismo del colegio presenta 2,448 nulos (7.27%), mientras que variables familiares como tenencia de computador (1.53%), nivel educativo del padre (1.26%) y de la madre (1.25%) muestran porcentajes menores.

Variables con datos faltantes mínimos (<1%)*:* Casi todas las variables del dataset presentan excelentes registros, con ausencias menores al 0.5%. Destacan variables de identificación institucional, información geográfica y características demográficas del estudiante, donde los valores nulos son prácticamente inexistentes (0% en 29 de las 51 variables).

4.2.2 Datos de la prueba Saber Pro

La base de datos de Saber Pro contiene información detallada a nivel individual sobre los estudiantes y sus contextos institucionales, académicos y familiares, así como sus resultados en las distintas competencias evaluadas por el ICFES, los detalles de todas las variables de la prueba Saber Pro se presentan en el anexo 6. Algunas variables relevantes incluyen:

- Identificación y ubicación: periodo de presentación (PERIODO), identificador único (ESTU_CONSECUTIVO), país y municipio de residencia, institución de educación superior, departamento y municipio de presentación.
- Información institucional y académica: nombre y tipo de institución (pública o privada), núcleo de formación, programa académico, nivel y metodología del programa, valor de matrícula, horas de trabajo semanal.
- Condiciones socioeconómicas: nivel educativo de los padres, tenencia de bienes (automóvil, lavadora, computador, acceso a internet), estrato de vivienda, tipo de financiación de la

matrícula.

- Resultados en el examen Saber Pro: puntajes y desempeños en las distintas competencias, entre ellas razonamiento cuantitativo, comunicación escrita, lectura crítica, inglés y competencias ciudadanas.

En total, la base de Saber Pro cuenta con 59 variables, lo que permite un análisis amplio sobre los desempeños académicos en educación superior.

El análisis revela tres grupos diferenciados de variables según el porcentaje de valores faltantes:

Variables con alta proporción de datos faltantes (> 25%)*:* Las variables del colegio donde terminaron su bachillerato presentan las mayores ausencias: código DANE, código del municipio (31.82%), y nombre del colegio (29.83%). Dado que estas variables corresponden a códigos de identificación institucional sin valor analítico directo para el estudio, fueron excluidas del análisis sin afectar la integridad del conjunto de datos ni la capacidad de responder a los objetivos planteados.

Variables con datos faltantes moderados (3-6%)*:* Las variables socioeconómicas muestran ausencias en estos puntos: tenencia de bienes como automóvil (5.76%), lavadora (5.49%), computador (5.35%) e internet (4.24%); estrato de vivienda (4.58%); horas semanales de trabajo (4.93%); y nivel educativo de madre (3.72%) y padre (3.66%).

Variables con datos faltantes mínimos (< 3%)*:* La mayoría de las variables (38 de 59) presentan excelente completitud con porcentajes inferiores al 1% o completitud total. Esto incluye identificación del estudiante, información de la institución de educación superior, programa académico y resultados en las pruebas, garantizando robustez en las variables clave del análisis.

4.2.3 Articulación entre las bases de datos

Un aspecto fundamental es que no existe una relación directa entre los identificadores de Saber 11 y Saber Pro. Para vincular los registros individuales entre ambas bases se utiliza el archivo de

llaves de vinculación proporcionado por el ICFES, el cual contiene identificadores únicos generados para cada estudiante, en cumplimiento de la Ley 1581 de 2012 sobre protección de datos personales.

El Sistema de llaves de vinculación del ICFES proporciona un archivo que permite establecer la correspondencia entre un mismo estudiante en ambas pruebas. Este archivo contiene cuatro variables fundamentales: estu_consecutivo_sb11 (identificador único del estudiante en Saber 11), periodo_sb11 (período en que presentó Saber 11), estu_consecutivo_sbpro (identificador único del mismo estudiante en Saber Pro), y periodo_sbpro (período en que presentó Saber Pro). Por ejemplo, un registro típico sería: el estudiante con ID SB11201120209353 que presentó Saber 11 en el período 20153 (segundo semestre de 2015) corresponde al estudiante con ID EK201830003181 que presentó Saber Pro en el período 20203 (segundo semestre de 2020). De esta manera, cada fila del archivo de llaves establece una relación única entre ambos sistemas de identificación, permitiendo rastrear la trayectoria del estudiante desde la educación media hasta la superior, incluso cuando los identificadores asignados en cada examen son completamente diferentes.

4.3 Preparación de los datos

4.3.1 Integración de las bases de datos Saber 11 y Saber Pro

La integración de ambas bases de datos permitió caracterizar a los estudiantes desde su ingreso a la educación superior, incorporando información socioeconómica, académica y contextual proveniente del examen Saber 11. Este proceso presenta importantes retos metodológicos debido al gran volumen de información la base de Saber 11 contenía cerca de 7 millones de registros desde 2010 hasta 2022 y a la ausencia de una relación directa entre ambos sistemas. Para vincular los registros de Saber 11 y Saber Pro se utilizó el archivo de llaves suministrado por el ICFES. La integración se realizó a partir de tres fuentes principales: la base de datos de Saber Pro (aproximadamente 35.000 registros), el archivo de llaves y la base de Saber 11.

4.3.2 Proceso de filtrado y vinculación

El proceso de integración se realizó en tres etapas secuenciales:

Etapa 1: La obtención de identificadores de estudiantes Saber Pro partió de la base de datos completa de estudiantes de la Universidad Javeriana con sus sedes Bogotá y Cali, Universidad Nacional de Colombia con sus sedes en Bogotá y Palmira que presentaron el examen Saber Pro entre 2018 y 2022, la cual contiene 35,000 registros aproximadamente. De esta base se extrajeron los identificadores únicos de los estudiantes (variable ESTU_CONSECUTIVO).

Etapa 2: Con la vinculación del archivo de llaves ICFES, se identificaron los códigos de los estudiantes correspondientes de la prueba saber Pro en la base de datos Saber 11. Este archivo actúa como puente entre ambos sistemas (variable estu_consecutivo_sb11).

Etapa 3: Filtrado de la base Saber 11: Dado el tamaño masivo de la base de datos Saber 11 (7 millones de registros), se implementó un proceso de filtrado por lotes (chunks) para gestionar eficientemente los recursos computacionales. Una vez se terminó este proceso se dejó únicamente el área de interés de administración y afines.

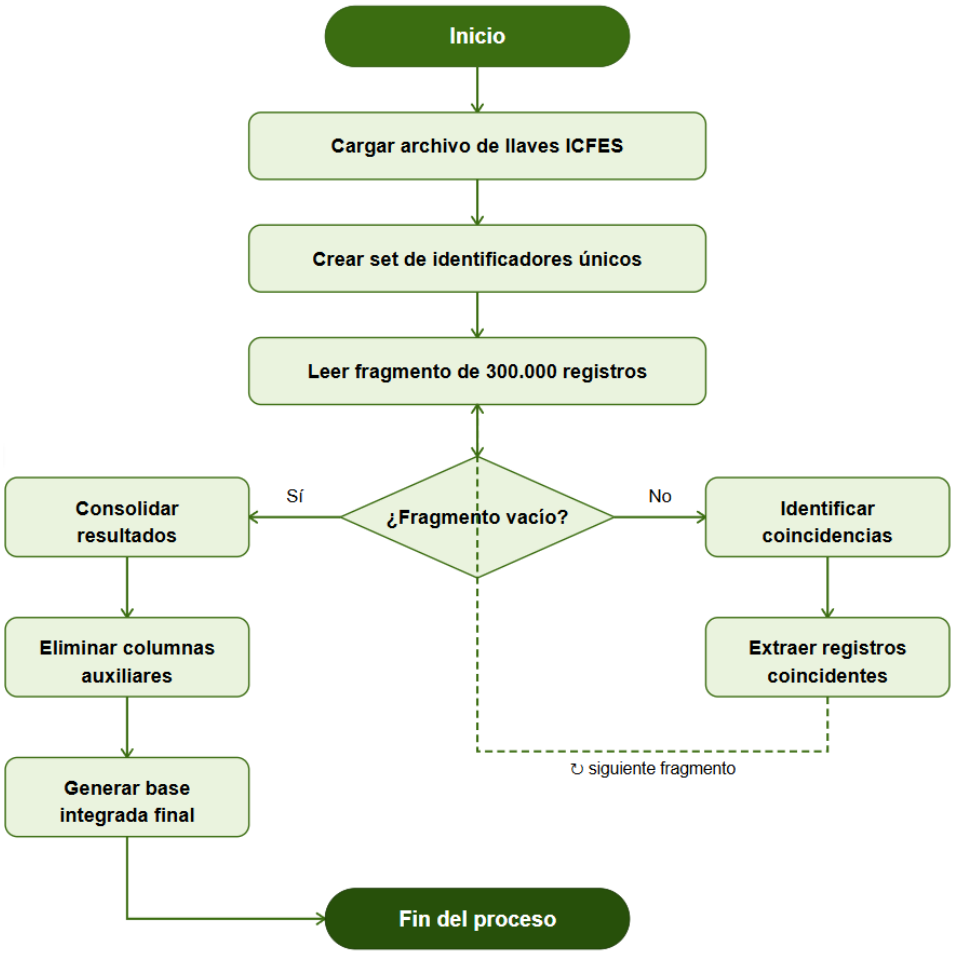


Figura 3. Proceso de vinculación.

Si bien esta etapa describe la lógica técnica del proceso de vinculación resulta igualmente importante documentar de manera transparente la evolución cuantitativa de la muestra a lo largo de cada filtro aplicado. Esta trazabilidad permite identificar en qué etapas se produce la mayor reducción de registros, delimitar con precisión el foco empírico del estudio y reconocer las implicaciones que la pérdida de observaciones puede tener sobre la validez externa de los resultados. La Figura 4 presenta el flujo completo de la muestra desde los registros iniciales hasta la base analítica final, indicando en cada paso tanto los registros que avanzan al siguiente filtro como los que son excluidos.

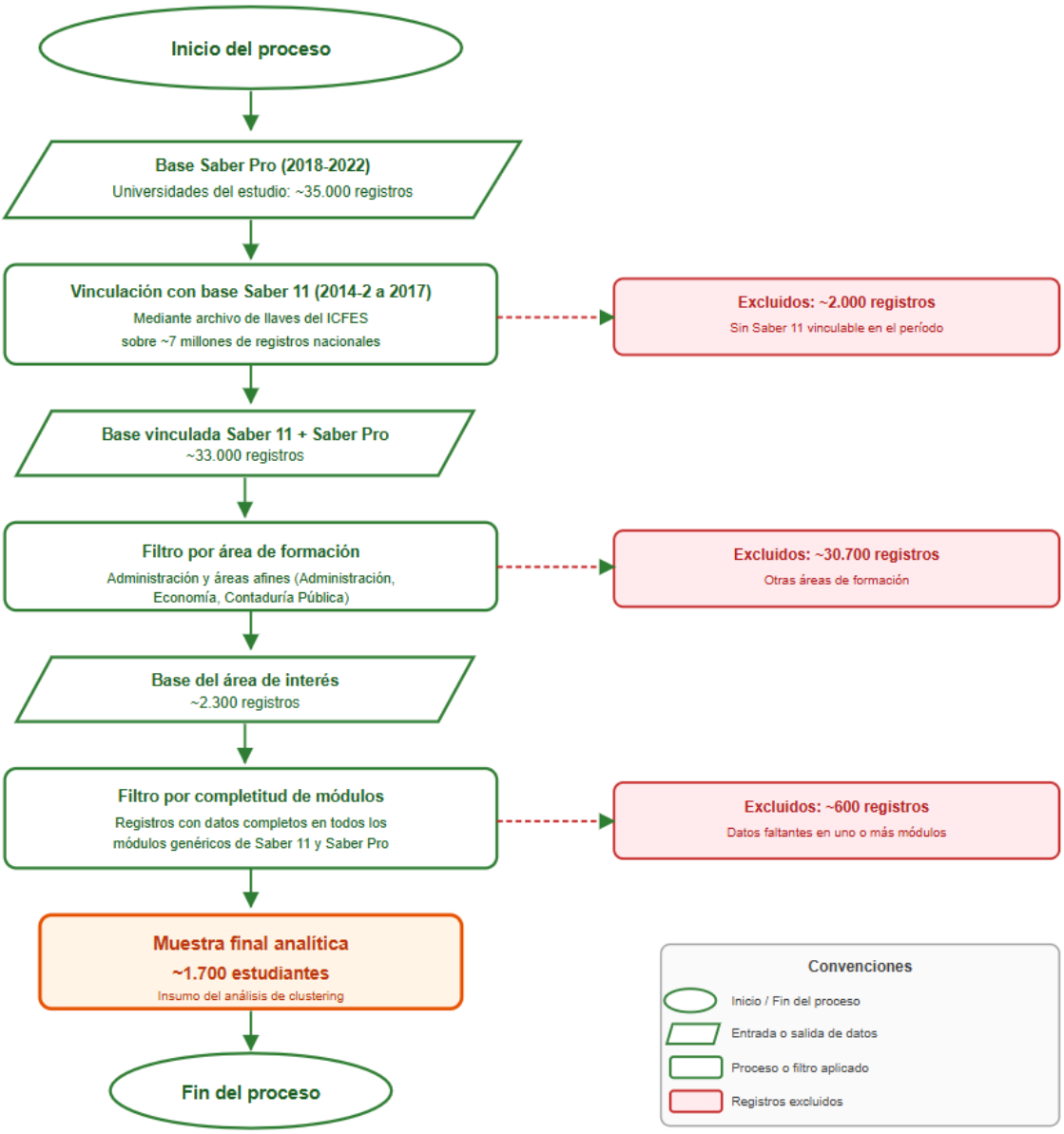


Figura 4. Flujo de la muestra desde los registros iniciales hasta la base analítica final.

El proceso parte de aproximadamente 35.000 registros de la base Saber Pro correspondientes a las cuatro sedes universitarias analizadas en el período 2018-2022. La primera etapa consistió en la vinculación con la base Saber 11 entre 2014-2 y 2017 mediante el archivo de llaves del ICFES, etapa en la cual se perdieron aproximadamente 2.000 registros, correspondientes a estudiantes para los que no fue posible establecer un vínculo entre ambas pruebas dentro del período definido. Aunque esta proporción es relativamente baja respecto al total inicial, las causas específicas de la no vinculación responder al cambio en la estructura del examen Saber 11 anterior a 2014-2, que no

quedan reflejadas en el archivo de llaves.

La segunda etapa, correspondiente al filtro por área de formación, es la que produce la mayor reducción del conjunto de datos: aproximadamente 30.700 registros fueron excluidos por no pertenecer a programas de Administración y áreas afines. Esta pérdida no constituye un sesgo metodológico sino una delimitación deliberada del alcance del estudio, dado que el objetivo es caracterizar trayectorias dentro de un campo disciplinar específico y no de la totalidad de la oferta académica de las cuatro sedes. Tras este filtro, la base se redujo a aproximadamente 2.300 estudiantes de Administración, Economía y Contaduría Pública.

La tercera etapa corresponde al filtro de los datos en los módulos evaluados tanto en Saber 11 como en Saber Pro. En esta etapa se perdieron aproximadamente 600 registros que presentaban valores nulos en uno o más módulos genéricos de cualquiera de las dos pruebas, lo cual resulta crítico para el análisis de clustering dado que el modelo requiere observaciones completas en todas las variables de entrada. Se optó por una eliminación por casos completos en lugar de aplicar técnicas de imputación, dado que los puntajes de los módulos constituyen la variable de análisis del estudio y no variables contextuales, siguiendo también algunos estudios guía, donde se usa únicamente los datos completos [16]; imputarlos implicaría fabricar el desempeño académico que precisamente se busca caracterizar, introduciendo correlaciones artificiales entre módulos que el algoritmo podría interpretar como patrones reales. Además, la concentración de valores faltantes en ciertas sedes sugiere un mecanismo no aleatorio que la imputación no corregiría, sino que podría amplificar, por lo tanto, la muestra analítica final está compuesta por aproximadamente 1.700 estudiantes.

4.3.3 Consideraciones técnicas

Para manejar eficientemente una base de datos de más de siete millones de registros, se implementó un procesamiento por lotes (chunking), que consiste en dividir el conjunto total en

fragmentos más pequeños que puedan ser cargados y procesados de forma secuencial. En este caso, se definieron chunks de 300.000 registros cada uno, un tamaño determinado con base en la capacidad de memoria RAM disponible, el tiempo de procesamiento óptimo y la necesidad de equilibrar eficiencia con uso de recursos. Esta estrategia permite trabajar con volúmenes de datos muy grandes sin sobrecargar la memoria, ya que cada lote se procesa, libera y da paso al siguiente.

4.3.4 Base de datos unida

El proceso de filtrado e integración descrito en las etapas anteriores permitió obtener una base final analítica de aproximadamente 1.700 registros, correspondientes a estudiantes de las cuatro sedes universitarias seleccionadas que presentaron el examen Saber Pro entre 2018 y 2022, contaban con registro vinculable en Saber 11 entre 2014-2 y 2017, pertenecen al área de Administración y áreas afines, y disponen de datos completos en todos los módulos genéricos de ambas pruebas. Esta base concentra la información necesaria para los análisis posteriores y constituye la unidad de análisis del estudio.

4.4 Modelado

La etapa de modelado tuvo como objetivo identificar agrupaciones naturales en los datos de desempeño académico mediante técnicas de clustering no supervisado. A continuación, se describen las decisiones metodológicas tomadas en cada aspecto del proceso.

Previo a la construcción de los modelos, se realizó un análisis exploratorio de los datos con el propósito de caracterizar el punto de partida del análisis y verificar la coherencia del conjunto de datos. Este análisis incluyó la distribución del puntaje promedio general, puntaje promedio en cada módulo de Saber 11 y Saber Pro, la cantidad de estudiantes por institución. Los resultados de esta exploración se compararon con la puntuación promedio más reciente reportada por el ICFES para cada prueba, correspondiente al año 2024, ya que todas las cifras de 2025 aún no han sido oficialmente entregadas al momento de realizar este análisis.

