ESTUDIO DEL DESEMPEÑO ACADÉMICO DE ESTUDIANTES COLOMBIANOS EN LAS PRUEBAS SABER 11 Y SABER PRO PARA LA ELECCIÓN VOCACIONAL Y PERMANENCIA UNIVERSITARIA

LINA NATALIA MORENO QUINTERO

UNIVERSIDAD PONTIFICIA BOLIVARIANA
ESCUELA INGENIERÍAS
FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN
MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN
MEDELLÍN
2019

ESTUDIO DEL DESEMPEÑO ACADÉMICO DE ESTUDIANTES COLOMBIANOS EN LAS PRUEBAS SABER 11 Y SABER PRO PARA LA ELECCIÓN VOCACIONAL Y PERMANENCIA UNIVERSITARIA

LINA NATALIA MORENO QUINTERO

Trabajo de grado para optar al título de Magíster en Tecnologías de la Información y la Comunicación

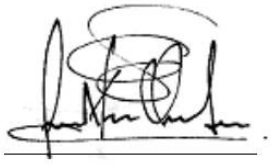
Asesor
Ana Isabel Oviedo Carrascal, PhD
Doctora en Ingeniería Electrónica

UNIVERSIDAD PONTIFICIA BOLIVARIANA
ESCUELA INGENIERÍAS
FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN
MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN
MEDELLÍN
2019

DECLARACIÓN ORIGINALIDAD

“Declaro que esta tesis (o trabajo de grado) no ha sido presentada para optar a un título, ya sea en igual forma o con variaciones, en esta o cualquier otra universidad”. Art. 82 Régimen Discente de Formación Avanzada, Universidad Pontificia Bolivariana.

FIRMA AUTOR



Lina Natalia Moreno Quintero

Ciudad y fecha

Medellín, 30 de noviembre de 2019.

Reconociendo a todos los que me han apoyado en la construcción del proyecto publicado como www.TuCarrera.co,
un portal que presta servicios de orientación vocacional para facilitar la elección consciente de una carrera profesional en los 32 departamentos del país.

Una ardilla siempre está en el bosque buscando algo para sí misma, ella es hábil y no le sirve cualquier cosa, quiere una nuez. La educación superior de Colombia es el bosque en el que no hay tantas nueces como jóvenes buscando.

Dedicado a mi mamá y a su carácter social como psicóloga clínica.
A mi papá quien sembró tanto amor en vida.
A Dante, Hanna, Vera y también a Río.

AGRADECIMIENTOS

Este interés investigativo surge de la necesidad de recorrer esfuerzos desde la academia para sustentar una prueba de orientación vocacional que acompañe la elección consciente e informada de una carrera profesional, procurando perfilar mejor a los jóvenes desde la educación media hacia la superior para combatir la deserción universitaria.

A Ana Isabel Oviedo Carrascal, PhD y a tantos docentes que dan a los estudiantes el conocimiento y tiempo que algunos solo dedicarían a sus hijos.

Al proyecto publicado como www.TuCarrera.co, un desarrollo que pertenece a Tnt Colombia SAS y cuyo objetivo primordial es mejorar la inclusión social en el sistema de educación superior de Colombia, conectando a los jóvenes de todos los departamentos del país con el 100% de la oferta en las carreras profesionales (desde técnicas hasta postdoctorados) avaladas por el Ministerio de Educación Nacional, además de generar reportes independientes que miden el estado real del Sistema de Educación Superior de Colombia.

Este proyecto está dedicado a:

Todas las objeciones que le caben al Sistema de Educación Superior de Colombia, cuya oferta en carreras profesionales proviene en un 62% del sector privado frente al 38% del público.

Los millones de jóvenes, en los 32 departamentos de Colombia que no reciben ningún tipo de orientación vocacional y se encuentran dentro de los indicadores que el país no se toma en serio.

Todos aquellos a quienes el Sistema de Educación Superior de Colombia forzó su desplazamiento desde las áreas rurales hacia alguna de las 5 capitales del país, en busca del sueño de profesionalizarse. El 70% de la oferta en carreras profesionales se concentra en Bogotá D.C., Medellín, Cali, Bucaramanga y Barranquilla, para el resto de los 28 de los 32 departamentos hay una asignación de tan solo el 30% del total de esta oferta.

Las personas en los departamentos de Quindío, Meta, Córdoba, Huila, Magdalena, Sucre, Cesar, Guajira, Chocó, Casanare, Caquetá, Putumayo, San Andrés y Providencia, Arauca, Amazonas, Guainía, Guaviare y Vichada,

para quienes, en cada uno de sus departamentos se concentra tan solo el 1% de la oferta total en carreras de educación superior en Colombia.

Aquellos que se enlistan en algún lado de la guerra sea en las Fuerzas Públicas de Colombia (Fuerza Militar o Policía Nacional) o en cualquier grupo insurgente, para quienes las actividades bélicas fueron la única escuela interesada en los jóvenes del campo.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ANEXOS

GLOSARIO

EDUCACIÓN MEDIA: alusiva en Colombia a los últimos dos años de formación del individuo para obtener el título de bachiller y posteriores a los cuatro años de educación básica secundaria, habiendo cumplido previamente con los primeros cinco años de formación correspondientes a la educación básica primaria.

EDUCACIÓN SUPERIOR: conformada por dos niveles principales (pregrado y posgrado), el primero de estos se habilita para quienes acrediten un título de bachiller en Colombia y a su vez se divide en los siguientes tres niveles; técnico profesional, tecnológico y profesional (relativo a los programas profesionales universitarios).

ICFES: Instituto Colombiano para la Evaluación de la Educación.

ORIENTACIÓN VOCACIONAL: acompañamiento profesional en el que se busca orientar a los individuos en la toma de decisiones relacionadas con la elección y el posterior ejercicio de una carrera o estudio.

PERMANENCIA UNIVERSITARIA: referente en este proyecto a los estudiantes de la educación superior de Colombia que una vez eligen una carrera profesional para cursar, permanecen en ellas al menos hasta completar un 75% de la carga académica de los programas o en promedio logran llegar mínimamente hasta el semestre 9 de sus estudios.

SABER 11: es la evaluación aplicada por el Ministerio de Educación Nacional (MEN) con el objetivo de medir las destrezas adquiridas por los estudiantes durante su formación básica y media.

SABER PRO: es el examen de estado para evaluar la calidad de la educación superior en Colombia.

RESUMEN

En Colombia, el Instituto Colombiano para la Evaluación de la Educación (ICFES), dependiente del Ministerio de Educación Nacional, analiza el desempeño de los estudiantes en varios momentos de la vida académica, este trabajo se enfoca específicamente cuando están finalizando la educación media y antes de graduarse de un programa profesional de pregrado de educación superior, los cuales están enmarcados en el contexto de las Pruebas Saber 11 y Saber Pro.

La presente investigación entrega un análisis sobre el desempeño de los estudiantes en las pruebas Saber 11 y Saber Pro y su relación con la elección vocacional, mediante técnicas de Ciencia de Datos. Con lo que se espera aportar un conocimiento para mejorar la comprensión de los factores clave que inciden en la elección y permanencia en el estudio de carreras profesionales en Colombia, buscando que el conocimiento generado sirva de soporte a las acciones encaminadas a mitigar el riesgo de deserción en la educación superior del país.

Los modelos empleados fueron los siguientes: 1) correlaciones para identificar las asociaciones entre las distintas variables de las Pruebas Saber 11 y Saber Pro, con el rendimiento académico; 2) Clustering con K-means para realizar agrupaciones que facilitan el perfilado de los estudiantes; 3) Reglas de asociación con “Apriori” para fortalecer las interpretaciones de las relaciones de asociación y consecuencias entre los datos.

A continuación, se presentan algunos aportes de esta investigación para la comprensión de los factores asociados a la elección y permanencia en una carrera profesional, destacando que los alumnos que se gradúan del bachillerato en Colombia se pueden clasificar dentro de los siguientes cuatro escenarios, en los que se analiza el déficit en las competencias que presentan

en la educación media versus las carreras universitarias que eligen y en las cuales suelen permanecer:

- Escenario 1. Poca competencia para las ciencias sociales y las matemáticas. Estos estudiantes suelen elegir y permanecer en carreras profesionales relacionadas con: educación, salud, medicina, enfermería, ciencias agropecuarias, bellas artes y diseño, normales superiores y comunicación, periodismo y publicidad.
- Escenario 2. Poca competencia para el lenguaje y las ciencias sociales. Estos estudiantes suelen elegir y permanecer en carreras profesionales relacionadas con: ingeniería y ciencias naturales y exactas.
- Escenario 3. Poca competencia para la biología y las matemáticas. Estos estudiantes suelen elegir y permanecer en carreras profesionales relacionadas con: derecho, psicología y humanidades.
- Escenario 4. Poca competencia para la biología y las ciencias sociales. Estos estudiantes suelen elegir y permanecer en carreras profesionales relacionadas con: administración, contaduría, economía y afines.

PALABRAS CLAVE: analíticas del aprendizaje, minería de datos, ciencia de datos, deserción, orientación vocacional, elección de carrera.

ABSTRACT

The Colombian Institute for the Education´s Evaluation (ICFES), under the Ministry of National Education, analyzes the Colombian student´s performance at various times in academic life, this work is focused on two moments in the student´s academical life, when they are finishing their middle school and about to get graduated from a professional degree program from higher education level, both moments are framed in the context of the Saber 11 and Saber Pro Tests.

This study analyzes the students' performance in the Saber 11 and Saber Pro tests and their relationship with vocational choice by using Data Science techniques. As a result of this research, it is expected to better understand the key factors associated to the choice and not drop out from the study of professional careers in Colombia, besides contributing with strategies to mitigate the risk of dropout in the country´s higher education system.

The models used were as it follows: 1) correlations to identify the associations between the data from “Saber 11” and “Saber Pro” Tests, with academic performance, 2) Clustering with K-means to identify different student´s profiles; and 3) association rules, known as “Apriori” to strengthen the interpretations of the relations and consequences between the data.

Below are some contributions of this research for the understanding of the factors associated with the election and permanence in a professional career, highlighting that graduated students from high school in Colombia can be classified within the following four scenarios, an analysis of the weaknesses that they present at high school against the university careers that they chose and have successfully completed up to 75%.

- Scenario 1. Weak skills at social sciences and mathematics. These students usually make a vocational choice related to education, health,

medicine, nursing, agricultural sciences, fine arts and design and communication, journalism and advertising.

- Scenario 2. Weak skills at language and social sciences. These students usually make a vocational choice related to engineering and natural and exact sciences.
- Scenario 3. Weak skills at biology and mathematics. These students usually make a vocational choice related to law, psychology and humanities.
- Scenario 4. Weak skills at biology and social sciences. These students usually make a vocational choice related to administration, accounting, economics and related.

KEY WORDS: learning analytics, data mining, data science, dropout, vocational guidance, career choice.

PARTE 1: FORMULACION DE PROYECTO

1. INTRODUCCIÓN

La elección de una carrera profesional es un reto que muchos hemos enfrentado de manera intuitiva en Colombia, contando con un sistema de educación superior que ofrece alrededor de 2,400 carreras profesionales de pregrado diferentes, en la que confluyen variables de tipo social, económico, institucional, académico, entre otras, de las cuales apenas un pequeño puñado de estas pasan por la mente de un joven al que se le pregunta ¿qué carrera vas a estudiar?.

En este contexto, se hace casi natural esperar que los pocos privilegiados que acceden al sistema de educación superior en Colombia, deserten masivamente de las carreras sobre las cuales algoritmos computacionales hubieran podido contribuir a una mirada más fundamentada para perfilar a los jóvenes desde la educación media hacia la superior.

Diversos estudios relacionados con las analíticas del aprendizaje en Colombia han puesto el foco sobre algún programa profesional en particular, región o grupo de variables específicas. En esta investigación se busca generar un aprendizaje que contribuya con una mirada holística en donde se aprenda sobre bases de entrenamiento que involucran datos para el 100% de las posibles carreras de pregrado sobre las cuales se puede elegir en el país.

Para lo anterior, se cuenta como punto de partida con un aproximado de 200 variables recolectadas en dos momentos de la vida académica de los colombianos: antes de graduarse de la educación media y ad-portas de culminar una carrera profesional. El proveedor de los datos es el Instituto Colombiano para la Evaluación de la Educación – ICFES.

La organización de esta investigación se divide en dos partes:

En la parte 1 se encuentra la formulación del proyecto, la cual inicia con el presente capítulo de introducción, seguido por el planteamiento del problema en donde se puede leer la justificación que motiva a este proyecto seguida de sus respectivos objetivos.

Continuando con la formulación del proyecto, se aprecia en el capítulo 4 el marco referencial, en el que se aterrizan los aspectos más relevantes a ser

tenidos en cuenta, seguido de los límites legales de este estudio y finalizando con el estado del arte en donde se exponen las principales investigaciones que sirvieron de punto de partida para la investigación.

La primera parte concluye con la metodología seleccionada para el desarrollo del problema de Ciencia de Datos planteado.

La parte 2, contiene la metodología que soporta al proyecto de ciencia de los datos, iniciando con el capítulo 6, en donde se gráfica un diseño de la solución, seguido por la comprensión del negocio y de los datos, para dar lugar a la preparación de los mismos y conducir al modelamiento y despliegue de los algoritmos seleccionados.

Al final se presentan y analizan los resultados obtenidos y se entregan las conclusiones y los aprendizajes para futuras investigaciones.

2. PLANTEAMIENTO DEL PROBLEMA

2.1 Problema

La tasa de deserción en la educación superior de Colombia es una de las más escandalosas de América Latina [1], una de las causas asociadas a este problema es la deficiente orientación profesional y vocacional que reciben los jóvenes en su educación media [2].

El Ministerio de Educación Nacional, a través del Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior –SPADIES-, reportó que la deserción por cohorte en Colombia, la cual se refiere a la deserción acumulada en cada semestre para un grupo de estudiantes que ingresaron a primer curso en un mismo periodo académico (cohorte) y se documenta para el nivel de formación universitario al décimo semestre, fue del 46.1% en el año 2015 [3].

No se ha demostrado que exista una articulación eficaz entre las políticas de la educación media y superior del Estado Colombiano, basta con comparar las bajas tasas de permanencia de los individuos en el Sistema de Educación Superior durante varios años, las cuales inician con una deserción por cohorte superior al 10% en el semestre 1 de formación universitaria y revelan una tendencia consistente al alza hasta ubicarse por encima del 40% de manera generalizada en el último semestre de la carrera [4].

La tasa de deserción universitaria por cohorte en Colombia desde 1998 hasta la actualidad, se ha consolidado con un patrón que no muestra mejoras significativas, es un fenómeno que más allá de las cifras estáticas que documenta el Ministerio de Educación Nacional como la tasa de inserción a la educación superior, deserción y graduados, amerita otro tipo de esfuerzos relacionados con las analíticas del aprendizaje, que permitan extraer conocimiento encaminado a comprender mejor la relación de los individuos con la educación media y superior del país.

A los jóvenes se les insta a escoger una carrera profesional en un rango de edad en el cual es complejo reconocer los factores claves del entorno social, económico, académico y cultural y a su vez asociarlos con la gran oferta en carreras profesionales avaladas por el Ministerio de Educación Nacional de Colombia, que en caso de ser integrados pueden contribuir a una elección más fundamentada en relación a qué carrera elegir y para quienes en el mejor de

los escenarios, nuestro sistema de educación pospuso para sus últimos años de la educación media un acompañamiento con orientación profesional, en el que el uso de tecnología no juega un papel protagónico.

El gran abanico de posibilidades a los que se enfrenta un joven a la hora de elegir una carrera profesional denota la necesidad de esfuerzos basados en el uso de la Ciencia de los Datos, para nutrir de aprendizajes fenómenos que giran en torno a la elección y permanencia en una carrera profesional en Colombia. Se propone en este trabajo un estudio del Desempeño Académico de Estudiantes Colombianos en las Pruebas Saber 11 - Saber Pro y su relación con la Elección Vocacional y Permanencia Universitaria.

Son muchos los retos de la educación en Colombia y aunados a ellos se evidencia que problemáticas tan complejas como perfilar a los individuos que están por culminar la educación media, hacia la educación superior se abordan con déficit en el uso de la tecnología.

2.2 Justificación

En un escenario ideal, la elección de una carrera profesional debe obedecer a un proyecto de vida y estar acompañada de orientación vocacional, siendo el sistema de educación superior de Colombia tan estructurado, a la hora de escoger una carrera profesional de pregrado, las opciones son abrumadoras para una joven al que se le pregunta ¿qué vas a estudiar?.

La oferta con cohorte al primer semestre del año 2019, daba la posibilidad de elegir entre 7.469 carreras profesionales, de las cuales por homonimia el asunto se reducía a 2.352 programas diferentes, ofrecidos por 323 instituciones de educación superior (contando cada seccional por aparte), en 3 niveles diferentes de formación (técnico, tecnológico y universitario), 8 áreas de conocimiento, 55 núcleos básicos de formación y 3 metodologías diferentes de educación (a distancia, presencial y virtual) [5].

Si a lo anterior se suma que no basta con reconocer la oferta y comprender la estructura del Sistema de Educación Superior de Colombia para elegir una carrera, sino que además sería de utilidad comprender como encajan el individuo con su personalidad, talentos, inteligencias, fortalezas y restricciones en todo esto, todos hemos tomado una decisión en cuánto a la carrera que elegimos en la que seguramente muchos de estos aspectos fueron desconocidos.

Para reconocer la necesidad de generar conocimiento sobre la elección vocacional y permanencia universitaria, se revisaron los planes de desarrollo de distintas alcaldías de Colombia, encontrando que la orientación vocacional es uno de los principales retos que asumen las Secretarías de Educación en el país.

Citando un referente del análisis expuesto en el párrafo anterior, se observan en la tabla 1 los indicadores de los estudiantes vinculados a programas de exploración y orientación vocacional de la Alcaldía de Medellín, consignados en la página 256 de su plan de desarrollo 2016-2019 [6].

Tabla 1. Indicadores de producto relacionados con el apoyo en la elección de una carrera profesional de la Alcaldía de Medellín 2016-2019.

Indicadores de producto

Nombre	Unidad	Línea de Base	Meta Plan	Logro acumulado 2019	Responsable
Estudiantes vinculados a programas de exploración y orientación vocacional	Número	851	10.000	10.851	Secretaría de Educación

Fuente: Alcaldía de Medellín. Plan de Desarrollo 2016-2019. Gaceta oficial.

Los beneficios esperados de este proyecto son el estudio de la relación entre las distintas variables recopiladas para miles de individuos en Colombia, cuando están a punto de graduarse de la educación media y ad-portas de culminar una carrera profesional de tipo universitaria en Colombia.

Con lo anterior se garantiza que los datos introducidos en la metodología seleccionada recogen información de individuos que efectivamente hicieron el tránsito de la educación media a la educación superior y además de esto superaron los ciclos de formación universitaria en al menos un 75% de completitud, lo cual facilitará la comprensión de las variables relacionadas con la elección y permanencia en estudios de formación profesional en el país.

3. OBJETIVOS

3.1 Objetivo General

Analizar la relación entre el desempeño académico de estudiantes colombianos en las Pruebas Saber 11 y Saber Pro como aporte para perfilar la elección vocacional y permanencia universitaria, mediante técnicas de Ciencia de Datos.

3.2 Objetivos Específicos

- Realizar un estudio analítico de las variables de estudiantes que están finalizando el bachillerato y son recogidas por el ICFES en las Pruebas Saber 11, a través de técnicas de analítica de datos.

- Realizar un estudio analítico de las variables de estudiantes que están finalizando la formación profesional y son recogidas por el ICFES en las Pruebas Saber Pro, a través de técnicas de analítica de datos.

- Buscar asociaciones entre el perfil de un bachiller y un estudiante adportas de culminar su carrera profesional, relacionando elementos que faciliten la elección y permanencia en una carrera profesional.

4. MARCO REFERENCIAL

4.1 Marco contextual

Se pretende realizar un estudio de la relación entre las variables con los protocolos de la calidad en la información del Instituto Colombiano para la Evaluación de la Educación (ICFES), dependiente del Ministerio de Educación Nacional, en dos momentos de la vida académica de varios estudiantes; cuando están finalizando la educación media y antes de graduarse de un programa profesional de pregrado de educación superior en Colombia, los cuales están enmarcados en el contexto de las Pruebas Saber 11 y Saber Pro, cuyos objetivos y propósitos, son entre otros [7]:

4.1.1 Pruebas Saber 11

Basados en el Decreto 869 de 2010, se citan los objetivos que se consideran más relevantes para la interpretación de esta prueba:

- Comprobar el grado de desarrollo de las competencias de los estudiantes que están por finalizar el grado undécimo de la educación media.
- Apoyar los procesos de selección y admisión que realizan las Instituciones de Educación Superior.
- Monitorear la calidad de la formación que ofrecen los establecimientos de educación media
- Producir información para la estimación del valor agregado de la educación superior.

4.1.2 Pruebas Saber Pro

Según el Decreto 3963 de 2009, los objetivos de las pruebas Saber Pro son:

- Comprobar el grado de desarrollo de las competencias de los estudiantes próximos a culminar los programas que ofrecen las instituciones de educación superior.
- Producir indicadores de valor agregado de la educación superior en relación con el nivel de competencias de quienes ingresan a este nivel.
- Servir de fuente de información para la construcción de indicadores de evaluación de la calidad de los programas e instituciones de educación

superior y del servicio público educativo, que soporten la cualificación de los procesos institucionales, la formulación de políticas y el proceso de toma de decisiones en todos los órdenes y componentes del sistema educativo.

4.2 Marco conceptual

Relación de los principales conceptos orientados a facilitar la comprensión del desarrollo de la presente investigación.

4.2.1 Orientación Vocacional

Para lo anterior, se busca apoyo en el área de conocimiento de las “Ciencias Sociales y Humanas” y su núcleo de formación “Psicología”, citando las perspectivas teóricas en Orientación Educativa, las cuales constan de 5 enfoques que hacen parte del compendio expuesto por Santana Vega [9]:

- El enfoque Psicométrico se caracteriza por la utilización de pruebas para diagnosticar el éxito académico y laboral del alumnado. La figura del orientador actúa al margen de la escuela y a la luz de un plan preestablecido.

- La perspectiva Clínico-médica, se centra en la intervención sobre las casuísticas puntuales identificadas como problemas. El papel del orientador se circunscribe al análisis de los casos-problema y a la prescripción de un plan de actuación. En contra de la escuela excluyente que se potencia desde esta perspectiva, surgen las “escuelas inclusivas” con el fin de atender y apoyar a todos los alumnos.

- El enfoque Humanista, auspiciado en la idea de humanizar las escuelas, promueve que el proceso de enseñanza-aprendizaje se desarrolle en un clima de libertad. El orientador sirve de apoyo al profesorado, que se convierte en el artífice de la práctica escolar.

- La perspectiva Sociológica determina que son las variables socioeconómicas y culturales las que establecen las trayectorias vitales del alumnado. La intervención del orientador y de otros profesionales puede llegar a subvertir las situaciones y a promover el cambio.

- El enfoque Didáctico, defensor a ultranza del carácter educativo de la orientación, considera al orientador un facilitador del proceso de enseñanza- aprendizaje.

Respecto a la orientación se destaca la importancia en diferenciar las siguientes expresiones:

- Orientación escolar: se acompaña al alumno en su proceso en la escuela.

- Orientación educativa: incluye la orientación escolar, pero se extiende a más allá de lo académico.

- Orientación profesional: contribuye a la elección de una carrera profesional, reconociendo al individuo y el contexto.

- Orientación vocacional: incluye la orientación profesional y el desarrollo de competencias para el acceso al ejercicio de una carrera.

La orientación vocacional tiene como objetivo relacionar al individuo con su entorno, despertar o potenciar sus intereses y enmarcarlos dentro de un contexto, sin embargo, culturalmente hemos adoptado la vocación como un llamado a cumplir una misión o un propósito, en ocasiones desprovisto de una argumentación objetiva, por lo cual valdría la pena analizar también qué es la orientación socio ocupacional.

En referencia al término vocación, el cual proviene de “vocatio” y se refiere a un llamado, históricamente asociado como un impulso que se acompaña de un esfuerzo sostenido para la consecución de un objetivo de carácter trascendental, dado el carácter indeterminado del cómo aparece la vocación, se dificulta su medición, cambio o construcción.

A continuación, se expone la orientación vocacional desde el punto de vista de distintos autores:

- La orientación no es un proceso puntual, sino continuo en el tiempo; no se dirige sólo a las personas con necesidades especiales, sino a todo el mundo. Se persiguen como objetivos: el desarrollo de la persona, y la prevención de problemas de toda índole; se interviene a través de programas [10].

- La orientación vocacional es un proceso que tiene como objetivo despertar intereses vocacionales, ajustar dichos intereses a la competencia laboral del sujeto y a las necesidades del mercado de trabajo, se plantea la importancia de la elección de un interés realista que le permita al sujeto alcanzar su meta laboral [11].

- La orientación no debe decidir por el individuo, es más bien una herramienta donde el otro reconoce sus aptitudes y se forma un concepto realista de sí mismo y de la situación en la que vive, involucrando a los padres en el proceso de sus hijos [12].

- Sugiere que se reemplace el concepto de orientar hacia una profesión por el de orientar para el ajuste al cambio. “Se debe enseñar a estudiar, a pensar, proveer a los jóvenes con recursos y técnicas para la expresión o la creación de conocimientos” [13].

- “La orientación es un proceso mediante el cual se ayuda y aconseja al individuo a fin de que logre una máxima ordenación interna y la mejor contribución a la sociedad.

  Lleva implícito el conocimiento de las aptitudes, intereses, rasgos de la personalidad y necesidades que siente el sujeto para su propia realización, a fin de poder aconsejarle acerca de sus problemas, asistirle en la formulación de planes y proyectos para aprovechar al máximo sus facultades, ayudarle a tomar decisiones y realizar las adaptaciones precisas para promover su ajuste y bienestar en la vida” [14].

- Establece la orientación vocacional como un proceso en el que se despiertan los intereses vocacionales y se ajustan estos a la competencia laboral del sujeto y a las necesidades del mercado laboral [15].

- Se resalta a la orientación como el arte de ayudar a aceptar y hacer uso de aquellas fuerzas naturales, que le son propias al desarrollo del estudiante. “El objeto de la orientación es guiar la transición de un campo de experiencia a otro. Una acertada orientación abre un área de conocimientos” [16].

4.2.2 Deserción y permanencia universitaria

La deserción universitaria hace referencia a la renuncia de un estudiante bien sea a una institución de educación superior o a la culminación los estudios de alguna carrera profesional en particular [17].

El Ministerio de Educación Nacional de Colombia considera como desertor a aquellos estudiantes que habiéndose matriculado previamente en el Sistema de Educación Superior, presentan una suspensión de sus estudios por un término superior a dos semestres consecutivos [17].

La deserción universitaria en Colombia trae como consecuencia la reducción del capital humano del país, dado que el no desarrollo de competencias profesionales en los ciudadanos, se traduce en una brecha en el nivel de ingresos a favor de los pocos que logran culminar con éxito un programa de educación superior y en detrimento de los que no concluyen su proyecto educativo [17].

La deserción relacionada con el tiempo se clasifica en [17]:

- Deserción precoz: individuo que previa admisión de la institución de educación superior no realiza el proceso de matrícula [17].
- Deserción temprana: abandono del proyecto educativo en los primeros semestres de la carrera profesional elegida [17].
- Deserción tardía: abandono del proyecto educativo en los últimos semestres de la carrera profesional elegida [17].

Entendiéndose la deserción como un fenómeno contrario a la permanencia universitaria, esta última expresión es asociada en el presente estudio a los individuos que logran llegar en promedio hasta el semestre 9 de la carrera profesional cursada y presentan las pruebas Saber Pro, habiendo culminado con éxito en un nivel igual o superior al 75%, la carga académica de sus carreras.

4.2.3 Ciencia de los Datos

Es un campo que se ocupa de la extracción del conocimiento a partir de volúmenes de datos que no se pueden comprender de forma intuitiva, para los cual se combinan métodos científicos, procesos y algoritmos que permiten perfilar, identificar patrones o clasificar datos que pueden ser de tipo estructurado o no estructurado.

Las técnicas de esta disciplina están relacionadas con las matemáticas, la estadística, el aprendizaje de máquinas, las ciencias de la información y la computación.

Una de sus áreas principales es el aprendizaje de máquinas, el cual se deriva de las ciencias de la computación y de la inteligencia artificial, se constituye como un conjunto de técnicas que agrupadas dentro de algoritmos permiten

generalizar conocimiento a partir de los datos con lo que se dispone o inferir comportamientos futuros basados en los datos históricos incorporados en los modelos.

El aprendizaje se puede abordar a través de métodos supervisados y no supervisados. Los métodos no supervisados facilitan el perfilado, la agrupación y/o asociación de los datos en búsqueda de información que permita identificar patrones y comportamientos al interior de los mismos, que por lo general se escapan de la intuición. Algunos de estos algoritmos son el Análisis de Componentes Principales para la identificación factores relevantes, K-means para realizar clustering, Apriori para encontrar reglas de asociación o relaciones de causa y efecto, entre otros.

Mientras que a través de los métodos supervisados, se puede identificar el valor de una variable de salida u objetivo a partir de una o varias de entrada, basándose en un histórico de variables de entrada y salida que conforman una ecuación, algunos de los algoritmos que facilitan dichos propósitos se dividen en clasificación (los Árboles de Decisión, las Redes Neuronales, la Máquina de Soporte Vectorial, la Regresión, los Métodos Bayesianos, los Métodos Basados en Vecinos, entre otros) y regresión (Lineal o Logística).

Dentro de la estadística, se destaca el método de Análisis de Componentes Principales, el cual es una técnica descriptiva de análisis multivariante, la cual facilita la reducción de la dimensionalidad de los datos, los siguientes conceptos fueron tomados en su mayoría del documento Máster en Técnicas Estadísticas Análisis Multivariante del profesor César Sánchez Sellero 2008-2009 [18].

El ACP permite pasar de una gran cantidad de variables interrelacionadas a unas pocas componentes principales. El método consiste en buscar las combinaciones lineales de las variables originales que representen lo mejor posible a la variabilidad presente en los datos.

Debido a lo anterior, con unas pocas combinaciones lineales, que serán las componentes principales, sería suficiente para entender la información contenida en los datos. Al mismo tiempo, la forma en que se construyen las componentes y su relación con unas u otras variables originales, sirven para comprender la estructura de correlación inherente a los datos [18].

4.2.4 Analíticas del aprendizaje

En este trabajo, la relación entre las pruebas de Estado y la elección vocacional se realizará a través de analíticas del aprendizaje implementando técnicas de minería de datos educativos, la cual según la International Educational Data Mining Society es una una disciplina emergente, que se ocupa del desarrollo de métodos para explorar datos únicos y cada vez más a gran escala obtenidos de entornos educativos, y utiliza dichos métodos para comprender mejor a los alumnos y los entornos en los que aprenden.

Las analíticas del aprendizaje son el resultado de la interacción entre diversas líneas de formación relacionadas con las ciencias del aprendizaje, la educación, las ciencias sociales y humanas, la estadística, la minería de datos y el relacionamiento de las personas con la tecnología [28].

Según la Sociedad para la Investigación en Analíticas de Aprendizaje (SoLAR), las analíticas de aprendizaje consisten en la medición, la recopilación, el análisis y la presentación de datos sobre los estudiantes y sus contextos, con el fin de comprender y optimizar el aprendizaje y los entornos en los que se produce [29].

En América Latina las analíticas del aprendizaje se configuran a nivel generalizado como un campo poco explorado dentro de la academia y es posible que esta apreciación se pueda extender hacia los encargados de la formulación de las políticas públicas [30].

Las analíticas del aprendizaje tienen la misión de devolver información a partir de datos extraídos en contextos relacionados con la educación, por lo cual su principal enfoque es el de retroalimentar a los estudiantes, habiendo previamente identificado a qué retos se enfrentan ellos, con el propósito de facilitarles una toma de decisiones que minimice mejor los resultados no esperados a partir de estas [30].

También se configuran como una herramienta que propende por una visión más revolucionaria del estudiante, su valor se halla en el potencial que tienen de retar la forma desintegrada como se evalúan los procesos educativos y su propuesta consiste en una visión basada en fuentes de información mejor interrelacionadas, holísticas e integradas, que con mayor precisión contribuyan a perfilar la retroalimentación que reciben los estudiantes en el sistema [31] [32].

4.3 Marco legal

Para cumplir con los objetivos del “Estudio del Desempeño Académico de Estudiantes Colombianos en las Pruebas Saber 11 - Saber Pro y su Relación con la Elección Vocacional”, se asume que se manejarán y procesarán datos sobre individuos que faciliten los objetivos de la investigación.

Algunos de los desafíos para tener en cuenta son la privacidad, el consentimiento informado, la transparencia, la localización e interpretación de los datos, los permisos para su uso y procesamiento, entre otros [35].

4.3.1 Panorama nacional

En Colombia, son dos las autoridades encargadas de la protección de los datos: la Superintendencia de Industria y Comercio (SIC), la cual es aplicable con sus disposiciones para el desarrollo de esta investigación y la Superintendencia Financiera de Colombia (SFC), reservada para supervisar la administración de datos proveniente de empresas financieras o crediticias adscritas a la SFC.

A continuación, se amplía información sobre la autoridad que define los límites legales en materia de protección de los datos, en la presente investigación:

“La Superintendencia de Industria y Comercio vela por el buen funcionamiento de los mercados a través de la vigilancia y protección de la libre competencia económica, de los derechos de los consumidores, del cumplimiento de aspectos concernientes con metrología legal y reglamentos técnicos, la actividad valuadora del país, y la gestión de las Cámaras de Comercio” [36].

“A su vez, es responsable por la protección de datos personales, administra y promueve el Sistema de Propiedad Industrial y dirime las controversias que se presenten ante afectaciones de derechos particulares relacionados con la protección del consumidor, asuntos de competencia desleal y derechos de propiedad industrial” [36].

A continuación, se presenta la legislación, definiciones y terminología relacionadas con el manejo de datos en Colombia:

La ley 1266 de 2008, tuvo como propósito configurarse como una directriz de principios generales aplicable a todas las categorías de los datos personales, sin embargo, no logró ir más allá de fijar los estándares básicos de protección para los datos financieros y comerciales, relacionados con el cálculo del riesgo crediticio de los individuos [37]. Por tal razón a través de la sentencia C-1011 de 2008, la Corte Constitucional acotó los alcances de esta ley [38].

Al ser considerada como una ley que parcialmente resolvía asuntos en materia de habeas data, a través de la sentencia C-748 de 2011, se evidencia que se da paso a “un sistema híbrido de protección en el que confluye una ley de principios generales con otras regulaciones sectoriales, que deben leerse en concordancia con la ley general, pero que introduce reglas específicas que atienden a la complejidad del tratamiento de cada tipo de dato” [39].

A partir de los cuestionamientos planteados anteriormente, surge la ley que dicta las disposiciones generales para la protección de datos personales, conocida como ley 1581 de 2012, la cual contiene pautas específicas para cada tipo de dato, y a su vez fue reglamentada parcialmente por el Decreto Nacional 1377 de 2013, en referencia a los aspectos afines a la autorización del titular de información para el tratamiento de sus datos personales [40].

A continuación, se exponen algunos conceptos aplicables a la protección de los datos en Colombia:

¿Qué es el derecho de Hábeas Data?

“El derecho de hábeas data es aquel que tiene toda persona de conocer, actualizar y rectificar la información que se haya recogido sobre ella en archivos y bancos de datos de naturaleza pública o privada [41]”.

“La Corte Constitucional lo definió como el derecho que otorga la facultad al titular de datos personales de exigir de las administradoras de esos datos el acceso, inclusión, exclusión, corrección, adición, actualización y certificación de los datos, así como la limitación en las posibilidades de su divulgación, publicación o cesión, de conformidad con los principios que regulan el proceso de administración de datos personales. Asimismo, ha señalado que este derecho tiene una naturaleza autónoma que lo diferencia de otras garantías con las que está en permanente relación, como los derechos a la intimidad y a la información [41]”.

¿Quién es el titular de la información?

“El titular de la información es la persona natural o jurídica a quien se refiere la información que reposa en un banco de datos. Ejemplo: Un usuario que celebró el contrato de prestación de servicio de comunicaciones [41]”.

¿Qué es el principio de interpretación integral de derechos constitucionales?

“La interpretación integral de derechos constitucionales, consiste en que la normas que rigen los datos personales se interpretarán en el sentido que se amparen otros derechos constitucionales, como son el hábeas data, el derecho al buen nombre, el derecho a la honra, el derecho a la intimidad y el derecho a la información. Asimismo, se refiere a que los derechos de los titulares se interpretarán en armonía con el derecho a la información y demás derechos constitucionales aplicables [41]”.

Con el propósito de facilitar la comprensión de los conceptos anteriores, se cita textualmente el artículo 1 de la ley 1581 de 2012, por la cual se dictan disposiciones generales para la protección de datos personales, tomado del Diario Oficial No. 48.587 de 18 de octubre de 2012 del Congreso de la República de Colombia [42].

“Artículo 1. Objeto. la presente ley tiene por objeto desarrollar el derecho constitucional que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos, y los demás derechos, libertades y garantías constitucionales a que se refiere el Artículo 15 de la constitución política. así como el derecho a la información consagrado en el Artículo 20 de la misma [42]”.

Dicha ley se rige por los principios de legalidad, finalidad, libertad, veracidad o calidad, transparencia, acceso y circulación restringida, seguridad y confidencialidad, los cuales se deben aplicar de manera armónica e integral para garantizar la política de protección de datos en Colombia [42].

Por otro lado, la ley 1712 de 2014, la cual tiene por objeto regular el derecho de acceso a la información pública, los procedimientos para el ejercicio y garantía del derecho y las excepciones a la publicidad de información, y constituye el marco general de la protección del ejercicio del derecho de acceso a la información pública en Colombia, también se considera pertinente para esta investigación [43].

La ley 1712 de 2014 se define a través del principio de máxima publicidad en la que se hace alusión a la universalidad del titular, lo cual significa que toda

información en posesión, bajo control o custodia de un sujeto obligado es pública y no podrá ser reservada o limitada sino por disposición constitucional o legal, también se define a través de los principios de la transparencia, buena fe, facilitación, no discriminación, gratuidad, celeridad, eficacia, calidad de la información, divulgación proactiva de la información y responsabilidad en el uso de la información [43].

4.3.2 Panorama global

La unión europea

El Reglamento general de protección de datos (RGPD) 2016/679, el cual derogó la directiva 95/46/EC, es atribuible a todos los países miembros de la Unión Europea (UE) y relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos, entró en vigor desde el 24 de mayo de 2016 [44].

Sin embargo, fue mucho antes que surgió la necesidad de facilitar el flujo de los datos entre los países miembros de la unión europea, por lo cual en el año 1995 se acoge la Directiva de Protección de Datos 95/46/EC, la cual adoptó los 7 principios de notificación, propósito, consentimiento, seguridad, transparencia, acceso y responsabilidad, que fueron recomendados en 1980 por la Organización de Cooperación y Desarrollo Económico (OCDE) [45] [46].

Dichos principios fueron concebidos como unas directrices para facilitar la protección de la privacidad y el flujo transfronterizo de los datos personales.

Como mandato base, se recomendó que los datos personales no deberían ser tratados en absoluto, excepto cuando se cumplieran ciertas condiciones. Estas condiciones se agruparon en tres categorías.

Transparencia, el individuo tiene derecho a conocer que sus datos serán recopilados y los propósitos y destinatarios de las actividades que sobre sus datos se lleven a cabo. Legitimidad, los intereses que se tengan sobre los datos personales de los individuos deben estrictamente estar enmarcados dentro de finalidades aceptadas por la ley y proporcionalidad, los datos recopilados no deben en ningún momento exceder los propósitos por los cuales fueron requeridos [46].

Latinoamérica

La Directiva de Protección de Datos 95/46/EC promulgada en el año 1995 en la unión europea, sirvió de base para las leyes en materia de protección de datos en Argentina, Uruguay, México, Perú, Costa Rica y Colombia [47].

En la actualidad en Chile y Paraguay, no están definidas las autoridades para la protección de los datos, a pesar de que en estos países se ha legislado en este campo [47].

Por otro lado, en el grupo de los países que le han otorgado al habeas data un estatus de derecho constitucional, se destacan Panamá, Honduras, Colombia, Ecuador, Perú, Argentina, Paraguay y Brasil [47].

Se estima que en Latinoamérica, un porcentaje superior a más de la mitad de los países han procurado adoptar normas en función de proteger los datos y la privacidad de los mismos [48].

Como referencia a las diferentes legislaciones en materia de protección de los datos, se evidencia al menos entre los 7 países más poblados de latino américa, México, Argentina, Colombia, Perú, Brasil, Chile y Venezuela, que a nivel generalizado exceptuando Venezuela, todos cuentan con esfuerzos en materia de normativas y legislación dirigidos a la protección de los datos [30] [49].

A nivel global, Estados Unidos, Canadá, Noruega, los países pertenecientes a la Unión Europea, Hong Kong, Corea del Sur y Australia, son los países que con mayor rigor han legislado sobre estos temas [49].

En una segunda posición se encuentran Argentina, Marruecos, Túnez, algunos países de Europa del Este, Grecia, China, Japón, Malasia y Nueva Zelanda, en una tercera posición; México, Costa Rica, República Dominicana, Colombia, Brasil, Perú, Chile, Uruguay, algunos países de África y Europa del Este, Rusia y Uzbekistán [49].
Por último, rotulados como limitados en materia de legislación en protección de datos, se encuentran, Honduras, Panamá, Islas Caimán, Trinidad y Tobago y algunos países de África y Asia, entre otros. Las zonas en donde se reconocen más países sin legislaciones específicas en materia de protección de datos se encuentran en África, Asia y Sur América [49].

4.4 Estado del arte

4.4.1 Aplicaciones Relacionadas con Orientación Profesional

Tratándose de variables que revelan capacidades de los individuos medidas dentro del contexto de la educación media y superior, se enfoca el estado del arte dentro de un marco que facilite el análisis desde la perspectiva profesional de John Holland (1919-2008), quizás el psicólogo más relevante en el ámbito de la orientación profesional de la historia contemporánea de Estados Unidos.

Holland trabajó durante 1953-1958 en un inventario de preferencias vocacionales, conocido como “Vocational Preference Inventory – VPI”, en este trabajo el autor asocia las preferencias vocacionales a diferentes tipos de personalidad atribuibles a un individuo y a su vez identificables a través de cada una de las letras que conforman la expresión “RICSEA”.

En 1966, se tomaron los principales perfiles del inventario de preferencias vocacionales VPI para las ocupaciones, asumiendo que estas podían clasificarse en igual número de grupos paralelos a los 6 tipos de personalidad establecidos por Holland y se usaron para configurar el código ocupacional de tres letras.

En 1969, se descubre un orden hexagonal para su inventario y reorganizan las categorías con la siguiente estructura RIASEC, en donde cada letra se distribuye a lo largo de los vértices de un hexágono y representa a un tipo de personalidad asociado a una letra, con los siguientes significados.

R: Realistic/Realístico.
I: Investigative/Investigativo.
A: Artistic/Artístico.
S: Social/Social.
E: Enterprising/Empresarial.
C: Conventional/Convencional.

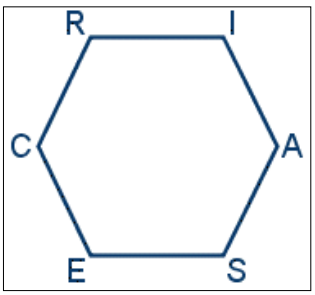


Figura 1. Modelo hexagonal de relación entre los tipos de personalidad y modelos ambientales.

Holland expuso que cada individuo tiene una combinación de diferentes tipos de personalidad y a través de sus teorías permite identificar una triada de estas.

Los tipos de personalidad que son encontrados para cada individuo a la luz de su propuesta, se analizan tomando en cuenta los dos tipos de personalidad más dominantes en el resultado, se evalúan las letras que representan al resultado y dado que la distribución de las letras responde a un modelo de relación entre los tipos de personalidad ubicados en el hexágono, se interpreta de la siguiente manera:

la consistencia en el resultado es alta, si los dos tipos de personalidad más relevantes resultantes son categorías adyacentes, media si estos son alternos y baja si se encuentran en posiciones opuestas.

En 1970, John Holland, recibe el encargo de crear un recurso para prestar orientación profesional a estudiantes universitarios de primer año.

Basándose en todos los estudios teóricos que había adelantado entre 1959-1969, desde 1970 hasta 1977 se ocupó de perfeccionar el Self-Directed Search (SDS), una guía auto aplicable para la planificación de la carrera que se puede expresar como un inventario de intereses vocacionales y profesionales, dicha guía se basó en su propia Teoría Tipológica y en el Vocational Preference Inventory (VPI), el cual fue uno de los primeros instrumentos de evaluación de intereses profesionales, creados por este reconocido autor.

El SDS en su Forma R (SDS-R) se compone de tres cuadernos (Holland, 1995a, 1995b, 1995c), diseñados para ser usados con estudiantes de High School (en Estados Unidos se refieren a los últimos 4 años de educación antes de la formación universitaria), estudiantes universitarios y adultos:

- El Cuaderno de Evaluación (The Assesement booklet) [55].
- El Descubridor de ocupaciones (The Occupations finder) / El Descubridor de Ocupaciones que es una versión abreviada del Diccionario de Códigos Ocupacionales (DOT; Gottfredson y Holland, 1989) en 1335 ocupaciones son clasificadas con un código de tres

letras de Holland [56].

- Usted y su carrera (You and your career) [57].

Hay otras versiones del SDS, cada una específica para una población determinada, las cuales son:

- La versión Fácil (Form E).
- La versión Planificación de la Carrera (Form PC).
- la versión de Exploración de la Carrera (Career Explorer)

El Self-Directed Search (SDS-R) y sus características fundamentales, según Holland y Rayman (1986):

- El incremento de la auto comprensión del número de alternativas vocacionales a tener en cuenta.
- La simplicidad en la puntuación del inventario.
- Posibilita la autoexploración comprensiva de las posibilidades ocupacionales y de las características personales.
- Permite la auto aplicación, corrección e interpretación (sin la presencia de un asesor).
- Presenta compatibilidad con muchos materiales de orientación vocacional.
- Es tanto un inventario de intereses como de personalidad.
- Organiza dentro de una teoría y relaciona las aspiraciones vocacionales, los intereses, las competencias y las auto estimaciones de una persona con una clasificación ocupacional.

La Teoría tipológica de Holland (1975, 1981, 1992, 1997), se traza el objetivo de explicar cómo las personas toman decisiones vocacionales y profesionales; que móviles tienen para querer modificar su situación laboral, qué causa tienen los cambios de sus intereses en las elecciones vocacionales, en función de alcanzar una satisfacción profesional [60].

Para esta teoría John Holland, se respalda en su convicción que las personas

con diferentes tipos de intereses reflejan a su vez distintos tipos de personalidad.

En referencia a la personalidad y el ambiente laboral, Holland en 1997 expresa que “el trabajo cambia a las personas, y las personas cambian los trabajos”.

En 1983 Howard Gardner, psicólogo también norteamericano plantea la teoría de las inteligencias múltiples, en la que redefine el concepto de inteligencia en el cual hasta el momento se basaba en las habilidades lingüísticas y lógico-matemáticas y en los resultados académicos de los individuos.

En su propuesta define 8 tipos posibles de inteligencias, como una red de conjuntos autónomos interrelacionados entre sí [62]:

- **Lingüística (o verbal-lingüística):** habilidad para utilizar con un dominio avanzado el lenguaje oral y escrito, así como para responder a él.
- **Lógico-matemática:** habilidad para el razonamiento complejo, la relación causa-efecto, la abstracción y la resolución de problemas.
- **Viso - espacial, corporal o quinestésica:** habilidad de utilizar el cuerpo para aprender y para expresar ideas y sentimientos. Incluye el dominio de habilidades físicas como el equilibrio, la fuerza, la flexibilidad y la velocidad.
- **Musical o rítmica:** habilidad de saber utilizar y responder a los diferentes elementos musicales (ritmo, timbre y tono).
- **Intrapersonal o individual:** habilidad de comprenderse a sí mismo y utilizar este conocimiento para operar de manera efectiva en la vida.
- **Interpersonal o social:** habilidad de interactuar y comprender a las personas y sus relaciones.
- **Naturalista:** habilidad para el pensamiento científico, para observar la naturaleza, identificar patrones y utilizarla de manera productiva.

El autor plantea que, así como los problemas tienen distintas formas de abordarse se puede sustentar que los individuos emplean distintas inteligencias.

En el 2006 en España se crea una adaptación española del Self-Directed Search (SDS-R), proyecto que fue titulado como: La elección vocacional y la planificación de la carrera.

En un estudio previo se habían detectado 68 ítems que debían ser rectificados por no ajustarse a criterios psicométricos y semánticos de España, por lo cual

los autores deciden llevar a cabo una adaptación con juicios de fiabilidad y validez [63].

Aplicando su trabajo a una muestra de 1.460 personas de diferentes niveles educativos, logran evidenciar mejoras en el índice de discriminación de los ítems y la consistencia interna de las escalas [63].

Con su trabajo los autores sustentan que llevaron a cabo distintos procedimientos para verificar la adaptación española con una validez aceptable del Self-Directed Search (SDS-R) de John Holland en cuanto a contenido, constructo y concurrente [63].

Los autores concluyen que investigación goza de la suficiente garantía científica y técnica para ser aplicada a la población española [63].

4.4.2 Aplicaciones de Analíticas del Aprendizaje en el estudio de exámenes de Estado

En el 2007, una investigación determina que, para cierto número de profesores de álgebra en colegios de secundaria en Colombia, se evidencia un déficit de conocimiento o competencias en cuanto a la materia que tienen a cargo, lo cual afecta su labor académica, a pesar de esto, ellos atribuyen el bajo rendimiento de sus estudiantes a otros factores del contexto y/o institucionales, pero no a su falta de capacidades [64].

En el 2007 Red Iberoamericana de Indicadores de Ciencia y Tecnología (RICYT), menciona que en ese año por cada 1000 personas en la fuerza laboral de América Latina y el Caribe, el número de investigadores fue de 1.96, mientras que dos años más tarde los países industrializados reunidos en la Organización para la Cooperación y el Desarrollo Económico (OCDE), presentaban un promedio de 7.3 para ese indicador [65].

Un análisis basado en el género de los datos de Segundo Estudio Regional Comparativo y Explicativo (SERCE) 2008 revela que los estudiantes hombres de sexto grado en Colombia, al igual que en El Salvador y Perú, tienen una diferencia significativa a favor de ellos frente al desempeño de las mujeres en Ciencias Naturales [66].

Mientras tanto, en países como Argentina, Cuba, Panamá, Paraguay, República Dominicana y Uruguay, no se encontraron diferencias significativas

discriminando los géneros a nivel del desempeño de los estudiantes evaluados [66].

En el 2010, se realiza un estudio titulado “La condición de la educación en matemáticas y ciencias naturales en América Latina y el Caribe”, el cual toma como premisa que las habilidades en matemáticas y ciencias naturales no se han desarrollado adecuadamente en los alumnos de los países analizados [67].

Explicando que en los esfuerzos de los gobiernos, padres de familia, educadores e investigadores en América Latina y el Caribe, se han descuidado el desarrollo de las capacidades cuantitativas y científicas de su población, dada la necesidad de priorizar el alfabetizar a los estudiantes en las etapas de preescolar, primaria y secundaria [67].

En el estudio se manifiesta que los rendimientos de los alumnos de países de América Latina y el Caribe, constantemente se encuentran por debajo de los reportados para estudiantes de Asia Oriental y de los países industrializados reunidos en la Organización para la Cooperación y el Desarrollo Económico (OCDE) [67].

En el año 2013, se llevó a cabo una investigación financiada por el Ministerio de Educación Nacional de Colombia, cuya finalidad fue la de encontrar patrones de deserción en dos instituciones de educación superior de San Juan de Pasto, La Universidad de Nariño (carácter público) y la Institución Universitaria CESMAG (carácter privado) [68].

