
# Análisis del Desempeño Académico del Examen de Estado para el Ingreso a la Educación Superior Aplicando Minería de Datos

**Vanessa Paola Blanco Villafañe**

Universidad Nacional de Colombia

Facultad de Ingeniería

Departamento de Ingeniería de Sistemas e industrial

Bogotá, Colombia

2015

# Análisis del Desempeño Académico del Examen de Estado para el Ingreso a la Educación Superior Aplicando Minería de Datos

**Vanessa Paola Blanco Villafañe**

Trabajo de investigación presentado como requisito para optar al título de:

**Magister en Ingeniería Sistemas y Computación**

Director:

Ing. Fabio Augusto González Osorio, PhD.

Profesor Titular

Universidad Nacional de Colombia

Facultad de Ingeniería

Departamento de Ingeniería de Sistemas e Industrial

Valledupar, Colombia

2015

*El aprendizaje es cualquier cambio que haga un sistema para adaptarse a su medio ambiente.*

*Herbert Simon.*

# Agradecimientos

Especial agradecimiento a la Universidad Nacional de Colombia y la Universidad Popular del Cesar por realizar estos procesos de aprendizaje desde las regiones. Al ICFES por permitir acceder a su repositorio de datos e incentivar a las propuestas de investigación en este campo.

A los profesores de la maestría, los cuales son de gran ejemplo por su formación y carácter en pro de la investigación, en especial al profesor Ph.D. Fabio González por compartir su tiempo y sus orientaciones.

A mi familia, en especial a mi madre e hijo, quienes con su amor y apoyo incondicional estuvieron siempre para dedicarme una sonrisa y un abrazo en los momentos apremiantes y de escasez de tiempo para con ellos.

A mis amigos, en especial a los que cursaron la maestría, sus observaciones siempre fueron un apoyo para el desarrollo del presente trabajo.

# Resumen

Este trabajo presenta un estudio de minería de datos en educación basado en un conjunto de datos de resultados del Examen de Estado para ingreso a la Educación Superior correspondiente al año 2012. El trabajo aplica técnicas de análisis de agrupamiento para construir un modelo descriptivo que permite entender de manera más clara la estructura de los datos. El documento presenta la descripción de los datos, los detalles de la metodología aplicada, basada en CRISP-DM, los modelos de agrupamiento usados, el diseño experimental y los resultados con su respectiva discusión.

**Palabras clave:** exámenes de estado, minería de datos en educación, análisis de conglomerados, CRISP-DM.

# Abstract

This document presents a educational data mining study based on a data set of results of the exam for admission to high education applied by the Colombian government corresponding to the year 2012. This work applies techniques for clustering analysis to build a descriptive model that allows to better understand the subjacent structure of data. The document presents the dataset description, the details of the applied methodology, based on CRISP-DM, the clustering models used, the experimental design with the corresponding results and discussion.

**Keywords:** state test scores, educational data mining, cluster analysis, CRISP-DM.

# Contenido

# Lista de figuras

**Pág.**

# Lista de tablas

**Pág.**

# Introducción

La minería de datos representa un avance computacional significativo en la obtención de información a partir de relaciones ocultas entre variables; esta disciplina tiene como objetivo la extracción de conocimiento útil de un alto volumen de datos en el cual inicialmente este conocimiento es desconocido, pero que al aplicar las técnicas de minería estas relaciones son descubiertas. La aplicación de las técnicas y herramientas de la minería de datos en los diferentes contextos educativos se conoce como minería de datos en educación (educational data mining) [45].

En Estados Unidos esta disciplina es considerada como la sexta tecnología de la información en la educación [8], y es usada para comprender y tener mejor conocimiento de los estudiantes, para evaluar su progreso y evaluar los entornos educativos en que aprenden. Ayuda a conocer cómo encontrar grupos de estudiantes con problemas similares, a identificar el éxito o el fracaso en las estrategias de enseñanza y genera un discernimiento más profundo del aprendizaje, entre otros aspectos [3, 4, 34, 39,42, 47].

En la actualidad hay un gran auge de la minería de datos en educación y la difusión de esta disciplina se debe principalmente a su divulgación a través de variadas conferencias, congresos , eventos y publicaciones; como por ejemplo, la Conferencia Internacional de Minería de Datos en la Educación iniciado en 2008 y la revista científica Journal of Educational Data Mining (JEDM) [4]; también se desarrollan conferencias de tecnología educativa como la Conferencia Internacional sobre Inteligencia Artificial en Educación (AIED), la Conferencia Internacional sobre Sistemas Inteligentes de Tutoría (ITS) y la Conferencia Internacional sobre Modelamiento de usuario, adaptación y personalización (UMAP).

La aplicación de la minería de datos en los sistemas educativos tiene requisitos específicos que no se presentan en otros dominios, principalmente en este tipo de

investigación se plantea la necesidad de tener en cuenta aspectos pedagógicos de los alumnos y el sistema educativo [46].

Estudiar el desempeño académico requiere aproximarse a su definición, el cual expresa cuantitativamente el resultado de la actividad de aprendizaje del profesor, el alumno y el entorno, este es cuantificado por las evaluaciones que se realizan en un proceso de aprendizaje; de esta manera, se define a un estudiante con un buen rendimiento académico como aquel que obtiene una alta calificación en la escala de medida utilizada en la evaluación o prueba, la obtención de altas calificaciones en el individuo representa la adquisición de los conocimientos establecidos en un proceso formativo y esto a su vez, influye como un factor de estatus de calidad de la institución educativa donde se forma el individuo.

En Colombia, la vigilancia de la calidad educativa corresponde al Ministerio de Educación Nacional (MEN) y es gestionada a través del Instituto Colombiano para la Evaluación de la Educación (ICFES), entidad que diseña, construye y aplica evaluaciones en los niveles educativos de básica primaria, básica secundaria, media vocacional y pregrado; los resultados de estas pruebas determinan el estado de las competencias de los estudiantes, a la vez que sirven de herramienta para el mejoramiento de la calidad educativa de las instituciones educativas.

En la actualidad, a este conjunto de pruebas se les conoce como Pruebas SABER y se aplican de la siguiente forma:

- En grado 3º de educación básica primaria, llamada SABER 3°.
- En grado 5º de educación básica primaria, llamada SABER 5°.
- En grado 9º de educación básica secundaria, llamada SABER 9°.
- En grado 11º de educación media, llamada SABER 11°.
- A nivel de pregrado, llamada SABER PRO.

La Prueba SABER 11°, conocida también como el examen de Estado de la educación media o examen de ingreso para la educación superior, se aplica dos veces al año, una durante el primer semestre para colegios calendario B y otra en el segundo semestre para colegios calendario A; este examen es obligatorio para poder cursar estudios en una

Institución de Educación Superior, pero, de acuerdo con la autonomía universitaria establecida en la Ley 30 de 1992, son las Instituciones de Educación Superior las que fijan los criterios para su uso.

El ICFES, como entidad encargada de la aplicación del examen SABER 11, obtiene datos de las Instituciones Educativas de educación media y de los estudiantes a través de los formularios de inscripción para la presentación de la prueba, estos registros representan un alto volumen de datos que en la actualidad supera los 500.000 registros en el calendario A.

Este trabajo busca aplicar las técnicas de minería de datos en los resultados de la Prueba SABER 11 2012-2 a un subconjunto de datos correspondiente al Departamento del Cesar para proponer un modelo descriptivo que permita encontrar relaciones entre los diferentes factores que permita aproximarnos desde el enfoque computacional al estudio del desempeño académico de los estudiantes de educación media que se presentan a este examen.

Para el logro de estos resultados se utiliza la metodología CRISP – DM, la cual describe un modelo de proceso jerárquico en un conjunto de tareas que proporcionan una descripción del ciclo de vida de un proyecto de minería de datos consistente en seis fases dinámicas: comprensión del negocio, comprensión de los datos, preparación de datos, modelado, evaluación y desarrollo. Siguiendo esta metodología se tomarán los resultados de la Prueba Saber 11 del año 2012 del calendario A para la aplicación de algunas técnicas de minería de datos que permitan proponer un modelo que describa en el subconjunto de datos los posibles factores que puedan influir en el rendimiento académico de los estudiantes del Departamento del Cesar; propuesto este modelo se realizará una evaluación sistemática y un análisis de los resultados obtenidos. Si bien el Estado busca propuestas en mejorar la calidad educativa, se pretende que los resultados de la investigación constituyan un aporte significativo al enfoque de la educación con calidad.

Las limitaciones que se puedan presentar a nivel técnico se relacionan con la capacidad del equipo informático disponible para procesar los datos y en cuanto a los datos, se pueden presentar limitaciones por valores faltantes y outliers.

El documento se organiza de la siguiente forma:

Capítulo 1, es este capítulo una introducción al concepto de la minería de datos y la aplicación de estas técnicas en los contextos educativos, de igual manera se exponen los conceptos del desempeño académico y las investigaciones realizadas en esta temática.

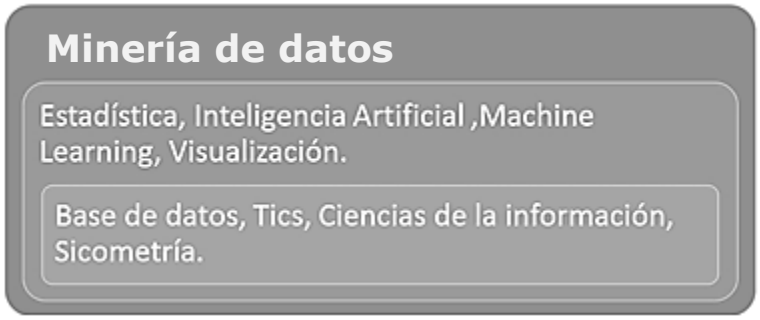
Los capítulos 2, 3, 4 y 5 están estrechamente relacionados a los objetivos específicos propuestos en el trabajo y la metodología CRISP-DM: Capítulo 2, comprensión del negocio; Capítulo 3, entendimiento de los datos; Capítulo 4, preparación de los datos; Capítulo 5, modelamiento.

# 1. Minería de datos y el contexto educativo

*En este capítulo se describe el concepto de minería de datos y la aplicación en el contexto educativo conocida como minería de datos en la educación (educational data mining). Seguidamente se da a conocer el concepto de desempeño académico, su relación con la minería de datos y algunos de los principales trabajos desarrollados referentes a este tema. Por último se presenta la conclusión del capítulo.*

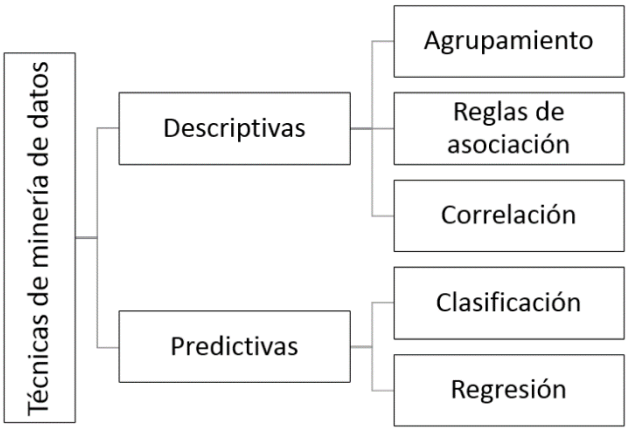
La minería de datos es una disciplina que tiene como objetivo la extracción de conocimiento útil de un conjunto de datos, donde inicialmente este conocimiento es desconocido pero que se encuentra en las relaciones implícitas de los atributos. El origen de la minería de datos se da hacia el año 1989 como una extensión de los fundamento de la inteligencia artificial y del análisis estadístico, sin embargo en el transcurrir de los años la minería de datos ha tenido la confluencia de otras disciplinas que han determinado la robustez de este concepto (ver Figura 1-1).

Figura 1-1: Confluencia de diferentes disciplinas en la minería de datos.




La minería de datos pretende resolver variadas tareas de investigación y para alcanzar estos retos la minería de datos determina un proceso iterativo a través de diferentes técnicas, en la Figura 1-2 se presentan algunas de ellas.

Figura 1-2: Técnicas de minería de datos



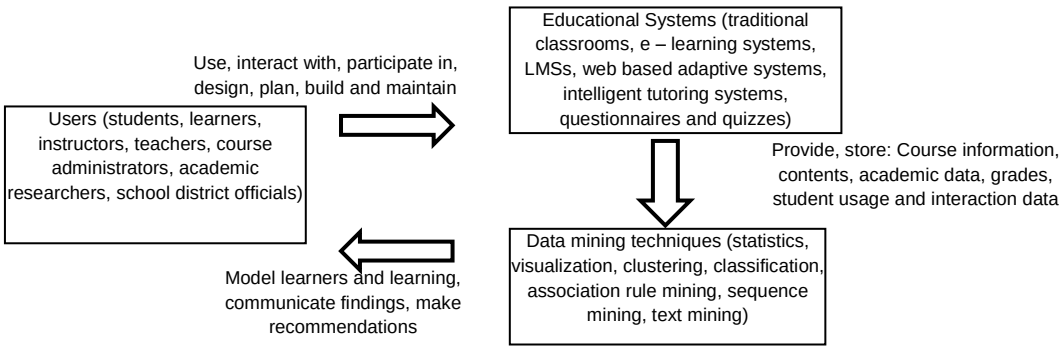

## 1.1 Minería de datos en la educación

La aplicación de los métodos y técnicas de la minería de datos en el contexto educativo es lo que se conoce como minería de datos en la educación o educational data mining, la cual es una disciplina emergente interesada en el desarrollo de métodos para la exploración de los tipos particulares de datos que provienen de los entornos educativos y el uso de esos métodos para entender mejor a los estudiantes y su entorno educativo [4].

La minería de datos en la educación es usada para comprender y tener mejor conocimiento de los estudiantes, para evaluar su progreso y los entornos educativos en que aprenden, ayuda a conocer a grupos en formación, a identificar el éxito o el fracaso en las estrategias de enseñanza y a generar un discernimiento más profundo del contexto educativo entre otros aspectos [46].

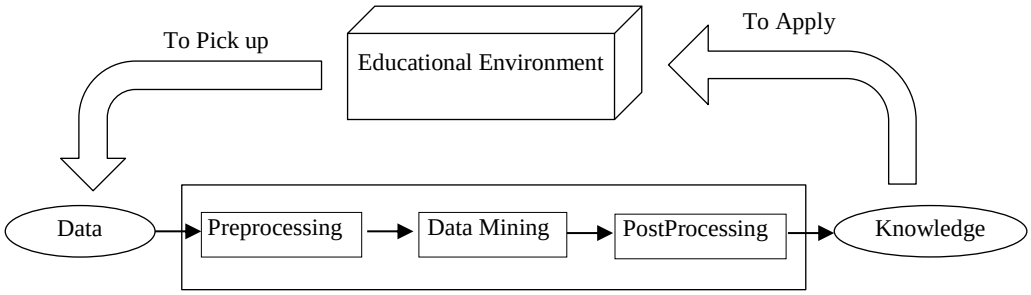
Para comprender como se desarrolla la minería de datos en la educación Romero & Ventura [45] presentan un esquema interactivo de este quehacer que se puede apreciar en la Figura 1-3.

Figura 1-3: El ciclo de Aplicación de la minería de datos en el sistema educativo. (Tomado de Romero & Ventura [45].




En la Figura 1-4 Ogor presenta otro modelo de cómo interactúa la minería de datos en el contexto educativo, en ambos modelos se observa la interacción de los datos de forma cíclica o que se retroalimenta, lo que supone un modelo que genera dinamismo propio de su naturaleza.

Figura 1-4.Interacción de la minería de datos en el contexto educativo. Tomado de [9].

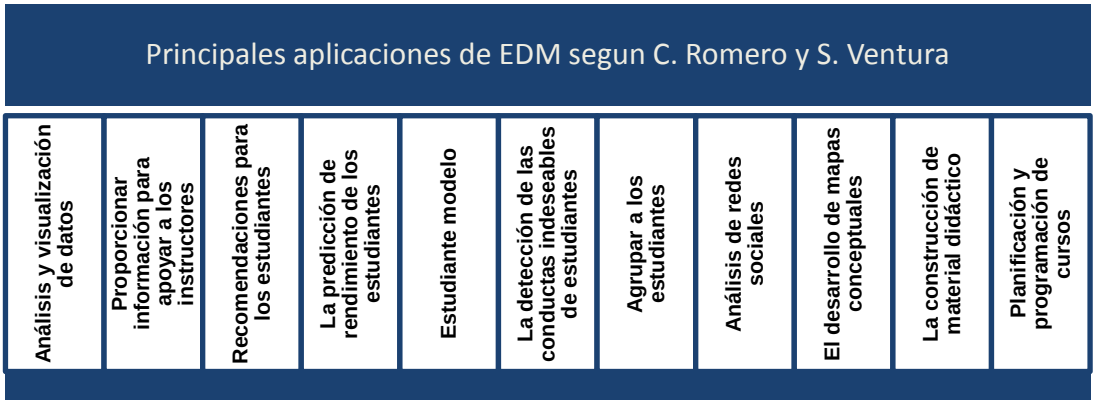



La minería de datos en la educación permite escudriñar los diferentes tipos de datos que provienen de los contextos educativos, su objetivo es analizar estos datos con el fin de apoyar las investigaciones y tener una mayor comprensión de los estudiantes y los contextos en que aprenden [29]. Las investigaciones llevadas a cabo en minería de datos en la educación se realizan con mayor frecuencia en el sistema educativo virtual y en menor medida en el sistema educativo tradicional o presencial, aunque estos dos sistemas difieren por los contextos en que se desarrollan, es posible aplicar las diferentes técnicas de minería en ambos casos. (Mr.Suhas G. Kulkarni, Mr.Ganesh C. Rampure, Mr.Bhagwat Yadav)

## 1.2 Aplicación de la minería de datos en la educación

Los investigadores Cristóbal Romero y Sebastián Ventura de la Universidad de Córdoba en España, proponen algunos campos de estudio aplicando la minería de datos en la educación, estos se ilustran en la Figura 1-5 y son aquellos de los cuales se encuentran los desafíos de investigación de este campo, es de resaltar el desafió de cómo se pueden generar los agrupamientos de los estudiantes según sus características.

Figura 1-5: Aplicaciones de minería de datos en la educación. Tomado de [1].




Los aportes logrados por la minería de datos en la educación abordan tres enfoques principales:

- En el alumno, para conocer cómo mejorar su aprendizaje
- En el docente, para tener una visión global de cómo desarrollar la clase
- En la institución educativa, para conocer las características de su entorno, hallar la eficiencia en sus servicios y desarrollar estrategias para elevar su nivel de calidad.

Entre los resultados de las investigaciones realizadas con la aplicación de las técnicas de minería de datos en la educación se encuentra desde el enfoque docente aquellos que han permitido apoyar a los tutores virtuales a ser conscientes de cómo los estudiantes progresan a través de un curso y como se puede realizar un seguimiento de los estudiantes para determinar quiénes tienen mayor riesgo de deserción; desde el enfoque de los estudiantes se ha logrado seleccionar grupos de trabajo con algunas características educacionales que aminoren el riesgo de fracaso escolar y desde el enfoque de las instituciones educativas se encuentran beneficios al hallar estrategias que faciliten la implementación de políticas administrativas y pedagógicas. Algunos de estos

aportes en estos tres enfoques de investigación educativa (estudiantes, docentes e institución) se pueden apreciar en la Tabla 1-1.

Tabla 1-1: Desafíos resueltos por la minería de datos en la educación. Elaboración Propia.

| AÑO | AUTOR | Entorno educativo | | Aporte a la población: | | | OBJETIVO | TÉCNICAS DATA MINING | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Basado en Web | Presencial | Estudiantes | Docentes | Institución Educativa | | Bayes | K-Means | Redes Neuronales | Arboles de decisión | Análisis discriminante | Maquinas de vectores soporte | Reglas de Asociación |
| 2005 | Merceron Agathe Yacef Kalina | X | | X | X | | Descubrir patrones que son pedagógicamente interesantes | | X | | X | | | X |
| 2006 | Superby JF | | X | X | X | | Estudiaron las correlaciones de diversos parámetros y su influencia en estudiantes universitarios | | | X | X | X | | |
| 2008 | Paulo Cortez and Alice Silva | | X | X | | X | Predicción del desempeño académico en secundaria con calificaciones de los alumnos de dos salones distintos | | | X | X | | X | |
| 2008 | Romero C, Ventura S, Espejo P., Hervás César | X | | | X | | Clasificar estudiantes comprando su rendimiento mediante diferentes técnicas | | | X | X | | | X |
| 2009 | Timarán P. | | X | | | X | Detectar patrones de bajo rendimiento académico en una universidad | | | | | | | X |
| 2010 | LS Affendey , IHM París , N. Mustapha , Md. Nasir Sulaiman y Z. Muda | | X | X | | | Extraer patrones útiles desde el archivo de los expedientes académicos de los estudiantes como indicadores de desempeño | | | X | X | | | |
| 2010 | C. Márquez-Vera , C. Romero And S. Ventura | | X | X | X | X | Predecir el fracaso escolar en estudiantes de secundaria. | | | | X | | | |
| 2012 | Antonenko Pavlo, Toy Serkan, Niederhauser D. | X | | | X | | Utilización de agrupamiento para inferir proceso de aprendizaje | | X | | | | | |
| 2013 | Lopez Guarin C. | | X | | | X | Identificación de características entre un grupo de estudiantes universitarios | X | X | | X | | | |
| 2015 | Harwati,, Ardita Permata Alfiani, Febriana Ayu Wulandari | | X | X | | | Identificar patrones rendimiento en estudiantes universitarios | | X | | | | | |
| 2015 | La Red Martínez, D. L., Karanik, M., Giovannini, M., y Pinto, N | | X | X | | | Determinación de perfiles de alumnos universitarios | | | | X | | | |

## 1.3 Desempeño Académico

El desempeño o rendimiento académico es un concepto físico que se obtiene de la interacción del profesor, el alumno y el entorno en que se desenvuelve; el desempeño académico se puede representar cualitativa y/o cuantitativamente, siendo esta última medida la más utilizada en su descripción y esta cuantificación se obtiene de las continuas evaluaciones o pruebas que se realizan en un proceso de aprendizaje. De esta manera, se define a un estudiante con buen rendimiento académico como aquel que obtiene una alta calificación en la escala de medida utilizada en la evaluación o prueba.