Esta caracterización inicial fue fundamental para contextualizar correctamente los perfiles que emergieron del clustering, recordando que las diferencias entre clústeres deben interpretarse en términos relativos dentro de una población académicamente selecta y no como una representación del sistema educativo colombiano en su conjunto.

Las variables utilizadas para la construcción de los clústeres corresponden exclusivamente a los puntajes obtenidos por los estudiantes en los módulos genéricos de ambas pruebas: los cinco módulos de Saber 11 (Lectura Crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés) y los cinco módulos genéricos de Saber Pro (Lectura Crítica, Razonamiento Cuantitativo, Competencias Ciudadanas, Comunicación Escrita e Inglés). Esta decisión fue debido a que incluir variables sociodemográficas, socioeconómicas o institucionales en la construcción de los clústeres habría sesgado la segmentación hacia diferencias estructurales entre las instituciones como el carácter público o privado de cada universidad en lugar de revelar patrones genuinos de desempeño académico. El objetivo fue que los clústeres identificarán exclusivamente las trayectorias de rendimiento de los estudiantes, dejando las variables contextuales para la etapa de validación externa.

4.4.1 Algoritmos evaluados e hiperparámetros

La Tabla 2 presenta los hiperparámetros configurados para cada algoritmo evaluado. En todos los casos se fijó el mismo número de grupos (4) para garantizar condiciones comparables entre modelos y que las métricas de validación interna reflejaran diferencias reales en la calidad del agrupamiento y no en el número de particiones.

Algoritmo	Hiperparámetro	Valor
K-Means	n_clusters	4
	random_state	100
Gaussian Mixture	n_components	4
	random_state	100
Fuzzy C-Means	c (número de clústeres)	4
	m (difusividad)	2
	error de convergencia	0.005
	máximo de iteraciones	1000
Clustering Aglomerativo	n_clusters	4
	linkage	ward (por defecto)
DBSCAN	ε (epsilon)	1.5
	min_samples	10
Spectral Clustering	n_clusters	4
	assign_labels	kmeans
	random_state	100

Tabla 2. Hiperparametros usados por cada algoritmo.

La selección de cada configuración responde a criterios técnicos específicos. Para K-Means, Gaussian Mixture y Spectral Clustering, el valor random_state=100 garantiza la reproducibilidad de los resultados ante la inicialización aleatoria propia de estos algoritmos. Para Fuzzy C-Means, el parámetro de difusividad m=2 es el valor estándar recomendado en la literatura para datos continuos, mientras que el error de convergencia de 0.005 y el máximo de 1000 iteraciones son valores convencionales que aseguran la convergencia del algoritmo sin costo computacional excesivo. Para el Clustering Aglomerativo se utilizó el criterio de enlace Ward, que minimiza la varianza intra-clúster al fusionar grupos, criterio coherente con el objetivo de K-Means y que facilita la comparación entre ambos métodos.

En el caso de DBSCAN, la selección de ε=1.5 y min_samples=10 respondió a un proceso de exploración empírica, dado que configuraciones con valores distintos resultaron en particiones con un número insuficiente de clústeres para ser evaluadas con las métricas seleccionadas. Este

comportamiento se debe a las características de los datos, DBSCAN requiere regiones de alta densidad separadas por zonas, condición que no se cumple en datos de desempeño académico con distribución continua y sin separaciones excesivas.

4.5 Evaluación

Una vez realizados los análisis iniciales, la etapa de evaluación tuvo como objetivo determinar cuál de los algoritmos comparados producía la segmentación más apropiada para los propósitos del estudio. Para evitar que esta selección dependiera de una única dimensión, se definieron los siguientes criterios complementarios de evaluación que, en conjunto, permiten valorar tanto la calidad técnica del agrupamiento como su utilidad práctica en el contexto educativo analizado:

En primer lugar, el desempeño en métricas de validación interna, evaluado mediante tres indicadores ampliamente empleados en la literatura de clustering: el coeficiente de Silhouette, que mide la cohesión interna y separación entre grupos; el índice de Davies-Bouldin, que penaliza la similitud entre clústeres cercanos; y el índice de Calinski-Harabasz, que evalúa la razón entre dispersión inter-clúster e intra-clúster. La combinación de estas tres métricas permite una evaluación más robusta que cualquier indicador individual [16]. En segundo lugar, el balance en la distribución de los grupos, entendido como la capacidad del algoritmo para producir clústeres con tamaños representativos y no soluciones en las que la mayoría de las observaciones se concentran en un único grupo y los demás contienen únicamente valores atípicos. Este criterio resulta especialmente relevante porque algoritmos como Spectral Clustering pueden optimizar las métricas internas a costa de generar particiones desequilibradas sin valor analítico [18]. En tercer lugar, la estabilidad de la solución, evaluada mediante un análisis de sensibilidad ante variaciones en el número de clústeres (k entre 3 y 6), con el fin de verificar que la segmentación seleccionada no fuese fruto de una configuración accidental sino el resultado de una estructura consistente en los datos. En cuarto lugar, la interpretabilidad pedagógica de los grupos resultantes, entendida como la

capacidad de los clústeres para describir perfiles de desempeño coherentes, jerárquicamente ordenados y lógicos desde un marco educativo. Un algoritmo que produzca clústeres sin estructura interpretable es inadecuado para los objetivos del estudio, independientemente de sus métricas internas.

Por último, la coherencia con la literatura previa, en particular con estudios de referencia sobre clustering aplicado a desempeño académico [16], donde se ha demostrado que ciertos algoritmos para datos educativos con distribución continua y solapamiento de competencias.

Estos criterios no se aplicaron de forma jerárquica estricta sino de manera conjunta y complementaria, considerando que un algoritmo plenamente apto debía cumplir aceptablemente en todos ellos. La aplicación de este marco a los seis algoritmos evaluados se presenta en la sección de resultados, donde se detallan los valores numéricos, las visualizaciones y el razonamiento que condujeron a la selección de K-Means con cuatro clústeres como solución final.

Dado que los módulos de Saber 11 y Saber Pro operan en escalas diferentes los módulos de Saber 11 reportan puntajes en una escala de 0 a 100 puntos, mientras que los módulos genéricos de Saber Pro lo hacen en una escala de 0 a 300, fue necesario aplicar un proceso de estandarización previo a la construcción de los modelos de agrupamiento. Sin esto, los puntajes de Saber Pro habrían dominado el cálculo de distancias por su mayor magnitud absoluta, sesgando los clústeres resultantes hacia esa prueba y restando peso a la información proveniente de Saber 11.

Para resolver este desbalance se aplicó la estandarización Z-score mediante la función StandardScaler de la librería scikit-learn, cuyo fundamento matemático se desarrolló en la sección 2.3.2. En el contexto específico de este estudio, la transformación se aplicó una única vez sobre el panel completo de la muestra final analítica de aproximadamente 1.700 estudiantes, considerando simultáneamente las diez variables de entrada del modelo: los cinco módulos de Saber 11 y los cinco módulos genéricos de Saber Pro. Bajo esta configuración, scikit-learn calcula la media y la

desviación estándar de cada módulo de forma independiente, usando como referencia las 1.700 observaciones de la muestra, y aplica la transformación columna por columna sin mezclar información entre módulos.

Esta propiedad es fundamental para los análisis posteriores que comparan trayectorias entre pruebas, como la matriz de transición y la comparativa entre puntajes promedio por clúster. No se aplicaron transformaciones adicionales debido a que StandardScaler resuelve por sí solo el problema de las escalas distintas, manteniendo además las propiedades estadísticas necesarias para los algoritmos de clustering aplicados.

Para determinar el número óptimo de clústeres se evaluaron configuraciones de k entre 2 y 6. La opción de k=2 fue descartada desde el inicio por razones interpretativas: una división de 2 reduciría los perfiles estudiantiles únicamente a alto y bajo desempeño. La selección final de k=4 se fundamentó en la combinación de tres métricas de validación interna, coeficiente de Silhouette, índice de Davies-Bouldin e índice de Calinski-Harabasz siguiendo el enfoque empleado por estudios similares [16]. Esta combinación de métricas permite evaluar simultáneamente la cohesión interna de los grupos y su separación entre sí, ofreciendo un criterio más robusto que cualquier métrica individual.

Con el fin de evaluar la estabilidad, se realizó un análisis de sensibilidad variando el número de clústeres entre k=3 y k=6. Los resultados del coeficiente de Silhouette para cada configuración fueron: k=3 obtuvo 0.1594, k=4 obtuvo 0.1486, k=5 obtuvo 0.1262 y k=6 obtuvo 0.1196. Si bien k=3 presenta el valor de Silhouette más alto, fue descartado por razones interpretativas: una segmentación en tres grupos reduciría la interpretación de los perfiles identificados, dificultando la caracterización de trayectorias intermedias que son precisamente las de mayor interés pedagógico en este estudio. La leve diferencia entre k=3 y k=4 en términos métricos, combinada con la mayor riqueza interpretativa de cuatro perfiles, justifica la selección de k=4 como la configuración óptima

para los objetivos del trabajo.

La evaluación de los modelos se realizó mediante la comparación sistemática de seis algoritmos de agrupamiento bajo las mismas condiciones: K-Means, Gaussian Mixture, Fuzzy C-Means, Clustering Aglomerativo, DBSCAN y Spectral Clustering, todos aplicados sobre el mismo conjunto de variables estandarizadas. Para K-Means se fijó una semilla de random_state=100, garantizando que la inicialización aleatoria de los centroides no afectará para que los resultados sean reproducibles y no cambian por efecto del azar entre una ejecución y otra y asegurando la reproducibilidad del análisis.

Se compararon los seis algoritmos mediante tres métricas de validación interna coeficiente de Silhouette, índice de Davies-Bouldin e índice de Calinski-Harabasz que permiten evaluar simultáneamente la cohesión interna de los grupos y su separación entre sí. El algoritmo con las mejores métricas internas Spectral Clustering generó agrupaciones sin estructura interpretable, lo que lo hizo inapropiado para los objetivos del estudio, debido a los criterios planteados. En consecuencia, se seleccionó K-Means con cuatro clústeres por ser el algoritmo que cumple satisfactoriamente con los criterios definidos: presenta el mayor índice de Calinski-Harabasz entre los algoritmos con particiones balanceadas, produce una distribución equilibrada de estudiantes entre los grupos, exhibe estabilidad en el análisis de sensibilidad de k, genera perfiles pedagógicamente interpretables, jerárquicamente ordenados y coincide con los hallazgos de la literatura previa sobre clustering aplicado a datos educativos [16].



Figura 5. Método de evaluación.

4.6 Despliegue

El despliegue de este proyecto se materializa en un reporte analítico obtenido a partir del modelo seleccionado. Esta etapa se desarrolla en dos momentos dentro del documento. El primero es el análisis exploratorio preliminar, presentado en la sección de resultados preliminares, donde se describe el comportamiento general de los datos mediante visualizaciones comparativas de puntajes por módulo. El segundo momento es el perfilamiento posterior al agrupamiento, donde el estudio se enriquece mediante la validación externa con el coeficiente de Cramér's V, que permite identificar qué variables sociodemográficas e institucionales presentan mayor asociación con los perfiles de desempeño, y cuyo desarrollo completo se presenta en la sección de resultados finales. En conjunto, ambos momentos materializan el propósito del trabajo: traducir los patrones identificados en el clustering en gráficos descriptivos comprensibles y contextualizados para las instituciones analizadas.

5. Resultados preliminares

Esta sección expone los resultados preliminares del estudio, orientados a ofrecer una comprensión inicial de los datos junto con las dinámicas generales observadas en las pruebas Saber 11 y Saber Pro. A través de distintos análisis exploratorios, se busca identificar tendencias, diferencias y comparaciones entre el país con la muestra analizada. Este acercamiento inicial no

solo permite reconocer la estructura y coherencia del conjunto de datos, sino que también proporciona las primeras vistas sobre los factores que podrían influir en el desempeño académico, constituyendo una base sólida para el análisis posterior.

5.1 Análisis exploratorio

En esta sección se presentan los resultados preliminares del análisis, con el objetivo de tener una claridad descriptiva de los datos antes de la aplicación de técnicas de agrupamiento. En particular, se examinan los puntajes obtenidos por los estudiantes en las pruebas Saber 11 y Saber Pro, así como los puntajes obtenidos únicamente por el grupo de estudio que son los estudiantes de las áreas de administración y afines. Este análisis permite identificar patrones generales, niveles de desempeño y posibles diferencias entre poblaciones, proporcionando un contexto sólido para la posterior implementación de métodos de clustering. De esta manera, se busca garantizar que las etapas posteriores del estudio se fundamenten en una comprensión clara y rigurosa de la estructura y comportamiento de los datos.

5.1.1 Resultados preliminares de la prueba Saber 11

Se tiene como objetivo ofrecer una caracterización inicial de los resultados mediante la visualización.

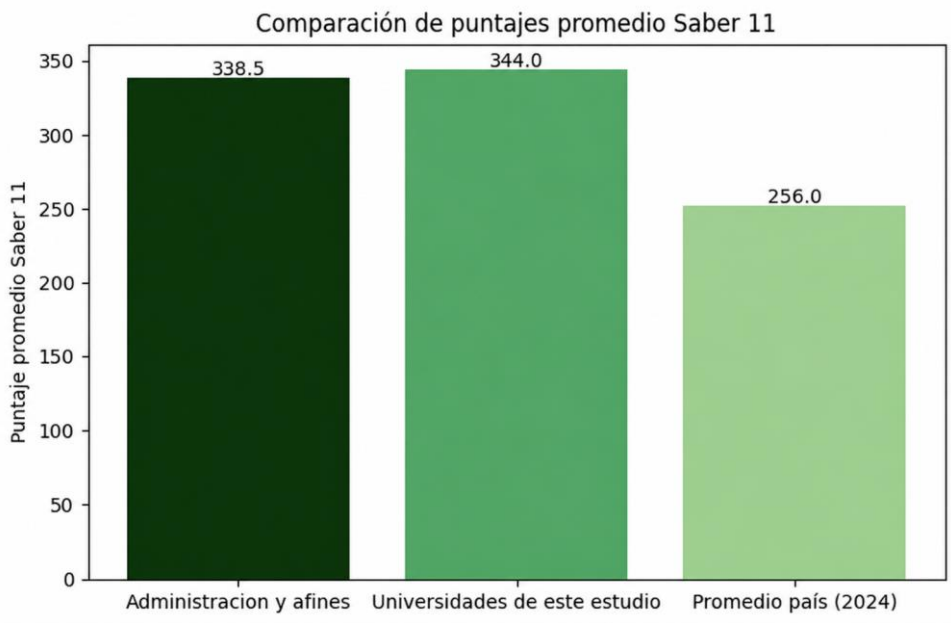


Figura 6. Distribución de puntaje Saber 11. Nota. Los puntajes corresponden al promedio global de Saber 11 en su escala original de 0 a 500 puntos.

Al comparar los puntajes promedio obtenidos en este estudio con el promedio nacional, se evidencia una diferencia considerable, donde los estudiantes de las universidades analizadas presentan un desempeño significativamente superior. Mientras que el promedio del país para el año 2024 se sitúa alrededor de 256 puntos, los estudiantes de estas instituciones alcanzan valores cercanos a 344 en promedio general y 338 en áreas de gestión (Administración y afines). Esta brecha sugiere que la muestra analizada no es representativa del conjunto nacional, sino que corresponde a un grupo con niveles de desempeño académico más altos. En este contexto, es importante señalar que un puntaje que podría considerarse “bajo” dentro de estas universidades, en realidad podría estar por encima del promedio nacional. Por tanto, la interpretación de los resultados debe realizarse únicamente bajo las universidades de interés, reconociendo que los estándares de desempeño dentro de estas instituciones son más exigentes en comparación con el contexto general del país.

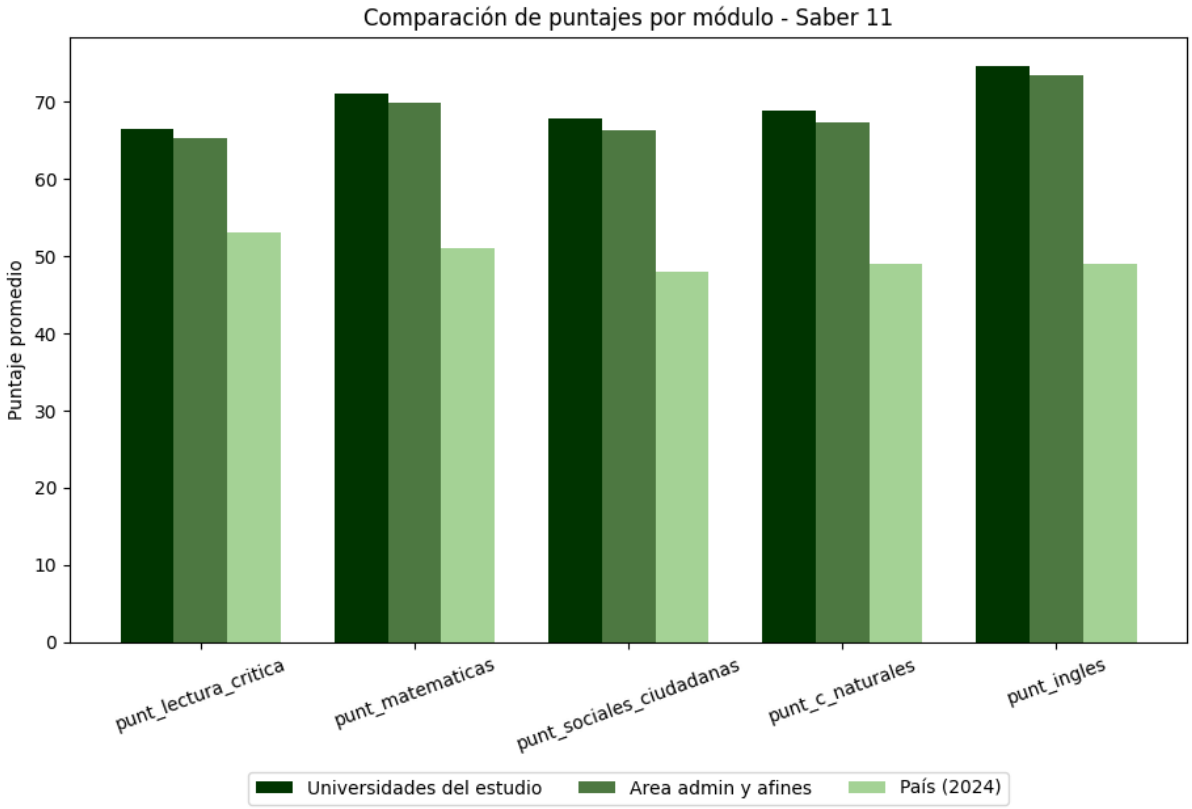


Figura 7. Comparación de puntajes por módulo Saber 11. Nota. Los puntajes están expresados en la escala original de cada módulo de Saber 11 (0-100 puntos). agrupan a los estudiantes de las cuatro sedes analizadas en el período 2014-2 a 2017. Los promedios no están estandarizados ni ponderados.

