Aplicación de técnicas Data Mining para el análisis del desempeño escolar en Cundinamarca (Colombia) 2015 a 2019

Lilian Daniela Suárez Riveros

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería Industrial
Decanatura de Ingeniería de Sistemas
Decanatura de Matemáticas
Maestría en Ciencia de Datos
Bogotá D.C., Colombia
2022

Aplicación de técnicas Data Mining para el análisis del desempeño escolar en Cundinamarca (Colombia) 2015 a 2019.

Lilian Daniela Suárez Riveros

Trabajo de grado para optar al título de
Magíster en Ciencia de Datos

Director
Wilmer Pineda Ríos
Magister en Ciencias Matemáticas Codirector

Codiretor
Iván Mauricio Mendivelso Ramírez
Magister en Antropología

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería Industrial
Decanatura de Ingeniería de Sistemas
Decanatura de Matemáticas
Maestría en Ciencia de Datos
Bogotá D.C., Colombia
2022

Publicado en 2022 por la Escuela Colombiana de Ingeniería Julio Garavito. Avenida 13 No 205-59 Bogotá. Colombia
TEL: +57 – 1 668 36 00

Reconocimiento o Agradecimientos

A la Escuela Colombiana de Ingeniería Julio Garavito, por apoyar mi formación profesional y personal; por permitirme ser parte de la institución como estudiante de la Maestría en Ciencia de Datos y como Monitora graduada de investigación, punto importante que me permitió completar con éxito mis estudios y crecer profesionalmente.

A la Infraestructura de Datos Espaciales y Estadísticos, de la Secretaría de Planeación de la gobernación de Cundinamarca, por facilitar los datos organizándolos acorde a los requerimientos de la presente investigación. En ese mismo sentido, es preciso reconocer a todas las entidades que tienen datos abiertos para ser utilizados en este tipo de estudios como es el caso del Instituto Colombiano para la Evaluación de la Educación, la Fiscalía General de la Nación, entre otras entidades.

A mis tutores, MSc. Wilmer Pineda Ríos e MSc. Iván Mauricio Meldivelso Ramírez, por confiar en mis capacidades; por su guía, aportes y entrega con el desarrollo de esta investigación; por su disponibilidad, apoyo y paciencia durante todo el proceso.

A la Ing. MSc. Sonia Jaimes por su gran apoyo desde la dirección del programa en el ámbito académico, pero especialmente, por incentivar mi crecimiento como personal y profesional, por su confianza durante estos años que compartimos.

A mis padres por siempre estar presentes y motivarme a crecer académicamente, sin dejar de lado mi crecimiento como persona integra en la vida profesional; por mostrarme el valor del trabajo duro y el amor. A mis hermanos por darme su alegría y el cariño, por su compañía y ser parte de mi motivación diaria.

A mis amigos que hicieron parte del proceso, por su interés en el desarrollo general del trabajo, avances y dificultades; por sus ánimos, gestos y palabras para seguir adelante.

Resumen

El objetivo de la investigación es identificar las variables que inciden en el desempeño escolar de las pruebas de Saber 11, en estudiantes de educación media, en Cundinamarca, durante los años 2015 a 2019, apoyado con modelamiento estadístico, usando fuentes de datos abiertas y disponibles del Instituto Colombiano para la Evaluación de la Educación ICFES y de la gobernación de Cundinamarca - Colombia. El desempeño escolar, en la presente investigación, se asume como la puntuación que los estudiantes obtienen al presentar las pruebas estandarizadas, que para el caso en estudio, corresponde con las aplicadas por el estado colombiano a través del ICFES, conocidas como pruebas Saber 11. La metodología empleada fue Cross Industry Standard Process for Data Mining (CRISP_DM). Los resultados mostraron que el modelo multinivel tiene una mayor simplicidad en su organización, tiempo de procesamiento y demás características para la información objeto de este estudio. Adicionalmente, se observó que las variables que tienen incidencia en la predicción del desempeño escolar son el género de estudiante, las horas del trabajo del estudiante, el apoyo del estado a la familia, si los padres tienen estudios terminados, si la madre trabaja, así como los recursos disponibles en la casa. En algunos municipios emergió como variable la obesidad del estudiantado como una variable opuesta al desempeño escolar.

Abstract

The objective of the research is to identify the variables that affect the school performance of Saber 11 tests, in middle school students, in Cundinamarca, during the years 2015 to 2019, supported by statistical modeling, using open and available data sources from the Colombian Institute for the Evaluation of Education ICFES and the government of Cundinamarca - Colombia. School performance, in the present investigation, is assumed as the score that students obtain when presenting standardized tests, which for the case under study, corresponds to those applied by the Colombian state through ICFES, known as Saber 11 tests. The methodology used was the Cross Industry Standard Process for Data Mining (CRISP_DM). The results showed that the multilevel model has greater simplicity in its organization, processing time and other characteristics for the information object of this study. Additionally, it was observed that the variables that have an impact on predicting school performance are the gender of the student, the student's work hours, state support for the family, if the parents have completed studies, if the mother works, as well as the resources available in the house. In some municipalities, student obesity emerged as a variable opposed to school performance.

Tabla de contenido

Lista de Figuras

Lista de Tablas

1 La investigación

1.1 Introducción

Con el fin de cerrar las brechas sociales, mejorar los índices de salud, de la calidad de la educación (CE), entre otras, varios países han incorporado, a través de sus lineamientos, políticas que procuran una mejor calidad de vida para su población. De hecho, hay evidencia que indica la existencia de una relación entre la salud y la educación, como condiciones necesarias y estructurales para el desarrollo de las condiciones sociales, económicas y culturales de la población de las naciones (Comisión Económica para América Latina y el Caribe, 2016).

En este sentido, enfocarse en la calidad educativa, necesariamente tiene que involucrar áreas con las cuales ésta se relaciona. Ahora bien, en las referencias consultadas para la presente investigación, no se identifica un consenso global en relación con el significado de la calidad educativa, sin embargo, entidades a nivel nacional e internacional, elaboran procesos de certificación de calidad de instituciones de educación, a través de sistemas de aseguramiento de la calidad, los cuales pretenden asegurar los procesos y procedimientos necesarios, que den cuenta de la calidad de la educación ofertada por la institución certificada.

Para el Ministerio de Educación Nacional, en Colombia, la calidad es entendida como:

> "… *el conjunto de atributos articulados, interdependientes, dinámicos, construidos por la comunidad académica como referentes y que responden a las demandas sociales, culturales y ambientales. Dichos atributos permiten hacer valoraciones internas y externas a las instituciones, con el fin de promover su transformación y el desarrollo permanente de sus labores formativas, académicas, docentes, científicas, culturales y de extensión*" (Ministerio de Educación Nacional de Colombia, 2019).

De otro lado, Reyes (2020), considera que el tema de la calidad educativa, hace referencia a la prestación del servicio con estándares de eficacia y eficiencia, como elementos adicionales a la existencia de saberes y competencias de las diversas comunidades sociales, que se entretejen en una sociedad; en ese mismo sentido, las personas que la integran, no solamente debe hacer parte de las prácticas sociales de la comunidad, sino que deben avanzar con el colectivo social, en las transformaciones que el contexto demanda de la sociedad, en las cuales, la educación debe constituirse en un eje fundamental (Lorna et al., 2014).

Los sistemas de aseguramiento de la calidad de la educación, se fundamentan en procesos y procedimientos que contienen indicadores, resultado de procesos de evaluación, desarrollados por la misma institución, y que tienen como fin la autorregulación, en pro del mejoramiento de ésta en diferentes dimensiones y niveles. En Colombia, el Ministerio de Educación Nacional (MEN), a través de la Comisión Nacional de Acreditación (CNA), es el encargado de certificar

y acreditar programas, para las instituciones que han hecho las actividades necesarias y suficientes (Peláez-Valencia et al., 2020), para pedir esta distinción en el nivel de la educación superior, así como la evaluación de procesos de autorregulación en los programas que oferta una institución (Ruiz & Moya, 2020), y los procesos de evaluación que se desarrollan en el aula (Montagud-Mascarell & Gandía-Cabedo, 2014).

La CE se ha asociado, en forma limitada, con el resultado de pruebas estandarizadas aplicadas por entidades regionales, nacionales e internacionales. Los resultados de estas pruebas se constituyeron en argumentos, para referirse a la calidad de la educación de la población en la que ha sido aplicada, como es el caso de las diversas expresiones de diferentes sectores en Colombia, con ocasión de los resultados de las pruebas Saber 11 (López-Vera et al., 2020), en la que diferentes voces de opinión buscan culpables, en tanto que otros señalan que *"…La causa de ese fracaso es que los gobiernos de las últimas décadas no han intervenido las variables de las que depende la calidad educativa..."* (Zubiria-Samper, 2022).

Luego, estudiar variables que intervienen de diferente forma, en el proceso de la calidad educativa y que están relacionados con otros sectores de la sociedad, es una necesidad vigente, no sólo para el campo académico, sino también para todo el sector que tiene que ver con toma de decisiones económicas, sociales, políticas entre otras, que afectan la calidad de vida de una sociedad.

1.2 Justificación

Las pruebas estandarizadas se han constituido como un elemento para construir argumentos tendientes a referirse a la CE. A nivel internacional, está el Program International Student Assessment (PISA), promovido por la Organización para la Cooperación y el Desarrollo Económicos (OCDE), cuyo propósito es medir las habilidades en lectura, matemáticas y ciencias, de los jóvenes de los países miembros, cuyo promedio de edad es de 15 años. No obstante, los países tienen sus propias entidades para efecto de medir el desempeño escolar a través de pruebas estandarizadas, en Colombia está el Instituto Colombiano para la Evaluación de la Educación (ICFES), en los Estados Unidos Mexicanos o México, el Plan Nacional para la Evaluación de los Aprendizajes (PLANEA), en Chile, el Sistema Nacional de Evaluación de Resultados de Aprendizaje (SIMCE), entre otros.

Específicamente, las pruebas Saber, en Colombia, miden algunas variables, que potencialmente, podrían aportar o contribuir con información que permita comprender o explicar el logro de aprendizaje, por parte de las personas que participan de las pruebas. Por lo anterior, las entidades territoriales en Colombia, retoman los informes y resultados propios de la aplicación de las pruebas, como un elemento más, necesario para la construcción de propuestas y estrategias en pro de mejorar la calidad de la educación a nivel local, regional o nacional.

De otro lado, partiendo del supuesto que el desempeño académico obtenido por los estudiantes en las pruebas estandarizadas, hace parte de la calidad de la educación en la región, con relación

a los países en las cuales se aplican y se tienen resultados, se precisa partir de estos datos, para procesarlos y construir la información, de alguna manera objetiva, que contribuya a comprender y explicar las variables que dan cuenta de la obtención del desempeño académico.

Por ejemplo, en el departamento de Cundinamarca, las tendencias de crecimiento demográfico son superiores respecto a todos los departamentos del país, de manera que, en los municipios cercanos al centro del departamento, la tasa de crecimiento promedio anual simple es del 4,5%. Además, para el 2018, estaba compuesto por un extenso territorio rural, concentrado en un 72% en los municipios y el 28% en los cascos urbanos. Algunos indicadores relacionados con la pobreza en Cundinamarca, muestran el 5,6% de analfabetismo, el 69,2% de trabajo informal, el 0,9% de paredes exteriores de viviendas con materiales inadecuados y 48,5% de bajo logro educativo (Gobernación de Cundinamarca, 2020).

Para trazar idearios y rutas políticas, con sustento financiero, a nivel del departamento que propendan por una mejor calidad de vida de la población, se requiere identificar variables y construir argumentos que contribuyan a identificar, no solamente las necesidades de la población, sino todas aquellas otras que subyace a la transformación deseada y necesaria de una sociedad.

Por lo anterior, esta investigación pretende contribuir con las entidades territoriales de Cundinamarca, con información objetiva, derivada del procesamiento de los datos provenientes de diferentes fuentes, aplicando diversas técnicas de modelamiento estadístico, en el contexto de la ciencia datos.

En este sentido, es preciso indicar que la dupla salud-educación, es el pilar estructural para el desarrollo de una nación. Diversos estudios han mostrado que pueden incidir en las condiciones sociales, económicas y culturales y en general, en la calidad de vida de sus pobladores. De hecho, para la Comisión Económica para América Latina y el Caribe, el incremento positivo en los índices educativos, representa a su vez un incremento positivo en los índices de salud, productividad y movilidad social, así como en la reducción de la pobreza y la construcción de una ciudadanía (2016). De igual manera, Reimers (2000), lo manifestó en el sentido de que los ingresos del hogar, el estrato socioeconómico y algunas otras variables de orden social, están relacionadas con la calidad educativa.

1.3 Planteamiento del problema

Como se ha advertido en líneas anteriores, el estudio de la CE reviste importancia al momento de trazar políticas de desarrollo nacionales. Asimismo, como se ha descrito, el desempeño académico de las pruebas Saber 11 es un indicador que hace parte de la percepción general relacionada con la calidad de la educación media.

La CE ofertada por las diferentes instituciones, se ve impactada por los resultados de los estudiantes que presentan la prueba Saber 11. Lo anterior, porque el ICFES establece una

categorización de las instituciones educativas, con base en los resultados en las pruebas Saber 11, para efecto de catalogarlas y diferenciarlas. Las categorías son: A+ (IG > 0,77), A ($0.72 < IG \leq 0.77$), B ($0.67 < IG \leq 0.72$), C ($0.62 < IG \leq 0.67$) y de D ($0 \leq IG \leq 0.62$) (ICFES, 2017). La categorización influye, entre otras formas, en la percepción que las instituciones educativas ofrecen a la población que acude a ellas, para beneficiarse de su oferta formativa.

Como resultado de la clasificación anterior, por ejemplo, Sapiens Research Group (SRG) publica los mejores colegios en Cundinamarca, acorde a criterios de categoría, calidad y acreditación. Según SRG, en este departamento, de los 1075, clasificaron 126 colegios con un índice superior a 0,78, en alguna de las diez categorías más altas (2021). Los colegios que están clasificados, en el ranking mencionado, están ubicados principalmente en los municipios cercanos a la ciudad de Bogotá (Chía, la Calera, Sopó, Tenjo y Zipaquirá). En la gráfica se logra identificar un crecimiento del año 2013 al año 2019, con un decrecimiento en los años 2020 y 2021, coincidiendo de alguna manera, con el periodo al que la humanidad ha estado sometida a la emergencia sanitaria debido a la COVID-19.

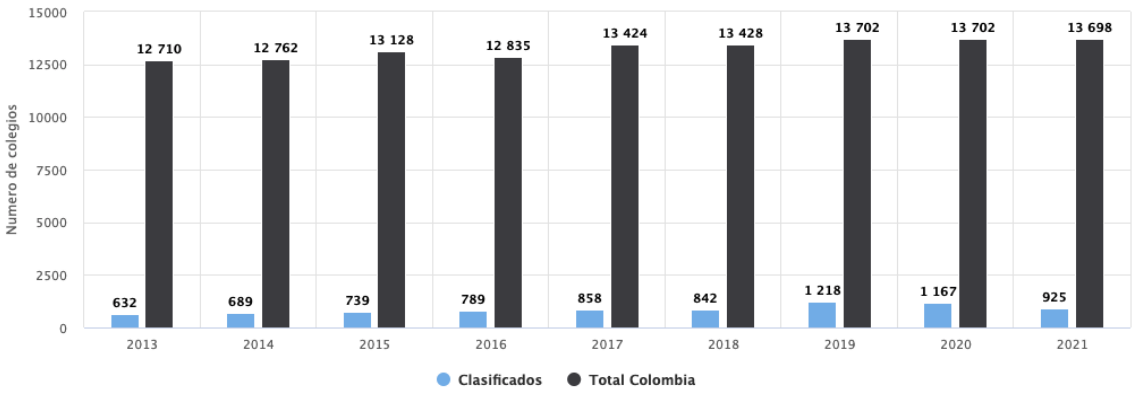


Figura 1. Colegios clasificados según el ranking Col-Sapiens 2021
Fuente. Sapiens Research Group (https://www.srg.com.co/losmejorescolegios-colsapiens)

Diversos estudios han señalado, que el nivel socioeconómico tiene una incidencia en la comprensión de las variables que subyacen al desempeño académico, (Abdul-Aziz et al., 2015; Ali et al., 2013; Benito et al., 2014; Chacón-Vargas & Roldán-Villalobos, 2021; Chaparro-Caso-López et al., 2016; Contreras et al., 2019; Guo et al., 2015; Hanin & Van-Nieuwenhoven, 2016; Lisboa- Bartholo & Da-Costa, 2016; Martínez-Mateus & TurriagoHoyos, 2015; Masci et al., 2018; Maulida & Kariyam, 2017; Rodríguez-Hernández et al., 2021; Salal & Abdullaev, 2020; Shah et al., 2019; Yopasá & Valbuena, 2019; Zhang & Campbell, 2014), excepto en Finlandia (Benito et al., 2014), país que goza de los mejores índices de calidad de vida y bajos índices de pobreza.

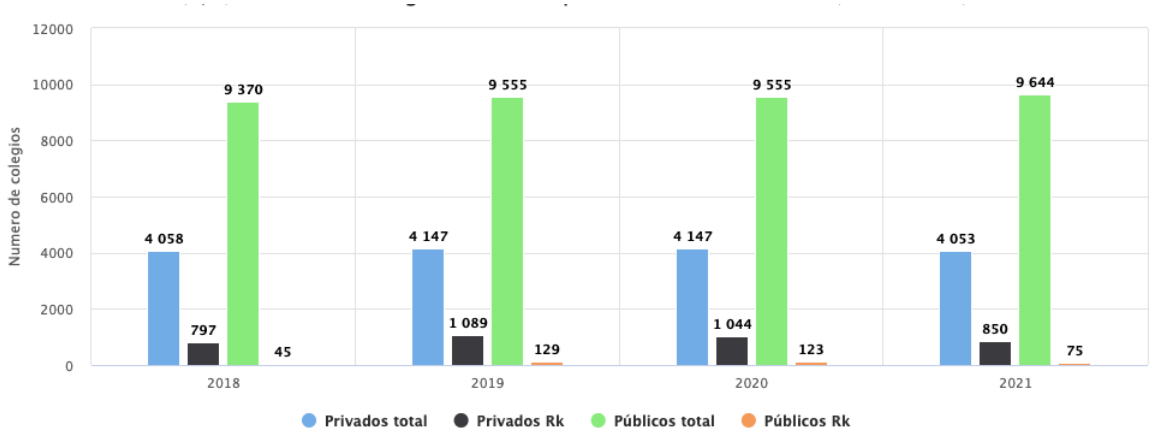


Figura 2. Número de colegios en Colombia por sector versus clasificados (2018-2021)
Fuente. Sapiens Research Group (https://www.srg.com.co/losmejorescolegios-colsapiens)

En la Figura 2 se identifica a nivel nacional, el número de colegios de carácter privado y público, así mismo el número de colegios clasificado en el Rankin de SRG de carácter privado y público. Se observa que, a pesar que el número de colegios públicos es más del doble de los privados, existe una enorme diferencia en cuanto a la clasificación en el ranking SRG, en comparación con colegios de carácter público. Es preciso indicar que, en Colombia, los colegios de carácter privado se sostienen con los ingresos económicos de las familias para la formación de sus jóvenes, a diferencia de los colegios públicos, que son parte del presupuesto nacional. Lo anterior, coincide con estudios, que señalan que el nivel socioeconómico tiene una incidencia en la CE y en consecuencia, en la obtención del desempeño académico.

En Cundinamarca el colegio Rochester, ubicado en el municipio de Chía, corresponde con el de mayor categoría en el departamento. Éste colegio tiene la categoría AAA+.

Con lo anterior, este trabajo pretende abordar como pregunta principal de investigación ¿Qué variables inciden, aplicando técnicas de modelamiento estadístico en información proveniente de fuentes de datos abiertas, en el desempeño escolar en las pruebas Saber 11 en los estudiantes de educación media de Cundinamarca – Colombia entre los años 2015 - 2019?

1.4 Objetivos

En el presente apartado se encuentran el objetivo general y los objetivos específicos sobre los que se orienta la presente propuesta investigación.

1.4.1 Objetivo general

Identificar las variables que inciden en el desempeño escolar de las pruebas de Saber 11, en estudiantes de educación media, en Cundinamarca, durante los años 2015 a 2019, apoyado con modelamiento estadístico, usando fuentes de datos abiertas y disponibles del Instituto Colombiano para la Evaluación de la Educación ICFES y de la gobernación de Cundinamarca - Colombia.

1.4.2 Objetivos específicos

- Depurar las bases de datos disponibles y abiertas, garantizando la integridad de los datos para su posterior procesamiento estadístico.
- Diseñar y construir un diccionario de datos, para proyectar el análisis de la información.
- Analizar diferentes modelos estadísticos, que resulten pertinentes para las bases de datos depuradas, que guarden coherencia con el objetivo de investigación.

1.5 El contexto de la investigación

El departamento de Cundinamarca, es una entidad territorial que hace parte de los 32 departamentos que conforman la República de Colombia. El departamento de Cundinamarca tiene una extensión de 22,605 km², fue creado en 1886, está constituido por 116 municipios distribuidos en 15 provincias, lo anterior no incluye el Bogotá Distrito Capital. Tiene un relieve variado, con alturas que van desde los 300 a 3500 metros sobre el nivel del mar. El departamento limita con otros departamentos como Boyacá, Tolima, Meta, Caldas y Huila. La Figura 3 muestra la división política del departamento, publicada por el Instituto Agustín Codazzi, sin embargo, es necesario precisar que no se incluye lo correspondiente al Distrito Capital de Bogotá (verde alargado en el mapa, con amarillo en la zona urbana), que representa aproximadamente el 2,1% del departamento. Bogotá, además de ser la capital de Colombia, es la capital del departamento.

Figura 3. Mapa político del departamento de Cundinamarca
Fuente. Instituto Geográfico Agustín Codazzi.

En los diferentes municipios del departamento se presentan delitos, suicidios, violencia, delitos sexuales, embarazos en menores, desplazamientos internos debido al conflicto, entre otros, aspectos que perturban la tranquilidad de algunas de las zonas urbanas y rurales. La Figura 4, corresponde con un esquema que da una comparación, en términos de correlación de las variables, a partir de los datos que se obtienen en la Fiscalía General de la Nación, el Instituto Nacional de Medicina Legal y Ciencias Forenses y la Unidad de Víctimas - Red Nacional de Información. Se identifica una alta correlación entre los delitos, cuando hay aspectos relacionados con la violencia y el delito sexual, así como el conflicto armado de la zona, junto con el desplazamiento de las personas, algunas otras correlaciones identifican el desplazamiento con delitos comunes, embarazos en jóvenes y en menor medida, con aspectos relacionados con la violencia. Las demás correlaciones no son concluyentes ni contundentes.

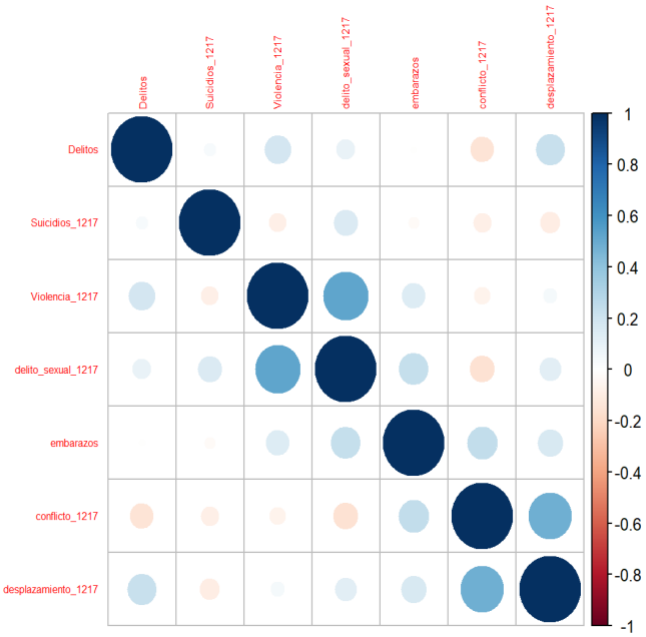
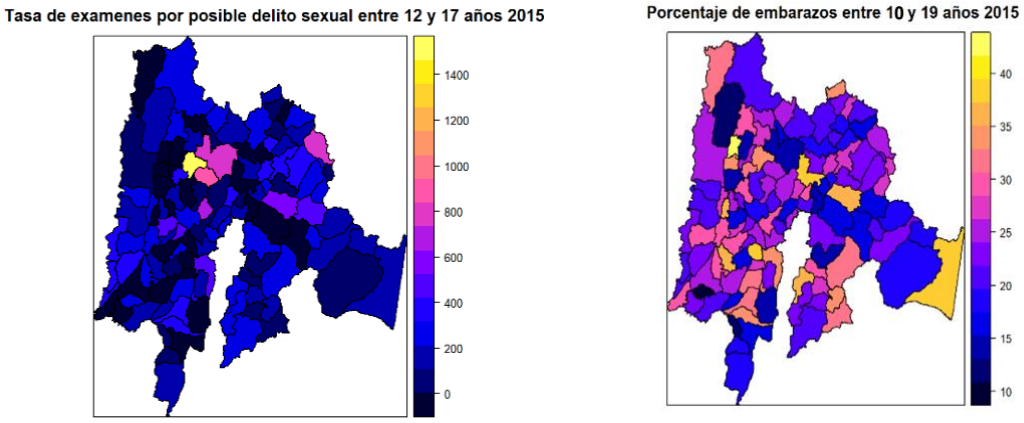


Figura 4. Esquema de correlaciones entre variables presentes en la base de datos de la Fiscalía General de la Nación.
Fuente. Elaboración propia con el software R.



(a) (b)

Figura 5. (a) Mapa de exámenes por delitos sexuales en jóvenes entre 12 y 17 años por municipio y (b) Porcentaje de embarazos en jóvenes, en edad escolar, con edades entre 10 y 19 años.
Fuente. Elaboración propia con el software R.