En otras palabras, el rendimiento académico es una expresión cuantitativa de la capacidad del estudiante que manifiesta lo que ha aprendido en el proceso de formación. No obstante, el estudio del rendimiento académico no se constituye en un proceso aislado y exclusivo del estudiante, este fenómeno representa la situación de todo un contexto socioeconómico e institucional donde se desenvuelve el estudiante y puede determinarse que cuando existe un bajo rendimiento académico el estudiante presenta frecuentemente la repetición de curso e incluso abandono escolar, son alumnos en los cuales el Estado invierte recursos, pero no obtiene el resultado esperado que es la promoción al grado siguiente [2].

## 1.4 Minería de datos y desempeño académico

El estudio del rendimiento académico es una de las aplicaciones más antiguas y populares de la minería de datos en la educación; la utilización de diferentes técnicas de minería como son las redes bayesianas, las reglas de asociación, la regresión, los arboles de decisión, el análisis de correlación, agrupamiento, entre otras, han generado avances en este campo.

El estudio del rendimiento académico con minería de datos puede darse a través de la predicción y de la descripción; el objetivo de la predicción es estimar el valor desconocido de una variable que describe al estudiante [3], [4],[5],[6] y [7].

El objetivo de la descripción es proporcionar la caracterización de un conjunto de datos en los cuales es posible encontrar subgrupos que permitan comprender y explicar el comportamiento del total de los datos.

A continuación se identifican estudios que relacionan la minería de datos en la educación y el rendimiento académico:

- Especificar las características de un estudiante o sus estados, tales como el conocimiento, la motivación, metacognición y actitudes; donde la capacidad de predecir el conocimiento y el futuro rendimiento de los estudiantes se incrementó con una exactitud hasta de un 48% [8] .

- Determinar que en entornos controlados el desempeño académico puede evaluarse fácilmente [9].

- Identificar las características del fracaso escolar en institutos [18] utilizando un árbol de decisión, se observa que la variable más relacionada con el fracaso escolar, es el rendimiento académico. *La minería de datos con respecto a los paradigmas tradicionales de investigación educacional como experimentos de laboratorio, estudios sociológicos o investigación de diseño ofrece más ventajas que cualquier otro método*.

- Estudios como el realizado por Baepler [10] en Scholarship of Teaching and Learning (SoTL), donde se diseña un algoritmo que envía señales de alerta al tutor por el rendimiento académico de los alumnos de cursos en línea.

- Pandey desarrolla un marco formado por factores demográficos, el rendimiento académico y los factores dimensionales (MUSTAS), subdivididos en tres dimensiones respectivamente: autoevaluación, la evaluación institucional y la evaluación externa. Todos estos factores ayudan a medir la dimensión del estudiante [11].

- Predecir el fracaso escolar en secundaria con la clasificación de los datos no balanceados por el reequilibrio de datos y utilizando la clasificación de coste

razonable con la aplicación de diez algoritmos: Cinco algoritmos de inducción de reglas tales como JRip, NNge, Oner, Prism y Ridor, y cinco algoritmos de árboles de decisión como J48, SimpleCart, ADTree, RandomTree y REPTree. Estos algoritmos fueron seleccionados porque se consideran como modelo de clasificación "caja blanca", es decir, que proporcionan una explicación para el resultado de la clasificación y se puede utilizar directamente para la toma de decisiones.[5].

En el ámbito local, los estudios en Colombia de la aplicación de la minería de datos en la educación se han desarrollado inicialmente en las universidades y se han aplicado para la detección de perfiles de bajo rendimiento y de fracaso estudiantil o deserción escolar. De esta manera Timarán en la Universidad de Nariño se desarrolló una herramienta computacional llamada TariYKDD para el descubrimiento de patrones de deserción estudiantil y bajo rendimiento académico, para lo cual utilizaron las tareas de Clasificación y Asociación [12]; de otra parte, el trabajo desarrollado por Pinzón en la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda bajo el estudio de variables demográficas del alumno con el registro de última matrícula del mismo semestre para el estudio de la deserción y las causas que lo generaron [13], recientemente, recientemente en la Universidad Nacional de Colombia se realizó un estudio con el fin de caracterizar y predecir los perfiles de rendimiento académico para algunos programas teniendo en cuenta los datos del sistema de información académica y los datos de admisión del estudiante [22].

## 1.5 Conclusiones del capítulo

En este capítulo se presentaron los conceptos y trabajos relacionados con minería de datos en la educación. Se inició con la descripción de minería de datos, después se introdujo a la minería de datos en la educación, el modelo y el ciclo de aplicación, campos de estudio y desafíos, como también algunos de los trabajos relacionados, nacionales e internacionales de la minería de datos y el desempeño académico.

# 2. SABER 11: Contexto de un examen estandarizado

*En este capítulo se describe el contexto donde se aplicará el proyecto, dado que está es una actividad que se requiere dentro de la metodología CRISP-DM donde se entiende el proyecto desde una perspectiva de negocio.*

La Prueba SABER 11 es realizada por el Instituto Colombiano para la Evaluación de la Educación (ICFES), entidad especializada en ofrecer servicios de evaluación de la educación en todos los niveles y en particular apoyar al Ministerio de Educación Nacional de Colombia (MEN) en la realización de los exámenes de Estado y en adelantar investigaciones sobre los factores que inciden en la calidad educativa, para ofrecer información pertinente y oportuna para contribuir al mejoramiento de la calidad de la educación. Algunas de sus funciones en cumplimiento de lo establecido mediante el artículo 12 de la Ley 1324 de 2009 y otras afines a las mismas, se encuentran:

- Establecer las metodologías y procedimientos que guían la evaluación externa de la calidad de la educación.
- Desarrollar la fundamentación teórica, diseñar, elaborar y aplicar instrumentos de evaluación de la calidad de la educación, dirigidos a los estudiantes de los niveles de educación básica, media y superior, de acuerdo con las orientaciones que para el efecto defina el Ministerio de Educación Nacional.
- Diseñar, implementar, administrar y mantener actualizadas las bases de datos con la información de los resultados alcanzados en las pruebas aplicadas y los factores asociados, de acuerdo con prácticas internacionalmente aceptadas.
- Organizar y administrar el banco de pruebas y preguntas, según niveles educativos y programas, el cual tendrá carácter reservado.
- Diseñar, implementar y controlar el procesamiento de información y la producción y divulgación de resultados de las evaluaciones realizadas, según las necesidades identificadas en cada nivel educativo.

- Impulsar y fortalecer la cultura de la evaluación de la calidad de la educación mediante la difusión de los resultados y de los análisis acerca de los factores que inciden en los mismos, y el desarrollo de actividades de formación en los temas que son de su competencia, en los niveles local, regional y nacional.

El ICFES es una entidad del Estado que vigila la calidad educativa en Colombia y dadas las funciones que le corresponden dispone de mecanismos que le permiten generar investigaciones de los fenómenos relacionados a la calidad en la educación, de igual forma deja abierta la posibilidad de que la comunidad en general participe de este proceso investigativo, por lo tanto, provee a través de su sitio FTP los registros de la aplicación de las pruebas desde al año 2000 para que las personas interesadas tengan acceso a ellos.

Los estudios y propuestas para el avance en calidad educativa constituyen para el Estado y la Nación una apuesta significativa que determina el desarrollo social, tecnológico y económico de un país. El plantear alternativas de estudio en el campo del rendimiento académico es un gran aporte para el entendimiento de este fenómeno que desde diversas áreas se ha analizado, tal como lo ha hecho la sicología, la pedagogía, las ciencias sociales, la medicina y las ciencias económicas. Abordar la investigación del rendimiento académico en la Prueba SABER 11 con técnicas de minería de datos constituye en un primer intento computacional en Colombia de describir aspectos de este fenómeno.

## 2.1 Evaluación de la situación

El examen de Estado es aplicado en Colombia desde 1968 a los estudiantes de Instituciones Educativas de grado 11, con el fin de conocer, entre otras variables, cómo se encuentra la calidad de la educación media en el país, además de ser requisito primordial para el ingreso de los jóvenes a la Educación Superior (ICFES). Este examen lo presentan tres tipos de población: bachilleres, validantes de bachillerato y personas

particulares; el ICFES dispone para el público en general solo de los resultados de los estudiantes adscritos a las instituciones educativas.

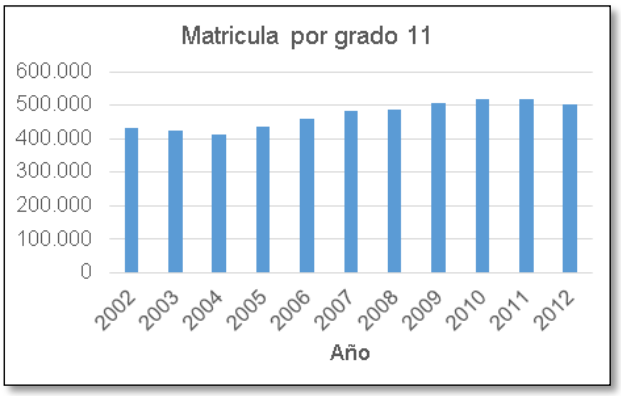
Desde el establecimiento de esta prueba como obligatoria y como instrumento de evaluación del desempeño académico para los estudiantes y de nivel de calidad educativa para las instituciones, este examen ha sufrido diversas modificaciones para adaptarse a los cambios en el entorno educativo y a las actualizaciones de los currículos académicos para la educación media, la más reciente modificación hecha a la prueba se dio para la segunda aplicación del año 2014, donde se consideraron las siguientes modificaciones: una nueva forma de calificación, la reestructuración de las áreas a evaluar, inclusión de una nueva competencia, reducción del número de preguntas y la utilización de preguntas abiertas. Es de aclarar que el conjunto de datos tomado corresponde al año 2012 y no le aplican estas modificaciones.

Con base en la Ley 1324 de 2009 y en la visión y misión del sector educativo como un todo, la misión del ICFES es “ofrecer el servicio de evaluación de la educación en todos sus niveles, y adelantar investigación sobre los factores que inciden en la calidad educativa, con la finalidad de ofrecer información para mejorarla”. Este examen es de suma importancia para el país, tanto a nivel individual, porque posibilita la iniciación en los estudios de educación superior, como institucional, porque representa un indicador de desempeño de calidad ante el Ministerio de Educación Nacional.

Para cumplir con este proceso, los colegios que ofrecen educación media tanto oficiales como no oficiales deben gestionar la inscripción de todos sus estudiantes de último grado para que presenten la prueba en las fechas establecidas de acuerdo al calendario en que se encuentre registrada la institución educativa (calendario A o Calendario B), para esto el ICFES posee un sistema de inscripción que permite autogestionar el proceso de inscripción de manera confiable y confidencial que permite diligenciar la información sociodemográfica del estudiante en el aplicativo web y por ende también se obtiene la información del colegio del inscrito. Las Instituciones que no realicen este proceso de manera oportuna con la totalidad de sus estudiantes no reciben los resultados de clasificación de planteles.

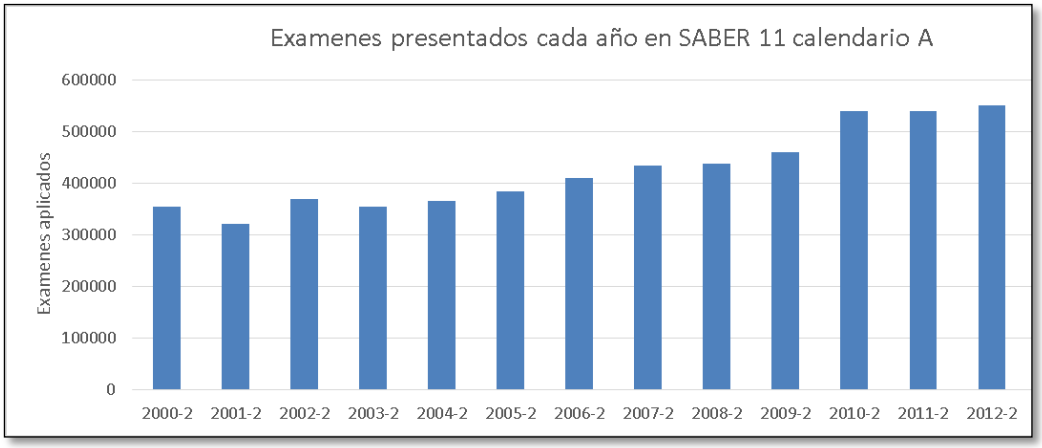
El Ministerio de Educación Nacional para el año 2012 reporta que los establecimientos educativos que ofrecen educación media son en totalidad 22.937, donde 13.126 son de tipo Oficial y 9.8110 son No Oficiales correspondientes al 57,23% y el 42,77% respectivamente, estos establecimientos representaron para el país en ese año una población estudiantil matriculada en grado 11 de aproximadamente de 502.603 jóvenes, véase en la Figura 2-1 el incremento de los estudiantes en grado 11 en los últimos 10 años.

Figura 2-1: Matrícula grado 11 2002 – 2012. Tomado del MEN




La población de estudiantes matriculados en grado 11 debe presentar la Prueba Saber 11 del ICFES y lo que se puede ver en la Figura 2-2 corresponde a los estudiantes de Calendario A que han presentado esta prueba entre los años 2000 -2012, representando un alto volumen de datos del contexto educativo de educación media en Colombia que se encuentran disponibles para ser analizados.

Figura 2-2: Total exámenes aplicados ICFES 2000 – 2012. Tomado del FTP ICFES.




Con base en los resultados del examen se desarrollan, entre otros, los procedimientos para:

- Selección, nivelación y prevención de la deserción en la educación superior.
- Monitoreo de la calidad de las instituciones educativas, a partir de los estándares básicos de competencias establecidos por el Ministerio de Educación Nacional (MEN).
- Dar información para estimación del valor agregado de la educación media y de la educación superior.

## 2.2 Objetivo de la prueba SABER 11

De acuerdo con el Decreto 869 de marzo de 2010, el examen de Estado de la educación media - ICFES SABER 11° tiene como fin comprobar el grado de desarrollo de las competencias de los estudiantes que están por terminar undécimo grado y que aspiran a ingresar a la educación superior, la cual tiene por objetivos:

1. Comprobar el grado de desarrollo de las competencias de los estudiantes que están por finalizar el grado undécimo de la educación media.
2. Proporcionar elementos al estudiante para la realización de su autoevaluación y el desarrollo de su proyecto de vida.

3. Proporcionar a las instituciones educativas información pertinente sobre las competencias de los aspirantes a ingresar a programas de educación superior, así como sobre los de quienes son admitidos que sirva como base para el diseño de programas de nivelación académica y prevención de la deserción en este nivel.
4. Monitorear la calidad de la educación de los establecimientos educativos del país, con fundamento en los estándares básicos de competencias y los referentes de calidad emitidos por el Ministerio de Educación Nacional.
5. Proporcionar información para el establecimiento de indicadores de valor agregado, tanto de la educación media como de la educación superior.
6. Servir como fuente de información para la construcción de indicadores de calidad de la educación, así como para el ejercicio de la inspección y vigilancia del servicio público educativo.
7. Proporcionar información a los establecimientos educativos que ofrecen educación media para el ejercicio de la autoevaluación y para que realicen la consolidación o reorientación de sus prácticas pedagógicas.
8. Ofrecer información que sirva como referente estratégico para el establecimiento de políticas educativas nacionales, territoriales e institucionales.

SABER 11 se concibe como:

- Una evaluación individual para efectos de admisión a la educación superior y otorgamiento de beneficios y
- Un instrumento de información sobre la calidad educativa.

## 2.3 Criterios de éxito del negocio

El Estado colombiano debe garantizar la cobertura y la calidad de la educación y se ha designado al ICFES como ente encargado de gestionar la calidad educativa en el país. Como parte de su estrategia esta entidad tiene a bien promover el uso de los datos de los resultados de los exámenes aplicados en los diferentes niveles educativos, por lo que la presentación de avances y propuestas en calidad educativa constituyen para el Estado una apuesta significativa que determina el avance social, tecnológico y económico de la Nación.

El plantear alternativas de estudio para el fenómeno del rendimiento académico en los estudiantes de educación media que presentan la Prueba SABER 11, es un aporte que permite apoyar iniciativas para la los planes de cobertura y calidad, de igual forma puede contribuir a programas de movilidad social y de estrategias para el acceso, la equidad y la pertinencia en la educación. El estudio del rendimiento académico representa un aporte en el avance de temas como la repitencia y el abandono escolar, un bajo rendimiento académico genera repetición en los alumnos y abandono escolar, son alumnos en los cuales el Estado invierte recursos, pero no obtiene el resultado esperado que es la promoción al grado siguiente (ICFES,[2]).

## 2.4 Objetivos de la minería de datos

Se pretende obtener un conjunto de datos de los resultados de la prueba SABER 11 correspondiente al Calendario A del año 2012 en el Departamento del Cesar, al cual se les aplique las actividades de comprensión y preprocesamiento según la técnica de minería de datos seleccionada lo requiera y que propicien la construcción de modelos descriptivos que posibiliten un mejor entendimiento en los resultados de la prueba SABER 11. Para ello se requerirá:

- Preparar un conjunto de datos apropiado para la aplicación de las diferentes técnicas de minería de datos.
- Proponer e implementar un modelo descriptivo para el agrupamiento de datos.
- Realizar una evaluación sistemática del modelo propuesto y un análisis de los resultados obtenidos.

Con estos objetivos se podrá obtener información que pueda contribuir a:

- Identificar características de los alumnos según el nivel de desempeño académico y
- Encontrar relaciones entre variables que permitan describir el nivel de desempeño académico.

## 2.5 Criterios de éxito de la minería de datos

Se determina como factor de éxito del modelo descriptivo, cuando los resultados o agrupamientos hallados sean comprensibles y permitan ser validados subjetivamente con el conjunto de datos procesado.

## 2.6 Evaluación inicial de herramientas y técnicas

Como herramienta de trabajo en el proceso de extracción de conocimiento se utilizará el software Rapid Miner (anteriormente, YALE, Yet Another Learning Environment) que trabaja con el encadenamiento de operadores a través de un entorno gráfico y tiene como características:

- Software de tipo Open-Source con licencia GNU GPL, basado en java.
- Posee alrededor de 400 operadores que pueden ser combinados.
- Capacidad de jerarquizar cadenas de operadores y de construir complejos árboles de operadores.

Esta herramienta cuenta con los algoritmos de minería de datos que se prevén para el desarrollo de este proyecto.

## 2.7 Conclusión

En este capítulo se presentó la etapa de entendimiento del contexto de negocio, este concepto proviene de la aplicación de la metodología CRISP-DM, donde el trabajo abordado se contempla desde una perspectiva de negocio del cliente y una perspectiva de minería de datos por el equipo desarrollador en donde se exponen los criterios de éxito del trabajo para satisfacer los objetivos propuestos.

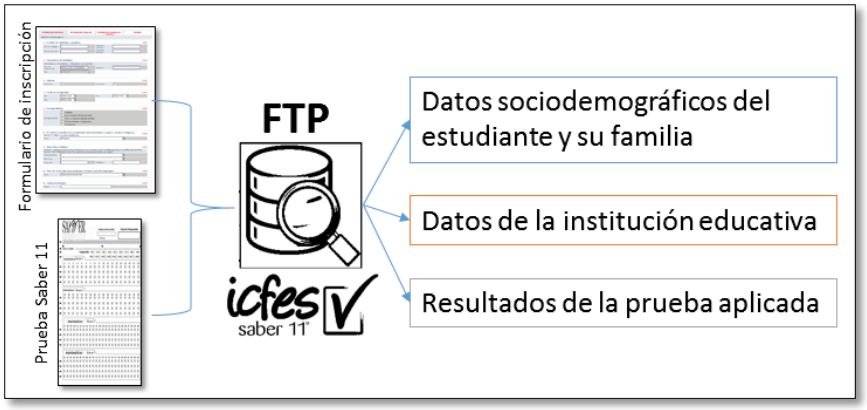
# 3. Entendimiento de los datos

*En este capítulo se realiza la descripción de los datos donde se aplican técnicas de visualización de datos como los histogramas, esto con el fin de realizar una exploración preliminar de los registros y verificar la calidad de los mismos. Se deben considerar la detección de datos faltantes y atípicos que puedan presentarse para obtener confiabilidad en el conjunto de datos a trabajar.*

## 3.1 Recolección de Datos

Los datos para este proyecto son obtenidos del sitio FTP del Instituto Colombiano para la Evaluación de la Educación (ftp://ftp.icfes.gov.co/) organismo que dispone para el público la información de los resultados nacionales de la Prueba Saber 11, entre otras pruebas, esto constituye la fuente primaria de información y un aseguramiento del 100% de fiabilidad en los datos. En los datos se encuentra la información socioeconómica del estudiante y su familia, información del colegio y los resultados obtenidos en la prueba, en la Figura 3-1 se puede apreciar el procedimiento para la obtención de los datos.

Figura 3-1: Base de datos FTP ICFES



Los datos se encuentran en un archivo plano en formato ZIP, las características del archivo con la información para el año 2012 -2 se encuentra en la Tabla 3-1. Estos datos son importados desde Excel para una mejor estructuración de la tabla. Las consideraciones a tener en cuenta al momento de cargar estos archivos a Excel, son las siguientes:

- Tipo de archivo: Archivo de texto plano (.txt)
- Codificación: UTF-8
- Símbolo delimitador de columnas: Pleca o Vertical bar (|) Hexa 007C

Tabla 3-1: Archivo FTP ICFES

| | |
|---|---|
| **Nombre Archivo FTP ICFES:** | SB11-20122-RGSTRO-CLFCCN-V1-0.zip |
| **Nombre Archivo compreso:** | SB11-20122-RGSTRO-CLFCCN-V1-0.txt |
| **Tamaño:** | ZIP: 24.3 MB<br>TXT: 187.6 MB |
| **Fecha de modificación:** | 29/01/2014 |

## 3.2 Descripción del conjunto de datos

La descripción de datos corresponde a los atributos contenidos en el archivo a utilizar para la tarea de minería, estos se presentan en la Tabla 3-2 y la Tabla 3-3. La dimensionalidad del conjunto de datos es multivariada, en lo que se encontró datos de tipo nominal y numérico.