Al analizar la figura 7, se observa que tanto el conjunto total de estudiantes de las universidades consideradas como el grupo de áreas de administración y afines presentan un desempeño consistentemente superior al promedio nacional en todas las competencias evaluadas. No obstante, es importante interpretar estos resultados teniendo en cuenta la forma en que el ICFES construye el puntaje global, el cual se basa en una suma ponderada donde los módulos de Lectura Crítica, Matemáticas, Ciencias Naturales y Sociales y Ciudadanas tienen un peso equivalente (23.1% cada uno), mientras que el módulo de Inglés tiene una menor incidencia (7.6%). Esto implica que las diferencias observadas en los módulos principales tienen un mayor impacto en el puntaje global que aquellas registradas en inglés. Sin embargo, desde la perspectiva del acceso a la educación superior, esta ponderación no necesariamente refleja los criterios de admisión de las universidades,

los cuales pueden variar según el programa académico. En particular, para áreas como administración y afines, es más común que se otorgue mayor relevancia a competencias como matemáticas e inglés, debido a su relación directa con las habilidades requeridas en estos campos. En consecuencia, el análisis por módulos permite una comprensión más precisa del perfil académico de los estudiantes, evidenciando que, aunque los puntajes globales pueden ser similares, las diferencias específicas en ciertas competencias pueden tener implicaciones importantes en los procesos de selección y desempeño académico posterior.

5.1.2 Resultados preliminares de la prueba Saber Pro

En esta sección se presentan los resultados preliminares del examen Saber Pro del análisis exploratorio, cuyo objetivo es describir de forma general el perfil de los estudiantes de Administración y áreas afines de las cuatro sedes universitarias analizadas y los patrones observados en su desempeño académico en ambas pruebas. Estos hallazgos iniciales permiten contextualizar el punto de partida del análisis de clustering y establecer una referencia.

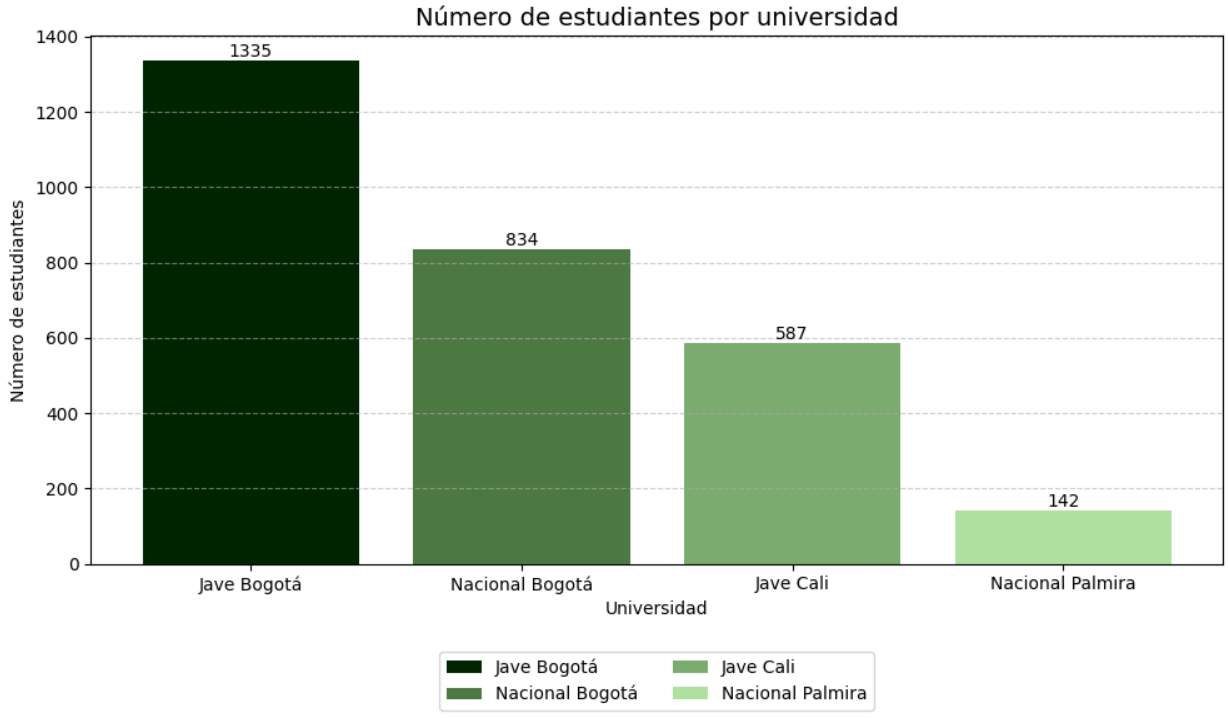


Figura 8. Cantidad de estudiantes por cada universidad en la muestra del estudio.

La distribución de estudiantes de Administración y áreas afines en la muestra final muestra que la Pontificia Universidad Javeriana sede Bogotá concentra la mayor cantidad de registros, seguida por la Universidad Nacional de Colombia sede Bogotá, la Pontificia Universidad Javeriana sede Cali y por último la Universidad Nacional de Colombia sede Palmira. Esta distribución es coherente con el tamaño y la oferta académica de cada institución: las sedes de Bogotá, al ser las más grandes y con mayor trayectoria en programas de gestión y administración, concentran naturalmente una proporción más alta de estudiantes. El desbalance entre sedes no constituye una limitación metodológica estricta para el clustering, dado que este tipo de análisis no requiere muestras equilibradas para la identificación de patrones. Sin embargo, sí debe tenerse en cuenta al momento de interpretar comparaciones entre instituciones, particularmente en el caso de la sede Palmira, cuyo tamaño muestral reducido limita la robustez de cualquier inferencia específica sobre esa sede y refuerza la necesidad de leer los resultados siempre bajo un contexto específico.

El promedio del puntaje en la prueba Saber Pro constituye una medida del desempeño académico de los estudiantes al momento de su egreso, permitiendo comparar de manera general. Este indicador resume el resultado global de la evaluación estandarizada y sirve como punto de referencia para el análisis.

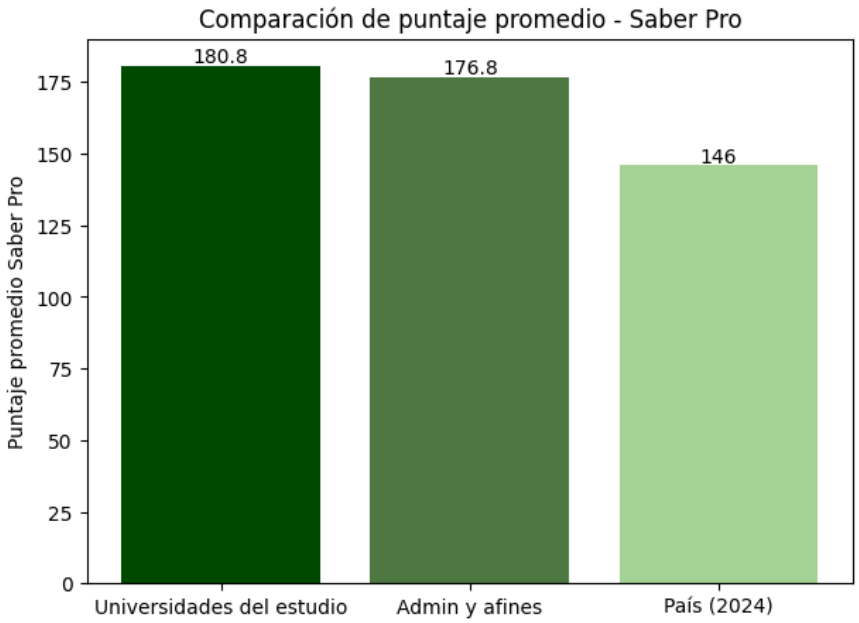


Figura 9. Comparación del puntaje promedio en Saber Pro. Nota. en su escala original (0-300 puntos por módulo)

En la figura 9 se observa que tanto el conjunto total de estudiantes de las universidades consideradas como el grupo perteneciente a áreas de administración y afines presentan resultados superiores al promedio nacional reportado para 2024, situado en 146 puntos. Esta diferencia evidencia que la muestra analizada también corresponde a un grupo con un nivel de desempeño académico más alto que el promedio nacional, lo cual es consistente con lo observado previamente en los resultados de Saber 11. Asimismo, la cercanía entre los promedios del grupo de administración y afines y el total de estudiantes sugiere que, a nivel general, no existen brechas significativas en el desempeño entre estas áreas dentro de las universidades analizadas.

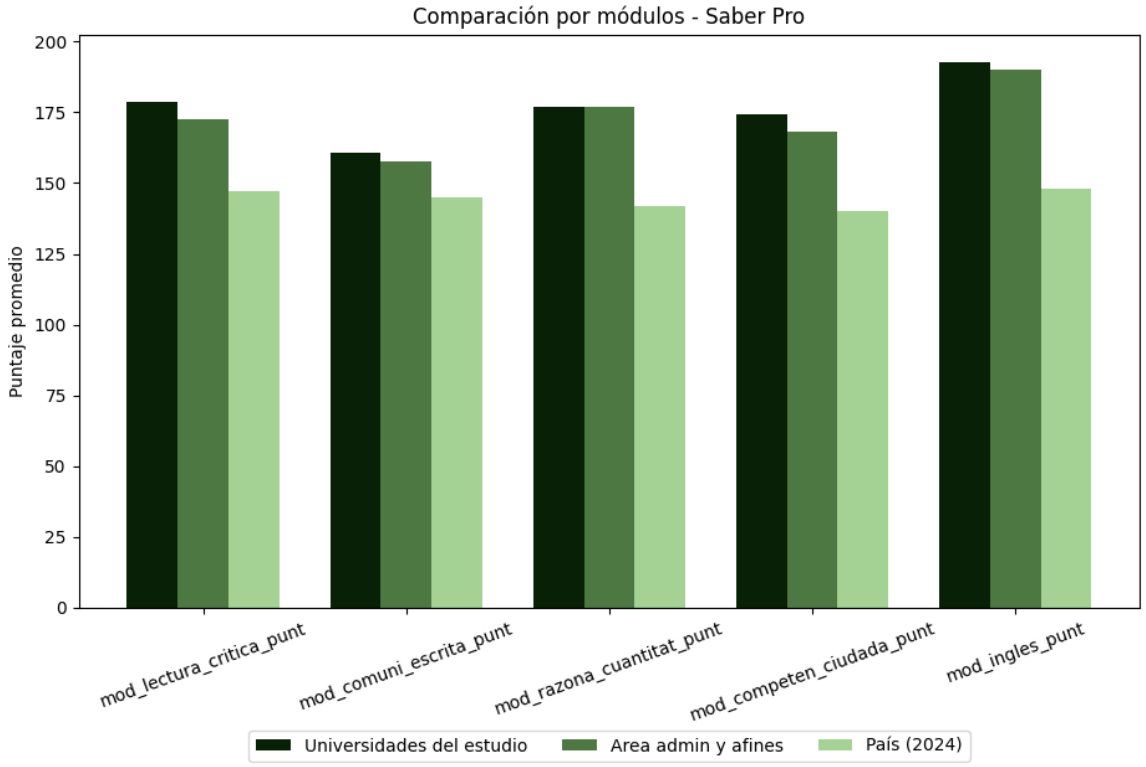


Figura 10. Distribución por cada módulo en prueba Saber Pro. Nota. Los puntajes están en la escala original de Saber Pro. "Universidades del estudio" y "Área admin y afines" agrupan a los estudiantes de las cuatro sedes analizadas en el período 2018-2022.

En esta En la figura 10 se observa que, a diferencia de Saber 11, todas las competencias evaluadas tienen el mismo peso dentro del desempeño general, lo que permite una comparación directa entre áreas sin la influencia de ponderaciones diferenciadas. Una hipótesis es que los estudiantes presenten fortalezas específicas según su formación, como un mayor desempeño en razonamiento cuantitativo para programas de administración y afines, o en lectura crítica para áreas como derecho y afines. En los resultados obtenidos, se evidencia que el conjunto total de estudiantes de las universidades analizadas presenta puntajes superiores al grupo de administración y afines en la mayoría de los módulos, con la excepción del módulo de razonamiento cuantitativo, donde ambos grupos alcanzan valores prácticamente iguales, lo cual resulta coherente para esta área de formación. Adicionalmente, al igual que en el caso de Saber 11, se mantiene una diferencia considerable frente al promedio nacional en todos los módulos, lo que fortalece la idea de que la

muestra analizada corresponde a estudiantes con un nivel de desempeño académico superior al del contexto general del país.

6. Resultados Finales

6.1 Clústeres obtenidos únicamente para el área de administración y afines

En esta sección se presentan los resultados obtenidos a partir de la aplicación de diferentes métodos de agrupamiento sobre el conjunto de datos analizado. Con el objetivo de identificar la técnica más adecuada para la formación de clústeres, se evaluaron seis algoritmos de agrupamiento, los cuales fueron comparados mediante métricas internas de validación que permiten medir la calidad y consistencia de los grupos generados.

La evaluación del desempeño de los modelos se realizó utilizando tres métricas ampliamente empleadas en el análisis de clustering: el coeficiente de Silhouette, el índice de Davies–Bouldin y el índice de Calinski–Harabasz. Estas métricas permiten analizar simultáneamente la cohesión interna de los clústeres y la separación entre grupos, proporcionando criterios cuantitativos para seleccionar el modelo más apropiado.

Las tres métricas tienen direcciones de interpretación distintas. Silhouette: valores entre -1 y 1, donde valores más altos indican mejor cohesión interna y separación entre grupos. Davies-Bouldin: valores menores indican mejor separación entre clústeres. Calinski-Harabasz: valores mayores indican mejor razón entre dispersión inter-clúster e intra-clúster. Todos los modelos fueron evaluados con k=4 sobre el mismo conjunto de variables estandarizadas, garantizando comparabilidad directa.

A continuación, se presentan los resultados comparativos entre los métodos evaluados, seguido de la interpretación del modelo seleccionado y el análisis de los clústeres obtenidos en relación con el desempeño académico de los estudiantes.

Modelo	Silhouette	Davies-Bouldin	Calinski-Harabasz
K-Means	0.15	1.81	394.65
Gaussian Mixture	0.14	4.44	155.26
Fuzzy C-Means	0.07	2.64	354.83
Agglomerative	0.10	2.07	335.26
DBSCAN	0.14	4.72	29.88
Spectral Clustering	0.41	0.57	19.2

Tabla 3. Comparación de métodos. Nota. todo el conjunto de datos estandarizado.

Los resultados de las métricas de validación interna (Tabla 3) son consistentes con la hipótesis metodológica planteada del análisis, dado que los datos académicos del estudio presentan una distribución continua y un alto grado de solapamiento entre niveles de desempeño, se esperaba obtener métricas internas de magnitud moderada y no valores que indicaran particiones perfectamente separables. Los valores observados de Silhouette (entre 0.07 y 0.15 para los algoritmos con particiones balanceadas) confirman esta expectativa y reflejan la naturaleza gradual del fenómeno académico evaluado, no una deficiencia en el ajuste de los modelos [16]. Las variables sociodemográficas e institucionales fueron reservadas para la etapa de validación externa mediante el coeficiente de Cramér's V, donde su rol es precisamente el de caracterizar y enriquecer los perfiles ya identificados, no el de construirlos.

En cuanto a los valores de las métricas, Una interpretación posible ante coeficientes de Silhouette entre 0.07 y 0.15 sería considerar que los algoritmos no logran identificar estructura en los datos. Sin embargo, esta lectura no es adecuada en el contexto de datos académicos continuos. Bajo el supuesto teórico de que el desempeño estudiantil no se distribuye en grupos discretos sino a lo largo de un continuo de competencias con fronteras graduales, y que por tanto los algoritmos de agrupamiento generaran transiciones suaves en lugar de particiones nítidamente separadas. Esta expectativa metodológica se confirma con los resultados obtenidos y es coherente con la evidencia reportada en estudios similares de clustering aplicado al rendimiento académico [16], donde valores

moderados de Silhouette son la norma para este tipo de poblaciones. Por tanto, los valores observados deben interpretarse como reflejo de la naturaleza del fenómeno analizado y no como una deficiencia en el desempeño de los modelos.

En cuanto al modelo de Spectral Clustering, si bien presenta las métricas internas más favorables con un coeficiente de Silhouette de 0.4123 y un índice de Davies-Bouldin de 0.5706, ambos superiores a los demás algoritmos evaluados este no fue seleccionado como modelo principal. Como se aprecia en la Figura 10, los grupos generados por este método se concentran casi en su totalidad en un único clúster masivo, con apenas unos pocos puntos aislados conformando los grupos restantes. Este patrón indica que el algoritmo no logró una partición significativa de los datos, sino que identificó valores atípicos como clústeres separados, lo que explica sus métricas favorables sin que estas reflejen una segmentación real y útil de la población estudiantil. En datos de desempeño académico con distribución continua y sin separaciones geométricas nítidas, Spectral Clustering tiende a comportarse de esta manera, optimizando la métrica sin producir grupos balanceados ni interpretables.

