Nota de Aceptación

Aprobado por el Comité de Trabajo de Grado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana para optar el título de Ingeniero de Sistemas y computación.

Dr. Hernán Camilo Rocha Niño
Decano de la Facultad de Ingeniería

Dr. Gerardo Mauricio Sarria
Director Carrera Ingeniería Sistemas y Computación.

Dr. Silvio Ricardo Timarán Pereira
Director(a) Trabajo

Dr. Gloria Inés Álvarez
Jurado 1

Mtr. David Arango Londoño
Jurado 2

Pontificia Universidad JAVERIANA Cali

Facultad de Ingeniería y Ciencias

Acta de Correcciones al Proyecto de Grado
Ingeniería de Sistemas y Computación

Fecha: 1 Julio 2021

Autores: Andrea Estefanía Timarán Buchely

Nombre del Proyecto de Grado: Detección de patrones de desempeño académico en las competencias genéricas de las pruebas Saber Pro.

Director: Silvio Ricardo Timarán Pereira

Como indica el artículo 2.27 de las Directrices de Trabajo de Grado, he verificado que los estudiantes indicados arriba han implementado todas las correcciones que los Jurados del Proyecto de Grado definieron que se efectuaran, como consta en el Acta de Calificación correspondiente.

_____

Firma de Director(a) del Proyecto de Grado

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería.
Ingeniería de Sistemas y Computación.
Proyecto de Grado.

Detección de patrones de desempeño académico en las competencias genéricas de las pruebas Saber Pro

Andrea Estefanía Timaran Buchely

Director: PhD. Silvio Ricardo Timaran Pereira

1 Junio 2021

Santiago de Cali, 1 Junio 2021.

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo Mauricio Sarria Montemiranda
Director Carrera de Ingeniería de Sistemas y Computación.
Cali.

Cordial Saludo.

Por medio de la presente me permito informarle que he revisado el proyecto de grado titulado "Detección de patrones de desempeño académico en las competencias genéricas de las pruebas Saber Pro " de la estudiante de Ingeniería de Sistemas y Computación Andrea Estefanía Timaran Buchely (cod: 8919722) del cual soy director y lo considero apto para ser presentado y sometido a consideración del jurado.

Atentamente,

PhD. Silvio Ricardo Timaran Pereira
Director de Trabajo de Grado
Profesor Titular del Departamento de Sistemas
Universidad de Nariño

Santiago de Cali, 1 Junio 2021.

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo Mauricio Sarria Montemiranda
Director Carrera de Ingeniería de Sistemas y Computación.
Cali.

Cordial Saludo.

Tengo el placer de presentar ante usted el proyecto de grado titulado "Detección de patrones de desempeño académico en las competencias genéricas de las pruebas Saber Pro ", para ser sometido a jurado.

Espero que el proyecto cumpla los requisitos académicos necesarios para su aprobación.

Atentamente,

Andrea Timaran Buchely

Andrea Estefanía Timaran Buchely
Código: 8919722

Resumen

Los estudios que se han realizado hasta el momento a nivel nacional, en el marco de las pruebas Saber Pro, se basan en información procesada mediante análisis estadístico, donde fundamentalmente se consideran variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones, que por lo general están ocultas y que únicamente se pueden descubrir utilizando un tratamiento más complejo de los datos, lo cual es posible con la Minería de Datos.

Por otra parte, en la Universidad Javeriana Cali no se han planteado investigaciones que analicen el desempeño de los estudiantes de los diferentes programas profesionales que ofrece esta institución en las competencias genéricas de las pruebas SABER PRO utilizando técnicas predictivas de minería de datos.

En este documento se presentan los resultados del proyecto de investigación cuyo objetivo fue utilizar técnicas predictivas de minería de datos para detectar patrones de desempeño en las competencias genéricas de las pruebas Saber Pro que presentaron los estudiantes de los programas profesionales de la Universidad Javeriana Cali en los años 2017 y 2018, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES.

La metodología utilizada fue CRISP-DM, la guía más ampliamente empleada en el desarrollo de proyectos de minería de datos, que contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

El conocimiento generado permitirá soportar la toma de decisiones de las directivas universitarias, con el fin de mejorar la calidad de la educación en la Universidad Javeriana Cali.

Palabras Clave:

- Competencias genéricas.
- Minería de datos.
- Patrones de Desempeño.
- Programas profesionales.
- Pruebas Saber Pro.

Abstract

The studies related to the Saber Pro Exam have been based on processed information through statistical analysis that considers variables and primary relations without a more in depth context that can only be seen by using advanced techniques, such as data mining. On the other hand, the Pontificia Universidad Javeriana Cali has not done nor proposed investigations with data mining related to the performance of students in the generic skills of the Saber Pro Exam.

In this paper we are reporting the results of an investigation on the generic skills of the Saber Pro Exams presented in 2017 and 2018 by the students at Pontificia Universidad Javeriana Cali. This investigation uses data mining techniques and its objective is to detect performance patterns based on socioeconomical, academical and institutional data stored in the ICFES database.

The methodology used is CRISP-DM, which is the most widely-used analytics model in data mining. This is split in six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment.

The results of the research can be used to support administrative decisions made by the directive board at the university to improve the quality of the education provided.

Keywords:

- Generic Skills.
- Data Mining.
- Performance Pattern.
- Undergraduate.
- Saber Pro Exam.

Índice general

Introducción

En toda institución de educación superior (IES), el desempeño académico de los estudiantes es el factor que determina la calidad de educación de dichas instituciones, es por eso, que el desempeño ha adquirido un gran interés, ya que permite mejorar el nivel de educación superior en las mismas. La calidad de la educación superior supone el esfuerzo continuo de las instituciones para cumplir en forma responsable con las exigencias propias de cada una de sus funciones. Estas funciones que, en última instancia pueden reducirse a docencia, investigación y proyección social, reciben diferentes énfasis en una institución u otra, dando lugar a distintos estilos de institución. [1]

La calidad de la educación superior es una prioridad. Ofrecerla es un deber de las instituciones. Para lograrlo, el Ministerio de Educación Nacional (MEN) y el Instituto Colombiano para la Evaluación de la educación (ICFES) definieron tres programas entrelazados: Estándares Mínimos de Calidad (EMC) para pregrado y posgrado, incentivos a la acreditación de excelencia, y exámenes de calidad.[2]

En Colombia, el Decreto 3963 de 2009, establece que las pruebas Saber Pro son instrumentos estandarizados para evaluación externa de la calidad de la educación superior; forman parte, con otros procesos y acciones de un conjunto de instrumentos que el Estado dispone para evaluar la calidad del servicio público educativo y ejercer su inspección y vigilancia [3].

En las últimas décadas, hay un gran interés por aplicar la minería de datos en los ambientes de la educación superior, lo que ha generado una nueva comunidad de investigación educativa denominada Minería de Datos Educativa (en inglés: Educational Data Mining) [4]. La Minería de Datos Educativa describe un campo de relacionado con la aplicación de minería de datos, aprendizaje automático y estadísticas a la información generada a partir de entornos educativos y así poder analizar y explorar los datos con la finalidad de comprender mejor a los estudiantes y los entornos en los que aprenden, logrando una mayor calidad en las intituciones educación superior.

El objetivo del presente trabajo fue detectar patrones de desempeño académico en las competencias genéricas de las Pruebas Saber Pro 2017 y 2018, presentadas por los estudiantes de la Universidad Javerina Cali en los años 2017 y 2018, utilizando técnicas de minería de datos. Se utilizó la metodología CRISP-DM, una de las guías más utilizadas en este tipo de proyectos. A partir de la información recolectada por el ICFES y almacenada en sus bases de datos, como lo son los factores socio-económicos, académicos e institucionales de los estudiantes y su núcleo familiar, se contruyó un conjunto inicial de datos que posteriormente fue limpiado y transformado para obtener un conjunto final de datos, a partir del cual se contruyó un modelo predictivo aplicando las técnicas de minería de datos y el modelo de clasificación por árboles de decisión. Se obtuvieron patrones asociados al buen o mal desempeño académico en las competencias genéricas de las pruebas Saber Pro. De igual manera se obtuvo un análisis exploratorio con los resultados obtenidos en estas pruebas que permitió

caracterizar a los estudiantes de la Universidad Javeriana Cali que presentaron las pruebas Saber Pro en el periodo comprendido entre 2017 y 2018, realizar un análisis de correlación con el fin de tener una clasificación de los diferentes programas de la Universidad de acuerdo con el desempeño promedio en cada competencia genérica, lo que se le denomino efecto del programa y finalmente establecer el efecto de las diferentes variables socioeconómicas consideradas en este estudio, sobre los puntajes obtenidos en las cinco competencias de las pruebas Saber Pro y en el puntaje Global.

El conocimiento generado en esta investigación, se contituye en información de calidad para soportar la toma de decisiones de las directivas universitarias en vía del mejoramiento de la calidad de la educación superior que se imparte.

Capítulo 1

Descripción del Problema

1.1. Planteamiento del Problema

Desde el año 2011, de acuerdo con los lineamientos Saber Pro, expedidos por el Instituto Colombiano para el fomento de la Educación Superior, todos los estudiantes de pregrado, sin importar el programa de formación profesional que cursen, deben presentar los módulos de competencias genéricas de la prueba Saber Pro que incluyen lectura crítica, razonamiento cuantitativo, competencias ciudadanas, comunicación escrita e inglés.

Según el decreto 3963 del 14 octubre de 2009 del Ministerio de Educación Nacional [3], uno de los objetivos del Examen de Estado de Calidad de la Educación Superior, es comprobar el grado de las competencias de los estudiantes próximos a culminar los programas académicos de pregrado que ofrecen las instituciones de educación superior. Las medidas de estas competencias son necesarias para el adecuado desempeño profesional o académico independientemente del programa que los estudiantes estén cursando. Las competencias genéricas se definen como aquellas competencias que son necesarias para el adecuado desempeño profesional; se consideran transversales, es decir, están presentes en todos los niveles de formación y su complejidad es progresiva a lo largo de toda su formación; por lo tanto, las deben desarrollar todas las personas, dependiendo del nivel de formación.

Cada competencia genérica evalúa distintos elemento. En lectura crítica se evalúa competencias relacionadas con la capacidad de leer de manera analítica y reflexiva. Implica comprender los planteamientos expuestos en un texto e identificar sus perspectivas y juicios de valor. Esto exige que el lector identifique y recupere información presente en uno o varios textos, construya su sentido global, establezca relaciones entre enunciados y evalúe su intencionalidad. Aborda los siguientes procesos: ubicar información, relacionar,construir la representación global del texto analizar y evaluar la relación entre procesos discursivos y contexto sociocultural [5].

En razonamiento cuantitativo se evalúa comprensión de conceptos básicos de matemáticas para analizar,modelar y resolver problemas, aplicando métodos y procedimientos cuantitativos y esquemáticos, basados las propiedades de los números y en las operaciones de las matemáticas; incluye, interpretación de datos, formulación y ejecución de problemas, evaluación y validación de resultados [6].

En competencias ciudadanas se evalúa los conocimientos y habilidades que posibilitan la construcción de marcos de comprensión del entorno, los cuales promueven el ejercicio de la ciudadanía

y la coexistencia inclusiva según la Constitución Política [7]. En comunicación escrita se evalúa la competencia para comunicar ideas por escrito referidas a un tema dado. Cada estudiante debe elaborar un escrito argumentativo según un tema planteado. Incluye las siguientes acciones expresar adecuadamente la intención comunicativa, dar coherencia y cohesión al texto, expresarse mediante un lenguaje apropiado, aplicando las reglas que rigen el lenguaje escrito [8]. Finalmente, en inglés se evalúa la competencia de los estudiantes para la comunicarse efectivamente en inglés, se alinea con el Marco Común Europeo de Referencia para las Lenguas que permite clasificar el desempeño en niveles, a saber -A1,A1,A2,B1 y B2 [9].

El puntaje global en estas pruebas se obtiene a partir de promedio de los puntajes conseguidos por un estudiante en cada competencia.

A pesar de que en la prueba Saber Pro no se pretende que los estudiantes de todas las formaciones desarrollen las competencias genéricas a un mismo nivel, ni aún las comunes a grupos de programas, sí es importante determinar cómo influyen los factores socioeconómicos, académicos e institucionales del estudiante para obtener un determinado nivel de desempeño de estas competencias genéricas.

Desde el año 2008, las competencias genéricas en las pruebas Saber Pro cuentan con un cuestionario socioeconómico, que se compone de preguntas cortas de selección múltiple que se responden en la hoja de respuestas y no se califican. Este cuestionario, a su vez, permite obtener información que podría ayudar a explicar los resultados obtenidos en el examen sobre los procesos de enseñanza y aprendizaje de los estudiantes. Por ejemplo, indaga por características del núcleo familiar (composición, estatus laboral y educativo), condiciones del hogar (dotación de bienes dentro de la vivienda, estrato socioeconómico, disponibilidad de conexión a internet y servicio de televisión por cable), así como el tiempo dedicado por la familia al entretenimiento [10].

Los estudios que se han realizado hasta el momento a nivel nacional, en el marco de las pruebas Saber Pro, se basan en información procesada mediante análisis estadístico, donde fundamentalmente se consideran variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones, que por lo general están ocultas y que únicamente se pueden descubrir utilizando un tratamiento más complejo de los datos, lo cual es posible con la minería de datos.

Por otra parte, en la Pontificia Universidad Javeriana Cali no se han planteado investigaciones que analicen el desempeño de los estudiantes de los diferentes programas profesionales que ofrece esta institución en las competencias genéricas de las pruebas Saber Pro utilizando técnicas predictivas de minería de datos.

1.1.1. Formulación

¿Cuáles son los patrones socioeconómicos, académicos e institucionales asociados al desempeño académico de los estudiantes de la Pontificia Universidad Javeriana Cali en las competencias

genéricas del Saber Pro presentadas en el año 2017 y 2018?

1.1.2. Sistematización

- ¿Cualés son los factores más adecuados para poder detectar patrones de desempeño académico en las pruebas Saber Pro?
- ¿Cualés serán los criterios que se optarán para seleccionar técnicas y algoritmos de minería de datos más apropiados para el descubrimiento de patrones de desempeño?
- ¿Cúales serán los aspectos a evaluar para interpretar los patrones encontrados ?

1.2. Objetivos

1.2.1. Objetivo General

Detectar patrones asociados al desempeño académico en las competencias genéricas de las pruebas Saber Pro presentadas por los estudiantes de la Pontificia Universidad Javeriana Cali en los años 2017 y 2018, a partir de los datos socioeconómicos, académicos e institucionales, almacenados en las bases de datos del ICFES, utilizando técnicas de minerá de datos.

1.2.2. Objetivos Específicos

- Identificar y seleccionar de las bases de datos del ICFES los datos socioeconómicos, académicos e institucionales de los estudiantes de la Pontificia Universidad Javeriana que presentaron las pruebas Saber Pro en los años 2017 y 2018.
- Seleccionar y aplicar las técnicas de minería de datos más apropiadas para el descubrimiento de patrones de desempeño académico en las competencias genéricas de las pruebas Saber Pro, utilizando una herramienta de software libre.
- Evaluar e interpretar los patrones obtenidos con el fin de determinar el conocimiento descubierto acerca de los factores socioeconómicos, académicos e institucionales asociados al desempeño de los estudiantes de la Pontificia Universidad Javeriana Cali, en las pruebas saber Pro.

1.3. Justificación

La educación es un derecho humano, un importante motor del desarrollo y uno de los instrumentos más eficaces para reducir la pobreza y mejorar la salud, y lograr la igualdad de género, la paz y la estabilidad [11]. La calidad de la educación superior en el mundo es un referente importante para el desarrollo de los países. Es así como el Banco Mundial asigna mayor importancia, para el crecimiento económico, a la calidad de la educación antes que a la cantidad de la misma; afirma, que la política pública de educación superior debe abordar el problema de aumentar el rendimiento

en la formación profesional.

En Colombia, el Decreto 3963 de 2009, establece que las pruebas Saber Pro son instrumentos estandarizados para evaluación externa de la calidad de la educación superior; forman parte, con otros procesos y acciones de un conjunto de instrumentos que el Estado dispone para evaluar la calidad del servicio público educativo y ejercer su inspección y vigilancia [3].

Las Pruebas Saber Pro evalúan las competencias necesarias para el adecuado desempeño profesional y/o académico, como también, las genéricas que debe tener todo egresado de la educación superior, independientemente del programa que haya cursado. Dichas pruebas se constituyen en fuente de información para validar la construcción de indicadores de evaluación de la calidad tanto de programas como de instituciones de educación superior y del servicio público educativo.

En este proyecto se utilizó técnicas predictivas de minería de datos para detectar patrones de desempeño académico en las competencias genéricas de las pruebas Saber Pro 2017 y 2018, presentadas por los estudiantes de los programas profesionales de la Pontificia Universidad Javeriana Cali, a partir de los datos socioeconómicos, académicos e institucionales almacenados en las bases de datos del ICFES. El conocimiento generado permitirá soportar la toma de decisiones de las directivas universitarias, con el fin de mejorar la calidad de la educación en la Pontificia Universidad Javeriana Cali.

Capítulo 2

Marco de Referencia

2.1. Áreas Temáticas

- Minería de Datos

2.2. Marco Teórico

2.2.1. Minería de datos

La minería de datos es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados. Empleando una amplia variedad de técnicas, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más. [12] Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. Se divide en 4 etapas: [13]

- Determinación de los objetivos: delimitar los objetivos que el cliente desea bajo la orientación del especialista en data mining.
- Preprocesamiento de los datos: selección, limpieza, enriquecimiento, reducción y transformación de las bases de datos.
- Determinación del modelo: Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
- Análisis de los resultados: Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica.

A la minería de datos se la conoce también como el descubrimiento de conocimiento en bases de datos KKD (del inglén Knowledge Discovery in Database). KDD es básicamente un proceso automático en el que se combinana descubrimiento y análisis. El proceso consiste en extraer patrones en forma de regla o funciones, a partir de los datos, para que el usuario los analice [14]. Es un proceso interactivo compuesto por varias fases de las cuales una de ellas es la minería e datos. Implica no solo obtener modelos o patrones (en la fase de minería de datos), sino seleccionar, limpiar, transformar los datos e interpretar y evaluar los patrones para convertirlos en conocimiento y de esta manera

puedan ser útiles para ayudar a la toma de decisiones efectivas en las organizaciones.[15][16][17] [18]

La minería de datos y la obtención de modelos se pueden concebir como aprendizaje a partir de datos. El aprendizaje puede ser surpervisado y no supervisado. Dentro de la minerá de datos se encuentran diferentes tipos de tareas, cada una de las cuales puede considerarse como un tipo de problema resuelto por un algoritmo de minería de datos. [19]

Se clasifican en tareas predictivas y descriptivas. Las tareas predictivas pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables o atributos denominadas variables independientes o predictivas. Son procesos de aprendizaje supervisado porque se parte de un conocimiento previo de los datos. Se supone que el usuario conoce con certeza las categorías de cada registro con el que se cuenta. En el aprendizaje supervisado, se pretende clasificar a los registros de datos en alguna de las categorías predefinidas. En el desarrollo de un modelo predictivo, la meta es crear un clasificador que pueda predecir la categoría de un registro basándose en los datos con los que cuenta. Estas tareas se fundamentan en la identificación de relaciones entre variables en eventos pasados, para luego explotar dichas relaciones y predecir posibles resultados en futuras situaciones. Al proceso de transferir el conocimiento al modelo se le conoce como entrenamiento. Los datos utilizados en este proceso se llaman conjunto de entrenamiento. Son ejemplos de este tipo de tareas de minería de datos la regresión, la clasificación y las series de tiempos entre otras.

Las tareas descriptivas identifican patrones que explican o resumen los datos. Sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Son procesos de aprendizaje no supervisados, ya que se buscan automáticamente grupos de valores para que después el usuario intente encontrar las correspondencias entre esos grupos seleccionados automáticamente y las categorías que le puedan ser de interés. En el aprendizaje no supervisado se cuenta con un conjunto de datos, pero no existe información sobre cómo agrupar los datos en conglomerados.

Basándose en similitudes entre los datos, se empiezan a desarrollar conglomerados o clústeres entre los datos hasta que comienzan a aparecer un conjunto de patrones diferenciables. Aquí no hay diferenciación entre variables independientes o predictores y variables dependientes. Entre las tareas descriptivas de minería de datos se pueden citar asociación, agrupamiento o clustering, patrones secuenciales y correlaciones.

2.2.2. Tareas de Minería de datos

- Clasificación: La clasificación de datos permite obtener resultados a partir de un proceso de aprendizaje supervisado. La clasificación de datos es el proceso por medio del cual se encuentra propiedades comunes entre un conjunto de objetos de una base de datos y se los cataloga en diferentes clases, de acuerdo con el modelo de clasificación. [20]

Este proceso se realiza en dos pasos: en el primer paso se construye un modelo, en el cual, cada tupla de un conjunto de tuplas de la base de datos tiene una clase conocida (etiqueta), determinada por uno de los atributos de la base de datos, llamado *atributo clase*. El conjunto de tuplas que sirve para construir el modelo se denomina *conjunto de entrenamiento* y se escoge randómicamente del total de tuplas de la base de datos. A cada tupla de este conjunto se denomina *ejemplo de entrenamiento* [18]. En el segundo paso, se usa el modelo para clasificar. Inicialmente, se estima la exactitud del modelo utilizando otro conjunto de tuplas de la base de datos, cuya clase es conocida, denominado *conjunto de prueba*. Este conjunto es escogido randómicamente y es independiente del conjunto de entrenamiento [18].

La exactitud del modelo, sobre el conjunto de prueba, es el porcentaje de ejemplos de prueba que son correctamente clasificadas por el modelo. Si la exactitud del modelo se considera aceptable, se puede usar para clasificar futuros datos o tuplas para los cuales no se conoce la clase a la cual pertenecen. Se han propuesto varios métodos de clasificación: rough sets, árboles de decisión, redes neuronales, Bayes, algoritmos genéticos entre otros.

El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender [18], [21]. Este modelo tiene su origen en los estudios de Aprendizaje de Máquina. Este es un método de aprendizaje supervisado que construye árboles de decisión a partir de un conjunto de casos o ejemplos denominados conjunto de entrenamiento (training set) extraídos de la base de datos. También se escoge un conjunto de prueba, cuyas características son conocidas, con el fin de evaluar el árbol.

- Clustering: El proceso de agrupar objetos físicos o abstractos en clases de objetos similares se llama clustering o clasificación no supervisada [16]. Básicamente, el clustering agrupa un conjunto de datos (sin un atributo de clase predefinido) basado en el principio de maximizar la similitud intraclase y minimizar la similitud interclase. El análisis de clustering ayuda a construir particiones significativas de un gran conjunto de objetos basado en la metodología "divide y conquista", la cual descompone un sistema de gran escala en pequeños componentes para simplificar el diseño y la implementación.

  La meta del clustering en una base de datos, es la partición de ésta en segmentos o clusters de registros similares que comparten un número de propiedades y son considerados homogéneos. Los registros en diferentes clusters son diferentes. Los clusters tienen una alta homogeneidad interna (dentro del cluster) y una alta heterogeneidad externa (entre clusters). Por homogeneidad se entiende que los registros en un cluster están próximos unos a otros, donde la proximidad se expresa por medio de una medida, dependiendo de la distancia de los registros al centro del segmento. Por heterogeneidad se entiende que los registros en diferentes segmentos no son similares de acuerdo a una medida de similaridad [22].

  Clustering, típicamente, permite descubrir subpoblaciones homogéneas. Por ejemplo, se aplica

a una base de datos de clientes, para mejorar la exactitud de los perfiles, identificando subgrupos de clientes que tienen un comportamiento similar al comprar.

El algoritmo de clustering segmenta una base de datos sin ninguna indicación por parte del usuario sobre el tipo de clusters que va a encontrar en la base de datos, desechando cualquier sesgo o intuición por parte del usuario y potenciando así, el verdadero descubrimiento de conocimiento. Por esta razón, al método de clustering, se lo denomina aprendizaje no supervisado. Uno de los algoritmos más utilizados para esta tarea es el K-means.

- Asociación: La tarea de asociación descubre patrones en forma de reglas, que muestran los hechos que ocurren frecuentemente juntos, en un conjunto de datos determinado. El problema fue formulado por Agrawal [20], y a menudo se referencia como el problema de canasta de mercado (market-basket). En este problema, se da un conjunto de ítems y una colección de transacciones que son subconjuntos (canastas) de estos ítems. La tarea es encontrar relaciones entre los ítems de esas canastas para descubrir reglas de asociación que cumplan unas especificaciones mínimas dadas por el usuario, expresadas en forma de soporte y confianza. Las cantidades de ítems comprados en una transacción no se toman en cuenta, lo que significa que cada ítem es una variable binaria representando si un ítem está presente o no en una transacción.

2.2.3. Modelo de Clasificación basado en árboles de decisión

El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender [18] [21]. Este modelo tiene su origen en los estudios de Aprendizaje de Máquina. Este es un método de aprendizaje supervisado que construye árboles de decisión a partir de un conjunto de casos o ejemplos denominados conjunto de entrenamiento (training set) extraídos de la base de datos. También se escoge un conjunto de prueba, cuyas características son conocidas, con el fin de evaluar el árbol. La calidad del árbol depende de la exactitud de la clasificación y del tamaño del árbol.

La idea básica de este modelo es la de construir los árboles de decisión en los que:

- Cada nodo no terminal está etiquetado con un atributo.
- Cada rama que sale de un nodo está etiquetada con un valor de ese atributo.
- Cada nodo terminal está etiquetado con un conjunto de casos, cada uno de los cuales satisface todos los valores de atributos que etiquetan el camino desde ese nodo al nodo inicial.

2.2.4. Algoritmos de árboles de decisión

Los árboles de decisión de un nivel o decisión stump (DS) son árboles que clasifican casos, basados en valores característicos. Cada nodo en un árbol de decisión de un nivel representa una

característica de un caso para ser clasificado, y cada rama representa un valor que el nodo puede tomar. Los casos son clasificados comenzando en el nodo raíz y se cataloga basándose en sus valores característicos. En el peor de los casos un árbol de decisión de un nivel puede reproducir el sentido más común, y puede hacerse mejor si la selección característica es particularmente informativa [23].

