MINERÍA DE DATOS Y TEXTO EN LAS PRUEBAS SABER PRO

Análisis de los factores asociados a los resultados del examen Saber Pro, de los estudiantes de Ingeniería Industrial de universidades en Colombia, usando técnicas de minería de datos y minería de texto.

Juan Camilo Parra Moreno y Paola Andrea Espinosa Orjuela

Proyecto de grado presentado como requisito para optar el título de Ingeniero e Ingeniera Industrial

Director

Henry Lamos Díaz

Ph.D en Matemática-Física

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga, Santander

2021

Agradecimientos

A mi familia.

Juan Camilo Parra Moreno

Agradecimientos

A mi familia, mi mamá Claudia Orjuela, papá José Telesforo Espinosa y hermano Felipe Espinosa, quienes fueron el apoyo incondicional y motivación en cada momento, por su amor incondicional, enseñanzas y esfuerzos.

a todas las amigas y amigos que a lo largo de la carrera fueron una pieza de aprendizaje y crecimiento mutuo.

A la Universidad Industrial de Santander, por permitirme hacer parte de esta prestigiosa institución y por todos los momentos compartidos en el campus universitario.

Paola Andrea Espinosa Orjuela

Tabla de contenido

Lista de tablas

Lista de figuras

Lista de apéndices

Los apéndices están adjuntos y puede visualizarlos en la base de datos de la biblioteca UIS

Apéndice A. Análisis bibliométrico

Apéndice B. Bases de datos de competencias genéricas

Apéndice C. Base de datos de competencias específicas

Apéndice D. Código Minería de Datos

Apéndice E. Formulario de encuesta

Apéndice F. Base de datos con resultados de encuesta

Apéndice G. Código Minería de textos

Apéndice H Artículo científico

Resumen

Título: Análisis de los factores asociados a los resultados del examen Saber Pro, de los estudiantes de Ingeniería Industrial de universidades en Colombia, usando técnicas de minería de datos y minería de texto.[1*]

Autor: Moreno Parra, Juan Camilo; Espinosa Orjuela, Paola Andrea[2*3*]

Palabras Clave: Minería de datos, Aprendizaje automático, Algoritmos, Analítica de datos, Metodología KDD, aprendizaje no supervisado, aprendizaje supervisado

Descripción: Las entidades del Estado Colombiano generan una gran cantidad de datos diariamente mediante sus procesos misionales, estos datos deben ser transformados en información útil que conduzca a una mejora en la eficiencia del gobierno en prácticas dirigidas al análisis y el perfeccionamiento del diseño de políticas públicas .En este artículo se explora la aplicación de la metodología KDD (Knowledge Discovery in Databases) a una base de datos que contiene los resultados del examen Saber Pro-2019 publicados por el ICFES. Se realizaron análisis estadísticos y matemáticos a las variables consideradas de interés, se ajustó un modelo de clasificación a Instituciones de Educación Superior (IES) , se agruparon los estudiantes por medio de una clasificación socioeconómica y se crearon los perfiles tipológicos de cada grupo. Adicionalmente, se estudió la relación entre la clasificación socioeconómica, la institución de educación superior y el puntaje global por medio de un árbol de decisión. Por último, se realizó un análisis de sentimientos y se construyeron nubes de palabras con base a los resultados de una encuesta relacionada con las pruebas saber Pro. Se concluye que la categoría socioeconómica si es una variable con significancia estadística sobre el puntaje global, sin embargo, su influencia se ve trivializada en comparación con el efecto de la institución de educación superior con el puntaje. También se observa que hay una mayor proporción de estudiantes con categorías socioeconómicas más altas conforme aumenta la categoría la universidad.

[1*] Trabajo de Grado

[2**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Díaz. Ph.D en Matemática-Física.

[3]

Abstract

Title: Application of data mining techniques to the Saber Pro 2019 test results in Industrial Engineering students.[4*]

Author(s): Moreno Parra, Juan Camilo; Espinosa Orjuela, Paola Andrea[5*6*]

Key Words: Data mining, Machine learning, Algorithms, Data analytics, KDD Methodology, unsupervised learning, supervised learning

Description: The Colombian State generates a large amount of data daily through his missionary processes, this data must be transformed into useful information that leads to improved government efficiency in practices aimed at the analysis and improvement of public policy design. This article explores the application of the KDD (Knowledge Discovery in Databases) methodology to a database containing the results of the Saber Pro-2019 tests published by the ICFES. Statistical and mathematical analyses were performed on the variables deemed of interest, a classification model was adjusted on the Higher Education Institutions (IES), students were grouped by socioeconomic classification and typological profiles were created for each group. Additionally, the relationship between socioeconomic classification, higher education institution and overall score was studied by means of a decision tree. Finally, a sentiment analysis was performed and word clouds were constructed based on the results of a survey related to the saber Pro tests. It is concluded that the socioeconomic category is a variable with statistical significance on the overall score; however, its influence is trivialized in comparison with the effect that the institution of higher education has over the global score. It is also observed that there is a greater proportion of students with higher socioeconomic categories as the university category increases.

[4*] Bachelor's degree

[5**] Faculty of Physical Mechanical Engineering. School of Industrial and Business Studies. Director: Henry Lamos Díaz. Ph.D in mathematics-physics.

[6]

Introducción

A nivel mundial, el desarrollo de pruebas estandarizadas se realiza con el fin de demostrar de manera cuantitativa lo que los estudiantes han aprendido y retenido en el entorno académico, estos resultados se utilizan como un estándar para medir el progreso de la institución en el desarrollo de habilidades y la generación de conocimiento en sus alumnos (Wray, 2016), concientizando a las instituciones educativas en cuanto su responsabilidad frente al logro y desempeño académico de sus estudiantes (Driscoll, Halcoussis, & Svorny, 2008).

A mitad del siglo XX, ante la necesidad de valorar el rendimiento de las instituciones educativas en Latinoamérica, se solicitó a un grupo de profesores preparar pruebas de selección múltiple y estandarizar los resultados en una curva, pero fue a partir de los años 60s que este tipo de pruebas comenzaron a ser usadas como exámenes de entrada en programas con alta demanda en educación superior como la Universidad Nacional Autónoma de México (UNAM) y el Instituto Tecnológico de Monterrey (ITM).

Por su parte, en 1968 en Colombia se fundó el Instituto Colombiano para el Fomento de la Educación Superior (ICFES), encargado de la elaboración de estas pruebas (Rizo, 2010) con su posterior aplicación en la evaluación de la educación en las etapas de básica primaria y secundaria. Finalmente, para el 2000, el examen se extiende a la evaluación profesional con la creación de las pruebas ECAES y en 2009 se agrupan todas las pruebas existentes en lo que hoy se conoce como las pruebas SABER. (“50 años del Icfes,” n.d.).

Dado que los resultado de dichas pruebas permiten realizar una medición de la calidad de los programas ofrecidos, las Instituciones de Educación Superior (IES) hacen uso de los resultados obtenidos para atraer nuevos estudiantes, sin contar que las políticas expedidas por el Ministerio

de Educación Nacional (MEN) son consideradas de acuerdo con el desempeño de las IES en las pruebas (Martínez Lobo, 2013) es por ello que para las IES sea de vital importancia conocer y entender cómo los resultados del examen SABER PRO pueden ser influenciados por ciertos factores, por ejemplo, por factores académicos y sociodemográficos, con el fin de enfocar sus recursos y políticas hacia la implementación de programas de apoyo que permitan mejorar sus resultados.

Una de las maneras más efectivas de evaluar cuáles son los factores que más influyen sobre los resultados es a través de la analítica de datos, que consisten en la aplicación de algoritmos para el análisis de medianos o grandes conjuntos de datos, que buscan extraer tendencias, patrones y relaciones entre variables que faciliten y optimicen la toma de decisiones (Ed & Goebel, 2013).

En consecuencia, este proyecto propone la aplicación de herramientas computacionales de analítica de datos, en conjunto con métodos estadísticos y minería de texto para valorar el rendimiento de los estudiantes del programa de Ingeniería Industrial en las pruebas SABER PRO con el fin de identificar las diferentes variables que afectan el rendimiento de sus estudiantes, y realizar una caracterización adecuada de los mismos y las variables que afectan su rendimiento con el fin de obtener información cuantitativa y cualitativa necesaria que faciliten a las entidades correspondientes el desarrollo e implementación de políticas educativas que favorezcan el programa y mejoren la calidad de la educación recibida por sus estudiantes. El cumplimiento de los objetivos planteados en el presente trabajo de investigación se puede observar en la Tabla 1.

Tabla 1

Cumplimiento de los objetivos del proyecto

Objetivos específicos	Numerales relacionados
1. Aplicar técnicas de ETL para la construcción de una base de datos con los resultados de las Pruebas Saber Pro-2019 de los estudiantes de ingeniería industrial, a partir de los datos publicados por el ICFES.	Sección 7.2
2. Identificar las técnicas adecuadas de minería de datos y métodos multivariados para el descubrimiento de patrones presentes en los datos.	Sección 7.5
3. Construir los perfiles tipológicos de los estudiantes de ingeniería industrial por medio de técnicas de *clustering.*	Sección 7.5.5
4. Crear un texto estructurado a partir de una encuesta acerca de los sentimientos y opiniones presentados por los estudiantes de ingeniería industrial de la UIS hacia las Pruebas Saber Pro-2019, por medio de técnicas de Procesamiento de Lenguajes Naturales (PLN).	Sección 8.1 y 8.2
5. Analizar los sentimientos expresados hacia la prueba Saber Pro-2019 por parte de los estudiantes de ingeniería industrial de la UIS.	Sección 8.2
6. Elaborar un artículo científico de carácter publicable con base a la investigación realizada que contenga los resultados del proyecto de investigación.	Apéndice H

1. Planteamiento del problema

Las entidades del Estado Colombiano generan una gran cantidad de datos diariamente mediante sus procesos misionales, estas organizaciones están obligadas a publicar sus documentos, lo que se conoce como Datos Abiertos, todo esto hace parte del modelo de implementación del e-Gobierno, regido por el decreto 2693 de 2012. La generación de estos documentos no es en sí

misma una garantía de mejoramiento, los datos que contienen deben ser transformados en información útil que conduzca a un crecimiento económico e innovador, así como a incrementar la eficiencia del gobierno en mejoras dirigidas al análisis, el perfeccionamiento en el diseño de políticas públicas, fomentar la transparencia y rendición de cuentas y de esta manera crear un impacto social positivo y notable frente a todos los individuos. (Gobierno de Colombia & MinTIC, 2016).

Datos Abiertos; es una iniciativa la cual promueve que todas las entidades del Estado publiquen y divulguen sus datos no sensibles de manera unificada y en formato abierto, con la intención de que sean usados por cualquier persona para desarrollar aplicaciones o servicios de valor agregado, hacer análisis e investigación, o ejercer control ciudadano. (Gobierno de Colombia & MinTIC, 2016).

Los resultados de las pruebas Saber Pro hacen parte de los documentos publicados por el Ministerio de Educación Nacional (MEN) como parte de la plataforma datos abiertos. El examen Saber Pro es un instrumento estandarizado para la evaluación externa de la calidad de la educación superior. El análisis de los resultados puede usarse para tomar decisiones y enfocar el gasto de los recursos de manera asertiva para el mejoramiento de la educación en el país, por parte del ministerio de educación como de las instituciones de educación superior. Sin embargo, esta no es una tarea sencilla, los datos publicados cuentan con numerosas variables de diferentes tipos; se maneja un volumen de datos significativo, que guardan relaciones complejas; hay datos con campos vacíos, guardan una amplia variabilidad, entre otras (Oviedo Carrascal & Jiménez Giraldo, 2019). Esto causa una amplia barrera para entender y transformar los datos en información útil.

El ministerio de educación por medio de las pruebas estandarizadas mide el grado de desarrollo de habilidades, conocimientos generales y particulares de cada uno de los estudiantes

por institución (ICFES, 2019). Un análisis superficial de los datos es insuficiente ya que los factores propios del plantel educativo sólo explican hasta el 29% del rendimiento por estudiante, se observa que existe una gran variabilidad en los resultados de los estudiantes de una misma institución educativa e incluso dentro de un mismo programa académico (Gil, Rodríguez, Sepúlveda, Rondón, & Gómez-Restrepo, 2013). Algunos autores atribuyen amplias diferencias a causa de las desigualdades en el entorno económico y sociodemográfico del estudiante (Ramírez & Teichler, 2014) (Gil et al., 2013).

Numerosos estudios se han realizado con el fin de analizar los resultados de las pruebas Saber Pro por parte de las instituciones de educación superior y unos más específicos por parte del ICFES; (Alexander, Perdomo, & Esther, 2014) (Vidal-Alegría & Timarán-Pereira, 2019) (Gil et al., 2013) (Pardo Franco, 2017) (Rodríguez Albor, Ariza Dau, & Ramos Ruíz, 2014), estos estudios usaron técnicas de estadística descriptiva tales como ANCOVA y ANOVA para regiones y carreras específicas debido a la dificultad de escalar estos procedimientos a las bases de datos publicadas por el ICFES; además, no se generan explicaciones con claridad a los fenómenos que pueden ser revelados con minería de datos. Otras investigaciones recientes revelan la utilidad del uso de nuevas tecnologías (Villafañe Blanco, 2015) (Oviedo Carrascal & Jiménez Giraldo, 2019) (Gonzalez Montes & Guillen Ibarra, 2019), en estos estudios se llegó a la conclusión, que las Instituciones de educación superior y los factores sociodemográficos juegan un rol influyente en los resultados individuales de los estudiantes y se crearon algunos modelos predictivos, pero estas investigaciones se realizaron para programas y regiones del país específicas que no se pueden generalizar porque estarían sujetos a sesgos muestrales. Hasta la fecha no se ha encontrado ninguna investigación que abarque la minería de texto como herramienta de apoyo al análisis de los resultados.

En nuestro país la ingeniería adopta numerosas especialidades y subespecialidades según las áreas del conocimiento que predominan en ella, según el Sistema Nacional de Información de la Educación Superior (SNIES), los programas académicos de ingeniería industrial activos cuentan con el mayor número de egresados, según el SNIES en el 2018 hubo 7 478 graduados, seguido de ingeniería civil con 5 727. En el caso de la Universidad Industrial de Santander según el reporte *La UIS en cifras*, hay 1589 estudiantes matriculados en programa de ingeniería industrial en el periodo 2019-2, es la carrera con el mayor número de estudiantes de la universidad, lo que la convierte en la población estudiantil homogénea más significativa en el entorno colombiano, además, es el foco de interés de uno de los principales actores de este proyecto, la Escuela de Estudios Industriales y Empresariales UIS.

Este proyecto busca estudiar las diferentes relaciones que guardan las variables estudiadas por el ICFES con los resultados de cada módulo de las pruebas genéricas y específicas, crear una perfilación de clasificación tipológica para los estudiantes de ingeniería industrial en Colombia y además, crear un texto estructurado a partir de encuestas realizadas a estudiantes que presentaron el examen Saber Pro del 2019, todo esto mediante técnicas de minería de datos y minería de texto. Esto con el fin de apoyar la misión de la plataforma Datos Abiertos del Ministerio de Tecnologías de la Información y Comunicaciones (MinTic) mediante la creación de conocimiento útil para el sector educativo y de investigación en el país, de interés para el estado y las instituciones educativas. El desarrollo e implementación de soluciones no hace parte del alcance de este proyecto, estas medidas se dejan en manos de individuos y entes con la experticia en la creación de políticas educativas, lo que sí busca es guiar y generar información útil y de fácil interpretación que sirva como base a estas.

2. Objetivos

2.1 Objetivo general

Analizar los factores asociados a los resultados de las Pruebas Genéricas y Específicas del examen Saber Pro 2019 publicados por el ICFES, de los estudiantes de Ingeniería Industrial en Colombia, usando técnicas de minería de datos y minería de texto.

2.2 Objetivos Específicos

Aplicar técnicas de ETL para la construcción de una base de datos con los resultados de las Pruebas Saber PRO 2019 de los estudiantes de ingeniería industrial, a partir de los datos publicados por el ICFES.

Identificar las técnicas adecuadas de minería de datos y métodos multivariados para el descubrimiento de patrones presentes en los datos.

Construir los perfiles tipológicos de los estudiantes de ingeniería industrial por medio de técnicas de *clustering.*

Crear un texto estructurado a partir de una encuesta acerca de los sentimientos y opiniones presentados por los estudiantes de ingeniería industrial de la UIS hacia las Pruebas Saber Pro 2019, por medio de técnicas de Procesamiento de Lenguajes Naturales (PLN).

Analizar los sentimientos expresados hacia la prueba Saber Pro 2019 por parte de los estudiantes de ingeniería industrial de la UIS.

Elaborar un artículo científico de carácter publicable con base a la investigación realizada que contenga los resultados del proyecto de investigación.

3. Metodología

3.1 Revisión bibliográfica

En esta etapa se hace una revisión y análisis de artículos científicos relacionados con el tema de la investigación, con el fin de estudiar las principales técnicas de minería de datos y textos, propuestos y utilizados hasta el momento.

3.2 Metodología KDD

3.2.1 Selección de datos

En esta etapa, se determinan las fuentes de datos a utilizar. Para el desarrollo de esta investigación se emplea la base de datos de los resultados de los componentes genéricos y específicos de la prueba Saber Pro del año 2019 tomados de la plataforma datos abiertos, de estudiantes del programa de ingeniería industrial de universidades en Colombia.

- Obtener una base de datos en bruto a partir de los resultados en la red gubernamental de Datos Abiertos publicados por el ICFES de las personas que aceptaron el tratamiento de sus datos personales no sensibles.
- Crear una base de datos exclusiva de estudiantes del programa de ingeniería industrial publicada por el ICFES.

8.2.2 Reprocesamiento

Una vez obtenida la base de datos, se procederá a preparar y limpiar los datos extraídos, con el fin de obtener una estructura de datos adecuada para la siguiente fase de transformación.

- Depurar la base de datos; sobreescritura en campos nulos o eliminación de estos.

3.2.3 Transformación

En esta fase, se requiere transformar las variables compuestas por texto, es decir, categóricas, en variables numéricas, para poder ser utilizadas en la fase posterior y también es necesario la normalización de los datos. En el cumplimiento de esta etapa se realizan las siguientes actividades:

- Codificación de las variables categóricas que pueden tener sólo dos tipos de valores, en una representación numérica binaria (de ceros y unos).
- Codificación las variables categóricas ordinales, que pueden tener más de dos tipos de valores, en valores numéricos ordinales que conserven su significado.
- Codificación de las variables categóricas nominales en numéricas por medio de procedimiento más complejos.

3.2.4 Minería de datos

Una vez realizada la fase de transformación, se procede a aplicar diferentes modelos de minería de datos, con el objetivo de extraer patrones desconocidos, "ocultos" en los datos, que son útiles para la investigación. Para esta fase se realizarán las siguientes actividades:

- Entrenar diferentes modelos de minería de datos.
- Analizar los modelos obtenidos, obtener información útil, relaciones y tendencias.
- Aplicar técnicas de *clustering* con el fin de agrupar a los estudiantes en grupos comunes.
- Analizar las métricas decisorias en la formación de los grupos con el fin de crear los perfiles tipológicos de los estudiantes.
- Clasificar a los estudiantes de ingeniería industrial de la Universidad Industrial de Santander de acuerdo a los perfiles tipológicos.

3.2.5 Interpretación y evaluación

En esta última fase de la metodología KDD, se procede a validar los diferentes modelos de minería de datos aplicados anteriormente, con el fin de verificar si las conclusiones generadas por estos modelos son válidas y satisfactorias.

3.3 Metodología KDT

3.3.1 Recopilación de información

En esta fase se realiza una encuesta a estudiantes de la Universidad Industrial de Santander de los programas de ingeniería industrial, que hayan presentado las pruebas Saber Pro 2019, con tres preguntas abiertas en las que se sugerirá que las respuestas tengan cincuenta palabras y se indague sobre sobre su opinión acerca de este examen, del rol que tuvo su universidad sobre su desempeño y por último la experiencia de haberla presentado. Estos resultados se obtendrán en formatos de bloques de texto que serán posteriormente analizados. De este texto se busca obtener como base la mayor cantidad de palabras para su posterior análisis de sentimientos.

3.3.2 Preprocesamiento de texto

Esta fase consiste en convertir el texto extraído de la encuesta en una secuencia bien definida de unidades. Se procede a aplicar todos los procesos respectivos del preprocesamiento de texto, entre ellos:

- Verificar la compatibilidad de los módulos de procesamiento de lenguaje para la lengua castellana, en caso de no existir soporte, transformar los textos al idioma inglés.
- Realizar limpieza de texto y filtrado.

3.3.3 Transformación del texto

Una vez completadas las tareas del preprocesamiento de texto, es necesario para el cumplimiento de esta fase generar características y a su vez seleccionar las más importantes con el fin de eliminar las que no sean relevantes o incluso redundantes para el proceso en cuestión.

3.3.4 Aplicación de técnicas de minería de texto

En esta última etapa se procede a realizar las siguientes actividades para llevar a cabo la metodología KDT:

- Aplicar técnica de análisis de sentimientos a los textos estructurados depurados.
- Analizar los resultados obtenidos a partir de la minería de texto y extraer información útil para mejorar el desempeño de los estudiantes.

4. Revisión de literatura

4.1 Análisis preliminar de la literatura

A partir de la búsqueda inicial de artículos en la base de datos Scopus, se considera que es necesario hacer claridad entre dos elementos de esta línea de investigación que son *Learning Analytics* (LA) y *Educational Data Mining* (EDM). El primero de ellos, hace referencia a la medición, recolección, análisis y reporte de datos con el propósito de optimizar el aprendizaje de los estudiantes (Tomasevic, Gvozdenovic, & Vranes, 2020), por otro lado, EDM se concentra en el uso de aplicaciones de minería de datos, *machine learning* (ML) y estadísticas, para generar información acerca de los patrones de aprendizaje del grupo de estudio en cuestión (Tomasevic et al., 2020). Tomando lo anterior en consideración, es necesario tener en cuenta las diferentes técnicas de minería de datos que han sido empleadas en el campo educacional, algunos de los

artículos encontrados que se enfocan en llevar a cabo este tipo revisiones y que, por ende, son de relevancia para el progreso de la actual investigación, se presentan a continuación.

Uno de los artículos consultados se titula *"An overview and comparison of supervised data mining techniques for student exam performance prediction"* (Tomasevic et al., 2020), en el que lleva a cabo una comparación de las técnicas de minería de datos empleadas en la predicción del desempeño de los estudiantes en los exámenes. Según los autores, las técnicas de DM supervisadas funcionan mejor en las tareas de predicción, que las no supervisadas y, por tanto, toman en consideración tres tipos de categorías de técnicas supervisadas: basadas en similitud, en modelación y en acercamientos probabilísticos, cuya metodología se describe a continuación (Tomasevic et al., 2020).

- Se hace la predicción por medio de la identificación de estudiantes con desempeños similares.
- Estimación por medio de la construcción de un modelo a partir de la correlación entre los datos de entrada en el aprendizaje.
- Adaptación de los resultados observados a una distribución estadística.

Los datos con los que se alimentan las máquinas de aprendizaje en esta investigación se dividieron en tres tipos, el desempeño histórico del estudiante, motivación estudiantil y datos demográficos. Parte de esta información se obtiene de manera manual, principalmente el último tipo de datos o puede provenir del uso de herramientas informáticas en la metodología educacional en la que se ve inmerso el estudiante, por ejemplo, plataformas virtuales, historial de búsqueda y material de aprendizaje (Tomasevic et al., 2020). Los resultados obtenidos por Tomasevic et al. (2020), reflejaron que la técnica de aprendizaje automático que llevó a cabo una mejor predicción del desempeño de los estudiantes fue la red neuronal artificial (*Artificial Neural Network* o ANN).

Otro de los artículos de relevancia encontrados se titula *"Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles"* (Aissaoui, El Madani, Oughdir, & Allioui, 2019), en él se considera un ambiente de formación virtual (*e-learning*) y se orienta a la creación de material de enseñanza personalizado de acuerdo con el estilo de aprendizaje de cada estudiante para potenciar el proceso de adquisición de conocimientos (Aissaoui et al., 2019). Los autores emplean el *Felder and Silverman Learning Style Model* (FSLSM) para identificar el estilo de aprendizaje de los estudiantes de acuerdo con las 4 categorías del FSLSM que son: procesamiento (activo / reflexivo), percepción (sensorial / intuitivo), entrada (visual / verbal) y comprensión (secuencial / global), posteriormente, se hacen agrupaciones de objetos por medio del algoritmo no supervisado *K-Modes* y finalmente se aplica el modelo de aprendizaje automático supervisado *Naive Bayes Classifier* para efectuar la predicción en tiempo real, la metodología empleada por los autores obtuvo un nivel de precisión del 89% (Aissaoui et al., 2019).

Además de estos estudios, y con ánimos de tener una mirada más global de las investigaciones realizadas en el área de EDM, se presenta el artículo *"Student performance análisis and prediction in classroom learning: A review of educational data mining studies"* en el que se hace una revisión de 140 estudios publicados entre los años 2000 y 2018, enfocados a clases presenciales identificando los predictores, los métodos empleados, el tiempo y el objetivo de la predicción (Khan & Ghosh, 2020). De acuerdo con el estudio algunos de los factores más populares son los antecedentes del estudiantes, su comportamiento y sus evaluaciones externas e internas; por el contrario, uno de los menos considerados son el conocimiento del área, adicionalmente, se determinó que los métodos más usados son la regresión y la clasificación; sin embargo, de todos los estudios efectuados sólo el 33% son capaces de predecir el desempeño de

los estudiantes y que el enfoque primario es la mejora del porcentaje de predicción de los modelos (Khan & Ghosh, 2020).

En el caso de la formación online (e-learning), otros autores también han llevado a cabo revisiones recientes, como es el caso de Rodrigues, Isotani, & Zárate (2018) con el artículo "*Educational Data Mining: A review of evaluation process in the e-learning*", en esta revisión los autores se centran en el proceso de evaluación, que es una cuestión vital cuando no se trata de educación en un salón de clases y se hace necesario determinar qué tipo de factores deben considerarse; según la revisión realizada por los autores en algunos trabajos se habla de la evaluación del estudiante a través de su experiencia, percepción y actitud durante el aprendizaje o en otros casos, de la motivación, el nivel de ansiedad o del rendimiento académico; más sin embargo, concuerdan en afirmar que la instrucción impartida de manera virtual requiere una gran variedad de aspectos de evaluación para asegurar la efectividad de las estrategias de enseñanza (Rodrigues, Isotani, & Zárate, 2018).

En la revisión de Rodrigues et al. (2018), se busca resolver dos preguntas: ¿Cuáles son las perspectivas y tendencias de minería de datos en la evaluación de la educación virtual? Y ¿Cuáles son los temas potenciales de investigación que reciben poca atención desde la perspectiva del e-learning? La primera de estas cuestiones encontró que los artículos analizados estaban orientados hacía (Rodrigues et al., 2018):

- Evaluación de las prácticas de e-learning en aulas de clase tradicional considerando las interacciones entre los actores educacionales, la gestión realizada por el administrador de la enseñanza con respecto a los resultados del educador y los riesgos latentes que puede conllevar el proceso de aprendizaje.

- Evaluación del uso de recursos multimedia y la proposición de estrategias que promuevan el aprendizaje colaborativo en lugar del individual, como el uso de foros y chats.
- Identificación del comportamiento del estudiante al momento de usar herramientas tecnológicas como material educativo recomendado o sistemas de tutoría inteligente. Igualmente, la agrupación de alumnos con características de aprendizaje similares, para proponer estrategias personalizadas más acordes con las necesidades didácticas y pedagógicas de cada estudiante.
- Desarrollo de modelos híbridos y colaborativos que puedan fortalecer los canales de comunicación entre estudiante y educador, que son construidos por medio de las técnicas de minería de datos.