La Figura 5a, presenta el mapa de Cundinamarca, con una distribución política, en la que se identifica el número de hechos que han sido reportados al Instituto Nacional de Medicina Legal y Ciencias Forenses, en cuánto a la tasa de exámenes por posibles delito sexual, en jóvenes adolescentes cuyas edades están entre 12 y 17 años. La Figura 5b, corresponde con el mapa

político de Cundinamarca, en el que se identifica el porcentaje de embarazos en población cuya edad oscila entre 10 y 19 años.

En las Figura 5a y 5b, a pesar que son del 2015, se identifica un número elevado de posibles delitos sexuales en jóvenes adolescentes en edad escolar y cuyas consecuencias pueden estar afectando, en diferente forma, el acceso a la educación de estos jóvenes, así como la calidad educativa que se oferta hacia ellos y en últimas, el desempeño escolar esperado. El mayor impacto por posibles delitos sexuales en Cundinamarca, está en el municipio de Vergara, ubicado en el noroccidente del departamento, seguido de los municipios vecinos como Pacho y Supatá y al nororiente con Villapinzón.

Adicionalmente, el porcentaje de embarazos en jóvenes adolescentes que están entre 10 y 19 años, tiene un mayor impacto en el municipio de Útica, así como también Puertos Salgar, Pirateobueno, Zipaquira, Guatama, entre otros.

1.6 Antecedentes

La revisión de trabajos previos se elabora, atendiendo dos criterios fundamentales: el primero, la identificación de tipo de técnicas estadísticas, en el modelamiento estadístico, para identificar las variables que permiten comprender o predecir la obtención del desempeño escolar, el segundo, se orienta a identificar las variables emergentes que, resultado de un análisis comprensivo de los artículos revisados, contribuyen en comprender o explicar el desempeño escolar.

Cabe mencionar que algunas investigaciones se enfocan en comprender y explicar el rendimiento académico, entendido como lo que el estudiante obtiene después de estar inmerso en un proceso educativo formal (Díaz-Cárdenas et al., 2019; González-Correa & Hernández-Ramírez, 2017; Mato-Vázquez & Muñoz-Cantero, 2010; Sánchez-Palacios & Zambrano-Vera, 2020). Algunos otros se orientan a estudiar el logro de aprendizaje, como aquello que el estudiante alcanza después de un proceso educativo (López-Vargas et al., 2011, 2012; O. Suárez & Mora, 2018). El desempeño escolar, en la presente investigación, se asume como la puntuación que los estudiantes obtienen al presentar las pruebas estandarizadas, que para el caso en estudio, corresponde con las aplicadas por el estado colombiano a través del ICFES, conocidas como pruebas Saber 11(2021).

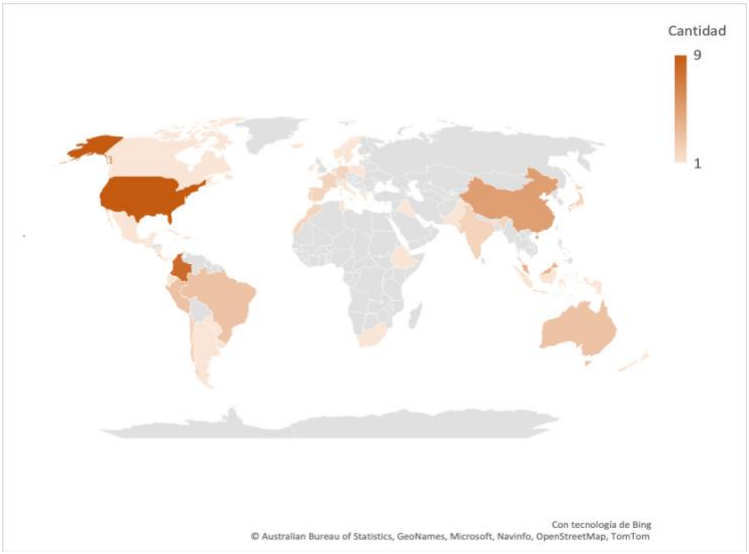


Figura 6. Distribución geográfica de los documentos consultados.
Fuente. Elaboración propia con Microsoft Excel®.

La Figura 6 muestra la georreferenciación del impacto de los artículos, de diferentes bases de datos, que han sido revisados para la presente investigación. Las bases de datos que la Universidad Escuela Colombiana Julio Garavito, a través de la internet, tiene habilitadas y a las que se tuvo acceso para la información son Elsevier (1), Google Scholar (6), IEEE (4), Scielo (2), Science Direct (5), Scopus (31), y Springer (1).

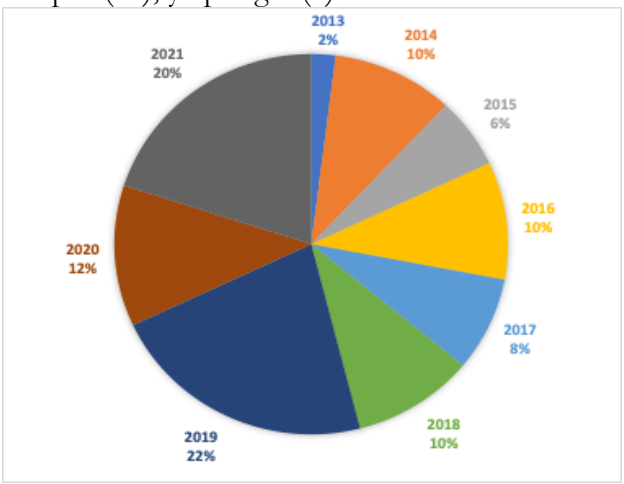


Figura 7. Distribución de artículos por año de publicación.
Fuente. Elaboración propia con Ms Excel ®.

La Figura 7 representa los artículos revisados, por año de publicación, en la que se evidencia que más del 53% son artículos publicados en los últimos tres años, menos del 30% fue publicado

entre 2016 y 2018, finalmente por su representatividad cerca del 18% de los artículos fueron publicados entre 2013 y 2015. Lo anterior demuestra que se ha tratado de recoger una mirada de los últimos años, en artículos publicados en revistas reconocidas en las bases de datos tradicionalmente indexadas y clasificadas.

1.6.1 Identificación de modelamiento estadístico en el estudio del desempeño escolar

A manera de síntesis, en la Figura 8, se identifica que los modelos de regresión son los más utilizados en la predicción del desempeño escolar. En el caso estudiado por Mineshita y otros (2021), encontraron cómo la exposición a pantallas (Televisión, computador o dispositivos móviles), podría tener un efecto negativo en la resequedad ocular, en el logro académico y la obesidad.

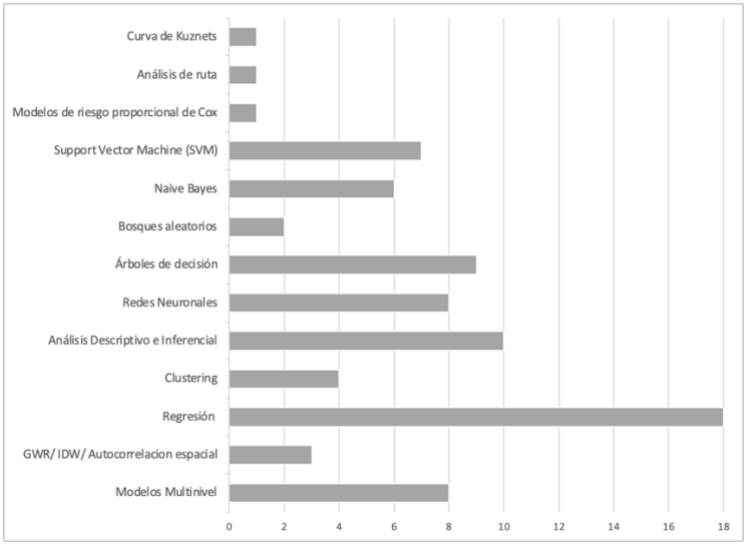


Figura 8. Distribución de técnicas estadísticas utilizadas en la comprensión o explicación del logro de aprendizaje. Fuente. Elaboración propia con Ms Excel®.

El segundo elemento epistemológico utilizado de preferencia, es el análisis descriptivo e inferencial, derivado del procesamiento estadístico de información para la comprensión del desempeño escolar (Chacón-Vargas & Roldán-Villalobos, 2021; Cvencek et al., 2017; Gaete-Rivas et al., 2021; Ibourk & Amaghouss, 2016; Khan & Ghosh, 2018; Rodríguez-De-Souza-Pajuelo et al., 2021; Wang et al., 2019), en algunos casos como complemento al modelamiento estadístico predictivo del desempeño escolar (Ali et al., 2013; Schuth et al., 2017; Zhang & Campbell, 2014).

Los árboles de decisión, las redes neuronales y los modelos multinivel, se constituyen en las terceras técnicas estadísticas que son utilizadas de preferencia, para la comprensión o explicación del desempeño escolar, tal como se observa en la Figura 8. Los árboles de decisión han sido

utilizados con un enfoque principal de comprensión en el desempeño escolar (Castrillón et al., 2020; Masci et al., 2018), así como articulado con otros modelamientos estadísticos para la explicación del desempeño escolar (Cornell-Farrow & Garrard, 2020; Rebai et al., 2019; Wandera et al., 2019; Xu et al., 2019; Yang et al., 2020).

Support Vector Machine, ha sido utilizado en conjunto con los árboles de decisión, así como complementado con redes neuronales (Xu et al., 2019), modelos de regresión (Shah et al., 2019) y clustering (Yang et al., 2020) en la comprensión o explicación del desempeño escolar. La utilización de Naive Bayes, permitió a Yang et al. (2020), identificar que los estudiantes que procrastinan tienen una menor probabilidad de tener un éxito en el desempeño escolar. Abdul-Aziz et al. (2015) apoyados en Naive Bayes encontraron que el género, los ingresos familiares y la ciudad de origen, son factores predictivos del desempeño escolar. El algoritmo clustering se ha utilizado en conjunto con otras técnicas estadísticas, para comprender o explicar el desempeño escolar (Kumari et al., 2018; Salal & Abdullaev, 2020; Yang et al., 2020), en particular el estudio adelantado por Chaparro Caso López et al. (2016), estableció dos conglomerados en el que la situación económica familiar, está relacionada con el desempeño escolar.

La revisión de trabajos adelantada da cuenta que la comunidad académica no tiene unanimidad en cuanto al uso de las técnicas estadísticas en forma estandarizada, para comprender o explicar el desempeño escolar. Contrario de lo anterior, la evidencia indica que múltiples técnicas estadísticas se articulan, como parte de la epistemología propia, que apoya desde el punto de vista cuantitativo, la comprensión o explicación del desempeño escolar.

1.6.2 Identificación de variables emergentes del desempeño escolar

La revisión adelantada para la presente investigación, se apoya en el Mapeamiento Informacional Bibliográfico (MIB), cuya finalidad se orienta a comprender e identificar las ideas del texto en forma sintética, es decir, se trata de reconfigurar en un nuevo y breve texto, el documento consultado, guardando la coherencia y sin perder de vista su objetivo, en pro de construir los antecedentes propios de un campo de investigación (Molina et al., 2012, 2013).

Se identificaron 16 variables, Figura 9, que guardan relación con el desempeño escolar, siendo estas: Naturaleza del colegio, Nivel socioeconómico, Género, Ubicación, Afecto, Antecedentes, Padres, Recursos materiales, Dependencia espacial, Relaciones con pares, Habilidades cognitivas, Alimentación, Características institucionales, Composición familiar, Experiencia docente y Edad. De igual manera, en la Figura 9 están representados los estudios, acordes al nivel educativo en los que se desarrollaron. Se observa cómo el nivel socioeconómico, es una de las variables que más se ha estudiado a nivel de básica, media y universidad, siendo la menos estudiada la alimentación.

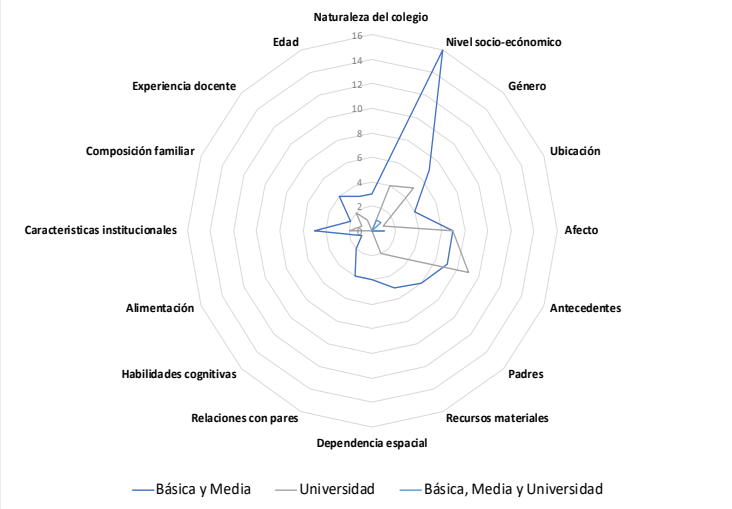


Figura 9. Variables emergentes en los estudios.
Fuente. Elaboración propia con Ms Excel ®.

La Tabla 1 se sintetizan los trabajos relacionados con algunas de las variables (nivel socio-económico, género, temas afectivos y emocionales, antecedentes de orden académico, características académicas de los padres) de la Figura 10.

Tabla 1. Trabajos relacionados con algunas de las variables.

Variable	Referencias
Nivel socio-económico relacionado con los aspectos económicos de las familias, así como el estrato socioeconómico en el que se ubican las viviendas, que de alguna manera, corresponde con el potencial económico de ingresos familiares.	(Benito et al., 2014; Chacón-Vargas & Roldán-Villalobos, 2021; Contreras et al., 2019; Lisboa- Bartholo & Da-Costa, 2016; Martínez-Mateus & TurriagoHoyos, 2015; Masci et al., 2018; Maulida & Kariyam, 2017; Rodríguez-Hernández et al., 2021; Salal & Abdullaev, 2020; Shah et al., 2019; Zhang & Campbell, 2014)
Género	(Chacón-Vargas & Roldán-Villalobos, 2021; Gaete-Rivas et al., 2021; Kumari et al., 2018; Lau et al., 2019; Lisboa- Bartholo & Da-Costa, 2016; Orjuela, 2014; Qazdar et al., 2019; Rodríguez-De-Souza-Pajuelo et al., 2021; Shah et al., 2019)
Temas afectivos y emocionales	(Cvencek et al., 2017; Dagnew, 2017; Froiland & Oros, 2014; Giannakas et al., 2021; Jovanović et al., 2021; Mineshita et al., 2021; Pollak & Parnell, 2018; Rebai et al., 2019; Rodríguez-De-Souza-Pajuelo et al., 2021; Xu et al., 2019; Yang et al., 2020)

Antecedentes de orden académico	(Chacón-Vargas & Roldán-Villalobos, 2021; Cornell-Farrow & Garrard, 2020; Febro, 2019; Hasan et al., 2018; Kumari et al., 2018; Lau et al., 2019; Maisarah-Samsudin et al., 2021; Maulida & Kariyam, 2017; Qazdar et al., 2019; Rebai et al., 2019; Rodríguez-De-Souza-Pajuelo et al., 2021; Rodríguez-Hernández et al., 2021; Tapasco-Alzate et al., 2020; Wang et al., 2019)
Características académicas de los padres	(Chacón-Vargas & Roldán-Villalobos, 2021; Cornell-Farrow & Garrard, 2020; Kumari et al., 2018; Lau et al., 2019; Masci et al., 2018; Maulida & Kariyam, 2017; Qiu & Wu, 2019; Salal & Abdullaev, 2020; Shah et al., 2019)

Fuente. Complementada a partir de Suárez et al. (2021)

Es necesario precisar que, si bien las variables son las de mayor frecuencia en los diferentes estudios, no necesariamente corresponden con el mayor peso en los modelos estadísticos que predicen el logro de aprendizaje de los estudiantes. Al parecer, la variable de mayor peso en la predicción del logro de aprendizaje, corresponde con los aspectos de orden socioeconómico de las familias.

1.7 Alcance y limitaciones

La presente investigación toma como fuente de información bases de datos de fuentes oficiales, por lo que la interpretación y alcance de las variables y el modelo encontrado, está limitada por los datos, sin embargo, los modelos estadísticos aplicados tienen relevancia en el estudio adelantado.

A continuación, se describe el alcance del presente proyecto de investigación.

- El modelo obtenido del procesamiento de los datos, se proyecta con base en las fuentes de información para los municipios del departamento de Cundinamarca – Colombia, lo anterior, dado que se tuvo la facilidad de acceso a los datos de diferentes fuentes de Cundinamarca, se optó a que este fuera la población de estudio.
- Colegios oficiales y no oficiales de calendario A y B.
- El modelo se enfoca en predecir las puntuaciones de la población estudiantil de educación media, cuando presenta las pruebas Saber 11, organizadas por el Estado colombiano, a través del Instituto colombiano para la Evaluación de la Educación ICFES.
- El modelo resultado de la investigación, es el mejor obtenido, luego de comparar cinco algoritmos ubicados en el contexto de Machine Learning y dos modelos generalizados multinivel.
- En razón a que las variables se normalizan y el modelo resultante es una regularización de la regresión lineal, el alcance del modelo es presentar el efecto positivo o negativo de las variables que lo componen, no siendo posible determinar el peso de las variables.

En las siguientes líneas se describen las limitaciones del presente proyecto investigación.

- La calidad, seguridad y validez de los datos utilizados para la presente investigación, están sujetos al rigor que cada una de las entidades tienen y siguen para su obtención, registro y almacenamiento.
- Los datos empleados están relacionados con los períodos previos a la pandemia (2015 a 2019) debido a la COVID-19, por lo que la capacidad predictiva de los modelos obtenidos, en consecuencia, del modelo seleccionado, estará sujeta a la realidad existente en los periodos estudiados; corresponderá a próximos estudios, analizar lo ocurrido en el periodo de la pandemia y de la post-pandemia.
- La calidad del procesamiento utilizado para los modelos que se obtienen, está sujeta a las librerías y algoritmos que se han retomado y adaptado, de las diferentes fuentes de desarrollo. Sin embargo, tanto las fuentes, como los algoritmos, tienen un reconocimiento en la comunidad académica relacionada con la ciencia de datos.

2 Fundamentos teóricos

En este apartado de la tesis se describen los elementos sobre los cuales se apoya la investigación, es decir, el conjunto de modelos estadísticos que soportan el procesamiento de la información de la que trata el estudio.

2.1 Minería de datos

En la actualidad, se generan y recolectan continuamente una enorme cantidad de datos heterogéneos, sin procesar, que han alcanzado el orden de petabytes o exabytes, resultado de los avances tecnológicos. De acuerdo con ello, la minería de datos es el análisis de la recopilación, limpieza, procesamiento, análisis y obtención de información valiosa, mediante el descubrimiento de patrones y relaciones ocultas o implícitas en los datos (Aggarwal, 2015).

El proceso de minería de datos está orientado mediante fases como recolección de datos, preprocesamiento de datos y proceso de análisis, donde el desafío de cada aplicación es único (Aggarwal, 2015). Por lo tanto, en la aplicación y ejecución de este proceso existen diferentes metodologías como Knowledge Discovery in Databases (KDD), Cross Industry Standard Process for Data Mining (CRISP-DM) y SEMMA (Sample, Explore, Modify, Model, Assess), que proveen una guía para llevar a cabo el proceso en forma sistemática y no trivial (Moine et al., 2011).

En la presente investigación se utiliza la metodología CRISP-DM, creada por un grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, que estructura el proceso en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Es importante resaltar que el proceso no es rígido y es jerárquico, donde cada fase está compuesta por tareas genéricas y especializadas (Moine et al., 2011).

2.2 Machine Learning

El Machine Learning es la ciencia que brinda a los computadores, la facultad de aprender de los datos, sin ser programados explícitamente, a través de la experiencia de realizar una tarea repetidamente (Ertel, 2017; Géron, 2017). El objetivo es construir modelos con buena capacidad de generalización, es decir, que predigan con precisión las etiquetas o valores numéricos de registros desconocidos. En este sentido, el conjunto de datos es dividido en: el conjunto de entrenamiento, que contiene el conocimiento que el algoritmo va a extraer y aprender, y el conjunto de prueba, que permite evaluar a través de una métrica de desempeño, si el algoritmo tiene poder de generalización con nuevos datos (Ertel, 2017).

Estos algoritmos se clasifican en cuatro categorías, según la cantidad y tipo de supervisión: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje

por refuerzo. Los algoritmos que se presentaran a continuación corresponden a aprendizaje supervisado.

2.2.1 Regresión lineal múltiple

El modelo de regresión lineal simple es una técnica de modelado estadístico que tiene como propósito determinar la relación de dependencia, entre una variable dependiente o respuesta Y, con una variable independiente, explicativa, regresora o covariable X, (Heumann et al., 2016), representado como:

$$Y = \beta_0 + \beta_1 X \quad (1)$$

donde β_0 es el intercepto y β_1 el parámetro de pendiente que indica el cambio del valor de *Y*, cuando el valor de *X* cambia en una unidad. Si β_1 es positivo, significa que el valor de Y incrementa, si *X* incrementa y si β_1 es negativo, significa que el valor de *Y* decrece, si *X* decrece.

Por otro lado, cada observación tiene un error, es decir, se desvía en ε de la aproximación lineal. Luego, el modelo de la ecuación (1) se expresaría como:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (2)$$

Considerando n observaciones (x_1, y_1) (x_2, y_2), …, (x_n, y_n), entonces $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ donde los errores ε_i son idénticamente distribuidos e independientes con media 0 y varianza constante σ^2, es decir, $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$ para todo *i*.

Para estimar los valores de los parámetros, se emplea el método Ordinary Least Squares (OLS) que busca minimizar la suma residual de cuadrados entre los valores observados en el conjunto de datos y los valores predichos por la aproximación lineal. Por ejemplo, para cada (x_i, y_i), el error es $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$. El problema estaría en minimizar los suma de los cuadrados de ε_i, en otras palabras,

$$\widehat{\beta_i} = min_{\beta_i} \sum_{i=1}^{n} {\varepsilon_i}^2 \quad (3)$$

Resolviendo el problema de optimización, se obtiene:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}; \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} \quad (4)$$

De momento, se ha asumido que la variable dependiente se ve afectada por una sola variable explicativa, pero en muchos casos, es afectada por más de una. De este modo, el modelo de regresión lineal múltiple es representado como:

$$Y = \beta_0 + \sum_{k=1}^{p} \beta_k X_k + \varepsilon \quad (5)$$

Considerando un conjunto de datos con n observaciones, cada observación satisface el modelo, es decir,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \text{ para todo } i = \{1,2,\cdots,n\} \quad (6)$$

La ecuación (6) puede ser expresada de forma matricial como $y = X\beta + \varepsilon$ donde

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

y suponiendo que $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2 I$. Para estimar los parámetros β se utiliza el método OLS, obteniendo que $\hat{\beta} = (X^T X)^{-1} X^T y$.

Los supuestos de la regresión lineal (Hoffmann, 2010) son:

- *Homocedasticidad:* La variación del error es constante para todas las combinaciones de las variables explicativas, es decir, $Var(\varepsilon_i | X_i) = \sigma^2$. La prueba de homocedasticidad Breusch-Pagan, puede ser utilizada para comprobar este supuesto.
- *Independencia:* Los valores de *Y* son independientes entre sí, es decir, las observaciones son independientes. La prueba de independencia Durbin-Watson, puede ser utilizada para comprobar este supuesto.
- *Normalidad:* Los errores se distribuyen normalmente con media cero y varianza constante. Las pruebas de normalidad Shapiro-Wilk y Kolmogorov-Smirnov, pueden ser utilizadas para comprobar este supuesto.

2.2.2 Ridge Regression

La regresión de Ridge, también conocida como la regularización de Tikhonov, es una forma regularizada de la regresión lineal, que utiliza el término de penalización L2 igual a

$$\alpha \sum_{i=1}^{p} \beta_i^2 \quad (7)$$

aplicado al método OLS. De modo que, en la regresión lineal simple $\hat{\beta} = min_{\beta_0,\beta_1} \left(\sum_{i=1}^{n} \varepsilon_i^2 + \alpha \sum_{i=1}^{p} \beta_i^2 \right)$ y en la regresión lineal múltiple $\hat{\beta} = (X^T X + \alpha I)^{-1} X^T y$. En consecuencia, el algoritmo no sólo se ajustará a los datos, sino también mantendrá los pesos

del modelo lo más pequeños posible (Géron, 2017), aunque se pierde interpretabilidad al agregar el término de regularización.

2.2.3 Lasso Regression

Least Absolute Shrinkage and Selection Operator Regression, también conocida como Lasso Regression, es una forma regularizada de la regresión lineal, que utiliza el término de penalización L1 igual a

$$\alpha \sum_{i=1}^{p} |\beta_i| \tag{8}$$

aplicado al método OLS. De modo que, en la regresión lineal simple $\hat{\beta} = min_{\beta_0,\beta_1}\left(\sum_{i=1}^{n} {\varepsilon_i}^2 + \alpha \sum_{i=1}^{p} |\beta_i|\right)$ y en la regresión lineal múltiple $\hat{\beta} = (X^T X + \alpha I)^{-1} X^T y$. Al igual que Ridge Regression, el algoritmo mantendrá los pesos del modelo lo más pequeños posible y pierde interpretabilidad, sin embargo, a diferencia de Ridge Regression, tiende a eliminar completamente los pesos de las características menos importantes, es decir, establecerlas en cero (Géron, 2017).

2.2.4 Decision Tree

Decision Tree es un método supervisado de Machine Learning, usado para clasificación y regresión, cuyo objetivo es construir un modelo con buena capacidad de generalización, mediante reglas de decisión derivadas de las variables explicativas. Algunos de los algoritmos empleados son CART y ID3, junto con las métricas para medir la calidad de la división que permite seleccionar la mejor (Ertel, 2017).