Para lo cual se eligieron datos socioeconómicos, académicos, disciplinares e institucionales, de los que disponían las dos instituciones, tanto en sus repositorios internos a través de sus oficinas de admisiones, como los procedentes de fuentes externas como el Instituto Colombiano para el Fomento de la Educación Superior (ICFES), Departamento Administrativo Nacional de Estadística (DANE), Sistema para la Prevención de la Deserción en la Educación Superior (SPADIES), entre otros [68].

Finalmente se llegó a un repositorio de 16 datos socio económicos, 11 que representaban resultados académicos y 1 variable de salida o de tipo clase, para un total de 2.136 estudiantes, registros equilibrados entre las dos instituciones seleccionadas en este proyecto [68].

Los resultados de este trabajo arrojaron en las primeras posiciones para explicar la deserción universitaria, a los factores académicos, en especial al bajo rendimiento de los estudiantes medido por un promedio de notas inferior a 2.4, situación que se presentó en el 19% de los casos de los estudiantes tomados en la muestra y a su vez, característica presente en el 34.8% de los desertores [68].

El riesgo de deserción sigue siendo alto, cuando el promedio de notas de los estudiantes es inferior a 3.1, pues el 18% de los estudiantes de la muestra, obtuvieron un promedio de notas entre 2.4 y 3.1 y el 32.8% de los desertores coincidieron con este desempeño académico [68].

El estudio revelo que vivir con la familia está asociado con la deserción estudiantil y con la finalidad de encontrar otros factores que explicarán el fenómeno de la deserción, se procedió a una poda de atributos para encontrar reglas más específicas, encontrando que tener un promedio de ICFES inferior a 48 puntos, un descuido en las materias de las Ciencias Básicas y proceder de la Costa Pacífica Nariñense, son variables que pueden incidir en la deserción [68].

En el 2014 Albor, Gustavo & Lorduy, Viviana & Dau, Marco, investigan cofinanciados por el Ministerio de Educación Nacional, en la Convocatoria denominada Realización de Estudios sobre Educación Superior: “Trabajos de Investigación sobre el Sistema de Educación Superior en 2011”, sobre la calidad de la educación presencial versus virtual, basándose en las pruebas Saber Pro del 2010 [69].

Los autores citados encuentran que los resultados de los estudiantes que son formados de manera presencial presentan mejor rendimiento en todas las áreas analizadas, también afirman que la capacidad de la universidad de transferir conocimiento a los estudiantes disminuye bajo la metodología a distancia [69].

Por otra parte, los autores identifican una brecha de género en la cual, los hombres obtienen mejor desempeño en todas las áreas en comparación con las mujeres y las variables socioeconómicas tienen una relación positiva con el desempeño de los estudiantes [69].

Continuando en el 2014, Cantillo, V., & García, L., realizan un estudio de las pruebas Saber 11 y Saber Pro para determinar como el género y otros factores

influencian los resultados, poniendo el foco en el programa de Ingeniera Civil [70].

La investigación determina que no existe diferencia entre el desempeño de los géneros en las Pruebas Saber 11, excepto que los hombres se destacan en Física mientras que las mujeres en habilidad verbal [70].

Mas adelante en la vida académica del estudiante, analizada a través de los resultados en las pruebas Saber Pro, los hombres presentan resultados más sobresalientes a nivel general con respecto a las mujeres y ellas solo se destacan en lo relacionado con competencias verbales [70].

En este trabajo se plantea que lo anterior puede explicarse por un sesgo de género en la estructuración del programa de Ingeniería Civil en el país, por lo cual los autores sugieren que los resultados que se alcanzan en esta prueba dependen de factores institucionales que permiten inferir que la formación y el desempeño del estudiante están sujetos a la calidad de la institución de educación superior [70].

Acotando en el 2014, se lleva a cabo una investigación titulada “Calidad institucional y rendimiento académico El caso de las universidades del Caribe Colombiano”, la cual se basa en los resultados de obtenidos por estudiantes que realizaron las pruebas Saber Pro en el Caribe colombiano, en el año 2009 [71].

Se encuentra que en algunas áreas de conocimiento el estrato socio económico del estudiante está directamente relacionado con su desempeño en las pruebas, sin embargo, contrario a lo que algunos estudio empíricos anteriores plantean, el citado estudio afirma que en la universidad la variable estrato socio económico deja de ser tan relevante como en etapas previas de la formación académica para el rendimiento académico [71].

Lo anterior se puede explicar debido al filtro en cuanto a las calidades y competencias, realizado por las instituciones de educación superior para la selección de sus alumnos y los retos que implican la permanencia en un programa de pregrado [71].

En este estudio además se encontró que los hombres se desempeñan mejor que las mujeres en las áreas de ingenierías y Economía, refiriéndose también

a que es un resultado que se ha documentado en todos los niveles de formación de la vida académica en el país [71].

Los autores sugieren la importancia de fortalecer en las mujeres las habilidades para las matemáticas desde la primaria, planteando que las brechas de genero a favor de los hombres, son tan importantes como los niveles de calidad de la universidad y los efectos de los factores socioeconómicos en el rendimiento académico de los universitarios [71].

Por último, la investigación sugiere incluir variables de tipo subjetivo que reconozcan al individuo y soportarse en las Ciencias Sociales y Humanas, en específico en la psicología y la sociología, para futuros estudios sobre el rendimiento académico de los estudiantes y poder ofrecer un acercamiento diferente que explique los procesos de elección de los individuos [71].

En el 2016 se realizó un estudio, basado en los resultados de las pruebas Saber Pro del segundo semestre del año 2011, titulado “Descubrimiento de patrones de desempeño académico en las competencias genéricas”, apoyado en la metodología Cross Industry Standard Process for Data Mining CRISP-DM [72].

El proyecto tomó como insumo los datos de estudiantes de programas en niveles de formación tecnológico y universitario, recopilados por el Instituto Colombiano para la Evaluación de la Educación - ICFES, a través de sus pruebas Saber Pro, las cuales evalúan independientemente del programa, cuatro competencias genéricas; lectura crítica, razonamiento cuantitativo, escritura e inglés [72].

Este estudio detalla los procesos de preparación de los datos en cuanto a la limpieza y transformación de estos, perfilándolos hacia modelos de clasificación basados en arboles de decisión, construidos de forma independiente para cada una de las cuatro competencias genéricas, en búsqueda de patrones relacionados con el desempeño académico de los estudiantes [72].

Apoyados en este trabajo, se evidencia a nivel generalizado que la variable tipo de acreditación de la institución, para la cual se asumen dos resultados posibles (de alta calidad o registro calificado), es supremamente determinante en el desempeño académico de los estudiantes, pues en el caso de que la institución cuente con acreditación de alta calidad en ocasiones sin necesidad

de más variables, se puede rápidamente inferir cual será el desempeño del estudiante en las competencias genéricas [72].

Por otro lado, a pesar de que esta investigación entrega muchos detalles de la preparación de los datos, queda faltando mayor precisión en la evaluación de los modelos, lo cual probablemente sugiere que las variables de entrada no tienen un poder explicativo muy contundente en el desempeño académico de los estudiantes, dado que se evidencia una gran relevancia de las instancias mal clasificadas en los cuatro modelos [72]

En el 2017 se publicó un estudio basado en una muestra de 805 estudiantes pertenecientes a facultades de Educación y ciclos complementarios de Escuelas Normales Superiores (ENS) en Colombia. En este trabajo se evidenció lo siguiente:

La mayoría de los estudiantes de la muestra, pertenecen a un estrato socioeconómico muy bajo y más del 70% de ellos se encuentran registrados en el Sistema de Selección de Beneficiarios Para Programas Sociales de Colombia (SISBEN), en su entorno familiar es constante una escasa formación, por lo cual al parecer la formación en el área de la educación es vista como una oportunidad de escalar a nivel social y lograr una estabilidad laboral que eventualmente se pudiera obtener [73].

A nivel nacional, latinoamericano y de otros países, estudios realizados sugieren que las motivaciones para escoger una carrera en esta área suelen no ser de tipo personal sino más bien estar condicionadas al valor o representación altruista de intervenir en la formación de los demás [73] [74] [75] [76] [77].

La situación anterior fue reivindicada por el estudio, al considerar que en Colombia no existen políticas públicas solidas que le den un estatus de dignidad social al ejercicio de la profesión docente, como una de las dificultades para lograr que personas pertenecientes a estratos socioeconómicos altos se interesen por estas carreras [73].

En el 2018 se realiza una investigación de minería de datos educativos que pretende analizar los factores económicos, sociales y demográficos que influyen en el desempeño de los estudiantes de ingeniería en Antioquia en las Pruebas Saber Pro [78].

Luego de aplicar distintos modelos de Aprendizaje de Maquina (Machine Learning), del trabajo del autor se rescata lo siguiente:

- El costo de la matrícula y el resultado en el módulo de inglés tienen una relación directa y positiva [78].
- Las variables estrato y valor de la matrícula, tienen relaciones de tipo inverso con la variable lectura crítica [78].
- Los mejores desempeños desde el punto de vista de los módulos evaluados en las Pruebas Saber Pro, los obtienen los estudiantes que optan por la modalidad de estudio presencial [78].
- los hombres tienen una dificultad generalizada para la comunicación escrita, reafirmando la postura de un estudio realizado por Cantillo, V., & García, L [78] [79].
- Las variables socio económicas y el género no fueron consistentes a la hora de explicar fenómenos [78].

4.4.3 Discusión

En el estado del arte se evidencia que los estudiantes de secundaria en Colombia presentan bajo rendimiento en álgebra y que existe un déficit en la calidad docente en este campo, sin embargo, estos últimos no se reconocen como parte del problema y además de esto, le atribuyen a otros factores ajenos al déficit en sus competencias docentes, el bajo rendimiento de sus alumnos.

También se deduce que la investigación no es un interés de los países de América Latina y el Caribe, pues este perfil está muy lejos de predominar en la fuerza laboral de estos países, mientras que en los países industrializados este indicador es mejor, se puede asumir que para la primera zona no hay una formación que se base en la producción de conocimiento.

Adicional a lo anterior, de forma generalizada se denota una falla en la formación base de los alumnos de países de América Latina el Caribe en Ciencias Naturales y Matemáticas, dado que se presume que se los recursos de estos gobiernos se han volcado principalmente en alfabetizar a sus ciudadanos.

Para el caso Colombia, el factor genero ha demostrado ser determinante a lo largo del ciclo educativo de los estudiantes, dado que varias investigaciones coinciden en que los hombres tienen una ventaja a su favor principalmente para las matemáticas, otros estudios reconocen que la supremacía de los

hombres en el rendimiento académico se extiende a todas las áreas de conocimiento de la educación superior, exceptuando aquellas en las que predominan las habilidades comunicativas.

También se afirma que la educación superior presencial en Colombia se asocia al buen rendimiento de los alumnos, contrario a las metodologías a distancia que no logran ser contundentes en la transferencia de conocimiento.

Algunos estudios explican las variables socio económicas como directamente relacionadas con el desempeño académico; sin embargo, en uno de ellos se expone que para el caso de la educación superior en Colombia es posible que los filtros de ingreso y las exigencias para la permanencia de los alumnos desdibujen el poder explicativo de estas variables en el rendimiento académico.

5. METODOLOGÍA

Esta investigación se basará en la metodología CRISP-DM (CRoss Industry Standard Process for Data Mining), la cual se encuentra representada en la siguiente imagen [8].

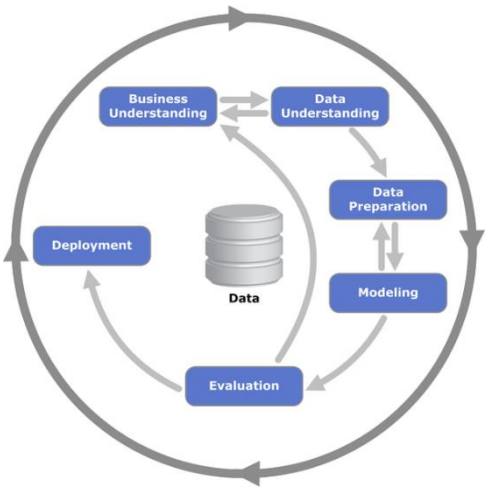


Figura 2. Diagrama de la Metodología CRISP-DM.

Fuente: https://es.wikipedia.org/wiki/Archivo:CRISP-DM_Process_Diagram.png

Está metodología proporciona un modelo estandarizado para la minería de datos, del cual se aplicarán los siguientes pasos en esta investigación:

Comprensión del negocio

Se empezará con el reconocimiento de información que facilite la comprensión sobre el sector de la educación superior en Colombia, aquí se tendrán en cuenta los reportes sobre el estado de la educación superior que realiza la firma independiente Tnt Colombia SAS a través de su proyecto publicado como TuCarrera.co, el cual toma como insumo información oficial del Gobierno de Colombia.

Comprensión de los datos

En este paso, se estudiarán las variables que servirán como punto de partida para la presente investigación, procurando reconocerlas dentro del contexto de la educación en Colombia.

Adicional a las variables relacionadas con las pruebas Saber 11 y Saber Pro, las cuales tienen como fuente de origen el ICFES, se revisarán algunos datos

de la teoría de los tipos de personalidad creada por el psicólogo norteamericano John Holland y las inteligencias múltiples de Howard Gardner, para explorar más adelante si estos pueden o no agregar valor a esta investigación.

Preparación de los datos

Se extraerá la totalidad de los datos que nos interesan, de las bases de datos en donde se encuentran.

Se aplicarán solo los pasos que sean necesarios procurando mejorar la calidad de los mismos:

- Integración de los datos
- Reducción de variables (eliminar variables irrelevantes o redundantes)
- Descripción estadística de los datos
- Limpieza de los datos
- Transformaciones
- Análisis de correlaciones
- Reducción de variables
- Balanceo de los datos

Modelado

Se escogerán los modelos más convenientes para extraer información a partir de los datos de acuerdo con los propósitos de esta investigación, los cuales pueden ser de tipo:

- Clustering (descubrir observaciones que tengan una separabilidad alta y una cohesión baja entre los grupos).
- Asociaciones y secuencias (encontrar afinidades e identificar los eventos más probables a ocurrir).
- Clasificación (determinar resultados esperados).
- Predicción (pronosticar eventos futuros basados en el aprendizaje de la maquina con datos históricos).

Evaluación

Partiendo de la validación cruzada para la base de entrenamiento y de testeo en el modelamiento, se usarán indicadores que permitan evaluar la calidad de los modelos que sean escogidos; a continuación, se presentan las medidas más usadas.

Para técnicas supervisadas, en ejercicios de clasificación, es decir predicción de categorías, algunas medidas son:

- La matriz de confusión
- La curva ROC
- Precisión
- Cobertura o Recall
- Exactitud.

Para técnicas no supervisadas como clustering algunas medidas son:

- Cohesión (que tienda a cero).
- Separabilidad (que tienda a infinito).
- Dunn Index (sirve el más grande).
- Davies and Bouldin Index (sirve el más pequeño).
- Silhouette Index (entre 0 y 1, entre más cercano a 1, más grande la separabilidad y más pequeña la cohesión).

En reglas de asociación las medidas usadas son soporte y confianza.

Despliegue

La fase conocida como “deployment” o despliegue de esta metodología puede considerar una estrategia para la incorporación de los modelos desarrollados en el sector analizado sin embargo, no es el caso de esta investigación, dado que se carece de injerencia sobre las prácticas en el sistema de educación del país.

Lo que si compete a este proyecto es mostrar los resultados de cada uno de los modelos desarrollados para los propósitos de este proyecto, cuyo alcance se describe en los objetivos del mismo.

La investigación será de tipo mixto ya que se analizarán variables con enfoques cualitativos y cuantitativos a partir de observaciones sobre estudiantes, recolectadas por el ICFES en dos momentos de la vida académica de los individuos.

PARTE 2: SOLUCIÓN Y DESARROLLO DE LA METODOLOGÍA CRISP-DM

6. DISEÑO DE LA SOLUCIÓN

Con el objetivo de analizar la relación entre el desempeño académico de estudiantes colombianos en las Pruebas Saber 11 y Saber Pro como aporte para perfilar la elección vocacional y permanencia universitaria, se diseña la siguiente solución analítica, donde se tienen 3 fuentes de datos: resultados de las pruebas Saber 11, resultados de las Pruebas Saber Pro y los identificadores únicos por estudiante que permiten enlazar ambos resultados.

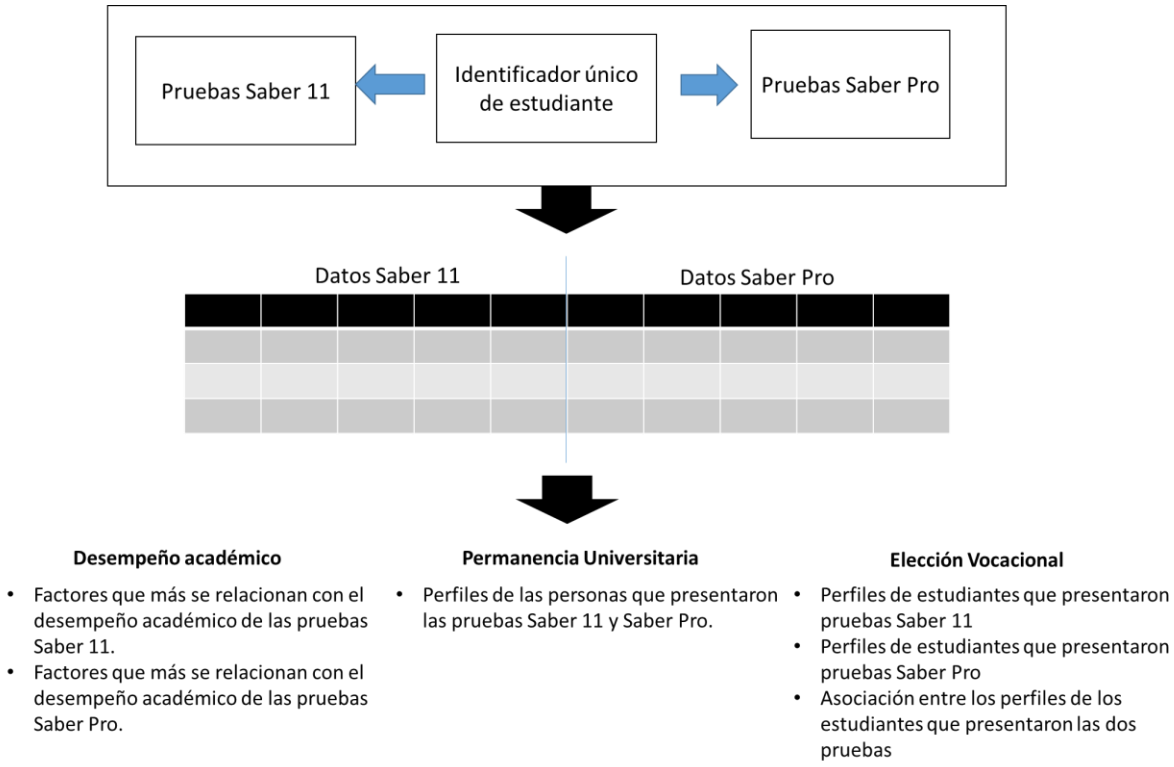


Figura 3. lustración 1. Diseño de la solución.

Mediante las 3 fuentes de datos, se construye una sábana unificada de datos, la cual es usada para crear varios modelos analíticos que permiten analizar permanencia universitaria y elección vocacional. En los capítulos siguientes se describe a detalle la creación de dichos modelos mediante la aplicación de la metodología CRISP-DM.

7. COMPRENSIÓN DEL NEGOCIO Y DE LOS DATOS

7.1 Descripción del Negocio

Se inicia con el reconocimiento de información que facilite la comprensión de información valiosa sobre el sistema de la educación de Colombia. Con estadísticas tomadas del proyecto publicado como www.TuCarrera.co, perteneciente a la empresa Tnt Colombia SAS, encargada de procesar datos abiertos del Ministerio de Educación Nacional y del Departamento Administrativo Nacional de Estadística DANE, para entregar un panorama independiente sobre la educación en Colombia, se deduce lo siguiente.

Partiendo de un tamaño de población estimado en 22.220.455 de colombianos entre los 15 y 44 años para el año 2015, el cual se consideran como ciudadanos con potencial para ingresar a la educación superior en el país, solo aproximadamente un 10% de ellos se matriculó en ese periodo en una carrera de educación superior en Colombia.

Si a lo anterior se suman las altas tasas de deserción de la educación superior en el país, el problema de cobertura y acceso a la educación se hace más escandaloso.

El 70% de la oferta en educación superior de Colombia se concentra en 5 zonas principalmente, Bogotá D.C, Antioquia, Bolívar, Santander y Valle del Cauca, para el resto de los 28 departamentos está disponible una oferta de un 30% en relación al sistema de educación superior, compuesto en su oferta por un total de 14.183 carreras profesionales a la cohorte de junio de 2019 (7.504 de estas de pregrado y 6.679 de posgrado).

Las carreras profesionales que más solicitudes de inscripción han recibido en los últimos años por parte de un interesado en ser admitido para un programa de educación superior en Colombia, son: medicina, enfermería, administración de empresas, contaduría pública, odontología, psicología, ingeniería industrial, derecho, ingeniería ambiental y veterinaria.

Por otro lado, se citan algunas disposiciones sobre las evaluaciones que el estado utiliza para medir la aprehensión de conocimiento por parte de los estudiantes a través de las Pruebas Saber 11 y Saber Pro y a su vez, valorar la capacidad de los colegios y las instituciones de educación superior de

transferir a sus alumnos el aprendizaje.

El ICFES es el Instituto Colombiano para la Evaluación de la Educación, tiene a su cargo el diseño, desarrollo y la aplicación de pruebas, así como el análisis y la divulgación de los resultados, a su vez se somete a las políticas, propósitos y usos de las evaluaciones que designa el Ministerio de Educación Nacional del país.

El Sistema Nacional de Evaluación Estandarizada del ICFES, evalúa las siguientes competencias en los estudiantes del Sistema educativo de Colombia, a través de sus pruebas:

- Saber 11:
  Lectura crítica
  Matemáticas
  Sociales y ciudadanas
  Ciencias naturales
  Inglés.

- Saber Pro:
  Lectura crítica
  Razonamiento cuantitativo
  Competencias ciudadanas
  Inglés.
  Comunicación escrita
  Módulos específicos (de acuerdo a la carrera profesional en curso)

7.2 Descripción de los Datos

Para este estudio se tienen las siguientes bases datos publicadas por el ICFES:

- Para las pruebas Saber 11 se tienen disponibles los datos de los periodos académicos desde el 2000 hasta el 2018.
- Para las pruebas Saber Pro, el ICFES tiene publicado los resultados desde el año 2006 hasta el año 2017. Es importante mencionar que no se han publicado los resultados de los años 2018 y 2019.
- Adicionalmente, se tiene una base de datos con 1,048,575 llaves que permiten comunicar los estudiantes de las pruebas Saber 11 de los años 2006-2018 con las pruebas Saber Pro de los años 2011-2017.

Como punto de partida, se cuenta con 1.048.575 de registros de estudiantes, para los cuales se tienen datos recolectados en aproximadamente 200 variables, que a su vez corresponden a dos momentos de la vida académica en Colombia, antes de terminar la educación media a través de la Prueba Saber 11 y ad-portas de culminar una carrera profesional de tipo universitaria por medio de las pruebas Saber Pro.

A continuación, se presenta el universo de los registros que se consideran insumos potenciales para esta investigación, basados en el número total de estudiantes; para quienes se cuente con llaves que comunica las variables recolectadas en las Pruebas Saber 11 y Saber Pro.

Tabla 2. Universo de registros potencialmente útiles.

		Saber Pro						
	Año	2012	2013	2014	2015	2016	2017	Total
Saber 11	2006	26,954	27,218	18,767	15,078	8,632	7,051	103,700
	2007	33,231	39,423	25,932	19,761	11,132	8,816	138,295
	2008	15,119	47,324	36,025	26,140	14,310	10,948	149,866
	2009	10,564	24,207	49,864	37,431	20,312	14,534	156,912
	2010	13,112	16,426	30,288	55,652	31,708	21,605	168,791
	2011	1,426	17,175	22,778	32,519	46,649	32,031	152,578
	2012	328	2,136	19,365	28,552	15,542	47,959	113,882
	2013	219	1,224	3,012	31,552	1,525	16,687	54,219
	2014	277	266	1,099	1,130	1,749	1,254	5,775
	2015	241	223	237	1,215	139	210	2,265
	2016	251	240	218	182	118	138	1,147
	2017	242	259	169	159	141	135	1,105
	2018	13	6	8	3	6	4	40
	Total	101,977	176,127	207,762	249,374	151,963	161,372	1,048,575

Cabe resaltar que en cada periodo se almacenaron variables diferentes, esto quiere decir que las bases de datos cambian de un periodo a otro, para seleccionar el conjunto de datos a evaluar en este trabajo, se identificaron aquellos periodos donde se almacenaron el mismo conjunto de variables, como se presenta en la siguiente tabla.

Tabla 3. Clasificación de cantidad de registros de estudiantes que comparten similitud entre las variables recolectadas.

		Saber Pro				
	Año-Periodo	2013	2014	2015	2016	2017
Saber 11	2008-2	28,920				
	2009-2		27,903			
	2010-2					
	2011-1					
	2011-2			33,106	140,325	
	2012-1					
	2012-2					
	Total	28,920	27,903	33,106	140,325	230,254

En la tabla se identifican algunos periodos donde se almacenaron las mismas variables para cada prueba, los cuales son:

- Saber 11 del año 2008 y saber Pro del año 2013.
- Saber 11 del año 2009 y saber Pro del año 2014.
- Saber 11 de los años 2010 - 2012 y saber Pro del año 2015.
- Saber 11 de los años 2010 - 2012 y saber Pro de los años 2016-2017.

Se decidió trabajar con el grupo de los datos que hacen referencia a los periodos más actuales y que a su vez cuentan con una gran cantidad de registros, equivalente a 140,325 estudiantes que presentaron la prueba Saber 11 a partir del semestre 2 del año 2010 hasta el mismo semestre del 2012 y para quienes los datos recogidos por la prueba Saber Pro pertenecen a los años 2016 y 2017. Esto corresponde al recuadro verde de la tabla anterior.

Los datos fueron integrados a través de las variables “11_estu_consecutivo” y “G_ESTU_CONSECUTIVO” las cuales contienen un código identificador para cada estudiante, comunicando las tablas de Saber 11 con las de Saber Pro.

En las siguientes secciones, se describen las variables a analizar, discriminando por aparte las que corresponden a las Pruebas Saber 11 y Saber Pro, a través de unas tablas en las que se toman en cuenta las siguientes convenciones:

- En la primera columna, se muestra un consecutivo ordenado de 1 a 183, para facilitar la posterior identificación de la variable a la que se hace alusión en el documento.

- La segunda columna, contiene el nombre de la variable con los prefijos “11” sí los datos fueron extraídos de las Pruebas Saber 11 y “G” o “E”, para los datos relacionados con las Pruebas Saber Pro, competencias genéricas o específicas respectivamente.

- En una tercera columna, se relaciona la explicación de la variable seguida (en filas) para algunos casos por un detalle de los atributos que dicha variable puede tomar.

- En cuarta posición, se relacionan las frecuencias de ocurrencia de los atributos de algunas de las variables dentro del set de datos.

- Por último, se distingue entre el tipo de variable sea string/cadena de caracteres o numérica.

Consideración importante: se consideran también como campos vacíos dentro de las variables, los representados por el texto “sin_info”.

7.2.1 Datos de las Pruebas Saber 11

A continuación, se describen los datos de las Pruebas Saber 11 agrupados según el contexto de los datos:

- La identificación de la prueba y el estudiante
- El colegio en el que estudió el inscrito
- La intención del estudiante en cuanto a la elección de una carrera profesional
- La presentación de la prueba
- El desempeño académico en la prueba
- Variables de tipo socioeconómico

7.2.1.1 La identificación de la prueba y el estudiante

Tabla 4. Descripción de variables Saber 11. Grupo 1.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
1	11_estu_exam_nombreexamen	Nombre y año del examen		string
2	11_periodo	Periodo de aplicación		numérica

		del examen, año seguido del semestre (solo el 5.25% de los registros corresponden al semestre 1)		
		2010	27.56%	
		2011	42.22%	
		2012	30.22%	
3	11_estu_consecutivo	Código consecutivo identificador del inscrito		string
4	11_estu_edad	edad		numérica
5	11_estu_tipo_documento	Tipo de documento de identidad del inscrito		string
		C- Cédula de ciudadanía	5.14%	
		E-Cédula de extranjería	0.04%	
		P-Pasaporte extranjero	0.01%	
		Q-Pasaporte colombiano	0.00%	
		R-Certificado de registraduría	1.37%	
		T-Tarjeta de identidad (se asume que mayoritariamente se habla de menores de edad)	93.43%	
		V-Por verificar	0.00%	
6	11_estu_pais_reside	Código del país de residencia del inscrito con codificación ISO 3166 alpha-2		string
		Colombia	99.96%	
		Otros países	0.04%	
7	11_estu_genero	Género del inscrito		string
		F-femenino	59.70%	
		M-masculino	39.96%	
		(en blanco)	0.34%	
8	11_estu_nacimiento_dia	fecha de nacimiento -dd-		numérica
9	11_estu_nacimiento_mes	fecha de nacimiento -mm		numérica
10	11_estu_nacimiento_anno	fecha de nacimiento -aaaa-		numérica
11	11_estu_cod_reside_mcpio	Código del municipio de residencia del inscrito		numérica
		1,096 valores diferentes (alta		

		cardinalidad)		
		(en blanco)	0.32%	
12	11_estu_reside_mcpio	Municipio de residencia del inscrito		string
		1,096 valores diferentes (alta cardinalidad)		
		(en blanco)	0.32%	
13	11_estu_reside_depto	Departamento de residencia del inscrito		string
		BOGOTÁ	22.85%	
		ANTIOQUIA	10.09%	
		VALLE	6.70%	
		CUNDINAMARCA	6.15%	
		SANTANDER	5.20%	
		ATLÁNTICO	5.15%	
		BOYACÁ	3.91%	
		CÓRDOBA	3.43%	
		BOLÍVAR	3.37%	
		NORTE SANTANDER	3.16%	
		(otros 23 departamentos, relacionados por separado)	29.65%	
		(en blanco)	0.32%	
14	11_estu_zona_reside	Zona de residencia del inscrito (puntos cardinales) 1- NORTE 2- ORIENTE 3- OCCIDENTE 4- SUR 5- CENTRO 6- NORORIENTE 7- SURORIENTE 8- NOROCCIDENTE 9-SUROCCIDENTE 10-UNICA (en blanco)	 6.7% 0.8% 3.6% 8.5% 3.8% 2.2% 2.4% 6.6% 4.5% 60.6% 0.3%	string
15	11_estu_area_reside	Área donde reside el inscrito		string
		Cabecera municipal	88.67%	
16	11_estu_trabaja	Trabaja actualmente (en el 95.59% de los casos los		string

		estudiantes no trabajan)		

7.2.1.2 El colegio en el que estudió el inscrito

Tabla 5. Descripción de variables Saber 11. Grupo 2.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
17	11_cole_cod_ICFES	Código interno del ICFES para los establecimientos educativos		numérica
18	11_cole_cod_DANE_institucion	Código DANE asignado al establecimiento educativo		numérica
		(en blanco)	0.22%	
19	11_cole_nombre_sede	Nombre de la sede educativa		string
20	11_cole_cod_mcpio_ubicacion	Código DANE del municipio donde está ubicada la Sede		numérica
21	11_cole_calendario	Calendario del establecimiento educativo		string
		A-Calendario A	91.95%	
		B-Calendario B	4.70%	
		F-Calendario flexible	3.35%	
22	11_cole_genero	Género de la población estudiantil que atiente el establecimiento educativo del inscrito		string
		M-Masculino	12.33%	
		F-Femenino	3.13%	
		X-Mixto	84.54%	
23	11_cole_naturaleza	Naturaleza del establecimiento educativo		string
		N-No oficial	42.27%	
		O-Oficial	57.73%	
24	11_cole_bilingue	El establecimiento educativo es bilingüe		numérica
		0-No	89.18%	
		1-Si	2.63%	
		(en blanco)	8.19%	

25	11_cole_jornada	Jornada del establecimiento educativo		string
		COMPLETA U ORDINARIA-Jornada completa u ordinaria	34.80%	
		MAÑANA-Jornada mañana	50.77%	
		NOCHE-Jornada noche	1.07%	
		SABTINA-DOMINICAL-Jornada sabatina-dominical	0.69%	
		TARDE-Jornada tarde	12.64%	
		UNICA-Jornada única	0.02%	
26	11_cole_caracter	Carácter del establecimiento educativo		string
		ACADEMICO	62.14%	
		ACADÉMICO Y TÉCNICO	17.39%	
		DESCONOCIDO	0.03%	
		NORMALISTA	3.78%	
		TÉCNICO	16.67%	
27	11_cole_valor_pension	Valor de la pensión pagada por el estudiante en el último año		string
		0-No paga pensión	54.71%	
		8-menos de 87,000 pesos	9.84%	
		9-entre 87,000 y menos de 120,000 pesos	5.46%	
		10-entre 120,000 y menos de 150,000 pesos	4.90%	
		11-entre 150,000 y menos de 250,000 pesos	11.45%	
		12-entre 250,000 pesos o más	13.26%	
		(en blanco)	0.39%	

7.2.1.3 La intención del estudiante en cuanto a la elección de una carrera profesional

Tabla 6. Descripción de variables Saber 11. Grupo 3.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
28	11_estu_ies_cod_d	Código de la institución de		numérica

	eseada	educación superior donde le gustaría estudiar		
		(alta cardinalidad, 294 códigos de institución diferentes)	10.34%	
		(en blanco)	89.66%	
29	11_estu_ies_deseada_nombre	Nombre de la institución de educación superior donde le gustaría estudiar		string
		(alta cardinalidad)	10.34%	
		(en blanco)	89.66%	
30	11_estu_ies_cod_mpio_deseada	Código del municipio de la institución de educación superior donde le gustaría estudiar		numérica
		(68 atributos diferentes están contenidos en esta variable)	10.34%	
		(en blanco)	89.66%	
31	11_estu_ies_mpio_deseada	Municipio de la institución de educación superior donde le gustaría estudiar		string
		(68 atributos diferentes están contenidos en esta variable)	10.34%	
		(en blanco)	89.66%	
32	11_estu_ies_dept_deseada	Departamento de la institución de educación superior donde le gustaría estudiar		string
		(29 atributos diferentes están contenidos en esta variable)	10.34%	
		(en blanco)	89.66%	
33	11_estu_carrdeseada_tipo	Tipo de carrera que desea estudiar el inscrito (ninguna, técnica, tecnológica, profesional)		string
		(en blanco)	90.32%	

7.2.1.4 La presentación de la prueba

Tabla 7. Descripción de variables Saber 11. Grupo 4.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
34	11_estu_exam_cod_mpio_presentacio	Código de municipio de presentación del examen		numérica
		(445 atributos diferentes están contenidos en esta variable)		
35	11_estu_mpio_presentacion	Municipio de presentación del examen		string
		(428 atributos diferentes están contenidos en esta variable)	100%	
36	11_estu_dept_presentacion	Departamento de presentación del examen		string
		(33 atributos diferentes están contenidos en esta variable)	100%	
37	11_estu_veces_estado	Veces que el estudiante ha presentado el examen de estado		numérica
		Presenta el examen por primera vez	82.71%	

7.2.1.5 El desempeño académico en la prueba

Tabla 8. Descripción de variables Saber 11. Grupo 5.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
38	11_punt_lenguaje	Puntaje del inscrito en lenguaje		numérica
39	11_punt_matematicas	Puntaje del inscrito en matemáticas		numérica
40	11_punt_c_sociales	Puntaje del inscrito en ciencias sociales		numérica
41	11_punt_filosofia	Puntaje del inscrito en filosofía		numérica
42	11_punt_biologia	Puntaje del inscrito en biología		numérica
43	11_punt_quimica	Puntaje del inscrito en química		numérica
44	11_punt_fisica	Puntaje del inscrito en física		numérica
45	11_punt_ingles	Puntaje del inscrito en inglés		numérica

46	11_desemp_ingles	Desempeño del inscrito en inglés según las bandas del Marco Común Europeo		string
		(A -) Nivel inferior	29.96%	
		A1 -Principiante	37.83%	
		A2 -Básico	16.63%	
		B1 -Pre -intermedio	11.85%	
		(B+) Supera al nivel B1	3.73%	
47	11_nombre_comp_flexible	Nombre del componente flexible (profundizaciones o interdisciplinar)		string
		MEDIO AMBIENTE	25.47%	
		PROFUNDIZACIÓN EN BIOLOGÍA	14.04%	
		PROFUNDIZACIÓN EN CIENCIAS SOCIALES	8.05%	
		PROFUNDIZACIÓN EN LENGUAJE	13.03%	
		PROFUNDIZACIÓN EN MATEMÁTICA	15.96%	
		VIOLENCIA Y SOCIEDAD	23.46%	
48	11_punt_comp_flexible	Puntaje del inscrito en componente flexible		numérica
49	11_desemp_comp_flexible	Desempeño del inscrito en componente flexible		string
		GB -Nivel de desempeño GB (grado básico)		
		I-Nivel de desempeño I		
		II -Nivel de desempeño II		
		III -Nivel de desempeño III		
		(en blanco)	48.93%	
50	11_estu_puesto	Puesto general del inscrito en el examen		numérica

7.2.1.6 Variables de tipo socioeconómico

Tabla 9. Descripción de variables Saber 11. Grupo 6.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
51	11_fami_nivel_sisben	Nivel de SISBEN en que está clasificada la familia		numérica
		1	21.80%	
		2	20.05%	
		3	5.64%	
		4	0.85%	
		5	51.60%	
		(en blanco)	0.06%	
52	11_fami_estrato_vivienda	Estrato socioeconómico de la residencia del estudiante según factura de energía		string
		1	18.92%	
		2	33.71%	
		3	29.24%	
		4, 5 o 6	15.96%	
		(en blanco)	2.17%	
53	NSE-11_ mensual fami_ing_fmiliar_ (esta variable es igual a "fami_ing_fmiliar_mensual-SMMLV)	Ingresos mensuales del hogar representado en salarios mínimos mensuales legales vigentes		string
		1	14.59%	
		2	37.23%	
		3	20.67%	
		4	13.98%	
		desde 5 hasta más de 10	13.46%	
		(en blanco)	0.07%	
54	11_fami_educa_padre	Nivel educativo del padre		string
55	11_fami_educa_madre	Nivel educativo de la madre		string
56	11_fami_ocup_padre	Ocupación del padre		string
57	11_fami_ocup_madre	Ocupación de la madre		string
58	11_fami_pisos_hog	Material de los pisos que		string

	ar	predomina en la vivienda		
59	11_fami_personas_hogar	Número de personas que conforman el hogar		string
60	11_fami_telefono_fijo	El hogar cuenta con servicio de teléfono fijo		string
61	11_fami_celular	Tiene servicio de teléfono móvil		string
62	11_fami_internet	El hogar cuenta con servicio de internet		string
63	11_fami_servicio_television	El hogar cuenta con servicio cerrado de televisión		string
64	11_fami_computador	Tiene computador en su hogar		string
65	11_fami_lavadora	El hogar cuenta con lavadora		string
66	11_fami_nevera	El hogar cuenta con nevera o enfriador		string
67	11_fami_horno	El hogar cuenta con horno eléctrico o a gas		string
68	11_fami_dvd	Cantidad de reproductores DVD con que cuenta su hogar		string
69	11_fami_microondas	El hogar cuenta con horno microondas		string
70	11_fami_automovil	Cantidad de automóviles particulares con que cuenta su hogar		string

7.2.2 Datos de las Pruebas Saber Pro

A continuación, se describen las variables relacionadas con las pruebas Saber Pro, agrupadas según la naturaleza de la información:

- La identificación de la prueba y el estudiante
- Identificación de minorías o condiciones especiales en el inscrito
- El colegio en donde estudió el inscrito
- La preparación para la prueba
- La elección vocacional del estudiante
- La presentación del examen y evaluación de competencias genéricas
- Variables de tipo socioeconómico
- Variables asociadas con el valor y pago de la matrícula de la carrera profesional
- La evaluación de competencias específicas

7.2.2.1 La identificación de la prueba y el estudiante

Tabla 10. Descripción de variables Saber Pro. Grupo 1.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
71	G_estu_tipodocumento	Tipo de Documento		string
		C- Cédula de ciudadanía	99.71%	
72	G_estu_nacionalidad	Nacionalidad		string
		Colombia	99.93%	
73	G_estu_genero	Género		string
		F-femenino	39.96%	
		M-masculino	60.04%	
74	G_estu_dia_nac	fecha de nacimiento -dd-		numérica
75	G_estu_mes_nac	fecha de nacimiento -mm-		numérica
76	G_estu_ano_nac	fecha de nacimiento -aaaa-		numérica
77	G_periodo	Periodo del examen		string
		20163 profesionales	53.81%	
		20173 profesionales	46.19%	
78	G_estu_consecutivo	Identificador estudiante		string
79	G_estu_depto_reside	Departamento de residencia		string
80	G_estu_cod_reside_depto	Código DANE del departamento de residencia		numérica
		99999 = extranjero		
81	G_estu_mcpio_reside	Municipio de residencia		string
82	G_estu_cod_mcpio_reside	Código DANE del municipio de residencia		numérica
		99999 = extranjero		
83	G_estu_areareside	Área de residencia		string
		Área Rural	5.08%	
		Cabecera Municipal	48.72%	
		(en blanco)	46.20%	
84	G_estu_estadocivil	Estado civil		string
		Casado	0.78%	

		Separado y/o viudo	0.08%	
		Soltero	51.26%	
		Unión libre	1.64%	
		(en blanco)	46.24%	
85	G_estu_tipodocumentosb11	Tipo de Documento del estudiante cuando presento las pruebas Saber 11		string
		(en blanco)	62.36%	

7.2.2.2 Identificación de minorías o condiciones especiales en el inscrito

Tabla 11. Descripción de variables Saber Pro. Grupo 2.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
86	G_estu_tieneetnia	El inscrito pertenece a alguna etnia		string
		Si	4.52%	
		No	95.48%	
87	G_estu_etnia	Detalle de la etnia		string
		(en blanco)	97.85%	
88	G_estu_limita_motriz	Limitación - motriz		string
		(en blanco)	99.97%	
89	G_estu_limita_invidente	Limitación - invidente		string
		(en blanco)	99.99%	
90	G_estu_limita_condicion Especial	Limitación - condición especial		string
		(en blanco)	100.00%	
91	G_estu_limita_sordo	Limitación - sordo		string
		(en blanco)	99.99%	
92	G_estu_limita_autismo	Limitación - autismo		string
		(en blanco)	100.00%	

7.2.2.3 El colegio en donde estudió el inscrito

Tabla 12. Descripción de variables Saber Pro. Grupo 3.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
93	G_estu_tituloobtenidobachiller	Título de bachiller obtenido		string
		Bachiller académico	72.87%	
		Bachiller pedagógico o normalista	3.00%	
		Bachiller técnico	24.02%	
		(en blanco)	0.11%	
94	G_estu_cole_termino	Nombre del colegio donde terminó bachillerato		string
		(en blanco)	30.35%	
95	G_estu_coddane_cole_termino	Código DANE del colegio donde terminó bachillerato		numérica
		(en blanco)	30.35%	
96	G_estu_cod_cole_mcpio_termino	Código DANE del municipio donde está ubicado el colegio donde terminó bachillerato		numérica
		(en blanco)	30.35%	
97	G_estu_otrocole_termino	Pregunta de respuesta abierta para escribir el nombre del colegio donde terminó bachillerato		string
		(en blanco)	79.41%	

7.2.2.4 La preparación para la prueba

Tabla 13. Descripción de variables Saber Pro. Grupo 4.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
98	G_estu_comocapacitoexamensb11	¿Cómo se preparó para el examen SABER 11?		string
		No realizó ninguna prueba de preparación	23.40%	

		Repasó por cuenta propia	67.24%	
		Tomó un curso de preparación	9.36%	
		(en blanco)	0.01%	
99	G_estu_cursodocentesies	Se preparó para el examen Saber Pro en su IES con docentes de la institución (número de horas)		string
		Menos de 20 horas	3.33%	
		Entre 20 y 30 horas	2.42%	
		Más de 30 horas	3.29%	
		(en cero o blanco)	90.96%	
100	G_estu_cursoiesapoyoexterno	Se preparó para el examen Saber Pro en un curso organizado por la institución con apoyo de un instituto de preparación de exámenes externos (número de horas)		string
		Menos de 20 horas		
		Entre 20 y 30 horas		
		Más de 30 horas		
		(en cero o blanco)	90.96%	
101	G_estu_cursoiesexterna	Se preparó para el examen Saber Pro en un instituto de preparación de exámenes (número de horas)		string
		Menos de 20 horas		
		Entre 20 y 30 horas		
		Más de 30 horas		
		(en cero o blanco)	90.96%	
102	G_estu_simulacrotipoicfes	¿Qué actividades desarrolló en el curso de preparación para el examen Saber Pro?: Simulacros con preguntas tipo ICFES		string
		No		
		Si		
		(en cero o blanco)	90.96%	

103	G_estu_actividadrefuerzoareas	¿Qué actividades desarrolló en el curso de preparación para el examen Saber Pro?: Clases de refuerzo en algunas áreas		string
		No		
		Si		
		(en cero o blanco)	90.96%	
104	G_estu_actividadrefuerzogeneric	¿Qué actividades desarrolló en el curso de preparación para el examen Saber Pro?, refuerzo en desarrollo de competencias genéricas		string
		No		
		Si		
		(en blanco)	90.96%	
		(en blanco)	62.36%	

7.2.2.5 La elección vocacional del estudiante

Tabla 14. Descripción de variables Saber Pro. Grupo 5.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
105	G_inst_cod_institucion	Código de la Institución de Educación Superior		numérica
106	G_inst_nombre_instituc	Nombre de la Institución de Educación Superior		string
107	G_estu_semestrecursa	Semestre que cursa actualmente el estudiante	0.05%	string
		Valores entre 0 y 8	20.30%	
		9	32.30%	
		10	39.23%	
		11 en adelante	8.12%	
		(en blanco)	0.05%	
108	G_estu_prgm_academico	Nombre del programa académico que estudia		string
109	G_estu_snies_prgm	Código SNIES del programa académico que estudia		numérica

	Académico			
110	G_cod-grup-ref-son17	Código creado para identificar el grupo de referencia al que pertenece el programa académico del estudiante		numérica
		6013	18.94%	
		202	0.16%	
		101	1.40%	
		8017	2.56%	
		508	3.25%	
		509	0.19%	
		6014	7.11%	
		5010	11.09%	
		6015	2.02%	
		303	10.70%	
		405	2.19%	
		5011	0.23%	
		7016	28.52%	
		406	3.48%	
		304	0.31%	
		5012	2.13%	
		407	5.72%	
111	G_gruporeferencia	Nombre del Grupo de Referencia al que pertenece el programa académico del estudiante		string
		ADMINISTRACIÓN Y AFINES	18.94%	
		BELLAS ARTES Y DISEÑO	0.16%	
		CIENCIAS AGROPECUARIAS	1.40%	
		CIENCIAS NATURALES Y EXACTAS	2.56%	
		CIENCIAS SOCIALES	3.25%	
		COMUNICACIÓN, PERIODISMO Y PUBLICIDAD	0.19%	
		CONTADURÍA Y AFINES	7.11%	
		DERECHO	11.09%	
		ECONOMIA	2.02%	
		EDUCACIÓN	10.70%	

		ENFERMERÍA HUMANIDADES INGENIERÍA MEDICINA NORMALES SUPERIORES PSICOLOGÍA SALUD	2.19% 0.23% 28.52% 3.48% 0.31% 2.13% 5.72%	
112	G_estu_prgm_cod municipio	Código del municipio donde se ofrece el programa académico		numérica
113	G_estu_prgm_muni cipio	Nombre del municipio donde se ofrece programa académico		string
114	G_estu_prgm_ departamento	Nombre del departamento donde se ofrece el programa académico		string
115	G_estu_nivel_prgm _ academico	Nivel del programa académico		string
116	G_estu_metodo_pr gm	Metodología del programa académico		string
117	G_estu_nucleo_pre grado	Nombre del núcleo de pregrado al que pertenece el programa académico		string
118	G_estu_inst_codm unicipio	Código del municipio donde está ubicada la IES		numérica
119	G_estu_inst_munici pio	Nombre del municipio donde está ubicada la IES		string
120	G_estu_inst_depart amento	Nombre del departamento donde está ubicada la IES		string
121	G_inst_caracter_ac ademico	Carácter académico de la IES		string
122	G_inst_origen	Naturaleza u origen de la IES		string

7.2.2.6 La presentación del examen y evaluación de competencias genéricas

Tabla 15. Descripción de variables Saber Pro. Grupo 6.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del	Tipo de variable

			atributo	
123	G_estu_cod_mcpio_presentacion	Código DANE del municipio presentación del examen		numérica
124	G_estu_mcpio_presentacion	Municipio de presentación del examen		string
125	G_estu_depto_presentacion	Código DANE del departamento del municipio de presentación del examen		string
126	G_estu_cod_depto_presentacion	Código del departamento del municipio de presentación del examen		numérica
127	G_razona_cuant_punt	Puntaje razonamiento cuantitativo, rango [0, 300]		numérica
128	G_razona_cuant_desem	Nivel de desempeño razonamiento cuantitativo, rango [1, 4]		numérica
129	G_razona_cuant_pnal	Percentil nacional razonamiento cuantitativo, rango [1, 100]		numérica
130	G_razona_cuant_pgref	Percentil por grupo de referencia razonamiento cuantitativo, rango [1, 100]		numérica
131	G_lect_critica_punt	Puntaje lectura crítica, rango [0, 300]		numérica
132	G_lect_critica_desem	Nivel de desempeño lectura crítica, rango [1, 4]		numérica
133	G_lect_critica_pnal	Percentil nacional lectura crítica, rango [1, 100]		numérica
134	G_lect_critica_pgref	Percentil por grupo de referencia lectura crítica, rango [1, 100]		numérica
135	G_compet_ciudad_punt	Puntaje competencias ciudadanas, rango [0, 300]		numérica
136	G_compet_ciudad_desem	Nivel de desempeño competencias ciudadanas, rango [1, 4]		numérica
137	G_compet_ciudad_pnal	Percentil nacional competencias ciudadanas,		numérica

		rango [1, 100]		
138	G_compet_ciudad_pgref	Percentil por grupo de referencia competencias ciudadanas, rango [1, 100]		numérica
139	G_ingles_punt	Puntaje inglés, rango [0, 300]		numérica
140	G_ingles_desem	Nivel de desempeño inglés		string
141	G_ingles_pnal	Percentil nacional inglés, rango [1, 100]		numérica
142	G_ingles_pgref	Percentil por grupo de referencia inglés, rango [1, 100]		numérica
143	G_com_escrita_punt	Puntaje comunicación escrita, rango [0, 300]	0.16%	numérica
144	G_com_escrita_desem	Nivel de desempeño comunicación escrita, rango [1, 4]	0.16%	numérica
145	G_com_escrita_pnal	Percentil nacional comunicación escrita, rango [1, 100]	0.16%	numérica
146	G_com_escrita_pgref	Percentil por grupo de referencia comunicación escrita, rango [1, 100]	0.16%	numérica
147	G_punt_global	Puntaje total obtenido		numérica
148	G_percentil_global	Percentil global en que se encuentra el evaluado, rango [1, 100]		numérica
149	G_estu_estadoinvestig	Identifica los usuarios que están en proceso de investigación en el ICFES		string
		publicar	99.94%	

7.2.2.7 Variables de tipo socioeconómico

Tabla 16. Descripción de variables Saber Pro. Grupo 7.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
150	G_fami_hogaractual	El hogar actual donde vive es permanente o temporal		string

		Permanente	79.04%	
		Temporal por estudio u otros	20.95%	
		(en blanco)	0.00%	
151	G_fami_cabezafamilia	El inscrito es jefe de hogar o cabeza de familia		string
		No	95.73%	
152	G_fami_numpersonasa cargo	Cuántas personas dependen económicamente del inscrito		string
		Ninguna	87%	
153	G_fami_educacion padre	Nivel educativo más alto alcanzado por el padre		string
154	G_fami_educacion madre	Nivel educativo más alto alcanzado por la madre		string
155	G_fami_ocupacion padre	Ocupación u oficio del padre		string
156	G_fami_ocupacion madre	Ocupación u oficio de la madre		string
157	nse-G_fami_estratovivienda	Estrato socioeconómico de la residencia del estudiante según factura de energía		numérica
		Valores entre 0 y 1	15.40%	
		2	31.56%	
		3	32.83%	
		Valores entre 4, 5 y 6	20.21%	
		(en blanco)	0.00%	
158	nse-G_fami_personas hogar	Cuántas personas conforman el hogar donde vive actualmente el inscrito, incluyéndolo a él		string
		Una	2.22%	
		Dos	10.58%	
		Tres	23.77%	
		Cuatro	32.52%	
		Cinco	18.93%	
		Seis	6.93%	
		Siete	2.60%	
		Ocho	1.24%	
		Nueve	0.45%	
		Diez	0.30%	

		Once	0.12%	
		Doce o mas	0.33%	
		(en blanco)	0.00%	
159	nse-G_fami_cuartos hogar	En total, ¿en cuántos cuartos duermen las personas de su hogar?		string
		Uno	5.26%	
		Dos	29.12%	
		Tres	46.71%	
		Cuatro	13.95%	
		Cinco	3.34%	
		Seis	0.89%	
		Siete	0.32%	
		Ocho	0.15%	
		Nueve	0.06%	
		Diez o mas	0.19%	
		(en blanco)	0.00%	
160	nse-G_fami_tieneinternet	El hogar cuenta con servicio o conexión a internet		string
		Si	86.34%	
161	nse-G_fami_tieneservicio tv	El hogar de inscrito cuenta con servicio cerrado de televisión		string
		Si	82.52%	
162	nse-G_fami_tiene computador	El hogar del inscrito cuenta con un computador		string
		Si	92.97%	
163	nse-G_fami_tiene lavadora	El hogar del inscrito cuenta con una máquina para lavar ropa		string
		Si	76.75%	
164	nse-G_fami_tienehorno microogas	El hogar del inscrito cuenta con horno microondas u horno eléctrico o a gas		string
165	nse-G_fami_tiene automovil	El hogar del inscrito cuenta con un automóvil particular		string
166	nse-G_fami_tiene motocicleta	El hogar del inscrito cuenta con una motocicleta		string

167	G_fami_numlibros	¿Cuántos libros físicos o electrónicos hay en su hogar excluyendo periódicos, revistas, directorios telefónicos y libros del colegio?		string
168	G_estu_dedicacion lecturadiaria	Usualmente, ¿cuánto tiempo al día dedica a leer por entretenimiento?		string
169	G_estu_dedicacion internet	Usualmente, ¿cuánto tiempo al día dedica a navegar en internet? Excluya actividades académicas		string
170	G_estu_horassemana trabaja	¿Cuántas horas trabajó usted durante la semana pasada?		string
171	G_estu_tiporemuneracion	¿Usted recibe algún tipo de remuneración por trabajar?		string
172	G_estu_inse_individual	Índice socioeconómico a nivel de estudiante, rango [0: 8,171,244,547]		numérica
173	G_estu_nse_individual	Nivel socioeconómico a nivel de estudiante		string
		NSE 1	19.04%	
		NSE 2	29.56%	
		NSE 3	25.74%	
		NSE 4	22.94%	
		(en blanco)	2.73%	

7.2.2.8 Variables asociadas con el valor y pago de la matrícula de la carrera profesional

Tabla 17. Descripción de variables Saber Pro. Grupo 8.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
174	G_estu_valormatricula universidad	Valor de la matrícula del último semestre cursado (sin considerar descuentos o becas)		string
		No pagó matrícula	0.59%	

		Menos de 500 mil	18.22%	
		Entre 500 mil y menos de 1 millón	12.74%	
		Entre 1 millón y menos de 2.5 millones	20.83%	
		Entre 2.5 millones y menos de 4 millones	19.94%	
		Entre 4 millones y menos de 5.5 millones	11.43%	
		Entre 5.5 millones y menos de 7 millones	5.83%	
		Más de 7 millones	10.37%	
		(en blanco)	0.05%	
175	G_estu_pagomatricula beca	El pago de matrícula es por beca		string
		Si	15.66%	
		No	84.17%	
		(en blanco)	0.17%	
176	G_estu_pagomatricula credito	El pago de matrícula es mediante crédito		string
		No	69.86%	
		Si	29.99%	
		(en blanco)	0.16%	
177	G_estu_pagomatricula padres	El pago de matrícula lo realizan los padres del estudiante		string
		No	26.12%	
		Si	73.75%	
		(en blanco)	0.13%	
178	G_estu_pagomatricula propio	El pago de matrícula es por recursos propios		string
		No	74.71%	
		Si	25.13%	
		(en blanco)	0.17%	
		No realizó ninguna prueba de preparación	23.40%	
		Repasó por cuenta propia	67.24%	
		Tomó un curso de preparación	9.36%	
		(en blanco)	0.01%	

7.2.2.9 La evaluación de competencias específicas

Tabla 18. Descripción de variables Saber Pro. Grupo 9.