Tabla 3-2: Descripción de fuente de datos

| | |
|---|---|
| **Archivo:** | SB11-2012-2-RGSTRO-CLFCCN-V1-0.txt |
| **Importado a:** | Excel |
| **Número de instancias:** | 550155 |
| **Número de atributos:** | 73 |

El diccionario de variables se encuentra en el Anexo C y su estructura se aprecia en algunos atributos que se exponen en la Tabla 3-3.

Tabla 3-3. Estructura del conjunto de datos SABER 11 2012-2.

| N° | ATRIBUTO | DESCRIPCIÓN | MIN | MAX | MEDIA | MODA |
|---|---|---|---|---|---|---|
| Datos sociodemográficos del estudiante y su familia | ESTU_EDAD | Edad del inscrito | 9 | 82 | 17 | 16 |
| | ESTU_GENERO | Genero del inscrito | | | | F |
| | ESTU_VECES_ESTDO | Veces que el estudiante ha presentado el examen de estado | 0 | 4 | | 0 |
| | ESTU_ESTRATO | Estrato socioeconómico de la residencia del estudiante según factura de energía | 1 | 8 | | 1 |
| | ESTU_TRABAJA | Trabaja actualmente. | 0 | 7 | | 0 |
| | FAMI_COD_EDUCA_PADRE / MADRE | Máximo nivel educativo alcanzado por el padre / madre | 0 | 99 | | 12 |
| | FAMI_COD_OCUP_PADRE / MADRE | Ocupación del padre/ madre | 1 | 26 | | 21 |
| | FAMI_ING_FMILIAR_MENSUAL | Ingresos mensuales representados en salarios mínimos mensuales. | 1 | 7 | | 2 |
| | FAMI_NIVEL_SISBEN | Nivel de SISBEN en que está clasificada la familia | 1 | 5 | | 1 |
| | ECON_CUARTOS | Número de habitaciones de la residencia | 1 | 10 | | 3 |
| | ECON_PERSONAS_HOGAR | Número de personas que conforman el hogar | 1 | 12 | | 4 |
| Datos de la institución educativa | COLE_CALENDARIO_COLEGIO | Calendario del colegio | | | | A |
| | COLE_GENERO_POBLACION | Población del colegio | | | | X |
| | COLE_NATURALEZA | Naturaleza del colegio | | | | O |
| | COLE_ES_BILINGUE | La institución educativa es bilingüe | 0 | 1 | | 0 |
| | COLE_INST_JORNADA | Jornada del colegio | | | | MAÑANA |
| | COLE_CARACTER_COLEGIO | Carácter del colegio | | | | ACADEMICO |
| | COLE_INST_VLR_PENSION | Valor de la pensión pagada por el estudiante en el último año | 0 | 12 | | 0 |
| Resultados de la prueba aplicada | LENGUAJE_PUNT | Puntaje en lenguaje | 0 | 93 | 46 | 44 |
| | MATEMATICAS_PUNT | Puntaje en matemáticas | 0 | 126 | 46 | 49 |
| | CIENCIAS_SOCIALES_PUNT | Puntaje en ciencias sociales | 0 | 111 | 44 | 42 |

## 3.3 Descripción de atributos

El conjunto de datos utilizado contiene 550.155 instancias con 75 atributos, 64 atributos categóricos y 11 atributos numéricos los cuales se describen en la Tabla 3-4 y la Tabla 3-5.

Tabla 3-4. Descripción atributos numéricos

| ATRIBUTOS NÚMERICOS (11) | | | | |
|---|---|---|---|---|
| **ATRIBUTO** | **PROMEDIO** | **DESVIACIÓN ESTÁNDAR** | **RANGO** | **MISSINGS** |
| ESTU_EDAD | 17.406 | 4.192 | [9.000 ; 82.000] | 1 |
| LENGUAJE_PUNT | 46.249 | 7.135 | [0.000 ; 93.000] | 0 |
| MATEMATICAS_PUNT | 45.614 | 11.181 | [0.000 ; 126.000] | 0 |
| CIENCIAS_SOCIALES_PUNT | 44.327 | 8.956 | [0.000 ; 111.000] | 0 |
| FILOSOFIA_PUNT | 40.367 | 8.815 | [0.000 ; 81.000] | 0 |
| BIOLOGIA_PUNT | 45.166 | 7.965 | [-1.000 ; 100.000] | 0 |
| QUIMICA_PUNT | 45.633 | 6.417 | [-1.000 ; 85.000] | 0 |
| FISICA_PUNT | 44.297 | 7.620 | [0.000 ; 112.000] | 0 |
| INGLES_PUNT | 43.703 | 10.254 | [-1.000 ; 100.000] | 0 |
| COMP_FLEX_PUNT | 31.251 | 22.127 | [0.000 ; 90.000] | 0 |

Tabla 3-5. Descripción atributos categóricos

| ATRIBUTOS CATEGÓRICOS (64) | | |
|---|---|---|
| **ATRIBUTO** | **MODA** | **MISSINGS** |
| ESTU_PAIS_RESIDE | CO (549956) | 0 |
| ESTU_GENERO | F (303199) | 93 |
| ESTU_DISC_SORDOCEGUERA | C (23) | 0 |
| ESTU_DISC_COGNITIVA | G (509) | 0 |
| ESTU_DISC_INVIDENTE | I (112) | 0 |
| ESTU_DISC_MOTRIZ | M (549) | 0 |
| ESTU_DISC_SORDOINTERPRETE | R (394) | 0 |
| ESTU_DISC_SORDONOINTERPRETE | S (146) | 0 |
| COLE_CALENDARIO_COLEGIO | A (463401) | 2102 |
| COLE_GENERO_POBLACION | X (517235) | 6390 |
| COLE_NATURALEZA | O (413234) | 0 |
| ESTU_EXAM_NOMBREEXAMEN | EXAMEN SABER 11- 2012 CAL A (550155) | 0 |
| PERIODO | 20122 (550155) | 0 |
| ESTU_NACIMIENTO_DIA | 5 (18660) | 102 |
| ESTU_NACIMIENTO_MES | 12 (49624) | 102 |
| ESTU_NACIMIENTO_ANNO | 1995 (206516) | 102 |
| ESTU_ETNIA | 1 (19496) | 0 |
| ESTU_CODIGO_RESIDE_MCPIO | 11001 (94427) | 94 |
| ESTU_ZONA_RESIDE | 10 (369866) | 30 |
| ECON_AREA_VIVE | 1 (429526) | 230 |
| INAC_COLEGIOTERMINO | 59 (1092) | 5163 |

| | | |
|---|---|---|
| COLE_CODIGO_COLEGIO | 59 (1093) | 0 |
| COLE_ES_BILINGUE | 0 (458380) | 86472 |
| COLE_INST_VLR_PENSION | 0 (427992) | 3387 |
| ESTU_VECES_ESTDO | 0 (525171) | 4031 |
| ESTU_EXAM_COD_MPIOPRESENTACION | 11001 (94616) | 0 |
| FAMI_COD_EDUCA_PADRE | 12 (116199) | 214 |
| FAMI_COD_EDUCA_MADRE | 12 (132748) | 214 |
| FAMI_COD_OCUP_PADRE | 21 (179417) | 214 |
| FAMI_COD_OCUP_MADRE | 22 (284133) | 214 |
| ESTU_ESTRATO | 1 (224952) | 485 |
| ECON_CUARTOS | 3 (226439) | 256 |
| FAMI_NIVEL_SISBEN | 1 (244136) | 214 |
| ECON_MATERIAL_PISOS | 2 (252039) | 214 |
| ECON_PERSONAS_HOGAR | 4 (158848) | 214 |
| ECON_SN_TELEFONIA | 0 (294752) | 213 |
| ECON_SN_CELULAR | 1 (512142) | 213 |
| ECON_SN_INTERNET | 0 (309991) | 213 |
| ECON_SN_SERVICIO_TV | 1 (337460) | 213 |
| ECON_SN_COMPUTADOR | 3 (289088) | 213 |
| ECON_SN_LAVADORA | 1 (340438) | 213 |
| ECON_SN_NEVERA | 1 (482891) | 213 |
| ECON_SN_HORNO | 0 (314973) | 213 |
| ECON_SN_DVD | 1 (364479) | 213 |
| ECON_SN_MICROHONDAS | 0 (402154) | 213 |
| ECON_SN_AUTOMOVIL | 0 (449274) | 213 |
| FAMI_ING_FMILIAR_MENSUAL | 2 (242176) | 214 |
| ESTU_TRABAJA | 0 (481137) | 3470 |
| ESTU_PUESTO | 978 (568) | 0 |
| ESTU_DISC_BAJAVISION | | 0 |
| ESTU_HORAS_TRABAJO | 20 (135) | 549934 |
| FTP_CONSECUTIVO | | 328 |
| ESTU_TIPODOCUMENTO | T (428073) | 0 |
| ESTU_RESIDE_MPIO_PRESENTACION | BOGOTÁ D.C. (94427) | 94 |
| ESTU_RESIDE_DEPT_PRESENTACION | BOGOTÁ (94427) | 94 |
| CODIGO_DANE | 1,11001E+11 (43397) | 1725 |
| COLE_INST_NOMBRE | INST EDUC CEFA (1093) | 0 |
| COLE_INST_JORNADA | MAÑANA (284440), | 0 |
| PLAN_CODIGODANEINSTITUCION | 1,11001E+11 (43159) | 4104 |
| COLE_CARACTER_COLEGIO | ACADEMICO (348661) | 20 |
| ESTU_EXAM_MPIO_PRESENTACION | BOGOTÁ D.C. (94616) | 0 |
| ESTU_EXAM_DEPT_PRESENTACION | BOGOTÁ (94616) | 0 |
| INGLES_DESEM | A- (311235) | 329 |
| COMP_FLEX_NOMBRE | MEDIO AMBIENTE (194538) | 0 |
| COMP_FLEX_DESEM | I (100537) | 0 |

## 3.4 Exploración de datos

El análisis exploratorio es una actividad que nos permite analizar en detalles algunas variables e identificar características de algunas subpoblaciones y formular algunas preguntas de interés con relación al objetivo de minería, por lo que se consideran los siguientes interrogantes:

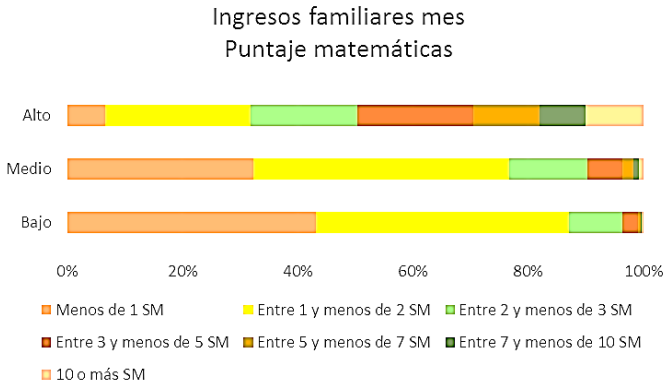
1. ¿El ingreso familiar influye negativa o positivamente en los resultados obtenidos por el estudiante? (Ver Tabla 3-6)
2. ¿Las características del plantel educativo influyen en los puntajes del resultado general?
3. ¿El género es preponderante en la obtención del resultado en alguna área en particular?
4. ¿El nivel educativo de los padres y específicamente el de la madre influye positivamente en los resultados de la prueba?
5. ¿El entorno socioeconómico de la región de residencia influye en el desempeño de la prueba?
6. ¿Qué nivel de desempeño académico se destaca en el conjunto de datos?

A continuación se realiza la exploración inicial en Excel para visualizar datos del conjunto total de la población que presento la prueba SABER 11 2012-2, correspondiente al calendario A, y realizar algunas conclusiones según las preguntas de interés planteadas, así:

Tabla 3-6. Ingresos familiares SABER 11 2012 - 2

| Ingreso familiar mensual | Frecuencia | Porcentaje | Porcentaje acumulado |
|---|---|---|---|
| Menos de 1 SM | 180122 | 32,8 | 32,8 |
| Entre 1 y menos de 2 SM | 241870 | 44,0 | 76,8 |
| Entre 2 y menos de 3 SM | 73435 | 13,4 | 90,1 |
| Entre 3 y menos de 5 SM | 32764 | 6,0 | 96,1 |
| Entre 5 y menos de 7 SM | 10827 | 2,0 | 98,1 |
| Entre 7 y menos de 10 SM | 5678 | 1,0 | 99,1 |
| 10 o más SM | 4954 | ,9 | 100,0 |
| Total | 549650 | 100,0 | |
| Perdidos | 182 | ,0 | |
| Total | 549832 | 100,0 | |

Figura 3-2: Ingreso familiar mensual SABER 11 2012 -2 - Puntaje matemáticas

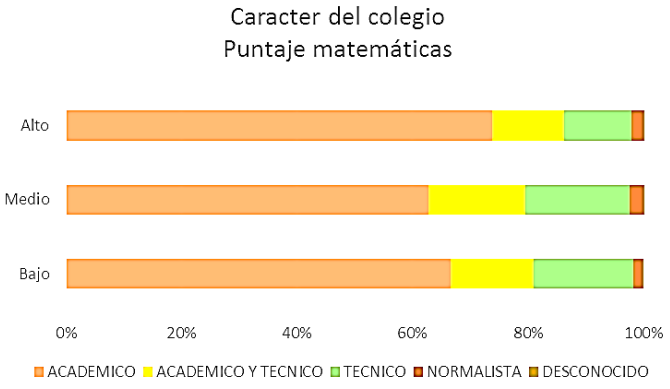



Se observa de la figura anterior que los ingresos familiares pueden influir positivamente en el logro de mejores resultados en el puntaje de la prueba, en este caso matemáticas, donde el 10% de estudiantes con puntajes altos provienen de hogares donde se perciben ingresos iguales o superiores a 10 salarios mínimos mensuales. Esta situación frente a un poco más del 40% de los puntajes bajos lo obtienen estudiantes donde el ingreso es menor a un salario mínimo, cabe resaltar que se presenta a estudiantes en donde sus hogares siguen percibiendo este mismo salario, aproximadamente un 7% de ellos logra resultados altos.

Tabla 3-7: Carácter del colegio SABER 11 2012 – 2.

| Carácter del colegio | Frecuencia | Porcentaje | Porcentaje acumulado |
|---|---|---|---|
| Vacios | 20 | ,0 | ,0 |
| ACADEMICO | 348437 | 63,4 | 63,4 |
| ACADEMICO Y TECNICO | 90396 | 16,4 | 79,8 |
| DESCONOCIDO | 307 | ,1 | 79,9 |
| NORMALISTA | 12703 | 2,3 | 82,2 |
| TECNICO | 97969 | 17,8 | 100,0 |
| Total | 549832 | 100,0 | |

Figura 3-3: Carácter del colegio SABER 11 2012 – 2

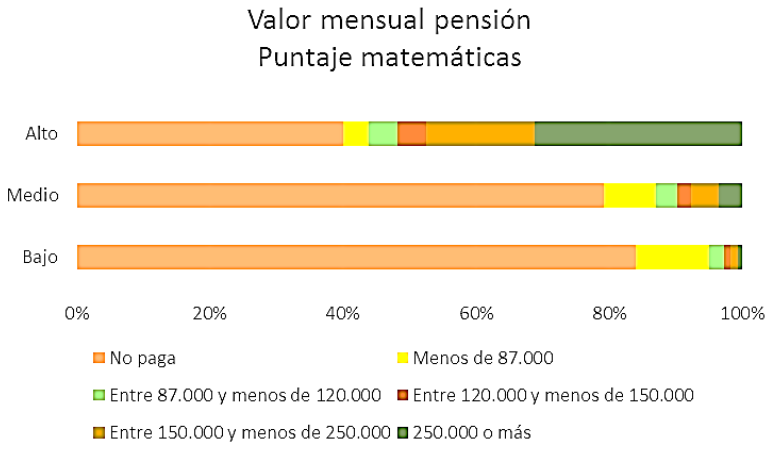


Se puede discernir de la Figura 3-3 que el carácter del colegio no es un factor preponderante en los resultados del examen.

Tabla 3-8. Valor mensual de pensión SABER 11 2012 – 2

| Valor Mensual de Pensión | Frecuencia | Porcentaje | Porcentaje acumulado |
|---|---|---|---|
| No paga | 433763 | 78,9 | 78,9 |
| Menos de 87.000 | 44104 | 8,0 | 86,9 |
| Entre 87.000 y menos de 120.000 | 16886 | 3,1 | 90,0 |
| Entre 120.000 y menos de 150.000 | 11197 | 2,0 | 92,1 |
| Entre 150.000 y menos de 250.000 | 22773 | 4,1 | 96,2 |
| 250.000 o más | 20909 | 3,8 | 100,0 |
| Total | 549632 | 100,0 | |
| Perdidos | 200 | ,0 | |
| Total | 549832 | 100,0 | |

Figura 3-4: Valor mensual de pensión vs puntaje en matemáticas SABER 11 2012 – 2

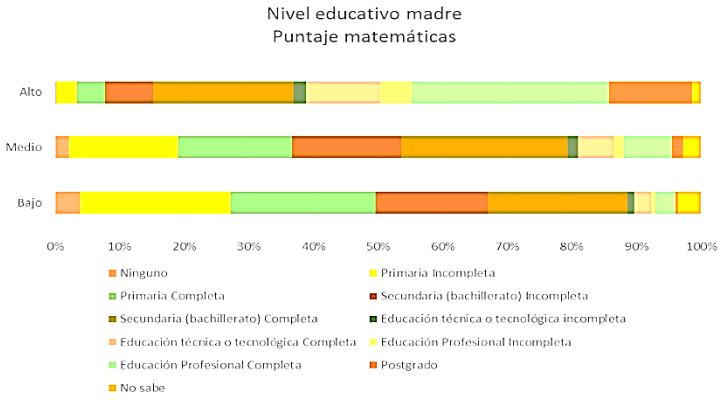



En la Tabla 3-8 y la Figura 3-4 se puede analizar que el 60% de los puntajes altos en el área de matemáticas proviene de colegios en donde se paga una pensión por encima de los $87.000 y un poco más del 30% de este grupo son colegios donde la mensualidad se ubica dentro de las más costosas.

Tabla 3-9. Nivel educativo de los padres SABER 11 2012 – 2.

| Nivel educativo | Nivel educativo del padre | | | Nivel educativo de la madre | | |
|---|---|---|---|---|---|---|
| | Frecuencia | Porcentaje | Porcentaje acumulado | Frecuencia | Porcentaje | Porcentaje acumulado |
| Ninguno | 21044 | 3,8 | 3,8 | 12683 | 2,3 | 2,3 |
| Primaria Incompleta | 106246 | 19,3 | 23,2 | 94269 | 17,1 | 19,5 |
| Primaria Completa | 97082 | 17,7 | 40,8 | 97895 | 17,8 | 37,3 |
| Secundaria (bachillerato) Incompleta | 77787 | 14,1 | 55,0 | 91936 | 16,7 | 54,0 |
| Secundaria (bachillerato) Completa | 121976 | 22,2 | 77,2 | 139422 | 25,4 | 79,4 |
| Educación técnica o tecnológica incompleta | 6599 | 1,2 | 78,4 | 8433 | 1,5 | 80,9 |
| Educación técnica o tecnológica Completa | 23112 | 4,2 | 82,6 | 29299 | 5,3 | 86,2 |
| Educación Profesional Incompleta | 7823 | 1,4 | 84,0 | 8523 | 1,6 | 87,8 |
| Educación Profesional Completa | 42446 | 7,7 | 91,7 | 41859 | 7,6 | 95,4 |
| Postgrado | 9518 | 1,7 | 93,4 | 9468 | 1,7 | 97,1 |
| No sabe | 36017 | 6,6 | 100,0 | 15863 | 2,9 | 100,0 |
| Total | 549650 | 100,0 | | 549650 | 100,0 | |
| Perdidos | 182 | ,0 | | 182 | ,0 | |
| **Total** | **549832** | **100,0** | | **549832** | **100,0** | |

Figura 3-5: Nivel educativo de la madre SABER 11 2012 – 2




Por la Tabla 3-8 y la Figura 3-5 se observa que el nivel educativo de la madre influye positivamente en la obtención de mejores resultados, el 50% de estos estudiantes tienen madres que han cursado programas universitarios de pregrado y postgrado, solo el 5% de estas madres no ha terminado el ciclo profesional.

Se encuentra en los datos que el 86,25% de los estudiantes tiene edades entre los 14 y 18 años de edad, y le sigue un porcentaje del 8,34% de estudiantes entre las edades de 19 y 23 años, considerada una extra edad para culminar el bachillerato, esto se denota en que los que presentan la prueba pertenecen también a instituciones con programas especiales de validación del bachillerato u otros como los sabatinos.

Se observa también que el 75,11% de los colegios en Colombia son de naturaleza oficial, con un 66,46% perteneciente al calendario A y el 43,98% de la jornada mañana, un 15% de la jornada de la tarde y un 7,99% en jornada completa.

Para el 2012-2 en el Cesar se presentaron a la prueba 11.522 estudiantes, lo que representa al 2% de la población total correspondiente a ese periodo; en el Cesar se encuentran 25 municipios y su capital Valledupar representa el 43% su población con un equivalente a 4.973 estudiantes. En la Tabla 3-10 se presentan algunas características de los estudiantes que aplicaron a SABER 11 para el año 2012-2 en este departamento.