Generalmente, el conjunto propuesto consiste en los siguientes cuatro pasos: Determinar la distancia métrica conveniente. Encontrar el k vecino más cercano usando la distancia métrica seleccionada. Aplicar la empaquetación de clasificación de los árboles de decisión de un nivel como entrenamiento de los k casos. Finalmente, la respuesta a la empaquetación de conjunto es la predicción para los casos de prueba.

- El algoritmo J48, el cual implementa al algoritmo C.45 [24], se basa en la utilización del criterio de ganancia de información (information gain). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido [25]. El parámetro más importante que se tiene en cuenta para la poda es el factor de confianza C (confidence level), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25 % y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños [26]. Otro parámetro utilizado para variar el tamaño del árbol es a través del factor M que especifica el mínimo número de instancias o registros por nodo del árbol [27].

- Logistic Model Tree (LMT) proporciona una descripción muy buena de los datos. Un LMT consiste básicamente en una estructura de un árbol de decisión con funciones de regresión logística en las hojas. Como en los árboles de decisión ordinarios, una prueba sobre uno de los atributos es asociado con cada nodo interno. Para enumerar los atributos con k valores, el nodo tiene k nodos hijos, y los casos son clasificados en las k ramas dependiendo del valor del atributo. Para atributos numéricos, el nodo tiene dos nodos hijos y la prueba consiste en comparar el valor del atributo con un umbral: un caso puede ser clasificar los datos menores en la rama izquierda mientras que los valores mayores en la rama derecha [23].

- Random Forest se basa en el desarrollo de muchos árboles de clasificación. Para clasificar un objeto desde un vector de entrada, se pone dicho vector bajo cada uno de los árboles del bosque. Cada árbol genera una clasificación, el bosque escoge la clasificación teniendo en cuenta el árbol más votado sobre todos los del bosque [28].

- Random Tree es un árbol dibujado al azar de un juego de árboles posibles. En este contexto “ al azar ” significa que cada árbol en el juego de árboles tiene una posibilidad igual de ser probado. Otro modo de decir esto consiste en que la distribución de árboles es “uniforme” [23].

- RepTree es un método de aprendizaje rápido mediante árboles de decisión. Construye un árbol de decisión usando la información de varianza y lo poda usando como criterio la reducción

del error. Solamente clasifica valores para atributos numéricos una vez. Los valores que faltan se obtienen partiendo las correspondientes instancias" [29]. Es un árbol de clasificación con modelo comprensible (reglas if then else). Construye un árbol de decisión usando la ganancia de información y realiza una poda de error reducido. Solamente ordena una vez los valores de los atributos numéricos. Los valores ausentes se manejan dividiendo las instancias correspondientes en segmentos [23].

2.3. Trabajos Relacionados

- *Descubrimieno de patrones de desempeño académico* [14].
  En este estudio la utilización de la técnica clasificación por árboles de decisión generó modelos predictivos para cada competencia genérica, que permitieron descubrir conocimiento para predecir tendencias sobre el desempeño académico de los actuales y futuros estudiantes de los programas profesionales que presentan las pruebas Saber Pro. Esto con el fin de identificar riesgos y oportunidades que ayuden a las instituciones gubernamentales y de educación superior a tomar decisiones para el mejoramiento de la educación superior en Colombia. En los patrones de desempeño académico descubiertos en las cuatro competencias genéricas de las pruebas Saber Pro 2011-2, se encontró que la acreditación institucional se constituye en factor importante asociado al desempeño académico de los estudiantes de programas profesionales en las pruebas Saber Pro 2011-2.

- *Estudio del Rendimiento Acaémico Aplicando Técnicas de Minería de Datos* [30].
  En este trabajo, el autor Beccera busca determinar el rendimiento académico de los estudiantes, de la Universidad Nacional de Loja, mediante la implementación de un Modelo Computacional a través de Técnicas de Minería de Datos, donde se propone la utilización de las mismas, para detectar cuáles son los factores (académicos, personales, socioeconómicos e institucionales) que influyen en el rendimiento académico de los estudiantes.

- *Predicción del Rendimiento Acaémico mediante minería de datos en estudiantes del primer ciclo de la escuela profesional de ingeniería de computación y sistemas* [31].
  El autor Yamao en su investigación, realizó predicciones a través de tres técnicas: regresión lineal, árbol de decisiones y support vector machines, y el mejor resultado que obtuvo es de 82.87 % utilizando árbol de decisiones. De los diferentes factores, los que más influyeron en el rendimiento académico fueron: nota de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios.

Capítulo 3

Metodología

3.1. Tipo De Estudio

Descriptivo

Los estudios descriptivos conciernen y son diseñados para describir la distribución de variables, sin considerar hipótesis causales o de otra naturaleza. Este proyecto es de tipo descriptivo bajo el enfoque cuantitativo, aplicado un diseño no experimental. Por ser una investigación que involucra la minería de datos, se utilizó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Se trata de un modelo de proceso de minería de datos que describe los enfoques comunes que utilizan los expertos en minería de datos. CRISP-DM es uno de los modelos utilizados, principalmente, en los ambientes académico e industrial y la guía de referencia más ampliamente utilizada en el desarrollo de este tipo de proyectos [19].

En un intento de normalización del proceso de minería de datos, de forma similar a como se hace en ingeniería software para normalizar el proceso de desarrollo software, surgieron a finales de los 90 dos metodologías principales: CRISP-DM y SEMMA (Sample, Explore, Modify, Model, and Assess). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase [32]. Azevedo y Santos [33] comparan ambas implementaciones y llega a la conclusión de que, aunque se puede establecer un paralelismo claro entre ellas, CRISP-DM es más completo porque tiene en cuenta la aplicación al entorno de negocio de los resultados, y por ello es la que se adoptó popularmente. En encuestas realizadas en KDNuggets en 2002, 2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada, cuatro veces más que SEMMA [32], y es la que se utilizó en esta investigación.

Aunque la metodología CRISP-DM para proyectos de minería de datos no es actual, es pertinente para comprender esta tecnología o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características. La metodología CRISP-DM, consiste en un conjunto de tareas que están organizadas en cuatro niveles de abstracción: fases, tareas generales, tareas especializadas e instancias de proceso. Dichos niveles están establecidos respetando jerarquías en tareas, inician en el nivel más general hasta llegar, finalmente, a los casos más específicos [34]. El modelo provee una representación completa del ciclo de vida de un proyecto de minería de datos. El proceso es dinámico e iterativo, por lo que la ejecución de los procesos no es estricta y con frecuencia se puede pasar de uno a otra, de atrás hacia delante y viceversa. Éstos dependen del resultado de cada fase o la planeación de la siguiente tarea a ejecutar. Cada fase se estructura en varias tareas generales, las tareas generales se proyectan en tareas específicas, donde finalmente se describen las acciones que

deben ser desarrolladas para situaciones definidas [35].

CRISP-DM está compuesta por seis fases: análisis del problema, análisis de los datos, preparación de los datos, modelado, evaluación y explotación [36] como se aprecia en la figura 3.1.

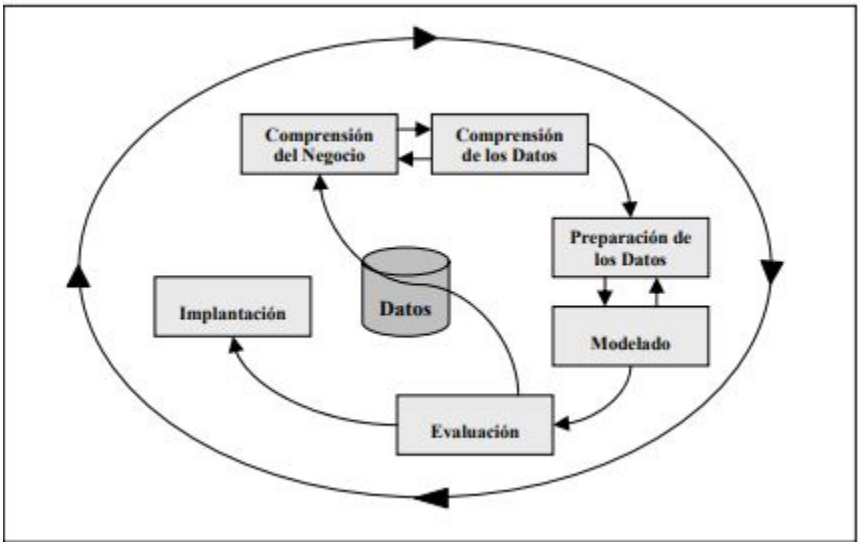


Figura 3.1: Fases de la metodología CRISP-DM

1. La fase de **Compresión del Negocio** se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición del problema de minería de datos, y un plan preliminar diseñado para alcanzar los objetivos.

2. La fase de **Comprensión de Datos** comienza con una colección inicial de datos y procesos con actividades con el objetivo de familiarizarse con los datos, identificar la calidad de los problemas, para descubrir las primeras señales dentro de los datos y detectar temas interesantes para poder formular hipótesis de información oculta.

3. La Fase de **Preparación de Datos** cubre todas las actividades para construir el conjunto de datos. Estas tareas son ejecutadas en múltiples oportunidades y sin orden. Las tareas incluyen selección y transformación de tablas, registros y atributos y limpieza de datos para las herramientas de modelado.

4. La fase de **Modelado** se seleccionan y aplican varias técnicas de modelado y se calibran los parámetros para obtener óptimos resultados. Hay varias técnicas que tienen requerimientos específicos para la forma de los datos, por lo que frecuentemente es necesario volver a la fase de preparación de datos. .

5. La fase de **Evaluación** se evalúa el modelo construido teniendo en cuenta el cumplimiento de los criterios de éxito y se interpretan los resultados.

6. La fase de **Implementación** o implementación el conocimiento obtenido se transforma en acciones dentro del proceso de negocio.

3.2. Datos

3.2.1. Base de datos inicial

Para el presente trabajo de grado se utilizó dos conjuntos datos almacenados en el repositorio del ICFES de las pruebas Saber Pro:

1. El primer conjunto de datos corresponde a la información de los estudiantes que presentaron las pruebas Saber Pro en el año 2017 con un total de 245593 registros y 107 atributos.

2. El segundo conjunto de datos corresponde a la información de los estudiantes que presentaron las pruebas Saber Pro en el año 2018 con un total de 237124 registros y 106 atributos.

Estos dos conjuntos de datos se integraron teniendo en cuenta los atributos comunes entre ellos y los equivalentes (con diferentes nombres, pero con valores iguales), como se muestra en la tabla 3.1 dejando un total de 101 atributos en común entre los dos conjuntos. Se dejaron únicamente los registros de los estudiantes de la Universidad Javeriana Cali. Como resultado se obtuvo el conjunto de datos inicial al cual se le denominó *sbpro_jave_inicial* , que contiene 2052 registros y 101 atributos [37].

#	Atributo	2017	2018	Observaciones
1	Estu_Tipodocumento	x	x	El atributo está presente en 2017 y 2018
2	Estu_Nacionalidad	x	x	El atributo está presente en 2017 y 2018
3	Estu_Genero	x	x	El atributo está presente en 2017 y 2018
4	Estu_Fechanacimiento	x	x	El atributo está presente en 2017 y 2018
5	Estu_Exterior	x	x	El atributo está presente en 2017 y 2018
6	Periodo	x	x	El atributo está presente en 2017 y 2018
7	Estu_Consecutivo	x	x	El atributo está presente en 2017 y 2018
8	Estu_Estadocivil		x	El atributo está presente únicamente en 2018
9	Estu_Estudiante	x	x	El atributo está presente en 2017 y 2018
10	Estu_Pais_Reside	x	x	El atributo está presente en 2017 y 2018
11	Estu_Tieneetnia	x	x	El atributo está presente en 2017 y 2018
12	Estu_Etnia	x	x	El atributo está presente en 2017 y 2018
13	Estu_Limita_Motriz	x		El atributo está presente únicamente en 2017
14	Estu_Limita_Invidente	x		El atributo está presente únicamente en 2017

#	Atributo	2017	2018	Observaciones
15	Estu_Limita_Condicionespecial	x		El atributo está presente únicamente en 2017
16	Estu_Limita_Sordo	x		El atributo está presente únicamente en 2017
17	Estu_Limita_Autismo	x		El atributo está presente únicamente en 2017
18	Estu_Depto_Reside	x	x	El atributo está presente en 2017 y 2018
19	Estu_Cod_Reside_Depto	x	x	El atributo está presente en 2017 y 2018
20	Estu_Mcpio_Reside	x	x	El atributo está presente en 2017 y 2018
21	Estu_Cod_Reside_Mcpio	x	x	El atributo está presente en 2017 y 2018
22	Estu_Areareside		x	El atributo está presente únicamente en 2018
23	Estu_Cole_Termino	x	x	El atributo está presente en 2017 y 2018
24	Estu_Coddane_Cole_Termino	x	x	El atributo está presente en 2017 y 2018
25	Estu_Cod_Cole_Mcpio_Termino	x	x	El atributo está presente en 2017 y 2018
26	Estu_Otrocole_Termino	x	x	El atributo está presente en 2017 y 2018
27	Estu_Tituloobtenidobachiller	x	x	El atributo está presente en 2017 y 2018
28	Estu_Valormatriculauext	x		El atributo está presente únicamente en 2017
29	Estu_Valormatriculauniversidad	x	x	El atributo está presente en 2017 y 2018
30	Estu_Pagomatriculaext	x		El atributo está presente únicamente en 2017
31	Estu_Pagomatriculabeca	x	x	El atributo está presente en 2017 y 2018
32	Estu_Pagomatriculacredito	x	x	El atributo está presente en 2017 y 2018
33	Estu_Pagomatriculapadres	x	x	El atributo está presente en 2017 y 2018
34	Estu_Pagomatriculapropio	x	x	El atributo está presente en 2017 y 2018
35	Estu_Comocapacitoexamensb11	x	x	El atributo está presente en 2017 y 2018
36	Estu_Cursodocentesies	x	x	El atributo está presente en 2017 y 2018
37	Estu_Cursoiesapoyoexterno	x	x	El atributo está presente en 2017 y 2018
38	Estu_Cursoiesexterna	x	x	El atributo está presente en 2017 y 2018
39	Estu_Simulacrotipoicfes	x	x	El atributo está presente en 2017 y 2018
40	Estu_Actividadrefuerzoareas	x	x	El atributo está presente en 2017 y 2018
41	Estu_Actividadrefuerzogeneric	x	x	El atributo está presente en 2017 y 2018
42	Estu_Paisdocumentosb11	x		El atributo está presente únicamente en 2017
43	Estu_Tipodocumentosb11	x	x	El atributo está presente en 2017 y 2018
44	Estu_Semestrecursa	x		El atributo está presente únicamente en 2017
45	Fami_Hogaractual	x	x	El atributo está presente en 2017 y 2018
46	Fami_Cabezafamilia	x	x	El atributo está presente en 2017 y 2018
47	Fami_Numpersonasacargo	x	x	El atributo está presente en 2017 y 2018
48	Fami_Educacionpadre	x	x	El atributo está presente en 2017 y 2018
49	Fami_Educacionmadre	x	x	El atributo está presente en 2017 y 2018
50	Fami_Ocupacionpadre	x		El atributo está presente únicamente en 2017
51	Fami_Ocupacionmadre	x		El atributo está presente únicamente en 2017
52	Fami_Trabajolaborpadre		x	El atributo está presente únicamente en 2018
53	Fami_Trabajolabormadre		x	El atributo está presente únicamente en 2018

#	Atributo	2017	2018	Observaciones
54	Fami_Estratovivienda	x	x	El atributo está presente en 2017 y 2018
55	Fami_Personashogar	x	x	El atributo está presente en 2017 y 2018
56	Fami_Cuartoshogar	x	x	El atributo está presente en 2017 y 2018
57	Fami_Tieneinternet	x	x	El atributo está presente en 2017 y 2018
58	Fami_Tieneserviciotv	x	x	El atributo está presente en 2017 y 2018
59	Fami_Tienecomputador	x	x	El atributo está presente en 2017 y 2018
60	Fami_Tienelavadora	x	x	El atributo está presente en 2017 y 2018
61	Fami_Tienehornomicroogas	x	x	El atributo está presente en 2017 y 2018
62	Fami_Tieneautomovil	x	x	El atributo está presente en 2017 y 2018
63	Fami_Tienemotocicleta	x	x	El atributo está presente en 2017 y 2018
64	Fami_Tieneconsolavideojuegos	x	x	El atributo está presente en 2017 y 2018
65	Fami_Cuantoscompartebaño	x	x	El atributo está presente en 2017 y 2018
66	Fami_Numlibros	x	x	El atributo está presente en 2017 y 2018
67	Estu_Dedicacionlecturadiaria	x	x	El atributo está presente en 2017 y 2018
68	Estu_Dedicacioninternet	x	x	El atributo está presente en 2017 y 2018
69	Estu_Horassemanatrabaja	x	x	El atributo está presente en 2017 y 2018
70	Estu_Tiporemuneracion	x	x	El atributo está presente en 2017 y 2018
71	Inst_Cod_Institucion	x	x	El atributo está presente en 2017 y 2018
72	Inst_Nombre_Institucion	x	x	El atributo está presente en 2017 y 2018
73	Estu_Prgm_Academico	x	x	El atributo está presente en 2017 y 2018
74	Estu_Snies_Prgmacademico	x	x	El atributo está presente en 2017 y 2018
75	Gruporeferencia	x	x	El atributo está presente en 2017 y 2018
76	Estu_Prgm_Codmunicipio	x	x	El atributo está presente en 2017 y 2018
77	Estu_Prgm_Municipio	x	x	El atributo está presente en 2017 y 2018
78	Estu_Prgm_Departamento	x	x	El atributo está presente en 2017 y 2018
79	Estu_Nivel_Prgm_Academico	x	x	El atributo está presente en 2017 y 2018
80	Estu_Metodo_Prgm	x	x	El atributo está presente en 2017 y 2018
81	Estu_Nucleo_Pregrado	x	x	El atributo está presente en 2017 y 2018
82	Estu_Inst_Codmunicipio	x	x	El atributo está presente en 2017 y 2018
83	Estu_Inst_Municipio	x	x	El atributo está presente en 2017 y 2018
84	Estu_Inst_Departamento	x	x	El atributo está presente en 2017 y 2018
85	Inst_Caracter_Academico	x	x	El atributo está presente en 2017 y 2018
86	Inst_Origen	x	x	El atributo está presente en 2017 y 2018
87	Estu_Privado_Libertad	x	x	El atributo está presente en 2017 y 2018
88	Estu_Cod_Mcpio_Presentacion	x	x	El atributo está presente en 2017 y 2018
89	Estu_Mcpio_Presentacion	x	x	El atributo está presente en 2017 y 2018
90	Estu_Depto_Presentacion	x	x	El atributo está presente en 2017 y 2018
91	Estu_Cod_Depto_Presentacion	x	x	El atributo está presente en 2017 y 2018
92	Mod_Razona_Cuantitat_Punt	x	x	El atributo está presente en 2017 y 2018

#	Atributo	2017	2018	Observaciones
93	Mod_Razona_Cuantitat_Desem	x	x	El atributo está presente en 2017 y 2018
94	Mod_Razona_Cuantitativo_Pnal	x	x	El atributo está presente en 2017 y 2018
95	Mod_Razona_Cuantitativo_Pgref	x	x	El atributo está presente en 2017 y 2018
96	Mod_Lectura_Critica_Punt	x	x	El atributo está presente en 2017 y 2018
97	Mod_Lectura_Critica_Desem	x	x	El atributo está presente en 2017 y 2018
98	Mod_Lectura_Critica_Pnal	x	x	El atributo está presente en 2017 y 2018
99	Mod_Lectura_Critica_Pgref	x	x	El atributo está presente en 2017 y 2018
100	Mod_Competen_Ciudada_Punt	x	x	El atributo está presente en 2017 y 2018
101	Mod_Competen_Ciudada_Desem	x	x	El atributo está presente en 2017 y 2018
102	Mod_Competen_Ciudada_Pnal	x	x	El atributo está presente en 2017 y 2018
103	Mod_Competen_Ciudada_Pgref	x	x	El atributo está presente en 2017 y 2018
104	Mod_Ingles_Punt	x	x	El atributo está presente en 2017 y 2018
105	Mod_Ingles_Desem	x	x	El atributo está presente en 2017 y 2018
106	Mod_Ingles_Pnal	x	x	El atributo está presente en 2017 y 2018
107	Mod_Ingles_Pgref	x	x	El atributo está presente en 2017 y 2018
108	Mod_Comuni_Escrita_Punt	x	x	El atributo está presente en 2017 y 2018
109	Mod_Comuni_Escrita_Desem	x	x	El atributo está presente en 2017 y 2018
110	Mod_Comuni_Escrita_Pnal	x	x	El atributo está presente en 2017 y 2018
111	Mod_Comuni_Escrita_Pgref	x	x	El atributo está presente en 2017 y 2018
112	Punt_Global	x	x	El atributo está presente en 2017 y 2018
113	Percentil_Global	x	x	El atributo está presente en 2017 y 2018
114	Estu_Inse_Individual	x	x	El atributo está presente en 2017 y 2018
115	Estu_Nse_Individual	x	x	El atributo está presente en 2017 y 2018
116	Estu_Nse_Ies	x	x	El atributo está presente en 2017 y 2018
117	Estu_Estadoinvestigacion	x	x	El atributo está presente en 2017 y 2018

Tabla 3.1: Análisis de atributos de los conjuntos de datos SABER PRO 2017-2018.

3.2.2. Base de datos final

La alta dimensionalidad es un problema para el descubrimiento de patrones con minería de datos [19]. Uno de los criterios utilizados para resolver este problema, es reducir el número de atributos a analizar, a través de la limpieza y transformación de datos. Teniendo en cuenta este criterio, al conjunto *sbpro_jave_inicial* se le aplicaron técnicas de limpieza y transformación dando como resultado el conjunto de datos denominado *sbpro_jave_final* compuesto por 2052 registros y 23 atributos, el cual sirvió de base para la fase de modelado.

Para obtener este repositorio inicialmente se realizó un análisis de la calidad de datos. Como resultado de este proceso, se efectuó una primera selección de atributos y se descartaron aquellos atributos con un alto porcentaje de valores nulos, por la imposibilidad de encontrar sus valores a

través de fuentes externas de datos como se muestra en la tabla 3.2; tampoco se tomaron en cuenta los atributos con valores distintos igual a 1 (significa que es una constante) como se observa en la tabla 3.3, atributos de identificación e irrelevantes 3.4 y atributos con mayoria en un solo valor 3.5.

Atributo	% Nulos
estu_otrocole_termino	89.035088
estu_cursodocentesies	77.339181
estu_cursoiesapoyoexterno	77.339181
estu_cursoiesexterna	77.339181
estu_simulacrotipoicfes	77.339181
estu_actividadrefuerzoareas	77.339181
estu_actividadrefuerzogeneric	77.339181
fami_tieneconsolavideojuegos	54.580897
fami_cuantoscomportebaño	54.678363
estu_tiporemuneracion	56.140351
estu_cole_termino	33.138402
estu_coddane_cole_termino	33.138402
estu_cod_cole_mcpio_termino	33.138402

Tabla 3.2: Atributos con alto porcentaje de nulos.

Atributo	Valor
estu_estudiante	Estudiante
inst_cod_institu cion	1702
inst_nombre_institucion	Pontificia Universidad Javeriana-Cali
estu_prgm_codmunicipio	76001
estu_prgm_municipio	Cali
estu_prgm_departamento	Valle
estu_nivel_prgm_academico	Universitario
estu_metodo_prgm	Presencial
estu_inst_codmunicipio	76001
estu_inst_municipio	Cali
estu_inst_departamento	Valle
inst_caracter_academico	Universidad
inst_origen	No oficial
estu_privado_libertad	N
estu_nse_ies	4

Tabla 3.3: Atributos con valores igual a 1 (constantes)

Atributo
estu_tipodocumento
estu_nacionalidad
estu_exterior
estu_pais_reside
estu_tieneetnia
estu_inse_individual
estu_comocapacitoexamensb11
estu_tipodocumentosb11

Tabla 3.4: Atributos de identificación e irrelevantes

Atributo	Mayoria en un solo valor
estu_tituloobtenidobachiller	bachiller académico
estu_horassemanatrabaja	0
estu_cod_mcpio_presentacion	76001
estu_mcpio_presentacion	Cali
estu_depto_presentacion	Valle
estu_cod_depto_presentacion	76

Tabla 3.5: Atributos con mayoria en un solo valor

Posteriormente, se realizó el proceso de transformación de datos, discretizando o generalizando los valores continuos a valores discretos y creando nuevos atributos con mayor semántica, reemplazando los existentes, obteniendo así atributos más representativos para el estudio y reduciendo el número de variables a tener en cuenta en la investigación.

Se discretizaron y generalizaron los valores numéricos de ciertos atributos teniendo en cuenta un rango de valores o que las frecuencias por cada valor sean proporcionales, para evitar sesgos, al construir los modelos de minería de datos.

Para el atributo *estu_fechanacimiento* se creó un atributo llamado *estu_edad* para calcular la edad del estudiante con la que presentó las pruebas saber pro teniendo en cuenta el atributo *periodo* que define en que año presentó dichas pruebas. Finalmente se creó un nuevo atributo denominado *estu_grupo_etareo*, cuyos valores se muestran en la tabla 3.6 a partir de el atributo *estu_edad*. Se eliminan los atributos *estu_fechanacimiento*, *estu_edad* y *periodo* dejando solo el atributo *estu_grupo_etareo*.

Estu_grupo_etareo	No. Estudiantes
[<=20]	154
[21]	441
[22]	468
[23]	377
[24]	227
[>=25]	385

Tabla 3.6: Valores discretizados del atributo *estu_edad*

Para el atributo *fami_estratovivienda* se creó un nuevo atributo denominado *cat_estrato*, cuyos valores se muestran en la tabla 3.7.

Fami_estratovivienda	Cat_estrato
1 y 2	Bajo
3 y 4	Medio
5 y 6	Alto

Tabla 3.7: Valores generalizados del atributo *fami_estratovivienda*

Para el atributo *estu_mcpio_reside*, se creó un atributo más general, que agrupe a estos en zonas. Este atributo se le denominó *estu_zona_procede*, con los valores que se muestran en la tabla 3.8. Se eliminan los atributos *estu_depto_reside*, *estu_cod_reside_depto*, *estu_mcpio_reside* y *estu_cod_reside_mcpio*.