No.	Variable	Explicación de la variable seguida por sus atributos	Frecuencia del atributo	Tipo de variable
179	E_estu_consecutivo	Id público del estudiante en Saber Pro		string
180	E_result_codigoprueba	Código de la prueba específica		numérica
		2001	2.02%	
		4001	11.41%	
		2912	4.79%	
		8001	1.26%	
		9001	2.57%	
		1002	1.90%	
		2611	6.73%	
		6004	11.05%	
		5002	11.76%	
		8002	16.72%	
		1003	9.10%	
		3003	3.43%	
		4003	2.40%	
		4005	11.09%	
		6003	1.21%	
		7003	0.33%	
		8003	0.22%	
		9003	0.29%	
		1004	0.39%	
		9004	0.46%	
		1005	0.88%	
181	E_result_nombreprueba	Nombre de la prueba específica		string
		ANÁLISIS ECONÓMICO	2.02%	
		ATENCIÓN EN SALUD	11.41%	
		DISEÑO DE OBRAS DE INFRAESTRUCTURA	4.79%	
		DISEÑO DE PROCESOS INDUSTRIALES	1.26%	
		DISEÑO DE SISTEMAS DE CONTROL	2.57%	

		DISEÑO DE SISTEMAS MECÁNICOS	1.90%	
		DISEÑO DE SISTEMAS PRODUCTIVOS Y LOGÍSTICOS	6.73%	
		ENSEÑAR	11.05%	
		FORMULACIÓN DE PROYECTOS DE INGENIERÍA	11.76%	
		GESTIÓN DE ORGANIZACIONES	16.72%	
		GESTIÓN FINANCIERA	9.10%	
		INTERVENCIÓN EN PROCESOS SOCIALES	3.43%	
		INVESTIGACIÓN EN CIENCIAS SOCIALES	2.40%	
		INVESTIGACIÓN JURÍDICA	11.09%	
		PENSAMIENTO CIENTÍFICO - CIENCIAS BIOLÓGICAS	1.21%	
		PENSAMIENTO CIENTÍFICO - CIENCIAS DE LA TIERRA	0.33%	
		PENSAMIENTO CIENTÍFICO - CIENCIAS FÍSICAS	0.22%	
		PENSAMIENTO CIENTÍFICO - MATEMÁTICAS Y ESTADÍSTICA	0.29%	
		PENSAMIENTO CIENTÍFICO - QUÍMICA	0.39%	
		PRODUCCIÓN AGRICOLA	0.46%	
		PRODUCCIÓN PECUARIA	0.88%	
182	E_result_puntaje	Puntaje de la prueba específica, rango [0, 300]		numérica
183	E_result_desem	Nivel de Desempeño, rango [1, 6]		numérica

8. PREPARACIÓN DE LOS DATOS

8.1 Resumen de los datos a analizar

A continuación, se presenta un resumen de las variables a analizar, la cuales fueron descritas en el capítulo anterior.

Tabla 19. Resumen de variables Saber 11.

Agrupadas según su relación con	**Rango de variables que representan a este ítem**
La identificación de la prueba y el estudiante	1-16
El colegio en donde estudió el inscrito	17-27
La intención del estudiante en cuanto a la elección de una carrera profesional	28-33
La presentación de la prueba	34-37
El desempeño académico en la prueba	38-50
Variables de tipo socioeconómico	51-70

Tabla 20. Resumen de variables Saber Pro.

Agrupadas según su relación con	Rango de variables que representan a este ítem
La identificación de la prueba y el estudiante	71-85
La identificación de minorías o condiciones especiales en el inscrito	86-92
El colegio en el que estudió el inscrito	93-97
La preparación para la prueba	98-104
La elección vocacional del estudiante	105-122
La presentación del examen y evaluación de competencias genéricas	123-149
Variables de tipo socioeconómico	150-173
Valor y pago de la matrícula	174-178
Evaluación de competencias específicas	179-183

8.2 Reducción de variables irrelevantes o redundantes

A continuación, se relacionan las variables que se eliminan en esta etapa del estudio, las explicaciones se encuentran a lo largo de este apartado en la

columna No. 5 de cada tabla “Razón para eliminar la variable”, haciendo alusión a las variables identificadas por un consecutivo del 1 al 183, asignado a cada dato en el capítulo 7 de este documento.

8.2.1 Datos de las pruebas Saber 11

Las siguientes 21 variables fueron desestimadas de entrada en esta investigación, relacionadas con diversos factores.

En la tabla 21, las variables eliminadas relacionas con la identificación de la prueba y del estudiante.

Tabla 21. Variables eliminadas en el paso 1, Saber 11. Grupo 1.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
1	11_estu_exam_nombreexamen	Nombre y año del examen (datos contenidos en otra variable)	string	Contenida en la variable No. 2 “Periodo de aplicación del examen”
3	11_estu_consecutivo	Código consecutivo identificador del inscrito	string	Código único
5	11_estu_tipo_documento	Tipo de documento de identidad del inscrito	string	Se considera representada en la variable No. 4 “Edad”, dado que Cédula de Ciudadanía y Tarjeta de Identidad suman el 98.57% de frecuencia en el dato.
6	11_estu_pais_reside	Código del país de residencia del inscrito con codificación ISO 3166 alpha-2	String	En el 99.96% de los casos los estudiantes residen en Colombia.
8	11_estu_nacimie	fecha de	numérica	Se deja la variable

	nto_dia	nacimiento -dd-		No. 4 “Edad” en reemplazo de estas
9	11_estu_nacimiento_mes	fecha de nacimiento -mm	numérica	
10	11_estu_nacimiento_anno	fecha de nacimiento -aaaa-	numérica	
11	11_estu_cod_reside_mcpio	Código del municipio de residencia del inscrito	numérica	Se deja la variable No. 36 “Departamento de presentación del examen”
		1,096 valores diferentes (alta cardinalidad)		
12	11_estu_reside_mcpio	Municipio de residencia del inscrito	string	
		1,096 valores diferentes (alta cardinalidad)		
13	11_estu_reside_depto	Departamento de residencia del inscrito	string	
14	11_estu_zona_reside	Zona de residencia del inscrito	string	Se deja la variable No. 15 “Área donde reside el inscrito”

En la tabla 22, las variables eliminadas relacionas con el colegio en el que estudió el inscrito

Tabla 22. Variables eliminadas en el paso 1, Saber 11. Grupo 2.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
17	11_cole_cod_ICFES	Código interno del ICFES para los establecimientos educativos	numérica	Identificador único para cada institución
18	11_cole_cod_DANE_institucion	Código DANE asignado al establecimiento educativo	numérica	

19	11_cole_nombre_sede	Nombre de la sede educativa	string	No aporta al proyecto
20	11_cole_cod_mcpio_ubicacion	Código DANE del municipio donde está ubicada la Sede	numérica	Se deja la variable No. 36 “Departamento de presentación del examen”

En la tabla 23, las variables eliminadas relacionas con la intención del estudiante en cuanto a la elección de una carrera profesional

Tabla 23. Variables eliminadas en el paso 1, Saber 11. Grupo 3.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
28	11_estu_ies_cod_deseada	Código de la institución de educación superior donde le gustaría estudiar (alta cardinalidad, 294 códigos de institución diferentes)	numérica	No aporta al proyecto
29	11_estu_ies_deseada_nombre	Nombre de la institución de educación superior donde le gustaría estudiar	string	No aporta al proyecto
30	11_estu_ies_cod_mpio_deseada	Código del municipio de la institución de educación superior donde le gustaría estudiar	numérica	Se deja la variable No. 32 “Departamento de la institución de educación superior donde le gustaría estudiar”
31	11_estu_ies_mpio_deseada	Municipio de la institución de educación superior donde le gustaría estudiar	string	

En la tabla 24, las variables eliminadas relacionas con la presentación de la prueba.

Tabla 24. Variables eliminadas en el paso 1, Saber 11. Grupo 4.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
34	11_estu_exam_cod_mpio_presentacio	Código de municipio de presentación del examen	numérica	Se deja la variable No. 36 “Departamento de presentación del examen”
35	11_estu_mpio_presentacion	Municipio de presentación del examen	string	

Para el caso de las variables agrupadas por su relación con el desempeño académico en la prueba, no se eliminaron variables de esa sección.

También se decidió trabajar inicialmente con todas las variables de tipo socioeconómico.

8.2.2 Datos de las pruebas Saber Pro

De este grupo se descartan de entrada 29 variables clasificadas según diferentes factores.

En la tabla 25, las variables eliminadas relacionas con la identificación de la prueba y el estudiante

Tabla 25. Variables eliminadas en el paso 1, Saber Pro. Grupo 1.

No.	Variable	Explicación de la variable seguida por sus atributos	Razón para eliminar la variable
72	G_estu_nacionalidad	Nacionalidad	El dato identifica a los estudiantes como colombianos en un valor cercano al 100%
78	G_estu_consecutivo	Identificador estudiante	Código único
79	G_estu_depto_reside	Departamento de residencia	Se deja la variable No. 114 “Departamento donde se ofrece

80	G_estu_cod_reside_depto	Código DANE del departamento de residencia	el programa académico”
81	G_estu_mcpio_reside	Municipio de residencia	
82	G_estu_cod_mcpio_reside	Código DANE del municipio de residencia	
85	G_estu_tipodocumentosb11	Tipo de Documento del estudiante cuando presentó las pruebas Saber 11	Se considera suficiente con conocer la edad del inscrito al momento de presentar las dos pruebas.

Para el caso de las variables agrupadas en la Identificación de minorías o condiciones especiales en el inscrito, todas se conservan.

En la tabla 26, las variables eliminadas relacionas con el colegio en donde estudió el inscrito.

Tabla 26. Variables eliminadas en el paso 1, Saber Pro. Grupo 3.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
94	G_estu_cole_termino	Nombre del colegio donde terminó bachillerato	string	No aporta al proyecto
95	G_estu_coddane_cole_termino	Código DANE del colegio donde terminó bachillerato	numérica	
96	G_estu_cod_cole_mcpio_termino	Código DANE del municipio donde está ubicado el colegio donde terminó bachillerato	numérica	
97	G_estu_otrocole_termino	Pregunta de respuesta abierta para escribir el nombre del colegio	string	

		donde terminó bachillerato		

También se mantienen todos los datos relacionados con la forma como el estudiante se preparó para la prueba.

En la tabla 27, las variables eliminadas relacionas con la elección vocacional del estudiante:

Tabla 27. Variables eliminadas en el paso 1, Saber Pro. Grupo 5.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
105	G_inst_cod_institucion	Código de la Institución de Educación Superior	numérica	No aporta al proyecto
106	G_inst_nombre_instituc	Nombre de la Institución de Educación Superior	string	
108	G_estu_prgm_academico	Nombre del programa académico que estudia	string	Alta cardinalidad, variable contenida dentro de la No. 111 “Nombre del Grupo de Referencia al que pertenece el programa académico del estudiante”
109	G_estu_snies_prgm academico	Código SNIES del programa académico que estudia	numérica	
112	G_estu_prgm_codmunicipio	Código del municipio donde se ofrece el programa académico	numérica	Se deja la variable No. 114 “Departamento donde se ofrece el programa académico”
113	G_estu_prgm_municipio	Nombre del municipio donde se ofrece programa académico	string	
117	G_estu_nucleo_pregrado	Nombre del núcleo de pregrado al que pertenece el programa	string	Alta cardinalidad 56 valores diferentes, variable representada en

		académico		“111-G_gruporeferencia”
118	G_estu_inst_codmunicipio	Código del municipio donde está ubicada la IES	numérica	Se deja la variable No. 114 “Departamento donde se ofrece el programa académico”
119	G_estu_inst_municipio	Nombre del municipio donde está ubicada la IES	string	
120	G_estu_inst_departamento	Nombre del departamento donde está ubicada la IES	string	

En la tabla 28, las variables eliminadas relacionas con la presentación del examen y evaluación de competencias genéricas.

Tabla 28. Variables eliminadas en el paso 1, Saber Pro. Grupo 6.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
123	G_estu_cod_mcpio_ presentacion	Código DANE del municipio presentación del examen	numérica	No aporta al proyecto
124	G_estu_mcpio_ presentacion	Municipio de presentación del examen	string	Se deja la variable No. 114 “Departamento donde se ofrece el programa académico”
125	G_estu_depto_ presentacion	Departamento del municipio de presentación del examen	string	
126	G_estu_cod_depto_ presentacion	Código del departamento del municipio de presentación del examen	numérica	
149	G_estu_estadoinvestig	Identifica los usuarios que están en proceso de investigación en el	String	La variable revela que en un valor cercano al 100% los estudiantes tienen el

		ICFES		estatus “publicar”

En la tabla 29, las variables eliminadas relacionas con la información de tipo socioeconómico.

Tabla 29. Variables eliminadas en el paso 1, Saber Pro. Grupo 7.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
172	G_estu_inse_individual	Índice socioeconómico a nivel de estudiante, rango [0: 8,171,244,547]	numérica	Contenida en la variable No. 173 “Nivel socioeconómico a nivel de estudiante”

Para el caso de las variables asociadas con el pago de la matrícula de la carrera profesional, todas se conservan.

Finalizando con las variables que se descartan en lo relacionado con la evaluación de competencias específicas, para completar las 30 variables eliminadas de entrada de las Pruebas Saber Pro.

Tabla 30. Variables eliminadas en el paso 1, Saber Pro. Grupo 9.

No.	Variable	Explicación de la variable seguida por sus atributos	Tipo de variable	Razón para eliminar la variable
179	E_estu_consecutivo	Id público del estudiante en Saber Pro	string	Identificador del estudiante
180	E_result_codigo prueba	Código de la prueba específica	numérica	Contenida en la variable No. 181 “Nombre de la prueba específica”

Al finalizar este paso, se conservan 49 de las 70 variables de las pruebas Saber 11 y 84 de las 113 variables de las pruebas saber pro, para un global de 133 variables que permanecen.

8.3 Descripción estadística de los datos

A continuación, se relaciona información que permite una exploración estadística de las variables relacionadas principalmente con el desempeño académico del inscrito en las Pruebas Saber 11 y Saber Pro, para ir al detalle de este análisis con todas las variables, favor remitirse al:

Anexo No. 1. Exploración estadística de los datos, disponible al final de este documento.

Con el objetivo de facilitar la citación de las variables de acá en adelante, se usará el número de la variable asignado a cada una de estas desde el capítulo 7 de este documento seguido por su nombre.

Cabe resaltar que todos los datos cuya fuente de origen son las Pruebas Saber 11, tienen el prefijo “11” en su nombre de variable, a su vez los prefijos “G” o “E”, están reservados para los nombres de las variables asociadas a las Pruebas Saber Pro; competencias genéricas o específicas respectivamente.

8.3.1 Variables de las Pruebas Saber 11

Variable: 4-11_estu_edad, tipo: float64.

Edad del inscrito

Tabla 31. Descripción estadística, variable No. 4.

Cuantiles		Estadística descriptiva	
Mínimo	9	Desviación estándar	1.71
5-th percentil	15	Coef. de variación	0.11
Q1	16	Kurtosis	152.06
Mediana	16	Promedio	16.29
Q3	17	MAD	0.77
95-th percentil	18	Skewness	9.82
Máximo	63	Sum	2285721
Range	54	Varianza	2.94
Rango intercuartil	1		

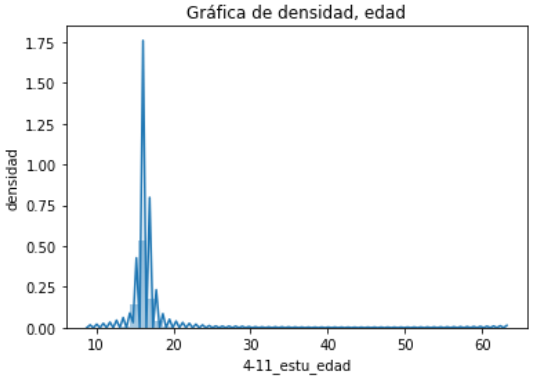


Figura 4. Gráfica de densidad, variable No. 4.

Variable: 38-11_punt_lenguaje, tipo: float64

Puntaje del inscrito en lenguaje

Tabla 32. Descripción estadística, variable No. 38.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	9.31
5-th percentil	39	Coef. de variación	0.17
Q1	47.41	Kurtosis	0.88
Mediana	53.5	Promedio	53.56
Q3	59	MAD	7.14
95-th percentil	69	Skewness	0.09
Máximo	102.62	Sum	7515900
Range	102.62	Varianza	86.59
Rango intercuartil	11.59		

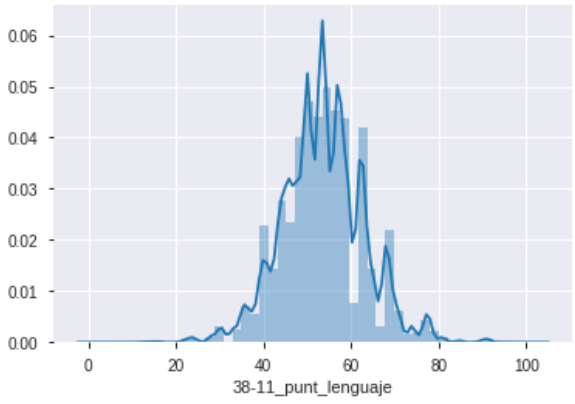


Figura 5. Gráfica de densidad, variable No. 38.

Variable: 39-11_punt_matematicas, tipo: float64

Puntaje del inscrito en matemáticas

Tabla 33. Descripción estadística, variable No. 39.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	11.82
5-th percentil	38	Coef. de variación	0.21
Q1	48	Kurtosis	1.17
Mediana	55	Promedio	55.15
Q3	63	MAD	9.10
95-th percentil	75.41	Skewness	0.44
Máximo	126	Sum	7739100
Range	126	Varianza	139.72
Rango intercuartil	15		

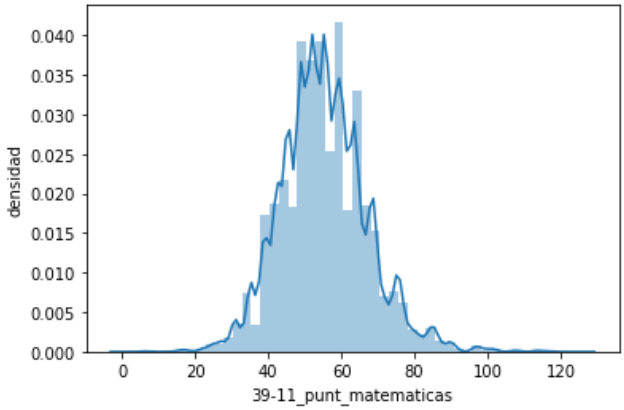


Figura 6. Gráfica de densidad, variable No. 39.

Variable: 40-11_punt_c_sociales, tipo: float64

Puntaje del inscrito en ciencias sociales

Tabla 34. Descripción estadística, variable No. 40.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	9.65
5-th percentil	37	Coef. de variación	0.18
Q1	46	Kurtosis	0.33
Mediana	53	Promedio	52.73
Q3	59.87	MAD	7.67
95-th percentil	68	Skewness	0.03
Máximo	107	Sum	7399900
Range	107	Varianza	93.05
Rango intercuartil	13.87		

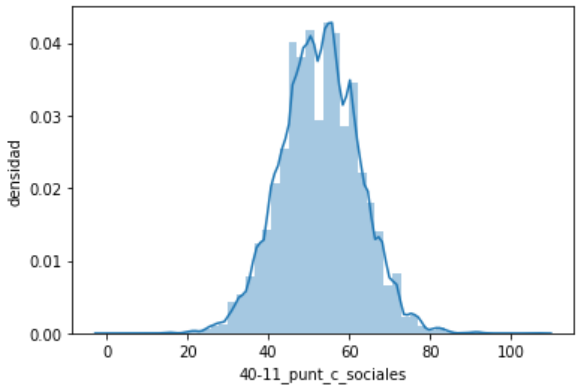


Figura 7. Gráfica de densidad, variable No. 40.

Variable: 41-11_punt_filosofia, tipo: float64

Puntaje del inscrito en filosofía

Tabla 35. Descripción estadística, variable No. 41.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	10.7
5-th percentil	32	Coef. de variación	0.22
Q1	42	Kurtosis	0.82
Mediana	49	Promedio	49
Q3	55	MAD	8.23
95-th percentil	68.09	Skewness	-0.03
Máximo	103	Sum	6875200
Range	103	Varianza	114.48
Rango intercuartil	13		

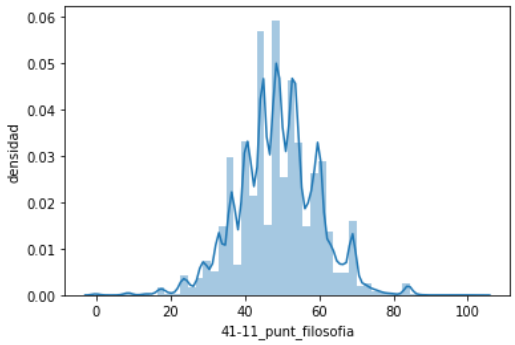


Figura 8. Gráfica de densidad, variable No. 41.

Variable: 42-11_punt_biologia, tipo: float64

Puntaje del inscrito en biología

Tabla 36. Descripción estadística, variable No. 42.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	9.54
5-th percentil	37.43	Coef. de variación	0.18
Q1	46.57	Kurtosis	0.88
Mediana	52	Promedio	52.62
Q3	58.6	MAD	7.45
95-th percentil	68	Skewness	0.18
Máximo	105.78	Sum	7383500
Range	105.78	Varianza	90.92
Rango intercuartil	12.03		

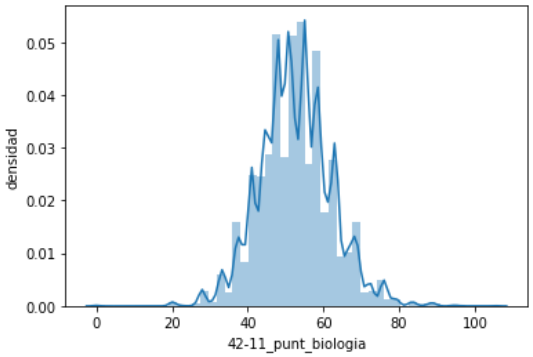


Figura 9. Gráfica de densidad, variable No. 42.

Variable: 43-11_punt_quimica, tipo: float64

Puntaje del inscrito en química

Tabla 37. Descripción estadística, variable No. 43.

Cuantiles		Estadística descriptiva	
Mínimo	0	Coef. de variación	0.18
5-th percentil	38	Kurtosis	2.39
Q1	46.19	Promedio	52.07
Mediana	51.58	MAD	6.90
Q3	57	Skewness	0.73
95-th percentil	67	Sum	7307200
Máximo	118.8	Varianza	84.22
Range	118.8		
Rango intercuartil	10.81		

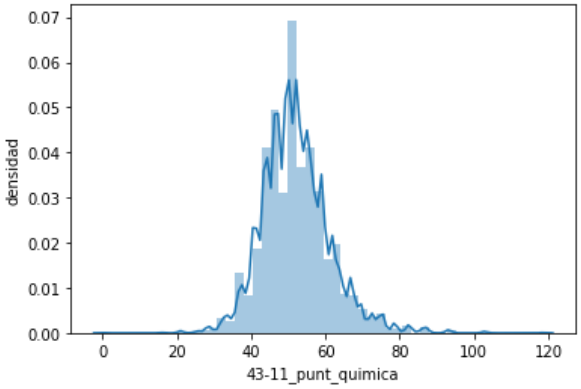


Figura 10. Gráfica de densidad, variable No. 43.

Variable: 44-11_punt_fisica, tipo: float64

Puntaje del inscrito en física

Tabla 38. Descripción estadística, variable No. 44.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	9.96
5-th percentil	35	Coef. de variación	0.20
Q1	45	Kurtosis	0.93
Mediana	51	Promedio	50.93
Q3	57.4	MAD	7.72
95-th percentil	67	Skewness	0.17
Máximo	124	Sum	7147200
Range	124	Varianza	99.17
Rango intercuartil	12.4		

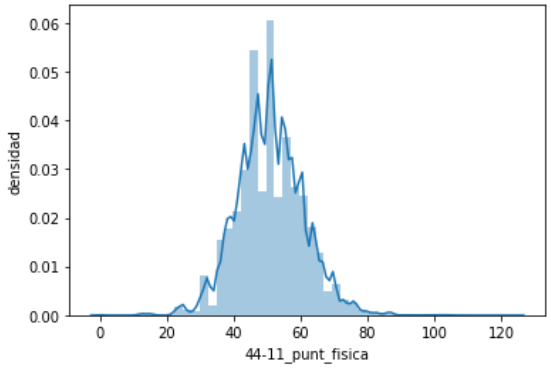


Figura 11. Gráfica de densidad, variable No. 44.

Variable: 45-11_punt_ingles, tipo: float64

Puntaje del inscrito en inglés

Tabla 39. Descripción estadística, variable No. 45.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	13.12
5-th percentil	36	Coef. de variación	0.25
Q1	43	Kurtosis	1.62
Mediana	49.66	Promedio	52.32
Q3	58	MAD	9.92
95-th percentil	79.81	Skewness	1.14
Máximo	117.29	Sum	7341800
Range	117.29	Varianza	172.05
Rango intercuartil	15		

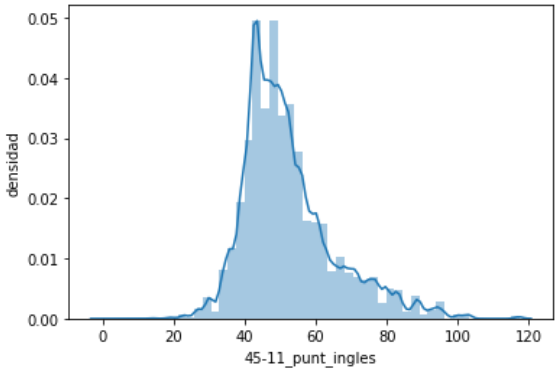


Figura 12. Gráfica de densidad, variable No. 45.

Variable: 48-11_punt_comp_flexible, tipo: float64

Tabla 40. Descripción estadística, variable No. 48.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	24.13
5-th percentil	4.2	Coef. de variación	0.85
Q1	5	Kurtosis	-1.78
Mediana	7	Promedio	28.42
Q3	53	MAD	23.54
95-th percentil	61.76	Skewness	0.18
Máximo	105.44	Sum	3987400
Range	105.44	Varianza	582.29
Rango intercuartil	48		

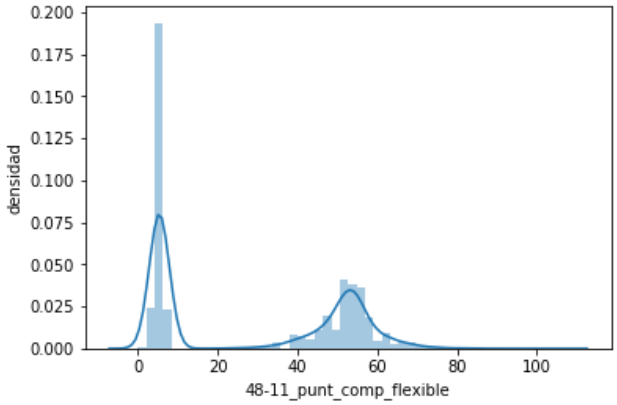


Figura 13. Gráfica de densidad, variable No. 48.

Variable: 50-11_estu_puesto

Puesto general del inscrito en el examen

Tabla 41. Descripción estadística, variable No. 50.

Cuantiles		Estadística descriptiva	
Mínimo	1	Desviación estándar	235
5-th percentil	12	Coef. de variación	0.89
Q1	73	Kurtosis	0.29
Mediana	192	Promedio	263.6
Q3	397	MAD	190.65
95-th percentil	758	Skewness	1.043
Máximo	1000	Sum	36988720
Range	999	Varianza	55227
Rango intercuartil	324		

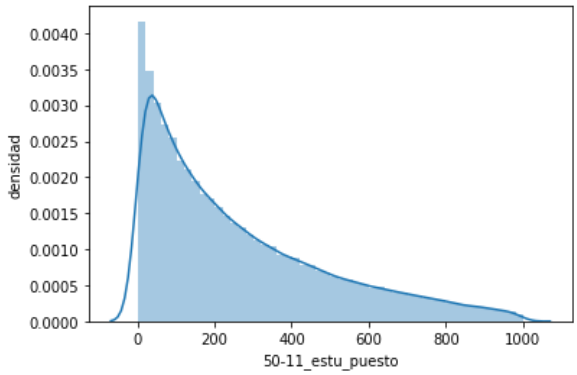


Figura 14. Gráfica de densidad, variable No. 50.

8.3.2 Variables de las Pruebas Saber Pro

Variable: 127-G_razona_cuantit_punt

Puntaje razonamiento cuantitativo

Tabla 42. Descripción estadística, variable No. 127.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	30.98
5-th percentil	107	Coef. de variación	0.20
Q1	137	Kurtosis	0.17
Mediana	158	Promedio	158.25
Q3	180	MAD	25.02
95-th percentil	208	Skewness	0.088
Máximo	300	Sum	22206023
Range	300	Varianza	959.97
Rango intercuartil	43		

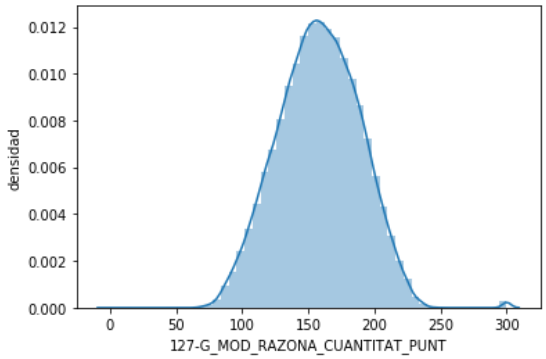


Figura 15. Gráfica de densidad, variable No. 127.

Variable: 131-G_lect_critica_punt

Puntaje lectura crítica

Tabla 43. Descripción estadística, variable No. 131.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	30.93
5-th percentil	104	Coef. de variación	0.20
Q1	135	Kurtosis	-0.28
Mediana	158	Promedio	156.98
Q3	180	MAD	25.19
95-th percentil	206	Skewness	-0.15
Máximo	300	Sum	22028333
Range	300	Varianza	956.4

Rango intercuartil	45

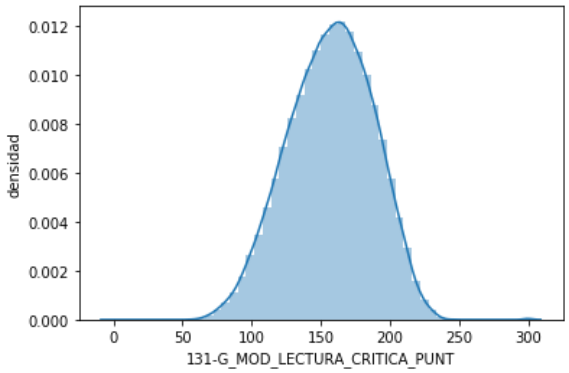


Figura 16. Gráfica de densidad, variable No. 131.

Variable: 135-G_compet_ciudad_punt

Puntaje competencias ciudadanas

Tabla 44. Descripción estadística, variable No. 135.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	32.05
5-th percentil	96	Coef. de variación	0.21
Q1	130	Kurtosis	0.12
Mediana	155	Promedio	152.23
Q3	175	MAD	25.96
95-th percentil	201	Skewness	-0.19
Máximo	300	Sum	21361608
Range	300	Varianza	1027.2
Rango intercuartil	45		

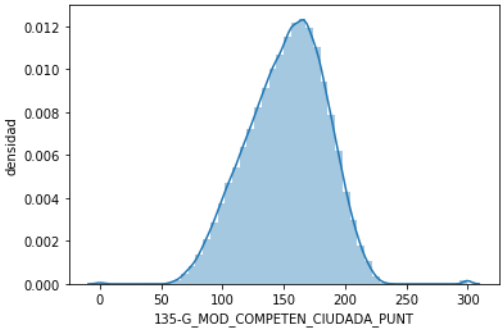


Figura 17. Gráfica de densidad, variable No. 135.

Variable: 139-G_ingles_punt

Puntaje inglés

Tabla 45. Descripción estadística, variable No. 139.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	32.15
5-th percentil	112	Coef. de variación	0.20
Q1	133	Kurtosis	0.78
Mediana	156	Promedio	158.27
Q3	182	MAD	26.54
95-th percentil	210	Skewness	0.23
Máximo	300	Sum	22209043
Range	300	Varianza	1033.7
Rango intercuartil	49		

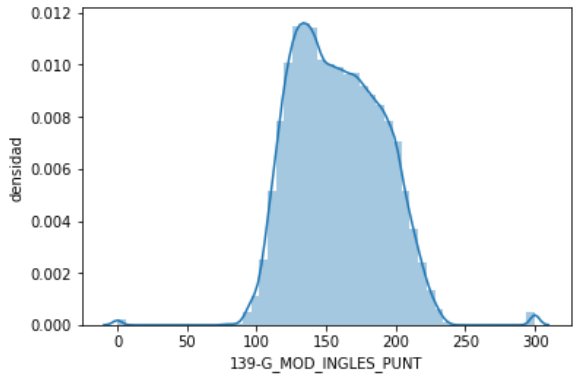


Figura 18. Gráfica de densidad, variable No. 139.

Variable: 143-G_com_escrita_punt

Puntaje comunicación escrita

Tabla 46. Descripción estadística, variable No. 143.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	36.91
5-th percentil	96	Coef. de variación	0.24
Q1	129	Kurtosis	3.83
Mediana	156	Promedio	151.47
Q3	168	MAD	26.78
95-th percentil	205	Skewness	-0.73
Máximo	300	Sum	21221000
Range	300	Varianza	1362.5
Rango intercuartil	39		

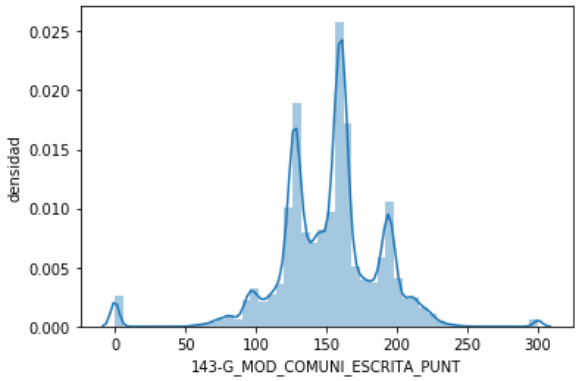


Figura 19. Gráfica de densidad, variable No. 143.

Variable: 147-G_punt_global

Puntaje total obtenido

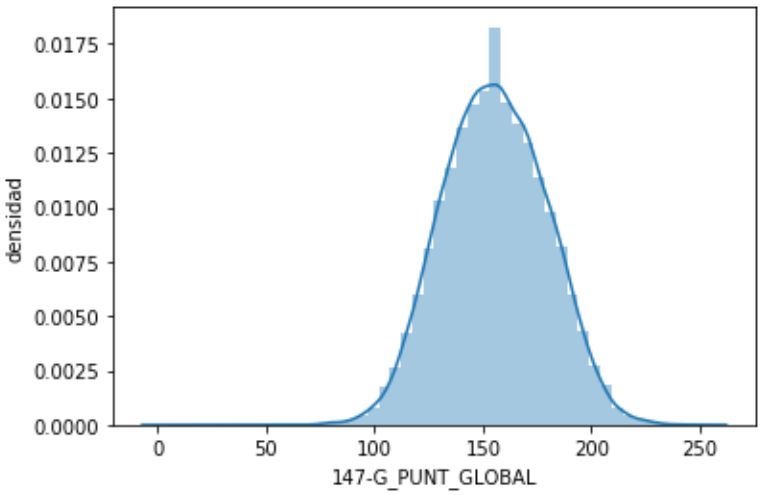
Tabla 47. Descripción estadística, variable No. 147.

Cuantiles

Mínimo	0
5-th percentil	117
Q1	138
Mediana	155
Q3	172
95-th percentil	194
Máximo	255
Range	255
Rango intercuartil	34

Estadística descriptiva

Desviación estándar	23.84
Coef. de variación	0.15
Kurtosis	-0.23
Promedio	155.39
MAD	19.34
Skewness	0.023
Sum	21805129
Varianza	568.14



182-E_result_puntaje

Puntaje de la prueba específica

Tabla 48. Descripción estadística, variable No. 182.

Cuantiles		Estadística descriptiva	
Mínimo	0	Desviación estándar	30.75
5-th percentil	105	Coef. de variación	0.20
Q1	135	Kurtosis	-0.18
Mediana	156	Promedio	156.02
Q3	178	MAD	24.82
95-th percentil	206	Skewness	-0.04
Máximo	300	Sum	21893077
Range	300	Varianza	945.33
Rango intercuartil	43		

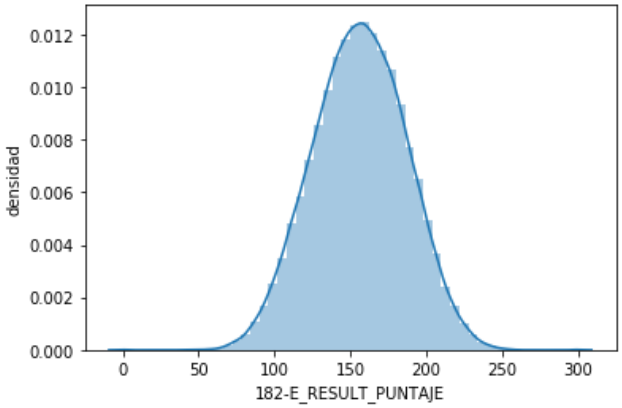


Figura 20. Gráfica de densidad, variable No. 182.

8.4 Limpieza de datos

A continuación, se presenta la revisión de registros duplicados, datos atípicos y datos nulos.

8.4.1 Registros duplicados

Se realizaron varias verificaciones con los códigos de ID que comunican las tablas de las pruebas Saber 11 con Saber Pro y que a su vez permiten la identificación única de cada estudiante, no se encontraron registros duplicados en la base de datos analizada para este proyecto.

8.4.2 Datos atípicos

Se revisaron los valores máximos y mínimos, encontrando que los datos más alejados de los valores centrales de cada variable eran perfectamente valores que podían ocurrir en distintos escenarios posibles dentro del contexto en el que se recolectó la información.

En el caso de las variables categóricas, los distintos atributos que toman se ajustan a las posibilidades consignadas por el ICFES en los informes con los que acompañan la divulgación de estos datos.

Para apoyar este apartado, se analizaron los distintos perfiles documentados para todas las variables a lo largo de este documento y en el Anexo 1. Exploración estadística de los datos, junto con gráficas de distribución de las variables, entre otras.

También se tuvieron en cuenta los siguientes dos indicadores:

- La medida de apuntamiento o de curtosis: la cual revela el grado de concentración de los valores de una variable alrededor de la zona central de distribución de frecuencias.

Para efectos de interpretación este indicador revela lo siguiente:

_Datos altamente concentrados alrededor de la media, curtosis > 0. Esta distribución se conoce como leptocúrtica.
_Datos que se distribuyen de forma normal, curtosis = 0. Esta distribución se conoce como mesocúrtica o normal.
_Datos dispersos que no se concentran alrededor de la media, curtosis < 0. Esta distribución se conoce como platicúrtica.

No se evidenciaron comportamientos de dispersión en las variables numéricas que ameriten una atención especial, dado que todos los valores se encontraron dentro de los posibles rangos de ocurrencia suministrados por el ICFES para la interpretación de los datos.

Tabla 49. Análisis de la dispersión para las variables numéricas.

Variable	Curtosis
Datos de las Pruebas Saber 11	

4-11_estu_edad	152.057446
38-11_punt_lenguaje	0.882831
39-11_punt_matematicas	1.168245
40-11_punt_c_sociales	0.32696
41-11_punt_filosofia	0.820055
42-11_punt_biologia	0.882055
43-11_punt_quimica	2.386496
44-11_punt_fisica	0.926392
45-11_punt_ingles	1.620589
48-11_punt_comp_flexible	-1.780101
50-11_estu_puesto	0.285857
59-11_fami_personas_hogar	3.931242

Datos de las Pruebas Saber Pro:

127-G_razona_cuant_punt	0.171396
128-G_razona_cuant_desem	-0.622135
129-G_razona_cuant_pnal	-1.061139
130-G_razona_cuant_pgref	-1.089924
131-G_lect_critica_punt	-0.282016
132-G_lectura_critica_desem	-0.692399
133-G_lect_critica_pnal	-1.11064
134-G_lect_critica_pgref	-1.12102
135-G_compet_ciudad_punt	0.122785
136-G_compet_ciudad_desem	-0.885293
137-G_compet_ciudad_pnal	-1.154561
138-G_compet_ciudad_pgref	-1.152053
139-G_ingles_punt	0.781
141-G_ingles_pnal	-1.082274
142-G_ingles_pgref	-1.09416
143-G_com_escrita_punt	3.830513
144-G_com_escrita_desem	-0.212039
145-G_com_escrita_pnal	-1.201244
146-G_com_escrita_pgref	-1.198091
147-G_punt_global	-0.225442
148-G_percentil_global	-1.071199
182-e_result_puntaje	-0.181
183-e_result_desemp	-0.854875

- El skewness: es un indicador de medida de asimetría estadística, útil para conocer el grado de simetría o asimetría que presenta una función de probabilidad de una variable aleatoria.

Para efectos de interpretación, esta medida de asimetría se lee de la siguiente manera. La distribución de la variable muestra:

_Una asimetría de cola derecha, cuando el skewness > 0.
_Una asimetría de cola izquierda, cuando el skewness < 0.

Lo anterior indica por citar algún ejemplo, que en el caso de la variable “4-11_estu_edad”, valores por encima de la media forman una asimetría positiva o de cola derecha, por lo cual es de esperarse que la mayor concentración de las edades se da en los números que se acercan más al rango inferior del dato. Es un público mayoritariamente joven el que presenta las Pruebas Saber 11.

Los resultados que se citan en la tabla a continuación se interpretaron en el contexto de los datos y no se encuentra ninguna alerta de valores atípicos.

Tabla 50. Medida de asimetría para todas las variables numéricas.

Variable	Swewness
Datos de las Pruebas Saber 11	
4-11_estu_edad	9.816357
38-11_punt_lenguaje	0.085531
39-11_punt_matematicas	0.435467
40-11_punt_c_sociales	0.032425
41-11_punt_filosofia	-0.032134
42-11_punt_biologia	0.176743
43-11_punt_quimica	0.729023
44-11_punt_fisica	0.17167
45-11_punt_ingles	1.138504
48-11_punt_comp_flexible	0.182684
50-11_estu_puesto	1.043059
59-11_fami_personas_hogar	1.37829
Datos de las Pruebas Saber Pro:	
127-G_razona_cuant_punt	0.087947
128-G_razona_cuant_desem	-0.340849
129-G_razona_cuant_pnal	-0.336281
130-G_razona_cuant_pgref	-0.263996
131-G_lect_critica_punt	-0.149253
132-G_lectura_critica_desem	-0.158492
133-G_lect_critica_pnal	-0.279125
134-G_lect_critica_pgref	-0.255657

135-G_compet_ciudad_punt	-0.189395
136-G_compet_ciudad_desem	-0.16935
137-G_compet_ciudad_pnal	-0.232826
138-G_compet_ciudad_pgref	-0.206753
139-G_ingles_punt	0.232577
141-G_ingles_pnal	-0.327017
142-G_ingles_pgref	-0.318433
143-G_com_escrita_punt	-0.725986
144-G_com_escrita_desem	-0.320863
145-G_com_escrita_pnal	-0.114013
146-G_com_escrita_pgref	-0.106807
147-G_punt_global	0.025821
148-G_percentil_global	-0.337352
182-e_result_puntaje	-0.040002
183-e_result_desemp	-0.076447
141-G_ingles_pnal	-0.327017
142-G_ingles_pgref	-0.318433
143-G_com_escrita_punt	-0.725986
144-G_com_escrita_desem	-0.320863
145-G_com_escrita_pnal	-0.114013
146-G_com_escrita_pgref	-0.106807
147-G_punt_global	0.025821
148-G_percentil_global	-0.337352
182-e_result_puntaje	-0.040002
183-e_result_desemp	-0.076447

8.4.3 Datos ausentes o nulos

La estimación porcentual global de los campos nulos en los datos que sirven de insumo para esta investigación es la siguiente, luego de eliminar 50 de las 183 variables por considerarse irrelevantes o redundantes, quedan 133 datos útiles y 140,324 registros, con los datos útiles se procede al siguiente calculo:

CV = Campos vacíos en los registros
TR = total de registros
TV = total de variables
PN = porcentaje de campos nulos

PN = CV / (TR * TV)
PN = 1,772,196 / (140,324 registros * 133 variables)

PN = 1,772,196 / 18,663,092

PN= 9.5% es el porcentaje de datos faltantes en todo el proyecto.

Consideraciones importantes: el atributo “sin_info” dentro de las variables se interpreta como un campo vacío.

Para las variables identificadas con la numeración desde la 160 hasta la 166 y 175 hasta la 178 se esperan dentro de sus atributos posibles, dos opciones (“Si” o “No”) sin embargo, con muy poca ocurrencia adicional a estas dos opciones, toman el valor de “0”, el cual fue interpretado y reemplazado por un “No”, que representa también la moda en todos los casos anteriores.

Se establecen los siguientes parámetros para el manejo de los datos nulos o ausentes, vistos para cada variable:

- Caso 1. Sí la cantidad de nulos o ausentes es inferior a 5%, se eliminan los registros vacíos o nulos.

- Caso 2. Sí el porcentaje de nulos es superior al 10%, se elimina la variable completa.

  El caso anterior se presentó para 11 variables.

- Caso 3. En el caso de estas variables, a pesar de contar con un porcentaje muy representativo de campos vacíos, se decidió llenar los registros en blanco con el texto “No_reporta”, con el objetivo de retener datos que representen a minorías.

- Caso 4. Sí el porcentaje de campos nulos o vacíos se encuentra entre un 5 y 10%, se imputa por la moda en variables categóricas, el promedio o la mediana, según convenga en variables numéricas o por algún otro método como el de vecinos cercanos. Este caso solo aplicó para la variable categórica: “24-11_cole biling”, la cual se imputó por la moda “0”, que presentaba una ocurrencia del 89.2% dentro del dato.

A continuación, se relacionan los porcentajes de los campos nulos encontrados en cada una de las siguientes variables, para los cuales se aplicaron los parámetros señalados anteriormente y en la última columna se cita el caso con el que se manejó la limpieza de los datos.

Tabla 51. Relación global de variables con campos nulos.

