Minería de Datos Educativos (EDM): análisis de los factores determinantes que influyeron en el desempeño de las pruebas SABER en Cundinamarca (Colombia) entre 2017 a 2021

Carlos Alberto Larrarte Torres

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería de Sistemas
Maestría Gestión de Información
Bogotá D.C., 14 de diciembre de 2022

Minería de Datos Educativos (EDM): análisis de los factores determinantes que influyeron el desempeño de las pruebas SABER en Cundinamarca (Colombia) entre 2017 a 2021

Carlos Alberto Larrarte Torres

Trabajo de investigación para optar al título de
Magíster en Gestión de Información

Directores
PhD. Victoria Eugenia Ospina Becerra
Candidata PhD. Fabiola del Toro Osorio

Jurados
Juan Ricardo Mozo Zapata
Director de la Infraestructura de Datos Espaciales y Estadísticos
Secretaría de Planeación de Cundinamarca

PhD. Dante Conti
Doctor en dirección de proyectos con énfasis en Minería y Modelado Basado en Datos para Mejora de Procesos

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería de Sistemas
Maestría en Gestión de Información
Bogotá D.C., 14 de diciembre de 2022

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

VIGILADA MINEDUCACIÓN

Publicado en 2023 por la Escuela Colombiana de Ingeniería Julio Garavito. Avenida 13 No 205-59 Bogotá. Colombia
TEL: +57 – 1 668 36 00

Página de aceptación del jurado

El trabajo de grado de maestría titulado “Minería de Datos Educativos (EDM): análisis de los factores determinantes que influyeron el desempeño de las pruebas SABER en Cundinamarca (Colombia) entre 2017 a 2021”, presentado por Carlos Alberto Larrarte Torres, cumple con los requisitos establecidos para optar al título de Magíster en Gestión de información.

_____

Juan Ricardo Mozo Zapata

_____

Dante Conti

_____

Victoria Eugenia Ospina Becerra

_____

Fabiola del Toro Osorio

Bogotá, D.C., 14 de diciembre de 2022

Resumen

Las implicaciones del rendimiento consistentemente bajo de los bachilleres colombianos a través de los últimos años tienen como consecuencia dificultades en asegurar su éxito al acceder a la educación superior, así como un impacto negativo en las oportunidades laborales para estos jóvenes y, consecuentemente, una menor calidad de vida a largo plazo en el contexto del mercado laboral, que es cada vez más competitivo. Las causas detrás de este fenómeno son numerosas y variadas, abarcando variables sociales, demográficas o económicas, dificultando la posibilidad de diseñar un plan de acción efectivo que intervenga los factores decisivos del sistema educativo que han resultado en este desenlace.

Según datos del CEINFES, el bajo rendimiento de los estudiantes en las pruebas Saber 11° desde el 2016 y el aumento importante en la abstención en la presentación de la prueba para el 2020 concluyen que el desempeño en las pruebas Saber 11° cayó en los últimos 5 años (Lesmes Martínez et al., 2021). La Organización para la Cooperación y el Desarrollo Económico (OCDE) menciona que Colombia es un país con serias deficiencias en su sistema educativo y sus resultados han estado por debajo del promedio de la OCDE, según las últimas pruebas PISA (Rey Ángel & Gobernador de Cundinamarca, 2016).

En Cundinamarca, entre el 2017 y el 2020, más del 50% de las instituciones educativas departamentales (IED) oficiales de los municipios no certificados se clasificaron en las categorías de resultados más bajos y se identificó un aumento

de 18.7% en el número de instituciones que ingresaron a esta clasificación. Igualmente, hubo una disminución de 9.1% en el número de IED categorizadas en los resultados más destacados para 2020 (Departamento de Cundinamarca et al., 2021).

El objetivo de este análisis se centra en investigar e implementar modelos de minería de datos, empleando algoritmos de aprendizaje supervisados y no supervisados, que permitan examinar los resultados de los estudiantes de Cundinamarca entre los años 2017 y 2021 en las pruebas Saber 11° reportadas por el Instituto Colombiano para la Evaluación de la Educación (ICFES). De esta manera, se identifican los principales determinantes asociados al entorno académico, social, económico y demográfico que estén relacionados con el desempeño del estudiante en las pruebas Saber 11°.

Abstract

The implications of the consistently low performance of Colombian high school graduates over the last few years have resulted in difficulties in ensuring their success in accessing higher education, as well as a negative impact on job opportunities for these young people and, consequently, a lower quality of life in the long term in the context of an increasingly competitive labor market. The causes behind this phenomenon are numerous and varied, encompassing social, demographic, or economic variables, making it difficult to design an effective action plan to intervene in the decisive factors of the educational system that have resulted in this outcome.

The low performance of students in the Saber 11° tests since 2016 and the significant increase in abstention in taking the test by 2020, according to data from CEINFES, conclude that performance in the Saber 11° tests fell in the last 5 years (Lesmes Martínez et al., 2021). The Organization for Economic Cooperation and Development (OECD) mentions that Colombia is a country with serious deficiencies in its education system and its results have been below the OECD average according to the latest PISA tests (Rey Ángel & Governor of Cundinamarca, 2016).

In Cundinamarca, between 2017 and 2020, more than 50% of the official departmental educational institutions (IEDs) of the non-certified municipalities were classified in the lowest result categories, an increase of 18.7% was identified in the

institutions that entered this classification. Likewise, there was a 9.1% decrease in the number of IEDs categorized in the most outstanding results by 2020 (Department of Cundinamarca et al., 2021).

The objective of this analysis focuses on investigating and implementing data mining models, employing supervised and unsupervised learning algorithms, that allow examining the results of Cundinamarca students between 2017 and 2021, contained in the Saber 11° tests reported by the Colombian Institute for the Evaluation of Education (ICFES). In this way, we identify the main determinants associated with the academic, social, economic and demographic environment that are related to student performance on the Saber 11° tests.

Índice General

ÍNDICE DE FIGURAS

Índice de Tablas

1 Introducción

1.1 Problemática (Justificación)

El Instituto Colombiano para Evaluación de la Educación (Icfes) evalúa la calidad de los servicios educativos de diferentes niveles en Colombia a través de exámenes estandarizados, cuyo contenido temático está definido por el Ministerio de Educación Nacional (MEN). El Icfes ha construido un mecanismo de evaluación conocido como el Sistema Nacional de Evaluación Externa Estandarizada (SNEE) basado en la evaluación de las mismas competencias para algunas áreas en todos los niveles educativos (competencias genéricas), esto con el fin de lograr que los resultados sean comparables entre los diferentes niveles de educación (Calderón García & Piñeros Rivera, 2020).

A partir de los planes de estudio diseñados por el MEN, se establecen los estándares mínimos de calidad con los que se espera que se eduquen los alumnos en las aulas de clases y se fijan las expectativas de lo que deberían aprender durante el curso de su educación primaria y secundaria. El Icfes ha diseñado la prueba Saber 11° de acuerdo con estos planes para evaluar las competencias que los estudiantes deberían haber adquirido durante su formación básica.(ICFES, 2022)

La prueba de Estado Saber 11° debe ser tomada por los estudiantes que se encuentren a puertas de finalizar el undécimo grado para obtener resultados oficiales de la culminación de su educación media y que, a su vez, les permitirá ingresar a la educación superior. Alternativamente, el examen también puede ser tomado por cualquiera que haya obtenido su título de bachiller o haya aprobado un programa de validación del bachillerato.

La estructura del examen se compone de 5 pruebas que incluyen: Lectura crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés. Adicionalmente, el examen incluye un cuestionario socioeconómico que incluye preguntas de selección múltiple que se responden en la misma hoja de respuestas, pero no se califican. Este cuestionario busca obtener información adicional sobre los estudiantes que ayuden a dar razón de los resultados desde las características del núcleo familiar, características del hogar y otras características de la familia del estudiante.

Los resultados de los últimos cinco años de las pruebas Saber 11, reflejan un retroceso en los indicadores de las pruebas de Estado de aproximadamente 6.1 puntos porcentuales. Aunque el desempeño de los colegios privados sigue siendo mejor que el de los oficiales, aún quedan por verse los efectos de la pandemia en la educación en próximos años. El informe del Observatorio de Realidades Educativas (ORE) reveló que, en 2016, el 24.9% de los estudiantes logró un buen desempeño en las 4 competencias evaluadas (matemáticas, lectura crítica,

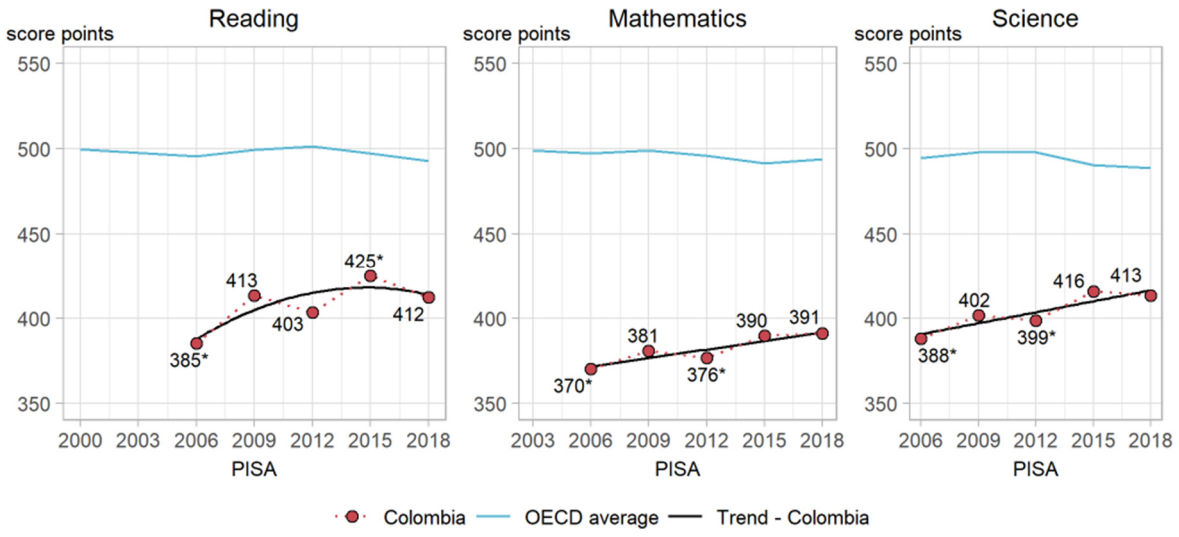
ciencias naturales y sociales y ciudadanas) frente al 18.8% alcanzado en el 2020 por los más de 500 mil estudiantes que presentaron la prueba ese año, de este 18.8% solo un 1,1% de los estudiantes fueron catalogados por el Icfes como óptimos en todas las competencias. Por otro lado, el porcentaje promedio de estudiantes que no ha logrado el desarrollo adecuado de ninguna de las competencias entre el 2016 y el 2020 es de 33%. Ante esta realidad, únicamente 2 de cada 10 estudiantes logran desarrollar las competencias evaluadas por el Icfes (Ruiz, 2021).

En otra encuesta realizada por el Centro de Investigación, Innovación y Desarrollo Tecnológico Orientado a la Gestión Académica – CEINFES, más de la mitad (56%) de los estudiantes consultados aseguran no estar preparados para presentar las pruebas Saber 11. Dicho estudio preguntó cuáles son las materias en los que creen tener mayores falencias: el 39,4% señaló matemáticas, seguida de inglés con un 19,6%, Ciencias Naturales con un 18,8%, Lectura Crítica con el 12,3% y Ciencias Sociales con el 9,9 %. La encuesta incluyó estudiantes de colegios públicos en el 63.4% y colegios privados 36.6% y evidencia un bajón recurrente desde el 2016 en los resultados de las competencias básicas, siendo cada vez menos las personas que obtienen puntajes satisfactorios con un promedio nacional que pasó de 260 puntos en el 2016 a 250 puntos en el 2020 (de 500 puntos posibles). Estos resultados son preocupantes si se tiene en cuenta que el 74% de los estudiantes que participaron en la encuesta considera que la

prueba es determinante para ingresar a la educación superior (Grupo Ceinfes, 2021).

En el ámbito internacional Colombia no tiene mejores resultados. De acuerdo con las pruebas PISA, en el 2018 los estudiantes de Colombia obtuvieron un rendimiento menor que la media de la OCDE en lectura, matemáticas y ciencias; cerca de 50% de los estudiantes alcanzaron por lo menos el Nivel 2 de competencia en lectura y ciencias, 35% alcanzaron por lo menos el mismo nivel de competencia en matemáticas, y casi 40% tuvieron un bajo nivel de logro en las tres materias (Markus, 2019).

Figura 1 Tendencias de rendimiento en lectura, matemáticas y ciencias



Fuente: (Markus, 2019)

Situación para Cundinamarca

El departamento de Cundinamarca está dividido en 15 provincias y 116 municipios de los cuales 108 son municipios no certificados. De acuerdo con el artículo 20 de la Ley 715 de 2001, un departamento es el conjunto de municipios que contiene un departamento dado, tanto certificados como no certificados, entendiéndose como municipio certificado a aquellos que cumplen con la capacidad técnica, administrativa y financiera para administrar de manera autónoma el sistema educativo en su territorio; para el caso de los municipios que no sean certificados, la prestación del servicio público de educación recae directamente sobre el departamento de origen.

La secretaria de Educación de Cundinamarca administra los 108 municipios no certificados, los cuales cuentan con una población cercana a 1.650.000 habitantes según el DANE, en donde el 33.5% es rural y 66.5% es urbana y en donde es importante destacar que, de los 108 municipios, 95 de ellos se consideran netamente rurales; siendo esta una variable por considerar en este tipo de análisis para la planeación, organización y administración del servicio educativo. Los 108 municipios no certificados cuentan con 275 IED (Instituciones Educativas Departamentales), de las cuales 140 están en zona urbana y 135 en zona rural, que prestan el servicio educativo a más de 228.000 estudiantes (Gobernación de Cundinamarca, 2020)

De acuerdo con los resultados de los estudiantes que presentaron las pruebas Saber 11 de las IED, éstas se clasifican en cinco categorías (A+, A, B, C, D); en donde la categoría A+ corresponde a los establecimientos con los resultados más destacados, mientras que la categoría D a los que obtienen los más bajos resultados. Para el departamento, el panorama para el desempeño en las pruebas Saber 11 no es diferente a la tendencia nacional; para el 2020, se observa que la mayoría de las instituciones educativas oficiales de los 108 municipios no certificados ubicados en la categoría inferiores C y D pasaron de un 35.81% en el 2017 a 54.52% en el 2020, con el agravante que las instituciones en las categorías superiores A+ y A viene disminuyendo, pasando de un 15.88% en el 2017 a un 6.69% en el 2020. Según el Banco Mundial, esto pudo ser resultado de la crisis sanitaria generada por el COVID-19 (Secretaría de Educación, 2021).

Al analizar otros posibles factores que puedan dar cuenta de cualquier fluctuación en el rendimiento en las pruebas Saber 11, a nivel nacional o departamental, cobran importancia las variables socioeconómicas y familiares que se recolectan a través del cuestionario socioeconómico. Para las pruebas Saber 11 del 2021, el Icfes reportó que hay una diferencia de 34 puntos en el puntaje global entre los estudiantes que tienen acceso a internet y los que no; igualmente, concluyeron que el 76% de los estudiantes que dedican más de 2 horas al día a la lectura tuvieron el desempeño más alto en la competencia de Lectura crítica (Ministerio de Educación Nacional, 2022).

Si bien los datos provistos por el Icfes son relevantes y sugieren que existen algunos factores determinantes para el éxito educativo, adicionales al servicio educativo en sí, se requiere un análisis mucho más profundo que pueda identificar la totalidad de variables influyentes y la proporción en la que impactan el desempeño en la prueba Saber 11. Adicionalmente, es importante poder priorizar intervenciones sobre aquellos factores que pueden modificarse a corto plazo y que tienen mayor probabilidad de cambiar el desempeño escolar de los estudiantes. De lograr la extracción, análisis y presentación de estos datos ante las entidades estatales competentes, se generaría una nueva herramienta para el diseño y ejecución de políticas públicas que optimicen los recursos de las entidades territoriales para intervenir la tendencia al decremento en el desempeño de los bachilleres en las pruebas Saber 11. Esto, teniendo en cuenta que el acceso al servicio público educativo, la permanencia en él y la garantía de la calidad de la educación en Colombia son responsabilidad de la nación, los departamentos y los municipios (artículo 4° de la Ley 115 de 1994 o Ley General de Educación). (Montes & Galvis, 2017)

Esta investigación académica recoge los datos abiertos, generados por el Instituto Colombiano para la Evaluación de la Educación ICFES, asociados a los resultados de las pruebas Saber 11 entre los años 2015 al 2021 de los 108 municipios no certificados del Departamento de Cundinamarca, evaluando las diferentes variables escolares y socioeconómicas para identificar su relación con los

resultados obtenidos. En la investigación, se aplicaron técnicas de análisis de datos que permitieron identificar posibles comportamientos que estén originando el continuo descenso de la calidad en la educación del departamento.

El objetivo de este documento es aportar información relevante que permita evaluar la cobertura y calidad educativas, generando estrategias a nivel municipio, como primer responsable del préstamo del servicio público de educación según la Constitución Política de 1991. Este trabajo de investigación responde a la pregunta: ¿Cuáles son los factores socioeconómicos y académicos que se correlacionan con los resultados de las pruebas Saber 11 en los últimos 5 años y que impactan la calidad del sistema educativo en los municipios no certificados de Cundinamarca, a partir del uso de técnicas de análisis de datos sobre fuentes de datos abiertas?

1.2 Objetivos y pregunta de investigación

El objetivo general de este trabajo es identificar los factores socioeconómicos y académicos que se correlacionan con los resultados de las pruebas Saber 11 en los últimos 5 años y que impactan la calidad del sistema educativo en los municipios no certificados de Cundinamarca, a partir del uso de técnicas de análisis de datos sobre fuentes de datos abiertas.

Para el desarrollo del objetivo general se proponen los siguientes objetivos específicos:

1. Identificar las fases de un modelo analítico de datos para la predicción de los puntajes globales de la prueba Saber 11° a partir de los factores identificados en el cuestionario socioeconómico de la misma prueba.
2. Seleccionar las fuentes de datos apropiadas para la extracción de datos relevantes en la caracterización de los factores socioeconómicos y académicos recolectados en las pruebas Saber 11° entre los años 2017 y 2021.
3. Estimar el impacto de diferentes métodos de limpieza de datos (imputación de datos o eliminación de valores faltantes) en la eficiencia de los modelos de predicción del puntaje global de la prueba Saber 11° a partir de los factores estipulados en el cuestionario socioeconómico de la misma prueba.
4. Evaluar la correlación de cada variable socioeconómica estipulada por el cuestionario socioeconómico de las pruebas Saber 11° y el puntaje global de la prueba Saber 11° utilizando diferentes algoritmos de aprendizaje de máquinas no supervisados para reducción de dimensionalidad (correlaciones entre variables, análisis factorial de datos mixtos y BORUTA).
5. Aplicar modelos de predicción basados en algoritmos supervisados de aprendizaje de máquinas (árboles de decisión, bosque aleatorio y regresión

logística) a partir de las variables socioeconómicas que se correlacionaron con el puntaje global de la prueba Saber 11, para dos escenarios con métodos de limpieza de datos diferentes (por imputación de datos o eliminación de valores faltantes).

Este trabajo responde a la pregunta de investigación: ¿Cuáles son los factores socioeconómicos y académicos que se correlacionan con los resultados de las pruebas Saber 11 en los últimos 5 años y que impactan la calidad del sistema educativo en los municipios no certificados de Cundinamarca, a partir del uso de técnicas de análisis de datos sobre fuentes de datos abiertas?

1.3 Metodología

Para desarrollar la pregunta de investigación y los objetivos planteados, se propuso una metodología de múltiples fases definidas a partir de la metodología CRISP-DM (Gironés et al., 2017), a saber: la primera fase es la revisión conceptual sobre la situación de evaluación estandarizada de la educación secundaria en Colombia a partir de la prueba Saber 11° diseñada por el ICFES. La segunda fase es la selección de fuentes de datos abiertos para obtener los registros correspondientes al cuestionario socioeconómico de la prueba para los años 2017 a 2021; a partir de esta selección la tercera fase involucra la limpieza de los datos por dos métodos y una vez se realiza la limpieza, se pasa a la selección de variables socioeconómicas relevantes a partir de los resultados de

algoritmos de reducción de dimensionalidad para estimación del nivel de correlación de cada una con el puntaje global de la prueba Saber 11. La siguiente fase está asociada al diseño de los modelos, a partir de algoritmos supervisados de aprendizaje de máquinas, con las variables relevantes previamente seleccionadas. Finalmente, la última fase está asociada al análisis de los resultados tomando el desempeño escolar general y en las pruebas Saber 11° como un proceso multicausal respaldado con la elaboración del presente documento para la presentación y visualización de los resultados.

En la primera fase, correspondiente al “Entendimiento del negocio”, se analizó la información que recopiló el Instituto Colombiano para la Evaluación de la Educación Superior (ICFES) de las pruebas que se realizan por año entre los años 2017 y 2021. También se hizo una revisión del estado del arte cubriendo la información disponible sobre los retos para la educación básica y secundaria en Colombia y América latina y la evolución en las mediciones estandarizadas de la educación específicamente para Colombia, incluyendo el impacto de la pandemia por COVID-19 en la educación escolar. A partir de esta búsqueda amplia, se detectó la situación problema del declive consistente en los resultados de la prueba Saber 11° desde el año 2014 y se formuló la pregunta de investigación para abordar la problemática.

En la fase “Comprensión de los datos”, se analizaron y normalizaron las fuentes de datos, identificando las variables cuantitativas y cualitativas de los datos

correspondientes a las variables socioeconómicas del cuestionario socioeconómico de la prueba Saber 11°. En la fase “Preparación de los datos” se limpiaron, unificaron, transformaron variables, se imputaron los valores nulos, se eliminaron valores atípicos y se hizo una selección de variables socioeconómicas en el cuestionario socioeconómico (reducción de dimensionalidad) en busca de todos los factores que pueden influir en la calidad educativa del departamento.

Desde esta fase, y en las siguientes, el proceso de análisis se tornó bidireccional; es decir, consistía en retornar a la etapa anterior y ajustar los datos y los modelos en caso de que los resultados no cumplieran con el objetivo del proyecto. En la fase “Modelado”, se diseñaron y probaron modelos de clasificación de datos para hacer finalmente una predicción de la variable puntaje global de la Prueba Saber 11 de acuerdo con las variables socioeconómicas seleccionadas.

En la fase de “Evaluación y despliegue del modelo”, se probó el modelo seleccionado, con varios grupos de datos en donde se midió la confiabilidad y certeza de la predicción, para finalmente concluir sobre los resultados que arrojan las diferentes corridas del modelo analítico y que permitan estructurar la información para utilización del usuario final y poder influir en la toma de decisiones de políticas públicas, que para efectos de este proyecto correspondería con la Gobernación de Cundinamarca.

1.4 Estructura del documento

El documento se organizó de la siguiente manera:

En Capítulo 1, describe el problema, define los objetivos, la metodología de minado y la pregunta a responder de la investigación.

En el Capítulo 2, se encuentra el marco teórico y referencial en los que se fundamentó el desarrollo de la investigación.

En el Capítulo 3, se describe como se realizó el análisis del problema, la recolección, preparación, limpieza, transformación y modelado de los datos.

En el Capítulo 4, se presentan los resultados de los modelos de selección de variables y clasificación de acuerdo con la variable de desempeño.

En el Capítulo 5, se presentan las conclusiones y alternativas de posibles trabajos futuros.

2 Marco teórico o estado del arte

El desarrollo de este proyecto aborda una metodología que se desarrolló por fases y toma diferentes elementos para la extracción, limpieza y análisis de los datos. A continuación, se presentará el estado del arte relevante para la ejecución de las diferentes fases del proyecto, resaltando su articulación con la ejecución de este.

2.1 La minería de datos

La minería de datos se utiliza en esta investigación para lograr dos objetivos principales para la gobernación de Cundinamarca; uno es construir modelos para predecir los valores numéricos para la variable objetivo o dependiente. Sin embargo, otro objetivo en esta investigación es identificar y entender las posibles relaciones existentes entre las demás variables o variables independientes con la variable objetivo. De aquí que, se hace necesario que la minería de datos sea aplicada a través de una metodología que incluya etapas de análisis y entendimiento del problema que se va a tratar, así como la elección de las técnicas apropiadas para obtener información de valor, tratada y aprovechable a través de la identificación de patrones, tendencias y elementos relevantes en los datos (Microsoft, 2019).

2.2 Técnicas y conceptos utilizados durante el proceso de minado

La minería de datos incluye procesos nada triviales como la preparación, limpieza, integración, transformación y modelado de datos. En esta investigación, durante el proceso de recolección de datos y extracción de información útil se utilizaron técnicas estadísticas en análisis estadístico para entender la distribución de variables tales como: la mediana, media aritmética, la varianza, la desviación estándar, la distribución normal sobre las variables que reflejan el desempeño de los estudiantes. Esto se realizó con el objetivo de identificar el grado de dispersión de los datos.

En la fase de preprocesamiento y limpieza de datos se utilizaron técnicas para identificar redundancia de registros, redundancia de variables, valores faltantes (tipos de datos faltantes) y valores atípicos. En la detección de valores atípicos y en los algoritmos de agrupación en clústeres (modelos no supervisados) se utilizan las distancias de similitud que cuantifican la semejanza o cercanía entre dos objetos estadísticos.

Adicionalmente, en esta fase (preprocesamiento) se identificaron y clasificaron los tipos de variables o atributos (cuantitativas o cualitativas), de esta clasificación se identifican los atributos numéricos (continuos o discretos) y categóricos (nominal u ordinal). Al estudiar el número de registros y los diferentes tipos de atributos se identificó la variable objeto del estudio o variable dependiente (desempeño del

estudiante) y las variables independientes, definidas como las que tienen razones potenciales de variación y afectan el comportamiento de la variable dependiente.

Como ya se mencionó, el objetivo de esta investigación es la selección de variables que son determinantes en el desempeño estudiantil en las pruebas Saber 11, en donde los conjuntos de datos de estas pruebas tienen en promedio 83 variables desde el 2016 hasta el 2021; para este número de variables es útil reducirlas con un algoritmo no supervisado (sin variable objetivo) antes de ejecutar algoritmos supervisados (con variable objetivo). Con este objetivo, se utilizaron estrategias para reducir los datos, utilizando una técnica llamada “la reducción de dimensionalidad” que permite eliminar las variables que no son relevantes en el análisis del desempeño estudiantil, reducir la complejidad del modelo predictivo y aumentar su desempeño. Para estos datos, al ser mixtos (categóricos y numéricos), se debe aplicar una técnica de análisis multivariante llamada FAMD (Análisis factorial de datos mixtos), la cual analiza la similitud entre variables conservando su naturaleza.

En los modelos predictivos o supervisados se utilizaron algoritmos de clasificación utilizando la variable objetivo (desempeño del estudiante / categórica), se corrieron modelos con Random Forest o “Bosques Aleatorios” en español, predicciones con regresiones logísticas y GBM (Gradient Boosting Machine), con el objetivo de predecir el desempeño del estudiante dentro los tres niveles definidos en la variable objetivo (ESTU_DESEMPEÑO).

Todas estas técnicas y conceptos se describen en las siguientes secciones de este documento.

2.3 Minería de datos para la educación (EDM)

Es una disciplina emergente que tiene como objetivo mejorar los resultados del aprendizaje mediante el uso de la minería de datos. Desarrollando métodos que ayudan a predecir el conocimiento de los estudiantes, la deserción y el cubrimiento escolar. Con las técnicas de minería de datos, una base de conocimiento explícito, habilidades analíticas y conocimiento del dominio se pueden descubrir tendencias y patrones ocultos. Estas tendencias y patrones forman los modelos predictivos que permiten ayudar a las organizaciones a descubrir información útil y luego guiar la toma de decisiones. (Cheng, 2017)

2.4 Metodología CRISP-DM

La metodología CRISP-DM es un proceso estándar de la industria para la minería de datos que desarrolló un modelo de seis fases; describe naturalmente el ciclo de vida del análisis de datos y está basado en la práctica y experiencia real del analista de datos. Es como el mapa de ruta que permite planear, ejecutar, verificar y aprender de los modelos desarrollados para estos análisis de datos, aplicando estrategias de calidad total (ciclo PHVA) y la visión de un proyecto de minería de datos como una secuencia de fases. Es importante mencionar que la revisión y el orden de ejecución de las fases es un aspecto clave en el desarrollo de un proyecto de calidad. Todas las fases son importantes y, al ignorar esto, se

terminan concentrando demasiados recursos al final del proyecto, por no haber hecho las cosas bien en las fases iniciales. (Gironés et al., 2017)

El modelo en sus seis fases secuenciales incluye tareas genéricas y específicas, que se interrelacionan y se ejecutan en un ciclo iterativo, hasta lograr el objetivo planeado y van a permitir dar respuesta a las siguientes preguntas (Gironés et al., 2017):

1. **Entendimiento del negocio** ¿Qué necesita el negocio?
2. **Comprensión de datos** ¿Qué datos tenemos/necesitamos? ¿Están limpios?
3. **Preparación de datos** ¿Cómo organizamos los datos para el modelado?
4. **Modelado** ¿Qué técnicas de modelado debemos aplicar?
5. **Evaluación** ¿Qué modelo cumple mejor con los objetivos de negocio?
6. **Despliegue** ¿Cómo acceden las partes interesadas a los resultados?

La metodología CRISP-DM nació en el seno de dos empresas, DaimlerChrysler y SPSS en 1999, desde entonces se convirtió en la metodología más común para proyectos de minería de datos, analítica y ciencia de datos (Data Science Process Alliance, 2020).

A continuación, se explican cada una de las fases:

2.4.1 Entendimiento del negocio

Esta fase se centra en la compresión de los objetivos y requisitos del análisis de datos, identificando las metas finales y los criterios de éxito desde una perspectiva de su utilidad para el negocio y las técnicas de la minería de datos, para lograr un plan de proyecto que contemple tecnologías y herramientas que incluya planes detallados por cada fase del proyecto. En esta fase, se realizan las actividades necesarias para abordar y apropiarse del conocimiento del problema objeto de estudio, que permiten determinar el punto de partida respecto a los objetivos, la disponibilidad de los recursos, incluyendo riesgos, contingencias y lograr un análisis costo-beneficio.

2.4.2 Comprensión de los datos

Corresponde al acercamiento inicial y familiarización de los datos, identificar su origen, la cantidad, la estructura, propiedades, clasificación, identificar los problemas de calidad, obtener conocimiento preliminar sobre los datos y descubrir las relaciones más evidentes que permitan definir las primeras hipótesis. De esta etapa de exploración, descripción, reconocimiento, relación entre los datos y verificación depende en gran parte que se cuente con un insumo favorable para la ejecución de las siguientes etapas.

2.4.3 Preparación de los datos

Del resultado de la etapa anterior, se inicia la preparación de los datos para la aplicación del modelo o técnica de minería de datos que se vaya a implementar.

Una regla general común menciona que entre el 60% y el 80% del proyecto es la preparación de datos (Dataversity, 2016). La definición de los datos se hará a través del uso de técnicas estadísticas, análisis de tablas, selección de propiedades y atributos, limpieza y transformación de datos. Es importante considerar también las posibles afectaciones o riesgos de acuerdo con la calidad de los datos, identificar las fuentes de origen relevantes. La finalidad de esta etapa es consolidar los datos orientados en los objetivos definidos en la fase de entendimiento de negocio.

2.4.4 Modelado

Esta fase está orientada a la selección, aplicación, prueba y evaluación de técnicas de modelado que respondan a los objetivos establecidos. Se seleccionan las técnicas de modelado y se determinan qué algoritmos probar. Algunos de los criterios a considerar para la elección del modelo, es que sea acorde al problema; que cumpla con los requisitos del problema y que cuente con los datos apropiados. Dependiendo de la técnica a usar, puede que sea necesario regresar a las fases anteriores para ajustar y dividir los datos en conjuntos de entrenamiento y pruebas. En la etapa de modelado se debe determinar el plan de monitoreo y calidad; generalmente se generan varios modelos que compiten entre sí y se deben interpretar los resultados en función del conocimiento del dominio y los criterios de éxito definidos y el diseño de la prueba, en caso de presentarse desviaciones se deben realizar las correcciones que respondan de manera

positiva a los objetivos planteados. En esta fase, se sugiere iniciar un proceso de iteración que incluya la construcción y evaluación de los modelos hasta que se considere que se ha encontrado el mejor modelo.

2.4.5 Evaluación del Modelo

Durante esta etapa, se realiza la evaluación del modelo seleccionado frente a los criterios establecidos en la definición de los objetivos. Se revisan las tareas ejecutadas en la construcción del modelo, así como la posible ocurrencia de fallas, ¿se pasó algo por alto?, ¿Se ejecutaron correctamente todos los pasos?, de la correcta evaluación depende la efectividad en el despliegue y el evitar los reprocesos. Así mismo en esta etapa se establece si se generaron conocimientos aún sin estar contemplados dentro de los planteamientos iniciales. Finalmente se definen los pasos a seguir dependiendo de los resultados, es decir si se retoman las fases anteriores, se abren nuevas líneas de trabajo con base a los descubrimientos obtenidos o si se continua la etapa de despliegue.

2.4.6 Despliegue

Esta es la última etapa del modelo CRISP-DM; busca transformar el conocimiento obtenido en insumo real y valioso para la toma de acciones, de acuerdo con la situación identificada y los objetivos planteados en la fase de compresión del negocio. Es recomendable realizar una retrospectiva del proyecto sobre lo que salió bien y lo que podría haber salido mejor y como mejorarlo en un futuro. La implementación de los resultados debe ser divulgada a los grupos de interés de

una manera entendible, un modelo no es particularmente útil a menos que se pueda acceder a sus resultados. Debe contar también con un plan de implementación, monitoreo, mantenimiento y medición durante la fase operativa del modelo, que permita identificar lecciones aprendidas y acciones de mejora, derivadas de la evaluación de todos los procesos realizados. (Gironés et al., 2017)

2.5 La estadística

El origen de la palabra Estadística proviene del vocablo “Estado”, pues adicional a la función de ejercer y administra el poder político; el estado era responsable de establecer registros de población, nacimientos, defunciones, impuestos, cosechas etc. Desde el establecimiento de las sociedades humanas organizadas ha existido la necesidad de tener datos sobre la población.

“Estadística: Es la ciencia que se encarga de la recolección, ordenamiento, representación, análisis e interpretación de datos generados en una investigación sobre hechos, individuos o grupos de estos, para deducir de ello conclusiones precisas o estimaciones futuras” (Salazar & del Castillo Santiago G, 2018).

2.5.1 Análisis estadístico

El proceso estadístico observa los datos y trata de analizarlos, con el fin de obtener explicaciones y predicciones sobre los fenómenos observados; por este motivo la minería de datos utiliza muchos conceptos que provienen de este campo del conocimiento. La estadística por su transversalidad comparte con la minería de

datos métodos como el muestreo, la exploración de datos, la inferencia estadística y la búsqueda de patrones.

En este estudio se van a examinar varios conceptos estadísticos y matemáticos que son útiles en cualquier proceso de análisis de datos que utiliza algoritmos de minería de datos. (Martin, 2012, p. 73)

2.5.2 Estadística muestral

Es la actividad de tomar una parte (muestra) de un set datos o población de acuerdo con un plan (muestreo), para obtener conclusiones resultado de un análisis que se puedan extender al total de la población de origen.

En la estadística, los datos son el resultado de las diferentes observaciones que se pueden hacer sobre una variable. Las observaciones como el mayor valor, el menor valor o los valores con mayor o menor frecuencia en un set de datos son propiedades de los valores que permiten definir la distribución de una variable. Es de especial interés en la estadística la zona central de una distribución y es allí donde aparecen las estimaciones más comunes: la mediana y la media aritmética, que miden en centro de una distribución, pero de manera distinta.

- **La mediana**: es el valor central de una distribución $Me = x\ (n+1)/2$, donde n es el número total de observaciones.
- **La media aritmética**: es su valor medio expresado como

$$x^{-} = 1/n \sum\nolimits_{-}(i = 1)\text{^}n\ x\_i$$

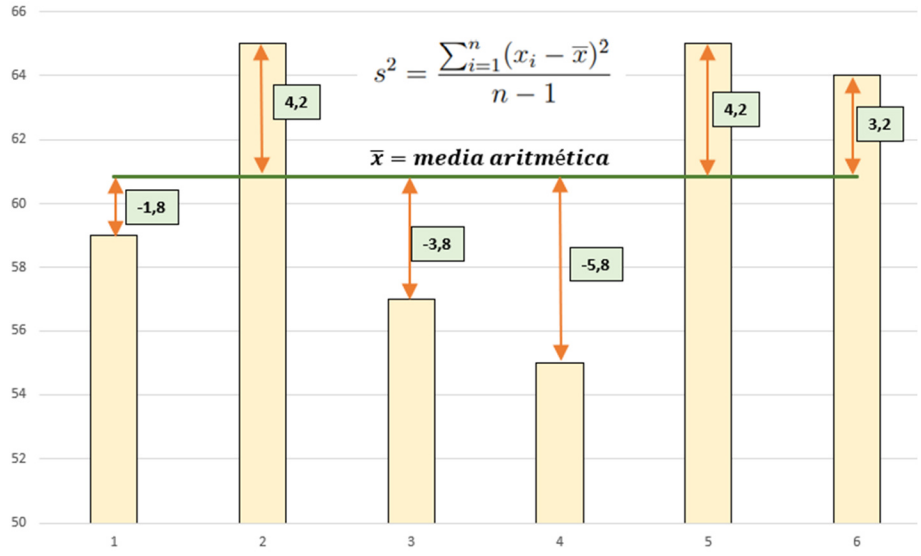
La media por sí sola no es suficiente para describir la distribución de una variable, por eso es conveniente acompañarla de la varianza, en donde entre menor sea la varianza indica que el set de datos es compacto y en consecuencia la media para este caso si es representativa. La varianza y la desviación estándar son medidas de dispersión de los datos, la dispersión se refiere a que si todos los valores de un set de datos son iguales no hay dispersión, en caso contrario hay dispersión. Estas medidas de dispersión son muy sensibles a los valores atípicos, que no son más que los valores que se alejan mucho de la media, por eso es recomendable hacer un tratamiento previo de estos valores atípicos (***Outliers***). (Martin, 2012, p. 73)

2.5.3 La varianza

Esta medida de una variable, entendida como la suma de los cuadrados de las diferencias entre cada valor y la media aritmética, mide la dispersión de los datos de la muestra $(x\_1, x\_2, x\_(3, \dots)\ x\_n)$ respecto a la media $(\bar{x})$, todo dividido entre el número total de observaciones menos 1. Esta dispersión es grande o pequeña, dependiendo de que tan cerca están los valores a la media. La varianza es la desviación estándar elevada al cuadrado.

$$\boldsymbol{S\text{^}2 = (\sum\_(i = 1)\text{^}n \blacksquare (x\_i - \bar{x})\text{^}2\ )/(n - 1)}$$

Figura 2 Calculo de la varianza



Fuente: Elaboración propia

2.5.4 La desviación estándar

Contrario a la varianza, la desviación estándar es simplemente la raíz cuadrada de la varianza, también retorna una medida de dispersión alrededor de la media, pero en la misma escala o unidades de la variable muestreada. Mientras **mayor** sea la desviación estándar, mayor es la dispersión de los datos.

$$s = \sqrt{(s\text{^}2\ )}$$

El valor de la aplicación de la desviación estándar como medida de dispersión, radica en que la mayoría de las desviaciones con respecto a la media, están dentro de un área o intervalo igual a una o dos veces la desviación estándar, es decir,

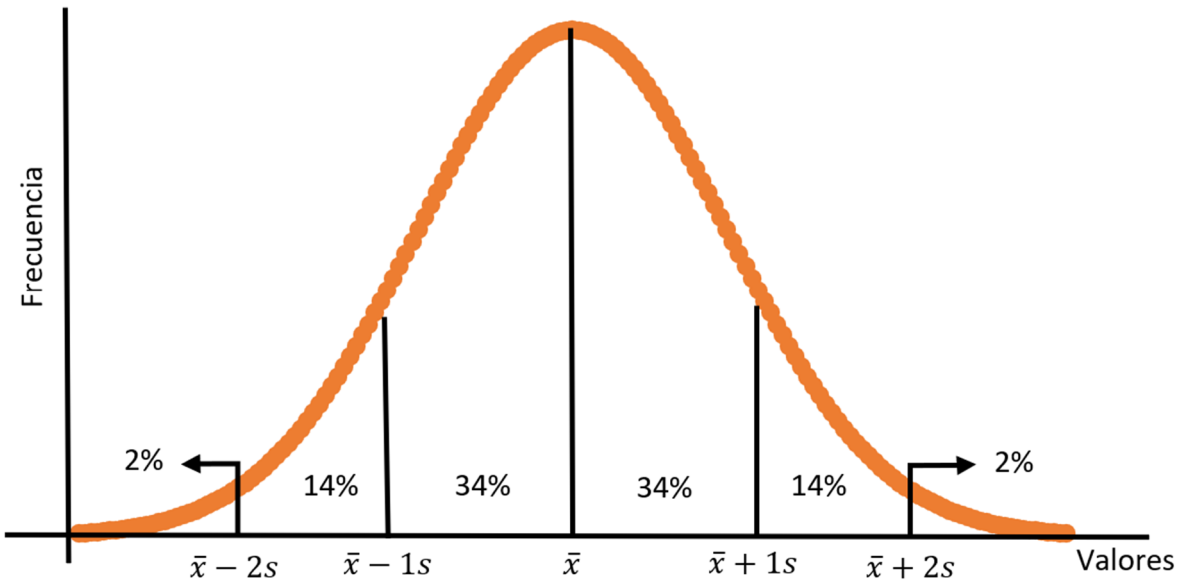
$(\bar{x} - s, \bar{x} + s\ )$ o $(\bar{x} - 2s, \bar{x} + 2s\ )$. (Gironés et al., 2017)

2.5.5 Distribución normal

Se dice que algo es normal cuando se encuentra en su estado natural, al estudiar las características de una población es necesario conocer si los valores observados son normales o los esperados. Generalmente lo normal se ubica cerca del valor medio.

Cuando un set de datos tiene una distribución “normal”, el 68.26% de las observaciones de la distribución están a menos de una desviación estándar de la media y en donde el 95.45% de los datos tienen observaciones que están a menos de dos desviaciones estándar de la media. En general, casi el 100% de las observaciones se encuentra a menos de 3 desviaciones estándar de la media 99.73%. (Martin, 2012, p. 73)

Figura 3 Distribución normal



Fuente: Elaboración propia

2.5.6 Valores atípicos (Outliers)

Son valores diferentes de los demás valores en un set de datos, son raros, distintos, son excepcionalmente lejanos del centro o de la característica principal de los datos. Los valores atípicos son importantes por su efecto en la media, no hay una regla para identificar los valores atípicos y es donde un experto en el dominio de los datos debe analizar los datos sin procesar para definir si es un valor atípico o no. Algunos libros definen los valores atípicos todos aquellos que son mayores 1.5 veces el valor del rango intercuartil que es la distancia entre el primer y tercer cuartiles(Q3-Q1); en donde un cuartil es una medida percentil que divide el total de 100% en cuatro partes iguales 25%,50%,75% y 100%.

Incluso con una comprensión profunda de los datos los valores atípicos pueden ser difíciles de definir. Se debe tener mucho cuidado de no quitar o cambiar valores apresuradamente especialmente si el tamaño de la muestra es pequeño (Kuhn & Johnson, 2013).

Los valores atípicos pueden tener muchas causas tales como (Jason Brownlee, 2020):

- Error de medida o entrada de datos.
- Problema del proceso.
- Corrupción de datos.
- Verdadera observación atípica.

2.6 Métricas de distancias o similitud

La medida de similitud en un contexto de minería de datos es una distancia con dimensiones que representan características de los objetos. Si la distancia es pequeña, dos objetos son muy similares mientras que si la distancia es grande observaremos un bajo grado de similitud. Estas métricas de distancia se utilizan tanto en el aprendizaje supervisado como en el no supervisado generalmente para calcular la similitud entre los puntos de datos.

La mayoría de los enfoques de agrupación, utilizan medidas de distancia para evaluar las similitudes o diferencias entre un par de objetos; las medidas de distancia más populares utilizadas son:

2.6.1 Distancia euclidiana

La distancia euclidiana se considera la métrica tradicional para problemas de geometría. Puede explicarse simplemente como la distancia ordinaria entre dos puntos. Calcula matemáticamente la raíz de las diferencias al cuadrado entre las coordenadas entre dos objetos, es una de las distancias más utilizadas con atributos numéricos. Uno de los problemas de esta técnica es que no tiene en cuenta las diferentes unidades de medida de las variables(Geeks for Geeks, 2020).

$$d = ((x_1 - y_1)^2 + (x_2 - y_2)^2)^{1/2} \qquad d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Donde m = número de dimensiones y x_1, y_1 los puntos del dato.

2.6.2 Distancia de Gauss

Para superar el problema de la distorsión generada por las diferentes unidades de medida de las variables, esta distancia normaliza todas las variables bajo la misma escala. La expresión para m dimensiones, donde los puntos tienen una desviación estándar σ, viene dada por la formula:

$$d(x, y) = \sqrt{\sum_{i=1}^{m} \left(\frac{x_i - y_i}{\sigma_i}\right)^2}$$

Esta distancia estadística no tiene en cuenta la correlación entre las variables, es decir; si las variables de trabajo son totalmente independientes, no habría dificultades pero en caso contrario esa correlación o influencia entre variables no queda bien reflejada.

Por ejemplo, las variables *entrenar* y *rendimiento* están correlacionadas y tienen una correlación positiva, es decir entre más entrenamiento siempre implica más rendimiento. Al comparar dos deportistas ignorando esta correlación se puede llegar a conclusiones erróneas. (Gironés et al., 2017)

2.6.3 Distancia de Mahalanobis

Esta medida de distancias se diferencia de la distancia euclídea y de Gauss en que tiene en cuenta la correlación entre variables aleatorias mediante la siguiente expresión para un espacio de m dimensiones es:

$$d(x,y) = \sqrt{(x_1 - y_1, \dots, x_m - y_m) \begin{pmatrix} \sigma_x^2 & \sigma(x,y) \\ \sigma(x,y) & \sigma_y^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - y_1 \\ x_m - y_m \end{pmatrix}}$$

donde σ_x^2 y σ_y^2 representan la varianza de las variables x e y, respectivamente, y $\sigma(x,y)$ la covarianza entre ambas variables. (Gironés et al., 2017). Al utilizar las distancias euclídeas, se da más importancia a la variable con menor varianza. El método Mahalanobis iguala la importancia de las variables y se puede incluir la varianza y la covarianza entre variables; esta es una distancia de medida basada en el análisis de los datos de donde proviene las variables objeto del estudio, en contraste con las distancias euclidianas y de Manhattan, que son independientes del conjunto de datos (Escobedo & Salas, 2008).

2.7 Entropía de la información

La entropía se define como falta de orden y previsibilidad, lo que parece una descripción adecuada de la diferencia entre los dos escenarios. Valores de entropía altos indican que el resultado es muy aleatorio y, en consecuencia, poco predecible; al relacionar este concepto con la ganancia de información, se obtiene una medida de qué tan relevante es una variable dentro de un set de datos. Es decir, un atributo con mucha ganancia de información juega un papel

preponderante en un set de datos con miras a predecir el atributo objetivo o clase. Su expresión matemática es la siguiente:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\, log2\ p(x_i)$$

donde X es una variable aleatoria discreta con un conjunto de posibles valores $X = \{x_i, \dots, x_n\}$ y $p(x_i)$ es la probabilidad de que la variable X tome el valor x_i.(Gironés et al., 2017)

2.8 Datos Abiertos

Según la Carta Internacional de Datos Abiertos (ODC Open Data Charter, por sus siglas en inglés), adoptada por Colombia desde 2016 como instrumento orientador en la generación y uso de datos, los datos abiertos son “datos digitales que son puestos a disposición con las características técnicas y jurídicas necesarias para que puedan ser usados, reutilizados y redistribuidos libremente por cualquier persona, en cualquier momento y en cualquier lugar” (Open Data Charter, 2015).

En Colombia, la Ley 1712 de 2014 sobre Transparencia y Acceso a la Información Pública Nacional, define los datos abiertos en el numeral sexto como “todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin

restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos" (Ministerio de Tecnologías de la Información y Comunicaciones, 2019).

De acuerdo con la ODC, se desarrollaron seis principios en el 2015 como un conjunto de normas acordadas a nivel mundial, para sentar las bases para el acceso, uso y publicación de los datos abiertos:

1. Abiertos por Defecto
2. Oportunos y Exhaustivos
3. Accesibles y Utilizables
4. Comparables e Interoperables
5. Para mejorar la Gobernanza y la Participación Ciudadana
6. Para el Desarrollo Incluyente y la Innovación

De la necesidad estratégica de disponer de métricas que evalúen la reutilización nace Meloda, que permite calificar información y evaluar su grado de reutilización. Las primeras versiones de este sistema de medición se remontan al año 2011, tras la iniciativa del investigador español Alberto Avella, la cual se publica por la Universidad Rey Juan Carlos de Madrid. En el año 2013 se actualiza la versión Meloda 2.0, en la que se evaluaban tres dimensiones: estándares técnicos, acceso y legal. Con el desarrollo de la metodología y con la aplicación a 200 conjuntos de datos originarios de portales de datos abiertos en España, se evidencia la necesidad de incluir la dimensión de modelo de datos y se emite la versión 3.10. La dimensión de estándares técnicos evalúa que los datos estén

almacenados en un formato que no sea privativo. El acceso evalúa que la información sea legible en formato automatizado; el aspecto legal se mide en términos de las restricciones o barreras legales que presente el acceso (license-free) y el modelo de los datos a publicar refleja la importancia que tiene la estructura de datos para poder procesar la información.

2.9 Tipos de atributos o variables

Una variable representa una dimensión de la realidad que se quiere analizar y en donde el valor de la variable se refiere a la medida que la variable presenta. La variación de la variable son los diferentes valores que asume una variable en un set de datos. Dependiendo del valor que representa las variables se agrupan en dos tipos de variables:

1. **Variable cuantitativa**: en donde el valor representa una cifra.
2. **Variable cualitativa**: en donde el valor representa una categoría. (Salazar & del Castillo Santiago G, 2018, p. 17)

2.9.1 Variables cuantitativas

Son aquellas que se expresan en valores numéricos y en donde se facilita hacer operaciones aritméticas con ellas (por ejemplo: precio, edad, etc.). Estas se dividen en dos tipos:

1. **Variable cuantitativa continua**: cuando el valor no representa ninguna interrupción, es decir puede tomar cualquier valor numérico (por ejemplo: estatura de una persona, área de un terreno).

2. **Variable cuantitativa discreta**: cuando el valor presenta una interrupción entre sus valores en donde los intermedios no existen. Variables que se determinan mediante conteos (por ejemplo: el número de hijos de una persona, número de clientes, número de pisos de un edificio) (Soles Ramos & Torrent Sellens, 2010).

2.9.2 Variables cualitativas

Son aquellas expresan atributos o características (por ejemplo: genero de las personas, el color de un objeto, nombre de un departamento) que no se pueden expresar con números. Estas también se dividen en tres grupos:

1. **Variable cualitativa nominal**: Son variables que no admiten un criterio de orden (por ejemplo: estado civil, departamento de una empresa, afiliación religiosa).

2. **Variable cualitativa ordinal**: Son variables no numéricas, en las que existe un orden o un puesto (por ejemplo: medallas de una prueba deportiva: oro, plata, bronce; nivel socioeconómico).

3. **Variable dicotómica**: Es un caso muy común de *variable cualitativa nominal* en donde solo se aceptan dos valores, se usa para señalar la presencia o ausencia de algo, o la afirmación o negación de algo (Soles Ramos & Torrent Sellens, 2010).

2.10 Variables independientes y variables dependientes

En una investigación es necesario identificar dos tipos de variables, una variable dependiente y otra independiente, la forma de identificarlas es:

1. **Variable independiente**: es una variable que representa una cantidad que se modifica, es sobre la que se prueba para demostrar una hipótesis, es la variable sobre la que se tiene control y se cambia sistemáticamente para analizar cómo afecta la variable dependiente.

2. **Variable dependiente**: es la variable que representa una cantidad cuyo valor cambia o depende de cómo se modifica la variable independiente; es la variable inestable y la que se pretende medir, el objetivo es analizar cómo se comporta frente a influencias que surgen de los cambios en las variables independientes (Soles Ramos & Torrent Sellens, 2010).