Estu_zona_procede	Estu_mcpio_reside
Norte	Alcala, Ansermanuevo, Argelia, Bolivar, Cartago, El Aguila, El Cairo, El Dovio, La Union, La Victoria, Obando, Roldanillo, Toro, Ulloa, Versalles, Zarzal
Centro	Andalucia, Buga, Bugalagrande, Calima, El Darien, El Cerrito, Ginebra, Guacari, Guadalajara De Buga, Restrepo, Riofrio, San Pedro, Trujillo, Tulua, Yotoco
Sur	Cali, Candelaria, Dagua, Florida, Jamundi, La Cumbre, Palmira, Pradera, Vijes, Yumbo
Occidente	Buenaventura
Oriente	Caicedonia, Sevilla
Otras	Fuera del departamento del Valle

Tabla 3.8: Atributo *estu_mcpio_reside* agrupado por zonas

De igual manera, para el atributo *estu_prgm_academico*, se creó un atributo que agrupe a los programas de la Universidad Javeriana Cali según su facultad. Este atributo se denominó *facultades*, con los valores que se muestran en la tabla 3.9. Se elimina el atributo *estu_prgm_academico*.

Facultades	Estu_prgm_academico
Ciencias Económicas Y Administrativas	Contaduría Pública, Economía, Administración De Empresas, Negocios Internacionales
Humanidades Y Ciencias Sociales	Arquitectura, Artes Visuales, Diseño De Comunicación Visual, Ciencia Política, Derecho, Psicología, Comunicación, Filosofía
Ciencias De La Salud	Medicina
Ingeniería Y Ciencias	Ingeniería Civil, Ingeniería De Sistemas Y Computación, Ingeniería Electrónica, Ingeniería Industrial, Matemáticas Aplicadas, Biología

Tabla 3.9: Atributo *estu_prgm_academico* agrupado por facultades

Para los cuatro atributos *estu_pagomatriculabeca*, *estu_pagomatriculacredito*, *estu_pagomatriculapadres*, *estu_pagomatriculapropio*, que almacenan las diferentes formas de pago de la matrícula, se creó un nuevo atributo denominado *forma_pago_matricula*, en donde se combinan los diferentes valores de estos atributos, teniendo en cuenta la presencia (Si) o ausencia (No) de estos, en cuatro dígitos binarios (binarizacion de atributos). Los valores de este atributo se muestran en la tabla 3.10. Se eliminaron los atributos *estu_pagomatriculabeca*, *estu_pagomatriculacredito*, *estu_pagomatriculapadres*, *estu_pagomatriculapropio* y *estu_valormatriculauniversidad*.

Beca	Crédito	Padres	Propios	Valores
No	No	No	Si	Recursos Propios
No	No	Si	No	Recursos Padres
No	No	Si	Si	Recursos Padres-Propios
No	Si	No	No	Recursos Crédito
No	Si	No	Si	Recursos Crédito-Propios
No	Si	Si	No	Recursos Crédito-Padres
No	Si	Si	Si	Recursos Crédito-Padres-Propios
Si	No	No	No	Recursos Beca
Si	No	No	Si	Recursos Beca-Propios
Si	No	Si	No	Recursos Beca-Padres
Si	No	Si	Si	Recursos Beca-Padres-Propios
Si	Si	No	No	Recursos Beca-Crédito
Si	Si	No	Si	Recursos Beca-Crédito-Propios
Si	Si	Si	No	Recursos Beca-Crédito-Padres
Si	Si	Si	Si	Recursos Beca-Crédito-Padres-Propios

Tabla 3.10: Valores nuevo atributo *forma_pago_matricula* binarizado

Se crearon unos índices para medir la condición de hacinamiento (*índice_hacinamiento*), de tics

(*índice_ tics*), electrodomésticos (*índice_ electro*) y transporte (*índice_ transporte*).

Para calcular el índice de hacinamiento se tuvo en cuenta lo expresado por Spicker, Alvarez, & Gordon [38] que dicen "el hacinamiento refiere a la relación entre el número de personas que habitan una vivienda o casa y el espacio o número de cuartos disponibles". Generalmente se aceptan los valores: hasta 2.4 - sin hacinamiento; de 2.5 a 4.9 - hacinamiento medio y de 5.0 o más - hacinamiento crítico.

Teniendo en cuenta estos conceptos, el índice de hacinamiento para cada estudiante se obtuvo dividiendo los valores de los atributos *fami_ personas_ hogar* entre *fami_ cuartos_ hogar*. Los valores de este nuevo atributo se pueden mirar en la tabla 3.11. Se eliminan los atributos *fami_ personas_ hogar* y *fami_ cuartos_ hogar*

Índice_ hacinamiento	***fami_ personas_ hogar / fami_ cuartos_ hogar***
Sin Hacinamiento	<=2.4
Hacinamiento Medio	Entre 2.5 y 4.9
Hacinamiento Crítico	>= 5.0

Tabla 3.11: Valores índice hacinamiento

Para calcular el valor del índice tics se obtuvo la sumatoria de los valores de la presencia (1) o ausencia (0) de los servicios de tecnologías de la información y la comunicación con que cuenta el hogar. Si el índice es 3, la condición tics es BUENA. Si el índice es 2, la condición tics es MEDIA y finalmente, si el índice es 0 o 1, la condición tics es MALA. En la tabla 3.12 se muestran las variables que intervienen en el cálculo del *índice_ tics*. Se eliminan los atributos *fami_ tienecomputado*, *fami_ tieneserviciotv* y *fami_ tieneinternet*.

Servicios	**Si**	**No**
fami_ tienecomputador	1	0
fami_ tieneserviciotv	1	0
fami_ tieneinternet	1	0

Tabla 3.12: Cálculo *índice_ tics*

De igual manera se procedió para el cálculo del índice de electrodomésticos presentes en el hogar. Si la sumatoria es 2 entonces el índice es BUENO. Si es 1 entonces el índice es REGULAR. Si no tiene electrodomésticos, entonces el índice es MALO. En la tabla 3.13 se muestran los atributos que intervienen en el cálculo de los valores para el nuevo atributo *índice_ electro*. Se eliminan los atributos *fami_ tienelavadora* y *fami_ tienehornomicroogas*.

Servicios	Si	No
fami_tienelavadora	1	0
fami_tienehornomicroogas	1	0

Tabla 3.13: Cálculo *índice_electro*

Para el cálculo del índice de transporte se tuvo en cuenta si el estudiante cuenta con automóvil particular o motocicleta para su transporte. Si tiene automóvil entonces el índice es BUENO. Si tiene motocicleta, entonces el índice es REGULAR. Si no tiene un medio de transporte particular, entonces el índice es MALO. En la tabla 3.14 se muestran los atributos que intervienen en el cálculo de los valores para el nuevo atributo *índice_transporte*. Se eliminan los atributos *fami_tieneautomovil* y *fami_tienemotocicleta*.

Servicios	Índice
fami_tieneautomovil	Bueno
fami_tienemotocicleta	Regular
No tiene	Malo

Tabla 3.14: Cálculo *índice_transporte*

Nota: Todos los atributos que fueron reemplazados por nuevos fueron eliminados.

Se eliminaron las llaves y atributos con pocos valores positivos 3.15.

Atributo
estu_consecutivo
fami_hogaractual
fami_cabezafamilia
fami_numpersonasacargo
estu_snies_prgmacademico

Tabla 3.15: Caption

Para cada módulo se creó un atributo de desempeño respecto al módulo con relación a la media nacional, sus valores son *Bajo la media* y *Sobre la media*. Los atributos antiguos se eliminaron como lo muestra la tabla 3.16.

Nuevo atributo	Atributos antiguos
desemp_global	punt_global percentil_global
desemp_leccritica	mod_lectura_critica_punt mod_lectura_critica_pnal mod_lectura_critica_pgref mod_lectura_critica_desem
desemp_comescrita	mod_comuni_escrita_punt mod_comuni_escrita_pnal mod_comuni_escrita_pgref mod_comuni_escrita_desem
desemp_razcuantitativo	mod_razona_cuantitativo_punt mod_razona_cuantitativo_pnal mod_razona_cuantitativo_pgref mod_razona_cuantitativo_desem
desemp_ingles	mod_ingles_punt mod_ingles_pnal mod_ingles_pgref mod_ingles_desem
desemp_compciudadanas	mod_competen_ciudada_punt mod_competen_ciudada_pnal mod_competen_ciudada_pgref mod_competen_ciudada_desem

Tabla 3.16: Atributos de desempeño respecto a la media nacional

Finalmente, con el fin de descubrir patrones asociados al rendimiento académico en cada una de las competencias genéricas, se construyó por cada competencia un repositorio de datos minables a partir del conjunto de datos *sbpro_jave_final*. La descripción de cada repositorio por competencia se muestra en la tabla 3.18. El diccionario de datos de este conjunto se muestra en la tabla 3.17.

#	Atributo	Descripción	Valores
1	estu_genero	Género	F - Femenino M - Masculino
2	estu_tituloobtenidobachiller	Título de bachiller obtenido	Bachiller académico Bachiller técnico Bachiller pedagógico o normalista
3	fami_educacionpadre	Nivel educativo	Ninguno Primaria incompleta Primaria completa Secundaria (Bachillerato) incompleta Secundaria (Bachillerato) completa

#	Atributo	Descripción	Valores
		más alto alcanzado por el padre	Técnica o tecnológica incompleta Técnica o tecnológica completa Educación profesional incompleta Educación profesional completa Postgrado No sabe
4	fami_educacionmadre	Nivel educativo más alto alcanzado por la madre	Ninguno Primaria incompleta Primaria completa Secundaria (Bachillerato) incompleta Secundaria (Bachillerato) completa Técnica o tecnológica incompleta Técnica o tecnológica completa Educación profesional incompleta Educación profesional completa Postgrado No sabe
5	fami_numlibros	¿Cuántos libros físicos o electrónicos hay en su hogar excluyendo periódicos, revis-, tas,directorios telefónicos y libros del colegio?	0 a 10 Libros 11 a 25 Libros 26 a 100 Libros Más de 100 libros
6	estu_dedicacionlecturadiaria	Usualmente, ¿cuánto tiempo o al día dedica a leer por entretenimiento?	No leo por entretenimiento 30 minutos o menos Entre 30 y 60 minutos Entre 1 y 2 horas Más de 2 horas
7	estu_dedicacioninternet	Usualmente, ¿cuánto tiempo al día dedica navegar en internet? Excluya actividades académicas	Menos de una hora Entre 1 y 3 horas Más de 4 horas
8	estu_nse_individual	Índice socioeconómico a nivel de estudiante	NSE1 NSE2 NSE3 NSE4
			<=20

#	Atributo	Descripción	Valores
	estu_grupo_etareo	Valores discretizados del atributo edad estudiantil	21 22 23 24 >= 25
10	cat_estrato	Valores generalizados del atributo fami_estratovivienda	Bajo Medio Alto
11	estu_zona_procede	Municipio de residencia agrupado por zonas	Norte Centro Sur Occidente Oriente Otras
12	hacinamiento	Índice hacinamiento	hacinamiento crítico hacinamiento medio sin hacinamiento
13	indice_tics	Posesión de medios tecnológicos	Malo Regular Bueno
14	indice_electro	Posesión de electrodomésticos	Malo Regular Bueno
15	indice_transporte	Posesión de medios de transporte	Malo Regular Bueno
16	desemp_global	Desempeño global con relación a la media nacional	Bajo la media Sobre la media
17	desemp_leccritica	Desempeño en lectura crítica con relación a la media nacional	Bajo la media Sobre la media
18	desemp_comescrita	Desempeño en comunicación escrita con relación a la media nacional	Bajo la media Sobre la media
19	desemp_ingles	Desempeño en inglés	Bajo la media

#	Atributo	Descripción	Valores
		con relación a la media nacional	Sobre la media
20	desemp_razcuantitativo	Desempeño en razonamiento cuantitativo con relación a la media nacional	Bajo la media Sobre la media
21	desemp_compciudadanas	Desempeño en competencia ciudadana con relación a la media nacional	Bajo la media Sobre la media
22	facultades	Facultad a la que pertenece el programa académico	Ciencias Economicas y Administrativas Ciencias de la Salud Humanidades y Ciencias Sociales Ingeniería y Ciencias
23	forma_pago_matricula	Procedencia de recursos para pago de matrícula	Recursos Beca Recursos Beca-Credito Recursos Beca-Credito-Padres Recursos Beca-Credito-Padres-Propios Recursos Beca-Credito-Propios Recursos Beca-Padres Recursos Beca-Padres-Propios Recursos Beca-Propios Recursos Credito Recursos Credito-Padres Recursos Credito-Padres-Propios Recursos Credito-Propios Recursos Padres Recursos Padres-Propios Recursos Propios

Tabla 3.17: Diccionario de datos del repositorio sbpro_jave_final.

Repositorio	Descripción
sbpro_jave_final_Lec	Repositorio para análisis de la competencia de lectura crítica. Contiene 2052 registros y 17 atributos.
sbpro_jave_final_Esc	Repositorio para análisis de la competencia de comunicación escrita. Contiene 2052 registros y 17 atributos.
sbpro_jave_final_Cua	Repositorio para análisis de la competencia de razonamiento cuantitativo. Contiene 2052 registros y 17 atributos.
sbpro_jave_final_Ing	Repositorio para análisis de la competencia de inglés. Contiene 2052 registros y 17 atributos.
sbpro_jave_final_Ciu	Repositorio para análisis de la competencia ciudadana. Contiene 2052 registros y 17 atributos.
sbpro_jave_final_Gen	Repositorio para análisis del desempeño general en las competencias. Contiene 2052 registros y 17 atributos.

Tabla 3.18: Repositorios por cada competencia

Capítulo 4

Resultados

4.1. Ánalisis exploratorio

Mediante análisis estadístico descriptivo con la calificación media obtenida en las pruebas por estudiantes de cada programa de la Universidad Javeriana Cali, se explica el desempeño académico que presentaron en las competencias genéricas de las pruebas Saber Pro en el período estudiado. Se estableció el "efecto programa", entendido como las diferencias de promedios observadas entre los diferentes programas académicos institucionales y que en gran parte está explicado por la calidad del programa. Se utilizó para la medición de este efecto la diferencia de medias estandarizadas de cada programa con aquel que alcanza el máximo puntaje en cada competencia, llamado estadístico d de Cohen. Dicho estadístico permitió elaborar un ranking de los programas a nivel institucional y ubicar la posición que ocupa cada programa de institución en cada una de las competencias. Con la misma metodología se determinó el "efecto de las variables socioeconómicas", que permitió establecer la asociación con estas variables con el desempeño en las cinco competencias genéricas de los diferentes programas que ofrece la Universidad Javeriana.

Inicialmente se describen las características socioeconómicas, las cuales pueden generar posibles brechas de rendimiento académico de los estudiantes de la Universidad Javeriana Cali que presentaron las pruebas Saber Pro en el período 2017-2018, a partir de la información contenida en los formularios de inscripción.

Posteriormente, y con el fin de tener una comprensión preliminar de la relación entre los datos, se realiza un análisis de correlación entre las cinco competencias genéricas y finalmente se establece el efecto que tiene en el desempeño de dichas competencias y en el puntaje global las variables de programas académicos y las socio-económicas.

4.1.1. Efecto del Programa y de Características Socioeconómicas en las Competencias Genéricas en el Contexto Institucional

Para establecer el "efecto programa" y de las variables socioeconómicas de la Universidad de Javeriana Cali en el contexto institucional, se elaboró un ranking de las Programas que participaron en las pruebas Saber Pro durante el período 2017 a 2018 y se estableció una clasificación de los diferentes grupos de las variables socioeconómicas.

Para la elaboración del ranking se utilizó el estadístico d de Cohen [39] que permite calcular el Tamaño del Efecto Programa a partir de las diferencias estandarizadas entre los promedios en cada uno de los programas frente al programa que alcanzó el máximo puntaje promedio en cada una de las cinco competencias genéricas y que se toma como referencia 4.1:

$$d_i = \frac{\overline{X_0} - \overline{X_i}}{S_{i0}} \tag{4.1}$$

Donde: $S_{i0} = \sqrt{\frac{(n_0-1)(S_0)^2+(n_i-1)(S_i)^2}{n_o+n_i-2}}$

d_i = Tamaño del Efecto del *programa i* con relación al referente.
$\overline{X_0}$ = Media del programa de referencia.
$\overline{X_i}$ = Media del programa i.
$(S_0)^2$ = Varianza del programa de referencia.
$(S_i)^2$ = Varianza del programa i.
N_0 = número de estudiantes que presentaron las pruebas del programa de referencia.
N_i = número de estudiantes que presentaron las pruebas del programa i.

Para interpretar las diferencias en unidades estándar, Cohen [39] propone la siguiente escala, obtenidas con el estadístico de la ecuación 4.1: en el intervalo [0.0, 0.2] diferencias triviales o irrelevantes que clasificamos como grupo A, en el intervalo (0.2, 0.5] diferentes pequeñas que se clasifican en el grupo B, en (0.5, 0.8] diferencias moderadas agrupadas en C y en el intervalo (0.8, infinito) diferencias grandes se clasifican en el grupo D.

4.1.2. Características Socioeconómicas

En el período 2017-2018, un total de 2052 estudiantes de la Universidad Javeriana Cali presentaron las pruebas Sabe Pro. La tabla 4.1 presenta las características socioeconómicas de dichos estudiantes.

VARIABLES SOCIOECONÓMICAS		N	%
Género	Femenino	1.138	55,5 %
	Masculino	914	44,5 %
Grupos de edad	<=20 años	154	7,5 %
	21 años	441	21,5 %
	22 años	468	22,8 %
	23 años	377	18,4 %
	24 años	227	11,1 %
	>= 25 años	385	18,8 %

VARIABLES SOCIOECONÓMICAS		N	%
Estrato social	Estrato 0	1	0,1 %
	Estrato 1	29	1,5 %
	Estrato 2	112	5,8 %
	Estrato 3	329	17,0 %
	Estrato 4	520	26,8 %
	Estrato 5	610	31,4 %
	Estrato 6	335	17,3 %
	Sin Estrato	5	0,3 %
Hogar actual	Es habitual o permanente	1.576	83,8 %
	Es temporal por razones de estudio u otra razón	305	16,2 %
Cabeza de familia	Si	77	4,1 %
	No	1.804	95,9 %
Personas a cargo	Ninguna	1.740	92,5 %
	Una	70	3,7 %
	Dos	33	1,8 %
	Tres	22	1,2 %
	Cuatro	11	0,6 %
	Cinco	4	0,2 %
	Seis	1	0,1 %
Hacinamiento	Hacinamiento crítico	1	0,0 %
	Hacinamiento medio	12	0,6 %
	Sin hacinamiento	2.039	99,4 %
Educación del Padre	Ninguno	12	0,6 %
	Primaria incompleta	56	2,9 %
	Primaria completa	29	1,5 %
	Secundaria incompleta	89	4,5 %
	Secundaria completa	248	12,5 %
	Técnica o Tecnológica incompleta	53	2,7 %
	Técnica o Tecnológica completa	176	9,0 %
	Educación Profesional incompleta	115	5,9 %
	Educación Profesional completa	655	33,4 %
	Postgrado	487	24,8 %
	No Aplica	13	0,7 %
	No sabe	31	1,6 %
Educación de la Madre	Ninguno	4	0,2 %
	Primaria incompleta	28	1,4 %
	Primaria completa	27	1,4 %
	Secundaria incompleta	96	4,9 %
	Secundaria completa	272	13,8 %
	Técnica o Tecnológica incompleta	48	2,4 %

VARIABLES SOCIOECONÓMICAS		N	%
	Técnica o Tecnológica completa	256	13,0 %
	Educación Profesional incompleta	153	7,8 %
	Educación Profesional completa	707	36,0 %
	Postgrado	360	18,3 %
	No Aplica	3	0,2 %
	No sabe	10	0,5 %
Índice TICS	Malo	224	10,9 %
	Regular	17	0,8 %
	Bueno	1.811	88,3 %
Índice Electrodomésticos	Malo	177	8,6 %
	Regular	163	7,9 %
	Bueno	1.712	83,4 %
Familia tiene motocicleta	Si	364	19,5 %
	No	1.498	80,5 %
Familia tiene automóvil	Si	1.570	80,7 %
	No	375	19,3 %
Dpto. de procedencia	Valle del Cauca	1.969	96,0 %
	Otros	83	4,0 %
	Total	2052	100 %

Tabla 4.1: Características Socioeconómicas de estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro en el período 2017-2018.

Por género, la mayoría son mujeres, con un 55.5 %; por edad el mayor porcentaje se encuentra en 21 (21.3 %) y 22 (22.3 %) años. Igualmente, la mayoría de estudiantes pertenecen a los estratos 4, 5 y 6, que sumados representan el 75.5 %. El hogar actual de la gran mayoría de estudiantes, equivalente al 83.8 %, es habitual o permanente y solamente el 4.1 % son cabeza de familia con un 7.5 % que tiene una o más personas a cargo. Casi la totalidad de estudiantes (99.4 %) no presenta situación de hacinamiento en su vivienda.

Un poco más de la mitad de los progenitores de los estudiantes que presentaron las pruebas cuentan con estudios profesionales o de posgrado, en las madres el 54.3 % y en padres el 58.2 %. Los índices de TICs y Electrodomésticos son buenos en el 88.3 % y 83.4 % de los hogares de los estudiantes, respectivamente. De igual manera el 19.5 % de las familias poseen motocicleta y el 80.7 % tiene automóvil. Finalmente, el 96 % es de origen de algún municipio del departamento del Valle.

4.1.3. Características Personales y Académicas

La tabla 4.2 presenta algunas características personales y académicas de los estudiantes de la Universidad Javeriana Cali que presentaron las pruebas Saber Pro en los años 2017 y 2018. La

mayoría de estudiantes, representados en un 88.4 %, son bachilleres en modalidad académica, pagan matrículas superiores a los 5.5 millones de pesos (87 %), un 15 % esta becado, 38.3 % paga su matrícula con crédito, el 14.5 % con sus ingresos y el 85.9 % recibe ayuda de sus padres. El 77.6 % de los estudiantes preparó la prueba Saber 11 haciendo repaso o mediante un curso, el 63 % tiene al menos 26 libros en su biblioteca, el 52.6 % dedica al menos 30 minutos a la lectura de libros y el 84.7 % se conecta al menos una hora a internet. Finalmente, el 55.5 % dedica en la semana algún tiempo a una actividad laboral.

VARIABLES PERSONALES Y ACADÉMICAS		N	%
Modalidad Bachiller	Bachiller académico	1.661	88,4 %
	Bachiller pedagógico o normalista	13	0,7 %
	Bachiller técnico	205	10,9 %
Valor de matrícula	No paga matrícula	8	0,4 %
	Menos de 500 mil	1	0,1 %
	Entre 500 mil y menos de 1 millón	7	0,4 %
	Entre 1 millón y menos de 2.5 millones	16	0,9 %
	Entre 2.5 millones y menos de 4 millones	88	4,7 %
	Entre 4 millones y menos de 5.5 millones	124	6,6 %
	Entre 5.5 millones y menos de 7 millones	721	38,4 %
	Más de 7 millones	914	48,6 %
Becado	Si	282	15,0 %
	No	1.597	85,0 %
Pago matrícula con crédito	Si	720	38,3 %
	No	1.159	61,7 %
Pago matrícula padres	Si	1.615	85,9 %
	No	264	14,1 %
Pago matrícula propio	Si	272	14,5 %
	No	1.607	85,5 %
Preparación Prueba Saber 11	No realizó ninguna prueba de preparación	422	22,4 %
	Repasó por cuenta propia	951	50,6 %
	Tomó un curso de preparación	507	27,0 %
Núm. libros	0 a 10 libros	242	12,9 %
	11 a 25 libros	455	24,2 %
	26 a 100 libros	756	40,2 %
	Más de 100 libros	428	22,8 %
Dedicación lectura diaria	No leo por entretenimiento	240	12,8 %
	30 minutos o menos	651	34,6 %
	Entre 30 y 60 minutos	634	33,7 %
	Entre 1 y 2 horas	248	13,2 %
	Más de 2 horas	108	5,7 %

VARIABLES PERSONALES Y ACADÉMICAS		N	%
Dedicación Internet	Menos de una hora	287	15,3 %
	Entre 1 y 3 horas	1.043	55,7 %
	Más de 4 horas	543	29,0 %
Horas semana trabaja	Ninguna	863	44,5 %
	Menos de 10 horas	262	13,5 %
	Entre 11 y 20 horas	219	11,3 %
	Entre 21 y 30 horas	188	9,7 %
	Más de 30 horas	406	20,9 %
	Total	2052	100 %

Tabla 4.2: Características Personales y Académicas de estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro en el período 2017-2018

4.1.4. Programas Académicas

La tabla 4.3 presenta la distribución por Programas Académicos de los estudiantes de la Universidad Javeriana Cali que presentaron las pruebas Saber Pro en los años 2017 y 2018.

No.	PROGRAMA	N	%
1	Administración De Empresas	356	17,3 %
2	Arquitectura	125	6,1 %
3	Artes Visuales	38	1,9 %
4	Biología	26	1,3 %
5	Ciencia Política	44	2,1 %
6	Comunicación	81	3,9 %
7	Contaduría Publica	49	2,4 %
8	Derecho	191	9,3 %
9	Diseño De Comunicación Visual	105	5,1 %
10	Economía	59	2,9 %
11	Filosofía	5	0,2 %
12	Ingeniería Civil	163	7,9 %
13	Ingeniería De Sistemas Y Computación	39	1,9 %
14	Ingeniería Electrónica	50	2,4 %
15	Ingeniería Industrial	180	8,8 %
16	Matemáticas Aplicadas	9	0,4 %
17	Medicina	194	9,5 %
18	Negocios Internacionales	186	9,1 %
19	Psicología	152	7,4 %
	Total	2052	100 %

Tabla 4.3: Características Personales y Académicas de estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro en el período 2017-2018

En el período 2017-2018 de los 2052 estudiantes que presentaron las pruebas Saber Pro la mayoría pertenecían al programa de Administración de Empresas (17.3 %) seguidos por estudiantes de Medicina, Derecho y Negocios Internacionales con porcentajes un tanto superiores al 9 %, los demás programas presentan participaciones inferiores a dicho valor.