Tabla 3-10. Distribución de los estudiantes SABER 11 2012-2 en el Dpto. del Cesar

| MUNICIPIO | TOTAL ESTUDIANTES | EDAD | GENERO | | N° VECES DE PRESENTADO EL EXAMEN | | | | PUESTO | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Media | F | M | 0 | 1 | 2 | 3 | Mínimo | Máximo | Moda |
| AGUACHICA | 1015 | 17 | 601 | 414 | 973 | 30 | 4 | 5 | 4 | 998 | 643 |
| AGUSTIN CODAZZI | 580 | 17 | 326 | 254 | 568 | 10 | 0 | 0 | 7 | 997 | 424 |
| ASTREA | 207 | 18 | 119 | 88 | 206 | 1 | 0 | 0 | 5 | 991 | 859 |
| BECERRIL | 166 | 17 | 75 | 91 | 163 | 2 | 1 | 0 | 64 | 998 | 710 |
| BOSCONIA | 395 | 18 | 215 | 180 | 391 | 1 | 1 | 0 | 59 | 999 | 848 |
| CHIMICHAGUA | 407 | 17 | 209 | 198 | 403 | 0 | 2 | 1 | 23 | 992 | 756 |
| CHIRIGUANA | 296 | 17 | 169 | 127 | 288 | 7 | 0 | 0 | 23 | 998 | 615 |
| CURUMANI | 407 | 17 | 223 | 184 | 404 | 2 | 1 | 0 | 23 | 998 | 580 |
| EL COPEY | 281 | 17 | 148 | 133 | 280 | 1 | 0 | 0 | 28 | 994 | 436 |
| EL PASO | 344 | 18 | 179 | 165 | 340 | 3 | 0 | 0 | 25 | 997 | 689 |
| GAMARRA | 126 | 17 | 60 | 66 | 124 | 0 | 0 | 0 | 87 | 996 | 921 |
| GONZALEZ | 34 | 17 | 17 | 17 | 33 | 0 | 1 | 0 | 151 | 983 | 440 |
| LA GLORIA | 139 | 17 | 75 | 64 | 137 | 2 | 0 | 0 | 31 | 998 | 210 |
| LA JAGUA DE IBIRICO | 390 | 17 | 202 | 188 | 345 | 38 | 1 | 0 | 40 | 1000 | 869 |
| LA PAZ | 243 | 16 | 124 | 119 | 232 | 1 | 8 | 0 | 7 | 996 | 497 |
| MANAURE | 95 | 16 | 56 | 39 | 90 | 3 | 0 | 0 | 8 | 968 | 190 |
| PAILITAS | 191 | 17 | 118 | 73 | 177 | 14 | 0 | 0 | 24 | 994 | 797 |
| PELAYA | 243 | 17 | 123 | 120 | 241 | 2 | 0 | 0 | 7 | 997 | 780 |
| PUEBLO BELLO | 110 | 17 | 58 | 52 | 110 | 0 | 0 | 0 | 24 | 994 | 207 |
| RIO DE ORO | 145 | 17 | 75 | 70 | 140 | 3 | 2 | 0 | 14 | 994 | 428 |
| SAN ALBERTO | 204 | 17 | 115 | 89 | 188 | 14 | 0 | 1 | 23 | 992 | 494 |
| SAN DIEGO | 166 | 17 | 85 | 81 | 155 | 5 | 0 | 0 | 39 | 991 | 156 |
| SAN MARTIN | 140 | 16 | 83 | 57 | 134 | 2 | 0 | 0 | 14 | 987 | 319 |
| TAMALAMEQUE | 225 | 17 | 107 | 118 | 225 | 0 | 0 | 0 | 52 | 995 | 908 |
| VALLEDUPAR | 4973 | 17 | 2764 | 2209 | 4834 | 104 | 4 | 1 | 1 | 1000 | 221 |
| **TOTAL** | **11522** | **425** | **6326** | **5196** | **11181** | **245** | **25** | **8** | | | |

Tabla 3-11. Distribución de los colegios SABER 11 2012-2 en el Dpto. del Cesar.

| MUNICIPIO | NATURALEZA | | CARÁCTER | | | | JORNADA | | | | | VALOR DE PENSIÓN | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NO OFICIAL | OFICIAL | ACADEMICO | ACADEMICO Y TECNICO | NORMALISTA | TECNICO | COMPLETA U ORDINARIA | MAÑANA | NOCHE | SABATINA - DOMINICAL | TARDE | NO PAGA | <$87 .000 | >$87.000 <$120.000 | >$120.000 <$150.000 | >$150.000 <$250.000 | >$250.000 |
| AGUACHICA | 184 | 831 | 522 | 4 | 0 | 489 | 0 | 448 | 5 | 116 | 446 | 819 | 64 | 107 | 13 | 0 | 0 |
| AGUSTIN CODAZZI | 108 | 472 | 268 | 254 | 0 | 58 | 0 | 335 | 50 | 52 | 143 | 480 | 54 | 44 | 0 | 0 | 0 |
| ASTREA | 0 | 207 | 206 | 0 | 0 | 1 | 0 | 120 | 50 | 0 | 37 | 53 | 154 | 0 | 0 | 0 | 0 |
| BECERRIL | 0 | 166 | 135 | 31 | 0 | 0 | 0 | 101 | 34 | 0 | 31 | 165 | 0 | 0 | 0 | 0 | 0 |
| BOSCONIA | 40 | 355 | 326 | 69 | 0 | 0 | 0 | 109 | 78 | 40 | 168 | 332 | 52 | 4 | 2 | 1 | 1 |
| CHIMICHAGUA | 0 | 407 | 222 | 28 | 0 | 157 | 0 | 243 | 25 | 0 | 139 | 403 | 0 | 0 | 0 | 0 | 0 |
| CHIRIGUANA | 14 | 282 | 106 | 145 | 0 | 45 | 13 | 209 | 14 | 14 | 46 | 286 | 9 | 0 | 0 | 0 | 0 |
| CURUMANI | 0 | 407 | 388 | 0 | 0 | 19 | 0 | 216 | 31 | 49 | 111 | 406 | 0 | 0 | 0 | 0 | 0 |
| EL COPEY | 22 | 259 | 168 | 58 | 0 | 55 | 48 | 127 | 50 | 0 | 56 | 256 | 25 | 0 | 0 | 0 | 0 |
| EL PASO | 1 | 343 | 173 | 127 | 0 | 44 | 1 | 277 | 30 | 0 | 36 | 339 | 0 | 0 | 1 | 2 | 1 |
| GAMARRA | 1 | 125 | 22 | 0 | 0 | 104 | 0 | 125 | 0 | 0 | 1 | 123 | 0 | 1 | 0 | 0 | 0 |
| GONZALEZ | 3 | 31 | 4 | 0 | 0 | 30 | 30 | 0 | 0 | 3 | 1 | 31 | 2 | 0 | 0 | 0 | 0 |
| LA GLORIA | 0 | 139 | 24 | 54 | 0 | 61 | 0 | 124 | 15 | 0 | 0 | 139 | 0 | 0 | 0 | 0 | 0 |
| LA JAGUA DE IBIRICO | 1 | 389 | 255 | 105 | 0 | 30 | 0 | 202 | 105 | 1 | 82 | 381 | 0 | 1 | 1 | 0 | 0 |
| LA PAZ | 32 | 211 | 113 | 101 | 0 | 29 | 13 | 191 | 0 | 2 | 37 | 209 | 5 | 20 | 1 | 1 | 5 |
| MANAURE | 1 | 94 | 73 | 0 | 0 | 22 | 72 | 22 | 0 | 1 | 0 | 90 | 3 | 0 | 0 | 0 | 0 |
| PAILITAS | 0 | 191 | 31 | 57 | 0 | 103 | 0 | 160 | 31 | 0 | 0 | 191 | 0 | 0 | 0 | 0 | 0 |
| PELAYA | 0 | 243 | 216 | 1 | 0 | 26 | 1 | 198 | 44 | 0 | 0 | 241 | 2 | 0 | 0 | 0 | 0 |
| PUEBLO BELLO | 3 | 107 | 3 | 42 | 0 | 65 | 65 | 0 | 0 | 3 | 42 | 107 | 3 | 0 | 0 | 0 | 0 |
| RIO DE ORO | 3 | 142 | 106 | 39 | 0 | 0 | 97 | 45 | 0 | 3 | 0 | 141 | 2 | 1 | 0 | 0 | 0 |
| SAN ALBERTO | 38 | 166 | 77 | 127 | 0 | 0 | 0 | 122 | 38 | 0 | 44 | 191 | 9 | 1 | 0 | 0 | 1 |
| SAN DIEGO | 1 | 165 | 66 | 100 | 0 | 0 | 1 | 165 | 0 | 0 | 0 | 159 | 0 | 0 | 0 | 1 | 0 |
| SAN MARTIN | 3 | 137 | 56 | 0 | 0 | 84 | 0 | 93 | 2 | 0 | 45 | 133 | 2 | 1 | 0 | 0 | 0 |
| TAMALAMEQUE | 0 | 225 | 3 | 51 | 0 | 171 | 75 | 148 | 1 | 0 | 1 | 225 | 0 | 0 | 0 | 0 | 0 |
| VALLEDUPAR | 1090 | 3883 | 3202 | 1090 | 1 | 680 | 590 | 2615 | 196 | 181 | 1391 | 3800 | 341 | 180 | 138 | 290 | 191 |
| **TOTAL** | **1545** | **9977** | **6765** | **2483** | **1** | **2273** | **1006** | **6395** | **799** | **465** | **2857** | **9700** | **727** | **360** | **156** | **295** | **199** |

Tabla 3-12. Distribución socioeconómica de los estudiantes SABER 11 2012-2 en el Dpto. del Cesar.

| MUNICIPIO | AREA DONDE VIVE | | NIVEL SISBEN | | | | | INGRESO FAMILIAR MENSUAL | | | | | | | MATERIAL DE LOS PISOS | | | | COMPUTADOR | | INTERNET | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | URBANO | RURAL | 1 | 2 | 3 | OTRO NIVEL | NO CLASIFICADO | <1 SM | 1-2 SM | 2-3 SM | 3-5 SM | 5-7 SM | 7-10 SM | >10 SM | TIERRA | CEMENTO | MADERA | BALDOSA | NO TIENE | SI TIENE | NO TIENE | SI TIENE |
| AGUACHICA | 949 | 66 | 597 | 182 | 5 | 10 | 221 | 354 | 495 | 116 | 38 | 6 | 3 | 3 | 64 | 723 | 10 | 218 | 589 | 403 | 707 | 308 |
| AGUSTIN CODAZZI | 546 | 34 | 490 | 58 | 8 | 3 | 21 | 283 | 239 | 46 | 9 | 3 | 0 | 0 | 26 | 360 | 56 | 138 | 304 | 270 | 331 | 249 |
| ASTREA | 154 | 53 | 196 | 10 | 1 | 0 | 0 | 25 | 156 | 21 | 5 | 0 | 0 | 0 | 51 | 142 | 3 | 11 | 165 | 42 | 185 | 22 |
| BECERRIL | 163 | 3 | 138 | 19 | 1 | 0 | 8 | 66 | 57 | 18 | 23 | 2 | 0 | 0 | 12 | 110 | 0 | 44 | 100 | 63 | 120 | 46 |
| BOSCONIA | 357 | 38 | 316 | 47 | 1 | 3 | 28 | 148 | 190 | 40 | 12 | 1 | 1 | 3 | 36 | 285 | 22 | 52 | 252 | 139 | 297 | 98 |
| CHIMICHAGUA | 209 | 198 | 377 | 17 | 2 | 0 | 11 | 304 | 80 | 18 | 4 | 1 | 0 | 0 | 122 | 271 | 3 | 11 | 332 | 72 | 366 | 41 |
| CHIRIGUANA | 185 | 111 | 246 | 40 | 0 | 0 | 10 | 135 | 121 | 26 | 13 | 0 | 1 | 0 | 53 | 210 | 6 | 27 | 194 | 100 | 238 | 58 |
| CURUMANI | 306 | 101 | 341 | 45 | 1 | 1 | 19 | 235 | 139 | 30 | 3 | 0 | 0 | 0 | 38 | 346 | 3 | 20 | 316 | 89 | 357 | 50 |
| EL COPEY | 238 | 43 | 251 | 18 | 1 | 1 | 10 | 169 | 89 | 20 | 1 | 1 | 0 | 1 | 17 | 242 | 13 | 9 | 228 | 52 | 239 | 42 |
| EL PASO | 106 | 238 | 333 | 2 | 1 | 0 | 8 | 148 | 122 | 32 | 28 | 11 | 2 | 1 | 72 | 225 | 2 | 45 | 222 | 120 | 254 | 90 |
| GAMARRA | 85 | 41 | 110 | 10 | 0 | 0 | 6 | 83 | 34 | 6 | 3 | 0 | 0 | 0 | 28 | 81 | 0 | 17 | 87 | 37 | 103 | 23 |
| GONZALEZ | 17 | 17 | 30 | 4 | 0 | 0 | 0 | 29 | 4 | 0 | 1 | 0 | 0 | 0 | 1 | 33 | 0 | 0 | 31 | 2 | 34 | 0 |
| LA GLORIA | 76 | 63 | 131 | 7 | 1 | 0 | 0 | 53 | 74 | 6 | 4 | 1 | 1 | 0 | 9 | 116 | 0 | 14 | 100 | 39 | 118 | 21 |
| LA JAGUA DE IBIRICO | 340 | 50 | 318 | 54 | 1 | 4 | 13 | 165 | 114 | 64 | 33 | 10 | 4 | 0 | 43 | 280 | 7 | 60 | 182 | 204 | 277 | 113 |
| LA PAZ | 206 | 37 | 174 | 29 | 5 | 7 | 28 | 26 | 174 | 36 | 5 | 2 | 0 | 0 | 9 | 180 | 2 | 52 | 127 | 116 | 167 | 76 |
| MANAURE | 93 | 2 | 87 | 7 | 0 | 0 | 1 | 73 | 16 | 5 | 1 | 0 | 0 | 0 | 11 | 78 | 0 | 6 | 69 | 24 | 85 | 10 |
| PAILITAS | 180 | 11 | 160 | 19 | 5 | 1 | 6 | 123 | 50 | 14 | 4 | 0 | 0 | 0 | 8 | 172 | 2 | 9 | 144 | 45 | 166 | 25 |
| PELAYA | 182 | 61 | 215 | 17 | 0 | 2 | 9 | 109 | 102 | 26 | 4 | 0 | 2 | 0 | 22 | 203 | 2 | 16 | 173 | 69 | 221 | 22 |
| PUEBLO BELLO | 92 | 18 | 83 | 11 | 1 | 0 | 15 | 62 | 36 | 8 | 4 | 0 | 0 | 0 | 17 | 92 | 0 | 1 | 74 | 36 | 98 | 12 |
| RIO DE ORO | 87 | 58 | 110 | 29 | 2 | 0 | 4 | 67 | 65 | 10 | 2 | 1 | 0 | 0 | 10 | 115 | 0 | 20 | 97 | 46 | 129 | 16 |
| SAN ALBERTO | 161 | 43 | 156 | 28 | 3 | 0 | 17 | 101 | 84 | 13 | 4 | 2 | 0 | 0 | 15 | 152 | 5 | 32 | 116 | 86 | 152 | 52 |
| SAN DIEGO | 103 | 63 | 114 | 12 | 0 | 2 | 38 | 85 | 62 | 16 | 3 | 0 | 0 | 0 | 18 | 138 | 4 | 6 | 100 | 59 | 120 | 46 |
| SAN MARTIN | 85 | 55 | 117 | 9 | 2 | 0 | 12 | 70 | 52 | 12 | 3 | 3 | 0 | 0 | 7 | 123 | 0 | 10 | 79 | 58 | 108 | 32 |
| TAMALAMEQUE | 113 | 112 | 223 | 2 | 0 | 0 | 0 | 128 | 80 | 17 | 0 | 0 | 0 | 0 | 36 | 186 | 0 | 3 | 177 | 48 | 183 | 42 |
| VALLEDUPAR | 4317 | 656 | 2243 | 832 | 81 | 27 | 1789 | 1167 | 2404 | 808 | 393 | 88 | 92 | 20 | 165 | 2793 | 140 | 1874 | 2346 | 2577 | 2799 | 2173 |

## 3.5 Verificación de la calidad de los datos

En esta sección se verifican la calidad de los datos correspondientes a los resultados de la Prueba para el Departamento del Cesar.

Para iniciar la verificación en la calidad de los datos se debe tener en cuenta algunos valores de atributos que se consideran No Válidos, esto definido por el diccionario de variables de la Prueba SABER 11 para el periodo 2012-2, a saber son seis atributos: *Tiene computador, Valor de la pensión, Educación de la madre, Educación del padre, Ocupación del padre y Ocupación de la madre*.

En la Tabla 3-13 se realiza la verificación para cada uno de estos atributos que presentan valores No Validos, los cuales no deben ser tenidos en cuenta ya que corresponden a codificaciones de periodos anteriores.

Tabla 3-13. Valores válidos y no válidos definidos por el ICFES en SABER 11 2012-2.

| ATRIBUTO | VALORES VÁLIDOS | FRECUENCIA | VALORES NO VÁLIDOS | FRECUENCIA |
|---|---|---|---|---|
| VALOR PENSIÓN | 0, 8, 9, 10, 11 Y 12 | 11437 | 1, 2, 3 Y 4 | 62 |
| TIENE COMPUTADOR | 0 Y 3 | 11400 | 1 | 121 |
| EDUCACIÓN DEL PADRE | 0, 9, 10, 11, 12, 13, 14, 15, 16, 17 Y 99 | 11384 | 1, 2, 3, 4, 5, 6, 7 Y 8 | 137 |
| EDUCACIÓN DE LA MADRE | 0, 9, 10, 11, 12, 13, 14, 15, 16, 17 Y 99 | 11390 | 1, 2, 3, 4, 5, 6, 7 Y 8 | 131 |
| OCUPACIÓN DEL PADRE | 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 Y 26 | 11350 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Y 12 | 171 |
| OCUPACIÓN DE LA MADRE | 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 Y 26 | 11350 | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Y 12 | 171 |

La verificación de la calidad de los datos específica la revisión de datos perdidos o missings en el conjunto de datos, esto se refiere a la verificación de valores faltantes y en la Tabla 3-14 se muestran los atributos que presentan missings para el Departamento del Cesar.

Tabla 3-14. Verificación datos perdidos o missings para el Dpto. del Cesar Prueba SABER 11 2012-2.

| Atributo | Missings | Observación |
|---|---|---|
| FAMI_COD_EDUCA_PADRE | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| FAMI_COD_EDUCA_MADRE | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| FAMI_COD_OCUP_PADRE | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| FAMI_COD_OCUP_MADRE | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_CUARTOS | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| FAMI_NIVEL_SISBEN | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_MATERIAL_PISOS | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_PERSONAS_HOGAR | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_TELEFONIA | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_CELULAR | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_INTERNET | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_SERVICIO_TV | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_COMPUTADOR | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_LAVADORA | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_NEVERA | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_HORNO | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_DVD | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_MICROHONDAS | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ECON_SN_AUTOMOVIL | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| FAMI_ING_FMILIAR_MENSUAL | 1 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| PLAN_CODIGODANEINSTITUCION | 9 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ESTU_ESTRATO | 11 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| COLE_GENERO_POBLACION | 12 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| INAC_COLEGIOTERMINO | 21 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ESTU_TRABAJA | 24 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| ESTU_VECES_ESTDO | 63 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| COLE_ES_BILINGUE | 1579 | Vacio, en blanco, ?; No puede ser calculado por otro atributo |
| COLE_INST_VLR_PENSION | 23 | Vacio, en blanco, ?; Puede ser calculado por otro atributo (COLE_INST_NOMBRE) |
| ESTU_HORAS_TRABAJO | 11512 | Vacio, en blanco, ?; Puede ser calculado por otro atributo (ESTU_TRABAJA) |
| COMP_FLEX_DESEM | 6795 | Vacio, en blanco, ?; Se refiere a la prueba con el componente flexible interdisciplinario, no se presenta nivel de desempeño |
| INGLES_DESEM | 1 | Vacio, en blanco,?; No existe un desempeño en inglés para puntaje de -1 |

Existen 6 atributos en el conjunto de datos relacionados a algún tipo de discapacidad: *sordoceguera, cognitiva, invidente, motriz, sordointerprete y sordonointerprete*, los cuales contienen registros vacíos pero que se consideran No Nulos, ya que estos campos contienen información relacionada a las discapacidades que pueden presentar los estudiantes y son del tipo SI y NO, y es vacío cuando la respuesta es NO; igual sucede con el atributo *Etnia*, que relaciona la etnia a la que pertenece en la que puede no pertenecer a ninguna etnia; y por último encontramos el atributo *Desempeño componente flexible* en la que se presenta el desempeño de la prueba cuando el estudiante escoge una prueba interdisciplinar.

Existen otros atributos, adicionales a los anteriores, que aunque se tienen la completitud de los valores representan códigos o ID, que designan numeraciones del tipo ordinal y

que pueden confundir al algoritmo de aprendizaje y no son tenidos en cuenta para la implementación del algoritmo, estos pueden verse en la Tabla 3-15.

Tabla 3-15. Atributos de tipo ID en SABER 2012-2.

| Atributo | Decisión de eliminación |
|---|---|
| COLE_CODIGO_COLEGIO<br>CODIGO_DANE<br>PLAN_CODIGODANEINSTITUCION<br><br>ESTU_CODIGO_RESIDE_MCPIO<br>ESTU_EXAM_COD_MPIOPRESENTACION | Es un número que identifica a la institución, asignado por Directorio de Establecimientos Educativos – DUE.<br><br>Es un número que identifica al Municipio. |
| FTP_CONSECUTIVO | Es un número consecutivo que identifica cada registro |
| PERIODO<br>ESTU_EXAM_NOMBREEXAMEN | Son datos que identifican un periodo y nombre de presentación del examen. |

## 3.6 Selección de datos

**Selección de atributos**

Se tiene que en el conjunto de datos existen 75 atributos y varios de ellos son atributos que contienen valores que representan numeraciones consecutivas y que se definen como atributos NO seleccionados, ellos son: periodo, código del municipio de residencia, código DANE de la institución, código del municipio y del departamento de presentación del examen, nombre del examen y código del colegio.

Por otra parte atributos como el tipo de documento, el departamento de residencia, el municipio de presentación del examen, el día, mes y año de nacimiento son atributos considerados no relevantes para el caso en estudio; atributos como las discapacidades representan menos del 0,06% representando un factor no relevante. Planteado esto se tienen en la Tabla 3-16 los atributos seleccionados para aplicar las tareas de procesamiento de datos.