Antes de presentar las visualizaciones de los clústeres, conviene aclarar qué representan los ejes de las figuras. Dado que el modelo de clustering trabaja simultáneamente sobre diez variables (los cinco módulos genéricos de Saber 11 y los cinco módulos genéricos de Saber Pro), no es posible visualizar directamente las agrupaciones en un plano bidimensional. Para superar esta limitación, se aplicó un Análisis de Componentes Principales (PCA), técnica que reduce las diez dimensiones originales a las dos que explican la mayor variabilidad conjunta de la muestra. Cada eje del gráfico corresponde, por tanto, a una combinación lineal de los puntajes originales con coeficientes calculados por el algoritmo.

El análisis de los pesos asignados por PCA permite interpretar el significado pedagógico de cada eje. La Componente 1 asigna pesos positivos y de magnitud similar a las diez variables, por lo que

puede interpretarse como un "nivel general de desempeño académico": los estudiantes ubicados en el extremo positivo de este eje presentan puntajes altos en todas las competencias de ambas pruebas, mientras que los del extremo negativo presentan puntajes bajos de forma generalizada. La Componente 2 muestra un patrón de contrastante en el que los puntajes de Saber 11 cargan principalmente con signo negativo y los de Saber Pro con signo positivo. Esto permite interpretar este eje como un "contraste de desempeño entre Saber 11 y Saber Pro": los estudiantes ubicados en el extremo positivo destacan relativamente más en Saber Pro que en Saber 11, mientras que los del extremo negativo presentan el patrón inverso.

En términos pedagógicos, esto significa que la posición de cada estudiante en el plano resume dos aspectos complementarios de su trayectoria académica: su nivel general de competencias (eje horizontal) y la dirección relativa de su cambio entre la educación media y la universitaria (eje vertical). Por tanto, un estudiante con valor positivo alto en este eje es alguien cuyo desempeño fue mejor en Saber Pro que en Saber 11, mientras que un estudiante con valor negativo es alguien que tuvo mejores resultados en Saber 11 que en Saber Pro.

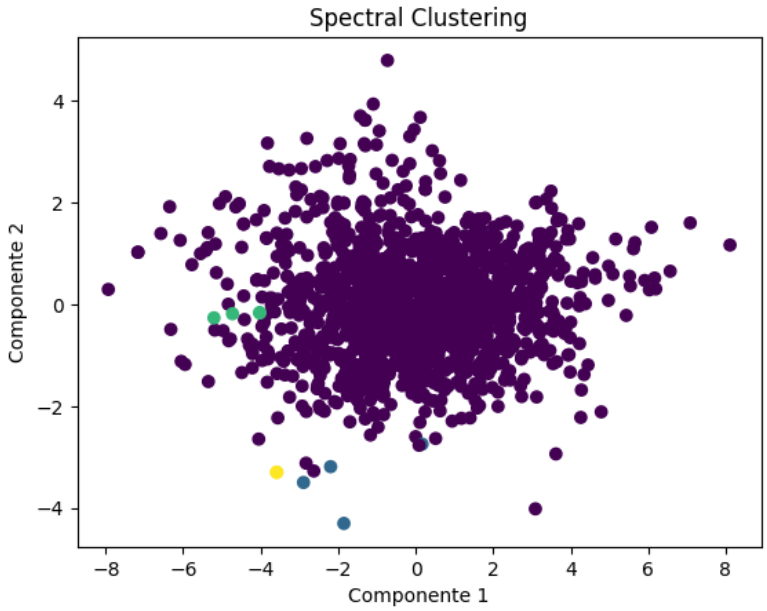


Figura 11. Clústeres obtenidos con Spectral.

La selección de K-Means con cuatro clústeres se fundamenta en la aplicación conjunta de los seis criterios definidos en la sección metodológica. K-Means obtuvo el mayor índice de Calinski-Harabasz (394.65) entre los algoritmos con particiones balanceadas, lo que indica una buena separación entre grupos en relación con su dispersión interna, y un índice de Davies-Bouldin de 1.81, razonable para datos educativos. Además, los cuatro clústeres generados presentan una jerarquía coherente y progresiva de desempeño que se mantiene tanto en Saber 11 como en Saber Pro, lo que permite caracterizar perfiles académicos diferenciados con sentido pedagógico real. Asimismo estudios previos de referencia en contextos muy similares respaldan esta elección, para análisis de clustering para clasificar el rendimiento académico en educación, encontraron que K-Means representa de manera consistente y eficiente las estructuras de desempeño estudiantil en comparación con otros métodos de agrupamiento, recomendando su uso en contextos donde la interpretabilidad de los perfiles es tan importante como la calidad métrica del modelo [16].

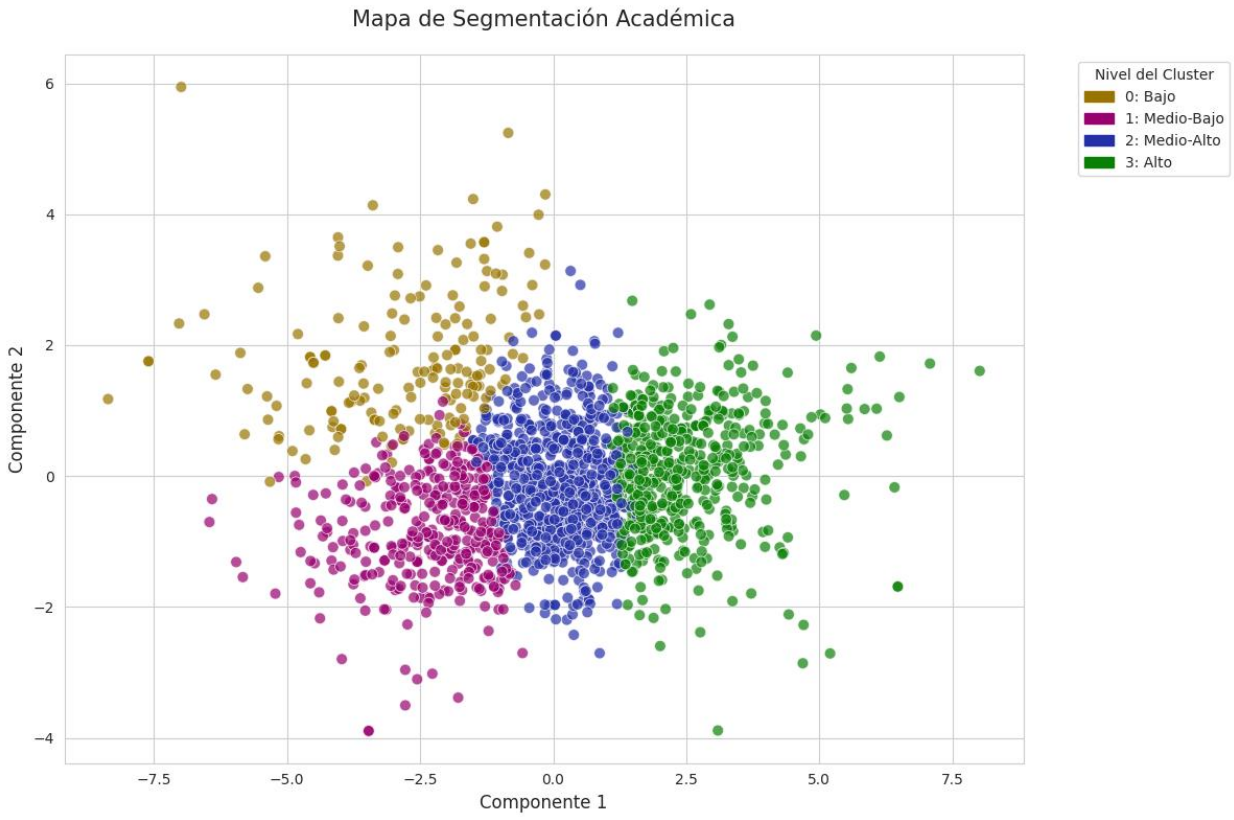


Figura 12. Clúster obtenido con K-means.

El mapa de dispersión muestra una nube de datos continua (figura 12), lo que explica por qué no vemos "huecos" o espacios vacíos entre los colores. En educación, el talento y el desempeño no se dividen tan fácilmente; no hay una barrera física que separe a un estudiante promedio de uno sobresaliente. Lo que se observa es una transición fluida de competencias, donde los estudiantes se solapan en sus puntajes. Por ello, el algoritmo K-Means actúa como un "organizador" que traza fronteras lógicas donde hay mayor densidad, permitiendo clasificar perfiles que, aunque están cerca en el gráfico, muestran puntuaciones distintas.

Antes de proceder con el análisis de los perfiles identificados, la Tabla 4 presenta la distribución de estudiantes entre los cuatro clústeres obtenidos, lo que permite evaluar el balance de la segmentación y contextualizar los análisis posteriores.

Cluster	N° Estudiantes	Porcentaje (%)
0	169	9.65
1	364	20.78
2	704	40.18
3	515	29.39

Tabla 4. Distribución de los clústeres.

Como se observa en la distribución, los cuatro clústeres presentan tamaños representativos, lo que indica que K-Means generó una segmentación con representación en todos los niveles, sin producir grupos vacíos o desproporcionados. Bajo el supuesto de que la muestra proviene de instituciones académicamente selectivas (Pontificia Universidad Javeriana y Universidad Nacional de Colombia), era razonable anticipar que la distribución de los clústeres no fuera uniforme y que el grueso de la muestra se concentrara en niveles intermedios-altos de desempeño. La concentración del 40.18% en el Clúster 2 (Avanzado) y el reducido tamaño del Clúster 0 (Inicial), con solo 9.65%, son consistentes con esto.

Los clústeres se ordenan jerárquicamente con base en los puntajes promedio de los centroides en todos los módulos de Saber 11 y Saber Pro simultáneamente, lo que significa que un estudiante pertenece al Clúster 0 porque sus puntajes en ambas pruebas son consistentemente los más bajos dentro de la muestra, no porque haya tenido un buen desempeño en Saber 11 y uno bajo en Saber Pro. Esta distinción es importante: los perfiles reflejan trayectorias de desempeño sostenidas a lo largo del tiempo, como se confirma en la Figura 12, donde la jerarquía entre clústeres se mantiene coherente en ambas pruebas.

La Figura 12 presenta los centroides estandarizados de cada clúster, es decir, el valor promedio de cada variable expresado en desviaciones estándar respecto a la media de la muestra. Esta representación permite comparar el perfil relativo de cada grupo independientemente de la escala original de cada módulo.

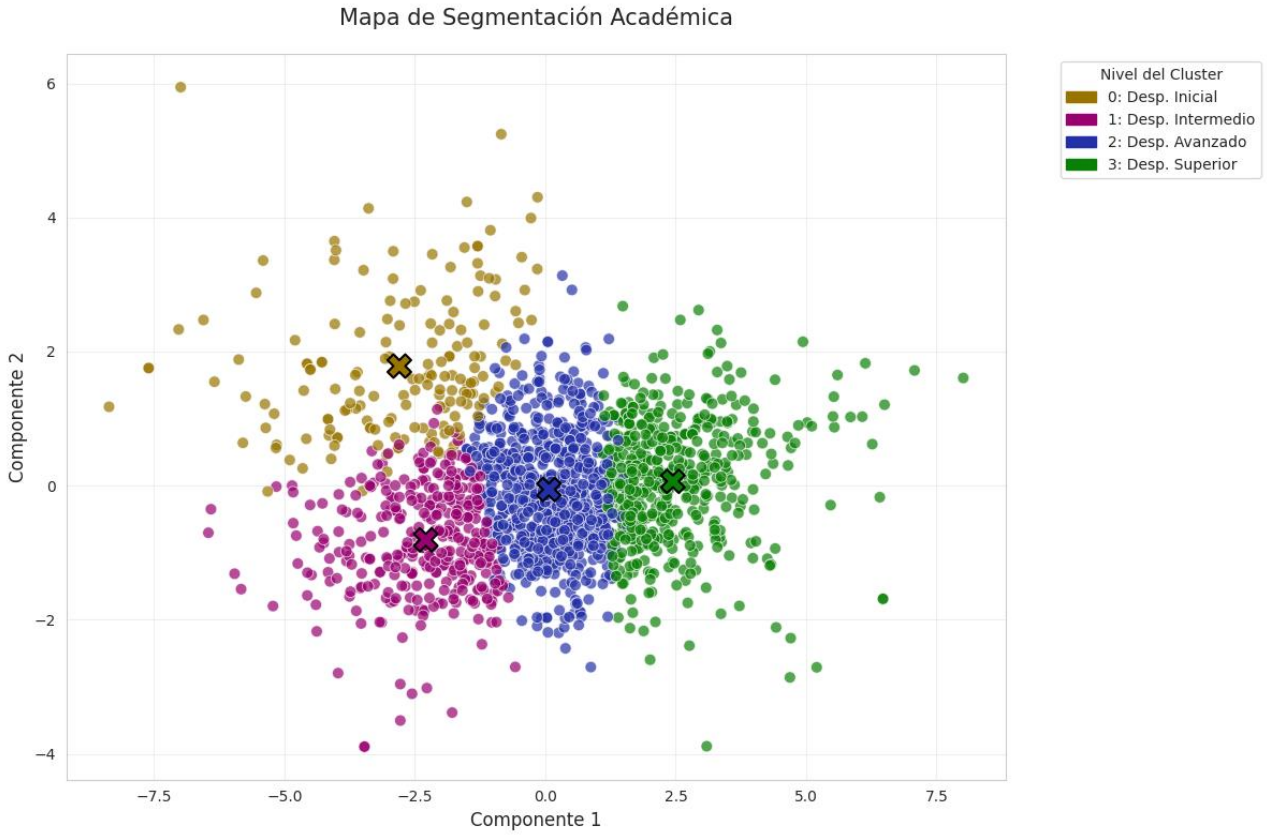


Figura 13. Centroides obtenidos con K-means. Nota. Los ejes corresponden a las dos primeras componentes principales del espacio de variables estandarizadas, utilizadas únicamente con fines de visualización

Los centroides estandarizados permiten caracterizar de manera clara los perfiles de desempeño de los clústeres identificados. Dado que los grupos fueron reordenados según el promedio total estandarizado, el clúster 0 corresponde al nivel de menor desempeño con un promedio de -0.90 desviaciones estándar por debajo de la media de la muestra, y el clúster 3 al de mayor desempeño con un promedio de +0.75 desviaciones estándar por encima. El clúster 0 presenta valores consistentemente negativos en todas las variables de Saber 11 y Saber Pro, evidenciando un bajo rendimiento académico generalizado a lo largo de toda la trayectoria. El clúster 1, con un promedio de -0.71, también se ubica por debajo de la media de la muestra, aunque de forma menos pronunciada, configurando un perfil de desempeño bajo-intermedio. El clúster 2, con un promedio de +0.03, deja ver un perfil académico un poco mejor que el intermedio y relativamente estable entre ambos momentos de evaluación. Finalmente, el clúster 3 exhibe valores positivos en todas las dimensiones de ambas pruebas, consolidándose como el grupo de mayor desempeño dentro de la muestra. En conjunto, estos resultados confirman que la segmentación obtenida captura trayectorias diferenciadas y ordenadas de desempeño académico, cuya composición y características sociodemográficas se analizan en detalle en la sección siguiente.

Con el propósito de facilitar la interpretación de los resultados y su comunicación práctica, los clústeres identificados fueron etiquetados siguiendo el criterio empleado por uno de los estudios guía [16], quienes denominaron sus grupos según niveles de rendimiento para permitir una identificación rápida de los perfiles estudiantiles. Es fundamental aclarar que las etiquetas Inicial, Intermedio, Avanzado y Superior son categorías analíticas construidas para comparar dentro de la muestra, y no valoraciones absolutas sobre la capacidad o el potencial de los estudiantes ni diagnósticos pedagógicos individuales. Como se señaló en la sección de resultados preliminares, los perfiles se ubican por encima del promedio nacional, por lo que estas denominaciones tienen sentido únicamente en el contexto relativo de las instituciones analizadas.

El Clúster 0, denominado Inicial, agrupa a los estudiantes cuyos puntajes promedio en ambas pruebas son los más bajos dentro de la muestra. Esto no implica una ausencia de capacidad, sino que sus resultados reflejan una base académica de ingreso más débil en relación con sus pares, que se mantiene en el egreso.

El Clúster 1, denominado Intermedio, corresponde a estudiantes que superan los puntajes del grupo anterior y muestran un desempeño más equilibrado entre módulos, aunque sin alcanzar la consistencia de los grupos superiores. Este grupo presenta la mayor volatilidad entre pruebas, con una proporción relevante de estudiantes que logran escalar al nivel Avanzado al egreso.

El Clúster 2, denominado Avanzado, agrupa a estudiantes con puntajes sólidos en la mayoría de los módulos evaluados, con un desempeño especialmente destacado en razonamiento cuantitativo, coherente con el perfil esperado para programas de Administración y áreas afines. Es el clúster más numeroso de la muestra, lo que sugiere que representa el perfil típico de los estudiantes de estas instituciones.

El Clúster 3, denominado Superior, corresponde al grupo de mayor desempeño en ambas pruebas. Sus integrantes mantienen puntajes consistentemente altos desde el ingreso hasta el egreso, con especial fortaleza en inglés, competencias ciudadanas y razonamiento cuantitativo. Dentro del contexto de la muestra, este grupo representa el techo de desempeño académico observado.

Es importante señalar que los clústeres de menor rendimiento deben interpretarse en su contexto como fue mencionado anteriormente, ya que corresponden a estudiantes de instituciones de alta selectividad como la Universidad Javeriana y la Universidad Nacional en sus 2 sedes (Cali y Palmira). Por ello, aunque presentan puntajes inferiores dentro de la muestra, no implican un bajo desempeño en términos absolutos a nivel nacional.