2.11 Tipos de tareas en minería de datos

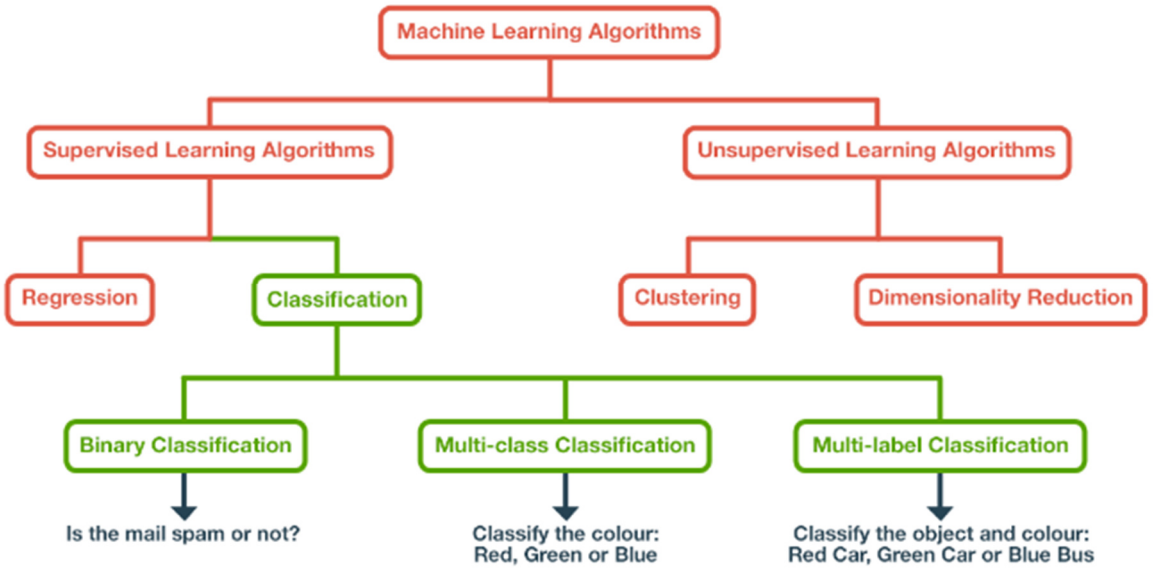
Las tareas o problemas principales que se pueden resolver con la minería de datos son la clasificación, agrupamiento y regresión; la selección de cual algoritmo

usar depende el tipo de variable de salida y el tipo de problema a solucionar. La regresión y la clasificación son algoritmos de aprendizaje supervisado mientras que el agrupamiento es no supervisado. La principal diferencia entre los algoritmos de regresión y clasificación es que los algoritmos de regresión se usan para predecir valores continuos como precio, salario, edad etc., y los algoritmos de clasificación se usan para predecir o clasificar valores discretos como hombre o mujer, verdadero o falso, spam o no spam etc.

2.11.1 Clasificación

La clasificación es una técnica en la que se categorizan los datos en un número determinado de clases. El objetivo principal de un problema de clasificación es identificar la categoría o clase a la que pertenecerá un nuevo dato. La clasificación es una técnica de aprendizaje supervisada en donde el algoritmo aprende del conjunto de datos y luego clasifica la nueva observación en una clase o grupo. A diferencia de la regresión, la variable de salida de una clasificación es una categoría o variable categórica, no un valor. La clasificación puede ser de tres tipos: clasificación binaria, clasificación multiclase y clasificación multietiqueta.

Figura 4 Algoritmos de clasificación



Fuente: (Bozkurt, 2021)

Un tipo de clasificación ampliamente estudiada es la binaria, es decir, con problemas en datos pertenecientes a dos clases que se relacionan entre sí como verdadero o falso. La clasificación por ser una técnica supervisada suele ser más precisa que las no supervisadas.

2.11.2 Regresión

Al igual que la clasificación es una tarea de aprendizaje inductivo, la regresión se basa en la observación y análisis de una variable. Los modelos de regresión predicen valores numéricos en lugar de etiquetas de clase discretas. Este tipo de modelos utilizan las características de los datos de entrada (variables independientes) y sus valores de salida son numéricos continuos (variables dependientes o de resultado) en donde aprenden de una asociación específica entre las entradas y las salidas correspondientes.

2.12 Tipología de algoritmos

Los algoritmos de minería de datos se dividen en dos grandes grupos: el aprendizaje supervisado y el aprendizaje no supervisado. El primero se refiere a la predicción con intervención humana, en donde los datos tienen algún tipo de etiquetado y el segundo no, donde los datos no tienen ningún tipo de etiqueta o clasificación previa.

2.12.1 Métodos supervisados

El método supervisado basa su proceso de aprendizaje en comportamientos o características (etiquetas) que se han visto en los datos ya almacenados. Por etiqueta se entiende la ocurrencia y relación del comportamiento de todos los atributos con un atributo especial que se conoce como atributo objetivo. Esto le permite al algoritmo predecir el atributo objetivo a un nuevo set de datos.

La dos grandes familias de algoritmos de aprendizaje supervisado son:

- **Algoritmos de clasificación**, indicados cuando el atributo objetivo es categórico, es decir, predice una categoría. Un ejemplo de clasificación es la identificación de correos spam.

- **Algoritmos de regresión**, indicados cuando el atributo objetivo es numérico, es decir, una regresión predice un número. Un ejemplo de

regresión es cuál va a ser el precio de un artículo o el número de reservas que se harán en un hotel.

2.12.2 Métodos no supervisados

El aprendizaje no supervisado usa algoritmos que basan su proceso de entrenamiento en datos históricos que no están etiquetados o no cuentan con clases previamente definidas, y no se conocen atributos objetivos, ya sean numéricos o categóricos. El fin es explorarlos para encontrar alguna relación, estructura o forma de organizarlos que finalmente permita agruparlos, también llamado *clustering o segmentación.*

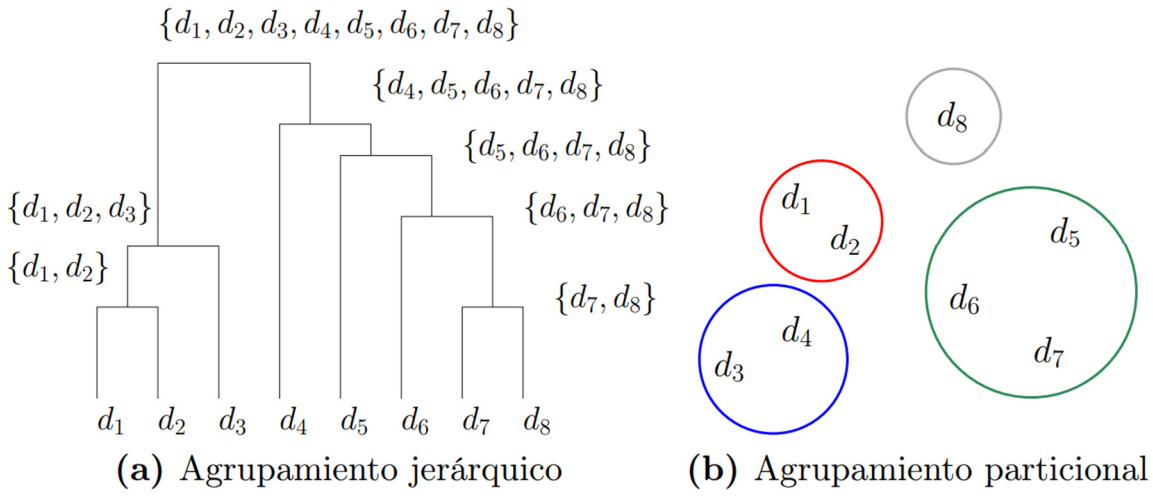
Hay dos algoritmos principales que permiten el agrupamiento:

1. **Los métodos jerárquicos**, es básicamente una técnica de agrupación no supervisada que implica la creación de grupos en un orden predefinido. Los grupos están ordenados de arriba a abajo. En este tipo de agrupamiento los clústeres similares se agrupan y se organizan de manera jerárquica tipo árbol, construyendo una estructura conocida como dendograma con distintos niveles de agrupación. El costo computacional del algoritmo es superior al que genera el agrupamiento particional, pero el dendograma que se obtiene es más rico que una partición, ya que se pueden generar varias particiones con solo variar el nivel de corte de la estructura jerárquica. Se divide en dos tipos llamados *agrupación jerárquica*

aglomerativa y *agrupación jerárquica divisiva*, y estas agrupaciones agrupan los pares de agrupaciones de los objetos de datos que hay en la jerarquía.

2. **Métodos particionales o no jerárquicos**, generan grupos de instancias que no responden a una organización jerárquica, el objetivo es partir o fusionar los objetos en grupos o clústeres de manera que el objeto pertenezca a uno de los clústeres posibles y en donde los clústeres sean disjuntos (no tienen elementos en común). Esta técnica agrupa los datos para maximizar o minimizar algunos criterios de evaluación.

Figura 5 Métodos de agrupamiento



(a) Agrupamiento jerárquico **(b)** Agrupamiento particional

Fuente: (Gironés et al., 2017)

2.13 Datos Faltantes

La falta de completitud de datos de dentro de cualquier estudio trae consigo implicaciones muy importantes para su análisis. Sin lugar a duda, la pérdida de datos necesariamente conlleva a la pérdida de información y a una menor precisión en la estimación de los parámetros de interés (Molenberghs et al., 2015).

En el análisis de datos una de las principales tareas es la identificación y tratamiento de la inexistencia de datos (datos faltantes o “missing data”). Entre los efectos que la ausencia de datos puede generar tenemos:

- Pueden limitar la capacidad para realizar tareas de ciencias de datos, en temas de conversión o visualización de datos.
- Puede reducir la validez estadística de los modelos, aumentando la probabilidad del error tipo II o también llamado error de tipo beta (β), que es cuando el investigador no rechaza la hipótesis nula siendo esta falsa en la población. (Aguilar Márquez et al., 2010).
- Los datos faltantes pueden reducir la representatividad de las muestras del conjunto de datos.
- Los datos que faltan pueden distorsionar la validez de los modelos de análisis y arrojar conclusiones no válidas.

Retomando la tarea de identificación y tratamiento de la falta de datos, los pasos a desarrollar son:

- Evaluar la existencia de valores perdidos (exploración).
- Excluir los valores ausentes.
- Recodificar los valores ausentes (imputación).

Es necesario identificar patrones de ausencia entre variables. El número y el patrón de ausencia nos ayudan a determinar la probabilidad de que sea aleatorio en lugar de sistemático y permitan definir las acciones que pueden llevarse a cabo:

- Eliminar las filas completas que contengan al menos un valor ausente.
- Eliminar las filas que contengan datos ausentes en alguna variable considerada clave para el análisis.
- Cambiar los valores ausentes a otro valor.
- Modificar los valores ausentes a un valor predeterminado.
- Realizar imputación de datos.

Si la cantidad de datos faltantes es muy pequeña en relación con el tamaño del conjunto de datos, la mejor estrategia para no sesgar el análisis puede ser omitir las pocas muestras con características faltantes. Las columnas de datos con demasiados valores faltantes no serán de mucho valor. Algunos especialistas en análisis de datos mencionan que la mejor estrategia es no contar con datos ausentes.

Sin embargo, la acción a tomar depende de cada situación, el omitir los puntos de datos disponibles priva al set de datos de cierta cantidad de información, y en donde es necesario hacer un análisis de valores perdidos con miras a buscar otras soluciones antes de eliminar puntos de datos potencialmente útiles en el conjunto de datos. Si bien algunas soluciones rápidas como la sustitución por promedios pueden estar bien en algunos casos, estos enfoques generalmente introducen sesgos en los datos.

2.13.1 Clasificación rápida de datos faltantes

Se identifican tres mecanismos que establecen datos ausentes:

1. MCAR (**Missing Completely at Random)** Cuando la probabilidad es estar ausentes siempre es la misma en todos los casos, es decir existen datos perdidos en forma **completamente aleatoria**. Este es el escenario ideal en caso de que falten datos.
2. MAR (**Missing at Random)** Se identifica una relación sistemática en los datos observados y la tendencia a los valores ausentes. Si la probabilidad de que falten es la misma solo dentro de los grupos definidos por los datos observados entonces los datos faltan al azar (MAR). MAR es más general y realista que MCAR. Los métodos modernos de datos faltantes generalmente parten de la suposición MAR.
3. MNAR (**Missing Not at Random)** Es la perdida de datos **no aleatoria**. Existe una relación entre la tendencia de perdida de datos y sus valores. La

falta de datos no aleatorios es un problema más grave y en este caso sería conveniente verificar más el proceso de recopilación de datos y tratar de comprender por qué falta la información.

Suponiendo que los datos sean MCAR, demasiados datos faltantes también pueden ser un problema. Por lo general un umbral máximo seguro es el 5 % del total para grandes conjuntos de datos. Si faltan datos para una característica (columna) o muestra(fila) determinada de más del 5 % entonces probablemente deba omitir esa característica o muestra (Molenberghs et al., 2015).

2.14 Reducción de Dimensionalidad

Muchos modelos de análisis de datos constan de cientos de variables y tener un número tan grande de características puede generar varios problemas. Se tiende a pensar que entre más números de variables se mejora la capacidad predictiva de los modelos; sin embargo, al aumentar la dimensionalidad de los conjuntos de datos se pierde el valor o la importancia en las distancias, en el sentido de que, para una instancia de la muestra dada, la instancia más cercana y la instancia más lejana están a distancias muy similares. Con un conjunto de datos con alta dimensionalidad se puede recurrir a dos caminos:

- Transformar los datos manteniendo la dimensionalidad, para que las distancias entre puntos sean significativas.

- Reducir la dimensionalidad, manteniendo toda la información útil que sea posible.

Si se opta por reducir la dimensionalidad, esto se puede hacer de dos formas que en ocasiones se pueden combinar: Construcción de Características y Selección de Características (Charte, 2017).

Existen varias técnicas para reducir la dimensionalidad de acuerdo con el tipo de datos como se puede apreciar en la tabla 1.

Tabla 1 Algunas técnicas de reducción de dimensionalidad

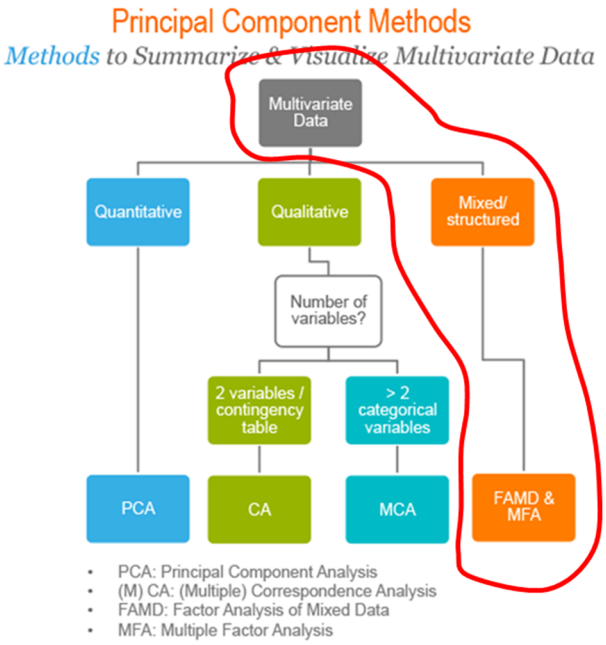
Tipo de Variables	Técnicas Recomendadas	Librerías
Numéricas	PCA, t-SNE	base, FactoMineR, Rtsne
Categóricas	CA, MCA	FactoMineR, ade4, epMCA
Mixtas (numéricas y categóricas)	FAMD, MFA	FactoMineR

Fuente: Elaboración propia

2.15 Análisis Factorial de Datos Mixtos (FAMD)

El análisis factorial de datos mixtos es una técnica estadística para la reducción de atributos, utilizada para explicar las correlaciones entre variables en un set o grupo de datos que se describe tanto con variables cuantitativas como cualitativas, de ahí su termino de mixto. En general el algoritmo FAMD es la combinación del PCA y MCA, es decir aplicar el análisis de componentes principales (PCA por sus siglas en inglés) para variables cuantitativas y el análisis de correspondencias principales (MCA por sus siglas en ingles) para variables cualitativas. En otras palabras, actúa

como PCA para variables cuantitativas y como MCA para variables cualitativas. Las variables cuantitativas y cualitativas se normalizan durante el análisis para equilibrar la influencia de cada conjunto de variables.



Fuente: (Charte, 2017)

2.16 Boruta

Boruta es otro algoritmo que está diseñado para realizar automáticamente la selección de variables de un conjunto de datos. Se origino como un paquete para R. Boruta es un método que utiliza Random Forest como algoritmo encubierto. El algoritmo genera en cada iteración una variable sombra, es decir, las variables no compiten entre sí, en cambio, compiten con una versión aleatoria de ellas. Se ajusta el modelo por Random Forest y se calculan las importancias relativas de

cada variable. Si una variable sistemáticamente queda por debajo de las sintéticas (ruido), será indicativo de que su aportación al modelo será dudosa y por tanto se elimina. El proceso es continuo hasta que todas las variables son aceptadas, rechazadas o se alcanza un número de iteraciones límite (Mazzanti S., 2020)

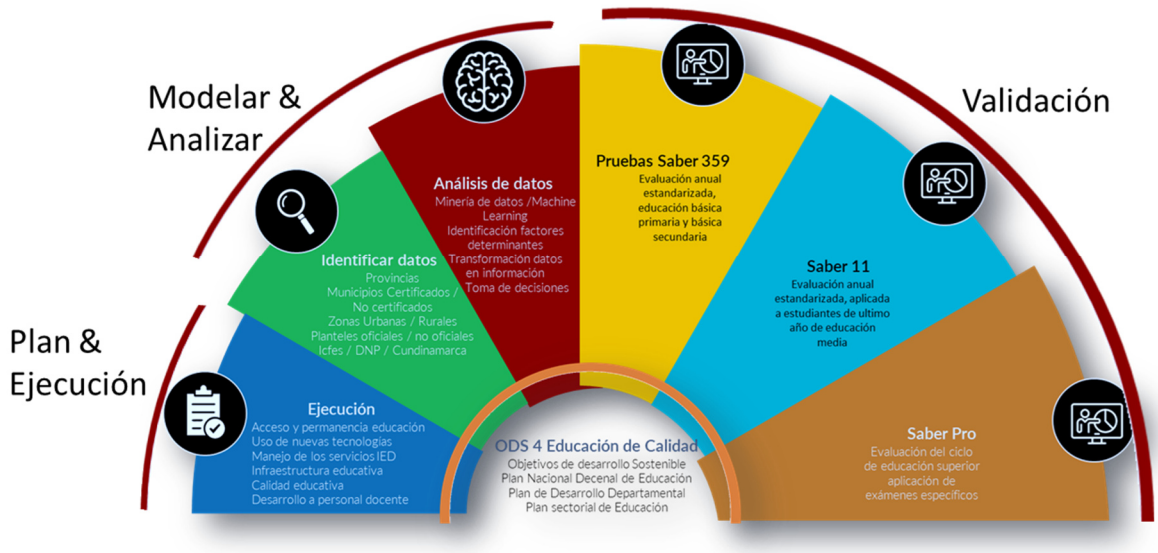
3 Metodología

3.1 Análisis del problema (Entendimiento del negocio)

3.1.1 Determinar los objetivos del negocio

Para efectos de este trabajo, se plantea implementar la primera versión de un marco de análisis de datos replicable, que permita cumplir con los objetivos de la secretaria de educación del departamento de Cundinamarca (ODS, planes etc.). Se busca generar un **ciclo continuo de ejecución y mejora**, que se alimenta de los datos del desempeño académico de los estudiantes pertenecientes a las entidades educativas oficiales del departamento en las pruebas Saber 11°, prueba que mide las competencias básicas definidas como referentes de calidad en Colombia. Este proceso se realiza con la meta de construir un proceso que identifica, modela y analiza los datos, validando la confiabilidad de estos modelos con pruebas del estado Saber 11°.

Figura 6 Modelo Cíclico para análisis de la calidad educativa



Fuente: elaboración propia

Esta investigación plantea el primer modelo, que incluye el análisis de los resultados de las pruebas Saber 11° entre 2017 y 2021, para los planteles oficiales, tanto urbanos como rurales, de los municipios no certificados de Cundinamarca.

3.1.2 Metodología de desarrollo

Este estudio se desarrolló bajo la metodología CRISP-DM que describe seis fases descritas en la sección 3.3 de este documento (*entendimiento del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue del modelo*)(Gironés et al., 2017).

En la fase de “Entendimiento del negocio”, se analiza la información que genera Instituto Colombiano para la Evaluación de la Educación Superior (ICFES) de las dos pruebas que se realizan por año entre los años 2017 y 2021.

En la fase “Comprensión de los datos”, se analizan y normalizan las fuentes de datos, identificando las variables cuantitativas y cualitativas de los datos que describen a cada uno de los estudiantes.

En la fase “Preparación de los datos”, se limpian, unifican, transforman variables, se imputan los valores nulos, se eliminan valores atípicos y se hace una selección de variables (reducción de dimensionalidad) en busca de todos los factores que pueden influir en la calidad educativa del departamento. Desde esta fase y las siguientes el proceso de análisis se torna bidireccional, es decir, permite retornar a la etapa anterior y ajustar los datos y los modelos en caso de que los resultados no cumplan con el objetivo del proyecto.

En la fase “Modelado”, se diseñan y prueban modelos de clasificación de estudiantes para, finalmente, hacer una predicción de acuerdo con las variables seleccionadas.

En la fase de “Evaluación y despliegue del modelo” se prueba el modelo, seleccionando varios conjuntos de datos en donde se mide la confiabilidad y certeza de la predicción, para concluir sobre los resultados que arrojan las diferentes corridas del modelo analítico y que permitan estructurar la información para uso del usuario final y poder influir en la toma de decisiones de políticas públicas.

3.1.3 Valoración de la situación

Para esta investigación se cuenta con las bases de datos de las diferentes pruebas de estado que realiza el ICFES para medir las competencias básicas que se encuentra formuladas por el Ministerio de Educación Nacional en los Estándares Básicos de Competencias; competencias que desarrollan los estudiantes a lo largo de su tránsito por el sistema educativo colombiano. Se encuentran pruebas para los niveles de educación básica (Saber 3°, Saber 5°, Saber 7°), media (Saber 7°, Saber 9°, Saber 11°) y superior (Saber TyT, Saber Pro) que buscan los siguientes objetivos:

- Comprobar el grado de desarrollo de las competencias de los estudiantes.
- Monitorear la calidad de la educación de los establecimientos educativos del país.
- Ser fuentes de información para la construcción indicadores y ejercicios de análisis, inspección y vigilancia del servicio público educativo.
- Ser referente estratégico para el establecimiento de políticas educativas.

- Proporcionar información a las instituciones educativas sobre las competencias de los estudiantes que aspiran a programas de educación superior.

Entre los beneficios que se obtienen de esta investigación está la identificación de las variables más determinantes y con mayor confiabilidad que influyen en el desempeño individual de los estudiantes; generando un método que permita replicar el proceso y poner en producción el modelo y la visualización de los datos. Se espera que la información resultante de este modelo sirva como referente estratégico para el establecimiento de políticas públicas educativas (consultas de funcionarios, reformas realizadas). Suministrando información adicional a los establecimientos educativos oficiales del departamento acerca de sus prácticas pedagógicas, para monitorear la calidad de la educación de los establecimientos educativos y el establecimiento de nuevos indicadores de valor agregado o el seguimiento de la eficiencia de los ya establecidos.

3.1.4 Determinar los objetivos de Data Mining

Las metas del proceso de minado son:

- Lograr los niveles adecuados de confiabilidad con los algoritmos para el tratamiento de valores ausentes para el conjunto de datos seleccionado de 67.049 registros de estudiantes.
- Identificar las principales variables o atributos que influyen en el desempeño académico de los estudiantes en las pruebas de estado Saber 11°.
- Predecir el desempeño académico de un estudiante al presentar la prueba Saber 11°, clasificándolo en uno de los tres grupos definidos para esta investigación.

Se estableció como criterio de éxito en esta investigación de minería de datos, desarrollar modelos de aprendizaje que permitan predicciones con una

confiabilidad superior al 80%. Este porcentaje de confiabilidad dependerá del algoritmo que se seleccione, la limpieza y transformación que se haga del conjunto de datos, tema que se aborda en el numeral que detalla la evaluación del modelo.

3.1.5 Realizar el plan de proyecto

Para darle un orden lógico al desarrollo de esta investigación, el trabajo se dividió en las siguientes etapas para facilitar su organización, secuencia de pasos y favorecer la curva de aprendizaje del proceso de minado de datos:

1. Identificación de las fuentes de datos y la selección de datos de acuerdo con el alcance de la investigación (años 2017 al 2021).
2. Filtros sobre el conjunto de datos seleccionado para lograr las muestras más representativas (Filtros por departamento, municipio y entidades educativas oficiales).
3. La transformación de los datos, eliminando atributos que no aportan información. Atributos tipo identificador, atributos constantes etc.
4. Preparar los datos ejecutando procesos de limpieza (valores, faltantes ausentes, conciliación de valores, valores atípicos).
5. Creación de dos escenarios: 1. Uno sin valores faltantes 2. Con datos imputados
6. Selección y pruebas de los modelos o técnicas de aprendizaje.
7. Generación y análisis de reportes obtenidos de los modelos a la luz de los objetivos de la investigación y las métricas de éxito definidas.
8. Presentación de resultados.

Durante esta investigación se utilizaron y evaluaron varios métodos, algoritmos y herramientas; todo con el objetivo de recomendar el mejor proceso para abordar el problema a solucionar y que facilite su replicación con nuevos conjuntos de datos. Por temas de disponibilidad de herramientas, algoritmos y facilidad de uso, se

utilizaron varias herramientas open source o de uso libre. Varias de ellas se aplicaron dependiendo de la fase de desarrollo en la que se encontraba el proyecto, así pues, en las fases de estructurar y limpiar los datos se utilizaron herramientas como Open Refine, R, y Knime. Por otro lado, al alcanzar las fases de modelado y evaluación del modelo se utilizaron herramientas como Weka, Orange, Knime y R.

Con respecto a las técnicas que se emplearon en los modelos de aprendizaje de esta investigación se utilizaron algoritmos supervisados y no supervisados, igualmente aplicados dependiendo de la fase de ejecución en la que se encontraba el proyecto:

- No supervisados
    - Reducción de dimensionalidad
- Supervisados
    - Clasificación
    - Regresión

3.2 Comprensión de los datos (análisis de datos)

En esta segunda fase de la metodología CRISP-DM se hace la recolección inicial de los datos que van a ser objeto del estudio. En esta actividad se define el tamaño de la muestra, se hace el primer reconocimiento del problema, se identifican las relaciones más evidentes y se proponen las primeras hipótesis.

3.2.1 Recolectar los datos iniciales

3.2.1.1 Sistema Nacional de Evaluación Estandarizada de la Educación

Desde la primera aplicación de la prueba Saber 11 º en 1968, el examen ha sufrido varios cambios a lo largo de los años, una en los años 80, una más en el

año 2000 y otra en el año 2014, pasando de ser un examen que solo se usaba como requerimiento para el ingreso a las instituciones de educación superior, a ser un medio para medir la calidad educativa de las instituciones educativas. Posteriormente, evoluciona para medir la calidad educativa por la adquisición de competencias y no la capacidad de memorizar del estudiante, es decir, no sólo mide el conocimiento del estudiante sino cómo sabe aplicar ese conocimiento.

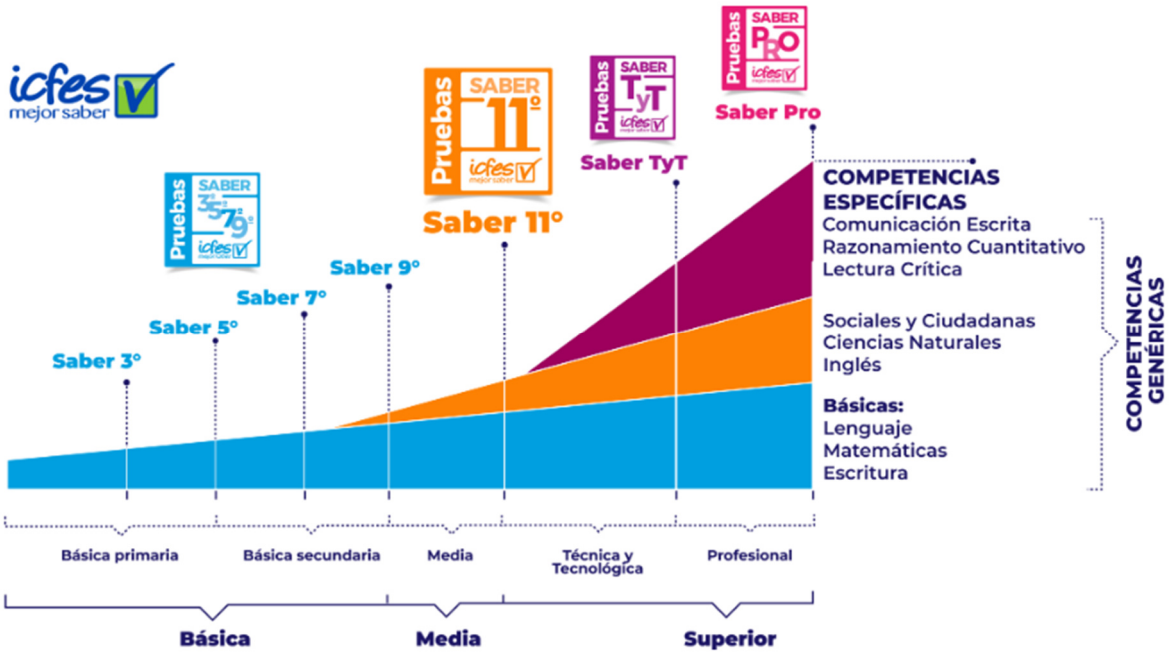
En el 2014, se tomaron decisiones en materia curricular y pedagógica, y se pasó de evaluar Lenguaje, Matemáticas, Biología, Física, Química, Ciencias Sociales, Filosofía e inglés a las cinco pruebas que hoy en día tiene el examen y se consolida Sistema Nacional de Evaluación Estandarizada.

Figura 7 Cambios en la prueba Saber 11°

Fuente: (ICFES, 2021b)

Las pruebas Saber evalúan el desempeño alcanzado por los estudiantes según las competencias básicas definidas por el Ministerio de Educación Nacional. Estas pruebas evalúan los resultados de los estudiantes al final de los ciclos de la educación básica y media. Saber 3º y 5º en la básica primaria, Saber 9º en el cierre de la educación básica secundaria, y Saber 11º al término de la educación media.(ICFES, 2021a)

Figura 8 El Sistema nacional de Evaluación Estandarizada



Fuente: (ICFES, 2021b)

3.2.1.2 Fuentes de información

Las fuentes de información básicamente se pueden clasificar en dos grupos: las fuentes primarias y las secundarias. Las fuentes primarias son las que se toman

como primera instancia, es decir son la fuente principal que ese usa en la investigación para obtener información. Las fuentes secundarias fueron analizadas y procesadas previamente por otras personas o instituciones.

Fuentes primarias: Son todos los datos y documentos que se encuentran en el sitio compartido en la página Web del ICFES llamada Data Icfes, portal destinado para publicar toda la información relacionada con las pruebas Saber (Saber 11, Saber TyT, Saber Pro Saber 3,5,9), diccionarios de datos, guías etc.

Fuentes secundarias: Información de: el Ministerio de Educación Nacional (MEN), el Departamento Administrativo Nacional de Estadística (DANE), el Instituto Colombiano para la Evaluación de la Educación Superior (ICFES), artículos de revistas científicas, estudios relacionados al desempeño académico usando minería de datos en entornos universitarios y de educación media.

3.2.1.3 Datos de la prueba Saber 11°

Las pruebas Saber 11° son una unidad de medida o un referente para ingresar a la educación superior; a pesar de que son el indicador principal, pero no el único, del aprendizaje de los estudiantes y de la calidad de la educación en general, partiendo del supuesto que estas predicen el desempeño académico del estudiante en la formación superior. Evalúan el grado de desarrollo de las competencias de los estudiantes que están por finalizar el grado undécimo de la educación media, las bases de datos de estas pruebas están disponibles en un

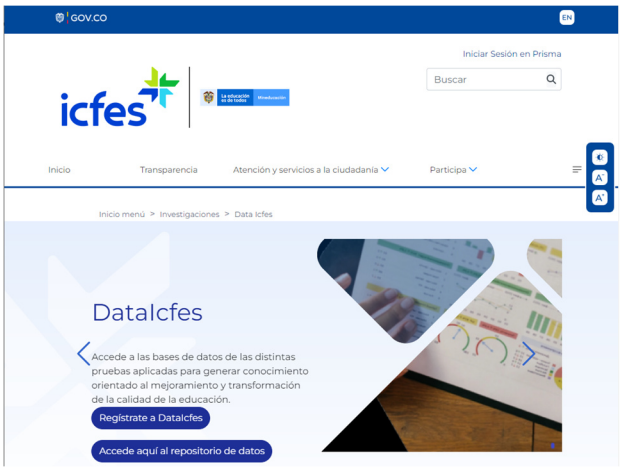
portal web de datos abiertos del ICFES que se llama “DataIcfes” y se encuentra ubicado en la URL https://www2.icfes.gov.co/data-icfes al cual se puede acceder mediante un usuario y contraseña entregado por el ICFES, previa solicitud formal por parte del investigador.

3.2.1.4 Portal DataIcfes

Es un espacio de almacenamiento en la nube en donde se encuentra la información de todas las pruebas que evalúa el ICFES; se accede solicitando una contraseña y el uso de la información se permite con fines investigativos. Se encuentran dispuestos los datos que se capturan durante la etapa de inscripción y los datos socioeconómicos que se diligencian durante la prueba.

En el sitio se alojan los datos de los resultados de las diferentes pruebas de estado Saber 11° desde el año 2000 a nivel de estudiante, se encuentran dos archivos por año con los resultados agrupados de acuerdo con los calendarios escolares A y B, por lo general en el primer semestre se evalúa el calendario “B”, en donde el grupo es más pequeño, comparado con el grupo del segundo semestre donde se evalúa el calendario “A”. Es importante resaltar que los datos que se consultan de este portal no permiten individualizar o perfilar a los titulares de la información que se ve reflejada en los conjuntos de datos.

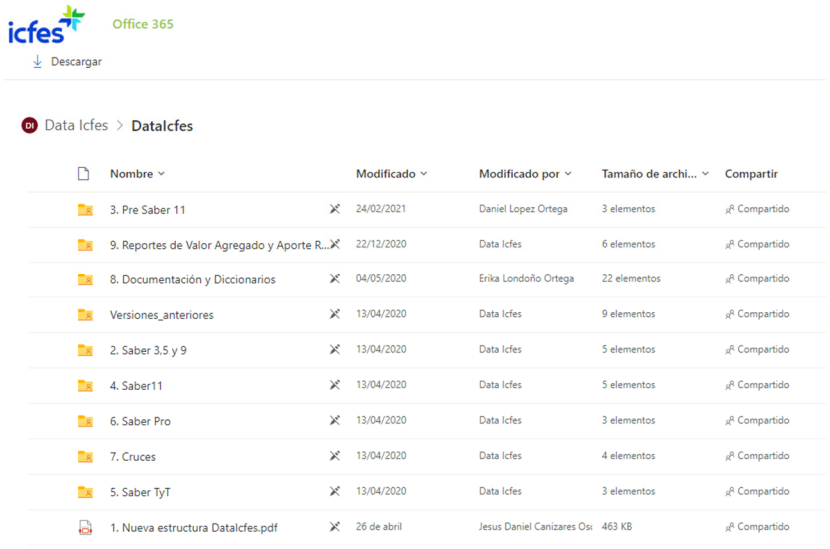
Figura 9 Portal web DataIcfes



Fuente: ICFES

Después de ingresar al portal se tiene acceso a la estructura de directorios con información relacionada a las pruebas Saber, como se puede apreciar en la siguiente figura:

Figura 10 Estructura de directorios portal web DataIcfes



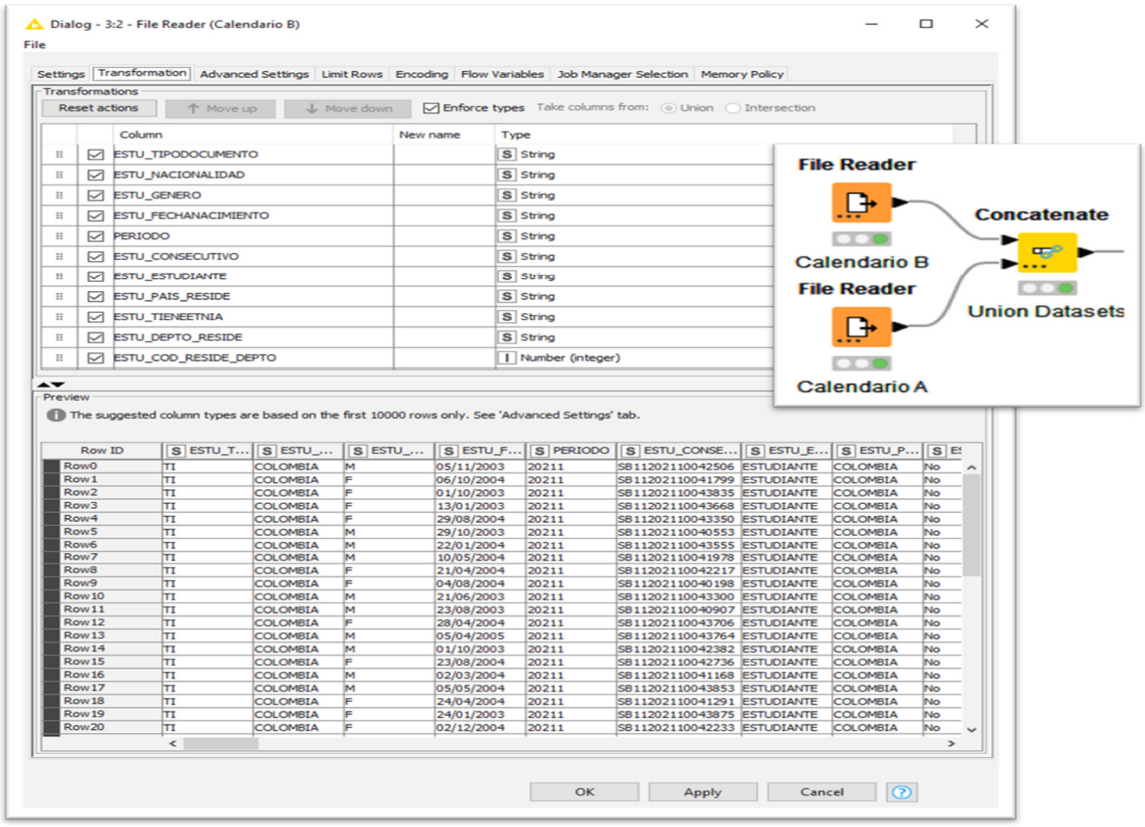
Fuente: Portal web DataIcfes

3.2.1.5 Extracción de los datos

Después de tener identificada la fuente de datos se procede a descargar los archivos con los resultados correspondientes a las pruebas Saber 11° entre los años 2017 y el 2021, es decir la información de cinco años. Para el objeto de esta investigación se utilizan 10 bases de datos, dos conjuntos de datos por año diferentes, uno con los resultados del calendario A y el otro con los resultados del calendario B.

Las bases de datos publicados por el ICFES en el repositorio DataIcfes, se encuentran en archivos de texto plano extensión ".txt". El proceso de visualización o apertura de los conjuntos de datos se puede hacer con varias herramientas como Microsoft Excel, Power BI o un block de notas; en donde es necesario tener identificado el separador o delimitador de cada uno de los campos para este tipo de archivos que puede estar entre los más comunes coma, punto y coma, dos puntos. Para los grupos de datos de esta investigación el separador utilizado es "¬", que se logra con la combinación de las teclas "alt gr" y la tecla de grado "°", y los archivos se procesaron e integraron usando la herramienta Knime.

Figura 11 Carga e integración de las fuentes de datos con KNIME



Fuente: elaboración propia

3.2.1.6 Universo de los datos y la muestra

De acuerdo con la evaluación de las variables reportadas en el cuestionario socioeconómico para cada uno de los años desde 2015 a 2021, se encontró que el ICFES realizó modificaciones sustanciales a la estructura del cuestionario, lo que causó que los años 2015 y 2016 fueran excluidos del estudio, por reportar variables socioeconómicas que a partir del año 2017 no se continuaron recolectando.

En esta investigación, se utilizaron los datos que aparecen en la página web del Instituto Colombiano para la Evaluación de la Educación Superior (ICFES) y se hizo el primer filtro en donde se tomaron los resultados de los estudiantes que presentaron la prueba Saber 11° entre los años 2017 al año 2021 incluyendo los dos calendarios escolares A y B. Construyendo una base de datos original que contiene 2.757.824 registros, Tabla 2.

Tabla 2 Fuentes de datos – Resultados Saber 11° años 2017-2021

Fuente de Datos	No. Atributos	No. Registros
SB11_20171	81	12,993
SB11_20172	82	546,261
SB11_20181	82	12,527
SB11_20182	83	549,934
SB11_20191	82	21,083
SB11_20192	82	546,212
SB11_20201	81	15,435
SB11_20202	81	504,872
SB11_20211	78	15,528
SB11_20212	82	532,979
Total		**2,757,824**

Fuente: Elaboración propia

El diccionario de datos de las variables del conjunto de datos se puede ver en el Anexo 2. Diccionario de datos pruebas Saber 11°

3.2.1.7 Descripción de los datos

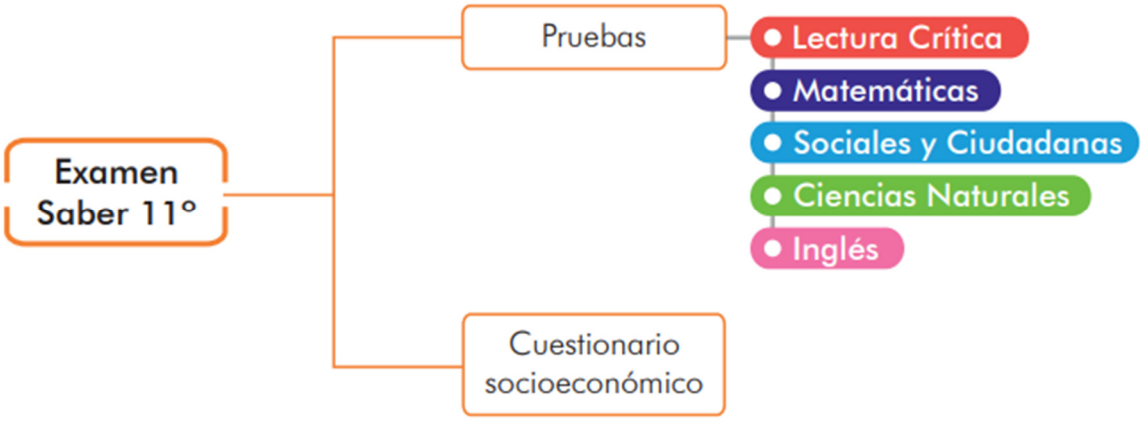
En las pruebas Saber 11° se evalúan los procesos de enseñanza y aprendizaje de los estudiantes en cinco áreas generales, para un total de 254 preguntas, en la Tabla 3 se aprecian el número de preguntas para cada prueba:

Tabla 3 Áreas generales que se evalúan en Saber 11°

Área para evaluar	No. de preguntas
Matemáticas	50
Lectura Crítica	41
Sociales y ciudadanas	50
Ciencias Naturales	58
Inglés	55
Total	**254**

Fuente: Elaboración propia

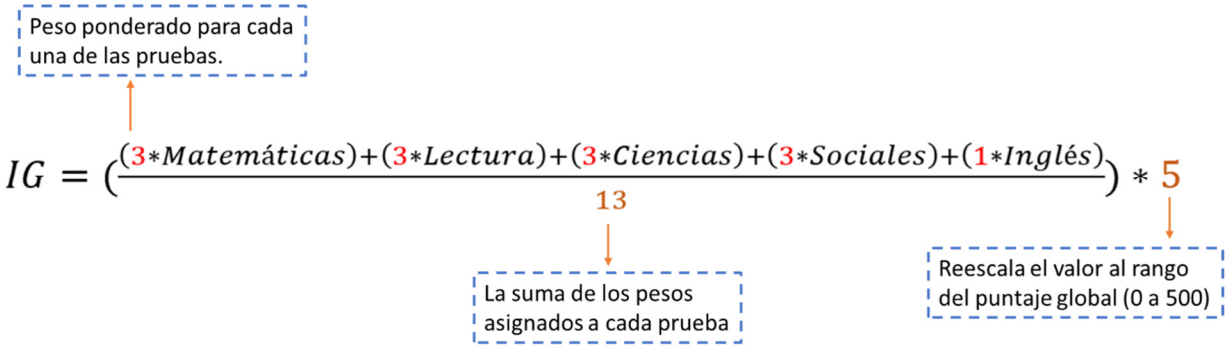
Figura 12 Estructura del examen Saber 11° (ICFES, 2022)



Fuente: (ICFES, 2022)

Se generan resultados a nivel individual de los estudiantes, en donde se evalúan cada una de las cinco pruebas genéricas antes mencionadas y se construye un puntaje global, construido a partir de un promedio ponderado de los puntajes en las cinco pruebas genéricas bajo la fórmula que se aprecia en la figura 13.

Figura 13 Calculo del puntaje global



Fuente: Elaboración propia

Adicional a las cinco pruebas presentadas los estudiantes deben responder un cuestionario socioeconómico de 24 preguntas, en donde se indaga tres aspectos generales como:

1. **Información familiar** del estudiante (como nivel educativo de los padres, su ocupación, servicios en el hogar como acceso a internet y servicios de televisión por cable) entre otros.
2. **Información académica** asociada al establecimiento educativo al que pertenece el estudiante, jornada, calendario, departamento y municipio de ubicación de la institución.
3. **Información personal** del estudiante donde se indaga lugar de residencia, el género, la edad y discapacidades del estudiante, entre otros.

Después del primer filtro se obtuvo un conjunto de datos con los resultados de los estudiantes entre los años de 2017 y 2021, originalmente conformado por 80

campos o variables y 2.757.824 registros. Con este grupo de datos se ejecutó un segundo filtro aplicando las variables que delimitan el alcance de este estudio, para obtener los resultados de los estudiantes pertenecientes a entidades educativas oficiales de los 108 municipios no certificados del departamento de Cundinamarca, generando una nueva una base de datos con 67.049 registros como se muestra en la tabla 4. Los municipios no certificados agrupados por regiones se pueden ver en el Anexo 1. Municipios no certificados.

Tabla 4 Resultados Saber 11° entidades educativas oficiales de municipios no certificados

Fuente de Datos	No. Atributos	No. Registros
SB11_2017	82	14155
SB11_2018	84	14010
SB11_2019	82	13656
SB11_2020	81	12704
SB11_2021	82	12524
Total		67049

Fuente: Elaboración propia

Para este nuevo grupo de datos fue necesario aplicar un proceso de transformación que incluyó:

- La eliminación las variables de tipo identificador tales como códigos de identificación ICFES, DANE, del estudiante, del municipio, del departamento, del colegio, de la sede etc.
- Inclusión del atributo EDAD que se genera a partir de la diferencia entre la fecha de presentación del examen y la fecha de nacimiento del estudiante.

- Adición de una variable categórica de tres estados (Insuficiente, Satisfactorio, Avanzado) que logró determinar el desempeño del estudiante, que es la variable objetivo o variable dependiente.
- Adición de una variable categórica con las 15 regiones que agrupan los 116 municipios del departamento de Cundinamarca.

La descripción detalla de las 80 variables del grupo de datos se pueden ver en el Anexo 2 Variables del conjunto de datos

3.2.2 Exploración de los datos (análisis del conjunto de datos)

El conjunto de datos con 80 variables tiene 48 variables categóricas y 32 variables numéricas. Una variable categórica es una variable con un número limitado de valores distintos o categorías (por ejemplo, jornada del colegio, nacionalidad del estudiante, género del estudiante, trabajo o labor del padre, educación de la madre etc.). Estas variables categóricas pueden ser **nominales, ordinales o booleanas**.

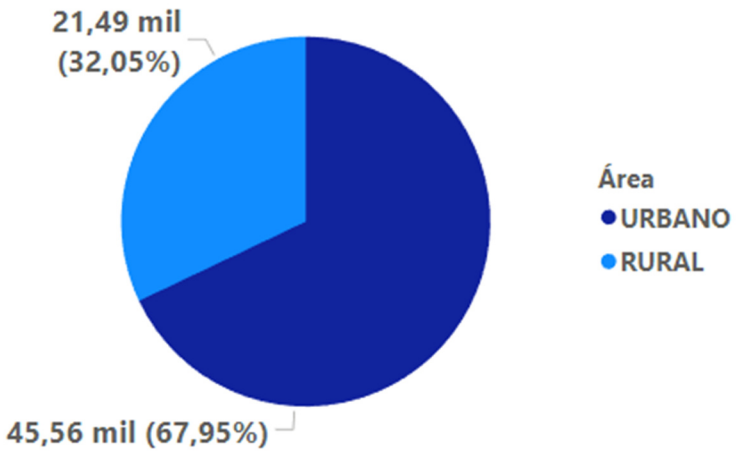
- **Variables nominales:** son aquellas que sus valores representan categorías sin clasificación propia o interna (por ejemplo, municipio de presentación del examen, área de ubicación del colegio etc.).
- **Variables ordinales:** son aquellas que sus valores representan categorías con alguna clasificación propia o interna (por ejemplo, puntaje de SISBEN, estrato de la vivienda, nivel socio económico del estudiante NSE etc.).

- **Variables booleanas o binarias:** son aquellas que sus valores solo pueden representar dos valores, que generalmente representa la presencia o ausencia de un concepto (por ejemplo, colegio bilingüe, la familia tiene automóvil, computador, servicio de internet etc.).

3.2.3 Analítica descriptiva de los datos

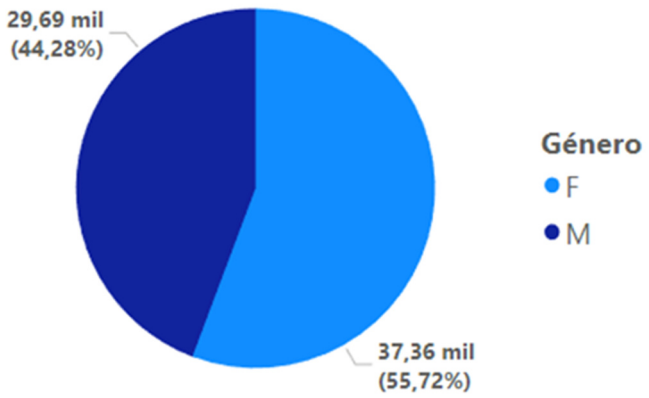
En esta sección se muestra una estadística descriptiva de algunas de las variables que intervienen en este estudio. En la figura 14 se puede observar que el 67.95% de los estudiantes se encuentran concentrados en la zona urbana y en la figura 15 se aprecia que el 55.72% de estos estudiantes son mujeres.

Figura 14 Área de ubicación del colegio



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

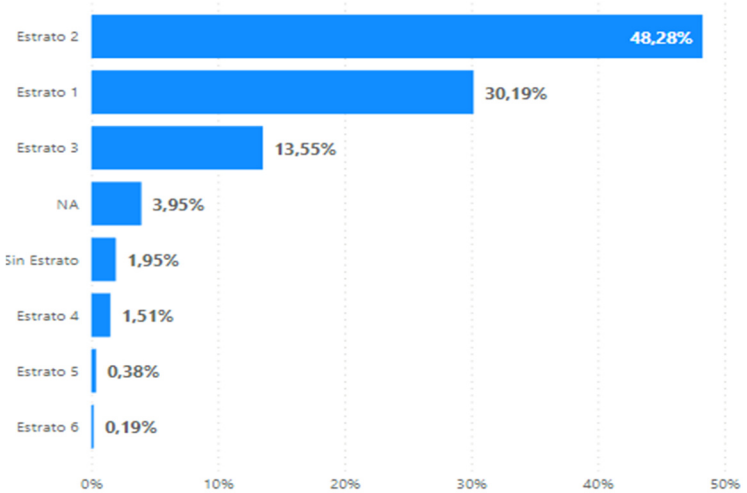
Figura 15 Género de los estudiantes del estudio



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

Del mismo modo, en la figura 16, se observa que en los estratos 1, 2 y 3 se encuentran concentrada la mayor parte de la muestra, equivalente al 86.02% y en donde solo el estrato 2 representa el 48.28% de la población estudiantil.