Algunas de las herramientas de aprendizaje automático supervisado y no supervisado empleadas en los artículos de esta revisión fueron el algoritmo *Predictive-A priori*, el proceso *Fuzzy Analytic Hierarchical* (FAHP), *Frequent Itemset* y reglas de asociación, clasificación *Multilayer Perceptron* y técnicas *Random Forest, K-Means y Self-Organizing Maps* (SOM), máquinas de soporte vectorial y el modelo oculto de Márkov, entre otros (Rodrigues et al., 2018). Finalmente, establecen la respuesta a la segunda pregunta, reconociendo temas de investigación futura los patrones de comportamiento del estudiante durante el proceso de aprendizaje, la investigación orientada a la cooperación y colaboración entre profesor y estudiante en el cumplimiento de logros, la identificación de los factores principales que influencian a los estudiantes al momento de usar recursos electrónicos educacionales y el mejoramiento del desempeño estudiantil por medio de la enfatización de sus habilidades y la identificación de sus deficiencias durante el aprendizaje (Rodrigues et al., 2018).

Otro estudio de revisión llevado a cabo determinó que en artículos realizados entre el año 1995 y 2005, las técnicas más populares de minería de datos son *clustering*, clasificación, patrones secuenciales, predicción y análisis de reglas de asociación; y el objetivo de mejora usual que se le da a la investigación es el mejoramiento de la enseñanza virtual o e-learning, ya que facilita la obtención de los datos de inicio de sesión y búsquedas del estudiante (Mohamad & Tasir, 2013). Sin embargo, uno de los aspectos que resalta la investigación es la carencia de incluir los aspectos colaborativos del aprendizaje que incluye estudiantes y profesores, por ejemplo, interacciones en foros o chats grupales (Mohamad & Tasir, 2013).

Los resultados de la investigación de Mohamad y Tasir (2013) concuerdan con los obtenidos por Aldowah, Al-Samarraie, & Fauzy (2019) en "*Educational data mining and learning analytics for 21st century higher education: A review and synthesis*" con lo que respecta a las técnicas de minería de datos. En él se identificaron 12 técnicas y se determinó el porcentaje de uso con respecto a la cantidad de artículos evaluados, dichas técnicas son: clasificación (26,25%), *clustering* (21,25%), minería de datos visual (15%), estadística (14,25%), minería por reglas de asociación (14%), regresión (10,25%), minería por patrones de secuencia (6,05%), minería de texto (4,75%), correlación (3%), detección de atípicos (2,25%), minería causal (1%) y estimación de densidad (1%) (Aldowah et al., 2019). Nuevamente, los autores determinaron que la aplicación de EDM es una manera efectiva de predecir patrones de interés para recrear modelos que promuevan ciertas actividades de aprendizaje y urgir a las instituciones de educación superior a incluir los análisis correspondientes en las áreas en las que sea factible la recolección de los datos para alimentar el modelo para predicciones en tiempo real (Aldowah et al., 2019).

5. Marco de antecedentes:

En diferentes universidades a nivel nacional, se han desarrollado trabajos en la modalidad de proyecto de investigación de pregrado y posgrado, que buscan analizar los resultados de las pruebas Saber Pro, enfocados en programas o regiones específicas del país. Se lleva a cabo una búsqueda web en los repositorios de algunas universidades nacionales como la Universidad Pontificia Bolivariana de Medellín, la Universidad Pedagógica y Tecnológica de Colombia, entre otras; y la biblioteca electrónica SciElo. En la búsqueda efectuada se emplearon palabras claves tales como: minería de datos, minería de texto y pruebas Saber Pro y; a continuación, se presentan algunos de los proyectos relevantes que contribuyen al desarrollo de esta investigación.

5.1 Antecedentes nacionales

En el año 2018, Jovanny Jiménez Giraldo, desarrolla un estudio de minería de datos educativos enfocado en el análisis de los factores económicos, sociales y demográficos que influyen en el desempeño de las pruebas Saber Pro en estudiantes de ingeniería en el departamento de Antioquia, aplicando diferentes modelos analíticos como *clustering*, selección de factores y predicción. En el experimento del *Clustering*, usando K-Means, se observó una separación clara entre los estudiantes que obtienen buenos resultados en inglés; (generalmente poseen una buena condición económica) y los que obtienen buen resultado en lectura crítica, (generalmente son de estratos bajos y pagan matrículas de bajo valor). Además, en el experimento de selección de factores se encontró que las variables más relevantes son: Cantidad de personas a cargo, metodología de enseñanza, hogar permanente, carácter académico de la institución y facilidades económicas como tener horno micro gas y motocicleta. Por último, a partir de 26 variables

determinadas en el ejercicio de selección de factores, se logró predecir el desempeño de las pruebas Saber Pro con una exactitud del 81% (Jiménez Giraldo, 2018).

Más adelante, en el año 2019, Camila González Montes y Sergio Guillen Ibarra desarrollaron un modelo para la caracterización y análisis de los resultados obtenidos en las pruebas Saber Pro del 2016 y 2017, de los programas de ingeniería en Colombia mediante la aplicación de minería de datos, con el fin de identificar el impacto de la calidad en el desempeño de las instituciones de educación superior y generar visualizaciones que permitan interpretar los resultados obtenidos por el modelo de manera sencilla. Este estudio, aplicó una metodología articulada en la técnica de Minería de datos para la clasificación y predicción. En los resultados obtenidos, se destaca la importancia de la acreditación institucional en universidades con programas de ingeniería industrial, ya que presentan un mejor desempeño que aquellas que no cuentan con dicha distinción, y las instituciones públicas cuentan con mejores resultados que las universidades privadas. Por otra parte, se identificó por varios métodos de Minería de Datos la capacidad de las instituciones a clasificarse entre ellas, con base en los resultados obtenidos, teniendo una precisión aproximada del 84% y 79% para los métodos jerárquico y no jerárquico respectivamente (González Montes y Guillen Ibarra, 2019).

Por su parte, José García González, Paola Sánchez Sánchez, Manuel Orozco y Sergio Obredor, llevaron a cabo en el año 2019 un estudio utilizando técnicas de extracción de Conocimiento para la predicción y análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia, con respecto al promedio académico por áreas genéricas mediante el uso de minería de datos, aplicando la metodología de extracción de conocimiento en bases de datos (KDD) se construyó una base de datos del desempeño académico del estudiante en áreas asociadas con los contenidos de dicha prueba y se utilizó redes neuronales como técnica para

la minería de datos, permitiendo la predicción de los resultados de la prueba Saber Pro con alta exactitud tanto en rangos cualitativos como cuantitativos. Además, se comprobó una correlación entre el desempeño académico y los resultados de Saber Pro. El estudio concluye que la metodología usada es una óptima guía para descubrir patrones ocultos en los datos y establecer estrategias de mejora de los resultados de estas pruebas que involucren el desempeño académico del estudiante (García González, Sánchez Sánchez, Orozco, y Obredor, 2019).

Del mismo modo, en el año 2017, Jorber Pardo Franco estudia los factores demográficos, académicos y socioeconómicos que influyen en los resultados del componente genérico de la prueba Saber Pro presentada por los estudiantes de Licenciatura en Matemáticas y Estadística desde el segundo semestre del 2011 al 2015. Dicho trabajo tuvo un enfoque cuantitativo, en el cual utilizaron herramientas estadísticas para la interpretación de los datos, tales como el análisis descriptivo y bidimensional, modelamiento por medio de los GAMLSS y RLO. Como aporte principal, identifica las principales variables que afectan los resultados de cada componente genérico (Pardo Franco, 2017).

5.2 Antecedentes internacionales

Como referentes de trabajo en Latinoamérica, se consultaron los repositorios de algunas de las universidades más importantes de América Latina, empleando palabras clave como: minería de datos, pruebas PISA (Programa para la Evaluación Internacional de Alumnos) y PSU (Prueba de Selección Universitaria). A partir de esta búsqueda, se presentan dos proyectos de grado de Chile del año 2017 relacionados al tema de investigación a tratar.

El primero de ellos, llevado a cabo por Felipe Bugueño, buscaba generar un modelo predictivo usando minería de datos sobre el perfil de los aspirantes a ingresar a la Facultad de Economía y Negocios de la Universidad de Chile. Entre las variables que se consideraron se

encontraba los resultados de las Notas de Enseñanza Media (NEM) y la Prueba de Selección Universitaria (PSU), que son exámenes estandarizados, empleados en la selección de candidatos para ingresar a un programa académico de educación superior; otras de las variables utilizadas incluían la edad, el sexo y las notas del colegio. Para la elaboración de la investigación, el autor, recurre a la metodología de minería de datos Knowledge Discovery in Databases (KDD), con técnicas de selección de variables, clústeres y combinación de modelos. Los resultados arrojaron que era posible predecir el 80,4% de los individuos que serían aceptados por la institución educativa (Bugueño, 2017).

El segundo proyecto, también del mismo año, llevado a cabo por Mauricio Miranda y Jheser Guzmán, trabajaba con datos sobre la deserción de los estudiantes de ingeniería de la Universidad Católica del Norte, de las ciudades Antofagasta y Coquimbo, mediante el uso de métodos de minería de datos. Los autores llevaron a cabo predicciones sobre la deserción estudiantil, empleando clasificadores de red bayesiana, árbol de decisión y red neuronal, y variables tales como: el puntaje en la prueba PSU, el NEM, los beneficios estudiantiles recibidos, las calificaciones semestrales de sus asignaturas, entre otras. Pudiendo predecir la deserción en más del 70% de los casos, y concluyendo que una de las variables con mayor peso eran los resultados de las pruebas PSU (Miranda y Guzmán, 2017).

A partir de los resultados de la búsqueda, se puede concluir que a nivel nacional e internacional existe un interés generalizado por los modelos predictivos empleando técnicas de minería de datos y que es un tema de investigación reciente del que se podría obtener herramientas de gestión para mejorar el desempeño académico de los estudiantes. Para esto, es necesario llevar a cabo un análisis de la información publicada por el Ministerio de Educación Nacional, de manera

que sea posible generar un valor agregado a esta línea investigativa y a las instituciones de educación superior en el país.

6. Marco Teórico

6.1 Minería de datos:

La minería de datos puede definirse simplemente como la extracción o minería de conocimiento de grandes cantidades de datos (Jiawei, Kamber, & Pei, 2014). Esta definición simple se puede seguir elaborando con la especificación clara de las características de las tareas de minería de datos. Por ejemplo, según (Frawley, Piatetsky- Shapiro, & Matheus, 1993) la minería de datos se puede ver como "la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos". (Berry & Linoff, 2000) definieron la minería de datos como "el proceso de exploración y análisis de grandes cantidades de datos con el fin de descubrir patrones y reglas significativas". A partir de las dos definiciones, se pueden observar los elementos en común de todas ellas; datos, conocimiento (información), tareas, algoritmos y resultados.

Aunque las definiciones difieren ligeramente entre investigadores, recae sobre nuestra comprensión intuitiva darle un marco común de referencia a la definición de esta disciplina relativamente nueva o aceptar el consenso dominante bajo el marco propuesto por el grupo de académicos AMC SIGKDD que definen la minería de datos como “El paso en el proceso KDD (Descubrimiento de Conocimiento en bases de datos) que consiste en aplicar algoritmos de análisis y detección de datos que, bajo limitaciones de eficiencia computacional aceptables, producen una enumeración particular de patrones (o modelos) sobre los datos” (Fayyad, Piatetsky-Shapiro, &

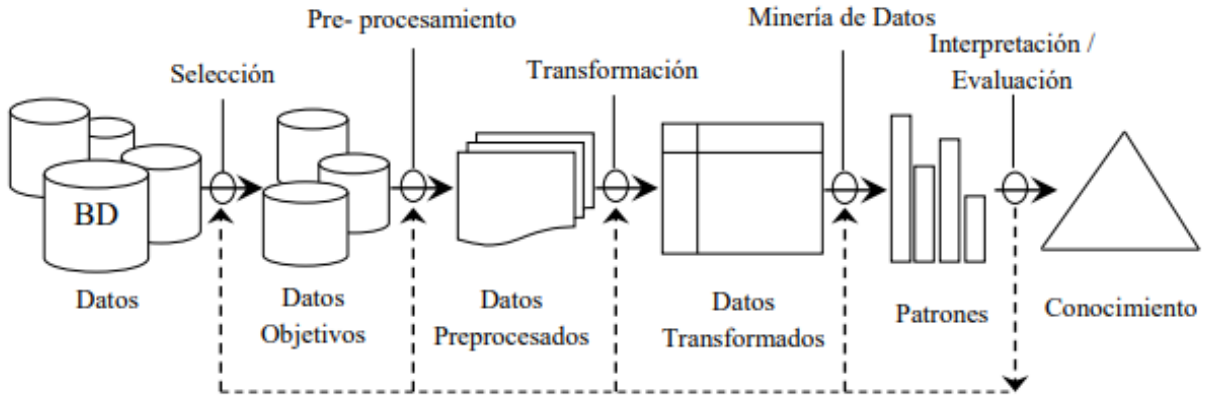
Smyth, 1996). Para el desarrollo de esta investigación se tratará a la minería de texto como un paso en el proceso KDD.

6.2 Proceso de Descubrimiento de Conocimiento en bases de datos KDD

KDD es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles en los datos. (Fayyad et al., 1996), en la figura 1 se puede observar esta metodología de manera abreviada.

Figura 1

Proceso KDD



Nota: adaptado de Beltran (2016)

Aquí, los datos son un conjunto de hechos (por ejemplo, casos en una base de datos), y el patrón es una expresión que describe un subconjunto de datos o un modelo aplicable al subconjunto.

Por lo tanto, extraer un patrón también designa ajustar un modelo a los datos; encontrar estructura a partir de datos; o, en general, hacer una descripción de alto nivel de un conjunto de datos (Beltran, 2016). El término proceso implica que el KDD comprende muchos pasos, que incluyen la preparación de datos, la búsqueda de patrones, la evaluación del conocimiento y el refinamiento de datos. Por no trivial, se designa que hay alguna búsqueda o inferencia involucrada;

es decir, no es un cálculo directo de cantidades predefinidas, como calcular el valor promedio de un conjunto de números. (Fayyad et al., 1996).

6.2.1 Pasos del proceso KDD

6.2.1.1 Selección de datos. Se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos de las fuentes de datos.

6.2.1.2 Preprocesamiento. Consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de forma necesaria para las fases posteriores. Además, utilizan diversas estrategias para gestionar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

6.2.1.3 Transformación. Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma más apropiada para la aplicación de las técnicas de minería de datos.

6.2.1.4 Minería de datos. Es la fase de modelamiento, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u “ocultos” en los datos. Los diferentes métodos serán descritos en el siguiente capítulo.

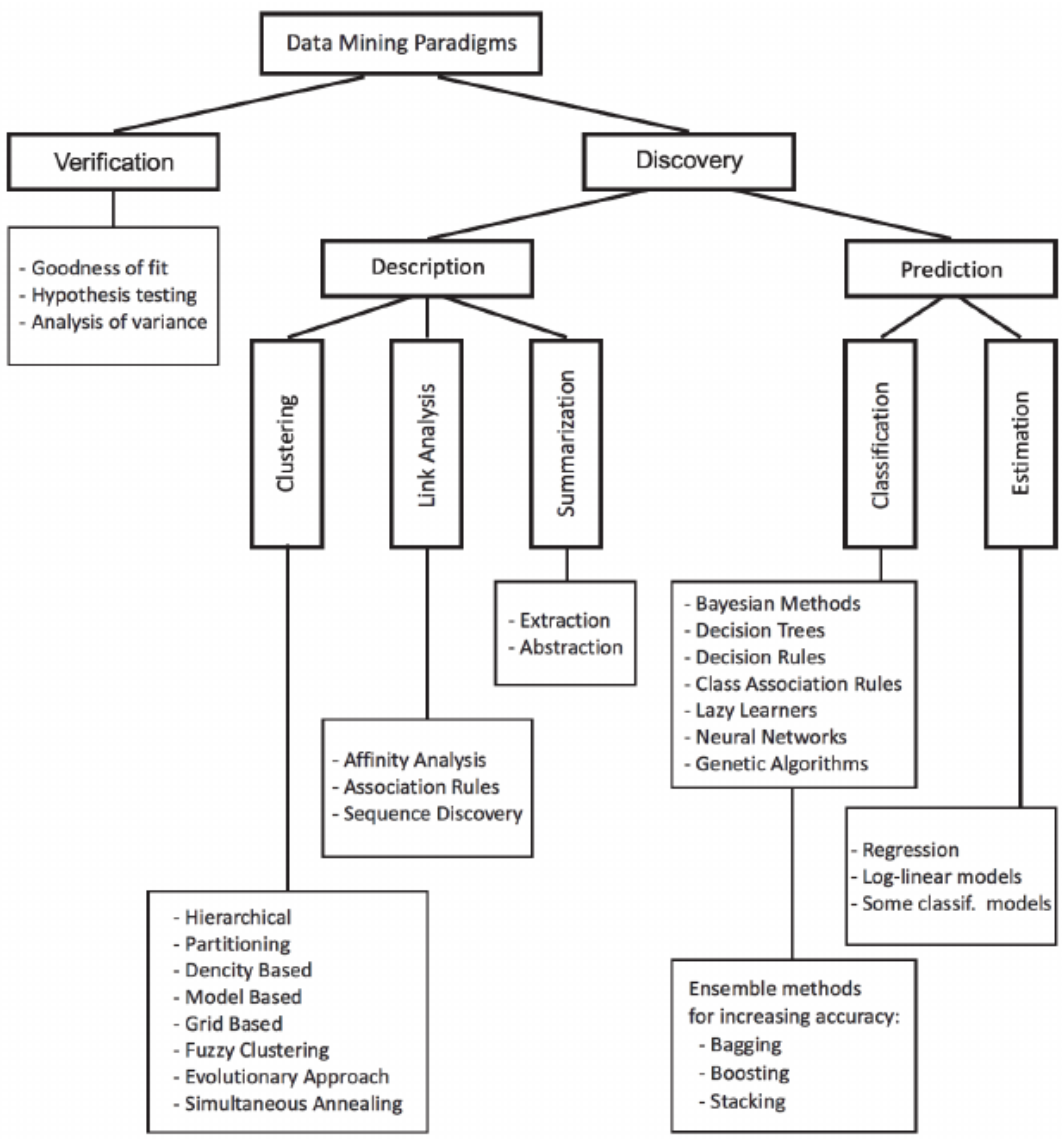
6.2.1.5 Interpretación y Evaluación. Se analizan los patrones obtenidos y que son realmente interesantes, así mismo se realiza la validación de los modelos (Santín & López, 2007)

6.3 Métodos de minería de datos

Los métodos de minería de datos se dividen esencialmente en dos tipos principales (Maimon & Rokach, n.d.) Además de subdivisiones adicionales, la clasificación propuesta por Mitov, Depaire, Ivanova, Vanhoof & Blagoev (2012) se puede evidenciar en la figura 2.

Figura 2

Diagrama de árbol métodos de minería de datos



Nota: adaptado de Mitov, Depaire, Ivanova, Vanhoof & Blagoev (2012).

- Orientado a la verificación (el sistema verifica la hipótesis del usuario)
- Orientado a la detección (el sistema encuentra nuevas reglas y patrones de forma autónoma) (Fayyad et al., 1996).

6.3.1 Métodos de verificación

Trata con la evaluación de una hipótesis propuesta por una fuente externa. Estos métodos incluyen los enfoques más comunes de las estadísticas tradicionales, como la prueba de bondad de ajuste, la prueba t y el análisis de varianza. Tales métodos no suelen estar asociados con la minería de datos porque la mayoría de los problemas de minería de datos están relacionados con el establecimiento de hipótesis en lugar de probar una conocida.

6.3.2 Métodos de descubrimiento

Son métodos que identifican automáticamente patrones en los datos. La rama del método de descubrimiento consiste en métodos de descripción versus métodos de predicción. (Mitov, Depaire, Ivanova, Vanhoof, & Blagoev, 2012)

6.3.2.1 Los métodos de minería de datos orientados a la descripción. Se centran en comprender cómo funcionan los datos subyacentes. Las principales orientaciones de estos métodos son la agrupación, el resumen y la visualización. Las principales direcciones de los métodos orientados a la descripción son la agrupación, el análisis de enlaces y el resumen.

6.3.2.1.1 Agrupamiento. Es el proceso de agrupar los datos en clases o grupos, de modo que los objetos dentro de un grupo tienen una gran similitud en comparación entre sí, pero son muy diferentes a los objetos en otros grupos. Las diferencias se evalúan en función de los valores de los atributos que describen los objetos, utilizando diferentes tipos de medidas de distancia.

6.3.2.1.2 Análisis de enlaces. Descubre relaciones entre datos. Se usa para 3 propósitos principales:

- Encontrar coincidencias en datos para patrones de interés conocidos.
- Encontrar anomalías donde se violan patrones conocidos.

- Descubrir nuevos patrones de interés. En esta dirección caen disciplinas como el análisis de afinidad, la minería de reglas de asociación y el descubrimiento de secuencias.

6.3.2.2 Métodos orientados a la predicción. Tienen como objetivo construir un modelo de comportamiento que pueda crear muestras nuevas y no observadas y que pueda predecir los valores de una o más variables relacionadas con la muestra. Aquí se obtienen dos ramas principales: clasificación y estimación. Estas dos formas de análisis de datos se utilizan para extraer modelos que describen clases de datos importantes o para predecir tendencias futuras de datos. La principal diferencia entre clasificación y estimación es que la *clasificación* asigna el espacio de entrada a clases predefinidas, mientras que los modelos de *estimación* asignan el espacio de entrada a un dominio de valor real. (Mitov et al., 2012)

6.3.2.2.1 Modelos de clasificación. Predicen etiquetas categóricas (discretas, desordenadas). La clasificación es el problema de identificar la subpoblación a la que pertenecen las nuevas observaciones, donde se desconoce la identidad de la subpoblación, sobre la base de un conjunto de datos de entrenamiento que contiene observaciones cuya subpoblación es conocida. Los nuevos elementos individuales se colocan en grupos en función de la información cuantitativa sobre una o más mediciones, rasgos o características, etc, y en función del conjunto de capacitación en el que ya se han establecido agrupaciones previamente decididas. Hay varios grupos grandes en los que pertenecen los clasificadores: métodos bayesianos, máquinas de vectores de soporte, árboles de decisión, reglas de decisión, reglas de asociación de clase, estudiantes perezosos, redes neuronales y algoritmos genéticos. Para aumentar la precisión recibida, se utiliza la técnica superior para métodos de conjunto o los llamados meta clasificadores como etapa superior. (Mitov et al., 2012)

6.3.2.2.2 Modelos de estimación. Construyen una función de valor continuo, o valor ordenado, que se utilizan como predictor (estimador). La técnica utilizada más común son los diferentes tipos de modelos de regresión (que involucran una sola variable predictiva o dos o más variables; regresión lineal o no lineal, etc.), mientras que también se utilizan otros modelos (como los modelos log-lineales que se aproximan discretamente a distribuciones de probabilidad multidimensional utilizando transformaciones logarítmicas). Algunos de los modelos de clasificadores también pueden ajustarse para usarse para la estimación (como árboles de decisión, redes neuronales, entre otros.) (Jiawei et al., 2014).

6.4 Minería de Texto

Es el descubrimiento por computadora de información nueva y previamente desconocida, mediante la extracción automática de información de diferentes recursos escritos. Un elemento clave es la vinculación de la información extraída para formar nuevos hechos o nuevas hipótesis para explorar más a fondo por medios de experimentación más convencionales. (Hearst, 1999)

La minería de texto es una variación del campo abarcado por la minería de datos, que trata de encontrar patrones interesantes de grandes bases de datos. La principal diferencia que existe entre estas dos es que en la minería de texto los patrones se extraen del texto en lenguaje natural en lugar de bases de datos estructuradas. Las bases de datos están diseñadas para que los programas las procesen automáticamente; el texto está escrito para que las personas lo entiendan, para facilitar esta transición se hace uso de herramientas de Procesamiento de Lenguaje Natural (LNP). (Witten, 2004).

6.5 Descubrimiento de Conocimiento en Textos (KDT)

En general, las fases principales del proceso de Descubrimiento de Conocimiento en Textos son las mismas que las del Descubrimiento de Conocimiento de bases de datos. Sin embargo,

existe una diferencia fundamental entre texto y datos que hace necesaria una revisión del proceso de descubrimiento de conocimiento: la falta de una estructura procesable automáticamente hace que las etapas de selección, preprocesamiento y transformación de los datos textuales sea imprescindible en el KDT, conformando todas ellas la fase de Preprocesamiento. (Justicia de la Torre, 2017). Debido a la naturaleza de los textos hace falta el preprocesamiento que es más complejo y riguroso. En la tabla 2 se puede visualizar la comparación entre las dos metodologías mencionadas.

Tabla 2

Comparación del proceso KDD y KDT.

	KDD	KDT
1	Comprender el dominio de la aplicación.	El usuario define qué conceptos le interesan.
2	Seleccionar un conjunto de datos objetivo.	Los textos se obtienen con herramientas de Recuperación de Información o de forma manual.
3,4	Limpieza, procesamiento y transformación de datos.	Se describen los conceptos y los textos serán analizados y representados mediante una Forma Intermedia (se eliminan *stop words*, palabras frecuentes poco relevantes…).
5	Desarrollo de modelos y construcción de hipótesis.	Identificación de los conceptos en la colección de textos.
6	Elección y ejecución de algoritmos adecuados de Minería de Datos.	Tareas de minería de textos.
7	Interpretación de resultados y visualización.	Interpretación de los resultados por un humano.

Nota: adaptado de Justicia de la Torre (2017).

6.6 Proceso de Descubrimiento de Conocimiento en Textos (KDT)

6.6.1 Preprocesamiento de texto

El preprocesamiento de texto se aplica a la recopilación de documentos que contienen datos no estructurados o semiestructurados. La tarea de procesamiento previo de texto convierte un archivo de texto sin procesar en una secuencia bien definida de unidades lingüísticamente significativas. Implica el siguiente tipo de procesamiento. (Gohil, 2015)

6.6.2 Limpieza de texto

Realiza tareas como la eliminación de anuncios de páginas web, la eliminación de tablas, figuras, etc.

6.6.3 Tokenización

Divide las oraciones en palabras eliminando espacios, comas, etc.

6.6.4 Filtrado

Elimina palabras que contienen poca o ninguna información de contenido, como artículos, conjunciones, preposiciones, etc. Las palabras que ocurren con mucha frecuencia también se eliminan.

6.6.5 Stemming

Es un proceso de transformación de la palabra a su raíz (forma normalizada). Construye una forma básica de palabras para identificar palabras por su raíz. P.ej. *ir* es raíz de *yendo*.

6.6.6 Etiquetado de parte del discurso (POS-Tagging)

Determina la categoría lingüística de la palabra. Asigna clase de palabra a cada token. En inglés, hay ocho clases de lexemas: sustantivo, pronombre, adjetivo, verbo, adverbio, preposición, conjunción e interjección, las librerías de programación están construidas sobre este idioma. Las

técnicas para el etiquetado POS son enfoques basados en el modelo oculto de Markov y enfoques basados en reglas.

6.6.7 Transformación de texto

Realiza la generación de características seguida de la tarea de selección de características. Esta primera tarea, representa documentos por las palabras que contienen y sus ocurrencias, donde el orden de las palabras no es significativo. Utiliza bolsas de palabras o modelo de espacio vectorial. Mientras que la selección de características es un proceso de selección de un subconjunto de características importantes para usar en la creación de modelos. Reduce la dimensionalidad al eliminar características redundantes e irrelevantes. (Gohil, 2015)

6.7 Aplicación de técnicas de minería de texto

Después de transformar el texto en una bolsa de palabras y un corpus, se puede aplicar cualquiera de las siguientes técnicas.