Tabla 3-16. Atributos seleccionados para preprocesamiento y modelamiento

| | Atributo | | |
|---|---|---|---|
| **1** | BIOLOGIA_PUNT | **26** | ECON_SN_SERVICIO_TV |
| **2** | CIENCIAS_SOCIALES_PUNT | **27** | ECON_SN_TELEFONIA |
| **3** | COLE_CALENDARIO_COLEGIO | **28** | ESTU_EDAD |
| **4** | COLE_CARACTER_COLEGIO | **29** | ESTU_ESTRATO |
| **5** | COLE_GENERO_POBLACION | **30** | ESTU_ETNIA |
| **6** | COLE_INST_JORNADA | **31** | ESTU_EXAM_MPIO_PRESENTACION |
| **7** | COLE_INST_NOMBRE | **32** | ESTU_GENERO |
| **8** | COLE_INST_VLR_PENSION | **33** | ESTU_PUESTO |
| **9** | COLE_NATURALEZA | **34** | ESTU_RESIDE_MPIO_PRESENTACION |
| **10** | COMP_FLEX_DESEM | **35** | ESTU_TRABAJA |
| **11** | COMP_FLEX_NOMBRE | **36** | ESTU_VECES_ESTDO |
| **12** | COMP_FLEX_PUNT | **37** | FAMI_COD_EDUCA_MADRE |
| **13** | ECON_AREA_VIVE | **38** | FAMI_COD_EDUCA_PADRE |
| **14** | ECON_CUARTOS | **39** | FAMI_COD_OCUP_MADRE |
| **15** | ECON_MATERIAL_PISOS | **40** | FAMI_COD_OCUP_PADRE |
| **16** | ECON_PERSONAS_HOGAR | **41** | FAMI_ING_FMILIAR_MENSUAL |
| **17** | ECON_SN_AUTOMOVIL | **42** | FAMI_NIVEL_SISBEN |
| **18** | ECON_SN_CELULAR | **43** | FILOSOFIA_PUNT |
| **19** | ECON_SN_COMPUTADOR | **44** | FISICA_PUNT |
| **20** | ECON_SN_DVD | **45** | FTP_CONSECUTIVO |
| **21** | ECON_SN_HORNO | **46** | INGLES_DESEM |
| **22** | ECON_SN_INTERNET | **47** | INGLES_PUNT |
| **23** | ECON_SN_LAVADORA | **48** | LENGUAJE_PUNT |
| **24** | ECON_SN_MICROHONDAS | **49** | MATEMATICAS_PUNT |
| **25** | ECON_SN_NEVERA | **50** | QUIMICA_PUNT |

## 3.7 Limpieza de datos

En la etapa de modelamiento se requieren actividades de limpieza de datos con el fin de alinearse con los requerimientos del algoritmo a aplicar, para este caso el algoritmo k-means es sensible a los datos faltantes; ya vista la sección de verificación de calidad de datos con respecto a los missings o datos faltantes en los que no se puede obtener estos valores por otros atributos y atendiendo al diccionario de variables en los valores para no tener en cuenta, ya que sería difícil de interpretar valores que no han sido previamente identificados, se concluye que el procedimiento a aplicar es ignorar estas instancias o registros, lo que conduce a ignorar 193 registros que constituyen el 1,67% del total correspondiente al Departamento del Cesar, quedando para el proceso de

modelamiento 11329 instancias con 50 atributos seleccionados para el estudio. En la Tabla 3-16 se muestran los resultados de este procedimiento.

## 3.8 Conclusión

En este capítulo se expuso la etapa de entendimiento de los datos y se cumple en parte con el primer objetivo de investigación que comprende los pasos de recolección de datos, exploración inicial y verificación de calidad de datos. Cabe resaltar que esta etapa es crucial, de mayor esfuerzo y la que determina el avance y validez de las demás etapas.

Por esta exploración, se concluye que para el puntaje en matemáticas influye positivamente el ingreso familiar mensual, a mayor ingreso mayor desempeño. No se visualiza una relación directa entre el carácter del colegio y el desempeño de la prueba, a diferencia del valor mensual pensión en que si se presenta una relación positiva de este atributo. Se refleja un alto porcentaje de asociación entre la escolaridad de la madre y el desempeño del alumno.

Por otra parte la verificación de los datos permite conocer el conjunto de datos en aspectos de completitud y validez, siendo el caso de identificar los casos de valores No Válidos o que no se deben tener en cuenta, para esta investigación son registros que no se tendrán en cuenta. La identificación de los atributos con valores perdidos o missing no pueden ser calculados por otros atributos.

Hay que tener presente que en la metodología CRISP-DM y en todo proyecto de minería de datos se tienen procesos iterativos en el cual las etapas de entendimiento del negocio y de los datos son etapas dinámicas que influyen directamente en la etapa de preparación de datos por lo que puede volverse a este capítulo para aplicar los procedimientos de limpieza.

# 4. Modelamiento y evaluación

*En este capítulo se trata la fase de modelamiento, la cual consiste en seleccionar una metodología de minería de datos que pueda aplicarse a los datos suministrados y obtener el resultado esperado, en esta fase puede ser necesario volver a la fase de preparación de datos para ajustar los datos de acuerdo a los requerimientos del algoritmo seleccionado. Se inicia con la selección del algoritmo a aplicar en el modelamiento que permita conseguir los objetivos propuestos en el trabajo.*

Para el desarrollo de este proyecto se aplicará un algoritmo no supervisado que realice agrupamiento de datos por su similitud en algunos atributos.

## 4.1 Selección de técnica de modelado

El modelamiento descriptivo considerado para este proceso se eligió según la literatura del estado del arte y se encuentra acorde a los objetivos de minería de datos propuesto, el algoritmo elegido es el de agrupamiento basado en centroides, el cual es una técnica descriptiva de aprendizaje no supervisado y un método de agrupamiento adaptativo que requiere conocer previamente un valor *k* que define el número de grupos a generar.

La minería de datos en la educación clasifica de manera general tres categorías de aplicación en este campo [48]: predicción, clustering y minería de relación. La predicción se ocupa de encontrar modelos que puedan predecir una variable a partir de otras, el clustering o agrupación permite encontrar grupos con características similares y la minería de relación se emplea en un gran conjunto de datos para encontrar relaciones ocultas entre variables. En la tabla 1.1 se puede apreciar que para los casos del desempeño académico las técnicas más utilizadas son el k-means [2, 10, 35] y los árboles de decisión; siendo el k-means un método de agrupamiento que resulta eficaz y que permite ser utilizado en conjuntos de datos de gran tamaño como el del caso en estudio.

## 4.2 Diseño experimental

Al elegir el algoritmo K-means y su implementación, se utiliza una herramienta computacional como el software RapidMiner, este software de código abierto lidera el campo de herramientas en minería de datos.

Al aplicar el algoritmo K-means al conjunto de datos es necesario que en ellos no existan datos faltantes y que estos se encuentren normalizados en el rango de 0 y 1. En la ejecución del algoritmo K-means se utiliza la fórmula de distancia euclidiana para evaluar la calidad de los grupos hallados y poder así minimizar la distancia interna de los grupos con respecto al centroide de cada uno, sin embargo la utilización del método visual para hallar el k en algunos casos no es claramente visible por lo que se prevé incluir un índice de medida intergrupal como lo es el índice de Davies Bouldin.

Dada las características de la herramienta seleccionada y el algoritmo elegido, el modelo a obtener debe incluir las variables que se encuentran en la base de datos del ICFES para el periodo 2012-2 las cuales contienen variados valores de tipo nominal, ordinal y real que deben ser normalizados para cumplir el requerimiento del algoritmo. Seguidamente en la elaboración del diseño experimental se tuvo en cuenta los 50 atributos seleccionados y los 11329 registros para el Departamento del Cesar.

## 4.3 Construcción del modelo

Construir un modelo descriptivo de minería de datos que se aplique en el campo educativo y que permita encontrar relaciones entre el desempeño académico de la prueba SABER 11 y las condiciones socioeconómicas del estudiante que presenta la prueba, implica seleccionar variables que contienen información del estudiante de tipo económico (15 atributos), familiar (6 atributos), personal (10 atributos), del desempeño académico en la prueba (12 atributos) y del colegio (7 atributos), para un total de 50 atributos que subyacen en la estructura de la base de datos del ICFES y que fueron depurados en la etapa de entendimiento de los datos. Siendo esta prueba de cobertura nacional, para la selección de los registros que representan a los estudiantes del Departamento del Cesar es específico el valor en el atributo de departamento de residencia del estudiante, considerando que el atributo municipio de presentación de la

prueba no es suficiente dada la flexibilidad del ICFES para el sitio de presentación de la prueba; las variables que determinan el colegio al cual pertenece el estudiante, es otro atributo que específica la pertenencia de éste al Departamento del Cesar.

Los estudiantes del Departamento del Cesar que presentaron la Prueba Saber 11 para el periodo 2012-2 corresponden a una población 11.329 estudiantes que cubre a sus 25 municipios incluyendo la capital, el municipio de Valledupar, quien cuenta para ese periodo con 4.897 estudiantes para un 43,23% del total del Departamento, porcentaje bastante alto ya que los otros 24 municipios cubren el 56,77% siendo Agustín Codazzi el siguiente municipio con mayor población con 976 estudiantes. Tomando esta distribución de la población es importante plantear subconjuntos de datos que incluyan o no el municipio de Valledupar en la construcción de los modelos y apreciar de cómo influyen los aspectos de la capital del Departamento como municipio certificado dentro y fuera del grupo tomado, para esto se proponen 3 subconjuntos de datos así: i) todo el Departamento, ii) solo el municipio de Valledupar y iii) el Departamento excluyendo a Valledupar.

La técnica de minería de datos mediante agrupamiento se aplicará en los subconjuntos de datos para establecer si hay diferencias importantes en términos de desempeño de los grupos formados y hallar los factores que determinan la separación de grupos producida por el algoritmo K-means. En consideración, para determinar cuál es el *k* o número de grupos que permite maximizar la agrupación del conjunto de datos se realizan repetidas iteraciones donde el valor *k* se modifica respectivamente y se evalúan los resultados con base al error cuadrático de cada iteración y se utiliza un método llamado el método del codo o Elbow, donde de manera visual se determina el mejor *k*, sin embargo, considerando que se puede llegar a presentar inconvenientes en este método visual, se utiliza adicionalmente un índice de medida interno llamado índice de Davies Bouldin, en donde su valor más bajo permite identificar el *k* ideal para el conjunto de datos.

Para este modelo tomamos valores de *k* entre 2 y 15 y los centroides iniciales son tomados de forma aleatoria con un máximo de 30 iteraciones. En cada iteración para un determinado *k* se recalculan los centroides de los grupos y seguidamente se reasigna

cada instancia a una agrupación para aquel en el que la instancia y el centroide sean lo más cercano. En la Figura 4-1, Figura 4-2 y Figura 4-3 se aprecian las iteraciones realizadas para hallar el *k* en los subconjuntos tomados.

Figura 4-1.Selección del número de grupos - Dpto. del Cesar

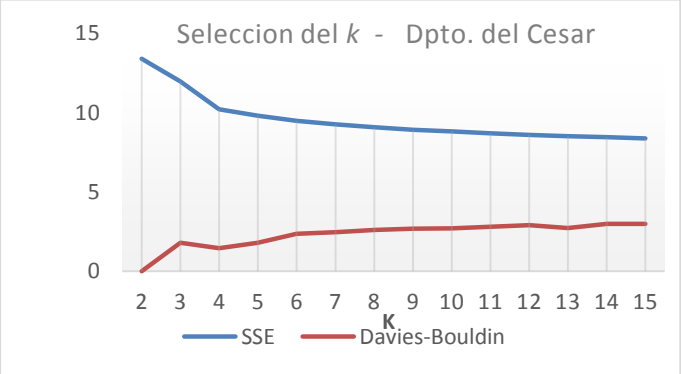



Figura 4-2. Selección del número de grupos - Municipio de Valledupar

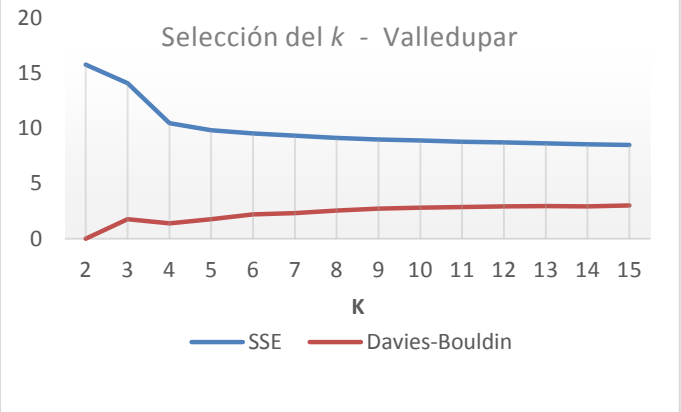
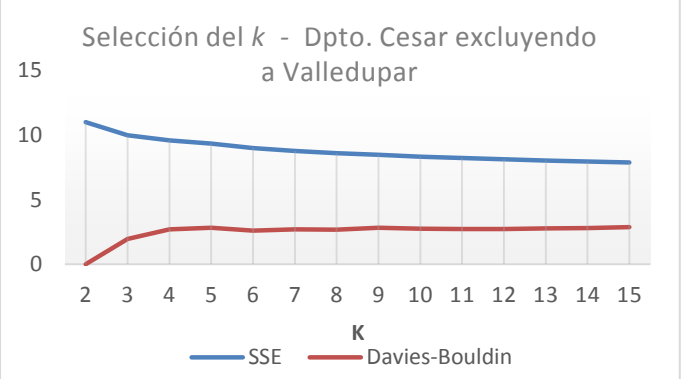



Figura 4-3. Selección del número de grupos - Dpto. del Cesar excluyendo Valledupar



Utilizar el método del codo tomando el SSE determina el k ideal para el conjunto de datos, sin embargo se enriquece la decisión del *k*, incorporando el índice de Davies Bouldin a las gráficas. De acuerdo a esto se toma el *k* para el conjunto de datos en el Departamento del Cesar y el Municipio de Valledupar con un valor de 4, por otra parte la exclusión del Municipio de Valledupar del conjunto de datos del Departamento arroja un *k* con valor de 3. A partir de estos grupos para cada conjunto de datos se espera caracterizar los perfiles de los estudiantes que presentaron la prueba y generar un valor agregado al análisis de los resultados de la Prueba SABER 2012-2 para el Departamento del Cesar.

## 4.4 Evaluación del modelo

La evaluación del modelo se hace por inspección de los grupos generados siendo un proceso subjetivo que requiere interpretar los resultados; sin embargo, teniendo en cuenta que se utilizó el índice de Davies Bouldin que nos da información sobre la separación inter-grupos este nos dice que los ejemplos se encuentran muy cerca de los centroides e internamente son grupos con alta similaridad. Para cada modelo se generó un subconjunto de datos correspondiente al tipo de agrupamiento a evaluar.

### 4.4.1 Agrupamiento Departamento del Cesar

En la Tabla 4-1 se observa que al obtener los grupos, el grupo 3 se caracteriza por tener el mejor desempeño en los puntajes de la prueba SABER 11 para el Departamento del Cesar, para el grupo 0 se encuentra por el contrario el menor desempeño. En la Tabla 4-2 a la

Tabla 4-5 se presentan las características en cada grupo obtenido.

Tabla 4-1. Agrupamiento Departamento del Cesar

| Grupos Cesar | Puesto Promedio | Puntaje Promedio | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Lenguaje | Matemáticas | Ciencias sociales | Filosofía | Biología | Química | Física | Inglés |
| Grupor_0 | 621 | 43,51 | 42,15 | 41,76 | 38,40 | 42,69 | 43,74 | 42,84 | 39,67 |
| Grupo_1 | 607 | 43,82 | 42,31 | 42,01 | 38,61 | 43,00 | 44,11 | 43,33 | 39,60 |
| Grupo_2 | 494 | 45,89 | 45,02 | 44,74 | 40,66 | 45,25 | 45,91 | 44,63 | 42,71 |
| Grupo_3 | 379 | 49,12 | 49,37 | 47,48 | 43,07 | 48,62 | 48,54 | 47,37 | 49,68 |

**Grupo 0**: En este grupo se encuentran los menores niveles de desempeño, y se asemeja con el grupo 1 en cuanto al alto porcentaje de estudiantes que pertenecen a colegios oficiales, los cuales son pocos los que cuentan con computador, internet, teléfono fijo y automóvil, con un 66% de madres de ocupación en el hogar, las cuales un 17% no terminaron la secundaria y solo un 3% alcanzo a terminar estudios superiores; se encuentra en este grupo las familias conformadas hasta por 7 personas (55%), diferenciándose del grupo 1 en que los hogares cuentan con lavadora y que un 53% de estos cuentan con ingresos mensuales entre 1 y 3 salarios mínimos, clasificados en un 86% en el estrato, un 12% en el estrato 2 y un 1% en el estrato 3.

**Grupo 1**: Se caracteriza por agrupar a estudiantes que pertenecen en un 95% a colegios de naturaleza oficial y un 22% son de carácter técnico, a este grupo pertenecen los hogares que no cuentan con lavadora y son pocos los que cuentan con computador, internet, teléfono fijo y automóvil. Son familias medianas que en un 51 % se encuentran conformadas hasta por 7 personas. La educación de sus madres, con respecto a los demás grupos, es donde se ubica el mayor porcentaje en los niveles de primaria completa, donde el 23,05% no culminaron la primaria y el 1,87% terminó estudios universitarios. Es en este grupo que se encuentra el mayor porcentaje (28%) de familias en la zona rural y el 61% de los hogares con ingresos mensuales de 1 salario mínimo con un 87% clasificado en estrato 1, un 12% en el estrato 2 y un 1% en el estrato 3.

**Grupo 2**: Se caracteriza por tener puntajes de mejor desempeño en relación a los grupos 0 y 1, y teniendo características similares a estos en cuanto a pertenecer el 99% a colegios oficiales con un 48% de carácter académico y un 62% perteneciente a la jornada de la mañana; se diferencia de los grupos 0 y 1en cuanto a que la distribución por estrato llega hasta el estrato 4 con el 1%, los porcentajes para el estrato 1 son del 40%, estrato 2 del 49% y el estrato 3 del 10%. Los hogares están conformados por un 42% en familias pequeñas de hasta 4 personas donde el 79% percibe ingresos mensuales entre 1 y 3 salarios mínimos, un 10% entre 3 y 7 salarios mínimos y un 1% con mayores ingresos a esto; se destaca de este grupo los menores porcentajes en cuanto a la no tenencia de computador, internet y teléfono fijo. Los padres alcanzan en un 38% a completar los estudios de secundaria y un 12% culmino una carrera profesional.

**Grupo 3**: El grupo con mejor desempeño en la prueba y que en un 99% pertenece a colegios no oficiales y el 1% a colegios oficiales, con el 49% perteneciente a la jornada completa. Se encuentra en este grupo el mayor porcentaje de estudiantes que presentaron la prueba por segunda ocasión, lo que puede dar a lugar a mejorar su nivel de desempeño. La educación de los padres llega a los niveles profesionales en el mayor porcentaje con respecto a los demás grupos en porcentajes del 38% y 39% para el padre y la madre respectivamente En este grupo se observa el escalonamiento de mayores porcentajes de agrupamiento para los estratos 3 (37%), 4(16%) y 6(1%). El grupo familiar se distribuye en hogares pequeños de hasta 4 personas en el que el 60% cuenta con ingresos mensuales entre 1 y 3 salarios mínimos, encontrando en este grupo el 8% de los hogares con ingresos mayores a 7 salarios mínimos y que cuentan con computador e internet en más del 80%.