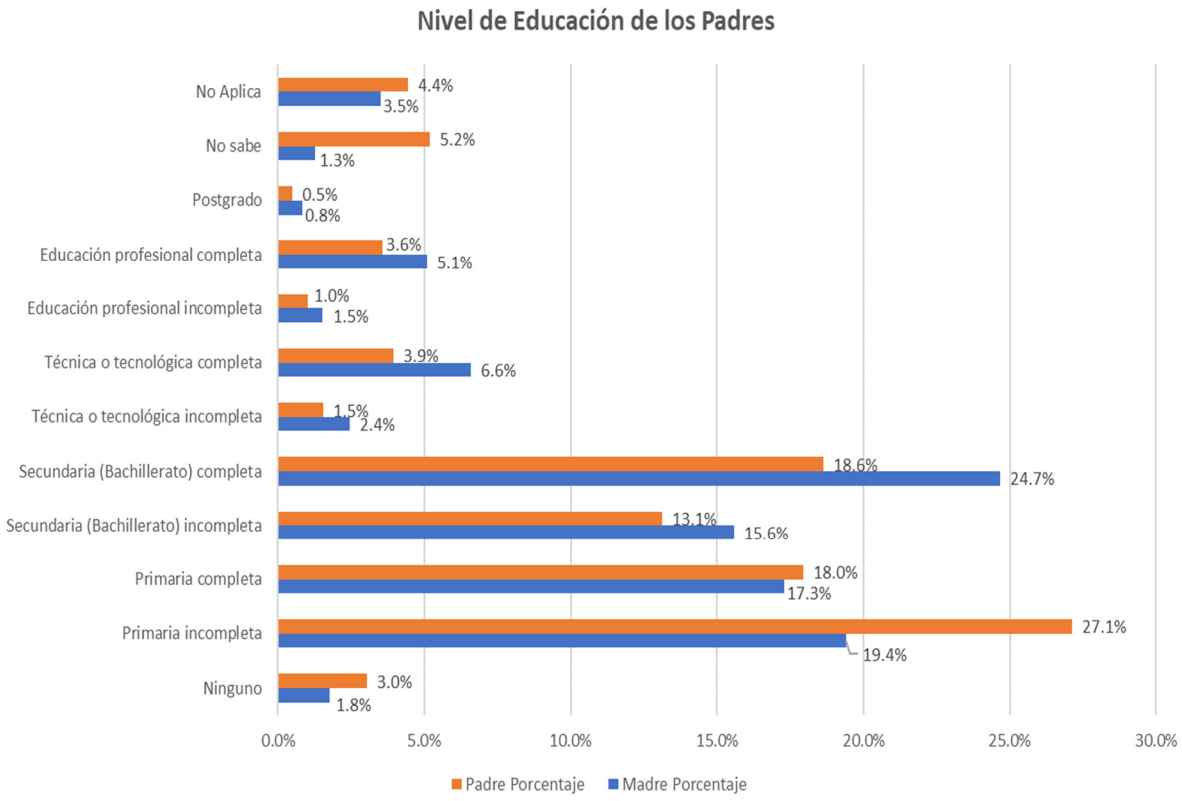
Figura 16 Estrato socioeconómico



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

En la figura 17 también se puede apreciar el nivel educativo de los padres en donde para este estudio, el 77% de los padres tienen un nivel educativo igual o inferior a la secundaria completa (Bachillerato) y que solo el 6% de las madres y el 4% de los padres lograron niveles superiores de estudio (Educación profesional completa y postgrado).

Figura 17 Nivel de educación de los padres



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

El NSE va de una escala de NSE 1 hasta NSE 4, donde el primer nivel hace referencia a niveles socioeconómicos bajos y se incrementa hasta el cuarto nivel, que corresponde a niveles socioeconómicos altos (Icfes, 2019).

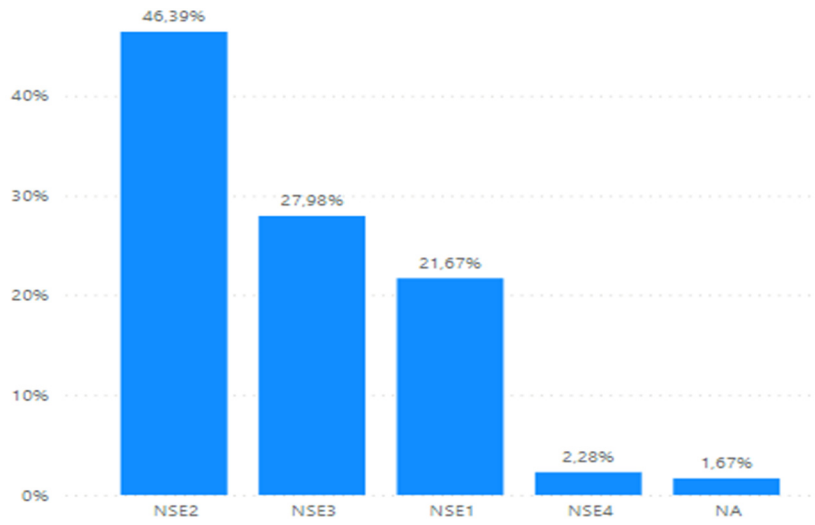
Tabla 5 Caracterización por NSE

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

Fuente: Icfes, 2019

Casi la mitad de la muestra objeto del análisis equivale al 46.39% y se encuentra en el NSE 2 (Nivel Socioeconómico) como se puede apreciar en la figura 18.

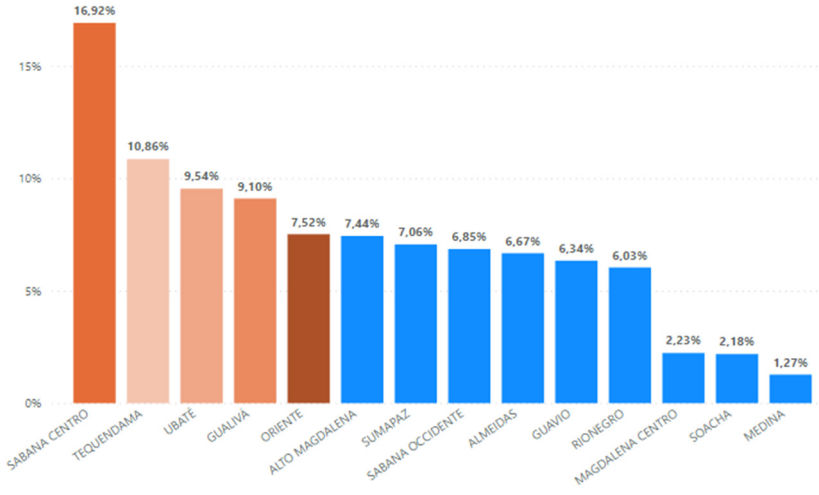
Figura 18 Nivel socioeconómico de los estudiantes



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

Utilizando el atributo de “Región” que se agregó al set de datos original se exploró la distribución de los estudiantes por región, en la figura 19 se evidencia que más del 50% de los estudiantes se encuentran concentrados en cinco (5) regiones de las 15 regiones que conforman el departamento de Cundinamarca.

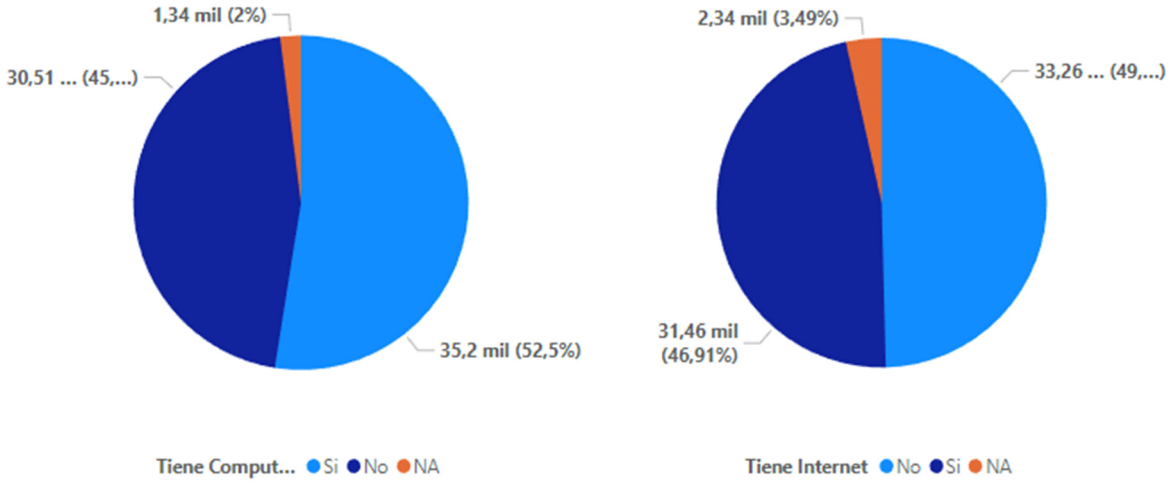
Figura 19 Concentración de estudiantes por regiones en Cundinamarca.



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

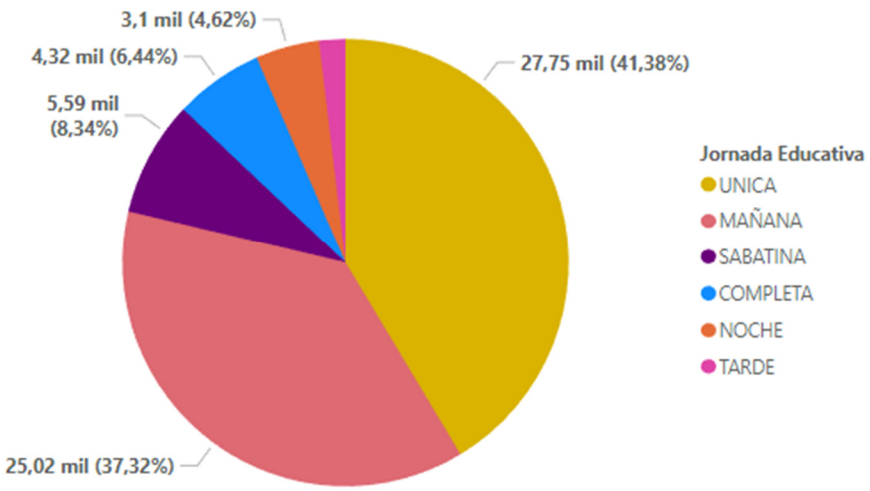
Adicionalmente se analizaron variables asociadas a la institución educativa, y al grupo familiar del estudiante que puedan llegar a reforzar su proceso de aprendizaje, tales como si el estudiante tiene acceso a un computador y a un servicio de internet. En la figura 20 se presentan que el 52.5% de los estudiantes tienen acceso a un computador y el 49.6% tiene acceso a un servicio de internet. De igual manera, en la figura 21 se refleja que el 41.38% de los estudiantes pertenecen a una jornada única y el 37.32% a la jornada de la mañana.

Figura 20 Acceso a servicios



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

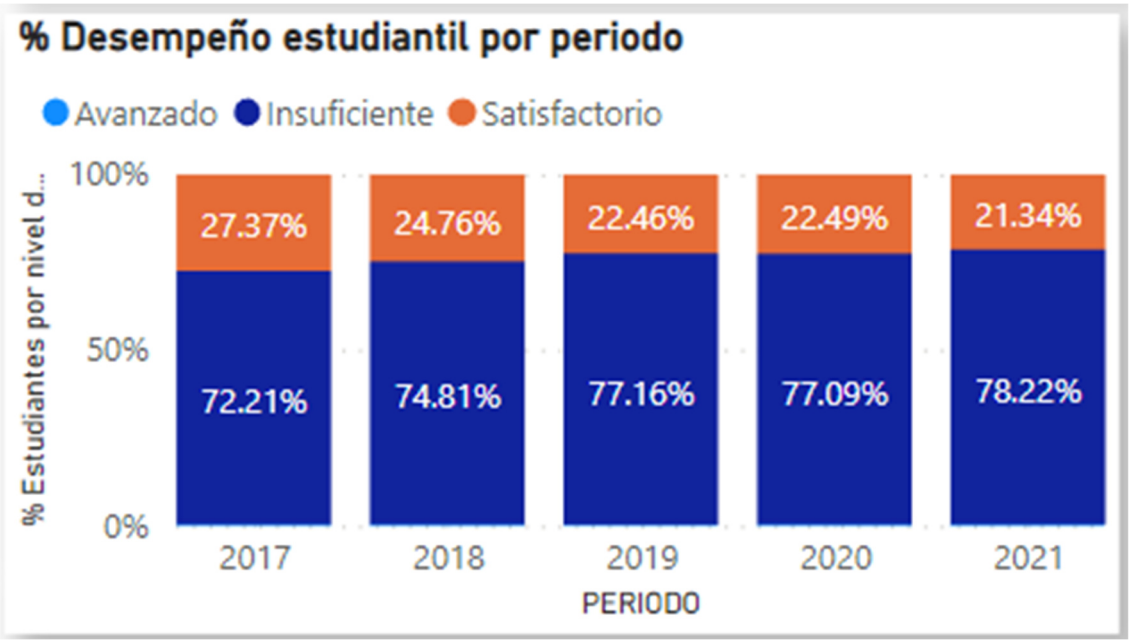
Figura 21 Jornada Instituciones educativas



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

También se explora el desempeño de los estudiantes a la luz de las tres categorías (Insuficiente, Satisfactorio y Avanzado) definidas como variable objetivo de este estudio. En la figura 22 se visualiza el desempeño de los estudiantes por año, en donde se evidencia la curva de crecimiento de la categoría insuficiente desde el año 2017 hasta el 2021.

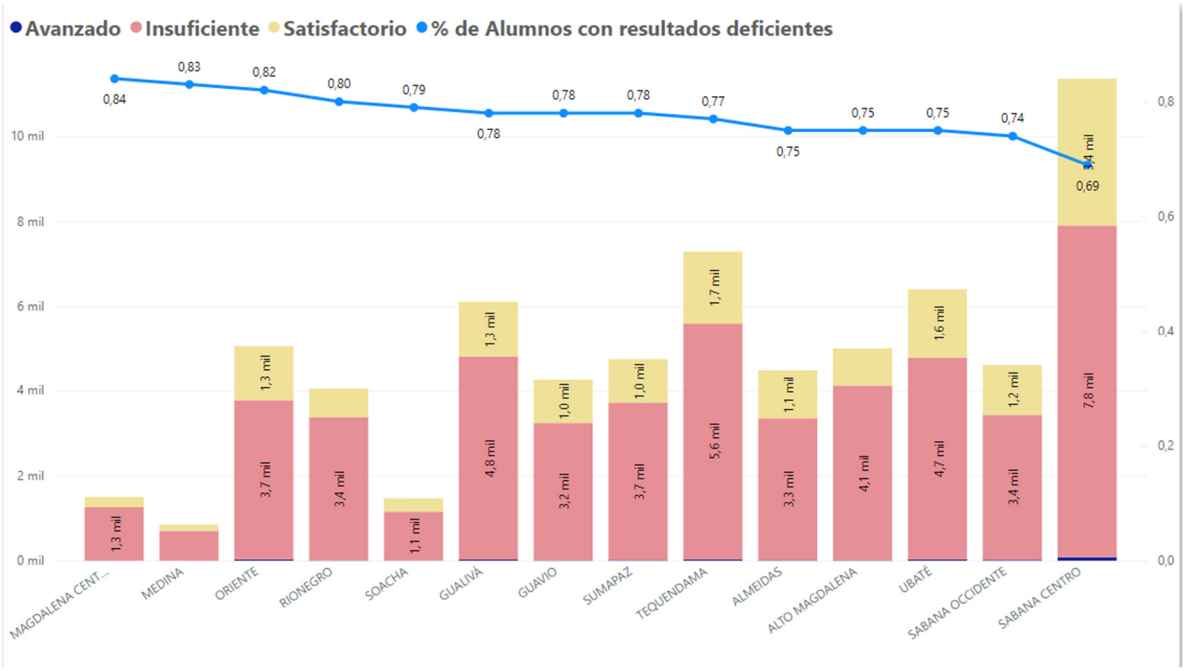
Figura 22 Desempeño estudiantil por año



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

En la figura 23, se visualizan la cantidad de estudiantes a nivel departamental (Cundinamarca), clasificados por región de acuerdo con su desempeño en las pruebas Saber 11°; en donde se visualiza que el 69% y el 84% (valor mínimo y máximo) de los estudiantes de las regiones “Sabana Centro” y “Magdalena Centro”, respectivamente, quedaron clasificados en la categoría "Insuficiente” en el periodo 2017-2021.

Figura 23 Desempeño por regiones de Cundinamarca en las pruebas Saber 11°



Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

3.2.4 Verificar la Calidad de los datos

Al momento de extraer información del conjunto de datos, la falta de calidad de datos representa uno de los problemas “ocultos” que ocupa un papel predominante en cualquier análisis de datos, y es necesario iniciar procesos de detección, eliminación o corrección de datos no válidos. Factores como los datos inconsistentes, errores humanos, valores faltantes, errores en la digitación, inconsistencias en los valores de los atributos, formatos inconsistentes, valores atípicos y extremos, caracteres extraños, etc., pueden causar ruido en los grupos de datos. Para el conjunto de datos seleccionado se realizaron las siguientes tareas:

- Discretizar los datos, que es el proceso de conversión de variables continuas a categóricas.
- Reducción de la dimensionalidad, que es el proceso de elegir los atributos determinantes en el análisis sin afectar la calidad del modelo.
- Balancear clases, que es el proceso de igualar las instancias de la clase de objetivo para evitar los problemas de la generalización de la información.
- Gestionar los valores ausentes (“missing data”).
- La gestión de datos fuera de rango (“outliers”), que es el proceso de identificar valores anormales o atípicos que pueden afectar las estimaciones estadísticas.
- Conciliar y hacer coincidir los datos, es el proceso de detectar y transformar los datos con tareas de depuración.
- Fusión y creación de nuevos atributos, es el proceso de crear o fusionar nuevos atributos a partir de los atributos existentes (edad, periodo, año de presentación, desempeño del estudiante).

Para la ejecución de estas tareas iniciales se seleccionó la herramienta de código abierto OpenRefine. Esta es una aplicación de escritorio que permite perfilar, limpiar y transformar un conjunto de datos a otros formatos, una actividad comúnmente conocida como data wrangling; esta aplicación usa un lenguaje de

programación llamado GREL (Google Refine Expression Languaje) con un gran número de funciones para realizar tareas de depuración avanzadas.

OpenRefine presenta la opción de “Detectar y transformar” formatos inconsistentes, espacios en blanco innecesarios, caracteres adicionales, errores tipográficos, etc., por ejemplo, fechas, textos y números. En el grupo de datos original se ejecutaron procesos con el objetivo de conciliar y hacer coincidir el dato solucionando problemas como:

1. Se eliminaron espacios en blanco extra al principio o al final del texto
2. Se eliminaron espacios en blanco extra entre palabras
3. La inclusión y exclusión de registros con valores determinados para un campo
4. El agrupamiento de campos por valores similares para detectar y corregir errores de ortografía y variaciones en los datos.

En este ejercicio se encuentran atributos que tienen valores que se refieren al mismo concepto, pero escritos o acentuados de manera diferente; por lo que es necesario revisar la consistencia de los valores de cada atributo y en donde se intervienen las siguientes variables:

- Los nombres de las instituciones educativas (COLE_NOMBRE ESTABLECIMIENTO).

- Los nombres de las sedes de las instituciones educativas (COLE_NOMBRE SEDE).
- El nombre del municipio donde está ubicada la Sede (COLE_MCPIO_UBICACION).
- El nivel Socioeconómico del evaluado (ESTU_INSE_INDIVIDUAL).
- El periodo de presentación del examen (PERIODO).

Figura 24 Limpieza de variables con Open Refine

Fuente: Open Refine

Se puede ver el detalle del proceso de limpieza con Open Refine en el Anexo 5 Resultados y variables intervenidas con Open Refine

Posteriormente, se aplicó la metodología MELODA (metric for releasing open data), la cual está orientada a la medición de la calidad de las fuentes de datos abiertos, en términos de reutilización y aprovechamiento máximo de los mismos.

Actualmente, se encuentra vigente la versión del modelo Meloda 5.0 que evalúa siete dimensiones y modifica los niveles y la ponderación correspondiente.
Para una descripción de las escalas de puntuación y la descripción de Meloda 5.0, por favor se ver en el Anexo 4 Dimensiones y niveles Meloda 5.0

En la versión 5.0 de Meloda se conservan las seis primeras dimensiones de Meloda 4.0 (el acceso a la información, el marco legal, el estándar técnico, el modelo de datos, la geolocalización es decir la inclusión de información geográfica en los datos y la frecuencia de actualización o tiempo transcurrido entre versiones), se modifica el nombre de la dimensión “modelo de datos” por “estandarización” y se incluyen dos nuevas dimensiones: una mide la difusión de las actividades asociadas a los datos abiertos y la otra analiza la reputación de los datos abiertos. Con todas estas dimensiones, la métrica proporciona una evaluación cuantitativa sobre sobre que tanto se pueden reutilizar los conjuntos de datos publicados.

Para verificar la calidad y la reutilización de los datos abiertos utilizados en esta investigación se aplicaron los dos modelos Meloda más recientes con sus diferentes dimensiones; la versión 4.13 y 5.0, en donde ambos ejercicios arrojaron una calificación de 34 (tabla 6) y 26 (tabla 8) respectivamente que permiten definir

que los datos abiertos tienen un nivel básico de reutilización de acuerdo con las metodologías aplicadas.

Tabla 6 Matriz MELODA 4.13 para los datos abiertos

	Legal	Acceso	Estándares Técnicos	Modelo de Datos	Geolocalización	Actualización
Nivel 5	Sin restricciones o solo atribuciones (CC BY 4.0) **(100%)**	Acceso completo (API o lenguaje de consulta) **(100%)**		Modelo de datos abierto **(100%)**	Información geográfica completa **(100%)**	Segundos, menor por minuto **(100%)**
Nivel 4	Reutilización comercial (CC BY-ND 4.0/CC BY-SA 4.0) **(90%)**	Acceso vía web con parámetros, URL con parámetros. **(90%)**	formato de estándar abierto, con metadatos (rdf, rss, json, xml) **(100%)**	Modelo local de datos abiertos **(90%)**	Coordenadas (sistema de coordenadas) **(90%)**	Minutos, 1 minuto a 1 hora (liberación por horas) **(90%)**
Nivel 3	Reutilización no comercial (CC BY-NC-ND 4.0/CC BY-NC 4.0/CC BY-NC-SA 4.0) **(25%)**	Acceso directo vía web URL unica **(50%)**	Formato estándar abierto (csv, txt, odb, odt, ods, WMS, xls y xlsx sin macro/fórmula) **(60%)**	Modelo de datos propio con especificaciones de campos (con una ontología, vocabularios disponibles).**(50%)**	Campo de texto complejo (el número de una calle se considera como un texto) **(50%)**	Horas, 1 a 24 horas (liberación diaria) **(70%)**
Nivel 2	Uso privado **(10%)**	Acceso vía web con registro **(10%)**	Formato estándar cerrado reutilizable (shp, xls y xlsx con macro/fórmula) **(35%)**	Modelo con campos de datos **(35%)**	Campo de texto simple (Id propio) **(30%)**	Días, 1 a 7 días (liberación semanal) **(40%)**
Nivel 1	Copyright **(0%)**	Sin acceso web o solicitud manual **(0%)**	Formato estándar cerrado no reutilizable (pdf image, doc) **(10%)**	Sin modelo publicado **(15%)**	No hay información geográfica **(15%)**	Mayor a una semana (mensual, semestral, anual) **(15%)**
	100%	**10%**	**100%**	**35%**	**30%**	**15%**

Calificación	Rangos de Reutilización
34	**Reutilización básica**

Fuente: elaboración propia

Tabla 7 Rangos de clasificación de la métrica Meloda 4.13

Rangos	Calificación
75-100	Reutilización avanzada
50-75	Reutilización avanzada con alguna característica mejorable
25-50	Reutilización básica
0-25	Inadecuado para reutilización

Fuente: Meloda 4.13 (Alberto Abella)

Tabla 8 Matriz Meloda 5.0 para datos abiertos

Dimensiones (máximo 61 puntos)	Niveles	Calificación
Licencia legal (máx. 6 puntos)	1. uso privado	3
	2. reutilización no comercial	
	3. reutilización comercial o sin restricciones	
Acceso a la información (máx. 6 puntos)	1. acceso a través de web o parámetros únicos URL al conjunto de datos	2
	2. acceso único a la web con parámetros referidos a datos individuales	
	3. API o lenguaje de consulta	
Estándar técnico (máx. 6 puntos)	1. estándar cerrado reutilizable o abierto no reutilizable	3
	2. estándar abierto reutilizable	
	3. estándar abierto con metadatos individuales	
Nivel de estandarización (máx. 10 puntos)	1. modelo propio de estandarización	1
	2. modelo de estandarización propio o ad hoc publicado (armonización)	
	3. estandarización local	
	4. estandarización global	
Contenido geolocalizado (máx. 6 puntos)	1. sin información geográfica	2
	2. campo de texto simple o complejo	
	3. con coordenadas o información geográfica completa	
Actualización de la frecuencia de datos (máx. 15 puntos)	1. por encima de un mes	1
	2. mensual. Con periodos de actualización entre 1 mes y 1 día	
	3. diaria. Con periodos de actualización entre 1 día y 1 hora	
	4. cada hora. Con periodos de actualización desde 1 hora a 1 minuto	
	5. en segundos. Periodo de actualización por debajo de 1 minuto	
Difusión (máx. 6 puntos)	1. comunicación/difusión no sistemática	2
	2. recursos disponibles sobre actualizaciones (p. ej., alimentación en redes sociales)	
	3. difusión proactiva/difusión push (información automatizada en determinado tiempo)	
Reputación (máx. 6 puntos)	1. no hay información sobre la reputación de la fuente de datos	2
	2. las estadísticas o informes se publican en función a las opiniones de los usuarios	
	3. rankings o indicadores basados en la reputación de la fuente de datos	
Total		**26**

Fuente: Elaboración propia

Para el uso de esta métrica se debe analizar el conjunto de datos desde la perspectiva de las siete dimensiones, que tienen una ponderación máxima por cada nivel, con la suma de las puntuaciones obtenidas se obtiene un puntaje que permite clasificar el conjunto de datos según su grado de reutilización:

Tabla 9 Rangos de clasificación de la métrica Meloda 5.0

Dimensiones (máximo 61 puntos)	8-23	24-47	48-61
Categoria MELODA 5	Inadecuado	Básico	Avanzado

Fuente: Elaboración propia

Así mismo con esta medición se puede identificar los aspectos a mejorar en cada dimensión.

3.3 Preparación de los datos

Es fundamental seleccionar las variables más adecuadas para pasar al algoritmo de aprendizaje y esto se puede lograr de varias formas: identificando atributos con gran aporte de información, eliminando atributos que aportan poca información, creando nuevos atributos a partir de los existentes. La preparación de los datos tiene el objetivo de adecuar los datos para facilitar su uso con algoritmos de clasificación, segmentación o regresión. En este proyecto, inicialmente se seleccionaron 14 data sets o conjuntos de datos correspondientes a los resultados de las pruebas Saber 11° entre los años 2015 y 2021, en donde cada año tiene 2 datasets, uno por cada calendario académico A y B. (Figura 25)

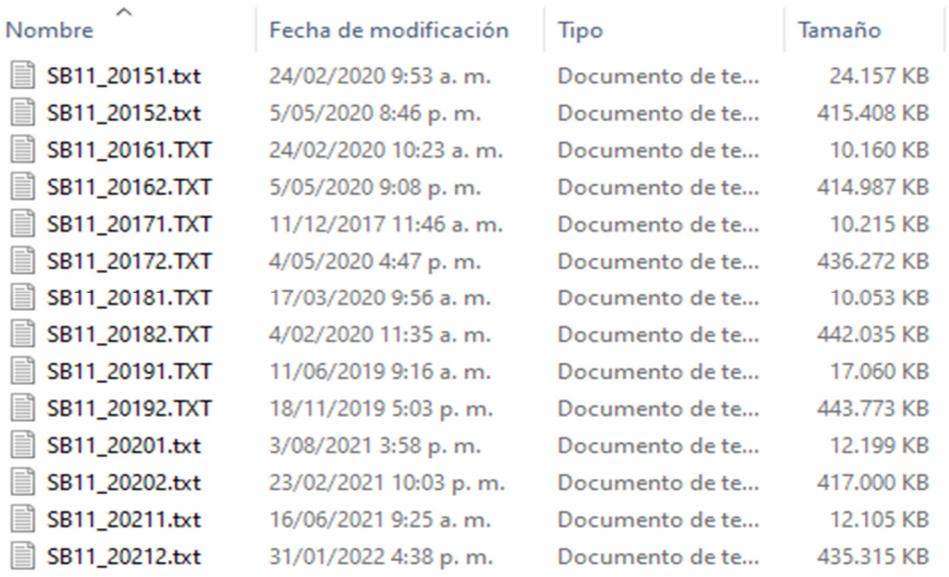
El proceso de preparación de los datos se ejecuta en 4 pasos, los cuales se resumen de la siguiente manera:

1. Se descargan 14 datasets originales (***Figura 25***)
    - ✓ Cada año tiene dos datasets (Calendario A y B)

- ✓ Calendario A tiene aproximadamente 15.000 registros
- ✓ Calendario B tiene aproximadamente 530.000 registros

2. Se aplican varios filtros a los datasets
   - ✓ Se seleccionan los estudiantes del departamento de Cundinamarca
   - ✓ Se seleccionan los municipios No certificados de Cundinamarca
   - ✓ Se seleccionan las entidades educativas oficiales
3. Se unifican los datasets en 7 archivos (***Figura 26***)
   - ✓ Se genera un dataset por año con un promedio de 13.000 registros
   - ✓ Cada dataset tiene entre 82 y 134 variables o atributos (***Figura 27***)
4. Se identifican los tipos de variables (***Figura 28***)
   - ✓ Variables numéricas aproximadamente 30%
   - ✓ Variables categóricas aproximadamente 70%

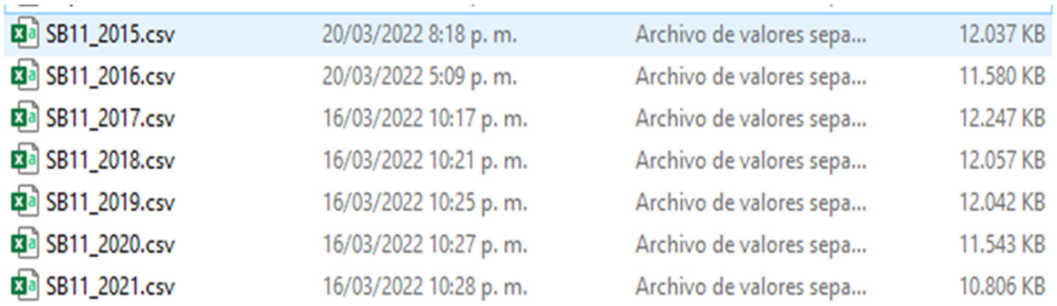
Figura 25 Grupos de datos iniciales de las pruebas Saber 11°



Nombre	Fecha de modificación	Tipo	Tamaño
SB11_20151.txt	24/02/2020 9:53 a. m.	Documento de te...	24.157 KB
SB11_20152.txt	5/05/2020 8:46 p. m.	Documento de te...	415.408 KB
SB11_20161.TXT	24/02/2020 10:23 a. m.	Documento de te...	10.160 KB
SB11_20162.TXT	5/05/2020 9:08 p. m.	Documento de te...	414.987 KB
SB11_20171.TXT	11/12/2017 11:46 a. m.	Documento de te...	10.215 KB
SB11_20172.TXT	4/05/2020 4:47 p. m.	Documento de te...	436.272 KB
SB11_20181.TXT	17/03/2020 9:56 a. m.	Documento de te...	10.053 KB
SB11_20182.TXT	4/02/2020 11:35 a. m.	Documento de te...	442.035 KB
SB11_20191.TXT	11/06/2019 9:16 a. m.	Documento de te...	17.060 KB
SB11_20192.TXT	18/11/2019 5:03 p. m.	Documento de te...	443.773 KB
SB11_20201.txt	3/08/2021 3:58 p. m.	Documento de te...	12.199 KB
SB11_20202.txt	23/02/2021 10:03 p. m.	Documento de te...	417.000 KB
SB11_20211.txt	16/06/2021 9:25 a. m.	Documento de te...	12.105 KB
SB11_20212.txt	31/01/2022 4:38 p. m.	Documento de te...	435.315 KB

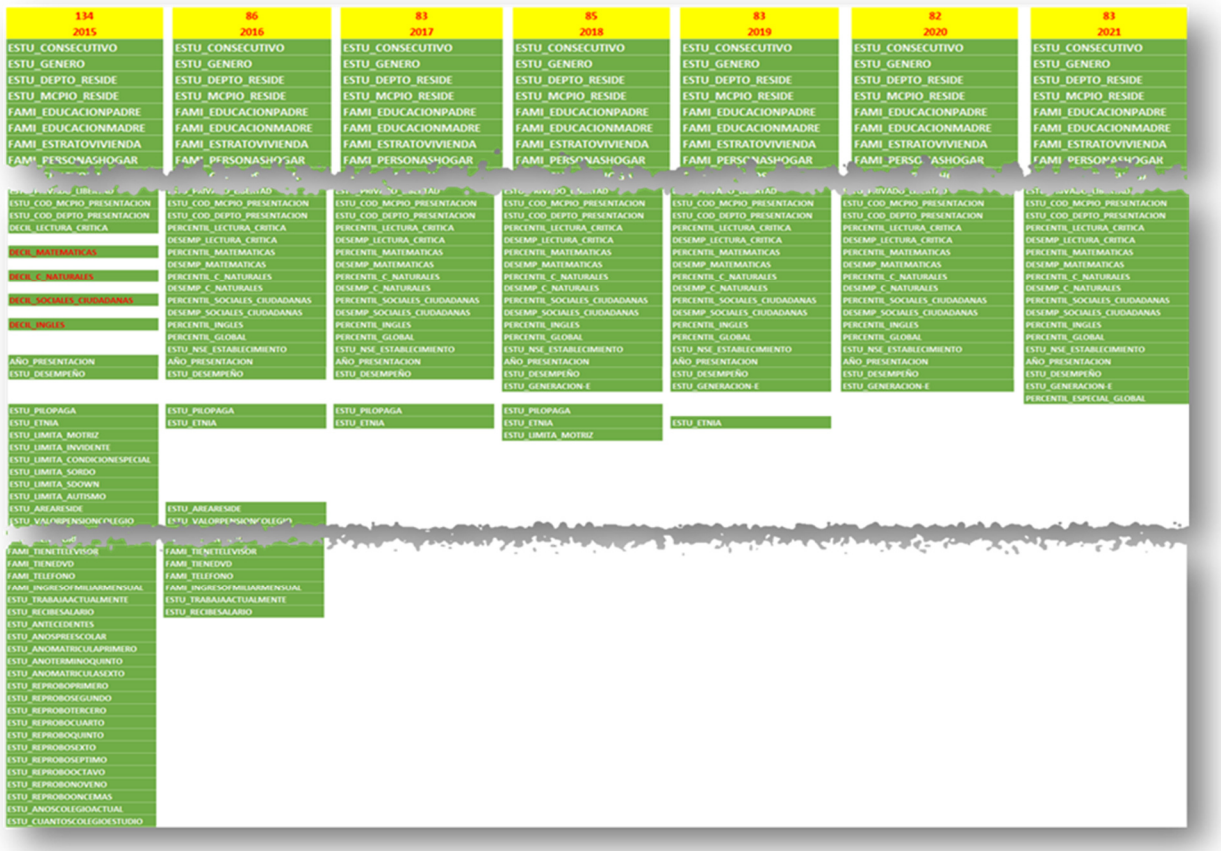
Fuente: Elaboración propia

Figura 26 Grupos de datos unificados



Fuente: Elaboración propia

Figura 27 Grupos de datos años 2015-2021



Fuente: Elaboración propia

Figura 28 Clasificación de los datos

Nominal

View original data and use it to make decision about other choices

Column	No. missings	Unique values
ESTU_TIPODOCUMENTO	0	4
ESTU_NACIONALIDAD	0	4
ESTU_GENERO	128	2
ESTU_FECHANACIMIENTO	0	100
PERIODO	0	2
ESTU_CONSECUTIVO	0	100
ESTU_ESTUDIANTE	0	1
ESTU_PAIS_RESIDE	0	4
ESTU_ETNIA	14680	8
ESTU_LIMITA_MOTRIZ	14730	1
ESTU_LIMITA_INVIDENTE	14737	1
ESTU_LIMITA_CONDICIONESPECIAL	14740	1
ESTU_LIMITA_SORDO	14742	1
ESTU_LIMITA_SDOWN	14737	1
ESTU_LIMITA_AUTISMO	14735	1
ESTU_DEPTO_RESIDE	0	7
ESTU_MCPIO_RESIDE	0	100
ESTU_AREARESIDE	5	2
ESTU_VALORPENSIONCOLEGIO	5	6
ESTU_VECESPRESENTOEXAMEN	5	4
FAMI_EDUCACIONPADRE	5	11
FAMI_EDUCACIONMADRE	5	11
FAMI_OCUPACIONPADRE	5	12
FAMI_OCUPACIONMADRE	5	12
FAMI_NUMHERMANOS	5	11
FAMI_ESTRATOVIVIENDA	0	6
FAMI_NIVELSISBEN	5	5
FAMI_PERSONASHOGAR	5	12
FAMI_CUARTOSHOGAR	5	10
FAMI_PISOSHOGAR	5	4
FAMI_TIENEINTERNET	5	2
FAMI_TIENECOMPUTADOR	5	2
FAMI_TIENELAVADORA	5	2
FAMI_TIENEMICROONDAS	5	2
FAMI_TIENEHORNO	5	2
FAMI_TIENEAUTOMOVIL	5	2
FAMI_TIENEDVD	5	2
FAMI_TELEFONO	5	2
FAMI_NUMLIBROS	5	4
FAMI_INGRESOFMILIARMENSUAL	5	6
ESTU_TRABAJAACTUALMENTE	5	3
ESTU_RECIBESALARIO	12250	2
COLE_NOMBRE_ESTABLECIMIENTO	0	100
COLE_GENERO	0	1
COLE_NATURALEZA	0	1
COLE_CALENDARIO	0	1
COLE_BILINGUE	1592	2
COLE_CARACTER	0	3
COLE_NOMBRE_SEDE	0	100
COLE_SEDE_PRINCIPAL	0	2
COLE_AREA_UBICACION	0	2
COLE_JORNADA	0	5
COLE_MCPIO_UBICACION	0	100
COLE_DEPTO_UBICACION	0	1
ESTU_PRIVADO_LIBERTAD	0	1
ESTU_MCPIO_PRESENTACION	0	48
ESTU_DEPTO_PRESENTACION	0	10
DESEMP_INGLES	0	5
ESTU_NSE_INDIVIDUAL	5	4
ESTU_ESTADOINVESTIGACION	0	2
ESTU_PILOPAGA	0	2
ESTU_DESEMPEÑO	0	4

Numeric

View original data and use it to make decision about other choices

Column	Minimum	Maximum
ESTU_COD_RESIDE_DEPTO	11	73
ESTU_COD_RESIDE_MCPIO	11001	73547
COLE_CODIGO_ICFES	8201	195610
COLE_COD_DANE_ESTABLECIMIENTO	1.25001E+11	5.25843E+11
COLE_COD_DANE_SEDE	1.25001E+11	5.25843E+11
COLE_COD_MCPIO_UBICACION	25001	25898
COLE_COD_DEPTO_UBICACION	25	25
ESTU_COD_MCPIO_PRESENTACION	5051	85440
ESTU_COD_DEPTO_PRESENTACION	5	85
PUNT_LECTURA_CRITICA	0	93
DECIL_LECTURA_CRITICA	1	10
PUNT_MATEMATICAS	0	100
DECIL_MATEMATICAS	1	10
PUNT_C_NATURALES	0	90
DECIL_C_NATURALES	1	10
PUNT_SOCIALES_CIUDADANAS	3	98
DECIL_SOCIALES_CIUDADANAS	1	10
PUNT_RAZONA_CUANTITATIVO	0	100
DECIL_RAZONA_CUANTITATIVO	1	10
PUNT_COMP_CIUDADANA	10	100
DECIL_COMP_CIUDADANA	1	10
PUNT_INGLES	0	100
DECIL_INGLES	1	10
PUNT_GLOBAL	60	426
ESTU_PUESTO	2	1000
ESTU_INSE_INDIVIDUAL	19.831	75.166
ESTU_DESEMPEÑO_NIVEL	1	4

Fuente: Pruebas ICFES Saber 11°, muestra seleccionada 2017-2021

3.3.1 Seleccionar los datos

Como se videncia en la figura 29, se decidió omitir los años 2015 y 2016 del análisis para aumentar la significancia de los resultados por su gran proporción de datos faltantes en la aplicación de los algoritmos.

Figura 29 Grupos de datos finalmente seleccionado años 2017-2021

Fuente: Elaboración propia

A partir de los datos que permanecieron, se clasificaron las variables socioeconómicas en cuatro grupos grandes o dimensiones según la temática sobre la que se indagaba, así: 1. Información personal (25 variables), 2. Información familiar (20 variables), 3. Información del colegio (17 variables) y 4. Información académica o resultados de la prueba (18 variables), para un total de 80 variables obtenidas del cuestionario socioeconómico de la Prueba Saber 11°, como se refleja en la tabla 10. Se limitó el análisis de estas variables únicamente para el Departamento de Cundinamarca y específicamente para los 108 municipios no certificados del departamento que se encuentran bajo control directo de la gobernación de Cundinamarca, en términos de políticas públicas y manejo de presupuesto. Adicionalmente, se determinó que el análisis se centraría en estudiantes graduados únicamente de instituciones oficiales y no privadas,

teniendo en cuenta que la revisión de la literatura evidenció que la disminución en los puntajes de la prueba Saber 11° fue más marcado en las instituciones oficiales.

Tabla 10 Variables iniciales seleccionadas

Estudiante	Institución Educativa	Información Familiar	Información de la prueba Saber 11
ESTU_CONSECUTIVO	COLE_NOMBRE_ESTABLECIMIENTO	FAMI_EDUCACIONPADRE	PUNT_LECTURA_CRITICA
ESTU_GENERO	COLE_BILINGUE	FAMI_EDUCACIONMADRE	PUNT_MATEMATICAS
ESTU_DEPTO_RESIDE	COLE_CARACTER	FAMI_ESTRATOVIVIENDA	PUNT_C_NATURALES
ESTU_MCPIO_RESIDE	COLE_NOMBRE_SEDE	FAMI_PERSONASHOGAR	PUNT_SOCIALES_CIUDADANAS
ESTU_MCPIO_PRESENTACION	COLE_SEDE_PRINCIPAL	FAMI_CUARTOSHOGAR	PUNT_INGLES
ESTU_DEPTO_PRESENTACION	COLE_AREA_UBICACION	FAMI_TIENEINTERNET	PUNT_GLOBAL
ESTU_INSE_INDIVIDUAL	COLE_JORNADA	FAMI_TIENECOMPUTADOR	DESEMP_INGLES
ESTU_NSE_INDIVIDUAL	COLE_MCPIO_UBICACION	FAMI_TIENELAVADORA	DESEMP_SOCIALES_CIUDADANAS
ESTU_ESTADOINVESTIGACION	COLE_CODIGO_ICFES	FAMI_TIENEAUTOMOVIL	DESEMP_MATEMATICAS
ESTU_AÑO_NACIMIENTO	COLE_COD_DANE_ESTABLECIMIENTO	FAMI_NUMLIBROS	DESEMP_C_NATURALES
ESTU_TIPODOCUMENTO	COLE_GENERO	FAMI_TRABAJOLABORPADRE	DESEMP_LECTURA_CRITICA
ESTU_NACIONALIDAD	COLE_CALENDARIO	FAMI_TRABAJOLABORMADRE	PERCENTIL_MATEMATICAS
ESTU_FECHANACIMIENTO	COLE_COD_DANE_SEDE	FAMI_TIENEHORNOMICROOGAS	PERCENTIL_C_NATURALES
ESTU_PAIS_RESIDE	COLE_COD_MCPIO_UBICACION	FAMI_TIENESERVICIOTV	PERCENTIL_LECTURA_CRITICA
ESTU_TIENEETNIA	COLE_COD_DEPTO_UBICACION	FAMI_TIENEMOTOCICLETA	PERCENTIL_SOCIALES_CIUDADANAS
ESTU_COD_RESIDE_DEPTO	COLE_DEPTO_UBICACION	FAMI_TIENECONSOLAVIDEOJUEGOS	PERCENTIL_INGLES
ESTU_COD_RESIDE_MCPIO	COLE_NATURALEZA	FAMI_COMELECHEDERIVADOS	PERCENTIL_GLOBAL
ESTU_DEDICACIONLECTURADIARIA		FAMI_COMECARNEPESCADOHUEVO	
ESTU_DEDICACIONINTERNET		FAMI_COMECEREALFRUTOSLEGUMBRE	
ESTU_HORASSEMANATRABAJA		FAMI_SITUACIONECONOMICA	PERIODO
ESTU_TIPOREMUNERACION			
ESTU_PRIVADO_LIBERTAD			
ESTU_COD_MCPIO_PRESENTACION			
ESTU_COD_DEPTO_PRESENTACION			
ESTU_NSE_ESTABLECIMIENTO			
25	17	20	18

Fuente: Elaboración propia

Después de haber seleccionado el número de atributos y registros del conjunto de datos inicial, se evaluaron y determinaron los atributos que no aportaban valor al modelo en términos de información, y se inició la limpieza de los datos, se seleccionaron los atributos determinantes en los resultados de los modelos de análisis, para finalmente tener el conjunto de datos óptimo en variables y registros.

3.3.2 Eliminación de atributos

Después de haber analizado las variables iniciales de manera empírica, se procedió con la eliminación de atributos que no contribuían con información al

modelo. Si una característica siempre contiene el mismo valor, no contribuye en absoluto al modelo por ser una columna de varianza cero. Con este tipo de variables constantes o cuasi constantes, el modelo de aprendizaje no aprenderá nada revelador o, peor aún, puede aprender de estos casos marginales o poco importantes y generar un sobreajuste del modelo.

De las variables iniciales de la tabla 10, y utilizando la librería de Python (fast-ML), se identificaron y eliminaron atributos con estas características y se muestran la figura 30.

Figura 30 Listado de variables constantes

	Desc	Var	Value	Perc
0	Constant	ESTU_ESTUDIANTE	ESTUDIANTE	100.000000
1	Constant	COLE_GENERO	MIXTO	100.000000
2	Constant	COLE_NATURALEZA	OFICIAL	100.000000
3	Constant	COLE_CALENDARIO	A	100.000000
4	Constant	COLE_COD_DEPTO_UBICACION	25	100.000000
5	Constant	COLE_DEPTO_UBICACION	CUNDINAMARCA	100.000000
6	Quasi Constant	ESTU_PRIVADO_LIBERTAD	N	99.991051
7	Quasi Constant	ESTU_ESTADOINVESTIGACION	PUBLICAR	99.956748
8	Quasi Constant	ESTU_NACIONALIDAD	COLOMBIA	99.753911
9	Quasi Constant	ESTU_PAIS_RESIDE	COLOMBIA	99.753911
10	Quasi Constant	ESTU_TIENEETNIA	No	99.568972
11	Quasi Constant	ESTU_DEPTO_RESIDE	CUNDINAMARCA	99.175230
12	Quasi Constant	ESTU_COD_RESIDE_DEPTO	25.0	99.175230

Fuente: Python (fast-ML)

Tabla 11 Nuevo listado de variables sin atributos constantes

Estudiante	Institución Educativa	Información Familiar	Información de la prueba Saber 11
ESTU_CONSECUTIVO	COLE_NOMBRE_ESTABLECIMIENTO	FAMI_EDUCACIONPADRE	PUNT_LECTURA_CRITICA
ESTU_GENERO	COLE_BILINGUE	FAMI_EDUCACIONMADRE	PUNT_MATEMATICAS
ESTU_MCPIO_RESIDE	COLE_CARACTER	FAMI_ESTRATOVIVIENDA	PUNT_C_NATURALES
ESTU_MCPIO_PRESENTACION	COLE_NOMBRE_SEDE	FAMI_PERSONASHOGAR	PUNT_SOCIALES_CIUDADANAS
ESTU_DEPTO_PRESENTACION	COLE_SEDE_PRINCIPAL	FAMI_CUARTOSHOGAR	PUNT_INGLES
ESTU_INSE_INDIVIDUAL	COLE_AREA_UBICACION	FAMI_TIENEINTERNET	PUNT_GLOBAL
ESTU_NSE_INDIVIDUAL	COLE_JORNADA	FAMI_TIENECOMPUTADOR	DESEMP_INGLES
ESTU_AÑO_NACIMIENTO	COLE_MCPIO_UBICACION	FAMI_TIENELAVADORA	DESEMP_SOCIALES_CIUDADANAS
ESTU_TIPODOCUMENTO	COLE_CODIGO_ICFES	FAMI_TIENEAUTOMOVIL	DESEMP_MATEMATICAS
ESTU_FECHANACIMIENTO	COLE_COD_DANE_ESTABLECIMIENTO	FAMI_NUMLIBROS	DESEMP_C_NATURALES
ESTU_COD_RESIDE_MCPIO	COLE_COD_DANE_SEDE	FAMI_TRABAJOLABORPADRE	DESEMP_LECTURA_CRITICA
ESTU_DEDICACIONLECTURADIARIA	COLE_COD_MCPIO_UBICACION	FAMI_TRABAJOLABORMADRE	PERCENTIL_MATEMATICAS
ESTU_DEDICACIONINTERNET		FAMI_TIENEHORNOMICROOGAS	PERCENTIL_C_NATURALES
ESTU_HORASSEMANATRABAJA		FAMI_TIENESERVICIOTV	PERCENTIL_LECTURA_CRITICA
ESTU_TIPOREMUNERACION		FAMI_TIENEMOTOCICLETA	PERCENTIL_SOCIALES_CIUDADANAS
ESTU_COD_MCPIO_PRESENTACION		FAMI_TIENECONSOLAVIDEOJUEGOS	PERCENTIL_INGLES
ESTU_COD_DEPTO_PRESENTACION		FAMI_COMELECHEDERIVADOS	PERCENTIL_GLOBAL
ESTU_NSE_ESTABLECIMIENTO		FAMI_COMECARNEPESCADOHUEVO	
		FAMI_COMECEREALFRUTOSLEGUMBRE	
		FAMI_SITUACIONECONOMICA	PERIODO
18	12	20	18

Fuente: Elaboración propia

Del grupo de variables “Información de la prueba Saber 11° (Anexo 2) se eliminaron los atributos relacionados con la variable objetivo y con los resultados de la prueba, relacionados en la siguiente tabla:

Tabla 12 Listado de variables relacionados con los resultados de la prueba

PUNT_LECTURA_CRITICA	**DESEMP_C_NATURALES**
PUNT_MATEMATICAS	**DESEMP_LECTURA_CRITICA**
PUNT_C_NATURALES	**PERCENTIL_MATEMATICAS**
PUNT_SOCIALES_CIUDADANAS	**PERCENTIL_C_NATURALES**
PUNT_INGLES	**PERCENTIL_LECTURA_CRITICA**
DESEMP_INGLES	**PERCENTIL_SOCIALES_CIUDADANAS**
DESEMP_SOCIALES_CIUDADANAS	**PERCENTIL_INGLES**
DESEMP_MATEMATICAS	**PERCENTIL_GLOBAL**
PUNT_GLOBAL	

Fuente: Elaboración propia

Posteriormente se eliminaron todas las variables que son identificadores (ID) para evitar que el modelo trate estas variables como valores numéricos, afectando el

rendimiento del modelo y no identificando la variable dependiente. Después de esta eliminación de atributos en la tabla 13, se tiene un nuevo grupo de datos con 43 variables.