4.1.5. Correlación de puntajes entre las Competencias Genéricas

A través del coeficiente de correlación de Pearson se establece como se asocian linealmente, entre sí, los puntajes obtenidos en las cinco competencias genéricas y el puntaje global por los estudiantes de los diferentes programas de la Potificia Universidad Javeriana Cali que presentaron las pruebas en el período 2017-2018. Los resultados se presentan en la tabla 4.4.

Competencia	Lectura Crítica	Comunicación Escrita	Razonamiento Cuantitativo	Inglés	Competencias Ciudadanas	Puntaje Global
Lectura Crítica	1	0,188**	0,436**	0,363**	0,560**	0,742**
Comunicación Escrita		1	0,119**	0,171**	0,169**	0,527**
Razonamiento Cuantitativo			1	0,298**	0,319**	0,625**
Inglés				1	0,344**	0,619**
Competencias Ciudadanas					1	0,720**
Puntaje Global						1

Tabla 4.4: Matriz de correlaciones de las Competencias genéricas.

**. La correlación es significativa en el nivel 0,01 (2 colas).

De acuerdo con la tabla anterior, todas las correlaciones resultan positivas y altamente significativas (p valor <0,01), esto último debido al tamaño de los datos. Siguiendo la clasificación de Cohen [39], para la interpretación del coeficiente de Pearson, se observó que lectura crítica presenta una correlación alta (r>0,5) con competencias ciudadanas, una correlación moderada (0,3<r<=0,5) con inglés y razonamiento cuantitativo y una correlación baja con comunicación escrita (r<=0,3). Esta

última competencia también presenta una correlación baja con el resto de competencias. Razonamiento cuantitativo tiene una correlación baja con inglés y moderada con competencias ciudadanas, mientras que estas últimas competencias presentan una correlación moderada.

Con el puntaje global todas las competencias presentan correlaciones altas (r>0,5), sin embargo, la comunicación escrita es la de más baja correlación y lectura escrita es la de mayor correlación.

De lo anterior se puede concluir que los resultados en la competencia de comunicación escrita, por sus correlaciones bajas, son un tanto independientes de los puntajes correspondientes a las demás competencias genéricas.

4.2. Efecto del Programa en las Competencias de la Prueba en el Contexto Institucional

4.2.1. Efecto del Programa en Lectura Crítica

Según la tabla 4.5, durante el período 2017 a 2018 los estudiantes de los programas de Filosofía, Matemáticas Aplicadas y Medicina presentaron el mejor desempeño en la competencia de Lectura Crítica, con diferencias irrelevantes en tamaño (d <= 0,2; grupo A). Los programas de Ingeniería de Sistemas y Computación, Ciencia Política, Economía y Derecho alcanzan diferencias moderadas (grupo B). Los demás programas alcanzan tamaños de efectos moderados o grandes (grupos C y D).

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
1	Filosofía	5	190,2	18,8	-	A
2	Matemáticas Aplicadas	9	189,9	23,7	0,01	A
3	Medicina	194	186,0	23,4	0,18	A
4	Ingeniería De Sistemas Y Computación	39	184,8	23,4	0,23	B
5	Ciencia Política	44	180,2	32,0	0,32	B
6	Economía	59	179,1	27,0	0,42	B
7	Derecho	191	177,0	32,0	0,41	B
8	Psicología	152	175,8	27,1	0,53	C
9	Ingeniería Electrónica	50	174,6	33,6	0,48	C
10	Comunicación	81	172,7	29,6	0,60	C
11	Ingeniería Civil	163	172,6	26,6	0,67	C
12	Ingeniería Industrial	180	170,9	27,0	0,72	C
13	Biología	26	168,3	29,9	0,77	C

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
14	Negocios Internacionales	186	167,5	29,3	0,78	C
15	Artes Visuales	38	166,7	28,9	0,84	D
16	Arquitectura	125	166,4	25,7	0,93	D
17	Diseño De Comunicación Visual	105	158,4	32,1	1,00	D
18	Contaduría Publica	49	158,1	27,0	1,21	D
19	Administración De Empresas	356	156,9	30,6	1,09	D
	General	2052	170,3	29,9	0,66 6	-

Tabla 4.5: Clasificación por Programas Académicos de la Pontificia Universidad de Javeriana Cali en Lectura Crítica de las pruebas SABER PRO 2017-2018.

4.2.2. Efecto del Programa en Comunicación Escrita

Como se observa en la tabla 4.6, en la competencia de Comunicación Escrita, durante el período 2017 a 2018, Filosofía alcanzó el mejor desempeño (grupo A), entre los programas académicos que ofrece la Universidad Javeriana, seguido por Derecho, Matemáticas Aplicadas, Economía, Ciencia Política y Medicina con diferencias pequeñas (d <= 0,5; grupo B). Los demás programas alcanzan tamaños de efecto moderado o grande en comparación al primer programa (grupos C y D).

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
1	Filosofía	5	184,6	6,4	-	A
2	Derecho	191	173,4	33,1	0,34	B
3	Matemáticas Aplicadas	9	171,8	62,5	0,25	B
4	Economía	59	171,8	62,5	0,25	B
5	Ciencia Política	44	170,8	30,6	0,47	B
6	Medicina	194	170,7	33,0	0,43	B
7	Psicología	152	166,6	35,0	0,52	C
8	Artes Visuales	38	165,1	37,2	0,55	C
9	Negocios Internacionales	186	165,0	27,9	0,71	C
10	Comunicación	81	164,2	38,5	0,54	C
11	Ingeniería Electrónica	50	162,6	26,2	0,87	C
12	Ingeniería Industrial	180	161,6	33,2	0,70	C
13	Biología	26	157,3	38,8	0,76	C

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
14	Administración De Empresas	356	156,7	30,1	0,93	D
15	Diseño De Comunicación Visual	105	156,2	36,1	0,80	D
16	Arquitectura	125	155,9	34,6	0,84	D
17	Ingeniería De Sistemas Y Computación	39	153,7	33,1	0,98	D
18	Ingeniería Civil	163	153,7	30,2	1,03	D
19	Contaduría Publica	49	152,9	29,4	1,12	D
	General	2052	162,5	33,4	0,6 6	-

Tabla 4.6: Clasificación por Programas Académicos de la Pontificia Universidad Javeriana Cali en Comunicación Escrita de las pruebas SABER PRO 2017-2018.

4.2.3. Efecto del Programa en Razonamiento Cuantitativo

En Razonamiento Cuantitativo el programa de Matemáticas Aplicadas alcanzó el mejor promedio en el período analizado (grupo A). Los programas de Ingenierías: Electrónica, Sistemas y Computación e Industrial ocupan los siguientes lugares con diferencias en el desempeño moderadas en tamaño (d <= 0,8; grupo C). El resto de programas presentan diferencias grandes (grupo D) en esta competencia al compararlas con el de Matemáticas Aplicadas (tabla 4.7).

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
1	Matemáticas Aplicadas	9	202,2	14,5	-	A
2	Ingeniería Electrónica	50	191,0	19,1	0,60	C
3	Ingeniería De Sistemas Y Computación	39	190,9	21,1	0,56	C
4	Ingeniería Civil	163	188,7	26,5	0,52	C
5	Ingeniería Industrial	180	186,4	22,2	0,72	C
6	Medicina	194	181,9	21,1	0,97	D
7	Economía	59	177,2	23,3	1,12	D
8	Biología	26	175,0	20,6	1,41	D
9	Negocios Internacionales	186	169,2	27,1	1,24	D
10	Administración De Empresas	356	164,7	25,3	1,49	D
11	Arquitectura	125	163,0	21,7	1,84	D
12	Filosofía	5	162,8	37,1	1,61	D

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
13	Contaduría Publica	49	162,7	23,3	1,77	D
14	Derecho	191	161,7	27,4	1,50	D
15	Psicología	152	155,1	26,8	1,79	D
16	Ciencia Política	44	152,5	25,5	2,06	D
17	Diseño De Comunicación Visual	105	152,4	24,7	2,06	D
18	Artes Visuales	38	149,4	28,6	1,98	D
19	Comunicación	81	146,8	28,9	1,99	D
	General	2052	169,3	28,1	1,17 6	-

Tabla 4.7: Clasificación por Programas Académicos de la Pontificia Universidad Javeriana Cali en Razonamiento Cuantitativa de las pruebas SABER PRO 2017-2018.

4.2.4. Efecto del Programa en Inglés

Los programas de Ingeniería de Sistemas y Computación y de Negocios Internacionales, obtuvieron el mejor desempeño en la prueba de inglés, durante el período estudiado, con diferencia no relevante (d <= 0,2; grupo A), y las posiciones siguientes son ocupadas por los programas de Filosofía, Medicina, Ingeniería Electrónica y Biología con diferencias pequeñas (d <= 0,5; grupo B). Los demás programas alcanzan tamaños de efecto superiores en los grupos C y D (tabla 4.8).

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
1	Ingeniería De Sistemas Y Computación	39	204,9	29,1	-	A
2	Negocios Internacionales	186	204,6	15,7	0,01	A
3	Filosofía	5	197,8	17,2	0,25	B
4	Medicina	194	193,3	24,0	0,46	B
5	Ingeniería Electrónica	50	193,0	29,7	0,40	B
6	Biología	26	192,7	39,3	0,36	B
7	Diseño De Comunicación Visual	105	186,9	21,3	0,76	C
8	Ingeniería Industrial	180	186,9	24,2	0,72	C
9	Psicología	152	185,6	29,3	0,66	C
10	Artes Visuales	38	184,6	25,7	0,74	C
11	Ciencia Política	44	184,3	28,6	0,71	C
12	Derecho	191	182,2	29,4	0,77	C

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
13	Economía	59	182,0	27,0	0,82	C
14	Ingeniería Civil	163	181,4	26,3	0,87	C
15	Matemáticas Aplicadas	9	181,2	30,7	0,81	D
16	Comunicación	81	179,1	32,3	0,82	D
17	Arquitectura	125	178,4	24,5	1,03	D
18	Administración De Empresas	356	177,1	28,3	0,98	D
19	Contaduría Publica	49	157,8	22,7	1,83	D
	General	2052	185,3	27,8	0,71 6	-

Tabla 4.8: Clasificación por Programas Académicos de la Pontificia Universidad Javeriana Cali en inglés de las pruebas SABER PRO 2017-2018.

4.2.5. Efecto del Programa en Competencias Ciudadanas

En las Competencias Ciudadanas los programas de mejor desempeño son Ciencia Política y Derecho con diferencias irrelevantes (d <= 0,2; grupo A). Las siguientes posiciones las ocupan, Economía, Medicina, Ingeniería de Sistemas y Computación, Filosofía, Ingeniería Electrónica, Ingeniería Civil y Psicología, en su orden, con tamaños de efecto pequeños (d <= 0,5; grupo B). Los programas restantes presentan diferencias moderadas o grandes en tamaño del efecto, clasificadas en los grupos C y D (tabla 4.9).

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
1	Ciencia Política	44	177,0	31,9	-	A
2	Derecho	191	175,1	30,6	0,06	A
3	Economía	59	169,1	26,5	0,27	B
4	Medicina	194	168,4	28,1	0,30	B
5	Ingeniería De Sistemas Y Computación	39	166,1	34,8	0,33	B
6	Filosofía	5	164,8	35,4	0,38	B
7	Ingeniería Electrónica	50	163,0	32,2	0,44	B
8	Ingeniería Civil	163	161,6	30,5	0,50	B
9	Psicología	152	161,4	31,0	0,50	B
10	Negocios Internacionales	186	157,9	31,6	0,60	C
11	Ingeniería Industrial	180	157,9	33,3	0,58	C
12	Matemáticas Aplicadas	9	155,1	36,1	0,67	C
13	Comunicación	81	154,4	32,2	0,70	C

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
14	Biología	26	154,2	28,1	0,75	C
15	Artes Visuales	38	152,3	29,9	0,80	C
16	Arquitectura	125	148,6	30,3	0,92	D
17	Contaduría Publica	49	146,3	25,6	1,07	D
18	Administración De Empresas	356	145,2	33,6	0,95	D
19	Diseño De Comunicación Visual	105	141,8	29,6	1,16	D
	General	2052	157,9	32,7	0,58	-

Tabla 4.9: Clasificación por Programas Académicos de la Pontificia Universidad Javeriana Cali en Competencias Ciudadanas de las pruebas SABER PRO 2017-2018.

4.2.6. Efecto del Programa en Puntaje Global

En el rendimiento global de la prueba o puntaje global los programas de mejor desempeño son Filosofía, Matemáticas Aplicadas, Ingeniería de Sistemas y Computación y Medicina, con tamaños de efecto no relevantes (d <= 0,2; grupo A). En los siguientes lugares se ubican los programas de Ingeniería Electrónica, Economía, Derecho, Ingeniería Industrial, Negocios Internacionales, Ciencia Política e Ingeniería Civil, en su orden, a diferencias pequeñas del primer lugar (d <= 0,5; grupo B). Los demás programas presentan diferencias moderadas o grandes en tamaño del efecto clasificados en los grupos C y D (tabla 4.10).

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
1	Filosofía	5	180,2	14,2	-	A
2	Matemáticas Aplicadas	9	180,0	23,0	0,01	A
3	Ingeniería De Sistemas Y Computación	39	179,5	20,1	0,04	A
4	Medicina	194	179,2	17,3	0,06	A
5	Ingeniería Electrónica	50	176,3	19,2	0,21	B
6	Economía	59	175,7	19,1	0,24	B
7	Derecho	191	173,5	21,2	0,32	B
8	Ingeniería Industrial	180	172,1	18,5	0,44	B
9	Negocios Internacionales	186	172,1	19,2	0,43	B
10	Ciencia Política	44	171,4	20,0	0,45	B
11	Ingeniería Civil	163	171,1	20,4	0,45	B

No.	Programa	N	Media	Desviación Estándar	d Cohen	Grupo
12	Biología	26	169,5	20,6	0,54	C
13	Psicología	152	168,7	19,1	0,60	C
14	Comunicación	81	162,9	22,1	0,80	C
15	Arquitectura	125	162,0	18,8	0,97	D
16	Artes Visuales	38	161,9	19,6	0,95	D
17	Administración De Empresas	356	159,2	20,6	1,02	D
18	Diseño De Comunicación Visual	105	158,9	18,7	1,15	D
19	Contaduría Publica	49	154,9	14,9	1,70	D
	General	2052	168,4	20,7	0,57	-

Tabla 4.10: Clasificación por Programas Académicos de la Pontificia Universidad Javeriana Cali en Puntaje Global de las pruebas SABER PRO 2017-2018.

4.3. Efecto de las Variables Socioecómicas en las Competencias Genéricas en el Contexto Institucional

A continuación, se establece el efecto de las diferentes variables socioeconómicas consideradas en este estudio, sobre los puntajes obtenidos en las 5 competencias de las Pruebas Saber Pro y en el puntaje Global, a partir de las diferencias estandarizadas de los promedios obtenidos en las pruebas, en los diferentes grupos que conforman dichas variables.

4.3.1. Efecto en Lectura Crítica

De la tabla 4.11 se concluye que en Lectura Crítica de las pruebas Saber Pro 2017-2018, los hombres presentan similar desempeño que las mujeres con diferencias irrelevantes (d <= 0,2). Por edad se observó mejor desempeño en los grupos de menor edad. No se observaron diferencias importantes entre los estratos sociales a excepción del estrato 0, como caso atípico. Igualmente el hecho de que el hogar actual del estudiante sea permanente o temporal no afecta el rendimiento en la prueba de Lectura Crítica, sin embargo el ser cabeza de familia afecta el rendimiento en esta competencia, pues como se observa en la tabla el estudiante que es cabeza de familia tiene más bajo rendimiento que aquel que no lo es, con diferencias moderadas, de la misma manera el tener personas a cargo hace que el rendimiento en la competencia sea inferior a los que no las tienen, excepto para el caso atípico de aquellos que tienen 6 personas a cargo. Los estudiantes con algún tipo de hacinamiento presentan, aunque con una diferencia pequeña, un mejor rendimiento en Lectura Crítica, que aquellos que no lo tienen. La educación del padre y la educación de la madre se relacionan positivamente con el rendimiento en la prueba, observándose que, a mayor nivel educativo de los progenitores del estudiante, mayor es el rendimiento del estudiante. Finalmente, los factores relacionados con

TICs, electrodomésticos y tener medio de transporte propio (motocicleta o vehículo) tienen efectos irrelevantes o pequeños en el rendimiento de la prueba de Lectura Crítica.

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
Género	Femenino	1.138	169,8	29,2	0,04	A
	Masculino	914	170,9	30,8	-	A
Grupos de edad	<=20 años	154	179,9	26,4	-	A
	21 años	441	174,4	29,2	0,19	A
	22 años	468	173,4	29,5	0,23	B
	23 años	377	170,0	28,7	0,35	B
	24 años	227	164,3	30,4	0,54	C
	>= 25 años	385	162,0	31,1	0,60	C
Estrato social	Estrato 0	1	207,0		-	A
	Estrato 1	29	170,2	24,9	1,48	D
	Estrato 2	112	174,6	24,6	1,32	D
	Estrato 3	329	168,2	28,7	1,35	D
	Estrato 4	520	170,4	29,7	1,23	D
	Estrato 5	610	170,9	29,9	1,21	D
	Estrato 6	335	167,7	34,6	1,13	D
	Sin Estrato	5	169,2	30,5	1,24	D
Hogar actual	Es habitual o permanente	1.576	168,9	30,1	0,19	A
	Es temporal por razones de estudio u otra razón	305	174,5	28,4	-	A
Cabeza de familia	Si	77	154,8	34,9	0,53	C
	No	1.804	170,4	29,5	-	A
Personas a cargo	Ninguna	1.740	170,9	29,5	0,75	C
	Una	70	153,9	30,4	1,29	D
	Dos	33	159,6	30,3	1,10	D
	Tres	22	152,7	37,0	1,09	D
	Cuatro	11	153,5	32,7	1,21	D
	Cinco	4	169,8	3,3	7,04	D
	Seis	1	193,0		-	A
Hacinamiento	Hacinamiento crítico	1	182,0		-	A
	Hacinamiento medio	12	180,4	28,5	0,06	A
	Sin hacinamiento	2.039	170,3	30,0	0,39	B
Educación del Padre	Ninguno	12	156,5	30,0	0,66	C
	Primaria incompleta	56	155,9	30,1	0,68	C
	Primaria completa	29	161,2	30,3	0,50	B
	Secundaria incompleta	89	163,1	27,8	0,45	B

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Secundaria completa	248	166,0	30,8	0,34	B
	Técnica o Tecnológica incompleta	53	167,4	29,2	0,30	B
	Técnica o Tecnológica completa	176	169,9	27,5	0,22	B
	Educación Profesional incompleta	115	174,1	29,7	0,07	A
	Educación Profesional completa	655	169,3	30,0	0,23	B
	Postgrado	487	176,3	29,9	-	A
	No Aplica	13	169,7	32,7	0,22	B
	No sabe	31	170,3	27,2	0,20	A
Educación de la Madre	Ninguno	4	160,3	26,1	0,75	C
	Primaria incompleta	28	161,1	19,4	0,83	D
	Primaria completa	27	153,3	26,7	0,96	D
	Secundaria incompleta	96	166,8	31,7	0,37	B
	Secundaria completa	272	165,1	28,9	0,45	B
	Técnica o Tecnológica incompleta	48	169,4	28,8	0,31	B
	Técnica o Tecnológica completa	256	167,4	31,1	0,35	B
	Educación Profesional incompleta	153	169,2	30,0	0,30	B
	Educación Profesional completa	707	171,4	30,4	0,22	B
	Postgrado	360	176,0	29,1	0,07	A
	No Aplica	3	143,7	48,8	1,17	D
	No sabe	10	178,1	23,1	-	A
Índice TICS	Malo	224	173,8	27,4	-	A
	Regular	17	168,9	28,6	0,18	A
	Bueno	1.811	169,9	30,2	0,13	A
Índice Electrodomésticos	Malo	177	176,1	25,6	-	A
	Regular	163	167,5	29,2	0,32	B
	Bueno	1.712	170,0	30,4	0,20	A
Familia tiene motocicleta	Si	364	165,1	30,8	0,21	B
	No	1.498	171,5	29,8	-	A
Familia tiene automóvil	Si	1.570	169,6	30,5	0,03	A
	No	375	170,6	28,7	-	A

VARIABLES SOCIOECONÓMICAS	N	Media	DT	d Cohen	Grupo

Tabla 4.11: Variables Sociodemográficas y desempeño académico en Lectura Crítica de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro 2017-2018.

4.3.2. Efecto en Comunicación Escrita

Como se observa en la tabla 4.12, en la competencia de Comunicación Escrita de las pruebas Saber Pro 2017-2018, las mujeres presentan mejor desempeño que los hombres con diferencias pequeñas (d <= 0,5). Por edad se observó que los más jóvenes (menores de 22 años) tienen mejor desempeño en esta competencia. Los estudiantes de estrato uno, presentaron mejor rendimiento que los de estratos restantes, los cuales no presentan entre ellos diferencias importantes. El hogar actual del estudiante (permanente o temporal) no es un factor que afecte el rendimiento en la prueba, pero el ser cabeza de familia si afecta el rendimiento en esta competencia, pues como se observa en la tabla el estudiante que es cabeza de familia tiene más bajo rendimiento que aquel que no lo es, con diferencias pequeñas, sin embargo el tener personas a cargo no hace que el rendimiento en la competencia sea inferior a los que no las tienen, con excepción del caso atípico, en aquellos que tienen 6 personas a cargo (1 solo caso) quien presentó mejor puntaje. Llama la atención que los estudiantes que presentan algún tipo de hacinamiento tienen un mayor rendimiento en la prueba de Escritura que aquellos que no lo tienen. La educación de los progenitores (padre y madre) de los estudiantes son factores que se relacionan positivamente con el rendimiento en la prueba, observándose que a mayor nivel educativo de ellos los estudiantes presentan un mejor desempeño. En general, los factores relacionados con TICs, electrodomésticos y tener medio de transporte propio (motocicleta o vehículo) no presentan efectos relevantes en el rendimiento de la prueba de Comunicación Escrita.

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
Género	Femenino	1.138	165,5	33,5	-	A
	Masculino	914	158,6	32,8	0,21	B
Grupos de edad	<=20 años	154	170,2	41,6	-	A
	21 años	441	166,1	34,4	0,11	A
	22 años	468	163,9	30,4	0,17	A
	23 años	377	161,1	31,6	0,26	B
	24 años	227	163,0	29,5	0,21	B
	>= 25 años	385	154,8	30,4	0,45	B
Estrato social	Estrato 0	1	153,0		0,56	C
	Estrato 1	29	175,2	39,9	-	A
	Estrato 2	112	159,5	31,5	0,47	B
	Estrato 3	329	159,8	35,4	0,43	B
	Estrato 4	520	164,0	34,8	0,32	B
	Estrato 5	610	163,2	30,5	0,39	B

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Estrato 6	335	160,8	33,2	0,43	B
	Sin Estrato	5	162,5	31,6	0,33	B
Hogar actual	Es habitual o permanente	1.576	161,9	33,2	0,11	A
	Es temporal por razones de estudio u otra razón	305	165,6	34,8	-	A
Cabeza de familia	Si	77	155,7	26,7	0,21	B
	No	1.804	162,8	33,7	-	A
Personas a cargo	Ninguna	1.740	163,0	33,9	1,21	D
	Una	70	156,2	30,0	1,14	D
	Dos	33	157,9	30,6	1,17	D
	Tres	22	159,1	24,6	1,50	D
	Cuatro	11	148,1	23,1	1,13	D
	Cinco	4	153,8	7,7	4,14	D
	Seis	1	122,0		-	A
Hacinamiento	Hacinamiento crítico	1	181,0		-	A
	Hacinamiento medio	12	167,2	33,1	0,42	B
	Sin hacinamiento	2.039	162,4	33,4	0,56	C
Educación del Padre	Ninguno	12	151,3	27,5	0,63	D
	Primaria incompleta	56	151,7	31,5	0,61	D
	Primaria completa	29	159,3	32,8	0,36	B
	Secundaria incompleta	89	160,1	29,3	0,37	B
	Secundaria completa	248	160,4	34,3	0,32	B
	Técnica o Tecnológica incompleta	53	158,0	26,6	0,47	B
	Técnica o Tecnológica completa	176	157,1	27,9	0,50	B
	Educación Profesional incompleta	115	160,8	31,2	0,33	B
	Educación Profesional completa	655	163,9	33,4	0,22	B
	Postgrado	487	165,8	36,0	0,15	A
	No Aplica	13	171,3	34,8	-	A
	No sabe	31	164,4	25,2	0,24	B
Educación de la Madre	Ninguno	4	143,0	49,5	0,91	D
	Primaria incompleta	28	158,7	32,8	0,40	B
	Primaria completa	27	148,7	31,5	0,76	C
	Secundaria incompleta	96	155,7	29,2	0,53	C
	Secundaria completa	272	159,7	29,3	0,38	B

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Técnica o Tecnológica incompleta	48	158,8	31,1	0,40	B
	Técnica o Tecnológica completa	256	161,3	32,8	0,29	B
	Educación Profesional incompleta	153	164,0	33,2	0,21	B
	Educación Profesional completa	707	163,5	33,5	0,22	B
	Postgrado	360	165,9	36,5	0,13	A
	No Aplica	3	142,0	11,3	1,49	D
	No sabe	10	170,8	20,6	-	A
Índice TICS	Malo	224	163,2	38,7	-	A
	Regular	17	153,0	33,9	0,27	B
	Bueno	1.811	162,5	32,6	0,02	A
Índice Electrodomésticos	Malo	177	165,5	32,6	-	A
	Regular	163	160,0	35,0	0,16	A
	Bueno	1.712	162,4	33,3	0,09	A
Familia tiene motocicleta	Si	364	157,0	31,8	0,20	A
	No	1.498	163,7	33,7	-	A
Familia tiene automóvil	Si	1.570	162,3	32,6	0,00	A
	No	375	162,2	36,0	-	A

Tabla 4.12: Variables Sociodemográficas y desempeño académico en Comunicación Escrita de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro 2017-2018.