Las métricas de clasificación son Gini, Entropy y Classification Error:

$$Entropy(S) = \sum_{x \in C} -p(x) log_2 p(x) \tag{9}$$

$$Gini(S) = 1 - \sum_{x \in C} p(x)^2 \tag{10}$$

$$Error(S) = 1 - max(p(x)) \tag{11}$$

donde S es el conjunto de datos, C es el conjunto de clases de S y $p(x)$, es la proporción de los números de elementos con clase x dentro del conjunto de elementos de S.

Las métricas de regresión son Mean Square Error (MSE) y Absolute Error (AE) para el nodo *m*:

$$MSE(X_m) = \frac{1}{N_m} \sum_{i \epsilon N_m} (y_i - \overline{y_m})^2 \quad (12)$$

$$AE(X_m) = \frac{1}{N_m} \sum_{i \epsilon N_m} |y_i - \overline{y_m}| \quad (13)$$

donde X_m es el conjunto de entrenamiento del nodo *m*, N_m el número de observaciones del nodo *m*, $\overline{y_m}$ es el promedio de las observaciones del nodo *m* y y_i son las observaciones del nodo *m*.

La construcción del Decision Tree para regresión se realiza mediante un proceso iterativo, para encontrar la mayor reducción posible de las métricas mencionadas, de la siguiente forma (Breiman et al., 1984):

a. El proceso iterativo inicia en la parte alta del árbol, donde todas las observaciones están dentro de la misma región.
b. Se identifica los puntos de posibles umbrales para cada una de las variables explicativas ($X_1, X_2, \cdots, X_p$). A diferencia de las variables categóricas, en las variables continuas, se utiliza discretización binaria, que convierte a los atributos estableciendo un umbral, es decir, se ordenan de menor a mayor los valores y el punto intermedio entre cada par, se escoge como los umbrales. En el caso de variables categóricas, los umbrales son los niveles de cada una.
c. Luego, se selecciona una métrica y se calcula el valor para cada posible división, generada en el paso anterior.
d. Se selecciona el umbral que obtuvo la mejor división, de acuerdo con la métrica, y si existen dos o más divisiones con el mismo valor, la elección será aleatoria.
e. Finalmente, se repite de forma iterativa, hasta alcanzar el criterio de parada, como número mínimo de observaciones por región.

Para predecir un nuevo dato, se debe recorrer el árbol entrenado, evaluado en cada nodo la condición establecida, hasta llegar a una de las hojas y calcular el promedio de los valores de la variable respuesta que pertenecen a la hoja.

Las características del algoritmo son (Tan et al., 2005):

- Es un método no paramétrico, es decir, no requiere suposiciones previas de las distribuciones con respecto a las clases y los otros atributos.
- Hallar un árbol óptimo es un problema, entonces debe utilizarse un enfoque heurístico.
- Son computacionalmente económicos y aplicarlo para calcular una clasificación es rápido.
- Pueden generar árboles muy complejos que no generalizan los datos y son inestables, es decir, una variación puede generar un árbol completamente diferente.
- A diferencia de otros algoritmos, es fácilmente interpretable.

2.2.5 Random Forest

El algoritmo de Random Forest, está compuesto de árboles de decisión, por lo que puede tener mayor estabilidad y precisión, cuyos métodos más comunes de aprendizaje en conjunto son: bagging, boosting y stacking (Sullivan, 2017).

El bagging es un método que busca la reducción de la varianza entre las predicciones, al sumar las salidas de dos o más clasificaciones o predicciones que se entrenan con diferentes muestras del conjunto de datos, aplicando los siguientes pasos (Sullivan, 2017):

a. Crear múltiples conjuntos de datos: Se forman conjuntos de datos con muestreo con reemplazo.
b. Construir múltiples modelos: Estos se entrenan con cada uno de los conjuntos del paso anterior.
c. Combinar los modelos: Cada modelo genera predicciones individuales y son combinadas con una técnica de agregación, eligiendo la clasificación más votada o el promedio de las clasificaciones.

Las características del algoritmo son (Sullivan, 2017):

- Se pude emplear para clasificación y regresión.
- Es posible que maneja una gran cantidad de datos en alta dimensionalidad y puede estimar de manera efectiva datos faltantes.
- No es tan bueno en la clasificación que, en la regresión, ya que no puede hacer predicciones más allá de rango de los datos de entrenamiento.
- Si los datos tienen demasiado ruido, tienden a ajustarse demasiado.

2.2.6 K-Nearest Neighbors

K-Nearest Neighbors, es un algoritmo que no crea un modelo de aprendizaje con los datos de entrenamiento, sino que construye el modelo al mismo tiempo en que se prueban los datos. El proceso iterativo de este, calcula la distancia o similitud entre cada ejemplo del conjunto de prueba y todos los ejemplos del conjunto de entrenamiento, selecciona los k vecinos más cercano y determina el valor, como el promedio de las predicciones o la clasificación más votada (Tan et al., 2005).

Las características del algoritmo son (Tan et al., 2005):

- Es parte de una técnica más general, conocida como aprendizaje basado en instancias, puesto que utiliza observaciones específicas del conjunto de entrenamiento para hacer predicciones.

- Es un algoritmo perezoso, debido a que no requiere la construcción de un modelo y por ello, puede ser bastante costoso computacionalmente.
- Puede producir predicciones incorrectas, sino se toman el tipo de distancia y los pasos de preprocesamiento adecuados.

2.3 Elementos teóricos adicionales que considerar

A continuación se enuncia el método cross validation y las métricas utilizadas en el contexto propio del Machine Learning.

2.3.1 Cross Validation

La construcción de modelos de Machine Learning, tiene como objetivo enseñarle al computador a clasificar o predecir, partiendo de un conjunto de datos. Sin embargo, si no se realiza el proceso correctamente, el modelo no reconocerá datos que sean diferentes a los que le fueron dados para entrenar, ya que se ajustan a la complejidad de estos. Este problema se conoce como *"overfitting"* (Ertel, 2017).

Por otro lado, según la Navaja de Ockham, se estipula que entre dos modelos que explican igualmente bien, se prefiere el más sencillo. Además, cuanto más complejo es el modelo, más detalles representa, pero así mismo, menos se transfiere el modelo a datos nuevos. Esto se observa en la Figura 11, donde se entrenaron árboles de decisión de varios tamaños y se calculó las tasas de error del conjunto de entrenamiento y el de prueba. De esta manera, la tasa de error en los datos de entrenamiento y de prueba, disminuyen hasta que el número de nodos es 55, e inmediatamente después, la tasa de error en el entrenamiento comienza a aumentar (Ertel, 2017).

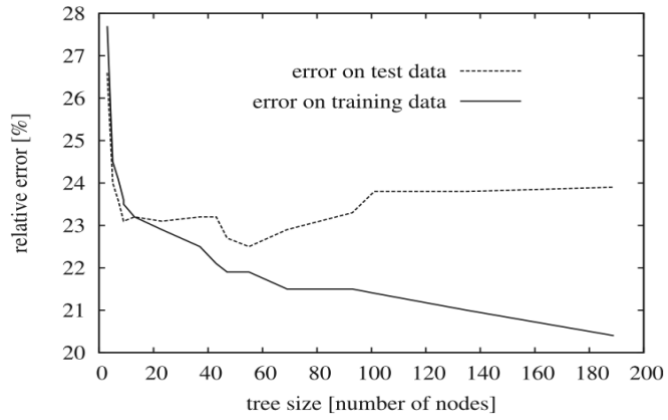


Figura 11. Curva de aprendizaje.
Fuente. Ertel (2017)

Por consiguiente, el método de *"cross validation"* es uno de los más utilizados para resolver el problema de *"overfitting"*, el cual pretende controlar el modelo, minimizando el error de aproximación del conjunto de datos, a través de hiperparámetros como la profundidad del árbol, el término de regularización, entre otros.

El "cross validation" es un método de optimización que funciona, dividiendo el conjunto de datos, en k bloques del mismo tamaño. El algoritmo se entrena k veces sobre $k-1$ bloques y testea con el bloque restante. Posteriormente, promedia los k errores y elige el que tenga el error medio más pequeño, para entrenar el modelo final.

2.3.2 Métricas

Las métricas de rendimiento, permiten observar el progreso de los modelos en las tareas de Machine Learning. A continuación, algunas métricas utilizadas para problemas de regresión en un conjunto de datos con n observaciones, donde $\hat{y}_i$ es la predicción y y_i es el valor real del *i*-ésimo valor:

a. Root Mean Square Error (RMSE):

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

b. Mean Absolute Error (MAE):

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

c. R^2:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{16}$$

2.4 Modelos lineales generalizados mixtos

Los modelos lineales tradicionales, conocidos como regresión lineal, han sido una técnica estadística ampliamente utilizada por su interpretabilidad y sus resultados teóricos. Sin embargo, en muchas aplicaciones, aparecen datos agrupados y relacionados donde esta técnica es inadecuada, ya que una suposición fundamental para su uso, es la independencia en las variables. Por ejemplo, en la investigación educativa, los puntajes de las pruebas de los estudiantes pueden estar relacionados con las instituciones educativas. De este problema, surgieron *los modelos lineales generalizados mixtos,* que permiten incluir este comportamiento mediante los efectos fijos, coeficientes determinados a propósito, que no varían entre niveles, y aleatorios, aquellos

tomados al azar de un conjunto de datos de posibles niveles (Correa Morales & Salazar Uribe, 2016).

Con el objetivo de exponer los modelos mixtos, se considerará el siguiente conjunto de datos sobre logaritmos de crecimiento de árboles de especie abeto, que contiene 12 observaciones por grupo (ambiente de control, ambiente rico en ozono) y 4 mediaciones del crecimiento de los días 152, 174, 201 y 277 (Tabla 2):

Tabla 2. Muestra de datos sobre árboles abeto.

Id	Grupo	y_{152}	y_{174}	y_{201}	y_{277}
1	Control	2,72	3,10	3,30	3,38
2	Control	3,30	3,90	4,34	4,96
3	Control	3,98	4,36	4,79	4,99
4	Control	4,36	4,77	5,10	5,30
⋮	⋮	⋮	⋮	⋮	⋮
21	Ozono	4,56	5,13	5,40	5,69
22	Ozono	5,16	5,49	5,74	6,05
23	Ozono	5,46	5,79	6,12	6,41
24	Ozono	4,52	4,91	5,04	5,71

Fuente: Correa Morales y Salazar Uribe (2016)

En la Figura 12, se observa que los gráficos de línea del grupo control por cada árbol, tienen un intercepto especifico, aunque sus pendientes son similares. Este tipo de datos puede ser analizado con estadísticos de resumen como pruebas de hipótesis, análisis de crecimiento y regresión lineal para cada sujeto. Sin embargo, es posible que se pierda información y no muestre la tendencia de los datos. Por lo tanto, es recomendable utilizar el modelo de dos etapas, el cual consta de ajustar el modelo de regresión para cada sujeto e interpretar la variabilidad de los coeficientes de las regresiones anteriores, usando los efectos fijos (Correa Morales & Salazar Uribe, 2016).

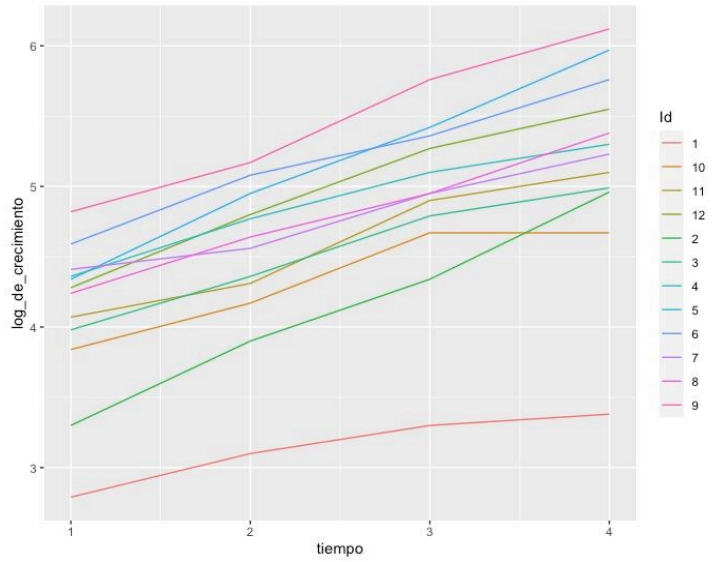


Figura 12. Gráfico de línea del crecimiento.
Fuente. Adecuación de gráfico de Correa Morales y Salazar Uribe (2016)

Supóngase que Y_{ij} es la respuesta para el *i*-ésimo sujeto en el tiempo X_{ij}, donde $i = 1,2,\ldots,m$ y $j = 1,2,\ldots,n_i$, de manera que $Y_i = [Y_{i1}, Y_{i2}, \cdots, Y_{in_i}]$ representa las respuestas del sujeto *i*. Luego, el modelo $Y_i = Z_i\beta_i + \varepsilon_i$, describe la variabilidad intra-sujeto donde Z_i es una matriz de covariables con dimensión $n_i\, x\, p$, β_i los coeficientes de la regresión con dimensión p y $\varepsilon_i \sim N(0, \Sigma_i)$ con Σ_i una matriz de varianzas y covarianzas.

Asimismo, se modela la variabilidad entre-sujetos con los β_i que se relacionan con los efectos fijos, es decir, $\beta_i = K_i\beta + b_i$ donde K_i es la matriz de covariables conocidas, β es un vector de parámetros de regresión desconocidos y $b_i \sim N(0, D)$ con D una matriz de varianzas y covarianzas. Para concluir, se combina las dos etapas en un solo modelo como (Correa Morales & Salazar Uribe, 2016):

$$Y_i = Z_iK_i\beta + Z_ib_i + \varepsilon_i \qquad (17)$$

$b_1, b_2, \cdots, b_m, \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m$ son mutuamente excluyentes

donde β son los efectos fijos, b_i y ε_i son los efectos aleatorios y D y Σ_i son los componentes de varianza. Además, $Cov(Y_i) = Z_iDZ_i^T$.

Retomando el ejemplo anterior, en la primera fase se formula el modelo $Y_i = Z_i\beta_i + \varepsilon_i$ donde $Z_i = \begin{bmatrix} 1 & 152 \\ 1 & 174 \\ 1 & 201 \\ 1 & 227 \end{bmatrix}$ y $\beta_i = \begin{bmatrix} \beta_{1i} \\ \beta_{2i} \end{bmatrix}$ con β_{1i} es el intercepto desconocido y β_{2i} es el intercepto

conocido para i. Posteriormente, en la segunda etapa, se relacionan los elementos de β_i y los efectos del tiempo con los grupos de la siguiente manera:

$$\beta_{1i} = \beta_0 + b_{1i} \tag{18}$$

donde b_{1i} es el efecto aleatorio que varía el intercepto y β_0 es la respuesta promedio al comienzo del tratamiento. A su vez, $\beta_{2i} = \beta_1 C_i + \beta_2 O_i$ son los efectos fijos, debido a que se asumen las pendientes iguales, donde β_1, β_2 son los efectos promedio del tiempo para los grupos,

$$C_i = \begin{cases} 1 & si\ pertenece\ al\ grupo\ control \\ 0 & en\ otro\ caso \end{cases}$$

y

$$O_i = \begin{cases} 1 & si\ pertenece\ al\ grupo\ rico\ en\ ozono \\ 0 & en\ otro\ caso \end{cases}$$

De manera que, de conformidad con la formulación del modelo, una opción adecuada para β, K_i y b_i sería:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$K_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & C_i & O_i \end{bmatrix}$$

$$b_i = \begin{bmatrix} \beta_i \\ 0 \end{bmatrix}$$

La formulación del modelo lineal mixto sería $Y_{ij} = \beta_0 + \beta_1 C_i x_j + \beta_2 O_i x_j + b_i + \varepsilon_{ij}$ con $b_i \sim N(0, \sigma^2)$ y $\varepsilon_i \sim N(0, \Sigma_i)$ mutuamente excluyentes, conocido como modelo de intercepto aleatorios.

Se considera el Criterio de Información de Akaike (AIC) (Touchon, 2021) como una medida de calidad del modelo definido como

$$\text{AIC} = -2\text{logL}\left(\hat{\beta}, \hat{\sigma}^2\right) + 2\text{p} \tag{19}$$

donde $logL\left(\hat{\beta}, \hat{\sigma}^2\right) = -\frac{n}{2} - \frac{n}{2} log\left(\frac{RSS}{n}\right) - \frac{n}{2}$ es la probabilidad logarítmica maximizada de un modelo de regresión con $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. El más pequeño AIC es el mejor, ya que el término que varía es el RSS, luego un modelo se ajusta bien cuando el RSS es bajo, y la penalidad

$2p$ donde p es el número de parámetros β, entonces es mejor que el p sea bajo. En otras palabras, un buen modelo tendrá el equilibrio entre un buen ajuste y uso de una pequeña cantidad de parámetros. Para comparar modelos es suficiente

$$AIC = nlog\left(\frac{RSS}{n}\right) + 2p \tag{20}$$

2.4.1 Modelo multinivel o lineal jerárquico generalizado

Los *modelos multinivel o lineal jerárquicos generalizados* son un tipo de modelo generalizado mixto, empleado para datos que tiene una estructura jerárquica y donde las variables pueden definirse en cada nivel. Esta técnica es útil debido a que existen realidades observadas por niveles, permite realizar análisis por nivel o por las interrelaciones entre niveles y la estructura anidada, ayuda a revelar relaciones que en una regresión lineal no permite (Acevedo-Álvarez, 2008).

Con el fin de presentar un ejemplo, se considera un estudio donde interesa conocer la relación entre la variable respuesta del desempeño escolar y otras variables explicativas, como el estatus socioeconómico, la calidad de la planta de profesores, la disponibilidad de materiales de trabajo, la alimentación, entre otros. Sin embargo, las diferencias detectadas pueden deberse a los colegios a los que pertenece cada estudiante, es decir, el desempeño puede variar de un colegio a otro y no considerar el efecto, que los colegios ofrecen, puede generar información incompleta y potencialmente engañosa (Acevedo-Álvarez, 2008).

Por lo tanto, se amplía la capacidad explicativa de la regresión lineal simple, planteando el desempeño escolar, con respecto al estatus socioeconómico como (Acevedo-Álvarez, 2008):

$$rend_{ij} = a_j + b_j * estatus + \varepsilon_{ij} \tag{21}$$

donde *i* representa a los estudiantes y *j* identifica los colegios a los que pertenecen. Además, a_j es el intercepto que proporciona información general de los colegios, b_j es el incremento de colegio a colegio y ε_{ij}, son los errores entre la predicción y el valor real.

Los modelos jerárquicos lineales tienen varias formas de representación matemática, por lo que se presentará el modelo más simple, conocido como modelo nulo, el cual únicamente cuenta con la variable respuesta y una constante. Retomando el ejemplo, el modelo cuenta con dos niveles, los estudiantes y los colegios, cuya ecuación es (Acevedo-Álvarez, 2008):

$$\begin{gathered} Y_{ij} = \beta_{oj} + \varepsilon_{ij} \\ \beta_{oj} = \beta_{00} + \mu_{0j} \\ \varepsilon_i \sim N(0, \sigma^2) \end{gathered} \tag{22}$$

donde Y_{ij} es la variable respuesta, es decir, el rendimiento del estudiante *i* dentro de la escuela *j*, β_{oj} representa la media general de la escuela *j*, β_{00} es la media general de todas las escuelas, ε_{ij} es el error entre la predicción y el valor real de cada estudiante y μ_{0j}, es el error entre la predicción y el valor real de cada colegio.

Ahora bien, teniendo en cuenta el modelo lineal generalizado mixto de la ecuación (17), el modelo multinivel puede formularse como

$$\begin{aligned} Y_i|b_i &\sim N(Z_iK_i\beta + Z_ib_i, \Sigma_i) \\ b_i &\sim N(0, D) \end{aligned} \tag{23}$$

aunque los modelos multinivel no siempre estas restringidos al caso normal (Correa Morales & Salazar Uribe, 2016).

Para estimar los parámetros se emplea el método de estimación por Máxima Verosimilitud (MV) (Gelman & Hill, 2006), por medio de la función de verosimilitud marginal definida como

$$L(\theta) = \prod_{i=1}^{m} \left\{ (2\pi)^{-\frac{n_i}{2}} |V_i(\alpha)|^{-\frac{1}{2}} e^{-\frac{1}{2}\left((Y_i - Z_iK_i\beta)^T V_i{}^{-1}(\alpha)(Y_i - Z_iK_i\beta)\right)} \right\} \tag{24}$$

donde $\theta = [\beta, \alpha]$, α es el vector de todos los componentes de varianza en D y Σ_i y $V_i = Z_i D Z_i{}^T$. Si α se conoce la estimación de β por MV sería

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{m} (Z_iK_i)^T V_i{}^{-1} Z_iK_i \right)^{-1} \sum_{i=1}^{m} (Z_iK_i)^T V_i{}^{-1} Z_iK_i \tag{25}$$

En general α no se conoce, por lo que, se estima como $\hat{\alpha}$ mediante MV maximizando con respecto a α, la función $L\left(\alpha, \hat{\beta}(\alpha)\right)$. Sin embargo, el estimador de MV tienden a ser sesgados, de modo que Patterson y Thompson (1971) plantearon la estimación por Máxima Verosimilitud Restringida (MVR) que corrige en relación con los efectos fijos para conseguir errores menores que el método MV (Correa Morales & Salazar Uribe, 2016).

2.4.2 Regresión multinivel gamma

La *regresión multinivel gamma* es un tipo de modelo generalizado mixto, para el análisis de positivos asimétricos, mediante la distribución gamma que ha sido utilizada principalmente para tiempos de sobrevivencia (Paula, 2013).

En este modelo se supone que $Y_1, Y_2, \cdots, Y_n$ son variables aleatorias, tales que $Y_i \sim Gamma(\mu_i, \varphi)$. En otras palabras, las variables tienen distinta media y el mismo coeficiente de variación $\varphi^{-\frac{1}{2}}$. Además, $g(\mu_i) = \eta_i$ donde $\eta_i = x_i{}^T \beta$ con $x_i = [x_{i1}, x_{i2}, \cdots, x_{ip}]$, son las variables explicativas y $\beta = [\beta_1, \beta_2, \cdots, \beta_p]$, el vector de parámetros de la represión multinivel gamma. Adicionalmente, se consideran las funciones de enlace más utilizadas como son la identidad, la logarítmica y la recíproca (Paula, 2013).

Considerando que la distribución de Y condicionado en b son independientes y obtenidos de una distribución de la familia exponencial, es decir,

$$f_{y_i|b}(y_i|b, \beta, \emptyset) = e^{\frac{y_i \eta_i + c(\eta_i)}{a(\phi)} + d(y_i, \emptyset)} \tag{26}$$

$$b \sim f_b(b|D)$$

Para estimar los parámetros se emplea el método de estimación por Máxima Verosimilitud (MV), mediante la función de verosimilitud dada por

$$L(\beta, \emptyset, D|Y) = \int \prod_{i=1}^{n} f_{y_i|b}(y_i|b, \beta, \emptyset) f_b(b|D) db \tag{27}$$

Esta función requiere de métodos numéricos para su resolución.

3 Metodología

En este capítulo se describe fundamentalmente el enfoque y el proceso metodológico. El primer caso, corresponde con el sustento propio, desde el modelamiento estadístico que da cuenta de cómo se utiliza el fundamento teórico descrito en el capítulo anterior para efecto del procesamiento de la información y en la segunda parte se describe la forma en cómo se abordó las diferentes etapas del proyecto investigación.

3.1 Enfoque metodológico

El enfoque metodológico de la presente investigación es cuantitativo (Hernández-Sampieri et al., 2014), basado en el uso de diversos modelos estadísticos (Breiman et al., 1984; Ertel, 2017; Heumann et al., 2016; Sullivan, 2017; Tan et al., 2005), previamente tratados en el capítulo anterior. A continuación, se indica cada uno de los modelos, especificando cómo se desarrolla su participación en la investigación. Los modelos que se utilizan para el modelamiento de los datos son: Regresión lineal múltiple (Heumann et al., 2016), Regresión Ridge (Ridge regression) (Géron, 2017), Regresión Lasso (Lasso regression) (Géron, 2017), Árboles de decisión (Decision tree) (Breiman et al., 1984; Ertel, 2017), Bosque Aleatorio (Random Forest) (Sullivan, 2017), Modelo multinivel o lineal jerárquico generalizado (Acevedo-Álvarez, 2008) y Modelo multinivel gamma (Paula, 2013).

3.2 Proceso metodológico

En este apartado se describe la forma en cómo se desarrolló el proyecto de investigación. El proyecto en general se consolida en cuatro etapas, apoyados en la metodología Cross Industry Standard Process for Data Mining (CRISP_DM), metodología propuesta al inicio del actual milenio (Chapman et al., 2000), que ha mostrado por más de dos décadas, tener pertinencia para el desarrollo de proyectos de ciencia de datos (Martinez-Plumed et al., 2021), con múltiples casos, tales como lo resume Marban, et al. (2009), lo concretó Salcedo-Parra y Galeano, en la implementación de un Ware House (2010), o recientemente Espinosa -Zuñiga la utilizó para la segmentación de un importante volumen de datos, en el contexto para la obtención de un modelo de segmentación de datos (2020).