6.7.1 Categorización

La categorización de texto es el problema de asignar automáticamente categorías predefinidas a documentos de texto de formato libre. La mayor dificultad de la categorización de texto es la alta dimensionalidad del espacio de características. Las aplicaciones de categorización de texto incluyen organización de documentos, filtrado de spam, categorización de SMS, categorización jerárquica de páginas web.

6.7.2 Agrupamiento

La técnica de agrupamiento utiliza medidas de similitud entre diferentes objetos, es decir, coloca elementos más similares en una clase y objetos diferentes en otra. Difiere de la categorización ya que los objetos se agrupan sin conocimiento previo de las clases.

6.7.3 Resumen

El resumen de texto trata de condensar el texto en una forma más corta con la retención de su información y significado general. Se puede clasificar en resumen abstracto y extractivo. Un resumen abstracto intenta desarrollar una comprensión de los conceptos clave en el texto y luego representar esos conceptos en lenguaje natural. Utiliza métodos lingüísticos para comprender, interpretar y describir el texto en una versión más corta. El resumen extractivo se realiza mediante la obtención de segmentos de texto clave basados en el análisis estadístico de las características del texto, como la frecuencia de palabras / frases, la ubicación o las palabras clave para ubicar las oraciones que se extraerán. (Gohil, 2015)

6.8 Las Pruebas Saber Pro

El Examen Saber Pro es un instrumento estandarizado para la evaluación externa de la calidad de la educación superior, este examen tiene los objetivos de comprobar el grado de desarrollo de las competencias de los estudiantes, producir indicadores de valor agregado y servir de fuente de información para la construcción de indicadores de evaluación de la calidad de los programas e instituciones de educación superior. (ICFES, 2019)

Con la publicación de la Ley 1324 y el Decreto 3963 de 2009, se estableció que la presentación de los exámenes de Estado de la educación superior era obligatoria para obtener el título del nivel de pregrado. Por lo anterior, este examen está orientado a los estudiantes que han aprobado el 75% de los créditos de sus respectivos programas de formación profesional universitaria.

6.8.1 Componentes y sesiones de las Pruebas Saber Pro

En la primera sesión, los evaluados deben presentar los cinco módulos genéricos:

- Comunicación escrita

- Razonamiento cuantitativo
- Lectura crítica
- Competencias ciudadanas

Adicionalmente, los estudiantes deben responder el cuestionario socioeconómico. Este cuestionario se compone de preguntas cortas de selección múltiple que se responden al final de la hoja de respuestas. El cuestionario socioeconómico, 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. (ICFES, 2019)

La segunda sesión está conformada por módulos temáticos. El examen Saber Pro cuenta con 40 módulos en total, el ICFES oferta estos módulos por combinatorias, es decir, agrupaciones de entre uno y tres módulos de acuerdo con el grupo de referencia asociado a cada programa. Conforme a la resolución 395 del 12 de junio del 2018, los grupos de referencia se definen de acuerdo con el Núcleo Básico del Conocimiento – NBC y el nivel de formación establecido para cada programa, de acuerdo con la clasificación del Sistema Nacional de Información de Educación Superior — SNIES del Ministerio de Educación Nacional. (ICFES, 2019)

7. Aplicación metodología KDD

7.1 Selección de datos

Para el estudio se usaron las bases de datos de los resultados del examen Saber Pro del año 2019. Estas se encuentran en dos archivos con extensión ".csv", que corresponden a los módulos de las pruebas genéricas (las cuales se usan para evaluar a todos los estudiantes), y las pruebas específicas (que dependen del tipo de carrera). Estas bases de datos se pueden descargar en el portal Datos Abiertos suministrados directamente por el ICFES.

La base de datos de pruebas genéricas está compuesta por 260756 entradas de datos y 105 columnas con información que las caracterizan. En la base de datos se encuentra la información personal de los estudiantes, los datos de contacto, datos académicos, socioeconómicos, la información sobre su Institución de Educación Superior (IES), la citación al examen, su resultado en las competencias genéricas e información adicional; discretizada por medio de diferentes tipos de variables.

Por su parte, en la base de datos de las pruebas específicas se encuentran los resultados de los diferentes módulos o subtemas que se evalúan en cada estudiante, dependiendo de su carrera profesional. La base de datos se compone de 401815 entradas en las filas y cinco columnas (el código del estudiante, el código de la prueba específica, el nombre de la prueba específica, el puntaje y el desempeño del estudiante). Para el caso de los estudiantes de ingeniería industrial, se evalúan tres módulos de competencias específicas:

- Formulación de proyectos de ingeniería.
- Diseño de sistemas productivos y logísticos.
- Pensamiento científico - matemáticas y estadística.

7.2 Preprocesamiento de datos

Previo a la utilización de la base de datos en el estudio, es indispensable realizar un proceso de filtrado, selección y limpieza de datos porque, a pesar de que la información del Icfes es totalmente confiable, esta posee la información general de todas las carreras, información que no es de utilidad para este estudio y datos erróneos o nulos que pueden alterar los resultados del análisis. Más adelante, se explicará de manera específica los criterios usados para la depuración de la información.

Para el proceso de filtrado de datos, el manejo de la base de datos, los cálculos y la posterior visualización de resultados; se usará el lenguaje de programación Python sobre la interfaz Jupyter Notebook.

En este proyecto se usan librerías de Python muy conocidas que permiten manejar grandes cantidades de datos organizados en tablas, listas y matrices, así como librerías para cálculo estadístico y Machine Learning. Las librerías que sirvieron como columna vertebral en el desarrollo de este proyecto fueron Pandas, Numpy y Scikit-Learn.

Como se mencionó previamente, la información de interés para este estudio se encuentra distribuida en dos archivos de bases de datos. Es por esta razón que, antes del proceso de filtrado y limpieza, es de utilidad tener la información de ambas bases de datos en una misma. Para hacerlo, los resultados de las pruebas genéricas ("*database*") y la de las pruebas específicas ("*database1*"), se unieron a través de la variable "*ESTU_CONSECUTIVO*", que es el código público del estudiante.

Teniendo definida la base de datos completa, es posible emplear filtros y limpieza sobre los datos, para obtener la fracción de la base de datos que se usará en el estudio. A continuación,

se describirán los filtros usados y los ajustes hechos, de modo que, no existan datos erróneos o nulos que afecten los resultados del análisis.

En primer lugar, se procede a eliminar datos que no serán usados en el estudio usando el método "*drop*" de Pandas, entre ellos:

- Los programas académicos diferentes al de ingeniería industrial.
- Los datos de los estudiantes que realizaron las pruebas en el exterior porque no realizaron el examen de pruebas específicas.
- Los datos que corresponden a las pruebas aplicadas en años diferentes al 2019 y que no corresponden al segundo periodo de evaluación en el año.
- Los casos que están en proceso de investigación en el Icfes.

También, se eliminan los puntajes de las pruebas genéricas y específicas que tienen valores nulos y, por ende, serán descartadas del estudio. Por último, se eliminan las columnas de las variables que corresponden principalmente a información general sin variabilidad de los estudiantes, la información de la IES y su programa académico, entre otros, ya que no son necesarias para los objetivos del estudio y posteriormente no se usarán en ningún proceso.

7.3 Transformación de datos

Después del preprocesamiento la base de datos contiene toda la información que es relevante para el estudio, sin embargo, la representación de algunas variables puede no ser eficiente y por esto, una mejora en la calidad de los datos y en su representación, es necesaria para reducir la dimensionalidad y simplificar la tabla de datos (Timarán y cols., 2016).

Se inicia con la codificación de las variables que pueden tener sólo dos tipos de valores, en una representación numérica binaria (de ceros y unos). A modo de ejemplo, la variable *"ESTU_GENERO"* que identifica el género del estudiante, se representa por medio de M o F, para

indicar si es masculino o femenino. Luego de la codificación, se representarán con un 0 o un 1. Lo mismo aplica para la variable *"ESTU_AREARESIDE"* y *"ESTU_PAGOMATRICULABECA"*, cuyos valores eran "Área rural" o "Cabecera municipal" y "Si" o "No", respectivamente; así como para las variables restantes de este tipo.

En el caso de las variables categóricas ordinales, que pueden tener más de dos tipos de valores, es conveniente realizar una representación numérica ordinal que conserve su significado. A continuación, se presentará la lista de variables que fueron codificadas para que sus posibles valores fueran representados por valores numéricos:

- *"ESTU_VALORMATRICULAUNIVERSIDAD"*
- *"ESTU_COMOCAPACITOEXAMENSB11"*
- *"ESTU_CURSOIESAPOYOEXTERNO"*
- *"ESTU_CURSOIESEXTERNA"*
- *"ESTU_CURSODOCENTESIES"*
- *"FAMI_EDUCACIONMADRE"*
- *"FAMI_EDUCACIONPADRE"*
- *"FAMI_ESTRATOVIVIENDA"*
- *"FAMI_CUANTOSCOMPARTEBAÑO"*
- *"FAMI_TRABAJOLABORPADRE"*
- *"FAMI_TRABAJOLABORMADRE"*
- *"ESTU_HORASSEMANATRABAJA"*

Por último, se transforman las variables nominales (*"INST_CARACTER_ACADEMICO", "INST_NOMBRE_INSTITUCION" y "INST_ORIGEN"*) en variables numéricas mediante el procedimiento denominado "Mean Encoding" (se crea una relación monótona entre la variable codificada y la variable de estudio "Puntaje Obtenido").

El procedimiento Mean Encoding permite la captura de la importancia de cada etiqueta de la variable categórica, por lo que se obtienen las características más predictivas.

Con respecto a la variable "*ESTU_INST_MUNICIPIO"* la transformación tiene en cuenta la frecuencia relativa y se usa para este caso el procedimiento denominado "Frequency Encoding", que representa la cantidad de estudiantes en cada municipio donde se ubica la IES.

Una vez codificadas las variables categóricas binarias, ordinales y nominales en variables numéricas, se da fin a la fase de transformación.

7.4 Análisis exploratorio

Inicialmente, se hallan las correlaciones de las variables con respecto al puntaje global. Al observar la figura 3 se concluye que:

La competencia de lectura crítica es la que mayor relación guarda con el puntaje global (0,81), lo cual indica que una buena comprensión lectora es el primer factor determinante en el éxito de un estudiante en las pruebas.

Por otro lado, la competencia de comunicación escrita es la que menor relación guarda con el puntaje global (0,41), incluso por debajo de una variable que no hace parte del cálculo del puntaje como lo es la Institución de Educación superior del estudiante (0,63), esto puede ser debido a que es la única competencia que no tiene una serie de respuestas correctas y es calificada por una persona con base en unos criterios establecidos.

Figura 3

Correlaciones de las variables con respecto al puntaje global

```
correlaciones
```

```
GLOBAL_PUNT                              1.000000
MOD_LECTURA_CRITICA_PUNT                 0.813600
MOD_RAZONA_CUANTITAT_PUNT                0.794702
MOD_COMPETEN_CIUDADA_PUNT                0.786644
MOD_MATEMATICAS_ESTADISTICAS_PNAL        0.769047
MOD_SISTEMAS_PRODUCTIVOS_PNAL            0.767412
MOD_FORMULACION_PROYECTOS_PNAL           0.747672
MOD_INGLES_PUNT                          0.741177
INST_NOMBRE_INSTITUCION_CAT              0.631911
MOD_COMUNI_ESCRITA_PUNT                  0.407428
INST_CARACTER_ACADEMICO_CAT              0.283407
FAMI_ESTRATOVIVIENDA_CAT                 0.274359
FAMI_EDUCACIONMADRE_CAT                  0.273063
FAMI_EDUCACIONPADRE_CAT                  0.268232
ESTU_PAGOMATRICULAPROPIO_CAT             0.243170
ESTU_VALORMATRICULAUNIVERSIDAD_CAT       0.223246
ESTU_HORASSEMANATRABAJA_CAT              0.200985
FAMI_TIENEMOTOCICLETA_CAT                0.200935
FAMI_TRABAJOLABORPADRE_CAT               0.178106
FAMI_TIENEHORNOMICROOGAS_CAT             0.175292
INST_ORIGEN_CAT                          0.172365
ESTU_AREARESIDE_CAT                      0.153602
FAMI_TIENECONSOLAVIDEOJUEGOS_CAT         0.153097
ESTU_PAGOMATRICULAPADRES_CAT             0.151804
FAMI_TIENEAUTOMOVIL_CAT                  0.144066
ESTU_PAGOMATRICULABECA_CAT               0.141554
FAMI_CUANTOSCOMPARTEBAÑO_CAT             0.137142
ESTU_PAGOMATRICULACREDITO_CAT            0.136541
FAMI_TRABAJOLABORMADRE_CAT               0.133032
FAMI_TIENECOMPUTADOR_CAT                 0.124198
FAMI_TIENEINTERNET_CAT                   0.123783
FAMI_TIENELAVADORA_CAT                   0.099015
ESTU_COMOCAPACITOEXAMENSB11_CAT          0.089875
ESTU_INST_MUNICIPIO_CAT                  0.075754
FAMI_TIENESERVICIOTV_CAT                 0.028816
ESTU_GENERO_CAT                          0.027445
Name: GLOBAL_PUNT, dtype: float64
```

Posteriormente, se procede a examinar con mayor detalle si el carácter académico y el origen de la institución producen diferencias estadísticamente significativas sobre el puntaje global cuando se tiene en cuenta la variación causada por el nombre de la institución, cabe recordar que una institución tiene únicamente un carácter académico y un origen asociado, esta prueba se hace por medio de un análisis ANOVA en la figura 4.

Figura 4

Análisis ANOVA de las variables carácter académico y origen de la institución

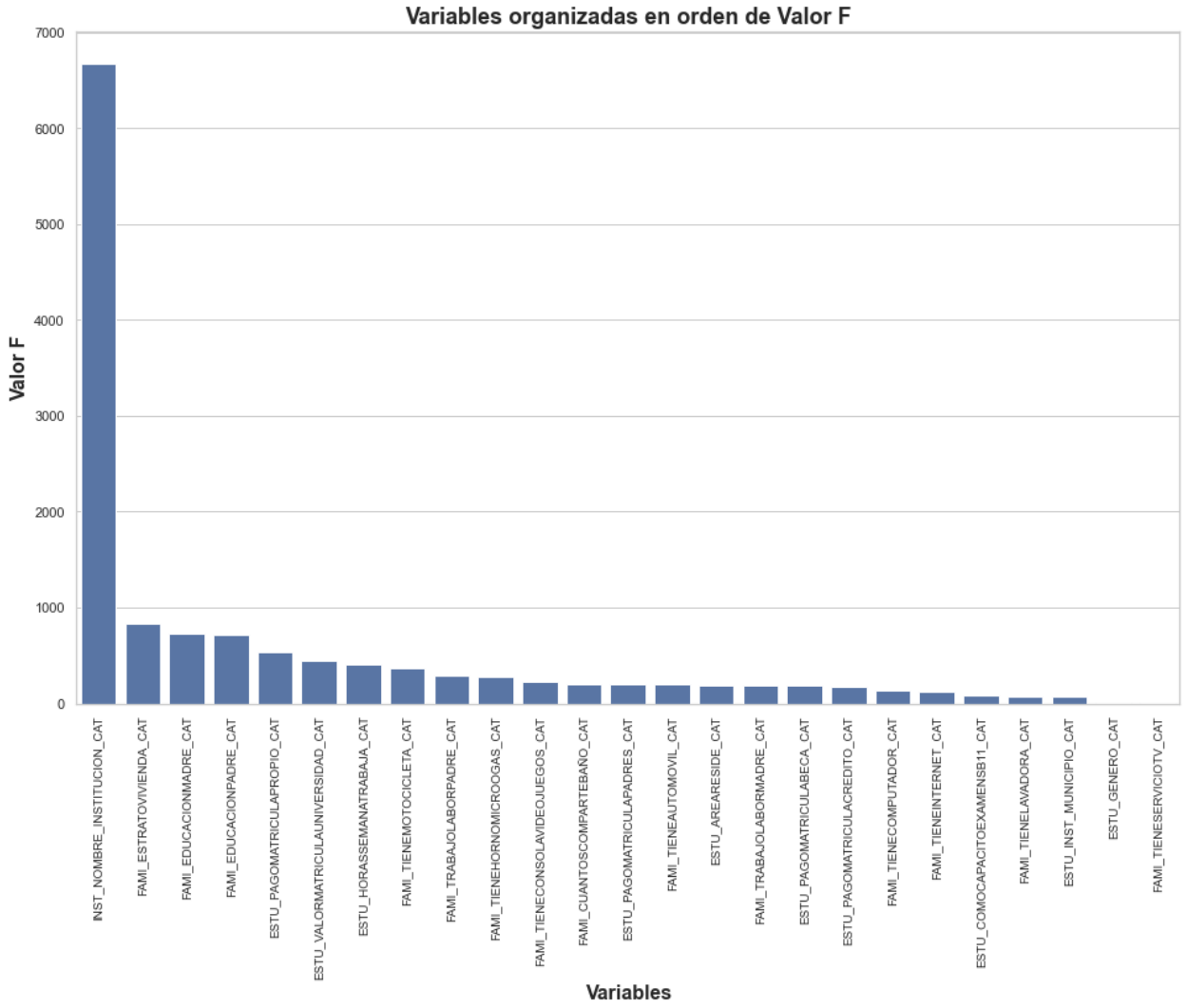
	df	sum_sq	mean_sq	F	PR(>F)
C(INST_NOMBRE_INSTITUCION_CAT)	82.0	1.780985e+06	21719.335007	80.250781	0.000000
C(INST_CARACTER_ACADEMICO_CAT)	3.0	1.529445e+03	509.814954	1.883716	0.129961
C(INST_ORIGEN_CAT)	4.0	4.368739e+02	109.218472	0.403551	0.806230
C(INST_NOMBRE_INSTITUCION_CAT):C(INST_CARACTER_ACADEMICO_CAT)	246.0	5.482609e+04	222.870277	0.823483	0.978851
C(INST_NOMBRE_INSTITUCION_CAT):C(INST_ORIGEN_CAT)	328.0	8.297833e+04	252.982725	0.934746	0.792250
C(INST_CARACTER_ACADEMICO_CAT):C(INST_ORIGEN_CAT)	12.0	2.015907e+03	167.992283	0.620715	0.826522
C(INST_NOMBRE_INSTITUCION_CAT):C(INST_CARACTER_ACADEMICO_CAT):C(INST_ORIGEN_CAT)	984.0	2.877703e+05	292.449442	1.080572	0.048981
Residual	8357.0	2.261766e+06	270.643285	NaN	NaN

Se observa de la tabla ANOVA que las dos variables (Carácter académico y origen de institución) no son significativas, sin embargo, existe una interacción de orden 3 significativa, que probablemente se deba a la fortaleza de la variable nombre de la institución. (Carácter académico valor-p = 0,13 y Origen institución valor-p = 0,81, $\alpha = 0,05$). En adelante sólo se considerará como variable de interés el nombre de la institución.

Con el fin de reducir el número de variables de estudio y facilitar el análisis, se aplica una técnica de selección de características; La selección de características univariantes; esta funciona mediante la selección de las mejores características basadas en pruebas estadísticas, en este caso se tomó en cuenta el valor F y el coeficiente de correlación de Spearman, estas estiman el grado de dependencia lineal entre dos variables aleatorias. Inicialmente se genera un diagrama de barras, en el cual se puede observar una comparación entre las variables y el valor f, en la figura 5.

Figura 5

Variables organizadas en orden de Valor F



Además, se calcula el coeficiente de correlación de Spearman a cada uno de los valores de la columna características. Estos valores se almacenan en una columna que se designa con el nombre *"Spearman"*. Como se puede visualizar en la figura 6.

Figura 6

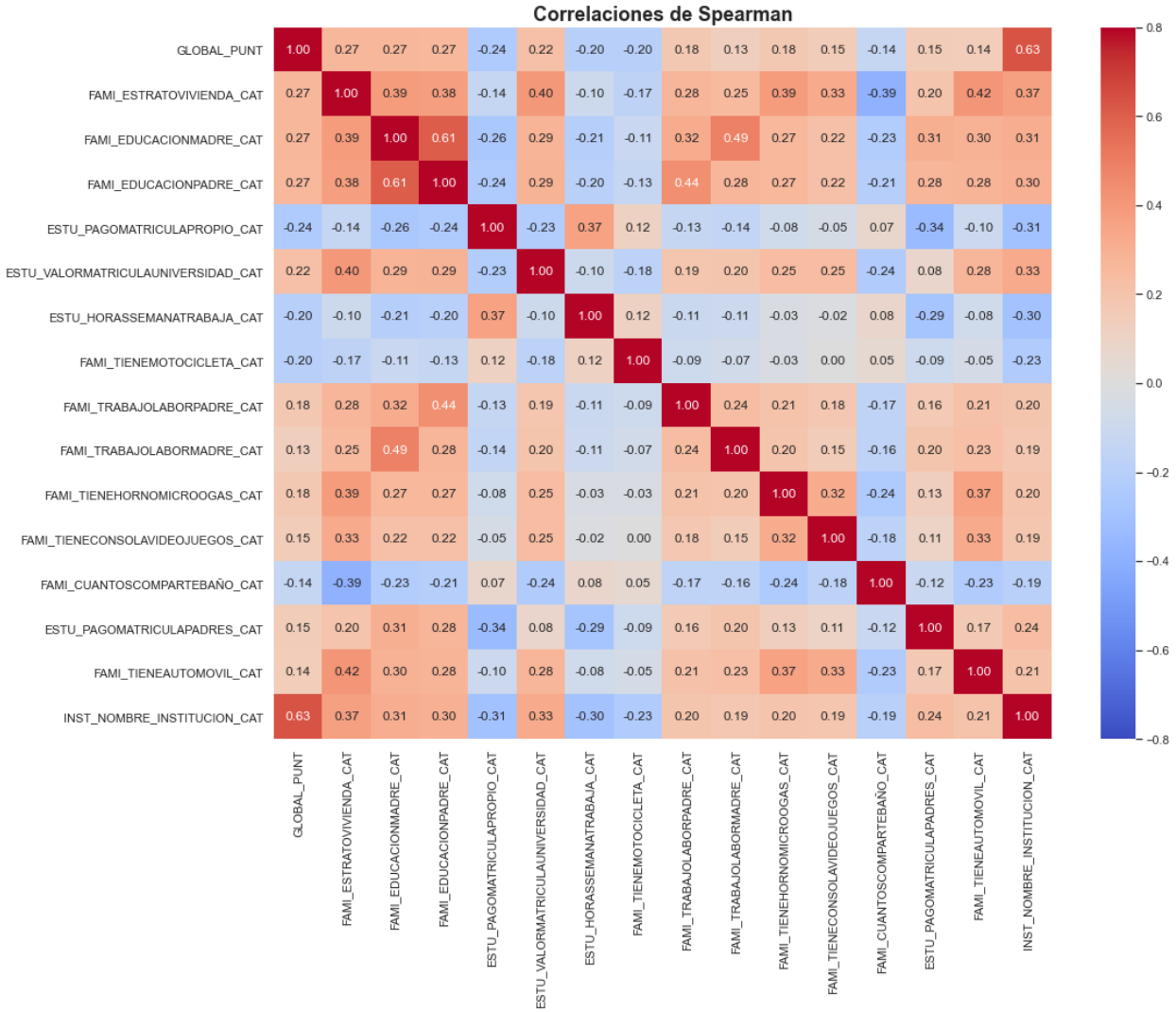
Selección de características (Valor F y coeficiente de correlación de Spearman)

	Caracteristicas	Valor F	Spearman
0	INST_NOMBRE_INSTITUCION_CAT	6676.757153	0.631911
1	FAMI_ESTRATOVIVIENDA_CAT	832.173714	0.274359
2	FAMI_EDUCACIONMADRE_CAT	725.750842	0.273063
3	FAMI_EDUCACIONPADRE_CAT	710.348987	0.268232
4	ESTU_PAGOMATRICULAPROPIO_CAT	538.459396	-0.243170
5	ESTU_VALORMATRICULAUNIVERSIDAD_CAT	446.238119	0.223246
6	ESTU_HORASSEMANATRABAJA_CAT	411.203777	-0.200985
7	FAMI_TIENEMOTOCICLETA_CAT	369.584505	-0.200935
8	FAMI_TRABAJOLABORPADRE_CAT	288.510905	0.178106
9	FAMI_TIENEHORNOMICROOGAS_CAT	273.398869	0.175292
10	FAMI_TIENECONSOLAVIDEOJUEGOS_CAT	223.072007	0.153097
11	FAMI_CUANTOSCOMPARTEBAÑO_CAT	204.399286	-0.137142
12	ESTU_PAGOMATRICULAPADRES_CAT	199.271425	0.151804
13	FAMI_TIENEAUTOMOVIL_CAT	195.405486	0.144066
14	ESTU_AREARESIDE_CAT	191.641838	0.153602
15	FAMI_TRABAJOLABORMADRE_CAT	191.326781	0.133032
16	ESTU_PAGOMATRICULABECA_CAT	183.750938	0.141554
17	ESTU_PAGOMATRICULACREDITO_CAT	168.335107	-0.136541
18	FAMI_TIENECOMPUTADOR_CAT	129.765898	0.124198
19	FAMI_TIENEINTERNET_CAT	127.168641	0.123783
20	ESTU_COMOCAPACITOEXAMENSB11_CAT	81.420243	-0.089875
21	FAMI_TIENELAVADORA_CAT	77.362005	0.099015
22	ESTU_INST_MUNICIPIO_CAT	74.086617	0.075754
23	ESTU_GENERO_CAT	9.819786	0.027445
24	FAMI_TIENESERVICIOTV_CAT	7.767877	0.028816

Se redujeron las variables de estudio de 25 a 15. De las 15 variables seleccionadas, 14 de ellas hacen parte de características socioeconómicas del estudiante, por tanto, se crea una nueva base de datos con el nombre de "Socioeconómicas" y se procede a examinar las relaciones que guardan estas variables de estudio bajo el criterio de correlación de Spearman, como se observa en la figura 4, con el fin de verificar que no exista multicolinealidad entre las variables y encontrar información útil.

Figura 7

Correlaciones de Spearman



Después de analizar las correlaciones de Spearman se puede concluir que:

- La mayor correlación que posee el puntaje global es la que guarda con la institución educativa (0,63) seguido del estrato de vivienda (0,27), existe un gran contraste entre la primera y la segunda, lo que indica que seguramente la predicción del puntaje global estará gobernada por el nombre de la institución en primera instancia.