4.3.3. Efecto en Razonamiento Cuantitativo

Según la tabla 4.13, en la prueba de Razonamiento Cuantitativo, los hombres presentan mejor desempeño que las mujeres con diferencias moderadas (d <= 0,8). De igual manera los más jóvenes (menores de 22 años) tienen mejor desempeño en esta competencia. Por estrato social se observaron diferencias entre no relevantes, pequeñas y grandes, sin presentar alguna tendencia. Los estudiantes con hogar actual permanente presentan más bajo desempeño en la prueba, con diferencias pequeñas, que aquellos en los cuales su hogar es temporal, mientras que el ser cabeza de familia no afecta el rendimiento en esta competencia, pues como se observa en la tabla el estudiante que es cabeza de familia tiene similar rendimiento que aquel que no lo es, y de otra parte el número de personas a cargo muestra diferencias pequeñas, moderadas y grandes en el desempeño de la competencia sin presentar, al igual que en estrato social, alguna tendencia. El hacinamiento en los hogares de los estudiantes no genera algún efecto en el rendimiento de Razonamiento Cuantitativo, si exceptuamos el caso atípico de hacinamiento crítico. Al igual que en las competencias anteriores el nivel de

educación tanto del padre como de la madre se relacionan positivamente con el rendimiento en la prueba, observándose que, a mayor nivel de educación de los padres, se observa mejor desempeño en la prueba. Mejores condiciones de TICs y electrodomésticos no producen mejor desempeño en la competencia y tener medio de transporte propio (motocicleta o vehículo) no tiene efectos importantes en el rendimiento en la Comunicación Escrita.

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
Género	Femenino	1.138	162,2	26,6	0,59	C
	Masculino	914	178,1	27,3	-	A
Grupos de edad	<=20 años	154	174,9	27,4	-	A
	21 años	441	171,8	29,1	0,11	A
	22 años	468	170,1	27,6	0,17	A
	23 años	377	169,3	26,9	0,21	B
	24 años	227	166,3	28,5	0,31	B
	>= 25 años	385	165,0	27,9	0,35	B
Estrato social	Estrato 0	1	146,0		0,93	D
	Estrato 1	29	170,1	29,3	0,12	A
	Estrato 2	112	172,0	24,0	0,07	A
	Estrato 3	329	166,7	27,8	0,25	B
	Estrato 4	520	169,5	27,5	0,15	A
	Estrato 5	610	170,3	28,7	0,11	A
	Estrato 6	335	167,0	30,1	0,22	B
	Sin Estrato	5	173,6	29,7	-	A
Hogar actual	Es habitual o permanente	1.576	167,5	27,6	0,22	B
	Es temporal por razones de estudio u otra razón	305	173,5	25,8	-	A
Cabeza de familia	Si	77	167,2	26,3	0,05	A
	No	1804	168,5	27,4	-	A
Personas a cargo	Ninguna	1.740	169,2	27,2	0,29	B
	Una	70	161,0	31,9	0,50	B
	Dos	33	154,1	23,8	0,96	D
	Tres	22	165,5	26,9	0,43	B
	Cuatro	11	154,3	29,6	0,77	C
	Cinco	4	166,5	21,2	0,50	B
	Seis	1	177,0		-	A
Hacinamiento	Hacinamiento crítico	1	201,0		-	A
	Hacinamiento medio	12	167,6	33,9	0,99	D
	Sin hacinamiento	2.039	169,3	28,0	1,13	D
Educación del Padre	Ninguno	12	165,4	28,7	0,24	B

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Primaria incompleta	56	159,5	28,9	0,44	B
	Primaria completa	29	164,5	28,3	0,27	B
	Secundaria incompleta	89	164,9	28,3	0,25	B
	Secundaria completa	248	164,9	28,3	0,26	B
	Técnica o Tecnológica incompleta	53	166,0	25,9	0,22	B
	Técnica o Tecnológica completa	176	168,7	24,8	0,13	A
	Educación Profesional incompleta	115	172,1	28,3	-	A
	Educación Profesional completa	655	169,8	26,9	0,09	A
	Postgrado	487	172,0	30,5	0,00	A
	No Aplica	13	171,9	33,8	0,01	A
	No sabe	31	167,4	29,0	0,17	A
Educación de la Madre	Ninguno	4	139,0	18,9	1,13	D
	Primaria incompleta	28	163,5	24,6	0,31	B
	Primaria completa	27	155,9	26,2	0,56	B
	Secundaria incompleta	96	163,6	25,7	0,31	B
	Secundaria completa	272	171,7	25,6	0,03	A
	Técnica o Tecnológica incompleta	48	164,8	29,7	0,26	B
	Técnica o Tecnológica completa	256	171,7	25,6	0,03	A
	Educación Profesional incompleta	153	166,7	27,7	0,20	A
	Educación Profesional completa	707	170,2	28,3	0,09	A
	Postgrado	360	172,7	29,9	-	A
	No Aplica	3	167,0	47,5	0,19	A
	No sabe	10	153,7	32,0	0,63	C
Índice TICS	Malo	224	175,9	26,2	-	A
	Regular	17	168,9	32,0	0,26	B
	Bueno	1.811	168,5	28,2	0,27	B
Índice Electrodomésticos	Malo	177	175,2	26,6	-	A
	Regular	163	163,3	27,7	0,44	B
	Bueno	1.712	169,3	28,1	0,21	B
Familia tiene motocicleta	Si	364	169,5	26,3	-	A
	No	1.498	169,5	28,3	0,00	A
Familia tiene automóvil	Si	1.570	168,9	28,3	0,00	A

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	No	375	168,7	28,8	0,01	A

Tabla 4.13: Variables Sociodemográficas y desempeño académico en Razonamiento Cuantitativo de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro 2017-2018.

4.3.4. Efecto en Inglés

En la competencia de inglés, según la tabla 4.14, las mujeres y los hombres presentan diferencias no relevantes (d <= 0,2), sin observar diferencias importantes por edad, solamente los estudiantes con edad iguales superiores a 25 años presentan desempeños más bajos con diferencias pequeñas. Se observa una tendencia en el desempeño de inglés según el estrato social, el rendimiento es mejor a medida que aumenta el estrato. El hogar actual del estudiante (permanente o temporal) no es un factor que afecte el desempeño en la prueba, pero el ser cabeza de familia si afecta el rendimiento en esta competencia, pues como se observa en la tabla los estudiantes cabeza de familia tiene más bajo rendimiento que aquellos que no lo son, con diferencias moderadas: Aunque se observa que el número personas a cargo, por parte de los estudiantes, afecta el desempeño en la prueba, con diferencias desde no relevantes a grandes en el desempeño de Inglés aunque no se visualiza una tendencia relacionada con este factor. El hacinamiento en los hogares de los estudiantes no genera algún efecto en el puntaje global de las pruebas, si se exceptúa el caso atípico de hacinamiento crítico, que presenta el mayor puntaje. Los niveles educativos tanto del padre como la madre de los estudiantes son factores que se relacionan positivamente con el rendimiento en la prueba, se observa mayores niveles educativos de estos implica mayor rendimiento. Buenas condiciones de TICs y electrodomésticos implicaron un mejor rendimiento en inglés, aunque con diferencias pequeñas o moderadas. De otra parte, tener medio de transporte propio (motocicleta o vehículo) no produce efectos importantes en el rendimiento en esta prueba.

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
Género	Femenino	1138	182,9	27,7	0,19	A
	Masculino	914	188,2	27,6	-	A
Grupos de edad	<=20 años	154	187,0	20,9	0,07	A
	21 años	441	188,1	25,7	0,02	A
	22 años	468	184,5	27,2	0,16	A
	23 años	377	188,7	26,7	-	A
	24 años	227	187,2	25,1	0,06	A
	>= 25 años	385	177,6	33,8	0,36	B
Estrato social	Estrato 0	1	165,0		1,29	D
	Estrato 1	29	167,2	25,2	1,20	D
	Estrato 2	112	173,6	31,6	0,87	D
	Estrato 3	329	175,5	25,1	0,86	D

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Estrato 4	520	183,1	26,9	0,53	C
	Estrato 5	610	189,8	26,5	0,27	B
	Estrato 6	335	196,9	24,7	-	A
	Sin Estrato	5	171,8	28,9	1,01	D
Hogar actual	Es habitual o permanente	1.576	183,7	27,2	-	A
	Es temporal por razones de estudio u otra razón	305	183,2	29,9	0,02	A
Cabeza de familia	Si	77	167,1	35,5	0,63	C
	No	1.804	184,4	27,0	-	A
Personas a cargo	Ninguna	1.740	184,8	27,2	0,03	A
	Una	70	165,6	27,5	0,67	C
	Dos	33	169,9	29,1	0,49	B
	Tres	22	185,5	35,8	-	A
	Cuatro	11	174,8	25,4	0,32	B
	Cinco	4	147,8	18,3	1,11	D
	Seis	1	161,0		0,68	C
Hacinamiento	Hacinamiento crítico	1	211,0		-	A
	Hacinamiento medio	12	192,7	30,4	0,60	C
	Sin hacinamiento	2.039	185,2	27,8	0,93	D
Educación del Padre	Ninguno	12	172,3	25,7	0,84	D
	Primaria incompleta	56	165,6	26,6	1,10	D
	Primaria completa	29	167,6	24,6	1,03	D
	Secundaria incompleta	89	174,5	28,6	0,74	C
	Secundaria completa	248	176,7	26,6	0,66	C
	Técnica o Tecnológica incompleta	53	176,9	20,3	0,67	C
	Técnica o Tecnológica completa	176	178,9	22,1	0,61	C
	Educación Profesional incompleta	115	185,7	30,2	0,31	B
	Educación Profesional completa	655	189,1	27,1	0,19	A
	Postgrado	487	194,0	25,8	-	A
	No Aplica	13	188,1	19,8	0,23	B
	No sabe	31	185,8	28,6	0,31	B
Educación de la Madre	Ninguno	4	158,8	27,1	1,57	D
	Primaria incompleta	28	161,7	24,6	1,47	D
	Primaria completa	27	163,2	27,0	1,31	D
	Secundaria incompleta	96	172,4	29,7	0,86	D

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Secundaria completa	272	174,5	25,6	0,90	D
	Técnica o Tecnológica incompleta	48	180,2	20,2	0,84	D
	Técnica o Tecnológica completa	256	183,1	26,4	0,55	C
	Educación Profesional incompleta	153	187,9	22,2	0,43	B
	Educación Profesional completa	707	190,5	27,8	0,26	B
	Postgrado	360	192,2	25,4	0,21	B
	No Aplica	3	165,3	17,6	1,41	D
	No sabe	10	197,6	23,9	-	A
Índice TICS	Malo	224	180,9	31,3	0,18	A
	Regular	17	174,9	20,7	0,40	B
	Bueno	1.811	185,9	27,3	-	A
Índice Electrodomésticos	Malo	177	180,3	30,8	0,25	B
	Regular	163	170,9	24,1	0,60	C
	Bueno	1.712	187,1	27,3	-	A
Familia tiene motocicleta	Si	364	175,9	29,1	0,46	B
	No	1.498	188,2	26,4	-	A
Familia tiene automóvil	Si	1.570	187,2	26,6	-	A
	No	375	178,4	29,8	0,32	B

Tabla 4.14: Variables Sociodemográficas y desempeño académico en inglés de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro 2017-2018.

4.3.5. Efecto en Competencias Ciudadanas

En la tabla 4.15 se observa que, en la prueba de Competencias Ciudadanas, los hombres presentan desempeño similar que las mujeres, siendo los estudiantes de 22 años o menos quienes mejor rinden en la prueba. No se observan diferencias importantes entre los estratos sociales a excepción del estrato 0, como caso atípico, que presenta el mejor rendimiento. El hecho de que el hogar actual del estudiante sea permanente o temporal no afecta el rendimiento en la prueba, sin embargo el ser cabeza de familia afecta el rendimiento en esta competencia, pues como se observa en la tabla el estudiante que es cabeza de familia tiene más bajo rendimiento que aquel que no lo es, con diferencias pequeñas y el número de personas a cargo de estudiantes no hace que el rendimiento en la competencia sea diferente, excepto para el caso atípico de aquellos que tienen 6 personas a cargo. El hacinamiento tampoco es un factor que afecte el desempeño en esta prueba, excepto en el caso atípico de hacinamiento crítico que presenta mejor rendimiento. Tanto la educación del padre y como el de madre se relacionan positivamente con el rendimiento en la prueba, observándose que, a mayor

nivel educativo de estos, los estudiantes presentan mayor rendimiento en la prueba. Finalmente, los factores relacionados con TICs, electrodomésticos y tener medio de transporte propio (motocicleta o vehículo) tienen efectos irrelevantes en el rendimiento de las Competencias Ciudadanas.

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
Género	Femenino	1.138	157,8	30,2	0,00	A
	Masculino	914	158,0	35,7	-	A
Grupos de edad	<=20 años	154	165,2	29,4	-	A
	21 años	441	163,8	31,6	0,04	A
	22 años	468	160,5	31,5	0,15	A
	23 años	377	158,0	31,7	0,23	B
	24 años	227	151,5	33,2	0,43	B
	>= 25 años	385	148,8	34,7	0,49	B
Estrato social	Estrato 0	1	182,0		-	A
	Estrato 1	29	164,2	26,6	0,67	C
	Estrato 2	112	159,5	30,9	0,73	C
	Estrato 3	329	154,8	32,3	0,84	D
	Estrato 4	520	159,2	32,3	0,70	C
	Estrato 5	610	159,3	32,8	0,69	C
	Estrato 6	335	156,5	32,8	0,78	C
	Sin Estrato	5	167,0	33,0	0,45	C
Hogar actual	Es habitual o permanente	1.576	156,7	32,9	0,15	A
	Es temporal por razones de estudio u otra razón	305	161,5	31,5	-	A
Cabeza de familia	Si	77	149,7	33,8	0,25	B
	No	1.804	157,8	32,6	-	A
Personas a cargo	Ninguna	1.740	158,6	32,5	1,27	D
	Una	70	146,7	30,1	1,77	D
	Dos	33	139,6	36,6	1,65	D
	Tres	22	142,4	34,2	1,68	D
	Cuatro	11	132,8	26,2	2,56	D
	Cinco	4	168,5	7,3	4,30	D
	Seis	1	200,0		-	A
Hacinamiento	Hacinamiento crítico	1	183,0		-	A
	Hacinamiento medio	12	164,4	33,0	0,56	C
	Sin hacinamiento	2.039	157,8	32,7	0,77	C
Educación del Padre	Ninguno	12	152,2	27,4	0,36	B
	Primaria incompleta	56	146,1	28,9	0,62	C
	Primaria completa	29	147,1	31,9	0,52	C

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Secundaria incompleta	89	155,6	32,6	0,32	B
	Secundaria completa	248	151,8	32,3	0,46	B
	Técnica o Tecnológica incompleta	53	153,9	31,0	0,37	B
	Técnica o Tecnológica completa	176	157,6	28,4	0,31	B
	Educación Profesional incompleta	115	161,2	36,3	0,15	A
	Educación Profesional completa	655	158,0	32,4	0,27	B
	Postgrado	487	163,5	32,7	0,11	A
	No Aplica	13	167,0	51,1	-	A
	No sabe	31	158,6	29,3	0,23	B
Educación de la Madre	Ninguno	4	171,0	24,2	-	A
	Primaria incompleta	28	153,0	25,7	0,71	C
	Primaria completa	27	145,1	25,7	1,01	D
	Secundaria incompleta	96	155,0	31,2	0,51	C
	Secundaria completa	272	155,2	31,2	0,51	C
	Técnica o Tecnológica incompleta	48	159,3	33,3	0,36	B
	Técnica o Tecnológica completa	256	154,8	32,3	0,50	B
	Educación Profesional incompleta	153	158,4	35,5	0,36	B
	Educación Profesional completa	707	158,4	32,1	0,39	B
	Postgrado	360	163,4	33,9	0,23	B
	No Aplica	3	130,0	45,8	1,19	D
	No sabe	10	165,1	23,6	0,25	B
Índice TICS	Malo	224	155,7	36,3	0,08	A
	Regular	17	152,0	34,6	0,19	A
	Bueno	1.811	158,2	32,2	-	A
Índice Electrodomésticos	Malo	177	156,3	37,1	0,07	A
	Regular	163	153,7	32,2	0,15	A
	Bueno	1.712	158,5	32,3	-	A
Familia tiene motocicleta	Si	364	154,8	31,2	0,14	A
	No	1.498	159,5	32,5	-	A
Familia tiene automóvil	Si	1.570	158,2	32,2	-	A
	No	375	157,9	33,0	0,01	A

VARIABLES SOCIOECONÓMICAS	N	Media	DT	d Cohen	Grupo

Tabla 4.15: Variables Sociodemográficas y desempeño académico en Competencias Ciudadanas de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro 2017-2018.

4.3.6. Efecto en Puntaje Global

De la tabla 4.16 se puede concluir que en el puntaje Global de las pruebas Saber Pro 2017-2018, los hombres presentan similar desempeño que las mujeres. Por edad se observa mejor rendimiento en los estudiantes con 22 años o menos, con una tendencia inversa, a menor edad mejor rendimiento. Se observan diferencias pequeñas y moderadas entre los estratos sociales sin una tendencia específica. El hecho de que el hogar actual del estudiante sea permanente o temporal no afecta el rendimiento global en la prueba, sin embargo el ser cabeza de familia afecta el puntaje global, pues como se observa en la tabla el estudiante que es cabeza de familia tiene más bajo rendimiento que aquel que no lo es, con diferencias pequeñas; de la misma manera el tener personas a cargo hace que el rendimiento en la competencia sea inferior a los que no las tienen, excepto para el caso atípico de aquellos que tienen 6 personas a cargo. El hacinamiento no es un factor que afecte el desempeño en esta prueba, excepto en el caso atípico de hacinamiento crítico que presenta mejor rendimiento. Los estudiantes que presentan algún tipo de hacinamiento presentan una diferencia pequeña en el puntaje global que aquellos que no lo tienen. Al igual que en las 5 competencias, tanto la educación del padre y como el de madre se relacionan positivamente con el rendimiento en las pruebas, observándose que, a mayor nivel educativo de estos, los estudiantes presentan mayor puntaje global en las pruebas. Finalmente, los factores relacionados con TICs, electrodomésticos y tener medio de transporte propio (motocicleta o vehículo) tienen efectos irrelevantes o pequeños en el puntaje global de las pruebas Saber Pro en el período 2017-2018.

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
Género	Femenino	1.138	167,1	20,1	0,14	A
	Masculino	914	170,1	21,4	-	A
Grupos de edad	<=20 años	154	175,3	18,1	-	A
	21 años	441	172,3	20,1	0,15	A
	22 años	468	170,0	20,3	0,27	B
	23 años	377	168,7	20,1	0,34	B
	24 años	227	166,1	19,5	0,48	B
	>= 25 años	385	160,5	21,7	0,71	C
Estrato social	Estrato 0	1	171,0		-	A
	Estrato 1	29	168,2	22,8	0,12	A
	Estrato 2	112	166,7	18,2	0,24	B
	Estrato 3	329	164,2	20,2	0,34	B
	Estrato 4	520	168,7	20,2	0,11	A

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Estrato 5	610	170,4	20,6	0,03	A
	Estrato 6	335	168,9	23,0	0,09	A
	Sin Estrato	5	162,4	14,5	0,59	C
Hogar actual	Es habitual o permanente	1.576	167,2	20,6	0,18	A
	Es temporal por razones de estudio u otra razón	305	170,9	19,7	-	A
Cabeza de familia	Si	77	158,1	22,4	0,50	B
	No	1804	168,2	20,3	-	A
Personas a cargo	Ninguna	1.740	168,7	20,2	0,12	A
	Una	70	156,6	19,0	0,76	C
	Dos	33	154,3	24,1	0,69	C
	Tres	22	161,0	22,1	0,45	B
	Cuatro	11	152,8	17,9	1,02	D
	Cinco	4	161,5	5,5	1,72	D
	Seis	1	171,0		-	A
Hacinamiento	Hacinamiento crítico	1	192,0		-	A
	Hacinamiento medio	12	174,4	20,9	0,84	D
	Sin hacinamiento	2.039	168,4	20,7	1,14	D
Educación del Padre	Ninguno	12	157,3	17,1	0,76	C
	Primaria incompleta	56	155,8	20,3	0,82	D
	Primaria completa	29	160,0	17,6	0,63	C
	Secundaria incompleta	89	161,5	21,6	0,56	C
	Secundaria completa	248	163,2	19,5	0,50	B
	Técnica o Tecnológica incompleta	53	164,7	16,6	0,42	B
	Técnica o Tecnológica completa	176	166,3	17,1	0,36	B
	Educación Profesional incompleta	115	171,0	21,9	0,12	A
	Educación Profesional completa	655	169,3	20,1	0,21	B
	Postgrado	487	173,7	21,8	-	A
	No Aplica	13	171,0	27,2	0,12	A
	No sabe	31	167,5	21,5	0,28	B
Educación de la Madre	Ninguno	4	155,3	20,0	0,83	D
	Primaria incompleta	28	159,6	15,3	0,64	C
	Primaria completa	27	153,4	15,4	0,94	D
	Secundaria incompleta	96	161,5	21,7	0,55	C
	Secundaria completa	272	163,5	18,8	0,48	B

VARIABLES SOCIOECONÓMICAS		N	Media	DT	d Cohen	Grupo
	Técnica o Tecnológica incompleta	48	166,3	21,2	0,33	B
	Técnica o Tecnológica completa	256	167,2	20,1	0,29	B
	Educación Profesional incompleta	153	168,4	20,7	0,23	B
	Educación Profesional completa	707	170,2	20,5	0,15	A
	Postgrado	360	173,4	21,7	-	A
	No Aplica	3	149,7	25,0	1,09	D
	No sabe	10	170,4	21,9	0,14	A
Índice TICS	Malo	224	169,1	21,3	-	A
	Regular	17	161,8	24,9	0,34	B
	Bueno	1.811	168,4	20,6	0,04	A
Índice Electrodomésticos	Malo	177	170,5	19,7	-	A
	Regular	163	162,5	18,5	0,41	B
	Bueno	1.712	168,8	21,0	0,08	A
Familia tiene motocicleta	Si	364	163,6	19,7	0,31	B
	No	1.498	169,9	20,8	-	A
Familia tiene automóvil	Si	1.570	168,7	20,5	-	A
	No	375	166,4	22,0	0,11	A

Tabla 4.16: Variables Sociodemográficas y desempeño académico en Puntaje Global de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas Saber Pro 2017-2018.

4.4. Modelado

Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables o atributos denominadas variables independientes o predictivas. Cuando esta clase es numérica se utiliza la tarea de regresión y cuando es categórica se utiliza la tarea de clasificación. La clasificación es una técnica predictiva que se aplica a problemas en los que hay que predecir nuevos datos para uno o más ejemplos que van acompañados de una salida denominada clase [19].

Para predecir los factores socioeconómicos, académicos e institucionales asociados al desempeño académico en las competencias genéricas de los estudiantes de la Universidad Javeriana Cali que presentaron las pruebas Saber Pro en los años 2017 y 2018, se seleccionó la tarea de clasificación porque la clase a predecir es de tipo categórica con los valores “Por encima de la media nacional” o “Por debajo de la media nacional”. Además, se escogió la clasificación basado en árboles de deci-

sión, como el modelo más adecuado para resolver el problema objeto de esta investigación, por ser probablemente el modelo más utilizado y popular por su simplicidad y facilidad para entender los resultados [18], [21]. La importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación. La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y sólo una hoja, asignando una única clase a la predicción [25].

Con esta técnica se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes de la Universidad Javeriana Cali, los factores socioeconómicos, académicos e institucionales asociados al buen (por encima de la media) o mal (por debajo de la media) desempeño académico en las pruebas Saber Pro, teniendo en cuenta, como atributo clase, el puntaje global obtenido en las pruebas Saber Pro y los puntajes en las competencias genéricas Lectura Crítica, Comunicación Escrita, Razonamiento Cuantitativo, Inglés y Competencias Ciudadanas.

Se evaluaron diferentes árboles o algoritmos de clasificación con la herramienta Weka, con el fin de seleccionar la técnica de árboles de decisión que mejor clasifique al conjunto de datos de desempeño global *sbpro_jave_final_Gen*. Los resultados se muestran en la tabla 4.17.

Algoritmo	Porcentaje correctamente clasificados
Decision Stump(árbol de decisión de un nivel)	53.02
J48	68,85
LMT (Logistic Model Tree)	62.37
Random Forest	57.89
Random Tree	3.89
RepTree	55.50

Tabla 4.17: Evaluación de diferentes técnicas de árboles de decisión.

De acuerdo con la tabla 4.17, el algoritmo con mayor exactitud fue J48. Por esa razón se escogió el algoritmo J48 para la construcción de los modelos de clasificación con árbol de decisión.

Por otra parte, se seleccionó el método más adecuado para entrenar y probar los modelos de clasificación. Existen diferentes formas de hacerlo [27]:

- Usar el conjunto de datos de entrenamiento (*Use training set*): se emplea todo el conjunto de datos para entrenar el modelo y después se prueba (esta técnica puede ser muy buena para ese conjunto de datos, pero puede ser poco precisa para nuevos datos).
- Proveer un conjunto de datos de prueba (*Supplied test set*): se emplea un conjunto de datos para entrenar y otro conjunto independiente al universo de los datos con los que se está trabajando para prueba (corriendo el riesgo que el conjunto de prueba no refleje o se corresponda con las características de los datos que se emplearon para entrenar el modelo).

- Porcentaje de Partición (*Percentage Split*): se emplea un % aleatorio de datos para entrenar y otro % para probar, este método difiere del anterior en que ambos conjuntos pertenecen al universo de datos con el que se está trabajando por lo que se elimina el riesgo que corre el anterior.

- Validación cruzada (*Cross validation*): Este mecanismo permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición [19]. Para este caso particular se utiliza el método de evaluación validación cruzada con n pliegues (n-fold cross validation). Esta es la opción por defecto y la más comúnmente utilizada. Este método consiste en dividir el conjunto de entrenamiento en n subconjuntos disjuntos de similar tamaño llamados pliegues (*folds*) de forma aleatoria. El número de subconjuntos se puede introducir en el campo *Folds*. Posteriormente se realizan n iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes n-1 (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último, se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones. Otra ventaja de la validación cruzada es que la varianza de los n errores de muestra parciales, permite estimar la variabilidad del método de aprendizaje con respecto al conjunto de datos. Comúnmente, se suelen utilizar 10 particiones (*10-fold cross validation*) [19].