La metodología CRISP_DM considera 6 momentos: 1) Comprensión del negocio o problema, 2) Comprensión de los datos, 3) Preparación de los datos, 4) Modelado, 5) Evaluación y 6) Implementación. Cada uno de los momentos se caracteriza por aspectos que resultan ser relevantes para el desarrollo del proyecto. A partir de Chapman, et. al (2000), se sintetiza a continuación, cada uno de esos momentos:

- *Comprensión del negocio o problema.* Se orienta a establecer de la manera más clara posible, el objetivo del proyecto, es decir, se fundamenta en un ejercicio cognitivo en el que se trata de discernir el problema a investigar, en el contexto propio de origen.
- *Comprensión de los datos.* Se trata de ubicar las fuentes primarias, elaborar la descripción propia, explorar y establecer la calidad y la necesidad potencial de imputar o descartar conjuntos de datos.
- *Preparación de los datos.* Se caracteriza por elaborar una serie de tareas orientadas a disponer los datos para su procesamiento, lo que está sujeto al tipo de técnicas de modelado a utilizar, transformación de datos a los que haya lugar, identificación de potenciales relaciones entre fuentes de datos diversas y, sobre todo, hacer la limpieza de los datos.
- *Modelado.* Está orientado a aplicar las técnicas de modelamiento estadístico definidas en el capítulo anterior. Este momento requiere conocer tanto las técnicas estadísticas, como la forma en que éstas se emplean.
- *Evaluación.* Se evalúa el resultado del momento anterior, al obtener el (los) modelo (s) del conjunto de datos estudiados, para atender el problema en el contexto propio de los datos.
- *Implementación.* Aunque éste no hace parte del presente proyecto de investigación, se orienta a establecer acciones que permitan abordar el problema en estudio.

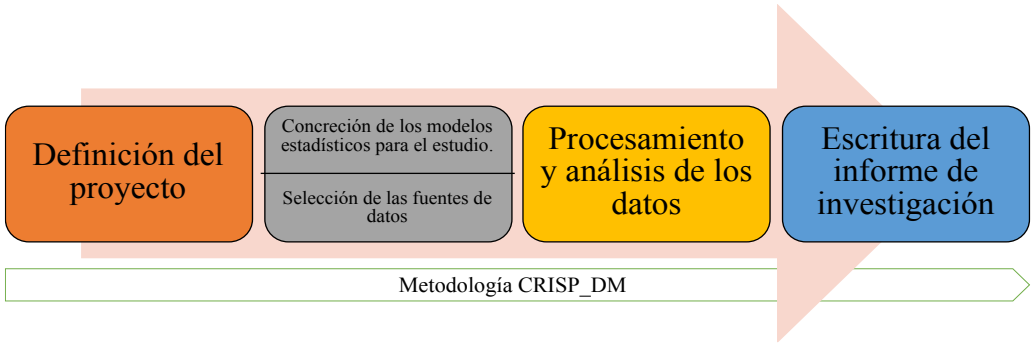


Figura 13. Etapas del proceso metodológico basado en la metodología CRISP_DM.
Fuente. Elaboración propia.

En la Figura 13 se identifican las cinco etapas que hicieron parte del proceso metodológico, a través de las cuales se desarrolla la presente investigación, que se sustenta sobre la metodología CRISP_MD antes descrita. A continuación, se desarrolla cada una de las etapas.

3.2.1 Definición del proyecto

En esta etapa el equipo de investigación entabla las relaciones necesarias a nivel de gestión interinstitucional, indagación de fuentes potenciales de información, así como las demás actividades que se concretaron en la formulación del proyecto, que dio lugar a la investigación y

que se describe en el capítulo 1. El proyecto se ajustó a la exigencia propia del programa de Maestría de Ciencia de Datos. Las siguientes dos etapas se desarrollaron casi que en paralelo.

3.2.2 Concreción de las categorías de estudio

Esta etapa se centra en describir los modelos estadísticos utilizados en la construcción de los modelos supervisados, en pro de establecer la predicción de las puntuaciones de los estudiantes de las pruebas Saber 11, en la población en estudio. En concreto, este apartado se desarrolla a en el capítulo dos de la presente tesis, denominado fundamentos teóricos.

3.2.3 Selección de fuentes de datos

Para abordar el problema en estudio, se acude a "fuentes primarias" de datos públicas de diferentes entidades, acorde a la periodicidad de cada entidad (semestral o anual). Los conjuntos de datos son:

- Resultados pruebas Saber 11 del semestre 2015-1 al 2019-2. Fuente Instituto Colombiano para la Evaluación de la Educación, ICFES.
- Indicadores de infancia y adolescencia del año 2015 al 2019. Fuente Sistema Nacional de Información de la Educación Superior (SNIES), Sistema de Salud Pública (SIVIGILA), Departamento Nacional de Planeación (DNP), Sistema de Matrícula (SIMAT), Unidad de Víctimas - Red Nacional de Información, Departamento Administrativo Nacional de Estadística (DANE) e Instituto Nacional de Medicina Legal y Ciencias Forenses.
- Conteo de las víctimas. Fuente Fiscalía General de la Nación. Del año 2015 a 2019.
- Nombres y código de los departamentos y municipios de Colombia. Fuente Departamento Administrativo Nacional de Estadística (DANE).
- Consolidado de educación del departamento de Cundinamarca. Fuente dirección de infraestructura de datos espaciales y estadísticos, de la Secretaría de Planeación de la gobernación de Cundinamarca.

Por lo anterior, se considera parte del momento de comprensión de los datos, toda vez que se hizo un barrido de los conjuntos de datos, estableciendo de manera inicial, un diccionario de datos junto con el conjunto de tablas, que fueron utilizados en la segunda parte de la comprensión de los datos y que, para el presente proyecto de investigación, se ubica en la etapa de procesamiento y análisis de los datos.

3.2.4 Procesamiento y análisis de los datos

Esta etapa se subdivide en 4 fases, para atender cada una de las actividades necesarias para la obtención del modelo que permita predecir de la mejor forma la puntuación de los estudiantes.

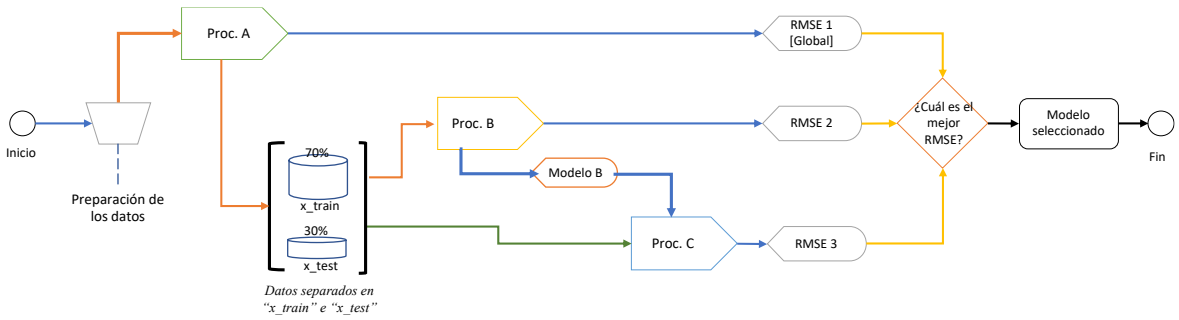


Figura 14. Etapas y flujo de procesamiento de la información.
Fuente. Elaboración propia.

Las fases son preparación de los datos, procesamiento A con 5 modelos, el procesamiento B, con el modelo multinivel y el procesamiento C, regresión multinivel gamma. La forma en cómo se procesa la información, se específica en la Figura 14. En primera instancia, se prepararon los datos, se procede al "procesamiento A" basado en los 5 modelos supervisados en el marco de Machine Learning. Terminada la fase A, se pasa a la fase B, teniendo como insumo la base de información organizada de la fase A, con el propósito de mantener los mismos datos y así poder comparar los modelos; el "procesamiento B", se soporta en el modelo multinivel o lineal jerárquico generalizado. En forma paralela se procede a la fase "procesamiento C", al igual que en la fase anterior, se apoya a en la base de datos resultado de la fase A, tomando como referencia el modelo multinivel gamma. Los modelos usados en las fases de procesamiento B y C están en el contexto de los modelos generalizados mixtos, ampliamente utilizados.

A continuación, se específica cada una de las fases con mayor detalle.

- *Fase preparación de los datos.*

Una vez se obtuvieron las bases de datos públicas, se procede a revisar los datos para establecer las variables que componen cada base de datos seleccionada. El proceso, un tanto laborioso, se inicia con la construcción de un diccionario de datos global, para en perspectiva, ver los datos y potenciales campos llaves entre ellos. A continuación, se revisan las tablas, para verificar que se ajustaran a los periodos en estudio, aquellos datos que no se ajustaban, se transformaron o adecuaron, para que el conjunto de datos quede bajo el mismo parámetro. Con los datos en forma de tablas en la hoja de cálculo, se procede a establecer los datos faltantes y el respectivo porcentaje, respecto a cada variable, con el propósito de definir si se descarta la variable, o se procede a imputar los datos faltantes. Se utiliza como criterio para imputar los datos de una variable o eliminar los registros con datos faltantes, como máximo el 20% de los datos faltantes.

Una vez se tuvieron los datos, se procede a ajustar el diccionario de datos, revisando nuevamente, que cada dato se ajuste al tipo de variable en el registro de datos (por ejemplo, que no queden datos tipo texto en datos que son numéricos).

Con las tablas se procede a diseñar el modelo entidad-relación, apoyados en Ms Access®, se efectuaron las consultas necesarias para el estudio, garantizando la integridad de los datos, así como evitando la redundancia.

- *Fase Procesamiento A "los 5 modelos"*

En esta fase de procesamiento se emplearon, por cada municipio, los cinco modelos provenientes de Machine Learning (Regresión lineal múltiple, Regresión Ridge (Ridge Regression), Regresión Lasso (Lasso Regression), Árboles de decisión (Decision Tree), Bosque Aleatorio (Random Forest)). En este primer procesamiento, lo que se espera encontrar por cada municipio, es un modelo de mejor RMSE, utilizando como base los modelos antes mencionados.

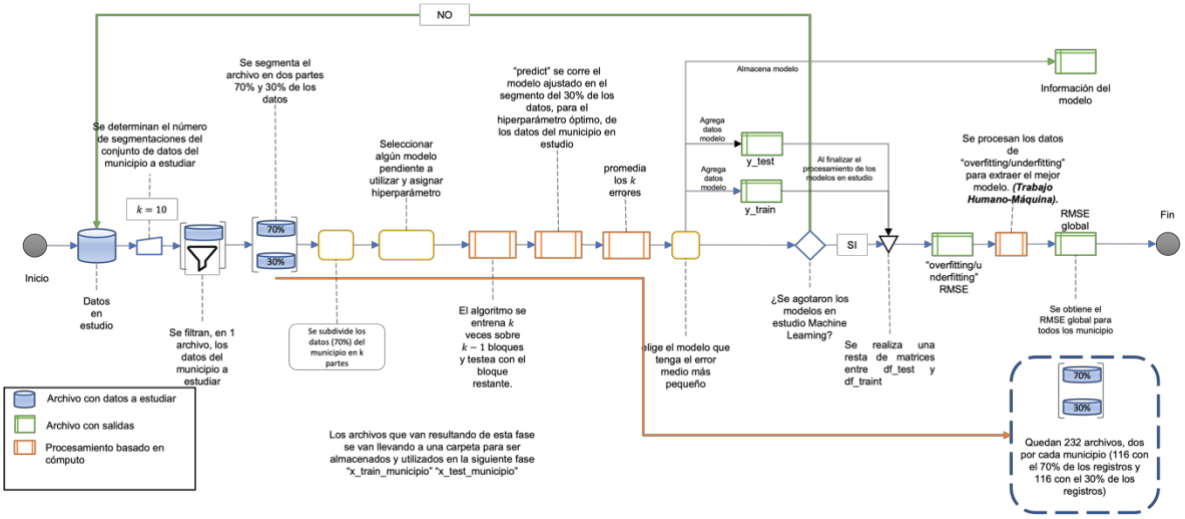


Figura 15. Proceso seguido para la obtención del RMSE global resultado del procesamiento de los datos en estudio con algoritmos de Machine Learning.
Fuente. Elaboración propia.

La Figura 15 describe el proceso seguido para determinar el modelo, en cada municipio, que predice las puntuaciones de los estudiantes en las pruebas Saber 11, a través de los algoritmos de Machine Learning, seleccionados para el caso en estudio.

Una de las decisiones, es establecer el número de segmentos del conjunto de datos del municipio que se va a estudiar, a juicio del equipo investigador fueron 10 ($k = 10$). Se tomaron los datos en estudio, segmentando el archivo solamente en los datos del municipio a estudiar. A continuación, el conjunto de datos seleccionado se subdivide e en dos partes: la primera con el 70% de la información utilizada para el entrenamiento ("train") del modelo y la segunda, con el 30% de la información utilizada para evaluar el modelo ("test"); dado que son 116 municipios, al final, quedan 232 archivos, la mitad de los cuales contenían el 70% de la información y la otra

mitad, la restante por cada municipio. Se toman los datos a utilizar para el entrenamiento del modelo y se subdivide en *k* partes (10).

Una vez seleccionado uno de los 5 modelos a utilizar, se asignan los hiperparámetros (Regresión lineal múltiple (sin hiperparámetro), Regresión Ridge (α), Regresión Lasso (α), Árboles de decisión (profundidad del árbol, criterio de división), Bosque Aleatorio (profundidad del árbol, criterio de división). Se procede entonces a procesar la información del archivo de entrenamiento, utilizando la librería scikit-learn en Python (Amat-Rodrigo, 2020), con el algoritmo escogido, seleccionando el parámetro óptimo. Posteriormente, se almacenan los modelos de los hiperparametros óptimos para cada municipio en forma independiente, de tal manera que se tuvieron 580 archivos.

Luego de seleccionar el modelo para el municipio, la información relacionada se almacena con el modelo, en un único archivo, de tal manera que cada vez que corre el algoritmo, se genere un nuevo registro. Con base en el modelo, se procede a modificar, agregando los datos derivados, en dos archivos denominados "*y_test*" e "*y_train*". La información del archivo "*y_test*" corresponde a una matriz, en la que las columnas son cinco, cada una de las cuales asignada a uno de los modelos y las filas, corresponden con la información de cada uno de los municipios; en cada una de las intersecciones, se almacena el valor RMSE que corresponde a cada municipio, con cada uno de los modelos. El archivo "*y_train*" se utiliza para almacenar, al igual que el archivo "*y_test*", la información relacionada con el RMSE, por cada municipio y por cada uno de los 5 modelos utilizados; es preciso indicar que, en los dos archivos, las columnas y filas se organizaron de la misma manera, es decir, que los modelos tienen el mismo orden en cada archivo y los municipios también conservan el mismo orden, para garantizar que en cada coordenada del vector, de cada matriz, está la información relacionada con el mismo modelo y el mismo municipio.

Una vez que se corrieron los cinco modelos en los 116 municipios, los archivos "*y_test*" y "*y_train*" tienen toda la información almacenada del ejercicio computacional, hecho hasta el momento. Con esos dos archivos se procede a hacer una resta, en valor absoluto, de las dos matrices que se representan en los archivos "*y_test*" y "*y_train*". Como resultado de la anterior operación, se verifica si existe overfitting o underfitting en cada modelo. Con este último archivo se efectua un trabajo de decisión "humano-máquina", que consiste en analizar los datos correspondientes a cada municipio, en los cinco modelos, para determinar cuál es el más óptimo. El criterio que se utiliza como factor de decisión, fue seleccionar el modelo con menor diferencia en el "*rmse(test)"* y "*rmse(train)",* en observancia de la capacidad de generalización. Con todo lo anterior, se procede a calcular el "RMSE global"; el cual debe ser tenido en cuenta en forma comparativa, para determinar el mejor modelo, una vez se desarrolla en la fase de procesamiento B y la fase de procesamiento C, finalizando de esta manera el procesamiento y apoyados en los algoritmos de Machine Learning.

Para avanzar a la siguiente fase de procesamiento B, se tomaron los 116 archivos que contienen cada uno el 70% de los datos del municipio, utilizados en esta fase anterior y se concatenaron,

para conformar un nuevo archivo con el 70% de los datos de todos los municipios en estudio. Del mismo modo, se toman los 116 archivos que contienen el 30% de cada uno de los datos relacionados con cada municipio y se concatenan, para conformar un archivo con el 30% del conjunto de datos en estudio.

Los anteriores dos archivos garantizan que los datos utilizados en la fase A para construir y evaluar el modelo, son los mismos que se utilizan en la fase de procesamiento B y C, con el propósito de obtener el modelo, usando el modelo multinivel y la regresión multinivel gamma.

- *Fase Preprocesamiento B "Modelo multinivel"*

Es necesario precisar que para la fase B, se garantiza que los datos en estudio corresponden con los mismos utilizados para la fase de procesamiento A. La Figura 16 muestra cada uno de los momentos que se tienen en cuenta para la fase de procesamiento B. Tomando como referencia el archivo con el 70% de los datos en estudio, se realiza el procesamiento, usando el paquete Ime4 (Bates et al., 2022), en el software R®, para aproximarse a encontrar el modelo a estudiar. Una vez hecho el procesamiento, se verifica que el proceso haya corrido, verificando nuevamente los parámetros, que se corren al menos dos veces, para establecer que las salidas sean las mismas. Una vez obtenido el modelo, se procede a seleccionar las variables más significativas, utilizando como criterio "*t value*".

Una vez seleccionadas las variables, estadísticamente significativas, del modelo multinivel, se determinan las variables asociadas, en cuatro casos: 1) colegio municipio, 2) colegio y variables asociadas a los municipios, 3) municipio y variables asociadas al colegio y 4) variables asociadas al colegio y variables asociadas al municipio. El resultado obtenido en cada uno de los procesamientos indicados anteriormente fue almacenado en un archivo, a partir del cual, se seleccionó a el mejor modelo, tomando como primer criterio el menor RMSE; en caso de coincidencia en varios RMSE, se debió utilizar como segundo criterio de decisión, el menor AIC.

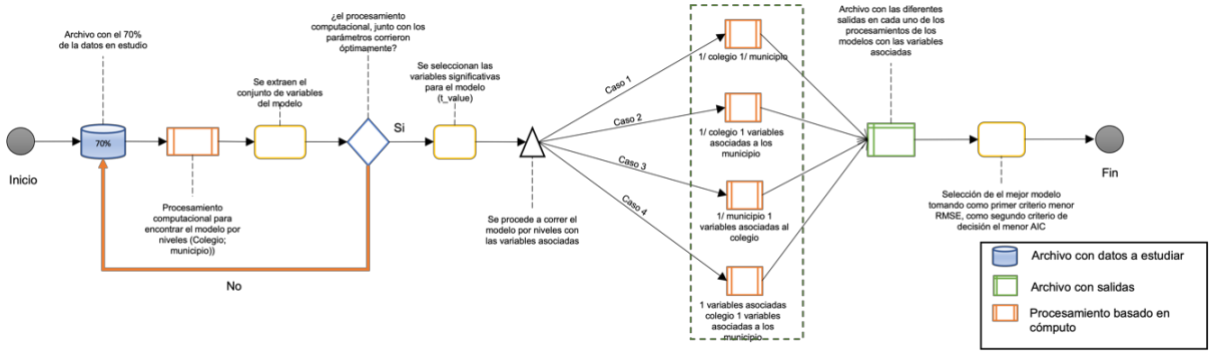


Figura 16. Momentos de la fase de procesamiento B para obtener el modelo multinivel.
Fuente. Elaboración propia.

El mejor modelo obtenido en la fase de procesamiento B, se guardó en un archivo, que se convirtió en el punto de partida para la fase de procesamiento C.

- *Fase Preprocesamiento C "Multinivel Gamma"*

La Figura 17 describe el proceso utilizado para la fase de procesamiento C, correspondiente al modelo de regresión multinivel gamma. Para este modelo, al igual que en la fase de procesamiento B, se parte del archivo que contiene el 70% de los datos en estudio, junto con el modelo que finalmente quedó en la fase de procesamiento B. A continuación, se utiliza el paquete Ime4(Bates et al., 2022), en el software R®, debidamente parametrizado, de acuerdo con el modelo de regresión multinivel gamma. El procesamiento se corre al menos dos veces, con el fin de verificar que las salidas fueran iguales, verificando los parámetros iniciales indicados en la parte anterior.

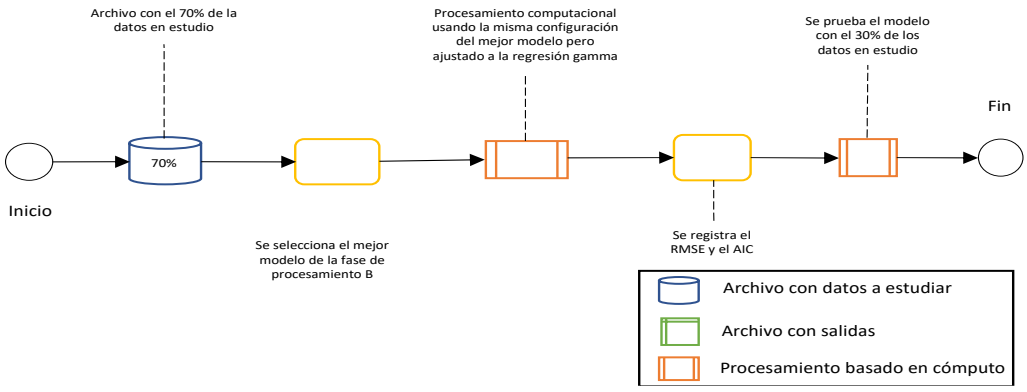


Figura 17. Diferentes momentos de la fase de procesamiento se para obtener el modelo de regresión multinivel gamma.
Fuente. Elaboración propia.

Una vez corrido el proceso, tal como se indica, se registró el RMSE y el AIC, para posteriormente, compararlos con los ya obtenidos. Tomando como referencia el modelo en esta fase de procesamiento C, nuevamente se corrió el modelo, con la información existente en el archivo con el 30% de los datos en estudio.

3.2.5 Escritura del informe de investigación

Esta etapa corresponde a la elaboración del informe del proyecto investigación, es decir, el presente documento.

- *Los datos de estudio*

Los datos obtenidos de los municipios del departamento de Cundinamarca – Colombia (N=116), en los periodos anunciados en número de registros son:

- Resultados pruebas SABER 11 del semestre 2015-1 al 2019-2. 188.707 observaciones.
- Indicadores de infancia y adolescencia del año 2015 al 2019. 586 observaciones.
- Conteo de las víctimas. Del año 2015 a 2019. 13.945 observaciones.
- Nombres y código de los departamentos y municipios de Colombia. 116 observaciones.
- Consolidado de educación del departamento de Cundinamarca. 580 observaciones.

- *Aspectos éticos*

El presente proyecto se apoya en fuentes de información pública, donde no aparecen registrados los nombres de las personas. Por lo anterior, se garantiza el anonimato de la información relacionada con los sujetos a la que le corresponde, en las diferentes fuentes de información, que hacen parte integral de este estudio, ajustándose a la normatividad vigente en Colombia (Congreso de la República de Colombia, 2012), así como a los parámetros que se exigen en las comunidades académicas a nivel internacional (Belmonte-Serrano, 2010).

4 Resultados

Este capítulo se centra en presentar los resultados obtenidos a lo largo de la investigación. Se organiza en cuatro segmentos: el primero se dedica a reportar todos los elementos iniciales en cuanto a la organización y procesamiento de los datos; el segundo, está orientado a reportar los resultados del procesamiento, con los algoritmos de Machine Learning; en el tercero, se ubican los resultados obtenidos al utilizar los modelos lineales generalizados y el cuarto, se centra en algunas reflexiones y análisis derivados de los dos anteriores segmentos.

4.1 Elementos iniciales

En esta sección se encuentran los resultados, así como los elementos básicos, en el proceso de alistamiento de los datos, para efecto de adelantar el procesamiento de la información con los modelos estadísticos seleccionados.

4.1.1 Hardware y software para el procesamiento

Para el procesamiento de la información, se tuvo acceso a un computador portátil marca Lenovo®, descritas en la Tabla 3.

Tabla 3. Características del ordenador usado en el procesamiento.

Parte	Descripción
Procesador	Ryzen 5
Pantalla	De 14" hasta FHD (1920x1080), antirreflejos, retroiluminación LED,
	220 nits, 16:9
Memoria	20 GB
Almacenamiento	SSD / SATA 6.0Gb/s, 2.5" ancho, 7mm alto
Tarjeta gráfica	AMD Radeon™
Sonido	Dolby Audio
Batería	Hasta 4,5 horas* con batería de 35 Wh
Dimensiones (An. × Pr. × Al.)	327,1 mm x 241 mm x 19,9 mm
Peso	A partir de 1,6 kg
Conectividad	802.11ac 1x1 Wi-Fi + Bluetooth 4.2

Fuente. El fabricante Lenovo® y manual del equipo.

El procesador Ryzen 5 tiene 4 núcleos de procesador y 8 subprocesos o hilos, con un reloj base de 3.4 GHz, cache total L1 (384Khz), L2 (2MB) y L3 (8MB), lanzado al mercado en noviembre de 2017. Presenta mejoras en el procesamiento respecto al procesador Intel Core 5 – 7400 en pruebas de juegos, creación de contenidos, realidad virtual, entre otras (https://www.amd.com/es/products/cpu/amd-ryzen-5-1400).

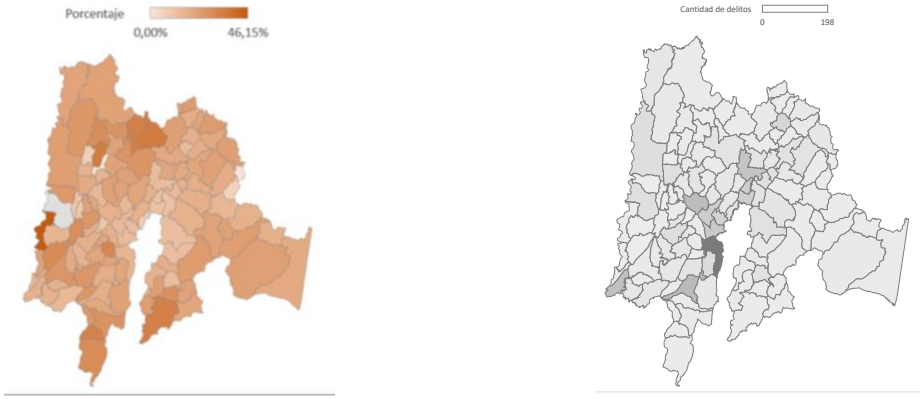
En relación con el software utilizado, el sistema operativo es Windows 11®, Microsoft Excel 2019®, R-4.1.1 para Windows (https://cran.r-project.org/bin/windows/base/), Python 3.9.7 (https://www.python.org/downloads/windows/), acceso a internet a través de Google Chrome ®, por banda ancha, vía fibra óptica.