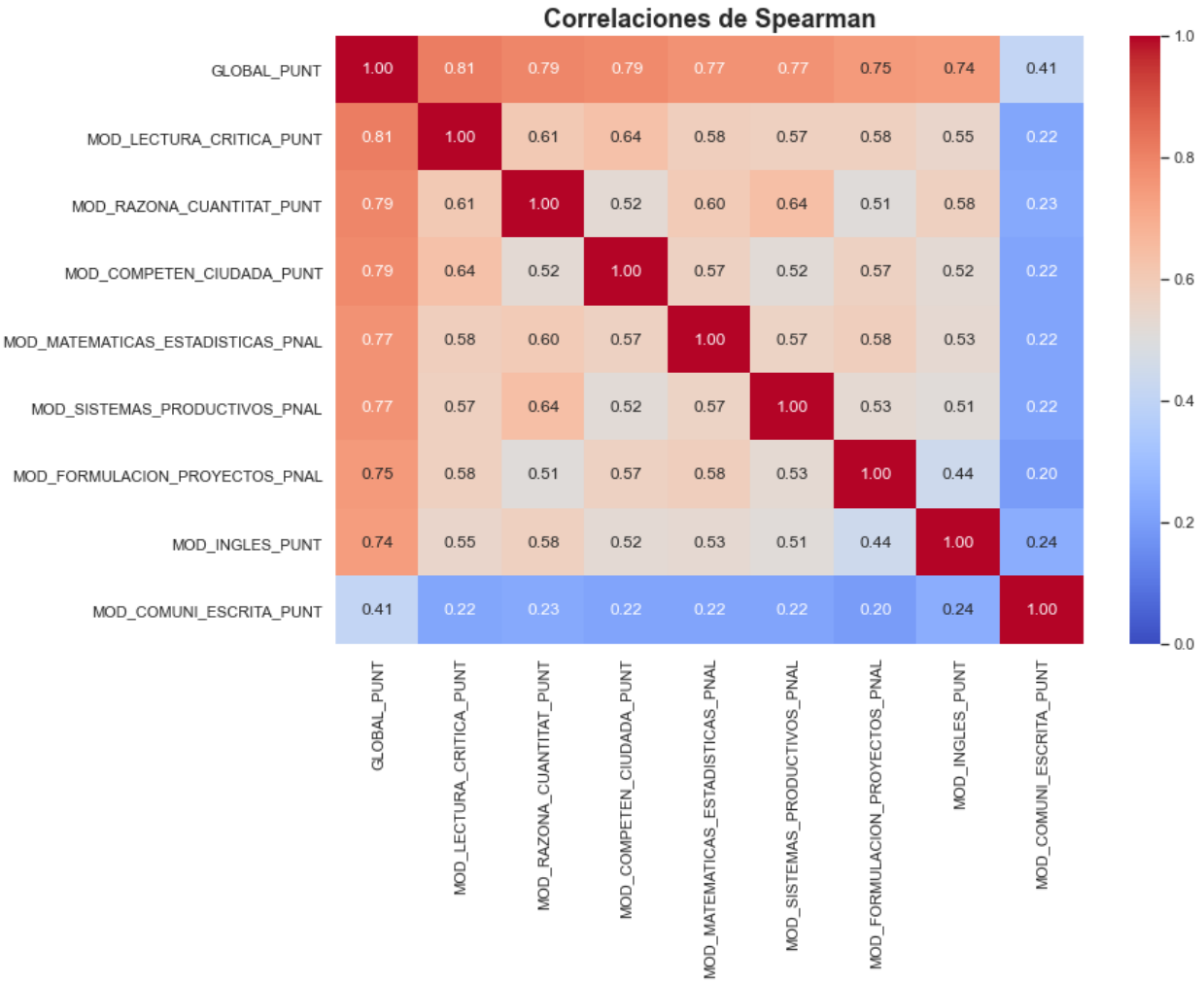
- Existe una correlación media entre el estrato de vivienda y la institución educativa, indicando que, en las instituciones educativas con mayor puntaje, se encuentra una mayor proporción de personas con altos estratos, lo que hace pensar que la alta correlación entre el estrato y el puntaje global sea causa de esta observación.
- Existe una correlación negativa ligera entre el puntaje global y las siguientes variables; el pago de matrícula propio, las horas que el estudiante trabaja a la semana, la posesión de motocicleta y el número de personas con las que comparte baño. De las dos primeras variables, se puede inferir que la calidad de la educación del estudiante se ve negativamente afectada por la disminución de las horas disponibles que este posee para su estudio, independiente fuera del salón de clases, sin embargo, para la tercera cabe recalcar que correlación no significa causación, y si en este caso es cuestión de causación permanece un interrogante, ¿cómo la posesión de una motocicleta afecta el desempeño académico?, y por último, la cuarta es una reflexión del estrato socioeconómico de la persona, que como vemos, guarda la segunda correlación más alta con el puntaje global.
- La posesión de motocicleta guarda una ligera correlación negativa con el estrato socioeconómico, los niveles educativos de los padres y la institución de educación superior. Pese a ser un bien económico parece que la motocicleta es un indicador de marginalidad, eso explicaría su relación con el puntaje global previamente discutido.
- El estrato socioeconómico guarda una relación moderada con el coste de la matrícula, la tenencia de electrodomésticos y automotores, como era de esperarse.

- Hay una correlación media entre el nivel de educación de la madre y la educación del padre indicando que es más probable que personas del mismo nivel educativo tengan hijos.
- El estrato socioeconómico guarda una correlación media con el nivel educativo del padre y de la madre al igual que con sus niveles de empleo, al igual que, como era de esperarse, el nivel educativo tanto del padre como de la madre guarda también una correlación con sus niveles de empleo.

También se realiza la correlación de Spearman para los diferentes módulos que componen el examen, como se puede observar en la figura 8.

Figura 8

Correlaciones de Spearman de los módulos evaluados



De la figura anterior se puede concluir que:

- El puntaje global guarda una fuerte relación con las competencias evaluadas (superior a 0,74), es decir, conociendo únicamente un puntaje en estas áreas se puede realizar un estimado del puntaje global sin estar tan alejado de la realidad, a excepción del módulo de comunicación escrita (correlación de 0,41), cabe recalcar que esta competencia es la única cuya calificación es realizada por una persona que se basa en algunos criterios

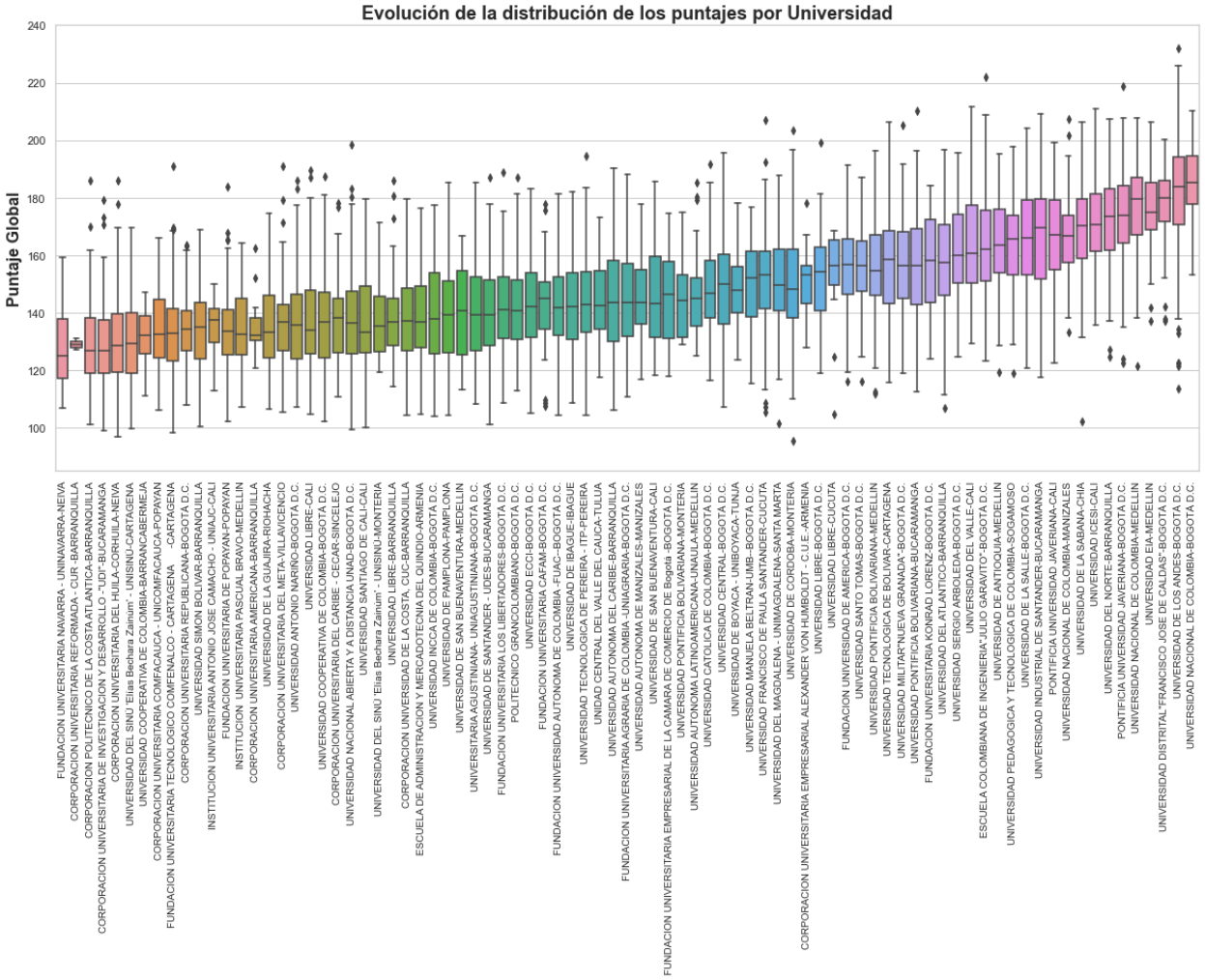
para cuantificar la calidad del texto, se podría asegurar que existe un mayor grado de subjetividad en esta calificación.

- La competencia de comunicación escrita es la que menor guarda relación con las demás competencias.
- El módulo de lectura crítica es la competencia que mayor relación guarda con el puntaje global y a su vez, guarda una relación significativa con las competencias ciudadanas, esto resalta la importancia de un buen nivel de lectura para obtener buenos resultados en las pruebas.

Con el fin de observar la evolución de la distribución de los puntajes por universidades, en la figura 9, se genera una gráfica de cajas cuyos valores en el eje horizontal son todas las instituciones educativas y en el eje vertical se presentan los valores del puntaje global.

Figura 9

Evolución de la distribución de los puntajes por universidades



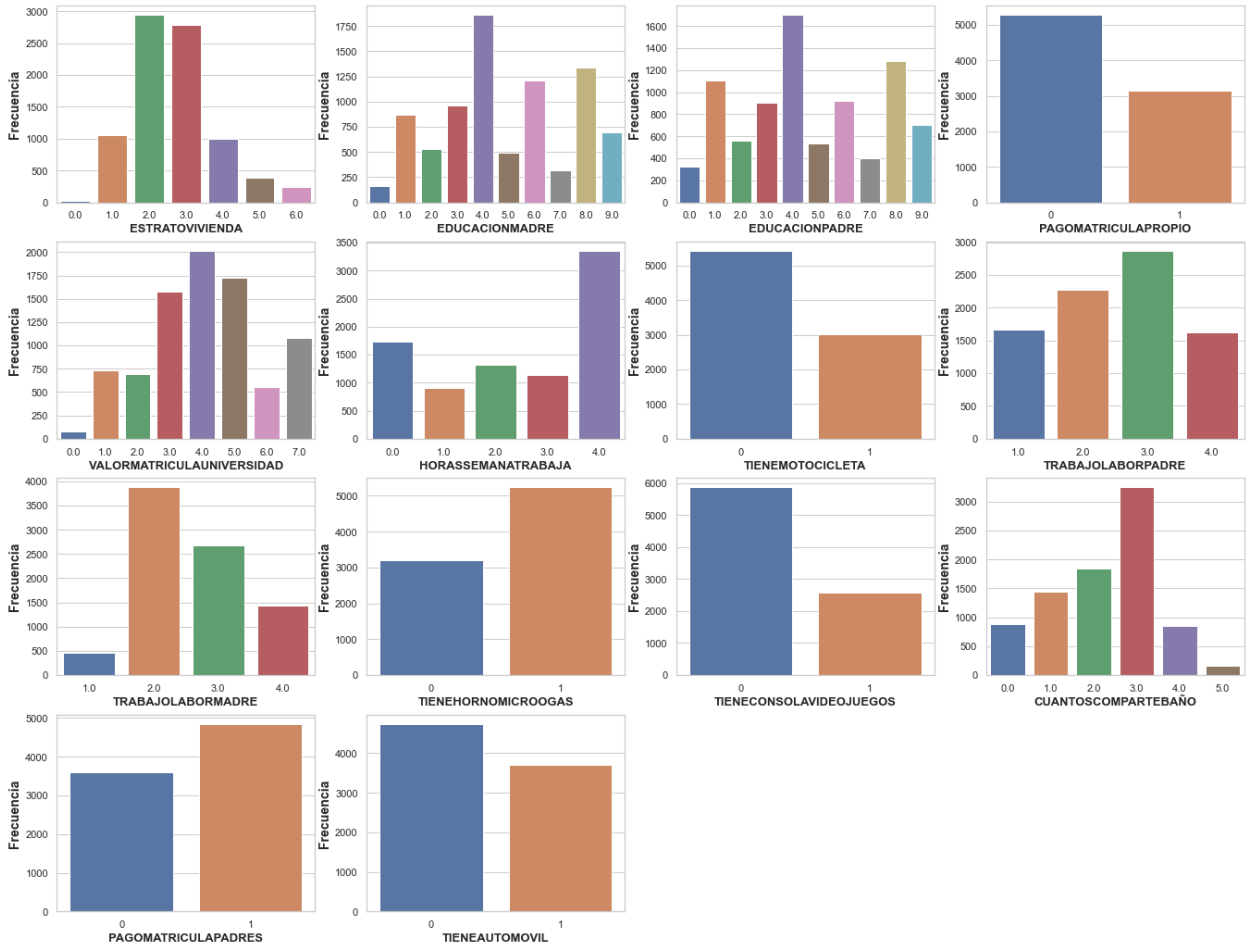
Como era de esperarse, las universidades con mejor desempeño en las pruebas Icfes Saber Pro, son las consideradas “élites” que dominan los rankings de calificaciones de universidades en Colombia, se observa que está liderada por universidades como la Nacional, los Andes, la Distrital, la EIA y la Javeriana. Lo que cabe recalcar es la diferencia abismal y clara que existe entre el desempeño en las instituciones, con las mejores universidades con una mediana superior a 180 y

las peores con una mediana que ronda a 125, teniendo en cuenta que la distribución de los puntajes tiene una desviación estándar de 21.85, es una diferencia de alrededor 2.5 desviaciones estándar.

Por consiguiente, en la figura 10 se observa la distribución univariada de las variables categóricas, con el fin de identificar cómo está compuesta la población.

Figura 10

Distribución de las variables categóricas



De la figura anterior, se puede deducir que:

- La mayoría de los estudiantes se encuentran en el estrato 2 y 3, y son muy pocos los que hacen parte del estrato 0, estos datos del estrato 0 dicen muy poco y agregan complejidad al análisis, por tal razón, se decide eliminarlos.
- El nivel más común de educación, tanto para los padres y las madres, es la secundaria completa, seguido de una carrera profesional, las mujeres parecen tener mejores niveles educativos en promedio.
- La mayoría de los estudiantes no pagan su propia matrícula y la mayoría de las matrículas se encuentran en el intervalo de 2.5 a 4 millones, seguido del intervalo de 4 a 5.5 millones, cabe recalcar que hay un número significativo de personas que paga una matrícula 7 millones.
- La mayor parte de los estudiantes trabaja más de 30h a la semana, muy por encima de los demás rubros, es seguido por el intervalo que no trabaja.
- Los padres de los estudiantes se encuentran en trabajos de mayor nivel en comparación con las madres, incluso teniendo en cuenta que las madres tienen ligeramente una mejor educación.
- La mayor parte de los estudiantes comparten baño con 3 a 4 personas, seguido por el intervalo de compartir baño con 2 personas.
- En su mayoría los estudiantes no poseen ni moto, ni carro y tampoco consola de videojuegos. Pero si cuentan con un horno microondas.

7.5 Minería de datos

7.5.1 Descripción del Algoritmo K-Means

El algoritmo K-Means divide un conjunto de N muestras X en K clústeres disjuntos C, cada uno descrito por la media μj de las muestras del clúster. Las medias se denominan comúnmente

los "centroides" de los clústeres; nótese que no son, en general, puntos de $\boldsymbol{X}$, aunque viven en el mismo espacio.

El algoritmo, pretende elegir los centroides que minimicen la **inercia**, o el criterio de **suma de cuadrados dentro del clúster**:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

En términos básicos, el algoritmo sigue los siguientes pasos:

El primer paso, elige los centroides iniciales, siendo el método más básico la elección de muestras del conjunto de datos. Tras la inicialización, K-Means consiste en un bucle entre los otros dos pasos:

El segundo paso asigna cada muestra a su centroide más cercano.

El tercer paso crea nuevos centroides tomando el valor medio de todas las muestras asignadas a cada centroide anterior. Se calcula la diferencia entre los centroides antiguos y los nuevos y el algoritmo repite estos dos últimos pasos hasta que este valor sea inferior a un umbral. En otras palabras, se repite hasta que los centroides no se muevan significativamente.

7.5.2 Aplicación del algoritmo K-Means

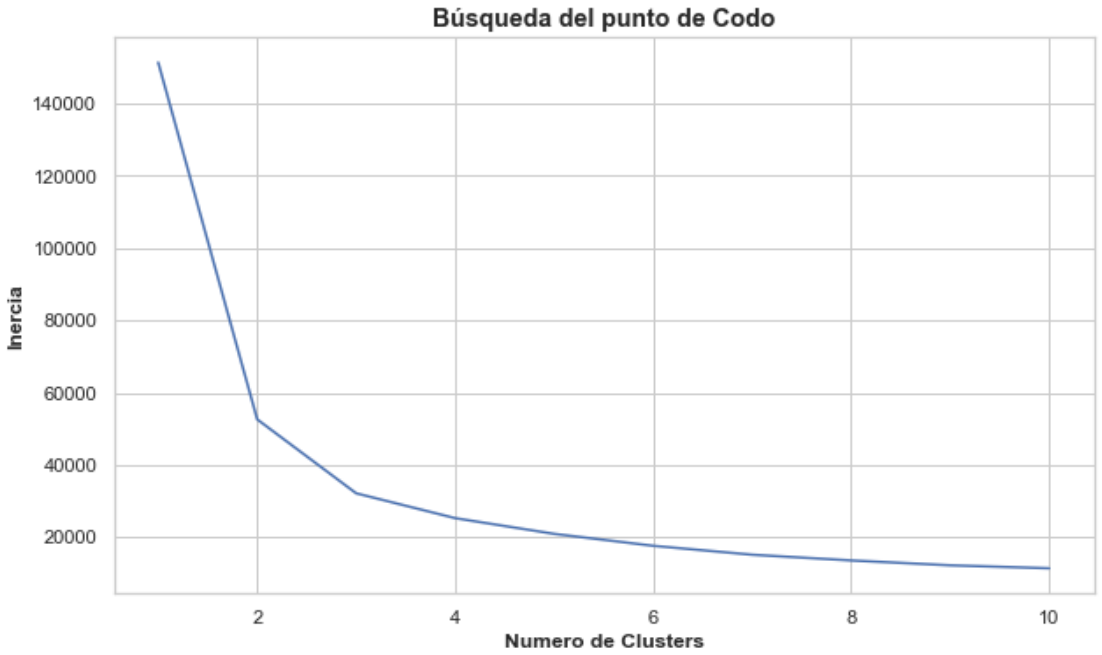
Se procede a aplicar una técnica de aprendizaje no supervisado llamada *clustering*, que busca agrupar las universidades con rendimientos similares teniendo en cuenta todas las competencias evaluadas por el ICFES.

Inicialmente se agrupa en forma de tabla los datos de los nombres de cada una de las instituciones evaluadas y cada uno de los módulos evaluados en la prueba, en filas y columnas respectivamente, con el fin de mostrar la puntuación media obtenida por cada institución en cada módulo evaluado.

El algoritmo de *clustering* usado fue K-Means, de la librería Scikit-Learn. Para determinar el número de clústeres óptimo se aplica el método del codo, que consiste en seleccionar el punto en el que la pendiente de la gráfica de "inercia vs número de clústeres" comienza a tornarse horizontal. En este caso en concreto, como se observa en la figura 11, el codo se encuentra al analizar 4 clústeres.

Figura 11

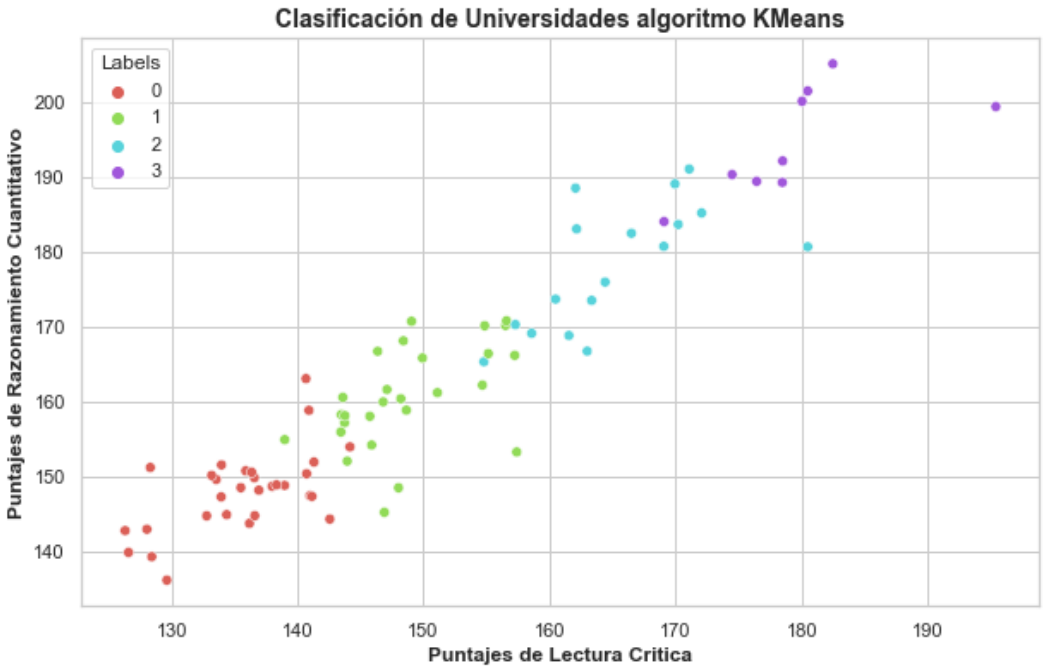
Búsqueda del punto de codo



Posteriormente, se utiliza el algoritmo K-Means para agrupar los datos, teniendo en cuenta que ya se conoce que el número de clústeres debe ser 4. Para comprender la agrupación de los datos en los 4 clústeres, se realiza una gráfica en la que sólo se tienen en cuenta las variables 'Razonamiento cuantitativo' y 'Lectura crítica' y están representados en la figura 12 por diferentes colores para lograr diferenciarlos.

Figura 12

Clasificación de Universidades por el Algoritmo K-Means



Ahora, se despliega una tabla que contiene la media del puntaje obtenido por los diferentes clústeres en cada módulo evaluado en el examen, en donde claramente se pueden apreciar las diferencias en las medias por competencias para los diferentes grupos de universidades, como es posible observar en la figura 13.

Figura 13

Media del puntaje obtenido por cada categoría IES en los diferentes módulos

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
MOD_RAZONA_CUANTITAT_PUNT	148.07	160.62	178.13	194.59
MOD_LECTURA_CRITICA_PUNT	135.71	148.73	165.13	179.50
MOD_COMPETEN_CIUDADA_PUNT	127.32	143.12	157.20	173.05
MOD_INGLES_PUNT	144.81	157.14	173.34	196.31
MOD_COMUNI_ESCRITA_PUNT	143.32	147.40	153.99	163.67
MOD_SISTEMAS_PRODUCTIVOS_PNAL	128.31	138.97	156.54	174.95
MOD_FORMULACION_PROYECTOS_PNAL	133.36	146.25	158.59	169.13
MOD_MATEMATICAS_ESTADISTICAS_PNAL	118.91	127.24	141.95	156.82

Continuando con la exploración de la información dentro de cada clúster obtenido, en la figura 14 se muestra el clúster de universidades clasificadas como "categoría 3" por el algoritmo, es decir, cuyo color es púrpura en la representación gráfica. 6 de estas 9 universidades son privadas y los nombres de estas instituciones generalmente lideran los rankings internacionales.

Figura 14

Clúster o categoría IES 3

```
X[X['Labels'] == 3].index.to_list()
```

```
['PONTIFICIA UNIVERSIDAD JAVERIANA-BOGOTÁ D.C.',
 'UNIVERSIDAD DE LA SABANA-CHIA',
 'UNIVERSIDAD DE LOS ANDES-BOGOTÁ D.C.',
 'UNIVERSIDAD DEL NORTE-BARRANQUILLA',
 'UNIVERSIDAD DISTRITAL"FRANCISCO JOSE DE CALDAS"-BOGOTÁ D.C.',
 'UNIVERSIDAD EIA-MEDELLIN',
 'UNIVERSIDAD ICESI-CALI',
 'UNIVERSIDAD NACIONAL DE COLOMBIA-BOGOTÁ D.C.',
 'UNIVERSIDAD NACIONAL DE COLOMBIA-MEDELLIN']
```

Por otro lado, la UIS se encuentra en la "categoría dos" de las instituciones de educación superior y es una de las que tiene mejor rendimiento dentro de este grupo. La clasificación del clúster dos se puede observar en la figura 15.

Figura 15

Clúster o categoría IES 2

```
X[X['Labels'] == 2].index.to_list()
```

```
['ESCUELA COLOMBIANA DE INGENIERIA"JULIO GARAVITO"-BOGOTÁ D.C.',
 'FUNDACION UNIVERSIDAD DE AMERICA-BOGOTÁ D.C.',
 'FUNDACION UNIVERSITARIA KONRAD LORENZ-BOGOTÁ D.C.',
 'PONTIFICIA UNIVERSIDAD JAVERIANA-CALI',
 'UNIVERSIDAD DE ANTIOQUIA-MEDELLIN',
 'UNIVERSIDAD DE LA SALLE-BOGOTÁ D.C.',
 'UNIVERSIDAD DEL ATLANTICO-BARRANQUILLA',
 'UNIVERSIDAD DEL VALLE-CALI',
 'UNIVERSIDAD INDUSTRIAL DE SANTANDER-BUCARAMANGA',
 'UNIVERSIDAD MILITAR"NUEVA GRANADA"-BOGOTÁ D.C.',
 'UNIVERSIDAD NACIONAL DE COLOMBIA-MANIZALES',
 'UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA-SOGAMOSO',
 'UNIVERSIDAD PONTIFICIA BOLIVARIANA-BUCARAMANGA',
 'UNIVERSIDAD PONTIFICIA BOLIVARIANA-MEDELLIN',
 'UNIVERSIDAD SANTO TOMAS-BOGOTÁ D.C.',
 'UNIVERSIDAD SERGIO ARBOLEDA-BOGOTÁ D.C.',
 'UNIVERSIDAD TECNOLOGICA DE BOLIVAR-CARTAGENA']
```

Dentro de las instituciones de "categoría 1", expuestas en la figura 16, se encuentran los nombres de universidades con menor trayectoria y la mayoría de las fundaciones universitarias

Figura 16

Clúster o categoría IES 1

```
X[X['Labels'] == 1].index.to_list()
```

```
['CORPORACION UNIVERSITARIA EMPRESARIAL ALEXANDER VON HUMBOLDT - C.U.E.-ARMENIA',
 'FUNDACION UNIVERSIDAD AUTONOMA DE COLOMBIA -FUAC--BOGOTÁ D.C.',
 'FUNDACION UNIVERSITARIA AGRARIA DE COLOMBIA -UNIAGRARIA-BOGOTÁ D.C.',
 'FUNDACION UNIVERSITARIA CAFAM-BOGOTÁ D.C.',
 'FUNDACION UNIVERSITARIA EMPRESARIAL DE LA CAMARA DE COMERCIO DE Bogotá -BOGOTÁ D.C.
 'FUNDACION UNIVERSITARIA LOS LIBERTADORES-BOGOTÁ D.C.',
 'POLITECNICO GRANCOLOMBIANO-BOGOTÁ D.C.',
 'UNIDAD CENTRAL DEL VALLE DEL CAUCA-TULUA',
 'UNIVERSIDAD AUTONOMA DE MANIZALES-MANIZALES',
 'UNIVERSIDAD AUTONOMA DEL CARIBE-BARRANQUILLA',
 'UNIVERSIDAD AUTONOMA LATINOAMERICANA-UNAULA-MEDELLIN',
 'UNIVERSIDAD CATOLICA DE COLOMBIA-BOGOTÁ D.C.',
 'UNIVERSIDAD CENTRAL-BOGOTÁ D.C.',
 'UNIVERSIDAD DE BOYACA - UNIBOYACA-TUNJA',
 'UNIVERSIDAD DE CORDOBA-MONTERIA',
 'UNIVERSIDAD DE IBAGUE-IBAGUE',
 'UNIVERSIDAD DE SAN BUENAVENTURA-CALI',
 'UNIVERSIDAD DE SANTANDER - UDES-BUCARAMANGA',
 'UNIVERSIDAD DEL MAGDALENA - UNIMAGDALENA-SANTA MARTA',
 'UNIVERSIDAD ECCI-BOGOTÁ D.C.',
 'UNIVERSIDAD FRANCISCO DE PAULA SANTANDER-CUCUTA',
 'UNIVERSIDAD LIBRE-BOGOTÁ D.C.',
 'UNIVERSIDAD LIBRE-CUCUTA',
 'UNIVERSIDAD MANUELA BELTRAN-UMB--BOGOTÁ D.C.',
 'UNIVERSIDAD PONTIFICIA BOLIVARIANA-MONTERIA',
 'UNIVERSIDAD TECNOLOGICA DE PEREIRA - ITP-PEREIRA',
 'UNIVERSITARIA AGUSTINIANA- UNIAGUSTINIANA-BOGOTÁ D.C.']
```

y, por último, en la figura 17, se encuentran las instituciones de educación superior de "categoría 0", la lista está conformada por corporaciones, fundaciones e instituciones universitarias, junto con algunas universidades menos reconocidas.