Consecutivo seguido por el nombre de la variable	Campos vacíos		Caso
	conteo	%	
13-11_estu_reside_depto	451	0.32%	1
15-11_estu_area_reside	50	0.04%	1
16-11_estu_trabaja	1238	0.88%	1
24-11_cole biling	11494	8.19%	4
27-11_cole_valor_pension	546	0.39%	1
32-11_estu_ies_dept_deseada	125815	89.66%	3
33-11_estu_carrdeseada_tipo	126737	90.32%	3
37-11_estu_estud-presentado-examen	213	0.15%	1
49-11_desemp_comp_flexible	68660	48.93%	2
51-11_fami_nivel_sisben	89	0.06%	1
52-11_fami_estrato_vivienda	3049	2.17%	1
53-11_fami_ing_fmiliar_mensual	101	0.07%	1
56-11_fami_ocup_padre	2758	1.97%	1
57-11_fami_ocup_madre	2758	1.97%	1
58-11_fami_pisos_hogar	158	0.11%	1
59-11_fami_personas_hogar	98	0.07%	1
60-11_fami_telefono_fijo	89	0.06%	1
61-11_fami_celular	88	0.06%	1
62-11_fami_internet	89	0.06%	1
63-11_fami_servicio_television	89	0.06%	1
64-11_fami_computador	2228	1.59%	1
65-11_fami_lavadora	89	0.06%	1
66-11_fami_nevera	88	0.06%	1
67-11_fami_horno	89	0.06%	1
68-11_fami_dvd	89	0.06%	1
69-11_fami_microondas	89	0.06%	1
70-11_fami_automovil	89	0.06%	1
79-G_estu_depto_reside	3	0.00%	1
83-G_estu_areareside	64827	46.20%	2
84-G_estu_estadocivil	64887	46.24%	2
86-G_estu_tieneetnia	4	0.00%	1
87-G_estu_etnia	137314	97.86%	3
88-G_estu_limita_motriz	140283	99.97%	3
89-G_estu_limita_invidente	140309	99.99%	3
90-G_estu_limita_condicionespecial	140323	100%	3
91-G_estu_limita_sordo	140307	99.99%	3
92-G_estu_limita_autismo	140324	100%	2
93-G_estu_tituloobtenidobachiller	72	0.05%	1
98-G_estu_comocapacitoexamensb11	11	0.01%	1

99-G_estu_cursodocentesies	69138	49.27%	2
100-G_estu_cursoiesapoyoexterno	69140	49.27%	2
101-G_estu_cursoiesexterna	69143	49.27%	2
102-G_estu_simulacrotipoicfes	69137	49.27%	2
103-G_estu_actividadrefuerzoareas	69138	49.27%	2
104-G_estu_actividadrefuerzogeneric	69140	49.27%	2
107-G_estu_semestrecursa	72	0.05%	1
143-G_comuni_escrita_punt	225	0.16%	1
144-G_comuni_escrita_desem	225	0.16%	1
145-G_comuni_escrita_pnal	225	0.16%	1
146-G_comuni_escrita_pgref	225	0.16%	1
150-G_fami_hogaractual	4	0.00%	1
151-G_fami_cabezafamilia	4	0.00%	1
152-G_fami_numpersonasacargo	4	0.00%	1
153-G_fami_educacionpadre	4	0.00%	1
154-G_fami_educacionmadre	4	0.00%	1
155-G_fami_ocupacionpadre	6	0.00%	1
156-G_fami_ocupacionmadre	5	0.00%	1
157-G_fami_estratovivienda	4	0.00%	1
158-G_fami_personashogar	4	0.00%	1
159-G_fami_cuartoshogar	4	0.00%	1
160-G_fami_tieneinternet	7	0.00%	1
161-G_fami_tieneserviciotv	6	0.00%	1
162-G_fami_tienecomputador	4	0.00%	1
163-G_fami_tienelavadora	5	0.00%	1
164-G_fami_tienehornomicroogas	1883	1.34%	1
165-G_fami_tieneautomovil	9	0.01%	1
166-G_fami_tienemotocicleta	3462	2.47%	1
167-G_fami_numlibros	4	0.00%	1
168-G_estu_dedicacionlecturadiaria	4	0.00%	1
169-G_estu_dedicacioninternet	132	0.09%	1
170-G_estu_horassemanatrabaja	4	0.00%	1
171-G_estu_tiporemuneracion	34288	24.43%	2
172-G_estu_inse_individual	2	0.00%	1
173-G_estu_nse_individual	2	0.00%	1
174-G_estu_valormatriculauniversidad	72	0.05%	1
175-G_estu_pagomatriculabeca	125	0.09%	1
176-G_estu_pagomatriculacredito	120	0.09%	1
177-G_estu_pagomatriculapadres	92	0.07%	1
178-G_estu_pagomatriculapropio	129	0.09%	1
Total	1772196		

Luego de la limpieza de los datos nulos, permanecen 122 variables y 132067 observaciones.

8.5 Correlaciones

Por efectos de visualización, en la tabla 52 se muestran las correlaciones solo de las variables numéricas, relacionadas con datos de las pruebas Saber 11. Para los estudiantes que presentan el examen de estado antes de graduarse de la educación media, se cuenta con 13 variables de tipo numérico, las cuales se representan en la siguiente tabla de correlación.

Tabla 52. Tabla de correlaciones, variables Saber 11.

nombre o prefijo identificador de la variable	2	4	38	39	40	41	42	43	44	45	48	50	59
2-11_periodo-anno	1.00	-0.03	-0.11	-0.05	-0.16	-0.28	-0.17	-0.19	-0.16	-0.05	-0.01	-0.10	-0.04
4-11_estu_edad	-0.03	1.00	-0.07	-0.08	-0.05	-0.04	-0.07	-0.07	-0.04	-0.02	0.02	0.13	0.01
38-11_punt_lenguaje	-0.11	-0.07	1.00	0.49	0.56	0.45	0.52	0.47	0.42	0.48	-0.03	-0.72	-0.08
39-11_punt_matematicas	-0.05	-0.08	0.49	1.00	0.52	0.39	0.54	0.56	0.55	0.53	-0.08	-0.72	-0.07
40-11_punt_c_sociales	-0.16	-0.05	0.56	0.52	1.00	0.51	0.57	0.52	0.46	0.49	-0.02	-0.69	-0.07
41-11_punt_filosofia	-0.28	-0.04	0.45	0.39	0.51	1.00	0.47	0.44	0.39	0.41	-0.01	-0.52	-0.04
42-11_punt_biologia	-0.17	-0.07	0.52	0.54	0.57	0.47	1.00	0.56	0.49	0.49	-0.04	-0.64	-0.07
43-11_punt_quimica	-0.19	-0.07	0.47	0.56	0.52	0.44	0.56	1.00	0.53	0.53	-0.06	-0.58	-0.05
44-11_punt_fisica	-0.16	-0.04	0.42	0.55	0.46	0.39	0.49	0.53	1.00	0.46	-0.06	-0.55	-0.05
45-11_punt_ingles	-0.05	-0.02	0.48	0.53	0.49	0.41	0.49	0.53	0.46	1.00	-0.05	-0.53	-0.10
48-11_punt_comp_flexible	-0.01	0.02	-0.03	-0.08	-0.02	-0.01	-0.04	-0.06	-0.06	-0.05	1.00	0.04	0.02
50-11_estu_puesto	-0.10	0.13	-0.72	-0.72	-0.69	-0.52	-0.64	-0.58	-0.55	-0.53	0.04	1.00	0.09
59-11_fami_personas_hogar	-0.04	0.01	-0.08	-0.07	-0.07	-0.04	-0.07	-0.05	-0.05	-0.10	0.02	0.09	1.00

Ninguna de las variables anteriores fue eliminada, pués se desea que todas las variables implicadas en esa asociación permanezcan en el modelo, dado que por ejemplo el puesto del estudiante en la prueba ayudará a simplificar algunas interpretaciones sobre el desempeño académico.

Por otra parte las asociaciones negativas que se dan entre los módulos evaluados y el puesto obtenido por el estudiante, denotan la forma como se debe interpretar esta última variable.

A menor puntuación en los módulos evaluados, mayor es el puesto global en el que el alumno de la educación media queda clasificado en las pruebas Saber 11, por lo tanto los estudiantes más sobresalientes, medidos por el desempeño académico ocuparán los primeros puestos dentro del rango que toma esta variable (0-1000).

Siguiendo con el método que nos facilita la selección de factores, se muestran las correlaciones de los datos numéricos de las Pruebas Saber Pro.

Tabla 53. Tabla de correlaciones, variables Saber Pro.

nombre o prefijo identificador de la variable	127	128	129	130	131	132	133	134	135	136	137	138	139	141	142	143	144	145	146	147	148	182	183
127-G_razona_cuant_punt	1.00	0.92	0.97	0.88	0.57	0.54	0.57	0.51	0.48	0.46	0.49	0.45	0.50	0.50	0.41	0.21	0.22	0.22	0.21	0.74	0.72	0.52	0.49
128-G_razona_cuant_desem	0.92	1.00	0.93	0.84	0.53	0.50	0.53	0.48	0.45	0.43	0.46	0.42	0.46	0.46	0.38	0.20	0.21	0.21	0.20	0.68	0.68	0.48	0.46
129-G_razona_cuant_pnal	0.97	0.93	1.00	0.89	0.57	0.54	0.57	0.52	0.47	0.46	0.49	0.45	0.49	0.50	0.41	0.21	0.21	0.22	0.21	0.73	0.73	0.51	0.49
130-G_razona_cuant_pgref	0.88	0.84	0.89	1.00	0.55	0.52	0.55	0.58	0.47	0.45	0.48	0.50	0.44	0.44	0.46	0.22	0.22	0.23	0.24	0.68	0.69	0.55	0.52
131-G_lect_critica_punt	0.57	0.53	0.57	0.55	1.00	0.93	0.98	0.95	0.66	0.63	0.67	0.63	0.54	0.54	0.49	0.28	0.29	0.30	0.28	0.82	0.82	0.57	0.55
132-G_lect_critica_desem	0.54	0.50	0.54	0.52	0.93	1.00	0.94	0.90	0.62	0.60	0.64	0.60	0.51	0.51	0.47	0.27	0.27	0.29	0.26	0.77	0.78	0.54	0.52
133-G_lect_critica_pnal	0.57	0.53	0.57	0.55	0.98	0.94	1.00	0.96	0.66	0.63	0.68	0.64	0.54	0.54	0.49	0.28	0.29	0.30	0.28	0.81	0.83	0.57	0.55
134-G_lect_critica_pgref	0.51	0.48	0.52	0.58	0.95	0.90	0.96	1.00	0.62	0.59	0.64	0.66	0.50	0.50	0.52	0.26	0.27	0.28	0.28	0.76	0.77	0.57	0.55
135-G_compet_ciudad_punt	0.48	0.45	0.47	0.47	0.66	0.62	0.66	0.62	1.00	0.93	0.97	0.94	0.50	0.49	0.46	0.27	0.29	0.28	0.26	0.78	0.77	0.54	0.51
136-G_compet_ciudad_desem	0.46	0.43	0.46	0.45	0.63	0.60	0.63	0.59	0.93	1.00	0.93	0.90	0.47	0.47	0.44	0.26	0.27	0.27	0.25	0.74	0.74	0.52	0.49
137-G_compet_ciudad_pnal	0.49	0.46	0.49	0.48	0.67	0.64	0.68	0.64	0.97	0.93	1.00	0.96	0.50	0.50	0.47	0.27	0.27	0.29	0.26	0.78	0.79	0.55	0.52
138-G_compet_ciudad_pgref	0.45	0.42	0.45	0.50	0.63	0.60	0.64	0.66	0.94	0.90	0.96	1.00	0.47	0.47	0.49	0.25	0.25	0.26	0.27	0.74	0.75	0.55	0.52
139-G_ingles_punt	0.50	0.46	0.49	0.44	0.54	0.51	0.54	0.50	0.50	0.47	0.50	0.47	1.00	0.96	0.91	0.25	0.25	0.27	0.25	0.75	0.73	0.44	0.42
141-G_ingles_pnal	0.50	0.46	0.50	0.44	0.54	0.51	0.54	0.50	0.49	0.47	0.50	0.47	0.96	1.00	0.95	0.25	0.25	0.27	0.25	0.74	0.74	0.44	0.42
142-G_ngles_pgref	0.41	0.38	0.41	0.46	0.49	0.47	0.49	0.52	0.46	0.44	0.47	0.49	0.91	0.95	1.00	0.24	0.24	0.25	0.26	0.68	0.68	0.44	0.42
143-G_com_escrita_punt	0.21	0.20	0.21	0.22	0.28	0.27	0.28	0.26	0.27	0.26	0.27	0.25	0.25	0.25	0.24	1.00	0.93	0.91	0.90	0.58	0.55	0.24	0.23
144-G_com_escrita_desem	0.22	0.21	0.21	0.22	0.29	0.27	0.29	0.27	0.29	0.27	0.27	0.25	0.25	0.25	0.24	0.93	1.00	0.93	0.92	0.56	0.54	0.25	0.24
145-G_com_escrita_pnal	0.22	0.21	0.22	0.23	0.30	0.29	0.30	0.28	0.28	0.27	0.29	0.26	0.27	0.27	0.25	0.91	0.93	1.00	0.99	0.57	0.55	0.26	0.25
146-G_com_escrita_pgref	0.21	0.20	0.21	0.24	0.28	0.26	0.28	0.28	0.26	0.25	0.26	0.27	0.25	0.25	0.26	0.90	0.92	0.99	1.00	0.54	0.53	0.25	0.24
147-G_punt_global	0.74	0.68	0.73	0.68	0.82	0.77	0.81	0.76	0.78	0.74	0.78	0.74	0.75	0.74	0.68	0.58	0.56	0.57	0.54	1.00	0.97	0.62	0.59
148-G_percentil_global	0.72	0.68	0.73	0.69	0.82	0.78	0.83	0.77	0.77	0.74	0.79	0.75	0.73	0.74	0.68	0.55	0.54	0.55	0.53	0.97	1.00	0.62	0.59
182-E_result_puntaje	0.52	0.48	0.51	0.55	0.57	0.54	0.57	0.57	0.54	0.52	0.55	0.55	0.44	0.44	0.44	0.24	0.25	0.26	0.25	0.62	0.62	1.00	0.94
183-E_result_desem	0.49	0.46	0.49	0.52	0.55	0.52	0.55	0.55	0.51	0.49	0.52	0.52	0.42	0.42	0.42	0.23	0.24	0.25	0.24	0.59	0.59	0.94	1.00

De las 23 variables numéricas expuestas en la tabla de correlaciones con los datos de las pruebas Saber Pro, se decide eliminar las siguientes 16, las cuales van a seguir representadas en el modelo a través de las variables que se relacionan en la columna 2 que se revela a continuación:

Tabla 54. Explicación de la eliminación de variables por correlación.

Variables eliminadas	Variable que contiene a las eliminadas y permanece en el modelo
128-G_razona_cuant_desem 129-G_razona_cuant_pnal 130-G_razona_cuant_pgref	127-G_razona_cuant_punt
132-G_lectura_critica_desem 133-G_lect_critica_pnal 134-G_lect_critica_pgref	131-G_lect_critica_punt
136-G_compet_ciudad_desem 137-G_compet_ciudad_pnal 138-G_compet_ciudad_pgref	135-G_compet_ciudad_punt

141-G_ingles_pnal 142-G_ingles_pgref 144-G_com_escrita_desem	139-G_ingles_punt
145-G_com_escrita_pnal 146-G_com_escrita_pgref	143-G_com_escrita_punt
148-G_percentil_global	147-G_punt_global
183-E_result_desem	182-E_result_puntaje

Luego de la eliminación de las variables por correlación, permanecen 106 variables y 132067 observaciones.

8.6 Datos finales para el modelamiento analítico

En las tablas de este apartado, se relacionan todas las variables que se tuvieron en cuenta en la etapa inicial de este proyecto y la cantidad que fueron eliminadas durante la preparación y limpieza de los datos, con las siguientes etiquetas para dar cuenta del criterio bajo el cual fueron eliminadas:

Paso 1. Variables irrelevantes o redundantes.
Paso 2. Variables con campos nulos o vacíos.
Paso 3. Variables contenidas en otra (correlación).

Tabla 55. Preparación de los datos Saber 11 – resumen de variables eliminadas

Agrupadas según su relación con	Número inicial de variables	Eliminadas			Número final de variables
		Paso 1	Paso 2	Paso 3	
La identificación de la prueba y el estudiante	16	11			5
El colegio en donde estudió el inscrito	11	4			7
La intención del estudiante en cuanto a la elección de una carrera profesional	6	4			2
La presentación de la prueba	4	2			2
El desempeño académico en la prueba	13		1		12

Variables de tipo socioeconómico	20				20
Total	70	21	1		48

Tabla 56. Preparación de los datos Saber Pro – resumen variables eliminadas.

Agrupadas según su relación con	Número inicial de variables	Eliminadas			Número final de variables
		Paso 1	Paso 2	Paso 3	
La identificación de la prueba y el estudiante	15	7	2		6
La identificación de minorías o condiciones especiales en el inscrito	7		1		6
El colegio en el que estudió el inscrito	5	4			1
La preparación para la prueba	7		6		1
La elección vocacional del estudiante	18	10			8
La presentación del examen y evaluación de competencias genéricas	27	5		15	7
Variables de tipo socioeconómico	24	1	1		22
Valor y pago de la matrícula	5				5
Evaluación de competencias específicas	5	2		1	2
Total	113	29	10	16	58

En la tabla 57 se muestra un resumen del global de las variables eliminadas en cada uno de los pasos descritos arriba.

Tabla 57. Preparación de los datos, consolidado variables eliminadas Saber 11 y Saber Pro.

Agrupadas según su	Número	Eliminadas	Número

relación con	inicial de variables	Paso 1	Paso 2	Paso 3	final de variables
Pruebas Saber 11	70	21	1		48
Pruebas Saber Pro	113	29	10	16	58
Total	183	50	11	16	106

Con todos los pasos aplicados anteriormente, se cuenta con 48 variables de las Pruebas Saber 11 y 58 de las Saber Pro, para un total de 106 variables y 132067 observaciones.

Como últimos ajustes realizados a las variables, partiendo de las 106 variables que se especifican en el párrafo anterior, se realizaron los cambios que se detallan a continuación:

Tabla 58. Ajustes adicionales sobre las variables.

Variable eliminada	Variable(s) creada(s) en remplazo de la(s) eliminada(s)
"2-11_periodo" Aplicación del examen, año seguido del semestre	"2-11_anno_num" Año de aplicación del examen
	"2-11_periodo_num" Semestre de aplicación del examen
"74-G_estu_dia_nac" Día de nacimiento del estudiante "75-G_estu_mes_nac" Mes de nacimiento del estudiante "76-G_estu_ano_nac" Año de nacimiento del estudiante	"76-G-creado_estu_edad" Edad del estudiante al presentar la Prueba Saber Pro
"158-G_fami_personashogar_num" Número de personas en el hogar "159-G_fami_cuartoshogar_num" Número de habitaciones en el hogar	"159-G_creado_fami_hacina" Índice de hacinamiento credo de la siguiente manera (158-G_fami_personashogar_num /159-G_fami_cuartoshogar_num)

Luego de este último paso, permanecen para efectos de los pasos del modelamiento que siguen a continuación, 49 variables de las Pruebas Saber 11 y 55 de las Pruebas Saber Pro. Para un total de 104 variables.

Además, se decide para efectos de reducir la dimensionalidad de este proyecto, reasignar nuevos atributos a las variables relacionadas con el departamento, las cuales son las siguientes tres:

- 32-11_estu_ies_dept_deseada (Departamento de la institución de educación superior donde le gustaría estudiar).
- 36-11_estu_dept_presentacion (Departamento de presentación del examen).
- 114- G_estu_prgm_departamento (Departamento donde se ofrece el programa académico).

La tabla a continuación muestra el nuevo atributo asignado a los 33 registros diferentes dentro de las variables asociadas al departamento, pasando de 33 opciones diferentes a 13, con el objetivo de reducir la dimensionalidad.

Tabla 59. Reducción de las dimensiones para variables relacionadas con el departamento.

Atributo original de las variables relacionadas con el departamento	Nuevos atributos asignados a las variables
AMAZONAS	Amazónica
CAQUETA	Amazónica
GUAINIA	Amazónica
GUAVIARE	Amazónica
PUTUMAYO	Amazónica
VAUPES	Amazónica
BOYACA	Andina
CAUCA	Andina
CUNDINAMARCA	Andina
HUILA	Andina
NARINO	Andina
NORTE SANTANDER	Andina
TOLIMA	Andina
ANTIOQUIA	Antioquia
ATLANTICO	Atlántico
BOGOTA	Bogotá
BOLIVAR	Bolívar
CESAR	Caribe
CORDOBA	Caribe

LA GUAJIRA	Caribe
MAGDALENA	Caribe
SUCRE	Caribe
CALDAS	Eje Cafetero
QUINDIO	Eje Cafetero
RISARALDA	Eje Cafetero
SAN ANDRES	San Andrés
ARAUCA	Llanos
CASANARE	Llanos
META	Llanos
VICHADA	Llanos
CHOCO	Choco
SANTANDER	Santander
VALLE	Valle del Cauca

Finalmente, se procede a una transformación de las variables categóricas, ya que la librería a emplearse para realizar la ciencia de los datos en este proyecto es scikit-learn de Python, como requisito todos los datos de entrada se deben ingresar de forma numérica, por lo cual procedemos de la siguiente manera:

Para las siguientes variables categóricas, se reemplazan sus atributos por ceros y unos, los cuales indican la presencia del atributo en la observación de la siguiente manera “0” = no y “1” = sí.

2-11_periodo_num
7-11_estu_genero
15-11_estu_area_reside
23-11_cole_naturaleza
24-11_cole biling
60-11_fami_telefono_fijo
61-11_fami_celular
62-11_fami_internet
63-11_fami_servicio_television
64-11_fami_computador
65-11_fami_lavadora
66-11_fami_nevera
67-11_fami_horno
68-11_fami_dvd
69-11_fami_microondas
70-11_fami_automovil
73-G_estu_genero

77-G_periodo
86-G_estu_tieneetnia
88-G_estu_limita_motriz
89-G_estu_limita_invidente
90-G_estu_limita_condicionespecial
91-G_estu_limita_sordo
150-G_fami_hogaractual
151-G_fami_cabezafamilia
161-G_fami_tieneserviciotv
162-G_fami_tienecomputador
163-G_fami_tienelavadora
164-G_fami_tienehornomicroogas
165-G_fami_tieneautomovil
166-G_fami_tienemotocicleta
175-G_estu_pagomatriculabeca
176-G_estu_pagomatriculacredito
177-G_estu_pagomatriculapadres
178-G_estu_pagomatriculapropio

La transformación anterior es justificada para el grupo de variables que se acaba de citar, dado cada uno de estos datos presentan solo dos atributos diferentes dentro de sus registros únicos.

A continuación, se relaciona un grupo de 34 variables categóricas para las cuales se procedió a crear variables ficticias de tipo dummy, pues cuentan con más de dos atributos en cada caso y no son ordinales, lo que dará origen a nuevas variables que son las combinaciones de la original con cada uno de sus registros únicos, para ser representadas en sus atributos por ceros y unos, con la convención explicada anteriormente.

21-11_cole_calendario
22-11_cole_genero
25-11_cole_jornada
26-11_cole_caracter
32-11_creada_estu_ies_dept_deseada
33-11_estu_carrdeseada_tipo
36-11_creada_estu_dept_presentacion
46-11_desemp_ingles
47-11_nombre_comp_flexible
54-11_fami_educa_padre

55-11_fami_educa_madre
71-G_estu_tipodocumento
87-G_estu_etnia
93-G_estu_tituloobtenidobachiller
98-G_estu_comocapacitoexamensb11
111-G_gruporeferencia
114-G_creada_estu_prgm_departamento
115-G_estu_nivel_prgm_academico
116-G_estu_metodo_prgm
121-G_inst_caracter_academico
122-G_inst_origen
140-G_ingles_desem
153-G_fami_educacionpadre
154-G_fami_educacionmadre
155-G_fami_ocupacionpadre
156-G_fami_ocupacionmadre
160-G_fami_tieneinternet
167-G_fami_numlibros
168-G_estu_dedicacionlecturadiaria
169-G_estu_dedicacioninternet
170-G_estu_horassemanatrabaja
173-G_estu_nse_individual
174-G_estu_valormatriculauniversidad
181-e_result_nombreprueba

Por las razones explicadas en el apartado “Creación de variables dummies”, se inicia el modelamiento de los datos con un total de 330 variables y 132067 observaciones, de las cuales 111 corresponden a las pruebas Saber 11 y 219 a las Saber Pro.

Cabe resaltar que para efectos interpretativos realmente existen 49 variables de las Pruebas Saber 11 y 55 de las Pruebas Saber Pro. Para un total de 104 variables, pues el número global paso de 104 a 330 variables dado que se crearon variables ficticias conocidas como dummy para poder introducir de forma numérica los datos categóricos a los modelos de ciencia de datos.

9. MODELAMIENTO Y EVALUACIÓN

Se inicia el modelamiento con las variables que permanecieron en el interés de este proyecto de investigación, luego de todos los procesos de preparación de los datos documentados en el capítulo anterior.

Los modelos analíticos desarrollados son:

- Selección de factores de mayor relación con el desempeño en las pruebas Saber 11 y Saber Pro.
- Definición de perfiles de estudiantes.
- Relaciones entre los perfiles de estudiantes de educación media y estudiantes de educación superior.

9.1 Selección de los factores de mayor relación con el desempeño de las Pruebas Saber 11 y Saber Pro

Para la identificación de factores asociados al rendimiento académico se aplica un análisis de correlaciones para identificar aquellas variables que tienen mayor relación con nuestras variables de interés. Para este fin se definen 4 variables de tipo objetivo, 2 de estas pertenecientes a las pruebas Saber 11 y 2 a las pruebas Saber Pro.

Se evalúa primero el comportamiento de los datos de las pruebas Saber 11 con relación a cada una de las dos variables objetivo de este grupo analizado. Seguido de la evaluación de las asociaciones entre todos los datos de las Pruebas Saber 11 y Saber Pro con las variables objetivo de este último examen.

Definiendo como regla de interpretación de las correlaciones lo siguiente, las variables que tengan una correlación en valor absoluto mayor a 0.3, están relacionadas con el desempeño académico de los estudiantes.

Cada uno de los ejercicios de correlación, iniciarán con el señalamiento de la variable definida como objetivo, acompañada de un diagrama de cajas y bigotes para facilitar la comprensión de la distribución del dato, seguidas de la tabla de correlación con el resto de las variables tenidas en cuenta.

9.1.1 Correlaciones con el desempeño académico en las pruebas Saber

11

Se definen en este apartado dos variables objetivo que representan consistentemente el rendimiento académico del estudiante en las Pruebas Saber 11.

Variable objetivo No. 1:

- 48-punt_comp_flexible (puntaje del inscrito en componente flexible rango [0,106]).

  El componente flexible consta de 2 categorías interdisciplinares, Medio Ambiente y Violencia y Sociedad y 4 de profundización; Biología, Ciencias Sociales, Leguaje y Matemáticas, para un total de 6 opciones de las cuales el estudiante escoge solo una de estas.

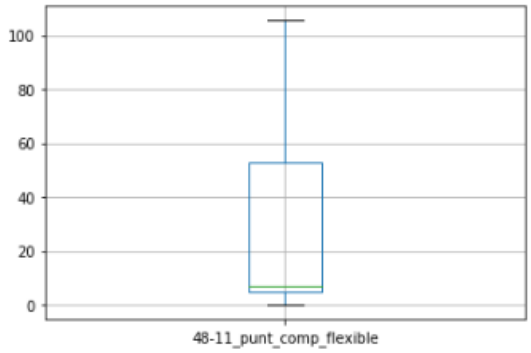


Figura 21. Diagrama de caja de la variable No. 48-Puntaje en el componente flexible de las pruebas Saber 11.

A continuación, se presenta la tabla de correlación de los datos con la variable definida arriba.

Tabla 60. Correlaciones de los datos Saber 11 con el rendimiento académico en la misma prueba.

	48-11_punt_comp_flexible
48-11_punt_comp_flexible	1.00
47_MEDIO AMBIENTE	0.61
47_VIOLENCIA Y SOCIEDAD	0.52
36_Antioquia	0.05
50-11_estu_puesto	0.04

15-11_estu_area_reside	0.04
36_Atlantico	0.04
36_Bolivar	0.03
2-11_periodo_num	0.03
36_Caribe	0.03
23-11_cole_naturaleza	0.03
46_A1	0.02
54_ninguno	0.02
25_SABATINA - DOMINICAL	0.02
46_A-	0.02
55_ninguno	0.02
57-11_fami_ocup_madre	0.02
22_F	0.02
54_primaria	0.02
55_primaria	0.02
59-11_fami_personas_hogar	0.02
21_Flexible	0.02
26_ACADEMICO Y TECNICO	0.02
56-11_fami_ocup_padre	0.02
36_Llanos	0.02
4-11_estu_edad	0.02
32_Antioquia	0.01
36_Amazonica	0.01
32_Atlantico	0.01
25_UNICA	0.01
21_A	0.01
25_TARDE	0.01
33_2	0.01
32_Caribe	0.01

55_No_reporta	0.01
33_0	0.01
26_TECNICO	0.01
36_SanAndres	0.01
25_NOCHE	0.01
32_Llanos	0.01
32_Bolivar	0.01
33_1	0.01
25_MANANA	0.01
26_DESCONOCIDO	0.01
32_Amazonica	0.01
54_bachiller	0.00
54_no_sabe	0.00
55_no_sabe	0.00
55_bachiller	0.00
32_No_reporta	0.00
33_3	0.00
16-11_estu_trabaja	0.00
32_SanAndres	0.00
32_Andina	0.00
32_EjeC	0.00
26_NORMALISTA	0.00
33_No_reporta	0.00
32_Choco	0.00
41-11_punt_filosofia	-0.01
46_A2	-0.01
61-11_fami_celular	-0.01
32_Bogota	-0.01
2-11_anno_num	-0.01

36_EjeC	-0.01
68-11_fami_dvd	-0.01
32_Valle	-0.01
36_Choco	-0.01
22_Mixto	-0.01
66-11_fami_nevera	-0.01
69-11_fami_microondas	-0.01
36_Andina	-0.01
32_Santander	-0.02
22_M	-0.02
24-11_cole biling	-0.02
26_ACADEMICO	-0.02
54_pregrado	-0.02
55_pregrado	-0.02
65-11_fami_lavadora	-0.02
25_COMPLETA U ORDINARIA	-0.02
67-11_fami_horno	-0.02
40-11_punt_c_sociales	-0.02
70-11_fami_automovil	-0.02
63-11_fami_servicio_television	-0.02
46_B+	-0.03
21_B	-0.03
62-11_fami_internet	-0.03
60-11_fami_telefono_fijo	-0.03
58-11_fami_pisos_hogar	-0.03
64-11_fami_computador	-0.03
27-11_cole_valor_pension	-0.03
36_Bogota	-0.03
54_posgrado	-0.03

55_posgrado	-0.03
38-11_punt_lenguaje	-0.03
36_Santander	-0.04
52-11_fami_estrato_vivienda	-0.04
46_B1	-0.04
36_Valle	-0.04
42-11_punt_biologia	-0.04
53-11_fami_ing_fmiliar_mensual	-0.05
51-11_fami_nivel_sisben	-0.05
45-11_punt_ingles	-0.05
44-11_punt_fisica	-0.06
37-11_estu_estud-presentado-examen	-0.06
43-11_punt_quimica	-0.06
7-11_estu_genero	-0.07
39-11_punt_matematicas	-0.08
47_PROFUNDIZACION EN CIENCIAS SOCIALES	-0.28
47_PROFUNDIZACION EN LENGUAJE	-0.37
47_PROFUNDIZACION EN BIOLOGIA	-0.38
47_PROFUNDIZACION EN MATEMATICA	-0.42

Siguiendo con las correlaciones de todas las variables de las pruebas Saber 11 con el desempeño académico en la misma Prueba, procedemos a valorar las relaciones de los datos con el puesto global obtenido por el estudiante en la misma prueba.

Variable objetivo No. 2:

- 50-puesto_est (puesto del estudiante según su desempeño global en la prueba, rango [1,1000]).

  Este dato revela el puesto obtenido por el estudiante de la educación media según su desempeño general en la prueba.

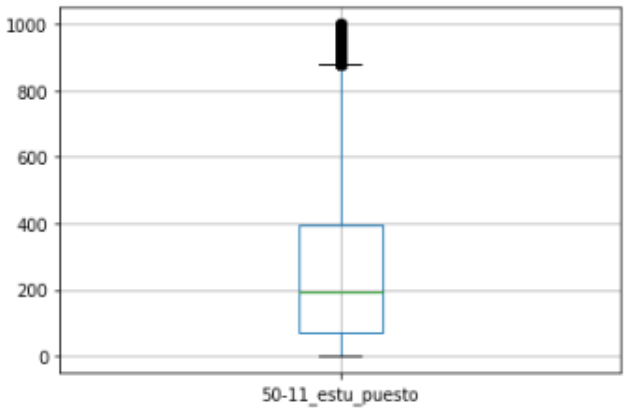


Figura 22. Diagrama de caja de la variable No. 50-Puesto global del estudiante en las pruebas Saber 11.

A continuación, se presenta la tabla de correlación de los datos con la variable definida arriba.

Tabla 61. Correlaciones de los datos Saber 11 con el puesto global del estudiante en la misma prueba.

	50-11_estu_puesto
50-11_estu_puesto	1
46_A-	0.41
23-11_cole_naturaleza	0.24
46_A1	0.15
57-11_fami_ocup_madre	0.14
25_MANANA	0.14
22_Mixto	0.13
4-11_estu_edad	0.13
56-11_fami_ocup_padre	0.13
54_primaria	0.13
55_primaria	0.12
36_Choco	0.12
21_Flexible	0.12
54_ninguno	0.12
55_ninguno	0.12

36_Caribe	0.11
15-11_estu_area_reside	0.11
25_NOCHE	0.10
59-11_fami_personas_hogar	0.09
47_MEDIO AMBIENTE	0.09
25_SABATINA - DOMINICAL	0.09
26_TECNICO	0.08
25_TARDE	0.07
16-11_estu_trabaja	0.06
47_VIOLENCIA Y SOCIEDAD	0.05
2-11_periodo_num	0.05
26_ACADEMICO Y TECNICO	0.05
36_Llanos	0.05
36_Amazonica	0.04
48-11_punt_comp_flexible	0.04
32_Caribe	0.04
54_bachiller	0.04
55_bachiller	0.04
36_Bolivar	0.04
36_Andina	0.04
55_No_reporta	0.04
32_Choco	0.03
26_NORMALISTA	0.03
33_1	0.03
32_Andina	0.02
54_no_sabe	0.02
55_no_sabe	0.02
33_0	0.02
36_SanAndres	0.02

47_PROFUNDIZACION EN LENGUAJE	0.02
32_Bolivar	0.02
32_Atlantico	0.01
33_2	0.01
32_Amazonica	0.01
36_Atlantico	0.01
32_Llanos	0.01
32_SanAndres	0.00
25_UNICA	0.00
36_Valle	0.00
32_Valle	0.00
26_DESCONOCIDO	0.00
33_No_reporta	0.00
32_No_reporta	0.00
32_EjeC	0.00
47_PROFUNDIZACION EN BIOLOGIA	-0.01
33_3	-0.01
32_Antioquia	-0.01
36_EjeC	-0.01
32_Santander	-0.01
21_A	-0.02
61-11_fami_celular	-0.03
32_Bogota	-0.04
47_PROFUNDIZACION EN CIENCIAS SOCIALES	-0.04
36_Antioquia	-0.04
36_Santander	-0.05
37-11_estu_estud-presentado-examen	-0.06
24-11_cole biling	-0.07
21_B	-0.08

66-11_fami_nevera	-0.09
22_F	-0.09
2-11_anno_num	-0.10
68-11_fami_dvd	-0.10
22_M	-0.10
26_ACADEMICO	-0.11
63-11_fami_servicio_television	-0.13
7-11_estu_genero	-0.14
65-11_fami_lavadora	-0.14
36_Bogota	-0.15
67-11_fami_horno	-0.15
47_PROFUNDIZACION EN MATEMATICA	-0.15
69-11_fami_microondas	-0.15
54_posgrado	-0.16
55_posgrado	-0.16
54_pregrado	-0.16
55_pregrado	-0.16
70-11_fami_automovil	-0.18
46_B+	-0.18
60-11_fami_telefono_fijo	-0.20
64-11_fami_computador	-0.21
62-11_fami_internet	-0.21
46_A2	-0.22
58-11_fami_pisos_hogar	-0.22
25_COMPLETA U ORDINARIA	-0.23
27-11_cole_valor_pension	-0.26
51-11_fami_nivel_sisben	-0.29
53-11_fami_ing_fmiliar_mensual	-0.31
52-11_fami_estrato_vivienda	-0.31

46_B1	-0.31
41-11_punt_filosofia	-0.52
45-11_punt_ingles	-0.53
44-11_punt_fisica	-0.55
43-11_punt_quimica	-0.58
42-11_punt_biologia	-0.64
40-11_punt_c_sociales	-0.69
38-11_punt_lenguaje	-0.72
39-11_punt_matematicas	-0.72

9.1.2 Correlaciones con el desempeño académico en las pruebas Saber Pro

Se definen en este apartado dos variables objetivo que representan consistentemente el rendimiento del estudiante en las Pruebas Saber Pro, acompañadas del diagrama de cajas y bigotes para facilitar la comprensión de la distribución de los datos y seguidas de la respectiva tabla de correlación.

Variable objetivo No. 1:

- 147-G_punt_global (puntaje total obtenido, rango [0,255]).

  Puntaje general obtenido por el estudiante en la evaluación de las competencias genéricas de las pruebas Saber Pro.

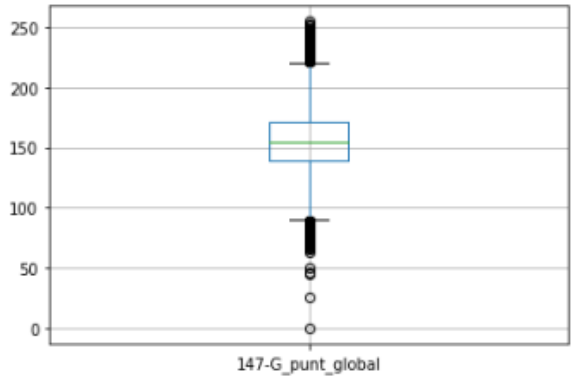


Figura 23. Diagrama de caja de la variable No. 147-Puntaje global del estudiante en las pruebas Saber Pro - competencias genéricas.

Correlaciones de todas las variables con el puntaje global obtenido por el estudiante en las pruebas Saber Pro – evaluación de competencias genéricas,

es decir comunes a todos los estudiantes sin discriminar los perfiles específicos de cada una de las carreras profesionales en curso.

Tabla 62. Correlaciones de todos los datos con el puntaje global obtenido en las competencias genéricas de las pruebas Saber Pro.

	147-G_punt_global
147-G_punt_global	1
131-G_lect_critica_punt	0.82
135-G_compet_ciudad_punt	0.78
139-G_ingles_punt	0.75
127-G_razona_cuant_punt	0.74
182-e_result_puntaje	0.62
39-11_punt_matematicas	0.61
40-11_punt_c_sociales	0.60
45-11_punt_ingles	0.60
143-G_com_escrita_punt	0.58
38-11_punt_lenguaje	0.57
42-11_punt_biologia	0.57
43-11_punt_quimica	0.52
140_B2	0.47
44-11_punt_fisica	0.47
41-11_punt_filosofia	0.45
53-11_fami_ing_fmiliar_mensual	0.38
52-11_fami_estrato_vivienda	0.37
46_B1	0.37
140_B1	0.36
157-G_fami_estratovivienda	0.35
51-11_fami_nivel_sisben	0.33
27-11_cole_valor_pension	0.32
173_NSE4	0.29

174_Mas de 7 millones	0.29
46_B+	0.29
121_UNIVERSIDAD	0.27
116_PRESENCIAL	0.27
25_COMPLETA U ORDINARIA	0.27
58-11_fami_pisos_hogar	0.26
62-11_fami_internet	0.26
64-11_fami_computador	0.25
60-11_fami_telefono_fijo	0.25
70-11_fami_automovil	0.22
167_MAS DE 100 LIBROS	0.22
164-G_fami_tienehornomicroogas	0.20
55_posgrado	0.20
54_posgrado	0.20
160_Si	0.20
69-11_fami_microondas	0.20
153_Postgrado	0.20
165-G_fami_tieneautomovil	0.19
36_Bogota	0.19
67-11_fami_horno	0.19
65-11_fami_lavadora	0.18
46_A2	0.18
55_pregrado	0.17
54_pregrado	0.17
154_Postgrado	0.17
167_26 A 100 LIBROS	0.17
21_B	0.16
63-11_fami_servicio_television	0.16
114_Bogota	0.16

162-G_fami_tienecomputador	0.16
107-G_estu_semestrecursa	0.15
111_INGENIERIA	0.15
154_Educacion profesional completa	0.15
153_Educacion profesional completa	0.14
73-G_estu_genero	0.14
7-11_estu_genero	0.13
47_PROFUNDIZACION EN MATEMATICA	0.13
26_ACADEMICO	0.13
111_MEDICINA	0.13
110-G_gruporeferencia	0.13
177-G_estu_pagomatriculapadres	0.13
68-11_fami_dvd	0.13
98_No realizo ninguna prueba de preparacion	0.13
22_M	0.13
163-G_fami_tienelavadora	0.13
24-11_cole biling	0.11
161-G_fami_tieneserviciotv	0.10
66-11_fami_nevera	0.10
156_Obrero o empleado de empresa particular	0.10
22_F	0.09
111_CIENCIAS NATURALES Y EXACTAS	0.09
93_Bachiller academico	0.09
181_DISENO DE PROCESOS INDUSTRIALES	0.09
122_NO OFICIAL - FUNDACION	0.09
111_ECONOMIA	0.09
181_ANALISIS ECONOMICO	0.09
36_Antioquia	0.09
114_Antioquia	0.08

122_OFICIAL NACIONAL	0.08
174_Entre 5.5 millones y menos de 7 millones	0.08
2-11_anno_num	0.08
181_DISENO DE SISTEMAS DE CONTROL	0.07
37-11_estu_estud-presentado-examen	0.07
87_No_reporta	0.07
170_0	0.07
169_Entre 1 y 3 horas	0.06
175-G_estu_pagomatriculabeca	0.06
153_Educacion profesional incompleta	0.06
154_Educacion profesional incompleta	0.06
122_OFICIAL MUNICIPAL	0.06
181_INVESTIGACION EN CIENCIAS SOCIALES	0.06
181_DISENO DE SISTEMAS PRODUCTIVOS Y LOGISTICOS	0.06
181_DISENO DE SISTEMAS MECANICOS	0.06
156_Pensionado	0.06
115_UNIVERSITARIO	0.06
47_PROFUNDIZACION EN CIENCIAS SOCIALES	0.05
111_DERECHO	0.05
181_INVESTIGACION JURIDICA	0.05
181_DISENO DE OBRAS DE INFRAESTRUCTURA	0.05
181_PENSAMIENTO CIENTIFICO - CIENCIAS DE LA TIERRA	0.05
181_PENSAMIENTO CIENTIFICO - CIENCIAS BIOLOGICAS	0.05
32_Bogota	0.05
174_Menos de 500 mil	0.04
181_PENSAMIENTO CIENTIFICO - CIENCIAS FISICAS	0.04
168_Mas de 2 horas	0.04
155_Pensionado	0.04

155_Obrero o empleado de empresa particular	0.04
181_FORMULACION DE PROYECTOS DE INGENIERIA	0.04
155_Obrero o empleado del gobierno	0.04
155_Patron o empleador	0.04
181_PENSAMIENTO CIENTIFICO - MATEMATICAS Y ESTADISTICA	0.04
156_Obrero o empleado del gobierno	0.04
36_Santander	0.04
111_HUMANIDADES	0.04
174_Entre 4 millones y menos de 5.5 millones	0.04
114_Santander	0.03
61-11_fami_celular	0.03
156_Patron o empleador	0.03
111_BELLAS ARTES Y DISENO	0.03
181_PENSAMIENTO CIENTIFICO - QUIMICA	0.03
173_NSE3	0.02
153_No sabe	0.02
111_COMUNICACION, PERIODISMO Y PUBLICIDAD	0.02
154_Tecnica o tecnologica completa	0.02
174_No paga matricula	0.02
32_Antioquia	0.02
153_Tecnica o tecnologica completa	0.02
168_Entre 30 y 60 minutos	0.02
168_Entre 1 y 2 horas	0.02
156_Trabajador familiar sin remuneracion	0.02
114_Valle	0.02
36_EjeC	0.02
71_CC	0.01
36_Valle	0.01
33_3	0.01

114_EjeC	0.01
32_Santander	0.01
155_Desempleado	0.01
71_PE	0.01
33_No_reporta	0.00
181_ATENCION EN SALUD	0.00
71_PC	0.00
32_EjeC	0.00
87_Sikuani	0.00
87_Tucano	0.00
87_Raizal	0.00
47_PROFUNDIZACION EN BIOLOGIA	0.00
32_No_reporta	0.00
26_DESCONOCIDO	0.00
90-G_estu_limita_condicionespecial	0.00
153_Tecnica o tecnologica incompleta	0.00
32_Valle	0.00
174_0	0.00
93_0	0.00
87_Comunidades Rom (Gitanas)	0.00
169_0	0.00
88-G_estu_limita_motriz	0.00
154_No sabe	0.00
87_Arhuaco	0.00
32_SanAndres	0.00
87_Pijao	0.00
87_Huitoto	0.00
47_PROFUNDIZACION EN LENGUAJE	0.00
89-G_estu_limita_invidente	0.00

87_Guambiano	0.00
168_No leo por entretenimiento	0.00
87_Cancuamo	0.00
71_CE	-0.01
154_Tecnica o tecnologica incompleta	-0.01
116_SEMI-PRESENCIAL	-0.01
87_Pasto	-0.01
111_CIENCIAS SOCIALES	-0.01
169_Mas de 4 horas	-0.01
71_TI	-0.01
87_Embera	-0.01
87_Paez	-0.01
181_PRODUCCION AGRICOLA	-0.01
87_Inga	-0.01
36_SanAndres	-0.01
98_0	-0.01
140_A2	-0.01
25_UNICA	-0.01
71_CR	-0.01
170_Entre 21 y 30 horas	-0.01
155_No aplica	-0.01
32_Llanos	-0.01
32_Amazonica	-0.01
170_Entre 11 y 20 horas	-0.01
87_Palenquero	-0.01
181_PRODUCCION PECUARIA	-0.02
122_OFICIAL DEPARTAMENTAL	-0.02
87_Zenu	-0.02
156_Jornalero o peon	-0.02

55_no_sabe	-0.02
155_Trabajador familiar sin remuneracion	-0.02
54_no_sabe	-0.02
33_2	-0.02
111_CIENCIAS AGROPECUARIAS	-0.02
33_0	-0.02
170_Mas de 30 horas	-0.02
32_Atlantico	-0.02
91-G_estu_limita_sordo	-0.02
155_Trabajador sin remuneracion en empresas o negocios de otros hogares	-0.02
32_Bolivar	-0.02
156_Trabajador sin remuneracion en empresas o negocios de otros hogares	-0.02
87_Otro grupo etnico minoritario	-0.02
32_Choco	-0.02
36_Atlantico	-0.02
87_Wayu	-0.02
111_PSICOLOGIA	-0.02
173_0	-0.02
150-G_fami_hogaractual	-0.03
93_Bachiller pedagogico o normalista	-0.03
156_Otra actividad u ocupacion	-0.03
155_Trabajador por cuenta propia	-0.03
155_Empleado domestico	-0.03
33_1	-0.03
32_Andina	-0.03
47_VIOLENCIA Y SOCIEDAD	-0.03
156_Desempleado	-0.03
156_No aplica	-0.03
87_Ninguno	-0.03

156_Trabajador por cuenta propia	-0.03
26_NORMALISTA	-0.04
116_DISTANCIA VITUAL	-0.04
55_No_reporta	-0.04
111_ENFERMERIA	-0.04
114_Amazonica	-0.04
121_INSTITUCION TECNOLOGICA	-0.04
87_Comunidad afrodescendiente	-0.05
114_Atlantico	-0.05
156_Otras	-0.05
77-G_periodo	-0.05
170_Menos de 10 horas	-0.05
36_Bolivar	-0.05
32_Caribe	-0.05
48-11_punt_comp_flexible	-0.05
168_30 minutos o menos	-0.05
155_Jornalero o peon	-0.05
36_Amazonica	-0.05
36_Llanos	-0.05
121_ESCUELA NORMAL SUPERIOR	-0.06
111_NORMALES SUPERIORES	-0.06
115_TECNOLOGIA	-0.06
114_Bolivar	-0.06
122_REGIMEN ESPECIAL	-0.06
154_Ninguno	-0.06
153_Secundaria (Bachillerato) incompleta	-0.06
55_bachiller	-0.06
181_INTERVENCION EN PROCESOS SOCIALES	-0.06
54_bachiller	-0.06

114_Llanos	-0.06
16-11_estu_trabaja	-0.06
174_Entre 500 mil y menos de 1 millon	-0.07
26_ACADEMICO Y TECNICO	-0.07
176-G_estu_pagomatriculacredito	-0.07
169_Menos de una hora	-0.07
36_Andina	-0.07
153_Ninguno	-0.07
21_A	-0.07
154_Secundaria (Bachillerato) completa	-0.07
98_Tomo un curso de preparacion	-0.07
153_Secundaria (Bachillerato) completa	-0.07
154_Secundaria (Bachillerato) incompleta	-0.07
98_Repaso por cuenta propia	-0.07
114_Andina	-0.07
167_11 A 25 LIBROS	-0.07
156_Empleado domestico	-0.08
25_SABATINA - DOMINICAL	-0.08
111_SALUD	-0.08
121_TECNICA PROFESIONAL	-0.08
25_NOCHE	-0.08
111_ADMINISTRACION Y AFINES	-0.08
93_Bachiller tecnico	-0.08
155_Otras	-0.08
26_TECNICO	-0.09
174_Entre 2.5 millones y menos de 4 millones	-0.09
25_TARDE	-0.09
21_Flexible	-0.09
181_GESTION DE ORGANIZACIONES	-0.09

181_GESTION FINANCIERA	-0.09
36_Choco	-0.09
114_Choco	-0.10
59-11_fami_personas_hogar	-0.11
173_NSE2	-0.11
153_Primaria completa	-0.11
111_CONTADURIA Y AFINES	-0.11
154_Primaria completa	-0.11
15-11_estu_area_reside	-0.11
47_MEDIO AMBIENTE	-0.11
151-G_fami_cabezafamilia	-0.12
159-G_creado_fami_hacina	-0.12
86-G_estu_tieneetnia	-0.12
55_ninguno	-0.13
153_Primaria incompleta	-0.13
154_Primaria incompleta	-0.13
54_ninguno	-0.13
4-11_estu_edad	-0.14
2-11_periodo_num	-0.14
55_primaria	-0.14
54_primaria	-0.14
22_Mixto	-0.14
111_EDUCACION	-0.14
178-G_estu_pagomatriculapropio	-0.15
56-11_fami_ocup_padre	-0.15
181_ENSENAR	-0.15
122_NO OFICIAL - CORPORACION	-0.15
36_Caribe	-0.16
76-G-creado_estu_edad	-0.16

114_Caribe	-0.16
25_MANANA	-0.17
166-G_fami_tienemotocicleta	-0.17
57-11_fami_ocup_madre	-0.17
152-G_fami_numpersonasacargo_num	-0.18
174_Entre 1 millon y menos de 2.5 millones	-0.19
160_No	-0.20
173_NSE1	-0.20
46_A1	-0.25
121_INSTITUCION UNIVERSITARIA	-0.25
167_0 A 10 LIBROS	-0.26
116_DISTANCIA	-0.27
23-11_cole_naturaleza	-0.29
46_A-	-0.35
140_A1	-0.37
140_A-	-0.41
50-11_estu_puesto	-0.72

Variable objetivo No. 2:

- 182-E_result_puntaje (Puntaje de la prueba específica, rango [0, 300]).

  Puntaje general obtenido por el estudiante en la evaluación de la competencia específica.

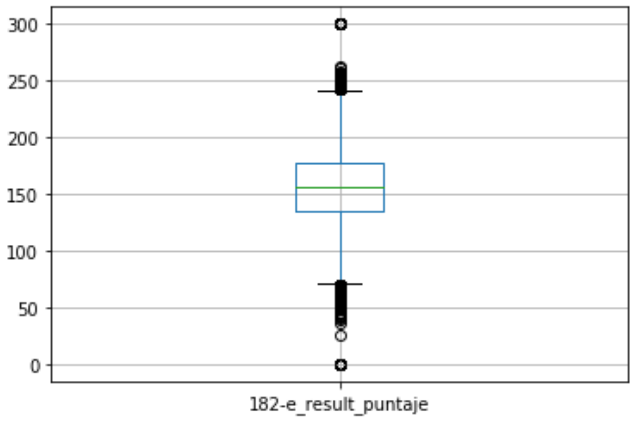


Figura 23. Diagrama de caja de la variable No. 182-Puntaje del estudiante en las pruebas Saber Pro – competencia específica.