Con el fin de profundizar en la comprensión de los grupos identificados por el método de KMeans, se presenta a continuación un análisis comparativo de las puntuaciones promedio por

clúster, para facilitar la comparación visual entre ambas pruebas, los puntajes se expresan en una escala común de 0 a 100, aplicando un reescalado únicamente con fines de visualización.

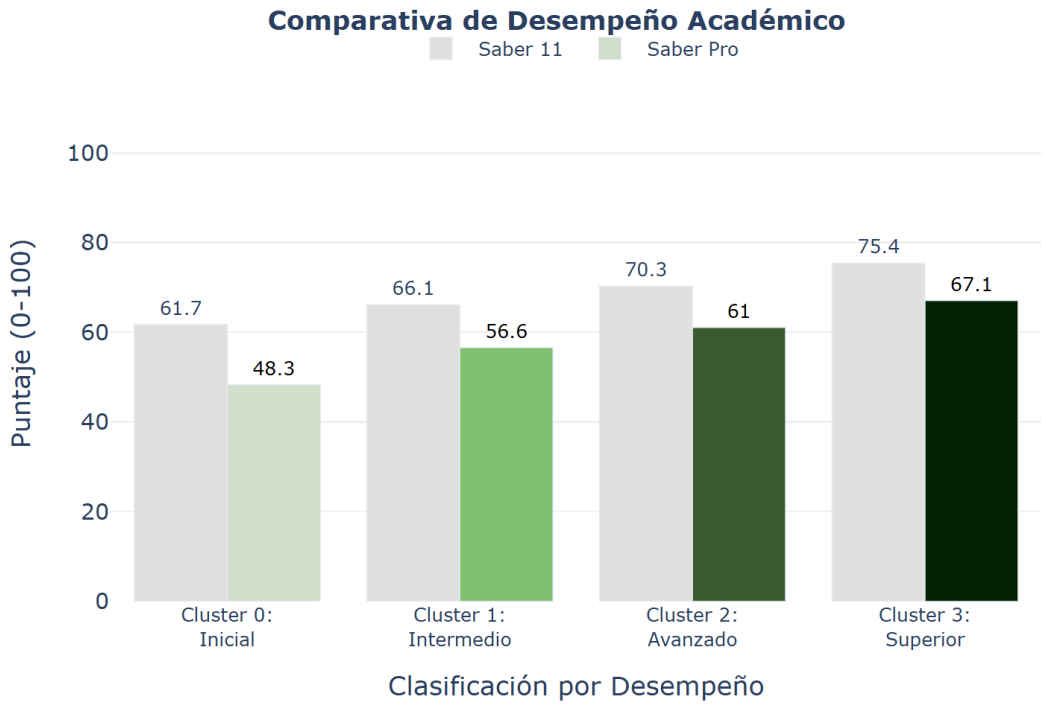


Figura 14. Comparativa entre puntaje de cada cluster. Nota. reescalado solo para esta visualización, no para el modelo.

Los puntajes están expresados en una escala 0-100 para facilitar la comparación visual entre Saber 11 y Saber Pro. Para construir esta escala, los puntajes de Saber Pro (originalmente en escala 0-300) fueron reescalados al rango 0-100. Estos valores son distintos a los centroides estandarizados utilizados internamente por el modelo K-Means (Figura 13).

Un hallazgo que merece atención es el descenso generalizado en los promedios entre Saber 11 y Saber Pro observable en todos los clústeres. Este fenómeno admite múltiples interpretaciones que no deben descartarse sin mayor evidencia. Una posibilidad es que la prueba Saber Pro, al ser percibida por muchos estudiantes como simplemente un requisito de grado, genere menor nivel de preparación y esfuerzo en comparación con Saber 11, cuyo resultado incide directamente en el acceso a la educación superior. Otra interpretación, igualmente válida, es que las universidades reciben estudiantes con un perfil académico sólido y logran mantenerlo a lo largo de la carrera, lo

cual ya representa un resultado positivo en sí mismo. Sin embargo, también se pueden dar casos en los que la formación universitaria puede tener un efecto nivelador para estudiantes que ingresan con una base más débil, mejorando su posición relativa al egreso. En cualquier caso, este descenso no debe leerse como una señal inequívoca de falta de valor agregado institucional, sino como un patrón que merece ser explorado con mayor profundidad en estudios futuros.

Con el fin de analizar la estabilidad del desempeño académico a lo largo de la trayectoria educativa, se construyó una matriz de transición que cruza el nivel de desempeño en Saber 11 con el clúster asignado en Saber Pro. Esta herramienta permite visualizar en qué medida el rendimiento de ingreso se mantiene, mejora o deteriora al egreso, ofreciendo una perspectiva longitudinal complementaria al análisis de clustering.

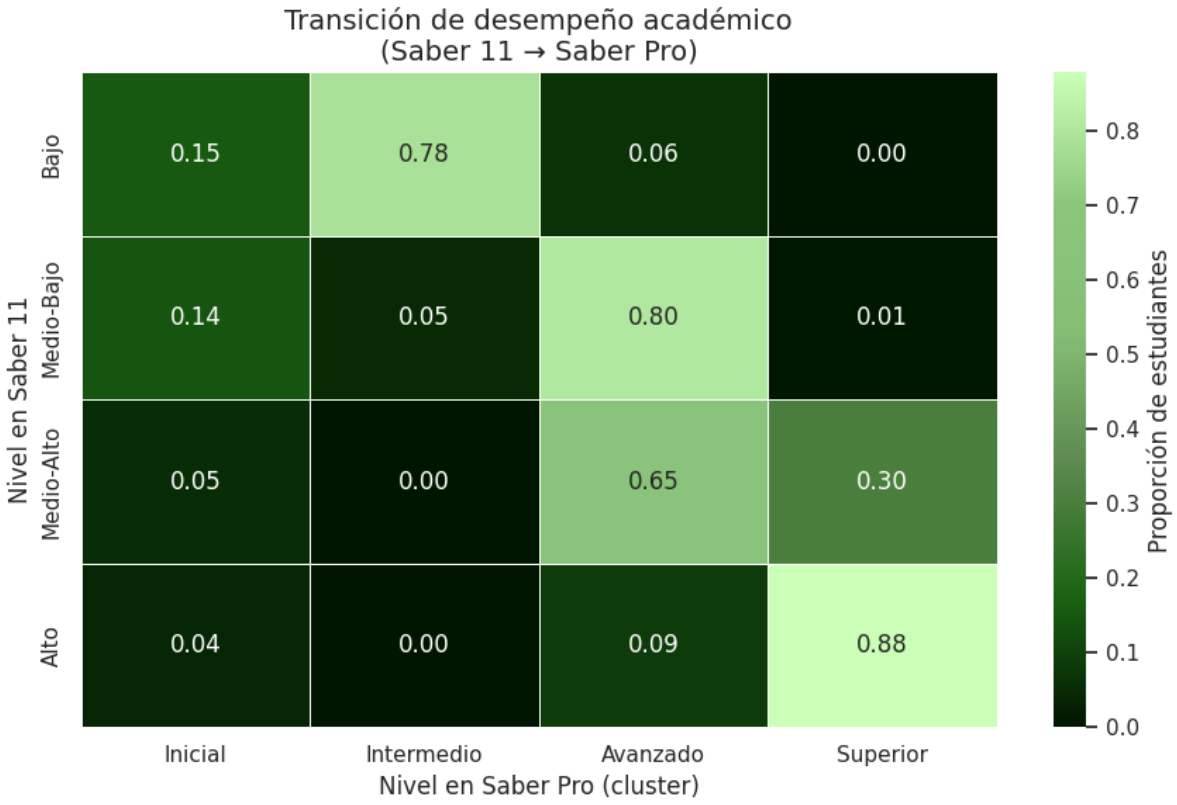


Figura 15. Matriz de transición de saber 11 a saber Pro.

Cada celda representa la proporción de estudiantes que, partiendo de un nivel de desempeño en

Saber 11 (filas), fueron asignados a un clúster específico en Saber Pro (columnas). Los valores de cada fila suman 1. Los niveles de desempeño en Saber 11 (Bajo, Medio-Bajo, Medio-Alto, Alto) corresponden a la clasificación oficial del ICFES por cuartiles del puntaje global [12]. Los clústeres en Saber Pro (Inicial, Intermedio, Avanzado, Superior) provienen del modelo K-Means con k=4 aplicado.

Para la creación de esta matriz se realizaron 2 pasos. Primero, los estudiantes fueron clasificados en cuatro niveles de desempeño en Saber 11 Bajo, Medio-Bajo, Medio-Alto y Alto utilizando cuartiles (como lo hace el ICFES oficialmente) sobre el puntaje global de esa prueba, sin ninguna relación con los clústeres obtenidos posteriormente. Segundo, cada estudiante fue ubicado en su clúster correspondiente, Inicial, Intermedio, Avanzado y Superior, resultado del proceso de clustering descrito anteriormente. Cada celda de la matriz representa la proporción de estudiantes de un nivel de entrada en Saber 11 que quedaron asignados a cada clúster de salida, donde los valores de cada fila suman 1.

Los resultados revelan un patrón claro de persistencia del desempeño en los extremos. Se observa que el 88% de los estudiantes con nivel Alto en Saber 11 se ubican en el clúster de mayor desempeño, mientras que el 78% de los estudiantes con nivel bajo se concentran en el clúster Intermedio. Estos resultados reflejan una alta consistencia en los niveles de desempeño, aunque corresponden a proporciones muestrales y no a probabilidades independientes. Este último dato es particularmente relevante: indica que incluso los estudiantes con los puntajes más bajos de ingreso dentro de esta muestra tienden a ubicarse en niveles intermedios al egreso, lo que podría interpretarse como un efecto positivo de la formación universitaria en este segmento. Los niveles medios de Saber 11 muestran también una alta persistencia: el 80% de los estudiantes Medio-Bajo permanece en el clúster Avanzado, y el 65% de los Medio-Alto se mantiene en ese mismo nivel, con un 30% que logra escalar al clúster Superior. En conjunto, la diagonal dominante de la matriz

confirma que el nivel de desempeño en Saber 11 es un marcador relevante de la trayectoria académica posterior.

6.2 Perfilamiento posterior al agrupamiento

Tras haber validado la consistencia técnica de la segmentación mediante métricas de cohesión y separación, este apartado se enfoca en analizar la composición de cada clúster identificado, describiendo sus características predominantes. Mientras que la fase anterior permitió organizar los datos basándose estrictamente en puntajes y distancias matemáticas, en esta etapa realizaremos una inmersión en las variables descriptivas. El objetivo es trascender la etiqueta numérica del clúster para identificar los comportamientos, rasgos sociodemográficos o patrones específicos que definen a cada grupo.

Para profundizar en la comprensión de los perfiles identificados, se realizó un proceso de validación externa de los clústeres a partir de variables sociodemográficas e institucionales que no fueron incluidas en la construcción del modelo, con el fin de identificar qué factores contextuales presentan mayor asociación con los grupos obtenidos. Para ello se utilizó el coeficiente de Cramér's V, una medida de asociación entre variables categóricas derivada del estadístico Chi-cuadrado que toma valores entre 0 y 1. Su ventaja frente al Chi-cuadrado simple es que normaliza el estadístico por el tamaño de la muestra y las dimensiones de la tabla de contingencia, lo que permite comparar la fuerza de asociación entre variables con diferente número de categorías de forma directa y consistente [19]. Este análisis tiene un carácter exploratorio y complementario dentro del estudio, orientado a identificar qué variables del contexto familiar, institucional y socioeconómico se relacionan en mayor medida con los perfiles de desempeño identificados. Los resultados completos con los valores exactos se presentan en la Tabla 5, ordenados de mayor a menor grado de asociación. Las variables con mayor asociación corresponden principalmente a características del contexto educativo y socioeconómico del estudiante, destacándose el núcleo de pregrado, el

departamento de ubicación del colegio, la institución de origen, el programa académico, el acceso a computador en el hogar, la naturaleza del colegio, la tenencia de automóvil y el estrato de la vivienda. Adicionalmente, se incluyeron las variables de nivel educativo del padre y de la madre, las cuales presentan asociaciones más moderadas, pero complementan la perspectiva sobre el capital educativo del hogar como factor contextual relevante.

#	Variable	Cramér's V
1	estu_nucleo_pregrado	0.2581
2	cole_depto_ubicacion	0.2181
3	inst_origen	0.2179
4	estu_depto_presentacion_sb11	0.2175
5	estu_prgm_academico	0.2155
6	fami_tienecomputador_sb11	0.2091
7	cole_naturaleza	0.1969
8	fami_tieneautomovil_sb11	0.1719
9	fami_estratovivienda_sb11	0.1593
10	estu_genero_sb11	0.1446
11	fami_educacionpadre_sb11	0.1406
12	fami_educacionmadre_sb11	0.1254

Tabla 5. Variables seleccionadas por valor de Cramer 's V.

El coeficiente de Cramér's V toma valores entre 0 y 1, donde 0 indica ausencia de asociación y 1 asociación perfecta. La interpretación práctica empleada en este estudio sigue los rangos convencionales: valores entre 0 y 0.1 corresponden a una asociación débil, entre 0.1 y 0.3 a una moderada, y superiores a 0.3 a una fuerte [19]. Todas las variables presentadas se cruzan con la pertenencia a los cuatro clústeres identificados en la muestra final del área de Administración y áreas afines.

6.3 Análisis de variables seleccionadas

A partir del ranking de asociación obtenido mediante Cramér's V, se seleccionaron las variables con mayor relevancia contextual para realizar un análisis descriptivo detallado por clúster. Este análisis no busca establecer relaciones causales entre estas variables y el desempeño académico, sino describir los patrones de distribución que emergen al cruzar los perfiles de clustering con el contexto de origen de los estudiantes, aportando una lectura más completa y contextualizada de los grupos identificados.

6.3.1 Análisis de núcleo de pregrado de los estudiantes del clúster

Se analiza la variable estu_nucleo_pregrado, la cual agrupa los programas académicos que tiene el área seleccionada de estudio. Aquí se puede evaluar si existen diferencias en los perfiles de desempeño académico según el campo de formación. A través de este análisis, se busca identificar posibles patrones de concentración de estudiantes en determinados clústeres según su carrera académica.

El análisis se restringe a los núcleos de pregrado asociados al área de gestión, administración y afines, de acuerdo con la clasificación del ICFES. Por esta razón, el número de carreras es reducido, lo cual permite una comparación más clara y enfocada entre los perfiles de desempeño identificados.

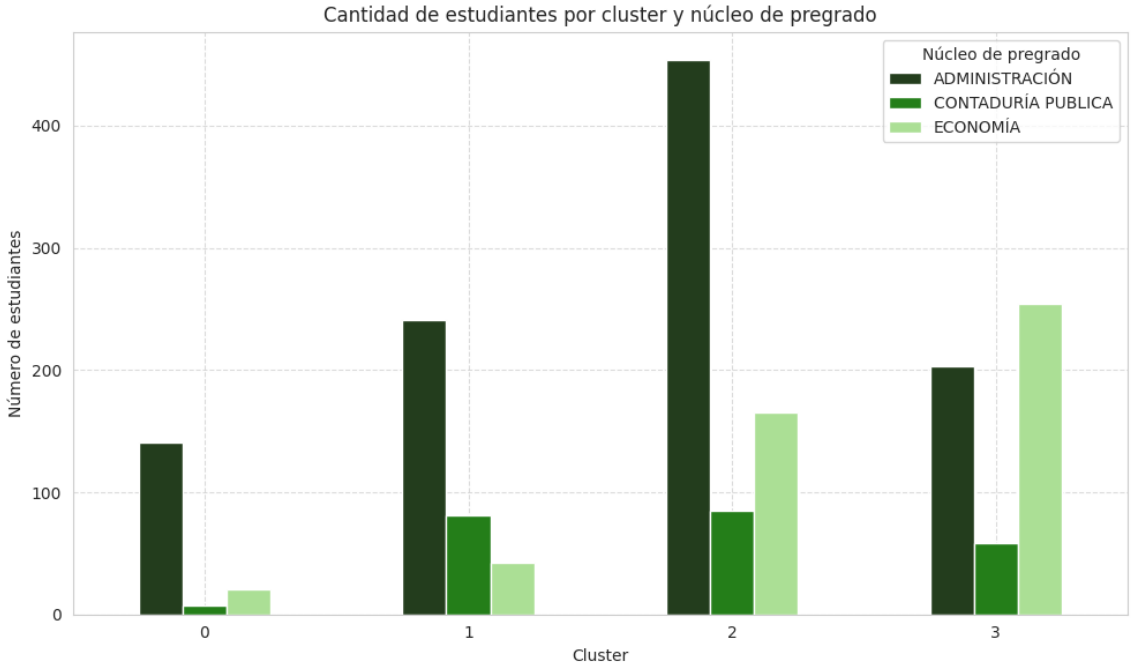


Figura 16. Cantidad de estudiantes de cada carrera en cada clúster.

Como se observa en la figura 16, hay una mayor cantidad de estudiantes en el clúster 2, y en general una mayor cantidad para la carrera de administración. Aunque en el peor clúster haya más estudiantes de administración no implica que estos sean los peores estudiantes ya que no solo tiene muchos más estudiantes que las demás carrera, sino que además la mayoría de estudiantes de administración están ubicados en el clúster con desempeño avanzado (clúster 2), sin embargo, los estudiantes de economía mantienen la punta como los mejores estudiantes de la muestra, quedando la mayoría en el clúster 3 (desempeño académico superior).