Figura 17

Clúster o categoría IES 0

```
X[X['Labels'] == 0].index.to_list()
```

```
['CORPORACION POLITECNICO DE LA COSTA ATLANTICA-BARRANQUILLA',
 'CORPORACION UNIVERSIDAD DE LA COSTA, CUC-BARRANQUILLA',
 'CORPORACION UNIVERSITARIA AMERICANA-BARRANQUILLA',
 'CORPORACION UNIVERSITARIA COMFACAUCA - UNICOMFACAUCA-POPAYAN',
 'CORPORACION UNIVERSITARIA DE INVESTIGACION Y DESARROLLO -"UDI"-BUCARAMANGA',
 'CORPORACION UNIVERSITARIA DEL CARIBE - CECAR-SINCELEJO',
 'CORPORACION UNIVERSITARIA DEL HUILA-CORHUILA-NEIVA',
 'CORPORACION UNIVERSITARIA DEL META-VILLAVICENCIO',
 'CORPORACION UNIVERSITARIA REFORMADA - CUR -BARRANQUILLA',
 'CORPORACION UNIVERSITARIA REPUBLICANA-BOGOTÁ D.C.',
 'ESCUELA DE ADMINISTRACION Y MERCADOTECNIA DEL QUINDIO-ARMENIA',
 'FUNDACION UNIVERSITARIA DE POPAYAN-POPAYAN',
 'FUNDACION UNIVERSITARIA NAVARRA - UNINAVARRA-NEIVA',
 'FUNDACION UNIVERSITARIA TECNOLOGICO COMFENALCO - CARTAGENA      -CARTAGENA',
 'INSTITUCION  UNIVERSITARIA PASCUAL BRAVO-MEDELLIN',
 'INSTITUCION UNIVERSITARIA ANTONIO JOSE CAMACHO - UNIAJC-CALI',
 'UNIVERSIDAD ANTONIO NARIÑO-BOGOTÁ D.C.',
 'UNIVERSIDAD COOPERATIVA DE COLOMBIA-BARRANCABERMEJA',
 'UNIVERSIDAD COOPERATIVA DE COLOMBIA-BOGOTÁ D.C.',
 'UNIVERSIDAD DE LA GUAJIRA-RIOHACHA',
 'UNIVERSIDAD DE PAMPLONA-PAMPLONA',
 'UNIVERSIDAD DE SAN BUENAVENTURA-MEDELLIN',
 "UNIVERSIDAD DEL SINÚ 'Elías Bechara Zainúm' - UNISINÚ-CARTAGENA",
 "UNIVERSIDAD DEL SINÚ 'Elías Bechara Zainúm' - UNISINÚ-MONTERIA",
 'UNIVERSIDAD INCCA DE COLOMBIA-BOGOTÁ D.C.',
 'UNIVERSIDAD LIBRE-BARRANQUILLA',
 'UNIVERSIDAD LIBRE-CALI',
 'UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA UNAD-BOGOTÁ D.C.',
 'UNIVERSIDAD SANTIAGO DE CALI-CALI',
 'UNIVERSIDAD SIMON BOLIVAR-BARRANQUILLA']
```

Una vez revisado los clústeres obtenidos, se crea una tabla donde se puede apreciar el porcentaje de los estratos que conforman cada uno de los grupos de las IES, las filas corresponden a las asignaciones de grupos (clústeres) y en sus columnas aparecen los distintos estratos económicos contemplados anteriormente, como se observa en la figura 18.

Figura 18

Porcentaje de los estratos que conforman cada clúster de universidades

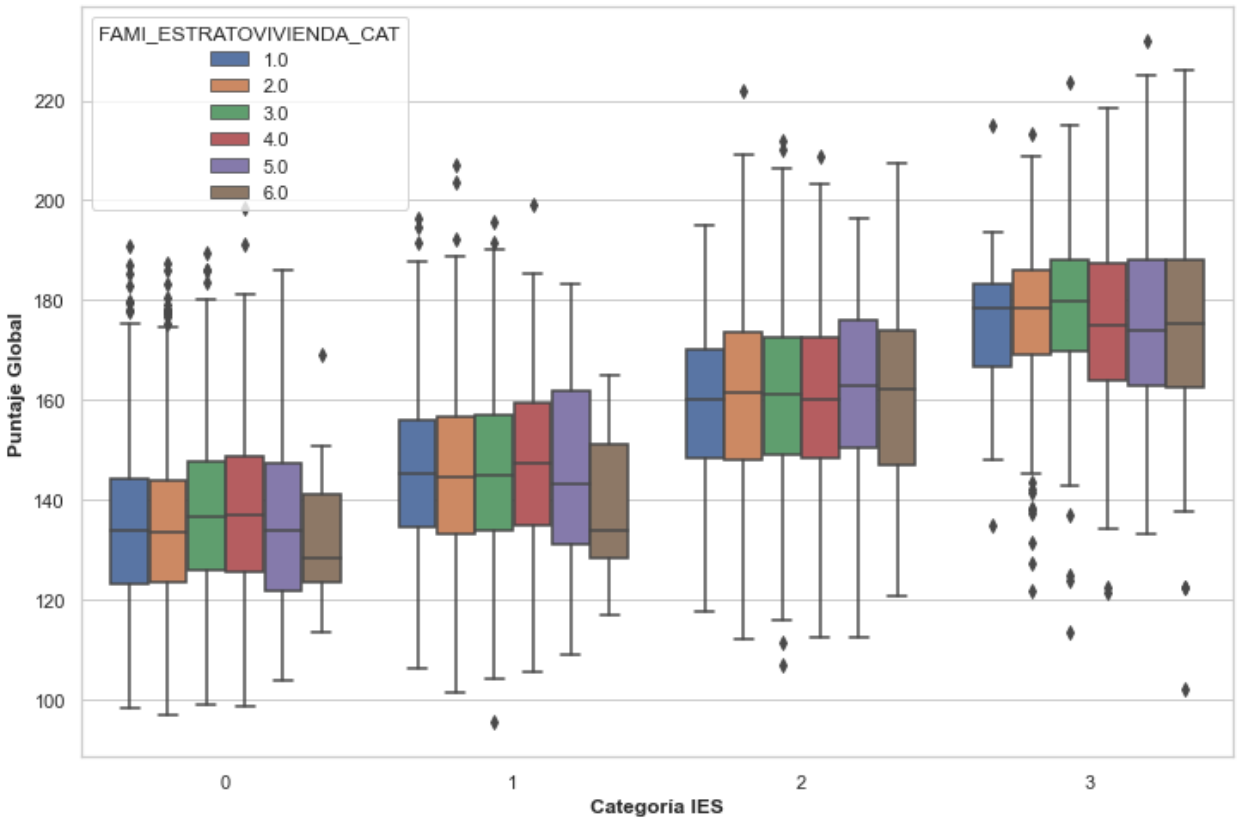
FAMI_ESTRATOVIVIENDA_CAT UNI_GRUPO	1.0	2.0	3.0	4.0	5.0	6.0
0	20.20	44.13	28.68	5.49	1.27	0.22
1	9.44	36.84	41.06	9.48	2.59	0.58
2	8.57	29.15	35.75	18.56	5.65	2.32
3	3.68	13.81	24.48	24.84	17.22	15.96

Si observamos detalladamente, los estratos de vivienda aumentan junto con la categoría de la universidad. El grupo 3 de las IES tiene mayores proporciones de estudiantes estratos 4, 5 y 6, la diferencia contrasta especialmente para el estrato 6, que está conformado en menor manera por estratos inferiores al 3. El grupo 0 está dominado por estudiantes de los estratos 1 , 2 y 3. El grupo 1 y 2 están conformados principalmente por estratos 2 y 3, sin embargo, el grupo 2 tiene una mayor proporción de estudiantes estrato 4. Estos resultados son conformes con el coeficiente de correlación de Pearson previamente calculado, donde se encuentra una correlación media entre el estrato y la institución educativa.

Con el fin de observar la influencia del estrato sobre el puntaje global, se genera una gráfica de cajas cuyos valores en el eje horizontal son los grupos (clústeres) asignados para las diferentes IES, en el eje vertical se presentan los valores del puntaje global y se separan en cajas respecto al estrato de vivienda. Se realiza el mismo procedimiento de la tabla y su respectiva gráfica para las variables 'FAMI_EDUCACIONMADRE_CAT', 'FAMI_EDUCACIONPADRE_CAT', 'ESTU_VALORMATRICULAUNIVERSIDAD_CAT'.

Figura 19

Influencia del estrato de vivienda de cada categoría IES sobre el puntaje global



En la figura 19, no se aprecia una influencia clara del estrato de un estudiante sobre su puntaje global. Como previamente se mencionó, la influencia de la IES es clara, el puntaje aumenta conforme se avanza en el grupo de las IES.

Ahora, se realiza el mismo procedimiento anterior para averiguar cómo afecta la educación de la madre en el puntaje global obtenido. Por ende, se genera una tabla donde las filas corresponden a los diferentes clústeres de universidades y en sus columnas aparecen los distintos niveles de educación de la madre, tal como se detalla en la figura 20.

Figura 20

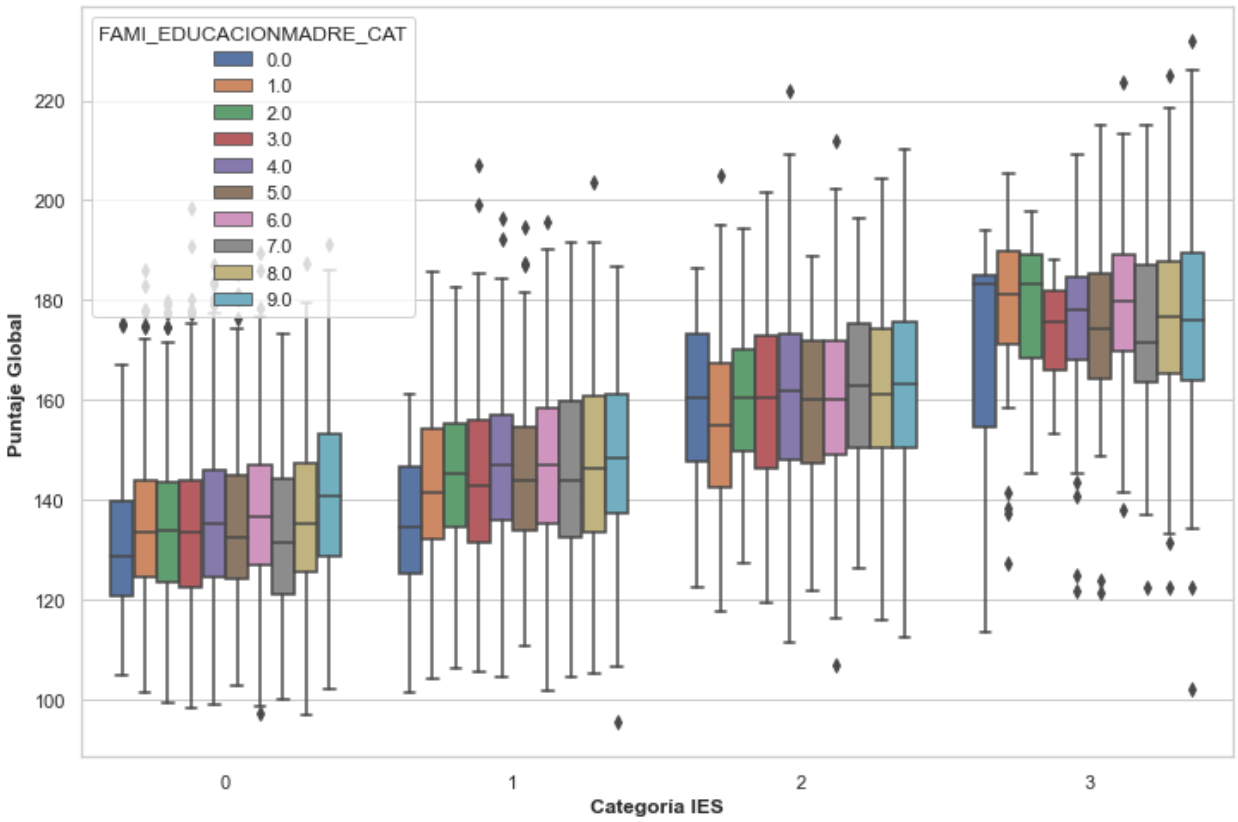
Nivel de educación de la madre de cada categoría IES

FAMI_EDUCACIONMADRE_CAT	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
UNI_GRUPO										
0	2.70	14.43	8.29	14.21	24.24	6.33	14.03	2.76	9.28	3.72
1	2.21	11.45	7.48	12.74	23.98	5.81	14.54	3.26	12.66	5.89
2	0.83	5.71	3.81	8.98	21.95	5.83	16.72	5.41	20.40	10.35
3	0.45	2.69	1.79	3.86	12.11	4.48	11.30	5.20	35.16	22.96

Al observar la tabla anterior, se puede encontrar que conforme se asciende en el grupo de IES, aumentan los niveles educativos de las madres de los estudiantes, llegando hasta 35% de madres profesionales y un 23% de madres con postgrado para el grupo 3 de las instituciones educativas, al igual que en la tabla de los estratos las diferencias no son amplias para los grupos 1 y 2.

Figura 21

Influencia del nivel educativo de la madre de cada categoría IES sobre el puntaje global



Se realiza el gráfico de caja y bigotes para poder visualizar la representación de la influencia del nivel educativo de la madre en cada grupo de universidades sobre el puntaje global, como se ilustra en la fig. 21. Después de analizarlo, se observa una tendencia leve al alza en los puntajes conforme se aumenta en el nivel educativo de la madre, sin embargo, esta tendencia es menos clara conforme ascendemos en el grupo de las instituciones educativas, hasta que en el tercer grupo se hace indistinguible. La influencia del grupo educativo ya fue previamente descrita. También se realiza el mismo procedimiento para la variable que mide el nivel educativo del padre, pero se observan las mismas tendencias y se realizan las mismas observaciones que para el nivel educativo de las madres, es decir, los comportamientos se replican.

Por último, para la variable del valor de la matrícula, el procedimiento es el mismo que se realiza en las anteriores variables analizadas, por ende, en la figura 22, se genera una tabla donde las filas corresponden a los diferentes clústeres de universidades y en sus columnas aparecen los distintos niveles del valor de la matrícula.

Figura 22

Valor de matrícula de la universidad de cada categoría IES

ESTU_VALORMATRICULAUNIVERSIDAD_CAT	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0
UNI_GRUPO								
0	0.40	1.43	10.58	39.48	34.79	12.26	0.93	0.12
1	0.25	6.98	9.57	6.56	31.79	37.43	7.35	0.08
2	2.26	22.84	5.41	5.12	7.26	24.99	18.08	14.04
3	1.70	11.39	2.78	4.66	0.27	0.72	3.59	74.89

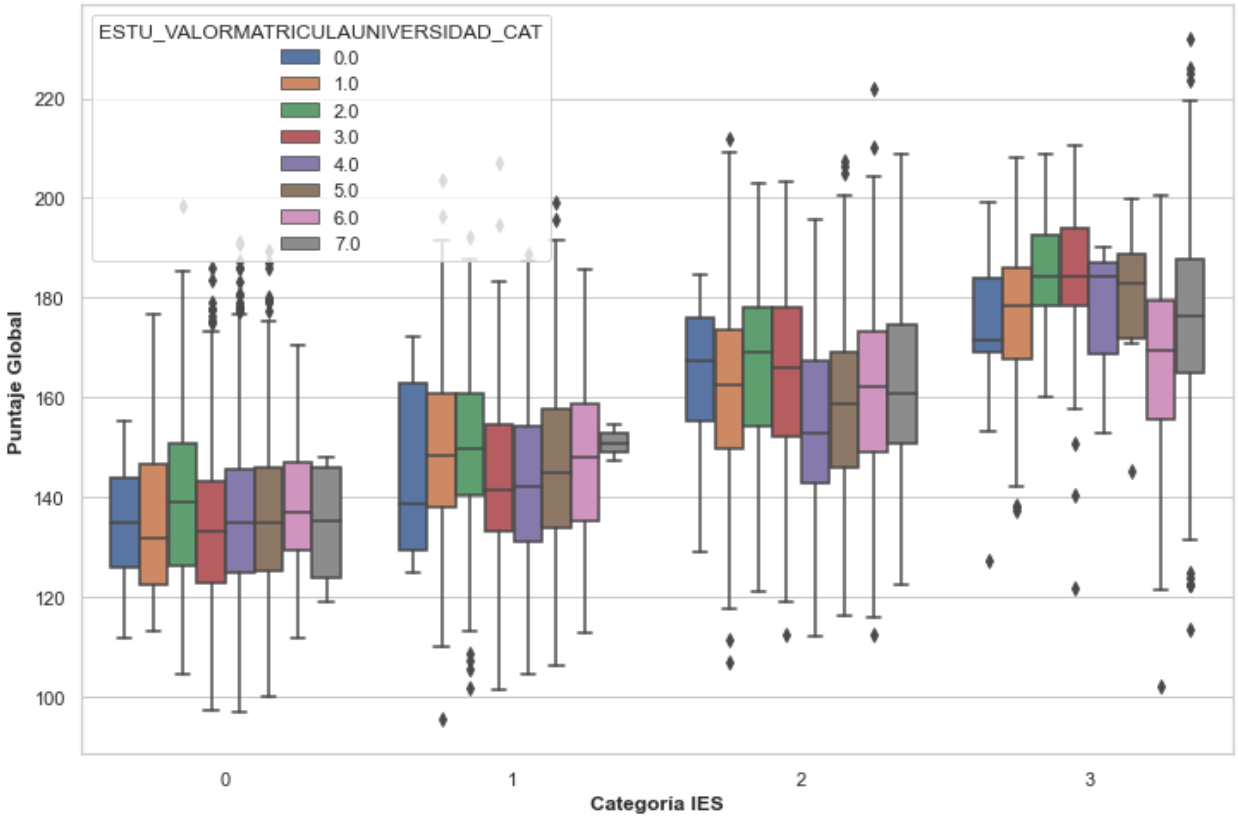
La anterior tabla brinda información interesante para la investigación: El grupo 3 de IES está en su gran mayoría compuesto por estudiantes cuyas matrículas superan los 7 millones de pesos (75%), un número significativo teniendo en cuenta que 3 de las 9 universidades en este grupo son públicas, el 11% siguiente en este grupo son estudiantes que pagan un monto menor a 500 mil pesos, posiblemente estudiantes con becas o de universidades públicas. El grupo 2 de IES está conformado por dos tipos de estudiantes, los que pagan matrículas muy bajas (23% paga por debajo de 500 mil), y otro grupo que paga matrículas por encima de los 4 millones (57%), el primer rubro son estudiantes con becas o de instituciones públicas con bajos ingresos. Para los grupos 0 y 1 se observa que los datos se congregan alrededor de los valores medios de matrícula (entre 1 millón y 4 millones), el grupo cero especialmente alrededor del millón de pesos y el grupo 1 alrededor de los 4 millones. En resumen; en el grupo 3 de las IES, las matrículas se ubican generalmente por encima de los 7 millones, para el grupo 2 se forman dos nichos, uno que paga muy poco dinero y

otro que paga matrículas alrededor de los 5 millones, en el grupo 1 las matrículas están alrededor de los 4 millones, y en el grupo 0 alrededor del millón de pesos.

En cuanto al gráfico de caja y bigotes representado en la figura 23, no se aprecia una influencia clara del coste de la matrícula sobre su puntaje global, se obtienen rendimientos similares dentro del mismo grupo de IES.

Figura 23

Influencia del valor de matrícula de cada categoría IES sobre el puntaje global



Agregando a lo anterior, con el fin de evaluar la significancia estadística de las variables previamente analizadas, se realiza un análisis Anova ilustrado en la figura 24.

Figura 24

Análisis Anova a las variables estrato vivienda, educación padre y madre y valor matrícula

	df	sum_sq	mean_sq	F	PR(>F)
C(UNI_GRUPO)	3.0	1.659588e+06	553195.879119	2013.425842	0.000000e+00
C(FAMI_ESTRATOVIVIENDA_CAT)	5.0	4.784199e+03	956.839861	3.482539	3.804098e-03
C(ESTU_VALORMATRICULAUNIVERSIDAD_CAT)	7.0	3.006743e+04	4295.347449	15.633456	1.661671e-20
C(FAMI_EDUCACIONMADRE_CAT)	9.0	1.660289e+04	1844.765340	6.714255	1.207068e-09
C(FAMI_EDUCACIONPADRE_CAT)	9.0	6.289490e+03	698.832181	2.543487	6.495638e-03
Residual	8378.0	2.301885e+06	274.753541	NaN	NaN

A pesar de no notar diferencias drásticas en las gráficas para algunas variables, según el análisis ANOVA, todas son estadísticamente significativas con un 1% de significancia.

7.5.3 Descripción del Algoritmo K-Modes

El algoritmo k-Modes es usado para agrupar variables categóricas. Define los clústeres basándose en el número de categorías que coinciden entre los puntos de datos. (Esto contrasta con el algoritmo k-Means, más conocido, que agrupa los datos numéricos basándose en la distancia euclidiana).

La causa de que el algoritmo k-Means no pueda agrupar objetos categóricos es su medida de disimilitud y el método utilizado para resolver el problema. Estos obstáculos pueden eliminarse realizando las siguientes modificaciones en el algoritmo k-Means:

1. Utilizando una medida de disimilitud simple para objetos categóricos,

2. Sustituyendo las medias de los clústeres por modos

3. Utilizando un método basado en la frecuencia para encontrar las modas para resolver el problema. La función de inercia pasa a llamarse función de costo.

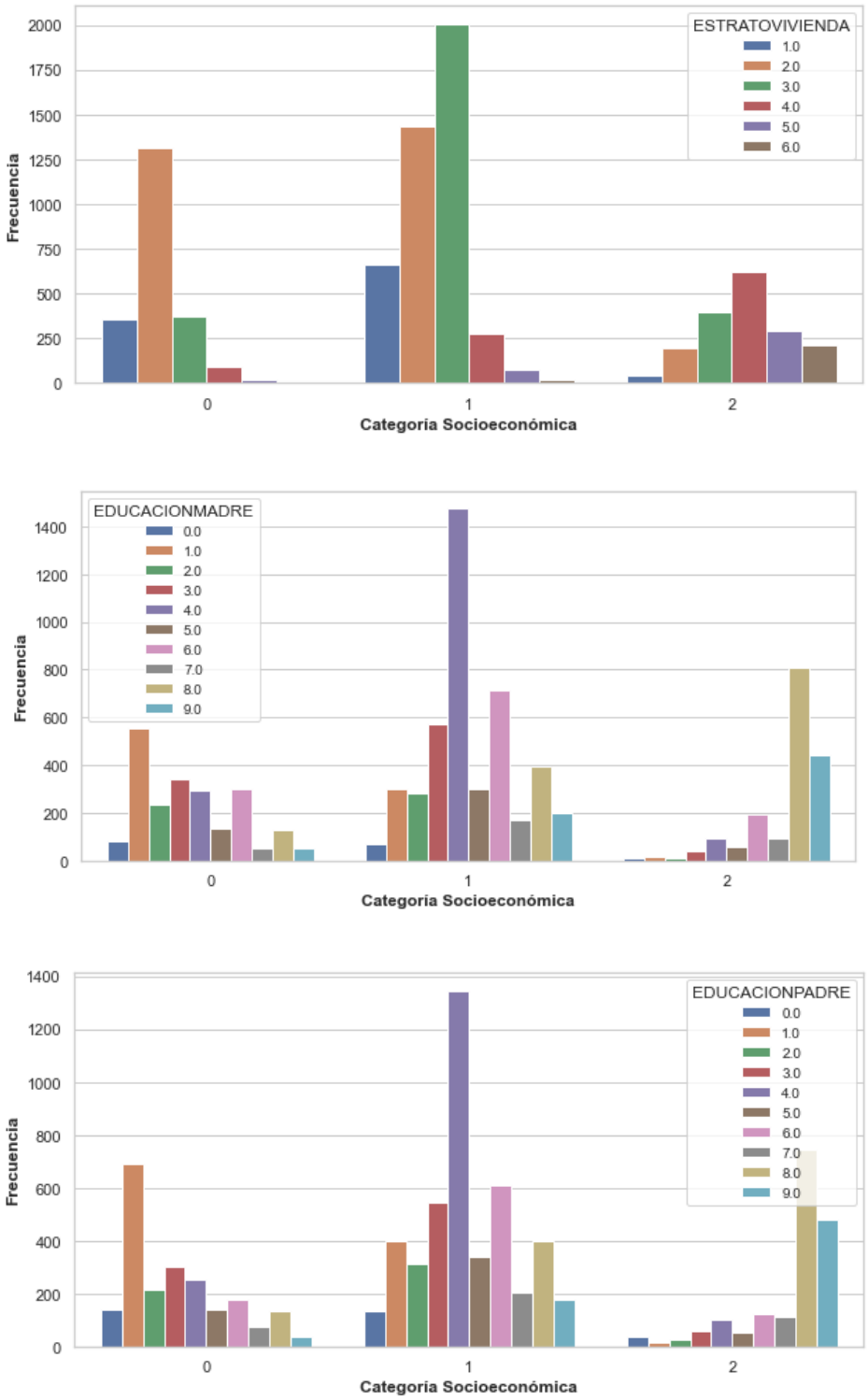
7.5.4 Aplicación del algoritmo K-Modes

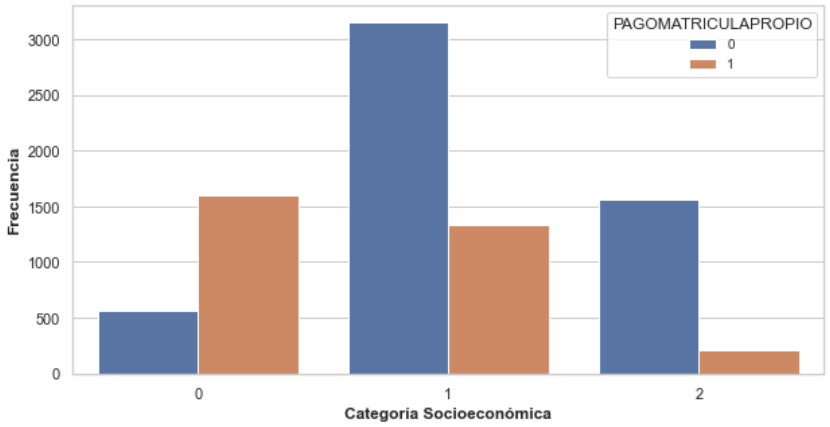
Ahora, con el fin de crear los perfiles tipológicos de los estudiantes de ingeniería industrial, se realiza una clasificación de los mismos por medio de sus variables socioeconómicas, usando aprendizaje no supervisado, por medio del algoritmo K-Modes de la librería K-Modes. Este es un análogo al algoritmo K-Means utilizado anteriormente, su única diferencia radica en que este se utiliza para analizar datos numéricos y no categóricos. Para esto, se agrupan las variables Socioeconómicas y se determina el número de clústeres óptimos encontrando el codo, en este caso en concreto, el codo se encuentra entre el 3 y 5 clúster, se opta por el número 3. Finalmente se realiza el algoritmo K-Modes directamente para el número de clústeres elegidos.