Tabla 13 Variables sin identificadores y constantes

Estudiante	Institución Educativa	Información Familiar	Información de la prueba Saber 11
ESTU_GENERO	COLE_NOMBRE_ESTABLECIMIENTO	FAMI_EDUCACIONPADRE	PERIODO
ESTU_MCPIO_RESIDE	COLE_BILINGUE	FAMI_EDUCACIONMADRE	
ESTU_MCPIO_PRESENTACION	COLE_CARACTER	FAMI_ESTRATOVIVIENDA	
ESTU_DEPTO_PRESENTACION	COLE_NOMBRE_SEDE	FAMI_PERSONASHOGAR	
ESTU_INSE_INDIVIDUAL	COLE_SEDE_PRINCIPAL	FAMI_CUARTOSHOGAR	
ESTU_NSE_INDIVIDUAL	COLE_AREA_UBICACION	FAMI_TIENEINTERNET	
ESTU_AÑO_NACIMIENTO	COLE_JORNADA	FAMI_TIENECOMPUTADOR	
ESTU_TIPODOCUMENTO	COLE_MCPIO_UBICACION	FAMI_TIENELAVADORA	
ESTU_FECHANACIMIENTO		FAMI_TIENEAUTOMOVIL	
ESTU_DEDICACIONLECTURADIARIA		FAMI_NUMLIBROS	
ESTU_DEDICACIONINTERNET		FAMI_TRABAJOLABORPADRE	
ESTU_HORASSEMANATRABAJA		FAMI_TRABAJOLABORMADRE	
ESTU_TIPOREMUNERACION		FAMI_TIENEHORNOMICROOGAS	
ESTU_NSE_ESTABLECIMIENTO		FAMI_TIENESERVICIOTV	
		FAMI_TIENEMOTOCICLETA	
		FAMI_TIENECONSOLAVIDEOJUEGOS	
		FAMI_COMELECHEDERIVADOS	
		FAMI_COMECARNEPESCADOHUEVO	
		FAMI_COMECEREALFRUTOSLEGUMBRE	
		FAMI_SITUACIONECONOMICA	
14	8	20	1

Fuente: Elaboración propia

3.3.3 Adicionando y transformando atributos

Finalmente se adiciono el atributo ESTU_EDAD que registra la edad del estudiante al momento de presentar la prueba y se obtuvo de la diferencia de fechas en el atributo ESTU_FECHANACIMIENTO y el atributo PERIODO, después del cálculo se eliminó el atributo ESTU_FECHANACIMIENTO. También se incluyó el atributo COLE_REGION que permitió agrupar los 108 municipios no certificados de Cundinamarca en las 15 provincias o regiones que tiene el departamento.

Para esta investigación se diseñó una clasificación de desempeño correspondiente a la suma de los puntajes que el estudiante obtuvo en cada área

y que determina si el desempeño del estudiante se clasifica cómo insuficiente, satisfactorio o avanzado. Todos los puntajes individuales de los estudiantes menores o iguales a 279 puntos se consideran “Insuficientes”, los puntajes entre 280 y 359 puntos se consideran “Satisfactorios” y los puntajes superiores a 359 puntos se consideran “Avanzados”. Esta clasificación se adaptó a partir de la categorización realizada por el Icfes de manera individual para cada área y que cataloga el desempeño del estudiante en una de 4 categorías (deficiente, mínimo, aceptable y sobresaliente) pero que no contaba con un equivalente diseñado por el Icfes para la calificación global.

El puntaje global en las pruebas Saber 11° van desde 0 a 500 puntos, a pesar de que el promedio del puntaje global en calendario A es de 250 puntos, mostrando una disminución de 2 puntos entre 2020 y 2021 (Ministerio de Educación Nacional, 2022)).

Aunque se tiene la tendencia que los puntajes superiores a esos 250 son positivos, en este estudio se definió como frontera para obtener un puntaje satisfactorio 280 puntos, suponiendo como objetivo de un estudiante ingresar a una universidad pública o privada donde los promedios que se exigen son de 320 puntos en el ICFES; o aquellos que están interesados en acceder al programa Generación E, necesitara un puntaje mayor a 349 puntos o 360 puntos si se está aspirando a una beca para cualquier carrera o universidad según el portal de Grupo Geard.

Al determinar esta clasificación, se realizó la recodificación de la variable numérica PUNT_GLOBAL a una variable categórica ESTU_DESEMPEÑO con tres posibles asignaciones (insuficiente, satisfactorio y avanzado) como se muestra en la tabla 14, la cual se utilizó para los análisis posteriores.

Tabla 14 Adición de atributo objetivo ESTU_DESEMPEÑO

Variable	Rango	Categoría
PUNT_GLOBAL	< = 279	Insuficiente
	> = 280 y < = 359	Satisfactorio
	> = 360	Avanzado

Fuente: Elaboración propia

La selección final con 44 atributos después del proceso de preparación de los datos se puede apreciar en la tabla 15.

Tabla 15 Variables creando y transformando atributos

Estudiante	Institución Educativa	Información Familiar	Información de la prueba Saber 11
ESTU_GENERO	COLE_NOMBRE_ESTABLECIMIENTO	FAMI_EDUCACIONPADRE	PERIODO
ESTU_MCPIO_RESIDE	COLE_BILINGUE	FAMI_EDUCACIONMADRE	
ESTU_MCPIO_PRESENTACION	COLE_CARACTER	FAMI_ESTRATOVIVIENDA	
ESTU_DEPTO_PRESENTACION	COLE_NOMBRE_SEDE	FAMI_PERSONASHOGAR	
ESTU_INSE_INDIVIDUAL	COLE_SEDE_PRINCIPAL	FAMI_CUARTOSHOGAR	
ESTU_NSE_INDIVIDUAL	COLE_AREA_UBICACION	FAMI_TIENEINTERNET	
ESTU_EDAD	COLE_JORNADA	FAMI_TIENECOMPUTADOR	
ESTU_TIPODOCUMENTO	COLE_MCPIO_UBICACION	FAMI_TIENELAVADORA	
ESTU_DEDICACIONLECTURADIARIA	COLE_REGION	FAMI_TIENEAUTOMOVIL	
ESTU_DEDICACIONINTERNET		FAMI_NUMLIBROS	
ESTU_HORASSEMANATRABAJA		FAMI_TRABAJOLABORPADRE	
ESTU_TIPOREMUNERACION		FAMI_TRABAJOLABORMADRE	
ESTU_NSE_ESTABLECIMIENTO		FAMI_TIENEHORNOMICROOGAS	
ESTU_DESEMPEÑO		FAMI_TIENESERVICIOTV	
		FAMI_TIENEMOTOCICLETA	
		FAMI_TIENECONSOLAVIDEOJUEGOS	
		FAMI_COMELECHEDERIVADOS	
		FAMI_COMECARNEPESCADOHUEVO	
		FAMI_COMECEREALFRUTOSLEGUMBRE	
		FAMI_SITUACIONECONOMICA	
14	9	20	1

Fuente: elaboración propia

3.3.4 Limpiar los datos

De acuerdo con los criterios especificados, se realizó la limpieza de los datos que permite corregir o eliminar registros inexactos, duplicados o faltantes en el conjunto de datos. Específicamente en:

- Igualar formatos

- Descartar campos o variables
- Corregir errores ortográficos
- Eliminar columnas duplicadas
- Eliminar registros no útiles

Para completar la información que falta se puede hacer tanto por la técnica de imputación de datos (en las variables en las que aplicaba) como por eliminación de valores faltantes. Esto permitió obtener dos conjuntos de datos provenientes de cada metodología de limpieza de datos, los cuales se utilizaron posteriormente para el desarrollo de los modelos analíticos de forma paralela.

3.3.4.1 Manejo de valores faltantes

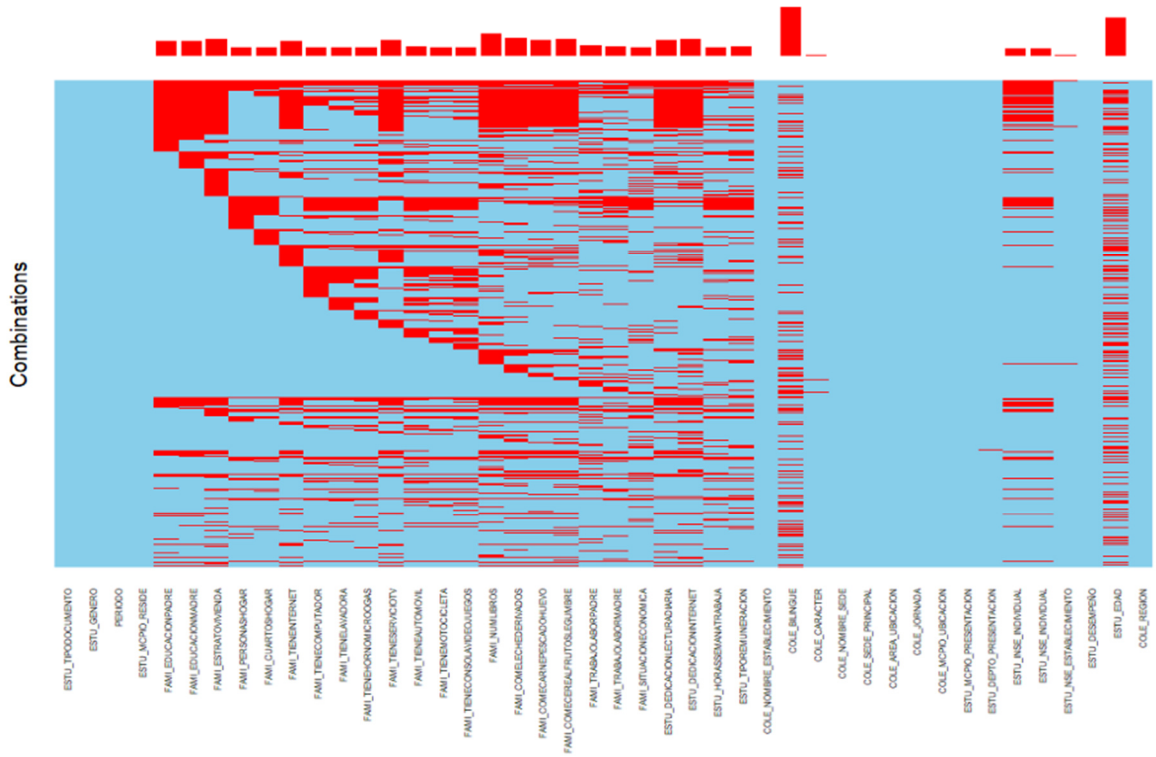
Este problema está relacionado con la forma en que se recolectan y obtienen los datos, pues esto puede afectar significativamente en los resultados del modelo de aprendizaje. Existen muchos métodos para sustituir datos faltantes, uno de ellos es trabajar solo con información completa; en este método se pueden 1) Omitir variables con datos faltantes. 2) Omitir registros completos con datos faltantes. Esta práctica no es la más apropiada, especialmente en conjuntos de datos pequeños, ya que se puede perder información y, así, afectar la representatividad de la muestra, y que puede terminar introduciendo sesgo. Otro método es la imputación de los datos faltantes, que consiste en reemplazar los valores faltantes

con valores predichos basándose en los valores validos de otras variables del conjunto de datos.

En esta investigación se hizo un análisis exploratorio de los valores faltantes; en el Anexo 6. Variables con valores faltantes se describen los valores faltantes del grupo de datos.

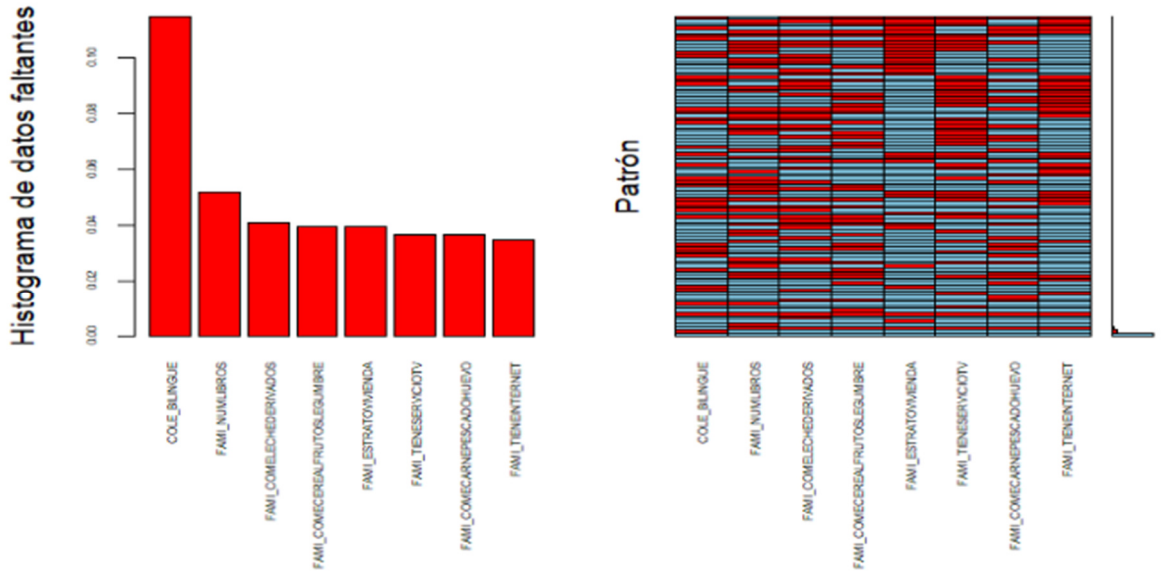
Se generó una gráfica con la distribución de datos faltantes del conjunto de datos utilizando VIM (Visualization and Imputation of Missing Values), que es una librería para visualización en R, ver figura 31 y 32.

Figura 31 Distribución de datos faltantes en el conjunto de datos seleccionado



Fuente: Elaboración propia R Studio

Figura 32 Variables con valores faltantes entre el 3.5% y el 12%



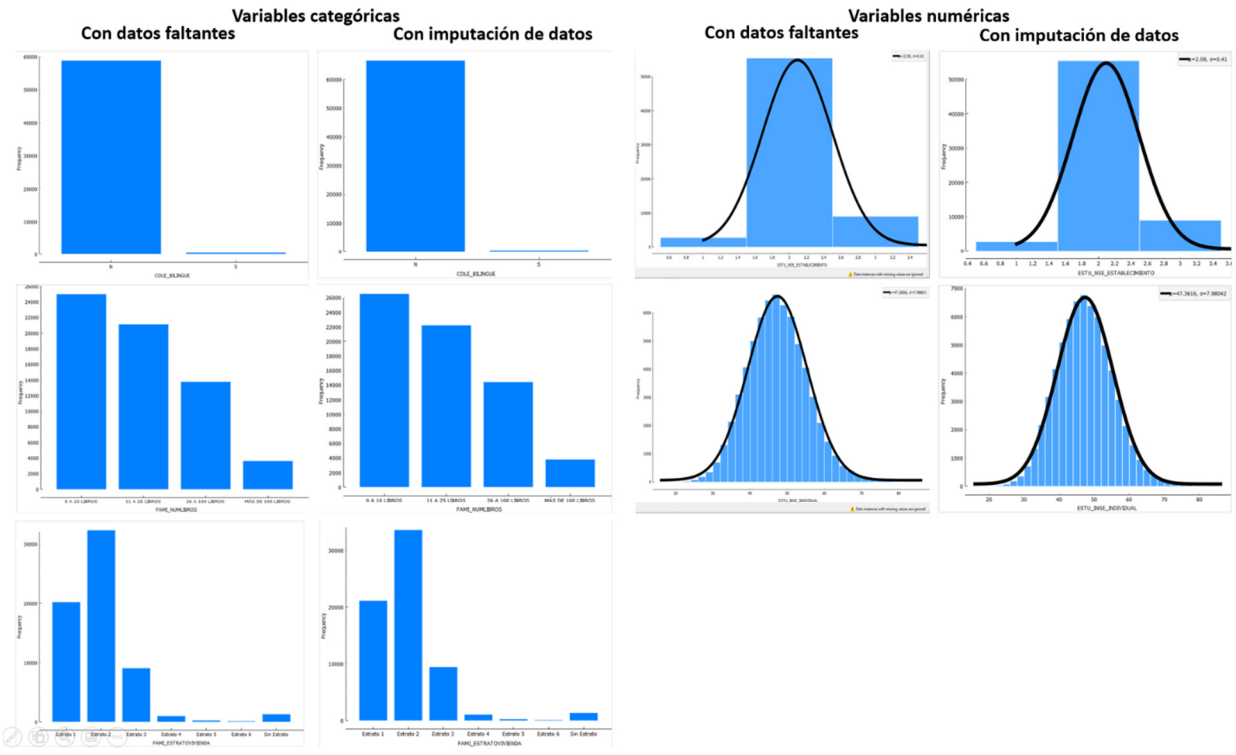
Fuente: elaboración propia R Studio

3.3.4.2 Imputación de los datos

El método de imputación de datos consiste en estimar los valores faltantes analizando los valores validos de otras variables o registros del conjunto de datos. Este ejercicio debe partir del análisis de datos y la imputación, puede ser simple, es decir usar el ultimo valor registrado, el promedio, el valor de moda, máximos y mínimos etc. Otra forma es utilizar ecuaciones o modelos para calcular los valores faltantes mirando el comportamiento de los valores vecinos, utilizando distribuciones normales, medias, regresiones, etc. En esta investigación se utilizaron los paquetes MICE, missForest para el tratamiento de datos ausentes. Anexo 7 Imputación de datos faltantes con MICE

Para comprobar la exactitud o precisión de la imputación se comparó la distribución estadística antes y después de la imputación de los atributos con mayor porcentaje de datos faltantes, idealmente esta imputación no debería cambiar la distribución de los datos. En este ejercicio, se evidenció que el algoritmo aplicado tiene un buen comportamiento ya que genera distribuciones normales (con forma de campana) muy parecidas en las variables numéricas con valores faltantes, así mismo, con las variables de tipo categórico también mantiene las clasificaciones originales, ver la figura 33.

Figura 33 Validación de la calidad de la imputación con MICE



Fuente: Elaboración propia

3.3.5 Formateo de los datos

En el apartado 3.1.5 ya se había desarrollado un primer formateo de datos, con la herramienta Open Refine conciliando valores de los atributos eliminado espacios, caracteres especiales etc. El conjunto de datos seleccionado está conformado por 41 variables categóricas y 3 variables numéricas. Este conjunto de datos tiene atributos con un porcentaje de datos ausentes hasta de un 12% para 33 de los 44 atributos que se van a utilizar en los modelos de predicción. Como ya se mencionó en apartados anteriores, se aplicaron procesos de imputación de datos para estos atributos con problemas de datos ausentes; para este ejercicio se emplearon dos algoritmos de imputación: MICE (imputación múltiple basado en ecuaciones encadenadas) y MissForest (predicción de datos faltantes mediante el método Random Forest).

En el caso de missForest que es la implementación en R (Software estadístico) del método Missing Forest, por una limitación del algoritmo, se deben formatear las variables categóricas con más de 53 categorías y convertir aquellas categorías que exceden este límite agrupándolas en una categoría llamada “Otros”. El conjunto de datos contiene 5 atributos que exceden el límite mencionado en categorías para aplicar el algoritmo missForest que se pueden apreciar en la figura 34.

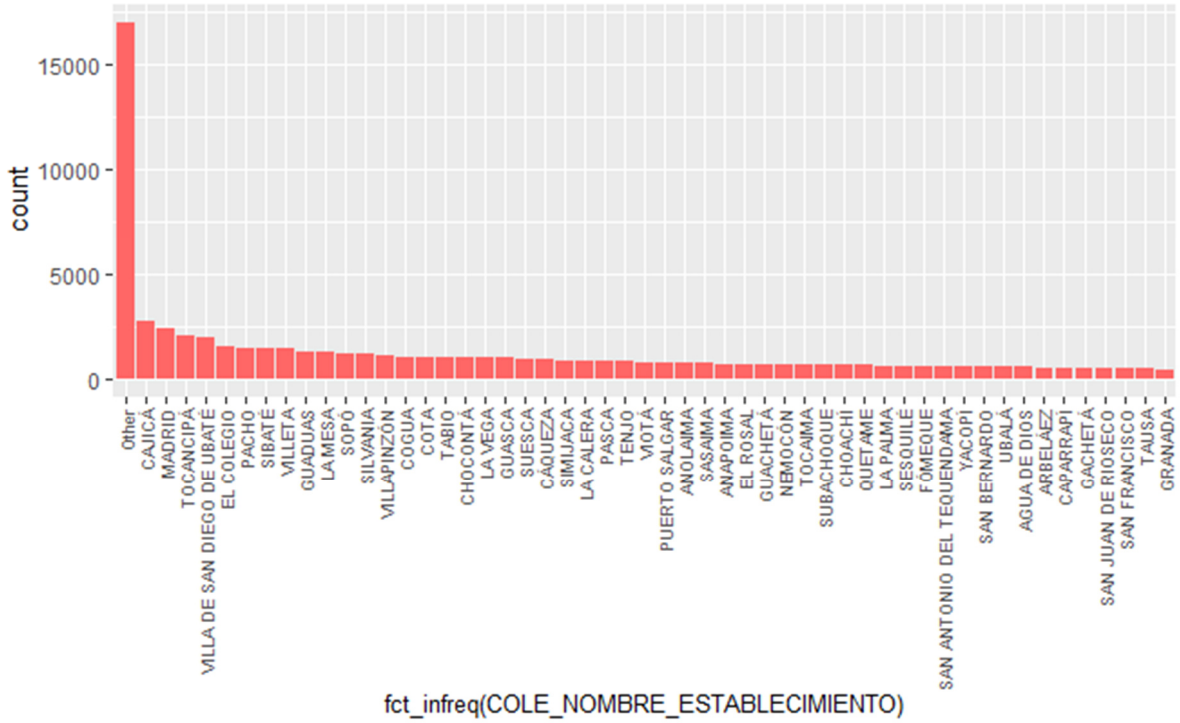
Figura 34 Atributos categóricos con muchas categorías

Feature	data_type_grp	num_unique_values	sample_unique_values	num_missing	perc_missing
ESTU_INSE_INDIVIDUAL	Numerical	65241	[46.74800699, 42.91297853, 45.59716613, 48.611...	1123	1.674894
COLE_NOMBRE_SEDE	Categorical	403	[INSTITUCION EDUCATIVA DEPARTAMENTAL ANTC	0	0
COLE_NOMBRE_ESTABLECIMIENTO	Categorical	323	[INSTITUCION EDUCATIVA DEPARTAMENTAL ANTC	0	0
ESTU_MCPIO_RESIDE	Categorical	174	[RICAURTE, FLANDES, GIRARDOT, JUNÍN, GACHETÁ	11	0.016406
COLE_MCPIO_UBICACION	Categorical	152	[RICAURTE, JUNIN, COTA, NARIÑO, ARBELAEZ, MA	0	0
ESTU_MCPIO_PRESENTACION	Categorical	94	[GIRARDOT, GACHETÁ, CHÍA, ARBELÁEZ, MADRID,	3	0.004474

Fuente: Elaboración propia

R gestiona las variables categóricas con factores y esta tarea se logra utilizando la librería “forcats” en R, que proporciona una serie de herramientas que solucionan problemas comunes con factores, incluido el método de agrupar las categorías más o menos frecuentes en un factor que generalmente se denomina “Otro/s”.

Figura 35 Variable categórica agrupada en "Otros"



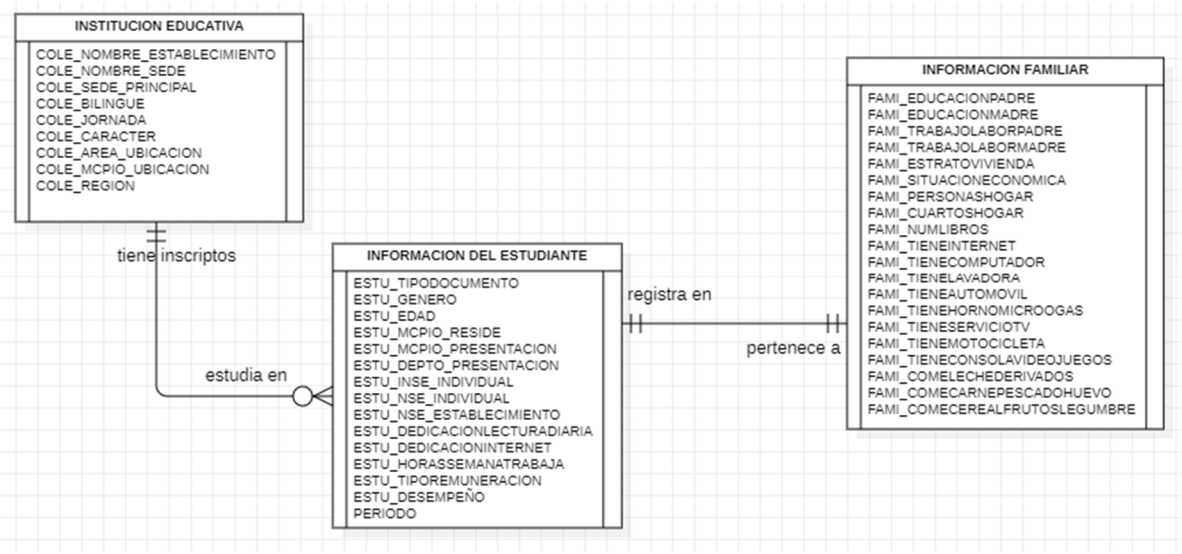
Fuente: elaboración propia en R

3.3.6 Estructurar los datos

En esta fase se afinan los datos a partir de los cuales se generan los modelos de minería de datos, se estructura el conjunto de datos para el análisis, el número y tipo de columnas, se establecen las particiones del conjunto de datos de entrenamiento y de pruebas, se define la variable objetivo o variable dependiente.

Después de ignorar los atributos que no generan valor al análisis, se estructura el conjunto de datos en un modelo lógico que en su primer análisis tiene 4 dimensiones conformadas con información: del estudiante, del colegio, de la familia y de las pruebas Saber 11°. Finalmente, se estructuran solo las tres dimensiones y atributos que pueden ayudar a responder las preguntas de esta investigación. En este modelo lógico se definen claves primarias y foranes como “ESTU_CONSECUTIVO”,”COLE_COD_DANE_ESTABLECIMIENTO”, con dos relaciones básicas de cardinalidad, la primera de uno a muchos y la otra de uno a uno, como se puede apreciar en la figura 36.

Figura 36 Modelo lógico con la estructura del conjunto de datos



Fuente: Elaboración propia

3.3.6.1 Reducción del número de variables

Siendo un objetivo de esta investigación la identificación de los factores determinantes que están relacionados con el desempeño de los estudiantes en las pruebas Saber 11°, es necesario identificar métricas que permitan identificar variables que se pueden eliminar del modelo sin sufrir una considerable perdida de información. Existen varios algoritmos para la selección de variables, por ejemplo, PCA (análisis de componentes principales) funciona muy bien con datos continuos, pero la mayoría de los conjuntos de datos son una combinación de variables continuas y categóricas. Para este estudio se utilizaron dos técnicas, FAMD (Factor Analysis of Mixed Data) por sus siglas en inglés y Boruta, este último por la cantidad de variables identificadas y su baja contribución que seleccionó FAMD.

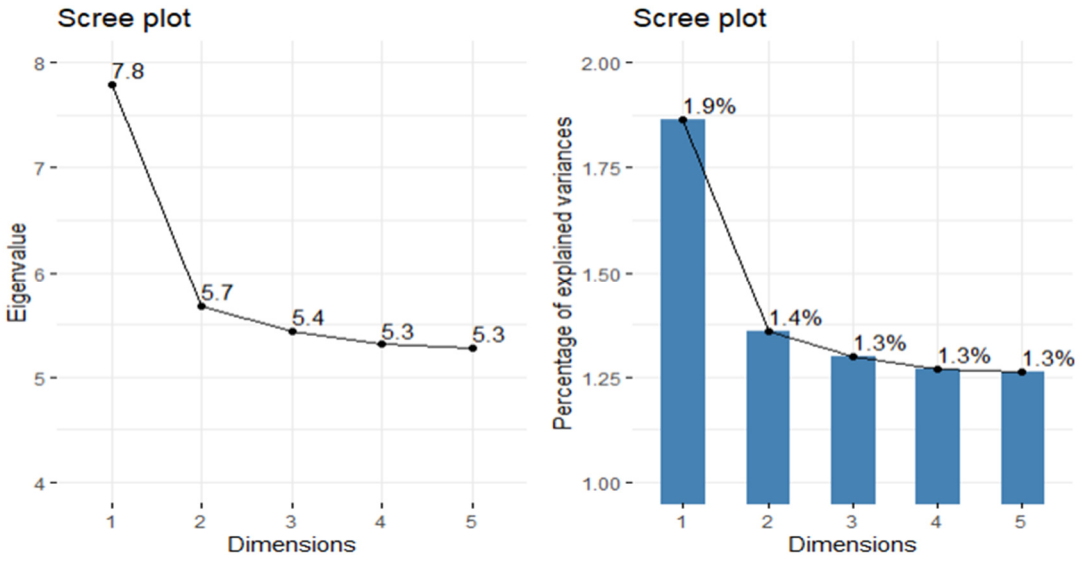
3.3.6.2 Análisis factorial de datos mixtos (FAMD)

Para estos universos de datos mixtos se recomiendan utilizar técnicas como el análisis factorial de datos mixtos (FAMD). FAMD es una combinación de dos técnicas PCA y MCA (análisis de correspondencia múltiple), muy adecuado para el análisis de múltiples variables categóricas, es decir evalúa tanto variables cuantitativas como cualitativas. El grupo de datos seleccionado está compuesto de 48 variables categóricas y 34 variables continuas.

Las dimensiones principales que genera esta técnica son combinaciones lineales de las variables originales que tratan de explicar la variación del conjunto de datos. En la figura 37 se visualiza el “Eigenvalue” y la varianza porcentual para cada

dimensión, en donde se visualiza la capacidad informativa de las variables del conjunto de datos. Todo dato “Eigenvalue” superior a 1 indica que la dimensión arroja más varianza que una de las variables originales. Este es el punto de partida para saber si las dimensiones se pueden usar en análisis posteriores. El grafico de la izquierda en la figura muestran que las tres primeras dimensiones tienen más varianza que cada una de las variables originales, mientras que el grafico de la derecha muestra que las tres dimensiones solo representan el 4.6% de la varianza total del conjunto de datos. Este valor tan bajo se puede originar debido a que las relaciones entre las variables no son lineales (Kassambara, 2017).

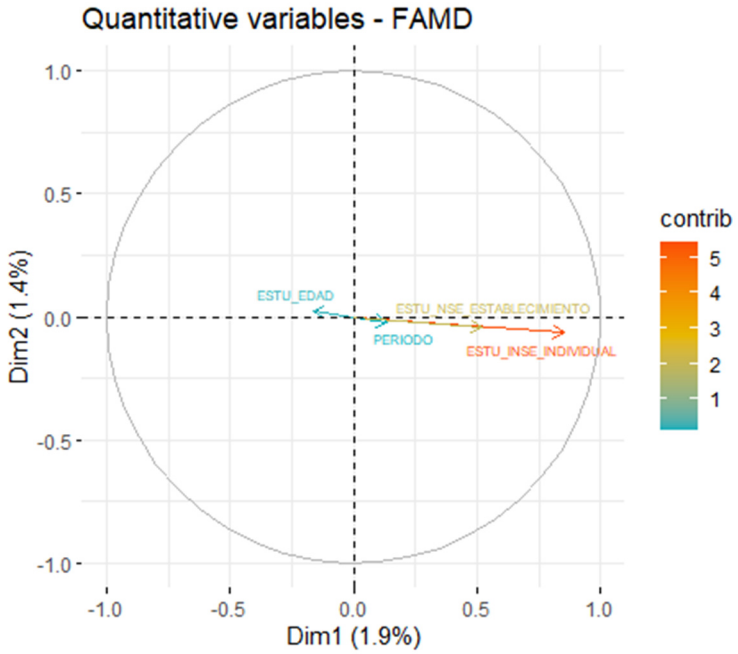
Figura 37 Selección de dimensiones en FAMD



Fuente: Librería FactoMineR

En la figura 38 se observa las correlaciones numéricas entre la Dimensión1 y la Dimensión2. Para este casó la proyección de la variable “ESTU_INSE_INDIVIDUAL” es buena ya que se encuentra muy cerca del borde del círculo, lo que indica que la Dimensión1 y la Dimension2 capturan muy bien la información de la variable. Las demás variables evidencian que se necesitan más de dos dimensiones para capturar la información contenida en cada variable; por eso, la proyección es menor que 1 y se encuentran muy cerca del origen del círculo. Cuanto más cerca este una variable del borde del círculo, más importante será dentro las dimensiones en estudio.

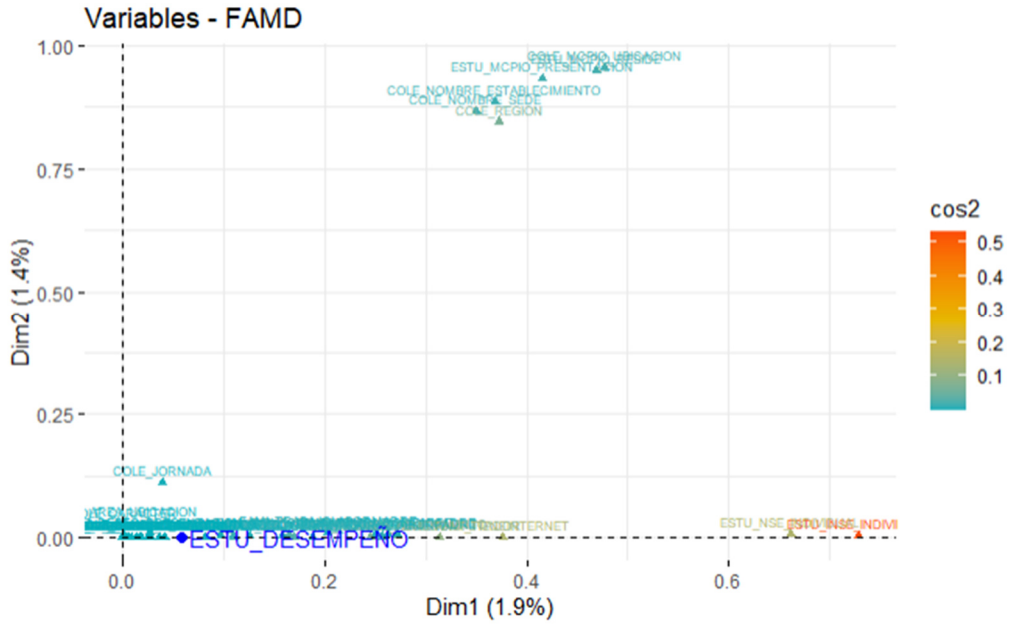
Figura 38 Correlación de variables numéricas en FAMD



Fuente: Librería FactoMineR

En la figura 39 se pueden visualizar variables cualitativas y cuantitativas. En esta grafica se aprecian las variables numéricas (ESTU_INSE_IND y ESTU_NSE_ESTABLECIMIENTO) están más estrechamente relacionada con la Dimensión1, mientras las variables categóricas se describen mejor con las dos dimensiones, a pesar de que están más relacionadas con la Dimensión2. Entre más alejadas estén las variables del origen, son más importantes para explicar la varianza de las dos dimensiones. La variable dependiente se encuentra sobre la Dimensión1, lo que indica que esta dimensión captura perfectamente la variación de esta variable.

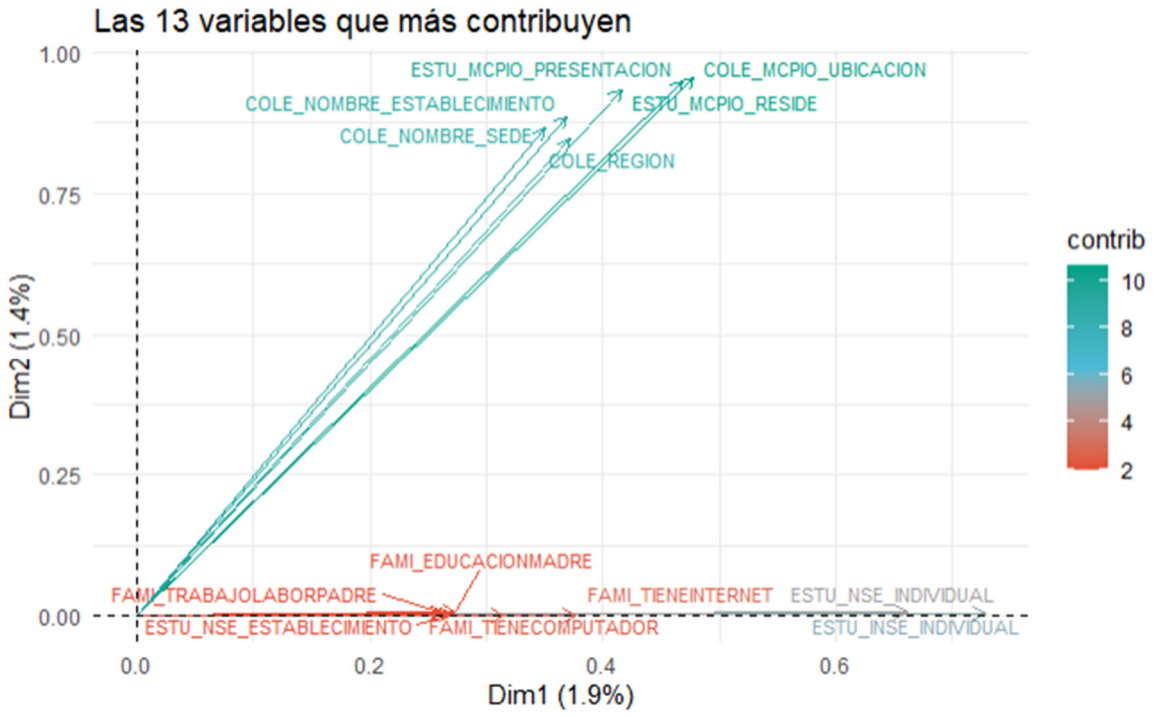
Figura 39 Variables numéricas y categóricas en FAMD



Fuente: Librería FactoMineR

En esta técnica, la contribución es hallar en qué proporción representa una variable la variación total capturada en una dimensión especifica. En la figura 40 se puede apreciar las 13 variables tanto cualitativas, como cuantitativas que más aportan al conjunto de datos.

Figura 40 Contribución de variables en FAMD

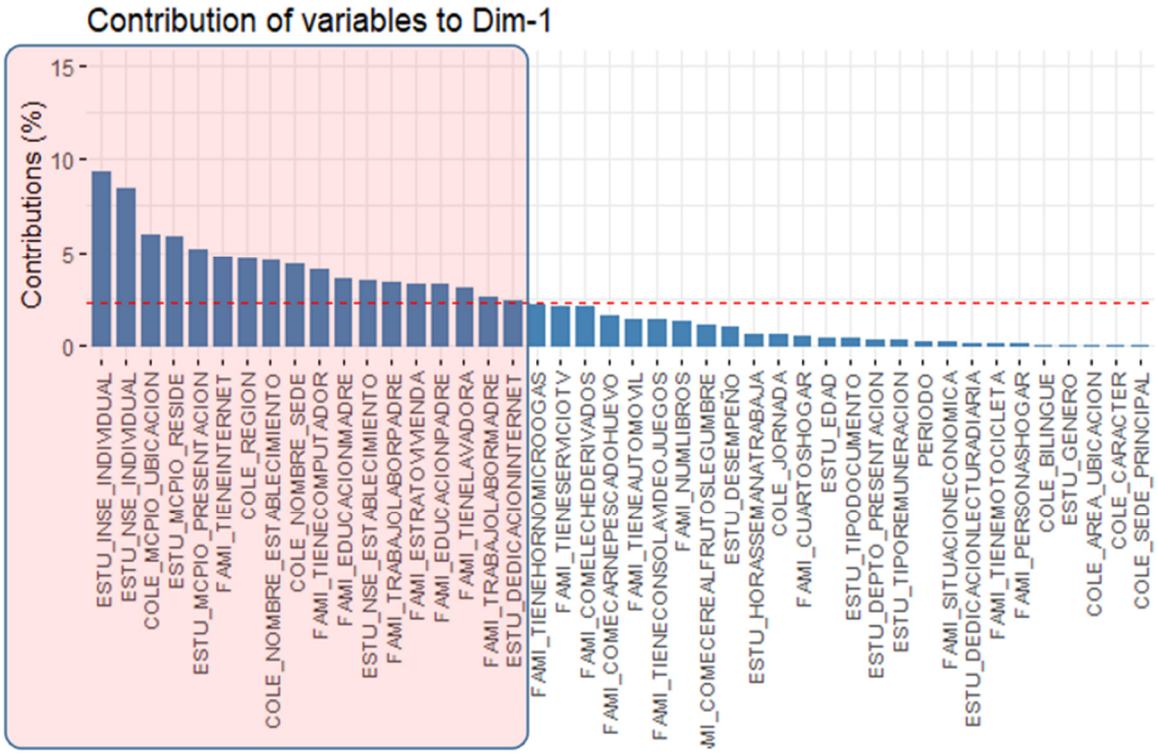


Fuente: Librería FactoMineR

Las variables principales que contribuyen a las primeras dimensiones pueden proporcionar conocimiento sobre qué variables son las que verdaderamente generan variaciones en el conjunto de datos, y pueden ayudar con la selección atributos para análisis posteriores. La línea roja discontinua indica la contribución

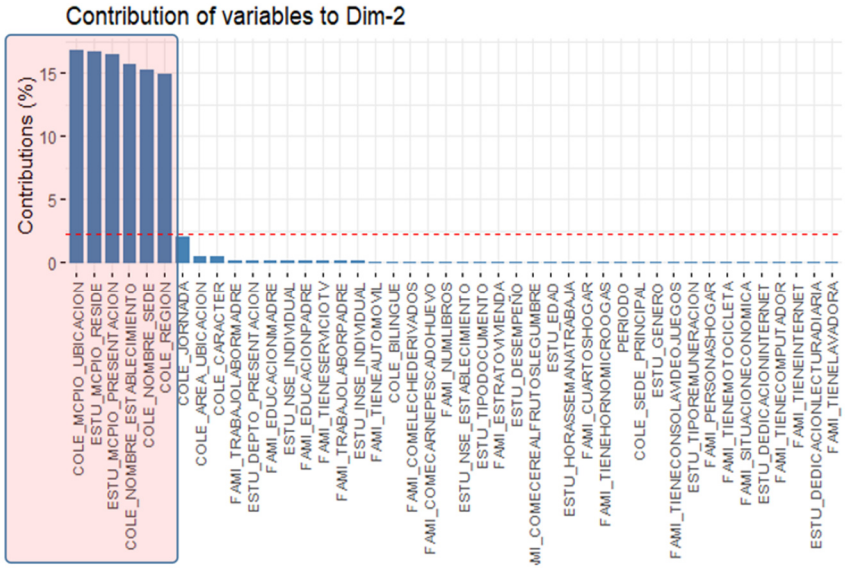
promedio esperada (contribución del 100% dividida en el número total de variables disponibles en el conjunto de datos). Por lo tanto, las variables que por encima de este límite se considerarían importantes para contribuir a la dimensión. En las figuras 41, 42 y 43 se aprecian algunas variables como COLE_MCPIO_UBICACION, FAMI_TIENEINTERNET, COLE_REGION, COLE_NOMBRE_SEDE que están por encima de la línea roja, están entre las variables más importantes en el conjunto de datos (Kassambara, 2017).

Figura 41 variables que contribuyen en la Dimensión1



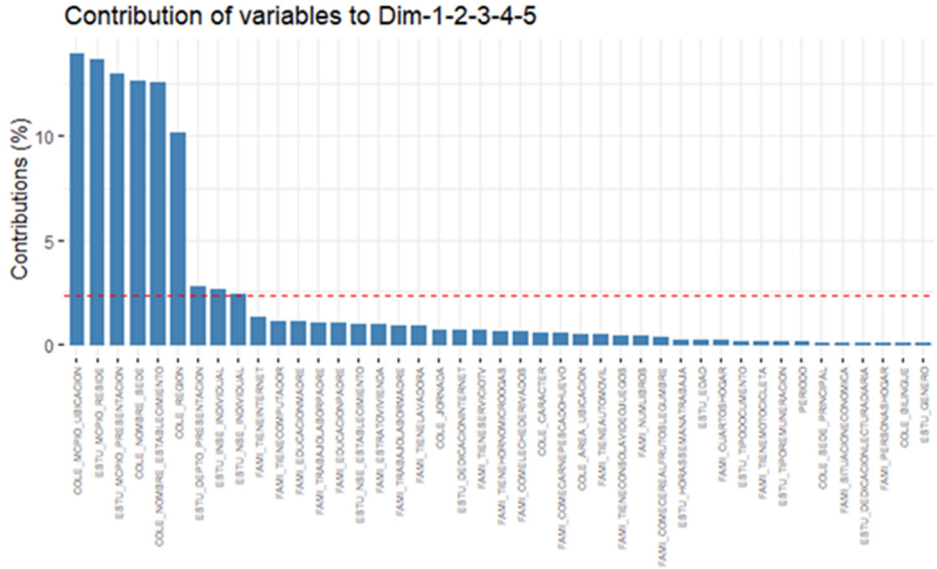
Fuente: Librería FactoMineR

Figura 42 Variables que contribuyen en la Dimensión2



Fuente: Librería FactoMineR

Figura 43 Contribución de variables en todas las dimensiones



Fuente: Librería FactoMineR

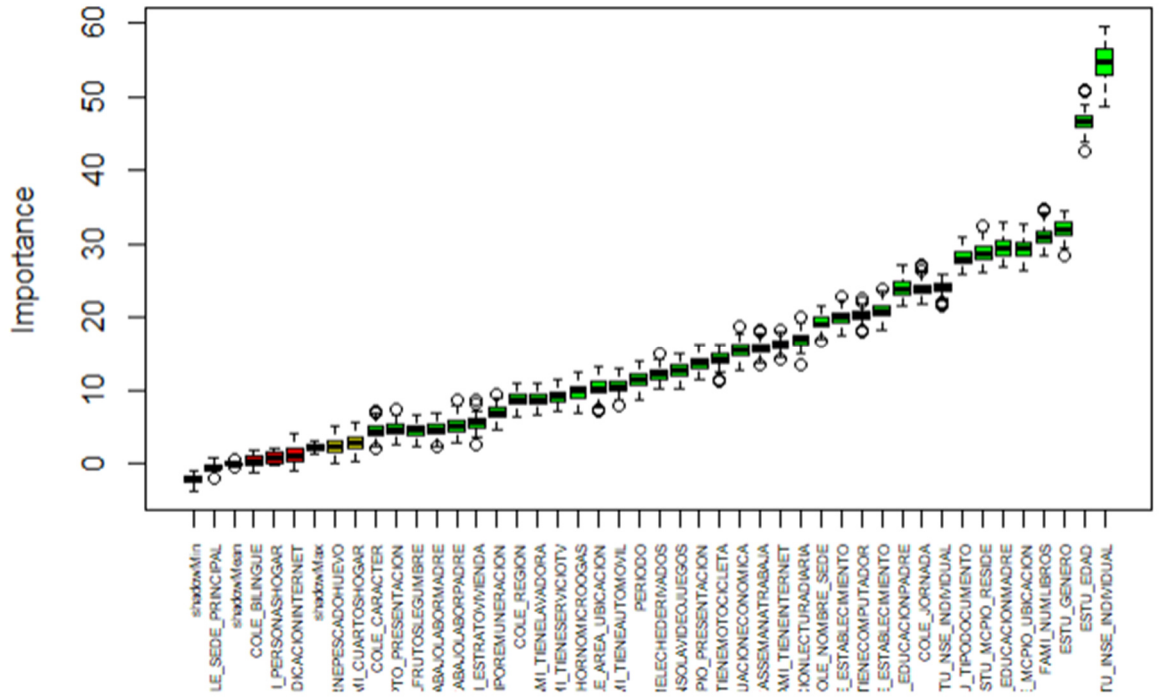
Anexo 8 Selección de variables con FAMD

3.3.6.3 Boruta (Bosques aleatorios)

Boruta es una técnica de selección de atributos o variables del conjunto de datos. La pregunta es ¿Cómo determinar las variables que realmente son útiles? Para esto, una forma de hacerlo es usando Boruta, este algoritmo crea una copia de cada variable que llama “variables sombra” y luego evalúa la importancia de las variables, extrayendo la importancia de cada característica de este modelo y manteniendo solo las características que están por encima de un umbral de importancia determinado.

En este caso de 44 atributos 5 de ellos se rechazan y 38 se confirman. En el grafico las cajas rojas amarillas y verdes representan puntajes Z de atributos rechazados, tentativos y confirmados respectivamente, como se refleja en la figura 44 y 45.

Figura 44 Gráfico de importancia de selección de variables utilizando el algoritmo BORUTA



Fuente: Boruta RStudio

Boruta performed 99 iterations in 19.0479 hours.

Tentatives roughfixed over the last 99 iterations.

38 ***attributes confirmed important****: COLE_AREA_UBICACION, COLE_CARACTER, COLE_JORNADA, COLE_MCPIO_UBICACION, COLE_NOMBRE_ESTABLECIMIENTO and 33 more;*

5 ***attributes confirmed unimportant****: COLE_BILINGUE, COLE_SEDE_PRINCIPAL, ESTU_DEDICACIONINTERNET, FAMI_COMECARNEPESCADOHUEVO, FAMI_PERSONASHOGAR;*

Figura 45 Reporte de variables seleccionadas con Boruta

	meanImp <dbl>	medianImp <dbl>	minImp <dbl>	maxImp <dbl>	normHits <dbl>	decision <fctr>
ESTU_TIPODOCUMENTO	28.1644776	28.0197982	25.8511861	30.9835300	1.0000000	Confirmed
ESTU_GENERO	32.0005498	31.9811885	28.5303248	34.5484684	1.0000000	Confirmed
PERIODO	11.5572440	11.6262313	8.7876431	14.0966969	1.0000000	Confirmed
ESTU_MCPIO_RESIDE	28.8264019	28.6594744	26.1619058	32.4123137	1.0000000	Confirmed
FAMI_EDUCACIONPADRE	23.9822184	23.8519170	21.5269706	27.1468130	1.0000000	Confirmed
FAMI_EDUCACIONMADRE	29.5271045	29.4097052	26.9543248	32.9940767	1.0000000	Confirmed
FAMI_ESTRATOVIVIENDA	5.5477895	5.5649088	2.5156309	8.6010044	1.0000000	Confirmed
FAMI_PERSONASHOGAR	0.8482379	0.8409162	-0.1905244	2.0280656	0.0000000	Rejected
FAMI_CUARTOSHOGAR	2.9165597	2.8132462	0.3229102	5.6894671	0.6767677	Confirmed
FAMI_TIENEINTERNET	16.3937668	16.3674033	14.2523705	18.3081755	1.0000000	Confirmed
FAMI_TIENECOMPUTADOR	20.3999931	20.4113276	18.0039771	22.7305242	1.0000000	Confirmed

Fuente: Boruta RStudio

Anexo 9 Selección de variables con BORUTA

3.4 Modelado

Con el objetivo de identificar las atributos personales, académicos y socioeconómicos determinantes asociados al desempeño de los estudiantes en las pruebas Saber 11° entre el 2017 y 2021 en el departamento de Cundinamarca, se exploraron y evaluaron herramientas de código abierto con algoritmos tanto predictivos como descriptivos. Antes de crear los modelos y después de que los datos estén preparados y adecuados, se debe asegurar que los modelos construidos a partir de los datos preparados funcionen apropiadamente para nuevos datos a clasificar, reducir, agrupar, etc., es decir que el modelo sea válido y se pueda usar en producción.

3.4.1 Seleccionar técnicas de modelado

Se utilizó cada uno de los dos conjuntos de datos obtenidos para la aplicación de algoritmos de reducción de dimensionalidad que permitieron encontrar correlaciones entre la variable Desempeño global y cada una de las 44 variables socioeconómicas. Se aplicaron tres algoritmos diferentes: Correlación entre variables, análisis factorial de datos mixtos (FAMD por sus siglas en inglés) y

BORUTA. Esta fase arrojó un conjunto de 24 variables determinantes según FAMD como se ve en la tabla 16, en las que se encontró una correlación significativa con la variable Desempeño Global. Por otro lado, el mismo proceso de selección de variables con el algoritmo BORUTA arrojó 38 variables, como se refleja en la tabla 17.