Clúster	ADMINISTRACIÓN	ECONOMÍA	CONTADURÍA PÚBLICA
0	0.834 %	0.124 %	0.041 %
1	0.662 %	0.115 %	0.223 %
2	0.645 %	0.234 %	0.121 %
3	0.394 %	0.493 %	0.113 %

Tabla 6. Proporción por núcleo de pregrado según clúster

El análisis del núcleo de pregrado evidencia diferencias significativas en la distribución de los estudiantes entre los clústeres. En particular, el programa de Administración concentra una alta

proporción de estudiantes en el clúster de menor desempeño (83.4%) como se mencionó anteriormente, mientras que su participación disminuye considerablemente en el clúster de mayor desempeño (39.4%). En contraste, el programa de Economía presenta el comportamiento opuesto, incrementando su participación del 12.4% en el clúster bajo al 49.3% en el clúster alto. Por su parte, Contaduría Pública muestra una distribución más equilibrada entre los clústeres.

6.3.2 Análisis del departamento donde está ubicado el colegio del estudiante y donde presentó el examen Saber 11

Con el fin de analizar posibles diferencias territoriales en el desempeño académico, se examina la variable departamento de ubicación del establecimiento educativo que coinciden con el departamento donde el estudiante presentó el examen Saber 11. Esta exploración busca identificar si existen patrones geográficos en la distribución de los estudiantes entre los clústeres.

También se tiene la variable de donde presentó el estudiante el examen Saber 11, aunque la información es prácticamente la misma al departamento donde está el colegio donde se graduó el estudiante por lo que estas 2 variables van de la mano y solo se usará una para el análisis la cual será el departamento del colegio donde se graduó.

La muestra analizada proviene principalmente de instituciones ubicadas en Bogotá y Cali, la distribución absoluta de estudiantes por departamento se encuentra influenciada por esta concentración geográfica. Por esta razón, el análisis se centra en medidas relativas, como proporciones y diferencias entre clústeres

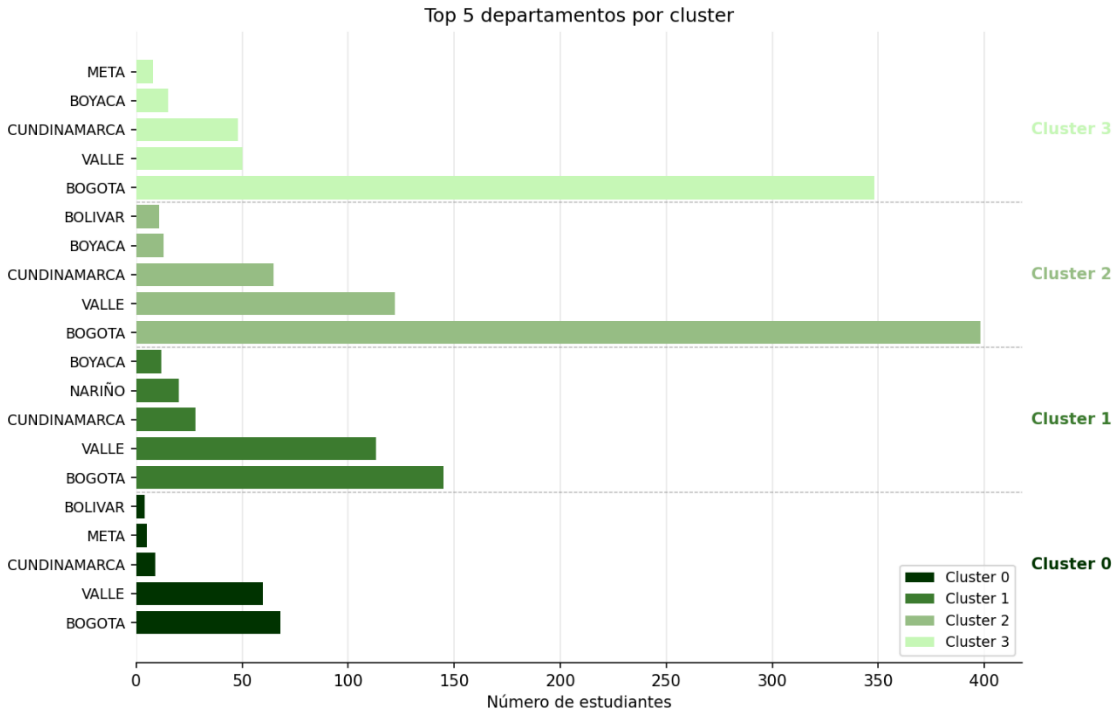


Figura 17. Cantidad de estudiantes de cada departamento en cada clúster.

Los cinco departamentos con mayor representación aparecen de forma consistente en todos los clústeres, lo que refleja su peso demográfico. Bogotá presenta un desempeño superior en todos los grupos considerados, siendo especialmente notable en el Cluster 2 (n=399) y el Cluster 3 (n=347), la capital concentra tanto a estudiantes de alto desempeño como a quienes presentan mayores dificultades, esto también debido a que su cantidad de estudiantes es mucho mayor que los demás departamentos. Valle del Cauca y Cundinamarca también aparecen en todos los clústeres, aunque con participaciones más moderadas.

El análisis de la variable departamento de ubicación del establecimiento educativo presenta limitaciones importantes asociadas a la composición de la muestra. En particular, dado que los datos provienen de instituciones con sedes principales en Bogotá y Cali, existe una sobrerrepresentación de estudiantes provenientes de estos territorios, lo que dificulta la comparación directa con otros departamentos que cuentan con un número reducido de observaciones, por lo que se hace difícil la interpretación de brechas territoriales.

Asimismo, en aquellos departamentos con baja cantidad de estudiantes, pequeñas variaciones en los resultados pueden generar asociaciones aparentemente fuertes con determinados clústeres, sin que estas reflejen patrones estructurales reales. Por esta razón, no es metodológicamente adecuado establecer conclusiones comparativas entre departamentos a partir de esta muestra.

En este sentido, se sugiere que futuros estudios orientados al análisis territorial del desempeño académico incorporen información proveniente de un mayor número de instituciones distribuidas en diferentes regiones del país, con el fin de garantizar una representación más equilibrada y permitir inferencias más robustas.

6.3.3 Análisis del tipo de origen de institución universitaria

Siguiendo con el orden de variables seleccionadas, se analiza la variable origen de la institución, el contexto institucional puede influir en las oportunidades de aprendizaje, recursos disponibles y exigencias académicas. En este sentido, se busca evaluar si existen patrones diferenciados en la distribución de los estudiantes entre los clústeres según el origen de su institución.

Una entidad oficial es aquella creada por el Estado, como la Universidad Nacional de Colombia, que es pública, autónoma y se financia con recursos estatales. Una corporación no oficial es una asociación de personas privadas sin ánimo de lucro que se unen democráticamente para un fin común, como un gremio de egresados. Una fundación no oficial es un patrimonio destinado por sus fundadores a un fin social, como la Pontificia Universidad Javeriana, sin que existan asociados que se beneficien directamente. En resumen, la Nacional es oficial y pública, mientras que la Javeriana es una fundación privada sin ánimo de lucro.

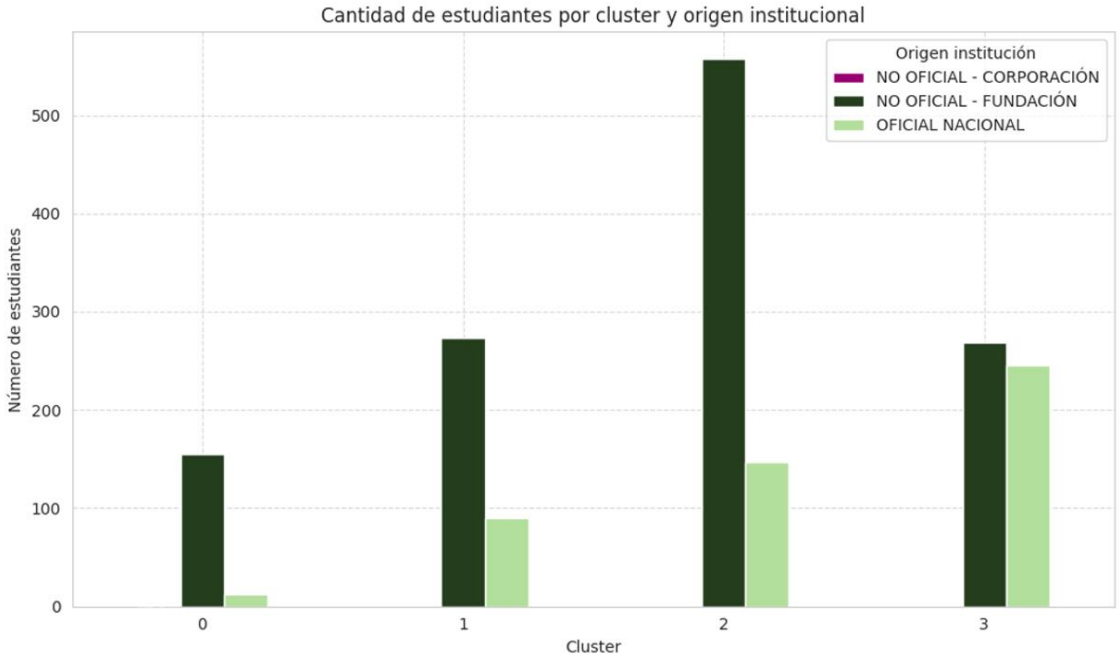


Figura 18. Estudiantes por clúster de universidad privada o pública.

Los resultados revelan una tendencia consistente a lo largo de todos los clústeres: las instituciones de origen No Oficial - Fundación, como la Pontificia Universidad Javeriana, concentran un mayor número de estudiantes en comparación con las instituciones Oficiales Nacionales, como la Universidad Nacional de Colombia. Esta diferencia es especialmente pronunciada en los clústeres intermedios y altos (Clústeres 1, 2 y 3).

Es importante señalar dos aspectos sobre estos resultados. Por un lado, el Clúster 0 concentra muy pocos estudiantes de la Universidad Nacional. El hecho de que la Javeriana también tenga representación en el Clúster Superior podría interpretarse superficialmente como una ventaja comparativa frente a la Nacional, pero esta lectura no es metodológicamente válida. Los tamaños de muestra son considerablemente distintos entre instituciones, existe una sobrerrepresentación de ciertas sedes y no se ha controlado por variables como la selectividad de ingreso, la composición de los programas académicos o las características sociodemográficas de los estudiantes de cada institución. Por tanto, no es posible establecer conclusiones comparativas robustas entre universidades a partir de estos datos, y cualquier afirmación en ese sentido excedería el alcance del

presente estudio.

6.3.4 Análisis del nivel educativo del padre y la madre

Con el fin de complementar el análisis, se examina la variable nivel educativo del padre y la madre. Esta variable permite aproximarse al capital educativo del hogar, el cual puede influir en el acompañamiento académico, las expectativas educativas y las oportunidades de aprendizaje del estudiante. Se busca identificar si existen diferencias en la distribución de los clústeres según el nivel de formación del padre o la madre.

6.3.5 Análisis complementario del perfil socioeconómico familiar por cluster

Como componente final del presente trabajo, se hizo un análisis socioeconómico familiar, usando por una parte las variables del estudio del padre y la madre, para hacer una investigación y con el fin de analizar de manera más precisa el nivel educativo de los padres, las variables categóricas originales fueron transformadas a una escala ordinal que respeta la progresión natural de los niveles educativos, desde “primaria” hasta “posgrado”. Esta transformación permitió tratar la educación como una variable cuantitativa ordenada, facilitando comparaciones entre individuos y grupos.

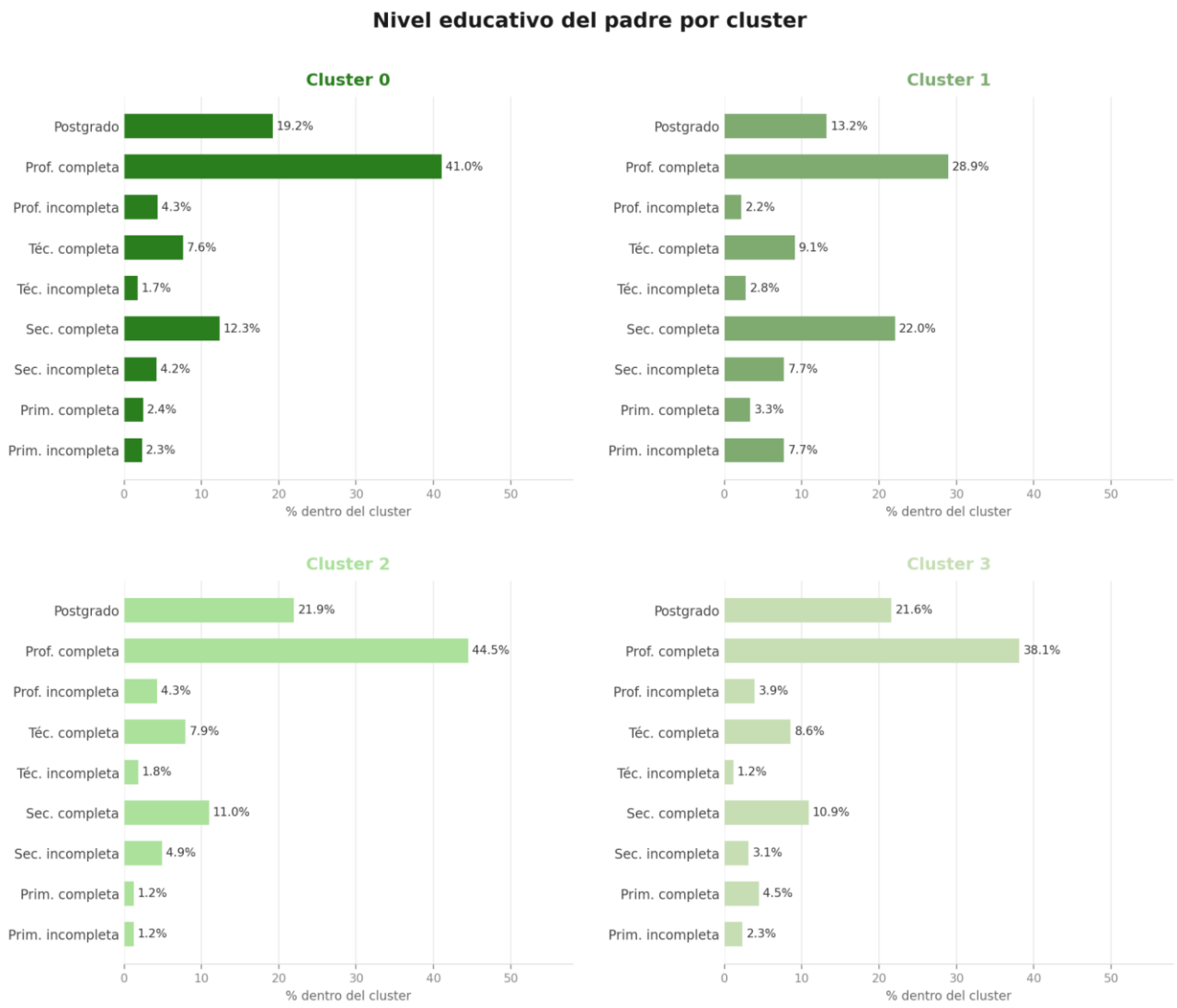


Figura 19. Nivel educativo del padre.

La distribución del nivel educativo del padre por clúster muestra un patrón interesante: aunque el clúster 0 identificado previamente como el de menor desempeño presenta una proporción importante de padres con educación profesional completa (alrededor del 41%), este no alcanza los niveles más altos observados en otros clústeres como el 2 o el 3, donde además se combina con mayores porcentajes de posgrado. Este resultado apunta que, si bien el nivel educativo de los padres es un factor relevante, no es suficiente por sí solo para explicar el rendimiento académico de los estudiantes en esta muestra. Es posible que en el clúster 0 afecten otros elementos que contrarrestan el efecto positivo esperado de la educación parental, como diferencias en la calidad de las

instituciones educativas, el contexto socioeconómico, el acceso a recursos académicos, o incluso factores individuales del estudiante.

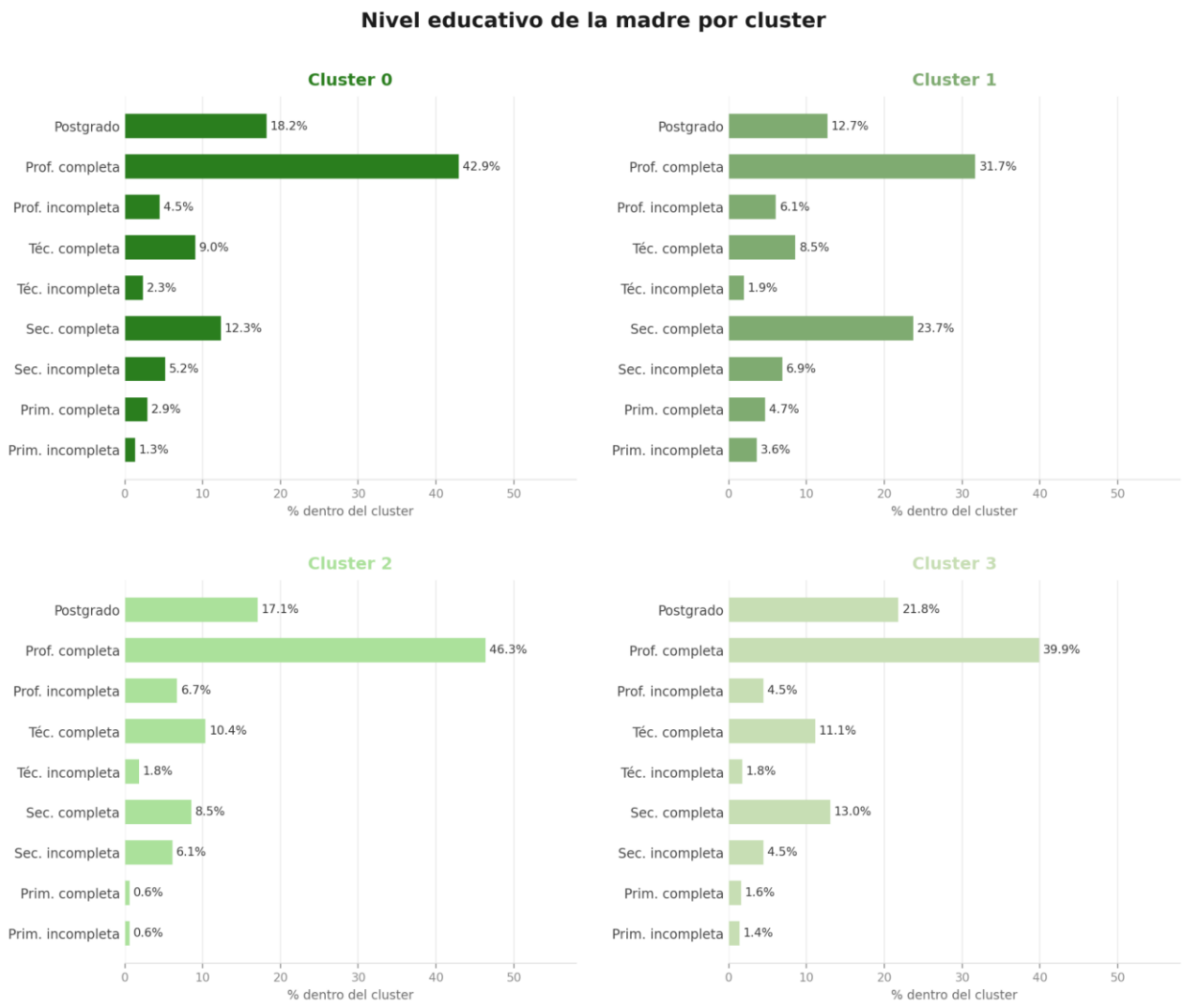


Figura 20. Nivel educativo de la madre.