Correlaciones de todas las variables con el puntaje global obtenido por el estudiante en las pruebas Saber Pro – evaluación de la competencia específica, de acuerdo con la carrera universitaria en curso.

Tabla 63. Correlaciones de todos los datos con el puntaje de la competencia específica a la carrera profesional en curso.

	182-e_result_puntaje
182-e_result_puntaje	1
147-G_punt_global	0.62
131-G_lect_critica_punt	0.57
135-G_compet_ciudad_punt	0.54
127-G_razona_cuant_punt	0.52
40-11_punt_c_sociales	0.46
139-G_ingles_punt	0.44
39-11_punt_matematicas	0.44
42-11_punt_biologia	0.43
38-11_punt_lenguaje	0.42
43-11_punt_quimica	0.38
45-11_punt_ingles	0.37
41-11_punt_filosofia	0.35
44-11_punt_fisica	0.34
140_B2	0.28
143-G_com_escrita_punt	0.24
46_B1	0.22
140_B1	0.22
121_UNIVERSIDAD	0.20
53-11_fami_ing_fmiliar_mensual	0.19
52-11_fami_estrato_vivienda	0.19
116_PRESENCIAL	0.19
157-G_fami_estratovivienda	0.18

174_Mas de 7 millones	0.17
51-11_fami_nivel_sisben	0.17
46_B+	0.17
27-11_cole_valor_pension	0.15
173_NSE4	0.14
25_COMPLETA U ORDINARIA	0.14
58-11_fami_pisos_hogar	0.14
167_MAS DE 100 LIBROS	0.14
60-11_fami_telefono_fijo	0.13
62-11_fami_internet	0.13
64-11_fami_computador	0.13
46_A2	0.12
167_26 A 100 LIBROS	0.12
55_posgrado	0.11
54_posgrado	0.11
36_Bogota	0.11
160_Si	0.11
153_Postgrado	0.11
174_Menos de 500 mil	0.10
111_MEDICINA	0.10
107-G_estu_semestrecursa	0.10
111_CIENCIAS NATURALES Y EXACTAS	0.10
154_Postgrado	0.10
162-G_fami_tienecomputador	0.10
70-11_fami_automovil	0.10
122_OFICIAL NACIONAL	0.09
114_Bogota	0.09
73-G_estu_genero	0.09
114_Antioquia	0.09

36_Antioquia	0.09
98_No realizo ninguna prueba de preparacion	0.09
7-11_estu_genero	0.09
47_PROFUNDIZACION EN MATEMATICA	0.09
164-G_fami_tienehornomicroogas	0.09
21_B	0.09
69-11_fami_microondas	0.09
67-11_fami_horno	0.09
55_pregrado	0.08
54_pregrado	0.08
65-11_fami_lavadora	0.08
22_M	0.08
175-G_estu_pagomatriculabeca	0.08
165-G_fami_tieneautomovil	0.08
63-11_fami_servicio_television	0.07
111_CONTADURIA Y AFINES	0.07
156_Obrero o empleado de empresa particular	0.07
181_GESTION FINANCIERA	0.06
154_Educacion profesional completa	0.06
153_Educacion profesional completa	0.06
24-11_cole biling	0.06
181_PENSAMIENTO CIENTIFICO - CIENCIAS BIOLOGICAS	0.06
122_OFICIAL MUNICIPAL	0.06
2-11_anno_num	0.06
68-11_fami_dvd	0.06
26_ACADEMICO	0.05
177-G_estu_pagomatriculapadres	0.05
181_PENSAMIENTO CIENTIFICO - CIENCIAS FISICAS	0.05
163-G_fami_tienelavadora	0.05

22_F	0.04
66-11_fami_nevera	0.04
181_PENSAMIENTO CIENTIFICO - MATEMATICAS Y ESTADISTICA	0.04
47_PROFUNDIZACION EN CIENCIAS SOCIALES	0.04
169_Entre 1 y 3 horas	0.04
181_INVESTIGACION JURIDICA	0.04
111_DERECHO	0.04
168_Mas de 2 horas	0.04
87_No_reporta	0.04
181_PENSAMIENTO CIENTIFICO - CIENCIAS DE LA TIERRA	0.04
161-G_fami_tieneserviciotv	0.04
153_Educacion profesional incompleta	0.03
122_OFICIAL DEPARTAMENTAL	0.03
37-11_estu_estud-presentado-examen	0.03
154_Educacion profesional incompleta	0.03
156_Trabajador familiar sin remuneracion	0.03
114_Valle	0.03
181_PENSAMIENTO CIENTIFICO - QUIMICA	0.03
155_Obrero o empleado de empresa particular	0.03
174_No paga matricula	0.03
32_Bogota	0.03
36_Valle	0.03
115_UNIVERSITARIO	0.03
156_Pensionado	0.03
174_Entre 5.5 millones y menos de 7 millones	0.03
170_0	0.02
32_Antioquia	0.02
93_Bachiller academico	0.02
153_No sabe	0.02

168_Entre 30 y 60 minutos	0.02
155_Pensionado	0.02
155_Obrero o empleado del gobierno	0.02
122_NO OFICIAL - FUNDACION	0.02
36_Santander	0.01
181_PRODUCCION AGRICOLA	0.01
111_HUMANIDADES	0.01
156_Obrero o empleado del gobierno	0.01
168_Entre 1 y 2 horas	0.01
155_Desempleado	0.01
111_CIENCIAS SOCIALES	0.01
36_EjeC	0.01
181_DISENO DE SISTEMAS PRODUCTIVOS Y LOGISTICOS	0.01
71_CC	0.01
114_Santander	0.01
170_Mas de 30 horas	0.01
61-11_fami_celular	0.01
155_No aplica	0.01
154_Tecnica o tecnologica completa	0.01
153_Tecnica o tecnologica completa	0.01
110-G_gruporeferencia	0.01
181_DISENO DE SISTEMAS MECANICOS	0.01
111_CIENCIAS AGROPECUARIAS	0.01
181_INVESTIGACION EN CIENCIAS SOCIALES	0.01
33_3	0.01
114_EjeC	0.01
155_Patron o empleador	0.01
32_Valle	0.00
111_EDUCACION	0.00

156_Patron o empleador	0.00
87_Tucano	0.00
32_EjeC	0.00
71_PE	0.00
90-G_estu_limita_condicionespecial	0.00
181_DISENO DE SISTEMAS DE CONTROL	0.00
71_PC	0.00
173_0	0.00
32_Santander	0.00
181_ENSENAR	0.00
32_No_reporta	0.00
47_PROFUNDIZACION EN LENGUAJE	0.00
26_DESCONOCIDO	0.00
33_No_reporta	0.00
87_Huitoto	0.00
181_PRODUCCION PECUARIA	0.00
87_Sikuani	0.00
93_Bachiller pedagogico o normalista	0.00
111_BELLAS ARTES Y DISENO	0.00
32_SanAndres	0.00
89-G_estu_limita_invidente	0.00
87_Cancuamo	0.00
88-G_estu_limita_motriz	0.00
87_Guambiano	0.00
87_Pasto	0.00
47_PROFUNDIZACION EN BIOLOGIA	0.00
150-G_fami_hogaractual	0.00
87_Pijao	0.00
116_SEMI-PRESENCIAL	0.00

87_Raizal	0.00
174_0	0.00
93_0	0.00
87_Paez	0.00
71_CE	0.00
87_Inga	0.00
87_Embera	0.00
169_0	0.00
25_UNICA	0.00
154_No sabe	-0.01
87_Comunidades Rom (Gitanas)	-0.01
87_Arhuaco	-0.01
153_Tecnica o tecnologica incompleta	-0.01
155_Trabajador familiar sin remuneracion	-0.01
156_Jornalero o peon	-0.01
156_Desempleado	-0.01
33_0	-0.01
87_Zenu	-0.01
71_CR	-0.01
173_NSE3	-0.01
71_TI	-0.01
168_No leo por entretenimiento	-0.01
174_Entre 4 millones y menos de 5.5 millones	-0.01
170_Entre 11 y 20 horas	-0.01
154_Tecnica o tecnologica incompleta	-0.01
170_Entre 21 y 30 horas	-0.01
140_A2	-0.01
181_ATENCION EN SALUD	-0.01
26_NORMALISTA	-0.01

155_Jornalero o peon	-0.01
32_Llanos	-0.01
98_0	-0.01
32_Amazonica	-0.01
87_Palenquero	-0.01
155_Trabajador por cuenta propia	-0.01
33_2	-0.01
181_DISENO DE PROCESOS INDUSTRIALES	-0.01
36_SanAndres	-0.01
55_no_sabe	-0.01
54_no_sabe	-0.01
91-G_estu_limita_sordo	-0.01
111_ENFERMERIA	-0.01
111_COMUNICACION, PERIODISMO Y PUBLICIDAD	-0.01
156_Trabajador sin remuneracion en empresas o negocios de otros hogares	-0.01
87_Otro grupo etnico minoritario	-0.02
156_No aplica	-0.02
181_FORMULACION DE PROYECTOS DE INGENIERIA	-0.02
32_Choco	-0.02
32_Andina	-0.02
169_Menos de una hora	-0.02
32_Atlantico	-0.02
155_Trabajador sin remuneracion en empresas o negocios de otros hogares	-0.02
32_Bolivar	-0.02
55_No_reporta	-0.02
47_VIOLENCIA Y SOCIEDAD	-0.02
114_Amazonica	-0.02
155_Empleado domestico	-0.02
87_Comunidad afrodescendiente	-0.02

33_1	-0.02
153_Secundaria (Bachillerato) incompleta	-0.02
93_Bachiller tecnico	-0.02
87_Ninguno	-0.02
87_Wayu	-0.02
156_Otra actividad u ocupacion	-0.02
174_Entre 500 mil y menos de 1 millon	-0.02
111_INGENIERIA	-0.03
26_ACADEMICO Y TECNICO	-0.03
156_Trabajador por cuenta propia	-0.03
111_NORMALES SUPERIORES	-0.03
115_TECNOLOGIA	-0.03
121_ESCUELA NORMAL SUPERIOR	-0.03
122_REGIMEN ESPECIAL	-0.03
181_ANALISIS ECONOMICO	-0.03
111_ECONOMIA	-0.03
154_Secundaria (Bachillerato) incompleta	-0.03
36_Amazonica	-0.03
48-11_punt_comp_flexible	-0.03
21_A	-0.03
16-11_estu_trabaja	-0.03
77-G_periodo	-0.03
36_Atlantico	-0.03
169_Mas de 4 horas	-0.03
181_DISENO DE OBRAS DE INFRAESTRUCTURA	-0.03
170_Menos de 10 horas	-0.04
154_Ninguno	-0.04
153_Ninguno	-0.04
26_TECNICO	-0.04

116_DISTANCIA VITUAL	-0.04
154_Secundaria (Bachillerato) completa	-0.04
156_Otras	-0.04
32_Caribe	-0.04
36_Andina	-0.04
156_Empleado domestico	-0.04
25_TARDE	-0.04
98_Repaso por cuenta propia	-0.04
36_Bolivar	-0.04
114_Andina	-0.04
55_bachiller	-0.04
54_bachiller	-0.04
168_30 minutos o menos	-0.04
121_INSTITUCION TECNOLOGICA	-0.04
25_SABATINA - DOMINICAL	-0.05
167_11 A 25 LIBROS	-0.05
181_INTERVENCION EN PROCESOS SOCIALES	-0.05
36_Llanos	-0.05
114_Atlantico	-0.05
114_Bolivar	-0.05
153_Secundaria (Bachillerato) completa	-0.05
153_Primaria incompleta	-0.05
154_Primaria completa	-0.05
25_NOCHE	-0.05
55_ninguno	-0.05
173_NSE2	-0.05
159-G_creado_fami_hacina	-0.05
54_ninguno	-0.05
114_Llanos	-0.05

111_ADMINISTRACION Y AFINES	-0.05
59-11_fami_personas_hogar	-0.05
153_Primaria completa	-0.06
176-G_estu_pagomatriculacredito	-0.06
15-11_estu_area_reside	-0.06
21_Flexible	-0.06
121_TECNICA PROFESIONAL	-0.06
154_Primaria incompleta	-0.06
181_GESTION DE ORGANIZACIONES	-0.06
155_Otras	-0.06
36_Choco	-0.06
98_Tomo un curso de preparacion	-0.06
55_primaria	-0.06
54_primaria	-0.07
111_PSICOLOGIA	-0.07
86-G_estu_tieneetnia	-0.07
151-G_fami_cabezafamilia	-0.07
178-G_estu_pagomatriculapropio	-0.07
114_Choco	-0.07
2-11_periodo_num	-0.07
56-11_fami_ocup_padre	-0.07
22_Mixto	-0.08
173_NSE1	-0.08
47_MEDIO AMBIENTE	-0.08
111_SALUD	-0.08
25_MANANA	-0.09
57-11_fami_ocup_madre	-0.09
4-11_estu_edad	-0.09
174_Entre 2.5 millones y menos de 4 millones	-0.10

166-G_fami_tienemotocicleta	-0.10
76-G-creado_estu_edad	-0.11
152-G_fami_numpersonasacargo_num	-0.11
160_No	-0.11
114_Caribe	-0.11
36_Caribe	-0.12
174_Entre 1 millón y menos de 2.5 millones	-0.13
122_NO OFICIAL - CORPORACION	-0.13
23-11_cole_naturaleza	-0.14
46_A1	-0.14
167_0 A 10 LIBROS	-0.17
121_INSTITUCION UNIVERSITARIA	-0.18
116_DISTANCIA	-0.19
140_A1	-0.23
140_A-	-0.23
46_A-	-0.23
50-11_estu_puesto	-0.54

9.2 Definición de perfiles de estudiantes

Para la definición de perfiles de estudiantes, se aplica un análisis de clustering mediante el algoritmo K-means traducido a K-medias en español. Este algoritmo se basa en la distancia euclídea, lo que significa que todas las variables que se le ingresan deben ser de tipo numérico, por lo cual se recurrió en el capítulo 8 de esta investigación a la conversión de los datos categóricos a numéricos a través de la creación de variables de tipo dummy.

La salida del modelo K-means:

Se obtienen de este modelo las agrupaciones de los registros conocidas como clústeres, de la siguiente manera:

- El data set completo con una columna en la que se identifica el clúster

asignado a cada uno de los registros

- Un data set que contiene los valores medios o centroides atribuibles a cada una de las variables para cada uno de los clústeres que se le solicitaron al modelo.

Lo anterior simplifica el proceso de perfilar grandes volúmenes de datos, que no se pueden agrupar de forma intuitiva, lo cual es justo lo que se requiere en este trabajo.

¿Cómo pedirle al modelo el número adecuado de clústeres?

En la modelación del algoritmo, para cada uno de los casos se recurrió a la gráfica conocida como la curva de codo, la cual revela la cantidad adecuada de clústeres que arrojan el mejor desempeño en la evaluación del modelo.

Anotación: el valor ideal de clústeres se encuentra en donde la gráfica señala un codo, dado que a partir de ese punto el incremento en la cantidad de agrupaciones no mejorará los indicadores de la evaluación del método.

Evaluación de los modelos de clustering

En el modelamiento de este algoritmo se busca que la separabilidad entre los clústeres generados sea tan grande como sea posible, para garantizar que los registros pertenecen a grupos diferentes y no deben estar dentro de una misma agrupación y a su vez, que la cohesión entre los registros que permanecen dentro de un mismo clúster revele que existe gran afinidad entre los mismos.

- El índice Davies and Bouldin: En su nivel óptimo debe ser lo más pequeño posible. Lo cual indica que la cohesión es pequeña (ideal) y la separabilidad es grande (ideal).

- El índice de Silueta es su nivel óptimo debe tender a 1. En el caso que se cumpla lo anterior, significa que en el modelo seleccionado se da una separabilidad grande entre los distintos clústeres y una cohesión pequeña, la cual indica que los registros que se clasificaron dentro de las mismas agrupaciones; en efecto corresponden a datos que tienen gran compatibilidad entre sí.

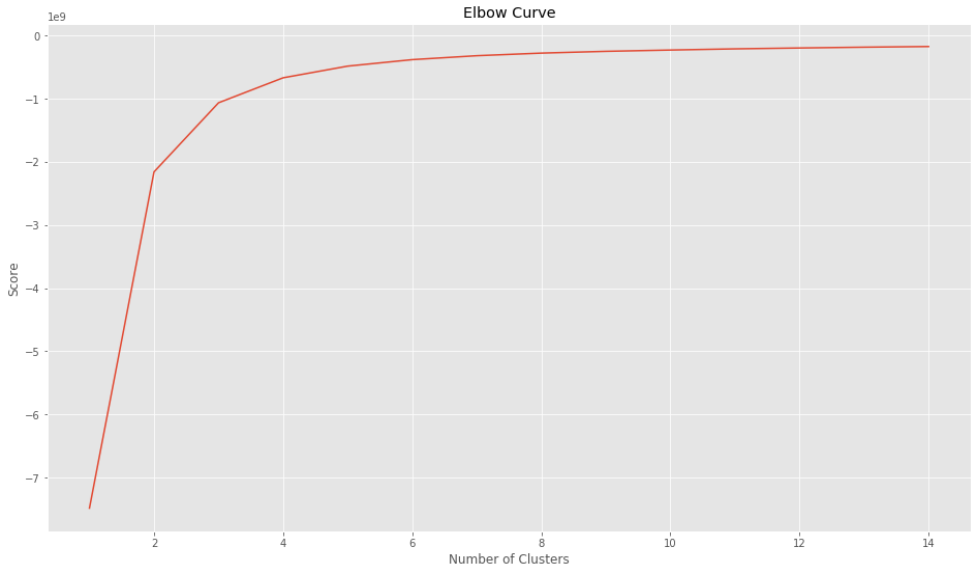
9.2.1 Clustering de los estudiantes que presentaron las pruebas Saber

11.

A continuación, se presenta un clustering para perfilar a los estudiantes cuando están en el último año de sus estudios de educación media. Para el modelamiento se ingresaron todos los datos de las Pruebas Saber 11, luego de surtir todos los pasos documentados en la preparación de los datos.

Se inició con la creación de la curva de codo para encontrar el número apropiado de las agrupaciones que se harán sobre los registros, buscando el mejor desempeño del modelo.

Figura 24. Curva de codo para la selección del número de clústeres. Datos de las Pruebas Saber 11.



Como se aprecia en la gráfica anterior, el número ideal de clústeres es igual a 3 sin embargo, se procedió a evaluar el modelo para los siguientes números de agrupaciones = 3, 4, 5 y 6.

A continuación, se relacionan los indicadores de desempeño del modelo, obtenidos para las distintas agrupaciones modeladas.

Tabla 64. Evaluación del modelo K-means para los datos de Saber 11.

Número de clústeres	Evaluación del modelo K-medias	
	Davies and Bouldin	Silueta

3	0.57	0.56	Mejor desempeño del modelo
4	0.62	0.50	
5	0.66	0.46	
6	0.71	0.42	

La interpretación de estos resultados se puede comprender guiándose por la última parte citada en la introducción a los modelos descriptivos de este proyecto.

La evaluación del modelo revela que no es necesario jugar con el rango de números de clústeres que arroja la curva de codo, dado que el punto óptimo se da en el quiebre de la misma. El cual es igual a 3 clústeres para el presente caso.

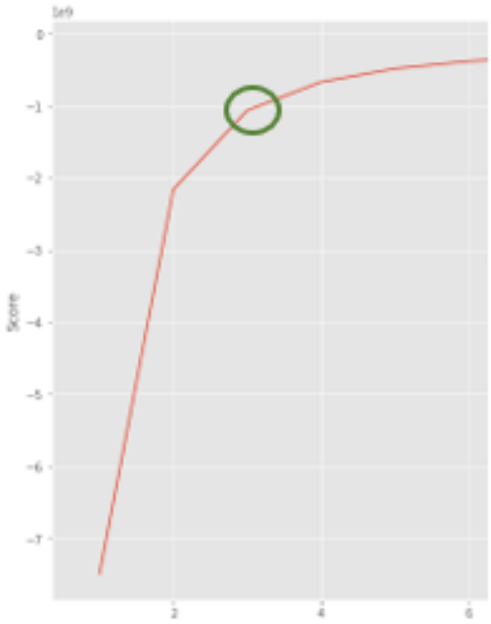


Figura 25. Punto óptimo del modelo k-means para las pruebas Saber 11

A continuación, en la tabla 65 se detalla el tamaño de los clusters.

Tabla 65. Perfilado del modelo k-means para Saber 11.

Número de clústeres	Cantidad de registros en cada clúster		
	6	4	3

			(número de clústeres seleccionados para el modelo)
1	42848	58230	39512
2	30758	36782	20035
3	22519	23465	72520
4	16650	13590	
5	11794		
6	7498		

A continuación, se detallan los centroides arrojados por el modelo para los datos de las pruebas Saber 11, indicando un nombre para cada grupo según las características del centriode.

Tabla 66. Centroides k-means. Modelo Saber 11.

	Clúster 1	**Cluster 2**	**Cluster 3**
Nombre del cluster	**Intermedios**	**Rezagados**	**Pilos**
Tamaño del cluster	**39512**	**20035**	**72520**
4-11_estu_edad	16.31	16.75	16.14
7-11_estu_genero	0.34	0.31	0.46
15-11_estu_area_reside	0.14	0.17	0.09
23-11_cole_naturaleza	0.67	0.75	0.48
27-11_cole_valor_pension	3.54	2.69	5.76
38-11_punt_lenguaje	49.83	42.58	58.74
39-11_punt_matematicas	49.76	41.88	61.91
40-11_punt_c_sociales	48.68	41.95	58.04
41-11_punt_filosofia	45.73	40.04	53.38
42-11_punt_biologia	48.88	42.91	57.46
43-11_punt_quimica	48.42	43.86	56.41
44-11_punt_fisica	47.05	42.60	55.47
45-11_punt_ingles	46.58	42.63	58.19
46_A-	0.27	0.45	0.06
46_A1	0.56	0.48	0.36
46_A2	0.13	0.05	0.27
46_B+	0.00	0.00	0.07
46_B1	0.04	0.01	0.25
47_Medio Ambiente	0.28	0.32	0.22
47_Profundizacion en Biología	0.13	0.14	0.14

47_Profundizacion en Ciencias Sociales	0.07	0.06	0.09
47_Profundizacion en Lenguaje	0.14	0.14	0.12
47_Profundizacion en Matemática	0.12	0.08	0.20
47_Violencia Y Sociedad	0.26	0.26	0.21
48-11_punt_comp_flexible	29.80	29.40	27.47
50-11_estu_puesto	352.00	705.79	91.26
51-11_fami_nivel_sisben	3.09	2.60	3.79
52-11_fami_estrato_vivienda	2.26	1.97	2.80
53-11_fami_ing_fmiliar_mensual	2.50	2.19	3.25

9.2.2 Clustering de los estudiantes que presentaron la pruebas Saber Pro

Se presenta en este apartado el clustering para perfilar a los estudiantes que presentaron las pruebas Saber Pro y sobre quienes, en el apartado anterior venimos analizando sus perfiles desde que estaban en la educación media. Para el modelamiento se ingresaron los datos de las Pruebas Saber Pro en búsqueda de cumplir con el objetivo citado arriba. Se inició con la creación de la curva de codo para encontrar el número apropiado de las agrupaciones que se harán sobre los registros, buscando el mejor desempeño del modelo.

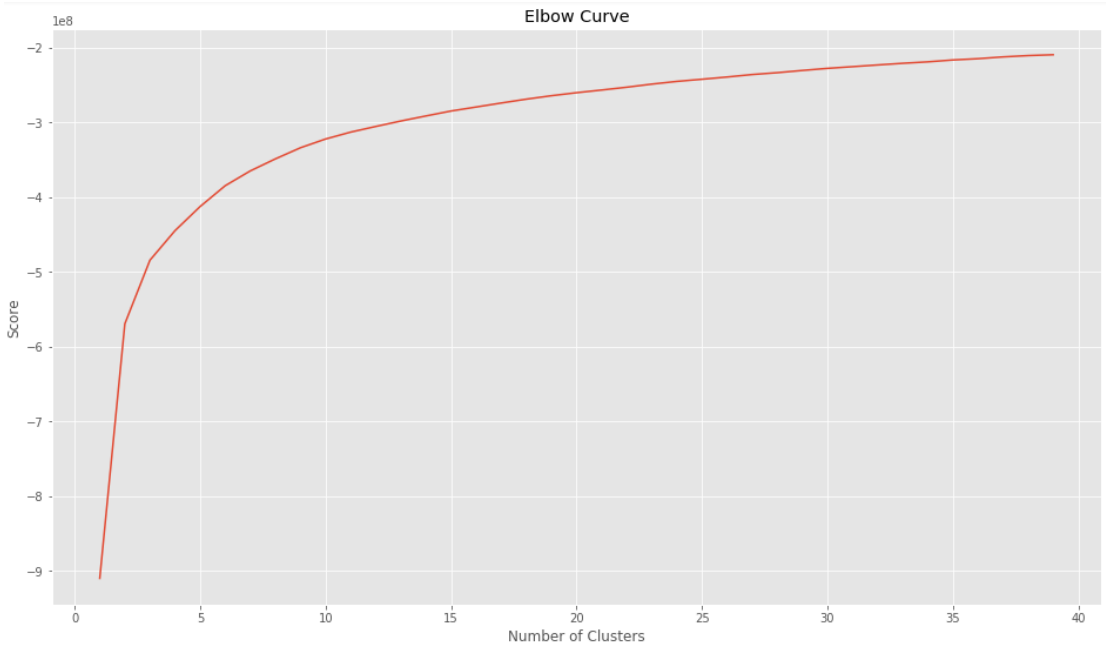


Figura 26. Curva de codo para la selección del número de clústeres. Datos de las Pruebas Saber Pro.

La información reportada a continuación muestra la dificultad del modelo para encontrar diferencias sustanciales entre los perfiles de los estudiantes que cursan una carrera profesional.

La evaluación óptima del modelo se da con dos clústeres, sin embargo se decide tomar la segunda mejor evaluación para poder desagregar mejor el conjunto de datos analizados.

Tabla 67. Evaluación del modelo K-means para los datos de Saber Pro.

Número de clústeres	Evaluación del modelo K-medias		
	Davies and Bouldin	Silueta	
2	1.20	0.31	
3	1.50	0.21	Se escoge el modelo con el segundo mejor desempeño
5	1.61	0.20	
6	1.58	0.16	
8	1.63	0.15	
10	1.58	0.15	

A continuación, se detalla el tamaño de los clusters en la tabla 68.

Tabla 68. Perfilado del modelo k-means para Saber Pro.

Número de clústeres	Cantidad de registros en cada clúster
	3 (número de clústeres seleccionados para el modelo)
1	56207
2	37690
3	38170

A continuación, en la tabla 69 se detallan los centroides arrojados por el modelo para los datos de las pruebas Saber Pro, teniendo en cuenta los que facilitan la diferenciación de los grupos encontrados.

Tabla 69. Centroides k-means. Modelo Saber Pro.

	Clúster 1	**Clúster 2**	**Clúster 3**
Nombre del cluster	**Desempeño medio**	**Desempeño alto**	**Desempeño bajo**
Tamaño del cluster	**56207**	**37690**	**38170**
73-G_estu_genero	0.38	0.50	0.33
76-G-creado_estu_edad	22.03	21.87	22.55
111_ADMINISTRACION Y AFINES	0.18	0.16	0.23
111_BELLAS ARTES Y DISENO	0.00	0.00	0.00
111_CIENCIAS AGROPECUARIAS	0.02	0.01	0.01
111_CIENCIAS NATURALES Y EXACTAS	0.02	0.05	0.01
111_CIENCIAS SOCIALES	0.03	0.03	0.03
111_COMUNICACION, PERIODISMO Y PUBLICIDAD	0.00	0.00	0.00
111_CONTADURIA Y AFINES	0.08	0.04	0.09
111_DERECHO	0.12	0.13	0.09
111_ECONOMIA	0.02	0.03	0.01
111_EDUCACION	0.10	0.06	0.16
111_ENFERMERIA	0.02	0.01	0.03
111_HUMANIDADES	0.00	0.00	0.00
111_INGENIERIA	0.29	0.35	0.21
111_MEDICINA	0.03	0.07	0.01
111_NORMALES SUPERIORES	0.00	0.00	0.01
111_PSICOLOGIA	0.02	0.01	0.03
111_SALUD	0.06	0.03	0.08
127-G_razona_cuant_punt	158.13	186.66	131.20
131-G_lect_critica_punt	158.29	188.33	125.09
135-G_compet_ciudad_punt	153.60	182.95	120.71
139-G_ingles_punt	154.66	190.62	132.29
140_A-	0.09	0.00	0.33
140_A1	0.28	0.03	0.45

140_A2	0.36	0.14	0.17
140_B1	0.24	0.48	0.05
140_B2	0.03	0.34	0.00
143-G_com_escrita_punt	152.78	172.26	129.27
147-G_punt_global	155.49	184.16	127.71
147-newCuantilPuestoGlobalSP	2.54	3.89	1.17
157-G_fami_estratovivienda	2.60	3.22	2.24
173_0	0.03	0.02	0.03
173_NSE1	0.19	0.10	0.28
173_NSE2	0.31	0.23	0.35
173_NSE3	0.27	0.26	0.23
173_NSE4	0.19	0.40	0.11
174_0	0.00	0.00	0.00
174_Entre 1 millon y menos de 2.5 millones	0.19	0.12	0.32
174_Entre 2.5 millones y menos de 4 millones	0.22	0.13	0.23
174_Entre 4 millones y menos de 5.5 millones	0.13	0.11	0.09
174_Entre 5.5 millones y menos de 7 millones	0.06	0.08	0.04
174_Entre 500 mil y menos de 1 millon	0.13	0.10	0.16
174_Mas de 7 millones	0.07	0.24	0.03
174_Menos de 500 mil	0.19	0.21	0.15
174_No paga matricula	0.00	0.01	0.00
181_ANALISIS ECONOMICO	0.02	0.03	0.01
181_ATENCION EN SALUD	0.11	0.12	0.12
181_DISENO DE OBRAS DE INFRAESTRUCTURA	0.05	0.06	0.04
181_DISENO DE PROCESOS INDUSTRIALES	0.01	0.02	0.00
181_DISENO DE SISTEMAS DE CONTROL	0.02	0.04	0.01

181_DISENO DE SISTEMAS MECANICOS	0.02	0.03	0.01
181_DISENO DE SISTEMAS PRODUCTIVOS Y LOGISTICOS	0.07	0.08	0.05
181_ENSENAR	0.11	0.07	0.16
181_FORMULACION DE PROYECTOS DE INGENIERIA	0.12	0.13	0.10
181_GESTION DE ORGANIZACIONES	0.16	0.13	0.21
181_GESTION FINANCIERA	0.10	0.06	0.11
181_INTERVENCION EN PROCESOS SOCIALES	0.03	0.02	0.05
181_INVESTIGACION EN CIENCIAS SOCIALES	0.02	0.03	0.01
181_INVESTIGACION JURIDICA	0.12	0.13	0.09
181_PENSAMIENTO CIENTIFICO - CIENCIAS BIOLOGICAS	0.01	0.02	0.01
181_PENSAMIENTO CIENTIFICO - CIENCIAS DE LA TIERRA	0.00	0.01	0.00
181_PENSAMIENTO CIENTIFICO - CIENCIAS FISICAS	0.00	0.01	0.00
181_PENSAMIENTO CIENTIFICO - MATEMATICAS Y ESTADISTICA	0.00	0.01	0.00

181_PENSAMIENTO CIENTIFICO - QUIMICA	0.00	0.01	0.00
181_PRODUCCION AGRICOLA	0.01	0.00	0.00
181_PRODUCCION PECUARIA	0.01	0.01	0.01
182-e_result_puntaje	156.95	183.12	128.65
182-NewCuantil_Epunt	2.54	3.46	1.53

9.2.3 Clustering con los datos de las pruebas Saber 11 y Saber Pro

Este modelo se basa en el perfilado conjunto de los estudiantes de educación media que hacen el tránsito a la educación superior, basado en los datos tomados en dos momentos de la vida académica de los estudiantes, los datos tenidos en cuenta corresponden a:

- Estudiantes que presentaron las pruebas Saber 11 antes de finalizar su educación media.
- Los mismos estudiantes anteriormente citados y para quienes también se tomaron en cuenta un número significativo de variables al momento en que estaban presentando la prueba de estado Saber Pro.

El punto óptimo de clústeres encontrado fue igual a 4, basados en la gráfica de la curva de codo que se muestra a continuación.

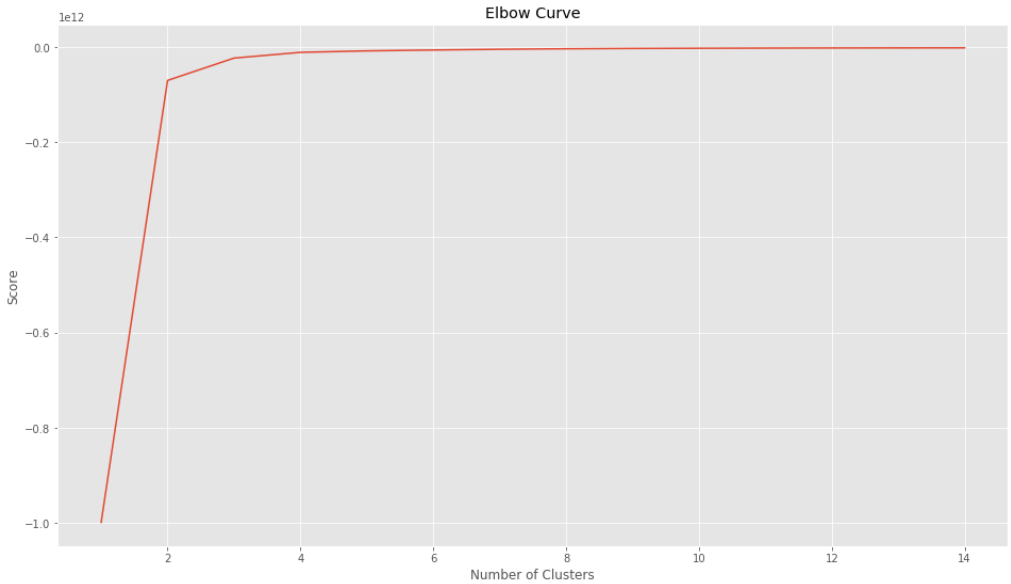


Figura 27. Curva de codo para la selección del número de clústeres. Datos de las Pruebas Saber 11 y Saber Pro.

El modelo K-MEANS, se puede medir a través de los siguientes dos indicadores, para los cuales se muestran los resultados configurando en el modelamiento de los datos un número de agrupaciones igual a 4.

- Davies and Bouldin = 0.35
- Silueta = 0.77

Los clústeres quedaron conformados con la siguiente cantidad de registros en cada uno de los casos:

Tabla 70. Perfilado del modelo k-means para Saber Pro.

Número de clústeres	Cantidad de registros en cada clúster
	4 (número de clústeres seleccionados para el modelo)
1	36424
2	41045
3	17777
4	36821

A continuación, se detallan los centroides arrojados por el modelo para perfilar de forma conjunta a los estudiantes de la educación media y superior, teniendo en cuenta los que facilitan la diferenciación de los grupos encontrados.

En este modelo se incluyeron de forma conjunta todos los datos de las pruebas Saber 11 y Saber Pro, sin embargo los centroides se van a mostrar de forma ordenada, primero todos los que corresponden a los datos de las Saber 11 y luego a los de Saber Pro.

<u>Saber 11: La identificación de la prueba y el estudiante</u>

Tabla 71. Centroides k-means. Modelo con todos los datos.

Agrupación de variables				
Clúster	1	2	3	4
4-11_estu_edad	16.38	16.08	16.36	16.37
7-11_estu_genero	0.27	0.59	0.34	0.35

Saber 11: El colegio en el que estudió el inscrito

Tabla 72. Centroides k-means. Modelo con todos los datos.

Agrupación de variables				
Clúster	1	2	3	4
23-11_cole_naturaleza	0.66	0.54	0.5	0.58
24-11_cole biling	0.02	0.02	0.03	0.03
27-11_cole_valor_pension	3.75	5.19	5.48	4.46

Saber 11: La intención del estudiante en cuanto a la elección de una carrera profesional

Para el tipo de carrera que desea estudiar el inscrito (ninguna, técnica, tecnológica, profesional), solo el 10.34% estudiantes eligieron una opción, dentro de la cual mayoritariamente todos seleccionaron “profesional”, por lo cual no hay diferencia entre los distintos clústeres para este criterio. Los demás atributos de esta variable tienen como centroide el número 0, en todos los casos.

Tabla 73. Caso de la variable: tipo de carrera deseada

33-Tipo de carrera que desea estudiar el inscrito				
Clúster	1	2	3	4
Profesional	0.09	0.09	0.09	0.09

De ese 10.34% del global de estudiantes analizados, que respondieron alguna opción en cuanto al interés en una carrera de educación superior al ser indagados al momento de presentar las Pruebas Saber 11, la mayoría de ellos deseaba en ese punto una carrera profesional en Bogotá.

Tabla 74. Centroides k-means. Modelo con todos los datos.

32-Departamento de la institución de educación superior donde le gustaría estudiar				
Clúster	1	2	3	4
Andina	0.01	0.01	0.01	0.01
Antioquia	0.01	0.01	0.01	0.01
Atlántico	0.01	0.01	0.01	0.01
Caribe	0.01	0.01	0.01	0.01
EjeC	0.01	0.01	0.01	0
Santander	0.01	0.01	0.01	0
Valle	0.01	0.01	0.01	0.01
Bogotá	0.03	0.04	0.04	0.04
No_reporta	0.9	0.9	0.9	0.9

Nota: verificar el capítulo 8 de este documento, las agrupaciones de los departamentos para interpretar los atributos de esta variable.

Saber 11: La presentación de la prueba

A continuación, se citan las zonas con resultados superiores a 0.01.

Tabla 75. Centroides k-means. Modelo con todos los datos.

36-Zona de presentación del examen				
Clúster	1	2	3	4
San Andrés	0	0	0	0
Choco	0.01	0	0.01	0.01
Amazónica	0.02	0.01	0.02	0.01
Llanos	0.03	0.05	0.03	0.04
Bolívar	0.04	0.03	0.04	0.04
Santander	0.05	0.07	0.06	0.03
Atlántico	0.05	0.05	0.04	0.06
EjeC	0.06	0.05	0.06	0.05
Valle	0.07	0.05	0.06	0.08
Antioquia	0.11	0.09	0.1	0.11
Caribe	0.12	0.1	0.14	0.11
Bogotá	0.21	0.24	0.2	0.24
Andina	0.23	0.24	0.25	0.22

Nota: verificar el capítulo 8 de este documento, las agrupaciones de los departamentos para interpretar los atributos de esta variable.

Saber 11: El desempeño académico en la prueba

Tabla 76. Centroides k-means. Modelo con todos los datos.

46- Desempeño del inscrito en inglés según las bandas del Marco Común Europeo				
Clúster	1	2	3	4
A-	0.22	0.12	0.19	0.2
A1	0.44	0.4	0.45	0.46
A2	0.17	0.24	0.19	0.18
B+	0.03	0.05	0.04	0.03
B1	0.13	0.19	0.13	0.12

47-Nombre del componente flexible (profundizaciones o interdisciplinar)				
Clúster	1	2	3	4
VIOLENCIA Y SOCIEDAD	0.23	0.16	0.35	0.27
MEDIO AMBIENTE	0.27	0.27	0.19	0.26
PROFUNDIZACION EN LENGUAJE	0.16	0.09	0.15	0.15
PROFUNDIZACION EN MATEMATICA	0.06	0.31	0.04	0.15
PROFUNDIZACION EN BIOLOGIA	0.22	0.13	0.09	0.09
PROFUNDIZACION EN CIENCIAS SOCIALES	0.07	0.05	0.18	0.08

Agrupación de variables				
Clúster	1	2	3	4
38-11_punt_lenguaje	52.85	55.39	54.03	52.2
39-11_punt_matematicas	52.66	59.92	53.12	53.58
40-11_punt_c_sociales	51.71	54.93	53.32	51.24
41-11_punt_filosofia	48.6	50.46	49.59	47.73
42-11_punt_biologia	51.85	55.3	51.94	50.95
43-11_punt_quimica	51.21	55.14	50.32	50.5
44-11_punt_fisica	49.59	54.43	49.3	49.38
45-11_punt_ingles	51.06	54.82	51.8	51.16
48-11_punt_comp_flexible	28.63	26.16	30.81	29.72

	50-11_estu_puesto	312.94	183.51	268.86	297.66	

Saber 11: Variables de tipo socioeconómico

Tabla 77. Centroides k-means. Modelo con todos los datos.

Agrupación de variables				
Clúster	1	2	3	4
51-11_fami_nivel_sisben	3.12	3.6	3.69	3.32
52-11_fami_estrato_vivienda	2.31	2.61	2.72	2.51
53-11_fami_ing_fmiliar_mensual	2.6	3.02	3.14	2.8
58-11_fami_pisos_hogar	3.21	3.4	3.45	3.31
59-11_fami_personas_hogar	4.6	4.43	4.39	4.53
60-11_fami_telefono_fijo	0.63	0.7	0.71	0.66
61-11_fami_celular	0.94	0.95	0.96	0.95
62-11_fami_internet	0.58	0.69	0.71	0.63
63-11_fami_servicio_television	0.7	0.76	0.79	0.74
64-11_fami_computador	0.72	0.82	0.83	0.75
65-11_fami_lavadora	0.73	0.8	0.81	0.76
66-11_fami_nevera	0.94	0.96	0.96	0.95
67-11_fami_horno	0.52	0.6	0.63	0.56
68-11_fami_dvd	0.7	0.73	0.75	0.73
69-11_fami_microondas	0.35	0.42	0.48	0.4
70-11_fami_automovil	0.29	0.39	0.43	0.35

Saber Pro: La elección vocacional

La variable a continuación nos da una clara interpretación de la carrera profesional elegida y cursada por el estudiante en una institución de educación superior del país, de quien estamos analizando los datos desde que se perfiló desde la educación media hacia la educación superior.

Tabla 78. Centroides k-means. Modelo con todos los datos.

111-Grupo de referencia				
Clúster	1	2	3	4
ADMINISTRACION Y AFINES	0	0	0	0.67
BELLAS ARTES Y DISENO	0.01	0	0	0
CIENCIAS AGROPECUARIAS	0.05	0	0	0

CIENCIAS NATURALES Y EXACTAS	0	0.08	0	0
CIENCIAS SOCIALES	0.12	0	0	0
COMUNICACION, PERIODISMO Y PUBLICIDAD	0.01	0	0	0
CONTADURIA Y AFINES	0	0	0	0.25
DERECHO	0	0	0.82	0
ECONOMIA	0	0	0	0.07
EDUCACION	0.39	0	0	0
ENFERMERIA	0.08	0	0	0
HUMANIDADES	0	0	0.02	0
INGENIERIA	0	0.92	0	0
MEDICINA	0.13	0	0	0
NORMALES SUPERIORES	0.01	0	0	0
PSICOLOGIA	0	0	0.16	0
SALUD	0.21	0	0	0

Saber Pro: La presentación del examen y evaluación de competencias genéricas.

Tabla 79. Centroides k-means. Modelo con todos los datos.

Agrupación de variables				
Clúster	1	2	3	4
127-G_razona_cuant_punt	148.45	174.86	149.09	154.33
131-G_lect_critica_punt	154.82	161.85	164.16	150.84
135-G_compet_ciudad_punt	149.18	154.46	164.13	147.44
139-G_ingles_punt	153.41	166.34	155.71	155.59
143-G_com_escrita_punt	150.69	151.14	158.13	149.38
147-G_punt_global	151.31	161.73	158.24	151.52

Saber Pro: Variables de tipo socio económico

Tabla 80. Centroides k-means. Modelo con todos los datos.

Clúster	1	2	3	4
173-Nivel socioeconómico del estudiante				
NSE1	0.25	0.16	0.12	0.21
NSE2	0.31	0.29	0.26	0.31

NSE3	0.23	0.27	0.29	0.26
NSE4	0.18	0.26	0.3	0.2

Saber Pro: La evaluación de competencias específicas

Tabla 81. Centroides k-means. Modelo con todos los datos.

Clúster	1	2	3	4
181-Nombre de la prueba específica (lo que permite identificar la carrera profesional elegida por el estudiante)				
ANALISIS ECONOMICO	0	0	0	0.07
ATENCION EN SALUD	0.42	0	0	0
DISENO DE OBRAS DE INFRAESTRUCTURA	0	0.15	0	0
DISENO DE PROCESOS INDUSTRIALES	0	0.04	0	0
DISENO DE SISTEMAS DE CONTROL	0	0.08	0	0
DISENO DE SISTEMAS MECANICOS	0	0.06	0	0
DISENO DE SISTEMAS PRODUCTIVOS Y LOGISTICOS	0	0.22	0	0
ENSENAR	0.4	0	0	0
FORMULACION DE PROYECTOS DE INGENIERIA	0.01	0.37	0	0.01
GESTION DE ORGANIZACIONES	0	0	0	0.6
GESTION FINANCIERA	0	0	0	0.33
INTERVENCION EN PROCESOS SOCIALES	0.09	0	0.08	0
INVESTIGACION EN CIENCIAS SOCIALES	0.04	0	0.1	0
INVESTIGACION JURIDICA	0	0	0.82	0
PENSAMIENTO CIENTIFICO - CIENCIAS BIOLOGICAS	0	0.04	0	0
PENSAMIENTO CIENTIFICO - CIENCIAS DE LA TIERRA	0	0.01	0	0
PENSAMIENTO CIENTIFICO - CIENCIAS FISICAS	0	0.01	0	0

PENSAMIENTO CIENTIFICO - MATEMATICAS Y ESTADISTICA	0	0.01	0	0
PENSAMIENTO CIENTIFICO - QUIMICA	0	0.01	0	0
PRODUCCION AGRICOLA	0.02	0	0	0
PRODUCCION PECUARIA	0.03	0	0	0
182-Puntaje obtenido en la prueba específica				
Resultado entre 0 y 300	156.09	156.56	157.01	155.31

9.3. Relaciones entre los perfiles de los estudiantes de la educación media y superior.

Se busca en las reglas de asociación denominadas “Apriori”, información que pueda ser útil para descubrir hechos que ocurran de forma conjunta y que a su vez revelen situaciones que se denominan las consecuencias de las apariciones condicionadas de dichos hechos.

Para este propósito se toma en cuenta el aprendizaje sobre el comportamiento de las variables una vez perfiladas en los modelos desarrollados hasta este punto, han mostrado algún poder explicativo.

Las siguientes dos variables también fueron tenidas en cuenta para las reglas de asociación:

- Cluster_S11{0,1,2}: Clúster generado en los datos de las Pruebas Saber 11, el cual arrojó el perfilado para los estudiantes de la educación media, sobre los cuales se trabajaron otros modelos con los datos que registraron su paso por la educación superior.

  Convención:

  Cluster_S11, atributo = 0: perfil de estudiante de la educación media bautizado como “**Intermedios**”.

  Cluster_S11, atributo = 1: perfil de estudiante de la educación media bautizado como “**Rezagados**”.

  Cluster_S11, atributo = 2: perfil de estudiante de la educación media bautizado como “**Pilos**”.

- Clúster_Spro {0,1,2}: Clúster generado en los datos de las Pruebas Saber Pro, el cual arrojó el perfilado para los estudiantes de la educación superior, sobre los cuáles se desarrollaron otros modelos con sus datos cuando estaban en la educación media.

  Convención:

  Clúster_Spro, atributo = 0: perfil de estudiante de la educación superior bautizado como "**Desempeño medio**".

  Clúster_Spro, atributo = 1: perfil de estudiante de la educación superior bautizado como "**Desempeño alto**".

  Clúster_Spro, atributo = 2: perfil de estudiante de la educación superior bautizado como "**Desempeño bajo**".

La tabla 82 presenta los resultados de las reglas de asociación para los perfiles.

Tabla 82. Reglas de asociación para los perfiles

Modelo Apriori Acá se acotaron las variables de entrada a solo los perfiles de los estudiantes en las pruebas Saber 11 y Saber Pro
1. Cluster_Spro=1 37690 ==> Cluster_S11=2 35547 <conf:(0.95)> lift:(1.73) lev:(0.11) [14984] conv:(8.88) 2. Cluster_S11=1 20035 ==> Cluster_Spro=2 16328 <conf:(0.81)> lift:(2.8) lev:(0.08) [10497] conv:(3.83) 3. Cluster_Spro=0 56207 ==> Cluster_S11=2 31972 <conf:(0.57)> lift:(1.04) lev:(0.01) [1119] conv:(1.05) 4. Cluster_S11=0 39512 ==> Cluster_Spro=0 20597 <conf:(0.52)> lift:(1.23) lev:(0.03) [3787] conv:(1.2) 5. Cluster_S11=2 72520 ==> Cluster_Spro=1 35547 <conf:(0.49)> lift:(1.73) lev:(0.11) [14984] conv:(1.41)

6. Cluster_Spro=2 38170 ==> Cluster_S11=0 17105 <conf:(0.45)> lift:(1.49) lev:(0.04) [5606] conv:(1.26)

7. Cluster_S11=2 72520 ==> Cluster_Spro=0 31972 <conf:(0.44)> lift:(1.04) lev:(0.01) [1119] conv:(1.03)

8. Cluster_S11=0 39512 ==> Cluster_Spro=2 17105 <conf:(0.43)> lift:(1.49) lev:(0.04) [5606] conv:(1.25)

9. Cluster_Spro=2 38170 ==> Cluster_S11=1 16328 <conf:(0.42)> lift:(2.8) lev:(0.08) [10497] conv:(1.47)

10. Cluster_Spro=0 56207 ==> Cluster_S11=0 20597 <conf:(0.37)> lift:(1.23) lev:(0.03) [3787] conv:(1.11)

10. DESPLIEGUE

A continuación, se entregan los análisis de resultados de los modelos aplicados en relación al estudio de:

- El rendimiento académico
- La permanencia en el sistema de educación superior.
- La elección vocacional.

10.1 Rendimiento académico en las Pruebas Saber 11 y Saber Pro

Para analizar el rendimiento académico en las pruebas, se aplicó un análisis de correlaciones para identificar aquellas variables que más están relacionadas con el desempeño en las pruebas, resaltando las variables con una correlación medida en valor absoluto superior a 0.3, se encontró lo siguiente.

10.1.1 Rendimiento en las pruebas Saber 11

Tomando en cuenta las correlaciones generadas para evaluar la relación entre los datos de las Pruebas Saber 11 con el **puntaje** obtenido por el estudiante en la opción que seleccionó entre los 6 componentes flexibles para presentar la prueba, se encontró que ninguna variable fuera de las que tienen que ver directamente con la escogencia del componente flexible, reveló una relación que amerite algún tipo de atención, dado que tenían valores muy bajos

Por otra parte, el grupo de las variables que presentan una correlación en valor absoluto mayor a (0.3) con el **puesto global** obtenido por el estudiante en las pruebas Saber 11, son las que se detallan a continuación, en donde cada uno de sus registros se acompaña del valor de la correlación con la variable puesto global.

Se destaca que las variables relacionadas con el rendimiento académico tienen una correlación negativa con el puesto global obtenido en la prueba, pues a mejor resultado en cada uno de los módulos menor es el puesto en la clasificación general de la prueba, lo que quiere decir que en esta variable definida como objetiva, los estudiantes con mejor desempeño en la prueba son aquellos cuyos puestos aparecen en los primeros resultados iniciando desde cero en un rango que va hasta 1000).

Datos de las Pruebas Saber 11:

- Fami_nivel_sisben, -0.29
- Fami_ing_fmiliar_mensual, -0.31
- Fami_estrato_vivienda, -0.31
- Clasificación del estudiante en nivel de inglés B1. Saber 11, -0.31

10.1.2 Rendimiento en las pruebas Saber Pro

A continuación, se detallan las variables de las Pruebas Saber 11 y Saber Pro, que mostraron una asociación mayor a 0.3 (medida por correlación en valor absoluto) con el rendimiento en las pruebas Saber Pro, seguidas del valor de la correlación cuando la variable objetivo es el **puntaje global** obtenido por el estudiante en las pruebas Saber Pro.