Tabla 16 Variables seleccionadas con FAMD

	Variables Seleccionadas con FAMD					
Cant	$contrib	Dim.1	Dim.2	Dim.	Dim.4	Dim.5
1	ESTU_INSE_INDIVIDUAL	6.13074693	1.68E+01	1.70E+01	1.66E+01	1.64E+01
2	ESTU_NSE_INDIVIDUAL	6.01784617	1.67E+01	1.63E+01	1.64E+01	1.61E+01
3	COLE_MCPIO_UBICACION	5.34027305	1.65E+01	1.55E+01	1.57E+01	1.50E+01
4	ESTU_MCPIO_RESIDE	4.49471843	1.53E+01	1.65E+01	1.56E+01	1.51E+01
5	ESTU_MCPIO_PRESENTACION	4.73458598	1.56E+01	1.64E+01	1.50E+01	1.46E+01
6	FAMI_TIENEINTERNET	4.77811434	1.49E+01	9.27E+00	1.13E+01	1.26E+01
7	COLE_REGION	0.35802963	1.41E-01	4.70E+00	6.81E+00	3.39E+00
8	COLE_NOMBRE_ESTABLECIMIENTO	9.35154688	7.50E-02	1.53E-01	8.93E-03	5.77E-01
9	COLE_NOMBRE_SEDE	8.48555055	1.01E-01	1.46E-01	1.20E-02	5.57E-01
10	FAMI_TIENECOMPUTADOR	4.8276792	1.72E-04	2.76E-02	1.75E-03	1.26E-01
11	COLE_JORNADA	3.50310613	1.13E-01	8.73E-02	5.62E-02	6.44E-01
12	COLE_CARACTER	4.03439989	6.03E-04	5.31E-02	1.04E-03	1.31E-01
13	COLE_AREA_UBICACION	3.38230485	8.20E-02	8.29E-02	2.04E-01	3.56E-01
14	ESTU_DEPTO_PRESENTACION	3.21864404	8.78E-02	6.44E-02	7.72E-02	4.45E-01
15	FAMI_TRABAJOLABORPADRE	2.61072641	1.63E-01	9.60E-02	1.30E-01	7.33E-01
16	FAMI_TRABAJOLABORMADRE	3.47345853	3.55E-02	1.75E-03	3.10E-03	1.21E-02
17	FAMI_EDUCACIONMADRE	3.30459117	2.38E-02	7.31E-02	2.92E-02	9.15E-02
18	ESTU_NSE_ESTABLECIMIENTO	0.51522862	1.95E+00	3.84E-01	1.62E-01	4.03E-01
19	FAMI_ESTRATOVIVIENDA	3.15231898	1.45E-04	1.86E-02	4.27E-03	3.93E-02
20	FAMI_EDUCACIONPADRE	0.01875296	4.74E-01	1.04E+00	1.11E+00	1.10E-01
21	FAMI_TIENELAVADORA	2.42080581	1.13E-03	4.16E-02	3.15E-02	1.31E-01
22	ESTU_DEDICACIONINTERNET	2.1222021	9.00E-02	6.11E-02	4.39E-03	1.71E-01
23	FAMI_TIENEHORNOMICROOGAS	2.1795042	1.13E-02	3.83E-03	5.70E-03	5.57E-02
24	FAMI_TIENESERVICIOTV	0.02388062	5.07E-01	8.27E-01	4.95E-02	8.44E-01

Fuente: FAMD RStudio

Tabla 17 Variables confirmadas con Boruta

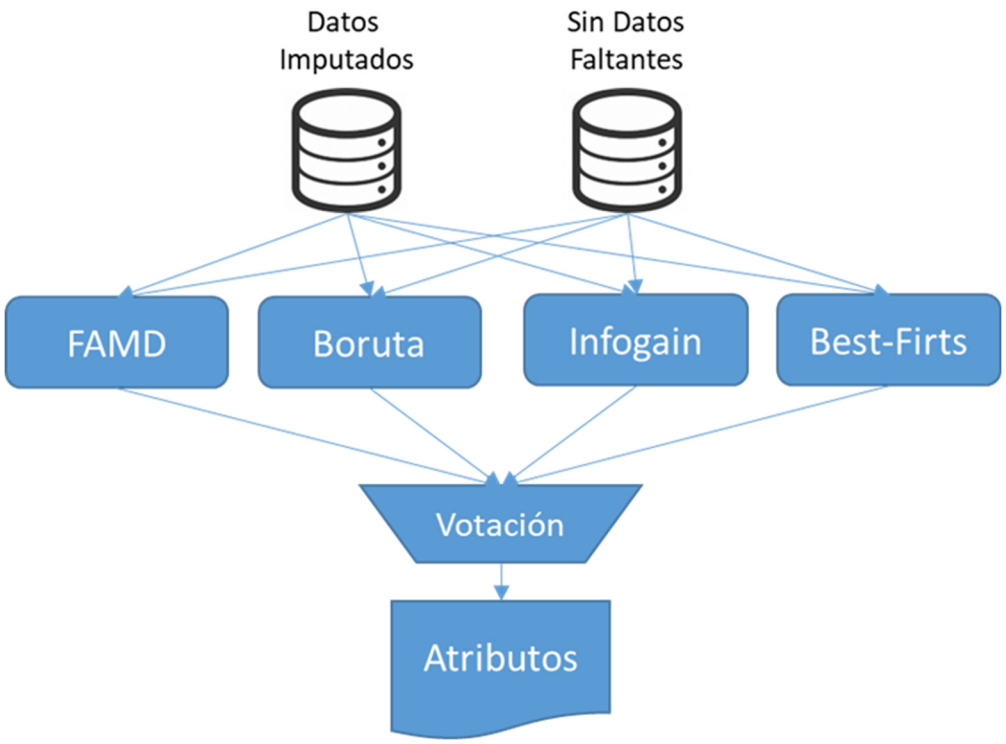
	Variables Seleccionadas con Boruta						
Cant	Variable	meanImp	medianImp	minImp	maxImp	normHits	decision
1	ESTU_INSE_INDIVIDUAL	54.6575102	54.8108373	48.6150049	59.6349935	1	Confirmed
2	ESTU_EDAD	46.6505575	46.5827998	42.7438369	50.8949835	1	Confirmed
3	ESTU_GENERO	32.0005498	31.9811885	28.5303248	34.5484684	1	Confirmed
4	FAMI_NUMLIBROS	31.0987275	31.0361182	28.4638557	34.673579	1	Confirmed
5	FAMI_EDUCACIONMADRE	29.5271045	29.4097052	26.9543248	32.9940767	1	Confirmed
6	COLE_MCPIO_UBICACION	29.4158834	29.4262802	26.4741325	32.7614033	1	Confirmed
7	ESTU_MCPIO_RESIDE	28.8264019	28.6594744	26.1619058	32.4123137	1	Confirmed
8	ESTU_TIPODOCUMENTO	28.1644776	28.0197982	25.8511861	30.98353	1	Confirmed
9	ESTU_NSE_INDIVIDUAL	24.0798024	24.2245771	21.6103311	25.8157747	1	Confirmed
10	COLE_JORNADA	24.0251753	23.9545376	21.9018969	27.2319215	1	Confirmed
11	FAMI_EDUCACIONPADRE	23.9822184	23.851917	21.5269706	27.146813	1	Confirmed
12	COLE_NOMBRE_ESTABLECIMIENTO	21.1006353	20.9373465	18.3825336	23.8530696	1	Confirmed
13	FAMI_TIENECOMPUTADOR	20.3999931	20.4113276	18.0039771	22.7305242	1	Confirmed
14	ESTU_NSE_ESTABLECIMIENTO	20.0393592	20.098174	17.6899801	22.8663843	1	Confirmed
15	COLE_NOMBRE_SEDE	19.400636	19.1663445	16.7049179	21.6699521	1	Confirmed
16	ESTU_DEDICACIONLECTURADIARIA	17.0242522	17.0997004	13.6444912	20.0700188	1	Confirmed
17	FAMI_TIENEINTERNET	16.3937668	16.3674033	14.2523705	18.3081755	1	Confirmed
18	ESTU_HORASSEMANATRABAJA	15.8877885	15.8377832	13.489533	18.2617744	1	Confirmed
19	FAMI_SITUACIONECONOMICA	15.6746365	15.6173195	12.7437965	18.785049	1	Confirmed
20	FAMI_TIENEMOTOCICLETA	14.3132106	14.3184941	11.1881214	16.3021962	1	Confirmed
21	ESTU_MCPIO_PRESENTACION	13.7035026	13.8188531	11.4240862	16.2568119	1	Confirmed
22	FAMI_TIENECONSOLAVIDEOJUEGOS	12.8144559	12.866711	10.2413117	15.1162729	1	Confirmed
23	FAMI_COMELECHEDERIVADOS	12.2239096	12.1763094	10.2363095	15.0922748	1	Confirmed
24	PERIODO	11.557244	11.6262313	8.7876431	14.0966969	1	Confirmed
25	FAMI_TIENEAUTOMOVIL	10.5755748	10.5734726	7.8759458	12.9423395	1	Confirmed
26	COLE_AREA_UBICACION	10.3890549	10.3106988	7.1785407	13.3019578	1	Confirmed
27	FAMI_TIENEHORNOMICROOGAS	9.8067626	9.8761694	6.978536	12.5317412	1	Confirmed
28	FAMI_TIENESERVICIOTV	9.1051651	9.1136633	7.2203595	11.5740837	1	Confirmed
29	FAMI_TIENELAVADORA	8.7879973	8.6892121	6.7946208	11.0713915	1	Confirmed
30	COLE_REGION	8.7665875	8.6507517	6.5522659	11.0170793	1	Confirmed
31	ESTU_TIPOREMUNERACION	6.9488309	6.9075389	4.6461907	9.5548377	1	Confirmed
32	FAMI_ESTRATOVIVIENDA	5.5477895	5.5649088	2.5156309	8.6010044	1	Confirmed
33	FAMI_TRABAJOLABORPADRE	5.233589	5.1205792	2.9904251	8.7181519	1	Confirmed
34	FAMI_TRABAJOLABORMADRE	4.7674007	4.7627335	2.3035377	7.0645514	0.959596	Confirmed
35	ESTU_DEPTO_PRESENTACION	4.7253668	4.6803002	2.7330694	7.4520436	1	Confirmed
36	FAMI_COMECEREALFRUTOSLEGUMBRE	4.6505599	4.7218584	2.4063102	6.6082007	0.989899	Confirmed
37	COLE_CARACTER	4.5423213	4.4301577	2.1260327	7.2888264	0.969697	Confirmed
38	FAMI_CUARTOSHOGAR	2.9165597	2.8132462	0.3229102	5.6894671	0.6767677	Confirmed

Fuente: Boruta RStudio

Con el objetivo de encontrar los factores del conjunto de datos con mayor relación o influencia con la variable dependiente, en esta investigación se ejecuta un sistema de votación que permita identificar las variables más seleccionadas por

las diferentes técnicas de ranqueo y selección de atributos, como se puede apreciar en la figura 46.

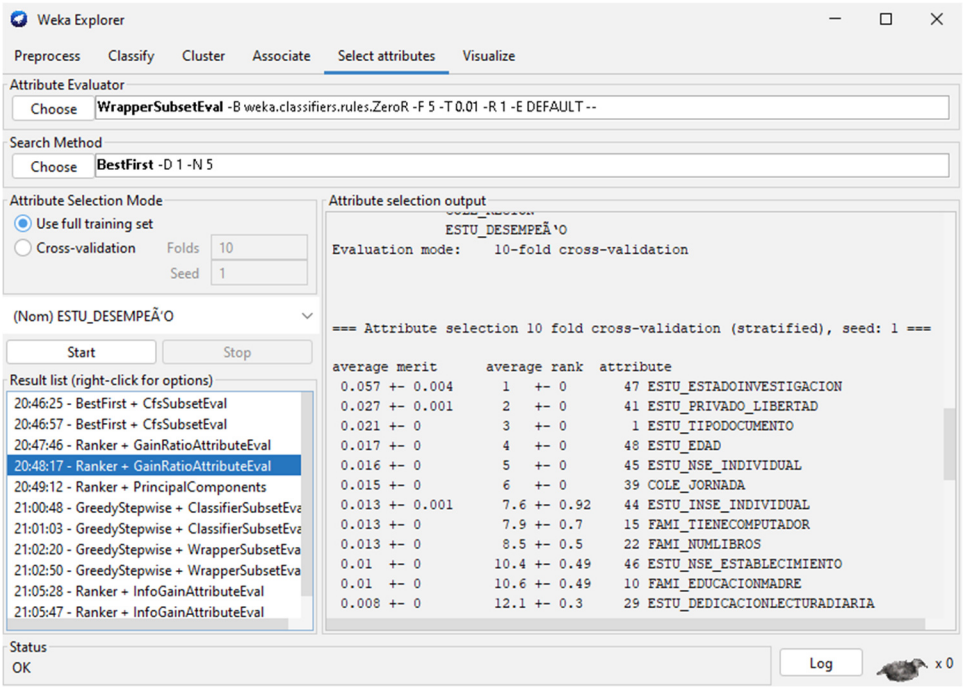
Figura 46 Sistema de votación de variables



Fuente: Elaboración propia

Dentro del sistema de votación para la selección de características se utiliza el atributo evaluador InfoGainAttributeEval de la herramienta WEKA (Waikato Environment for Knowledge Analysis) como se visualiza en la figura 47, para correr varios algoritmos que seleccionan subconjuntos de variables y poder incluirlos en el sistema de selección y votación de atributos.

Figura 47 Algoritmos para ranqueo y selección de variables



Evaluator: weka.attributeSelection.InfoGainAttributeEval
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation: DataTable-weka.filters.unsupervised.attribute.Remove-R44-60
Instances: 67049

Ranked	attributes:
0.0328606	48 ESTU_EDAD
0.0322809	44 ESTU_INSE_INDIVIDUAL
0.0280624	39 COLE_JORNADA
0.027154	10 FAMI_EDUCACIONMADRE
0.0268325	45 ESTU_NSE_INDIVIDUAL
0.0231949	9 FAMI_EDUCACIONPADRE
0.0227712	22 FAMI_NUMLIBROS
0.0172388	1 ESTU_TIPODOCUMENTO
0.0169441	40 COLE_MCPIO_UBICACION
0.0162475	29 ESTU_DEDICACIONLECTURADIARIA
0.0150824	8 ESTU_MCPIO_RESIDE
0.0139689	33 COLE_NOMBRE_ESTABLECIMIENTO
0.0132859	23 FAMI_COMELECHEDERIVADOS
0.0129888	15 FAMI_TIENECOMPUTADOR
0.0129363	36 COLE_NOMBRE_SEDE
0.0111532	31 ESTU_HORASSEMANATRABAJA
0.0111243	26 FAMI_TRABAJOLABORPADRE
0.0109929	42 ESTU_MCPIO_PRESENTACION
0.0095843	27 FAMI_TRABAJOLABORMADRE
0.0095199	30 ESTU_DEDICACIONINTERNET
0.0081604	11 FAMI_ESTRATOVIVIENDA
0.0078498	46 ESTU_NSE_ESTABLECIMIENTO
0.0078132	14 FAMI_TIENEINTERNET
0.0072531	49 COLE_REGION
0.0065859	3 ESTU_GENERO
0.005172	25 FAMI_COMECEREALFRUTOSLEGUMBRE
0.0050291	24 FAMI_COMECARNEPESCADOHUEVO
0.0049962	28 FAMI_SITUACIONECONOMICA
0.004307	32 ESTU_TIPOREMUNERACION
0.0039838	16 FAMI_TIENELAVADORA
0.0039144	20 FAMI_TIENEMOTOCICLETA

Evaluator: weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search: weka.attributeSelection.BestFirst -D 1 -N 5
Relation: DataTable-weka.filters.unsupervised.attribute.Remove-R44-60
Instances: 67049

number of folds (%)	attribute
10(100 %)	1 ESTU_TIPODOCUMENTO
10(100 %)	3 ESTU_GENERO
10(100 %)	9 FAMI_EDUCACIONPADRE
10(100 %)	10 FAMI_EDUCACIONMADRE
10(100 %)	15 FAMI_TIENECOMPUTADOR
10(100 %)	22 FAMI_NUMLIBROS
10(100 %)	23 FAMI_COMELECHEDERIVADOS
10(100 %)	29 ESTU_DEDICACIONLECTURADIARIA
10(100 %)	31 ESTU_HORASSEMANATRABAJA
10(100 %)	39 COLE_JORNADA
10(100 %)	45 ESTU_NSE_INDIVIDUAL
10(100 %)	46 ESTU_NSE_ESTABLECIMIENTO
10(100 %)	47 ESTU_ESTADOINVESTIGACION
10(100 %)	48 ESTU_EDAD

Fuente: Weka

[Anexo 10 Selección de variables con Weka]

Con las técnicas anteriores se han mostrado resultados parciales de la aplicación de cada algoritmo en el proceso de selección variables. Posteriormente, se procedió a ordenar por grado de importancia el reporte de atributos de cada técnica, logrando determinar las variables que más se repitan en la mayoría de las técnicas como se aprecia en la tabla 18, estas variables están en orden descendente y son las más representativas o correlacionadas con la variable dependiente.

Tabla 18 Lista de variables por técnica y orden de importancia

Naive Bayes	InfoGain	BORUTA	FAMD	Random Forest
ESTU_INSE_INDIVIDUAL	ESTU_EDAD	ESTU_INSE_INDIVIDUAL	ESTU_INSE_INDIVIDUAL	ESTU_EDAD
FAMI_ESTRATOVIVIENDA	ESTU_INSE_INDIVIDUAL	ESTU_EDAD	ESTU_NSE_INDIVIDUAL	ESTU_INSE_INDIVIDUAL
COLE_MCPIO_UBICACION	COLE_JORNADA	ESTU_GENERO	COLE_MCPIO_UBICACION	FAMI_EDUCACIONPADRE
FAMI_COMELECHEDERIVADOS	FAMI_EDUCACIONMADRE	FAMI_NUMLIBROS	ESTU_MCPIO_RESIDE	ESTU_MCPIO_RESIDE
FAMI_TRABAJOLABORMADRE	ESTU_NSE_INDIVIDUAL	FAMI_EDUCACIONMADRE	ESTU_MCPIO_PRESENTACION	FAMI_EDUCACIONMADRE
FAMI_TIENEINTERNET	FAMI_EDUCACIONPADRE	COLE_MCPIO_UBICACION	FAMI_TIENEINTERNET	COLE_MCPIO_UBICACION
ESTU_EDAD	FAMI_NUMLIBROS	ESTU_MCPIO_RESIDE	COLE_REGION	COLE_JORNADA
FAMI_TRABAJOLABORPADRE	ESTU_TIPODOCUMENTO	ESTU_TIPODOCUMENTO	COLE_NOMBRE_ESTABLECIMIENTO	COLE_NOMBRE_ESTABLECIMIENTO
FAMI_TIENECOMPUTADOR	COLE_MCPIO_UBICACION	ESTU_NSE_INDIVIDUAL	COLE_NOMBRE_SEDE	COLE_NOMBRE_SEDE
ESTU_MCPIO_RESIDE	ESTU_DEDICACIONLECTURADIARIA	COLE_JORNADA	FAMI_TIENECOMPUTADOR	ESTU_NSE_INDIVIDUAL
ESTU_MCPIO_PRESENTACION	ESTU_MCPIO_RESIDE	FAMI_EDUCACIONPADRE	COLE_JORNADA	FAMI_NUMLIBROS
FAMI_TIENELAVADORA	COLE_NOMBRE_ESTABLECIMIENTO	COLE_NOMBRE_ESTABLECIMIENTO	COLE_CARACTER	ESTU_MCPIO_PRESENTACION
FAMI_EDUCACIONMADRE	FAMI_COMELECHEDERIVADOS	FAMI_TIENECOMPUTADOR	COLE_AREA_UBICACION	ESTU_DEDICACIONLECTURADIARIA
FAMI_TIENEAUTOMOVIL	FAMI_TIENECOMPUTADOR	ESTU_NSE_ESTABLECIMIENTO	ESTU_DEPTO_PRESENTACION	FAMI_TRABAJOLABORMADRE
COLE_NOMBRE_SEDE	COLE_NOMBRE_SEDE	COLE_NOMBRE_SEDE	FAMI_TRABAJOLABORPADRE	FAMI_TRABAJOLABORPADRE
FAMI_TIENEHORNOMICROOGAS	ESTU_HORASSEMANATRABAJA	ESTU_DEDICACIONLECTURADIARIA	FAMI_TRABAJOLABORMADRE	ESTU_HORASSEMANATRABAJA
ESTU_NSE_ESTABLECIMIENTO	FAMI_TRABAJOLABORPADRE	FAMI_TIENEINTERNET	FAMI_EDUCACIONMADRE	ESTU_TIPODOCUMENTO
COLE_JORNADA	ESTU_MCPIO_PRESENTACION	ESTU_HORASSEMANATRABAJA	ESTU_NSE_ESTABLECIMIENTO	ESTU_GENERO
FAMI_EDUCACIONPADRE	FAMI_TRABAJOLABORMADRE	FAMI_SITUACIONECONOMICA	FAMI_ESTRATOVIVIENDA	FAMI_COMELECHEDERIVADOS
ESTU_DEPTO_PRESENTACION	ESTU_DEDICACIONINTERNET	FAMI_TIENEMOTOCICLETA	FAMI_EDUCACIONPADRE	ESTU_NSE_ESTABLECIMIENTO
FAMI_TIENECONSOLAVIDEOJUEGOS	FAMI_ESTRATOVIVIENDA	ESTU_MCPIO_PRESENTACION	FAMI_TIENELAVADORA	COLE_REGION
FAMI_NUMLIBROS	ESTU_NSE_ESTABLECIMIENTO	FAMI_TIENECONSOLAVIDEOJUEGOS	ESTU_DEDICACIONINTERNET	FAMI_COMECARNEPESCADOHUEVO
FAMI_PERSONASHOGAR	FAMI_TIENEINTERNET	FAMI_COMELECHEDERIVADOS	FAMI_TIENEHORNOMICROOGAS	FAMI_ESTRATOVIVIENDA
COLE_CARACTER	COLE_REGION	PERIODO	FAMI_TIENESERVICIOTV	FAMI_TIENEMOTOCICLETA
COLE_NOMBRE_ESTABLECIMIENTO	ESTU_GENERO	FAMI_TIENEAUTOMOVIL	FAMI_COMELECHEDERIVADOS	ESTU_TIPOREMUNERACION
ESTU_TIPODOCUMENTO	FAMI_COMECEREALFRUTOSLEGUMBRE	COLE_AREA_UBICACION	FAMI_COMECARNEPESCADOHUEVO	FAMI_SITUACIONECONOMICA
FAMI_CUARTOSHOGAR	FAMI_COMECARNEPESCADOHUEVO	FAMI_TIENEHORNOMICROOGAS	FAMI_TIENECONSOLAVIDEOJUEGOS	FAMI_TIENECOMPUTADOR
ESTU_DEDICACIONLECTURADIARIA	FAMI_SITUACIONECONOMICA	FAMI_TIENESERVICIOTV	FAMI_TIENEAUTOMOVIL	FAMI_CUARTOSHOGAR
COLE_AREA_UBICACION	ESTU_TIPOREMUNERACION	FAMI_TIENELAVADORA	FAMI_NUMLIBROS	FAMI_COMECEREALFRUTOSLEGUMBRE
ESTU_NSE_INDIVIDUAL	FAMI_TIENELAVADORA	COLE_REGION	FAMI_COMECEREALFRUTOSLEGUMBRE	ESTU_DEDICACIONINTERNET
FAMI_SITUACIONECONOMICA	FAMI_TIENEMOTOCICLETA	ESTU_TIPOREMUNERACION	ESTU_HORASSEMANATRABAJA	FAMI_TIENEINTERNET
ESTU_GENERO	ESTU_DEPTO_PRESENTACION	FAMI_ESTRATOVIVIENDA	FAMI_CUARTOSHOGAR	FAMI_PERSONASHOGAR
FAMI_TIENEMOTOCICLETA	FAMI_PERSONASHOGAR	FAMI_TRABAJOLABORPADRE	ESTU_EDAD	PERIODO
ESTU_DEDICACIONINTERNET	FAMI_TIENEAUTOMOVIL	FAMI_TRABAJOLABORMADRE	ESTU_TIPODOCUMENTO	COLE_CARACTER
FAMI_COMECEREALFRUTOSLEGUMBRE	PERIODO	ESTU_DEPTO_PRESENTACION	ESTU_TIPOREMUNERACION	COLE_AREA_UBICACION
ESTU_HORASSEMANATRABAJA	FAMI_CUARTOSHOGAR	FAMI_COMECEREALFRUTOSLEGUMBRE	PERIODO	FAMI_TIENESERVICIOTV
ESTU_TIPOREMUNERACION	FAMI_TIENEHORNOMICROOGAS	COLE_CARACTER	FAMI_SITUACIONECONOMICA	ESTU_DEPTO_PRESENTACION
FAMI_COMECARNEPESCADOHUEVO	COLE_AREA_UBICACION	FAMI_CUARTOSHOGAR	ESTU_DEDICACIONLECTURADIARIA	FAMI_TIENELAVADORA
FAMI_TIENESERVICIOTV	COLE_CARACTER	FAMI_COMECARNEPESCADOHUEVO	FAMI_TIENEMOTOCICLETA	FAMI_TIENEAUTOMOVIL
PERIODO	FAMI_TIENECONSOLAVIDEOJUEGOS	ESTU_DEDICACIONINTERNET	FAMI_PERSONASHOGAR	FAMI_TIENECONSOLAVIDEOJUEGOS
COLE_REGION	FAMI_TIENESERVICIOTV	FAMI_PERSONASHOGAR	ESTU_GENERO	FAMI_TIENEHORNOMICROOGAS
COLE_SEDE_PRINCIPAL	COLE_SEDE_PRINCIPAL	COLE_SEDE_PRINCIPAL	COLE_SEDE_PRINCIPAL	COLE_SEDE_PRINCIPAL

Fuente: Elaboración propia

Al tomar el 50% de las variables presentadas la tabla 18 y ordenarlas por el número de repeticiones, es decir aquellas que fueron identificas en más de una ocasión por las diferentes técnicas aplicadas para encontrar los atributos más determinantes para predecir el comportamiento de la variable dependiente; se generaron grupos de atributos que fueron seleccionadas por 5,4,3 y 2 de las cinco (5) técnicas utilizadas de selección de variables. El resultado de este sistema de votación se visualiza en la figura 19.

Tabla 19 Lista de variables por técnica y repetición en cada técnica

Random Forest	InfoGain	BORUTA	FAMD	Naive Bayes	Total Repeticiones
ESTU_INSE_INDIVIDUAL	ESTU_INSE_INDIVIDUAL	ESTU_INSE_INDIVIDUAL	ESTU_INSE_INDIVIDUAL	ESTU_INSE_INDIVIDUAL	5
FAMI_EDUCACIONPADRE	FAMI_EDUCACIONPADRE	FAMI_EDUCACIONPADRE	FAMI_EDUCACIONPADRE	FAMI_EDUCACIONPADRE	5
ESTU_MCPIO_RESIDE	ESTU_MCPIO_RESIDE	ESTU_MCPIO_RESIDE	ESTU_MCPIO_RESIDE	ESTU_MCPIO_RESIDE	5
FAMI_EDUCACIONMADRE	FAMI_EDUCACIONMADRE	FAMI_EDUCACIONMADRE	FAMI_EDUCACIONMADRE	FAMI_EDUCACIONMADRE	5
COLE_MCPIO_UBICACION	COLE_MCPIO_UBICACION	COLE_MCPIO_UBICACION	COLE_MCPIO_UBICACION	COLE_MCPIO_UBICACION	5
COLE_JORNADA	COLE_JORNADA	COLE_JORNADA	COLE_JORNADA	COLE_JORNADA	5
COLE_NOMBRE_SEDE	COLE_NOMBRE_SEDE	COLE_NOMBRE_SEDE	COLE_NOMBRE_SEDE	COLE_NOMBRE_SEDE	5
ESTU_MCPIO_PRESENTACION	ESTU_MCPIO_PRESENTACION	ESTU_MCPIO_PRESENTACION	ESTU_MCPIO_PRESENTACION	ESTU_MCPIO_PRESENTACION	5
ESTU_NSE_ESTABLECIMIENTO	ESTU_NSE_ESTABLECIMIENTO	ESTU_NSE_ESTABLECIMIENTO	ESTU_NSE_ESTABLECIMIENTO	ESTU_NSE_ESTABLECIMIENTO	5
ESTU_EDAD	ESTU_EDAD	ESTU_EDAD		ESTU_EDAD	4
COLE_NOMBRE_ESTABLECIMIENTO	COLE_NOMBRE_ESTABLECIMIENTO	COLE_NOMBRE_ESTABLECIMIENTO	COLE_NOMBRE_ESTABLECIMIENTO		4
ESTU_NSE_INDIVIDUAL	ESTU_NSE_INDIVIDUAL	ESTU_NSE_INDIVIDUAL	ESTU_NSE_INDIVIDUAL		4
FAMI_NUMLIBROS	FAMI_NUMLIBROS	FAMI_NUMLIBROS		FAMI_NUMLIBROS	4
FAMI_TRABAJOLABORMADRE	FAMI_TRABAJOLABORMADRE		FAMI_TRABAJOLABORMADRE	FAMI_TRABAJOLABORMADRE	4
FAMI_TRABAJOLABORPADRE	FAMI_TRABAJOLABORPADRE		FAMI_TRABAJOLABORPADRE	FAMI_TRABAJOLABORPADRE	4
FAMI_COMELECHEDERIVADOS	FAMI_COMELECHEDERIVADOS	FAMI_COMELECHEDERIVADOS		FAMI_COMELECHEDERIVADOS	4
FAMI_ESTRATOVIVIENDA	FAMI_ESTRATOVIVIENDA		FAMI_ESTRATOVIVIENDA	FAMI_ESTRATOVIVIENDA	4
	FAMI_TIENEINTERNET	FAMI_TIENEINTERNET	FAMI_TIENEINTERNET	FAMI_TIENEINTERNET	4
	FAMI_TIENECOMPUTADOR	FAMI_TIENECOMPUTADOR	FAMI_TIENECOMPUTADOR	FAMI_TIENECOMPUTADOR	4
ESTU_DEDICACIONLECTURADIARIA	ESTU_DEDICACIONLECTURADIARIA	ESTU_DEDICACIONLECTURADIARIA			3
ESTU_HORASSEMANATRABAJA	ESTU_HORASSEMANATRABAJA	ESTU_HORASSEMANATRABAJA			3
ESTU_TIPODOCUMENTO	ESTU_TIPODOCUMENTO	ESTU_TIPODOCUMENTO			3
ESTU_GENERO		ESTU_GENERO			2
COLE_REGION			COLE_REGION		2
	ESTU_DEDICACIONINTERNET		ESTU_DEDICACIONINTERNET		2
			ESTU_DEPTO_PRESENTACION	ESTU_DEPTO_PRESENTACION	2
			FAMI_TIENEHORNOMICROOGAS	FAMI_TIENEHORNOMICROOGAS	2
		FAMI_TIENECONSOLAVIDEOJUEGOS		FAMI_TIENECONSOLAVIDEOJUEGOS	2
			FAMI_TIENELAVADORA	FAMI_TIENELAVADORA	2
FAMI_COMECARNEPESCADOHUEVO					1
			COLE_CARACTER		1
		FAMI_TIENEMOTOCICLETA			1
			COLE_AREA_UBICACION		1
				FAMI_PERSONASHOGAR	1
				FAMI_TIENEAUTOMOVIL	1
		FAMI_SITUACIONECONOMICA			1
23	23	23	23	23	

Fuente: Elaboración propia

Con las variables con mayores repeticiones seleccionadas de acuerdo a la tabla 19, se construyeron grupos de datos que se utilizaron como los dataset para probar los modelos de predicción objeto de este estudio que se trataran más adelante en este documento.

3.4.2 Generar el plan de prueba

El objetivo del algoritmo de aprendizaje o modelo es que capture la esencia del problema a resolver y lo generalice correctamente, para esto es necesario que funcione a futuro con datos de prueba o test en un ambiente de producción. Para lograr que el modelo sea más robusto el grupo de datos de entrenamiento es mayor que el grupo de datos de prueba y generalmente estos se seleccionan aleatoriamente. Se debe eliminar cualquier dependencia con los datos con lo que se entrenó, para evitar los dos problemas más comunes de estos modelos, el sobreajuste o overfitting y el subajuste o underfitting.

El sobreajuste es el problema de sobreentrenar un modelo y termine respondiendo solo a las propiedades de los datos con lo que fue entrenado y sea incapaz de generar niveles de acierto con otros juegos de datos. Este proceso solo aplica para algoritmos de aprendizaje supervisado, en donde se usa un conjunto de datos para entrenar el modelo y otro grupo de datos para medir para la precisión alcanzada por el modelo, y es necesario iniciar un proceso iterativo de entrenamiento y prueba del modelo con diferentes grupos de datos, hasta alcanzar

los niveles o porcentajes de precisión y capacidad de predicción aceptables (Brownlee, 1967).

A medida que el grupo original de datos es más grande, se debe reducir el grupo de datos de prueba, lo que genera un grupo de datos de entrenamiento más grande y a su vez genera un modelo más robusto. Para esta investigación, se partieron los datos originales en cuatro grupos para construir los modelos de predicción:

1. **Escenario 1:** Datos originales sin valores faltantes, se tomaron los datos entre el 2017 y el 2020 (4 años) para entrenamiento y los datos del 2021 para pruebas.
2. **Escenario 2:** Datos originales sin valores faltantes, se tomaron los datos entre el 2017 y el 2019 (tres años) para entrenamiento y los datos del 2020 y 2021 (dos años) para pruebas.
3. **Escenario 3:** Datos originales imputados, se tomaron los datos entre el 2017 y el 2020 (4 años) para entrenamiento y los datos del 2021 para pruebas.
4. **Escenario 4:** Datos originales imputados, se tomaron los datos entre el 2017 y el 2019 (tres años) para entrenamiento y los datos del 2020 y 2021 (dos años) para pruebas.

3.4.3 Construir el modelo

En esta fase, se utilizaron los dos conjuntos de datos obtenidos por los dos métodos de limpieza de datos, es decir el conjunto de datos sin datos faltantes y el conjunto de datos con datos imputados para la posterior aplicación de los modelos de predicción. Este proceso se realizó con la utilización de algoritmos supervisados de aprendizaje de máquinas en los que se emplearon Árboles de decisión, bosque aleatorio y regresión logística. Se estipuló una meta de porcentaje de precisión del 80% para los modelos predictivos; en el apartado anterior se definieron los escenarios de datos con los que se van a construir los modelos elegidos y posteriormente evaluar su confiabilidad en la predicción del rendimiento de los estudiantes en la prueba Saber 11°. De acuerdo con los objetivos definidos para el proceso de minado de datos, esta sección se va a desarrollar en dos partes, una por cada objetivo para diferenciar la construcción y la parametrización de los modelos.

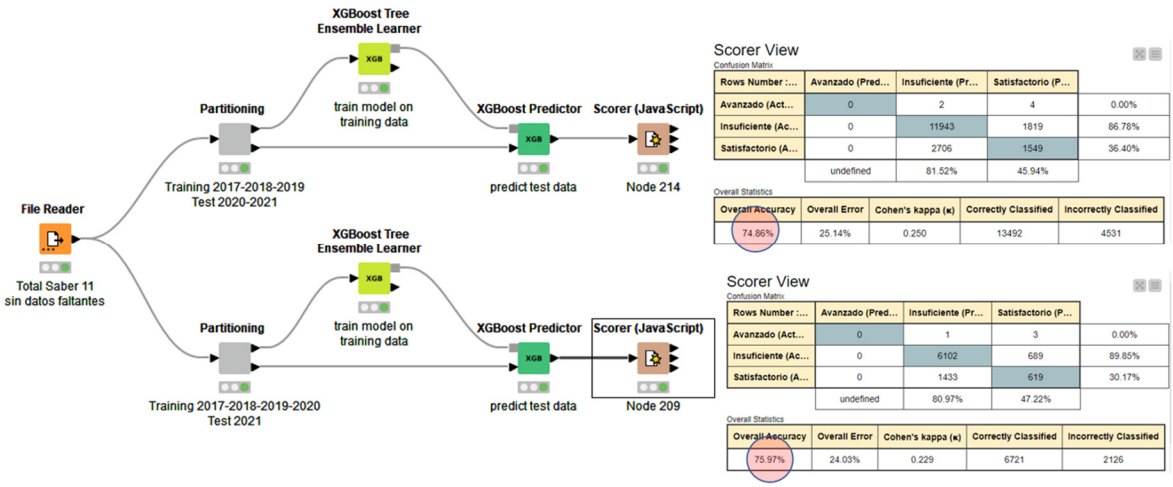
3.4.3.1 Escenario sin datos faltantes

Una vez que se tiene del conjunto de datos original, se eliminaron los datos faltantes generando un nuevo conjunto de datos con 47.147 registros y se procede a aplicar modelos de predicción aplicando las características que resultaron del proceso de reducción de dimensionalidad tratado en el numeral 3.4.1.

XGBoost es un método de aprendizaje automático supervisado para clasificación y regresión. XGBoost es la abreviatura de las palabras inglesas "extreme gradient

boosting" (refuerzo de gradientes extremo), este método se basa en arboles de decisión y está diseñado para conjuntos de datos grandes y complejos. Este algoritmo arrojó una precisión de 74.86% y 75.97% y un valor Kappa aceptable entre 0.229 y 0.250 para los dos conjuntos de datos de Prueba y Training que se definieron respectivamente. Ver figura 48.

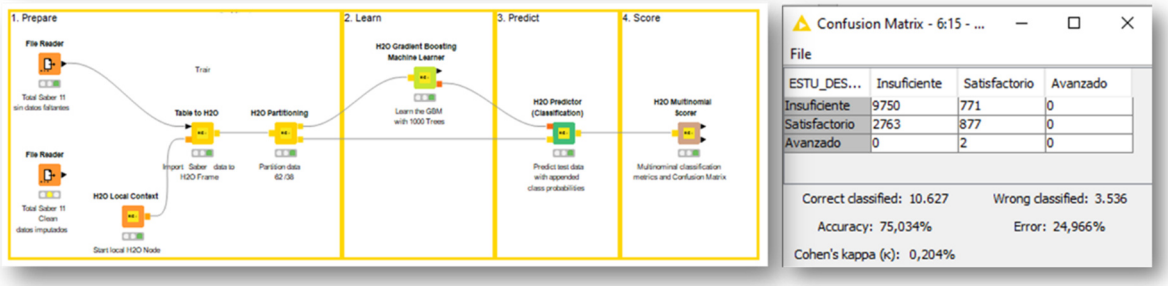
Figura 48 Algoritmo XGBoost sin datos faltantes



Fuente: Elaboración propia con Knime

Otro método de predicción utilizado es el algoritmo de aumento de gradiente (Gradient Boosting Machine GBM) que genera secuencialmente árboles de regresión, arrojando una precisión de 75.034% y factor leve Kappa de 0.204%. Ver figura 49.

Figura 49 Algoritmo Gradient Boosting Machine GBM sin datos faltantes

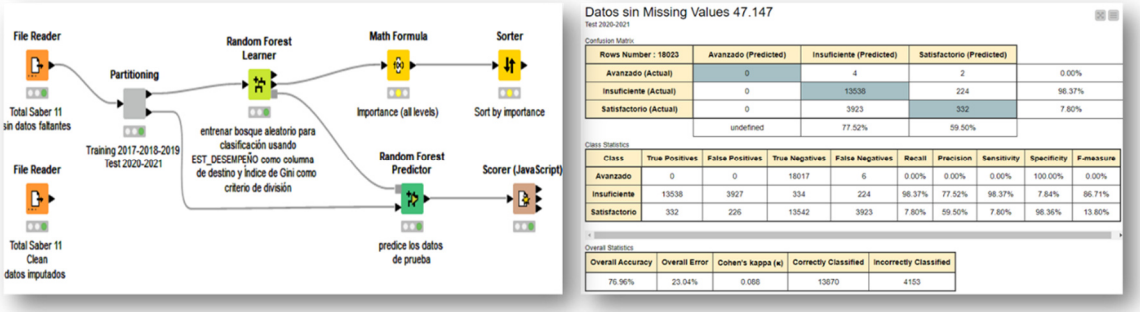


Fuente: Elaboración propia con Knime

[Anexo 11 Selección de variables con Gradrient Boosting Machine GBM]

Otra técnica de árboles de decisión es el algoritmo de predicción Random Forest o bosques aleatorios en español aplicándolo al set de datos arrojó una precisión de 76.96% y factor insignificante Kappa de 0.088%. Ver figura 50.

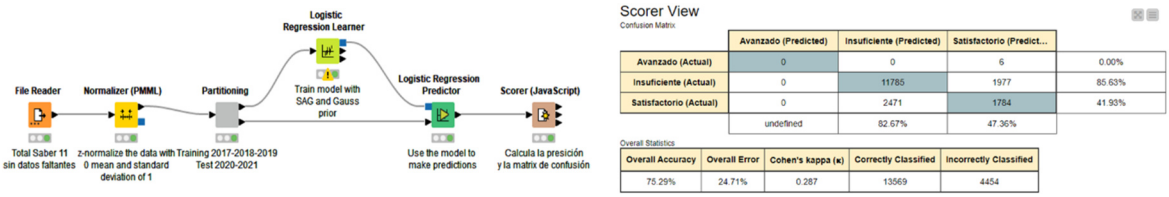
Figura 50 Algoritmo Random Forest sin datos faltantes



Fuente: Elaboración propia con Knime

La regresión logística es otro tipo de análisis de regresión para predecir o estimar el resultado de una variable dependiente en función de las variables independientes, generando una precisión de 75.29% y factor Kappa aceptable de 0.287%. Ver figura 51.

Figura 51 Algoritmo de regresión logística sin datos faltantes



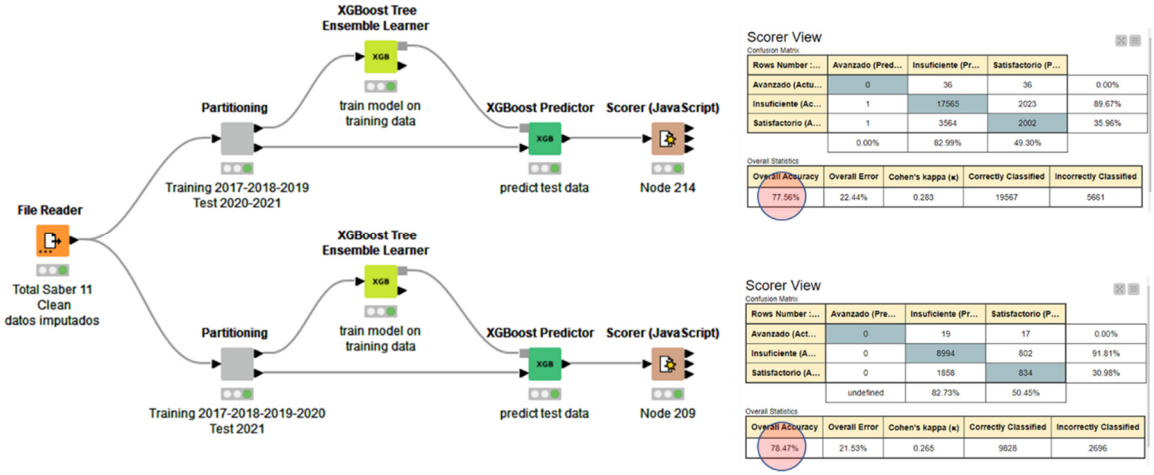
Fuente: Elaboración propia con Knime

3.4.3.2 Escenario con datos imputados

Por otro lado, con el conjunto de datos original en lugar de eliminar los datos faltantes estos se imputaron; es decir se sustituyeron por otros manteniendo el conjunto de datos original con 67.049 registros y se procede a aplicar los mismos modelos de predicción del numeral anterior.

El algoritmo XGBoost para datos imputados arrojó una precisión de 77.56% y 78.47% y un valor Kappa aceptable entre 0.265 y 0.283 para los dos conjuntos de datos de Prueba y Training que se definieron respectivamente. Ver figura 52.

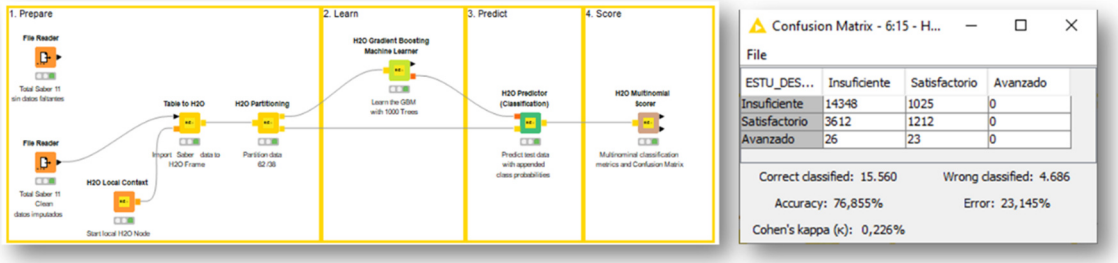
Figura 52 Algoritmo XGBoost con datos imputados



Fuente: Elaboración propia con Knime

Utilizando el algoritmo de aumento de gradiente (Gradient Boosting Machine GBM) arrojó una precisión de 76.885% y factor aceptable Kappa de 0.226%. Ver figura 53.

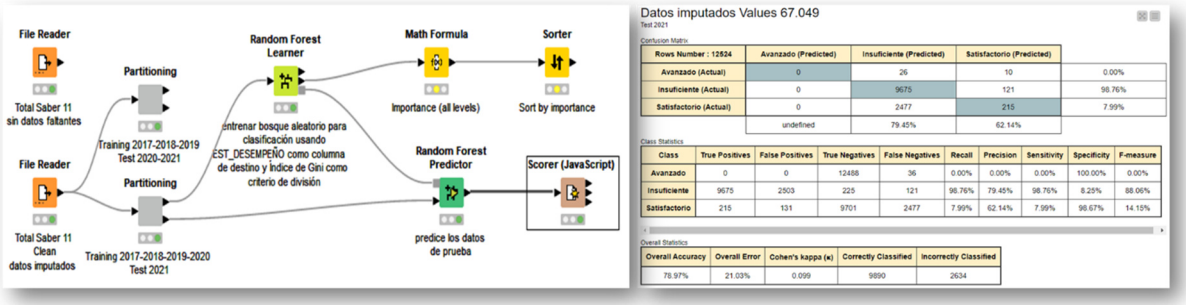
Figura 53 Algoritmo Gradient Boosting Machine GBM con datos imputados



Fuente: Elaboración propia con Knime

El algoritmo de predicción Random Forest con conjunto de datos imputados arrojó una precisión de 78.97% y factor leve Kappa de 0.099%. Ver figura 54.

Figura 54 Algoritmo Random Forest con datos imputados

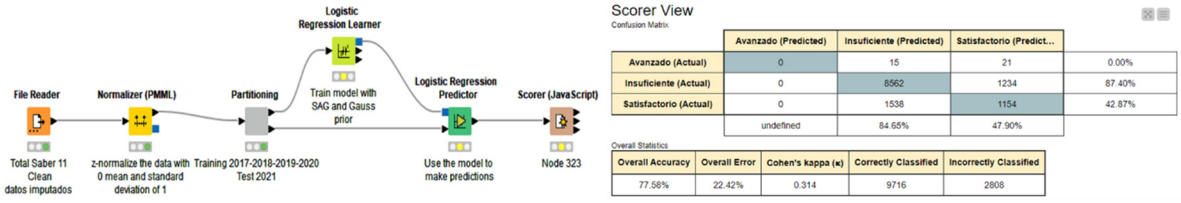


Fuente: Elaboración propia con Knime

Anexo 12 Selección de variables con Random Forest

La regresión logística con datos imputados generó una precisión de 77.58% y factor Kappa aceptable de 0.314%. Ver figura 55.

Figura 55 Algoritmo de regresión logística con datos imputados



Fuente: Elaboración propia con Knime

En la siguiente tabla se muestra la precisión de cada algoritmo aplicado en los dos conjuntos de datos elegidos (sin datos faltantes y con datos imputados). En donde se evidencia los mejores resultados con el Random Forest.

Tabla 20 Resultados modelos de predicción

Técnica de predicción	Sin Datos Faltantes	Técnica de predicción	Con datos imputados
GBM	75.03%	GBM	76.86%
Regresión logística	75.29%	Regresión logística	77.58%
XGBoost	75.97%	XGBoost	78.47%
Random Forest	76.96%	Random Forest	78.97%

Fuente: Elaboración propia

3.4.4 Evaluar el modelo

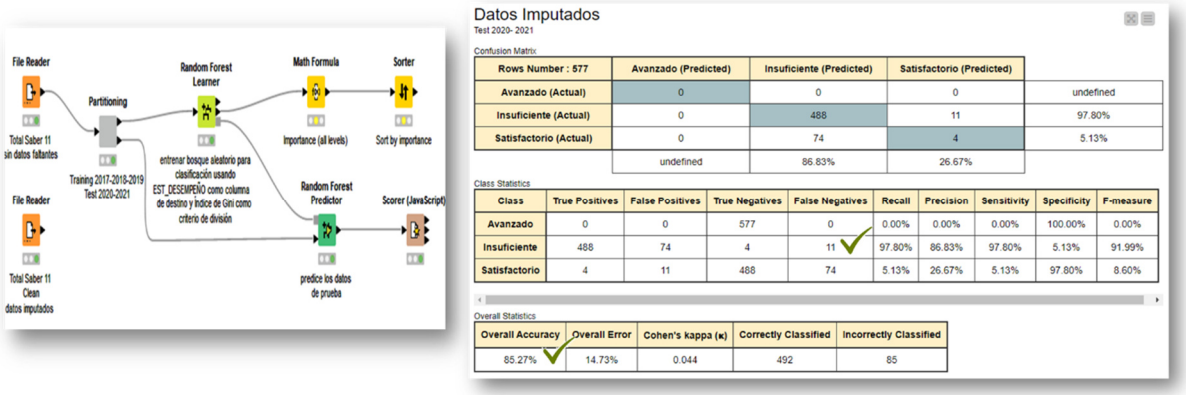
En esta fase, se utilizaron múltiples escenarios en los que el conjunto de datos podía corresponder a los registros de diferentes años (2017 a 2021) o únicamente uno de los años, esto con el fin de evaluar las fluctuaciones en el porcentaje de precisión de acuerdo con el conjunto de datos utilizados y detectar el mejor escenario para la aplicación del modelo predictivo. A partir de estos resultados, se identificaron y se discutieron las fortalezas y debilidades del modelo, tomando en cuenta cada una de las fases ya mencionadas y se realizó una comparación crítica de los escenarios evaluados de acuerdo con su desempeño y la relevancia del conjunto de reglas de decisión.

Puede verse que a pesar de que no hay grandes diferencias entre los resultados entre los dos conjuntos de datos, es decir sin datos faltantes e imputados; si se observa que las mejores predicciones ocurren en el conjunto de datos con datos imputados. De las cuatro técnicas aplicadas el algoritmo de Random Forest es el que mejores resultados ofrece, con una precisión general que ronda el 79% y el 79.45% específicamente en la categoría de desempeño estudiantil “Insuficiente”

con una Recall o sensibilidad de 98.76%, que es uno de los objetivos planteados al inicio de esta investigación.

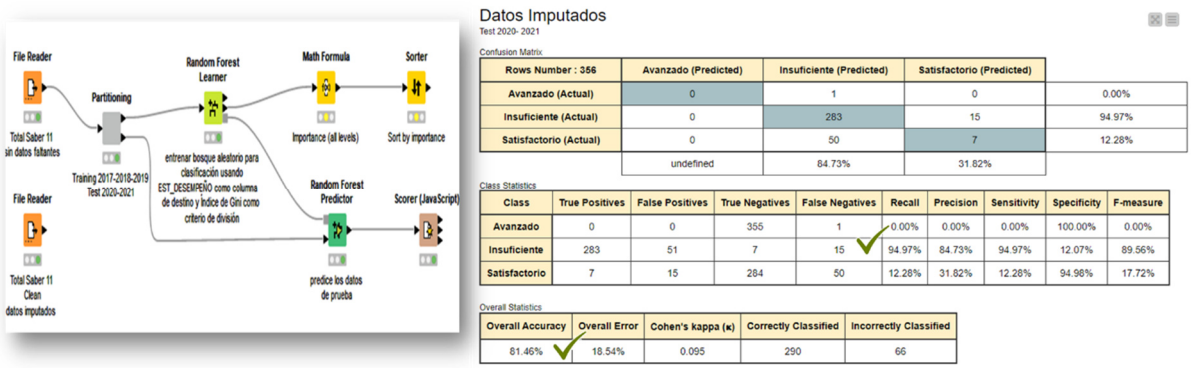
Al aplicar este modelo predictivo de Random Forest en las regiones y municipios del departamento de Cundinamarca con los más bajos desempeños en las pruebas Saber 11 entre el año 2017 y 2021, se probó con las regiones Magdalena Centro y Medina, arrojando una precisión de 85.27% y 81.46% respectivamente. Ver figuras 56 y 57.