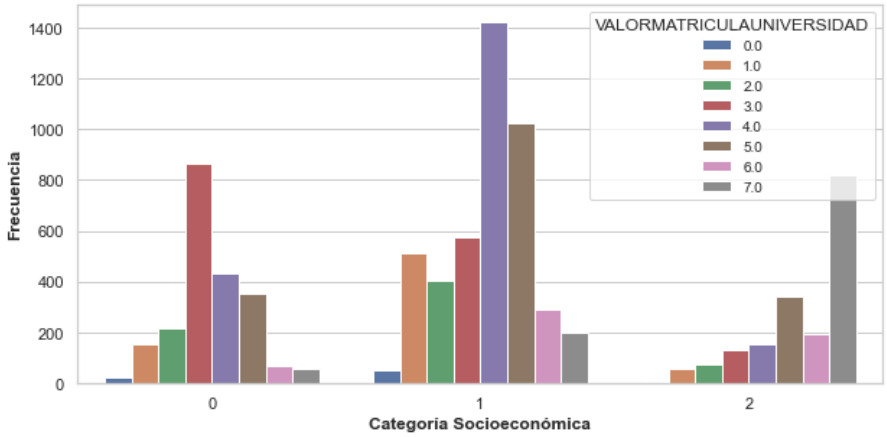
Para visualizar los datos dentro de cada clúster, se genera una gráfica de barras para cada una de las variables que contienen la información socioeconómica de los evaluados; cada gráfica presenta un conteo por clústeres de la cantidad de evaluados pertenecientes a cada variable analizada (Estrato, Estudios Madre, Estudios Padre, etc.). En los siguientes diagramas de barras, el número de clúster es llamado “categoría socioeconómica” para facilidad de análisis, como se puede observar de forma agrupada en la figura 25.

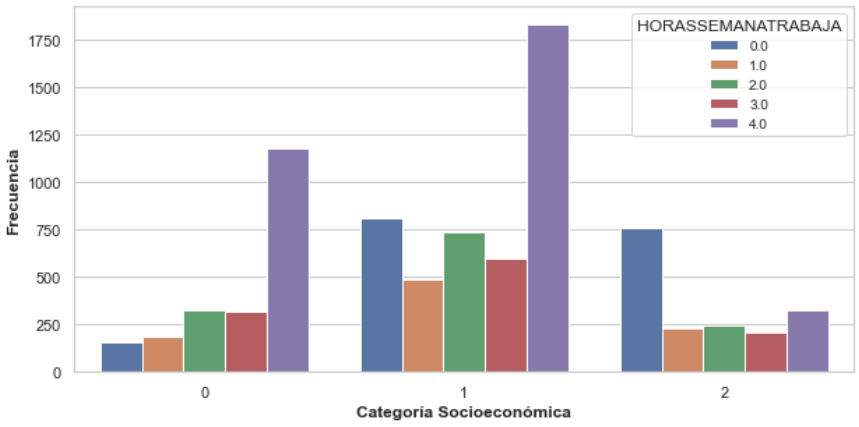
Figura 25

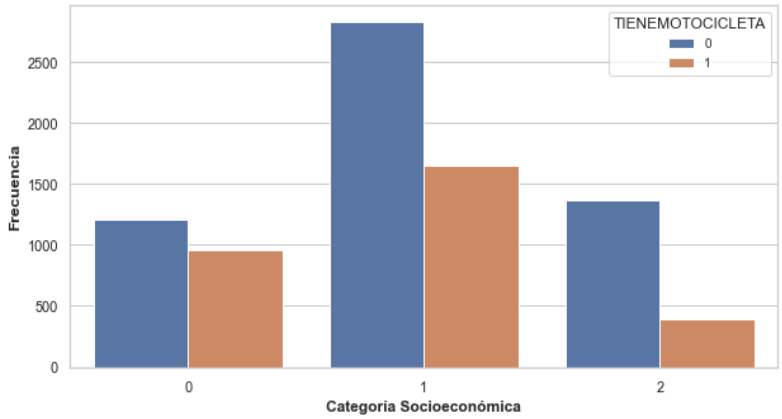
Conteo por clúster de los estudiantes pertenecientes a cada variable socioeconómica analizada

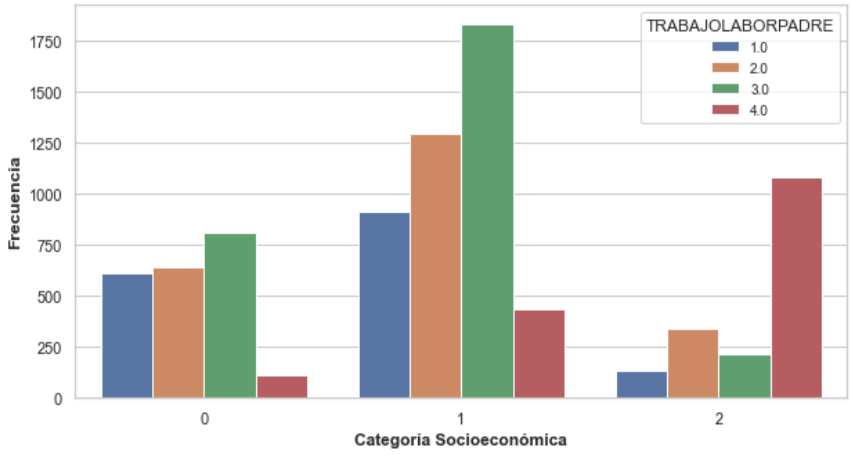


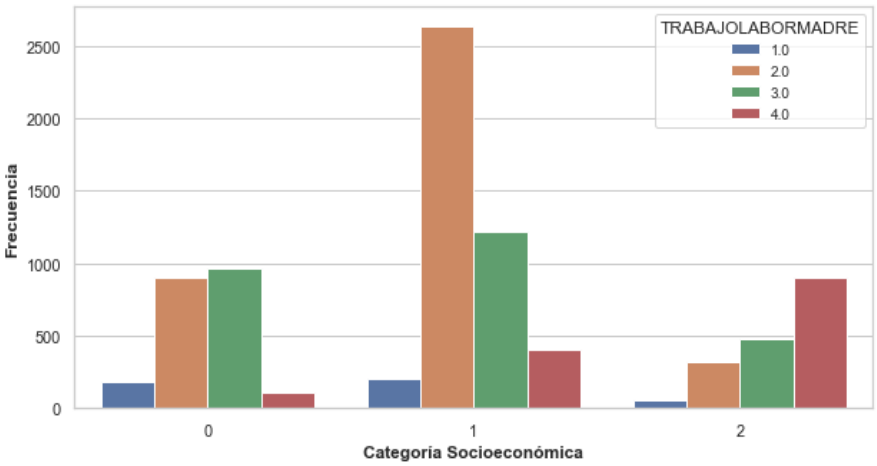

PAGOMATRICULAPROPIO
0
1
Frecuencia
3000
2500
2000
1500
1000
500
0
0
1
2
Categoría Socioeconómica

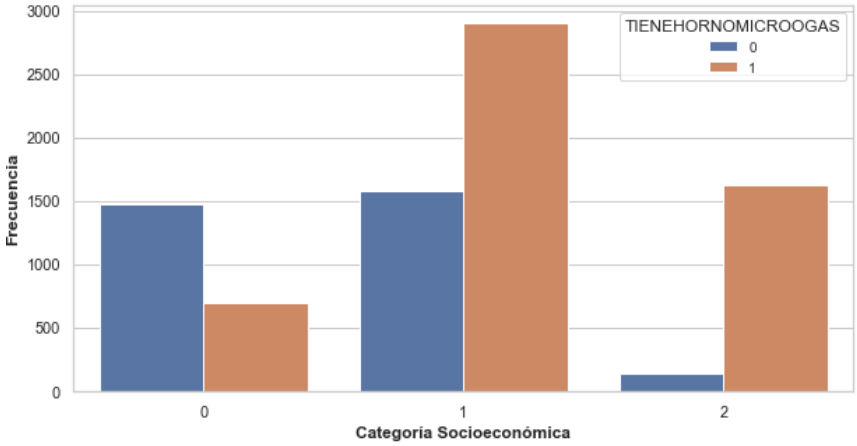

VALORMATRICULAUNIVERSIDAD
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
Frecuencia
1400
1200
1000
800
600
400
200
0
0
1
2
Categoría Socioeconómica

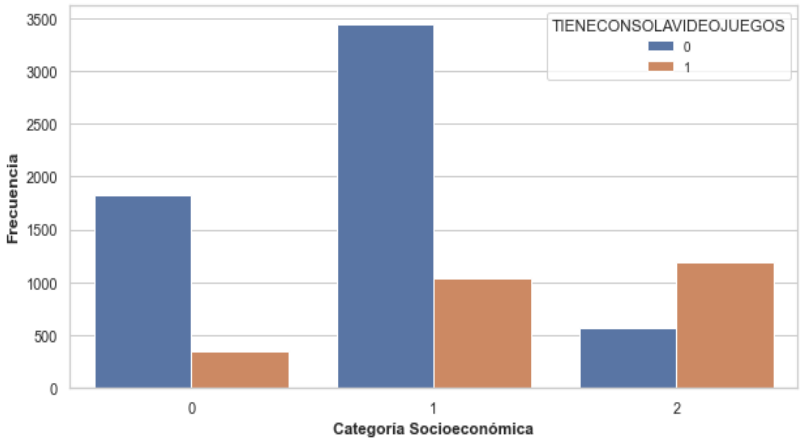

HORASSEMANATRABAJA
0.0
1.0
2.0
3.0
4.0
Frecuencia
1750
1500
1250
1000
750
500
250
0
0
1
2
Categoría Socioeconómica

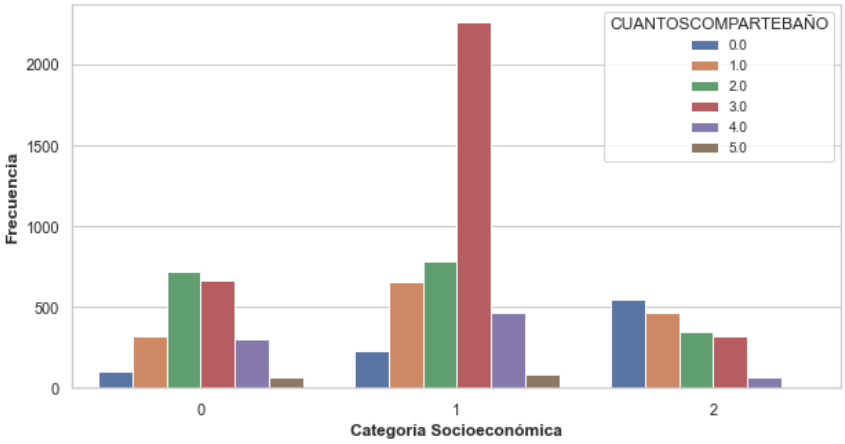

TIENEMOTOCICLETA
0
1
Frecuencia
2500
2000
1500
1000
500
0
0
1
2
Categoría Socioeconómica

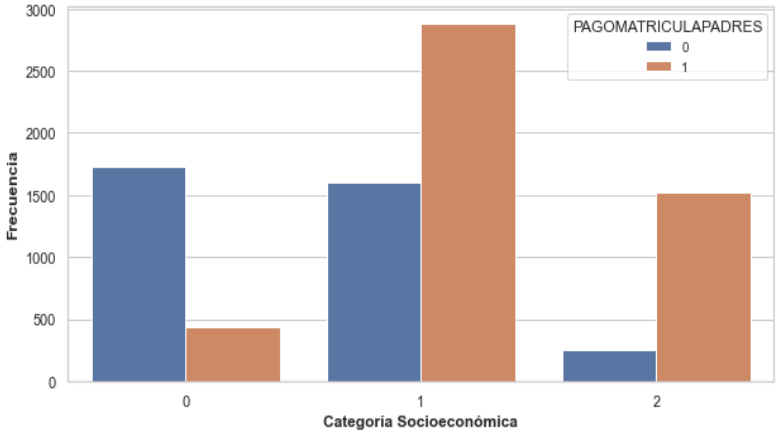

TRABAJOLABORPADRE
1.0
2.0
3.0
4.0
Frecuencia
1750
1500
1250
1000
750
500
250
0
0
1
2
Categoría Socioeconómica

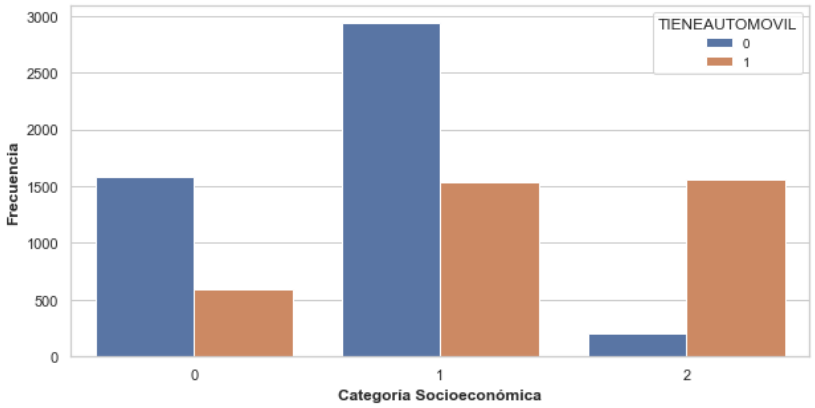

TRABAJOLABORMADRE
1.0
2.0
3.0
4.0
Frecuencia
2500
2000
1500
1000
500
0
0
1
2
Categoría Socioeconómica


TIENEHORNOMICROOGAS
0
1
Frecuencia
3000
2500
2000
1500
1000
500
0
0
1
2
Categoría Socioeconómica


TIENECONSOLAVIDEOJUEGOS
0
1
Frecuencia
3500
3000
2500
2000
1500
1000
500
0
0
1
2
Categoría Socioeconómica


CUANTOSCOMPARTEBAÑO
0.0
1.0
2.0
3.0
4.0
5.0
Frecuencia
2000
1500
1000
500
0
0
1
2
Categoría Socioeconómica





A partir de los diagramas de barras generados, se puede concluir:

- Conforme aumenta la Categoría Socioeconómica del estudiante también aumenta el estrato de su vivienda, el nivel educativo de sus padres, el valor de su matrícula, el nivel laboral de sus padres, la probabilidad de que tenga horno, consola de videojuegos y automóvil y la probabilidad de que los padres paguen su matrícula.
- Conforme disminuye la Categoría Socioeconómica aumenta el número de horas que trabaja, la probabilidad de que tenga motocicleta, la probabilidad que el estudiante pague su propia matrícula y el número de personas con las que comparte baño.

7.5.5 Perfiles tipológicos de los estudiantes de ingeniería industrial en Colombia

Para la creación de los perfiles tipológicos de los estudiantes conforme su categoría socioeconómica, se calcula la mediana de cada una de las variables para cada categoría en la figura 26.

Figura 26

Mediana de las variables socioeconómicas

SE_CAT	0	1	2
GLOBAL_PUNT	138.0	146.0	162.0
FAMI_ESTRATOVIVIENDA_CAT	2.0	3.0	4.0
FAMI_EDUCACIONMADRE_CAT	3.0	4.0	8.0
FAMI_EDUCACIONPADRE_CAT	3.0	4.0	8.0
ESTU_PAGOMATRICULAPROPIO_CAT	1.0	0.0	0.0
ESTU_VALORMATRICULAUNIVERSIDAD_CAT	3.0	4.0	6.0
ESTU_HORASSEMANATRABAJA_CAT	4.0	3.0	1.0
FAMI_TIENEMOTOCICLETA_CAT	0.0	0.0	0.0
FAMI_TRABAJOLABORPADRE_CAT	2.0	3.0	4.0
FAMI_TRABAJOLABORMADRE_CAT	2.0	2.0	4.0
FAMI_TIENEHORNOMICROOGAS_CAT	0.0	1.0	1.0
FAMI_TIENECONSOLAVIDEOJUEGOS_CAT	0.0	0.0	1.0
FAMI_CUANTOSCOMPARTEBAÑO_CAT	2.0	3.0	1.0
ESTU_PAGOMATRICULAPADRES_CAT	0.0	1.0	1.0
FAMI_TIENEAUTOMOVIL_CAT	0.0	0.0	1.0
UNI_GRUPO	0.0	1.0	2.0

A continuación, se muestran los perfiles socioeconómicos una vez agrupados.

7.5.5.1 Categoría socioeconómica 0. El estudiante medio que pertenece a la categoría socioeconómica 0 tiene las siguientes características:

- Puntúa 141 en las pruebas Saber Pro
- Es estrato 2
- Su madre completó el Bachillerato

- Su padre completó el Bachillerato
- Pago su propia matrícula
- Pagan entre 1 y 2.5 millones de matrícula
- Trabajan entre 21 y 30 h a la semana
- No tiene motocicleta (pero es más probable que la tenga vs otros grupos)
- Su padre es dueño de un negocio pequeño o tiene un trabajo de tipo auxiliar administrativo
- Su madre trabaja en el hogar, no trabaja o estudia o trabaja como personal de limpieza, mantenimiento, seguridad o construcción o es vendedora o trabaja en atención al público
- Tiene horno microondas o gas (pero es menos probable vs otros grupos)
- No tiene consola de videojuegos (y es menos probable que la tenga vs otros grupos)
- Comparte baño con 3 0 4 personas
- No tiene automóvil (y es menos probable que la tenga vs otros grupos)

7.5.5.2 Categoría Socioeconómica 1. El estudiante medio que pertenece a la categoría socioeconómica 1 tiene las siguientes características:

- puntúa 146 en las pruebas Saber Pro
- Es estrato 2
- Su madre completó el Bachillerato
- Su padre completó el Bachillerato
- Sus padres pagaron su matrícula
- Pagan entre 2.5 y 4 millones de matrícula
- Trabajan entre 21 y 30 h a la semana

- No tiene motocicleta
- Su padre trabaja por cuenta propia (por ejemplo, plomero, electricista) o es pensionado o es operario de máquinas o conduce vehículo
- Su madre trabaja en el hogar, no trabaja o estudia o trabaja como personal de limpieza, mantenimiento, seguridad o construcción o es vendedora o trabaja en atención al público.
- Tiene horno microondas o gas
- No tiene consola de videojuegos
- Comparte baño con 3 0 4 personas
- No tiene automóvil

7.5.5.3 Categoría Socioeconómica 2. El estudiante medio que pertenece a la categoría socioeconómica 2 tiene las siguientes características:

- Puntúa 161 en las pruebas Saber Pro
- Estudia en una IES categoría 2
- Es estrato 3
- Su madre completó el Bachillerato
- Su padre completó el Bachillerato
- Sus padres pagaron su matrícula
- Pagan entre 5.5 y 7 millones de matrícula
- Trabajan entre 11 y 20h a la semana
- No tiene motocicleta (Y es menos probable que la tenga vs Otros grupos)
- Su padre trabaja por cuenta propia (por ejemplo, plomero, electricista) o es pensionado o es operario de máquinas o conduce vehículo

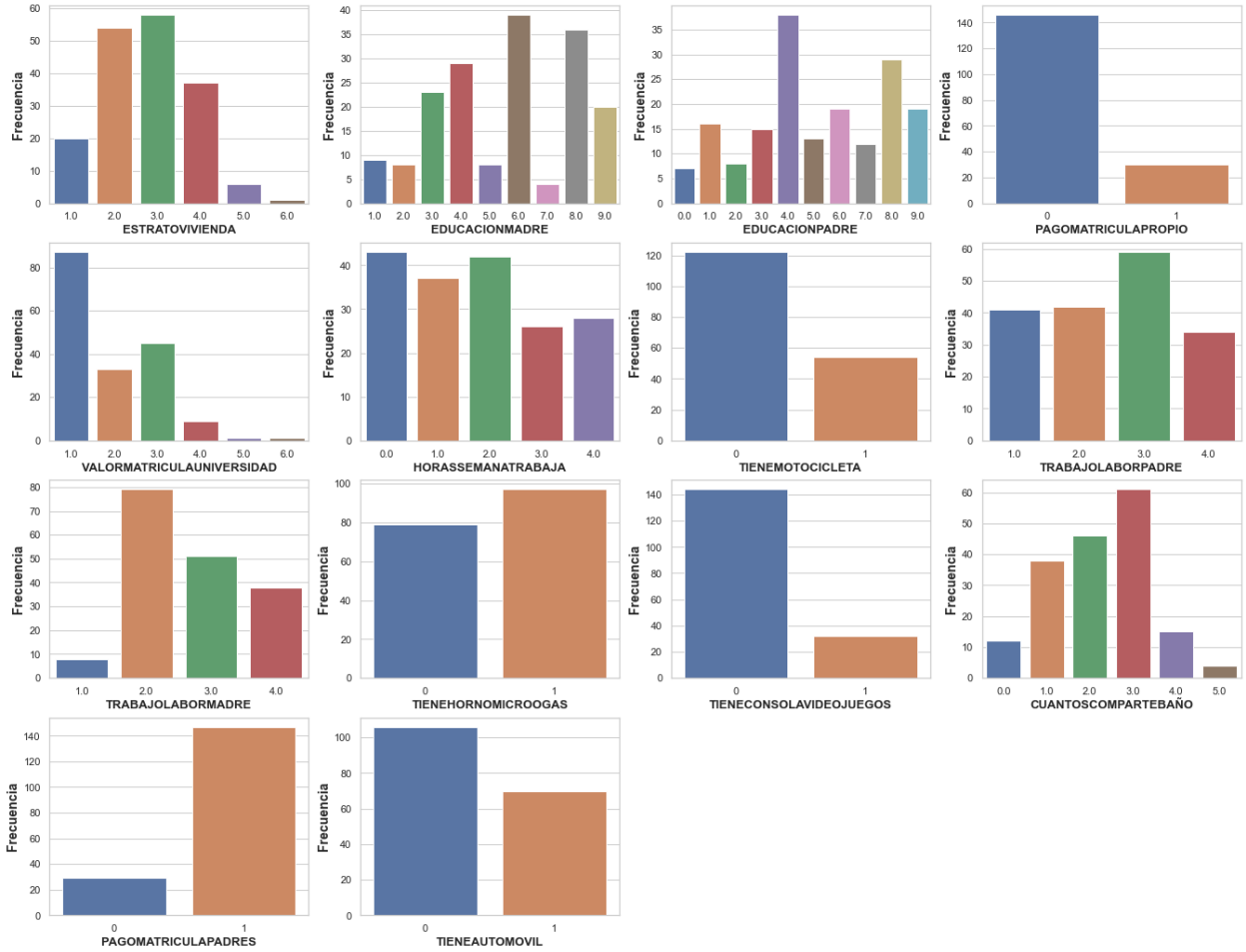
- Su madre es dueña de un negocio pequeño o tiene un trabajo de tipo auxiliar administrativo o trabaja por cuenta propia (por ejemplo, plomero, electricista) o es pensionada o es operaria de máquinas o conduce vehículos.
- Tiene horno microondas o gas (y es más probable que lo tenga vs Otros grupos)
- No tiene consola de videojuegos(Pero es más probable que la tenga vs Otros grupos)
- Comparte baño con 2 personas
- Tiene automóvil

7.5.6 Clasificación de estudiantes UIS

De igual manera, se procede a examinar las variables Socioeconómicas, pero esta vez, únicamente de los estudiantes UIS. En la figura 27, se observa la representación para cada variable socioeconómica.

Figura 27

Frecuencia de estudiantes UIS para cada variable socioeconómica



Los estudiantes en la UIS tienen una distribución de variables socioeconómicas similar a la población de estudiantes de todo el país, sin embargo, se identifican algunas diferencias:

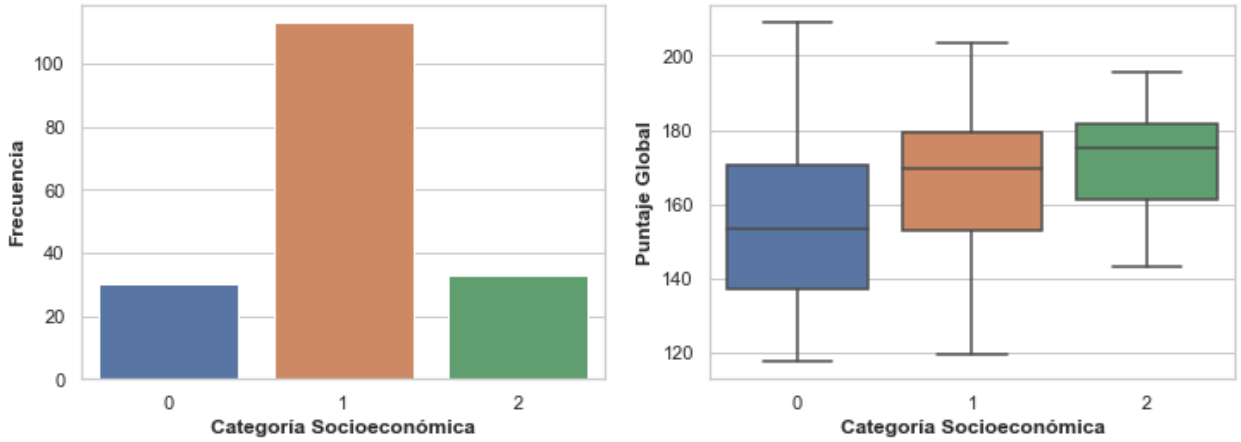
- Un reducido número de estudiantes con estratos altos (5 y 6).
- Una menor proporción de estudiantes que pagan su matrícula.
- Costos de matrícula bajos, hay una cantidad mínima de estudiantes con matrícula superior a los 5.5 millones.
- Un menor número de horas de trabajo.

- Una ligera mejora en el nivel educativo en los padres.

A continuación, se clasifica a los estudiantes de la UIS, de acuerdo a sus perfiles tipológicos por medio del algoritmo K-Modes. Para observar la clasificación, en la figura 28, se realiza un diagrama de barras, junto con un diagrama de cajas, para observar los puntajes en los diferentes niveles socioeconómicos.