Datos de las pruebas Saber 11:

- Cole_valor_pension, 0.32
- Clasificación del estudiante en nivel de inglés B1. Saber 11, 0.37
- Clasificación del estudiante en nivel de inglés A-. Saber 11), -0.35
- Fami_ing_fmiliar_mensual, 0.38
- Fami_estrato_vivienda, 0.37
- Fami_nivel_sisben, 0.33

Datos de las pruebas Saber Pro:

- Clasificación del estudiante en nivel de inglés B2. Saber Pro, 0.47
- Clasificación del estudiante en nivel de inglés A-. Saber Pro, 0.36
- Fami_estratovivienda, 0.35
- Clasificación del estudiante en nivel de inglés A1. Saber Pro, -0.37
- Clasificación del estudiante en nivel de inglés A-. Saber Pro, -0.41
- Estu_puesto, -0.72

Para el caso de las variables que mostraron una asociación mayor a 0.3 (medida por correlación en valor absoluto) con la siguiente variable definida como objetivo: el **rendimiento de la prueba específica** relacionada con la carrera profesional en curso (pruebas Saber Pro).

Descartando todas las relacionadas con el desempeño académico, ninguna de las variables diferentes a estás mostró una asociación que amerite ser resaltada.

10.2 Permanencia universitaria

Para estudiar la permanencia universitaria, se presentan a continuación los resultados del clustering con todos los datos, es decir los resultados de Saber 11 y Saber Pro, con esto identificamos los perfiles de aquellos estudiantes que al finalizar el bachillerato, inician una carrera profesional y completan al menos el 75% de esta, ya que llegan hasta presentar las pruebas Saber Pro cuando se encuentran en promedio en el semestre No. 9 de sus programas. Por lo anterior, no se catalogan ni como desertores de la educación media ni desertores tempranos de la educación superior.

Todas las interpretaciones se basan en los centroides o valores promedio para cada variable en los resultados del clustering. En la tabla del perfilado de los estudiantes en 4 agrupaciones, se decidió asignar un campo con el rótulo "Nombre del clúster", para facilitar la interpretación de los resultados.

Tabla 83. Rendimiento académico, elección vocacional y permanencia - Interpretación del modelo de K-means con todos los datos.

Clúster	**1**	**2**	**3**	**4**
Nombre del clúster	**Diversos**	**Realísticos**	**Políticos y Sociales**	**Administrativos**
Cantidad de estudiantes agrupados en cada perfil	**36424**	**41045**	**17777**	**36821**
Elección vocacional y permanencia **La carrera profesional elegida por el estudiante, sobre el cual se analiza su tránsito desde la educación media hasta la superior.**				
111-Grupo de referencia				
17 posibilidades que agrupan al 100% de las 556 opciones diferentes de carreras profesionales escogidas por los estudiantes analizados.	_Educación _Salud _Medicina _Ciencias Sociales _Enfermería _Ciencias Agropecuarias _Bellas Artes y Diseño _Comunicación, Periodismo y Publicidad	_Ingeniería _Ciencias Naturales y Exactas	_Derecho _Psicología _Humanidades	_Administración y afines _Contaduría y afines _Economía

	_Normales Superiores			
Estudiantes al momento de presentar las pruebas Saber 11				
La identificación de la prueba y del estudiante				
Edad promedio	16.38	16.08	16.36	16.37
Género 1=masculino 0=femenino	0.27	0.59	0.34	0.35
	Mujer	Hombre	Mujer	Mujer
El desempeño académico en la prueba				
Resultados altos en:	Lenguaje, matemáticas y Biología	Matemáticas, Lenguaje y Química	Lenguaje, Ciencias Sociales y Matemáticas	Matemáticas, Lenguaje y Ciencias Sociales
Resultados bajos en:	Filosofía, Física e inglés	Filosofía, Física e inglés	Física, Filosofía y Química	Filosofía, Física y Química
46- Desempeño del inscrito en inglés (según las bandas del Marco Común Europeo)				
Se identifican más con:	A1, A-	A1, A2	A1, A-, A2	A1, A2
Se identifican menos con:	B1, B+	A-, B+	B1, B+	B+, A-
Nivel de inglés:	Bajo	Alto	Intermedio	Intermedio
47-Nombre del componente flexible (profundizaciones o interdisciplinar)				
Interés en:	Medio Ambiente	Matemática	Violencia y Sociedad	Violencia y Sociedad
	Violencia y Sociedad	Medio Ambiente	Medio Ambiente	Medio Ambiente
Desinterés en:	Ciencias Sociales	Lenguaje	Biología	Biología
	Matemática	Ciencias Sociales	Matemática	Ciencias Sociales
Variables de tipo socioeconómico (Interpretación conjunta de todas las variables de este tipo)				
Clasificación socioeconómica:	Baja	Intermedia-Alta	Alta	Intermedia-Baja
Estudiantes al momento de presentar las pruebas Saber Pro				
identificación del estudiante				
Edad promedio	22.3	21.96	22.14	22.16
La evaluación de competencias específicas				
Se les facilita:	Lectura crítica e inglés	Razonamiento cuantitativo e inglés	Lectura crítica y competencias ciudadanas	Inglés y razonamiento cuantitativo
Se les dificulta:	Competencias ciudadanas y razonamiento cuantitativo	Comunicación escrita y competencias ciudadanas	Inglés y razonamiento cuantitativo	Competencias ciudadanas y comunicación escrita

Puntaje global	Bajo	Alto	Intermedio	Intermedio-bajo
Variables de tipo socioeconómico				
173-Nivel socioeconómico del estudiante				
Agrupación de varias variables	Bajo	Intermedio-alto	Alto	Intermedio-bajo
La elección vocacional y la evaluación de la competencia específica				
181-Nombre de la prueba específica (Esta variable refuerza también la interpretación de la elección y el perfil vocacional del estudiante, pues a través de las Pruebas Saber Pro se evalúa 1 de las 21 competencias específicas que determina el ICFES, según la carrera profesional escogida).				
Resultados más frecuentes:	_Atención en salud _Enseñar	_Formación en proyectos de ingeniería _Diseño de sistemas productivos y logísticos	_Investigación jurídica	_Gestión de organizaciones
182-Puntaje obtenido en la prueba específica				
Puntaje obtenido promedio	156.09	156.56	157.01	155.31

10.3 Elección vocacional

Para estudiar la elección vocacional se realizaron 3 modelos analíticos:

- Perfiles de estudiantes que presentaron las pruebas Saber 11
- Perfiles de estudiantes que presentaron las pruebas Saber Pro
- Asociación entre los perfiles.

10.3.1 Perfiles de estudiantes que presentaron las pruebas Saber 11

La siguiente tabla, muestra una interpretación de los 3 perfiles diferentes encontrados en los estudiantes de bachillerato, basada en el modelo de clustering con los datos de las pruebas Saber 11. En la tabla que sigue a continuación, se decidió asignar un campo con el rótulo “Nombre del clúster”, para facilitar la interpretación de los resultados.

Tabla 84. Interpretación de k-means. Modelo Saber 11.

Clúster			**1**	**2**	**3**
Nombre del clúster			**Intermedios**	**Rezagados**	**Pilos**
Cantidad de estudiantes perfilados en cada uno de los grupos			**39514**	**20035**	**72520**
Identificación del estudiante		Edad	Media	Alta	Baja
		Genero	Femenino	Femenino	Masculino
Identificación del colegio		Naturaleza del colegio	Oficial	Oficial	Privada
		Valor de la pensión del colegio	Media	Baja	Alta
Desempeño en la prueba Saber 11					
Puntaje		Lenguaje	Medio	Bajo	Alto
		Matemáticas	Medio	Bajo	Alto
		Ciencias Sociales	Medio	Bajo	Alto
		Filosofía	Medio	Bajo	Alto
		Biología	Medio	Bajo	Alto
		Química	Medio	Bajo	Alto
		Física	Medio	Bajo	Alto
		Inglés	Medio	Bajo	Alto
interés del estudiante, medido con la elección de 1 componente flexible	Profundización en:	matemática	Medio	bajo	alto
	Otras opciones para elegir el componente flexible:	Medio ambiente	Medio	alto	Bajo
Puntaje		Componente flexible	Alto	Medio	Bajo
Puesto del estudiante en las pruebas Saber 11		puesto promedio (el promedio general del dato es 263)	352.00 Rango II {250-499}	705.79 Rango III {500-749}	91.23 Rango I {001-249}

Variables de tipo socioeconómico	Nivel de SISBEN de la familia	Medio	Bajo	Alto
	Estrato de la vivienda	Medio	Bajo	Alto
	Ingreso familiar mensual	Medio	Bajo	Alto

10.3.2 Perfiles de estudiantes que presentaron las pruebas Saber Pro

La siguiente tabla, muestra una interpretación de los 3 perfiles diferentes encontrados en los estudiantes de la educación superior, basada en el modelo de clustering con los datos de las pruebas Saber Pro. En la tabla que sigue a continuación, se decidió asignar un campo con el rótulo “Nombre del clúster”, para facilitar la interpretación de los resultados.

Tabla 85. Interpretación de k-means. Modelo Saber Pro.

Clúster		**1**	**2**	**3**
Nombre del clúster		**Desempeño medio**	**Desempeño alto**	**Desempeño bajo**
Cantidad de estudiantes perfilados en cada uno de los grupos		**56207**	**37690**	**38170**
Identificación del estudiante	Edad	Media	Baja	Alta
	Genero	femenino	masculino	Femenino
Grupo de referencia de la carrera profesional escogida (elección vocacional)		Ingenierías, Administración, Derecho	Ciencias exctas, Ingenierías, Administración, Derecho y Medicina	Administración, Ingeniería y Educación
Desempeño académico en las competencias genéricas Saber Pro	Razonamiento cuantitativo	Medio	Alto	Bajo
	Lectura crítica	Medio	Alto	Bajo
	Competencias ciudadanas	Medio	Alto	Bajo
	Inglés	Medio	Alto	Bajo

	Comunicación escrita	Medio	Alto	Bajo
Nivel socioeconómico	Estrato de la vivienda	Medio	Alto	Bajo
	Nivel Socio Económico	Medio	Alto	Bajo
	Valor de la matrícula universitaria	Medio	Alto	Bajo
Elección vocacional asociada la competencia específica que se evalúa de acuerdo con la carrera profesional elegida y desempeño académico en dicha competencia	Elección vocacional	Gestión de las organizaciones	Gestión de las organizaciones	Gestión de las organizaciones
		Investigación jurídica	Investigación jurídica	Enseñar
		Formulación de proyectos de ingeniería	Formulación de proyectos de ingeniería	Atención en Salud
		Enseñar	Atención en Salud	Formulación de proyectos de ingeniería
		Atención en Salud		
	Desempeño académico	Medio	Alto	Bajo

10.3.3 Asociaciones entre los perfiles

Finalmente, se presentan las siguientes **asociaciones entre los perfiles** de estudiantes que presentan las Pruebas Saber 11 y Saber Pro con el modelo Apriori, resaltando todas aquellas que tienen como consecuente al perfil encontrado en las pruebas Saber Pro.

Tabla 86. Asociaciones entre los perfiles.

Regla	Interpretación
Cluster_S11=1 ==> Cluster_Spro=2 <conf:(0.81)>	El 81% de estudiantes bautizados como “Rezagados” en los perfiles que se realizaron de la educación media, están asociados al perfil de “Desempeño bajo” en la educación superior.

Cluster_S11=0 ==> Cluster_Spro=0 <conf:(0.52)>	El 52% de estudiantes bautizados como “Intermedios” en los perfiles que se realizaron de la educación media, están asociados al perfil de “Desempeño medio” en la educación superior.
Cluster_S11=2 ==> Cluster_Spro=1 <conf:(0.49)>	El 49% de estudiantes bautizados como “Pilos” en los perfiles que se realizaron de la educación media, están asociados al perfil de “Desempeño alto” en la educación superior.
Cluster_S11=2 ==> Cluster_Spro=0 <conf:(0.44)>	El 44% de estudiantes bautizados como “Pilos” en los perfiles que se realizaron de la educación media, están asociados al perfil de “Desempeño medio” en la educación superior.
Cluster_S11=0 ==> Cluster_Spro=2 <conf:(0.43)>	El 43% de estudiantes bautizados como “Intermedios” en los perfiles que se realizaron de la educación media, están asociados al perfil de “Desempeño bajo” en la educación superior.

Estas relaciones se resumen en la figura 28, la cual revela las asociaciones entre los perfiles de estudiantes de la educación media asociados a los perfiles de la educación superior.

Los perfiles relacionados anteriormente, contienen información sobre el paso de los estudiantes por la educación media (datos de las pruebas Saber 11) y registran también el tránsito de esos estudiantes, por la educación superior (datos de las pruebas Saber Pro).

Por lo tanto, este trabajo contiene el perfilado de los estudiantes de la educación media y superior relacionado con:

El rendimiento académico
La elección vocacional
La permanencia en la educación superior

Entre muchas otras variables de tipo socioeconómico y demás contenidas en los modelos previamente desarrollados.

Se justifica mencionar que esta investigación se basa en las observaciones de los estudiantes que una vez eligieron su carrera profesional, lograron avanzar mínimamente en el 75% de los créditos académicos de su programa (requisito para presentar las pruebas Saber Pro) y además, los perfiles de los estudiantes de la educación superior analizados, cursaban en promedio el semestre 9 de sus carreras profesionales.

Lo anterior significa que esta investigación contiene aportes rescatables para estudiar los fenómenos asociados a la elección y permanencia en la educación superior, como modelo inverso para comprender la deserción.

A continuación, se presentan las asociaciones entre los perfiles de los estudiantes de la educación media y superior, para las observaciones de estudiantes colombianos tenidas en cuenta.

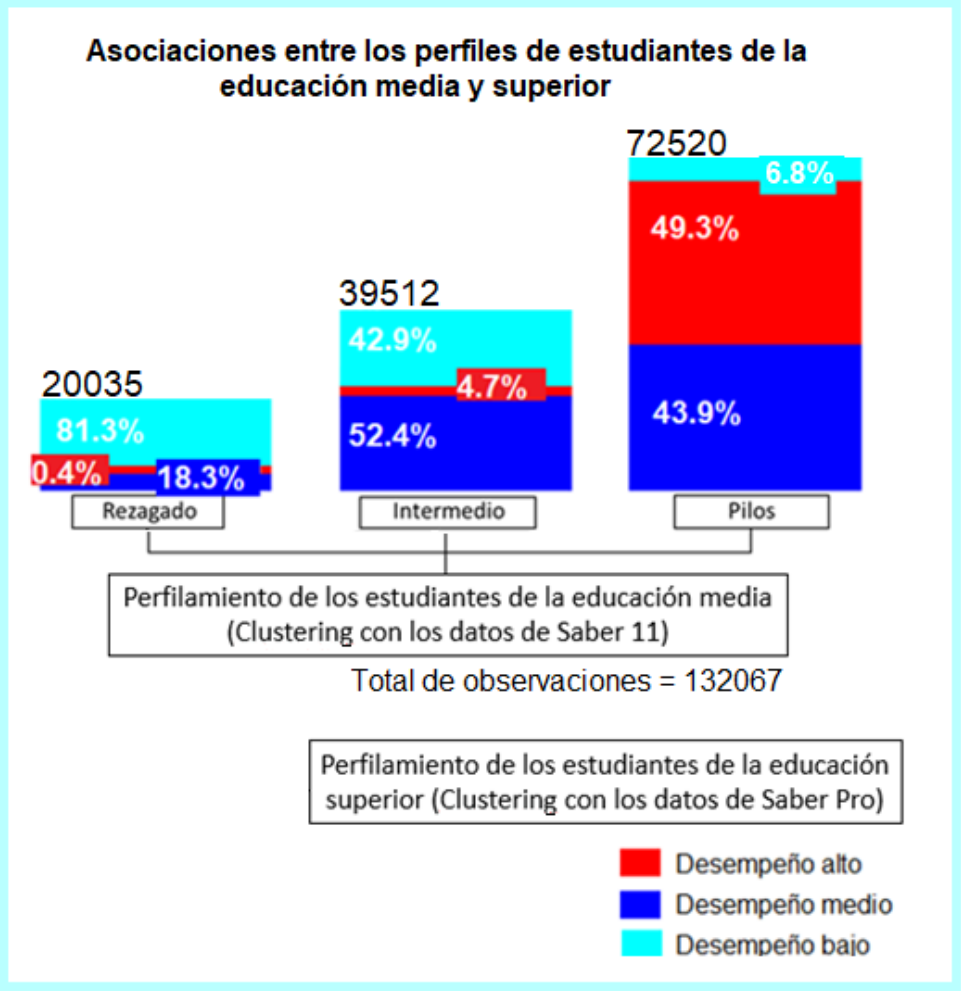


Figura 28. Asociaciones entre los perfiles de los estudiantes de la educación media y superior.

11. CONCLUSIONES

El objetivo del presente estudio fue analizar la relación entre el desempeño académico de estudiantes colombianos en las Pruebas Saber 11 y Saber Pro como aporte para perfilar la elección vocacional y permanencia universitaria, mediante técnicas de Ciencia de Datos.

En cumplimiento con lo anterior, se tomaron en cuenta múltiples datos que luego de ser depurados, permitieron extraer información a partir de las observaciones de 132,067 estudiantes sobre los cuales se exploraron distintas características de tipo socioeconómico, académicas, vocacionales y hasta de condiciones especiales como limitaciones físicas o emocionales o pertenencia a grupos étnicos o minorías.

En relación a algunas **estadísticas de los datos**, se resalta que solo un 1.9%, del total de los estudiantes sobre los cuales se hizo la extracción de la información mediante técnicas de analítica y ciencia de los datos y sobre quienes se tiene la garantía que pasaron por la educación media llegando hasta la superior, estaba catalogado dentro de algún grupo étnico como comunidad afrodescendiente, Emberá, Wayú, Raizal (perteneciente al pueblo indígena del Archipiélago de San Andrés, Providencia y Santa Catalina en Colombia).

Complementando lo expresado anteriormente, solo un 0.03% del total de los estudiantes, manifestó tener alguna limitación motriz y ninguno reportó una condición como el autismo. Condiciones como limitaciones de tipo visual o auditiva fueron de ocurrencia menos frecuente que la primera citada en este párrafo, lo cual revela una incipiente inclusión de las minorías en los registros de los datos abiertos analizados.

En los siguientes departamentos se concentró el 70.4% del global de las instituciones de educación superior en donde los estudiantes observados eligieron y al menos se sabe que completaron la carga académica en mínimamente un 75% en relación a una carrera profesional: Bogotá (32%), Antioquia (11.6%), Atlántico (6.9%), Santander (6%), Valle (5.8%), Boyacá (2.9%), Bolívar (2.7%) y Norte de Santander (2.5%).

El país cuenta con 32 departamentos y Bogotá como Distrito Capital aparte, lo cual significa que en los 25 departamentos no mencionados en el párrafo anterior, hacen presencia solo un 29.6% del total de las sedes de las instituciones de educación superior donde los estudiantes observados cursaron sus carreras profesionales, denotando un condicionamiento en el acceso a la educación superior para los jóvenes de las áreas rurales del país.

En relación con el **rendimiento académico** de los estudiantes, la **elección vocacional** y la **permanencia** en la educación superior, se destaca que el 90% de todos los estudiantes de bachillerato analizados, quienes cursaban el último año de su formación no tuvieron una respuesta para las siguientes inquietudes:

- Municipio de la institución de educación superior donde le gustaría estudiar.

- Departamento de la institución de educación superior donde le gustaría estudiar.

- Tipo de carrera que desea estudiar el inscrito (ninguna, técnica, tecnológica, profesional).

Lo cual revela una falta de determinación en los jóvenes que se perfilan desde la educación media hacia la superior.

Por otra parte, los alumnos que se gradúan del bachillerato en Colombia se pueden clasificar dentro de los siguientes cuatro escenarios, en los que se analiza el déficit en las competencias que presentan en la educación media versus las carreras universitarias que eligen y en las cuales suelen permanecer:

- Escenario 1. Poca competencia para las ciencias sociales y las matemáticas. Estos estudiantes suelen elegir y permanecer en carreras profesionales relacionadas con: educación, salud, medicina, ciencias sociales, enfermería, ciencias agropecuarias, bellas artes y diseño, normales superiores y comunicación, periodismo y publicidad.

- Escenario 2. Poca competencia para el lenguaje y las ciencias sociales. Estos estudiantes suelen elegir y permanecer en carreras profesionales relacionadas con: ingeniería y ciencias naturales y exactas.

- Escenario 3. Poca competencia para la biología y las matemáticas. Estos estudiantes suelen elegir y permanecer en carreras profesionales relacionadas con: derecho, psicología y humanidades.

- Escenario 4. Poca competencia para la biología y las ciencias sociales. Estos estudiantes suelen elegir y permanecer en carreras profesionales relacionadas con: administración, contaduría, economía y afines.

Para finalizar, se presentan las asociaciones encontradas entre los perfiles de los estudiantes de la educación media y superior, basadas en el rendimiento académico, la elección vocacional y la permanencia en las carreras profesionales en las que se matricularon (en las cuales permanecieron mínimamente hasta completar en un 75% de la carga académica).

Para los estudiantes de **educación media** se encontraron 3 perfiles, cuando aún no se han graduado del bachillerato, haciendo referencia al desempeño del alumno en los módulos evaluados en las pruebas Saber 11.

- Los "**Rezagados**", tienen la edad más avanzada, pertenecen principalmente a los estratos más bajos (nivel bajo de Sisbén) y por lo general obtienen puntajes bajos en todos los componentes evaluados.

- Los "**Intermedios**", por lo general pertenecen a los estratos medios y obtienen puntajes de tipo medio en todos los componentes evaluados.

- Los "**Pilos**", por lo general provienen de colegios privados, suelen pagar pensiones en el rango superior en los colegios en donde estudian en bachillerato, son de edades bajas, pertenecen a estratos altos y obtienen puntajes altos en todos los componentes evaluados.

  Presentan un interés en las matemáticas, medido a través del componente elegido para la profundización en las pruebas Saber 11.

Los anteriores 3 perfiles, se asociaron con los siguientes tipos de estudiantes de la **educación superior**:

- Los "**Desempeño bajo**", estudiantes cuya elección vocacional se relaciona principalmente con programas de Administración o Educación.

  Obtienen en promedio un desempeño bajo en el componente específico de las pruebas Saber Pro, el cual mide la competencia en el área afín a

la carrera elegida, la cual a su vez guarda afinidad con los siguientes temas: Gestión de las organizaciones, Enseñanza, Atención en Salud, Formulación de proyectos de ingeniería.

Presentan en promedio un resultado bajo en todos los módulos evaluados en las Pruebas Saber Pro.

Son los estudiantes que pagan en promedio las matrículas más bajas en las instituciones de educación superior donde adelantan sus estudios.

- Los "**Desempeño medio**", estudiantes cuya elección vocacional se relaciona principalmente con programas de Ingenierías, Administración o Derecho.

  Obtienen en promedio un desempeño medio en el componente específico de las pruebas Saber Pro, el cual mide la competencia en el área afín a la carrera elegida, la cual a su vez guarda afinidad con los siguientes temas: Gestión de las organizaciones, Enseñanza, Atención en Salud, Formulación de proyectos de ingeniería e Investigación jurídica.

  Presentan en promedio un resultado medio en todos los módulos evaluados en las Pruebas Saber Pro.

- Los "**Desempeño alto**", estudiantes cuya elección vocacional se relaciona principalmente con Ciencias exactas, Ingeniería, Administración, Derecho y Medicina.

  Obtienen en promedio un desempeño excelente en el componente específico de las pruebas Saber Pro, el cual mide la competencia en el área afín a la carrera elegida, además de un alto desempeño en racionamiento cuantitativo, acompañado de un alto nivel de inglés.

  Presentan en promedio un desempeño alto en todos los módulos evaluados en las Pruebas Saber Pro.

Las asociaciones entre los perfiles de los estudiantes de la educación media y superior indican que:

- Los estudiantes de la educación media "**Rezagados**", cuando hacen el tránsito hacia la educación superior, una vez que están cursando en promedio el semestre 9 de sus carreras profesionales, se perfilan consistentemente en el grupo "**Desempeño bajo**", ligeramente se

asocian con el perfil definido como “Desempeño medio” y están muy lejos del grupo de “Desempeño alto”.

A pesar que todos los estudiantes analizados en este proyecto se catalogan como estables dentro del sistema de educación superior, en cuanto a la capacidad de permanecer en las carreras profesionales que cursan, este grupo sí bien el menor dentro de los 3 perfiles de los estudiantes de la educación media revela consistentemente que las bases de la educación superior se desarrollan en la educación media y un estudiante con un desempeño bajo en el bachillerato, difícilmente va a dar sorpresas positivas en cuanto a su desempeño académico en la carrera profesional elegida.

- Los estudiantes de educación media “**Intermedios**”, cuando hacen el tránsito hacia la educación superior, una vez que están cursando en promedio el semestre 9 de sus carreras profesionales, se perfilan mayoritariamente en los grupos de “**Desempeño bajo**” y “**Desempeño medio**”, siendo muy ligera la relevancia de estos dentro del grupo de “Desempeño alto” en la educación superior.

  Lo anterior denota una diferenciación entre los niveles medios de los perfiles de la educación media y superior, en la que se entiende que no basta con un rango “intermedio” en la educación media, pues estos estudiantes no logran un rango similar en la educación superior, sino que mayoritariamente descienden hacia el nivel más bajo de desempeño medido a través de los componentes flexibles y el específico, evaluados en las pruebas Saber Pro.

  Estos estudiantes tampoco logran reconocerse dentro del grupo de “Desempeño alto” en la educación superior, su participación ahí es incipiente.

- Los estudiantes de educación media “**Pilos**”, cuando hacen el tránsito hacia la educación superior, una vez que están cursando en promedio el semestre 9 de sus carreras profesionales, se perfilan mayoritariamente en los grupos de “**Desempeño alto**” y “**Desempeño medio**”, siendo muy ligera la relevancia de estos dentro del grupo de “Desempeño bajo” en la educación superior.

  Lo anterior indica que unas condiciones socioeconómicas estables (alto poder adquisitivo) contribuyen a un desempeño alto desde la educación media hasta la superior, pues estas categorías están mayoritariamente asociadas a los jóvenes de estratos más altos.

12. RECOMENDACIONES Y/O PROPUESTAS PARA INVESTIGACIONES FUTURAS

- Se inició este proyecto con un total de 183 variables de las pruebas Saber 11 y Saber Pro, concluyendo que existe mucho ruido y poco poder explicativo de la mayoría de estas, dado que muchas muestran colinealidad entre sí y excesiva redundancia.

  Las variables que resultaron más redundantes en este proyecto fueron las que se reconocieron en la descripción de los datos como las relacionadas con aspectos socioeconómicos del estudiante y su familia, tanto para las Pruebas Saber 11 como para las Pro.

- Para perfilar las condiciones socioeconómicas de los estudiantes cuando están en la educación media y superior, se pueden conservar variables como el estrato de la vivienda familiar y el SISBEN (datos de Saber 11) y nuevamente el estrato de la vivienda y el nivel socioeconómico del estudiante (datos de Saber Pro), pues en ambos casos resultó que estos datos tenían más carácter para reconocer a los estudiantes que un sinnúmero de variables que se comportaban de una forma similar a las destacadas en este apartado.

- Las variables catalogadas en las Pruebas Saber Pro como “La identificación de minorías o condiciones especiales en el inscrito”, denotan una inclusión muy incipiente de estas poblaciones en el Sistema de Educación Superior, la cual no pudo ser interpretada por ninguno de los modelos desarrollados.

- La variable puntaje en el componente flexible de las Pruebas Saber 11, es presentada por el ICFES como un solo dato que recoge información de los resultados en la categoría sea de tipo interdisciplinar o de profundización, escogida por el estudiante.

  Dichas categorías tienen distintos rangos posibles para su resultado. Se recomienda separar en dos variables el puntaje en el componente flexible dependiendo de la elección del estudiante (componente flexible o interdisciplinar), dado que los rangos que manejan son muy distintos entre sí.

- Teniendo en cuenta las descripciones de los perfiles de los estudiantes de la educación media, se recomienda crear políticas de apoyo para los

perfiles que no logran obtener desempeños altos en las pruebas Saber Pro, enfocados a políticas de inclusión de estudiantes de estratos bajos y apoyo académico a largo de su carrera en educación superior.

Se espera que este trabajo permita que programas como “Ser Pilo Paga” y “Generación E” reconozcan acciones orientadas a mejorar el déficit en el desempeño académico asociado a los estudiantes de estratos bajos, los cuales transitan desde la educación media hacia la educación superior, identificándose mayoritariamente en el bachillerato y luego ad-portas de culminar una carrera de educación superior, con los resultados académicos más bajos entre todos los estudiantes observados.

- Los distintos actores de la educación en Colombia (Secretarías de Educación, Colegios, Instituciones de Educación Superior, padres de familia, entre otros) deben replantearse la ligereza como los jóvenes escogen una carrera profesional en el país, sobre quienes se espera que a tan temprana edad tomen decisiones asociadas a la elección consciente y permanencia en el Sistema de Educación Superior del país, sin que sea un requisito el haber sido parte de algún programa de orientación vocacional serio.

El conocimiento construido aquí a través de la Ciencia de los Datos para comprender las asociaciones entre los perfiles de los estudiantes de la educación media y los de la educación superior, se configura como la base para la creación de una prueba de orientación vocacional, adaptada a Colombia y su sistema de educación media y superior. Dicha prueba será puesta a disposición del país a través del proyecto conocido como www.TuCarrera.co y busca facilitar la elección consciente de una carrera profesional en Colombia, mitigando el riesgo de deserción.

Bibliografía

[1] Ferreyra, M., Avitabile, C., Álvarez, J., Haimovich, F., Urzúa, S. (2017) Momento decisivo: La educación superior en América Latina y el Caribe. Washington, DC: Banco Mundial. Disponible en /openknowledge.worldbank.org/bitstream/handle/10986/26489/9781464810145.pdf?sequence=2&isAllowed=y

[2] Alcaldía de Medellín (2017). Observatorio de la Deserción en la Educación Superior-ODES. Boletín 5 julio 2017. Disponible en http://www.sapiencia.gov.co/wp-content/uploads/2017/11/5_JULIO_BOLETIN_ODES_DESERCION_EN_LA_EDUCACION_SUPERIOR.pdf

[3] Gobierno de Colombia, Ministerio de Educación Nacional. Estadísticas Deserción y Graduación 2015. Disponible en https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-357549_recurso_3.pdf

[4] Gobierno de Colombia, Ministerio de Educación Nacional. Estadísticas e Indicadores
de Deserción Estudiantil, cifras 1998-2013. Disponible en https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_archivo_pdf_indicadores_permanencia.pdf

[5] Gobierno de Colombia, Ministerio de Educación Nacional (2018). Sistema Nacional de Educación Superior – SNIES. Estadísticas de programas. Cohorte al 17.11.18.

[6] Alcaldía de Medellín (2018). Secretaria de Gestión Humana y Servicio a la Ciudadanía. Coordinación Archivo Central. Plan de Desarrollo 2016-2019. Gaceta oficial. 29 de junio de 2016. Disponible en https://www.medellin.gov.co/irj/go/km/docs/pccdesign/SubportaldelCiudadano_2/PlandeDesarrollo_0_15/Publicaciones/Shared%20Content/GACETA%20OFICIAL/2016/Gaceta%204383/GACETA%204383.pdf

[7] Instituto Colombiano para la Evaluación de la Educación (ICFES), dependiente del Ministerio de Educación Nacional. 14 de mayo de 2019. Disponible en http://www.icfes.gov.co/web/guest/bases-de-datos

[8] The CRISP-DM process model (1999), http://www.crisp-dm.org/

[9] Santana Vega, L. E. (2009). Orientación educativa e intervención psicopedagógica. Cambian los tiempos, cambian las responsabilidades profesionales. Madrid. Pirámide.

[10] Bisquerra. R. (1990). Métodos de Investigación Educativa. España, Edit. La Editorial.

[11] Anthony & Cols. (1984), citado por Galilea. V. (2000) Orientación vocacional. Disponible en http://www.sie.es/crl/archivo_pdf/ORIENTACION%20VOCACIONAL.pdf.

[12] Castillo. G. (1983). Los Padres y la Orientación Profesional de sus Hijos. España, Edit. Educación NT.

[13] Cortada de Kohan. N. (1977). El Profesor y la Orientación Vocacional. México, Edit. Trillas.

[14] Gosabez. A. (1977). Orientación y Tratamiento Pedagógico. España, Edit.Cincel.

[15] Cols. S. (1975). La Tarea Docente. Argentina, Edit. Marymar.

[16] Fritz. K y Gardner. R. (1971). El Consejo Psicológico En los Momentos Cruciales de la Vida. España, Edit. Luis Moracle S.A.

[17] Guzmán, C., Durán, D., Franco, J., Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2009). Deserción estudiantil en la educación superior colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención.

[18] Sánchez, C. (2009). Máster en Técnicas Estadísticas. Análisis Multivariante.

[19] Bisquerra, R., Sarriera, J. C., & Matínez, F. (2009). Introdução à estatística: enfoque informático com o pacote estatístico SPSS. Bookman Editora.

[20] Pérez, C. (2004). Técnicas de análisis multivariante de datos. Pearson Educación, S.A.

[21] Anderson, T.W. (2003). An introduction to multivariate statistical analysis. Wiley.

[22] Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill.

[23] Seber, G.A.F. (1984). Multivariate observations. Wiley.

[24] Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. Biometrika, pág. 58, 453-467.

[25] Johnson, R.A. y Wichern, D.W. (1982). Applied multivariate statistical analysis. Prentice-Hall. Mardia,

[26] K.V., Kent, J.T. y Bibby, J.M. (1979). Multivariate analysis. Academic Press.

[27] Oded Maimon and Lior Rokach (2010). Data Mining and Knowledge Discovery Handbook. Springer, New York. ISBN 978-0-387-09823-4.

[28] Dawson, S., Gašević, D., Siemens, G., & Joksimović, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge (pp. 231–240). New York, NY: ACM. https://doi.org/10.1145/ 2567574.2567585.

[29] Long, P. D., Siemens, G., Conole, G., & Gašević, D. (Eds.). (2011). Proceedings of the 1st International Conference on Learning Analytics and Knowledge (LAK ’11). New York, NY: ACM.

[30] Lim, C. P., y Tinio, V. L. (Eds.). (2018). Analíticas de aprendizaje para el Sur Global. Quezon City, Filipinas: Fundación para la Formación en Tecnologías de la Información y el Desarrollo.

[31] Gašević, D. (2018). Include us all! Directions for adoption of learning analytics in the global south. In C. P.

[32] Pea, R. (2014). The Learning Analytics Workgroup: A report on building the field of learning analytics for personalized learning at scale. Stanford, CA: Stanford University. Disponible en https://ed.stanford.edu/sites/default/files/law_report_complete_ 09-02-2014.pdf

[33] Corrin, L., & de Barba, P. (2014). Exploring students' interpretation of feedback delivered through learning analytics dashboards. In Proceedings of the 31st Annual ASCILITE Conference (2014) (pp. 629–633).

[34] Buckingham Shum, S., & Deakin Crick, R. (2016). Learning analytics for 21st century competencies. Journal of Learning Analytics, 3(2), 6–21. https://doi.org/10.18608/jla.2016.32.2

[35] Steiner, C. M., Kickmeier-Rust, M. D., & Albert, D. (2015). Let's talk ethics: Privacy and data protection framework for a learning analytics toolbox. Disponible en http://css-kmi.tugraz.at/mkrwww/leas-box/ downloads/LAKEthics15.pdf

[36] Superintendencia de Industria y Comercio. Manejo de información personal, 'Habeas data'. 21 de septiembre de 2019. Disponible en http://www.sic.gov.co/mision-y-vision.

[37] Gobierno de Colombia. Función Pública, Ley 1266 de 2008. Disponible en https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=34488,

[38] República de Colombia. Corte Constitucional, Proyecto de ley estatutaria de habeas data y manejo de información contenida en bases de datos personales, Sentencia C-1011/08. Disponible en http://www.corteconstitucional.gov.co/relatoria/2008/C-1011-08.htm,

[39] República de Colombia. Corte Constitucional, Proyecto de ley estatutaria de habeas data y protección de datos personales, Sentencia C-748/11. Disponible en http://www.corteconstitucional.gov.co/relatoria/2011/c-748-11.htm.

[40] Gobierno de Colombia. Ministerio de Tecnologías de la Información y las

Comunicaciones, Decreto 1377 de 2013. Disponible en https://www.mintic.gov.co/portal/604/articles-4274_documento.pdf

[41] Superintendencia de Industria y Comercio. Manejo de información personal, 'Habeas data'. 21 de septiembre de 2019. Disponible en http://www.sic.gov.co/manejo-de-informacion-personal.

[42] Gobierno de Colombia. Alcaldía de Bogotá, Disposiciones generales para la protección de datos personales, Ley 1581 de 2012. Disponible en https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=49981.

[43] Gobierno de Colombia. Ministerio de Tecnologías de la Información y las Comunicaciones, Ley 1712 de 2014. Disponible en https://www.mintic.gov.co/portal/604/articles-7147_documento.pdf.

[44] Oficina de Publicaciones de la Unión Europea. Directiva de Protección de Datos, Reglamento general de protección de datos 2016/679. 27 de abril de 2016. Disponible en https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_es.

[45] Oficina de Publicaciones de la Unión Europea. Directiva de Protección de Datos 95/46/EC. 24 de octubre de 1995. Disponible en https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:31995L0046&from=EN.

[46] Organización para la Cooperación y el Desarrollo Económicos (OCDE). Directrices sobre protección de la privacidad y flujos transfronterizos de datos personales. Disponible en https://www.oecd.org/sti/ieconomy/15590267.pdf.

[47] Diaz, P., Jackson, M., & Motz, R. (2015). Learning analytics y protección de datos personales: Recomendaciones [Learning analytics and personal data protection: Recommendations]. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação [Proceedings of the Brazilian Congress of Informatics in Education] (p. 981). Disponible en http://br-ie.org/pub/index.php/wcbie/article/view/6199.

[48] Tobon, C. (2015). Data privacy laws in Latin America: An overview. International Law News, 44(2). Retrieved from http://www.americanbar.org/publications/international_law_news/2015/spring/data_privacy_laws_latin_america_overview.html.

[49] DLA Piper. (2017). Global data protection laws of the world - Full handbook. Disponible en https://www.dlapiperdataprotection.com/.

[50] Holland, J. L. (1965). Manual for the Vocational Preference In- ventory. Iowa City, IA, E.U.: Educational Research Asso- ciates.

[51] Holland, J. L. (1959). A theory of vocational choice. Journal of Counseling Psychology, 6, 35-45.

[52] Holland, J. L. (1975). La elección vocacional. Teoría de las carreras. México: Trillas.

[53] Holland, J. L. (1977). Manual for the Vocational Preference Inven- tory. Palo Alto, California, E.U.: Consulting Psychologist Press.

[54] Holland, J. L. (1992). Making vocational choices: A theory of voca- tional personalities and work environments, 2nd. ed. Odessa, FL, E.U.: Psychological Assessment Resources.

[55] Holland, J. L., Powell, A. B., & Fritzsche, B. A. (1994). The self-directed search (SDS). Odessa, FL: Psychological Assessment Resources.

[56] Holland, J. L. (1995a). Self-directed search. Cuaderno de evaluación. Una guía para la planificación educacional y vocacional. Odessa FL, E.U.: Psychological Assessment Resources.

[57] Holland, J. L. (1995b). Self-directed search. Descubridor de ocupaciones. Odessa FL, E.U.: Psychological Assessment Resources.
[58] Holland, J. L. (1995c). Self-directed search. Usted y su carrera. Odessa FL, E.U.: Psychological Assessment Resources.

[59] Holland, J. L. (1997). Making vocational choices: A theory of voca- tional personalities and work environments, 3rd. ed. Odessa, FL, E.U.: Psychological Assessment Resources.

[60] Holland, J. L. (1981). Técnica de la elección vocacional. Tipos de personalidad y modelos ambientales. México: Trillas.

[61] Martínez Vicente, J. M. (2007). El asesoramiento vocacional y profesional

a través del Self-Directed Search (SDS).

[62] Gardner, H. (1983). La Teoría de las Múltiples Inteligencias.

[63] Vicente, J. M. M., & Fernández, F. V. (2006). La elección vocacional y la planificación de la carrera. Adaptación española del Self-Directed Search (SDS-R) de Holland [The vocational choice and the career planning. Spanish adaptation of Holland's Self-Directed Search (SDS-R)]. Psicothema, 18(1), 117-122.

[64] Agudelo-Valderrama, Cecilia, Barbara Clarken y Alan J. Bishop. 2007. "Explanations of Attitutdes to Change: Colombian Mathematics Teachers" Conceptions of the Crucial Determinants to their Teaching Practices of Beginning Algebra." Journal of Mathematics Teacher Education.

[65] Manual de Santiago (2007). Manual de Indicadores de Internacionalización de la Ciencia y de la
Tecnología. Red Iberoamericana de Indicadores de ciencia y tecnología (RICYT).

[66] LLECE–UNESCO. Segundo Estudio Regional Comparativo y Explicativo (SERCE), 2008.

[67] Valverde, Gilbert y Emma Näslund-Hadley (2010), La condición de la educación en matemáticas y ciencias naturales en América Latina y el Caribe, Washington DC, Banco Interamericano de Desarrollo-División de Educación.

[68] Tamirán-Pereira, R., Calderón-Romero, A., Jiménez-Toledo, J.: Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. Revista Vínculos, Vol. 10, No. 1, pp. 373–383 (2013).

[69] Albor, Gustavo & Lorduy, Viviana & Dau, Marco. (2014). Calidad de la educación superior a distancia y virtual: Un análisis de desempeño académico en Colombia. Investigación & Desarrollo. 22. 79-119. 10.14482/indes.22.1.6079

[70] Cantillo, V., & García, L. (2014). Gender and Other Factors Influencing the Outcome of a Test to Assess Quality of Education in Civil Engineering in Colombia. Journal of Professional Issues in Engineering Education Practise, (2004), 1–7. http://doi.org/10.1061/(ASCE)EI.1943-5541.0000194

[71] Rodríguez Albor, G., Ariza Dau, M., & Ramos Ruíz, J. L. (2014). Calidad institucional y rendimiento académico El caso de las universidades del Caribe Colombiano. Perfiles Educativos, 36(143), 10–29. http://doi.org/10.1016/S0185-2698(14)70607-5

[72] Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A. y Alvarado Pérez, J. C. (2016). Desempeño académico y competencias genéricas en la formación de profesionales. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 19-62). Bogotá: Ediciones Universidad Cooperativa de Colombia. doi: http://dx.doi.org/10.16925/9789587600490

[73] Said-Hung, Elias, & Gratacós, Gloria, & Valencia Cobos, Jorge (2017). Factores que influyen en la elección de las carreras de pedagogía en Colombia. Educação e Pesquisa, 43(1), undefined-undefined. [fecha de consulta 21 de septiembre de 2019]. ISSN: 1517-9702. Disponible en: http://www.redalyc.org/articulo.oa?id=298/29849949003

[74] García, Sandra et al. Tras la excelencia docente: cómo mejorar la calidad de la educación para todos los colombianos. Bogotá: Fundación Compartir, 2014.

[75] Avendaño, Cecilia & González, Rodrigo. Motivos para ingresar a las carreras de Pedagogía de los estudiantes de primer año de la Universidad de Concepción. Estudios Pedagógicos, Valdivia, n. 38, n. 2, p. 21-33, 2012.

[76] RYAN Richard; DECI, Edward. Self-determination theory and the facilitation of intrinsic motivation, social development, and wellbeing. American Psychologist, v. 55, n. 1, p. 68-78, jan. 2000.

[77] Gratacós, Gloria. Estudio sobre las motivaciones en la elección de ser maestro, 2014. 357 p. Tesis (Doctorado) – Universidad Internacional de Cataluña, Barcelona, 2014.

[78] Oviedo Carrascal, Ana & Giraldo, Jovanny. (2018). Minería de datos educativos: análisis de los factores económicos, sociales y demográficos que influyen en el desempeño de las Pruebas Saber-Pro en estudiantes de ingeniería en Antioquia.

[79] Cantillo, V., y García, L. Gender and Other Factors Influencing the Outcome of a Test to Assess Quality of Education in Civil Engineering in Colombia. Journal of Professional Issues in Engineering Education Practise, 1–7, 2014.