Figura 56 Modelo de predicción Random Forest con la Región Magdalena Centro



Fuente: Random Forest Knime

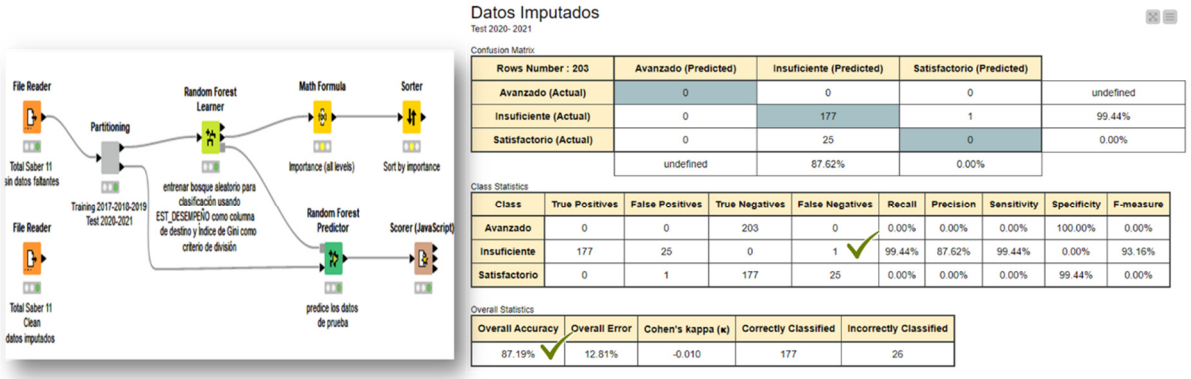
Figura 57 Modelo de predicción Random Forest con la Región Medina



Fuente: Random Forest Knime

Posteriormente se probó con el municipio San Juan de Rioseco de la región Magdalena Centro, arrojando una precisión de 87.19%, ver figura 58**.**

Figura 58 Modelo de predicción Random Forest con el municipio San Juan de Rioseco



Fuente: Random Forest Knime

3.5 Evaluación y despliegue (Evaluación del modelo)

Después de recorrer todo del proceso de minado desde la recolección de los datos, su exploración, la limpieza y el manejo de datos faltantes se logró reducir el conjunto de datos de 82 variables a 43 variables en la primera fase; en la segunda

fase se ejecuta el proceso de reducción de dimensionalidad logrando identificar las variables más representativas y haciendo una nueva de reducción de atributos pasando de 43 a 22 variables, que se listan en la tabla 20, es decir se logró un más de un 50% en la reducción de la dimensionalidad.

Tabla 21 Listado de los factores determinantes en las pruebas Saber 11

$contrib	Total Repeticiones
ESTU_INSE_INDIVIDUAL	5
FAMI_EDUCACIONPADRE	5
ESTU_MCPIO_RESIDE	5
FAMI_EDUCACIONMADRE	5
COLE_MCPIO_UBICACION	5
COLE_JORNADA	5
COLE_NOMBRE_SEDE	5
ESTU_MCPIO_PRESENTACION	5
ESTU_NSE_ESTABLECIMIENTO	5
ESTU_EDAD	4
COLE_NOMBRE_ESTABLECIMIENTO	4
ESTU_NSE_INDIVIDUAL	4
FAMI_NUMLIBROS	4
FAMI_TRABAJOLABORMADRE	4
FAMI_TRABAJOLABORPADRE	4
FAMI_COMELECHEDERIVADOS	4
FAMI_ESTRATOVIVIENDA	4
FAMI_TIENEINTERNET	4
FAMI_TIENECOMPUTADOR	4
ESTU_DEDICACIONLECTURADIARIA	3
ESTU_HORASSEMANATRABAJA	3
ESTU_TIPODOCUMENTO	3

Fuente: Elaboración propia

Se hace un perfilamiento de los dos grupos más representativos en los que se agruparon los estudiantes de acuerdo con su desempeño en las pruebas Saber

11°, evaluando algunas de las variables identificadas como determinantes en este estudio.

- Los Estudiantes con el nivel de desempeño “INSUFICIENTE”, que son el 75.81% del total del conjunto de datos; es el grupo donde el mayor porcentaje de evaluados provienen del área urbana (50.73%) y el 62.18% pertenecen al estrato 1 y 2, sus familias están compuestas por grupos pequeños de 3 a 4 integrantes, en donde más del 79% de esos hogares no cuenta con de computador ni de internet y solo el 58% tienen entre 0 y 25 libros. Este grupo es el que menos tiempo le dedica a la lectura (30 min o menos al día), donde más del 55% pertenecen a los niveles socioeconómicos NSE1 y NSE2 y en el cual el nivel educativo del padre más común es primaria incompleta y de la madre es secundaria completa. Este grupo está más representado por mujeres que por hombres (44.26% y 31.55% respectivamente).

- Los estudiantes con el nivel de desempeño “SATISFACTORIO”, que equivale al 23.92% del total de la muestra; en donde los estratos más comunes son 1 el 2 (19%). Más del 29% de los evaluados de este grupo cuentan con servicio de computador e internet y, en sus hogares, el 21% tiene entre 0 y 100 libros. El género de los estudiantes se distribuye

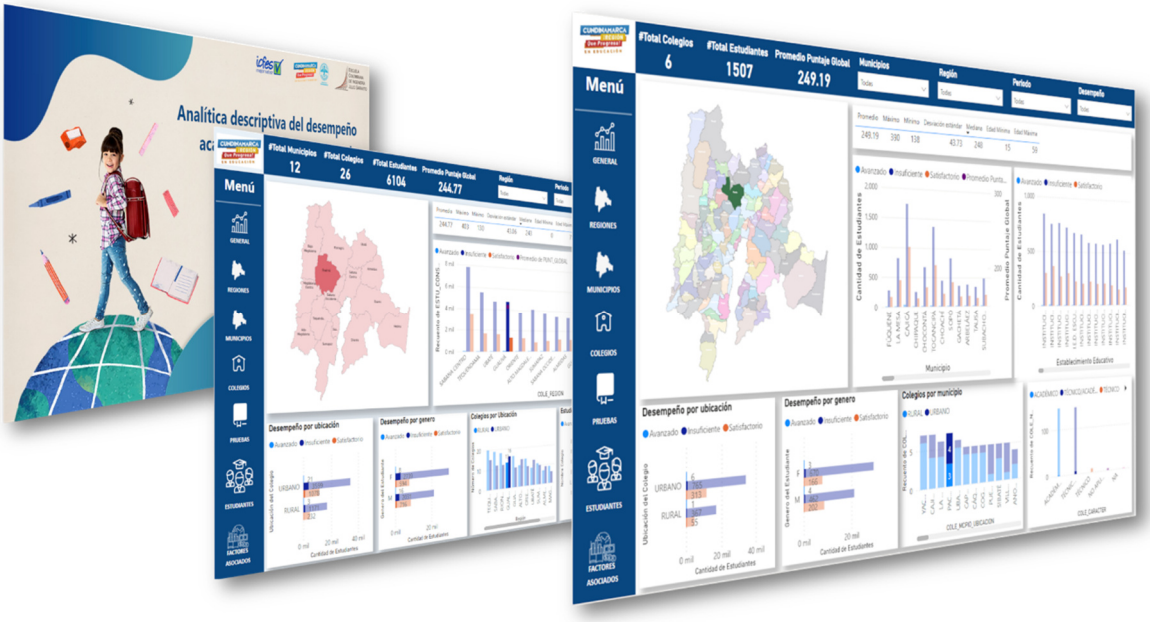
equitativamente entre mujer y hombre (11% y 12% respectivamente). El nivel educativo más común de la madre y el padre es secundaria completa.

- Para esta investigación no es determinante hacer experimentos sobre el conjunto de datos sin datos faltantes, o el conjunto de datos imputados al no existir mayores diferencias en los resultados de los modelos de predicción aplicados.

3.6 Fase de implementación

Después de organizar y analizar los datos, creo una herramienta en Power BI que permite visualizar y hacer un seguimiento a nivel departamental, regional y municipal de los resultados obtenidos durante los años de estudio, en los municipios no certificados y las instituciones educativas oficiales, donde se evidencia la relación de los resultados en las pruebas con las variables socioeconómicas de cada estudiante capturadas en la inscripción de la prueba Saber 11°. Esta herramienta permite hacer el análisis a nivel departamental, regional, por municipios, por entidad educativa, por materia o área evaluada, por estudiantes que presentaron la prueba y finalmente por los factores determinantes asociados a los resultados de la prueba, como se puede apreciar en la figura 59.

Figura 59 Interfaz de herramienta de visualización y análisis en PowerBI



Fuente: Elaboración propia

En la figura 60, se presenta el desempeño estudiantil a nivel institución educativa oficial por área de ubicación, el carácter o tipo de educación (académico o técnico), jornada de la institución y el listado de instituciones ordenadas por desempeño en la prueba. También se observa el desempeño por área evaluada (Ciencias, inglés, lectura, matemáticas y sociales) y el comportamiento de algunas de las variables socioeconómicas más representativas (tiene internet, tiene computador, estrato socioeconómico entre otras) todo evaluado a nivel regional, por municipio, periodo evaluado (años) y niveles de desempeño definidos en la variable de estudio.

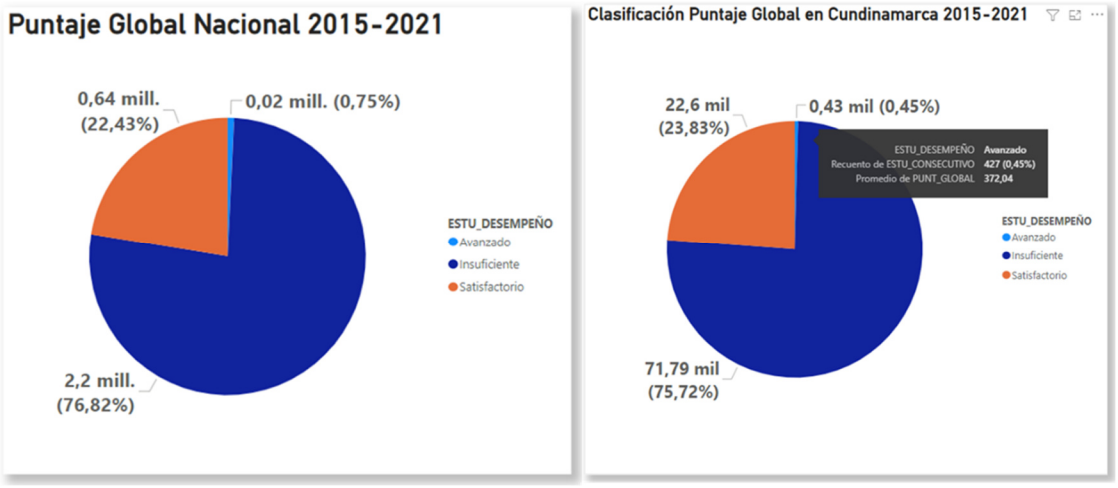
Figura 60 Interfaz por pruebas, factores asociados e instituciones educativas.

Fuente: Elaboración propia

4 Resultados y contribución

Esta investigación permitió identificar la misma tendencia de comportamiento en el desempeño de los estudiantes en las pruebas Saber 11°, a nivel nacional, departamental y municipal. A nivel nacional, durante los periodos en evaluación (2015-2021) más del 76% de la población estudiantil quedo clasificada en el nivel insuficiente, es decir con un puntaje global menor o igual a 279 puntos y más del 22% se encuentra clasificado en un nivel satisfactorio (puntajes entre 280 y 359 puntos) y solo el 0.75% se ubica en el nivel avanzado con un puntaje igual o mayor a 360 puntos. A nivel departamental, Cundinamarca presenta una distribución porcentual de 75.72%, 23.83% y 0.45% (Insuficiente, Satisfactorio y Avanzado respectivamente) para los mismos periodos de evaluación.

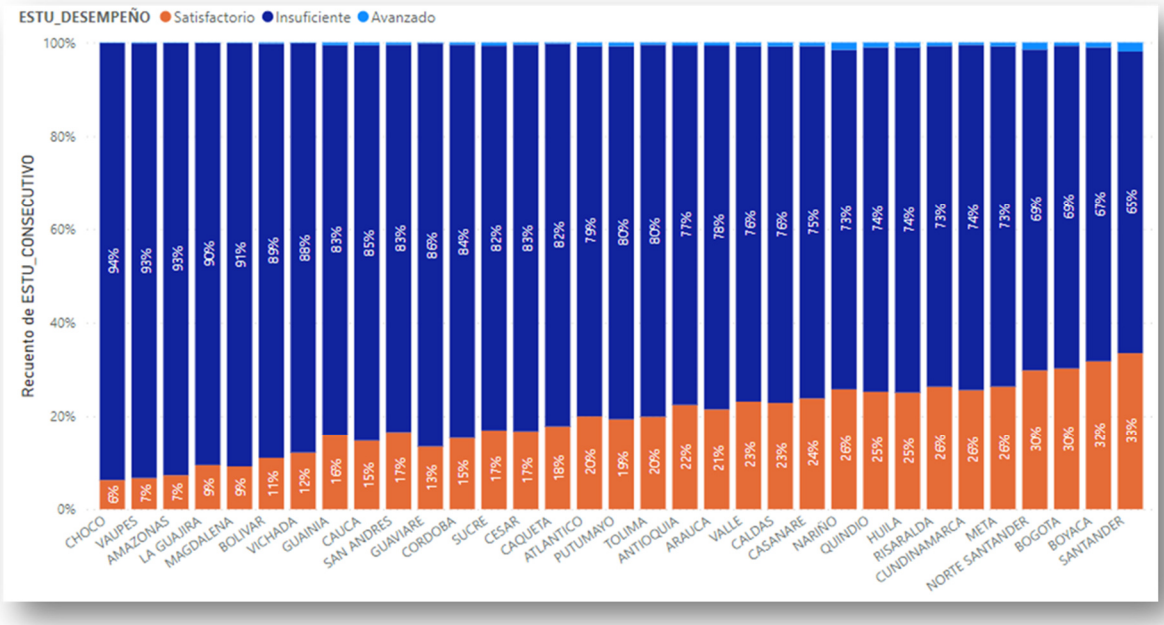
Figura 61 Desempeño en la prueba Saber 11 Nacional Vs Cundinamarca



Fuente: Elaboración propia PowerBI

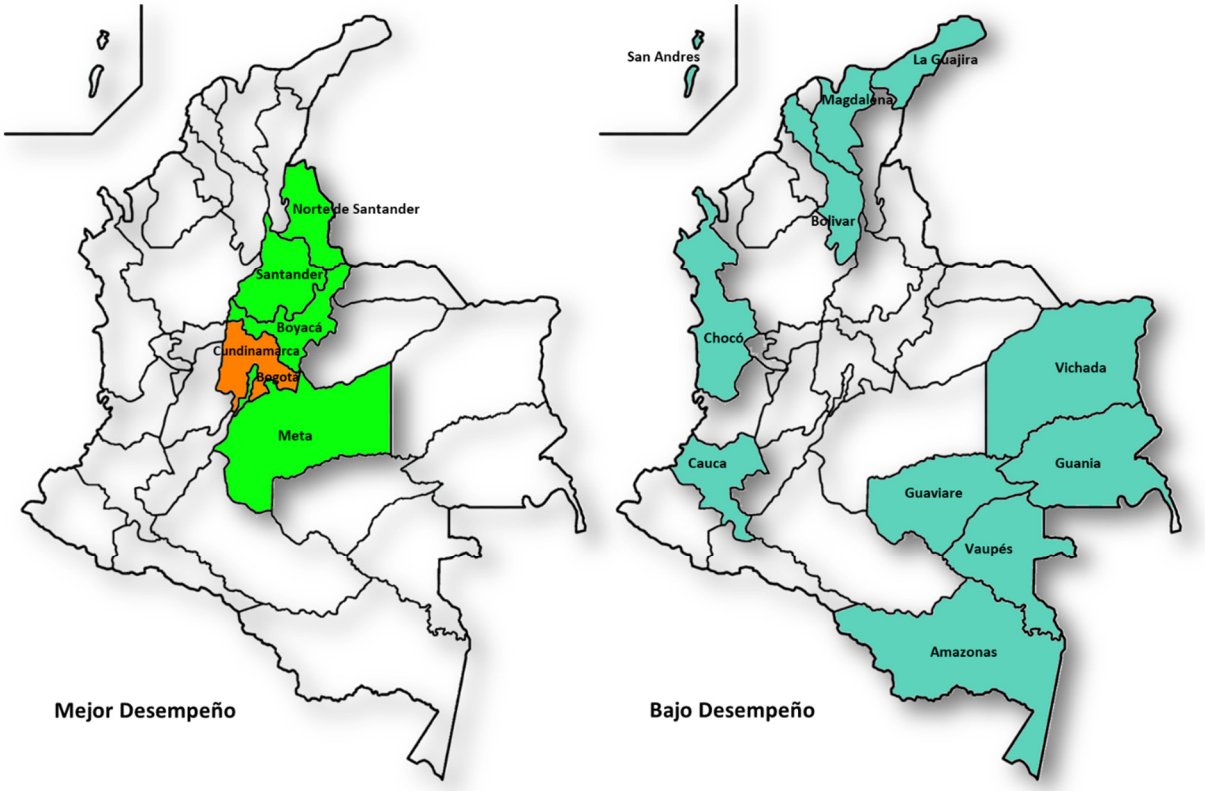
Bogotá y los departamentos de Cundinamarca, Meta, Boyacá, Santander y Norte de Santander son los que obtienen los mejores niveles de desempeño en las pruebas Saber 11, es decir tienen 74% o menos de estudiantes clasificados en la categoría “Insuficiente”. En contraste, los departamentos de Chocó, Vaupés, Amazonas, La Guajira, Magdalena, Bolívar tiene un porcentaje del 89% o más de estudiantes clasificados en la misma categoría “Insuficiente” en el mismo periodo de tiempo analizado, en la figura 62 y 63 se evidencian los departamentos con los desempeños más altos y bajos en las pruebas.

Figura 62 Departamentos clasificados de acuerdo con el desempeño en la prueba Saber 11°.



Fuente: Elaboración propia PowerBI

Figura 63 Niveles de desempeño más altos y bajos por departamento



Fuente: Elaboración propia

Entre las características seleccionadas, se analizaron el comportamiento de variables relacionadas con la vivienda del estudiante; como el estrato socioeconómico, si se cuenta con el servicio de computador y conexión a internet, el número de libros que se tiene en la vivienda y el tiempo diario de dedicación a la lectura, al igual que el tiempo diario de dedicación a navegar en internet. Este análisis arroja información sobre qué tan determinantes son estas variables en los resultados de las pruebas Saber 11° en Cundinamarca para el periodo de tiempo

en estudio, comparando cada variable en sus valores opuestos y analizando las variaciones de la variable dependiente (ESTU_DESEMPEÑO).

De acuerdo con el análisis realizado, uno de los atributos más relacionado con los resultados en la prueba Saber 11 es el nivel educativo de los padres de los estudiantes que presentaron la prueba. Las cifras revelan que, de los estudiantes con padres con un nivel académico bajo (primaria incompleta o completa), solo el 15% obtuvieron resultados satisfactorios o superiores con un promedio en el puntaje global de 239 puntos, contrario a los resultados de los estudiantes con padres con un nivel educativo alto, donde el 58% obtuvieron resultados satisfactorios o superiores con un promedio en el puntaje global de 281 puntos. De acuerdo con estos resultados, los estudiantes con padres con un bajo nivel educativo presentan desempeños más bajos en la prueba Saber 11°.

Otro aspecto importante es la cantidad de libros del estudiante y las horas diarias de dedicación a la lectura, en promedio el 38% los alumnos que tiene más de 100 libros en su casa tienen resultados “satisfactorios” o “avanzados” en la prueba, opuesto a los estudiantes con 10 libros o menos, donde solo el 16% obtuvieron resultados “satisfactorios” o “avanzados”. Igualmente, el 40% de los estudiantes que le dedican más de 2 horas diarias a la lectura obtuvieron resultados “satisfactorios” o “avanzados” en la prueba con un promedio en el puntaje global de 267 puntos, por el contrario, solo el 20% de los estudiantes que le dedican 30

minutos o menos a la lectura diaria obtuvieron puntajes “satisfactorios” o “avanzados” con un promedio en el puntaje global de 245 puntos.

Adicional a los análisis anteriores, al evaluar si se cuenta con el servicio de computador, una conexión a internet y la cantidad de tiempo que se dedica a internet. Los resultados demuestran solo una diferencia alrededor de 10 puntos porcentuales en los resultados “insuficiente” o “satisfactorio” en las pruebas de los estudiantes con acceso a computador, internet y dedicación a internet y aquellos que carecen de estos beneficios. Esto evidencia muy poca relación entre estas características y los resultados obtenidos.

5 Conclusiones y recomendaciones

Las técnicas y modelos de minería de datos aplicados en esta investigación ayudan a entender el panorama nacional, departamental y municipal de la calidad de la educación en Colombia; específicamente en las provincias o regiones y municipios del departamento de Cundinamarca en los últimos 5 años. Este estudio reafirma y evidencia el descenso continuo de la calidad de la educación en los municipios no certificados del departamento, situación confirmada por los resultados de las pruebas Saber 11° entre los años 2017-2021 como indicador de calidad.

Al inicio de esta investigación se planteaba la hipótesis que la situación podía empeorar a causa de la pandemia del Covid-19; con la evaluación de los resultados se concluye que incluso antes de la pandemia ya se presentaba este descenso en el desempeño estudiantil de los estudiantes del departamento y la pandemia ha profundizado la crisis. Con la propagación del virus y la suspensión de clases presenciales, aumento el riesgo del abandono escolar, especialmente de los estudiantes en hogares más vulnerables, que disponían de menos recursos (conectividad, dispositivos electrónicos, etc.) en donde los estudiantes pasaron tres cuartas partes de los 200 días del año académico en casa (marzo 2020-junio 2021).

Con la intención de determinar las causas de los resultados en las pruebas Saber 11° en Cundinamarca, en donde se compararon factores familiares, individuales y del colegio del estudiante que se identificaron como de gran incidencia en el puntaje obtenido por cada estudiante. Solo el 23% en promedio de los estudiantes analizados lograron superar o compensar las condiciones asociadas durante los 5 años de análisis que permitió obtener resultados “satisfactorios” o “avanzados” de acuerdo con la variable objetivo-definida, contrario al 76% de los estudiantes que no lograron los resultados esperados en la prueba.

Esta investigación evidencia las diferencias en la calidad de la educación y los efectos que generan en quienes reciben educación de menor calidad en su núcleo familiar. Aunque no es el objeto de esta investigación, los resultados de los estudiantes en las pruebas Saber es mejor en colegios privados que en los públicos. Es evidente la brecha de la calidad entre la educación pública y la privada; diferencias en temas como la infraestructura educativa son determinantes a la hora de presentar una prueba como las del ICFES.

En principio los resultados de esta investigación evidencian, mediante el análisis estadístico y modelos de minería de datos, que los estudiantes de hogares en situación de pobreza o estratos bajos tienen una tendencia a continuar manteniendo la misma situación socioeconómica en la que viven; como parte de un fenómeno asociado al bajo nivel educativo. Esta situación está determinada de

algún modo según los resultados de las pruebas, por el estrato socioeconómico, el nivel educativo, e ingresos de los padres o del hogar del estudiante.

Este análisis empírico de la relación entre un bajo nivel socioeconómico y el bajo nivel académico, en donde la percepción del círculo vicioso entre la pobreza y el bajo nivel educativo es la misma pobreza, generando ciclos repetidos de pobreza - baja educación – pobreza, y estos se explican en gran parte por la ausencia o mínimo nivel educativo que se repite en padres e hijos de generación en generación. En estos hogares, los niños deben trabajar para cubrir sus necesidades, en donde la exclusión, la privación de recursos monetarios y las desigualdades los obligan a interrumpir sus estudios para sobrevivir, y trabajar resulta más rentable, mientras que la educación es un costo económico adicional.

Los resultados en las pruebas confirman que existe una fuerte correlación entre el nivel académico de los padres; en donde a mayor nivel educativo de los padres mejor el desempeño de los hijos en las pruebas Saber 11°.

Después de evaluar los resultados de las pruebas para los 5 años de este estudio, otro hallazgo interesante en esta investigación es que, a pesar de tener el acceso a un servicio de internet y el tiempo de dedicación a navegar en él, acompañado de la tenencia de un computador no representa mayores ventajas que determinen mejores resultados en las pruebas Saber 11°. Se esperaría que estos tiempos de transformación digital, que es la aplicación de tecnologías digitales en los

diferentes entornos de la sociedad humana incluida la educación, brindaran herramientas a los estudiantes que les permitiera mejorar su desempeño en la prueba. Si bien es cierto que es importante fomentar y fortalecer el acceso a tecnologías que permitan a los estudiantes acceder a información académica, la tarea y reflexión para un trabajo futuro es identificar si este comportamiento está asociado a un problema común de la mayoría de los padres hoy en día, que es el buen uso y aprovechamiento de estas nuevas tecnologías por parte de sus hijos.

La lectura es otra variable determinante hallada en esta investigación, entendiendo la lectura como instrumento indispensable para el acceso al conocimiento: si un alumno no comprende lo que lee, pierde motivación en los estudios. La observación y la reflexión permite concluir que un estudiante con nivel de compresión lectora bajo o deficiente tiene mayor probabilidad de obtener malos resultados en las pruebas, este efecto se puede apreciar en los resultados de la prueba de “LECTURA_CRITICA” en donde el promedio del puntaje en la prueba fue de 74, 62 y 49 puntos para los desempeños “Avanzado”, “Satisfactorio” e “Insuficiente” respectivamente.

El DANE señala que el índice de lectura en Colombia es de 2.7 libros al año, comparado con Chile donde este índice es de 5.4; de ahí la importancia y la responsabilidad de fomentar la lectura y continuar desarrollando plataformas como “Cundinamarca lee” y lograr alcanzar un punto de equilibrio que permita que los

estudiantes de estratos bajos accedan a los libros y estos dejen de ser un lujo y, en muchas ocasiones, compitan con cubrir las necesidades básicas del hogar.

Al comparar las provincias o regiones de Cundinamarca con los más altos y más bajos desempeños de los estudiantes en las pruebas Saber 11°, se identifican grandes diferencias en varias categorías definidas por la gobernación de Cundinamarca dentro del plan de desarrollo departamental 2020-2024.

En la provincia Magdalena Centro, que tuvo los menores resultados en las pruebas Saber 11° de acuerdo con los resultados de este estudio, con más del 72% de la población ubicada en zona rural, tiene un índice de competitividad (en una escala de 0 a 10 puntos porcentuales) del 2.97%, contrario al 8.16% que obtuvo la provincia con mejor desempeño en la prueba. Este índice contiene temas como instituciones, infraestructura, educación básica, media y superior y salud entre otros.

Esto refleja que Magdalena Centro es una de las provincias con menos competitividad a 5.19 puntos de Sabana Centro, que es la provincia con mejor nivel de competitividad y con los mejores resultados en las pruebas Saber 11°. Por otro lado, al analizar los servicios públicos en promedio esta provincia solo tiene un 62% y 37% de cobertura en acueducto y alcantarillado respectivamente vs un 97.61% y 82.50% de Sabana Centro. En temas de infraestructura, la penetración en banda ancha corresponde solamente al 2%, contrario a la de Sabana Centro

que es del 13.57%. Esto evidencia que la provincia de Magdalena Centro aún enfrenta grandes retos de cobertura y desempeño en las categorías antes mencionadas, los cuales están generando un efecto en los resultados de sus estudiantes en las pruebas Saber 11°.

Para concluir, es evidente que existen diferencias en la calidad de la educación básica y media, quienes reciben la educación de menor calidad se enfrentan a actuar en una sociedad muy competitiva con menos herramientas y más restricciones; en donde continuar sus estudios en la educación superior es menos probable y el desempeño junto a sus ingresos en el mercado laboral será menor.

Según lo desarrollado a lo largo del estudio, de las conclusiones anteriores se proponen las siguientes líneas de trabajo, con el objetivo de desarrollar el capital humano como eje central de cualquier estrategia de solución:

- Continuar el estudio que contribuya al esclarecimiento del proceso social que asocia el bajo nivel educativo y la pobreza.
- Continuar el estudio para determinar y validar los resultados de las pruebas a nivel regional o provincial, haciendo un análisis comparativo evaluando las categorías de población con condiciones especiales, competitividad, actividad económica, servicios públicos, internet, educación, salud y seguridad.

- Probar y validar los métodos y modelos predictivos planteados en esta investigación con las pruebas Saber 11° del 2022, incluyendo el Índice Sintético de Calidad Educativa (ISCE) como otro instrumento para medir la calidad educativa en el departamento, que incluye cuatro componentes progreso, desempeño, eficiencia y ambiente escolar.

Referencias bibliográficas

Aguilar Márquez, A., Altamira Ibarra, J., & García León, O. (2010). *Introducción a la Inferencia Estadística* (1ra ed.). Pearson Educación.

Bozkurt, E. (2021). Machine Learning Classification Algorithms with Codes. In *Analytics Vidhya*. https://blog.quantinsti.com/machine-learning-classification/?utm_term=&utm_campaign=&utm_source=adwords&utm_medium=ppc&hsa_acc=2144065910&hsa_cam=16770050358&hsa_grp=&hsa_ad=&hsa_src=x&hsa_tgt=&hsa_kw=&hsa_mt=&hsa_net=adwords&hsa_ver=3&gclid=CjwKCAjwu_mSB

Brownlee, J. (1967). Master Machine Learning Algorithms. *Angewandte Chemie International Edition, 6(11), 951–952.* https://machinelearningmastery.com/master-machine-learning-algorithms/

Brownlee J. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Books.Google.Co.In. https://books.google.com.co/books/about/Data_Preparation_for_Machine_Learning.html?id=uAPuDwAAQBAJ&redir_esc=y

Calderón García, A., & Piñeros Rivera, M. A. (2020). *Guía de orientación Saber 11° 2020-1*. ICFES. https://www2.icfes.gov.co/documents/39286/2171114/Gu%C3%ADa+de+orientaci%C3%B3n+Saber+11.%C2%B0+2020-1.pdf

Charte, D. (2017). *Reducción de la dimensionalidad en problemas de clasificación con Deep Learning: Análisis y propuesta de herramienta en R*. https://www.researchgate.net/publication/318888351_Reduccion_de_la_dimensionalidad_en_problemas_de_clasificacion_con_Deep_Learning_Analisis_y_propuesta_de_herramienta_en_R

Cheng, Jiechao. (2017). *Data-Mining Research in Education.*

Data Science Process Alliance. (2020). *CRISP-DM - Data Science Process Alliance*. Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Dataversity. (2016). *Survey Shows Data Scientists Spend Most of Their Time Cleaning Data*. Data Topics. https://www.dataversity.net/survey-shows-data-scientists-spend-time-cleaning-data/#

DEPARTAMENTO DE CUNDINAMARCA, SECRETARÍA DE EDUCACIÓN, & DIRECCIÓN CALIDAD EDUCATIVA. (2021). *ANÁLISIS PRUEBAS SABER GRADO 11 VIGENCIA 2020*. https://www.cundinamarca.gov.co/wcm/connect/dce39daf-ea41-4938-a7d9-debf9175518e/AN%C3%81LISIS+SABER+11-+2020+%281%29.pdf?MOD=AJPERES&CVID=nD5GGso

Escobedo, M. T., & Salas, J. A. (2008). P. ch. mahalanobis y las aplicaciones de su distancia estadistica. *CULCyT*, *5*(January 2008), 13–20. https://www.researchgate.net/publication/28249208_P_Ch_Mahalanobis_y_las_aplicaciones_de_su_distancia_estadistica

Geeks for Geeks. (2020). *Measures of Distance in Data Mining*. https://www.geeksforgeeks.org/measures-of-distance-in-data-mining/?ref=gcse

Gironés, J., Casas, J., Minguillón, J., & Caihuelas, R. (2017). Minería de datos: Modelos y Algoritmos. *Editorial UOC*, *1*, 1–271. https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part

Gobernación de Cundinamarca. (2020). *CARACTERIZACIÓN Y PERFIL DEL SECTOR EDUCATIVO 2020*.

Grupo Ceinfes. (2021). *¿LOS ESTUDIANTES SE SIENTEN PREPARADOS PARA LAS PRUEBAS SABER? – Ceinfes*. CEINFES. https://ceinfes.com/los-estudiantes-no-se-sienten-preparados-para-las-pruebas-saber/

Molenberghs, G., Garret, F., Kenward, M. G., Tsiatis, A., & Verbeke, G. (2015). *Handbook of Missing Data Methodology*. https://es.scribd.com/document/398942452/Handbook-of-Missing-Data-Methodology

Icfes, Instituto Colombiano para la Evaluación de la Educación (2019). *SABER AL DETALLE- EDICIÓN 4*.

ICFES. (2021). *Informe Nacional de resultados del examen Saber 11 2020*. *I*, 13–14.

ICFES. (2021). *Resultados agregados examen Saber 11° - 2021*.

ICFES. (2022). *Pruebas Saber 11 Guía de orientación 2022-1*.

Kassambara, A. (2017). *Practical guide to principal component methods in R*. 155.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Markus, S. (2019). Programme For International Studen Assesment (PISA) Results from PISA 2018. *OECD*.

Martin, Z. H. (2012). Método de análisis de datos: apuntes. *Journal of Visual Languages & Computing*, *11*(3), 176.

Mazzanti S. (2020). *Boruta Explained Exactly How You Wished Someone Explained to You*. Towards Data Science. https://towardsdatascience.com/boruta-explained-the-way-i-wish-someone-explained-it-to-me-4489d70e154a

Microsoft. (2019). *Conceptos de minería de datos | Microsoft Docs*. Microsoft. https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions

Ministerio de Educación Nacional. (2022). *Icfes presentó a la comunidad educativa el Informe de los Resultados agregado Saber 11 en 2021*. Ministerio de Educación Nacional. https://www.mineducacion.gov.co/portal/salaprensa/Noticias/409545:Icfes-presento-a-la-comunidad-educativa-el-Informe-de-los-Resultados-agregado-Saber-11-en-2021

Ministerio de Tecnologías de la Información y Comunicaciones. (2019). *Guía para el uso y aprovechamiento de Datos Abiertos en Colombia*. 42. https://herramientas.datos.gov.co/sites/default/files/Guia de Datos Abiertos de Colombia.pdf

Montes, M. A. M., & Galvis, L. A. R. (2017). Desarrollo del sistema educativo en municipios certificados y no certificados en Colombia en 2005, 2008 y 20111. *Https://Doi.Org/10.18175/Vys8.1.2017.07*, *8*(1), 103–128. https://doi.org/10.18175/VYS8.1.2017.07

Open Data Charter. (2015). *Principios - International Open Data Charter*. Carta Internacional de Datos Abiertos. https://opendatacharter.net/principles-es/

Rey Ángel, J. E., & Gobernador de Cundinamarca. (2016). *PLAN DE DESARROLLO CUNDINAMARCA 2016-2020*. https://www.cundinamarca.gov.co/wcm/connect/2a9dd7d1-d693-414a-94cd-37fe5f901e7d/PLAN+DE+DESARROLLO+VERSION+FINAL.pdf?MOD=AJPERES&CVID=lDlW39U

Ruiz, J. (2021). *El reto de la formación integral: una mirada a las Pruebas Saber 11 2020 | Contexto*. Contexto. https://contextomedia.com/el-reto-de-la-formacion-integral-una-mirada-a-las-pruebas-saber-11-2020/

Salazar, C., & del Castillo Santiago G. (2018). *FUNDAMENTOS BÁSICOS DE ESTADÍSTICA*.

Secretaría de Educación. (2021). *ANÁLISIS PRUEBAS SABER GRADO 11 VIGENCIA 2020*. https://www.cundinamarca.gov.co/wcm/connect/dce39daf-ea41-4938-a7d9-debf9175518e/ANÁLISIS+SABER+11-+2020+%281%29.pdf?MOD=AJPERES&CVID=nD5GGso

Soles Ramos, I., & Torrent Sellens, J. (2010). *Técnicas de análisis de datos para la empresa*.

Anexos

Anexo 1. Municipios no certificados de Cundinamarca

Provincias	Municipio
Almeidas	Chocontá
	Machetá
	Manta
	Sesquilé
	Suesca
	Tibirita
	Villapinzón
Alto Magdalena	Agua de Dios
	Guataquí
	Jerusalén
	Nariño
	Nilo
	Ricaurte
	Tocaima
Bajo Magdalena	Caparrapí
	Guaduas
	Puerto Salgar
Gualivá	Albán
	La Peña
	La Vega
	Nimaima
	Nocaima
	Quebradanegra
	San Francisco
	Sasaima
	Supatá
	Útica
	Vergara
	Villeta
Guavio	Gachalá
	Gachetá
	Gama
	Guasca
	Guatavita

Provincias	Municipio
Rionegro	El Peñón
	La Palma
	Pacho
	Paime
	San Cayetano
	Topaipí
	Villagómez
	Yacopí
Sabana Centro	Cajicá
	Cogua
	Cota
	Gachancipá
	Nemocón
	Sopó
	Tabio
	Tenjo
	Tocancipá
Sabana Occidente	Bojacá
	El Rosal
	Madrid
	Subachoque
	Zipacón
Soacha	Sibaté
Sumapaz	Arbeláez
	Cabrera
	Granada
	Pandi
	Pasca
	San Bernardo
	Silvania
	Tibacuy
	Venecia
Tequendama	Anapoima
	Anolaima

Magdalena Centro	Junín
	La Calera
	Ubalá
	Beltrán
	Bituima
	Chaguaní
	Guayabal de Síquima
	Pulí
	San Juan de Rioseco
	Vianí
Oriente	Cáqueza
	Chipaque
	Choachí
	Fómeque
	Fosca
	Guayabetal
	Gutiérrez
	Quetame
	Ubaque
	Une

Ubaté	Apulo
	Cachipay
	El Colegio
	La Mesa
	Quipile
	San Antonio del Tequendama
	Tena
	Viotá
Medina	Carmen de Carupa
	Cucunubá
	Fúquene
	Guachetá
	Lenguazaque
	Simijaca
	Susa
	Sutatausa
	Tausa
	Villa de San Diego de Ubaté
	Medina
	Paratebueno

Anexo 2. Variables del conjunto de datos

GRUPO	ATRIBUTO	Tipo	Descripción
Información del Estudiante	ESTU_CONSECUTIVO	Nominal	Id público del inscrito (SB11)
	ESTU_GENERO	Nominal	Género
	ESTU_DEPTO_RESIDE	Nominal	Departamento residencia
	ESTU_MCPIO_RESIDE	Nominal	Municipio de residencia
	ESTU_MCPIO_PRESENTACION	Nominal	Municipio de presentación
	ESTU_DEPTO_PRESENTACION	Nominal	Departamento de presentación
	ESTU_INSE_INDIVIDUAL	Numérico	Indice nivel socioeconomico
	ESTU_NSE_INDIVIDUAL	Nominal	Nivel socioeconómico estudiante
	ESTU_ESTADOINVESTIGACION	Nominal	Estado de los resultados para los evaluados
	ESTU_AÑO_NACIMIENTO	Numérico	Año de nacimiento del estudiante
	ESTU_TIPODOCUMENTO	Nominal	Tipo de documento
	ESTU_NACIONALIDAD	Nominal	Nacionalidad
	ESTU_FECHANACIMIENTO	Fecha	Fecha de nacimiento
	ESTU_PAIS_RESIDE	Nominal	País de residencia
	ESTU_TIENEETNIA	Booleano	Pertenece a un grupo étnico minoritario
	ESTU_COD_RESIDE_DEPTO	Nominal	Código Dane del departamento de residencia
	ESTU_COD_RESIDE_MCPIO	Nominal	Código Dane del municipio de residencia
	ESTU_DEDICACIONLECTURADIARIA	Nominal	Horas dedicadas a la lectura por semana
	ESTU_DEDICACIONINTERNET	Nominal	Horas de uso de internet por día
	ESTU_HORASSEMANATRABAJA	Nominal	Horas de trabajo por semana
	ESTU_TIPOREMUNERACION	Nominal	Tipo de remuneración por las horas trabajadas
	ESTU_PRIVADO_LIBERTAD	Booleano	Condición judicial del evaluado
	ESTU_COD_MCPIO_PRESENTACION	Nominal	Código Dane del municipio presentación del examen

	ESTU_COD_DEPTO_PRESENTACION	Nominal	Código Dane del departamento de presentación del examen
	ESTU_NSE_ESTABLECIMIENTO	Numérico	Nivel socioeconómico del establecimiento
	ESTU_DESEMPEÑO	Nominal	Desempeño en la prueba del estudiante
	ESTU_GENERACION-E	Nominal	Cumple con los requisitos necesarios para ser beneficiario del programa Generación E
Información de la Entidad Educativa	COLE_NOMBRE_ESTABLECIMIENTO	Nominal	Nombre del establecimiento
	COLE_BILINGUE	Booleano	Educación bilingüe
	COLE_CARACTER	Nominal	Carácter
	COLE_NOMBRE_SEDE	Nominal	Nombre de la sede
	COLE_SEDE_PRINCIPAL	Booleano	Indica si la sede es la principal
	COLE_AREA_UBICACION	Nominal	Área de ubicación del colegio
	COLE_JORNADA	Nominal	Jornada
	COLE_MCPIO_UBICACION	Nominal	Municipio de ubicación
	COLE_CODIGO_ICFES	Nominal	Código Icfes del colegio
	COLE_COD_DANE_ESTABLECIMIENTO	Nominal	Código Dane del colegio
	COLE_GENERO	Nominal	Género de la población de estudiantes del colegio
	COLE_CALENDARIO	Nominal	Calendario académico
	COLE_COD_DANE_SEDE	Nominal	Código de la sede
	COLE_COD_MCPIO_UBICACION	Nominal	Código de ubicación del colegio en el municipio
	COLE_COD_DEPTO_UBICACION	Nominal	Código de ubicación del colegio en el departamento
	COLE_DEPTO_UBICACION	Nominal	Departamento de ubicación
Información Familiar	FAMI_EDUCACIONPADRE	Nominal	Nivel educativo del padre
	FAMI_EDUCACIONMADRE	Nominal	Nivel educativo de la madre
	FAMI_ESTRATOVIVIENDA	Nominal	Estrato socioeconómico de la vivienda
	FAMI_PERSONASHOGAR	Nominal	Número de personas en el hogar
	FAMI_CUARTOSHOGAR	Nominal	Número de cuartos (habitaciones)

	FAMI_TIENEINTERNET	Booleano	Disponibilidad de Internet
	FAMI_TIENECOMPUTADOR	Booleano	Tiene computadora
	FAMI_TIENELAVADORA	Booleano	Tiene lavadora
	FAMI_TIENEAUTOMOVIL	Booleano	Tiene automóvil
	FAMI_NUMLIBROS	Nominal	Cantidad de libros, revistas y similares en el hogar
	FAMI_TRABAJOLABORPADRE	Nominal	Labor del trabajo del padre
	FAMI_TRABAJOLABORMADRE	Nominal	Labor del trabajo de la madre
	FAMI_TIENEHORNOMICROOGAS	Booleano	Tiene horno
	FAMI_TIENESERVICIOTV	Booleano	Servicio de televisión cerrada
	FAMI_TIENEMOTOCICLETA	Booleano	Tiene motocicleta
	FAMI_TIENECONSOLAVIDEOJUEGOS	Booleano	Tiene consola de videojuegos
	FAMI_COMELECHEDERIVADOS	Nominal	Veces por semana que consume derivados lácteos
	FAMI_COMECARNEPESCADOHUEVO	Nominal	Veces por semana que consume pescado y huevo
	FAMI_COMECEREALFRUTOSLEGUMBRE	Nominal	Veces por semana que consume cereales – frutas – legumbres
	FAMI_SITUACIONECONOMICA	Nominal	Percepción de la situación económica respecto al año anterior
Información de la prueba Saber 11	PUNT_LECTURA_CRITICA	Numérico	Puntajes y desempeño en las áreas de lectura crítica, matemática, ciencias naturales, ciencias sociales y ciudadanas, e inglés
	PUNT_MATEMATICAS	Numérico	
	PUNT_C_NATURALES	Numérico	
	PUNT_SOCIALES_CIUDADANAS	Numérico	
	PUNT_INGLES	Numérico	
	PUNT_GLOBAL	Numérico	
	DESEMP_INGLES	Numérico	
	DESEMP_SOCIALES_CIUDADANAS	Numérico	
	DESEMP_MATEMATICAS	Numérico	
	DESEMP_C_NATURALES	Numérico	
	DESEMP_LECTURA_CRITICA	Numérico	
	PERCENTIL_MATEMATICAS	Numérico	Percentiles de la prueba
	PERCENTIL_C_NATURALES	Numérico	
	PERCENTIL_LECTURA_CRITICA	Numérico	

	PERCENTIL_SOCIALES_CIUDADANAS	Numérico	
	PERCENTIL_INGLES	Numérico	
	PERCENTIL_GLOBAL	Numérico	
	PERIODO	Nominal	Periodo de presentación

Anexo 3. Diccionario de datos pruebas Saber 11°

CAMPO	DESCRIPCIÓN DEL CAMPO	OPCIONES DE RESPUESTA
INFORMACIÓN PERSONAL		
ESTU_TIPODOCUMENTO	Tipo de Documento	CC – Cédula de ciudadanía
		CE – Cédula extranjera
		CR – Certificado registraduría
		CCB – Certificado de cabildo
		NES – Número establecido por la SE
		PC – Pasaporte colombiano
		PE – Pasaporte extranjero
		RC – Registro civil
		PEP– Permiso Especial de Permanencia
		TI – Tarjeta de identidad
ESTU_NACIONALIDAD	Nacionalidad	Texto
ESTU_GENERO	Género	F - Femenino
		M - Masculino
ESTU_FECHANACIMIENTO	Fecha de Nacimiento	[DD/MM/AAAA]
PERIODO	Periodo de presentación del examen	20191 - Corresponde a la aplicación de Saber11 Calendario B 20194 - Corresponde a la aplicación de Saber11 Calendario A
ESTU_CONSECUTIVO	Id público del inscrito (SB11)	Texto
ESTU_ESTUDIANTE	Indica si el inscrito realizó la inscripción por medio de un colegio (estudiante) o fue de manera particular (individual)	ESTUDIANTE
ESTU_PAIS_RESIDE	País donde reside actualmente	Texto
ESTU_TIENEETNIA	¿Pertenece usted a un grupo étnico minoritario?	No
		Si
		Arhuaco
		Cancuamo
		Comunidad afrodescendiente
		Comunidades Rom (Gitanas)
		Cubeo
		Emberá
		Guambiano

ESTU_ETNIA	¿Cuál es el grupo étnico minoritario al que pertenece?	Huitoto
		Inga
		Paez
		Palenquero
		Pasto
		Pijao
		Raizal
		Sikuani
		Tucano
		Wayúu
		Zenú
		Otro grupo étnico minoritario
		Ninguno
ESTU_LIMITA_MOTRIZ	Se inscribió indicando que tiene discapacidad -Motriz	X – Indica que se inscribió con discapacidad
ESTU_LIMITA_INVIDENTE	Se inscribió indicando que tiene discapacidad -Invidente	X – Indica que se inscribió con discapacidad
ESTU_LIMITA_CONDICIONESPECIAL	Se inscribió indicando que tiene discapacidad - Condición Especial	X – Indica que se inscribió con discapacidad
ESTU_LIMITA_SORDO	Se inscribió indicando que tiene discapacidad - Sordo	X – Indica que se inscribió con discapacidad
ESTU_LIMITA_SDOWN	Se inscribió indicando que tiene discapacidad – Síndrome de Down	X – Indica que se inscribió con discapacidad
ESTU_LIMITA_AUTISMO	Se inscribió indicando que tiene discapacidad - Autismo	X – Indica que se inscribió con discapacidad
INFORMACIÓN DE CONTACTO		
ESTU_DEPTO_RESIDE	Departamento de residencia	Texto
ESTU_COD_RESIDE_DEPTO	Código Dane del departamento de residencia	Numérica [99999 extranjero]
ESTU_MCPIO_RESIDE	Municipio de Residencia	Texto
ESTU_COD_RESIDE_MCPIO	Código Dane del municipio de residencia	Numérica [99999 extranjero]
ESTU_AREARESIDE	Área de residencia	Área Rural Cabecera Municipal
ESTU_VALORPENSIONCOLEGIO	Valor mensual de la pensión que paga actualmente	No paga Pensión Menos de 87.000 Entre 87.000 y menos de 120.000 Entre 120.000 y menos de 150.000 Entre 150.000 y menos de 250.000 250.000 o más

ESTU_VECESPRESENTOEXAMEN	¿Cuántas veces ha presentado el examen SB11? Para ESTUDIANTES	Ninguna vez Una vez Dos veces Tres veces o más
INFORMACIÓN SOCIOECONÓMICA		
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de la vivienda según recibo de energía eléctrica	Estrato 1
		Estrato 2
		Estrato 3
		Estrato 4
		Estrato 5
		Estrato 6
		Sin Estrato
FAMI_NUMHERMANOS	¿Cuántas hermanas y hermanos tiene usted en total?	Ninguno Uno Dos Tres Cuatro Cinco Seis Siete Ocho Nueve Más de nueve
FAMI_PERSONASHOGAR	¿Cuántas personas conforman el hogar donde vive actualmente, incluido usted?	1 a 2
		3 a 4
		5 a 6
		7 a 8
		9 o más
FAMI_PISOSHOGAR	¿Cuál es el material de los pisos que predomina en su vivienda?	Cemento, gravilla, ladrillo. Madera burda, tabla, tablón. Madera pulida, baldosa, tableta, mármol, alfombra. Tierra, arena.
FAMI_CUARTOSHOGAR	En total, ¿en cuántos cuartos duermen las personas de su hogar?	Uno
		Dos
		Tres
		Cuatro
		Cinco
		Seis o más
		Ninguno

FAMI_EDUCACIONPADRE	Nivel educativo más alto alcanzado por el padre	Primaria incompleta
		Primaria completa
		Secundaria (Bachillerato) incompleta
		Secundaria (Bachillerato) completa
		Técnica o tecnológica incompleta
		Técnica o tecnológica completa
		Educación profesional incompleta
		Educación profesional completa
		Postgrado
		No Aplica
		No sabe
FAMI_EDUCACIONMADRE	Nivel educativo más alto alcanzado por la madre	Ninguno
		Primaria incompleta
		Primaria completa
		Secundaria (Bachillerato) incompleta
		Secundaria (Bachillerato) completa
		Técnica o tecnológica incompleta
		Técnica o tecnológica completa
		Educación profesional incompleta
		Educación profesional completa
		Postgrado
		No Aplica
		No sabe
FAMI_OCUPACIONPADRE	Ocupación u oficio del padre	Empleado con cargo como director o gerente general Empleado de nivel auxiliar o administrativo Empleado de nivel directivo Empleado de nivel técnico o profesional Empleado obrero u operario Empresario Hogar Otra actividad u ocupación Pensionado Pequeño empresario Profesional independiente Trabajador por cuenta propia

FAMI_OCUPACIONMADRE	Ocupación u oficio de la madre	Empleado con cargo como director o gerente general Empleado de nivel auxiliar o administrativo Empleado de nivel directivo Empleado de nivel técnico o profesional Empleado obrero u operario Empresario Hogar Otra actividad u ocupación Pensionado Pequeño empresario Profesional independiente Trabajador por cuenta propia
FAMI_TRABAJOLABORPADRE	Señale aquella labor que sea más similar al trabajo que realizó el padre durante la mayor parte del último año:	Es agricultor, pesquero o jornalero.
		Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial.
		Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo, tienda, papelería, etc.
		Es operario de máquinas o conduce vehículos (taxista, chofer).
		Es vendedor o trabaja en atención al público.
		Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente).
		Trabaja como personal de limpieza, mantenimiento, seguridad o construcción. Trabaja como profesional (por ejemplo, médico, abogado, ingeniero).
		Trabaja en el hogar, no trabaja o estudia.
		Trabaja por cuenta propia (por ejemplo, plomero, electricista).
		Pensionado.
		No sabe.
		No aplica
FAMI_TRABAJOLABORMADRE	Señale aquella labor que sea más similar al trabajo que realizó la madre durante la mayor parte del último año:	Es agricultor, pesquero o jornalero.
		Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial.
		Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo, tienda, papelería, etc.
		Es operario de máquinas o conduce vehículos (taxista, chofer).
		Es vendedor o trabaja en atención al público.
		Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente).
		Trabaja como personal de limpieza, mantenimiento, seguridad o construcción. Trabaja como profesional (por ejemplo, médico, abogado, ingeniero).
		Trabaja en el hogar, no trabaja o estudia.