Figura 28

Clasificación de los estudiantes de la UIS de acuerdo a sus perfiles tipológicos



La mayoría de los estudiantes de la UIS están clasificados dentro de la categoría 1. También, se observa una mejora en el rendimiento conforme se avanza en la categoría socioeconómica, sin embargo, se puede observar que el mayor puntaje fue obtenido por un estudiante dentro de la categoría 0, esta categoría es la que tiene una mayor dispersión en sus datos, esta dispersión también disminuye conforme se avanza en la categoría.

Por otro lado, continuando con el análisis socioeconómico de los estudiantes de ingeniería industrial, se genera una tabla (figura 29), que muestra los porcentajes de cada categoría socioeconómica que componen cada grupo de IES, en donde se observa que conforme

se avanza en la categoría de universidad, aumenta la proporción de estudiantes de mayor categoría socioeconómica.

Figura 29

Porcentaje de cada categoría socioeconómica que compone cada grupo de IES.

SE_CAT	0	1	2
UNI_GRUPO			
0	39.82	53.51	6.67
1	23.81	63.28	12.91
2	13.74	56.40	29.86
3	7.62	26.10	66.28

Agregando a lo anterior, en la figura 30 y 31 se realizan dos diagramas de cajas para las categorías socioeconómicas y de IES respectivamente, para observar su relación con el puntaje global. Después, en la figura 32, se hace un diagrama de cajas para las combinaciones de estas dos variables con el fin de observar detalladamente la relación que guardan con el puntaje.

Figura 30

Influencia de la categoría socioeconómica sobre el puntaje global de los estudiantes de Ingeniería Industrial de la UIS

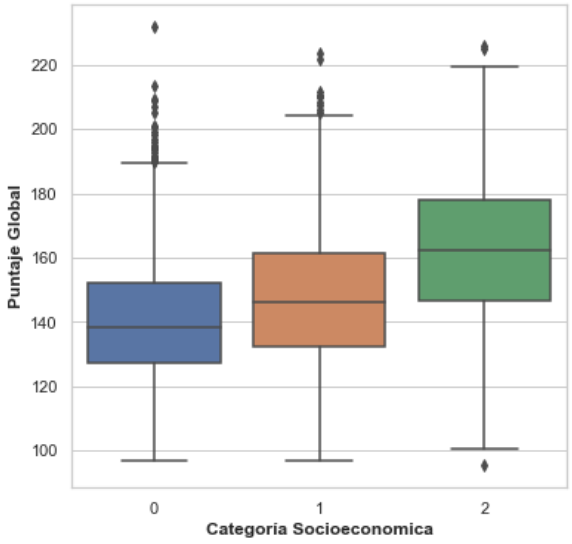


Figura 31

Influencia de la categoría IES sobre el puntaje global de los estudiantes de Ingeniería Industrial de la UIS

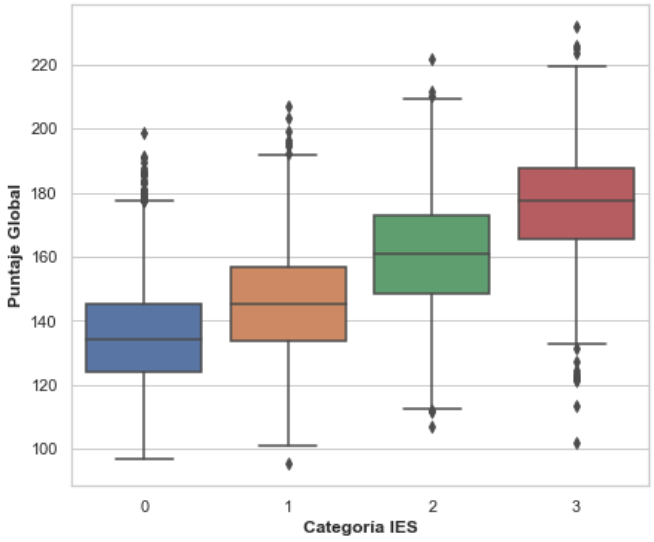
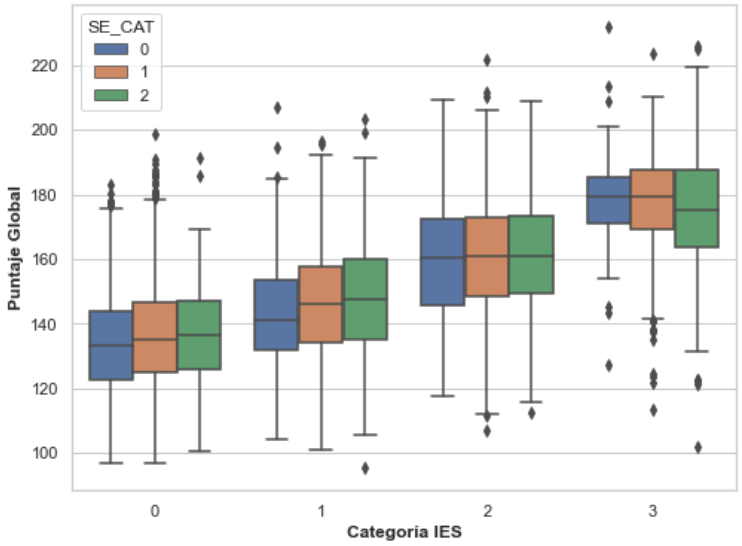


Figura 32

Relación entre la categoría IES y la variable socioeconómica con el puntaje global de los estudiantes de Ingeniería Industrial de la UIS



A primera vista, según la gráfica 30 parecería que la categoría socioeconómica tiene una clara influencia sobre el puntaje global, pero al ver de cerca la gráfica 32 observamos que esta influencia no es tan significativa, si bien dentro de los 3 primeros grupos de IES hay una leve tendencia a la alza en el puntaje global conforme aumenta la categoría socioeconómica, gran parte de la relación inicial es resultado de que conforme aumenta el nivel socioeconómico, hay una mayor probabilidad de pertenecer a una mejor IES, es decir hay una sobrerrepresentación de los niveles socioeconómicos más altos en las mejores IES.

Con el fin de explorar la significancia estadística de la variable socioeconómica, en la figura 33, se realiza un análisis ANOVA junto con la categoría de IES, en el cual se obtiene como resultado que la categoría socioeconómica si es estadísticamente significativa, sin embargo, al observar el valor F de las variables se observa que tiene una influencia leve en comparación con la categoría IES.

Figura 33

Análisis Anova categoría socioeconómica

	df	sum_sq	mean_sq	F	PR(>F)
C(UNI_GRUPO)	3.0	1.659588e+06	553195.879119	1980.431278	0.000000e+00
C(SE_CAT)	2.0	9.033028e+03	4516.514097	16.169039	9.802972e-08
C(UNI_GRUPO):C(SE_CAT)	6.0	4.215619e+03	702.603223	2.515307	1.962927e-02
Residual	8400.0	2.346381e+06	279.331015	NaN	NaN

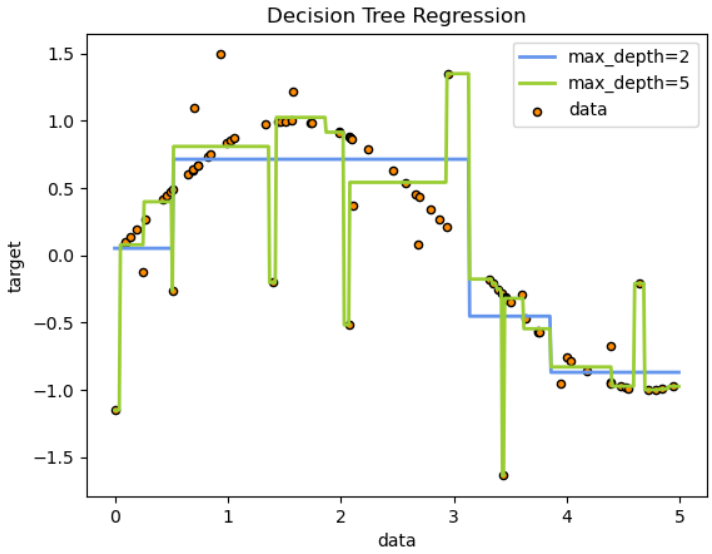
7.5.7 Descripción de Árbol de decisión

Los árboles de decisión (DT) son un método de aprendizaje supervisado no paramétrico que se utiliza para la clasificación y la regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo mediante el aprendizaje de reglas de decisión simples inferidas a partir de las características de los datos. Un árbol puede verse como una aproximación constante a trozos.

Por ejemplo, en el siguiente caso ilustrado en la figura 34, los árboles de decisión aprenden de los datos para aproximar una curva senoidal con un conjunto de reglas de decisión si-entonces-sí. Cuanto más profundo sea el árbol, más complejas serán las reglas de decisión y más ajustado será el modelo.

Figura 34

Ejemplo de árbol de decisión



Algunas ventajas de los árboles de decisión son:

- Son fáciles de entender e interpretar. Los árboles pueden visualizarse.
- Requiere poca preparación de los datos. Otras técnicas suelen requerir la normalización de los datos, la creación de variables ficticias y la eliminación de los valores en blanco. Sin embargo, hay que tener en cuenta que este módulo no admite valores perdidos.
- El costo de utilizar el árbol (es decir, de predecir los datos) es logarítmico en el número de puntos de datos utilizados para entrenar el árbol.
- Puede manejar tanto datos numéricos como categóricos.
- Capaz de manejar problemas de múltiples salidas.
- Utiliza un modelo de caja blanca. Si una situación determinada es observable en un modelo, la explicación de la condición se explica fácilmente mediante la lógica booleana. En cambio, en un modelo de caja negra (por ejemplo, en una red neuronal artificial), los resultados pueden ser más difíciles de interpretar.

- Es posible validar un modelo mediante pruebas estadísticas. Eso permite dar cuenta de la fiabilidad del modelo.

7.5.8 Aplicación de Árbol de decisión

Con el fin de mejorar la comprensión sobre las variables que afectan el puntaje global de los estudiantes, se ajusta un árbol de decisión regresor teniendo en cuenta las variables académicas (Categorías IES) y las variables socioeconómicas (Categoría Socioeconómica)

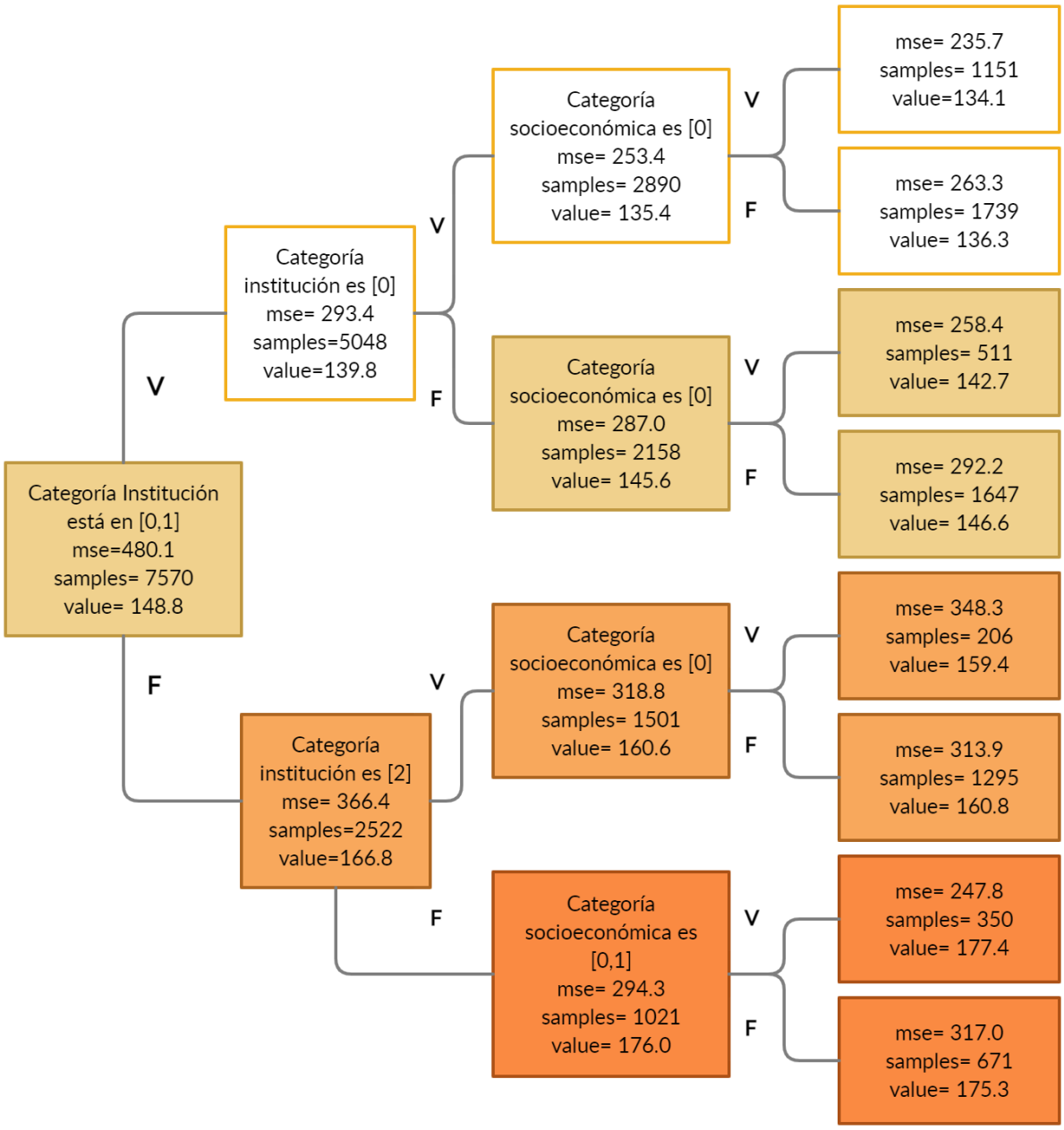
Este modelo de aprendizaje automático supervisado realiza abstracciones fáciles de comprender cuando se tiene un número limitado de variables. Para esto, se importan las librerías necesarias para poder realizar la predicción de los datos entre los dos grupos de clústeres obtenidos anteriormente (‘SE_CAT’ y ‘UNI_GRUPO’) respecto a los puntajes globales obtenidos en las pruebas.

Esta predicción se realiza gracias a Scikit-Learn, la cual permite entrenar un cierto número de elementos del conjunto de datos X (x_train) y mientras que otro cierto número de datos de este mismo conjunto de datos se guardan para probar la precisión del modelo obtenido, estos datos son x_test.

Por consiguiente, en la figura 35, se genera una gráfica con formato png del árbol de decisión con una profundidad máxima para el árbol de 3 niveles de decisión.

Figura 35

Árbol de decisión



La jerarquía de los nodos es decreciente, es decir, los primeros nodos toman las decisiones más trascendentales. El análisis y resumen del árbol se presenta a continuación:

Primer Nodo (Categoría IES está en [0,1]), separa a los estudiantes entre V: Categoría IES 0, 1 y F: Categoría IES 2, 3

El determinante más importante del resultado esperado de un estudiante depende de esta pregunta, si es verdadera, se espera un resultado global de 139.8 en las pruebas, si es falsa 166.7, una diferencia de 26.9 puntos un poco más que una desviación estándar (21.8).

Segundo Nodo A (Categoría IES es [0]), separa a los estudiantes entre V: Categoría IES 0 y F: Categoría IES 1

No tan significativo como el anterior nodo, si C Categoría IES = 0 se espera un resultado 135.4 y si Categoría IES = 1 se espera 145.7, una diferencia de 10 puntos en el puntaje global.

Segundo Nodo B (Categoría IES es [2]), separa a los estudiantes entre V: Categoría IES 2 y F: Categoría IES 3

Una diferencia más amplia que en el Nodo 2A, si Categoría IES = 2 se espera un resultado 160.4 y si Categoría IES = 3 se espera 176.1, la diferencia es de 15.7 puntos. El número esperado para la categoría IES 3 está muy por encima de la media (148.75) la diferencia es de 1.25 veces la desviación estándar, se podría decir que con solo saber que un estudiante hace parte de una institución categoría 3 su resultado estará por encima del 89% de los demás resultados ($Z = 1.25$, $p = 0.894$).

Tercer Nodo A, B, C (Categoría socioeconómica es [0]), separa a los estudiantes entre V: Categoría socioeconómica 0 y F: Categoría socioeconómica 1, 2

Este nodo separa a los estudiantes con categoría socioeconómica 0 de los demás, en los tres nodos se observan las siguientes variaciones de 2.2, 3.9 y 1.8, son pequeñas en comparación con las variaciones resultantes de Categoría IES, curiosamente los 3 nodos dividen las ramas en 0.5 indicando que las diferencias más significativas se dan en la Categoría socioeconómica 0, es

decir para las IES de categorías 0, 1, 2 tener una categoría socioeconómica mayor a 0 ofrece una ligera mejora en los puntajes.

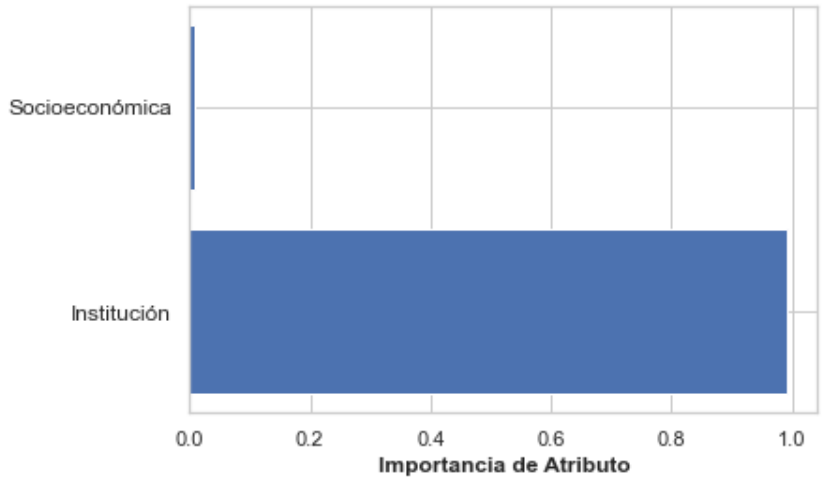
Tercer Nodo D (Categoría socioeconómica es [0,1]), separa a los estudiantes entre V: Categoría socioeconómica 0, 1 y F: Categoría socioeconómica 2

Este nodo separa a los estudiantes con categoría socioeconómica 2 de los demás en las instituciones de categoría 3, para las Categorías socioeconómicas 0 y 1 se espera un puntaje de 177.4 y para la Categoría socioeconómica 2 se espera 175.4. Este es el único grupo de IES, donde una menor Categoría socioeconómica, lleva a una mejora de los puntajes obtenidos, es decir, para las IES de categorías 3 tener una Categoría socioeconómica menor a 2 ofrece una ligera mejora en los puntajes.

Se procede a acceder a los puntajes de importancia de las variables dentro del árbol de decisión para contrastar la influencia de la categoría socioeconómico con la de la IES.

Figura 36

Importancia de atributos en el árbol de decisión



Al analizar la figura 36 se concluye que, aunque la categoría socioeconómica si es una variable útil para calcular el puntaje esperado de los estudiantes, no es tan relevante como la

institución de educación superior a la que pertenecen. También se observa que hay una mayor proporción de estudiantes con mejor categoría socioeconómica a medida que se asciende en las categorías socioeconómicas.

8. Minería de texto

8.1 Recopilación de información

Para recopilar la información necesaria para realizar el proceso de análisis de sentimientos, se realiza una encuesta virtual por medio de la plataforma Google Forms, con tres preguntas de tipo abiertas que indagan sobre la opinión de los estudiantes de Ingeniería Industrial de la Universidad Industrial de Santander, que presentaron las pruebas Saber Pro en el año 2019, las preguntas formuladas se encuentran en el apéndice E, en el cual se sugiere que las respuestas tengan más de cincuenta (50) palabras, con el fin de recolectar la mayor cantidad de texto posible. La encuesta es enviada directamente desde el correo de la Escuela de Estudios Industriales y Empresariales, a las bases de datos de los egresados que presentaron las pruebas Saber Pro en el año 2019 y a los estudiantes activos que actualmente estén cursando noveno y décimo semestre.

Para el desarrollo de las actividades de minería de texto, se usó el lenguaje de programación Python 3, sobre el entorno web computacional Jupyter Notebook, en el apéndice G es posible observarlo detalladamente.

8.2 Preprocesamiento de texto

Para el desarrollo de estas actividades se usó el lenguaje de programación Python 3 sobre el entorno web computacional Jupyter Notebook

Inicialmente, se abre el archivo con la información recolectada en la fase anterior, se puede observar que es un archivo con extensión xlsx y no se tiene en cuenta la columna “Marca temporal” la cual indica la fecha y la hora en la que fue respondida la encuesta y no es relevante para el estudio, es decir, en total se tienen 3 columnas y cada una de ellas consta de 54 filas, cada fila corresponde a la respuesta de un encuestado.

Después, se renombran las columnas de la base de datos bajo las etiquetas “P1”, “P2”, “P3”, correspondientes a las 3 preguntas realizadas a los encuestados; esto se realiza para simplificar el llamado de cada columna debido a que los nombres originales eran muy extensos: ¿Cuál es su opinión respecto al examen Saber Pro?, ¿Qué opina del rol que tuvo su universidad en el desempeño de las pruebas Saber Pro?, Describa cómo fue su experiencia al momento de presentar las pruebas Saber Pro, ahora se llamarán “P1”, “P2” y “P3” respectivamente.

Después de importar y organizar la base de datos que contiene las respuestas, se procede a limpiar el texto, esto es de suma importancia en el análisis de las respuestas porque es la se encarga de limpiar los datos escritos por cada estudiante para después juntar todas las 54 respuestas en un solo bloque de texto continuo. Dicha limpieza consta de conversión de todo el texto a minúscula, eliminación de signos de puntuación, eliminación de símbolos y comillas y conversión de punto y aparte en un espacio.

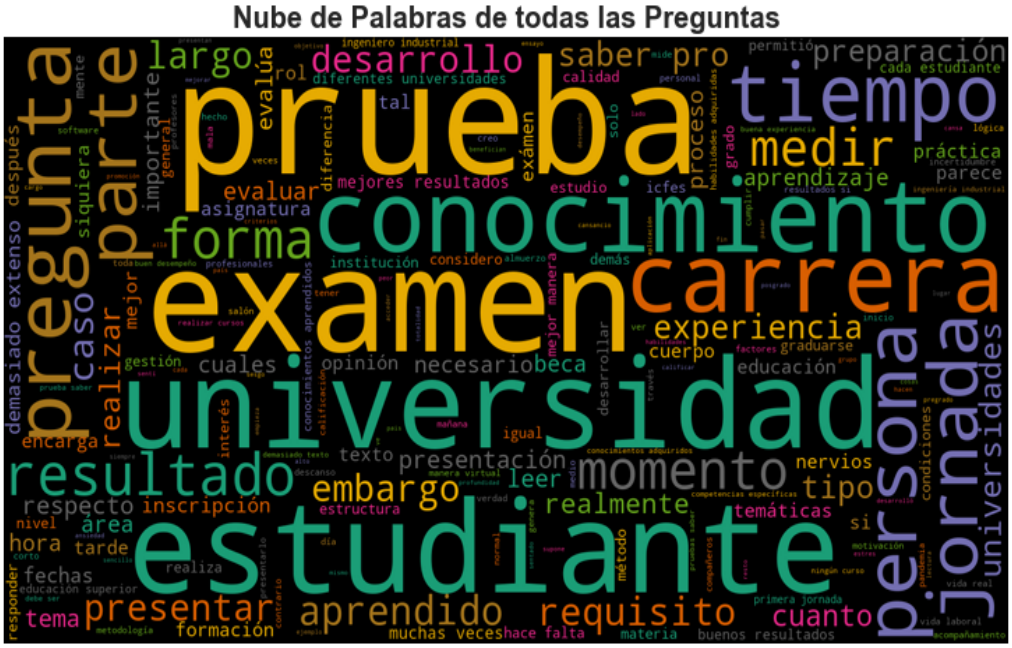
“re” es la librería regular expressions, su función re.sub() permite reemplazar caracteres seleccionados por espacios, buscando evitar que estos caracteres sean tenidos en cuenta. Para ejecutar la limpieza de los datos se crea una función llamada text_cleaner, esta función será aplicada a cada una de las preguntas realizadas a los encuestados. El texto obtenido de cada una de las preguntas se llama extract_p1, extract_p2 y extract_3 para las preguntas P1, P2 y P3

respectivamente. Posteriormente se crea una nueva estructura de datos de pandas con las respuestas limpias, y se le da el nombre de extractos.

Después, se genera un gráfico llamado Wordcloud o en español nube de palabras. Dicho gráfico, es un objeto de representación visual sencillo pero potente para el tratamiento de textos, que muestra la palabra más frecuente con letras más grandes y más gruesas, y con diferentes colores. Cuanto menor sea el tamaño de la palabra, menor será su importancia. El argumento stopwords = spanish_stopwords limpia las palabras comunes y signos de puntuación que se utilizan en el lenguaje español y que podrían resultar en una malinterpretación de la información debido a su alta frecuencia de uso en los textos. Primero se genera una nube de palabras para todo el texto respondido por los estudiantes, independientemente de la pregunta, como se observa en la figura 37.

Figura 37

Nube de palabras de todas las preguntas

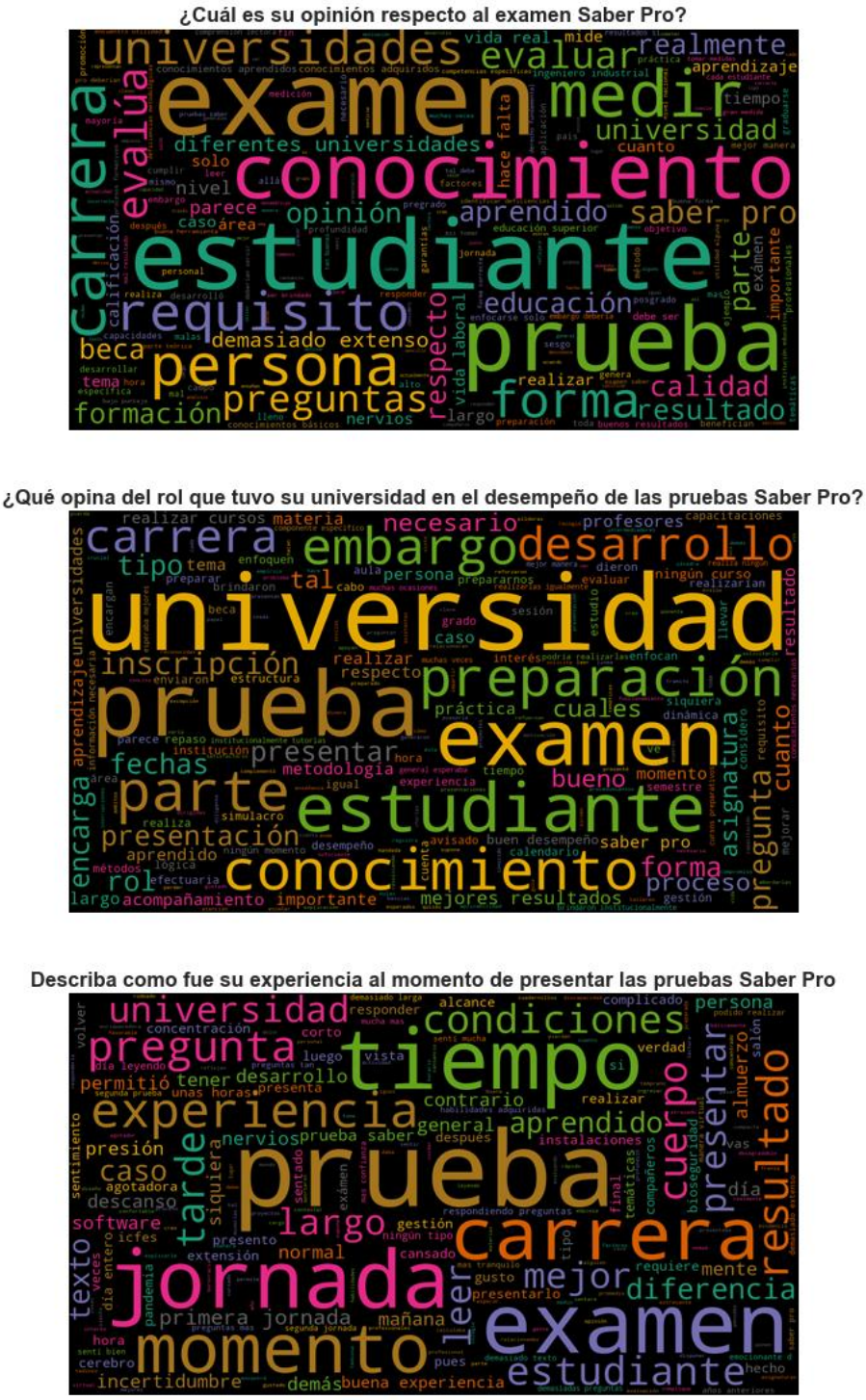


Las palabras más utilizadas fueron; prueba, examen, universidad, estudiante, conocimiento, pregunta, tiempo, carrera, entre otras. Las palabras "prueba y examen", son el sujeto del cual se está indagando en la encuesta, "universidad y estudiante" son los evaluados, "pregunta" es un constituyente del examen, "conocimiento" es lo que se evalúa y, por último, "tiempo y carrera", que son parámetros del examen. También se pueden observar algunas palabras interesantes como lo son; "demasiado extenso/ demasiado texto", "nervios", "requisito", "grado/graduarse" y "almuerzo", de las cuales se concluye que algunos estudiantes únicamente ven el examen como un requisito para su graduación, es percibido como extenso, genera nervios y el almuerzo parece jugar cierto rol en él.