Es preciso indicar que, el tiempo del equipo para el procesamiento de los datos en la fase de procesamiento A "los 5 modelos" - Machine Learning superó las 24 horas, en la fase de procesamiento B "Modelo multinivel", el tiempo fue de aproximadamente 3 horas en cada uno y, en la fase de procesamiento C "Multinivel gamma", el tiempo superó las 24 horas. Un hecho importante para tener en cuenta en el procesamiento es la necesidad de evitar que el equipo entre en estado invernadero, toda vez que detiene el proceso y requiere reinicio de toda la actividad, lo que puede ocasionar retrasos.

4.1.2 Entendimiento de los datos

En esta sección los datos fueron tomados tal cual, desde las bases de datos públicas ICFES, gobernación de Cundinamarca, Fiscalía General de la Nación, y otras fuentes indicadas en el apartado metodológico.

El departamento de Cundinamarca cuenta con 116 municipios, de los cuales siete concentran aproximadamente el 50% de la población. Los municipios son Soacha, Facatativá, Fusagasugá, Zipaquirá, Chía, Mosquera y Madrid. Estos siete municipios guardan una relación directa con la ciudad de Bogotá. No es objeto de análisis en este estudio, pero resultaría interesante en un siguiente estudio, aproximarse a evaluar y valorar la conexión multidimensional entre los 7 municipios y la ciudad de Bogotá.



(a) (b)

Figura 18. La figura (a) muestra el porcentaje de embarazos por municipio en población adolescente entre los 10 y 19 años, la figura (b) muestra la frecuencia de delitos, año 2018, por municipios.
Fuente. Elaboración propia con Ms Excel®.

La Figura 18, en su parte (a), identifica que, en el municipio de Beltrán, la tasa de embarazos supera el 45%, en jóvenes adolescentes con edades de 10 y 19 años. En su parte (b), muestra cómo el municipio de Soacha se configura como el de mayor cantidad de frecuencia de delitos en 2018.

La Figura 19 muestra la distribución de estudiantes que presentaron las pruebas Saber 11, por municipio, entre los años 2015 y 2019, así como la diferencia de los puntajes entre la naturaleza oficial y no oficial de las instituciones ($[Dif] = [promedio\ puntaje\ global\ no\ oficial] - [promedio\ de\ puntaje\ oficial]$). Los municipios que aparecen en la Figura 19, corresponden solamente a aquellos que tienen instituciones educativas oficiales y no oficiales, por lo anterior, no aparecen los municipios que solamente tienen un tipo de institución.

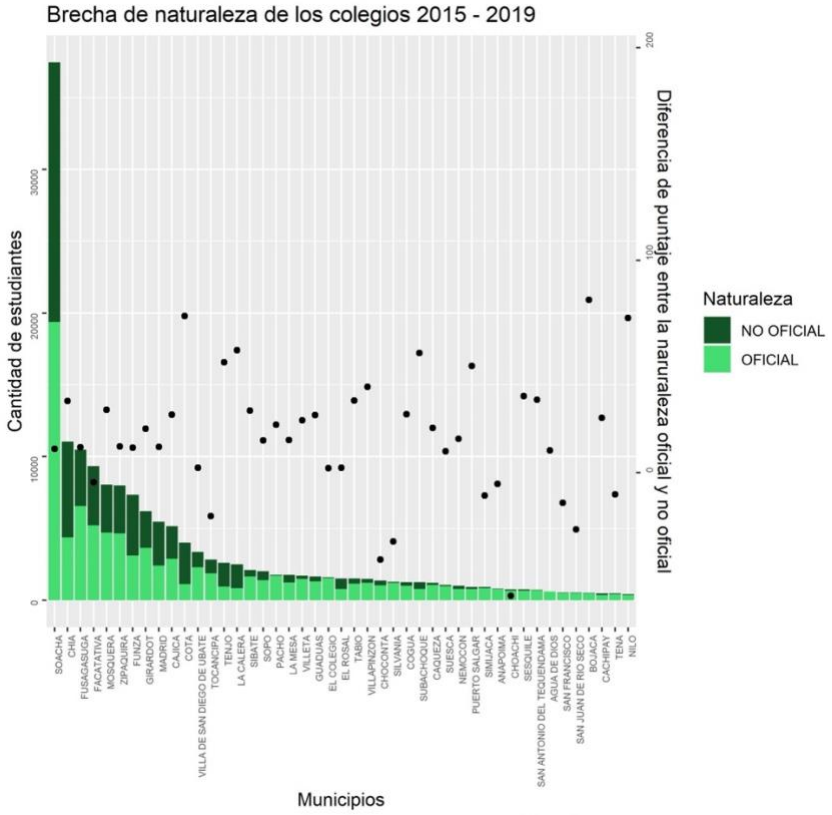


Figura 19. Distribución de estudiantes y la diferencia de los puntajes entre la naturaleza, oficial y no oficial, de las instituciones educativas.
Fuente. Elaboración propia con R®.

En esta figura, el municipio de Soacha se diferencia claramente de los demás municipios del departamento de Cundinamarca, en cuanto a la población que lo conforma y que presentan las pruebas Saber 11.

De lo anterior, junto con toda la información que se recolecta desde las diferentes fuentes, se logra identificar que las personas que presentan las pruebas Saber 11, están inmersos en un contexto social conformado por múltiples variables, provenientes de diferentes fuentes y de las cuales, algunas de ellas hacen parte de este estudio, toda vez que están cuantificadas y disponibles.

4.1.3 Alistamiento de los datos

En esta fase se preparan adecuadamente los datos, para aplicar las técnicas estadísticas y de Machine Learning. Esto implica seleccionarlos, añadir los existentes, imputarlos y darles el formato adecuado, para la fase de modelado.

Como se observa en la Tabla 4, de los datos de resultados de los estudiantes en las pruebas Saber 11, se revisó la distribución de las variables y se unieron los resultados de todo el periodo de estudio, en una hoja de cálculo.

Tabla 4. Distribución de variables del ICFES entre el periodo 2015-1 al 2019-2.

VARIABLES	Presencia de variables en cada periodo									
	2015-1	2015-2	2016-1	2016-2	2017-1	2017-2	2018-1	2018-2	2019-1	2019-2
ESTU_LIMITA_MOTRIZ	X	X				X	X	X		
ESTU_LIMITA_INVIDENTE	X	X				X	X			
FAMI_OCUPACIONPADRE	X	X	X	X						
FAMI_OCUPACIONMADRE	X	X	X	X						
FAMI_TRABAJOLABORPADRE					X	X	X	X	X	X
FAMI_TRABAJOLABORMADRE					X	X	X	X	X	X
FAMI_TIENEHORNOMICROOGAS					X	X	X	X	X	X
FAMI_TIENEMICROONDAS	X	X	X	X						
FAMI_TIENEHORNO	X	X	X	X						
ESTU_TRABAJAACTUALMENTE	X	X	X	X						
ESTU_RECIBESALARIO	X	X	X	X						
ESTU_TIPOREMUNERACION					X	X	X	X	X	X

Fuente: Elaboración propia a partir de los datos existentes en las bases de datos del ICFES.

Lo primero que se determina es el porcentaje de datos faltantes por variable, para lo cual se utiliza la función *apply*[1] en el software R®. A continuación, se eliminan las variables que tenían 80% o más de valores faltantes y se verifica cuáles variables podrían ser equiparables entre las existentes en las bases de datos, debido a que, en algunos casos, la información es comparable.

Una vez identificadas las variables equiparables, se procede a transformarlas, utilizando la hoja de cálculo, tomando como referencia las categorías del estudio, obteniéndose la Tabla 5. Finalmente, se eliminan todas las observaciones que contenían algún dato faltante y que representaban el 7,32%.

Tabla 5. Transformación de las variables del ICFES.

Variable para el estudio	Variables existentes en base de datos
FAMI_TELEVISOR ('Si', 'No') Esta acción redujo los valores faltantes al 2,722739 %.	FAMI_TIENETELEVISOR y FAMI_TIENESERVICIOTV. Se unieron, ya que se trasponen los años en los que está la información.

[1] La función utilizada se describe como apply (X = is.na(tabla), MARGIN = 2, FUN = mean).

Variable para el estudio	Variables existentes en base de datos
ESTU_TIENEREMUNERACION ('Si', 'No') Esta acción redujo los valores faltantes al 36,05431%.	Debido a que ESTU_RECIBESALARIO no se sabe si es efectivo o en especie, los niveles se cambiaron 'Si, en efectivo', 'Si, en especie' y 'Si, en efectivo y especie' a 'Si'. Finalmente, se unieron las variables ESTU_RECIBESALARIO y ESTU_TIPOREMUNERACION con el nombre ESTU_TIENEREMUNERACION.
ESTU_APOYOGOBIERNO ('NO', 'SI') Esta acción redujo los valores faltantes al 0,3916124 %.	ESTU_GENERACION-E y ESTU_PILOPAGA. Se unieron, ya que se trasponen los años en esta información. Esto debido a que un programa dejo de existir y el otro empezó a funcionar en ese mismo año. Finalmente, se unieron las variables con el nombre ESTU_APOYOGOBIERNO.
FAMI_TIENEMICROOHORNO ('Si', 'No') Esta acción redujo los valores faltantes al 1,265984%.	FAMI_TIENEHORNOMICROOGAS, FAMI_TIENEMICROONDAS y FAMI_TIENEHORNO. Se unieron, ya que se trasponen los años en los que está la información. Por la definición de las variables, se puso 'Si' en caso de que FAMI_TIENEMICROONDAS o FAMI_TIENEHORNO fuera 'Si' y 'No' en caso de que FAMI_TIENEMICROONDAS o FAMI_TIENEHORNO tuviera 'No'. Finalmente, se unieron las variables con el nombre FAMI_TIENEMICROOHORNO.
ESTU_HORASTRABAJO ('No', 'Si, menos de 20h a la semana', Si, 20h o más a la semana') Esta acción redujo los valores faltantes al 1,227299%.	Se unió ESTU_HORASSEMANATRABAJA con ESTU_TRABAJAACTUALMENTE de la siguiente manera: se cambió '0' por 'No'; cambiar 'menos de 10h' y 'entre 11 y 20h' por 'Si, menos de 20h a la semana'; cambiar 'entre 21 y 30h' y 'más de 30 h' por 'Si, 20h o más a la semana'.
FAMI_TRABAJAPADRE ('Empleado de nivel directivo', 'Empleado de nivel auxiliar o administrativo', 'Empleado de nivel técnico o profesional', 'Empleado obrero, conductor u operario', 'Empresario', 'Hogar, no trabaja o estudia', 'Otra actividad u ocupación', 'Pensionado', 'Pequeño empresario', 'Profesional independiente', 'Trabajador por cuenta propia', 'No sabe', 'No aplica'). Esta acción redujo los valores faltantes al 1,434499%.	'Empleado con cargo como director o gerente general' unir con 'Empleado de nivel directivo' de la variable FAMI_OCUPACIONPADRE.
	Empleado de nivel auxiliar o administrativo' de FAMI_OCUPACIONPADRE unir con 'Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente)' de FAMI_TRABAJOLABORPADRE.
	Empleado de nivel técnico o profesional' de FAMI_OCUPACIONPADRE unir con 'Trabaja como profesional (por ejemplo, médico, abogado, ingeniero)' de FAMI_TRABAJOLABORPADRE.
	Empleado obrero u operario' de FAMI_OCUPACIONPADRE unir con 'Es operario de máquinas o conduce vehículos (taxista, chofer)' y 'Trabaja como personal de limpieza, mantenimiento, seguridad o construcción' de FAMI_TRABAJOLABORPADRE.
	Empresario' de FAMI_OCUPACIONPADRE unir con 'Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial' de FAMI_TRABAJOLABORPADRE.
	Hogar' de FAMI_OCUPACIONPADRE unir con 'Trabaja en el hogar, no trabaja o estudia' de FAMI_TRABAJOLABORPADRE.
	Otra actividad u ocupación' de FAMI_OCUPACIONPADRE unir con 'Es vendedor o trabaja en atención al público' y 'Es agricultor, pesquero o jornalero' de FAMI_TRABAJOLABORPADRE.
	Pensionado' de FAMI_OCUPACIONPADRE unir con 'Pensionado' de FAMI_TRABAJOLABORPADRE.
	Pequeño empresario' de FAMI_OCUPACIONPADRE unir con 'Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo, tienda, papelería, etc.' de FAMI_TRABAJOLABORPADRE.
	Profesional independiente' de FAMI_OCUPACIONPADRE.
	Trabajador por cuenta propia' de FAMI_OCUPACIONPADRE unir con 'Trabaja por cuenta propia (por ejemplo, plomero, electricista)' de FAMI_TRABAJOLABORPADRE.
	No sabe' de FAMI_TRABAJOLABORPADRE.
	No aplica' de FAMI_TRABAJOLABORPADRE.
FAMI_TRABAJAMADRE ('Empleado de nivel directivo', 'Empleado de nivel auxiliar o administrativo', 'Empleado de nivel técnico o profesional', 'Empleado obrero, conductor u operario', 'Empresario', 'Hogar, no trabaja o estudia', 'Otra actividad u ocupación', 'Pensionado', 'Pequeño empresario', 'Profesional independiente', 'Trabajador por cuenta propia', 'No sabe', 'No aplica').	Empleado con cargo como director o gerente general' unir con 'Empleado de nivel directivo' de la variable FAMI_OCUPACIONMADRE.
	Empleado de nivel auxiliar o administrativo' de FAMI_OCUPACIONMADRE unir con 'Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente)' de FAMI_TRABAJOLABORMADRE.
	Empleado de nivel técnico o profesional' de FAMI_OCUPACIONMADRE unir con 'Trabaja como profesional (por ejemplo, médico, abogado, ingeniero)' de FAMI_TRABAJOLABORMADRE.
	Empleado obrero u operario' de FAMI_OCUPACIONMADRE unir con 'Es operario de máquinas o conduce vehículos (taxista, chofer)' y 'Trabaja como personal

Variable para el estudio	Variables existentes en base de datos
Esta acción redujo los valores faltantes al 1,313147%.	de limpieza, mantenimiento, seguridad o construcción' de FAMI_TRABAJOLABORMADRE.
	Empresario' de FAMI_OCUPACIONMADRE unir con 'Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial' de FAMI_TRABAJOLABORMADRE.
	Hogar' de FAMI_OCUPACIONMADRE unir con 'Trabaja en el hogar, no trabaja o estudia' de FAMI_TRABAJOLABORMADRE.
	Otra actividad u ocupación' de FAMI_OCUPACIONMADRE unir con 'Es vendedor o trabaja en atención al público' y 'Es agricultor, pesquero o jornalero' de FAMI_TRABAJOLABORMADRE.
	Pensionado' de FAMI_OCUPACIONMADRE unir con 'Pensionado' de FAMI_TRABAJOLABORMADRE.
	Pequeño empresario de FAMI_OCUPACIONMADRE unir con 'Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo, tienda, papelería, etc.' de FAMI_TRABAJOLABORMADRE.
	Profesional independiente' de FAMI_OCUPACIONMADRE.
	Trabajador por cuenta propia de FAMI_OCUPACIONMADRE unir con 'Trabaja por cuenta propia (por ejemplo, plomero, electricista)' de FAMI_TRABAJOLABORMADRE.
	No sabe' de FAMI_TRABAJOLABORMADRE.
	No aplica' de FAMI_TRABAJOLABORMADRE.

Fuente: Elaboración propia.

En cuanto a los indicadores de infancia y adolescencia, al igual que los datos anteriores, se determinó el porcentaje de datos faltantes por variable, se eliminaron las variables que tenían 80% o más de valores faltantes y las que se consideraron irrelevantes para el estudio, quedando distribuidos los datos faltantes como se muestra en la Figura 21.

Posteriormente, se aplicó el algoritmo K-Nearest Neighbors (KNN) para imputar los valores faltantes de los indicadores con $k = 2,3,4,5$. Con el fin de seleccionar la mejor imputación, se realizaron las gráficas de densidad para cada variable, en el conjunto de datos inicial y en cada uno de los conjuntos de datos imputados. De acuerdo con ello, se observó que, en general, las gráficas de densidad de las variables, con respecto a los k, fueron similares, excepto para la cobertura del acueducto (Figura 20). Por lo tanto, para realizar la imputación se eligió $k = 3$.

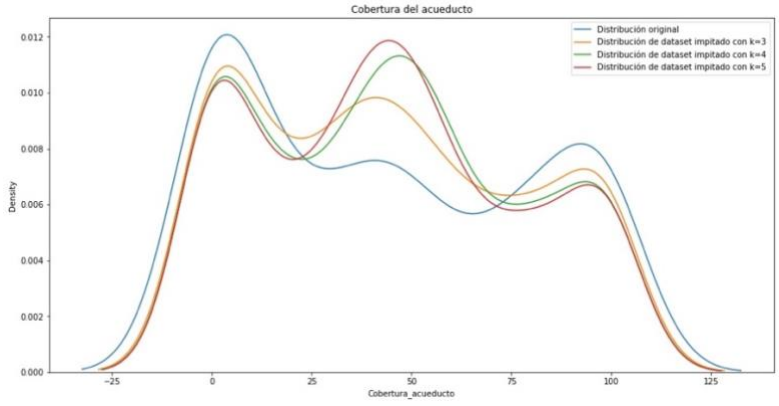


Figura 20. Gráfica de densidad de la cobertura del acueducto.
Fuente: Elaboración propia.

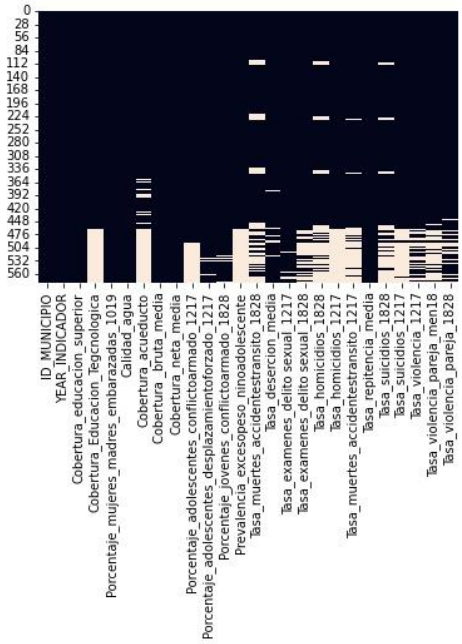


Figura 21. Datos faltantes de los indicadores de infancia y adolescencia.
Fuente: Elaboración propia.

De los resultados de las pruebas Saber 11 a nivel colegio, se seleccionan las variables consideradas como relevantes para el estudio y se completa la información de la ubicación de los colegios (rural, urbana), por medio de sistema de consulta de las instituciones educativas del país que toma los datos del Directorio Único de Establecimientos Educativos (DUE).

Por medio del registro de víctimas de la Fiscalía General de la Nación, se realizó un conteo de los casos registrados por municipio en Cundinamarca y por año del 2015 al 2019, para construir el conjunto de datos de delitos.

Finalmente, se eliminaron las variables de los indicadores que hacían referencia a personas mayores de 18 años y que correspondían a niveles de educación, diferente a la educación media, obteniendo las siguientes variables para cada tabla dentro de la base de datos.

- ID_MUNICIPIO: Código del DANE del municipio.
- NOM_MUNICIPIO: Nombre del municipio.
- YEAR: Año de presentación del SABER 11.
- ID_INSTITUCION_EDUCATIVA: Código DANE de las instituciones educativas.
- NOM_INSTITUCION_EDUCATIVA: Nombre de las instituciones educativas.
- INST_NATURALEZA: Oficial es la categoría para colegios públicos y no oficial es la categoría para colegios privados.

- INST_CALENDARIO: Calendario escolar de los colegios. El A empieza en el mes de febrero y finaliza actividades en el mes de noviembre y el calendario B, que empieza en el mes de septiembre y finaliza actividades en el mes de junio.
- INST_AREA_UBICACION: Zona en la que se encuentra los colegios, es decir, urbana o rural.
- MEDIAB: Cobertura bruta de la educación media por municipio.
- ESTU_GENERO: Genero del estudiante.
- ESTU_TIENEETNIA: Representa si el estudiante pertenece a una etnia o no.
- ESTU_HORASTRABAJO: Cantidad de horas que trabaja el estudiante (no; si, más de 20 horas; si, menos de 20 horas).
- ESTU_APOYOGOBIERNO: El estudiante recibió algún apoyo económico del estado una vez se graduó.
- ESTU_PRIVADO_LIBERTAD: Representa si el estudiante esta privado de su libertad.
- ESTU_INSE_INDIVIDUAL: Índice de Nivel Socioeconómico por estudiante.
- FAMI_EDUCACIONMADRE: Nivel educativo de la madre del estudiante.
- FAMI_TRABAJAMADRE: Ocupación de la madre del estudiante.
- FAMI_EDUCACIONPADRE: Nivel educativo del padre del estudiante.
- FAMI_TRABAJAPADRE: Ocupación de la madre del estudiante.
- FAMI_ESTRATOVIVIENDA: Estrato socioeconómico.
- FAMI_PERSONASHOGAR: Cantidad de personas en el hogar.
- FAMI_CUARTOSHOGAR: Número de cuartos en el hogar.
- FAMI_TIENEINTERNET: Acceso a internet del estudiante.
- FAMI_TIENECOMPUTADOR: Acceso a un computador del estudiante.
- FAMI_TIENELAVADORA: Acceso a una lavadora del estudiante.
- FAMI_TIENEAUTOMOVIL: Acceso a automóvil del estudiante.
- FAMI_TELEVISOR: Acceso a un televisor del estudiante.
- FAMI_TIENEMICROOHORNO: Acceso a microondas u horno del estudiante.
- FAMI_NUMLIBROS: Número de libros del estudiante.
- COBERTURA_EDUCACION_SUPERIOR: Cobertura de educación superior.
- COBERTURA_EDUCACION_TECNOLOGICA: Cobertura de educación tecnológica.
- PORCENTAJE_MUJERES_MADRES_EMBARAZADAS_1019: Porcentaje de mujeres embarazadas entre 10 a 19 años por municipio.
- CALIDAD_AGUA: Índice de la calidad del agua por municipio.
- COBERTURA_ACUEDUCTO: Cobertura del agua por municipio.
- COBERTURA_BRUTA_MEDIA: Cobertura bruta de la educación media por municipio.
- COBERTURA_NETA_MEDIA: Cobertura neta de la educación media por municipio.
- PORCENTAJE_ADOLESCENTES_CONFLICTOARMADO_1217: Porcentaje de adolescentes afectados por el conflicto armada entre 12 a 17 años por municipio.

- PORCENTAJE_ADOLESCENTES_DESPLAZAMIENTOFORZADO_1217: Porcentaje de adolescentes afectados por el desplazamiento forzado entre 12 a 17 años por municipio.
- PREVALENCIA_EXCESOPESO_NINOADOLESCENTE: Prevalencia de exceso de peso de niños y adolecientes por municipio.
- TASA_DESERCION_MEDIA: Tasa de deserción en la educación media por municipio.
- TASA_EXAMENES_DELITO SEXUAL_1217: Tasa por 100.000 habitantes con valoraciones de adolescentes de 12 a 17 años que se sospeche han sido víctimas de violencia sexual.
- TASA_HOMICIDIOS_1217: Tasa de homicidios entre los 12 y 17 años por municipio.
- TASA_MUERTES_ACCIDENTESTRANSITO_1217: Tasa de accidentes de tránsito entre los 12 y 17 años por municipio.
- TASA_REPITENCIA_MEDIA: Tasa de repitencia en la educación media por municipio.
- TASA_SUICIDIOS_1217: Tasa de suicidios entre los 12 y 17 años por municipio.
- TASA_VIOLENCIA_1217: Tasa de violencia entre los 12 y 17 años por municipio.
- CANTIDAD_DELITOS: Cantidad de delitos por municipio.
- PUNT_GLOBAL: Puntaje global del estudiante.

En relación con ESTU_APOYOGOBIERNO es una variable que tiene una doble mirada. De un lado puede ser interpretada como consecuencia de obtener buenos puntajes en el desempeño escolar de las pruebas Saber 11, pero también puede ser una variable de causa toda vez que puede constituirse en una forma motivacional extrínseca del estudiantado para obtener mejores desempeños y en consecuencia acceder al beneficio como se ha dicho.

Trabajos como los de Rodríguez-De-Souza-Pajuelo et al. (2021), Rebai et al. (2019), entre otros, han logrado establecer como la motivación, intrínseca o extrínseca, es un factor que interviene en la obtención de mejores desempeños escolares de las pruebas Saber 11, por lo que considerar las diferentes variables existentes en las bases de datos que de alguna manera se constituyen en un factor motivacional resulta siendo relevante para el presente estudio.

Por otro lado, la calidad de vida está asociada con las comodidades, entendido como el acceso al consumo de equipamientos y electrodomésticos (Conde Gutiérrez del Álamo, 2009), de manera que se incluyeron la lavadora, el microondas, el televisor, el automóvil, el internet, el computador y cantidad de libros para estudiar como la calidad de vida podría influir en el desempeño escolar de las pruebas Saber 11.

Una vez surtida toda la preparación de los datos descritas anteriormente, queda filtrada la base de datos a utilizar en el presente estudio. En la Tabla 6, se describe la cantidad de variables y de observaciones.

Tabla 6. Datos del estudio.

Entidad	Base de datos inicial		Base de datos después del alistamiento	
	Variables	Observaciones	Variables	Observaciones
DANE – Departamento	2	33	2	33
DANE - Municipios	3	1.120	3	1.120
ICFES - Estudiantes	158	188.707	39	174.693
ICFES – Colegio	20	2.997	7	859
DIVERSAS FUENTES – Indicadores*	11	586	26	580
FISCALIA	25	13.945	26	580
CONSOLIDADO DE EDUCACIÓN	15	580	14	580

* Esta base de datos implicó cambio de formato, en razón a que varias variables venían catalogadas en una misma que, al desagregarla, da el número correcto de variables, lo que implica no es que hayan crecido, sino que se organizaron adecuadamente.
Fuente. Elaboración propia.