Por las ventajas que ofrece la validación cruzada, se escogió este método para el entrenamiento y prueba de los modelos de clasificación construidos para las diferentes competencias que evalúa las pruebas Saber pro. Se utilizó 10 particiones (10-fold cross validation) teniendo en cuenta lo recomendado por Hernández,Ramírez y Ferri [19].

Una vez seleccionado el algoritmo y el método para el entrenamiento y prueba de los modelos, se procedió a construir los diferentes árboles de decisión con la herramienta WEKA [27] y su algoritmo J48, el cual implementa al algoritmo C.45 [24]. El algoritmo J48 se basa en la utilización del criterio de ganancia de información (information gain). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido [25]. El parámetro más importante que se tuvo en cuenta para la poda fue el factor de confianza C (confidence level), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25 % y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños [26]. Otro parámetro utilizado para variar el tamaño del árbol fue a través del factor M que especifica el mínimo número de instancias o registros por nodo del árbol [27]. Se escogieron como clase los puntajes obtenidos por los estudiantes en cada una de las competencias Saber Pro, las cuales fueron discretizados en los valores *"por encima de la media nacional"*, y *"por debajo de la media nacional"*. En la tabla 4.18 se muestra la distribución de los estudiantes de *sbpro_jave_final*

con respecto a cada uno de los atributos clase que se escogieron para las pruebas Saber Pro.

Clase	Bajo la media	%	Sobre la media	%	Media nacional
Desemp_global	964	47	1088	53	168,44
Desemp_leccritica	916	45	1136	55	170,32
Desemp_comescrita	1074	52	978	48	162,48
Desemp_razcuantitativo	934	46	1118	54	169,32
Desemp_ingles	922	45	1130	55	185,25
Desemp_compciudadanas	890	43	1162	57	157,90

Tabla 4.18: Clases de *sbrpro_jave_final*.

Se generaron diferentes modelos de árboles de decisión por cada una de las competencias genéricas que evalúan las pruebas Saber Pro, con el fin de escoger el árbol de decisión que mejor clasifique a los estudiantes y con mayor nivel de interpretabilidad de los patrones asociados al desempeño académico. Por esta razón, se configuraron dos valores para el factor de confianza C en 25 %, 50 %, combinándolos con dos valores para el factor M en 2.5 % (52 ejemplos) y 5 % (104 ejemplos). Además, se aplicó un proceso de pospoda para dejar las ramas y por ende las reglas más representativas, que son aquellas que sobrepasan un mínimo soporte del 2.5 % y una confianza del 60 %.

4.4.1. Patrones Asociados al Desempeño Global en las Pruebas Saber Pro 2017-2018

Para la construcción de los árboles de decisión para el descubrimiento de los patrones asociados al desempeño general de los estudiantes de la Universidad Javeriana Cali que presentaron las pruebas SaberPro entre los años 2017 y 2018, se utilizó el conjunto de datos *sbpro_jave_final_Gen* descrito en la tabla 3.18. En la tabla 4.19 se muestran los diferentes árboles construidos con su porcentaje de exactitud.

Árbol	C	M	% Exactitud
Gen_c25m52	25	52	59,16
Gen_c25m104	25	104	58,33
Gen_c50m52	50	52	58,33
Gen_c50m104	50	104	57,55

Tabla 4.19: Selección mejor árbol puntaje global del Saber Pro.

El mejor árbol fue construido con los parámetros C=0.25 y M=52 para la prepoda y con soporte mayor o igual al 2.5 % para la postpoda. En la figura 4.1 se muestra el árbol de clasificación obtenido :

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

facultades = Ciencias Economicas y Administrativas
|   estu_grupo_etario = [22]: Bajo la Media (111.0/53.0)
|   estu_grupo_etario = [21]: Sobre la Media (99.0/40.0)
|   estu_grupo_etario = [23]: Bajo la Media (107.0/51.0)
|   estu_grupo_etario = [24]: Bajo la Media (81.0/31.0)
|   estu_grupo_etario = [>=25]: Bajo la Media (217.0/70.0)
facultades = Ingenieria y Ciencias: Sobre la Media (467.0/181.0)
facultades = Humanidades y Ciencias Sociales
|   fami_educacionmadre = Postgrado: Sobre la Media (132.0/51.0)
|   fami_educacionmadre = Educacion profesional incompleta: Sobre la Media (62.0/23.0)
|   fami_educacionmadre = Secundaria (Bachillerato) completa: Bajo la Media (92.0/40.0)
|   fami_educacionmadre = Educacion profesional completa
|   |   fami_numlibros = 26 A 100 LIBROS: Sobre la Media (105.23/51.82)
|   |   fami_numlibros = MAS DE 100 LIBROS: Sobre la Media (74.87/27.58)
|   |   fami_numlibros = 11 A 25 LIBROS: Bajo la Media (52.61/23.2)
|   fami_educacionmadre = Tecnica o tecnologica completa: Bajo la Media (81.0/26.0)
facultades = Ciencias de la Salud: Sobre la Media (194.0/45.0)

Number of Leaves :      21

Size of the tree :  25

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1214               59.1618 %
Incorrectly Classified Instances       838               40.8382 %

Total Number of Instances             2052
```

Figura 4.1: Mejor Árbol para el puntaje global de Saber Pro podado

4.4.2. Patrones Asociados al Desempeño en la Competencia de Lectura Crítica en las Pruebas Saber Pro 2017-2018

Para la construcción de los árboles de decisión para el descubrimiento de los patrones asociados al desempeño en la competencia de Lectura Crítica de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas SaberPro entre los años 2017 y 2018, se utilizó el conjunto de datos *sbpro_jave_final_Lec* descrito en la tabla 3.18. En la tabla 4.20 se muestran los diferentes árboles construidos con su porcentaje de exactitud.

Árbol	C	M	% Exactitud
Lec_c25m52	25	52	59,55
Lec_c25m104	25	104	59,99
Lec_c50m52	50	52	59,79
Lec_c50m104	50	104	58,82

Tabla 4.20: Selección mejor árbol competencia Lectura Crítica del Saber Pro.

El mejor árbol fue construido con los parámetros C=0.25 y M=104 para la prepoda y con un soporte mayor o igual al 5 % para la postpoda. En la figura 4.2 se muestra el árbol de clasificación obtenido.

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

facultades = Ciencias Economicas y Administrativas
|   estu_grupo_etario = [22]: Bajo la Media (111.0/50.0)
|   estu_grupo_etario = [23]: Bajo la Media (107.0/40.0)
|   estu_grupo_etario = [>=25]: Bajo la Media (217.0/81.0)
facultades = Ingeniería y Ciencias
|   indice_transporte = BUENO: Sobre la Media (331.0/135.0)
|   indice_transporte = MALO: Sobre la Media (107.0/31.0)
facultades = Humanidades y Ciencias Sociales: Sobre la Media (741.0/315.0)
facultades = Ciencias de la Salud: Sobre la Media (194.0/45.0)

Number of Leaves  :         11

Size of the tree :  14

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1231               59.9903 %
Incorrectly Classified Instances       821               40.0097 %
```

Figura 4.2: Mejor Árbol para la competencia de Lectura Crítica podado

4.4.3. Patrones Asociados al Desempeño en la Competencia de Comunicación Escrita en las Pruebas Saber Pro 2017-2018

Para la construcción de los árboles de decisión para el descubrimiento de los patrones asociados al desempeño en la competencia de Comunicación Escrita de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas SaberPro entre los años 2017 y 2018, se utilizó

el conjunto de datos *sbpro_jave_final_Esc* descrito en la tabla 3.18. En la tabla 4.21 se muestran los diferentes árboles construidos con su porcentaje de exactitud.

Árbol	C	M	% Exactitud
Esc_c25m52	25	52	55,40
Esc_c25m104	25	104	54,82
Esc_c50m52	50	52	55,65
Esc_c50m104	50	104	54,53

Tabla 4.21: Selección mejor árbol competencia Comunicación Escrita del Saber Pro.

El mejor árbol fue construido con los parámetros C=0.50 y M=52 para la prepoda y con un soporte mayor o igual al 2.5 % para la postpoda. En la figura 4.3 se muestra el árbol de clasificación obtenido.

4.4.4. Patrones Asociados al Desempeño en la Competencia de Razonamiento Cuantitativo en las Pruebas Saber Pro 2017-2018

Para la construcción de los árboles de decisión para el descubrimiento de los patrones asociados al desempeño en la competencia de Razonamiento Cuantitativo de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas SaberPro entre los años 2017 y 2018, se utilizó el conjunto de datos *sbpro_jave_final_Cua* descrito en la tabla 3.18. En la tabla 4.22 se muestran los diferentes árboles construidos con su porcentaje de exactitud.

Árbol	C	M	% Exactitud
Cua_c25m52	25	52	68,56
Cua_c25m104	25	104	68,85
Cua_c50m52	50	52	67,15
Cua_c50m104	50	104	68,95

Tabla 4.22: Selección mejor árbol competencia Razonamiento Cuantitativo del Saber Pro.

El mejor árbol fue construido con los parámetros C=0.25 y M=104 para la prepoda y con un soporte mayor o igual al 2.5 % para la postpoda. En la figura 4.4 se muestra el árbol de clasificación obtenido.

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

facultades = Ciencias Economicas y Administrativas
|  forma_pago_matricula = Recursos Padres
|  |  fami_educacionmadre = Postgrado: Sobre la Media (69.0/32.0)
|  |  fami_educacionmadre = Educacion profesional completa: Bajo la Media (124.54/47.0)
|  forma_pago_matricula = Recursos Credito-Padres: Sobre la Media (98.45/42.3)
facultades = Ingenieria y Ciencias: Bajo la Media (467.0/195.0)
facultades = Humanidades y Ciencias Sociales
|  forma_pago_matricula = Recursos Padres
|  |  fami_educacionmadre = Postgrado: Bajo la Media (67.0/32.0)
|  |  fami_educacionmadre = Educacion profesional completa
|  |  |  cat_estrato = ALTO: Sobre la Media (67.0/28.0)
|  |  |  cat_estrato = MEDIO: Bajo la Media (54.0/26.0)
|  forma_pago_matricula = Recursos Credito-Padres
|  |  estu_dedicacionlecturadiaria = 30 minutos o menos: Sobre la Media (61.0/28.0)
|  |  estu_dedicacionlecturadiaria = Entre 30 y 60 minutos: Bajo la Media (73.0/27.0)
facultades = Ciencias de la Salud
|  fami_educacionpadre = Postgrado: Sobre la Media (72.0/22.0)
|  fami_educacionpadre = Educacien profesional completa: Sobre la Media (56.0/19.0)

Number of Leaves :        65

Size of the tree :  73

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1142               55.653  %
Incorrectly Classified Instances       910               44.347  %
```

Figura 4.3: Mejor Árbol para la competencia de Comunicación Escrita podado

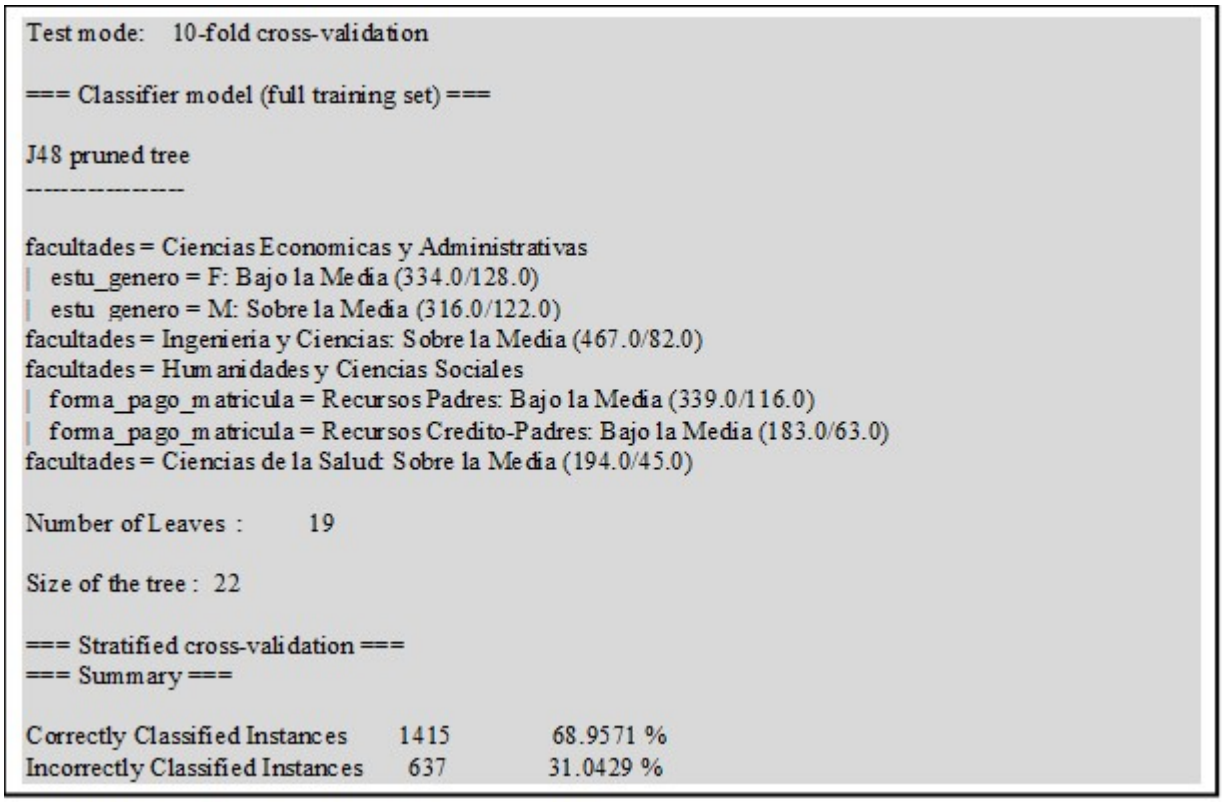


Figura 4.4: Mejor Árbol para la competencia de Razonamiento Cuantitativo podado

4.4.5. Patrones Asociados al Desempeño en la Competencia de Inglés en las Pruebas Saber Pro 2017-2018

Para la construcción de los árboles de decisión para el descubrimiento de los patrones asociados al desempeño en la competencia de Inglés de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas SaberPro entre los años 2017 y 2018, se utilizó el conjunto de datos *sbpro_jave_final_Ing* descrito en la tabla 3.18. En la tabla 4.23 se muestran los diferentes árboles construidos con su porcentaje de exactitud.

Árbol	C	M	% Exactitud
Ing_c25m52	25	52	64,08
Ing_c25m104	25	104	63,25
Ing_c50m52	50	52	62,67
Ing_c50m104	50	104	62,71

Tabla 4.23: Selección mejor árbol competencia Inglés del Saber Pro.

El mejor árbol fue construido con los parámetros C=0.25 y M=52 para la prepoda y con un

soporte mayor o igual al 2.5 % para la postpoda. En la figura 4.5 se muestra el árbol de clasificación obtenido.

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

estu_nse_individual = NSE4
|   indice_electro = BUENO: Sobre la Media (1316.0/470.0)
|   indice_electro = MALO
|   |   estu_genero = F: Bajo la Media (73.0/27.0)
|   |   estu_genero = M: Sobre la Media (61.0/23.0)
|   indice_electro = REGULAR: Bajo la Media (74.0/28.0)
estu_nse_individual = NSE2: Bajo la Media (252.0/76.0)
estu_nse_individual = NSE1: Bajo la Media (57.0/21.0)
estu_nse_individual = NSE3
|   fami_numlibros = 26 A 100 LIBROS: Bajo la Media (81.61/24.24)
|   fami_numlibros = 11 A 25 LIBROS: Bajo la Media (57.85/27.58)

Number of Leaves  :         10

Size of the tree :  14

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1315               64.0838 %
Incorrectly Classified Instances       737               35.9162 %
```

Figura 4.5: Mejor Árbol para la competencia de Inglés podado

4.4.6. Patrones Asociados al Desempeño en la Competencia Ciudadanas en las Pruebas Saber Pro 2017-2018

Para la construcción de los árboles de decisión para el descubrimiento de los patrones asociados al desempeño en la competencia de Competencias Ciudadanas de los estudiantes de la Pontificia Universidad Javeriana Cali que presentaron las pruebas SaberPro entre los años 2017 y 2018, se utilizó el conjunto de datos *sbpro_jave_final_Ciu* descrito en la tabla 3.18. En la tabla 4.24 se muestran los diferentes árboles construidos con su porcentaje de exactitud.

Árbol	C	M	% Exactitud
Ciu_c25m52	25	52	58,91
Ciu_c25m104	25	104	57,30
Ciu_c50m52	50	52	57,60
Ciu_c50m104	50	104	56,23

Tabla 4.24: Selección mejor árbol en Competencias Ciudadanas del Saber Pro.

El mejor árbol fue construido con los parámetros C=0.25 y M=52 para la prepoda y con un soporte mayor o igual al 2.5 % para la postpoda. En la figura 4.6 se muestra el árbol de clasificación obtenido.

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

facultades = Ciencias Economicas y Administrativas
|   estu_grupo_etario = [22]: Bajo la Media (111.0/52.0)
|   estu_grupo_etario = [21]: Sobre la Media (99.0/33.0)
|   estu_grupo_etario = [23]: Bajo la Media (107.0/51.0)
|   estu_grupo_etario = [24]: Bajo la Media (81.0/34.0)
|   estu_grupo_etario = [>=25]: Bajo la Media (217.0/82.0)
facultades = Ingeniería y Ciencias: Sobre la Media (467.0/183.0)
facultades = Humanidades y Ciencias Sociales: Sobre la Media (741.0/313.0)
facultades = Ciencias de la Salud: Sobre la Media (194.0/54.0)