		Trabaja por cuenta propia (por ejemplo, plomero, electricista).
		Pensionado.
		No sabe.
		No aplica
FAMI_NIVELSISBEN	Puntaje de SISBEN en el que está clasificada su familia	Nivel 1 Nivel 2 Nivel 3 Está clasificada en otro nivel del SISBEN No está clasificada por el SISBEN
FAMI_TIENEINTERNET	¿Su hogar cuenta con servicio o conexión a internet?	¿No conexión a internet?
FAMI_TIENESERVICIOTV	¿Su hogar cuenta con servicio cerrado de televisión?	No
		Si
FAMI_TIENECOMPUTADOR	¿Cuáles de los siguientes bienes posee su hogar?: Computador	No
		Si
FAMI_TELEFONO	¿Con cuáles de los siguientes servicios públicos, privados o comunales cuenta su hogar?: Teléfono (fijo)	No Si
FAMI_TIENELAVADORA	¿Cuáles de los siguientes bienes posee su hogar?: Máquina lavadora de ropa	No
		Si
FAMI_TIENEHORNOMICROOGAS	¿Cuáles de los siguientes bienes posee su hogar?: Horno Microondas u Horno eléctrico o a gas	No
		Si
FAMI_TIENEHORNO	¿Cuáles de los siguientes bienes posee su hogar?: Horno eléctrico o a gas	No Si
FAMI_TIENEDVD	¿Cuáles de los siguientes bienes posee su hogar?: DVD	No Si
FAMI_TIENEMICROONDAS	¿Cuáles de los siguientes bienes posee su hogar?: Horno microondas	No Si
FAMI_TIENEAUTOMOVIL	¿Cuáles de los siguientes bienes posee su hogar?: Automóvil particular	No
		Si
FAMI_TIENEMOTOCICLETA	¿Cuáles de los siguientes bienes posee su hogar?: Motocicleta	No
		Si
FAMI_TIENECONSOLAVIDEOJUEGOS	¿Cuáles de los siguientes bienes posee su hogar?: Consola para juegos electrónicos (PlayStation, Xbox, Nintendo, etc.)	No Si
		0 a 10 LIBROS

FAMI_NUMLIBROS	¿Cuántos libros físicos o electrónicos hay en su hogar excluyendo periódicos, revistas, directorios telefónicos y libros del colegio?	11 a 25 LIBROS
		26 a 100 LIBROS
		MÁS DE 100 LIBROS
FAMI_INGRESOFMILIARMENSUAL	¿Cuál es el total de ingresos mensuales de su hogar, en términos de salarios mínimos (SMMLV)?	Menos de 1 SMLV Entre 1 y menos de 2 SMLV Entre 2 y menos de 3 SMLV Entre 3 y menos de 5 SMLV Entre 5 y menos de 7 SMLV Entre 7 y menos de 10 SMLV 10 o más SMLV
FAMI_COMELECHEDERIVADOS	¿Cuántas veces por semana se comen los siguientes alimentos en su hogar? Leche o derivados (queso, yogurt, etc.)	1 o 2 veces por semana
		3 a 5 veces por semana
		Nunca o rara vez comemos eso
		Todos o casi todos los días
FAMI_COMECARNEPESCADOHUEVO	¿Cuántas veces por semana se comen los siguientes alimentos en su hogar? Carne (pollo, pavo, res, cordero, cerdo, conejo, etc.), pescados o huevos	1 o 2 veces por semana
		3 a 5 veces por semana
		Nunca o rara vez comemos eso
		Todos o casi todos los días
FAMI_COMECEREALFRUTOSLEGUMBRE	¿Cuántas veces por semana se comen los siguientes alimentos en su hogar? Cereales (avena, granola), frutos secos (almendras, maní) o legumbres (frijoles, garbanzos, lentejas)	1 o 2 veces por semana
		3 a 5 veces por semana
		Nunca o rara vez comemos eso
		Todos o casi todos los días
FAMI_SITUACIONECONOMICA	Con respecto al año inmediatamente anterior, la situación económica de su hogar es:	Igual
		Mejor
		Peor
ESTU_DEDICACIONLECTURADIARIA	Usualmente, ¿cuánto tiempo al día dedica a leer por entretenimiento?	No leo por entretenimiento
		30 minutos o menos
		Entre 30 y 60 minutos
		Entre 1 y 2 horas
		Más de 2 horas
	Usualmente, ¿cuánto tiempo al día dedica a navegar en internet? Excluya actividades académicas	No Navega Internet
		30 minutos o menos

ESTU_DEDICACIONINTERNET		Entre 30 y 60 minutos
		Entre 1 y 3 horas
		Más de 3 horas
ESTU_HORASSEMANATRABAJA	¿Cuántas horas trabajó usted durante la semana pasada?	0
		Menos de 10 horas
		Entre 11 y 20 horas
		Entre 21 y 30 horas
		Más de 30 horas
ESTU_TIPOREMUNERACION	¿Usted recibe algún tipo de remuneración por trabajar?	No
		Si, en efectivo
		Si, en especie
		Si, en efectivo y especie
ESTU_HORASSEMANATRABAJA	¿Cuántas horas trabajó usted durante la semana pasada?	0 Menos de 10 horas Entre 11 y 20 horas Entre 21 y 30 horas Más de 30 horas
ESTU_TRABAJAACTUALMENTE	¿Trabaja usted actualmente?	No Si, 20 horas o más a la semana Si, menos de 20 horas a la semana
ESTU_RECIBESALARIO	¿Recibe algún salario por trabajar?	No Si
COLE_CODIGO_ICFES	Código Icfes de la sede-jornada	Numérica
COLE_COD_DANE_ESTABLECIMIENTO	Código Dane del Establecimiento Educativo	Numérica
COLE_NOMBRE_ESTABLECIMIENTO	Nombre del Establecimiento	Texto
COLE_GENERO	Indica el género de la población del Establecimiento.	FEMENINO
		MASCULINO
		MIXTO
COLE_NATURALEZA	Indica la naturaleza del Establecimiento	NO OFICIAL
		OFICIAL
COLE_CALENDARIO	Calendario académico del Establecimiento	A
		B
		OTRO
	Indica si el Establecimiento es bilingüe o no	N - No

COLE_BILINGUE		S - Sí
COLE_CARACTER	Indica el carácter del Establecimiento	ACADÉMICO
		TÉCNICO
		TÉCNICO/ACADÉMICO
		NO APLICA
COLE_COD_DANE_SEDE	Código Dane de la Sede	Numérica
COLE_NOMBRE_SEDE	Nombre de la Sede	Texto
COLE_SEDE_PRINCIPAL	¿Esta es la sede principal del Establecimiento Educativo?	N - No
		S - Sí
COLE_AREA_UBICACION	Área de ubicación de la Sede	RURAL
		URBANO
COLE_JORNADA	Jornada de la Sede	COMPLETA
		MAÑANA
		NOCHE
		SABATINA
		TARDE
		UNICA
COLE_COD_MCPIO_UBICACION	Código Dane del municipio donde está ubicada la Sede	Numérica
COLE_MCPIO_UBICACION	Nombre del municipio donde está ubicada la Sede	Texto
COLE_COD_DEPTO_UBICACION	Código Dane del departamento de la Sede	Numérica
COLE_DEPTO_UBICACION	Nombre del departamento donde está ubicada la Sede	Texto
DATOS DE CITACIÓN DEL EXAMEN		
ESTU_PRIVADO_LIBERTAD	¿Se encuentra privado de la libertad?	N - No
		S - Sí
ESTU_COD_MCPIO_PRESENTACION	Código Dane del municipio presentación del examen	Numérica
ESTU_MCPIO_PRESENTACION	Municipio de presentación del examen	Texto
ESTU_DEPTO_PRESENTACION	Departamento de presentación del examen	Texto
ESTU_COD_DEPTO_PRESENTACION	Código Dane del departamento de presentación del examen	Numérica
RESULTADOS		
PUNT_LECTURA_CRITICA	Puntaje en lectura crítica	Numérica
PERCENTIL_LECTURA_CRITICA	Percentil lectura crítica	Numérica

DESEMP_LECTURA_CRITICA	Desempeño en lectura crítica	Numérica
PUNT_MATEMATICAS	Puntaje en matemáticas	Numérica
PERCENTIL_MATEMATICAS	Percentil matemáticas	Numérica
DESEMP_MATEMATICAS	Desempeño matemáticas	Numérica
PUNT_C_NATURALES	Puntaje en ciencias naturales	Numérica
PERCENTIL_C_NATURALES	Percentil ciencias naturales	Numérica
DESEMP_C_NATURALES	Desempeño ciencias naturales	Numérica
PUNT_SOCIALES_CIUDADANAS	Puntaje sociales y ciudadanas	Numérica
PERCENTIL_SOCIALES_CIUDADANAS	Percentil sociales y ciudadanas	Numérica
DESEMP_SOCIALES_CIUDADANAS	Desempeño sociales y ciudadanas	Numérica
PUNT_INGLES	Puntaje inglés. *El valor de -1 indica que el estudiante no contestó ninguna pregunta*	Numérica
PERCENTIL_INGLES	Percentil en inglés	Numérica
DESEMP_INGLES	Desempeño en inglés	A-
		A1
		A2
		B+
		B1
PUNT_GLOBAL	Puntaje total obtenido	Numérica
PERCENTIL_GLOBAL	Percentil global en que se encuentra el evaluado	Numérica
ESTU_INSE_INDIVIDUAL	Índice Socioeconómico del evaluado	Numérica
ESTU_NSE_INDIVIDUAL	Nivel Socioeconómico del evaluado	NSE1
		NSE2
		NSE3
		NSE4
ESTU_NSE_ESTABLECIMIENTO	Nivel Socioeconómico del Establecimiento	Numérica
ESTU_ESTADOINVESTIGACION	Identifica el estado de los resultados para los evaluados	PUBLICAR
		NO SE COMPROBO IDENTIDAD DEL EXAMINADO
		VALIDEZ OFICINA JURÍDICA - La validez de los resultados está condicionada a la conclusión de la

		actuación administrativa que adelanta la Oficina Asesora Jurídica del Icfes por la presunta comisión de falta o conducta prohibida, en los términos de la Resolución 631/2015 del ICFES.
ESTU_GENERACIONE	Identifica si el evaluado cumple con los requisitos necesarios para ser beneficiario del programa Generación E	GENERACION E - EXCELENCIA DEPARTAMENTAL
		GENERACION E - EXCELENCIA NACIONAL
		GENERACION E - GRATUIDAD
		NO

Anexo 4. Dimensiones y niveles Meloda 5.0

Dimensiones (máximo 61 puntos)	*Niveles*
Licencia legal (máx. 6 puntos)	1. uso privado 2. reutilización no comercial 3. reutilización comercial o sin restricciones
Acceso a la información (máx. 6 puntos)	1. acceso a través de web o parámetros únicos URL al conjunto de datos 2. acceso único a la web con parámetros referidos a datos individuales 3. API o lenguaje de consulta
Estándar técnico (máx. 6 puntos)	1. estándar cerrado reutilizable o abierto no reutilizable 2. estándar abierto reutilizable 3. estándar abierto con metadatos individuales
Nivel de estandarización (máx. 10 puntos)	1. modelo propio de estandarización 2. modelo de estandarización propio o ad hoc publicado (armonización) 3. estandarización local 4. estandarización global
Contenido geolocalizado (máx. 6 puntos)	1. sin información geográfica 2. campo de texto simple o complejo 3. con coordenadas o información geográfica completa
Actualización de la frecuencia de datos (máx. 15 puntos)	1. por encima de un mes 2. mensual. Con periodos de actualización entre 1 mes y 1 día 3. diaria. Con periodos de actualización entre 1 día y 1 hora 4. cada hora. Con periodos de actualización desde 1 hora a 1 minuto 5. en segundos. Periodo de actualización por debajo de 1 minuto
Difusión (máx. 6 puntos)	1. comunicación/difusión no sistemática 2. recursos disponibles sobre actualizaciones (p. ej., alimentación en redes sociales) 3. difusión proactiva/difusión push (información automatizada en determinado tiempo)
Reputación (máx. 6 puntos)	1. no hay información sobre la reputación de la fuente de datos 2. las estadísticas o informes se publican en función a las opiniones de los usuarios 3. rankings o indicadores basados en la reputación de la fuente de datos

Anexo 5. Proceso de limpieza con Open Refine

Figura 64 Variable COLE_NOMBRE_ESTABLECIMIENTO (21 agrupamientos)

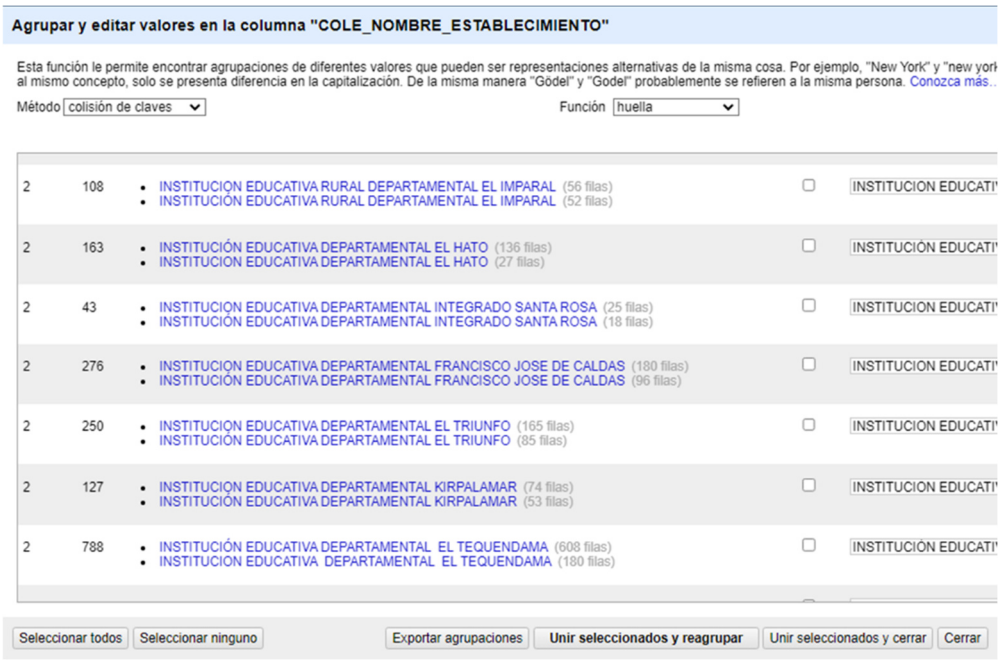


Figura 65 Variable COLE_NOMBRE_SEDE (19 agrupamientos)

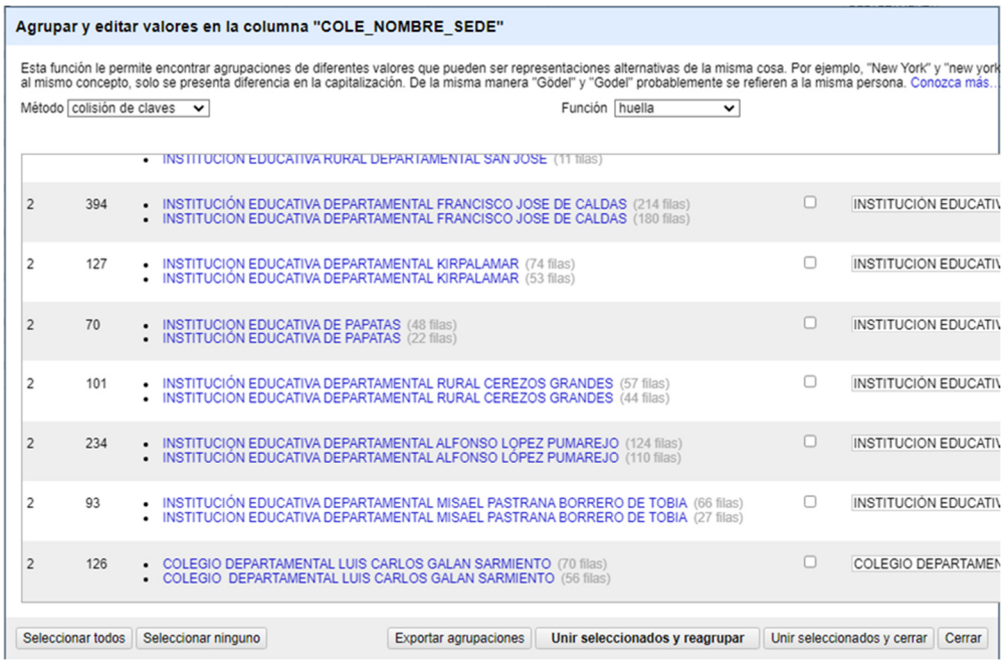


Figura 66 Variable COLE_MCPIO_UBICACION (39 agrupamientos)

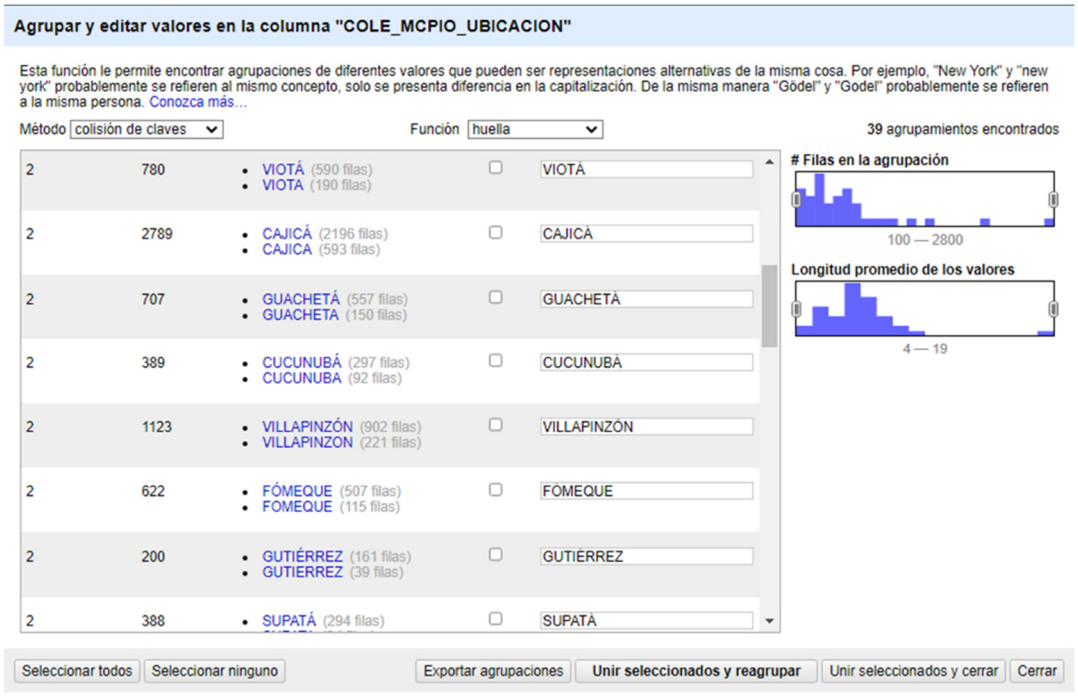
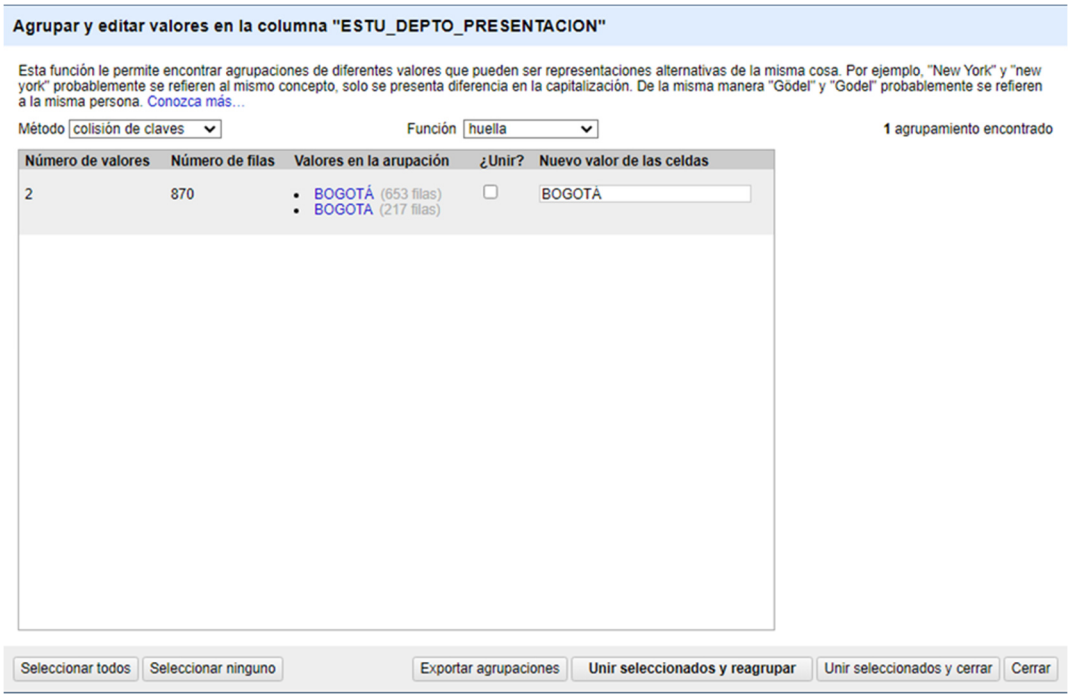


Figura 67 Variable ESTU_DPTO_PRESENTACION (1 agrupamiento)


Agrupar y editar valores en la columna "ESTU_DEPTO_PRESENTACION"
Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Gödel" y "Godel" probablemente se refieren a la misma persona. Conozca más...
Método colisión de claves
Función huella
1 agrupamiento encontrado
Número de valores
Número de filas
Valores en la arupación
¿Unir?
Nuevo valor de las celdas
2
870
BOGOTÁ (653 filas)
BOGOTA (217 filas)
BOGOTÀ
Seleccionar todos
Seleccionar ninguno
Exportar agrupaciones
Unir seleccionados y reagrupar
Unir seleccionados y cerrar
Cerrar

Anexo 6. Variables con valores faltantes

#	Grupo	Atributo	Tipo	Cantidad Valores faltantes	Porcentaje	Cant x Grupo
1	Información de la Entidad Educativa	COLE_BILINGUE	Booleano	7722	11.517%	7751
2		COLE_CARACTER	Nominal	29	0.043%	
3	Información de la prueba Saber 11	DESEMP_INGLES	Numérico	2	0.003%	8
4		PERCENTIL_GLOBAL	Numérico	2	0.003%	
5		PERCENTIL_INGLES	Numérico	2	0.003%	
6		PUNT_INGLES	Numérico	2	0.003%	
7	Información del Estudiante	ESTU_COD_DEPTO_PRESENTACION	Nominal	3	0.004%	10041
8		ESTU_COD_MCPIO_PRESENTACION	Nominal	3	0.004%	
9		ESTU_COD_RESIDE_DEPTO	Nominal	11	0.016%	
10		ESTU_COD_RESIDE_MCPIO	Nominal	11	0.016%	
11		ESTU_DEDICACIONINTERNET	Nominal	2576	3.842%	
12		ESTU_DEDICACIONLECTURADIARIA	Nominal	2416	3.603%	
13		ESTU_DEPTO_PRESENTACION	Nominal	3	0.004%	
14		ESTU_DEPTO_RESIDE	Nominal	11	0.016%	
15		ESTU_GENERO	Nominal	1	0.001%	
16		ESTU_HORASSEMANATRABAJA	Nominal	1223	1.824%	
17		ESTU_INSE_INDIVIDUAL	Numérico	1123	1.675%	
18		ESTU_MCPIO_PRESENTACION	Nominal	3	0.004%	
19		ESTU_MCPIO_RESIDE	Nominal	11	0.016%	
20		ESTU_NSE_ESTABLECIMIENTO	Numérico	24	0.036%	
21		ESTU_NSE_INDIVIDUAL	Nominal	1123	1.675%	
22		ESTU_TIENEETNIA	Booleano	115	0.172%	
23		ESTU_TIPOREMUNERACION	Nominal	1384	2.064%	
24	Información Familiar	FAMI_COMECARNEPESCADOHUEVO	Nominal	2473	3.688%	38007
25		FAMI_COMECEREALFRUTOSLEGUMBRE	Nominal	2650	3.952%	
26		FAMI_COMELECHEDERIVADOS	Nominal	2774	4.137%	
27		FAMI_CUARTOSHOGAR	Nominal	1262	1.882%	
28		FAMI_EDUCACIONMADRE	Nominal	2245	3.348%	
29		FAMI_EDUCACIONPADRE	Nominal	2283	3.405%	
30		FAMI_ESTRATOVIVIENDA	Nominal	2649	3.951%	
31		FAMI_NUMLIBROS	Nominal	3483	5.195%	

32		FAMI_PERSONASHOGAR	Nominal	1309	1.952%	
33		FAMI_SITUACIONECONOMICA	Nominal	1301	1.940%	
34		FAMI_TIENEAUTOMOVIL	Booleano	1358	2.025%	
35		FAMI_TIENECOMPUTADOR	Booleano	1340	1.999%	
36		FAMI_TIENECONSOLAVIDEOJUEGOS	Booleano	1334	1.990%	
37		FAMI_TIENEHORNOMICROOGAS	Booleano	1302	1.942%	
38		FAMI_TIENEINTERNET	Booleano	2339	3.488%	
39		FAMI_TIENELAVADORA	Booleano	1233	1.839%	
40		FAMI_TIENEMOTOCICLETA	Booleano	1261	1.881%	
41		FAMI_TIENESERVICIOTV	Booleano	2481	3.700%	
42		FAMI_TRABAJOLABORMADRE	Nominal	1387	2.069%	
43		FAMI_TRABAJOLABORPADRE	Nominal	1543	2.301%	
	Total					55807

Anexo 7. Imputación de datos faltantes con el paquete MICE in R

#Leer el conjunto de datos
df <- read.csv ("D:/AppFiles/Universidad/4. Saber11/IDData/SB11_TotalCleanSFechaSID.csv", encoding = "UTF-8", header = TRUE, stringsAsFactors = T)

#Veamos el encabezado del conjunto de datos.
head(df)

Description: df [6 x 44]

	ESTU_TIPODOCUMENTO <fctr>	ESTU_GENERO <fctr>	PERIODO <int>	ESTU_MCPIO_RESIDE <fctr>	FAMI_EDUCACIONPADRE <fctr>
1	CC	F	2017	Other	No sabe
2	TI	F	2017	Other	Primaria incompleta
3	CC	F	2017	Other	No sabe
4	TI	F	2017	Other	Secundaria (Bachillerato) completa
5	TI	M	2017	Other	Secundaria (Bachillerato) incompleta
6	TI	F	2017	Other	Secundaria (Bachillerato) completa

6 rows | 1-6 of 44 columns

#Verificar los datos en busca de valores faltantes
sapply (df, function(x) sum(is.na(x)))

```
> sapply(df, function(x) sum(is.na(x)))
          ESTU_TIPODOCUMENTO                ESTU_GENERO                   PERIODO 
                           0                           1                           0 
          ESTU_MCPIO_RESIDE            FAMI_EDUCACIONPADRE          FAMI_EDUCACIONMADRE 
                           0                        2283                        2245 
       FAMI_ESTRATOVIVIENDA           FAMI_PERSONASHOGAR             FAMI_CUARTOSHOGAR 
                        2649                        1309                        1262 
          FAMI_TIENEINTERNET          FAMI_TIENECOMPUTADOR           FAMI_TIENELAVADORA 
                        2339                        1340                        1233 
  FAMI_TIENEHORNOMICROOGAS          FAMI_TIENESERVICIOTV           FAMI_TIENEAUTOMOVIL 
                        1302                        2481                        1358 
       FAMI_TIENEMOTOCICLETA     FAMI_TIENECONSOLAVIDEOJUEGOS               FAMI_NUMLIBROS 
                        1261                        1334                        3483 
    FAMI_COMELECHEDERIVADOS     FAMI_COMECARNEPESCADOHUEVO FAMI_COMECEREALFRUTOSLEGUMBRE 
                        2774                        2473                        2650 
        FAMI_TRABAJOLABORPADRE          FAMI_TRABAJOLABORMADRE          FAMI_SITUACIONECONOMICA 
                        1543                        1387                        1301 
 ESTU_DEDICACIONLECTURADIARIA       ESTU_DEDICACIONINTERNET          ESTU_HORASSEMANATRABAJA 
                        2416                        2576                        1223 
          ESTU_TIPOREMUNERACION      COLE_NOMBRE_ESTABLECIMIENTO                  COLE_BILINGUE 
                        1384                           0                        7722 
                COLE_CARACTER                COLE_NOMBRE_SEDE           COLE_SEDE_PRINCIPAL 
                          29                           0                           0 
             COLE_AREA_UBICACION                 COLE_JORNADA              COLE_MCPIO_UBICACION 
                           0                           0                           0 
          ESTU_MCPIO_PRESENTACION       ESTU_DEPTO_PRESENTACION          ESTU_INSE_INDIVIDUAL 
                           0                           3                        1123 
       ESTU_NSE_INDIVIDUAL         ESTU_NSE_ESTABLECIMIENTO                  ESTU_DESEMPEÑO 
                        1123                          24                           0 
                  ESTU_EDAD                  COLE_REGION 
                        5977                           0 
> |
```

#Mirar la estructura del conjunto de datos.
str(df)

```
data.frame':    67049 obs. of  44 variables:
 $ ESTU_TIPODOCUMENTO           : Factor w/ 7 levels "CC","CE","CR",..: 1 7 1 7 7 7 7 1 7 7 ...
 $ ESTU_GENERO                  : Factor w/ 2 levels "F","M": 1 1 1 1 2 1 1 1 1 2 ...
 $ PERIODO                      : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
 $ ESTU_MCPIO_RESIDE            : Factor w/ 53 levels "AGUA DE DIOS",..: 27 27 27 27 27 27 27 27 27 27 ...
 $ FAMI_EDUCACIONPADRE          : Factor w/ 12 levels "Educación profesional completa",..: 5 8 5 9 10 9 4 8 8 10 ...
 $ FAMI_EDUCACIONMADRE          : Factor w/ 12 levels "Educación profesional completa",..: 5 7 10 8 9 10 8 8 10 9 ...
 $ FAMI_ESTRATOVIVIENDA         : Factor w/ 7 levels "Estrato 1","Estrato 2",..: 1 3 1 2 1 1 2 1 3 3 ...
 $ FAMI_PERSONASHOGAR           : Factor w/ 5 levels "1 a 2","3 a 4",..: 3 3 3 2 3 2 2 2 NA 2 ...
 $ FAMI_CUARTOSHOGAR            : Factor w/ 6 levels "Cinco","Cuatro",..: 3 3 5 6 5 5 3 5 5 5 ...
 $ FAMI_TIENEINTERNET           : Factor w/ 2 levels "No","Si": 1 2 1 2 2 2 2 1 2 2 ...
 $ FAMI_TIENECOMPUTADOR         : Factor w/ 2 levels "No","Si": 2 1 1 2 2 1 1 1 2 2 ...
 $ FAMI_TIENELAVADORA           : Factor w/ 2 levels "No","Si": NA 1 2 2 2 2 2 2 2 2 ...
 $ FAMI_TIENEHORNOMICROOGAS     : Factor w/ 2 levels "No","Si": NA 1 2 1 2 2 1 1 1 1 ...
 $ FAMI_TIENESERVICIOTV         : Factor w/ 2 levels "No","Si": 1 1 2 2 2 2 2 1 1 2 ...
 $ FAMI_TIENEAUTOMOVIL          : Factor w/ 2 levels "No","Si": NA 1 1 1 1 1 1 1 2 2 ...
 $ FAMI_TIENEMOTOCICLETA        : Factor w/ 2 levels "No","Si": NA 2 1 1 2 2 1 1 2 2 ...
 $ FAMI_TIENECONSOLAVIDEOJUEGOS : Factor w/ 2 levels "No","Si": NA 1 1 1 2 1 1 1 1 2 ...
 $ FAMI_NUMLIBROS               : Factor w/ 4 levels "0 A 10 LIBROS",..: 1 2 1 2 2 2 1 1 1 2 2 ...
 $ FAMI_COMELECHEDERIVADOS      : Factor w/ 4 levels "1 o 2 veces por semana",..: 4 3 4 1 2 1 2 3 1 2 ...
 $ FAMI_COMECARNEPESCADOHUEVO   : Factor w/ 4 levels "1 o 2 veces por semana",..: 4 1 4 2 2 2 1 1 2 1 ...
 $ FAMI_COMECEREALFRUTOSLEGUMBRE: Factor w/ 4 levels "1 o 2 veces por semana",..: 3 2 3 1 2 1 2 2 2 1 ...
 $ FAMI_TRABAJOLABORPADRE       : Factor w/ 13 levels "Es agricultor, pesquero o jornalero",..: 13 13 7 10 3 9 6 1 3 6 ...
 $ FAMI_TRABAJOLABORMADRE       : Factor w/ 13 levels "Es agricultor, pesquero o jornalero",..: 12 13 10 10 6 12 12 12 12 12 ...
 $ FAMI_SITUACIONECONOMICA      : Factor w/ 3 levels "Igual","Mejor",..: 1 1 1 2 3 2 1 2 1 2 ...
 $ ESTU_DEDICACIONLECTURADIARIA : Factor w/ 5 levels "30 minutos o menos",..: 1 1 5 5 3 3 1 1 3 1 ...
 $ ESTU_DEDICACIONINTERNET      : Factor w/ 5 levels "30 minutos o menos",..: 4 3 1 1 4 3 3 4 2 2 ...
 $ ESTU_HORASSEMANATRABAJA      : Factor w/ 5 levels "0","Entre 11 y 20 horas",..: 1 1 1 1 5 5 5 1 1 1 ...
 $ ESTU_TIPOREMUNERACION        : Factor w/ 4 levels "No","Si, en efectivo",..: 1 1 1 1 2 1 2 1 1 1 ...
 $ COLE_NOMBRE_ESTABLECIMIENTO  : Factor w/ 53 levels "I.E.D. ALONSO RONQUILLO",..: 13 13 13 13 13 13 13 13 13 13 ...
 $ COLE_BILINGUE                : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ COLE_CARACTER                : Factor w/ 4 levels "ACADÉMICO","NO APLICA",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ COLE_NOMBRE_SEDE             : Factor w/ 53 levels "I.E.D. ALONSO RONQUILLO - SEDE PRINCIPAL",..: 17 17 17 17 17 17 17 17 17 17 ...
 $ COLE_SEDE_PRINCIPAL          : Factor w/ 2 levels "N","S": 2 2 2 2 2 2 2 2 2 2 ...
 $ COLE_AREA_UBICACION          : Factor w/ 2 levels "RURAL","URBANO": 2 2 2 2 2 2 2 2 2 2 ...
 $ COLE_JORNADA                 : Factor w/ 6 levels "COMPLETA","MAÑANA",..: 3 2 3 2 2 3 3 3 2 2 ...
 $ COLE_MCPIO_UBICACION         : Factor w/ 53 levels "AGUA DE DIOS",..: 26 26 26 26 26 26 26 26 26 26 ...
 $ ESTU_MCPIO_PRESENTACION      : Factor w/ 53 levels "ALBÁN","ANAPOIMA",..: 21 21 21 21 21 21 21 21 21 21 ...
 $ ESTU_DEPTO_PRESENTACION      : Factor w/ 23 levels "ANTIOQUIA","BOGOTÁ",..: 12 12 12 12 12 12 12 12 12 12 ...
 $ ESTU_INSE_INDIVIDUAL         : num  46.7 42.9 45.6 48.6 56.1 ...
 $ ESTU_NSE_INDIVIDUAL          : Factor w/ 4 levels "NSE1","NSE2",..: 2 2 2 2 3 2 2 1 3 3 ...
 $ ESTU_NSE_ESTABLECIMIENTO     : int  2 2 2 2 2 2 2 2 2 2 ...
 $ ESTU_DESEMPEÑO               : Factor w/ 4 levels "Avanzado","Insuficiente",..: 2 3 3 2 3 3 3 2 3 4 ...
 $ ESTU_EDAD                    : int  NA 17 NA 17 17 18 17 NA 17 17 ...
 $ COLE_REGION                  : Factor w/ 14 levels "ALMEIDAS","ALTO MAGDALENA",..: 2 2 2 2 2 2 2 2 2 2 ...
```

Imputación
#Ahora que el conjunto de datos está listo para la imputación llamaremos al paquete de MICE.
library(mice)
init = mice (df, maxit=0)
meth = init$method
predM = init$predictorMatrix

#Para imputar los valores faltantes el paquete MICE usa un algoritmo de tal manera que usa información de otras variables
#en el conjunto de datos para predecir e imputar los valores faltantes.
set.seed(103)
imputed = mice (df, method="cart", predictorMatrix=predM, m=5)

```
iter imp variable
  1   1  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  1   2  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  1   3  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  1   4  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  1   5  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  2   1  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  2   2  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  2   3  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  2   4  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  2   5  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  3   1  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  3   2  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  3   3  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  3   4  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  3   5  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  4   1  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  4   2  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  4   3  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  4   4  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  4   5  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  5   1  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  5   2  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  5   3  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  5   4  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
  5   5  ESTU_GENERO  ESTU_TIENEETNIA  ESTU_DEPTO_RESIDE  FAMI_EDUCACIONPADRE  FAMI_EDUCACIONMADRE  FAMI_ESTRATOVIVIENDA  FAMI_PERSONASHOGAR  FAMI_CUARTOSHOGAR  FAMI_TIENEINTERNET  FAMI_TIENECOMPUTADOR  FAMI_TIENELAVADORA  FAMI
```

##Reporte
reporte <- summary(imputed)

```
Class: mids
Number of multiple imputations:  5
Imputation methods:
           ESTU_TIPODOCUMENTO              ESTU_NACIONALIDAD                   ESTU_GENERO
                           ""                             ""                        "cart"
                      PERIODO               ESTU_PAIS_RESIDE                ESTU_TIENEETNIA
                           ""                             ""                        "cart"
            ESTU_DEPTO_RESIDE             ESTU_MCPIO_RESIDE            FAMI_EDUCACIONPADRE
                        "cart"                             ""                        "cart"
          FAMI_EDUCACIONMADRE          FAMI_ESTRATOVIVIENDA             FAMI_PERSONASHOGAR
                        "cart"                        "cart"                        "cart"
            FAMI_CUARTOSHOGAR            FAMI_TIENEINTERNET           FAMI_TIENECOMPUTADOR
                        "cart"                        "cart"                        "cart"
           FAMI_TIENELAVADORA       FAMI_TIENEHORNOMICROOGAS           FAMI_TIENESERVICIOTV
                        "cart"                        "cart"                        "cart"
          FAMI_TIENEAUTOMOVIL          FAMI_TIENEMOTOCICLETA  FAMI_TIENECONSOLAVIDEOJUEGOS
                        "cart"                        "cart"                        "cart"
               FAMI_NUMLIBROS        FAMI_COMELECHEDERIVADOS     FAMI_COMECARNEPESCADOHUEVO
                        "cart"                        "cart"                        "cart"
FAMI_COMECEREALFRUTOSLEGUMBRE          FAMI_TRABAJOLABORPADRE         FAMI_TRABAJOLABORMADRE
                        "cart"                        "cart"                        "cart"
      FAMI_SITUACIONECONOMICA  ESTU_DEDICACIONLECTURADIARIA        ESTU_DEDICACIONINTERNET
                        "cart"                        "cart"                        "cart"
       ESTU_HORASSEMANATRABAJA          ESTU_TIPOREMUNERACION    COLE_NOMBRE_ESTABLECIMIENTO
                        "cart"                        "cart"                            ""
                 COLE_BILINGUE                 COLE_CARACTER                COLE_NOMBRE_SEDE
                        "cart"                        "cart"                            ""
           COLE_SEDE_PRINCIPAL              COLE_AREA_UBICACION                  COLE_JORNADA
                           ""                             ""                            ""
         COLE_MCPIO_UBICACION            ESTU_PRIVADO_LIBERTAD         ESTU_MCPIO_PRESENTACION
                           ""                             ""                            ""
       ESTU_DEPTO_PRESENTACION            PUNT_LECTURA_CRITICA       PERCENTIL_LECTURA_CRITICA
                        "cart"                        "cart"                            ""
        DESEMP_LECTURA_CRITICA                 PUNT_MATEMATICAS           PERCENTIL_MATEMATICAS
                           ""                        "cart"                            ""
             DESEMP_MATEMATICAS                 PUNT_C_NATURALES           PERCENTIL_C_NATURALES
                           ""                        "cart"                            ""
             DESEMP_C_NATURALES       PUNT_SOCIALES_CIUDADANAS PERCENTIL_SOCIALES_CIUDADANAS
                           ""                        "cart"                            ""
    DESEMP_SOCIALES_CIUDADANAS                    PUNT_INGLES                PERCENTIL_INGLES
                        "cart"                        "cart"                        "cart"
                  DESEMP_INGLES                   PUNT_GLOBAL                PERCENTIL_GLOBAL
                        "cart"                        "cart"                        "cart"
          ESTU_INSE_INDIVIDUAL            ESTU_NSE_INDIVIDUAL          ESTU_NSE_ESTABLECIMIENTO
                        "cart"                        "cart"                        "cart"
       ESTU_ESTADOINVESTIGACION                 ESTU_DESEMPEÑO                     ESTU_EDAD
                           ""                             ""                        "cart"
                    COLE_REGION
                           ""
```

```{r}
str(reporte)
```

Classes 'mipo.summary' and 'data.frame': 19 obs. of 6 variables:
$ term   : Factor w/ 19 levels "(Intercept)",..: 1 2 3 4 5 6 7 8 9 10 ...
$ estimate: num 68.46 1.16 1.14 1.15 1.14 ...
$ std.error: num 10.78652 0.00371 0.00557 0.00577 0.00644 ...
$ statistic: num 6.35 311.55 205.43 198.66 176.39 ...
$ df   : num 13.89 53.61 6.24 8.92 5.67 ...
$ p.value : num 1.88e-05 0.00 3.40e-13 0.00 8.11e-12 ...

Anexo 8. Selección de variables con FAMD

Dimension 1		
Link between the variable and the continuous variables (R-square)		
==========		
	correlation	p.value
ESTU_INSE_INDIVIDUAL	0.8535525	0.00E+00
ESTU_NSE_ESTABLECIMIENTO	0.5201991	0.00E+00
PERIODO	0.134068	1.96E-266
ESTU_EDAD	-0.1662157	0.00E+00
Link between the variable and the categorical variable (1-way anova)		
==========		
	R2	p.value
ESTU_TIPODOCUMENTO	0.02672750900	0.000000E+00
ESTU_MCPIO_RESIDE	0.46883290900	0.000000E+00
FAMI_EDUCACIONPADRE	0.25075520500	0.000000E+00
FAMI_EDUCACIONMADRE	0.27291681900	0.000000E+00
FAMI_ESTRATOVIVIENDA	0.25745109600	0.000000E+00
FAMI_CUARTOSHOGAR	0.03706640500	0.000000E+00
FAMI_TIENEINTERNET	0.37611045800	0.000000E+00
FAMI_TIENECOMPUTADOR	0.31430837200	0.000000E+00
FAMI_TIENELAVADORA	0.24558801200	0.000000E+00
FAMI_TIENEHORNOMICROOGAS	0.16979883900	0.000000E+00
FAMI_TIENESERVICIOTV	0.16533459900	0.000000E+00
FAMI_TIENEAUTOMOVIL	0.11011777300	0.000000E+00
FAMI_TIENECONSOLAVIDEOJUEGOS	0.11028328100	0.000000E+00
FAMI_NUMLIBROS	0.09815262200	0.000000E+00
FAMI_COMELECHEDERIVADOS	0.15858593900	0.000000E+00
FAMI_COMECARNEPESCADOHUEVO	0.12609054800	0.000000E+00
FAMI_COMECEREALFRUTOSLEGUMBRE	0.08214248700	0.000000E+00
FAMI_TRABAJOLABORPADRE	0.26350554300	0.000000E+00
FAMI_TRABAJOLABORMADRE	0.20339410900	0.000000E+00
ESTU_DEDICACIONINTERNET	0.18859794600	0.000000E+00
ESTU_HORASSEMANATRABAJA	0.04273713000	0.000000E+00

COLE_NOMBRE_ESTABLECIMIENTO	0.36885783600	0.000000E+00
COLE_NOMBRE_SEDE	0.35017045200	0.000000E+00
COLE_JORNADA	0.04013996500	0.000000E+00
COLE_MCPIO_UBICACION	0.47762867900	0.000000E+00
ESTU_MCPIO_PRESENTACION	0.41604515500	0.000000E+00
ESTU_DEPTO_PRESENTACION	0.02789304800	0.000000E+00
ESTU_NSE_INDIVIDUAL	0.66108458600	0.000000E+00
ESTU_DESEMPEÑO	0.05868916200	0.000000E+00
COLE_REGION	0.37224900400	0.000000E+00
ESTU_TIPOREMUNERACION	0.01948745600	8.885686E-286
FAMI_SITUACIONECONOMICA	0.01452926500	8.295075E-214
ESTU_DEDICACIONLECTURADIARIA	0.00918839200	1.268136E-132
FAMI_TIENEMOTOCICLETA	0.00642888100	4.819344E-96
FAMI_PERSONASHOGAR	0.00617005000	1.634092E-88
COLE_BILINGUE	0.00237015500	1.780172E-36
ESTU_GENERO	0.00187686900	3.144200E-29
COLE_AREA_UBICACION	0.00186047000	5.477610E-29
COLE_CARACTER	0.00146098900	4.132940E-21

Dimension 2		
Link between the variable and the continuous variables (R-square)		
==========		
	correlation	p.value
ESTU_EDAD	0.02587057	2.09E-11
PERIODO	-0.02435178	2.86E-10
ESTU_NSE_ESTABLECIMIENTO	-0.04487732	3.04E-31
ESTU_INSE_INDIVIDUAL	-0.06523085	3.90E-64
Link between the variable and the categorical variable (1-way anova)		
==========		
	R2	p.value
COLE_MCPIO_UBICACION	0.955486756	0.00E+00
ESTU_MCPIO_RESIDE	0.94917767	0.00E+00
ESTU_MCPIO_PRESENTACION	0.93394438	0.00E+00
COLE_NOMBRE_ESTABLECIMIENTO	0.887312799	0.00E+00

COLE_NOMBRE_SEDE	0.866674066	0.00E+00
COLE_REGION	0.846413582	0.00E+00
COLE_JORNADA	0.110642399	0.00E+00
COLE_AREA_UBICACION	0.028752128	0.00E+00
COLE_CARACTER	0.026930897	0.00E+00
FAMI_TRABAJOLABORMADRE	0.009248293	1.36E-125
ESTU_DEPTO_PRESENTACION	0.00798608	1.08E-99
FAMI_EDUCACIONMADRE	0.006415711	1.24E-85
ESTU_NSE_INDIVIDUAL	0.005717448	5.22E-83
FAMI_TIENESERVICIOTV	0.005109331	1.14E-76
FAMI_EDUCACIONPADRE	0.004982491	3.84E-65
FAMI_TRABAJOLABORPADRE	0.004652158	1.05E-59
FAMI_TIENEAUTOMOVIL	0.003268476	1.16E-49
COLE_BILINGUE	0.002195906	6.45E-34
FAMI_COMELECHEDERIVADOS	0.002146602	5.00E-31
FAMI_COMECARNEPESCADOHUEVO	0.002017416	3.72E-29
FAMI_NUMLIBROS	0.001901432	1.78E-27
ESTU_TIPODOCUMENTO	0.001520233	9.60E-20
FAMI_ESTRATOVIVIENDA	0.001349172	2.37E-17
FAMI_COMECEREALFRUTOSLEGUMBRE	0.001063585	2.20E-15
ESTU_HORASSEMANATRABAJA	0.000780653	1.16E-10
ESTU_DESEMPEÑO	0.000730577	1.30E-10
FAMI_CUARTOSHOGAR	0.000646943	3.08E-08
FAMI_TIENEHORNOMICROOGAS	0.000643834	4.99E-11
COLE_SEDE_PRINCIPAL	0.000451355	3.76E-08
FAMI_TIENECONSOLAVIDEOJUEGOS	0.000296879	8.13E-06
ESTU_GENERO	0.000293064	9.29E-06
ESTU_TIPOREMUNERACION	0.000240856	1.06E-03
FAMI_TIENEMOTOCICLETA	0.000141477	2.07E-03

"$var" "Results for the variables"					
$coord					
	Dim.1	**Dim.2**	**Dim.3**	**Dim.4**	**Dim.5**
PERIODO	1.80E-02	5.93E-04	6.78E-04	5.92E-09	1.77E-03
ESTU_INSE_INDIVIDUAL	7.29E-01	4.26E-03	8.33E-03	4.74E-04	3.05E-02

ESTU_NSE_ESTABLECIMIENTO	2.71E-01	2.01E-03	9.51E-05	1.65E-04	6.37E-04
ESTU_EDAD	2.76E-02	6.69E-04	5.42E-03	1.63E-03	1.07E-02
ESTU_TIPODOCUMENTO	2.67E-02	1.52E-03	5.02E-03	2.29E-03	4.86E-03
ESTU_GENERO	1.88E-03	2.93E-04	2.23E-04	7.60E-05	4.97E-05
ESTU_MCPIO_RESIDE	4.69E-01	9.49E-01	8.85E-01	8.73E-01	8.51E-01
FAMI_EDUCACIONPADRE	2.51E-01	4.98E-03	3.50E-03	4.10E-03	2.35E-02
FAMI_EDUCACIONMADRE	2.73E-01	6.42E-03	4.74E-03	2.98E-03	3.40E-02
FAMI_ESTRATOVIVIENDA	2.57E-01	1.35E-03	3.97E-03	1.55E-03	4.83E-03
FAMI_PERSONASHOGAR	6.17E-03	1.36E-04	1.42E-03	2.36E-03	1.85E-03
FAMI_CUARTOSHOGAR	3.71E-02	6.47E-04	4.70E-03	1.93E-03	9.17E-04
FAMI_TIENEINTERNET	3.76E-01	9.78E-06	1.50E-03	9.30E-05	6.66E-03
FAMI_TIENECOMPUTADOR	3.14E-01	3.42E-05	2.88E-03	5.55E-05	6.90E-03
FAMI_TIENELAVADORA	2.46E-01	8.23E-06	1.01E-03	2.27E-04	2.08E-03
FAMI_TIENEHORNOMICROOGAS	1.70E-01	6.44E-04	2.08E-04	3.03E-04	2.94E-03
FAMI_TIENESERVICIOTV	1.65E-01	5.11E-03	3.32E-03	2.33E-04	9.05E-03
FAMI_TIENEAUTOMOVIL	1.10E-01	3.27E-03	4.61E-03	2.63E-03	1.74E-03
FAMI_TIENEMOTOCICLETA	6.43E-03	1.41E-04	1.91E-04	7.68E-03	1.51E-02
FAMI_TIENECONSOLAVIDEOJUEGOS	1.10E-01	2.97E-04	2.76E-05	4.98E-04	2.02E-03
FAMI_NUMLIBROS	9.82E-02	1.90E-03	2.55E-03	2.90E-04	1.08E-03
FAMI_COMELECHEDERIVADOS	1.59E-01	2.15E-03	3.65E-03	1.78E-04	1.46E-03
FAMI_COMECARNEPESCADOHUEVO	1.26E-01	2.02E-03	1.14E-03	4.39E-03	1.60E-02
FAMI_COMECEREALFRUTOSLEGUMBRE	8.21E-02	1.06E-03	1.90E-03	1.26E-03	2.77E-03
FAMI_TRABAJOLABORPADRE	2.64E-01	4.65E-03	4.50E-03	1.08E-02	1.88E-02
FAMI_TRABAJOLABORMADRE	2.03E-01	9.25E-03	5.22E-03	6.91E-03	3.87E-02
FAMI_SITUACIONECONOMICA	1.45E-02	8.47E-05	5.14E-04	7.45E-04	1.15E-04
ESTU_DEDICACIONLECTURADIARIA	9.19E-03	1.37E-05	3.37E-04	7.37E-04	1.82E-03
ESTU_DEDICACIONINTERNET	1.89E-01	6.39E-05	2.26E-03	1.67E-03	6.93E-03
ESTU_HORASSEMANATRABAJA	4.27E-02	7.81E-04	8.86E-03	2.11E-03	1.12E-03
ESTU_TIPOREMUNERACION	1.95E-02	2.41E-04	6.43E-03	1.06E-03	7.61E-04
COLE_NOMBRE_ESTABLECIMIENTO	3.69E-01	8.87E-01	8.93E-01	7.98E-01	7.69E-01
COLE_BILINGUE	2.37E-03	2.20E-03	7.27E-04	6.92E-05	1.60E-03
COLE_CARACTER	1.46E-03	2.69E-02	5.67E-02	5.89E-02	5.79E-03
COLE_NOMBRE_SEDE	3.50E-01	8.67E-01	8.97E-01	8.26E-01	7.96E-01
COLE_SEDE_PRINCIPAL	9.79E-06	4.51E-04	8.85E-03	2.33E-03	7.94E-03
COLE_AREA_UBICACION	1.86E-03	2.88E-02	4.50E-02	2.63E-03	4.46E-02
COLE_JORNADA	4.01E-02	1.11E-01	2.09E-02	8.62E-03	2.13E-02
COLE_MCPIO_UBICACION	4.78E-01	9.55E-01	9.26E-01	8.82E-01	8.67E-01
ESTU_MCPIO_PRESENTACION	4.16E-01	9.34E-01	8.45E-01	8.35E-01	7.92E-01