Cada conjunto de datos provenientes de las diferentes fuentes tiene algunas variables que se necesitan repetidas, para poder hacer la consulta que integre los datos. Por ejemplo, se requieren los datos vinculados con cada uno de los años en estudio, así como la identificación clara del municipio, colegio o estudiantes, por lo que, para la consulta final, quedó un total de 49 variables y 174.693 observaciones.

4.1.4 El modelo entidad relación del conjunto de datos

Como resultado de la fase de preparación de datos, se obtuvieron 7 conjuntos de datos, que se integraron en una base de datos relacional, mediante el sistema de gestión de bases de datos Microsoft Access®. La Figura 22 identifica el modelo entidad relación, diseñado para el presente estudio. Las tablas son Tb_Departamentos, Tb_Municipios, Tb_Indicadores, Tb_Delitos, Tb_Cobertura_por_nivel, Tb_Institución_educativa y Tb_Estudiantes.

En la preparación de los datos, se tuvo especial atención en que las llaves de cada una de las tablas tuviesen el mismo formato, para garantizar la integridad de la información. La tabla Tb_Departamentos el ID_Departamento, tiene el mismo formato que ID_Departamento de la tabla Tb_Municipios, asimismo en las demás que se identifican en la Figura 22.

Además, en la Figura 22 se puede identificar que un departamento puede tener *n* municipios, un municipio puede tener *n* indicadores, *n* delitos, tener observaciones de cobertura por nivel, *n* instituciones educativas y en una institución pueden existir *n* estudiantes.

Cuando el modelo entidad relación, con los datos debidamente listados, se logra estructurar en el módulo de Microsoft Access®, se garantiza la integridad de la información y, en consecuencia, que el resultado de la consulta no va a tener datos perdidos.

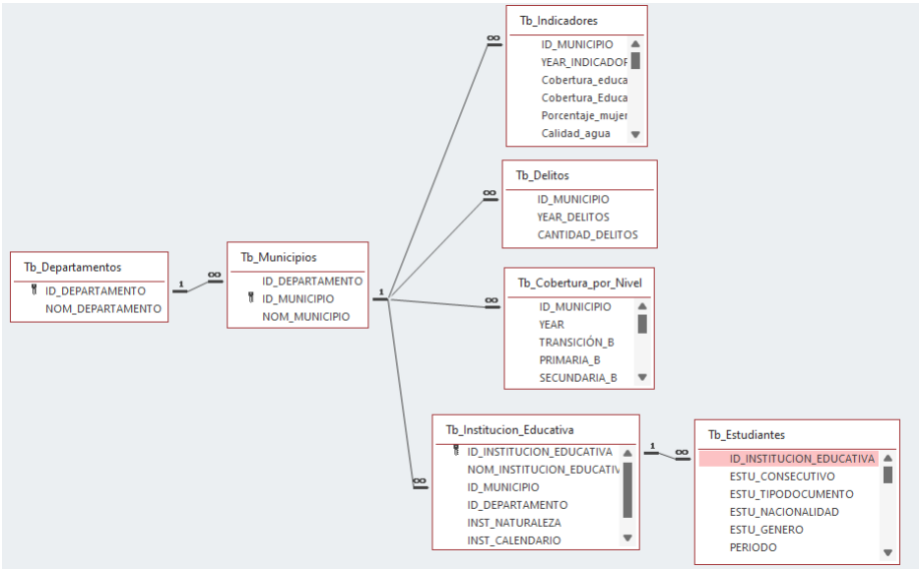


Figura 22. Base de datos relacional.
Fuente: Elaboración propia.

4.2 Procesamiento A "los 5 modelos"

Para esta actividad se empleó Python®, una herramienta de código abierto para análisis de datos, particularmente la librería Scikit-learn de Machine Learning, que incluye funciones para el ajuste de modelos, preprocesamiento de datos, selección de modelos, evaluación de modelos, entre otras.

A través del código en *Jupyter* presentado en el Anexo 1, para cada subconjunto de datos de las observaciones por municipio, se aplicaron las siguientes técnicas de Machine Learning:

A. *Regresión lineal múltiple*
   Se utiliza la función LinearRegression() cuyos parámetros son:
   - fit_intercept: Si calcula la intersección en este modelo. Por defecto es verdadero.
   - normalize: Si es verdadero, los regresores, covariables o variables independientes se normalizarán con Z-score antes de la regresión. Por defecto es falso.

B. *Regresión Ridge*
   Se utiliza la función Ridge() cuyos parámetros son:
   - alpha: hiperparámetro de regularización con el término de penalización L2, que controla la complejidad.
   - fit_intercept: Si calcula la intersección en este modelo. Por defecto es verdadero.
   - normalize: Si es verdadero, los regresores, covariables o variables independientes se normalizarán con Z-score, antes de la regresión. Por defecto es falso.

C. *Regresión Lasso*
   Se utiliza la función Lasso() cuyos parámetros son:

- alpha: hiperparámetro de regularización con el término de penalización L1, que controla la complejidad.
- fit_intercept: Si calcula la intersección en este modelo. Por defecto es verdadero.
- normalize: Si es verdadero, los regresores, covariables o variables independientes se normalizarán con Z-score, antes de la regresión. Por defecto es falso.

D. Decision Tree

Se utiliza la función DecisionTreeRegressor() cuyos parámetros son:

- criterion: función de FMSE y AE que mide la calidad de la división.
- max_depth: la máxima profundidad del árbol.

E. Random Forest

Se utiliza la función RandomForestRegressor() cuyos parámetros son:

- criterion: función de FMSE y AE que mide la calidad de la división.
- max_depth: la máxima profundidad del árbol

Adicionalmente, se guardaron los RMSE, los modelos y las divisiones en conjunto de entrenamiento y de prueba, siendo utilizados para la aplicación de las técnicas estadísticas de modelo multinivel y regresión multinivel gamma, para hacer los modelos comparables.

Para la elección de los hiperparámetros, se utilizó el método de "cross validation" mediante la métrica RMSE y R^2, de manera que se obtienen los mejores parámetros para cada modelo y municipio. Luego, se calcula la métrica de RSME, para los modelos del conjunto de entrenamiento y de prueba, como se muestra en las Tabla 7 y 8.

Tabla 7. RMSE del conjunto de prueba.

MUNICIPIO	Linear Regression	Ridge	Lasso	Decision Tree Regressor	Random Forest
25001	40,18	36,62	37,11	53,23	41,98
25019	6249085236385,87	32,58	33,26	46,75	34,15
25035	1852599219894,53	36,70	36,67	52,08	37,32
25040	3507584395410,43	36,58	36,57	54,09	38,43
25053	7946433895793,52	34,99	35,44	55,35	38,21
25086	187568130041377,00	34,46	34,65	45,74	32,67

Fuente: Elaboración propia.

Tabla 8. RMSE del conjunto de entrenamiento.

MUNICIPIO	Linear Regression	Ridge	Lasso	Decision Tree Regressor	Random Forest
25001	34,66	37,11	36,55	2,22	16,49
25019	29,62	34,51	33,21	1,43	13,83
25035	33,93	35,99	36,44	0	14,65
25040	35,48	36,36	36,80	0,23	14,79
25053	32,44	35,22	34,66	3,55	14,92
25086	21,33	33,64	34,141	0	14,07

Fuente: Elaboración propia

En consecuencia, se calcula el valor absoluto de la diferencia entre el RMSE del conjunto de prueba y, el RMSE del conjunto de entrenamiento (Tabla 9), para escoger el mejor modelo para cada municipio, con base en si los modelos presentan overfitting o underfiting y su capacidad de generalización.

Tabla 9. Overfitting y Underfitting.

MUNICIPIO	Linear Regression	Ridge	Lasso	Decision Tree Regressor	Random Forest
25001	5,52	0,49	0,55	51,01	25,49
25019	6249085236356,25	1,94	0,04	45,32	20,32
25035	1852599219860,59	0,71	0,23	52,08	22,67
25040	3507584395374,95	0,22	0,24	53,86	23,64
25053	7946433895761,07	0,24	0,78	51,80	23,29
25086	187568130041356,00	0,81	0,51	45,74	18,59

Fuente: Elaboración propia

Conforme a lo anterior, se seleccionan los mejores modelos, con sus hiperparámetros separados en cuatro grupos, según el volumen de estudiantes en cada municipio, obteniendo la Tabla 10, donde α aumenta, cuando la cantidad de estudiantes aumenta para cada municipio y el modelo Ridge Regression, tiene los valores más grandes de α.

Tabla 10. Mejores modelos por municipio con hiperparámetro.

Grupo	ID_MUNICIPIO	NOM_MUNICIPIO	N° estudiantes	Modelo	α
4	25095	BITUIMA	94	ridge	5000000000
4	25086	BELTRAN	128	ridge	352,74
4	25506	VENECIA	289	ridge	352,74
4	25779	SUSA	397	ridge	152,69
4	25489	NIMAIMA	243	ridge	115,51
4	25491	NOCAIMA	375	ridge	87,37
4	25019	ALBAN	315	ridge	66,09
4	25407	LENGUAZAQUE	419	ridge	66,09
4	25483	NARIÑO	123	ridge	66,09
4	25518	PAIME	227	ridge	66,09
4	25662	SAN JUAN DE RIO SECO	500	ridge	66,09
4	25777	SUPATA	338	ridge	66,09
4	25851	UTICA	236	ridge	66,09
4	25053	ARBELAEZ	569	ridge	50
3	25200	COGUA	1195	ridge	50
4	25328	GUAYABAL DE SIQUIMA	263	ridge	50
4	25436	MANTA	239	ridge	50
4	25524	PANDI	210	ridge	50
4	25596	QUIPILE	365	ridge	50
4	25612	RICAURTE	416	ridge	50
4	25745	SIMIJACA	870	ridge	50
4	25841	UBAQUE	382	ridge	50
4	25035	ANAPOIMA	772	ridge	37,82
4	25599	APULO	391	ridge	37,82
4	25154	CARMEN DE CARUPA	449	ridge	37,82
4	25178	CHIPAQUE	415	ridge	37,82
4	25317	GUACHETA	676	ridge	37,82
4	25335	GUAYABETAL	320	ridge	37,82
4	25339	GUTIERREZ	212	ridge	37,82
4	25535	PASCA	840	ridge	37,82
4	25658	SAN FRANCISCO	497	ridge	37,82

Grupo	ID_MUNICIPIO	NOM_MUNICIPIO	N° estudiantes	Modelo	α
3	25740	SIBATE	1963	ridge	37,82
3	25743	SILVANIA	1166	ridge	37,82
4	25781	SUTATAUSA	287	ridge	37,82
4	25862	VERGARA	280	ridge	37,82
4	25878	VIOTA	827	ridge	37,82
4	25001	AGUA DE DIOS	597	ridge	28,61
4	25120	CABRERA	327	ridge	28,61
3	25151	CAQUEZA	1153	ridge	28,61
4	25168	CHAGUANI	174	ridge	28,61
4	25293	GACHALA	302	ridge	28,61
4	25295	GACHANCIPA	580	ridge	28,61
3	25402	LA VEGA	1022	ridge	28,61
3	25486	NEMOCON	948	ridge	28,61
4	25580	PULI	112	ridge	28,61
4	25793	TAUSA	464	ridge	28,61
4	25797	TENA	448	ridge	28,61
4	25839	UBALA	577	ridge	28,61
4	25867	VIANI	237	ridge	28,61
4	25871	VILLAGOMEZ	141	ridge	28,61
3	25873	VILLAPINZON	1446	ridge	28,61
4	25040	ANOLAIMA	894	ridge	21,64
4	25099	BOJACA	469	ridge	21,64
4	25123	CACHIPAY	451	ridge	21,64
4	25181	CHOACHI	698	ridge	21,64
4	25224	CUCUNUBA	404	ridge	21,64
4	25258	EL PEÑON	231	ridge	21,64
4	25279	FOMEQUE	642	ridge	21,64
4	25312	GRANADA	496	ridge	21,6
4	25394	LA PALMA	719	ridge	21,64
4	25530	PARATEBUENO	415	ridge	21,64
4	25592	QUEBRADANEGRA	240	ridge	21,64
3	25875	VILLETA	1589	ridge	21,64
4	25885	YACOPI	614	ridge	21,64
3	25245	EL COLEGIO	1481	ridge	16,37
3	25322	GUASCA	1028	ridge	16,37
4	25324	GUATAQUI	125	ridge	16,37
4	25368	JERUSALEN	145	ridge	16,37
4	25438	MEDINA	384	ridge	16,37
3	25817	TOCANCIPA	2595	ridge	16,37
4	25297	GACHETA	554	ridge	12,38
4	25653	SAN CAYETANO	355	ridge	12,38
3	25758	SOPO	1867	ridge	12,38
4	25898	ZIPACON	299	ridge	12,38
2	25307	GIRARDOT	5771	ridge	9,37
2	25126	CAJICA	4533	ridge	7,08
3	25377	LA CALERA	2159	ridge	7,08
2	25473	MOSQUERA	7526	ridge	7,08
3	25799	TENJO	2451	ridge	7,08
2	25175	CHIA	9641	ridge	5,36
2	25290	FUSAGASUGA	9794	ridge	5,36
4	25807	TIBIRITA	152	lasso	4,19
2	25286	FUNZA	6893	ridge	4,05
1	25754	SOACHA	34808	ridge	3,07
4	25805	TIBACUY	217	lasso	2,14
4	25326	GUATAVITA	292	lasso	1,50
4	25426	MACHETA	330	lasso	1,41
4	25488	NILO	411	lasso	1,27
4	25281	FOSCA	401	lasso	1,18
4	25398	LA PEÑA	334	lasso	1,01
4	25845	UNE	456	lasso	0,83
4	25288	FUQUENE	431	lasso	0,81
4	25572	PUERTO SALGAR	884	lasso	0,56
4	25148	CAPARRAPI	511	lasso	0,56
4	25649	SAN BERNARDO	611	lasso	0,49

Grupo	ID_MUNICIPIO	NOM_MUNICIPIO	N° estudiantes	Modelo	α
4	25372	JUNIN	389	lasso	0,48
4	25645	SAN ANTONIO DEL TEQUENDAMA	684	lasso	0,46
4	25823	TOPAIPI	301	lasso	0,44
4	25718	SASAIMA	723	lasso	0,43
4	25815	TOCAIMA	686	lasso	0,432
4	25299	GAMA	190	lasso	0,41
4	25594	QUETAME	584	lasso	0,40
4	25736	SESQUILE	711	lasso	0,29
3	25260	EL ROSAL	1384	lasso	0,29
3	25772	SUESCA	1035	lasso	0,26
3	25183	CHOCONTA	1277	lasso	0,19
3	25785	TABIO	1433	lasso	0,18
3	25513	PACHO	1696	lasso	0,18
3	25320	GUADUAS	1564	lasso	0,16
3	25769	SUBACHOQUE	1220	lasso	0,13
3	25843	VILLA DE SAN DIEGO DE UBATE	3182	lasso	0,10
3	25386	LA MESA	1677	lasso	0,09
2	25214	COTA	3567	lasso	0,06
2	25430	MADRID	5144	lasso	0,05
2	25899	ZIPAQUIRA	7532	lasso	0,02
2	25269	FACATATIVA	8597	lasso	0,01

Fuente: Elaboración propia

Adicionalmente, en la Figura 23, se puede observar que los municipios están distribuidos, en su totalidad, en los modelos Ridge Regression y Lasso Regression.

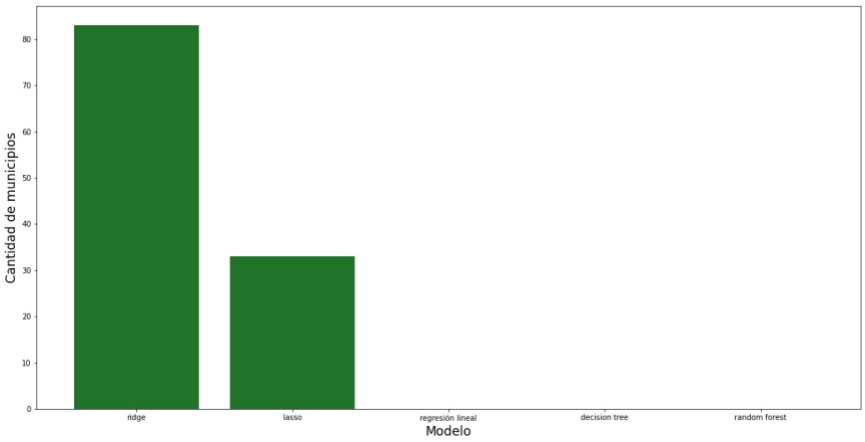


Figura 23. Cantidad de municipios asignados a cada modelo de Machine Learning.
Fuente: Elaboración propia

Luego, se unieron las predicciones de cada municipio, conforme al mejor modelo encontrado y se calculó el RMSE global, para el conjunto de entrenamiento, obteniendo un valor de 35,2399; en el conjunto de prueba, el valor de RMSE global es 36,6837.

4.3 Modelos generalizados

En esta sección se presentan los dos modelos generalizados que se adoptan en el estudio: el

modelo multinivel jerárquico o lineal generalizado y el segundo, que hace referencia al modelo de regresión multinivel gamma.

4.3.1 Modelo multinivel o lineal jerárquico generalizado

El modelo multinivel, se aplica para dos niveles (colegios, municipios), con todas las variables del conjunto de datos (modelo #1), posteriormente se eliminan las variables que se muestran en la Tabla 11, que no resultaron significativas, de acuerdo con el *t value*, para aplicar finalmente, el mismo modelo (modelo #2).

Tabla 11. Variables no significativas para modelo multinivel.

Variable	t value
CANTIDADDELITOS	-0,69
MEDIAB	-1,65
Coberturaeducacionsuperior	1,09
CoberturaEducacionTegcnologica	-0,76
Coberturaacueducto	-1,75
Coberturabrutamedia	-1,31
Tasahomicidios1217	1,14
Tasadesercionmedia	0,98
Tasarepitenciamedia	-0,11
Prevalenciaexcesopesoninoadolescente	0,30
Porcentajeadolescentesconflictoarmado1217	0,19
Tasamuertesaccidentestransito1217	1,79
Porcentajemujeresmadresembarazadas1019	1,31
Tasasuicidios1217	0,53
Tasaviolencia1217	-0,29
Tasaviolenciaparejamen18	-0,34
ESTUTIENEETNIASi	-1,88

Fuente: Elaboración propia

Posteriormente, se aplica el modelo multinivel de dos niveles (colegios, municipios), considerando que los resultados varíen por variables relacionadas con los colegios (modelo #3), y las variables relacionadas con los municipios (modelo #4). Por último, se examina el modelo de dos niveles (colegios, municipios), teniendo en cuenta las variables relacionadas con los colegios y municipios simultáneamente (modelo # 5). De todos los casos estudiados para el modelo multinivel, se obtuvo la Tabla 12, con el AIC y el RMSE, para el conjunto de entrenamiento y de prueba para cada modelo.

Tabla 12. Resultados RMSE de modelos multinivel.

Número de modelo	RMSE entrenamiento	RMSE prueba	AIC
1	35,9366	36,2575	1.223.827
2	35,9414	36,2594	1.223.798
3	35,9404	36,2572	1.223.715
4	35,9254	36,2524	1.223.793
5	35,9243	36,2497	1.223.714

Fuente: Elaboración propia

Según el menor AIC y los menores RMSE, el mejor modelo multinivel corresponde a los dos niveles (colegios, municipios), teniendo en cuenta las variables relacionadas con los colegios y municipios simultáneamente.

En la Tabla 12 se observa que el modelo cinco, es el que tiene el RMSE en relación con los otros modelos. En la Tabla 13, se ubican las variables que hacen parte del modelo multinivel 5.

Tabla 13. Variables del modelo #5 modelo multinivel.

Variable comparativa	**Efectos fijos**	
	Variables	**Estimate**
	(Intercept)	243,27
YEAR2015	YEAR2016	9,36
	YEAR2017	9,76
	YEAR2018	4,11
ESTUGENEROF	ESTUGENEROM	11,23
ESTUHORASTRABAJONo	ESTUHORASTRABAJOSi; 20 horas o más a la semana	-12,18
	ESTUHORASTRABAJOSi; menos de 20 horas a lasemana	-8,74
ESTUAPOYOGOBIERNOSNO	ESTUAPOYOGOBIERNOSI	6,49
	ESTUINSEINDIVIDUAL	6,59
FAMIEDUCACIONPADRE Educación profesional completa	FAMIEDUCACIONPADREEducación profesional incompleta	4,75
	FAMIEDUCACIONPADRENinguno	-10,27
	FAMIEDUCACIONPADRENo sabe	-2,44
	FAMIEDUCACIONPADREPostgrado	5,92
	FAMIEDUCACIONPADREPrimaria completa	-6,81
	FAMIEDUCACIONPADREPrimaria incompleta	-6,79
	FAMIEDUCACIONPADRESecundaria (Bachillerato)completa	-4,13
	FAMIEDUCACIONPADRESecundaria (Bachillerato)incompleta	-5,61
	FAMIEDUCACIONPADRETécnica o tecnológica completa	2,26
FAMIEDUCACIONMADREEducación profesional completa	FAMIEDUCACIONMADREEducación profesional incompleta	3,01
	FAMIEDUCACIONMADRENinguno	-15,22
	FAMIEDUCACIONMADRENo sabe	-7,78
	FAMIEDUCACIONMADREPostgrado	6,52
	FAMIEDUCACIONMADREPrimaria completa	-8,72
	FAMIEDUCACIONMADREPrimaria incompleta	-9,90
	FAMIEDUCACIONMADRESecundaria (Bachillerato)completa	-4,62
	FAMIEDUCACIONMADRESecundaria (Bachillerato)incompleta	-5,73
	FAMIEDUCACIONMADRETécnica o tecnológica completa	1,72
FAMITRABAJAMADREEmpleado de nivel auxiliar o administrativo	FAMITRABAJAMADREEmpleado obrero; conductoru operario	2,96
	FAMITRABAJAMADREEmpresario	-3,38
	FAMITRABAJAMADREHogar; no trabaja o estudia	2,19
	FAMITRABAJAMADRENo sabe	-4,18
	FAMITRABAJAMADREOtra actividad u ocupación	1,27
	FAMITRABAJAMADREPensionado	-3,60
	FAMITRABAJAMADREPequeño empresario	3,05
	FAMITRABAJAMADRETrabajador por cuenta propia	1,31
FAMITRABAJAPADREEmpleado de nivel auxiliar o administrativo	FAMITRABAJAPADREEmpleado de nivel técnico oprofesional	2,50
	FAMITRABAJAPADREEmpleado obrero; conductoru operario	3,61
	FAMITRABAJAPADREEmpresario	-1,99
	FAMITRABAJAPADREHogar; no trabaja o estudia	-1,75
	FAMITRABAJAPADRENo aplica	4,61
	FAMITRABAJAPADRENo sabe	3,90
	FAMITRABAJAPADREOtra actividad u ocupación	4,06
	FAMITRABAJAPADREPensionado	2,60
	FAMITRABAJAPADREPequeño empresario	5,95
	FAMITRABAJAPADREProfesional independiente	4,17
	FAMITRABAJAPADRETrabajador por cuenta propia	5,06
FAMIPERSONASHOGAR1 a 2	FAMIPERSONASHOGAR3 a 4	1,29
	FAMIPERSONASHOGAR9 o más	-1,91
	FAMICUARTOSHOGARCuatro	3,05

Variable comparativa	Efectos fijos	
	Variables	Estimate
	FAMICUARTOSHOGARDiez o más	-2,03
	FAMICUARTOSHOGARDos	8,32
	FAMICUARTOSHOGARTres	6,31
	FAMICUARTOSHOGARUno	5,80
FAMITIENEINTERNETNo	FAMITIENEINTERNETSi	-0,71
FAMITIENECOMPUTADORNo	FAMITIENECOMPUTADORSi	1,45
FAMITIENELAVADORANo	FAMITIENELAVADORASi	-1,55
FAMITIENEAUTOMOVILNo	FAMITIENEAUTOMOVILSi	-3,25
FAMITELEVISORNo	FAMITELEVISORSi	-3,33
FAMITIENEMICROOHORNONo	FAMITIENEMICROOHORNOSi	-3,16
FAMINUMLIBROS1 A 10 LIBROS	FAMINUMLIBROS11 A 25 LIBROS	6,04
	FAMINUMLIBROS26 A 100 LIBROS	11,89
	FAMINUMLIBROSMÁS DE 100 LIBROS	14,94

Fuente. Elaboración propia con R®

El modelo multinivel, descrito en la tabla anterior, se convierte en el más representativo del presente estudio, toda vez que retoma muchas de las variables que emergieron en la parte del procesamiento A, pero que, en su organización, ejecución y demás características tuvo una mayor simpleza. Este tipo de modelamiento se ha utilizado exitosamente, para la determinación de las variables que permiten comprender y explicar las puntuaciones en pruebas estandarizadas (Benito et al., 2014; Froiland & Oros, 2014; Lisboa- Bartholo & Da-Costa, 2016; Murillo & Carrillo, 2021; Schuth et al., 2017; Yopasá & Valbuena, 2019; Zhang & Campbell, 2014).

En el modelo cinco, aparece la variable género, en la que el hombre se ve más favorecido en las puntuaciones, con respecto a la mujer, hecho que coincide con estudios previos en Colombia (Orjuela, 2014), así como en algunas regiones de Brasil (Lisboa- Bartholo & Da-Costa, 2016), Asia (Abdul-Aziz et al., 2015; Ali et al., 2013; Guo et al., 2015; Kumari et al., 2018; Lau et al., 2019), Centroamérica (Chacón-Vargas & Roldán-Villalobos, 2021).