Tabla 4-2. Características del colegio - Dpto. del Cesar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 | Grupo_3 |
|---|---|---|---|---|
| Colegio Oficial | 95% | 96% | 99% | 1% |
| Colegio No oficial | 5% | 4% | 1% | 99% |
| Colegio académico | 60% | 59% | 48% | 81% |
| Colegio académico y técnico | 18% | 19% | 33% | 8% |
| Colegio técnico | 21% | 22% | 19% | 10% |
| Jornada completa | 2% | 8% | 2% | 49% |
| Jornada mañana | 60% | 56% | 62% | 22% |
| Jornada Tarde | 25% | 24% | 31% | 11% |
| Jornada noche | 8% | 8% | 5% | 5% |
| Jornada sabatina | 4% | 4% | 0% | 13% |
| No paga pensión | 93% | 94% | 98% | 1% |
| Paga pensión hasta por $150.000 | 7% | 6% | 2% | 60% |
| Paga pensión entre $150.000 y $250.000 | 0% | 0% | 0% | 39% |

Tabla 4-3. Características del estudiante - Dpto. del Cesar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 | Grupo_3 |
|---|---|---|---|---|
| Género femenino | 54% | 57% | 54% | 53% |
| Género masculino | 46% | 43% | 46% | 47% |
| Adolescente | 85% | 83% | 93% | 91% |
| Joven | 11% | 13% | 3% | 4% |
| Adulto | 2% | 2% | 1% | 2% |
| Primera vez del examen | 98% | 99% | 97% | 96% |
| Segunda vez del examen | 2% | 1% | 3% | 4% |
| Tercera vez del examen | 0% | 0% | 0% | 0% |
| No trabaja | 91% | 92% | 97% | 94% |
| Trabaja y no recibe pago | 6% | 6% | 2% | 4% |
| Trabaja y recibe pago | 3% | 3% | 1% | 2% |
| Sisben Nivel 1 | 86% | 84% | 43% | 9% |
| Sisben Nivel 2 | 8% | 8% | 25% | 12% |
| No está clasificado en Sisben | 5% | 8% | 29% | 76% |
| Zona Rural | 25% | 28% | 8% | 3% |
| Zona Urbana | 75% | 72% | 92% | 97% |
| Estrato 1 | 86% | 87% | 40% | 10% |
| Estrato 2 | 12% | 12% | 49% | 34% |
| Estrato 3 | 1% | 1% | 10% | 37% |
| Estrato 4 | 0% | 0% | 1% | 16% |
| Estrato 6 | 0% | 0% | 0% | 1% |

Tabla 4-4.Características de educación de la madre - Dpto. del Cesar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 | Grupo_3 |
|---|---|---|---|---|
| Sin ninguna educación | 4,93% | 5,12% | 1,04% | 0,34% |
| Primaria incompleta | 21,48% | 23,05% | 4,68% | 2,18% |
| Primaria completa | 22,16% | 27,05% | 9,02% | 4,45% |
| Secundaria incompleta | 17,34% | 14,27% | 12,12% | 6,39% |
| Secundaria completa | 23,89% | 22,55% | 37,94% | 23,87% |
| Técnica/Tecnológica incompleta | 0,62% | 0,87% | 2,77% | 3,36% |
| Técnica/Tecnológica completa | 3,70% | 2,07% | 9,32% | 10,34% |
| Profesional incompleta | 0,45% | 0,34% | 2,54% | 3,28% |
| Profesional completa | 2,86% | 1,87% | 11,96% | 38,82% |
| Postgrado | 0,14% | 0,20% | 2,27% | 4,79% |
| No sabe educación de la madre | 2,44% | 2,63% | 6,35% | 2,18% |

Tabla 4-5. Características del grupo familiar - Dpto. del Cesar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 | Grupo_3 |
|---|---|---|---|---|
| Familia pequeña hasta 4 personas | 31% | 37% | 42% | 52% |
| Familia mediana hasta 7 personas | 55% | 51% | 51% | 45% |
| Familia grande hasta 12 personas | 13% | 12% | 7% | 3% |
| Ingreso familiar mensual de 1SM | 45% | 61% | 11% | 4% |
| Ingreso familiar mensual entre 1 y 3 SM | 53% | 38% | 79% | 60% |
| Ingreso familiar mensual entre 3 y 7 SM | 2% | 1% | 10% | 28% |
| Ingreso familiar mensual desde 7 SM | 0% | 0% | 1% | 8% |
| Piso de Cemento, gravilla, ladrillo | 83% | 79% | 49% | 26% |
| Piso de Madera pulida, baldosa, tableta, mármol, alfombra | 8% | 5% | 45% | 70% |
| Piso de Tierra, arena | 8% | 15% | 1% | 1% |
| No tiene automóvil | 88% | 96% | 63% | 42% |
| No tiene celular | 3% | 5% | 3% | 2% |
| No tiene computador | 82% | 91% | 8% | 9% |
| No tiene internet | 96% | 97% | 21% | 20% |
| No tiene lavadora | 0% | 100% | 15% | 12% |
| No tiene nevera | 8% | 26% | 2% | 2% |
| No tiene servicio cerrado tv | 50% | 72% | 23% | 21% |
| No tiene teléfono fijo | 92% | 91% | 43% | 22% |

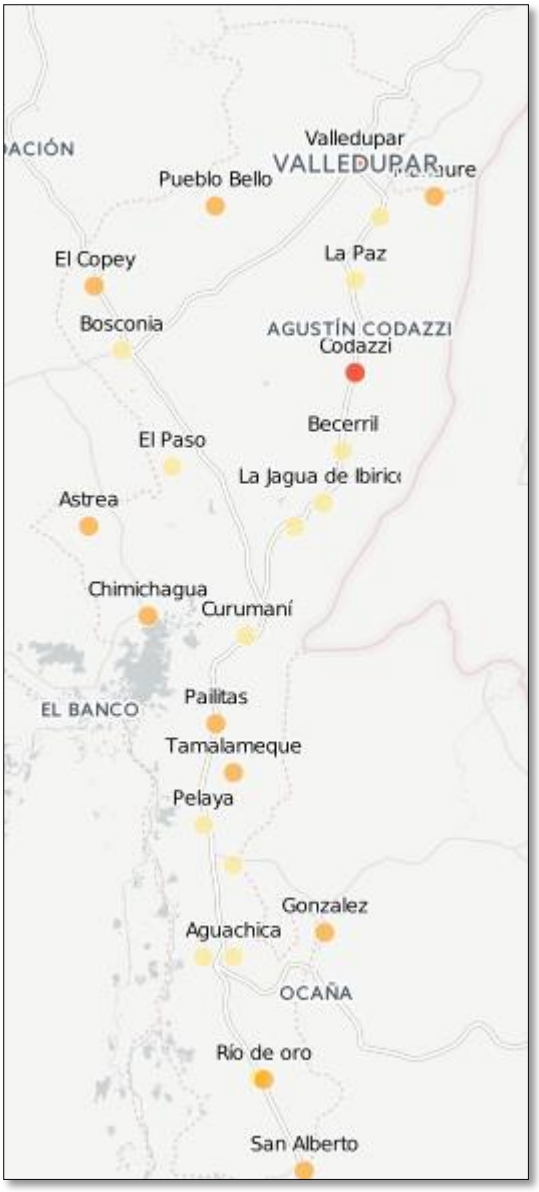
En la Tabla 4-6 se observa la distribución por municipios en cada grupo obtenido a nivel del Departamento. Al hacer un análisis de los datos contenidos en esta tabla se aprecia que el mayor porcentaje de los municipios del Cesar se ubican en el grupo 0, el cual corresponde al grupo con menor nivel de desempeño en la prueba, para el grupo 3 que corresponde al grupo con mejor nivel de desempeño se ubican pocos municipios y el que mayor porcentaje lo representa el municipio de Valledupar que como entidad certificada y capital del Departamento aporta con un 20% de los estudiantes a este grupo que presentaron la prueba, le sigue en participación el Municipio de Aguachica con un 10,5%, La Paz con un 11,34% y Agustín Codazzi con un 7,34%, cabe destacar que estos municipios a nivel departamental le siguen en importancia después de la capital.

Para el grupo 0 se ubican la mayoría de los municipios, siendo este grupo en donde se ubican la mayor parte de la zona rural.

Tabla 4-6. Distribución de municipios del Dpto. del Cesar por grupos obtenidos

| MUNICIPIO | N° Estudiantes | Grupo_0 | | Grupo_1 | | Grupo_2 | | Grupo_3 | |
|---|---|---|---|---|---|---|---|---|---|
| | | % | N° Estudiantes del municipio | % | N° Estudiantes del municipio | % | N° Estudiantes del municipio | % | N° Estudiantes del municipio |
| AGUACHICA | 976 | 33% | 323 | 31% | 307 | 25% | 243 | 11% | 103 |
| AGUSTIN CODAZZI | 572 | 18% | 105 | 37% | 209 | 38% | 216 | 7% | 42 |
| ASTREA | 207 | 43% | 90 | 46% | 96 | 10% | 21 | 0% | 0 |
| BECERRIL | 162 | 43% | 69 | 27% | 43 | 31% | 50 | 0% | 0 |
| BOSCONIA | 391 | 47% | 182 | 29% | 112 | 20% | 80 | 4% | 17 |
| CHIMICHAGUA | 402 | 44% | 176 | 49% | 197 | 7% | 29 | 0% | 0 |
| CHIRIGUANA | 292 | 50% | 147 | 28% | 83 | 21% | 61 | 0% | 1 |
| CURUMANI | 405 | 53% | 216 | 34% | 139 | 12% | 50 | 0% | 0 |
| EL COPEY | 280 | 25% | 69 | 61% | 171 | 10% | 29 | 4% | 11 |
| EL PASO | 342 | 52% | 177 | 25% | 86 | 23% | 78 | 0% | 1 |
| GAMARRA | 124 | 46% | 57 | 38% | 47 | 15% | 19 | 1% | 1 |
| GONZALEZ | 33 | 3% | 1 | 97% | 32 | 0% | 0 | 0% | 0 |
| LA GLORIA | 139 | 57% | 79 | 27% | 38 | 16% | 22 | 0% | 0 |
| LA JAGUA DE IBIRICO | 382 | 48% | 184 | 16% | 62 | 36% | 136 | 0% | 0 |
| LA PAZ | 238 | 39% | 92 | 20% | 48 | 30% | 71 | 11% | 27 |
| MANAURE | 93 | 13% | 12 | 73% | 68 | 14% | 13 | 0% | 0 |
| PAILITAS | 185 | 36% | 67 | 49% | 91 | 15% | 27 | 0% | 0 |
| PELAYA | 241 | 49% | 118 | 41% | 98 | 10% | 25 | 0% | 0 |
| PUEBLO BELLO | 110 | 37% | 41 | 49% | 54 | 13% | 14 | 1% | 1 |
| RIO DE ORO | 141 | 29% | 41 | 50% | 70 | 20% | 28 | 1% | 2 |
| SAN ALBERTO | 200 | 32% | 64 | 40% | 79 | 28% | 56 | 1% | 1 |
| SAN DIEGO | 156 | 41% | 64 | 29% | 45 | 29% | 46 | 1% | 1 |
| SAN MARTIN | 136 | 47% | 64 | 32% | 43 | 21% | 28 | 1% | 1 |
| TAMALAMEQUE | 225 | 43% | 97 | 46% | 104 | 11% | 24 | 0% | 0 |
| VALLEDUPAR | 4897 | 21% | 1035 | 26% | 1253 | 33% | 1628 | 20% | 981 |
| **TOTAL** | **11329** | **3570** | | **3575** | | **1190** | | **2994** | |

Figura 4-4.Mapa de distribución de grupos en el Departamento del Cesar




## 4.4.2 Agrupamiento Municipio de Valledupar

En la  se observa que al obtener los grupos, el grupo 3 se caracteriza por tener el mejor desempeño en los puntajes de la prueba SABER 11 para el Municipio de Valledupar, para el grupo 0 se encuentra por el contrario el menor desempeño.  En la Tabla 4-2 a la se presentan las características en cada grupo obtenido.

Tabla 4-7. Agrupamiento Municipio de Valledupar

| Grupos Valledupar | Puesto Promedio | Puntaje Promedio | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Lenguaje | Matemáticas | Ciencias sociales | Filosofía | Biología | Química | Física | Ingles |
| Grupo_0 | 603 | 43,86 | 41,97 | 42,51 | 39,28 | 43,00 | 44,08 | 43,20 | 39,67 |
| Grupo_1 | 533 | 45,06 | 43,71 | 43,74 | 40,48 | 44,73 | 45,52 | 44,12 | 41,66 |
| Grupo_2 | 467 | 46,53 | 45,48 | 45,43 | 41,43 | 45,69 | 46,41 | 45,30 | 43,57 |
| Grupo_3 | 328 | 50,32 | 50,84 | 48,74 | 44,19 | 49,84 | 49,57 | 48,45 | 51,75 |

**Grupo 0**: El grupo con menor nivel de desempeño. Se caracteriza este grupo por pertenecer el 96% a colegios oficiales con un 75% de carácter académico con un 62% en la jornada de la mañana. El 89% de los estudiantes se encuentra en Sisben Nivel 1y es entre todos los grupo el de mayor porcentaje (25%) perteneciente a la zona rural clasificado en un 97% como estrato 1. Los hogares en un 55% se encuentran conformados por familias de hasta 7 personas con el 46% de ellos percibiendo ingresos mensuales de 1 salario mínimo, un 53% entre 1 y 3 salarios mínimos y solo un 1% entre 3 y 7 salarios mínimos. Este grupo presenta los mayores porcentajes en cuanto a la carencia de bienes y servicios como la no tenencia de automóvil (93%), de computador (88%) e internet (96%) como el de telefonía fija que para el caso se encuentra que el 86% del grupo no tiene.

**Grupo 1**: En este grupo el 91% pertenece a colegios oficiales con un 44% de carácter académico con el 54% perteneciente a la jornada de la mañana y el 39% en la jornada de la tarde. El 18% y 47% se encuentran en Sisben Nivel 1 y 2 respectivamente; la distribución por estratos se encuentra en el estrato 1 (8%), estrato 2 (80%) y estrato 3 (11%). Los hogares se encuentran conformados por familias medianas en un 46% y perciben ingresos mensuales de 1 salario mínimo (20%), entre 1 y 3 salarios mínimos (77%) y entre 3 y 7 salarios mínimos el 3%. Es uno de los mayores porcentajes en la carencia de bienes y servicios como el computador (77%), el internet (98%) y telefonía fija (66%). Sobresale que el 43% de las madres y el 44% de los padres tienen una educación secundaria completa y solo un 5% y 6% respectivamente alcanzo a culminar estudios universitarios; el 48% de las madres se ocupan exclusivamente de las ocupaciones del hogar.

**Grupo 2**: En este grupo observamos que el 97% pertenece a colegios oficiales, con un 49% de carácter académico y un 61% se encuentra en la jornada de la mañana, el 93% de los estudiantes se encuentra en la zona urbana distribuidos en los estratos 1 (33%), estrato 2(49%), estrato 3(17%) y un 2% en el estrato 4. Los hogares se encuentran conformados en un 54% por familias medianas de hasta 7 personas con ingresos familiares mensuales entre 1 y 3 salarios mínimos en el 79% y un 1% percibe más de 7 salarios mínimos, son bastantes bajos los porcentajes relacionados con la no tenencia de bienes y servicios como el computador(2%) e internet (7%). El 42% de los padres

termino la secundaria y un 10% culmino estudios superiores, solo un 4% tiene nivel de primaria incompleta; el 43% de las madres se dedica a ocupaciones en el hogar junto a un 30% de padres que trabajan por cuenta propia. Este grupo presenta el segundo mejor desempeño en la prueba.

**Grupo 3**: Se observa en este grupo el mejor desempeño, superando hasta por 4 puntos el promedio nacional. Solo el 1% pertenece a colegios oficiales, con un 88% de carácter académico y el 64% en jornada completa, situándose el 52% en pagos de pensión entre $150.000 y $250.000. Se distribuyen en los niveles de estrato 1 (5%), estrato 2 (26%), estrato 3 (44%), estrato4 (20%) y estrato 6 (1%). Los hogares se encuentran conformados familias pequeñas (49%) y familias medianas (48%) con ingresos mensuales en un 54% entre 1 y 3 salarios mínimos, el 34% percibe entre 3 y 7 salarios mínimos mensuales y un 11% más de 7 salarios mínimos; los padres poseen en un 45% educación profesional completa y solo un 2% no alcanzo a culminar la primaria, la ocupación de la madre en el hogar alcanza un 28%, con madre y padre que trabajan por cuenta propia en un 12% y 15% respectivamente. El 80% de los pisos que constituyen la vivienda es de madera pulida, baldosa, mármol o alfombra. El hecho que el 5% del grupo presentó la prueba por segunda vez puede haber influido en los resultados de desempeño.

Tabla 4-8. Características del colegio - Municipio de Valledupar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 | Grupo_3 |
|---|---|---|---|---|
| Colegio Oficial | 96% | 91% | 97% | 1% |
| Colegio No oficial | 4% | 9% | 3% | 99% |
| Colegio académico | 75% | 44% | 49% | 88% |
| Colegio académico y técnico | 11% | 20% | 19% | 5% |
| Colegio técnico | 14% | 36% | 32% | 7% |
| Jornada completa | 0% | 1% | 0% | 64% |
| Jornada mañana | 62% | 54% | 61% | 19% |
| Jornada Tarde | 28% | 39% | 36% | 4% |
| Jornada noche | 5% | 3% | 3% | 6% |
| Jornada sabatina | 5% | 3% | 0% | 7% |
| No paga pensión | 95% | 91% | 95% | 1% |
| Paga pensión hasta por $150.000 | 5% | 9% | 5% | 47% |
| Paga pensión entre $150.000 y $250.000 | 0% | 0% | 0% | 52% |

Figura 4-5 Colegios por grupo - Valledupar

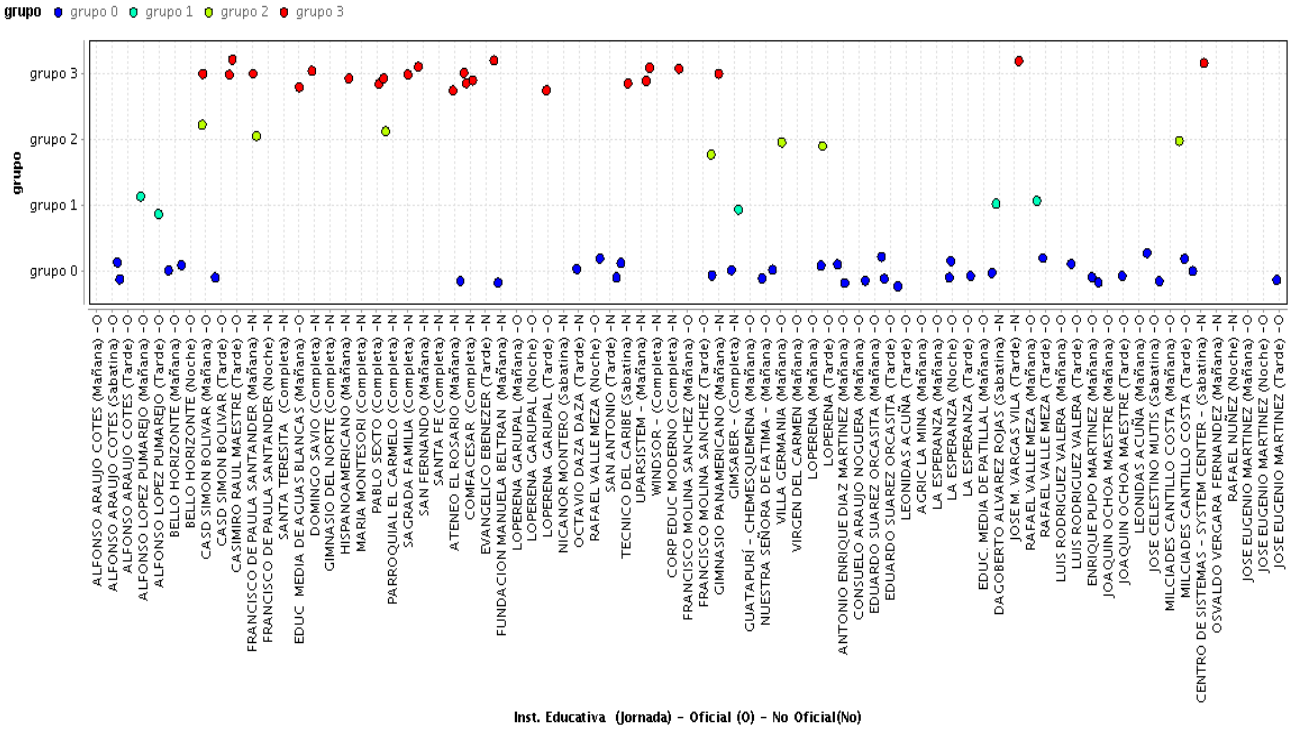



Figura 4-6.Distribución porcentual en el estrato 1 de los colegios - Valledupar

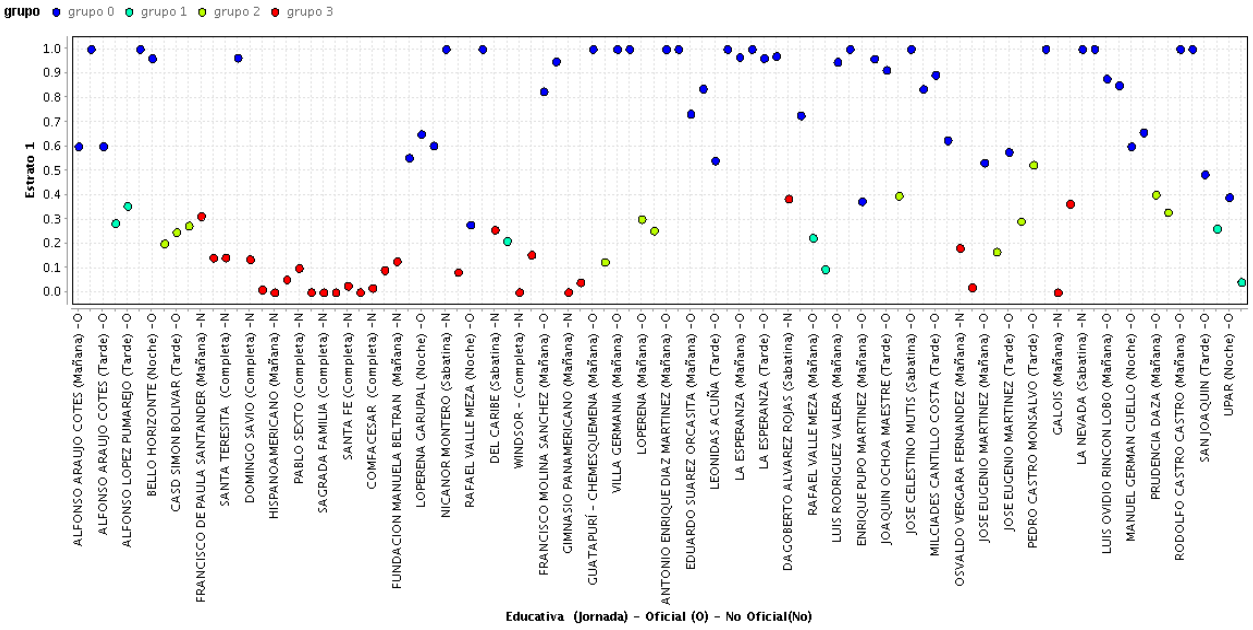


Tabla 4-9. Características del estudiante - Municipio de Valledupar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 | Grupo_3 |
|---|---|---|---|---|
| Género femenino | 56% | 55% | 57% | 53% |
| Género masculino | 44% | 45% | 43% | 47% |
| Adolescente | 86% | 94% | 93% | 93% |
| Joven | 11% | 4% | 4% | 3% |
| Adulto | 2% | 0% | 1% | 2% |
| Primera vez del examen | 99% | 98% | 98% | 95% |
| Segunda vez del examen | 1% | 2% | 2% | 5% |
| Tercera vez del examen | 0% | 0% | 0% | 0% |
| No trabaja | 93% | 97% | 97% | 97% |
| Trabaja y no recibe pago | 4% | 2% | 2% | 2% |
| Trabaja y recibe pago | 3% | 2% | 1% | 1% |
| Sisben Nivel 1 | 89% | 18% | 35% | 3% |
| Sisben Nivel 2 | 1% | 47% | 23% | 7% |
| No está clasificado en Sisben | 9% | 33% | 39% | 87% |
| Zona Rural | 25% | 10% | 7% | 2% |
| Zona Urbana | 75% | 90% | 93% | 98% |
| Estrato 1 | 97% | 8% | 33% | 5% |
| Estrato 2 | 2% | 80% | 49% | 26% |
| Estrato 3 | 0% | 11% | 17% | 44% |
| Estrato 4 | 0% | 0% | 2% | 20% |
| Estrato 6 | 0% | 0% | 0% | 1% |