Number of Leaves  :        9

Size of the tree :  11

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1209               58.9181 %
Incorrectly Classified Instances       843               41.0819 %
```

Figura 4.6: Mejor Árbol para la competencia de Competencias Ciudadanas podado

Capítulo 5

Evaluación e Interpretación de Resultados

En esta sección se evalúan los modelos de clasificación basados en árboles de decisión obtenidos por cada competencia y se interpretan los resultados obtenidos en la etapa de modelamiento.

Para evaluar o estimar el coste de los modelos de clasificación construidos, se utilizó la matriz de confusión, también llamada tabla de contingencia. La matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, o sea en términos prácticos permite ver qué tipos de aciertos y errores está teniendo el modelo a la hora de pasar por el proceso de aprendizaje con los datos.

La matriz de confusión representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila i, i =1...n constituyen el número de instancias que realmente pertenecen a la clase i. Similarmente la sumatoria de los ejemplos o registros en cada columna j, j = 1...n son las instancias que ha predicho el algoritmo al valor j de la clase. Los valores en la diagonal son los aciertos y se los conoce, en el caso de que el atributo clase tenga dos valores, como verdaderos positivos (*True positive* TP) y verdaderos negativos (*True negative* TN) (ejemplos que pertenecían a la clase i de la fila i y fueron clasificados correctamente en la clase i) y el resto son los errores de clasificación conocidos como falsos positivos (*False positive* FP) y falsos negativos (*False negative* FN) (ejemplos que pertenecían a la clase i de la fila i y fueron clasificados incorrectamente en otra). En la tabla 5.1 se muestra la representación de la matriz de confusión.

		Prediction	
		Positive	**Negative**
Actual	**Positive**	TP	FN
	Negative	FP	TN

Tabla 5.1: Matriz de Confusión.

Con estos datos se puede calcular varias métricas que van a ayudar a saber que tan bien está funcionando el modelo y en caso de que no esté funcionando muy bien va a ayudar a darse cuenta el porqué.

- **Exactitud (*Accuracy*):** es la proporción entre las predicciones correctas que ha hecho el modelo y el total de predicciones. Su fórmula es 5.1.

$$Ac = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

- **Precisión (*Precision*):** Se refiere a lo cerca que está el resultado de una medición del valor verdadero. Da la calidad de la predicción: ¿qué porcentaje de los que se ha dicho que son la clase positiva, en realidad lo son? Se representa por la proporción entre los verdaderos positivos predichos por el algoritmo y todos los casos positivos predichos (ver la matriz de confusión 5.1 en la columna de Positive Prediction). Se calcula según la ecuación 5.2.

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

- **Sensibilidad o Exhaustividad (*Recall o Sensitivity*):** También se conoce como Tasa de Verdaderos Positivos (*True Positive Rate-TPR*). Informa sobre la cantidad que el modelo de clasificación es capaz de identificar. Da la cantidad: ¿Qué porcentaje de la clase positiva el modelo ha sido capaz de identificar? Es la proporción de positivos reales predichos por el algoritmo entre todos los casos positivos reales. (ver la matriz de confusión 5.1 en la fila de Actual Positive). Se calcula según la ecuación 5.3.

$$TPR = Recall = \frac{TP}{TP + FN} \tag{5.3}$$

  TPR o Recall toma valores en el rango 0 (para el peor clasificador posible) y 1 (para el clasificador ideal).

- **Especificidad (*Specificity*):** También se conoce como Tasa de Verdaderos Negativos (True Negative Rate -TNR). Expresa cuan bien puede el modelo detectar esa clase. Es la proporción de negativos reales predichos por el algoritmo entre todos los casos negativos reales. (ver la matriz de confusión 5.1 en la fila de Actual Negative). Se calcula según la ecuación 5.4.

$$Specificity = \frac{TN}{TN + FP} \tag{5.4}$$

- **Tasa de Falsos Positivos (*False Positive rate-FPR*):** Es la proporción de casos negativos que fueron erróneamente clasificados como positivos por el algoritmo. Se calcula según la ecuación 5.5.

$$FPR = \frac{FP}{FP + TN} \tag{5.5}$$

  También FPR se puede calcular como FPR = 1-especificidad. El FPR toma valores en el rango 0 (para el clasificador ideal) y 1 (para el peor clasificador posible).

- **Tasa de Falsos Negativos (*False negative rate-FNR*):** Es la proporción de casos positivos que fueron erróneamente clasificados como negativos por el algoritmo. Se calcula según la ecuación 5.6.

$$FNR = \frac{FN}{FN + TP} \tag{5.6}$$

  El FPR toma valores en el rango 0 (para el clasificador ideal) y 1 (para el peor clasificador posible).

- **F1 Score (*F-measure*):** Es la media armónica entre la precisión y sensibilidad (*Precision* y *Recall* en una sola métrica). Es de gran utilidad cuando la distribución de las clases es desigual. Se calcula según la ecuación 5.7.

$$F1Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{5.7}$$

- **Matthews correlation coefficient (*MCC*):** El coeficiente de correlación de Matthews se utiliza como una medida de la calidad de las clasificaciones binarias (de dos clases). El coeficiente tiene en cuenta los verdaderos y falsos positivos y verdaderos y falsos negativos. En general, se considera una medida equilibrada que puede utilizarse incluso si las clases son de tamaños muy diferentes. El MCC es, en esencia, un coeficiente de correlación entre las clasificaciones binarias observadas y predichas; devuelve un valor entre -1 y +1. Un coeficiente de +1 representa una predicción perfecta, 0 no es mejor que una predicción aleatoria y -1 indica un desacuerdo total entre la predicción y la observación. Sin embargo, si MCC no es igual a -1, 0 o +1, no es un indicador confiable de cuán similar es un predictor a la conjetura aleatoria porque MCC depende del conjunto de datos [40].

Si bien no existe una manera perfecta de describir la matriz de confusión de verdaderos y falsos positivos y negativos con un solo número, el coeficiente de correlación de Matthews generalmente se considera una de las mejores medidas de este tipo.

Se calcula según la ecuación 5.8.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{5.8}$$

- **Curva o Espacio ROC (*Receiver Operating Characteristic*):** Es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario. También se conoce a esta gráfica como la representación de la razón o proporción de verdaderos positivos (TPR) frente a la razón o proporción de falsos positivos (FPR) según se varía el umbral de discriminación (valor a partir del cual se decide que un caso es un positivo). Un espacio bajo ROC (ROC área o AUROC) se define por FPR y TPR como ejes X e Y respectivamente, y representa los intercambios entre verdaderos positivos (en principio, beneficios) y falsos positivos (en principio, costes) [41].

  Una diagonal divide el espacio ROC. Los puntos por encima de la diagonal representan los buenos resultados de clasificación. En cambio, los puntos por debajo de la línea representan los resultados pobres. La salida de un predictor consistentemente pobre simplemente podría ser invertida para obtener un buen predictor.

- **Precision-Recall Curva (*PRC area*):** Es una representación gráfica de la sensibilidad (recall) frente a la precisión. Un espacio Precision-Recall (PRC área) se define por el Recall en el eje X y la Precisión en el eje Y. Muy útil cuando se trabaja con conjunto de datos muy sesgados, porque da una imagen más informativa del desempeño de un algoritmo.

5.1. Evaluación e Interpretación de Resultados en el Desempeño Global en las Pruebas Saber Pro

Analizando los resultados sobre el desempeño global de los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro 2017-2018, obtenidos en el árbol de decisión que se muestra en la figura 4.1, se puede observar que este clasifica correctamente a 1214 instancias, que corresponde a una exactitud del 59,1 % y 838 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 40.9 %.

Evaluando el modelo con la matriz de confusión, obtenido con la herramienta Weka, de la figura 5.1, este predice correctamente a 747 casos de estudiantes cuyo desempeño global está sobre la media

(TP) y a 467 casos que están bajo la media (TN). Por otra parte, 341 casos cuyo desempeño esta sobre la media, el modelo los clasifica incorrectamente como bajo la media (FN) y 497 casos cuyo desempeño está bajo la media, el modelo los clasifica incorrectamente sobre la media (FP).

Total Number of Instances 2052

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,687	0,516	0,600	0,687	0,641	0,175	0,618	0,629	Sobre la Media
	0,484	0,313	0,578	0,484	0,527	0,175	0,618	0,572	Bajo la Media
Weighted Avg.	0,592	0,421	0,590	0,592	0,587	0,175	0,618	0,602	

=== Confusion Matrix ===

a	b	<-- classified as
747	341	a = Sobre la Media
497	467	b = Bajo la Media

Figura 5.1: Matriz de Confusión desempeño global en Saber Pro.

Para el caso de los estudiantes que están sobre la media en el puntaje global, el modelo tiene una precisión de predicción de 0.60, lo que quiere decir que, del total de casos predichos que están sobre la media, el 60 % son correctos. La sensibilidad (TPR) y Recall del modelo, es de 0.687, lo que indica que el modelo clasifica correctamente al 68.7 % de los estudiantes que realmente están sobre la media. Por otra parte, la tasa de Falsos Positivos (FP Rate) del modelo es de 0.516, lo que significa que el 51.6 % de estudiantes que estaban bajo la media fueron clasificados como sobre la media. El F-measure es de 0.641 lo que significa que la media armónica entre la precisión y el recall de los que están sobre la media es del 64.1 %. En la combinación de estas medidas se aprecia un mejor desempeño del modelo para los que están sobre la media.

Para el caso de los estudiantes que están bajo la media en el puntaje global, el modelo tiene una precisión de predicción de 0.578, lo que quiere decir que del total de casos predichos que están bajo la media, el 57.8 % son correctos. La especificidad (TNR) y Recall del modelo, es de 0.484, lo que indica que el modelo clasifica correctamente al 48.4 % de los estudiantes que realmente están bajo la media. Por otra parte, la tasa de Falsos Negativos (FN Rate) del modelo es de 0.313, lo que significa que el 31.3 % de estudiantes que estaban sobre la media fueron clasificados como bajo la media. El F-measure es de 0.527 lo que significa que la media armónica entre la precisión y el recall de los que están bajo la media es del 52.7 %. En la combinación de estas medidas se aprecia un desempeño moderado del modelo para los que están bajo la media.

De acuerdo a la tabla 4.18, el modelo construido para detectar patrones de rendimiento global en las pruebas Saber Pro de los estudiantes de la Universidad Javeriana Cali no está excesivamen-

te desbalanceado ya que hay una diferencia pequeña de casos entre los que están sobre la media (53 %) y los que están bajo la media (47 %) y es de 124 casos (6 %). Por esta razón, dentro de las métricas de evaluación calculadas anteriormente, se puede decir que el modelo tiene una exactitud del 59.1 % y que predice mejor a los estudiantes que están sobre la media que a los que están bajo la media. Esto también lo muestra en la relación entre el Recall y la Precisión dada en el PRC área, donde para los estudiantes que están sobre la media es de 0,629 y los que están bajo la media es de 0.572. El coeficiente de correlación de Mathews MCC del modelo es de 0.175, lo que indica que hay una relación débil entre lo predicho y lo observado, es decir una baja calidad en la predicción. Finalmente, en cuanto a las áreas, el área ROC del modelo por ser mayor que 0.5 indica que el modelo tiene un desempeño bueno en la clasificación de los estudiantes de la Universidad Javeriana Cali con respecto al puntaje global obtenido en las pruebas Saber Pro 2017-2018.

Con respecto a los patrones de desempeño global en las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018 y que se pueden observar en la figura 4.1, las siguientes reglas son la interpretación de estos patrones. Para escoger los patrones más representativos, se tuvo en cuenta aquellos que superen un soporte mínimo del 2.5 % y una confianza mínima del 60 %:

- **Regla 1:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los 21 años entonces su desempeño global en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 4.8 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 59.6 % de los 99 estudiantes que tienen estas características, son correctamente clasificados y el 5.4 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).
- **Regla 2:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los 24 años entonces su desempeño global en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 3.9 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 61.7 % de los 81 estudiantes que tienen estas características, son correctamente clasificados y el 5.2 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).
- **Regla 3:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los mayores o iguales que 25 años entonces su desempeño global en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 10.6 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 67.7 % de los 217 estudiantes que tienen estas características, son correctamente clasificados y el 15.2 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).
- **Regla 4:** Si el estudiante es de la facultad de Ingeniería y Ciencias entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 22.8 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta

manera. El 61.2 % de los 467 estudiantes que tienen estas características, son correctamente clasificados y el 26.3 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 5:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales y el nivel de educación de la madre es postgrado entonces su desempeño global en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 6.4 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 61.4 % de los 132 estudiantes que tienen estas características, son correctamente clasificados y el 7.4 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 6:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales y el nivel de educación de la madre es profesional incompleta entonces su desempeño global en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 3.0 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 62.9 % de los 62 estudiantes que tienen estas características, son correctamente clasificados y el 3.6 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 7:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales, el nivel de educación de la madre es profesional completa y la familia tiene más de 100 libros entonces su desempeño global en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 3.6 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 63.5 % de los 74 estudiantes que tienen estas características, son correctamente clasificados y el 4.3 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 8:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales y el nivel de educación de la madre es técnica o tecnológica entonces su desempeño global en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 3.9 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 67.9 % de los 81 estudiantes que tienen estas características, son correctamente clasificados y el 5.7 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 9:** Si el estudiante es de la facultad de Ciencias de la Salud entonces su desempeño global en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 9.5 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 76.8 % de los 194 estudiantes que tienen estas características, son correctamente clasificados y el 13.7 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

5.2. Evaluación e Interpretación de Resultados en el Desempeño en la Competencia de Lectura Crítica de las Pruebas Saber Pro

Analizando los resultados sobre el desempeño en la competencia de lectura crítica de los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro 2017-2018, obtenidos en el árbol de decisión que se muestra en la figura 4.2, se puede observar que este clasifica correctamente a 1231 instancias, que corresponde a una exactitud del 60 % y 821 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 40 %.

Evaluando el modelo con la matriz de confusión, obtenido con la herramienta Weka, de la figura 5.2, este predice correctamente a 864 casos de estudiantes cuyo desempeño en lectura crítica está sobre la media (TP) y a 367 casos que están bajo la media (TN). Por otra parte, 272 casos cuyo desempeño esta sobre la media, el modelo los clasifica incorrectamente como bajo la media (FN) y 549 casos cuyo desempeño está bajo la media, el modelo los clasifica incorrectamente sobre la media (FP).

Total Number of Instances 2052

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,761	0,599	0,611	0,761	0,678	0,173	0,592	0,631	Sobre la Media
	0,401	0,239	0,574	0,401	0,472	0,173	0,592	0,516	Bajo la Media
Weighted Avg.	0,600	0,439	0,595	0,600	0,586	0,173	0,592	0,580	

=== Confusion Matrix ===

a	b	<-- classified as
864	272	a = Sobre la Media
549	367	b = Bajo la Media

Figura 5.2: Matriz de Confusión del desempeño en lectura crítica en Saber Pro.

Para el caso de los estudiantes que están sobre la media en la competencia de lectura crítica, el modelo tiene una precisión de predicción de 0.611, lo que quiere decir que, del total de casos predichos que están sobre la media, el 61 % son correctos. La sensibilidad (TPR) y Recall del modelo, es de 0.761, lo que indica que el modelo clasifica correctamente al 76.1 % de los estudiantes que realmente están sobre la media. Por otra parte, la tasa de Falsos Positivos (FP Rate) del modelo es de 0.599, lo que significa que el 59.9 % de estudiantes que estaban bajo la media fueron clasificados como sobre la media. El F-measure es de 0.678 lo que significa que la media armónica entre la precisión y el recall de los que están sobre la media es del 67.8 %. En la combinación de estas medidas se aprecia un mejor desempeño del modelo para los que están sobre la media.

Para el caso de los estudiantes que están bajo la media en la competencia de lectura crítica, el modelo tiene una precisión de predicción de 0.574, lo que quiere decir que del total de casos predichos que están bajo la media, el 57.4 % son correctos. La especificidad (TNR) y Recall del modelo, es de 0.401, lo que indica que el modelo clasifica correctamente al 40.1 % de los estudiantes que realmente están bajo la media. Por otra parte, la tasa de Falsos Negativos (FN Rate) del modelo es de 0.239, lo que significa que el 23.9 % de estudiantes que estaban sobre la media fueron clasificados como bajo la media. El F-measure es de 0.472 lo que significa que la media armónica entre la precisión y el recall de los que están bajo la media es del 47.2 %. En la combinación de estas medidas se aprecia un peor desempeño del modelo para los que están bajo la media.

De acuerdo a la tabla 4.18, el modelo construido para detectar patrones de desempeño en la competencia de lectura crítica en las pruebas Saber Pro de los estudiantes de la Universidad Javeriana Cali no está excesivamente desbalanceado ya que hay una diferencia pequeña de casos entre los que están sobre la media (55 %) y los que están bajo la media (45 %) y es de 220 casos (10 %). Por esta razón, dentro de las métricas de evaluación calculadas anteriormente, se puede decir que el modelo tiene una exactitud del 60 % y que predice mejor a los estudiantes que están sobre la media que a los que están bajo la media. Esto también lo muestra en la relación entre el Recall y la Precisión dada en el PRC área, donde para los estudiantes que están sobre la media es de 0,631 y los que están bajo la media es de 0.516. El coeficiente de correlación de Mathews MCC del modelo es de 0.173, lo que indica que hay una relación débil entre lo predicho y lo observado, es decir una baja calidad en la predicción. Finalmente, en cuanto a las áreas, el área ROC del modelo por ser mayor que 0.5 indica que el modelo tiene un desempeño bueno en la clasificación de los estudiantes de la Universidad Javeriana Cali con respecto al desempeño en la competencia de lectura crítica obtenido en las pruebas Saber Pro 2017-2018.

Con respecto a los patrones de desempeño en la competencia de lectura crítica en las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018 y que se pueden observar en la figura 4.2, las siguientes reglas son la interpretación de estos patrones. Para escoger los patrones más representativos, se tuvo en cuenta aquellos que superen un soporte mínimo del 2.5 % y una confianza mínima del 55 %:

- **Regla 1:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los 22 años entonces su desempeño en la competencia de lectura crítica en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 5.4 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 55 % de los 111 estudiantes que tienen estas características, son correctamente clasificados y el 6.7 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).
- **Regla 2:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los 23 años entonces su desempeño en la competencia de lectura crítica en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 5.2 % del

total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 62.6 % de los 107 estudiantes que tienen estas características, son correctamente clasificados y el 7.3 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 3:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los mayores o iguales que 25 años entonces su desempeño en la competencia de lectura crítica en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 10.6 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 62.7 % de los 217 estudiantes que tienen estas características, son correctamente clasificados y el 14.8 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 4:** Si el estudiante es de la facultad de Ingeniería y Ciencias y el índice de transporte es bueno (cuenta con automóvil o motocicleta) entonces su desempeño en la competencia de lectura crítica en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 16.1 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 59.2 % de los 331 estudiantes que tienen estas características, son correctamente clasificados y el 17.3 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 5:** Si el estudiante es de la facultad de Ingeniería y Ciencias y el índice de transporte es malo (no cuenta ni automóvil ni motocicleta) entonces su desempeño en la competencia de lectura crítica en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 5.2 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 71 % de los 107 estudiantes que tienen estas características, son correctamente clasificados y el 6.7 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 6:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales entonces su desempeño en la competencia de lectura crítica en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 36.1 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 57.5 % de los 741 estudiantes que tienen estas características, son correctamente clasificados y el 37.5 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 7:** Si el estudiante es de la facultad de Ciencias de la Salud entonces su desempeño en la competencia de lectura crítica en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 9.5 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 76.8 % de los 194 estudiantes que tienen estas características, son correctamente clasificados y el 13.1 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

5.3. Evaluación e Interpretación de Resultados en el Desempeño en la Competencia de Comunicación Escrita de las Pruebas Saber Pro

Analizando los resultados sobre el desempeño en la competencia de comunicación escrita de los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro 2017-2018, obtenidos en el árbol de decisión que se muestra en la figura 4.3, se puede observar que este clasifica correctamente a 1142 instancias, que corresponde a una exactitud del 56 % y 910 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 44 %.

Evaluando el modelo con la matriz de confusión, obtenido con la herramienta Weka, de la figura 5.3, este predice correctamente a 689 casos de estudiantes cuyo desempeño en comunicación escrita está bajo la media (TN) y a 453 casos que están sobre la media (TP). Por otra parte, 385 casos cuyo desempeño está bajo la media, el modelo los clasifica incorrectamente como sobre la media (FN) y 525 casos cuyo desempeño está sobre la media, el modelo los clasifica incorrectamente bajo la media (FP).

Total Number of Instances 2052

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,642	0,537	0,568	0,642	0,602	0,106	0,560	0,564	Bajo la Media
	0,463	0,358	0,541	0,463	0,499	0,106	0,560	0,528	Sobre la Media
Weighted Avg.	0,557	0,452	0,555	0,557	0,553	0,106	0,560	0,547	

=== Confusion Matrix ===

a	b	<-- classified as
689	385	a = Bajo la Media
525	453	b = Sobre la Media

Figura 5.3: Matriz de Confusión del desempeño en comunicación escrita en Saber Pro.

Para el caso de los estudiantes que están bajo la media en la competencia de comunicación escrita, el modelo tiene una precisión de predicción de 0.568, lo que quiere decir que del total de casos predichos que están bajo la media, el 56.8 % son correctos. La especificidad (TNR) y Recall del modelo, es de 0.642, lo que indica que el modelo clasifica correctamente al 64.2 % de los estudiantes que realmente están bajo la media. Por otra parte, la tasa de Falsos Negativos (FN Rate) del modelo es de 0.537, lo que significa que el 53.7 % de estudiantes que estaban sobre la media fueron clasificados como bajo la media. El F-measure es de 0.602 lo que significa que la media armónica entre la precisión y el recall de los que están bajo la media es del 60.2 %. En la combinación de estas medidas se aprecia un mejor desempeño del modelo para los que están bajo la media.

De acuerdo a la tabla 4.18, el modelo construido para detectar patrones de desempeño en la competencia de comunicación escrita en las pruebas Saber Pro de los estudiantes de la Universidad Javeriana Cali no está excesivamente desbalanceado ya que hay una diferencia pequeña de casos entre los que están bajo la media (52 %) y los que están sobre la media (48 %) y es de 96 casos (4 %). Por esta razón, dentro de las métricas de evaluación calculadas anteriormente, se puede decir que el modelo tiene una exactitud del 56 % y que predice mejor a los estudiantes que están bajo la media que a los que están sobre la media. Esto también lo muestra en la relación entre el Recall y la Precisión dada en el PRC área, donde para los estudiantes que están bajo la media es de 0,564 y los que están sobre la media es de 0.528. El coeficiente de correlación de Mathews MCC del modelo es de 0.106, lo que indica que hay una relación débil entre lo predicho y lo observado, es decir una baja calidad en la predicción. Finalmente, en cuanto a las áreas, el área ROC del modelo por ser mayor que 0.5 indica que el modelo tiene un desempeño bueno en la clasificación de los estudiantes de la Universidad Javeriana Cali con respecto al desempeño en la competencia de comunicación escrita obtenido en las pruebas Saber Pro 2017-2018.

Con respecto a los patrones de desempeño en la competencia de comunicación escrita en las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018 y que se pueden observar en la figura 4.3, las siguientes reglas son la interpretación de estos patrones. Para escoger los patrones más representativos, se tuvo en cuenta aquellos que superen un soporte mínimo del 2.5 % y una confianza mínima del 55 %:

- **Regla 1:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas, el pago de la matrícula se hace con recursos de los padres y el nivel de educación de la madre es profesional completa entonces su desempeño en la competencia de comunicación escrita en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 6 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 62.1 % de los 124 estudiantes que tienen estas características, son correctamente clasificados y el 7.2 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 2:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y el pago de la matrícula se hace con recursos tanto de los padres como de crédito entonces su desempeño en la competencia de comunicación escrita en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 4.8 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 57.1 % de los 98 estudiantes que tienen estas características, son correctamente clasificados y el 5.7 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 3:** Si el estudiante es de la facultad de Ingeniería y Ciencias y el índice de transporte es bueno (cuenta con automóvil o motocicleta) entonces su desempeño en la competencia de comunicación escrita en las pruebas Saber Pro tiene mayor probabilidad de estar bajo

la media. El 22.8 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 58.2 % de los 467 estudiantes que tienen estas características, son correctamente clasificados y el 25.3 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 4:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales, el pago de la matrícula se hace con recursos de los padres, el nivel de educación de la madre es profesional completa y su estrato es alto entonces su desempeño en la competencia de comunicación escrita en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 3.3 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 58.2 % de los 67 estudiantes que tienen estas características, son correctamente clasificados y el 4 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 5:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales, el pago de la matrícula se hace con recursos tanto de los padres como de crédito y su dedicación diaria a la lectura es de 30 a 60 minutos entonces su desempeño en la competencia de comunicación escrita en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 3.6 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 63 % de los 73 estudiantes que tienen estas características, son correctamente clasificados y el 4.3 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

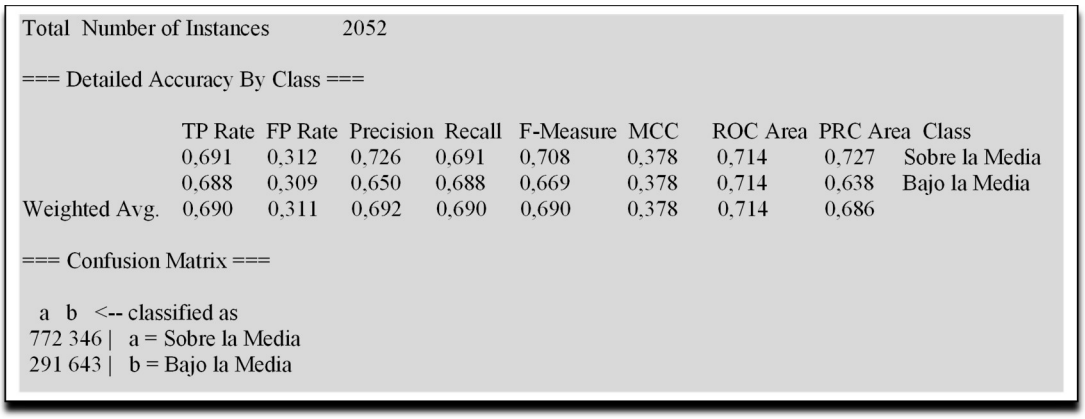
- **Regla 6:** Si el estudiante es de la facultad de Ciencias de la Salud y el nivel de educación del padre es postgrado entonces su desempeño en la competencia de comunicación escrita en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 3.5 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 69.4 % de los 72 estudiantes que tienen estas características, son correctamente clasificados y el 5.1 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 7:** Si el estudiante es de la facultad de Ciencias de la Salud y el nivel de educación del padre es profesional completo entonces su desempeño en la competencia de comunicación escrita en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 2.7 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 66.1 % de los 56 estudiantes que tienen estas características, son correctamente clasificados y el 3.8 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

5.4. Evaluación e Interpretación de Resultados en el Desempeño en la Competencia de Razonamiento Cuantitativo de las Pruebas Saber Pro

Analizando los resultados sobre el desempeño en la competencia de razonamiento cuantitativo de los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro 2017-2018, obtenidos en el árbol de decisión que se muestra en la figura 5, se puede observar que este clasifica correctamente a 1415 instancias, que corresponde a una exactitud del 69 % y 637 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 31 %.

Evaluando el modelo con la matriz de confusión, obtenido con la herramienta Weka, de la figura 5.4, este predice correctamente a 772 casos de estudiantes cuyo desempeño en razonamiento cuantitativo está sobre la media (TP) y a 643 casos que están bajo la media (TN). Por otra parte, 346 casos cuyo desempeño esta sobre la media, el modelo los clasifica incorrectamente como bajo la media (FN) y 291 casos cuyo desempeño está bajo la media, el modelo los clasifica incorrectamente sobre la media (FP).



Total Number of Instances 2052

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,691	0,312	0,726	0,691	0,708	0,378	0,714	0,727	Sobre la Media
	0,688	0,309	0,650	0,688	0,669	0,378	0,714	0,638	Bajo la Media
Weighted Avg.	0,690	0,311	0,692	0,690	0,690	0,378	0,714	0,686	

=== Confusion Matrix ===

a	b	<-- classified as
772	346	a = Sobre la Media
291	643	b = Bajo la Media

Figura 5.4: Matriz de Confusión del desempeño en razonamiento cuantitativo en Saber Pro.

Para el caso de los estudiantes que están sobre la media en la competencia de razonamiento cuantitativo, el modelo tiene una precisión de predicción de 0.726, lo que quiere decir que, del total de casos predichos que están sobre la media, el 72.6 % son correctos. La sensibilidad (TPR) y Recall del modelo, es de 0.691, lo que indica que el modelo clasifica correctamente al 69.1 % de los estudiantes que realmente están sobre la media. Por otra parte, la tasa de Falsos Positivos (FP Rate) del modelo es de 0.312, lo que significa que el 31.2 % de estudiantes que estaban bajo la media fueron clasificados como sobre la media. El F-measure es de 0.708 lo que significa que la media armónica entre la precisión y el recall de los que están sobre la media es del 70.8 %. En la combinación de estas medidas se aprecia un mejor desempeño del modelo para los que están sobre la media.

Para el caso de los estudiantes que están bajo la media en la competencia de razonamiento cuantitativo, el modelo tiene una precisión de predicción de 0.650, lo que quiere decir que del total de casos predichos que están bajo la media, el 65 % son correctos. La especificidad (TNR) y Recall del modelo, es de 0.688, lo que indica que el modelo clasifica correctamente al 68.8 % de los estudiantes que realmente están bajo la media. Por otra parte, la tasa de Falsos Negativos (FN Rate) del modelo es de 0.309, lo que significa que el 30.9 % de estudiantes que estaban sobre la media fueron clasificados como bajo la media. El F-measure es de 0.669 lo que significa que la media armónica entre la precisión y el recall de los que están bajo la media es del 66.9 %. En la combinación de estas medidas se aprecia también un buen desempeño del modelo para los que están bajo la media.