Al analizar la distribución del nivel educativo de la madre, se observa un patrón similar al encontrado en el caso del padre, pero con una diferencia clave: la brecha entre el clúster de menor desempeño (clúster 0) y los de mayor desempeño es menos pronunciada. En todos los clústeres predomina la educación profesional completa, con valores relativamente altos y bastante cercanos entre sí (alrededor del 40%–46%), así como una presencia consistente de posgrados. Esto podría decir que, dentro de esta muestra existe una relativa homogeneidad en el nivel educativo materno. En otras palabras, incluso los estudiantes con menor desempeño provienen en muchos casos de

hogares donde la madre tiene un nivel educativo alto, lo que reduce la capacidad de esta variable para discriminar entre clústeres.

Sin embargo, es importante recordar que aun en el clúster inicial o con menor desempeño, se tienen unos resultados académicos más altos que el promedio a nivel nacional, por lo que si se hiciera una comparación más amplia con todo el país es posible que estudiantes como los de esta muestra si demostraran que la educación de los padres marca una gran diferencia frente a los padres que no han tenido las mismas oportunidades, ya que todos estos estudiantes tienen puntajes muy por encima del nacional.

También se incorporó un análisis del perfil socioeconómico familiar de los estudiantes en cada clúster como suplemento del análisis, con el propósito de identificar posibles asociaciones entre el contexto del hogar y el desempeño académico. Para ello se consideraron tres variables del Saber 11: el estrato de vivienda, la tenencia de computador y la tenencia de automóvil, las cuales funcionan como aproximaciones indirectas a las condiciones materiales del hogar.

La figura presenta, para cada clúster, la distribución porcentual de tres variables socioeconómicas: el estrato de vivienda, desagregado en bajo (estratos 1 y 2), medio (estratos 3 y 4) y alto (estratos 5 y 6); la tenencia de computador en el hogar; y la tenencia de automóvil. Cada porcentaje debe leerse de forma independiente dentro de su clúster, es decir, representa la proporción de estudiantes de ese grupo que pertenece a determinada categoría con respecto al total de estudiantes válidos del mismo clúster. Así, por ejemplo, un valor del 20% en estrato bajo para un clúster indica que el 20% de los estudiantes de ese grupo reportó vivir en un hogar de estrato 1 o 2, sin que esto guarde relación directa con los porcentajes de los demás clústeres.

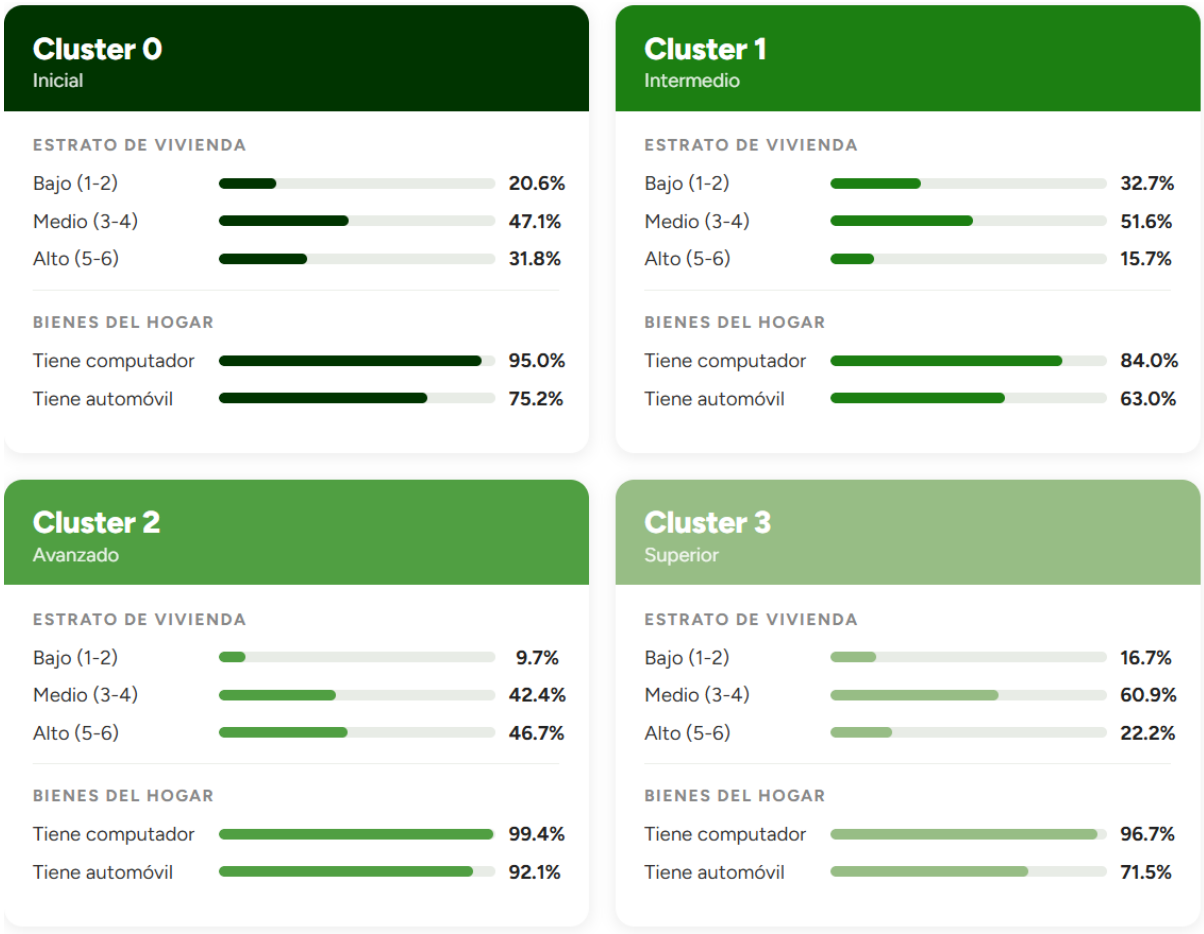


Figura 21. Perfil socioeconómico familiar por clúster.

Los resultados muestran que el clúster 2 (avanzado) concentra la mayor proporción de estudiantes provenientes de estratos altos (5-6) con un 46.7%, así como los porcentajes más elevados de hogares con computador (99.4%) y automóvil (92.1%). En contraste, el clúster 1 (intermedio) presenta el perfil más vulnerable, con la mayor proporción de estrato bajo (32.7%) y los valores más bajos en tenencia de bienes. Los clústeres 0 (inicial) y 3 (superior) muestran perfiles intermedios, aunque con composiciones distintas: el clúster 3 tiene mayor concentración en estrato medio (60.9%), mientras que el clúster 0 presenta una distribución más equilibrada entre los tres niveles de estrato.

No obstante, es importante señalar que estos hallazgos deben interpretarse con cautela. La muestra utilizada en este estudio proviene de una población específica de estudiantes universitarios,

lo que limita su representatividad frente a la diversidad socioeconómica del país. Colombia presenta altos niveles de desigualdad regional y social, por lo que para establecer conclusiones más robustas sobre la relación entre condiciones socioeconómicas y desempeño académico sería necesario contar con una muestra de mayor tamaño y cobertura geográfica que capture adecuadamente esa heterogeneidad.

6.4 Discusión

Los resultados obtenidos permiten responder las tres preguntas de investigación planteadas. Frente a la primera pregunta, sobre qué patrones de desempeño académico se pueden identificar en los estudiantes de Administración y áreas afines de las cuatro sedes universitarias analizadas, el estudio evidenció la existencia de cuatro perfiles diferenciados Inicial, Intermedio, Avanzado y Superior que se mantienen coherentes tanto en Saber 11 como en Saber Pro, confirmando que el desempeño académico de esta población sigue trayectorias relativamente estables a lo largo del tiempo. Frente a la segunda pregunta, sobre qué perfiles homogéneos emergen al agrupar a los estudiantes según sus puntajes, K-Means con cuatro clústeres produjo una segmentación interpretable y pedagógicamente útil, donde el perfil Avanzado concentra la mayor proporción de estudiantes (40.18%), lo que es coherente con el carácter selectivo de las instituciones analizadas. Frente a la tercera pregunta, sobre qué asociación presentan las variables sociodemográficas e institucionales con los perfiles identificados, la validación externa mediante Cramér's V mostró que variables como el programa académico, el tipo de institución, el estrato de vivienda y el acceso a tecnología en el hogar presentan asociaciones relevantes con los perfiles de desempeño, aunque estas asociaciones no deben interpretarse como relaciones causales sino como patrones descriptivos que caracterizan el contexto de origen de cada grupo.

Un aspecto que merece aclaración es la decisión de analizar conjuntamente a la Pontificia Universidad Javeriana y la Universidad Nacional de Colombia en lugar de realizar un análisis

separado por institución. Esta decisión responde a tres criterios. Primero, el objetivo del estudio no fue comparar instituciones sino identificar perfiles de trayectoria académica dentro del área de Administración y áreas afines; los perfiles obtenidos son una propiedad del estudio educativo analizado y no una característica institucional específica. Segundo, separar el clustering por sede habría implicado trabajar con muestras de tamaño muy desigual, lo cual habría comprometido la estabilidad estadística de los algoritmos de agrupamiento. Tercero, los análisis post-clustering presentados en la sección 6.3 muestran cómo se distribuyen las distintas variables institucionales y sociodemográficas entre los perfiles identificados, incluyendo la institución de origen, lo que ofrece al lector transparencia sobre la composición de cada grupo sin pretender establecer comparaciones causales entre universidades.

Los hallazgos son consistentes con lo reportado en la literatura sobre trayectorias académicas en educación superior. La persistencia del desempeño observada en la matriz de transición, donde los estudiantes de los extremos tienden a mantenerse en sus niveles de ingreso al egreso, coincide con patrones similares de estabilidad en el rendimiento documentados para estudiantes de la Universidad Nacional asociados a condiciones de ingreso y contexto socioeconómico [1]. Al mismo tiempo, la mayor volatilidad observada en los niveles intermedios, donde un porcentaje relevante de estudiantes logra escalar al clúster Superior al egreso, es coherente con la noción de valor agregado institucional ampliamente estudiada en Colombia [29]: la universidad no solo recibe estudiantes con un perfil determinado, sino que puede potenciar trayectorias que de otro modo habrían permanecido estancadas, especialmente en segmentos con condiciones de ingreso intermedias [32]. Este diálogo con la literatura es relevante porque los estudios previos han mostrado que entre tres y cuatro de cada cinco estudiantes universitarios en Colombia obtienen mejores puntajes de lo esperado en Saber Pro considerando su desempeño previo en Saber 11 [31], lo que refuerza la idea de que las trayectorias académicas no están totalmente determinadas por el

punto de partida.

La asociación encontrada entre las variables socioeconómicas y los perfiles de desempeño, aunque no fue el eje central del estudio, es consistente con la evidencia sobre desigualdad de origen en educación superior. Como señalan Chaves-Manzano y Ordoñez-López en uno de los textos guía de la investigación, las pruebas estandarizadas tienden a reflejar desigualdades estructurales preexistentes más que a medirlas de forma neutra [2]. En este sentido, el uso de analítica educativa como la empleada en este trabajo tiene un valor descriptivo importante, pero también conlleva una responsabilidad que ha sido destacada en la literatura especializada, que advierten que los sistemas de analítica del aprendizaje pueden reproducir o amplificar desigualdades existentes si los datos no se manejan de forma correcta o los perfiles identificados se utilizan para clasificar definitivamente a los estudiantes en lugar de orientar decisiones [25]. En la misma línea se tienen otros estudios que demostraron que las brechas de desempeño asociadas a características sociodemográficas persisten incluso después de controlar el comportamiento académico, que el contexto de origen tiene un peso estructural que las instituciones deben reconocer explícitamente en sus políticas de ingreso [26].

En cuanto a las implicaciones institucionales, los perfiles identificados pueden servir como punto de partida para que las cuatro sedes universitarias analizadas presten atención sobre las condiciones de ingreso de sus estudiantes y diseñen estrategias de acompañamiento diferenciadas desde los primeros semestres, reconociendo que el contexto socioeconómico condiciona, pero no determina la trayectoria académica. No obstante, es necesario señalar con claridad las limitaciones de generalización de estos hallazgos. La muestra analizada corresponde a estudiantes de programas de Administración y áreas afines de cuatro sedes universitarias con perfiles institucionales y estudiantiles particulares, todos ellos con niveles de desempeño consistentemente superiores al promedio nacional. Por tanto, los patrones identificados no son extrapolables al conjunto del sistema educativo colombiano ni a otras áreas de formación, y cualquier aplicación de estos

resultados en contextos distintos requeriría una validación específica. Una comparación más robusta entre instituciones requeriría además controlar por factores como el tamaño de la muestra, pérdida de registros por datos faltantes en los módulos evaluados, la selectividad de ingreso y la conformación de los programas, lo que constituye una línea natural de trabajo futuro que permitiría avanzar desde la descripción de patrones hacia inferencias más sólidas sobre los factores que influyen en las trayectorias educativas en la educación superior colombiana.

6.5 Conclusiones

El presente trabajo logró identificar, mediante técnicas de clustering no supervisado aplicadas sobre datos de las pruebas Saber 11 del periodo 2014-2 a 2017 y Saber Pro del período 2018-2022, cuatro perfiles diferenciados de desempeño académico en estudiantes de Administración y áreas afines de la Pontificia Universidad Javeriana y la Universidad Nacional de Colombia. La metodología CRISP-DM permitió estructurar de manera rigurosa todo el proceso, desde la integración de más de siete millones de registros hasta la obtención de una base depurada de aproximadamente 1.700 estudiantes del área de interés, vinculados a través del archivo de llaves del ICFES. Su principal utilidad radica en ofrecer un camino metodológico replicable para el análisis longitudinal del desempeño académico en educación superior colombiana, que pueda ser retomado, ampliado o mejorado por investigaciones futuras que busquen profundizar en la comprensión de las trayectorias estudiantiles más allá de las mediciones tradicionales.

El algoritmo K-Means con cuatro clústeres resultó ser el más adecuado aun sin ser el mejor en todas métricas internas, que fueron modestas en todos los modelos evaluados, reflejando la naturaleza continua y el solapamiento esperado de unos datos reales de desempeño académico, se eligió principalmente por su interpretabilidad y utilidad pedagógica respaldada en literatura. Los cuatro perfiles obtenidos, denominados Inicial, Intermedio, Avanzado y Superior, guardan una coherencia entre la estandarización de Saber 11 y Saber Pro, lo que valida la capacidad del modelo

para capturar trayectorias diferenciadas dentro de la muestra. También, se observó una diferencia entre ambas pruebas en todos los perfiles, patrón que no debe interpretarse como pérdida de aprendizaje ni como evidencia de menor esfuerzo, dado que Saber 11 y Saber Pro no son instrumentos equivalentes en escala, contenido ni momento del desarrollo académico. Este patrón constituye una línea de investigación futura relevante que requeriría un diseño metodológico específicamente orientado a su análisis.

La validación externa mediante el coeficiente de Cramér's V demostró que, aunque las variables socioeconómicas y contextuales no fueron utilizadas en la construcción de los clústeres, sí presentan una asociación con los perfiles de desempeño identificados. El programa académico, la institución de origen, el acceso a tecnología en el hogar, el estrato de vivienda y el tipo de universidad fueron las variables con mayor poder de asociación. Esto aporta evidencia sobre la relación entre las condiciones de origen y los perfiles de desempeño, subrayando que las diferencias observadas no son puramente académicas, sino que están asociadas al contexto del estudiante. El análisis por programa evidenció que Economía concentra proporcionalmente más estudiantes en el clúster Superior, mientras que Administración, siendo el programa más numeroso, distribuye la mayoría de sus estudiantes en el clúster Avanzado. El contraste entre instituciones privadas y públicas debe leerse con cautela dado el desbalance en el tamaño de las muestras y la ausencia de controles por selectividad de ingreso, por lo que no es posible establecer conclusiones comparativas robustas entre universidades a partir de estos datos.