Tabla 4-10. Características de educación de los padres - Municipio de Valledupar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 | Grupo_3 |
|---|---|---|---|---|
| **Educación de la madre** | | | | |
| Sin ninguna educación | 3,69% | 1,02% | 0,88% | 0,23% |
| Primaria incompleta | 16,99% | 4,85% | 4,04% | 1,58% |
| Primaria completa | 23,35% | 12,30% | 8,51% | 3,38% |
| Secundaria incompleta | 19,55% | 13,54% | 12,18% | 4,28% |
| Secundaria completa | 27,84% | 42,78% | 41,97% | 19,26% |
| Técnica/Tecnológica incompleta | 0,97% | 1,13% | 0,95% | 3,60% |
| Técnica/Tecnológica completa | 3,24% | 8,35% | 8,73% | 11,15% |
| Profesional incompleta | 0,34% | 0,90% | 1,91% | 3,27% |
| Profesional completa | 1,25% | 4,51% | 9,68% | 45,83% |
| Postgrado | 0,06% | 0,23% | 1,17% | 4,84% |
| No sabe educación de la madre | 2,73% | 10,38% | 9,98% | 2,59% |
| **Educación del padre** | | | | |
| Sin ninguna educación | 6,82% | 1,47% | 1,54% | 0,45% |
| Primaria incompleta | 16,59% | 5,19% | 4,40% | 2,36% |
| Primaria completa | 20,80% | 10,61% | 8,80% | 3,83% |
| Secundaria incompleta | 17,56% | 10,27% | 9,24% | 4,62% |
| Secundaria completa | 26,14% | 43,57% | 42,77% | 19,93% |
| Técnica/Tecnológica incompleta | 0,74% | 2,26% | 1,10% | 2,25% |
| Técnica/Tecnológica completa | 2,78% | 6,32% | 6,68% | 9,01% |
| Profesional incompleta | 0,57% | 0,79% | 1,83% | 3,38% |
| Profesional completa | 2,67% | 6,09% | 10,86% | 44,82% |
| Postgrado | 0,11% | 0,00% | 1,17% | 5,52% |
| No sabe educación del padre | 5,23% | 13,43% | 11,59% | 3,83% |

Tabla 4-11. Características del grupo familiar - Municipio de Valledupar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 | Grupo_3 |
|---|---|---|---|---|
| Familia pequeña hasta 4 personas | 34% | 43% | 36% | 49% |
| Familia mediana hasta 7 personas | 55% | 46% | 54% | 48% |
| Familia grande hasta 12 personas | 11% | 11% | 9% | 3% |
| Ingreso familiar mensual de 1SM | 46% | 20% | 11% | 2% |
| Ingreso familiar mensual entre 1 y 3 SM | 53% | 77% | 79% | 54% |
| Ingreso familiar mensual entre 3 y 7 SM | 1% | 3% | 10% | 34% |
| Ingreso familiar mensual desde 7 SM | 0% | 0% | 1% | 11% |
| Piso de Cemento, gravilla, ladrillo | 83% | 61% | 45% | 17% |
| Piso de Madera pulida, baldosa, tableta, mármol, alfombra | 7% | 35% | 50% | 80% |
| Piso de Tierra, arena | 8% | 1% | 1% | 0% |
| No tiene automóvil | 93% | 88% | 65% | 34% |
| No tiene celular | 4% | 5% | 3% | 2% |
| No tiene computador | 88% | 77% | 2% | 6% |
| No tiene internet | 96% | 98% | 7% | 14% |
| No tiene lavadora | 56% | 43% | 16% | 8% |
| No tiene nevera | 14% | 7% | 1% | 0% |
| No tiene servicio cerrado tv | 66% | 58% | 25% | 18% |
| No tiene teléfono fijo | 86% | 66% | 28% | 17% |

### 4.4.3 Agrupamiento Departamento del Cesar excluyendo el Municipio de Valledupar

Se muestran en la Tabla 4-12 los resultados obtenidos para el subconjunto de datos del Departamento del Cesar excluyendo Valledupar.

Tabla 4-12. Agrupamiento Departamento del Cesar excluyendo Valledupar

| Agrupamiento Cesar sin municipio de Valledupar | Puesto Promedio | Puntaje Promedio | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Lenguaje | Matemáticas | Ciencias sociales | Filosofía | Biología | Química | Física | Ingles |
| Grupo_0 | 636 | 43,3 | 42,0 | 41,3 | 38,0 | 42,4 | 43,5 | 42,6 | 39,3 |
| Grupo_1 | 622 | 43,6 | 42,1 | 41,5 | 38,1 | 42,7 | 43,8 | 43,1 | 39,4 |
| Grupo_2 | 518 | 45,4 | 45,0 | 44,1 | 39,7 | 44,9 | 45,5 | 44,2 | 42,2 |

En la Tabla 4-12 se observa que al obtener los grupos, el grupo 2 se caracteriza por tener el mejor desempeño en los puntajes de la Prueba SABER 11 para el Departamento del Cesar excluyendo al Municipio de Valledupar, para el grupo 0 se encuentra por el contrario, el menor desempeño. En la Tabla 4-13 a la Tabla 4-16 se presentan las características en cada grupo obtenido.

**Grupo 0**: Siendo el grupo de menor nivel de desempeño en la prueba se caracteriza por que el 95% del grupo pertenece a colegios oficiales, con un 55% de carácter académico, solo un 3% es de jornada completa y el 61% de la jornada de la mañana; el 90% se encuentra en estrato 1 con un 27% en la zona rural y un 89% está clasificada en Sisben Nivel 1, el 11% de los estudiantes trabaja siendo este el mayor porcentaje de todos los grupos y solo el 3% de estos recibe un pago por su trabajo. El 55% de las familias está conformada por hasta 7 personas, percibiendo ingresos familiares en el 51% de los casos de un salario mínimo, el 47% percibe entre 1 y 3 salarios mínimos alcanzando solo un 2% a recibir ingresos entre 3 y 7 salarios mínimos; se caracteriza por que todos los hogares cuentan con lavadora y carecen de bienes y servicios como el automóvil (89%), el computador (84%), internet (97%) y telefonía fija (95%). Los porcentajes en el nivel de estudios primarios incompleto se encuentran en 25% y 27% para la madre y el padre respectivamente, con un 23% de primaria completa para ambos padres. El más alto porcentaje de ocupación de la madre en el hogar se encuentra en este grupo con un 69%, mientras que en el padre el 48% se dedica a trabajar por cuenta propia.

**Grupo 1**: En este grupo encontramos los hogares que no cuentan con lavadora y carecen de bienes y servicios con los más altos porcentajes, como el automóvil (96%), computador (92%), internet (98%), telefonía fija (96%); el 20% de los pisos de sus viviendas está hecho de tierra y arena, donde el 69% de las familias reciben ingresos mensuales de 1 salario mínimo, y el 30% entre 1 y 3 salarios mínimos; en este grupo se encuentran el 91% de los estudiantes sisbenizados en el nivel 1 con el más alto porcentaje (29%) que viven en la zona rural estratificada y el 92% en estrato1. Los niveles de educación de los padres son muy similares al grupo 0. Es de resaltar que el 12% de los colegios de este grupo se encuentra en la jornada completa.

**Grupo 2**: Este agrupamiento es el de mejor nivel de desempeño cerca del promedio nacional para esta aplicación. El 87% de los colegios de este grupo son oficiales y el 13% son no oficiales donde el 14% paga pensión hasta por $150.000 y un 1% entre $150.000 y $250.000. El 3% realizó el examen por segunda vez y solo el 10% vive en la zona rural, el 54% pertenece a estrato 1, el 41% a estrato 2 y un 4% al estrato 3. El 50% de las familias están conformadas por hasta 4 personas, el 45% por hasta 7 personas donde el 12% percibe ingresos mensuales de 1 salario mínimo, el 77% entre 1 y 3 salarios mínimos. Los porcentajes en los niveles de educación de los padres son bajos para lo que son estudios primarios (entre el 6% y el 13%) y son más altos los de educación profesional en relación a los otros grupos (16% para la madre y 14% para el padre), el 43% de las madres se ocupan exclusivamente de las tareas del hogar, el 19% trabaja por cuenta propia y el 9% está empleada a nivel técnico o profesional.

Tabla 4-13. Características del colegio - Dpto. del Cesar excluyendo Valledupar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 |
|---|---|---|---|
| Colegio Oficial | 95% | 96% | 87% |
| Colegio No oficial | 5% | 4% | 13% |
| Colegio académico | 55% | 55% | 53% |
| Colegio académico y técnico | 19% | 19% | 27% |
| Colegio técnico | 26% | 26% | 20% |
| Jornada completa | 3% | 12% | 4% |
| Jornada mañana | 61% | 53% | 58% |
| Jornada Tarde | 21% | 21% | 26% |
| Jornada noche | 5% | 4% | 5% |
| Jornada sabatina | 10% | 11% | 7% |
| No paga pensión | 92% | 94% | 85% |
| Paga pensión hasta por $150.000 | 8% | 6% | 14% |
| Paga pensión entre $150.000 y $250.000 | 0% | 0% | 1% |

Tabla 4-14. Características del estudiante - Dpto. del Cesar excluyendo Valledupar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 |
|---|---|---|---|
| Género femenino | 54% | 56% | 53% |
| Género masculino | 46% | 44% | 47% |
| Adolescente | 84% | 81% | 91% |
| Joven | 12% | 15% | 5% |
| Adulto | 2% | 2% | 1% |
| Primera vez del examen | 98% | 98% | 96% |
| Segunda vez del examen | 2% | 1% | 3% |
| Tercera vez del examen | 0% | 0% | 0% |
| No trabaja | 89% | 91% | 96% |
| Trabaja y no recibe pago | 8% | 7% | 3% |
| Trabaja y recibe pago | 3% | 3% | 1% |
| Sisben Nivel 1 | 89% | 91% | 58% |
| Sisben Nivel 2 | 7% | 5% | 21% |
| No está clasificado en Sisben | 3% | 4% | 18% |
| Zona Rural | 27% | 29% | 10% |
| Zona Urbana | 73% | 71% | 90% |
| Estrato 1 | 90% | 92% | 54% |
| Estrato 2 | 10% | 8% | 41% |
| Estrato 3 | 0% | 0% | 4% |
| Estrato 4 | 0% | 0% | 0% |
| Estrato 6 | 0% | 0% | 0% |

Tabla 4-15. Características de los padres - Dpto. del Cesar excluyendo Valledupar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 |
|---|---|---|---|
| **Educación de la madre** | | | |
| Sin ninguna educación | 5,85% | 6,11% | 1,40% |
| Primaria incompleta | 24,80% | 27,87% | 6,13% |
| Primaria completa | 22,58% | 30,22% | 10,16% |
| Secundaria incompleta | 15,82% | 12,87% | 12,91% |
| Secundaria completa | 22,13% | 16,92% | 31,66% |
| Técnica/Tecnológica incompleta | 0,66% | 0,57% | 4,38% |
| Técnica/Tecnológica completa | 2,68% | 1,53% | 9,29% |
| Profesional incompleta | 0,49% | 0,22% | 3,33% |
| Profesional completa | 3,13% | 1,92% | 15,65% |
| Postgrado | 0,12% | 0,26% | 3,86% |
| No sabe educación de la madre | 1,73% | 1,53% | 1,23% |
| **Educación del padre** | | | |
| Sin ninguna educación | 9,48% | 8,72% | 2,45% |
| Primaria incompleta | 27,44% | 29,39% | 9,35% |
| Primaria completa | 23,16% | 30,05% | 10,51% |
| Secundaria incompleta | 12,65% | 11,21% | 12,68% |
| Secundaria completa | 18,05% | 13,30% | 30,08% |
| Técnica/Tecnológica incompleta | 0,33% | 0,13% | 4,09% |
| Técnica/Tecnológica completa | 1,40% | 1,35% | 8,41% |
| Profesional incompleta | 0,62% | 0,17% | 3,04% |
| Profesional completa | 2,97% | 2,27% | 13,55% |
| Postgrado | 0,25% | 0,22% | 2,98% |
| No sabe educación del padre | 3,67% | 3,18% | 2,86% |

Tabla 4-16. Características del grupo familiar - Dpto. del Cesar excluyendo Valledupar

| ATRIBUTO | Grupo_0 | Grupo_1 | Grupo_2 |
|---|---|---|---|
| Familia pequeña hasta 4 personas | 31% | 35% | 50% |
| Familia mediana hasta 7 personas | 55% | 52% | 45% |
| Familia grande hasta 12 personas | 14% | 13% | 5% |
| Ingreso familiar mensual de 1SM | 51% | 69% | 12% |
| Ingreso familiar mensual entre 1 y 3 SM | 47% | 30% | 77% |
| Ingreso familiar mensual entre 3 y 7 SM | 2% | 1% | 11% |
| Ingreso familiar mensual desde 7 SM | 0% | 0% | 1% |
| Piso de Cemento, gravilla, ladrillo | 84% | 77% | 57% |
| Piso de Madera pulida, baldosa, tableta, mármol, alfombra | 6% | 2% | 35% |
| Piso de Tierra, arena | 10% | 20% | 2% |
| No tiene automóvil | 89% | 96% | 58% |
| No tiene celular | 2% | 6% | 2% |
| No tiene computador | 84% | 92% | 4% |
| No tiene internet | 97% | 98% | 23% |
| No tiene lavadora | 0% | 100% | 13% |
| No tiene nevera | 10% | 29% | 3% |
| No tiene servicio cerrado tv | 47% | 71% | 20% |
| No tiene teléfono fijo | 95% | 96% | 56% |

# 5. Conclusiones y recomendaciones

## 5.1 Conclusiones

Este proyecto tuvo como finalidad realizar un análisis del desempeño académico en el Examen de Estado para el ingreso a la Educación Superior, conocido en la actualidad como SABER 11, aplicando técnicas de minería de datos en educación, siguiendo la metodología CRISP-DM y la aplicación de algoritmo de agrupamientos. El principal objetivo es aportar información que permita describir e identificar las características que diferencian los grupos hallados en el conjunto de datos correspondiente a la aplicación del examen para el año 2012-2 en el Departamento del Cesar – Colombia.

En este trabajo, el alto volumen de datos y las características propias del algoritmo k-means determinaron un reto en la fase de preparación de datos al aplicarles las actividades de comprensión y preprocesamiento que exige el algoritmo y poder así construir un modelo descriptivo que caracterice los grupos de estudiantes que obtuvieron los diferentes niveles de desempeño en la prueba.

Se obtuvieron resultados consistentes con la literatura consultada, en donde se observa que el entorno socioeconómico del estudiante incide en los resultados de su desempeño académico, encontrándose que a mayor nivel socioeconómico del estudiante y su familia, mayor es el puntaje en los resultados de la prueba. Es interesante observar que los grupos con una alta población en la zona rural tienden a presentar menor nivel de desempeño y son grupos en los cuales se presenta que el nivel de escolaridad de los padres llega solo a un nivel de escolaridad de básica primaria y en pocos casos de secundaria.

Un hallazgo interesante es el bajo rendimiento en el área de filosofía para el Departamento del Cesar, en los que prevalece para todos los grupos, el más bajo nivel de desempeño de los resultados. Se evidencia también que los estudiantes con padres con estudios en posgrado tienen mayores posibilidades de obtener mayores puntajes en la prueba SABER 11.

Los objetivos planteados fueron abordados en su totalidad al aplicar los siguientes procesos:

- Se utilizó CRISP- DM como metodología para proyecto de minería de datos.
- Se conformó el conjunto de datos para la aplicación del algoritmo k-means.
- Se realizó la preparación de datos requerida por el algoritmo implementado.
- Se implementó la técnica k-means que posibilitó el agrupamiento sin supervisión del conjunto de datos y obtener así el modelo descriptivo.
- Se realizó una evaluación sistemática del modelo propuesto y un análisis de los resultados obtenidos.

## 5.2 Recomendaciones

El tema tratado en esta investigación plantea variados interrogantes que pueden generar iniciativas para otros trabajos de investigación desde diferentes disciplinas; de igual forma se puede integrar este modelo para emprender estudios como el de la deserción y el desempeño académico, por ejemplo, en los programas de educación superior.

A futuro se puede fortalecer esta investigación con la inclusión de datos relacionados con los recursos didácticos o multimediales con que cuenta el colegio, el tamaño de los grupos por clase y a la actividad docente; de este modo se puede obtener resultados más precisos con relación a grupos similares en sus características socioeconómicas y que generan diferente niveles de desempeño académico.

La aplicación de otras técnicas de minería de datos, como las reglas de asociación y/o los árboles de decisión, permitirá obtener modelos alternativos que faciliten entender de mejor manera los patrones y las interrelaciones exhibidas en los datos.

# A. Anexo: Glosario

| | |
|---|---|
| EDM | Educacional Data Mining |
| CRISP – DM | Cross Industry Standard Process for Data Mining |
| MEN | Ministerio de Educación - Colombia |
| ICFES | Instituto Colombiano para la Evaluación de la Educación |
| SABER 11 | Examen de Estado de la educación Media - Colombia. |
| CARTODB | Sitio Web para el diseño de mapas personalizados, https://cartodb.com/ |

# B. Anexo: Diccionario de variables

| N° | VARIABLE | CATEGORÍA/RANGO/EJEMPLO | | DESCRIPCION |
|---|---|---|---|---|
| 1 | PERIODO | 20122 | | Periodo de aplicación del examen |
| 2 | ESTU_CONSECUTIVO | ejemplo: | "SABER11201 22005707" | Código de identificación de la persona en la base de datos de Investigación |
| 3 | ESTU_EDAD | rango: | [10,90] | Edad del inscrito |
| 4 | ESTU_TIPODOCUMENTO | valores posibles: | 6 | Tipo de documento de identidad del inscrito |
| | | C | | Cédula de ciudadanía |
| | | E | | Cédula de extranjería |
| | | P | | Pasaporte extranjero |
| | | R | | Certificado de registraduría |
| | | T | | Tarjeta de identidad |
| | | Q | | Pasaporte colombiano |
| 5 | ESTU_PAIS_RESIDE | ejemplo: | "CO" | Código del país de residencia del inscrito con codificación ISO 3166 alpha-2 |
| 6 | ESTU_GENERO | valores posibles: | 2 | Género del inscrito |
| | | F | | Femenino |
| | | M | | Masculino |
| 7 | ESTU_NACIMIENTO_DIA | valores posibles: | [1,31] | Día de nacimiento del inscrito |
| 8 | ESTU_NACIMIENTO_MES | valores posibles: | [1,12] | Mes de nacimiento del inscrito |
| 9 | ESTU_NACIMIENTO_ANNO | valores posibles: | [1930,2003] | Año de nacimiento del inscrito |
| 10 | ESTU_ETNIA | valores posibles: | 18 | Etnia a la que pertenece el inscrito |
| | | 1 | | Comunidades negras |
| | | 2 | | Raizal (isleño) |
| | | 3 | | Paez |
| | | 4 | | Sikuani |
| | | 5 | | Arhuaco |
| | | 6 | | Emberá |
| | | 7 | | Guambiano |
| | | 8 | | Pijao |
| | | 9 | | Wayúu |
| | | 10 | | Zenú |
| | | 11 | | Pasto |
| | | 12 | | Cancuamo |
| | | 13 | | Inga |
| | | 14 | | Tucano |
| | | 15 | | Huitoto |
| | | 16 | | Cubeo |
| | | 17 | | Comunidad Rom(gitana) |

| | | | | |
|---|---|---|---|---|
| | | 99 | | Otro |
| 11 | ESTU_DISC_BAJAVISION | B | | Indicador de discapacidad - Baja visión |
| 12 | ESTU_DISC_SORDOCEGUERA | C | | Indicador de discapacidad - Sordoceguera |
| 13 | ESTU_DISC_COGNITIVA | G | | Indicador de discapacidad - Cognitiva |
| 14 | ESTU_DISC_INVIDENTE | I | | Indicador de discapacidad - Invidente |
| 15 | ESTU_DISC_MOTRIZ | M | | Indicador de discapacidad - Motriz |
| 16 | ESTU_DISC_SORDOINTERPRETE | R | | Indicador de discapacidad - Sordo y requiere intérprete de señas |
| 17 | ESTU_DISC_SORDONOINTERPRETE | S | | Indicador de discapacidad - Sordo y NO requiere intérprete de señas |
| 18 | ESTU_CODIGO_RESIDE_MCPIO | valores posibles: | [5001,95001] | Código del municipio de residencia del estudiante |
| 19 | ESTU_RESIDE_MPIO | ejemplo: | "BOGOTÁ D.C." | Municipio de residencia del inscrito |
| 20 | ESTU_RESIDE_DEPT | ejemplo: "NARIÑO" | | Departamento de residencia del inscrito |
| 21 | ESTU_ZONA_RESIDE | valores posibles: | [1,10] | Zona de residencia del inscrito |
| | | 1 | | Norte |
| | | 2 | | Oriente |
| | | 3 | | Occidente |
| | | 4 | | Sur |
| | | 5 | | Centro |
| | | 6 | | Nororiente |
| | | 7 | | Suroriente |
| | | 8 | | Noroccidente |
| | | 9 | | Suroccidente |
| | | 10 | | Unica |
| 22 | ECON_AREA_VIVE | valores posibles: | [1,2] | Área donde vive el inscrito |
| | | 1 | | Cabecera municipal |
| | | 2 | | Rural |
| 23 | COLE_CODIGO_COLEGIO | valores posibles: | [182,170621] | Código asignado a la institución |
| 24 | COLE_CODIGO_DANE_ESTAB | ejemplo: "115403000207" | | Código DANE asignado al establecimiento educativo |
| 25 | COLE_CODIGO_DANE_SEDE | ejemplo: "115403000011" | | Código DANE asignado a la sede de la institución |
| 26 | COLE_INST_NOMBRE | ejemplo: | "COL COMPUSOCIAL" | Nombre del colegio en que terminó |
| 27 | COLE_ZONALOCALIZACION | valores posibles: | [1,10] | Zona de presentación del examen |
| | | 1 | | Norte |
| | | 2 | | Oriente |
| | | 3 | | Occidente |
| | | 4 | | Sur |
| | | 5 | | Centro |
| | | 6 | | Nororiente |
| | | 7 | | Suroriente |
| | | 8 | | Noroccidente |
| | | 9 | | Suroccidente |
| | | 10 | | Unica |
| 28 | COLE_UBICACIONPLANTEL | valores posibles: 2 | | Ubicación del plantel en el municipio |
| | | R | | Zona rural |
| | | U | | Perímetro urbano |