ESTU_DEPTO_PRESENTACION	2.79E-02	7.99E-03	2.55E-01	3.62E-01	1.79E-01
ESTU_NSE_INDIVIDUAL	6.61E-01	5.72E-03	7.91E-03	6.38E-04	2.94E-02
COLE_REGION	3.72E-01	8.46E-01	5.04E-01	6.02E-01	6.66E-01
$contrib					
	Dim.1	**Dim.2**	**Dim.3**	**Dim.4**	**Dim.5**
PERIODO	0.2307135	1.04E-02	1.25E-02	1.12E-07	3.35E-02
ESTU_INSE_INDIVIDUAL	9.3515469	7.50E-02	1.53E-01	8.93E-03	5.77E-01
ESTU_NSE_ESTABLECIMIENTO	3.4734585	3.55E-02	1.75E-03	3.10E-03	1.21E-02
ESTU_EDAD	0.3546232	1.18E-02	9.97E-02	3.08E-02	2.02E-01
ESTU_TIPODOCUMENTO	0.343069	2.68E-02	9.24E-02	4.32E-02	9.21E-02
ESTU_GENERO	0.0240911	5.16E-03	4.10E-03	1.43E-03	9.41E-04
ESTU_MCPIO_RESIDE	6.0178462	1.67E+01	1.63E+01	1.64E+01	1.61E+01
FAMI_EDUCACIONPADRE	3.218644	8.78E-02	6.44E-02	7.72E-02	4.45E-01
FAMI_EDUCACIONMADRE	3.5031061	1.13E-01	8.73E-02	5.62E-02	6.44E-01
FAMI_ESTRATOVIVIENDA	3.3045912	2.38E-02	7.31E-02	2.92E-02	9.15E-02
FAMI_PERSONASHOGAR	0.0791975	2.40E-03	2.61E-02	4.45E-02	3.50E-02
FAMI_CUARTOSHOGAR	0.475777	1.14E-02	8.65E-02	3.63E-02	1.74E-02
FAMI_TIENEINTERNET	4.8276792	1.72E-04	2.76E-02	1.75E-03	1.26E-01
FAMI_TIENECOMPUTADOR	4.0343999	6.03E-04	5.31E-02	1.04E-03	1.31E-01
FAMI_TIENELAVADORA	3.152319	1.45E-04	1.86E-02	4.27E-03	3.93E-02
FAMI_TIENEHORNOMICROOGAS	2.1795042	1.13E-02	3.83E-03	5.70E-03	5.57E-02
FAMI_TIENESERVICIOTV	2.1222021	9.00E-02	6.11E-02	4.39E-03	1.71E-01
FAMI_TIENEAUTOMOVIL	1.4134499	5.76E-02	8.48E-02	4.96E-02	3.30E-02
FAMI_TIENEMOTOCICLETA	0.0825198	2.49E-03	3.52E-03	1.45E-01	2.85E-01
FAMI_TIENECONSOLAVIDEOJUEGOS	1.4155743	5.23E-03	5.08E-04	9.37E-03	3.82E-02
FAMI_NUMLIBROS	1.2598676	3.35E-02	4.69E-02	5.46E-03	2.04E-02
FAMI_COMELECHEDERIVADOS	2.0355776	3.78E-02	6.72E-02	3.36E-03	2.76E-02
FAMI_COMECARNEPESCADOHUEVO	1.6184732	3.55E-02	2.09E-02	8.27E-02	3.04E-01
FAMI_COMECEREALFRUTOSLEGUMBRE	1.0543647	1.87E-02	3.49E-02	2.37E-02	5.24E-02
FAMI_TRABAJOLABORPADRE	3.3823049	8.20E-02	8.29E-02	2.04E-01	3.56E-01
FAMI_TRABAJOLABORMADRE	2.6107264	1.63E-01	9.60E-02	1.30E-01	7.33E-01
FAMI_SITUACIONECONOMICA	0.1864948	1.49E-03	9.45E-03	1.40E-02	2.18E-03
ESTU_DEDICACIONLECTURADIARIA	0.1179404	2.41E-04	6.20E-03	1.39E-02	3.45E-02
ESTU_DEDICACIONINTERNET	2.4208058	1.13E-03	4.16E-02	3.15E-02	1.31E-01
ESTU_HORASSEMANATRABAJA	0.5485653	1.38E-02	1.63E-01	3.96E-02	2.13E-02
ESTU_TIPOREMUNERACION	0.2501371	4.24E-03	1.18E-01	1.99E-02	1.44E-02
COLE_NOMBRE_ESTABLECIMIENTO	4.734586	1.56E+01	1.64E+01	1.50E+01	1.46E+01

COLE_BILINGUE	0.0304228	3.87E-02	1.34E-02	1.30E-03	3.03E-02
COLE_CARACTER	0.018753	4.74E-01	1.04E+00	1.11E+00	1.10E-01
COLE_NOMBRE_SEDE	4.4947184	1.53E+01	1.65E+01	1.56E+01	1.51E+01
COLE_SEDE_PRINCIPAL	0.0001257	7.95E-03	1.63E-01	4.38E-02	1.50E-01
COLE_AREA_UBICACION	0.0238806	5.07E-01	8.27E-01	4.95E-02	8.44E-01
COLE_JORNADA	0.5152286	1.95E+00	3.84E-01	1.62E-01	4.03E-01
COLE_MCPIO_UBICACION	6.1307469	1.68E+01	1.70E+01	1.66E+01	1.64E+01
ESTU_MCPIO_PRESENTACION	5.340273	1.65E+01	1.55E+01	1.57E+01	1.50E+01
ESTU_DEPTO_PRESENTACION	0.3580296	1.41E-01	4.70E+00	6.81E+00	3.39E+00
ESTU_NSE_INDIVIDUAL	8.4855505	1.01E-01	1.46E-01	1.20E-02	5.57E-01
COLE_REGION	4.7781143	1.49E+01	9.27E+00	1.13E+01	1.26E+01
$cos2					
	Dim.1	**Dim.2**	**Dim.3**	**Dim.4**	**Dim.5**
PERIODO	3.23E-04	3.52E-07	4.60E-07	3.51E-17	3.13E-06
ESTU_INSE_INDIVIDUAL	5.31E-01	1.81E-05	6.95E-05	2.25E-07	9.30E-04
ESTU_NSE_ESTABLECIMIENTO	7.32E-02	4.06E-06	9.05E-09	2.71E-08	4.05E-07
ESTU_EDAD	7.63E-04	4.48E-07	2.94E-05	2.67E-06	1.14E-04
ESTU_TIPODOCUMENTO	1.19E-04	3.85E-07	4.21E-06	8.76E-07	3.94E-06
ESTU_GENERO	3.52E-06	8.59E-08	4.97E-08	5.77E-09	2.47E-09
ESTU_MCPIO_RESIDE	4.23E-03	1.73E-02	1.51E-02	1.47E-02	1.39E-02
FAMI_EDUCACIONPADRE	5.72E-03	2.26E-06	1.11E-06	1.53E-06	5.03E-05
FAMI_EDUCACIONMADRE	6.77E-03	3.74E-06	2.05E-06	8.09E-07	1.05E-04
FAMI_ESTRATOVIVIENDA	1.10E-02	3.03E-07	2.63E-06	4.00E-07	3.89E-06
FAMI_PERSONASHOGAR	9.52E-06	4.62E-09	5.05E-07	1.40E-06	8.52E-07
FAMI_CUARTOSHOGAR	2.75E-04	8.37E-08	4.42E-06	7.42E-07	1.68E-07
FAMI_TIENEINTERNET	1.41E-01	9.56E-11	2.25E-06	8.65E-09	4.44E-05
FAMI_TIENECOMPUTADOR	9.88E-02	1.17E-09	8.32E-06	3.08E-09	4.77E-05
FAMI_TIENELAVADORA	6.03E-02	6.78E-11	1.02E-06	5.14E-08	4.31E-06
FAMI_TIENEHORNOMICROOGAS	2.88E-02	4.15E-07	4.33E-08	9.17E-08	8.65E-06
FAMI_TIENESERVICIOTV	2.73E-02	2.61E-05	1.10E-05	5.44E-08	8.19E-05
FAMI_TIENEAUTOMOVIL	1.21E-02	1.07E-05	2.12E-05	6.93E-06	3.03E-06
FAMI_TIENEMOTOCICLETA	4.13E-05	2.00E-08	3.67E-08	5.90E-05	2.27E-04
FAMI_TIENECONSOLAVIDEOJUEGOS	1.22E-02	8.81E-08	7.63E-10	2.48E-07	4.06E-06
FAMI_NUMLIBROS	3.21E-03	1.21E-06	2.16E-06	2.81E-08	3.88E-07
FAMI_COMELECHEDERIVADOS	8.38E-03	1.54E-06	4.45E-06	1.06E-08	7.06E-07
FAMI_COMECARNEPESCADOHUEVO	5.30E-03	1.36E-06	4.30E-07	6.42E-06	8.58E-05
FAMI_COMECEREALFRUTOSLEGUMBRE	2.25E-03	3.77E-07	1.20E-06	5.27E-07	2.55E-06

FAMI_TRABAJOLABORPADRE	5.79E-03	1.80E-06	1.69E-06	9.77E-06	2.94E-05
FAMI_TRABAJOLABORMADRE	3.45E-03	7.13E-06	2.27E-06	3.98E-06	1.25E-04
FAMI_SITUACIONECONOMICA	1.06E-04	3.58E-09	1.32E-07	2.77E-07	6.61E-09
ESTU_DEDICACIONLECTURADIARIA	2.11E-05	4.66E-11	2.84E-08	1.36E-07	8.28E-07
ESTU_DEDICACIONINTERNET	8.89E-03	1.02E-09	1.28E-06	7.00E-07	1.20E-05
ESTU_HORASSEMANATRABAJA	4.57E-04	1.52E-07	1.96E-05	1.11E-06	3.16E-07
ESTU_TIPOREMUNERACION	1.27E-04	1.93E-08	1.38E-05	3.73E-07	1.93E-07
COLE_NOMBRE_ESTABLECIMIENTO	2.62E-03	1.51E-02	1.54E-02	1.22E-02	1.14E-02
COLE_BILINGUE	5.62E-06	4.82E-06	5.29E-07	4.79E-09	2.56E-06
COLE_CARACTER	7.11E-07	2.42E-04	1.07E-03	1.16E-03	1.12E-05
COLE_NOMBRE_SEDE	2.36E-03	1.44E-02	1.55E-02	1.31E-02	1.22E-02
COLE_SEDE_PRINCIPAL	9.58E-11	2.04E-07	7.84E-05	5.42E-06	6.30E-05
COLE_AREA_UBICACION	3.46E-06	8.27E-04	2.02E-03	6.91E-06	1.99E-03
COLE_JORNADA	3.22E-04	2.45E-03	8.72E-05	1.49E-05	9.07E-05
COLE_MCPIO_UBICACION	4.39E-03	1.76E-02	1.65E-02	1.50E-02	1.44E-02
ESTU_MCPIO_PRESENTACION	3.33E-03	1.68E-02	1.37E-02	1.34E-02	1.21E-02
ESTU_DEPTO_PRESENTACION	3.54E-05	2.90E-06	2.97E-03	5.95E-03	1.46E-03
ESTU_NSE_INDIVIDUAL	1.46E-01	1.09E-05	2.09E-05	1.36E-07	2.88E-04
COLE_REGION	1.07E-02	5.51E-02	1.95E-02	2.79E-02	3.41E-02
$coord.sup					
	Dim.1	**Dim.2**	**Dim.3**	**Dim.4**	**Dim.5**
ESTU_DESEMPEÑO	0.0586892	0.000731	0.0025	0.002434	0.000661
$cos2.sup					
	Dim.1	**Dim.2**	**Dim.3**	**Dim.4**	**Dim.5**
ESTU_DESEMPEÑO	0.0011481	1.78E-07	2.09E-06	1.97E-06	1.46E-07

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	7.790709	1.863806	1.863806
Dim.2	5.676286	1.357963	3.221769
Dim.3	5.436093	1.300501	4.52227
Dim.4	5.310757	1.270516	5.792786
Dim.5	5.28053	1.263285	7.05607

Anexo 9. Selección de variables con Boruta (Random Forest)

```
df <- read.csv ("D:/AppFiles/Universidad/4. Saber11/Data/SB11_ImputeData_OCt17.csv", encoding = "UTF-8", header =
TRUE, stringsAsFactors = T)
library (Boruta)
set.seed(123)
boruta.train <- Boruta (ESTU_DESEMPEÑO~., data = df, doTrace = 2)

final.boruta <- TentativeRoughFix(boruta.train)
print(final.boruta)

Boruta performed 99 iterations in 19.0479 hours.
Tentatives roughfixed over the last 99 iterations.
 38 attributes confirmed important: COLE_AREA_UBICACION, COLE_CARACTER, COLE_JORNADA,
COLE_MCPIO_UBICACION, COLE_NOMBRE_ESTABLECIMIENTO and
33 more;
 5 attributes confirmed unimportant: COLE_BILINGUE, COLE_SEDE_PRINCIPAL, ESTU_DEDICACIONINTERNET,
FAMI_COMECARNEPESCADOHUEVO,
FAMI_PERSONASHOGAR;
```

Variable	meanImp	medianImp	minImp	maxImp	normHits	decision
FAMI_PERSONASHOGAR	0.8482379	0.8409162	-0.1905244	20.280.656	0	Rejected
COLE_BILINGUE	0.4277407	0.4897597	-12.701.510	18.787.543	0	Rejected
COLE_SEDE_PRINCIPAL	-0.3756405	-0.5241672	-19.295.377	0.8714085	0	Rejected
ESTU_INSE_INDIVIDUAL	546.575.102	548.108.373	486.150.049	596.349.935	1,0000000	Confirmed
ESTU_EDAD	466.505.575	465.827.998	427.438.369	508.949.835	1,0000000	Confirmed
ESTU_GENERO	320.005.498	319.811.885	285.303.248	345.484.684	1,0000000	Confirmed
FAMI_NUMLIBROS	310.987.275	310.361.182	284.638.557	346.735.790	1,0000000	Confirmed
FAMI_EDUCACIONMADRE	295.271.045	294.097.052	269.543.248	329.940.767	1,0000000	Confirmed
COLE_MCPIO_UBICACION	294.158.834	294.262.802	264.741.325	327.614.033	1,0000000	Confirmed
ESTU_MCPIO_RESIDE	288.264.019	286.594.744	261.619.058	324.123.137	1,0000000	Confirmed
ESTU_TIPODOCUMENTO	281.644.776	280.197.982	258.511.861	309.835.300	1,0000000	Confirmed
ESTU_NSE_INDIVIDUAL	240.798.024	242.245.771	216.103.311	258.157.747	1,0000000	Confirmed
COLE_JORNADA	240.251.753	239.545.376	219.018.969	272.319.215	1,0000000	Confirmed
FAMI_EDUCACIONPADRE	239.822.184	238.519.170	215.269.706	271.468.130	1,0000000	Confirmed
COLE_NOMBRE_ESTABLECIMIENTO	211.006.353	209.373.465	183.825.336	238.530.696	1,0000000	Confirmed
FAMI_TIENECOMPUTADOR	203.999.931	204.113.276	180.039.771	227.305.242	1,0000000	Confirmed
ESTU_NSE_ESTABLECIMIENTO	200.393.592	200.981.740	176.899.801	228.663.843	1,0000000	Confirmed
COLE_NOMBRE_SEDE	194.006.360	191.663.445	167.049.179	216.699.521	1,0000000	Confirmed
ESTU_DEDICACIONLECTURADIARIA	170.242.522	170.997.004	136.444.912	200.700.188	1,0000000	Confirmed
FAMI_TIENEINTERNET	163.937.668	163.674.033	142.523.705	183.081.755	1,0000000	Confirmed
ESTU_HORASSEMANATRABAJA	158.877.885	158.377.832	134.895.330	182.617.744	1,0000000	Confirmed
FAMI_SITUACIONECONOMICA	156.746.365	156.173.195	127.437.965	187.850.490	1,0000000	Confirmed

FAMI_TIENEMOTOCICLETA	143.132.106	143.184.941	111.881.214	163.021.962	1,0000000	Confirmed
ESTU_MCPIO_PRESENTACION	137.035.026	138.188.531	114.240.862	162.568.119	1,0000000	Confirmed
FAMI_TIENECONSOLAVIDEOJUEGOS	128.144.559	128.667.110	102.413.117	151.162.729	1,0000000	Confirmed
FAMI_COMELECHEDERIVADOS	122.239.096	121.763.094	102.363.095	150.922.748	1,0000000	Confirmed
PERIODO	115.572.440	116.262.313	87.876.431	140.966.969	1,0000000	Confirmed
FAMI_TIENEAUTOMOVIL	105.755.748	105.734.726	78.759.458	129.423.395	1,0000000	Confirmed
COLE_AREA_UBICACION	103.890.549	103.106.988	71.785.407	133.019.578	1,0000000	Confirmed
FAMI_TIENEHORNOMICROOGAS	98.067.626	98.761.694	69.785.360	125.317.412	1,0000000	Confirmed
FAMI_TIENESERVICIOTV	91.051.651	91.136.633	72.203.595	115.740.837	1,0000000	Confirmed
FAMI_TIENELAVADORA	87.879.973	86.892.121	67.946.208	110.713.915	1,0000000	Confirmed
COLE_REGION	87.665.875	86.507.517	65.522.659	110.170.793	1,0000000	Confirmed
ESTU_TIPOREMUNERACION	69.488.309	69.075.389	46.461.907	95.548.377	1,0000000	Confirmed
FAMI_ESTRATOVIVIENDA	55.477.895	55.649.088	25.156.309	86.010.044	1,0000000	Confirmed
FAMI_TRABAJOLABORPADRE	52.335.890	51.205.792	29.904.251	87.181.519	1,0000000	Confirmed
FAMI_TRABAJOLABORMADRE	47.674.007	47.627.335	23.035.377	70.645.514	0,959596	Confirmed
ESTU_DEPTO_PRESENTACION	47.253.668	46.803.002	27.330.694	74.520.436	1,0000000	Confirmed
FAMI_COMECEREALFRUTOSLEGUMBRE	46.505.599	47.218.584	24.063.102	66.082.007	0,989899	Confirmed
COLE_CARACTER	45.423.213	44.301.577	21.260.327	72.888.264	0,969697	Confirmed
FAMI_CUARTOSHOGAR	29.165.597	28.132.462	0.3229102	56.894.671	0,6767677	Confirmed
FAMI_COMECARNEPESCADOHUEVO	24.348.295	23.637.914	0.1644522	52.091.766	0,5555556	Rejected
ESTU_DEDICACIONINTERNET	12.320.651	10.658.058	-0.9758305	40.342.058	0,1313131	Rejected

Anexo 10. Selección de variables con Weka

=== Run information ===
Evaluator: weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search: weka.attributeSelection.BestFirst -D 1 -N 5
Relation: DataTable-weka.filters.unsupervised.attribute.Remove-R44-60
Instances: 67049
Attributes: 50

- ESTU_TIPODOCUMENTO
- ESTU_NACIONALIDAD
- ESTU_GENERO
- PERIODO
- ESTU_PAIS_RESIDE
- ESTU_TIENEETNIA
- ESTU_DEPTO_RESIDE
- ESTU_MCPIO_RESIDE
- FAMI_EDUCACIONPADRE
- FAMI_EDUCACIONMADRE
- FAMI_ESTRATOVIVIENDA
- FAMI_PERSONASHOGAR
- FAMI_CUARTOSHOGAR
- FAMI_TIENEINTERNET
- FAMI_TIENECOMPUTADOR
- FAMI_TIENELAVADORA
- FAMI_TIENEHORNOMICROOGAS
- FAMI_TIENESERVICIOTV
- FAMI_TIENEAUTOMOVIL
- FAMI_TIENEMOTOCICLETA
- FAMI_TIENECONSOLAVIDEOJUEGOS
- FAMI_NUMLIBROS
- FAMI_COMELECHEDERIVADOS
- FAMI_COMECARNEPESCADOHUEVO
- FAMI_COMECEREALFRUTOSLEGUMBRE
- FAMI_TRABAJOLABORPADRE
- FAMI_TRABAJOLABORMADRE
- FAMI_SITUACIONECONOMICA
- ESTU_DEDICACIONLECTURADIARIA
- ESTU_DEDICACIONINTERNET
- ESTU_HORASSEMANATRABAJA
- ESTU_TIPOREMUNERACION
- COLE_NOMBRE_ESTABLECIMIENTO
- COLE_BILINGUE
- COLE_CARACTER
- COLE_NOMBRE_SEDE
- COLE_SEDE_PRINCIPAL
- COLE_AREA_UBICACION
- COLE_JORNADA
- COLE_MCPIO_UBICACION
- ESTU_PRIVADO_LIBERTAD
- ESTU_MCPIO_PRESENTACION
- ESTU_DEPTO_PRESENTACION
- ESTU_INSE_INDIVIDUAL
- ESTU_NSE_INDIVIDUAL
- ESTU_NSE_ESTABLECIMIENTO
- ESTU_ESTADOINVESTIGACION

ESTU_EDAD
COLE_REGION
ESTU_DESEMPEÃ‘O
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 738.
Merit of best subset found: 0.048

Attribute Subset Evaluator (supervised, Class (nominal): 50 ESTU_DESEMPEÑO):
CFS Subset Evaluator
Including locally predictive attributes

Selected attributes: 1,3,9,10,15,22,23,29,31,39,45,46,47,48 : 14
ESTU_TIPODOCUMENTO
ESTU_GENERO
FAMI_EDUCACIONPADRE
FAMI_EDUCACIONMADRE
FAMI_TIENECOMPUTADOR
FAMI_NUMLIBROS
FAMI_COMELECHEDERIVADOS
ESTU_DEDICACIONLECTURADIARIA
ESTU_HORASSEMANATRABAJA
COLE_JORNADA
ESTU_NSE_INDIVIDUAL
ESTU_NSE_ESTABLECIMIENTO
ESTU_ESTADOINVESTIGACION
ESTU_EDAD

=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

number of folds (%) attribute
10(100 %) 1 ESTU_TIPODOCUMENTO
0(0 %) 2 ESTU_NACIONALIDAD
10(100 %) 3 ESTU_GENERO
0(0 %) 4 PERIODO
0(0 %) 5 ESTU_PAIS_RESIDE
0(0 %) 6 ESTU_TIENEETNIA
0(0 %) 7 ESTU_DEPTO_RESIDE
0(0 %) 8 ESTU_MCPIO_RESIDE
10(100 %) 9 FAMI_EDUCACIONPADRE
10(100 %) 10 FAMI_EDUCACIONMADRE
0(0 %) 11 FAMI_ESTRATOVIVIENDA
0(0 %) 12 FAMI_PERSONASHOGAR
0(0 %) 13 FAMI_CUARTOSHOGAR
0(0 %) 14 FAMI_TIENEINTERNET
10(100 %) 15 FAMI_TIENECOMPUTADOR
0(0 %) 16 FAMI_TIENELAVADORA

```
  0(  0 %)    17 FAMI_TIENEHORNOMICROOGAS
  0(  0 %)    18 FAMI_TIENESERVICIOTV
  0(  0 %)    19 FAMI_TIENEAUTOMOVIL
  0(  0 %)    20 FAMI_TIENEMOTOCICLETA
  0(  0 %)    21 FAMI_TIENECONSOLAVIDEOJUEGOS
 10(100 %)    22 FAMI_NUMLIBROS
 10(100 %)    23 FAMI_COMELECHEDERIVADOS
  0(  0 %)    24 FAMI_COMECARNEPESCADOHUEVO
  0(  0 %)    25 FAMI_COMECEREALFRUTOSLEGUMBRE
  0(  0 %)    26 FAMI_TRABAJOLABORPADRE
  0(  0 %)    27 FAMI_TRABAJOLABORMADRE
  0(  0 %)    28 FAMI_SITUACIONECONOMICA
 10(100 %)    29 ESTU_DEDICACIONLECTURADIARIA
  0(  0 %)    30 ESTU_DEDICACIONINTERNET
 10(100 %)    31 ESTU_HORASSEMANATRABAJA
  0(  0 %)    32 ESTU_TIPOREMUNERACION
  0(  0 %)    33 COLE_NOMBRE_ESTABLECIMIENTO
  0(  0 %)    34 COLE_BILINGUE
  0(  0 %)    35 COLE_CARACTER
  0(  0 %)    36 COLE_NOMBRE_SEDE
  0(  0 %)    37 COLE_SEDE_PRINCIPAL
  0(  0 %)    38 COLE_AREA_UBICACION
 10(100 %)    39 COLE_JORNADA
  0(  0 %)    40 COLE_MCPIO_UBICACION
  0(  0 %)    41 ESTU_PRIVADO_LIBERTAD
  0(  0 %)    42 ESTU_MCPIO_PRESENTACION
  0(  0 %)    43 ESTU_DEPTO_PRESENTACION
  4( 40 %)    44 ESTU_INSE_INDIVIDUAL
 10(100 %)    45 ESTU_NSE_INDIVIDUAL
 10(100 %)    46 ESTU_NSE_ESTABLECIMIENTO
 10(100 %)    47 ESTU_ESTADOINVESTIGACION
 10(100 %)    48 ESTU_EDAD
  0(  0 %)    49 COLE_REGION
```

```
=== Run information ===

Evaluator:    weka.attributeSelection.GainRatioAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     DataTable-weka.filters.unsupervised.attribute.Remove-R44-60
Instances:    67049

=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 50 ESTU_DESEMPEÃ‘O):
        Gain Ratio feature evaluator

Ranked attributes:
 0.056254   47 ESTU_ESTADOINVESTIGACION
 0.026837   41 ESTU_PRIVADO_LIBERTAD
 0.020696    1 ESTU_TIPODOCUMENTO
 0.01741    48 ESTU_EDAD
 0.016419   45 ESTU_NSE_INDIVIDUAL
```

0.014579 39 COLE_JORNADA
0.013036 15 FAMI_TIENECOMPUTADOR
0.012868 22 FAMI_NUMLIBROS
0.012411 44 ESTU_INSE_INDIVIDUAL
0.009781 46 ESTU_NSE_ESTABLECIMIENTO
0.009653 10 FAMI_EDUCACIONMADRE
0.008408 29 ESTU_DEDICACIONLECTURADIARIA
0.008195 9 FAMI_EDUCACIONPADRE
0.007819 14 FAMI_TIENEINTERNET
0.007218 23 FAMI_COMELECHEDERIVADOS
0.006649 3 ESTU_GENERO
0.006342 31 ESTU_HORASSEMANATRABAJA
0.006004 7 ESTU_DEPTO_RESIDE
0.005915 2 ESTU_NACIONALIDAD
0.005915 5 ESTU_PAIS_RESIDE
0.004852 11 FAMI_ESTRATOVIVIENDA
0.004499 16 FAMI_TIENELAVADORA
0.004319 30 ESTU_DEDICACIONINTERNET
0.004036 20 FAMI_TIENEMOTOCICLETA
0.003924 28 FAMI_SITUACIONECONOMICA
0.003888 33 COLE_NOMBRE_ESTABLECIMIENTO
0.00378 43 ESTU_DEPTO_PRESENTACION
0.003723 36 COLE_NOMBRE_SEDE
0.003603 32 ESTU_TIPOREMUNERACION
0.003419 40 COLE_MCPIO_UBICACION
0.00338 26 FAMI_TRABAJOLABORPADRE
0.00336 27 FAMI_TRABAJOLABORMADRE
0.003083 8 ESTU_MCPIO_RESIDE
0.002838 24 FAMI_COMECARNEPESCADOHUEVO
0.002785 25 FAMI_COMECEREALFRUTOSLEGUMBRE
0.002405 19 FAMI_TIENEAUTOMOVIL
0.002228 42 ESTU_MCPIO_PRESENTACION
0.002018 49 COLE_REGION
0.001493 17 FAMI_TIENEHORNOMICROOGAS
0.001457 34 COLE_BILINGUE
0.001307 21 FAMI_TIENECONSOLAVIDEOJUEGOS
0.00123 38 COLE_AREA_UBICACION
0.001184 12 FAMI_PERSONASHOGAR
0.000984 35 COLE_CARACTER
0.000912 13 FAMI_CUARTOSHOGAR
0.000871 37 COLE_SEDE_PRINCIPAL
0.000803 4 PERIODO
0.000728 18 FAMI_TIENESERVICIOTV
0.000578 6 ESTU_TIENEETNIA

=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

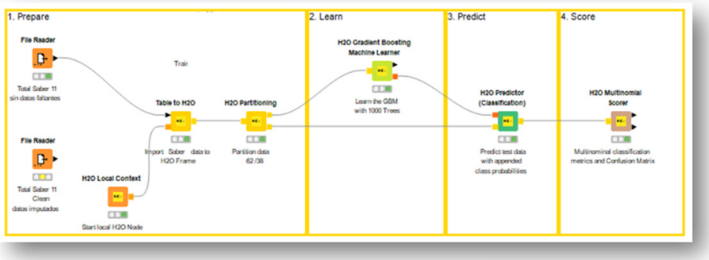
average merit average rank attribute
0.057 +- 0.004 1 +- 0 47 ESTU_ESTADOINVESTIGACION
0.027 +- 0.001 2 +- 0 41 ESTU_PRIVADO_LIBERTAD
0.021 +- 0 3 +- 0 1 ESTU_TIPODOCUMENTO
0.017 +- 0 4 +- 0 48 ESTU_EDAD
0.016 +- 0 5 +- 0 45 ESTU_NSE_INDIVIDUAL
0.015 +- 0 6 +- 0 39 COLE_JORNADA
0.013 +- 0.001 7.6 +- 0.92 44 ESTU_INSE_INDIVIDUAL

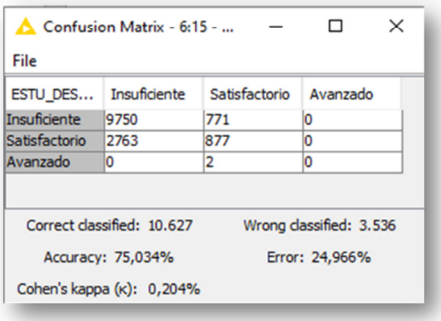
0.013 +- 0 7.9 +- 0.7 15 FAMI_TIENECOMPUTADOR
0.013 +- 0 8.5 +- 0.5 22 FAMI_NUMLIBROS
0.01 +- 0 10.4 +- 0.49 46 ESTU_NSE_ESTABLECIMIENTO
0.01 +- 0 10.6 +- 0.49 10 FAMI_EDUCACIONMADRE
0.008 +- 0 12.1 +- 0.3 29 ESTU_DEDICACIONLECTURADIARIA
0.008 +- 0 12.9 +- 0.3 9 FAMI_EDUCACIONPADRE
0.008 +- 0 14 +- 0 14 FAMI_TIENEINTERNET
0.007 +- 0 15.2 +- 0.6 23 FAMI_COMELECHEDERIVADOS
0.007 +- 0 16.6 +- 0.66 3 ESTU_GENERO
0.006 +- 0 18 +- 1.18 31 ESTU_HORASSEMANATRABAJA
0.006 +- 0.001 18.2 +- 1.78 2 ESTU_NACIONALIDAD
0.006 +- 0.001 18.7 +- 1.9 7 ESTU_DEPTO_RESIDE
0.006 +- 0.001 19 +- 1.41 5 ESTU_PAIS_RESIDE
0.005 +- 0 20.5 +- 0.81 11 FAMI_ESTRATOVIVIENDA
0.005 +- 0 21.9 +- 0.7 16 FAMI_TIENELAVADORA
0.004 +- 0 23.1 +- 0.3 30 ESTU_DEDICACIONINTERNET
0.004 +- 0 24.7 +- 1.42 20 FAMI_TIENEMOTOCICLETA
0.004 +- 0 25.4 +- 1.02 33 COLE_NOMBRE_ESTABLECIMIENTO
0.004 +- 0 25.8 +- 0.98 28 FAMI_SITUACIONECONOMICA
0.004 +- 0 27 +- 1.67 43 ESTU_DEPTO_PRESENTACION
0.004 +- 0 27.4 +- 0.8 36 COLE_NOMBRE_SEDE
0.004 +- 0 28.7 +- 1.42 32 ESTU_TIPOREMUNERACION
0.003 +- 0 30.2 +- 0.6 40 COLE_MCPIO_UBICACION
0.003 +- 0 31 +- 0.77 26 FAMI_TRABAJOLABORPADRE
0.003 +- 0 31.6 +- 0.66 27 FAMI_TRABAJOLABORMADRE
0.003 +- 0 33 +- 0 8 ESTU_MCPIO_RESIDE
0.003 +- 0 34.3 +- 0.46 24 FAMI_COMECARNEPESCADOHUEVO
0.003 +- 0 34.7 +- 0.46 25 FAMI_COMECEREALFRUTOSLEGUMBRE
0.002 +- 0 36.1 +- 0.3 19 FAMI_TIENEAUTOMOVIL
0.002 +- 0 36.9 +- 0.3 42 ESTU_MCPIO_PRESENTACION
0.002 +- 0 38.1 +- 0.3 49 COLE_REGION
0.001 +- 0 39.7 +- 0.64 17 FAMI_TIENEHORNOMICROOGAS
0.002 +- 0 40.4 +- 2.06 34 COLE_BILINGUE
0.001 +- 0 40.9 +- 0.7 21 FAMI_TIENECONSOLAVIDEOJUEGOS
0.001 +- 0 41.8 +- 1.08 38 COLE_AREA_UBICACION
0.001 +- 0 42.2 +- 0.75 12 FAMI_PERSONASHOGAR
0.001 +- 0 44 +- 0.45 35 COLE_CARACTER
0.001 +- 0 45.3 +- 0.46 13 FAMI_CUARTOSHOGAR
0.001 +- 0 46.1 +- 1.22 37 COLE_SEDE_PRINCIPAL
0.001 +- 0 46.9 +- 0.7 4 PERIODO
0.001 +- 0 48 +- 0.63 18 FAMI_TIENESERVICIOTV
0.001 +- 0 48.6 +- 0.8 6 ESTU_TIENEETNIA

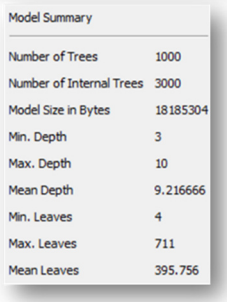
Anexo 11. Gradient Boosting Machine (GBM)

Escenario sin datos faltantes (47,147 Registros)

Row ID	Relative Importance	Scaled ...	Percent...
ESTU_MCPIO_PRESENTACION	4,808.014	1	0.103
FAMI_TRABAJOLABORPADRE	4,671.771	0.972	0.1
ESTU_MCPIO_RESIDE	3,790.503	0.788	0.081
FAMI_TRABAJOLABORMADRE	3,505.535	0.729	0.075
COLE_MCPIO_UBICACION	3,323.547	0.691	0.071
FAMI_EDUCACIONPADRE	3,142.959	0.654	0.067
FAMI_EDUCACIONMADRE	2,749.202	0.572	0.059
COLE_NOMBRE_ESTABLECIMIENTO	2,439.234	0.507	0.052
ESTU_EDAD	1,966.434	0.409	0.042
COLE_NOMBRE_SEDE	1,782.98	0.371	0.038
ESTU_DEDICACIONLECTURADIARIA	1,446.777	0.301	0.031
PERIODO	1,231.599	0.256	0.026
FAMI_NUMLIBROS	1,193.359	0.248	0.026
ESTU_INSE_INDIVIDUAL	1,002.438	0.208	0.021
ESTU_DEDICACIONINTERNET	984.008	0.205	0.021
ESTU_GENERO	962.647	0.2	0.021
COLE_JORNADA	867.05	0.18	0.019
FAMI_CUARTOSHOGAR	805.427	0.168	0.017
FAMI_COMELECHEDERIVADOS	744.119	0.155	0.016
FAMI_SITUACIONECONOMICA	659.044	0.137	0.014
ESTU_HORASSEMANATRABAJA	618.6	0.129	0.013
FAMI_COMECEREALFRUTOSLEGUMBRE	600.831	0.125	0.013
FAMI_PERSONASHOGAR	507.474	0.106	0.011
FAMI_ESTRATOVIVIENDA	474.885	0.099	0.01
FAMI_COMECARNEPESCADOHUEVO	421.564	0.088	0.009
ESTU_TIPOREMUNERACION	264.656	0.055	0.006
FAMI_TIENEMOTOCICLETA	226.677	0.047	0.005
COLE_REGION	202.686	0.042	0.004
ESTU_NSE_INDIVIDUAL	167.641	0.035	0.004
FAMI_TIENESERVICIOTV	143.529	0.03	0.003
FAMI_TIENECOMPUTADOR	126.159	0.026	0.003
FAMI_TIENEHORNOMICROOGAS	119.561	0.025	0.003
FAMI_TIENELAVADORA	106.771	0.022	0.002
FAMI_TIENEAUTOMOVIL	98.084	0.02	0.002
ESTU_TIPODOCUMENTO	87.06	0.018	0.002
FAMI_TIENEINTERNET	86.24	0.018	0.002
COLE_CARACTER	84.394	0.018	0.002
FAMI_TIENECONSOLAVIDEOJUEGOS	73.172	0.015	0.002
COLE_AREA_UBICACION	70.061	0.015	0.002
ESTU_NSE_ESTABLECIMIENTO	59.276	0.012	0.001
COLE_SEDE_PRINCIPAL	20.376	0.004	0
ESTU_DEPTO_PRESENTACION	3.009	0.001	0

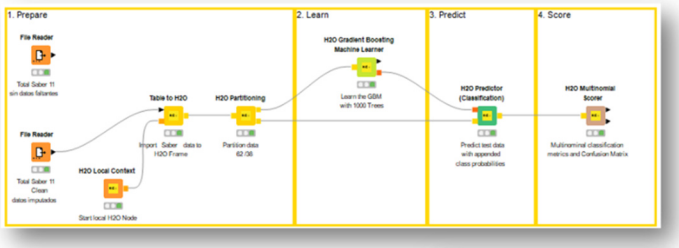


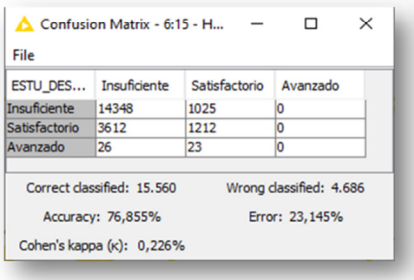


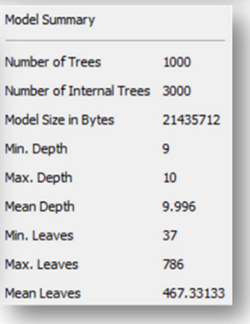


Escenario con datos imputados (67,049 Registros)

Row ID	Relative Importance	Scaled Importance	Percentage
FAMI_TRABAJOLABORPADRE	5,922.122	1	0.095
ESTU_MCPIO_PRESENTACION	5,902.76	0.997	0.094
ESTU_MCPIO_RESIDE	5,218.312	0.881	0.083
COLE_MCPIO_UBICACION	4,693.322	0.793	0.075
FAMI_TRABAJOLABORMADRE	4,386.753	0.741	0.07
FAMI_EDUCACIONPADRE	3,819.748	0.645	0.061
ESTU_EDAD	3,705.642	0.626	0.059
FAMI_EDUCACIONMADRE	3,667.847	0.619	0.059
COLE_NOMBRE_ESTABLECIMIENTO	3,082.469	0.521	0.049
COLE_NOMBRE_SEDE	2,449.09	0.414	0.039
ESTU_DEDICACIONLECTURADIARIA	2,202.141	0.372	0.035
PERIODO	1,644.888	0.278	0.026
COLE_JORNADA	1,635.543	0.276	0.026
FAMI_NUMLIBROS	1,606.893	0.271	0.026
ESTU_INSE_INDIVIDUAL	1,421.019	0.24	0.023
ESTU_GENERO	1,287.492	0.217	0.021
ESTU_DEDICACIONINTERNET	1,270.668	0.215	0.02
FAMI_COMELECHEDERIVADOS	990.913	0.167	0.016
FAMI_CUARTOSHOGAR	948.565	0.16	0.015
ESTU_HORASSEMANATRABAJA	921.901	0.156	0.015
FAMI_SITUACIONECONOMICA	857.794	0.145	0.014
FAMI_COMECEREALFRUTOSLEGUMBRE	774.186	0.131	0.012
FAMI_ESTRATOVIVIENDA	588.129	0.099	0.009
FAMI_COMECARNEPESCADOHUEVO	568.273	0.096	0.009
FAMI_PERSONASHOGAR	562.564	0.095	0.009
ESTU_TIPOREMUNERACION	336.403	0.057	0.005
FAMI_TIENEMOTOCICLETA	296.837	0.05	0.005
COLE_REGION	284.742	0.048	0.005
FAMI_TIENESERVICIOTV	207.587	0.035	0.003
ESTU_NSE_INDIVIDUAL	196.088	0.033	0.003
FAMI_TIENECOMPUTADOR	171.339	0.029	0.003
FAMI_TIENEHORNOMICROOGAS	157.106	0.027	0.003
ESTU_TIPODOCUMENTO	147.21	0.025	0.002
FAMI_TIENEAUTOMOVIL	130.901	0.022	0.002
FAMI_TIENECONSOLAVIDEOJUEGOS	111.846	0.019	0.002
COLE_CARACTER	108.453	0.018	0.002
FAMI_TIENELAVADORA	100.551	0.017	0.002
FAMI_TIENEINTERNET	89.453	0.015	0.001
COLE_AREA_UBICACION	72.726	0.012	0.001
ESTU_NSE_ESTABLECIMIENTO	60.329	0.01	0.001
COLE_SEDE_PRINCIPAL	21.049	0.004	0
ESTU_DEPTO_PRESENTACION	2.411	0	0



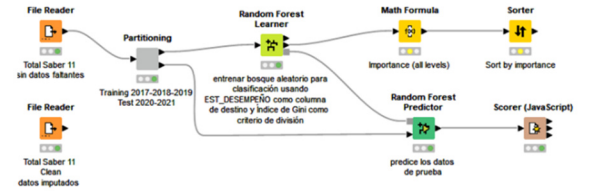




Anexo 12. Random Forest Learner

Escenario sin datos faltantes (47,147 Registros)

Row ID	#splits ...	#splits ...	#splits ...	#candi...	#candi...	#candi...	importance (all levels)
ESTU_EDAD	69	108	181	69	140	278	2.423
ESTU_INSE_INDIVIDUAL	73	108	161	83	153	280	2.16
FAMI_NUMLIBROS	49	99	139	86	157	288	1.683
COLE_JORNADA	37	73	152	83	138	293	1.494
FAMI_EDUCACIONMADRE	32	61	130	64	128	294	1.419
ESTU_DEDICACIONLECTURADIARIA	25	70	157	69	136	292	1.415
ESTU_NSE_INDIVIDUAL	39	50	86	65	140	289	1.255
COLE_MCPIO_UBICACION	31	60	66	57	145	289	1.186
FAMI_EDUCACIONPADRE	25	45	111	60	124	286	1.168
ESTU_MCPIO_RESIDE	21	41	45	63	129	305	0.799
FAMI_COMELECHEDERIVADOS	16	46	59	76	170	261	0.707
ESTU_GENERO	7	27	98	64	139	266	0.672
COLE_NOMBRE_SEDE	11	31	41	56	163	289	0.528
COLE_NOMBRE_ESTABLECIMIENTO	11	36	39	74	158	267	0.523
FAMI_SITUACIONECONOMICA	4	17	66	63	160	322	0.375
COLE_REGION	3	21	54	71	152	278	0.375
FAMI_TRABAJOLABORMADRE	9	7	44	85	147	294	0.303
FAMI_ESTRATOVIVIENDA	10	9	26	71	139	291	0.295
FAMI_TIENECOMPUTADOR	7	7	24	52	148	261	0.274
ESTU_MCPIO_PRESENTACION	3	13	30	77	151	274	0.235
FAMI_TRABAJOLABORPADRE	3	6	43	79	139	306	0.222
ESTU_TIPODOCUMENTO	2	10	28	71	141	282	0.198
ESTU_NSE_ESTABLECIMIENTO	3	5	34	78	137	278	0.197
FAMI_TIENEMOTOCICLETA	0	13	31	65	147	297	0.193
FAMI_TIENEINTERNET	5	7	14	71	140	288	0.169
FAMI_COMECEREALFRUTOSLEGUMBRE	0	6	24	68	146	296	0.122
ESTU_HORASSEMANATRABAJA	2	5	16	75	142	273	0.12
FAMI_COMECARNEPESCADOHUEVO	1	3	15	77	120	282	0.091
ESTU_DEDICACIONINTERNET	2	3	10	73	136	276	0.086
FAMI_CUARTOSHOGAR	0	3	16	77	117	280	0.083
PERIODO	0	2	11	80	141	281	0.053
ESTU_DEPTO_PRESENTACION	0	4	4	74	136	278	0.044
ESTU_TIPOREMUNERACION	0	0	12	63	149	295	0.041
FAMI_PERSONASHOGAR	0	0	11	89	152	292	0.038
FAMI_TIENECONSOLAVIDEOJUEGOS	0	1	7	66	138	320	0.029
FAMI_TIENESERVICIOTV	0	1	3	66	150	290	0.017
FAMI_TIENEAUTOMOVIL	0	1	2	57	126	257	0.016
COLE_CARACTER	0	0	4	82	151	288	0.014
FAMI_TIENELAVADORA	0	1	1	75	139	301	0.011
FAMI_TIENEHORNOMICROOGAS	0	0	2	83	140	287	0.007
COLE_SEDE_PRINCIPAL	0	0	1	71	148	273	0.004
COLE_AREA_UBICACION	0	0	1	72	148	283	0.004



Datos sin Missing Values 47.147

Test 2020-2021

Confusion Matrix

Rows Number : 18023	Avanzado (Predicted)	Insuficiente (Predicted)	Satisfactorio (Predicted)	
Avanzado (Actual)	0	4	2	0.00%
Insuficiente (Actual)	0	13538	224	98.37%
Satisfactorio (Actual)	0	3923	332	7.80%
	undefined	77.52%	59.50%	

Class Statistics

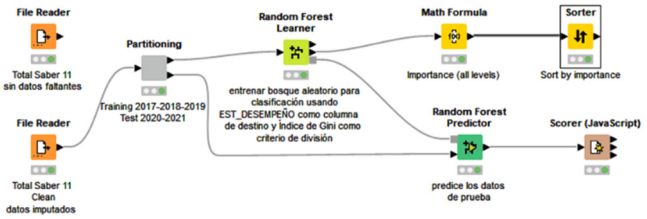
Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
Avanzado	0	0	18017	6	0.00%	0.00%	0.00%	100.00%	0.00%
Insuficiente	13538	3927	334	224	98.37%	77.52%	98.37%	7.84%	86.71%
Satisfactorio	332	226	13542	3923	7.80%	59.50%	7.80%	98.36%	13.80%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
76.96%	23.04%	0.088	13870	4153

Escenario con datos imputados (67,049 Registros)

Row ID	#splits ...	#splits ...	#splits ...	#candi...	#candi...	#candi...	imp...
ESTU_EDAD	61	110	195	61	137	294	2.466
COLE_JORNADA	64	100	155	71	135	259	2.241
ESTU_INSE_INDIVIDUAL	53	99	147	67	154	277	1.965
FAMI_NUMLIBROS	47	73	157	82	132	291	1.666
FAMI_EDUCACIONMADRE	40	60	113	66	137	299	1.422
ESTU_NSE_INDIVIDUAL	42	46	81	60	112	269	1.412
ESTU_DEDICACIONLECTURADIARIA	29	47	121	86	120	267	1.182
FAMI_EDUCACIONPADRE	27	41	98	70	135	285	1.033
COLE_MCPIO_UBICACION	13	75	85	56	167	266	1.001
ESTU_TIPODOCUMENTO	27	41	86	72	125	323	0.969
ESTU_MCPIO_RESIDE	18	45	59	72	151	277	0.761
FAMI_COMELECHEDERIVADOS	16	32	72	82	138	297	0.669
ESTU_GENERO	0	29	108	64	129	302	0.582
COLE_NOMBRE_ESTABLECIMIENTO	10	34	44	84	151	309	0.487
FAMI_TIENECOMPUTADOR	17	11	23	70	135	301	0.401
COLE_NOMBRE_SEDE	6	14	42	62	130	293	0.348
ESTU_HORASSEMANATRABAJA	3	26	30	68	144	315	0.32
FAMI_SITUACIONECONOMICA	0	12	51	88	139	280	0.268
FAMI_ESTRATOVIVIENDA	7	18	16	83	153	259	0.264
FAMI_TRABAJOLABORPADRE	2	10	38	70	151	262	0.24
COLE_REGION	0	8	42	81	121	274	0.219
FAMI_TRABAJOLABORMADRE	1	12	38	75	149	311	0.216
ESTU_DEDICACIONINTERNET	2	8	36	75	147	295	0.203
ESTU_MCPIO_PRESENTACION	2	10	23	69	139	262	0.189
FAMI_TIENEMOTOCICLETA	0	9	34	71	144	291	0.179
FAMI_TIENEINTERNET	7	9	7	87	138	279	0.171
ESTU_NSE_ESTABLECIMIENTO	4	4	17	68	129	282	0.15
FAMI_COMECEREALFRUTOSLEGUMBRE	1	6	17	66	146	271	0.119
PERIODO	0	3	15	69	133	298	0.073
FAMI_COMECARNEPESCADOHUEVO	0	2	13	70	142	276	0.061
ESTU_TIPOREMUNERACION	1	1	10	80	158	272	0.056
FAMI_TIENELAVADORA	0	3	4	63	136	283	0.036
FAMI_CUARTOSHOGAR	0	1	6	74	147	290	0.027
FAMI_TIENECONSOLAVIDEOJUEGOS	0	1	4	69	148	283	0.021
FAMI_PERSONASHOGAR	0	0	4	69	151	306	0.013
FAMI_TIENESERVICIOTV	0	0	3	65	179	334	0.009
FAMI_TIENEAUTOMOVIL	0	0	2	79	171	276	0.007
COLE_AREA_UBICACION	0	0	2	66	146	277	0.007
ESTU_DEPTO_PRESENTACION	0	0	1	67	148	264	0.004
FAMI_TIENEHORNOMICROOGAS	0	0	1	73	153	279	0.004
COLE_CARACTER	0	0	0	61	137	298	0
COLE_SEDE_PRINCIPAL	0	0	0	69	163	274	0



Confusion Matrix

Rows Number : 25228	Avanzado (Predicted)	Insuficiente (Predicted)	Satisfactorio (Predicted)	
Avanzado (Actual)	0	53	19	0.00%
Insuficiente (Actual)	0	19251	338	98.27%
Satisfactorio (Actual)	0	5002	565	10.15%
	undefined	79.20%	61.28%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
Avanzado	0	0	25156	72	0.00%	0.00%	0.00%	100.00%	0.00%
Insuficiente	19251	5055	584	338	98.27%	79.20%	98.27%	10.36%	87.71%
Satisfactorio	565	357	19304	5002	10.15%	61.28%	10.15%	98.18%	17.41%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
78.55%	21.45%	0.120	19816	5412