En cuanto a las implicaciones institucionales derivadas de este trabajo, los perfiles identificados pueden servir para diseñar estrategias de acompañamiento académico diferenciadas a lo largo de la trayectoria universitaria, reconociendo que el contexto socioeconómico y las condiciones de ingreso condicionan, pero no determinan, los resultados de egreso. A modo de ejemplo concreto, los estudiantes provenientes de colegios no bilingües, cuyos puntajes en el módulo de inglés en

Saber 11 tienden a ubicarse por debajo del resto de la muestra, podrían beneficiarse de ajustes específicos, como orientar su plan de estudios a reforzar esta competencia desde los primeros semestres, dado que el inglés es una habilidad transversal con peso relevante también en Saber Pro. Esta línea de aplicación, además, abre la puerta para el desarrollo futuro de otros modelos para las universidades como estimar el riesgo de deserción temprana, donde los perfiles de ingreso aquí caracterizados podrían integrarse como variables predictoras. En general, los resultados respaldan el interés de incorporar la información de ingreso como una variable contextual dentro de los sistemas de información institucional, entendida siempre como un orientador y no como un criterio de clasificación definitiva de los estudiantes.

Finalmente, se recomienda que estudios futuros amplíen la cobertura geográfica e institucional, incorporen más programas académicos, exploren modelos con mayor capacidad inferencial, profundicen en el análisis del patrón de diferencia entre pruebas y examinen las causas detrás de la diferencia de datos faltantes observada en la Universidad Nacional de Colombia. El camino metodológico aquí trazado aspira a servir de base para que otros trabajos pasen de una simple medición estandarizada hacia una comprensión verdaderamente multidimensional de la educación colombiana.

7. Anexos

Anexo 1. Ficha técnica del entorno de análisis

Componente	Detalle
General	
Lenguaje	Python 3
Entorno de ejecución	Google Colaboratory
Librerías	
pandas	2.2.2
scikit-learn	1.6.1
numpy	2.0.2
matplotlib	3.10.0
seaborn	0.13.2
skfuzzy	0.5.0
Recursos	
Notebook compartido	https://github.com/diegocaballero10/TDG

Anexo 2. Especificaciones técnicas de los algoritmos

Modelo	Librería	Clase	Hiperparámetros configurados
K-Means	scikit-learn 1.6.1	sklearn.cluster.KMeans	n_clusters=4, random_state=100
Gaussian Mixture	scikit-learn 1.6.1	sklearn.mixture.GaussianMixture	n_components=4, random_state=100
Fuzzy C-Means	skfuzzy 0.5.0	skfuzzy.cluster.cmeans	c=4, m=2, error=0.005, maxiter=1000
Clustering Aglomerativo	scikit-learn 1.6.1	sklearn.cluster.AgglomerativeClustering	n_clusters=4, linkage='ward'
DBSCAN	scikit-learn 1.6.1	sklearn.cluster.DBSCAN	eps=1.5, min_samples=10
Spectral Clustering	scikit-learn 1.6.1	sklearn.cluster.SpectralClustering	n_clusters=4, assign_labels='kmeans', random state=100

Anexo 3. Configuración de estandarización

Componente	Librería	Clase	Configuración
Estandarización	scikit-learn 1.6.1	sklearn.preprocessing.StandardScaler	Parámetros por defecto

Anexo 4. Métricas de validación interna

Métrica	Librería	Función
Coeficiente de Silhouette	scikit-learn 1.6.1	sklearn.metrics.silhouette_score
Índice Davies-Bouldin	scikit-learn 1.6.1	sklearn.metrics.davies_bouldin_score
Índice Calinski-Harabasz	scikit-learn 1.6.1	sklearn.metrics.calinski_harabasz_score

Anexo 5. Métrica de asociación

Métrica	Librería	Función
Coeficiente de Cramér's V	scipy 1.16.3	scipy.stats.chi2_contingency

Anexo 6. Diccionario Saber Pro

#	Variable	Descripción
1	periodo	Período académico de presentación
2	estu_consecutivo	Código consecutivo del estudiante
3	estu_tipodocumento	Tipo de documento de identidad
4	estu_pais_reside	País de residencia del estudiante
5	estu_cod_reside_depto	Código del departamento de residencia
6	estu_depto_reside	Departamento de residencia
7	estu_cod_reside_mcpio	Código del municipio de residencia
8	estu_mcpio_reside	Municipio de residencia
9	estu_coddane_cole_termino	Código DANE del colegio donde terminó bachillerato
10	estu_cod_cole_mcpio_termino	Código del municipio del colegio donde terminó bachillerato
11	estu_cod_depto_presentacion	Código del departamento de presentación del examen
12	inst_cod_institucion	Código de la institución de educación superior
13	inst_nombre_institucion	Nombre de la institución
14	inst_caracter_academico	Carácter académico de la institución
15	estu_nucleo_pregrado	Núcleo básico de conocimiento del programa
16	estu_inst_departamento	Departamento de la institución
17	estu_inst_codmunicipio	Código del municipio de la institución
18	estu_inst_municipio	Municipio de la institución
19	estu_prgm_academico	Nombre del programa académico
20	estu_prgm_departamento	Departamento del programa académico
21	estu_prgm_codmunicipio	Código del municipio del programa
22	estu_prgm_municipio	Municipio del programa académico
23	estu_nivel_prgm_academico	Nivel del programa (pregrado, posgrado, etc.)
24	estu_metodo_prgm	Metodología del programa (presencial, distancia, etc.)
25	estu_valormatriculauniversidad	Valor de la matrícula universitaria
26	estu_depto_presentacion	Departamento de presentación del examen
27	estu_cod_mcpio_presentacion	Código del municipio de presentación
28	estu_mcpio_presentacion	Municipio de presentación del examen
29	estu_pagomatriculabeca	Pago de matrícula con beca
30	estu_pagomatriculacredito	Pago de matrícula con crédito
31	estu_horassemanatrabaja	Horas semanales que trabaja el estudiante

#	Variable	Descripción
32	estu_snies_prgmacademico	Código SNIES del programa académico
33	estu_privado_libertad	Indicador de privación de libertad
34	estu_nacionalidad	Nacionalidad del estudiante
35	estu_estudiante	Tipo de estudiante
36	estu_genero	Género del estudiante
37	estu_cole_termino	Nombre del colegio donde terminó bachillerato
38	estu_pagomatriculapadres	Pago de matrícula por parte de los padres
39	estu_estadoinvestigacion	Estado de investigación del registro
40	estu_fechanacimiento	Fecha de nacimiento del estudiante
41	estu_pagomatriculapropio	Pago de matrícula con recursos propios
42	estu_tipodocumentosb11	Tipo de documento registrado en Saber 11
43	fami_educacionpadre	Nivel educativo del padre
44	fami_tieneautomovil	¿La familia tiene automóvil?
45	fami_tienelavadora	¿La familia tiene lavadora?
46	fami_estratovivienda	Estrato socioeconómico de la vivienda
47	fami_tienecomputador	¿La familia tiene computador?
48	fami_tieneinternet	¿La familia tiene internet?
49	fami_educacionmadre	Nivel educativo de la madre
50	inst_origen	Origen / tipo de institución (pública o privada)
51	mod_razona_cuantitat_punt	Puntaje módulo Razonamiento Cuantitativo
52	mod_comuni_escrita_punt	Puntaje módulo Comunicación Escrita
53	mod_comuni_escrita_desem	Desempeño módulo Comunicación Escrita
54	mod_ingles_desem	Desempeño módulo Inglés
55	mod_lectura_critica_punt	Puntaje módulo Lectura Crítica
56	mod_ingles_punt	Puntaje módulo Inglés
57	mod_competen_ciudada_punt	Puntaje módulo Competencias Ciudadanas
58	result_nombreprueba	Nombre de la prueba
59	result_puntaje	Puntaje global del resultado

Anexo 7. Diccionario Saber 11

#	Variable	Descripción
1	PERIODO	Período académico de presentación del examen
2	ESTU_TIPODOCUMENTO	Tipo de documento de identidad del estudiante
3	ESTU_CONSECUTIVO	Código consecutivo del estudiante
4	COLE_AREA_UBICACION	Área de ubicación del colegio (urbana o rural)
5	COLE_BILINGUE	Indica si el colegio es bilingüe
6	COLE_CALENDARIO	Calendario escolar del colegio (A o B)
7	COLE_CARACTER	Carácter académico del colegio
8	COLE_COD_DANE_ESTABLECIMIENTO	Código DANE del establecimiento educativo
9	COLE_COD_DANE_SEDE	Código DANE de la sede del colegio
10	COLE_COD_DEPTO_UBICACION	Código del departamento de ubicación del colegio
11	COLE_COD_MCPIO_UBICACION	Código del municipio de ubicación del colegio
12	COLE_CODIGO_ICFES	Código ICFES del colegio
13	COLE_DEPTO_UBICACION	Departamento de ubicación del colegio
14	COLE_GENERO	Género al que atiende el colegio (mixto, masculino, femenino)
15	COLE_JORNADA	Jornada escolar del colegio
16	COLE_MCPIO_UBICACION	Municipio de ubicación del colegio
17	COLE_NATURALEZA	Naturaleza del colegio (oficial o no oficial)
18	COLE_NOMBRE_ESTABLECIMIENTO	Nombre del establecimiento educativo
19	COLE_NOMBRE_SEDE	Nombre de la sede del colegio
20	COLE_SEDE_PRINCIPAL	Indica si es la sede principal del colegio
21	ESTU_COD_DEPTO_PRESENTACION	Código del departamento donde presentó el examen
22	ESTU_COD_MCPIO_PRESENTACION	Código del municipio donde presentó el examen
23	ESTU_COD_RESIDE_DEPTO	Código del departamento de residencia del estudiante
24	ESTU_COD_RESIDE_MCPIO	Código del municipio de residencia del estudiante
25	ESTU_DEPTO_PRESENTACION	Departamento donde presentó el examen
26	ESTU_DEPTO_RESIDE	Departamento de residencia del estudiante
27	ESTU_ESTADOINVESTIGACION	Estado de investigación del registro
28	ESTU_ESTUDIANTE	Tipo de estudiante (estudiante regular, etc.)
29	ESTU_FECHANACIMIENTO	Fecha de nacimiento del estudiante
30	ESTU_GENERO	Género del estudiante
31	ESTU_MCPIO_PRESENTACION	Municipio donde presentó el examen

#	Variable	Descripción
32	ESTU_MCPIO_RESIDE	Municipio de residencia del estudiante
33	ESTU_NACIONALIDAD	Nacionalidad del estudiante
34	ESTU_PAIS_RESIDE	País de residencia del estudiante
35	ESTU_PRIVADO_LIBERTAD	Indica si el estudiante está privado de la libertad
36	FAMI_CUARTOSHOGAR	Número de cuartos en el hogar del estudiante
37	FAMI_EDUCACIONMADRE	Nivel educativo de la madre
38	FAMI_EDUCACIONPADRE	Nivel educativo del padre
39	FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de la vivienda familiar
40	FAMI_PERSONASHOGAR	Número de personas en el hogar
41	FAMI_TIENEAUTOMOVIL	¿La familia tiene automóvil?
42	FAMI_TIENECOMPUTADOR	¿La familia tiene computador?
43	FAMI_TIENEINTERNET	¿La familia tiene acceso a internet?
44	FAMI_TIENELAVADORA	¿La familia tiene lavadora?
45	DESEMP_INGLES	Nivel de desempeño en la prueba de Inglés
46	PUNT_INGLES	Puntaje obtenido en la prueba de Inglés
47	PUNT_MATEMATICAS	Puntaje obtenido en la prueba de Matemáticas
48	PUNT_SOCIALES_CIUDADANAS	Puntaje obtenido en Sociales y Ciudadanas
49	PUNT_C_NATURALES	Puntaje obtenido en Ciencias Naturales
50	PUNT_LECTURA_CRITICA	Puntaje obtenido en Lectura Crítica
51	PUNT_GLOBAL	Puntaje global del examen Saber 11

Referencias

[1] Gallego Vega, L. E. (2021). Gallego Vega, L. E. (2021). *Saltos y brechas en la UNAL*. Editorial Universidad Nacional de Colombia. Prefacio | SPRO1.knit

[2] Chaves-Manzano, H. R., & Ordoñez-López, I. D. P. (2020). *Cavilaciones sobre la evaluación y la calidad educativa en Colombia*. Revista Arbitrada Interdisciplinaria Koinonía. Cavilaciones sobre la evaluación y la calidad educativa en Colombia

[3] OECD. (2020). *PISA 2018 results (Volume V): Effective policies, successful schools*. OECD Publishing. https://doi.org/10.1787/ca768d40-en

[4] Demarchi Sánchez, G. D. (2023). *Factores que intervienen en el desempeño de los estudiantes en las Pruebas de Estado Saber Pro*. El Ágora U.S.B., 23(2), 490–505. https://doi.org/10.21500/16578031.6224

[5] Congreso de Colombia. (1992). *Ley 30 de 1992: Por la cual se organiza el servicio público de la educación superior*. Diario Oficial No. 40.700. Ley 30 de 1992 - Gestor Normativo - Función Pública

[6] Consejo Nacional de Acreditación. (2020). *Acuerdo 02 de 2020: Por el cual se actualiza el modelo de acreditación en alta calidad*. CNA. articles-402848_documento.pdf

[7] Universidad Nacional de Colombia. (2021). *Misión y visión – Sede Palmira*. Misión y visión

[8] Pontificia Universidad Javeriana. (2022). *Sección institucional*. javeriana.edu.co/institucional

[9] Pontificia Universidad Javeriana Cali. (2024). *Misión, visión y proyecto educativo institucional*. https://www.javerianacali.edu.co/

[10] Universidad Nacional de Colombia. (2024). *Presentación institucional y oferta académica – Sede Palmira*. https://palmira.unal.edu.co/

[11] Torres Chaves, L. P., Álvarez García, M. A., Cantillo Guerrero, J., Usme Ramos, M. C., Mena Rodríguez, E., & Labrador Mancilla, G. A. (2024). *La evaluación en Colombia, ruta para la calidad educativa: desafíos y retos*. *Ciencia Latina Revista Científica Multidisciplinar*. https://ciencialatina.org/index.php/cienciala/article/view/15562

[12] Instituto Colombiano para la Evaluación de la Educación (ICFES). (2025, abril). *Guía de orientación del Examen Saber 11° – Aplicación 2025-2*. 30-abril-guia-orientacion-saber-11o-2025-2.pdf

[13] Instituto Colombiano para la Evaluación de la Educación (ICFES). (2024, diciembre). *Guía de orientación del Examen Saber Pro – Módulos de competencias genéricas, aplicación 2025-1*. 02-Diciembre_Guía-de-Orientación-Módulos-Genéricos-Saber-Pro-2025-1.pdf

[14] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer. https://www.statlearning.com/

[15] Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Prentice Hall.

[16] Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). *Clustering analysis for classifying student academic performance in higher education*. *Applied Sciences, 12*(19), 9467. https://doi.org/10.3390/app12199467

[17] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

[18] Von Luxburg, U. (2007). *A tutorial on spectral clustering*. *Statistics and Computing, 17*(4), 395–416. A tutorial on spectral clustering | Statistics and Computing | Springer Nature Link

[19] Agresti, A. (2018). *An introduction to categorical data analysis* (3rd ed.). John Wiley & Sons. https://www.wiley.com/en-ie/An+Introduction+to+Categorical+Data+Analysis%2C+3rd+Edition-p-9781119405269

[20] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS. [PDF] CRISP-DM 1.0: Step-by-step data mining guide | Semantic Scholar

[21] Martínez Rizo, F. (2010). *Las pruebas de rendimiento y sus consecuencias*. Ministerio de Educación Nacional de Colombia. Las pruebas de rendimiento y sus consecuencias - ..::Ministerio de Educación Nacional de Colombia::..

[22] Colombia. (2012). *Ley 1581 de 2012: Por la cual se dictan disposiciones generales para la protección de datos personales*. Diario Oficial No. 48.587. Ley 1581 de 2012 - Gestor Normativo - Función Pública

[23] Instituto Colombiano para la Evaluación de la Educación (ICFES). (2025, marzo 26). *Transparencia y acceso a la información pública*. https://www.icfes.gov.co/transparencia-y-acceso-a-la-informacion-publica/

[24] Instituto Colombiano para la Evaluación de la Educación (ICFES). (2026). *Informe nacional de resultados Saber 11° 2024*. https://www.icfes.gov.co/informes-de-resultados

[25] Dollinger, M., Liu, D., Arthars, N., & Lodge, J. (2019). *Working together in learning analytics towards the co-creation of value*. *Journal of Learning Analytics, 6*(2), 10–26.

https://doi.org/10.18608/jla.2019.62.2

[26] Rienties, B., Cross, S., & Zdrahal, Z. (2020). *Implementing a learning analytics intervention and evaluation framework: What works? Assessment & Evaluation in Higher Education, 45*(6), 842–856. Implementing a Learning Analytics Intervention and Evaluation Framework: What Works? | Springer Nature Link

[27] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. https://link.springer.com/book/9780387310732

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[29] Balcázar, C. F., & Ñopo, H. (2016). *Broken gears: The value added of higher education on teachers' academic achievement*. *Higher Education, 72*(3), 341–361. https://doi.org/10.1007/s10734-015-9960-0

[30] Gómez Soler, S. C., Bernal Nisperuza, G. L., & Herrera Idárraga, P. (2020). *Test preparation and students' performance: The case of the Colombian high school exit exam*. *Cuadernos de Economía, 39*(79), 31–72. https://revistas.unal.edu.co/index.php/ceconomia/article/download/77106/73508

[31] Burgos Flórez, M. A., Ruales Suárez, K. J., Bastidas García, Y. E., & Ortiz Benavides, E. (2019). *Cálculo del valor agregado generado por la Universidad de Nariño en relación a las pruebas Saber 11 - Saber Pro 2010-2014*. *Tendencias, 20*(2), 203–226.

https://doi.org/10.22267/rtend.192002.125

[32] Universidad del Norte, Instituto de Estudios en Educación. (2020). *¿Qué es el "valor agregado" en la educación universitaria? Intellecta: Periodismo Científico de la Universidad del Norte*. https://www.uninorte.edu.co/web/intellecta/que-es-el-valor-agregado-en-la-educacion-universitaria