| | | | | |
|---|---|---|---|---|
| 29 | COLE_CALENDARIO_COLEGIO | valores posibles: | 3 | Calendario del colegio |
| | | A | | Calendario A |
| | | B | | Calendario B |
| | | F | | Calendario flexible |
| 30 | COLE_GENERO_POBLACION | valores posibles: | 3 | Población del colegio |
| | | F | | Femenino |
| | | M | | Masculino |
| | | X | | Mixto |
| 31 | COLE_NATURALEZA | valores posibles: | 2 | Naturaleza del colegio |
| | | N | | No oficial |
| | | O | | Oficial |
| 32 | COLE_ES_BILINGUE | valores posibles: | 2 | La institución educativa es bilingüe |
| | | 0 | | No |
| | | 1 | | Si |
| 33 | COLE_INST_JORNADA | valores posibles: | 5 | Jornada del colegio |
| | | COMPLETA U ORDINARIA | | Jornada completa u ordinaria |
| | | MAÑANA | | Jornada mañana |
| | | NOCHE | | Jornada noche |
| | | SABATINA-DOMINICAL | | Jornada sabatina-dominical |
| | | TARDE | | Jornada tarde |
| 34 | COLE_CARACTER_COLEGIO | valores posibles: | 5 | Carácter del colegio |
| | | ACADEMICO | | Carácter académico |
| | | AVADEMICO Y TECNICO | | Carácter académico y técnico |
| | | DESCONOCIDO | | Carácter desconocido |
| | | NORMALIST | | Carácter normalista |
| | | TECNICO | | Carácter técnico |
| 35 | COLE_INST_VLR_PENSION | valores posibles: 7 | | Valor de la pensión pagada por el estudiante en el último año |
| | | 0 | | No Paga Pensión |
| | | 8 | | Menos de 87.000 Pesos |
| | | 9 | | Entre 87.000 y menos de 120.000 Pesos |
| | | 10 | | Entre 120.000 y menos de 150.000 Pesos |
| | | 11 | | Entre 150.000 y menos de 250.000 Pesos |
| | | 12 | | Entre 250.000 pesos o más |
| | | *valores entre 1 y 7 | | no deben ser tenidos en cuenta, corresponden a codificaciones de periodos anteriores |
| 36 | ESTU_VECES_ESTDO | rango: | [0,4] | Veces que el estudiante ha presentado el examen de estado |
| 37 | ESTU_EXAM_COD_MPIO PRESENTACION | valores posibles: | [5001,95001] | Código de municipio de presentación del examen |
| 38 | ESTU_EXAM_MPIO_PRESENTACION | ejemplo: | "BOGOTÁ D.C." | Municipio de presentación del examen |
| 39 | ESTU_EXAM_DEPT_PRESENTACION | ejemplo: "NARIÑO" | | Departamento de presentación del examen |
| 40 | ESTU_EXAM_NOMBREEXAMEN | EXAMEN SABER 11- 2012 CAL A | | Nombre del examen presentado |
| 41 | FAMI_COD_EDUCA_PADRE / | valores posibles: 11 | | Máximo nivel educativo alcanzado por el padre |
| | | 0 | | Ninguno |

| | | | | |
|---|---|---|---|---|
| | FAMI_COD_EDUCA_MADRE | 9 | | Primaria incompleta |
| | | 10 | | Primaria completa |
| | | 11 | | Secundaria (bachillerato) incompleta |
| | | 12 | | Secundaria (bachillerato) completa |
| | | 13 | | Educación técnica o tecnológica incompleta |
| | | 14 | | Educación técnica o tecnológica completa |
| | | 15 | | Educación profesional incompleta |
| | | 16 | | Educación profesional completa |
| | | 17 | | Postgrado |
| | | 99 | | No sabe |
| | | *valores entre 2 y 8 | | no deben ser tenidos en cuenta, corresponden a codificaciones de periodos anteriores |
| 42 | FAMI_COD_OCUP_PADRE / FAMI_COD_OCUP_MADRE | valores posibles: 12 | | Ocupación del padre |
| | | 13 | | Empresario |
| | | 14 | | Pequeño empresario |
| | | 15 | | Empleado con cargo como director o gerente general |
| | | 16 | | Empleado de nivel directivo |
| | | 17 | | Empleado de nivel técnico o profesional |
| | | 18 | | Empleado de nivel auxiliar o administrativo |
| | | 19 | | Empleado obrero u operario |
| | | 20 | | Profesional Independiente |
| | | 21 | | Trabajador por cuenta propia |
| | | 22 | | Hogar |
| | | 23 | | Pensionado |
| | | 26 | | Otra actividad u ocupación |
| | | *valores entre 0 y 12, 24 y 25 | | no deben ser tenidos en cuenta, corresponden a codificaciones de periodos anteriores |
| 43 | ESTU_ESTRATO | valores posibles: 7 | | Estrato socioeconómico de la residencia del estudiante según factura de energía |
| | | 1 | | Estrato 1 |
| | | 2 | | Estrato 2 |
| | | 3 | | Estrato 3 |
| | | 4 | | Estrato 4 |
| | | 5 | | Estrato 5 |
| | | 6 | | Estrato 6 |
| | | 8 | | Vive en una zona rural donde no hay estratificación económica |
| 44 | ECON_CUARTOS | rango: | [1,10] | Número de habitaciones de la residencia |
| 45 | FAMI_NIVEL_SISBEN | valores posibles: 5 | | Nivel de SISBEN en que está clasificada la familia |
| | | 1 | | Nivel 1 |
| | | 2 | | Nivel 2 |
| | | 3 | | Nivel 3 |
| | | 4 | | Clasificado en otro nivel |
| | | 5 | | No está clasificado |
| 46 | ECON_MATERIAL_PISOS | valores posibles: 4 | | Material de los pisos que predomina en la vivienda |
| | | 1 | | Tierra, arena |
| | | 2 | | Cemento, gravilla, ladrillo |
| | | 3 | | Madera burda, tabla o tablón |
| | | 4 | | Madera pulida, baldosa, tableta, mármol, alfombra |

| | | | | |
|---|---|---|---|---|
| 47 | ECON_PERSONAS_HOGAR | rango: | [1,12] | Número de personas que conforman el hogar |
| 48 | ECON_SN_TELEFONIA | valores posibles: 2 | | El hogar cuenta con servicio de teléfono fijo |
| | | 0 | | No |
| | | 1 | | Si |
| 49 | ECON_SN_CELULAR | valores posibles: 2 | | Cantidad de celulares con que cuenta su hogar |
| | | 0 | | No |
| | | 1 | | Uno |
| 50 | ECON_SN_INTERNET | valores posibles: 2 | | El hogar cuenta con conexión a internet |
| | | 0 | | No |
| | | 1 | | Si |
| 51 | ECON_SN_SERVICIO_TV | valores posibles: 2 | | El hogar cuenta con Servicio cerrado de televisión |
| | | 0 | | No |
| | | 1 | | Si |
| 52 | ECON_SN_COMPUTADOR | valores posibles: 2 | | Tiene computador en su hogar |
| | | 0 | | No |
| | | 3 | | Si |
| | | *valores 1 y 2 | | no deben ser tenidos en cuenta, corresponden a codificaciones de periodos anteriores |
| 53 | ECON_SN_LAVADORA | valores posibles: 2 | | El hogar cuenta con lavadora |
| | | 0 | | No |
| | | 1 | | Si |
| 54 | ECON_SN_NEVERA | valores posibles: 2 | | El hogar cuenta con nevera o enfriador |
| | | 0 | | No |
| | | 1 | | Si |
| 55 | ECON_SN_HORNO | valores posibles: 2 | | El hogar cuenta con horno eléctrico o a gas |
| | | 0 | | No |
| | | 1 | | Si |
| 56 | ECON_SN_DVD | valores posibles: 2 | | Cantidad de reproductores DVD con que cuenta su hogar |
| | | 0 | | No |
| | | 1 | | Uno |
| 57 | ECON_SN_MICROHONDAS | valores posibles: 2 | | El hogar cuenta con horno microondas |
| | | 0 | | No |
| | | 1 | | Si |
| 58 | ECON_SN_AUTOMOVIL | valores posibles: 2 | | Cantidad de automóviles particulares con que cuenta su hogar |
| | | 0 | | No |
| | | 1 | | Uno |
| 59 | FAMI_ING_FMILIAR_MENSUAL | valores posibles: 7 | | Ingresos mensuales representado en salarios mínimos mensuales |
| | | 1 | | Menos de 1 SM |
| | | 2 | | Entre 1 y Menos de 2 SM |
| | | 3 | | Entre 2 y Menos de 3 SM |
| | | 4 | | Entre 3 y Menos de 5 SM |
| | | 5 | | Entre 5 y Menos de 7 SM |
| | | 6 | | Entre 7 y Menos de 10 SM |
| | | 7 | | 10 o más SM |
| 60 | ESTU_TRABAJA | valores posibles: 3 | | Trabaja actualmente |
| | | 0 | | No |

| | | | | | |
|---|---|---|---|---|---|
| | | 6 | | Si y recibe algún pago o salario por trabajar | |
| | | 7 | | Si y no recibe pago o salario por trabajar | |
| 61 | ESTU_HORAS_TRABAJO | rango: | [8,20] | Número de horas que trabaja a la semana | |
| 62 | LENGUAJE_PUNT | numérico | | Puntaje en lenguaje | |
| 63 | MATEMATICAS_PUNT | numérico | | Puntaje en matemáticas | |
| 64 | CIENCIAS_SOCIALES_PUNT | numérico | | Puntaje en ciencias sociales | |
| 65 | FILOSOFIA_PUNT | numérico | | Puntaje en filosofía | |
| 66 | BIOLOGIA_PUNT | numérico | | Puntaje en biología | |
| 67 | QUIMICA_PUNT | numérico | | Puntaje en química | |
| 68 | FISICA_PUNT | numérico | | Puntaje en física | |
| 69 | INGLES_PUNT | numérico | | Puntaje en inglés | |
| 70 | INGLES_DESEM | valores posibles: | 6 | Desempeño en inglés según las bandas del Marco Común Europeo | |
| | | B+ | | Supera al nivel B1 | |
| | | B1 | | Pre-intermedio | |
| | | A2 | | Básico | |
| | | A1 | | Principiante | |
| | | A- | | Nivel inferior | |
| 71 | COMP_FLEX_NOMBRE | valores posibles: | 6 | Nombre de componente flexible (profundizaciones o interdisciplinar) | |
| | | PROFUNDIZACIÓN EN BIOLOGÍA | | Profundización en biología | |
| | | PROFUNDIZACIÓN EN CIENCIAS SOCIALES | | Profundización en ciencias sociales | |
| | | PROFUNDIZACIÓN EN LENGUAJE | | Profundización en lenguaje | |
| | | PROFUNDIZACIÓN EN MATEMÁTICA | | Profundización en matemática | |
| | | MEDIO AMBIENTE | | Medio ambiente | |
| | | VIOLENCIA Y SOCIEDAD | | Violencia y sociedad | |
| 72 | COMP_FLEX_PUNT | numérico | | | Puntaje en componente flexible |
| 73 | COMP_FLEX_DESEM | valores posibles: | | 4 | Desempeño en componente flexible |
| | | GB | | | Nivel de desempeño GB |
| | | I | | | Nivel de desempeño I |
| | | II | | | Nivel de desempeño II |
| | | III | | | Nivel de desempeño III |
| 74 | ESTU_PUESTO | rango: | | [1,1000] | Puesto general examen |

# Bibliografía


[1] L. Affendey y I. Paris, «Ranking of Influencing Factors in Predicting Students' Academic Performance», *Information Technology Journal*, 2010.

[2] P. D. Antonenko, S. Toy, y D. S. Niederhauser, «Using cluster analysis for data mining in educational technology research», *Educational Technology Research and Development*, vol. 60, n.º 3, pp. 383-398, feb. 2012.

[3] P. Baepler y C. J. Murdoch, «Academic analytics and data mining in higher education», *International Journal for the Scholarship of Teaching and Learning*, 2010.

[4] D. M. Baker Ryan, «Home | International Educational Data Mining Society», 2011. [En línea]. Disponible en: http://www.educationaldatamining.org/.

[5] J. D. Barón, «La brecha de rendimiento académico de Barranquilla», 2010.

[6] F. Barrera-Osorio, «Calidad de la educación básica y media en Colombia: diagnóstico y propuesta», 2012.

[7] M. Beikzadeh, «Data mining application in higher learning institutions», *Informatics in Education*, vol. 7, n.º 1, pp. 31-54, 2008.

[8] M. Bienkowski, M. Feng, y B. Means, «Enhancing teaching and learning through educational data mining and learning analytics: An issue brief», *Department of Education's (ED) Office of Educational Technology*, pp. 1-57, 2012.

[9] P. Bradley, U. Fayyad, y C. Reina, «Scaling EM (Expectation-Maximization) Clustering to Large Databases», 1998.

[10] L. L. P. Cadena, «Aplicando minería de datos al marketing educativo», 2011.

[11] T. Calders y M. Pechenizkiy, «Introduction to the special section on educational data mining», *ACM SIGKDD Explorations Newsletter*, vol. 13, n.º 2, p. 3, may 2012.

[12] M. Castro y M. Lizasoain, «Las técnicas de modelización estadística en la investigación educativa: minería de datos, modelos de ecuaciones estructurales y modelos jerárquicos lineales», *Revista Española de Pedagogía*, pp. 131-149, 2012.

[13] R. Chaturvedi y C. I. Ezeife, «Mining the Impact of Course Assignments on Student Performance», *educationaldatamining.org*.

[14] P. Cortez y A. Silva, «Using data mining to predict secondary school student performance», *University of Minho*, vol. 2003, n.º 2000, 2008.

[15] I. Davidson, «Understanding K-Means Non-hierarchical Clustering», 2002.

[16] D. L. Davies y D. W. Bouldin, «A cluster separation measure.», *IEEE transactions on pattern analysis and machine intelligence*, vol. 1, n.º 2, pp. 224-227, 1979.

[17] J. L. Díaz, M. Herrera, J. Izquierdo, y R. Pérez-garcía, «The tasks of pre and post-processing in Data Mining applied to a real world problem», *International Environmental Modelling and Software Society*, 2010.

[18] Á. Galindo y H. García, «Minería de Datos en la Educación», *universidad Carlos II de Madrid*, 2010.

[19] D. Garcıa-Saiz y M. Zorrilla, «Towards the development of a classification service for predicting students' performance», *educationaldatamining.org*.

[20] A. Gaviria y J. Barrientos, «Características del plantel y calidad de la educación en Bogotá», *Coyuntura social*, n.º 25, pp. 81-98, 2001.

[21] P. Golding y O. Donaldson, «Predicting academic performance», *Frontiers in Education Conference*, pp. 21-26, 2006.

[22] L. Guarín, «Data Mining Model to Predict Academic Performance at the Universidad Nacional de Colombia» Maestría thesis, Universidad Nacional de Colombia. *bdigital.unal.edu.co*. 2013

[23] W. Hämäläinen, «Descriptive and Predictive Modelling Techniques for Educational Technology», *Licentiate thesis, Department of Computer Science, …*, 2006.

[24] J. Han, *Data mining : concepts and techniques*, Second Edi. 2012.

[25] M. Jiménez, «La predicción del rendimiento académico: regresión lineal versus regresión logística», *Psicothema*, vol. 12, 2000.

[26] S. P. Jiménez, J. J. Puldón, y R. A. E. Andrade, «Modelo clustering para el análisis en la ejecución de procesos de negocio», *Investigacion Operacional*, vol. 33, n.º 3, pp. 210-221, 2012.

[27] S. Kardan y C. Conati, «A framework for capturing distinguishing user interaction behaviours in novel interfaces», en *EDM 2011 - Proceedings of the 4th International Conference on Educational Data Mining*, 2011, pp. 159-168.

[28] S. G. Kulkarni, G. C. Rampure, y B. Yadav, «Understanding Educational Data Mining ( EDM )», *international journal of electronics and computer science engineering*, pp. 773-777, 1956.

[29] A. Kumar, «Implication Of Classification Techniques In Predicting Student's Recital», *International Journal of Data Mining & Knowledge Management Process*, vol. 1, n.º 5, pp. 41-51, sep. 2011.

[30] S.-H. Liao, P.-H. Chu, y P.-Y. Hsiao, «Data mining techniques and applications – A decade review from 2000 to 2011», *Expert Systems with Applications*, vol. 39, n.º 12, pp. 11303-11311, sep. 2012.

[31] S. Liu, J. Kim, S. Macskassy, y E. Shaw, «Predicting Group Programming Project Performance using SVN Activity Traces», *educationaldatamining.org*.

[32] M. Lopez, J. Luna, C. Romero, y S. Ventura, «Classification via Clustering for Predicting Final Marks Based on Student Participation in Forums.», *International Educational Data …*, pp. 4-7, 2012.

[33] J. Luan, «Data Mining and Knowledge Management in Higher Education-Potential Applications.», 2002.

[34] C. Marquez-Vera, «Predicting School Failure Using Data Mining», *… Data Mining, …*, n.º December, 2011.

[35] A. Merceron y K. Yacef, «Educational data mining: a case study», *Proceeding of the 2005 conference on Artificial*, 2005.

[36] A. Merceron y K. Yacef, «Interestingness measures for association rules in educational data», en *The 1st International Conference on Educational Data Mining*, 2008.

[37] B. Minaei-bidgoli, D. A. D. A. Kashy, G. Kortemeyer, y W. F. W. F. Punch, «Predicting student performance : an application of data mining methods with the educational web-based system lon-capa», en *33rd Annual Frontiers in Education, 2003. FIE 2003.*, 2003, vol. 1, pp. 1-6.

[38] R. Navarro, «El rendimiento académico: concepto, investigación y desarrollo», *… sobre Calidad, Eficacia y Cambio en Educación*, vol. 1, 2003.

[39] E. Ogor, «Student academic performance monitoring and evaluation using data mining techniques», *Electronics, Robotics and Automotive Mechanics …*, pp. 0-5, 2007.

[40] E. Olmos, «El rendimiento estudiantil: una metodología para su medición», *Revista Economía*, vol. 13, pp. 7-25, 1997.

[41] J. Orjuela, «Determinantes individuales de desempeño en las pruebas de estado para educación media y superior en colombia 1», *Saber Investigar*, pp. 1-13, 2008.

[42] U. Pandey y B. Bhardwaj, «Data Mining as a Torch Bearer in Education Sector», *arXiv preprint arXiv:1201.5182*, pp. 115-125, 2012.

[43] R. T. Pereira, «Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos», *iiis.org*, 2007.

[44] M. Ramaswami y R. Bhaskaran, «A CHAID based performance prediction model in educational data mining», *arXiv preprint arXiv:1002.1144*, vol. 7, n.º 1, pp. 10-18, 2010.

[45] C. Romero y S. Ventura, «Educational data mining: A survey from 1995 to 2005», *Expert Systems with Applications*, vol. 33, n.º 1, pp. 135-146, jul. 2007.

[46] C. Romero y S. Ventura, «Educational data mining: a review of the state of the art», *Systems, Man, and Cybernetics, Part C: …*, vol. 40, n.º 6, pp. 601-618, 2010.

[47] C. Romero, S. Ventura, P. Espejo, y C. Hervás, «Data Mining Algorithms to Classify Students.», *educationaldatamining.org*, 2008.

[48] R. B. Sachin y M. S. Vijay, «A Survey and Future Vision of Data Mining in Educational Field», *2012 Second International Conference on Advanced Computing & Communication Technologies*, pp. 96-100, ene. 2012.

[49] A. Sarmiento Gomez, «Educación, compromiso de todos. Situación de la Educación preescolar, básica, media y superior en Colombia.» .

[50] R. W. Sembiring, S. Sembiring, y J. M. Zain, «An efficient dimensional reduction method for data clustering», *Bulletin of Mathematics*, vol. 04, n.º 01, pp. 43-58, 2012.

[51] J. Superby, «Determination of factors influencing the achievement of the first-year university students using data mining methods», *educationaldatamining.org*, 2006.

[52] P. Tan, «Introduction to data mining», vol. Chap 8, pp. 1-108, 2007.

[53] N. Thai-Nghe y L. Drumond, «Factorization techniques for predicting student performance», *and Challenges (In …*, pp. 1-30, 2011.

[54] P. Timarán, «Detección de patrones de bajo rendimiento académico y deserción estudiantil con técnicas de minería de datos», *Memorias de la VIII Conferencia Iberoamericana en…*, 2009.

[55] J. Tourón, «La predicción del rendimiento académico: procedimientos, resultados e implicaciones», *Revista Española de Pedagogía*, 1985.

[56] S. pal3 Umesh Kumar Pandey1, Brijesh Kumar Bhardwaj2, «Data Mining as a Torch Bearer in Education Sector», *Technical Journal of LBSIMDS*, 2005.

[57] UNESCO, «Datos Mundiales de Educación VII Ed. 2010/11», 2010.

[58] S. K. Yadav, B. Bharadwaj, y S. Pal, «Mining Education Data to Predict Student's Retention: A comparative Study», *International Journal of Computer Science and Information Security*, p. 5, mar. 2012.

[59] S. Zhang, C. Zhang, y Q. Yang, *Data preparation for data mining*, vol. 17, n.º 5-6. 2003.

[60] R. S. J. D. R. Baker y K. Yacef, «The state of educational data mining in 2009: A review and future visions», *Journal of Educational Data Mining*, 2009.

[61] La Red Martínez, D. L., Karanik, M., giovannini, M., y Pinto, N. Perfiles de Rendimiento Académico: Un Modelo basado en Minería de datos. Campus Virtuales, Vol. IV, num. 1, pp. 12-30. Consultado el [12/11/2015] en www.revistacampusvirtuales.es 2015.

[62] Harwati,, Ardita Permata Alfiani, Febriana Ayu Wulanda. Mapping Student's Performance Based on Data Mining Approach (A Case Study). Available online at www.sciencedirect.com. 2015