ANEXOS

Anexo 1. Exploración estadística de los datos

Generación del reporte de perfilado de todas las variables, con las líneas de código en Python:

```
pip install pandas-profiling
report = pandas_profiling.ProfileReport(df)
```

Reporte

Dataset info	
Number of variables	183
Number of observations	140324
Total Faltantes (%)	7.3%
Total size in memory	95.2 MiB
Average record size in memory	711.1 B

Variables types	
Numérica	24
Categórica	142
Boolean	0
Date	0
Text (Únicos)	0
Rejected	17
Unsupported	0

Warnings

- 1-11_estu_exam_nombreexamen has a high cardinality: 140227 distinct values Warning
- 3-11_estu_consecutivo has a high cardinality: 140227 distinct values Warning
- 11-11_estu_cod_reside_mcpio has a high cardinality: 1097 distinct values Warning
- 12-11_estu_reside_mcpio has a high cardinality: 1014 distinct values Warning
- 17-11_cole_cod_icfes has a high cardinality: 9712 distinct values Warning
- 18-11_cole_cod_dane_institucion has a high cardinality: 252 distinct values Warning
- 19-11_cole_nombre_sede has a high cardinality: 8123 distinct values Warning
- 20-11_cole_cod_mcpio_ubicacion has a high cardinality: 1095 distinct values Warning
- 28-11_estu_ies_cod_deseada has 125815 / 89.7% Faltantes values Faltantes
- 28-11_estu_ies_cod_deseada has a high cardinality: 295 distinct values Warning
- 29-11_estu_ies_deseada_nombre has 125815 / 89.7% Faltantes values Faltantes
- 29-11_estu_ies_deseada_nombre has a high cardinality: 248 distinct values Warning
- 30-11_estu_ies_cod_mpio_deseada has 125815 / 89.7% Faltantes values Faltantes

- 30-11_estu_ies_cod_mpio_deseada has a high cardinality: 69 distinct values Warning
- 31-11_estu_ies_mpio_deseada has 125815 / 89.7% Faltantes values Faltantes
- 31-11_estu_ies_mpio_deseada has a high cardinality: 69 distinct values Warning
- 32-11_estu_ies_dept_deseada has 125815 / 89.7% Faltantes values Faltantes
- 33-11_estu_carrdeseada_tipo has 126737 / 90.3% Faltantes values Faltantes
- 34-11_estu_exam_cod_mpio_presentacio has a high cardinality: 445 distinct values Warning
- 35-11_estu_mpio_presentacion has a high cardinality: 428 distinct values Warning
- 49-11_desemp_comp_flexible has 68660 / 48.9% Faltantes values Faltantes
- 52-11_fami_estrato_vivienda has 3049 / 2.2% Faltantes values Faltantes
- 56-11_fami_ocup_padre has 2758 / 2.0% Faltantes values Faltantes
- 57-11_fami_ocup_madre has 2758 / 2.0% Faltantes values Faltantes
- 64-11_fami_computador has 2228 / 1.6% Faltantes values Faltantes
- 74-G_estu_dia_nac is highly correlated with 8-11_estu_nacimiento_dia (ρ = 0.98162) Rejected
- 75-G_estu_mes_nac is highly correlated with 9-11_estu_nacimiento_mes (ρ = 0.97838) Rejected
- 78-G_estu_consecutivo has a high cardinality: 139844 distinct values Warning
- 81-G_estu_mcpio_reside has a high cardinality: 915 distinct values Warning
- 82-G_estu_cod_reside_mcpio has a high cardinality: 976 distinct values Warning
- 87-G_estu_etnia has 72491 / 51.7% Faltantes values Faltantes
- 88-G_estu_limita_motriz has 75460 / 53.8% Faltantes values Faltantes
- 89-G_estu_limita_invidente has 75486 / 53.8% Faltantes values Faltantes
- 90-G_estu_limita_condicionespecial has 75500 / 53.8% Faltantes values Faltantes
- 91-G_estu_limita_sordo has 75484 / 53.8% Faltantes values Faltantes
- 92-G_estu_limita_autismo has 75502 / 53.8% Faltantes values Faltantes
- 94-G_estu_cole_termino has 22622 / 16.1% Faltantes values Faltantes
- 94-G_estu_cole_termino has a high cardinality: 7493 distinct values Warning
- 95-G_estu_coddane_cole_termino has 22622 / 16.1% Faltantes values Faltantes
- 95-G_estu_coddane_cole_termino has a high cardinality: 225 distinct values Warning
- 96-G_estu_cod_cole_mcpio_termino has 22622 / 16.1% Faltantes values Faltantes
- 96-G_estu_cod_cole_mcpio_termino has a high cardinality: 1058 distinct values Warning
- 97-G_estu_otrocole_termino has 59751 / 42.6% Faltantes values Faltantes
- 97-G_estu_otrocole_termino has a high cardinality: 19807 distinct values Warning
- 99-G_estu_cursodocentesies has 69138 / 49.3% Faltantes values Faltantes
- 100-G_estu_cursoiesapoyoexterno has 69140 / 49.3% Faltantes values Faltantes
- 101-G_estu_cursoiesexterna has 69143 / 49.3% Faltantes values Faltantes
- 102-G_estu_simulacrotipoicfes has 69137 / 49.3% Faltantes values Faltantes
- 103-G_estu_actividadrefuerzoareas has 69138 / 49.3% Faltantes values Faltantes
- 104-G_estu_actividadrefuerzogeneric has 69140 / 49.3% Faltantes values Faltantes
- 105-G_inst_cod_institucion has a high cardinality: 322 distinct values Warning
- 106-G_inst_nombre_institucion has a high cardinality: 325 distinct values Warning
- 108-G_estu_prgm_academico has a high cardinality: 556 distinct values Warning
- 109-G_estu_snies_prgmacademico has a high cardinality: 2839 distinct values Warning
- 112-G_estu_prgm_codmunicipio has a high cardinality: 184 distinct values Warning
- 113-G_estu_prgm_municipio has a high cardinality: 184 distinct values Warning
- 117-G_estu_nucleo_pregrado has a high cardinality: 56 distinct values Warning
- 118-G_estu_inst_codmunicipio has a high cardinality: 121 distinct values Warning
- 119-G_estu_inst_municipio has a high cardinality: 122 distinct values Warning
- 123-G_estu_cod_mcpio_presentacion has a high cardinality: 113 distinct values Warning
- 124-G_estu_mcpio_presentacion has a high cardinality: 114 distinct values Warning
- 128-G_razona_cuant_desem is highly correlated with 127-G_razona_cuant_punt (ρ = 0.91938) Rejected
- 129-G_razona_cuant_pnal is highly correlated with 128-G_razona_cuant_desem (ρ = 0.93122) Rejected
- 132-G_lectura_critica_desem is highly correlated with 131-G_lect_critica_punt (ρ = 0.93464) Rejected
- 133-G_lect_critica_pnal is highly correlated with 132-G_lectura_critica_desem (ρ = 0.93602) Rejected
- 134-G_lect_critica_pgref is highly correlated with 133-G_lect_critica_pnal (ρ = 0.95989) Rejected

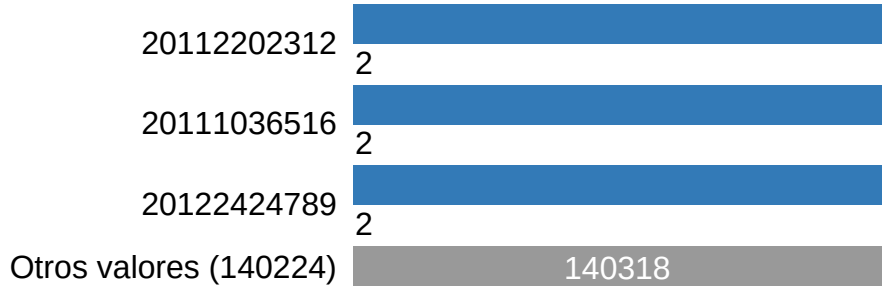
- 136-G_compet_ciudad_desem is highly correlated with 135-G_compet_ciudad_punt (ρ = 0.92888) Rejected
- 137-G_compet_ciudad_pnal is highly correlated with 136-G_compet_ciudad_desem (ρ = 0.93182) Rejected
- 138-G_compet_ciudad_pgref is highly correlated with 137-G_compet_ciudad_pnal (ρ = 0.96363) Rejected
- 141-G_ingles_pnal is highly correlated with 139-G_ingles_punt (ρ = 0.96073) Rejected
- 142-G_ingles_pgref is highly correlated with 141-G_ingles_pnal (ρ = 0.9487) Rejected
- 143-G_com_escrita_punt has 2205 / 1.6% zeros Zeros
- 144-G_com_escrita_desem is highly correlated with 143-G_com_escrita_punt (ρ = 0.93129) Rejected
- 145-G_com_escrita_pnal is highly correlated with 144-G_com_escrita_desem (ρ = 0.92739) Rejected
- 146-G_com_escrita_pgref is highly correlated with 145-G_com_escrita_pnal (ρ = 0.99032) Rejected
- 148-G_percentil_global is highly correlated with 147-G_punt_global (ρ = 0.97408) Rejected
- 164-G_fami_tienehornomicroogas has 1883 / 1.3% Faltantes values Faltantes
- 166-G_fami_tienemotocicleta has 3462 / 2.5% Faltantes values Faltantes
- 171-G_estu_tiporemuneracion has 34288 / 24.4% Faltantes values Faltantes
- 172-G_estu_inse_individual has a high cardinality: 97379 distinct values Warning
- 179-e_estu_consecutivo has a high cardinality: 139844 distinct values Warning
- 183-e_result_desemp is highly correlated with 182-e_result_puntaje (ρ = 0.93529) Rejected

Variables

1-11_estu_exam_nombreexamen

Categórica

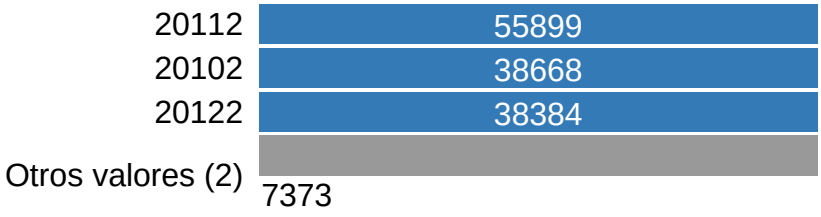
Valores distintos	140227
Únicos (%)	99.9%
Faltantes (%)	0.0%
Faltantes (n)	0



2-11_periodo-anno

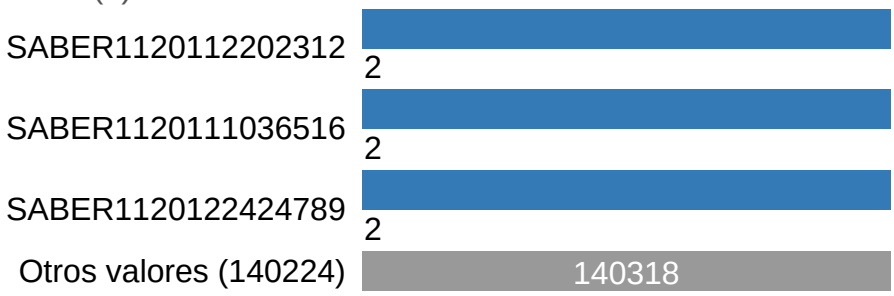
Categórica

Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



3-11_estu_consecutivo
Categórica

Valores distintos	140227
Únicos (%)	99.9%
Faltantes (%)	0.0%
Faltantes (n)	0

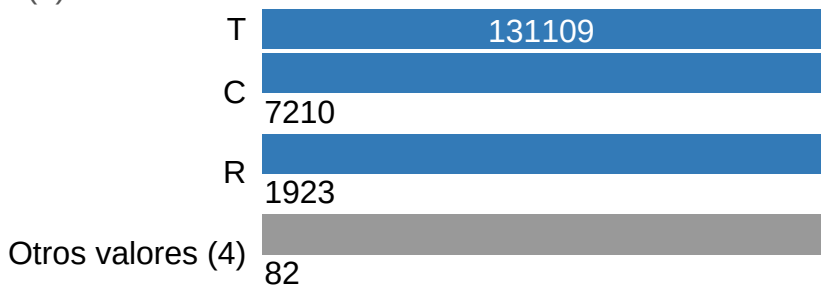


4-11_estu_edad
Numérica

Valores distintos	52
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	16.289
Minimum	9
Maximum	63
Zeros (%)	0.0%

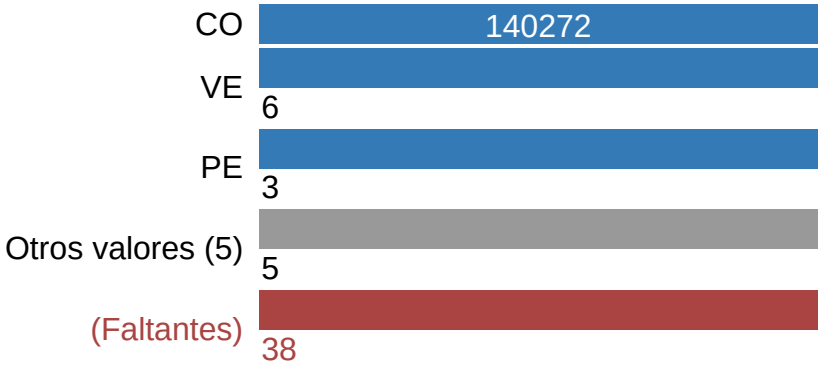
5-11_estu_tipo_documento
Categórica

Valores distintos	7
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



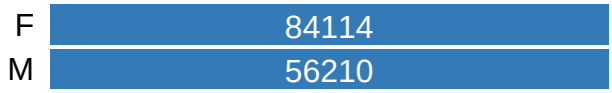
6-11_estu_pais_reside
Categórica

Valores distintos	9
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	38



7-11_estu_genero
Categórica

Valores distintos	2
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



8-11_estu_nacimiento_dia
Numérica

Valores distintos	31
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	15.698
Minimum	1
Maximum	31
Zeros (%)	0.0%

9-11_estu_nacimiento_mes
Numérica

Valores distintos	12
Únicos (%)	0.0%

Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	6.6463
Minimum	1
Maximum	12
Zeros (%)	0.0%

10-11_estu_nacimiento_anno

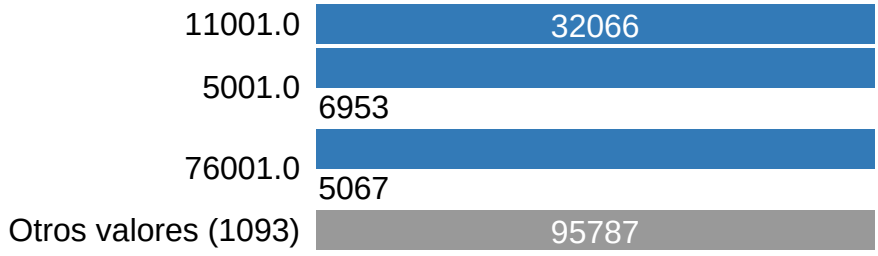
Numérica

Valores distintos	55
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	1994.2
Minimum	1949
Maximum	2004
Zeros (%)	0.0%

11-11_estu_cod_reside_mcpio

Categórica

Valores distintos	1097
Únicos (%)	0.8%
Faltantes (%)	0.3%
Faltantes (n)	451

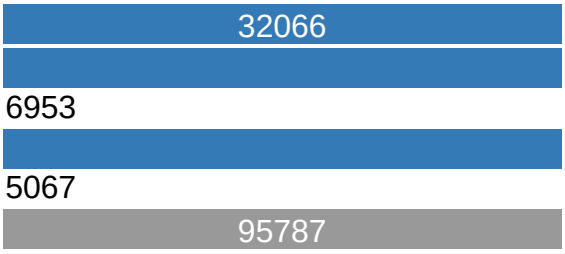


12-11_estu_reside_mcpio

Categórica

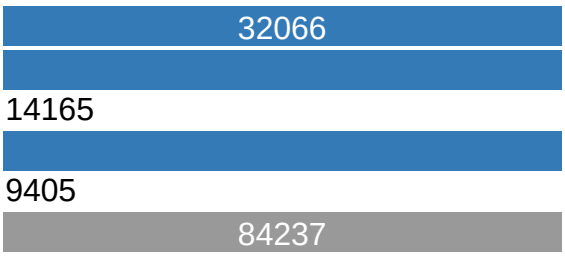
Valores distintos	1014
Únicos (%)	0.7%
Faltantes (%)	0.3%

Faltantes (n) 451



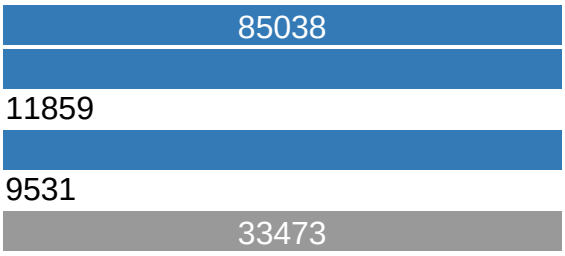
13-11_estu_reside_depto
Categórica

Valores distintos	34
Únicos (%)	0.0%
Faltantes (%)	0.3%
Faltantes (n)	451



14-11_estu_zona_reside
Categórica

Valores distintos	11
Únicos (%)	0.0%
Faltantes (%)	0.3%
Faltantes (n)	423

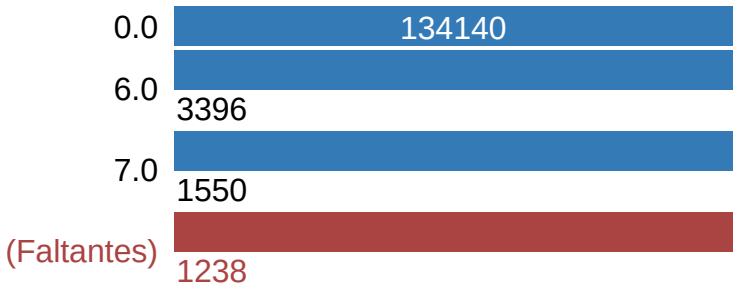


15-11_estu_area_reside
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	50

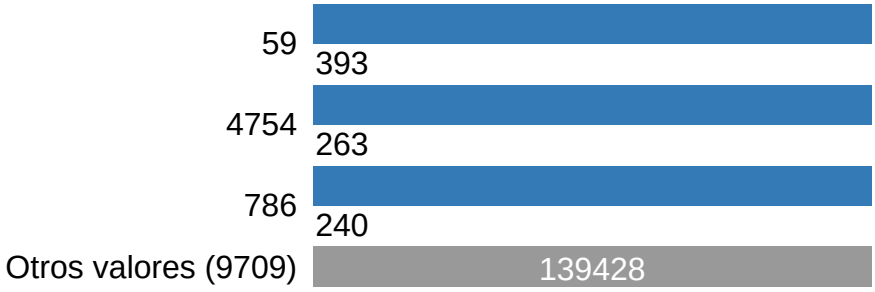
16-11_estu_trabaja
Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.9%
Faltantes (n)	1238



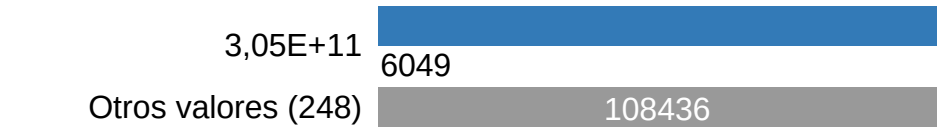
17-11_cole_cod_icfes
Categórica

Valores distintos	9712
Únicos (%)	6.9%
Faltantes (%)	0.0%
Faltantes (n)	0



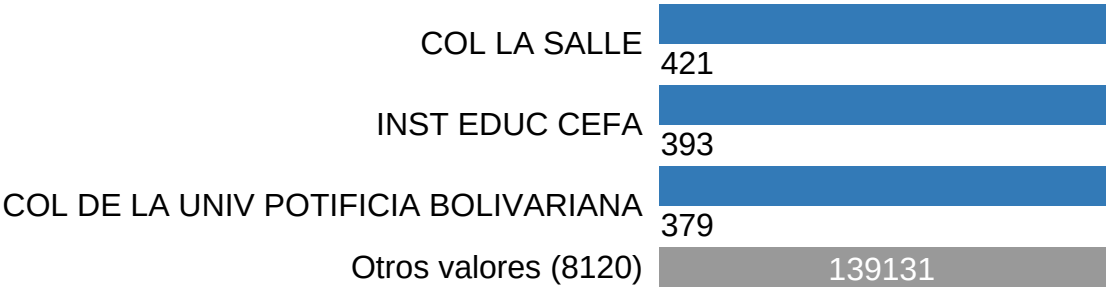
18-11_cole_cod_dane_institucion
Categórica

Valores distintos	252
Únicos (%)	0.2%
Faltantes (%)	0.2%
Faltantes (n)	306



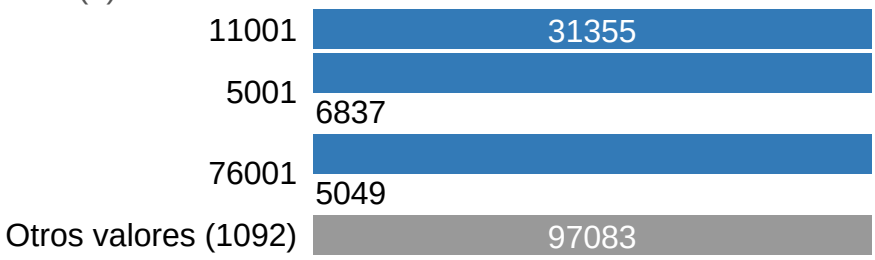
19-11_cole_nombre_sede
Categórica

Valores distintos	8123
Únicos (%)	5.8%
Faltantes (%)	0.0%
Faltantes (n)	0



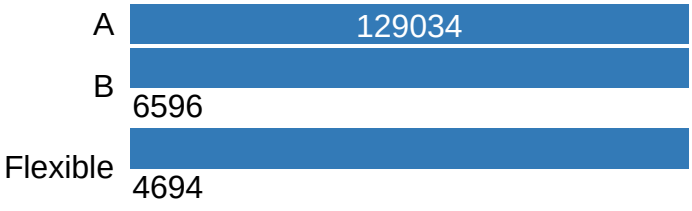
20-11_cole_cod_mcpio_ubicacion
Categórica

Valores distintos	1095
Únicos (%)	0.8%
Faltantes (%)	0.0%
Faltantes (n)	0



21-11_cole_calendario
Categórica

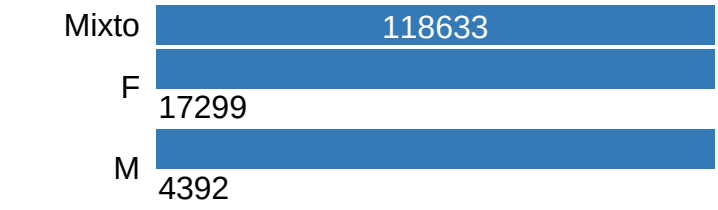
Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



22-11_cole_genero

Categórica

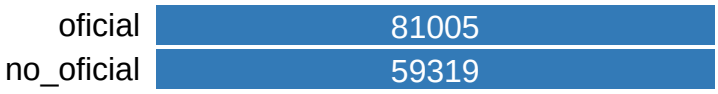
Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



23-11_cole_naturaleza

Categórica

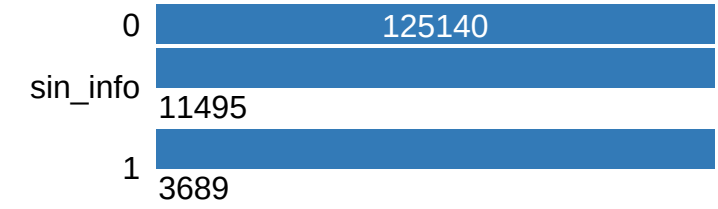
Valores distintos	2
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



24-11_cole bilingüe

Categórica

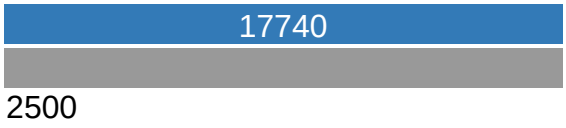
Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



25-11_cole_jornada

Categórica

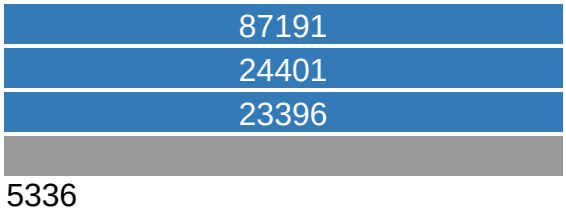
Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



26-11_cole_caracter

Categórica

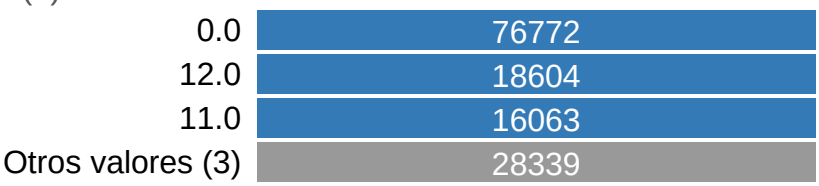
Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



27-11_cole_valor_pension

Categórica

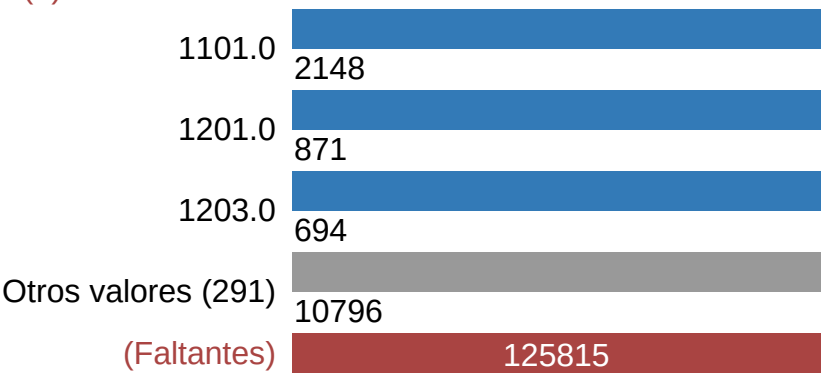
Valores distintos	7
Únicos (%)	0.0%
Faltantes (%)	0.4%
Faltantes (n)	546



28-11_estu_ies_cod_deseada

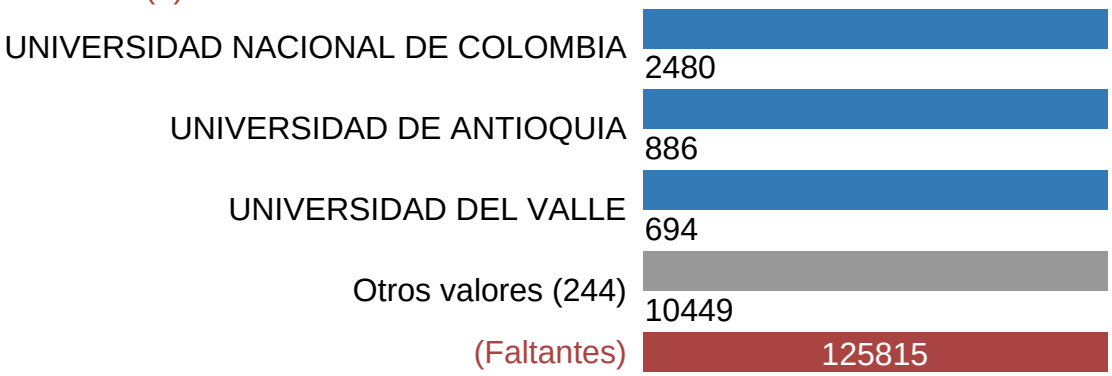
Categórica

Valores distintos	295
Únicos (%)	0.2%
Faltantes (%)	89.7%
Faltantes (n)	125815



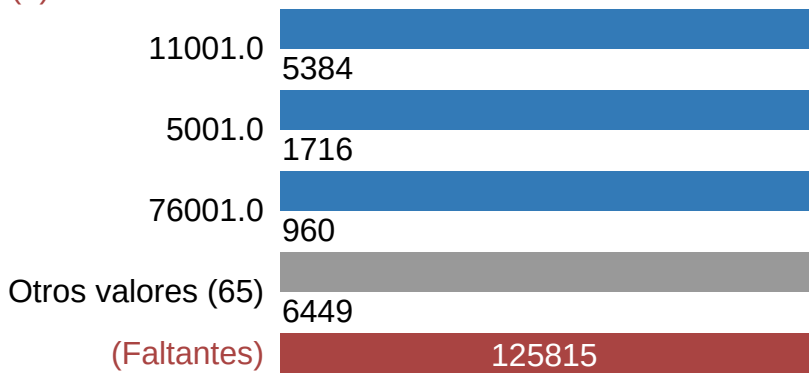
29-11_estu_ies_deseada_nombre
Categórica

Valores distintos	248
Únicos (%)	0.2%
Faltantes (%)	89.7%
Faltantes (n)	125815



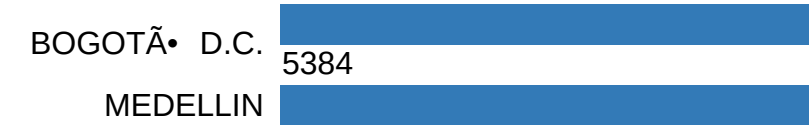
30-11_estu_ies_cod_mpio_deseada
Categórica

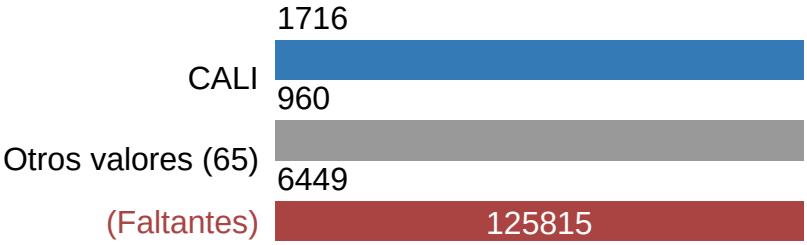
Valores distintos	69
Únicos (%)	0.0%
Faltantes (%)	89.7%
Faltantes (n)	125815



31-11_estu_ies_mpio_deseada
Categórica

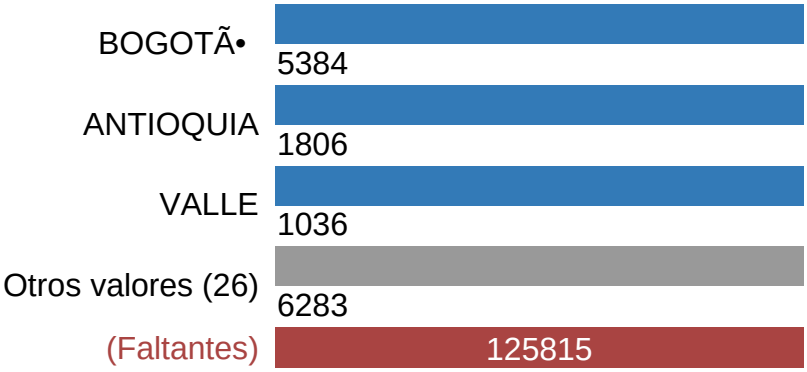
Valores distintos	69
Únicos (%)	0.0%
Faltantes (%)	89.7%
Faltantes (n)	125815





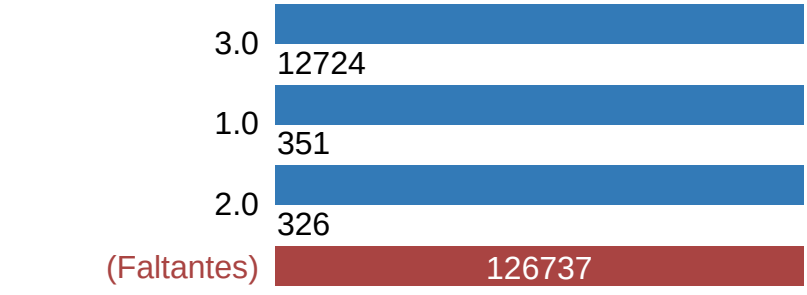
32-11_estu_ies_dept_deseada
Categórica

Valores distintos	30
Únicos (%)	0.0%
Faltantes (%)	89.7%
Faltantes (n)	125815



33-11_estu_carrdeseada_tipo
Categórica

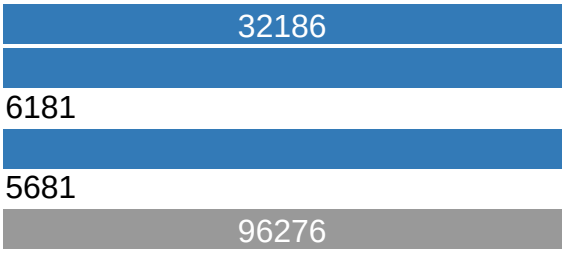
Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	90.3%
Faltantes (n)	126737



34-11_estu_exam_cod_mpio_presentacio
Categórica

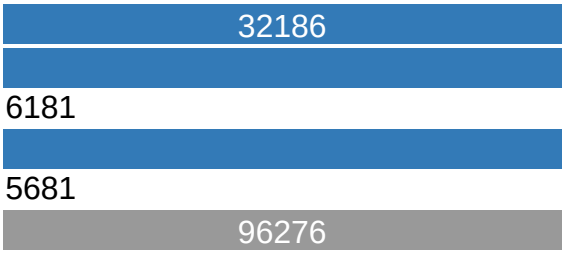
Valores distintos	445
Únicos (%)	0.3%
Faltantes (%)	0.0%

Faltantes (n) 0



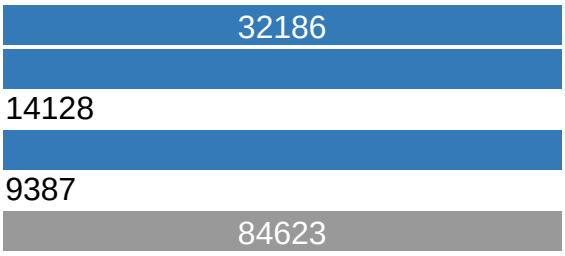
35-11_estu_mpio_presentacion
Categórica

Valores distintos	428
Únicos (%)	0.3%
Faltantes (%)	0.0%
Faltantes (n)	0



36-11_estu_dept_presentacion
Categórica

Valores distintos	33
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



37-11_estu_estud-presentado-examen
Categórica

Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.2%
Faltantes (n)	213

38-11_punt_lenguaje

Numérica

Valores distintos	188
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	53.561
Minimum	0
Maximum	102.62
Zeros (%)	0.0%

39-11_punt_matematicas

Numérica

Valores distintos	230
Únicos (%)	0.2%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	55.152
Minimum	0
Maximum	126
Zeros (%)	0.0%

40-11_punt_c_sociales

Numérica

Valores distintos	224
Únicos (%)	0.2%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	52.734
Minimum	0
Maximum	107

Zeros (%)	0.0%

41-11_punt_filosofia
Numérica

Valores distintos	200
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	48.995
Minimum	0
Maximum	103
Zeros (%)	0.1%

42-11_punt_biologia
Numérica

Valores distintos	197
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	52.618
Minimum	0
Maximum	105.78
Zeros (%)	0.0%

43-11_punt_quimica
Numérica

Valores distintos	196
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	52.074
Minimum	0
Maximum	118.8
Zeros (%)	0.0%

44-11_punt_fisica
Numérica

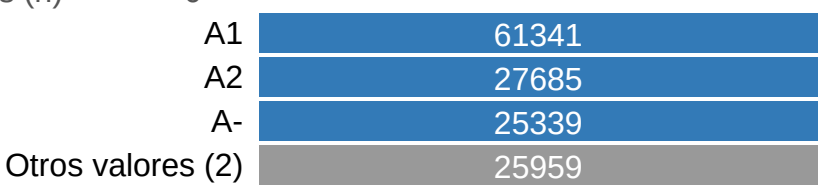
Valores distintos	212
Únicos (%)	0.2%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	50.933
Minimum	0
Maximum	124
Zeros (%)	0.0%

45-11_punt_ingles
Numérica

Valores distintos	202
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	52.32
Minimum	0
Maximum	117.29
Zeros (%)	0.0%

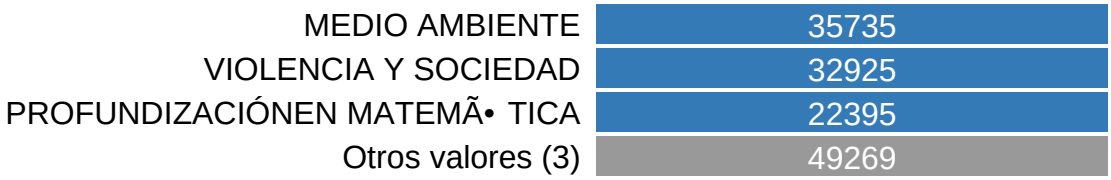
46-11_desemp_ingles
Categórica

Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



47-11_nombre_comp_flexible
Categórica

Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



48-11_punt_comp_flexible

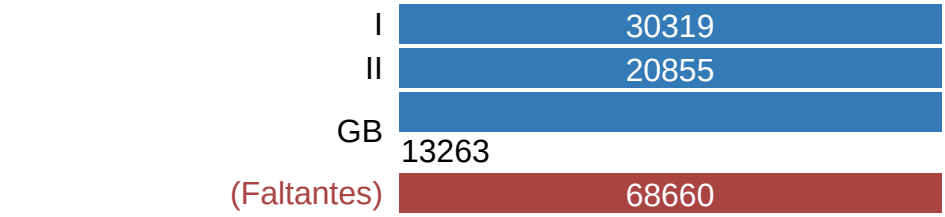
Numérica

Valores distintos	164
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	28.416
Minimum	0
Maximum	105.44
Zeros (%)	0.2%

49-11_desemp_comp_flexible

Categórica

Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	48.9%
Faltantes (n)	68660



50-11_estu_puesto

Numérica

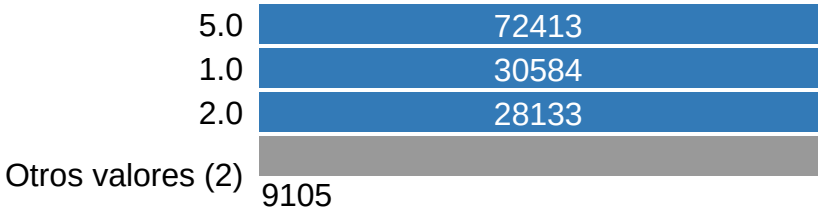
Valores distintos	1000
Únicos (%)	0.7%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%

Infinite (n)	0
Mean	263.6
Minimum	1
Maximum	1000
Zeros (%)	0.0%

51-11_fami_nivel_sisben

Categórica

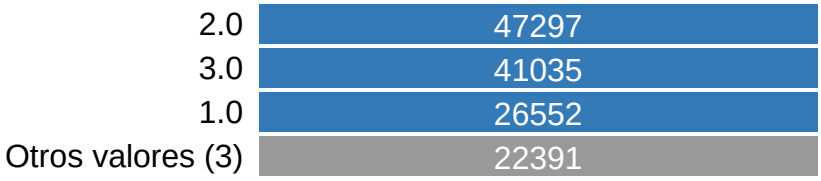
Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89



52-11_fami_estrato_vivienda

Categórica

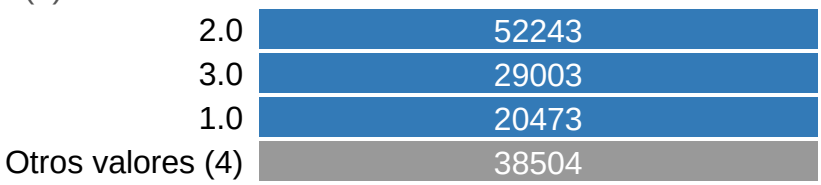
Valores distintos	7
Únicos (%)	0.0%
Faltantes (%)	2.2%
Faltantes (n)	3049



53-11_fami_ing_fmiliar_mensual

Categórica

Valores distintos	8
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	101



54-11_fami_educa_padre
Categórica

Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0

bachiller	41496
pregrado	38133
primaria	29170
Otros valores (3)	31525

55-11_fami_educa_madre
Categórica

Valores distintos	7
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0

bachiller	41440
pregrado	38076
primaria	28944
Otros valores (4)	31864

56-11_fami_ocup_padre
Categórica

Valores distintos	13
Únicos (%)	0.0%
Faltantes (%)	2.0%
Faltantes (n)	2758

21.0	34205
19.0	20420
26.0	15105
Otros valores (9)	67836

57-11_fami_ocup_madre
Categórica

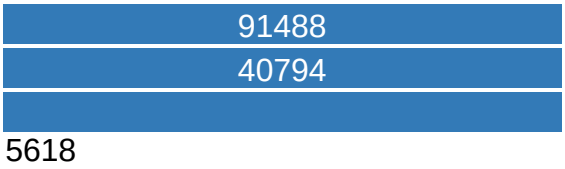
Valores distintos	13
Únicos (%)	0.0%
Faltantes (%)	2.0%
Faltantes (n)	2758

22.0	56844
21.0	14172

58-11_fami_pisos_hogar

Categórica

Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	158



59-11_fami_personas_hogar

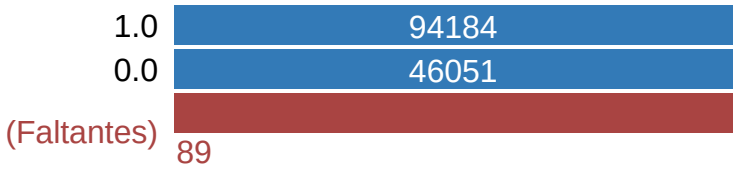
Numérica

Valores distintos	13
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	98
Infinite (%)	0.0%
Infinite (n)	0
Mean	4.4941
Minimum	1
Maximum	12
Zeros (%)	0.0%

60-11_fami_telefono_fijo

Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89

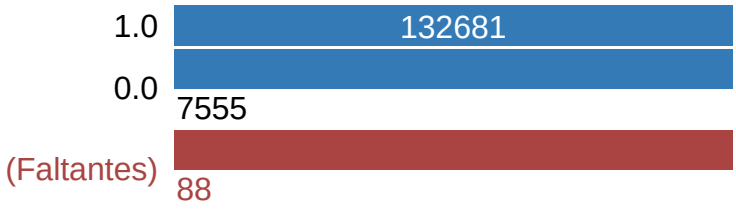


61-11_fami_celular

Categórica

Valores distintos	3

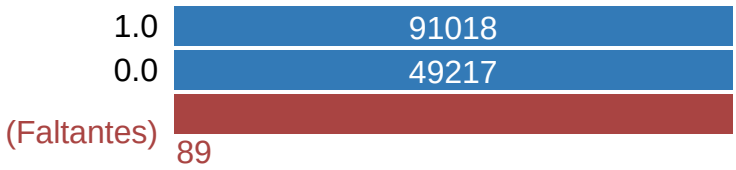
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	88



62-11_fami_internet

Categórica

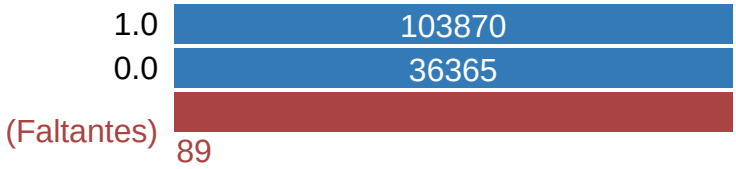
Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89



63-11_fami_servicio_television

Categórica

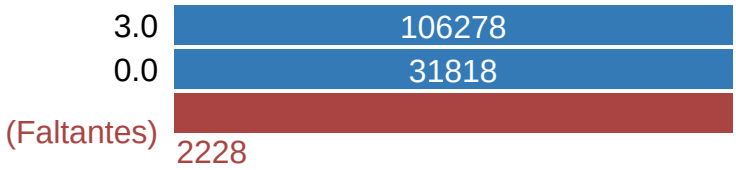
Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89



64-11_fami_computador

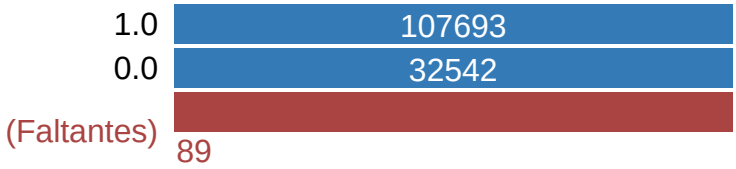
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	1.6%
Faltantes (n)	2228



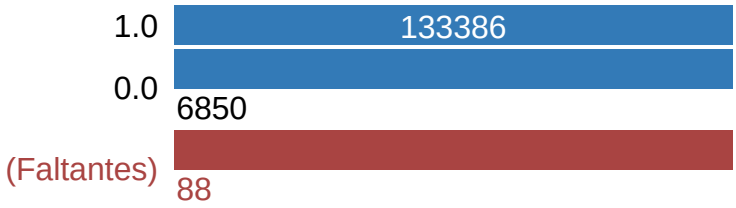
65-11_fami_lavadora
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89



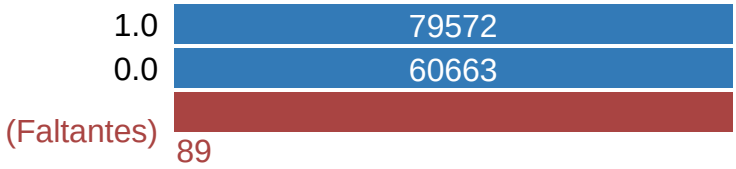
66-11_fami_nevera
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	88



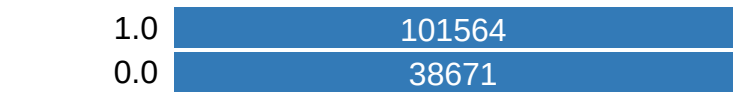
67-11_fami_horno
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89



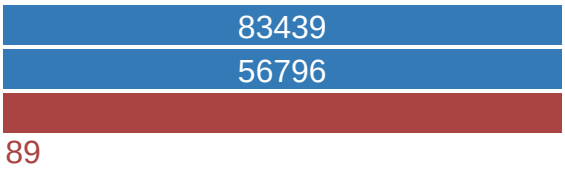
68-11_fami_dvd
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89



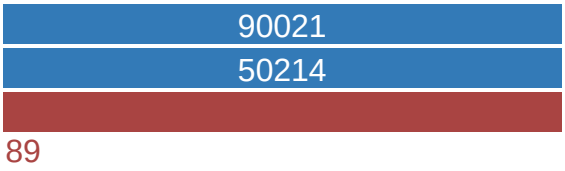
69-11_fami_microondas
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89



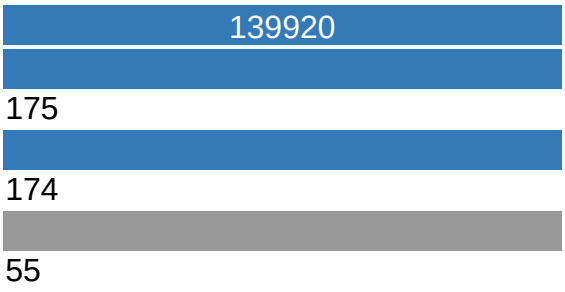
70-11_fami_automovil
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	89



71-G_estu_tipodocumento
Categórica

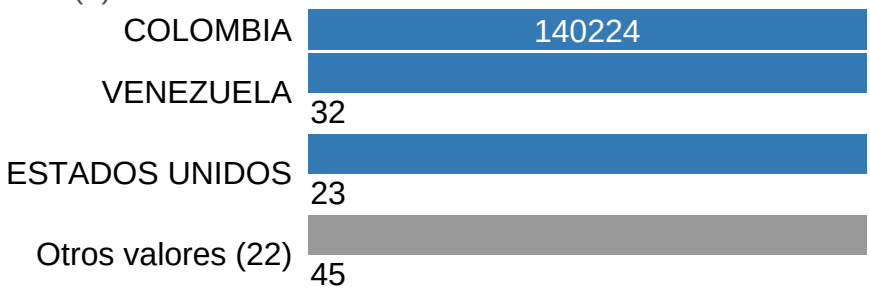
Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



72-G_estu_nacionalidad
Categórica

Valores distintos	25

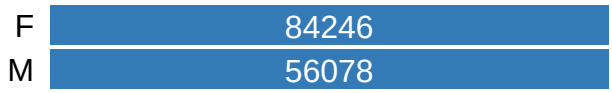
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



73-G_estu_genero

Categórica

Valores distintos	2
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



~~74-G_estu_dia_nac~~

Highly correlated

This variable is highly correlated with 8-11_estu_nacimiento_dia and should be ignored for analysis

Correlation 0.98162

~~75-G_estu_mes_nac~~

Highly correlated

This variable is highly correlated with 9-11_estu_nacimiento_mes and should be ignored for analysis

Correlation 0.97838

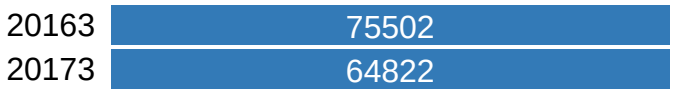
76-G_estu_ano_nac

Numérica

Valores distintos	53
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	1994.2
Minimum	1949
Maximum	2002
Zeros (%)	0.0%

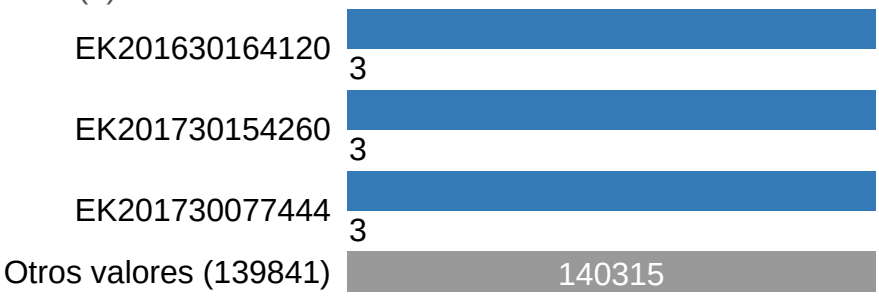
77-G_periodo
Categórica

Valores distintos	2
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



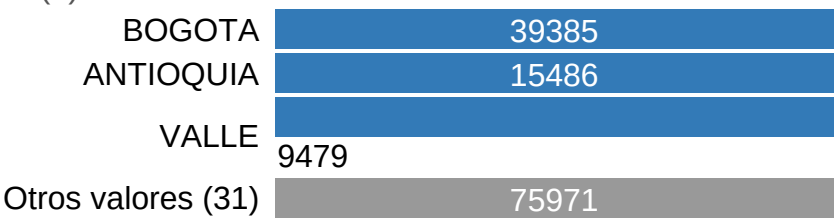
78-G_estu_consecutivo
Categórica

Valores distintos	139844
Únicos (%)	99.7%
Faltantes (%)	0.0%
Faltantes (n)	0



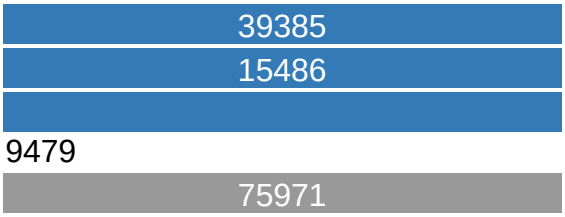
79-G_estu_depto_reside
Categórica

Valores distintos	35
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	3



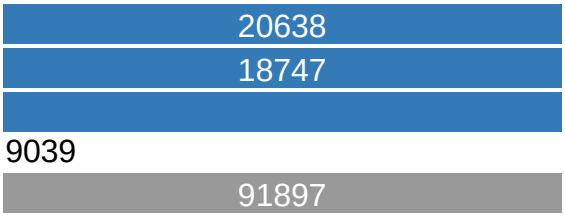
80-G_estu_cod_reside_depto
Categórica

Valores distintos	35
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	3



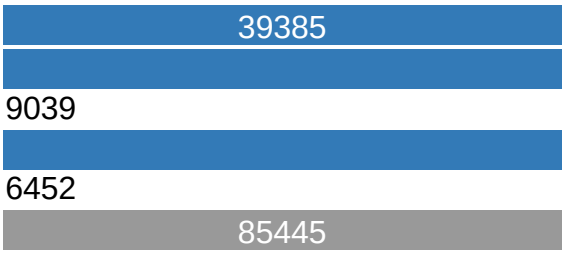
81-G_estu_mcpio_reside
Categórica

Valores distintos	915
Únicos (%)	0.7%
Faltantes (%)	0.0%
Faltantes (n)	3



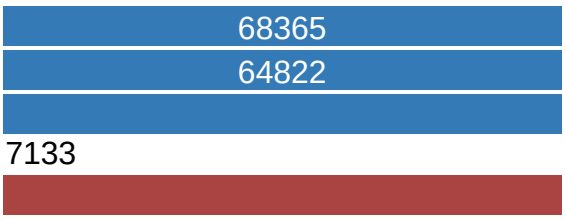
82-G_estu_cod_reside_mcpio
Categórica

Valores distintos	976
Únicos (%)	0.7%
Faltantes (%)	0.0%
Faltantes (n)	3



83-G_estu_areareside
Categórica

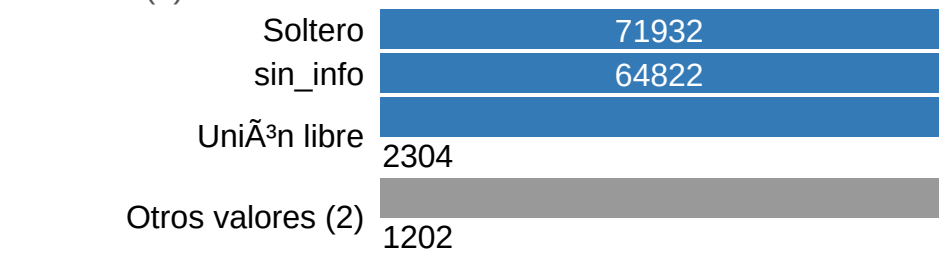
Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



4

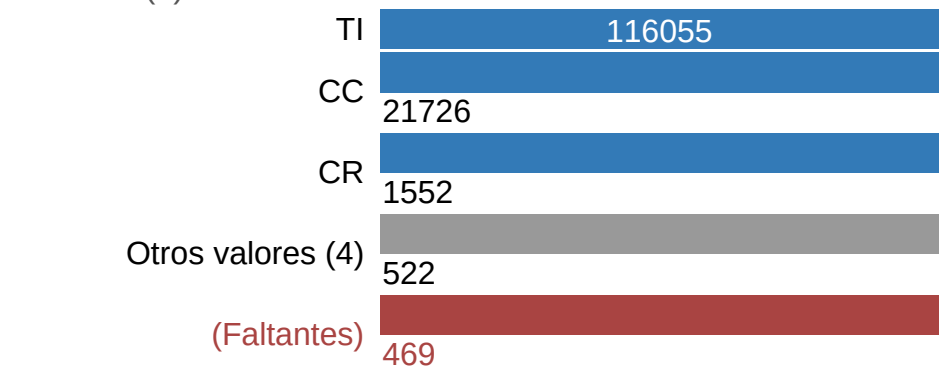
84-G_estu_estadocivil
Categórica

Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	64



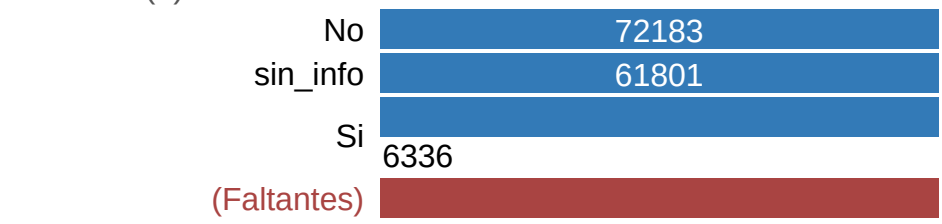
85-G_estu_tipodocumentosb11
Categórica

Valores distintos	8
Únicos (%)	0.0%
Faltantes (%)	0.3%
Faltantes (n)	469



86-G_estu_tieneetnia
Categórica

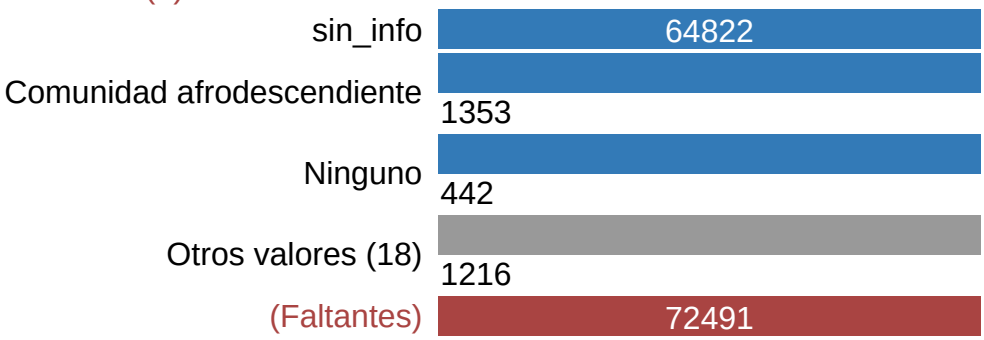
Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



4

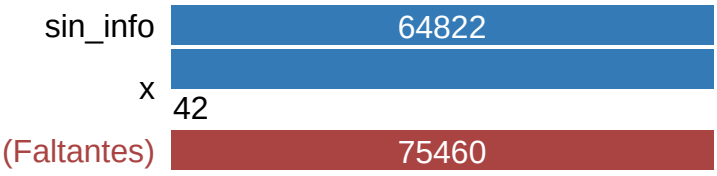
87-G_estu_etnia
Categórica

Valores distintos	22
Únicos (%)	0.0%
Faltantes (%)	51.7%
Faltantes (n)	72491



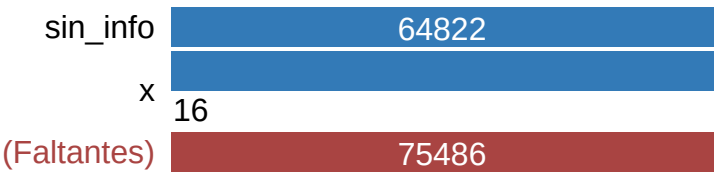
88-G_estu_limita_motriz
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	53.8%
Faltantes (n)	75460



89-G_estu_limita_invidente
Categórica

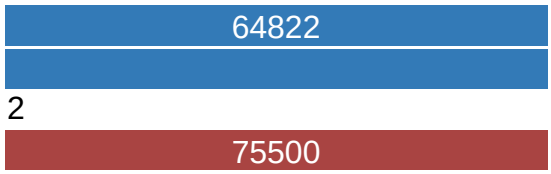
Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	53.8%
Faltantes (n)	75486



90-G_estu_limita_condicionespecial
Categórica

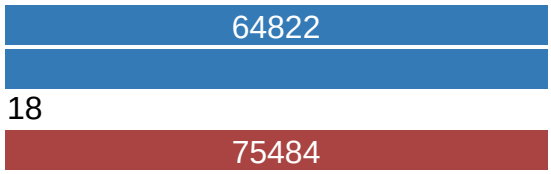
Valores distintos	3

Únicos (%)	0.0%
Faltantes (%)	53.8%
Faltantes (n)	75500



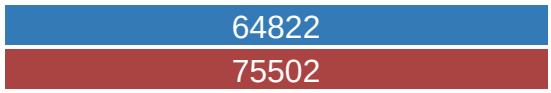
91-G_estu_limita_sordo

Categórica	
Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	53.8%
Faltantes (n)	75484