Los puntajes que obtienen los estudiantes en las pruebas estandarizadas Saber 11, tienen una influencia positiva, al relacionarlos con la educación finalizada, tanto de los padres como de las madres. Respecto a lo anterior, el hecho que la mujer no tenga una formación completa tiene una mayor influencia negativa, en los puntajes que logran los jóvenes estudiantes en las pruebas. Llama además la atención, que cuando la madre trabaja de manera independiente, como empresaria según los datos fuentes, su influencia en los resultados que logran los estudiantes, es negativa.

Cuando los estudiantes de grado 11, adicional a su estudio, trabajan, presentan una afectación negativa en las puntuaciones alcanzadas en las pruebas Saber 11; se observa que quienes trabajan menos de 20 horas por semana, tienen menor efecto negativo, que los que deben trabajar más de 20 horas semanales.

De otro lado, cuando las familias van incrementando el número de integrantes, va creciendo el efecto negativo sobre las puntuaciones alcanzadas. Lo anterior coincide con estudios previos (Chacón-Vargas & Roldán-Villalobos, 2021; Masci et al., 2018; Qiu & Wu, 2019), en los que la configuración familiar y su participación, afectan las puntuaciones obtenidas por el estudiantado.

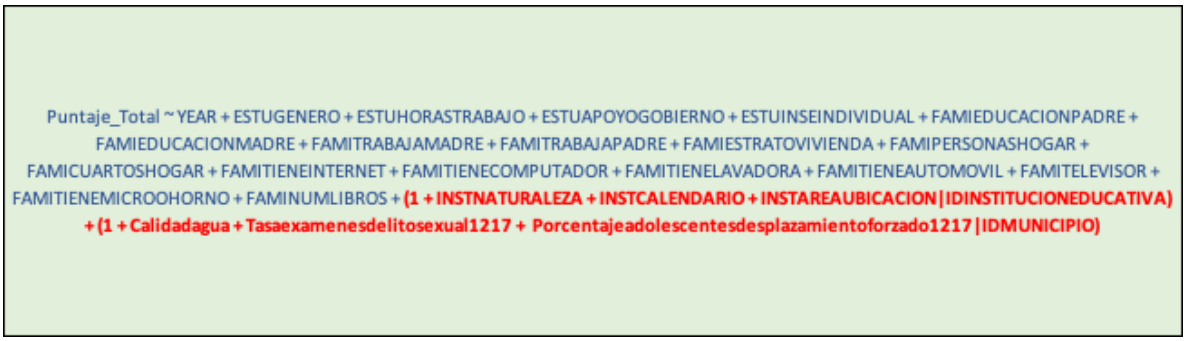


Figura 24. Resumen del modelo multinivel, en el que están las variables con efectos fijo (Azul) y las variables aleatorias (rojas).
Fuente. Elaboración propia con R®

La Figura 24 presenta a manera de síntesis, el modelo cinco, identificándose las variables con efecto fijo y con efecto aleatorio.

En resumen, en el departamento de Cundinamarca el puntaje global de la prueba Saber 11, es afectada negativamente cuando el estudiante trabaja, no tiene los recursos materiales[2] para su actividad educativa, el nivel educativo de los padres no está completo o es nulo, la madre trabaja como empresaria y el género del estudiante es femenino.

4.3.2 Regresión multinivel gamma

El modelo de regresión multinivel gamma se construye con base en el mejor modelo encontrado en la sección 4.3.1., es decir, con dos niveles (colegios, municipios) y teniendo en cuenta las variables relacionadas con los colegios y municipios simultáneamente. Por consiguiente, se obtiene la Tabla 14, donde se muestra el AIC y el RMSE para el conjunto de entrenamiento y el de prueba.

Tabla 14. Resultados del modelo de regresión multinivel gamma.

RMSE entrenamiento	RMSE prueba	AIC
36,00263	36,30028	1.225.149

Fuente. Elaboración propia con R®.

El multinivel gamma tiene un RMSE equiparable con el modelo multinivel, dado que, al correrlo con las variables obtenidas en el modelo multinivel, se presentan cambios al descrito en la sección anterior, lo que ratifica el modelo multinivel de la sección anterior, como el más representativo en el presente estudio.

[2] Entiéndase por recursos materiales como el conjunto de variables presentes en el estudio que al parecer mejoran las condiciones del ambiente de las personas (microondas, televisión, computador, lavadora, automóvil, número de libros e internet).

4.4 Algunos casos en el estudio en el modelamiento Machine Learning

En esta sección se toman 4 municipios, en los que aplica el modelamiento mediante Machine Learning, con el objeto de analizarlos a partir de las variables obtenidas en los resultados. Los municipios Soacha, Cajicá, Sibaté y Gachancipá, fueron seleccionados, tomando como criterio el número de estudiantes. En las siguientes subsecciones se describe cada uno de los casos.

4.4.1 El municipio de Soacha

Figura 25. División política del departamento de Cundinamarca en el que se señalan los cuatro municipios seleccionados para el análisis.
Fuente. Ingeominas.

El municipio de Soacha cuenta con uno de los mayores presupuestos a nivel nacional, está conexo con la zona urbana de la ciudad de Bogotá (Figura 25), tiene un área de 184 km^2 (menos del 1% de Cundinamarca), es el municipio con mayor número de estudiantes en el departamento y es el municipio, no ciudad capital, que más recepciona migrantes[3], de origen nacional y extranjero, con mayores necesidades de alimentación, salud, vivienda entre otras.

[3] Infografía de refugiados inmigrantes venezolanos en Bogotá y región al 1 de enero de 2021. Dirección electrónica https://www.r4v.info consultada el 10/06/2022.

Es el municipio que tiene la mayor población en Cundinamarca, en el 2018 tenía aproximadamente 660.179 habitantes, para el 2021 la población ascendió a 782.632 habitantes, es decir, tuvo un incremento ligeramente por debajo del 19%. Su tasa de crecimiento absoluta promedio simple, fue de 20.144 habitantes nuevos por año, lo que implica una tasa de crecimiento promedio anual simple del 5,1%[4]. La población rural es menos del 1%, de modo que el municipio tiene un carácter predominantemente urbano. La población urbana alcanzó el 99,21%, de acuerdo con el censo 2018. El gran número de esa población obedece a migración. Según el DANE, en el 2018 el número de hogares era de 210.423, de los cuales 29.670 están en Índice de Pobreza Multidimensional (IPM); la proporción de personas con Necesidades Básicas Insatisfechas (NBI) es de 5.3%. El número de embarazos de adolescentes está en el orden del 15,83%, con la mayor frecuencia de delitos en el departamento de Cundinamarca.

En cuanto al modelo resultado del presente estudio, en el municipio de Soacha se puede considerar que si aumenta la cantidad de delitos reportados, el porcentaje de adolescentes afectados por el conflicto armado como lo prevé el plan de desarrollo de Cundinamarca (Gobernación de Cundinamarca - Nicolás García-Bustos, 2020), el porcentaje de adolescentes afectados por el desplazamiento forzado, la tasa de delitos sexuales, la tasa de homicidio y tasa de violencia puntaje global en las pruebas Saber 11, disminuye. Así mismo, si la ubicación del colegio es rural, la naturaleza del colegio es no oficial, el género del estudiante es femenino, el estudiante no tiene acceso a recursos materiales y los padres no cuentan con un nivel de educación superior, también se ven afectados negativamente los resultados del desempeño académico de las pruebas Saber 11.

De otro lado, si la ubicación del colegio es urbana, la naturaleza del colegio es oficial, el género del estudiante es masculino, el estudiante tiene acceso a recursos materiales y los padres cuentan con un nivel de educación superior, se ven afectados positivamente los resultados de desempeño académico de las pruebas Saber 11.

En el modelo se logra identificar cómo la migración, por las diferentes causas como el conflicto armado o el desplazamiento forzado, por ejemplo, afectan el desempeño académico de las pruebas Saber 11. Las variables mencionadas, son diferenciales en el municipio, como se indicó en su caracterización, las cuales, al parecer, se constituyen en una causa que afecta socialmente los resultados en educación.

En relación con la incidencia del género, ya había sido reportado previamente por Orjuela (2014), Ariza et al. (2021), quienes utilizaron un modelo de regresión, para revisar la asociación estadística entre el acceso a las tecnologías a nivel individual y el rendimiento académico; en el análisis, los autores señalan que la interacción entre el computador e Internet, al parecer, sugieren una relación positiva entre los mayores puntajes y el hecho de tener un ordenador en casa. En

[4] Ésta tasa, del 5.1% para Soacha, es mucho más alta que la de las regiones o países con alto crecimiento económico y poblacional (Según datos del Banco Mundial, China, por ejemplo, creció en 2018 al 0,5%; México, al 1,1%; Brasil, al 0,8%, y Colombia, al 1,5 %.).

ese mismo sentido, es preciso indicar que en Europa (Shah et al., 2019), África (Qazdar et al., 2019) y Asia (Guo et al., 2015), también se ha encontrado que el género es una variable que incide en los resultados de pruebas estandarizadas, con estudiantes de educación media.

4.4.2 El municipio de Cajicá

El municipio de Cajicá (Figura 25), fundado en el año 1537, está en la *provincia centro* del departamento de Cundinamarca, tiene una extensión aproximada de 51 km^2. Recientemente ha tenido una expansión urbana, en razón al desplazamiento de población que trabaja en la ciudad de Bogotá, asimismo, algunas industrias se han desplazado a zonas cercanas. Está ubicado 17 km al norte, de la zona urbana del distrito capital, es uno de los municipios más habitados (82.244 personas en 2018, con un crecimiento de aproximadamente 17,5%) de la provincia (junto con Chía y Zipaquirá), según lo reporta el DANE.

El municipio tiene un total de 26.416 hogares, de los cuales 1.796 están en los IPM. La proporción de personas con NBI es de 3,1%. La frecuencia relacionada con los delitos, lo ubica al nivel de Zipaquirá y Chía, con una frecuencia media. El porcentaje de embarazos en adolescentes, cuyas edades están entre 10 y 19 años, es de 11,62%.

En este municipio, de acuerdo con el modelo resultado de este estudio, si la prevalencia de exceso de peso en niños y adolescentes, la tasa de deserción, la tasa de violencia y la cantidad de delitos aumenta, el puntaje global disminuye y si la ubicación del colegio es rural, la naturaleza del colegio es no oficial, el género del estudiante es femenino, el estudiante trabaja, el estudiante no tiene acceso a recursos materiales, la calidad del agua es baja y los padres tienen un nivel de educación bajo, afecta negativamente el desempeño académico de las pruebas Saber 11.

En este municipio se identifica cómo podría afectar negativamente el exceso de peso de la niñez y los adolescentes, en relación con los resultados globales en las pruebas Saber 11. El índice de masa corporal, con los temas de la obesidad, ha venido siendo objeto de estudio y recomendaciones, por parte Ministerio de Salud en Colombia (Ministerio de Salud y Protección Social, 2005). En un estudio cuantitativo apoyado en un modelo probit ($h_i = \alpha + X'\beta + \varepsilon$) que estima la variable salud obesidad (h_i), a partir de un vector matricial (X), establece que los hombres con mayores posibilidades económicas tienen mayor índice de obesidad. De hecho, Acosta (2013), señala que el tema de la obesidad es pandémico, que no es espontáneo, que éste va surgiendo gradualmente y puede tener diversas consecuencias. Para el caso en estudio, se ha identificado cómo la obesidad, en el caso del municipio de Cajicá, se comporta como una variable que afecta los resultados de los adolescentes en las pruebas Saber 11.

De otro lado, al igual que en el municipio de Soacha, se identifica cómo el nivel educativo de los padres afecta proporcionablemente los resultados de los estudiantes en las pruebas saber 11. Este hecho, coincide con resultados encontrados en diversos estudios en América del Sur (Lisboa- Bartholo & Da-Costa, 2016; Martínez-Mateus & TurriagoHoyos, 2015; Murillo & Carrillo, 2021; Orjuela, 2014; Rodríguez-Hernández et al., 2021; Yopasá & Valbuena, 2019),

Centroamérica (Chacón-Vargas & Roldán-Villalobos, 2021; Chaparro-Caso-López et al., 2016), América del Norte (Benito et al., 2014), Asia (Abdul-Aziz et al., 2015; Ali et al., 2013; Guo et al., 2015; Maulida & Kariyam, 2017; Salal & Abdullaev, 2020; Zhang & Campbell, 2014), Europa (Hanin & Van-Nieuwenhoven, 2016; Shah et al., 2019) y Oceanía (Masci et al., 2018).

4.4.3 El municipio de Sibaté

El municipio de Sibaté fue fundado del 24 de noviembre de 1967, posee un área de 126 km², pertenece a la provincia de Soacha, tiene una temperatura promedio de 14°C. Para el año 2018, lo habitaban aproximadamente 32.803 personas, que conformaban un total de 10.819 hogares, de los cuales 1.461 estaban en el IPM. Las personas con NBI es del orden de 1.947.

Acorde al modelo, si la prevalencia de exceso de peso en niños y adolescentes, la tasa de repitencia, el porcentaje de adolescentes afectados por el desplazamiento forzado, la tasa de delitos sexuales, la tasa de homicidio y tasa de violencia aumenta, el puntaje global disminuye; de otro lado, si la ubicación del colegio es rural, la naturaleza del colegio es oficial, el género del estudiante es femenino, el estudiante no tiene acceso a recursos materiales y padres con un nivel de educación bajo, afecta negativamente el desempeño académico de las pruebas Saber 11.

En este municipio se presenta una variable similar a la del municipio de Cajicá, relacionadas con el sobrepeso, del mismo modo, aparece una variable vinculada con el municipio de Soacha "desplazamiento forzado". Quiere decir esto que, el desplazamiento o migración obligatoria de las personas, afecta negativamente las puntuaciones totales que obtienen los adolescentes que se presentan en las pruebas Saber 11. La variable relacionada con la educación de padres aparece en este municipio, al igual que en Cajicá y Soacha. Llama la atención que la ubicación de la institución educativa en la zona rural, afecta negativamente el puntaje obtenido por los adolescentes; lo anterior puede ser debido a que en la temporada de cosechas, los estudiantes deben participar de la actividad agrícola, casi de manera obligatoria, junto con la familia, lo que implica el ausentismo en la institución educativa, además de las otras situaciones que se vivencian en el campo, como lo reporta Hernández Bonilla para el espectador en 2018[5].

En el aspecto relacionado con la escuela rural, se precisa indicar que la brecha en relación con las instituciones educativas citadinas es muy alta, lo que se profundiza aún más, con la falta de recursos, que como ya se advierte en el modelo presentado en este estudio, afecta en forma negativa los resultados en las pruebas saber 11.

[5] La difícil situación de las escuelas rurales en Colombia. Dirección electrónica https://www.elespectador.com/colombia-20/conflicto/la-dificil-situacion-de-las-escuelas-rurales-en-colombia-article/ consultada el 10/06/2022.

4.4.4 El municipio de Gachancipá

El municipio de Gachancipá, que hace parte de la provincia Sabana centro, fue fundado el 5 de septiembre de 1612. Tiene un área aproximada de 44 km², una temperatura promedio de 44°C, está ubicada a una altura de 2568m sobre el nivel del mar. Limita con el Nemocón, Guatavita, Sesquilé y Tocancipá. Basa su economía en la agricultura y en la ganadería, recientemente por su cercanía a la ciudad de Bogotá, contó con la creación de la zona industrial, convirtiéndose éste en un nuevo eje de desarrollo.

Para el 2018, el municipio estaba conformado por 5.401 hogares con una población de 16.633 personas. De los 5.401 hogares, 600 están en los niveles de IPM. La proporción de personas con NBI es de 4,6%. En el municipio Gachancipá se presenta una tasa de embarazos de 20,54%, en adolescentes cuyas edades están entre 10 y 19 años.

De acuerdo con el modelamiento, la prevalencia de exceso de peso en niños y adolescentes, el porcentaje de mujeres embarazadas, la tasa de delitos sexuales y la tasa de violencia, afecta negativamente el puntaje global; adicionalmente, si la naturaleza del colegio es oficial, el género del estudiante es femenino y el estudiante tiene trabajo, afecta negativamente el desempeño académico de las pruebas Saber 11.

4.5 A manera de cierre

Después de abortar los datos a través de los procesamientos A, B y C descritos en el capítulo de la metodología, se identifica que el procesamiento A y B guardan similitud entre sí. Sin embargo, el procesamiento B "Modelo multinivel", tiene una mayor simplicidad en su organización, tiempo de procesamiento y demás características para la información objeto de este estudio.

Ratificar como variable representativa en los modelos, el género del joven estudiante que presenta las pruebas Saber 11; si la persona trabaja, también tienen un efecto negativo; asimismo, cómo afecta el que sus padres hayan terminado algún estudio pos secundario (técnico, tecnológico o superior). Una información empírica que emerge del estudio es que, si la madre trabaja como empresaria, tiene un efecto negativo sobre las puntuaciones que el estudiante alcanza en las pruebas saber 11.

De manera adicional, resulta ser relevante para continuar las indagaciones, el sobrepeso de los jóvenes adolescentes y su efecto negativo en las puntuaciones que alcanzan en las pruebas Saber 11. Si bien, hay estudios que indican que el sobrepeso afecta multidimensionalmente a las personas, en este estudio, al parecer, da indicios de que el sobrepeso puede estar afectando negativamente el desempeño académico de las pruebas Saber 11 de los estudiantes, en las instituciones educativas de educación media.

5 Conclusiones, recomendaciones y trabajos futuros

Este apartado presenta las conclusiones a las que se llega en la investigación, en coherencia con la pregunta que direccionó la investigación, los objetivos y los resultados. Esta sección se organiza en dos aportados: el primero referido a las conclusiones, el segundo con algunas reflexiones derivadas de la investigación y el tercero destinado a mirar en prospectiva el potencial de trabajos futuros.

5.1 Conclusiones

Para la presente investigación se seleccionaron cinco modelos estadísticos vinculados a Machine Learning (Regresión lineal), Regresión Ridge (Ridge regression), Regresión Lasso (Lasso regression), Árboles de decisión (Decision tree), Bosque Aleatorio (Random Forest), y dos modelos vinculados a los modelos de regresión generalizados, a saber: Modelo de regresión multinivel y multinivel gamma. Los datos abiertos a los que su tuvo acceso, se procesaron en tres formas (1) Modelamiento Machine Learning [RMSE global de 36,68370], (2) modelo de regresión multinivel [RMSE global de 36,24970], y (3) modelo de regresión multinivel gamma [RMSE global de 36,00263]; encontrándose diferencias relativas porcentuales absolutas de menos del 2,00% entre los diferentes RMSE. Adicionalmente, la mayor complejidad en el procesamiento computacional está en el procesamiento relacionado con el modelamiento Machine Learning, la cual demandó cerca de ocho veces más el tiempo que el procesamiento para el modelo multinivel, entre otros aspectos.

Por lo anterior, para esta investigación, utilizar el modelo de regresión multinivel resulta ser el más adecuado por su simplicidad y comportamiento similar a los otros modelos en relación con sus resultados, específicamente con el RSME y el AIC. Esta conclusión coincide con la revisión adelantada en la presente investigación, en la que este modelamiento es el más recurrente en las publicaciones consultadas.

La investigación se orientó a identificar las variables que inciden en el desempeño escolar de las pruebas de Saber 11 de estudiantes de educación media en los años 2015 a 2019 del departamento de Cundinamarca-Colombia. Los resultados muestran que las variables que tienen efecto fijos en la predicción del desempeño escolar son el género de estudiante, las horas del trabajo del estudiante, el apoyo del estado a la familia, si los padres tienen estudios terminados, si la madre trabaja, así como los recursos disponibles en la casa. En algunos municipios emergió como una variable la obesidad del estudiantado como una variable opuesta al desempeño escolar.

Con lo anterior, no solamente se han identificado las variables predictoras, con efectos fijo, del desempeño escolar de las pruebas de Saber 11 sino que emergen, como dato empírico, variables que requieren mayor estudio para su comprensión y evaluación del impacto, como la obesidad.

Finalmente, en relación con proceso metodológico resulta plausible indicar la pertinencia de la metodología Cross Industry Standard Process for Data Mining (CRISP_DM) para el alistamiento, procesamiento de los datos y en general para garantizar un adecuado cuidado en cuanto al Volumen, Variedad, Veracidad y Valor de los datos. La Metodología CRISP_DM permite al investigador en Ciencia de Datos tener un primer orden para abordar el estudio.

5.2 Recomendaciones

En cuanto al procesamiento de los datos, se requiere señalar que disponer de un ordenador con un buen procesador, un disco de estado sólido, quizá con tarjeta de vídeo independiente y un buen tamaño de memoria RAM, propicia tiempos de procesamiento con menores fallos y en consecuencia mayor rapidez. Para el presente estudio se utilizó un equipo con 24 Gb de memoria RAM, lo que permitió al equipo procesar los datos sin fallos, excepto cuando el equipo entró en modo reposo.

Para la gobernación de Cundinamarca, así como para los gobiernos locales a nivel municipal, propiciar políticas y programas educativos orientados a favorecer la igualdad de género en los colegios puede incrementar la puntuación en el desempeño escolar de las pruebas de Saber 11 del estudiantado en general, toda que son las mujeres las que se ven más afectadas negativamente sus puntuaciones.

A nivel departamental, promover la alfabetización en educación hacia la educación media y a terminar nivel superior de los padres y madres, o la población adulta (que potencialmente serán padres), permite prever que sus hijos e hijas jóvenes, en un futuro mediano y largo plazo, que presenten las pruebas Saber 11 obtendrán mejores desempeños en las pruebas Saber 11.

Diseñar y ejecutar programas complementarios para mitigar las Necesidades Básicas Insatisfechas, así como los Niveles de Pobreza Multinivel redundará en que las personas que egresen de la educación media y presenten las pruebas Saber 11 obtengan mejores resultados en las pruebas Saber 11.

Al parecer la obesidad será una variable, que no solo afecte la salud de los jóvenes, sino que va a incidir negativamente en el desempeño escolar de las pruebas de Saber 11 de la población estudiantil que presente las pruebas Saber 11, por lo anterior anticipar, contener, mitigar y quizá evitar que esto suceda puede ser relevante, para lo que generar políticas, programas y demás estrategias multi-institucionales a diferentes niveles podría resultar plausible.

5.1 Trabajos futuros

Se requiere avanzar en nuevas investigaciones que continúen comparando diversas posibilidades de modelamiento estadístico, propiciando criterios que hagan apuestas a la eficiencia, eficacia en pro de obtener información fiable y oportuna para el sector que lo requiera.

De otro lado, se necesitan nuevas investigaciones que profundicen en la comprensión y explicación, no solamente el desempeño escolar de las pruebas de Saber 11, sino la calidad de la educación. En ese mismo sentido, replicar este estudio en los diferentes departamentos del país, permite tener una cartografía de las variables incidentes en la obtención del desempeño escolar de las pruebas de Saber 11. Como una necesidad adicional a estudiar es determinar lo ocurrido en los años de la pandemia debido a la COVID-19.

Bibliografía

Abdul-Aziz, A., Hafieza-Ismail, N., Ahmad, F., & Hassan, H. (2015). A framework for students' academic performance analysis using naïve bayes classifier. *Jurnal Teknologi*, *75*(3), 13–19. https://doi.org/10.11113/jt.v75.5037

Acevedo-Álvarez, R. (2008). *Los modelos jerárquicos lineales: fundamentos básicos para su uso y aplicación*. Universidad de Costa Rica.

Acosta, K. (2013). La obesidad y su consentracion segun nivel socioeconomico en Colombia. *Revista Econoíma Del Rosario*, *170*(2), 171–200.

Aggarwal, C. C. (2015). An Introduction to Data Mining. In *Data Mining* (pp. 1–26). Springer. https://doi.org/10.1007/978-3-319-14142-8

Ali, S., Haider, Z., Munir, F., Khan, H., & Awais, A. (2013). Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus. *American Journal of Educational Research*, *1*(8), 283–289. https://doi.org/10.12691/education-1-8-3

Amat-Rodrigo, J. (2020). *Machine learning con Python y Scikit-learn*. Ciencia de Datos Punto Net. https://www.cienciadedatos.net/documentos/py06_machine_learning_python_scikitlearn.html

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

Bates, D., Maechler, M., Bolker, B., Walker, S., Haubo, R., Christensen, B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., & Green, P. (2022). *Linear Mixed-Effects Models using "Eigen" and S4 - Package 'lme4'* (B. Bolker, Ed.). CRAN.

Belmonte-Serrano, M. Á. (2010). Requisitos éticos en los proyectos de investigación. Otra oveja negra. *Seminarios de La Fundacion Espanola de Reumatologia*, *11*(1), 7–13. https://doi.org/10.1016/j.semreu.2009.09.005

Benito, R., Alegre, M., & Gonzàlez-Balletbò, I. (2014). School Segregation and Its Effects on Educational Equality and Efficiency in 16 OECD Comprehensive School Systems. *Comparative Education Review*, *58*(1), 104–134. https://doi.org/10.1086/672011

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees* (1st ed.). https://doi.org/10.1201/9781315139470

Castrillón, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación Universitaria*, *13*(1), 93–102. https://doi.org/10.4067/s0718-50062020000100093

Chacón-Vargas, É., & Roldán-Villalobos, G. (2021). Factores que inciden sobre el rendimiento académico de los estudiantes de primer ingreso del curso Matemática General del Instituto Tecnológico de Costa Rica. *Uniciencia*, *35*(1), 265–283. https://doi.org/10.15359/ru.35-1.16

Chaparro-Caso-López, A., González-Barbera, C., & Caso-Niebla, J. (2016). Familia y rendimiento académico: configuración de perfiles estudiantiles en secundaria. *Revista Electronica de Investigacion Educativa*, *18*(1), 53–68.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.O step-by-step data mining guide*. NCR System Engineering Copenhagen, DaimlerChrysler AG, SPSS and Verzekeringen en Bank Groep B.V.