De acuerdo a la tabla 4.18, el modelo construido para detectar patrones de desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro de los estudiantes de la Universidad Javeriana Cali no está excesivamente desbalanceado ya que hay una diferencia pequeña de casos entre los que están sobre la media (54.5 %) y los que están bajo la media (45.5 %) y es de 184 casos (9 %). Por esta razón, dentro de las métricas de evaluación calculadas anteriormente, se puede decir que el modelo tiene una exactitud del 69 % y que predice por igual a los estudiantes que están sobre la media y los que están bajo la media, si tomamos en cuenta el Recall. Si tomamos en cuenta la relación entre el Recall y la Precisión dada en el PRC área, cuyo valor para los estudiantes que están sobre la media es de 0,727 y los que están bajo la media es de 0.638, el modelo tiene un mejor desempeño para los estudiantes que están sobre la media. El coeficiente de correlación de Mathews MCC del modelo es de 0.378, lo que indica que hay una correlación débil entre lo predicho y lo observado, es decir una baja calidad en la predicción. Finalmente, en cuanto a las áreas, el área ROC del modelo es de 0.714 y por ser mayor que 0.5 indica que el modelo tiene un desempeño bueno en la clasificación de los estudiantes de la Universidad Javeriana Cali con respecto al desempeño en la competencia de razonamiento cuantitativo obtenido en las pruebas Saber Pro 2017-2018.

Con respecto a los patrones de desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018 y que se pueden observar en la figura 4.4, las siguientes reglas son la interpretación de estos patrones. Para escoger los patrones más representativos, se tuvo en cuenta aquellos que superen un soporte mínimo del 5 % y una confianza mínima del 60 %:

- **Regla 1:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y es de sexo femenino entonces su desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 16.3 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 61.7 % de los 334 estudiantes que tienen estas características, son correctamente clasificados y el 22.1 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).
- **Regla 2:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y es

de sexo masculino entonces su desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 15.4 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 61.4 % de los 316 estudiantes que tienen estas características, son correctamente clasificados y el 17.4 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 3:** Si el estudiante es de la facultad de Ingeniería y Ciencias entonces su desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 22.8 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 82.4 % de los 467 estudiantes que tienen estas características, son correctamente clasificados y el 34.4 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 4:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales y el pago de matrícula se hace con recursos de los padres entonces su desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 16.5 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 65.8 % de los 339 estudiantes que tienen estas características, son correctamente clasificados y el 23.9 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 5:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales y el pago de matrícula se hace con recursos tanto de los padres como de crédito entonces su desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 8.9 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 65.6 % de los 183 estudiantes que tienen estas características, son correctamente clasificados y el 12.8 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 6:** Si el estudiante es de la facultad de Ciencias de la Salud entonces su desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 9.5 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 76.8 % de los 194 estudiantes que tienen estas características, son correctamente clasificados y el 13.3 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

5.5. Evaluación e Interpretación de Resultados en el Desempeño en la Competencia de Inglés de las Pruebas Saber Pro

Analizando los resultados sobre el desempeño en la competencia de inglés de los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro 2017-2018, obtenidos en el árbol de decisión

que se muestra en la figura 4.5, se puede observar que este clasifica correctamente a 1315 instancias, que corresponde a una exactitud del 64 % y 737 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 36 %.

Evaluando el modelo con la matriz de confusión, obtenido con la herramienta Weka, de la figura 5.5, este predice correctamente a 898 casos de estudiantes cuyo desempeño en inglés está sobre la media (TP) y a 417 casos que están bajo la media (TN). Por otra parte, 232 casos cuyo desempeño esta sobre la media, el modelo los clasifica incorrectamente como bajo la media (FN) y 505 casos cuyo desempeño está bajo la media, el modelo los clasifica incorrectamente sobre la media (FP).

Total Number of Instances 2052

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,795	0,548	0,640	0,795	0,709	0,264	0,617	0,619	Sobre la Media
	0,452	0,205	0,643	0,452	0,531	0,264	0,617	0,563	Bajo la Media
Weighted Avg.	0,641	0,394	0,641	0,641	0,629	0,264	0,617	0,594	

=== Confusion Matrix ===

a	b	<-- classified as
898	232	a = Sobre la Media
505	417	b = Bajo la Media

Figura 5.5: Matriz de Confusión del desempeño en inglés en Saber Pro.

Para el caso de los estudiantes que están sobre la media en la competencia de inglés, el modelo tiene una precisión de predicción de 0.640, lo que quiere decir que, del total de casos predichos que están sobre la media, el 64 % son correctos. La sensibilidad (TPR) y Recall del modelo, es de 0.795, lo que indica que el modelo clasifica correctamente al 79.5 % de los estudiantes que realmente están sobre la media. Por otra parte, la tasa de Falsos Positivos (FP Rate) del modelo es de 0.548, lo que significa que el 54.8 % de estudiantes que estaban bajo la media fueron clasificados como sobre la media. El F-measure es de 0.709 lo que significa que la media armónica entre la precisión y el recall de los que están sobre la media es del 70.9 %. En la combinación de estas medidas se aprecia un mejor desempeño del modelo para los que están sobre la media.

Para el caso de los estudiantes que están bajo la media en la competencia de inglés, el modelo tiene una precisión de predicción de 0.643, lo que quiere decir que del total de casos predichos que están bajo la media, el 64.3 % son correctos. La especificidad (TNR) y Recall del modelo, es de 0.452, lo que indica que el modelo clasifica correctamente al 45.2 % de los estudiantes que realmente están bajo la media. Por otra parte, la tasa de Falsos Negativos (FN Rate) del modelo es de 0.205, lo que significa que el 20.5 % de estudiantes que estaban sobre la media fueron clasificados como

bajo la media. El F-measure es de 0.531 lo que significa que la media armónica entre la precisión y el recall de los que están bajo la media es del 53.1 %. En la combinación de estas medidas se aprecia un peor desempeño del modelo para los que están bajo la media.

De acuerdo a la tabla 4.18, el modelo construido para detectar patrones de desempeño en la competencia de inglés en las pruebas Saber Pro de los estudiantes de la Universidad Javeriana Cali no está excesivamente desbalanceado ya que hay una diferencia pequeña de casos entre los que están sobre la media (55 %) y los que están bajo la media (45 %) y es de 208 casos (10 %). Por esta razón, dentro de las métricas de evaluación calculadas anteriormente, se puede decir que el modelo tiene una exactitud del 64 % y que predice mejor a los estudiantes que están sobre la media que a los que están bajo la media, si tomamos en cuenta el Recall. Si tomamos en cuenta la relación entre el Recall y la Precisión dada en el PRC área, cuyo valor para los estudiantes que están sobre la media es de 0,619 y los que están bajo la media es de 0.563, el modelo tiene un mejor desempeño para los estudiantes que están sobre la media. El coeficiente de correlación de Mathews MCC del modelo es de 0.264, lo que indica que hay una correlación débil entre lo predicho y lo observado, es decir una baja calidad en la predicción. Finalmente, en cuanto a las áreas, el área ROC del modelo es de 0.617 y por ser mayor que 0.5 indica que el modelo tiene un desempeño bueno en la clasificación de los estudiantes de la Universidad Javeriana Cali con respecto al desempeño en la competencia de inglés obtenido en las pruebas Saber Pro 2017-2018.

Con respecto a los patrones de desempeño en la competencia de inglés en las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018 y que se pueden observar en la figura 4.5, las siguientes reglas son la interpretación de estos patrones. Para escoger los patrones más representativos, se tuvo en cuenta aquellos que superen un soporte mínimo del 2.5 % y una confianza mínima del 60 %:

- **Regla 1:** Si el nivel socioeconómico del estudiante pertenece al grupo NSE4 (según los descriptores del nivel socioeconómico del ICFES: *"Los estudiantes que presentan Saber PRO pertenecientes a este nivel se caracterizan por la posesión de electrodomésticos como el horno microondas o a gas y la lavadora, así como por su acceso a internet y la tenencia de computador y automóvil particular. En cuanto a la educación de los padres se observan niveles completos de educación superior profesional. No obstante, en el caso del nivel educativo de la madre se encuentran niveles incompletos en educación profesional. Se observa un espectro amplio de ocupaciones tanto de padres como de madres, tales como dueño de negocios pequeños, operario de maquinaria, conductor de vehículos, vendedor, pensionado, trabajo auxiliar administrativo, así como trabajo profesional. En este NSE es característico un valor del semestre mayor a 3 millones")* y el índice de electrodomésticos es bueno entonces su desempeño en la competencia de inglés en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 64.1 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 64.3 % de los 1316 estudiantes que tienen estas características, son correctamente clasificados y el 75.7 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 2:** Si el nivel socioeconómico del estudiante pertenece al grupo NSE4, el índice de electrodomésticos es malo y es de sexo femenino entonces su desempeño en la competencia de inglés en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 3.6 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 63 % de los 73 estudiantes que tienen estas características, son correctamente clasificados y el 4.9 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 3:** Si el nivel socioeconómico del estudiante pertenece al grupo NSE4, el índice de electrodomésticos es malo y es de sexo masculino entonces su desempeño en la competencia de inglés en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 3 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 62.3 % de los 61 estudiantes que tienen estas características, son correctamente clasificados y el 3.4 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

- **Regla 4:** Si el nivel socioeconómico del estudiante pertenece al grupo NSE4 y el índice de electrodomésticos es regular entonces su desempeño en la competencia de inglés en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 3.6 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 62.2 % de los 74 estudiantes que tienen estas características, son correctamente clasificados y el 4.9 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 5:** Si el nivel socioeconómico del estudiante pertenece al grupo NSE2 *("Los estudiantes que presentan Saber PRO pertenecientes a este nivel muestran características como la tenencia del servicio de internet y de computador, así como la posesión de lavadora y servicio de televisión. En los hogares de estos estudiantes es característico no tener horno microondas o a gas. A su vez, se observan distintos niveles educativos característicos: en el caso de los padres va desde la primaria completa hasta la secundaria (bachillerato) completa y en el caso de las madres, principalmente corresponde a secundaria (bachillerato) completa. Cabe señalar que, para ambos casos, se observa en menor medida la secundaria incompleta. En cuanto a la ocupación de los padres, se destacan labores como la agricultura, la pesca o la jornalería; trabajar como personal de limpieza, mantenimiento o seguridad; tener un negocio pequeño y trabajar por cuenta propia, entre otros. En el NSE 2 y NSE 3 no se identifica una tendencia clara en el pago del valor del semestre, el cual se ubica en los distintos rangos")* entonces su desempeño en la competencia de inglés en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 12.3 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 69.8 % de los 252 estudiantes que tienen estas características, son correctamente clasificados y el 18.8 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 6:** Si el nivel socioeconómico del estudiante pertenece al grupo NSE1 *("Los estudiantes que presentan Saber PRO pertenecientes a este nivel se caracterizan por la ausencia en sus*

hogares de un horno microondas o de gas. Se observa una carencia predominante del servicio de internet, sin embargo, algunos estudiantes en este nivel tienen acceso a dicho servicio. Otra característica importante en este NSE es la ausencia generalizada de elementos como la consola de videojuegos y el automóvil particular. En cuanto al nivel educativo de los padres, la primaria es característica de este nivel, tanto para el padre como para la madre. Adicionalmente, es característico compartir baño con más de 3 personas. Un rasgo relevante de este NSE es el valor del semestre, el cual corresponde a no pago del semestre") entonces su desempeño en la competencia de inglés en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 2.8 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 63.2 % de los 57 estudiantes que tienen estas características, son correctamente clasificados y el 3.9 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 6:** Si el nivel socioeconómico del estudiante pertenece al grupo NSE3 *("Los estudiantes que presentan Saber PRO pertenecientes a este nivel se caracterizan por el acceso al servicio de internet, seguido por la posesión de un horno microondas o a gas. Para este nivel es relevante la caracterización del nivel educativo de los padres. Por parte de los padres, los niveles educativos van desde secundaria completa hasta técnico o tecnológico completo; en el caso de las madres, se observan los mismos niveles que para el padre e incluso nivel incompleto de educación profesional. Adicionalmente, es característico la tenencia de lavadora, computador y servicio de televisión. Respecto a la ocupación de los padres, se observan labores como ser dueño de un negocio, conductor de vehículos u operario de maquinaria y pensionados. En el NSE 2 y NSE 3 no se identifica una tendencia clara en el pago del valor del semestre, el cual se ubica en los distintos rangos")* y el número de libros que posee la familia está entre 26 y 100 libros entonces su desempeño en la competencia de inglés en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 3.9 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 70.4 % de los 81 estudiantes que tienen estas características, son correctamente clasificados y el 6.1 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

5.6. Evaluación e Interpretación de Resultados en el Desempeño en Competencias Ciudadanas de las Pruebas Saber Pro

Analizando los resultados sobre el desempeño en competencias ciudadanas de los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro 2017-2018, obtenidos en el árbol de decisión que se muestra en la figura 4.6, se puede observar que este clasifica correctamente a 1209 instancias, que corresponde a una exactitud del 59 % y 843 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 41 %.

Evaluando el modelo con la matriz de confusión, obtenido con la herramienta Weka, de la figura 5.6, este predice correctamente a 941 casos de estudiantes cuyo desempeño en competencias

ciudadanas está sobre la media (TP) y a 268 casos que están bajo la media (TN). Por otra parte, 221 casos cuyo desempeño esta sobre la media, el modelo los clasifica incorrectamente como bajo la media (FN) y 622 casos cuyo desempeño está bajo la media, el modelo los clasifica incorrectamente sobre la media (FP).

Total Number of Instances 2052

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,810	0,699	0,602	0,810	0,691	0,129	0,582	0,622	Sobre la Media
	0,301	0,190	0,548	0,301	0,389	0,129	0,582	0,500	Bajo la Media
Weighted Avg.	0,589	0,478	0,579	0,589	0,560	0,129	0,582	0,569	

=== Confusion Matrix ===

a	b	<-- classified as
941	221	a = Sobre la Media
622	268	b = Bajo la Media

Figura 5.6: Matriz de Confusión del desempeño en competencia ciudadana en Saber Pro.

Para el caso de los estudiantes que están sobre la media en competencias ciudadanas, el modelo tiene una precisión de predicción de 0.602, lo que quiere decir que, del total de casos predichos que están sobre la media, el 60.2 % son correctos. La sensibilidad (TPR) y Recall del modelo, es de 0.810, lo que indica que el modelo clasifica correctamente al 81 % de los estudiantes que realmente están sobre la media. Por otra parte, la tasa de Falsos Positivos (FP Rate) del modelo es de 0.699, lo que significa que el 69.9 % de estudiantes que estaban bajo la media fueron clasificados como sobre la media. El F-measure es de 0.691 lo que significa que la media armónica entre la precisión y el recall de los que están sobre la media es del 69.1 %. En la combinación de estas medidas se aprecia un mejor desempeño del modelo para los que están sobre la media.

Para el caso de los estudiantes que están bajo la media en competencias ciudadanas, el modelo tiene una precisión de predicción de 0.548 lo que quiere decir que del total de casos predichos que están bajo la media, el 54.8 % son correctos. La especificidad (TNR) y Recall del modelo, es de 0.301, lo que indica que el modelo clasifica correctamente al 30.1 % de los estudiantes que realmente están bajo la media. Por otra parte, la tasa de Falsos Negativos (FN Rate) del modelo es de 0.190, lo que significa que el 19 % de estudiantes que estaban sobre la media fueron clasificados como bajo la media. El F-measure es de 0.389 lo que significa que la media armónica entre la precisión y el recall de los que están bajo la media es del 38.9 %. En la combinación de estas medidas se aprecia un peor desempeño del modelo para los que están bajo la media.

De acuerdo a la tabla 4.18, el modelo construido para detectar patrones de desempeño en competencias ciudadanas en las pruebas Saber Pro de los estudiantes de la Universidad Javeriana Cali

no está excesivamente desbalanceado ya que hay una diferencia pequeña de casos entre los que están sobre la media (56.6 %) y los que están bajo la media (43.4 %) y es de 272 casos (13 %). Por esta razón, dentro de las métricas de evaluación calculadas anteriormente, se puede decir que el modelo tiene una exactitud del 59 % y que predice mejor a los estudiantes que están sobre la media que a los que están bajo la media, si tomamos en cuenta el Recall. Si tomamos en cuenta la relación entre el Recall y la Precisión dada en el PRC área, cuyo valor para los estudiantes que están sobre la media es de 0,622 y los que están bajo la media es de 0.5, el modelo también tiene un mejor desempeño para los estudiantes que están sobre la media. El coeficiente de correlación de Mathews MCC del modelo es de 0.129, lo que indica que hay una correlación débil entre lo predicho y lo observado, es decir una baja calidad en la predicción. Finalmente, en cuanto a las áreas, el área ROC del modelo es de 0.582 y por ser mayor que 0.5 indica que el modelo tiene un desempeño bueno en la clasificación de los estudiantes de la Universidad Javeriana Cali con respecto al desempeño en competencias ciudadanas obtenido en las pruebas Saber Pro 2017-2018.

Con respecto a los patrones de desempeño en competencias ciudadanas en las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018 y que se pueden observar en la figura 4.6, las siguientes reglas son la interpretación de estos patrones. Para escoger los patrones más representativos, se tuvo en cuenta aquellos que superen un soporte mínimo del 2.5 % y una confianza mínima del 53 %:

- **Regla 1:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y el estudiante pertenece al grupo etario de 22 años entonces su desempeño en competencias ciudadanas en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 5.4 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 53.2 % de los 111 estudiantes que tienen estas características, son correctamente clasificados y el 6.6 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).
- **Regla 2:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y el estudiante pertenece al grupo etario de 21 años entonces su desempeño en competencias ciudadanas en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 4.8 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 66.7 % de los 99 estudiantes que tienen estas características, son correctamente clasificados y el 5.7 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).
- **Regla 3:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y el estudiante pertenece al grupo etario de 24 años entonces su desempeño en competencias ciudadanas en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 3.9 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 58 % de los 81 estudiantes que tienen estas características, son correctamente clasificados y el 5.3 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).

- **Regla 4:** Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y el estudiante pertenece al grupo etario de mayores o iguales que 25 años entonces su desempeño en competencias ciudadanas en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 10.6 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 62.2 % de los 217 estudiantes que tienen estas características, son correctamente clasificados y el 15.2 % del total de estudiantes observados que están bajo la media cumplen este patrón (ver tabla 4.18).
- **Regla 5:** Si el estudiante es de la facultad de Ingeniería y Ciencias entonces su desempeño en competencias ciudadanas en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 22.8 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 60.8 % de los 467 estudiantes que tienen estas características, son correctamente clasificados y el 24.4 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).
- **Regla 6:** Si el estudiante es de la facultad de Humanidades y Ciencias Sociales entonces su desempeño en competencias ciudadanas en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 36.1 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 57.8 % de los 741 estudiantes que tienen estas características, son correctamente clasificados y el 36.8 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).
- **Regla 7:** Si el estudiante es de la facultad de Ciencias de la Salud entonces su desempeño en competencias ciudadanas en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 9.5 % del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 72.2 % de los 194 estudiantes que tienen estas características, son correctamente clasificados y el 12 % del total de estudiantes observados que están sobre la media cumplen este patrón (ver tabla 4.18).

Capítulo 6

Conclusiones

En esta investigación se detectaron patrones asociados al desempeño académico en las competencias genéricas de las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018. Para la construcción de los modelos predictivos, se escogió la tarea de aprendizaje supervisado clasificación, la técnica basada en árboles de decisión y la metodología CRISP-DM.

Se realizó un análisis exploratorio de datos estableciendo el efecto "programa" que analiza el desempeño de los diferentes programas de la institución en cada una de las competencias genéricas que evaluaron en las pruebas Saber Pro en los años 2017 y 2018. Se destacaron los programas de Filosofía, Matemáticas Aplicadas y Medicina por tener el mejor desempeño en la competencia de lectura crítica. En la competencia de comunicación escrita el mejor desempeño fue para el programa de Filosofía, aunque por el mayor número de estudiantes le acompaña Derecho. En la competencia de razonamiento cuantitativo el mejor desempeño fue para el programa de Matemáticas Aplicadas, aunque por el mayor número de estudiantes estaría también Ingeniería Electrónica. En la competencia de inglés los mejores desempeños fueron para Ingeniería de Sistemas y Computación y Negocios Internacionales. Para competencias ciudadanas, los mejores desempeños fueron para los programas de Ciencia Política y Derecho. En general, tomando como referencia el puntaje global obtenido en las pruebas Saber Pro, los mejores desempeños fueron para los programas de Filosofía, Matemáticas Aplicadas, Ingeniería de Sistemas y Computación y Medicina, resaltando estos dos últimos programas por el mayor número de estudiantes que presentaron las pruebas.

En cuanto a la calidad de los modelos construidos y teniendo en cuenta la métrica F-measure (F1-score) que mide la relación entre lo predicho y lo observado (la precisión y el recall), los mejores modelos corresponden a las competencias de inglés (0.709) y razonamiento cuantitativo (0.708) y el peor modelo, a la competencia de comunicación escrita (0.602).

Para el descubrimiento de patrones en cada competencia se tuvo en cuenta aquellos que sobrepasaron un soporte mínimo del total de casos evaluados y una confianza mínima con respecto al número de casos que cumplían el patrón descubierto. El resto de los patrones fueron descartados.

Entre las variables predictoras asociadas a los patrones descubiertos en la competencia de lectura crítica, están la facultad, el grupo etario y el índice de transporte del estudiante, como tres variables importantes asociadas al buen o bajo desempeño académico de los estudiantes de la Universidad

Javeriana Cali en esta competencia.

Entre las variables predictoras asociadas al buen o bajo desempeño académico de los estudiantes de la Universidad Javeriana Cali en la competencia de comunicación escrita están la facultad, la forma de pago de la matrícula, la educación de la madre, el estrato y el tiempo diario que el estudiante dedica a la lectura. Entre las variables predictoras asociadas al buen o bajo desempeño académico de los estudiantes de la Universidad Javeriana Cali en la competencia de razonamiento cuantitativo están la facultad, el sexo y la forma de pago de la matrícula.

Entre las variables predictoras asociadas al buen o bajo desempeño académico de los estudiantes de la institución en la competencia de inglés están el nivel socioeconómico al cual pertenece el estudiante, el índice de electrodomésticos del hogar, el sexo y el número de libros que están a disponibilidad del estudiante en la casa.

Entre las variables predictoras asociadas al buen o bajo desempeño académico de los estudiantes de la institución en competencias ciudadanas están la facultad y el grupo etario al cual pertenece el estudiante.

Finalmente, entre las variables predictoras de los patrones descubiertos asociados al puntaje global obtenido por los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro 2017-2018, están la facultad, el grupo etario, la educación de la madre y el número de libros disponibles en la casa del estudiante, como cuatro variables importantes asociadas al buen o bajo desempeño académico de los estudiantes de la Universidad Javeriana Cali.

Cabe destacar a los estudiantes de la facultad de Ciencias de la Salud de la Universidad Javeriana Cali, cuyo desempeño académico en las pruebas Saber Pro en todas las competencias esta sobre la media.

Se plantea como trabajos futuros complementar este estudio utilizando otras técnicas de minería de datos como reglas de asociación y agrupamiento que permitan relacionar que atributos se presentan juntos asociados al desempeño académico de los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro y cómo se agrupan los individuos de acuerdo a su rendimiento en dichas pruebas.

Bibliografía

[1] CNA, "¿Qué significa calidad en la educación superior?¿Cómo se determina? ." [Online]. Available: https://www.mineducacion.gov.co/CNA/1741/article-187264.html

[2] M. de Educación Nacional de Colombia, "Calidad en Educación Superior." [Online]. Available: https://www.mineducacion.gov.co/1621/article-87349.html

[3] A. Uribe and M. Vélez, "Decreto 3963 de 2009 - Gestor Normativo Función Pública." [Online]. Available: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=37606

[4] C. H. Menacho Chiok, "Predicción del rendimiento académico aplicando técnicas de minería de datos." [Online]. Available: http://revistas.lamolina.edu.pe/index.php/acu/article/view/811/pdf_43

[5] ICFES, "Lectura Crítica," 2020. [Online]. Available: https://www.icfes.gov.co/documents/20143/1519985/Infografia+de+lectura+critica+saber+pro+2020.pdf

[6] ——, "Razonamiento Cuantitativo," 2020. [Online]. Available: https://www.icfes.gov.co/documents/20143/1519985/Infografia+de+razonamiento+cuantitativo+saber+pro+2020.pdf

[7] ——, "Competencias Ciudadanas," 2020. [Online]. Available: https://www.icfes.gov.co/documents/20143/1519985/Infografia+de+competencias+ciudadanas+saber+pro+2020.pdf

[8] ——, "Comunicación Escrita," 2020. [Online]. Available: https://www.icfes.gov.co/documents/20143/1519985/Infografia+de+comunicacion+escrita+saber+pro+2020.pdf

[9] ——, "Inglés," 2020. [Online]. Available: https://www.icfes.gov.co/documents/20143/1519985/Infografia+de+ingles+saber+pro+2020.pdf

[10] ICFES, "Guía de orientación de módulos de competencias genéricas," 2020. [Online]. Available: https://www2.icfes.gov.co/documents/20143/1891934/Guia+de+orientacion+de+Modulos+genericos+Saber+Pro-2020.pdf/a0f24d6f-d82e-cf94-0c1b-2c6fb3acddb0?t=1597776761084

[11] B. Mundial, "Educación." [Online]. Available: https://www.bancomundial.org/es/topic/education/overview

[12] SAS, "¿Qué es la minería de datos?" [Online]. Available: https://www.sas.com/es_co/insights/analytics/data-mining.html

[13] Sinnexus, "Datamining (Minería de datos)." [Online]. Available: https://www.sinnexus.com/business_intelligence/datamining.aspx

[14] S. Timarán, I. Hernández, S. Caicedo, A. Hidalgo, and J. Alvarado, *DESCUBRIMIENTO DE PATRONES DE DESEMPEÑO ACADÉMICO*. [Online]. Available: https://repository.ucc.edu.co/bitstream/20.500.12494/1039/2/LIBRO_DESCUBRIMIENTO_PATRONES_DESDEMPE%C3%91O.pdf

[15] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," (1994).

[16] M.-S. Chen, J. Han, and P. Yu, *Data mining: an overview from a database perspective*. IEEE, (1996).

[17] G. Piatetsky-Shapiro, R. Brachman, T. Khabaza, W. Kloesgen, and E. Simoudis, *An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications*. AAAI, (1996).

[18] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan, (2001).

[19] J. Hernández, J. Ramírez, and C. Ferri, *Introducción a la Minería de Dato*. Pearson, (2005).

[20] R. Agrawal, S. Ghosh, T. Imielinski, and B. Iyer, "An interval classifier for database mining applications." *18th International Conference on Very Large Data Bases*, (1992). [Online]. Available: https://www.researchgate.net/publication/221311447_An_Interval_Classifier_for_Database_Mining_Applications

[21] K.-U. Sattler and O. Dunemann, "Sql database primitives for decision tree classifiers." (2002). [Online]. Available: https://www.researchgate.net/publication/2523553_SQL_Database_Primitives_for_Decision_Tree_Classifiers

[22] P. Cabena, P. Hadjnian, R. Stadler, J. Verhees, and A. Zanasi, *Discovering Data Mining: From Concept to Implementation*. IBM, (1998).

[23] P. Vizcaino, "Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de weka," (2008). [Online]. Available: http://www.konradlorenz.edu.co/images/stories/suma_digital_sistemas/2009_01/final_paula_andrea.pdf

[24] R. Quinlan, *Programs for machine learning*, (1993).

[25] E. Hernández and R. Lorente, "Minera de datos aplicada a la detección de cáncer de mama." (2009). [Online]. Available: http://www.it.uc3m.es/jvillena/irc/practicas/08-09/14.pdf

[26] M. García and A. Álvarez, "Análisis de datos en weka – pruebas de selectividad." (2010). [Online]. Available: http://www.it.uc3m.es/~jvillena/irc/practicas/06-07/28.pdf

[27] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., (2011).

[28] F. Soltero and D. Bodas, "Clasificadores inductivos para el posicionamiento web." [Online]. Available: http://profesionaldelainformacion.com/contenidos/2005/enero/1.pdf

[29] A. Agudo, J. Alonso, and R. Tejero, "Evaluación de modelos para predicción meteorológica." [Online]. Available: http://www.it.uc3m.es/jvillena/irc/practicas/04-05/21mem.pdf

[30] D. Becerra, "Estudio del Rendimiento Académico Aplicando Técnicas de Minería de Datos." [Online]. Available: https://dspace.unl.edu.ec/jspui/bitstream/123456789/10988/1/Becerra%20Encarnaci%c3%b3n%2c%20Darwin%20Andr%c3%a9s.pdf

[31] E. Yamao, "Predicción del Rendimiento Acaémico mediante minería de datos en estudiantes del primerciclo de la escuela profesional de ingeniería de computación y sistemas." [Online]. Available: https://repositorio.usmp.edu.pe/bitstream/handle/20.500.12727/3555/yamao_e.pdf?sequence=3&isAllowed=y

[32] Singular, "Crisp-dm: La metodología para poner orden en los proyectos." [Online]. Available: https://www.sngular.com/es/data-science-crisp-dm-metodologia/

[33] A. Azevedo and M. Santos, "Kdd, semma and crisp-dm: A parallel overview." (2008). [Online]. Available: https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview

[34] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0*. SPSS, (2000). [Online]. Available: https://www.the-modeling-agency.com/crisp-dm.pdf

[35] D. Larose and C. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, 2nd ed. Wiley, (2014). [Online]. Available: https://doc.lagout.org/Others/Data%20Mining/Discovering%20Knowledge%20in%20Data_%20An%20Introduction%20to%20Data%20Mining%20%282nd%20ed.%29%20%5BLarose%20%26%20Larose%202014-06-30%5D.pdf

[36] J. Gallardo, "Metodología para la definición de requisitos en proyectos de data mining (er-dm)." (2009). [Online]. Available: http://oa.upm.es/1946/1/JOSE_ALBERTO_GALLARDO_ARANCIBIA.pdf

[37] ICFES, "Diccionario de variables saber pro periodo 2012-2018." [Online]. Available: https://www.icfes.gov.co/documents/20143/518379/Diccionario%20saber%20pro%202012%20-%202018%20genericas.pdf

[38] P. Spicker, S. Alvarez, and D. Gordon, "hh." [Online]. Available: http://biblioteca.clacso.edu.ar/gsdl/collect/clacso/index/assoc/D9393.dir/h.pdf

[39] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. IEA, (1988). [Online]. Available: http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf

[40] S. Boughorbe, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using matthews correlation coefficient metric," (2017). [Online]. Available: https://journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0177678

[41] J. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," (2010). [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20736804/