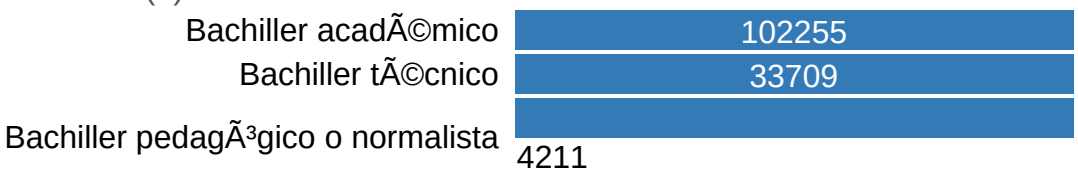
92-G_estu_limita_autismo

Categórica	
Valores distintos	2
Únicos (%)	0.0%
Faltantes (%)	53.8%
Faltantes (n)	75502



93-G_estu_tituloobtenidobachiller

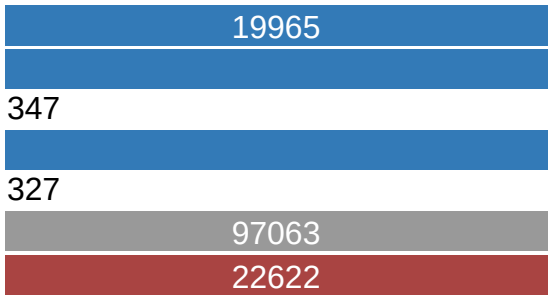
Categórica	
Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	72



94-G_estu_cole_termino

Categórica	
Valores distintos	7493

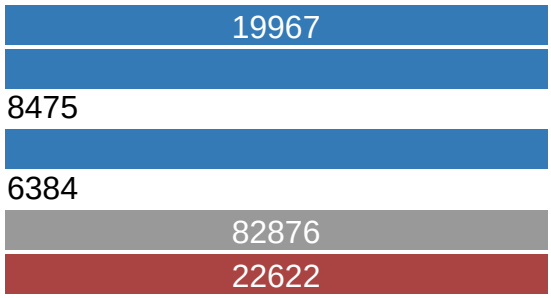
Únicos (%)	5.3%
Faltantes (%)	16.1%
Faltantes (n)	22622



95-G_estu_coddane_cole_termino

Categórica

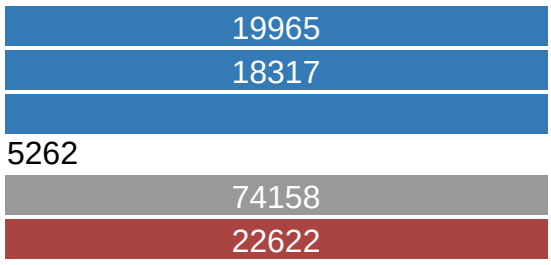
Valores distintos	225
Únicos (%)	0.2%
Faltantes (%)	16.1%
Faltantes (n)	22622



96-G_estu_cod_cole_mcpio_termino

Categórica

Valores distintos	1058
Únicos (%)	0.8%
Faltantes (%)	16.1%
Faltantes (n)	22622

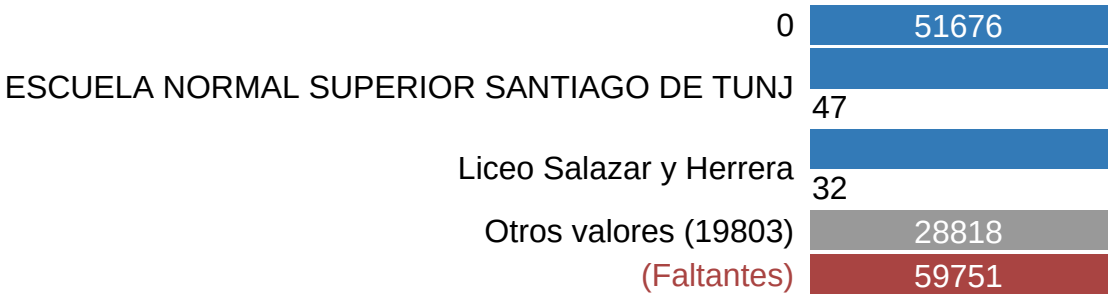


97-G_estu_otrocole_termino

Categórica

Valores distintos	19807
Únicos (%)	14.1%

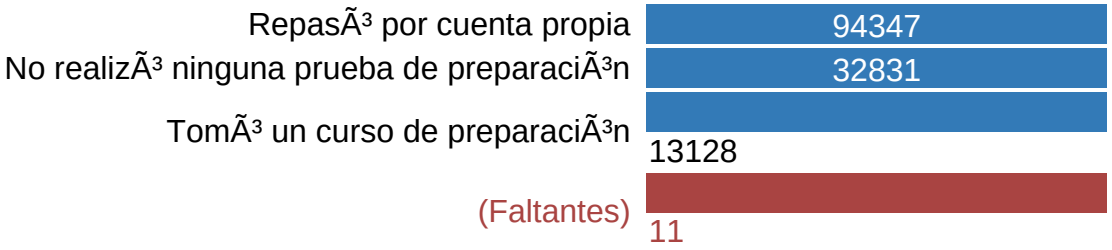
Faltantes (%)	42.6%
Faltantes (n)	59751



98-G_estu_comocapacitoexamensb11

Categórica

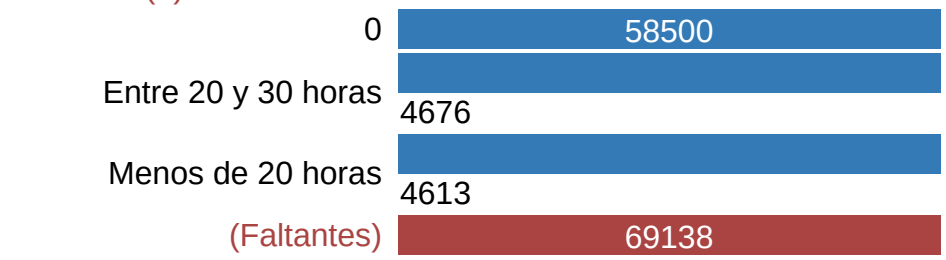
Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	11



99-G_estu_cursodocentesies

Categórica

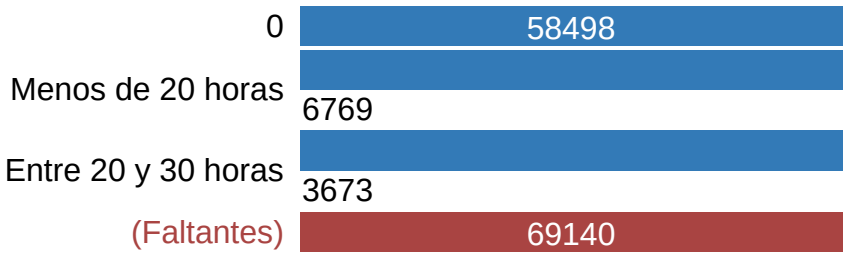
Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	49.3%
Faltantes (n)	69138



100-G_estu_cursoiesapoyoexterno

Categórica

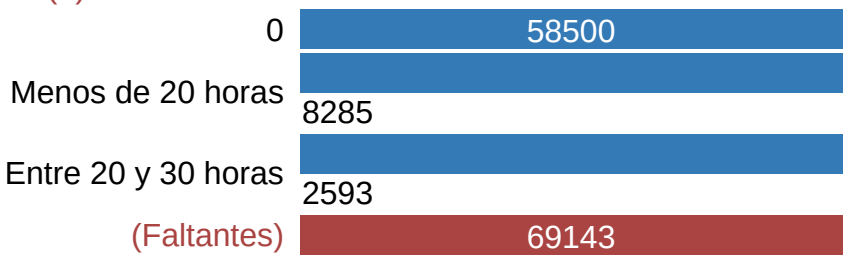
Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	49.3%
Faltantes (n)	69140



101-G_estu_cursoiesexterna

Categórica

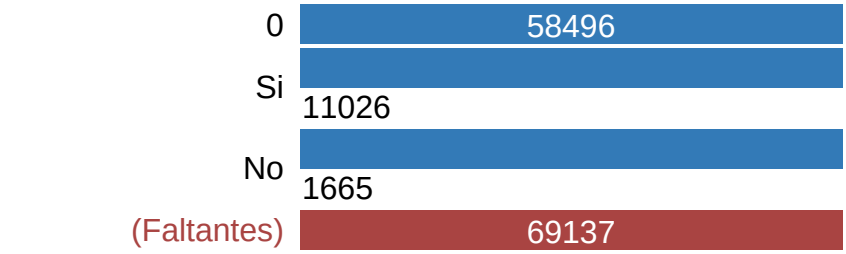
Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	49.3%
Faltantes (n)	69143



102-G_estu_simulacrotipoicfes

Categórica

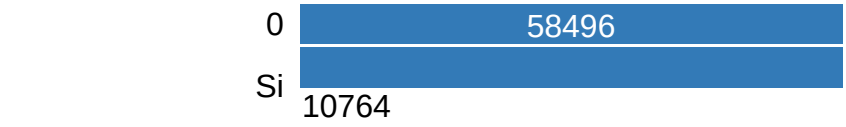
Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	49.3%
Faltantes (n)	69137

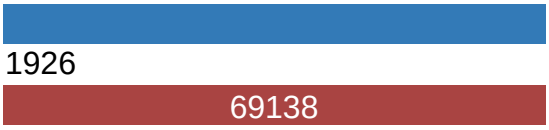


103-G_estu_actividadrefuerzoareas

Categórica

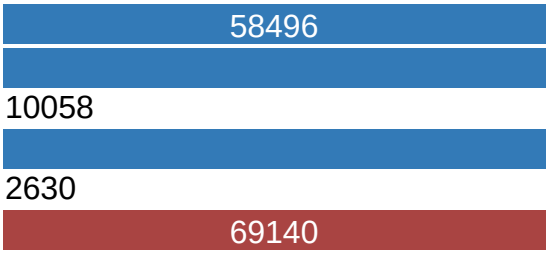
Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	49.3%
Faltantes (n)	69138





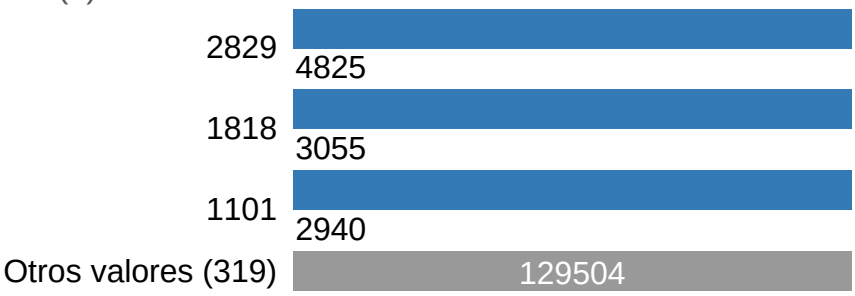
104-G_estu_actividadrefuerzogeneric
Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	49.3%
Faltantes (n)	69140



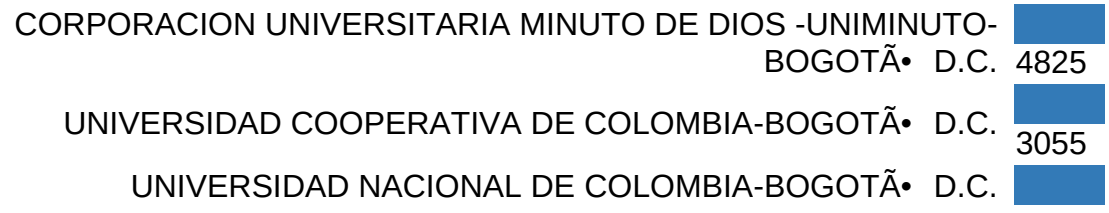
105-G_inst_cod_institucion
Categórica

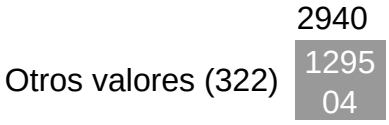
Valores distintos	322
Únicos (%)	0.2%
Faltantes (%)	0.0%
Faltantes (n)	0



106-G_inst_nombre_institucion
Categórica

Valores distintos	325
Únicos (%)	0.2%
Faltantes (%)	0.0%
Faltantes (n)	0





107-G_estu_semestrecursa

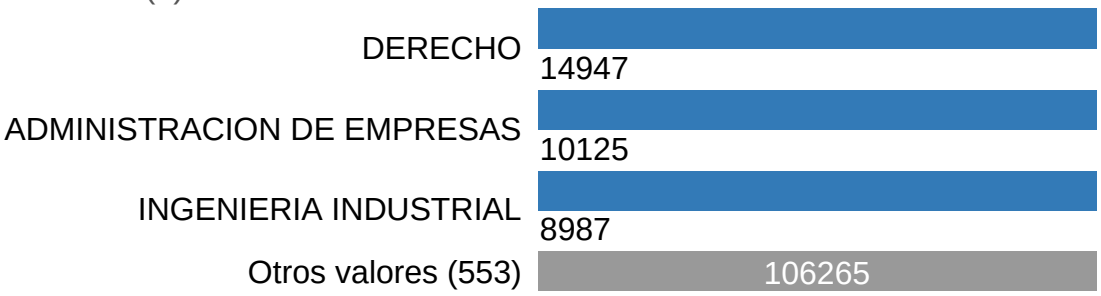
Categórica

Valores distintos	24
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	72

108-G_estu_prgm_academico

Categórica

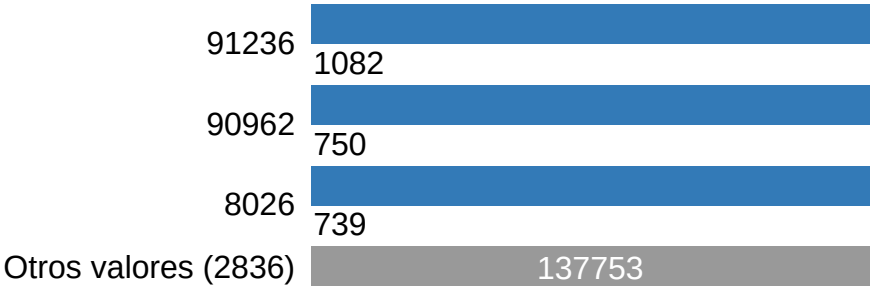
Valores distintos	556
Únicos (%)	0.4%
Faltantes (%)	0.0%
Faltantes (n)	0



109-G_estu_snies_prgmacademico

Categórica

Valores distintos	2839
Únicos (%)	2.0%
Faltantes (%)	0.0%
Faltantes (n)	0



110-G_gruporeferencia

Categórica

Valores distintos	17
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0

111-G_gruporeferencia

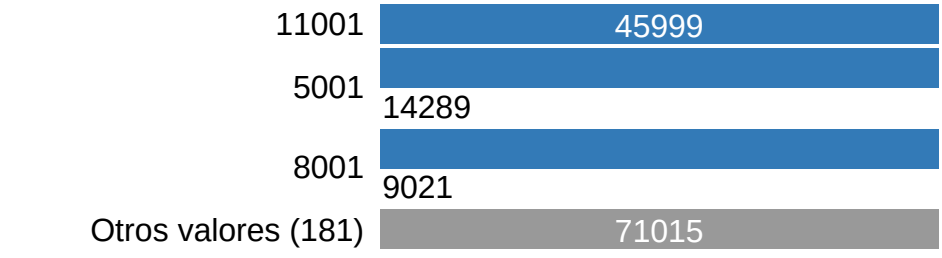
Categórica

Valores distintos	17
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0

112-G_estu_prgm_codmunicipio

Categórica

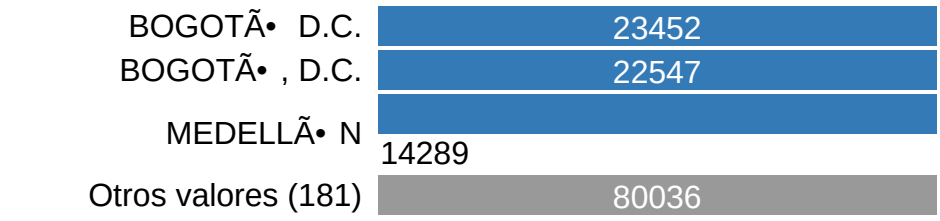
Valores distintos	184
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0



113-G_estu_prgm_municipio

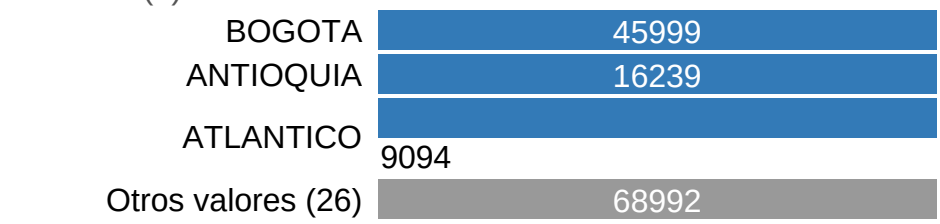
Categórica

Valores distintos	184
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0



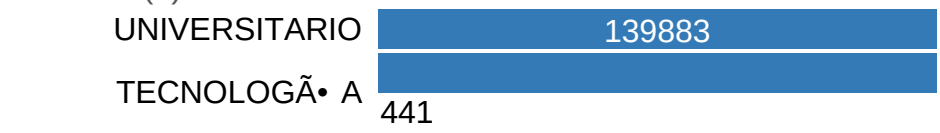
114-G_estu_prgm_departamento
Categórica

Valores distintos	29
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



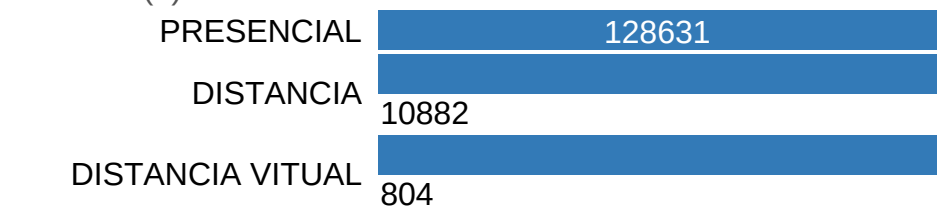
115-G_estu_nivel_prgm_academico
Categórica

Valores distintos	2
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



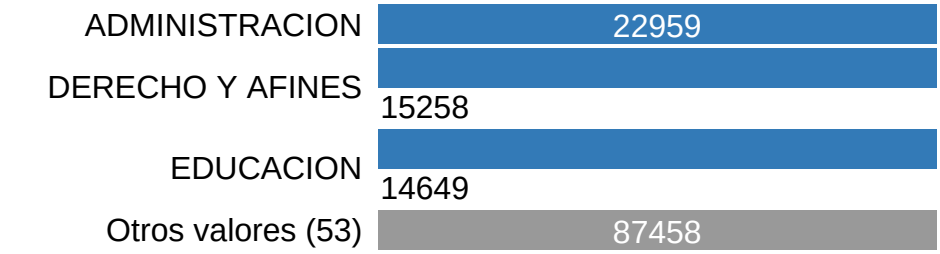
116-G_estu_metodo_prgm
Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



117-G_estu_nucleo_pregrado
Categórica

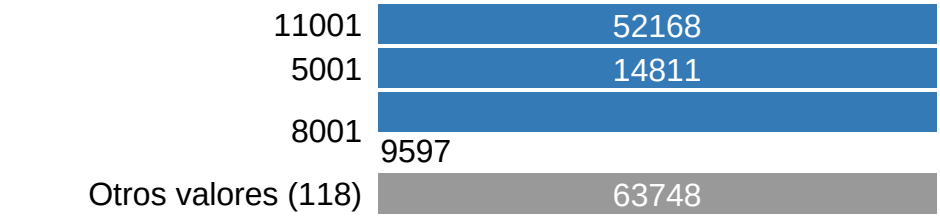
Valores distintos	56
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



118-G_estu_inst_codmunicipio

Categórica

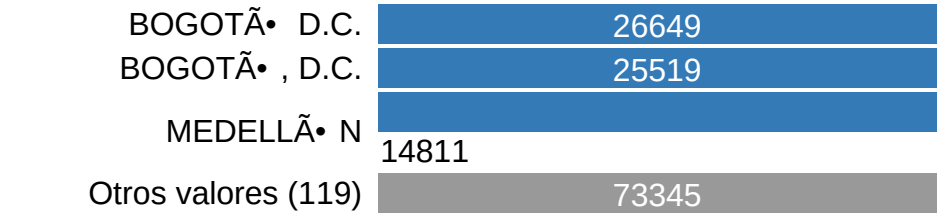
Valores distintos	121
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0



119-G_estu_inst_municipio

Categórica

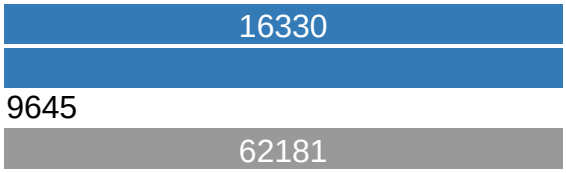
Valores distintos	122
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0



120-G_estu_inst_departamento

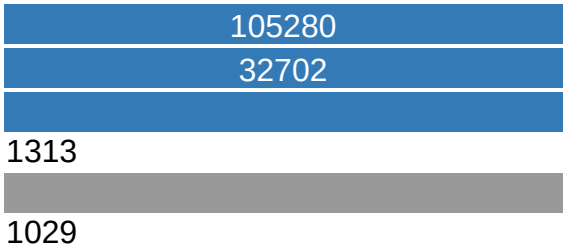
Categórica

Valores distintos	28
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



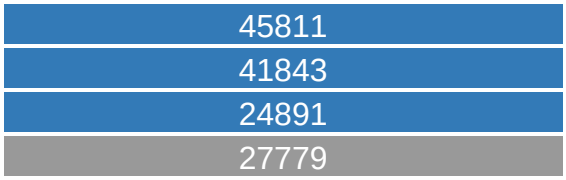
121-G_inst_caracter_academico
Categórica

Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



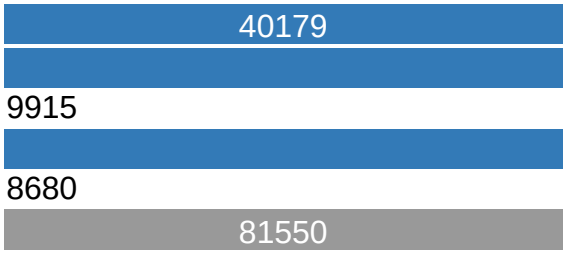
122-G_inst_origen
Categórica

Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



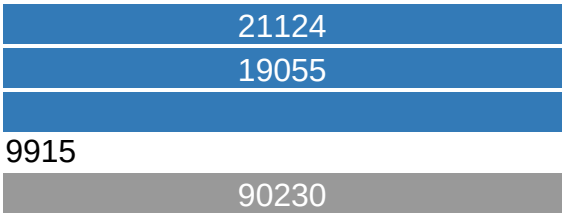
123-G_estu_cod_mcpio_presentacion
Categórica

Valores distintos	113
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0



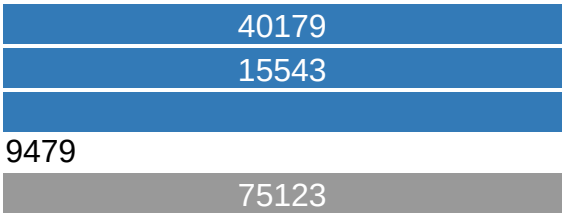
124-G_estu_mcpio_presentacion
Categórica

Valores distintos	114
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0



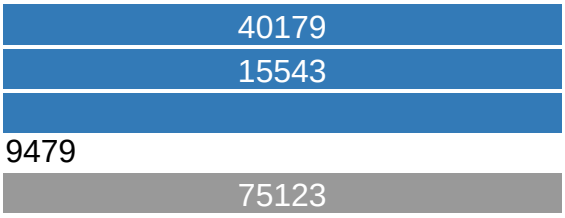
125-G_estu_depto_presentacion
Categórica

Valores distintos	33
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



126-G_estu_cod_depto_presentacion
Categórica

Valores distintos	33
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



127-G_razona_cuant_punt
Numérica

Valores distintos	181
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0

Infinite (%)	0.0%
Infinite (n)	0
Mean	158.25
Minimum	0
Maximum	300
Zeros (%)	0.0%

~~128-G_razona_cuant_desem~~
Highly correlated
This variable is highly correlated with 127-G_razona_cuant_punt and should be ignored for analysis

Correlation	0.91938

~~129-G_razona_cuant_pnal~~
Highly correlated
This variable is highly correlated with 128-G_razona_cuant_desem and should be ignored for analysis

Correlation	0.93122

130-G_razona_cuant_pgref
Numérica

Valores distintos	100
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	57.105
Minimum	1
Maximum	100
Zeros (%)	0.0%

131-G_lect_critica_punt
Numérica

Valores distintos	190
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	156.98
Minimum	0
Maximum	300

Zeros (%)	0.0%

~~132-G_lectura_critica_desem~~
Highly correlated
This variable is highly correlated with 131-G_lect_critica_punt and should be ignored for analysis
Correlation 0.93464
~~133-G_lect_critica_pnal~~
Highly correlated
This variable is highly correlated with 132-G_lectura_critica_desem and should be ignored for analysis
Correlation 0.93602
~~134-G_lect_critica_pgref~~
Highly correlated
This variable is highly correlated with 133-G_lect_critica_pnal and should be ignored for analysis
Correlation 0.95989
135-G_compet_ciudad_punt
Numérica

Valores distintos	178
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	152.23
Minimum	0
Maximum	300
Zeros (%)	0.0%

~~136-G_compet_ciudad_desem~~
Highly correlated
This variable is highly correlated with 135-G_compet_ciudad_punt and should be ignored for analysis
Correlation 0.92888
~~137-G_compet_ciudad_pnal~~
Highly correlated
This variable is highly correlated with 136-G_compet_ciudad_desem and should be ignored for analysis
Correlation 0.93182
~~138-G_compet_ciudad_pgref~~
Highly correlated

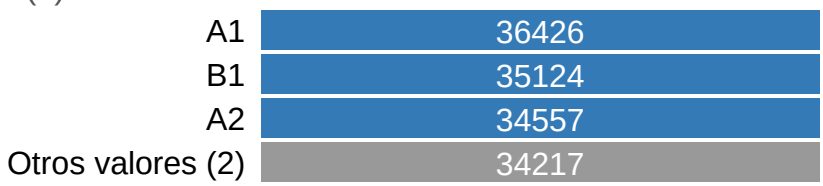
This variable is highly correlated with 137-G_compet_ciudad_pnal and should be ignored for analysis
Correlation 0.96363

139-G_ingles_punt
Numérica

Valores distintos	163
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	158.27
Minimum	0
Maximum	300
Zeros (%)	0.1%

140-G_ingles_desem
Categórica

Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



~~141-G_ingles_pnal~~
Highly correlated
This variable is highly correlated with 139-G_ingles_punt and should be ignored for analysis
Correlation 0.96073

~~142-G_ingles_pgref~~
Highly correlated
This variable is highly correlated with 141-G_ingles_pnal and should be ignored for analysis
Correlation 0.9487

143-G_com_escrita_punt
Numérica

Valores distintos	211
Únicos (%)	0.2%
Faltantes (%)	0.2%

Faltantes (n)	225
Infinite (%)	0.0%
Infinite (n)	0
Mean	151.47
Minimum	0
Maximum	300
Zeros (%)	1.6%

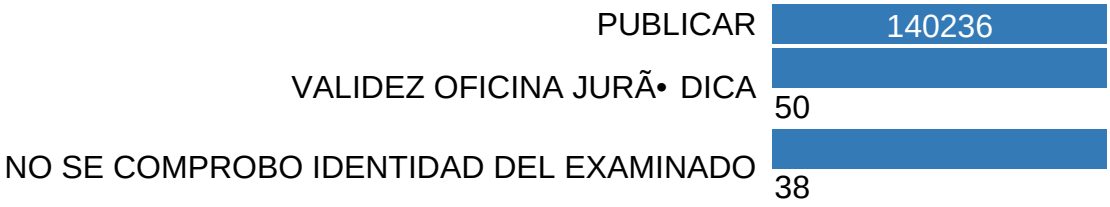
~~144-G_com_escrita_desem~~
Highly correlated
This variable is highly correlated with 143-G_com_escrita_punt and should be ignored for analysis
Correlation 0.93129
~~145-G_com_escrita_pnal~~
Highly correlated
This variable is highly correlated with 144-G_com_escrita_desem and should be ignored for analysis
Correlation 0.92739
~~146-G_com_escrita_pgref~~
Highly correlated
This variable is highly correlated with 145-G_com_escrita_pnal and should be ignored for analysis
Correlation 0.99032
147-G_punt_global
Numérica

Valores distintos	191
Únicos (%)	0.1%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	155.39
Minimum	0
Maximum	255
Zeros (%)	0.0%

~~148-G_percentil_global~~
Highly correlated
This variable is highly correlated with 147-G_punt_global and should be ignored for analysis
Correlation 0.97408

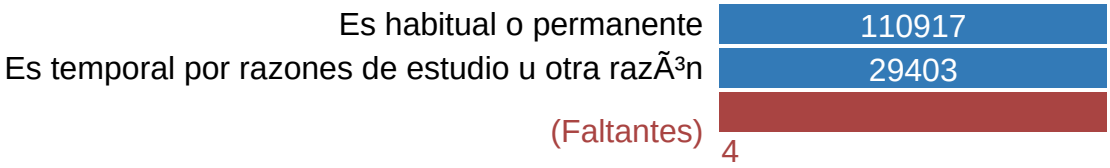
149-G_estu_estadoinvestigacion
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



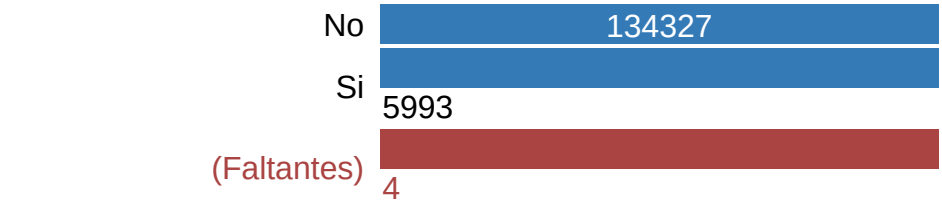
150-G_fami_hogaractual
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



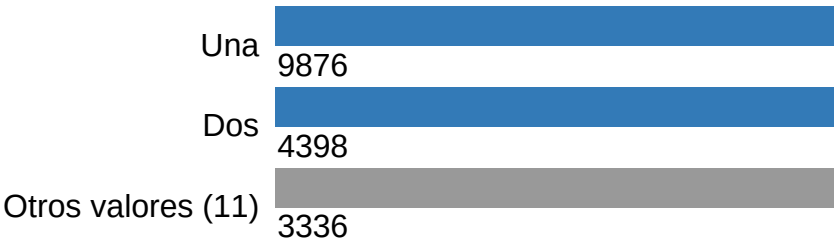
151-G_fami_cabezafamilia
Categórica

Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



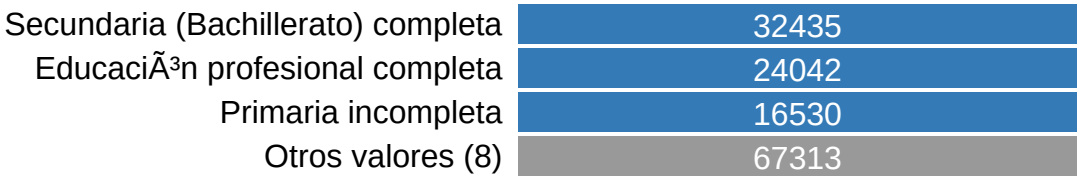
152-G_fami_numpersonasacargo
Categórica

Valores distintos	15
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



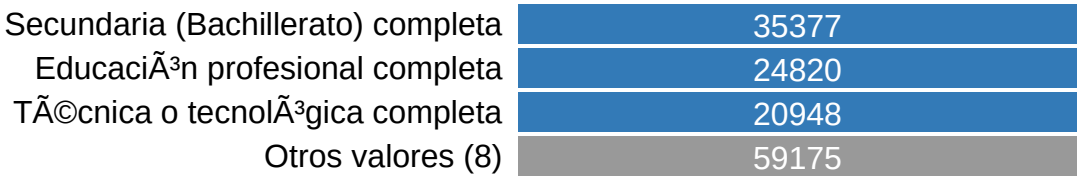
153-G_fami_educacionpadre
Categórica

Valores distintos	12
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



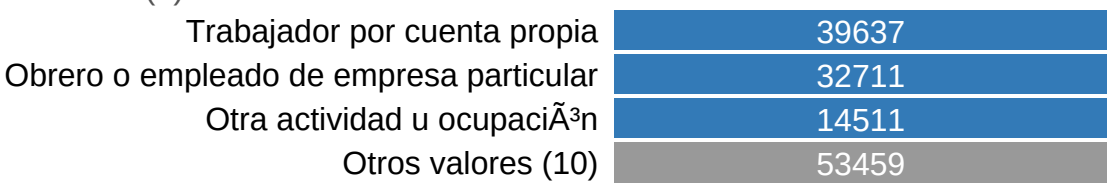
154-G_fami_educacionmadre
Categórica

Valores distintos	12
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



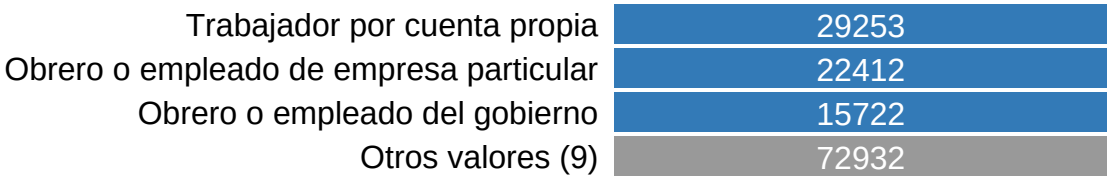
155-G_fami_ocupacionpadre
Categórica

Valores distintos	14
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	6



156-G_fami_ocupacionmadre
Categórica

Valores distintos	13
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	5



157-G_fami_estratovivienda

Categórica

Valores distintos	8
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4

158-G_fami_personashogar

Categórica

Valores distintos	14
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4

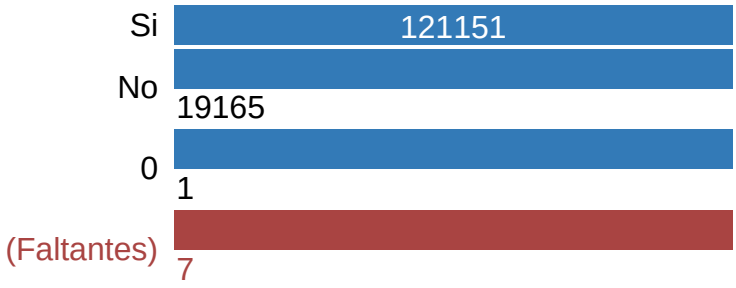
159-G_fami_cuartoshogar

Categórica

Valores distintos	13
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4

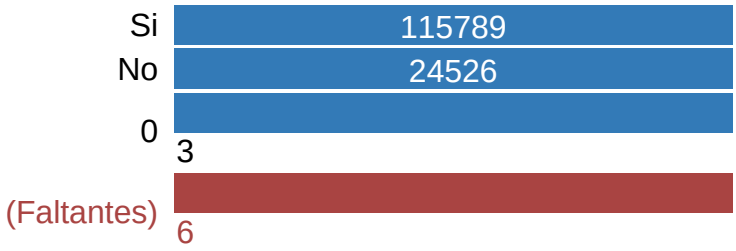
160-G_fami_tieneinternet
Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	7



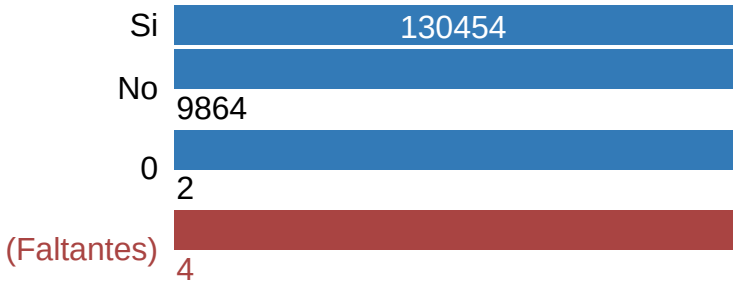
161-G_fami_tieneserviciotv
Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	6



162-G_fami_tienecomputador
Categórica

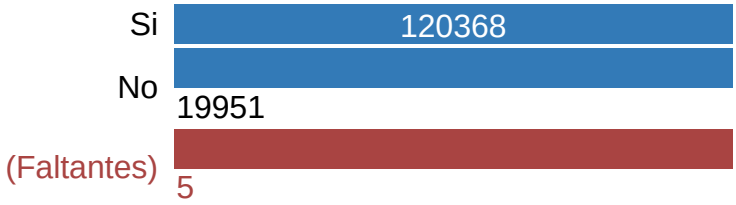
Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



163-G_fami_tienelavadora

Categórica

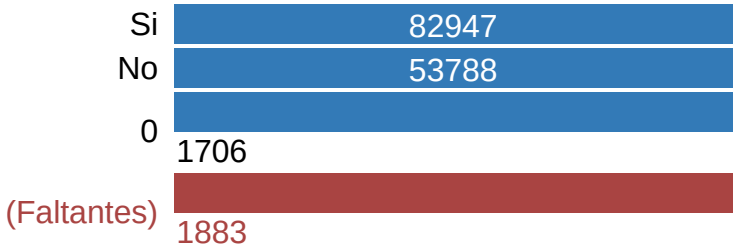
Valores distintos	3
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	5



164-G_fami_tienehornomicroogas

Categórica

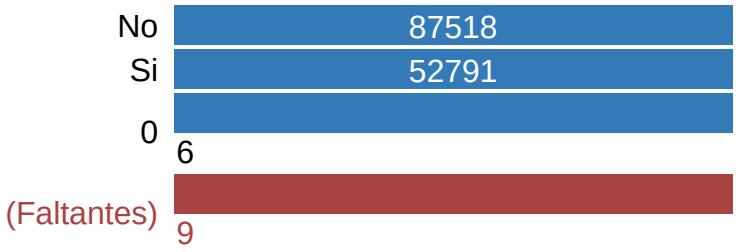
Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	1.3%
Faltantes (n)	1883



165-G_fami_tieneautomovil

Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	9

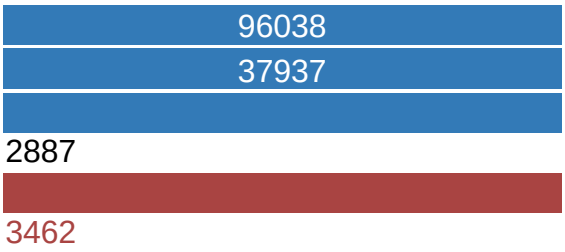


166-G_fami_tienemotocicleta

Categórica

Valores distintos	4
Únicos (%)	0.0%

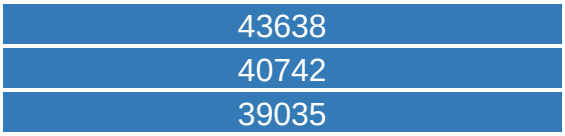
Faltantes (%)	2.5%
Faltantes (n)	3462



167-G_fami_numlibros

Categórica

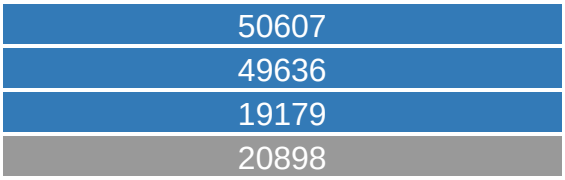
Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



168-G_estu_dedicacionlecturadiaria

Categórica

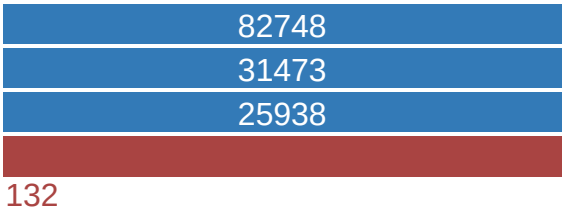
Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4



169-G_estu_dedicacioninternet

Categórica

Valores distintos	5
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	132

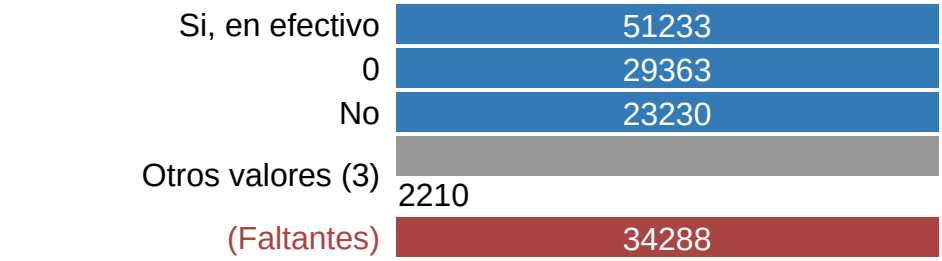


170-G_estu_horassemanatrabaja
Categórica

Valores distintos	6
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	4

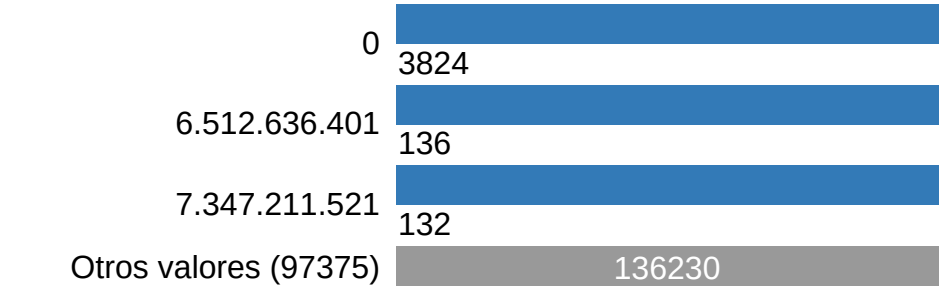
171-G_estu_tiporemuneracion
Categórica

Valores distintos	7
Únicos (%)	0.0%
Faltantes (%)	24.4%
Faltantes (n)	34288



172-G_estu_inse_individual
Categórica

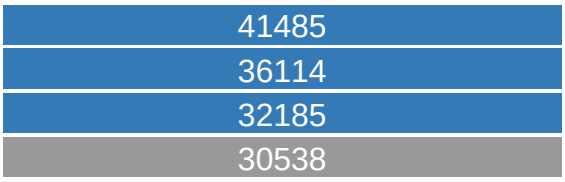
Valores distintos	97379
Únicos (%)	69.4%
Faltantes (%)	0.0%
Faltantes (n)	2



173-G_estu_nse_individual
Categórica

Valores distintos	6
Únicos (%)	0.0%

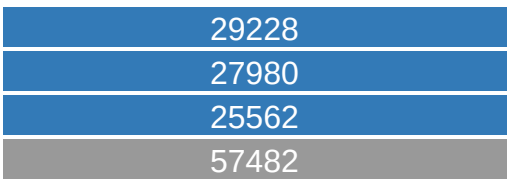
Faltantes (%)	0.0%
Faltantes (n)	2



174-G_estu_valormatriculauniversidad

Categórica

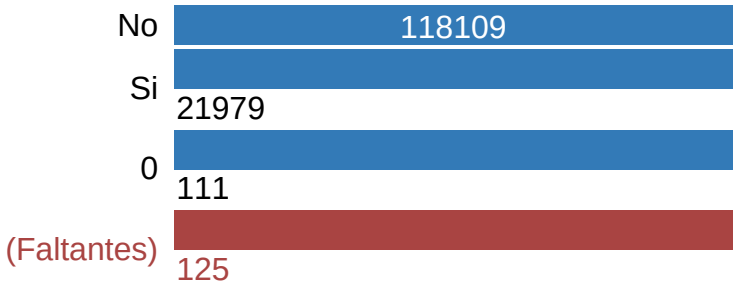
Valores distintos	11
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	72



175-G_estu_pagomatriculabeca

Categórica

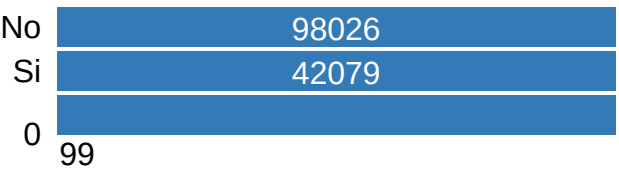
Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	125



176-G_estu_pagomatriculacredito

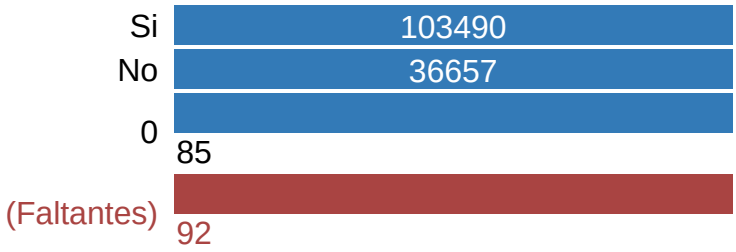
Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	120



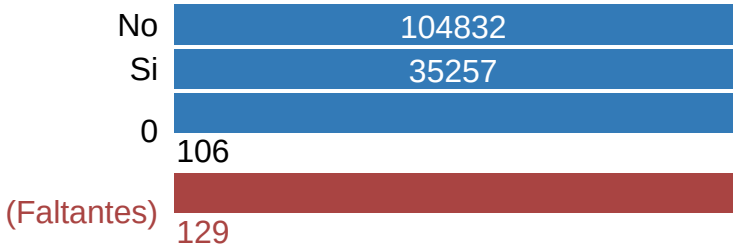
177-G_estu_pagomatriculapadres
Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	92



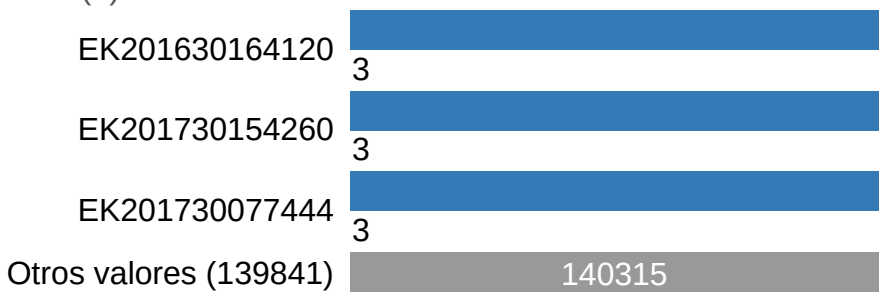
178-G_estu_pagomatriculapropio
Categórica

Valores distintos	4
Únicos (%)	0.0%
Faltantes (%)	0.1%
Faltantes (n)	129



179-e_estu_consecutivo
Categórica

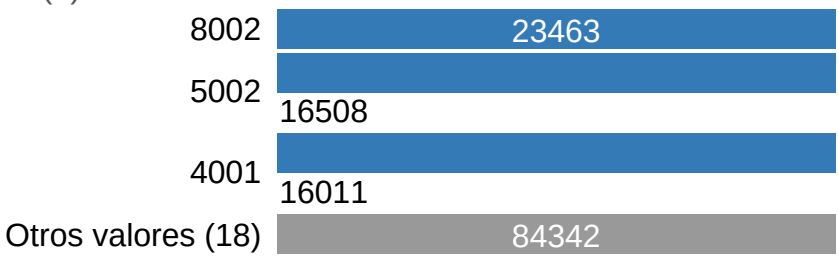
Valores distintos	139844
Únicos (%)	99.7%
Faltantes (%)	0.0%
Faltantes (n)	0



180-e_result_codigoprueba

Categórica

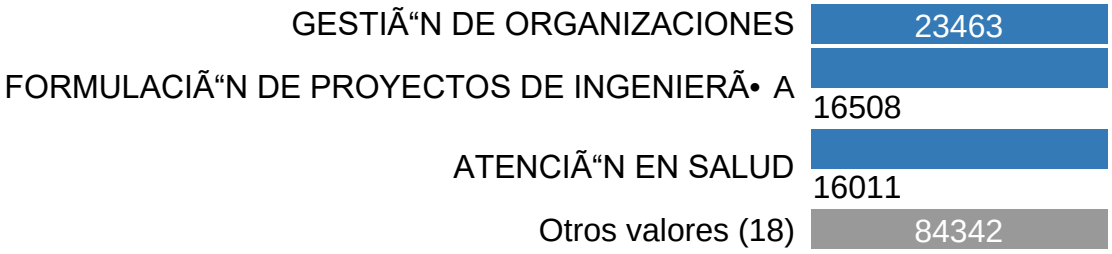
Valores distintos	21
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



181-e_result_nombreprueba

Categórica

Valores distintos	21
Únicos (%)	0.0%
Faltantes (%)	0.0%
Faltantes (n)	0



182-e_result_puntaje

Numérica

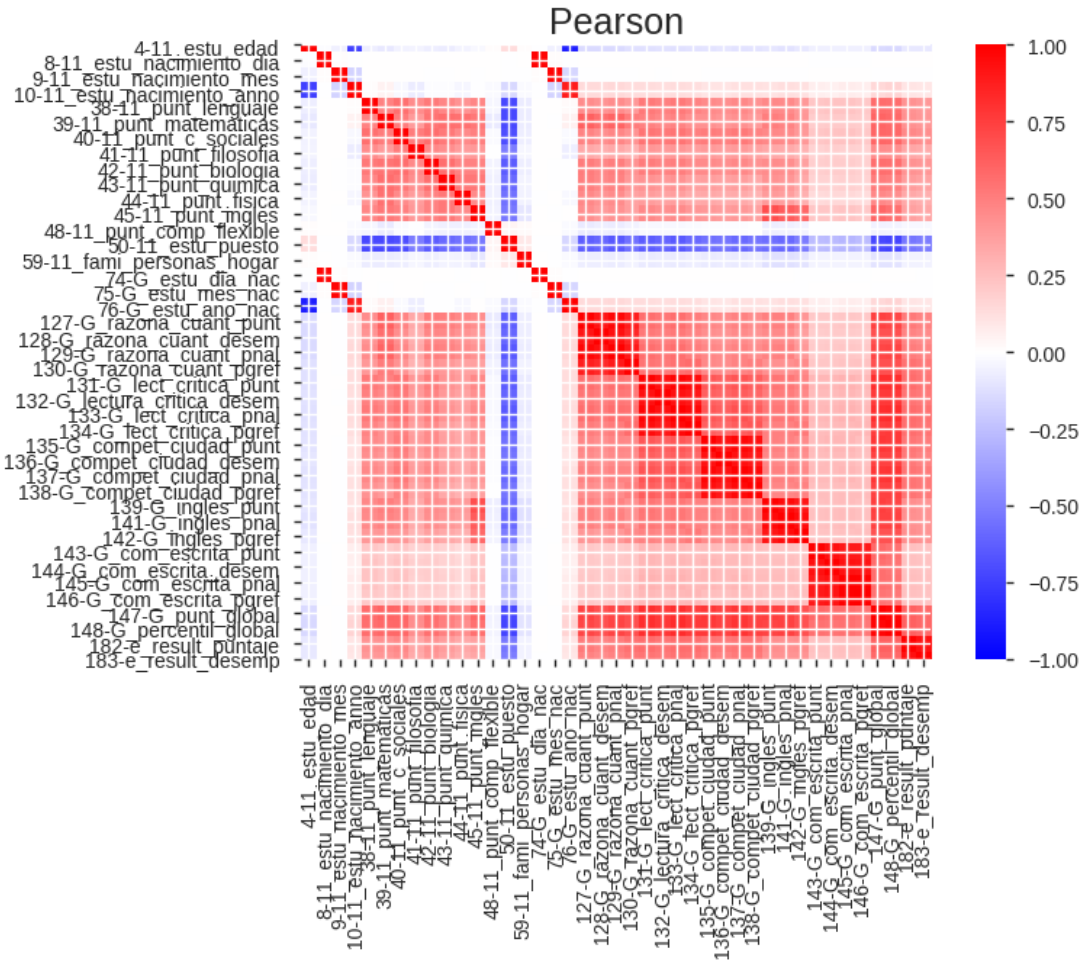
Valores distintos	221
Únicos (%)	0.2%
Faltantes (%)	0.0%
Faltantes (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	156.02
Minimum	0
Maximum	300
Zeros (%)	0.0%

~~183-e_result_desemp~~

Highly correlated

This variable is highly correlated with 182-e_result_puntaje and should be ignored for analysis
Correlation 0.93529

Correlations



Correlations