Comisión Económica para América Latina y el Caribe. (2016). Desarrollo Social Inclusivo, una nueva generación de políticas para superar la pobreza y reducir la desigualdad en América Latina y Caribe. *Cepal*, 304.

Conde Gutiérrez del Álamo, F. (2009). Análisis sociológico del sistema de discursos. Centro de Investigaciones Sociológicas.

Congreso de la República de Colombia. (2012). Ley 1581 de 2012. *Por La Cual Se Dictan Disposiciones Generales Para La Protección de Datos Personales*.

Contreras, D., Delgadillo, J., & Riveros, G. (2019). Is home overcrowding a significant factor in children's academic performance? Evidence from Latin America. *International Journal of Educational Development*, *67*, 1–17. https://doi.org/10.1016/j.ijedudev.2019.01.006

Cornell-Farrow, S., & Garrard, R. (2020). Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia. *Communications in Statistics Case Studies Data Analysis and Applications*, *6*(2), 228–246. https://doi.org/10.1080/23737484.2020.1752849

Correa Morales, J. C., & Salazar Uribe, J. C. (2016). *Introducción a los modelos mixtos* (1st ed.). Universidad Nacional de Colombia.

Cvencek, D., Fryberg, S., Covarrubias, R., & Meltzoff, A. (2017). Self-Concepts, Self-Esteem, and Academic Achievement of Minority and Majority North American Elementary School Children. *Child Development*, *89*(4), 1099–1109. https://doi.org/10.1111/cdev.12802

Dagnew, A. (2017). The relationship between students' attitudes towards school, values of education, achievement motivation and academic achievement in gondar secondary schools, Ethiopia. *Research in Pedagogy*, *7*(1), 30–42. https://doi.org/10.17810/2015.46

Díaz-Cárdenas, S., Arrieta-Vergara, K., & Simancas-Pallares, M. (2019). Adicción a Internet y rendimiento académico de estudiantes de odontología. *Revista Colombiana de Psiquiatria*, *48*(4), 198–207. https://doi.org/10.1016/j.rcp.2018.03.002

Ertel, W. (2017). Introduction to Artificial Intelligence. In *Predictive Toxicology* (Segunda Ed). Springer. https://doi.org/10.1007/978-3-319-58487-4 Library

Espinosa-Zúñiga, J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería Investigación y Tecnología*, *XXI*(1), 1–17. https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Febro, J. (2019). Utilizing feature selection in identifying predicting factors of student retention. *International Journal of Advanced Computer Science and Applications*, *10*(9), 269–274. https://doi.org/10.14569/ijacsa.2019.0100934

Froiland, J., & Oros, E. (2014). Intrinsic motivation, perceived competence and classroom engagement as longitudinal predictors of adolescent reading achievement. *Educational Psychology*, *34*(2), 119–132. https://doi.org/10.1080/01443410.2013.822964

Gaete-Rivas, D., Olea, M., Meléndez-Illanes, L., Granfeldt, G., Sáez, K., Zapata-lamana, R., & Cigarroa, I. (2021). Hábitos alimentarios y rendimiento académico en escolares chilenos de quinto a octavo año básico. *Revista Chilena de Nutrición*, *48*(1), 41–50. https://doi.org/10.4067/S0717-75182021000100041

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multinivel/Hierarchical Models*. Cambridge University Press.

George-Reyes, C. (2020). Pruebas Estandarizadas Y Calidad De La Educacion En México. *Universidad y Sociedad Revista Científica de La Universidad de Cienfuegos*, *12*(4), 418–425.

Géron, A. (2017). Hands-on Machine Learning with Scikit-Learn and TensorFlow. In N. Tache (Ed.), *O'Reilly Media, Inc* (1st ed., Vol. 53, Issue 9). O'Reilly Media.

Giannakas, F., Troussas, C., Voyiatzis, I., & Sgouropoulou, C. (2021). A deep learning classification framework for early prediction of team-based academic performance. *Applied Soft Computing*, *106*. https://doi.org/10.1016/j.asoc.2021.107355

Gobernación de Cundinamarca. (2020). *Plan de Desarrollo Departamental*.

Gobernación de Cundinamarca - Nicolás García-Bustos. (2020). *Plan departamental de desarrollo 2020-2024 Cundinamarca región que progresa*. Gobernación de Cundinamarca.

González-Correa, A., & Hernández-Ramírez, M. (2017). Diferencias entre los nivels de ansiedad en estudiantes de pregrado de Ingeniería de la universidad de Antioquia, 2017. *Séptima Conferencia Latinoaméricana Sobre El Abandono En La Educación Superior*.

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting Students Performance in Educational Data Mining. *International Symposium on Educational Technology*, 125–128. https://doi.org/10.1109/ISET.2015.33

Hanin, V., & Van-Nieuwenhoven, C. (2016). The influence of motivational and emotional factors in mathematical learning in secondary education. *Revue Europeenne de Psychologie Appliquee*, *66*(3), 127–138. https://doi.org/10.1016/j.erap.2016.04.006

Hasan, R., Palaniappan, S., Rafiez-Abdul, A., Mahmood, S., & Uddin-Sarker, K. (2018). Student Academic Performance Prediction by using Decision Tree Algorithm. *4th International Conference on Computer and Information Sciences (ICCOINS)*, 1–5. https://doi.org/10.1109/ICCOINS.2018.8510600

Hernández-Sampieri, R., Fernández-Collado, C., & Baptista-Lucio, M. (2014). *Metodología de la investigación* (J. Mares-Chacon, Ed.). McGraw Hill Intereamericana Editores S.A.

Heumann, C., Schomaker, M., & Shalabh. (2016). Linear Regression. In *Introduction to Statistics and Data Analysis* (pp. 249–295). https://doi.org/10.1007/978-3-319-46162-5

Hoffmann, J. (2010). Linear Regression Analysis: Applications and Assumptions. *Brigham Young University, Provo*, 1–285.

Ibourk, A., & Amaghouss, J. (2016). Convergence éducative et déterminants socioéconomiques: Analyse spatiale sur des données marocaines. *Mondes En Developpement*, *176*(4), 93–116. https://doi.org/10.3917/med.176.0093

ICFES. (2017). Clasificación de establecimientos y sedes. *Mineducación*.

Instituto Colombiano para la Evaluación de la Educación (ICFES). (2021). *Informe nacional de resultados del examen Saber 11° 2020*.

Jovanović, J., Saqr, M., Joksimović, S., & Gašević, D. (2021). Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success. *Computers and Education*, *172*(April), 1–13. https://doi.org/10.1016/j.compedu.2021.104251

Khan, A., & Ghosh, S. (2018). Data mining based analysis to explore the effect of teaching on student performance. *Education and Information Technologies*, *23*(4), 1677–1697. https://doi.org/10.1007/s10639-017-9685-z

Kumari, P., Jain, P., & Pamula, R. (2018). An Efficient use of Ensemble Methods to Predict Students Academic Performance. *4th Int'l Conf. on Recent Advances in Information Technology*. https://doi.org/10.1109/RAIT.2018.8389056

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, *1*(982). https://doi.org/10.1007/s42452-019-0884-7

Lisboa- Bartholo, T., & Da-Costa, M. (2016). Evidence of a school composition effect in Rio de Janeiro public schools. *Ensaio*, *24*(92), 498–521. https://doi.org/10.1590/S0104-40362016000300001

López-Vargas, O., Hederich-Martínez, C., & Camargo-Uribe, A. (2011). Estilo cognitivo y logro académico. *Educación y Educadores*, *14*(1), 67–82.

López-Vargas, O., Hederich-Martínez, C., & Camargo-Uribe, A. (2012). Logro de aprendizaje en ambientes hipermediales: Andamiaje autorregulador y estilo cognitivo. *Revista Latinoamericana de Psicologia*, *44*(2), 13–25. https://doi.org/10.14349/rlp.v44i2.1028

López-Vera, C., Vargas-Peñaloza, M., Gómez-Rodriguez, F., Rico-Marin, J., & Escandón-Wittsack, J. (2020). *Informe Nacional de resultados del examen Saber 11 2020* (M. Bravo-Osorio & P. Cifuentes-Velasquez, Eds.). Instituto Colombiano para el Fomento de la Educación Superior ICFES.

Lorna, A., Donders, G., Simon, E., López, O., Madden, C., Morrone, A., Puddephatt, A., Throsby, D., & Wagner, A. (2014). *Indicadores de cultura para el desarrollo. Manual metodológico. Patrimonio, Relevancia de la dimensión para la cultura y el desarrollo* (G. A. y M. Medici, Ed.). Organización de las Naciones Unidas para la Educación.

Maisarah-Samsudin, N., Milleana-Shaharudin, S., Filza-Sulaiman, N., Mohd-Fuad, M., Fareezuan-Zulfikri, M., & Hila-Zainuddin, N. (2021). Modeling student's academic performance during Covid-19 based on classification in support vector machine. *Turkish Journal of Computer and Mathematics Education*, *12*(5), 1798–1804. https://doi.org/10.17762/turcomat.v12i5.2190

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). Data Mining and Knowledge Discovery in Real Life Applications. In J. Ponce & A. Karahoca (Eds.), *IntechOpen* (Issue February, p. 436). I-Tech.

Martínez-Mateus, W., & TurriagoHoyos, Á. (2015). Análisis de distribución geográfica y espacial de los resultadosde las Pruebas Saber 11 del Instituto Colombiano para el Fomento de la Educación Superior -ICFES-. 2005-2012. Colombia. *Cuadernos Latinoamericanos de Administración*, *11*(21), 39–50. https://doi.org/10.18270/cuaderlam.v11i21.1618

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Masci, C., Johnes, G., & Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, *269*(3), 1072–1085. https://doi.org/10.1016/j.ejor.2018.02.031

Mato-Vázquez, V., & Muñoz-Cantero, M. (2010). Efectos Generales De Las Variables Actitud Y Ansiedad Sobre El Rendimiento En Matemáticas En Alumnos De Attitude and Anxiety Towards Mathematics and Student Achievement. *Ciencias Psicologicas*, *IV*(1), 27–40.

Maulida, J., & Kariyam. (2017). Students academic performance based on behavior. *AIP Conference Proceedings*, *1911*(December 2017). https://doi.org/10.1063/1.5016003

Mineshita, Y., Kim, H., Chijiki, H., Nanba, T., Shinto, T., Furuhashi, S., Oneda, S., Kuwahara, M., Suwama, A., & Shibata, S. (2021). Screen time duration and timing: effects on obesity, physical activity, dry eyes, and learning ability in elementary school children. *BMC Public Health*, *21*(422). https://doi.org/10.1186/s12889-021-10484-7

Ministerio de Educación Nacional de Colombia. (2019). Decreto 1330 de 2019. *Ministerio de Educación Nacional*, 32.

Ministerio de Salud y Protección Social. (2005). *Guía de atención de la obesidad*. 15.

Moine, J. Mi., Haedo, A., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. *XIII Workshop de Investigadores En Ciencias de La Computación*, 278–281.

Molina, A., Pérez, M., Castaño, N., Bustos, E., Suárez, O., & Sánchez, M. (2012). Mapeamiento informacional bibliográfico en el campo de la enseñanza de las ciencias, contexto y diversidad cultural: el caso del Journal Cultural Studies in Science Education (CSSE). *Revista EDUCyT*, *Extraordin*, 1997–222.

Molina, A., Pérez, R., Bustos, E., Castaño, C., Suaréz, O., & Sánchez, M. (2013). Mapeamento informacional bibliográfico de enfoques e campos temáticos da diversidade cultural : o caso dos journal CSSE , Sci . Edu . e Sci & Campos Temáticos de la diversidad cultural : el caso de las. *Atas Do IX Encontro Nacional de Pesquisa Em Educação Em Ciências – IX ENPEC*, 1–8.

Montagud-Mascarell, M. D., & Gandía-Cabedo, J. L. (2014). Virtual learning environment and academic outcomes: Empirical evidence for the teaching of Management Accounting. *Revista de Contabilidad-Spanish Accounting Review*, *17*(2), 108–115. https://doi.org/10.1016/j.rcsar.2013.08.003

Murillo, J., & Carrillo, S. (2021). Incidencia de la Segregación Escolar por Nivel Socioeconómico en el Rendimiento Académico. Un Estudio desde Perú. *Archivos Analíticos de Políticas Educativas*, *29*(49), 3–11. https://doi.org/10.14507/epaa.29.5129

Orjuela, J. (2014). Análisis del Desempeño Estudiantil en las Pruebas de Estado para Educación Media en Colombia mediante Modelos Jerárquicos Lineales. *Ingeniería*, *18*(2). https://doi.org/10.14483/udistrital.jour.reving.2013.2.a04

Paula, G. (2013). Modelos de regressão com apoio computacional. In *Universidade de São Paulo*. Universidade de São Paulo.

Peláez-Valencia, L. E., Trefftz, H., & Delgado-González, I. A. (2020). Acreditación Internacional de Carreras de Ingeniería. *Educación En Ingeniería*, *15*(29), 28–33. https://doi.org//dx.doi.org/10.26507/rei.v15n29.1044

Pollak, M., & Parnell, D. (2018). An Interdisciplinary Analysis of Course Meeting Frequency, Attendance and Performance. *Journal of the Scholarship of Teaching and Learning*, *18*(3), 132–152. https://doi.org/10.14434/josotl.v18i3.23752

Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies*, *24*(6), 3577–3589. https://doi.org/10.1007/s10639-019-09946-8

Qiu, X., & Wu, S. sheng. (2019). Contextual variables of student math proficiency and their geographic variations in Missouri. *Applied Geography*, *109*, 102040. https://doi.org/10.1016/j.apgeog.2019.102040

Rebai, S., Ben Yahia, F., & Essid, H. (2019). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, *70*(August 2018), 100724. https://doi.org/10.1016/j.seps.2019.06.009

Reimers, F. (2000). Educación , desigualdad y opciones de política en América Latina en el siglo XXI. *Revista Latinoamericana de Estudios Educativos*, 11–42.

Rodríguez-De-Souza-Pajuelo, A. A., Tarazona-Luján, A. F., & Reyes-Bossio, M. (2021). Physical activity enjoyment and self-efficacy in school performance of 11-17-year-old students at educational institutions in Lima. *Journal of Physical Education and Sport*, *21*(3), 2183–2189. https://doi.org/10.7752/jpes.2021.s3278

Rodríguez-Hernández, C., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, *2*, 100018. https://doi.org/10.1016/j.caeai.2021.100018

Ruiz, J., & Moya, S. (2020). Evaluación de las competencias y de los resultados de aprendizaje en destrezas y habilidades en los estudiantes de Grado de Podología de la Universidad de Barcelona. *Educacion Medica*, *21*(2), 127–136. https://doi.org/10.1016/j.edumed.2018.08.007

Salal, Y., & Abdullaev, S. (2020). Deep Learning based Ensemble Approach to Predict Student Academic Performance: Case Study. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. https://doi.org/10.1109/ICISS49785.2020.9316044

Salcedo-Parra, O., Galeano, R., & Rodriguez, L. (2010). Metodología crisp para la implementación Data Warehouse. *Revista Tecnura*, *14*(26), 35–48.

Sánchez-Palacios, B., & Zambrano-Vera, G. (2020). La Ciberadicción En El Rendimiento Académico De Los Estudiantes De Educación Básica Superior De La Escuela Cicerón Robles Velásquez, 2019. *Revista Caribeña de Ciencias Sociales*, *2020–08*.

Sapiens Research. (2021). *Ranking mejores Colegios-Cundinamarca, Colombia 2020-2021*. Sapiens Research.

Schuth, E., Köhne, J., & Weinert, S. (2017). The influence of academic vocabulary knowledge on school performance. *Learning and Instruction*, *49*, 157–165. https://doi.org/10.1016/j.learninstruc.2017.01.005

Shah, M., Kaistha, M., & Gupta, Y. (2019). Student Performance Assessment and Prediction System using Machine Learning. *4th International Conference on Information Systems and Computer Networks, ISCON 2019*, 386–390. https://doi.org/10.1109/ISCON47742.2019.9036250

Suárez, L., Pineda, W., & Mendivelso, I. (2021). Técnicas estadísticas y logro de aprendizaje: revisión bibliográfica. *Eco Matemático*, *12*(2), 112–124. https://doi.org/10.22463/17948231.3323

Suárez, O., & Mora, C. (2018). Efecto de una secuencia didáctica basada en los estilos de aprendizaje y el aprendizaje activo en el logro de aprendizaje de cinemática. *Latin-American Journal of Physics Education*, *12*(4), 1–10.

Sullivan, W. (2017). *Machine Learning for Begginners Guide Algorithms: Supervised & Unsupervised Learning, Decision Tree & Random Forest Introduction*. CreateSpace Independent Publishing Platform.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*. Pearson Educación, Inc.

Tapasco-Alzate, O., Ruiz-Ortega, F., Osorio-García, D., & Ramírez-Ramírez, D. (2020). El historial académico de secundaria como factor predictor del rendimiento universitario. Caso de estudio. *Revista Colombiana de Educación*, *1*(81), 147–169. https://doi.org/10.17227/rce.num81-7530

Touchon, J. C. (2021). Applied Statistics with R. In *Oxford Scholarship*. https://doi.org/10.1093/oso/9780198869979.001.0001

Wandera, H., Marivate, V., & Sengeh, M. (2019). Predicting national school performance for policy making in South Africa. *6th International Conference on Soft Computing and Machine Intelligence, ISCMI 2019*, 23–28. https://doi.org/10.1109/ISCMI47871.2019.9004323

Wang, Y., Pei, F., Zhai, F., & Gao, Q. (2019). Academic performance and peer relations among rural-to-urban migrant children in Beijing: Do social identity and self-efficacy matter? *Asian Social Work and Policy Review*, *13*(3), 263–273. https://doi.org/10.1111/aswp.12179

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*(April), 166–173. https://doi.org/10.1016/j.chb.2019.04.015

Yang, Y., Hooshyar, D., Pedaste, M., Wang, M., Huang, Y.-M., & Lim, H. (2020). Predicting course achievement of university students based on their procrastination behaviour on Moodle. *Soft Computing*, *24*(24), 18777–18793. https://doi.org/10.1007/s00500-020-05110-4

Yopasá, A., & Valbuena, F. (2019). *Resultados de las Pruebas iCFES en Ciencias sociales a través del análisis espacial*.

Zhang, D., & Campbell, T. (2014). An examination of the impact of teacher quality and "Opportunity Gap" on student Science Achievement in China. *International Journal of Science and Mathematics Education*, *13*(3), 489–513. https://doi.org/10.1007/s10763-013-9491-z

Zubiria-Samper, J. (2022, February). ¿Por qué es tan baja la calidad de la educación en Colombia? *El Espectador*.

Abreviaciones

CE	Calidad de la educación
CNA	Comisión Nacional de Acreditación
MEN	Ministerio de Educación Nacional
PISA	Program International Student Assessment
OCDE	Organización para la Cooperación y el Desarrollo Económicos
ICFES	Instituto Colombiano para la Evaluación de la Educación
PLANEA	Plan Nacional para la Evaluación de los Aprendizajes
SIMCE	Sistema Nacional de Evaluación de Resultados de Aprendizaje
KDD	Knowledge Discovery in Databases
CRISP-DM	Cross Industry Standard Process for Data Mining
SEMMA	Sample, Explore, Modify, Model, Assess
OLS	Ordinary Least Squares
SNIES	Sistema Nacional de Información de la Educación Superior
SIVIGILA	Sistema de Salud Pública
DNP	Departamento Nacional de Planeación
SIMAT	Sistema de Matrícula
DANE	Departamento Administrativo Nacional de Estadística
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
AIC	Criterio de Información de Akaike
IPM	Índice de Pobreza Multidimensional
NBI	Necesidades Básicas Insatisfechas
MV	Máxima Verosimilitud
MVR	Máxima Verosimilitud Restringida
MSE	Mean Square Error
AE	Absolute Error

Anexo 1

Código generado para la ejecución de los modelos de Machine Learning.

```
for i in range(len(municipios)):
    # Filtrar por municipio
    tabla_municipio = table[table.IDMUNICIPIO == municipios[i]]
    tabla_municipio = tabla_municipio.drop(columns = ['IDMUNICIPIO'])
    tabla_municipio.IDINSTITUCIONEDUCATIVA = tabla_municipio.IDINSTITUCIONEDUCATIVA.astype('category')
    elemento = tabla_municipio.elemento
    # Generar x,y
    x = tabla_municipio.drop(['PUNTGLOBAL','elemento'],axis = 1)
    y = tabla_municipio[['elemento','PUNTGLOBAL']]
    # Get dummies para categoricas
    x = pd.get_dummies(x,drop_first=True)
    columns = x.columns
    # Estandarizar
    x = pd.DataFrame(MinMaxScaler().fit_transform(x), columns = columns)
    # Guardar indices
    x = pd.concat([pd.DataFrame(columns = ['elemento']),x],axis = 1)
    x['elemento'] = list(elemento)
    # Separar por data training and data test
    x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.3,random_state = 0)
    # Guardar datasets de training and testing
    X.iloc[i,0] = x_train
    X.iloc[i,1] = y_train
    X.iloc[i,2] = x_test
    X.iloc[i,3] = y_test
    # Eliminar indices
    x_train = x_train.drop(columns = ['elemento'])
    x_test = x_test.drop(columns = ['elemento'])
    y_test = y_test.PUNTGLOBAL
    y_train = y_train.PUNTGLOBAL
    # Regresion Lineal
    ml = LinearRegression()
    ml.fit(x_train,y_train)
    y_pred = ml.predict(x_test)
    df_test.iloc[i,0] = mean_squared_error(y_test,y_pred, squared = False)
    df_train.iloc[i,0] = mean_squared_error(y_train,ml.predict(x_train), squared = False)
    df_model.iloc[i,0] = ml
    # Ridge
    ridgecv = RidgeCV(alphas = alphas, scoring = 'neg_mean_squared_error', normalize = False)
    ridgecv.fit(x_train, y_train)
    ridge = Ridge(alpha = ridgecv.alpha_)
    ridge.fit(x_train, y_train)
    y_pred = ridge.predict(x_test)
    df_test.iloc[i,1] = mean_squared_error(y_test,y_pred, squared = False)
    df_train.iloc[i,1] = mean_squared_error(y_train,ridge.predict(x_train), squared = False)
    df_model.iloc[i,1] = ridge
    # Lasso
    lassocv = LassoCV(alphas = None, cv = 10, max_iter = 100000, normalize = False)
    lassocv.fit(x_train, y_train)
    lasso = Lasso(max_iter = 100000, normalize = False)
    lasso.set_params(alpha=lassocv.alpha_)
    lasso.fit(x_train, y_train)
    y_pred = lasso.predict(x_test)
    df_test.iloc[i,2] = mean_squared_error(y_test,y_pred, squared = False)
    df_train.iloc[i,2] = mean_squared_error(y_train,lasso.predict(x_train), squared = False)
    df_model.iloc[i,2] = lasso
```

Anexo 2

Interacción con la comunidad académica

Eco matemático ISSN: 1794-8231 (Impreso), E-ISSN: 2462-8794 (En línea) Volumen 12 (2) Julio-Diciembre de 2021, páginas 112-124

Técnicas estadísticas y logro de aprendizaje: revisión bibliográfica

Statistical techniques and learning achievement: literature review.

Lilian Daniela Suárez-Riveros[a], Wilmer Pineda-Ríos[b], Iván Mauricio Mendivelso-Ramírez[c]

[a*]Maestranda en Ciencia de Datos, Matemática - Universidad Escuela Colombiana Ingeniería Julio Garavito. lilian.suarez@mail.escuelaing.edu.co. ORCID: 0000-0002-8329-0765.
[b]PhD (c) en Estadística, MSc en Matemáticas, Matemático - Universidad Nacional de Colombia. Docente catedra - Universidad Escuela Colombiana Ingeniería Julio Garavito. wilmer.pineda@escuelaing.edu.co. ORCID: 0000-0001-7774-951X.
[c]MSc en Antropología Social – Universidad de los Andes, Estadístico - Universidad Nacional de Colombia. Docente catedra - Universidad Escuela Colombiana Ingeniería Julio Garavito. ivan.mendivelso@escuelaing.edu.co

Forma de citar: Suárez-Riveros, L.D, Pineda-Ríos, W, Mendivelso-Ramírez, I.M. (2021), Técnicas estadísticas y logro de aprendizaje: revisión bibliográfica. *Eco Matemático,* 12 (2), 112-124

Recibido: 19/04/2021
Aceptado: 16/06/2021

Palabras clave

Mapeamiento informacional bibliográfico, modelos estadísticos, métodos estadísticos, logro de aprendizaje.

Resumen: El objetivo de este escrito fue describir las diferentes técnicas estadísticas que han sido empleados para comprender o explicar el logro de aprendizaje, en estudiantes en diferentes niveles educativos. Desde el punto de vista teórico se consolidaron las categorías a priori, provenientes de las técnicas estadísticas (Modelos Multinivel, Modelos geoespaciales, Regresión, Clustering, Análisis Descriptivo, Redes Neuronales, Árboles de decisión, Bosques aleatorios, NaiveBayes y Support Vector Machine), así como la conceptualización de Logro de Aprendizaje. El enfoque metodológico para la revisión se hizo a partir del mapeamiento informacional bibliográfico. Entre los resultados se encontraron 50 documentos de diferentes bases de datos (Elsevier (1), Google Scholar (6), IEEE (4), Scielo (2), ScienceDirect (5), Scopus (31), y Springer (1)), que estudian diferentes regiones del mundo (Asia (17), América del sur (13), América del norte (8), Europa (6), África (5), Oceanía (4), Centro América (3), junto con la orientación a explicar (17), comprender (31) o comprender y explicar (2).Adicionalmente, se identificó un conjunto de variables emergentes en los diferentes reportes, entre las que se encuentra, con mayor relevancia, el nivel socioeconómico, género, afectividad, antecedentes y características y posibilidades de los padres.

*Autor para correspondencia: lilian.suarez@mail.escuelaing.edu.co

DOI: 10.22463/17948231.3323