Después, en la figura 38, se genera el respectivo gráfico de nube de palabras para cada una de las preguntas.

Figura 38

Nube de palabras para cada una de las preguntas



De la nube de palabras obtenida para la pregunta 1, es evidente que la palabra más utilizada por los estudiantes es “examen” junto con “estudiante” y “conocimiento”. Estas palabras por sí solas no ofrecen una interpretación acerca de la opinión de los estudiantes encuestados. Sin embargo, si se analizan las palabras escritas con menos frecuencia se pueden encontrar palabras como “deberían” ,”debería”, “mejorar”, “largo”; estas últimas palabras pueden ser un indicador de que la opinión de los encuestados acerca de las pruebas Saber Pro no es del todo positiva, sin embargo, para realmente analizar si el pensamiento colectivo de los encuestados es positivo o negativo, más adelante se realiza un análisis de la polaridad del sentimiento de las respuestas de los estudiantes. las palabras que resultan interesantes son: “calidad”, “aprendizaje”, “vida real”, “vida laboral”, “realmente”, “beca”, “hace falta”, “demasiado extenso”, “importante” y “posgrado”. A partir de estas palabras se concluye que hay un sentimiento generalizado de falta de interés, parcialmente causado por la carencia de beneficios al obtener un buen puntaje, también se observa que algunos estudiantes no creen que esta evaluación mida las habilidades usadas en la vida laboral y el mundo real.

De la nube de palabras resultante para la segunda pregunta, se puede evidenciar que la palabra más utilizada por los encuestados fue “universidad”, esto es un resultado evidente ya que es el sujeto sobre el cual recae directamente la pregunta. Las palabras más interesantes son: “buena experiencia”, “buen desempeño”, “bueno”, “ningún curso”, “mejores resultados”, “acompañamiento” e “inscripción”. A partir de estas palabras se concluye, que hay un número de estudiantes que considera que la UIS hizo un buen acompañamiento, sin embargo, hay algunas carencias en cuestiones de cursos.

Por último, para la pregunta tres, en la nube de palabras obtenida se puede evidenciar que las palabras más frecuentemente utilizadas son: “prueba”, “tiempo” y “jornada”. Las palabras más

interesantes son: “agotadora”, “presión”, “demasiado larga”, “nervios”, “largo”, “día entero” y “concentración”. A partir de estas palabras se concluye, que algunos estudiantes encuentran las pruebas agotadoras, ya sea por el número de horas del examen, o el esfuerzo mental requerido para completar todos los módulos.

Para complementar los resultados obtenidos en las nubes palabras, se realiza un análisis del sentimiento de los estudiantes encuestados. Para esto se utiliza una inteligencia artificial perteneciente al submódulo SentimentIntensityAnalyzer de la librería vaderSentiment.vaderSentiment de Python. Los resultados obtenidos del análisis de sentimiento son valores entre -1 y 1 para cada respuesta escrita por un estudiante, donde -1 es una opinión muy negativa y 1 es una opinión muy positiva. Se aplica la inteligencia artificial obtenida llamada ai_sentiment; a cada una de las preguntas: extract_p1, extract_p2 y extract_p3 para las preguntas 1, 2 y 3 respectivamente. Se revisa si los resultados de la inteligencia artificial son los esperados y se obtiene efectivamente cada una de las respuestas de los encuestados con una ponderación que representa su sentimiento positivo o negativo referente a las pruebas Saber Pro.

Finalmente, para mostrar gráficamente los resultados obtenidos del análisis de sentimiento, se crean histogramas mostrando la frecuencia con la que se obtienen diferentes valores de sentimiento de los encuestados para cada una de las preguntas realizadas. En cada una de estas gráficas los datos a los cuales se le analizará su repetitividad son las ponderaciones obtenidas acerca del sentimiento de los encuestados ante cada una de las preguntas realizadas.

Figura 39

Histograma para la pregunta 1

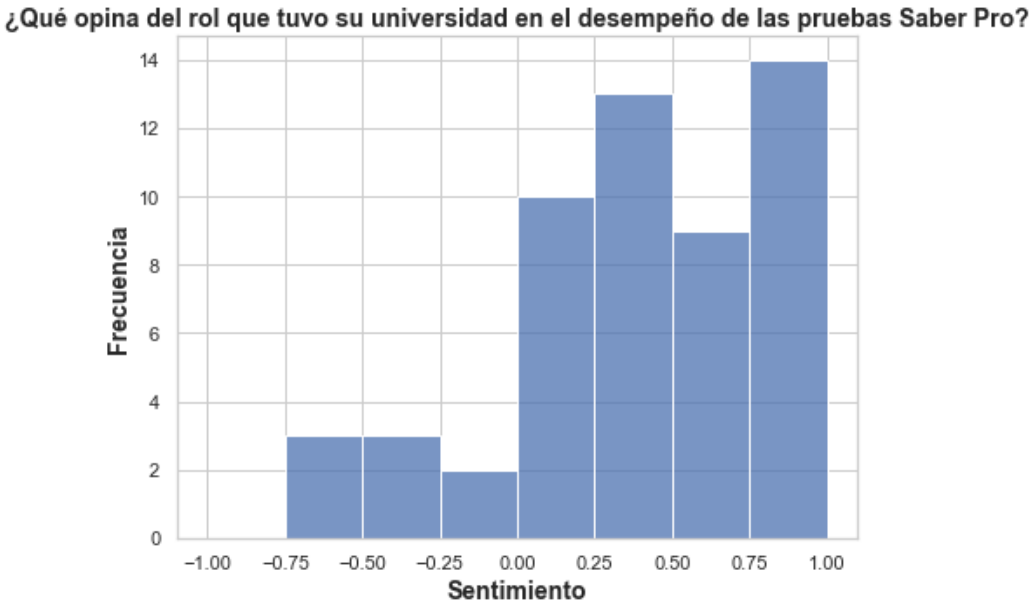
Acerca de la opinión de los estudiantes con respecto al examen en la figura 39, se concluye que la mayoría de los encuestados tienen una percepción ligeramente positiva (sentimiento cercano a 0 por la derecha), seguido de los intervalos [-0.5: 0.25] y [0.75: 1]. Se concluye que la mayoría de los estudiantes tienen sentimientos neutros hacia las pruebas en general, un grupo siente una mediana aversión y otro, una opinión muy positiva hacia estas. Las opiniones sobre estas pruebas son diversas, sin embargo, tienden a ser neutrales. Un ejemplo de las opiniones positivas del examen es:

- “me parece una buena herramienta para evaluar la calidad de la educación superior en Colombia, pero en igual medida para evaluar los conocimientos que como estudiantes adquirimos y desarrollamos en nuestra institución educativa igualmente por medio de este examen se puede conocer la calidad de profesionales que salen al mundo laboral a poner en práctica los conocimientos adquiridos durante la carrera”.

- “es útil para medir las competencias de los estudiantes y comparar los resultados y las habilidades que se adquieren en las diferentes universidades del país permite al gobierno dar direcciones para mejorar las políticas educativas y una mejor administración de los recursos permite a las instituciones de educación mejorar”.
- “considero que este examen es necesario para realizar una retroalimentación de los conocimientos, pero no deberían hacerlo tan extenso”.

Figura 40

Histograma para la pregunta 2



Respecto a la opinión de los estudiantes con el rol que tuvo la UIS en sus pruebas como se observa en la figura 40, las opiniones son mayoritariamente positivas, el intervalo con mayor número de integrantes es el de [0.75: 1] seguido por [0.25: 0.5], aunque hay entrevistados con una opinión negativa, estos son pocos en comparación con el total. Algunas de las respuestas son:

- “Pues, la universidad se encarga de prepararnos a lo largo de 10 semestres, pero ya el examen y los resultados dependen de nosotros como estudiantes. En general la

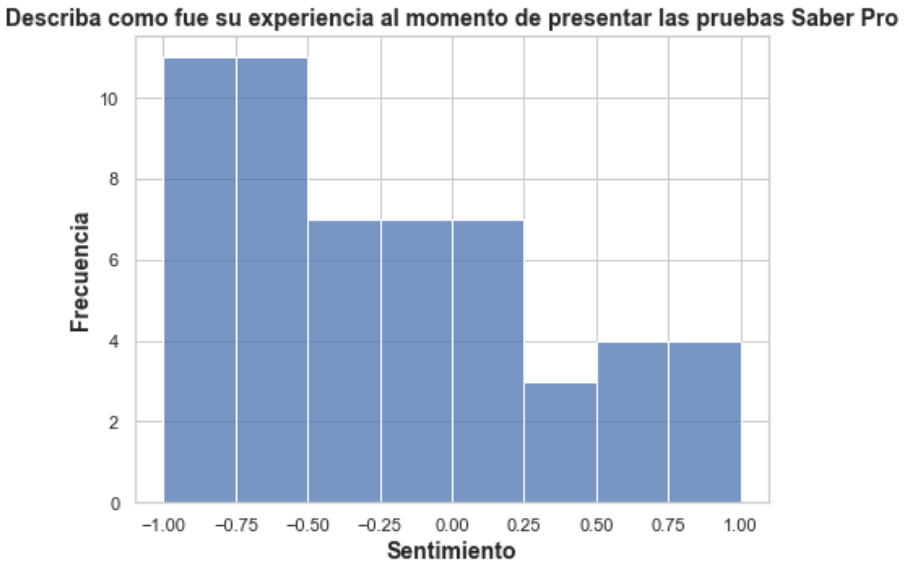
universidad fue diligente con todos los procedimientos e inscripción necesaria para que pudiésemos presentar el examen”

- “La universidad a lo largo de la carrera proporciona bases para la presentación de esas pruebas. Sin embargo, en cuanto a preparación específica creo que se está quedando corta en el apoyo hacia los estudiantes, en este caso podría ser la propia Escuela. En mi caso preparé el examen por mi propia cuenta con repasos y demás y el resultado para mí fue satisfactorio. Las charlas que la universidad proporciona para los componentes generales son básicas y sirven”
- “La universidad me dio las herramientas y los conocimientos necesarios para presentar las pruebas, una formación académica durante diez semestres, aunque considero que el desempeño también adquiere un carácter muy individual que depende del tipo de persona y el compromiso asumido. Las universidades ven medido su rendimiento mediante este tipo de pruebas”

Finalmente se realizó la misma representación gráfica para la tercera pregunta realizada y los resultados obtenidos de los sentimientos de los estudiantes hacia esta, se encuentran en el siguiente histograma:

Figura 41

Histograma para la pregunta 3



A partir del histograma para la pregunta 3 ilustrado en la figura 41, resulta evidente que la mayor parte de los estudiantes tienen una percepción negativa acerca de su experiencia en el Saber Pro. Los valores de sentimiento se encuentran en todo el rango posible de valores (-1 a 1), sin embargo, existe una evidente tendencia hacia una opinión negativa por parte de los estudiantes encuestados, esto se puede analizar de las dos últimas barras, cuyos intervalos con el mayor número de encuestados son [-1: -0.75] y [-0.75: -0.5]. Algunas de las respuestas dadas son:

- “No fue muy buena, por el contrario, de lo más desagradable, las instalaciones, la persona a cargo de cuidar la prueba fue un desastre totaaaal”
- “nada de eso tuvo que pasar, existe mucha improvisación en cuanto a este tipo de pruebas, y más cuando es alguien que no está preparada para explicarle a personas con algún tipo de discapacidad o lento aprendizaje”
- “Estresante y de incertidumbre. Las preguntas que encontré en el examen realizado en 2019-2 fueron ambiguas, muchas con concepto político y connotación despectiva

frente a minorías. Algunas preguntas estaban relacionadas con campos totalmente diferentes a la ingeniería industrial, teniendo que realizar conjeturas de temáticas como química, geología y física. Sentí que estas pruebas realmente no reflejan mi proceso de aprendizaje, ni ponen en práctica habilidades que son valiosas hoy en el mundo laboral. Tener que disponer de un día de descanso para una actividad como estas me parece inadecuado."

- "La experiencia es volver a presentar una prueba saber común y corriente, con la diferencia de que vas sin la presión de un resultado. Así como vas más relajado, también te esfuerzas menos. Sales de la prueba con la certeza de que cumpliste el requisito de grado, pero en lo personal nunca me interesó conocer mis resultados ya que estos no definían mis habilidades como profesional."

Por último, en la tabla 14 se genera la media obtenida para el sentimiento de cada una de las preguntas realizadas, esto se genera utilizando el método groupby de la librería pandas y aplicando la función "mean" de la librería numpy para obtener sus promedios.

Figura 42

Evolución de la distribución de los puntajes por universidades

Pregunta	Pregunta 1	Pregunta 2	Pregunta 3
Sentimiento	0.067	0.393	-0.228

Después de obtener el promedio de los sentimientos por pregunta, y a partir de las nubes de palabras se infiere que:

- Acerca del examen en general, los entrevistados en promedio no exhiben emociones fuertes hacia él, se manifiesta una disconformidad acerca de su magnitud, falta de congruencia con

la vida laboral y carencia de estímulos para un buen desempeño, y por otro lado se recalca que es un requisito necesario para medir la calidad de la educación.

- En cuanto al rol de la universidad, los entrevistados muestran en promedio, una actitud medianamente positiva, se remarca el acompañamiento en cuanto al proceso de inscripción y a la formación brindada, también se menciona la presencia o falta de cursos para la preparación.
- En cuanto a la experiencia de presentar las pruebas, los entrevistados en promedio comunican una opinión ligeramente negativa, en la nube se hace mención al cansancio, las jornadas, las condiciones, se menciona que la prueba es larga, estos factores parecen ser los que hacen de la prueba una experiencia no tan amena.

El artículo científico con los resultados obtenidos del presente trabajo de investigación se adjunta en el Apéndice H.

9. Conclusiones

A partir de un análisis univariado se identifica a la institución de educación superior como el principal factor determinante del resultado de un estudiante sobre las pruebas Saber Pro. Las IES pueden ser clasificadas en 4 grupos y la UIS hace parte del segundo mejor de estos.

El estrato socioeconómico y la educación de los padres son los factores socioeconómicos más determinantes en el resultado de un estudiante sobre las pruebas Saber Pro. Todas estas características pueden ser agrupadas mediante una caracterización tipológica que dan como resultados un grupo de perfiles determinados que agrupan a la población estudiantil.

El análisis univariado sobre la categoría socioeconómica permite observar la influencia significativa que esta tiene sobre el puntaje global, sin embargo, al añadir la categoría de IES al análisis, se nota que, aunque la categoría socioeconómica si es estadísticamente significativa, gran parte de la varianza de esta es causada por la categoría de IES y se ve reflejada sobre la variable anterior debido a la sobre representación de estudiantes con mejor categoría socioeconómica en las mejores IES. El árbol de decisión permite concluir que si bien, el estrato socioeconómico importa, su efecto es mínimo relativo al efecto causado por la IES.

La UIS comparada frente a las universidades del segundo grupo tiene una composición socioeconómica dominada por el nivel medio (1), con menor proporción de estudiantes del nivel alto. Comparada con las tendencias en general, las diferencias socioeconómicas si tienen una clara influencia sobre el puntaje individual de cada estudiante.

La opinión de los estudiantes de la UIS en promedio hacia el examen Saber pro es medianamente positiva y no exhiben emociones fuertes hacia él, se manifiesta una disconformidad acerca de su magnitud, falta de congruencia con la vida laboral y carencia de estímulos para un

buen desempeño, y por otro lado se recalca que es un requisito necesario para medir la calidad de la educación. En cuanto al rol de la universidad, demuestran una actitud medianamente positiva, se remarca el acompañamiento en cuanto al proceso de inscripción y a la formación brindada, también se menciona la presencia o falta de cursos para la preparación. Y frente a la experiencia de presentar las pruebas, los entrevistados comunican una opinión ligeramente negativa, se hace mención del cansancio, las jornadas, las condiciones, se menciona que la prueba es larga, estos factores parecen ser los que hacen de la prueba una experiencia no tan amena.

10. Recomendaciones

Para futuras investigaciones, existen oportunidades de aplicar diferentes algoritmos de aprendizaje automático y de mejorar el proceso de ingeniería de atributos

Se recomienda continuar la línea de investigación dentro de la Escuela de Estudios Industriales y Empresariales relacionada con las pruebas Saber y la analítica de datos.

Referencias bibliográficas

50 años del Icfes. (n.d.). Retrieved from https://www.icfes.gov.co/50-icfes

Aissaoui, O. El, El Madani, Y. E. A., Oughdir, L., & Allioui, Y. El. (2019). Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles. Procedia Computer Science, 148, 87–96. https://doi.org/10.1016/j.procs.2019.01.012

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 37(January), 13–49. https://doi.org/10.1016/j.tele.2019.01.007

Alexander, W., Perdomo, M., & Esther, T. (2014). *Factores Asociados Al Rendimiento En Las Pruebas Saber Pro En Estudiantes De Ingeniería Civil En Universidades Colombianas.* Innovaciencia, 2(1), 17–24. Retrieved from http://revistas.udes.edu.co/site/index.php/innovaciencia/article/download/234/pdf_13

Beltran, B. (2016). Minería de Datos. In Benemérita Universidad Autónoma de Puebla.

Berry, M. J. A., & Linoff, G. S. (2000). Mastering Data Mining. New York: John Wiley & Sons, Inc.

Bugueño, F. (2017). Modelo Predictivo para la Selección de Postulantes Destacados a una Institución de Educación Superior (Universidad de Chile). Recuperado de

http://repositorio.uchile.cl/handle/2250/147565

Driscoll, D., Halcoussis, D., & Svorny, S. (2008). Gains in standardized test scores: Evidence of diminishing returns to achievement. Economics of Education Review, 27(2), 211–220. https://doi.org/10.1016/j.econedurev.2006.10.002

Ed, P. P., & Goebel, R. (2013). Advances in Data Mining. https://doi.org/10.1007/978-3-642-14400-4

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery in databases and data mining. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1079, 623–632. https://doi.org/10.1007/3-540-61286-6_186

Frawley, W. J., Piatetsky- Shapiro, G., & Matheus, C. J. (1993). Minería de datos : técnicas y herramientas. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 723 LNAI(3), 407–426. https://doi.org/10.1007/3-540-57253-8_65

García González, J. R., Sánchez Sánchez, P. A., Orozco, M., & Obredor, S. (2019). Knowledge capture for the prediction and analysis of results of the quality test of higher education in Colombia. Formacion Universitaria, 12(4), 55–62. https://doi.org/10.4067/S0718-

50062019000400055

Gil, F. A., Rodríguez, V. A., Sepúlveda, L. A., Rondón, M. A., & Gómez-Restrepo, C. (2013). Impacto de las facultades de medicina y de los estudiantes sobre los resultados en la prueba nacional de calidad de la educación superior (SABER PRO). Revista Colombiana de Anestesiología, 41(3), 196–204. https://doi.org/10.1016/j.rca.2013.04.003

Gobierno de Colombia, & MinTIC. (2016). Guía de estándares de calidad e interoperabilidad de los datos abiertos del gobierno de Colombia. 50. Retrieved from https://herramientas.datos.gov.co/sites/default/files/A_guia_de_estandares_final_0.pdf

Gohil, L. (2015). Text Mining: Process and Techniques. International Journal of Innovative Research in Computer Science and Technology, (3), 70–72.

Gonzalez Montes, C., & Guillen Ibarra, S. A. (2019). Análisis por minería de datos del impacto de los sistemas de calidad de las instituciones de educación superior en los resultados de las pruebas Saber Pro enfocado a los programas de Ingeniería Industrial (Vol. 1). https://doi.org/10.1017/CBO9781107415324.004

González Montes, C., & Guillen Ibarra, S. A. (2019). Análisis por minería de datos del impacto de los sistemas de calidad de las instituciones de educación superior en los resultados de las pruebas saber pro enfocado a los programas de ingeniería industrial /. En Universidad Tecnológica de Bolívar. Recuperado de Universidad Tecnológica de Bolívar website:

https://repositorio.utb.edu.co/handle/20.500.12585/3180

Hearst, M. A. (1999). Untangling text data mining. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics -, 3–10. https://doi.org/10.3115/1034678.1034679

ICFES. (2019). Guía de orientación Saber Pro. 45. Retrieved from http://contratacion.icfes.gov.co/documents/20143/1518930/Guia+de+orientacion+modulos+de+competencias+genericas+saber+pro+2019.pdf/3fe99e8b-229a-c4e8-3aed-f4b719460c51

Jiawei, H., Kamber, M., & Pei, J. (2014). Data mining: Data mining concepts and techniques. In Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013. https://doi.org/10.1109/ICMIRA.2013.45

Jiménez Giraldo, J. (2018). Minería de datos educativos: análisis de los factores económicos, sociales y demográficos que influyen en el desempeño de las Pruebas Saber-Pro en estudiantes de ingeniería en Antioquia. En Universidad Pontificia Bolivariana. Recuperado de Escuela de Ingenierías website: https://repository.upb.edu.co/handle/20.500.11912/4317

Justicia de la Torre, M. del C. (2017). Nuevas técnicas de minería de textos: Aplicaciones. Retrieved from http://hdl.handle.net/10481/46975

Khan, A., & Ghosh, S. K. (2020). Student performance analysis and prediction in classroom

learning: A review of educational data mining studies. En Education and Information Technologies. https://doi.org/10.1007/s10639-020-10230-3

Maimon, O., & Rokach, L. (n.d.). Decomposition Methodology for Knowledge Discovery and Data Mining. In Data Mining and Knowledge Discovery Handbook (pp. 981–1003). https://doi.org/10.1007/0-387-25465-X_46

Martínez Lobo, D. S. (2013). Análisis de la relación entre las pruebas saber pro y cursos realizados por estudiantes de licenciatura en matemáticas utilizando correlación canónica (Vol. 1). Universidad Industrial de Santander.

Miranda, M. A., & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos Analysis of Dropouts of University Students using Data Mining Techniques (Universidad Católica del Norte; Vol. 10). https://doi.org/10.4067/S0718-50062017000300007

Mitov, I., Depaire, B., Ivanova, K., Vanhoof, K., & Blagoev, D. (2012). Automatic Metadata Generation and Digital Cultural Heritage. Access to Digital Cultural Heritage: Innovative Applications of Automated Metadata Generation, (June), 203-.

Mohamad, S. K., & Tasir, Z. (2013). Educational Data Mining: A Review. Procedia - Social and Behavioral Sciences, 97, 320–324. https://doi.org/10.1016/j.sbspro.2013.10.240

Oviedo Carrascal, A. I., & Jiménez Giraldo, J. (2019). Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO. Revista Politécnica, 15(29), 128–140. https://doi.org/10.33571/rpolitec.v15n29a10

Pardo Franco, J. A. (2017). *Factores Demograficos, Academicos Y Socioeconomicos Que Influyen En Los Resultados Del Componente Generico De La Prueba Saber Pro*. Duitama - Boyacá, Colombia: Universidad Pedagógica y Tecnológica de Colombia.

Pardo Franco, J. A. (2017). Factores demográficos, académicos y socioeconómicos que influyen en los resultados del componente genérico de la Prueba Saber Pro Caso : Licenciatura En Matemáticas y Estadística de la Universidad Pedagógica y Tecnológica De Colombia Facultad Seccional Du. En Repositorio de la Universidad Pedagogica y Tecnologica de Colombia. Recuperado de Universidad Pedagógica y Tecnológica de Colombia website: http://repositorio.uptc.edu.co/handle/001/2648

Poveda-Ramos, G. (1993). La ingeniería en Colombia. Historia de Las Ciencias En Colombia. Ingeniería e Historia de Las Técnicas, (1), 35–46.

Ramírez, C. E., & Teichler, T. U. (2014). Factores socioeconómicos y educativos asociados con el desempeño académico, según nivel de formación y género de los estudiantes que presentaron la prueba Saber Pro 2009. 26. Retrieved from https://www.icfes.gov.co/documents/20143/233983/Factores+socioeconomicos+y+educatio

s+asociados+a+desempeno+academico+segun+formacion+y+genero+Saber+Pro+2009.pdf

Rizo, F. M. (2010). Assessment practice in policy context: Latin american countries. International Encyclopedia of Education, 479–485. https://doi.org/10.1016/B978-0-08-044894-7.00302-X

Rodrigues, M. W., Isotani, S., & Zárate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. Telematics and Informatics, 35(6), 1701–1717. https://doi.org/10.1016/j.tele.2018.04.015

Rodríguez Albor, G., Ariza Dau, M., & Ramos Ruíz, J. L. (2014). Calidad institucional y rendimiento académico. Perfiles Educativos, 36(143), 10–29. https://doi.org/10.1016/s0185-2698(14)70607-5

Santín, D., & López, C. (2007). Minería de datos: técnicas y herramientas. Thomson Paraninfo.

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia. doi: http://dx.doi.org/10.16925/9789587600490

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised

data mining techniques for student exam performance prediction. Computers and Education, 143(August 2019), 103676. https://doi.org/10.1016/j.compedu.2019.103676

Vidal-Alegría, F. A., & Timarán-Pereira, S. R. (2019). Análisis de resultados en Pruebas Saber Pro: caso Institución Universitaria Colegio Mayor del Cauca. Ventana Informatica, (38), 51–64. https://doi.org/10.30554/ventanainform.38.2859.2018

Villafañe Blanco, P. V. (2015). Análisis del Desempeño Académico del Examen de Estado para el Ingreso a la Educación Superior Aplicando Minería de Datos. 85. Retrieved from http://www.bdigital.unal.edu.co/51414/1/39004913.2015.pdf

Witten, I. (2004). Text Mining. In The Practical Handbook of Internet Computing. https://doi.org/10.1201/9780203507223.ch14

Wray, J. B. (2016). Abstract Principals' Perspectives on the Effect of Standardized Testing on Teaching and Learning.