

# STEM SIN FRONTERAS DE GÉNERO: UN ANÁLISIS CUANTITATIVO DE LOS CONSTRUCTOS SOCIO COGNITIVOS Y SU IMPACTO EN LA PARTICIPACIÓN FEMENINA EN STEM, APOYADO EN TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

---

## STEM WITHOUT GENDER BOUNDARIES: A QUANTITATIVE ANALYSIS OF SOCIO-COGNITIVE CONSTRUCTS AND THEIR IMPACT ON FEMALE PARTICIPATION IN STEM, SUPPORTED BY MACHINE LEARNING TECHNIQUES

Juan Sebastian Gutiérrez Rivera Maestría en Ciencia de datos y analítica<sup>1</sup>,

<sup>1</sup> Universidad EAFIT sede Medellín Colombia, [jsgutierrr@eafit.edu.co](mailto:jsgutierrr@eafit.edu.co)

**Abstract**– This study, framed within the international 4U project “STEM Without Gender Borders: Strategies for Inclusive Education”, aimed to explore the perceptions of university students enrolled in STEM programs regarding fourteen cognitive-social constructs, with a specific focus on gender differences. The central motivation stems from the persistent underrepresentation of women in STEM fields, a phenomenon documented globally and nationally, which requires innovative approaches that combine quantitative analysis with data science methodologies.

The research was conducted at Universidad EAFIT (Medellín, Colombia) with a sample of 1007 undergraduate engineering students, of which 665 valid responses were analyzed. Data were collected through a structured survey consisting of 97 variables, 67 of which used Likert scales linked to constructs such as self-efficacy, self-regulation, intrinsic and extrinsic motivation, sense of belonging, expectations of success, and perceived compatibility with STEM, among others.


The CRISP-DM methodology was used to guide the analysis process, including data cleaning, internal consistency validation (Cronbach’s alpha), dimensionality reduction using Principal Component Analysis (PCA), and modeling through unsupervised machine learning algorithms (K-means and DBSCAN). Model evaluation was carried out using internal validation metrics such as the silhouette coefficient.

Results showed that clustering techniques did not identify clearly separable groups, as the data tended to form a single semi-elliptical cluster with low silhouette scores ( $<0.5$ ), suggesting high perceptual homogeneity among students. However, inferential analysis of mean differences revealed statistically significant gender-based differences in 9 out of the 14 constructs studied, highlighting dimensions such as self-efficacy, self-regulation, and sense of belonging, where female students reported lower perceptions in some cases.

It is concluded that, although clustering algorithms did not uncover latent subgroups, the study provides evidence of gender-based perceptual differences that should be taken into account for targeted interventions. The significance of these findings lies in their potential to inform more equitable and evidence-based educational policies. The research suggests that female students do not experience a uniform disadvantage, but rather specific differences that can be addressed through construct-specific strategies.

**Keywords**–STEM, Constructos Sociocognitivos, Género, Brecha, Aprendizaje Automático.

# STEM sin fronteras de género: un análisis cuantitativo de los constructos sociocognitivos y su impacto en la participación femenina en STEM, apoyado en técnicas de aprendizaje automático.

Juan Sebastian Gutiérrez Rivera Maestría en Ciencia de datos y analítica<sup>1</sup>,

<sup>1</sup> Universidad EAFIT sede Medellín Colombia, [jsgutierrr@eafit.edu.co](mailto:jsgutierrr@eafit.edu.co)

*Este estudio, enmarcado en el proyecto internacional 4U “STEM sin fronteras de género: Estrategias para una Educación Inclusiva”, tuvo como propósito explorar la percepción de estudiantes universitarios en programas STEM sobre catorce constructos sociocognitivos, con énfasis en las diferencias por género. La motivación central radica en la brecha de participación femenina en áreas STEM, fenómeno documentado a nivel mundial y nacional, cuya comprensión demanda enfoques que integren técnicas cuantitativas y metodologías de análisis de datos.*

*La investigación se llevó a cabo en la Universidad EAFIT (Medellín, Colombia) con una muestra de 1007 estudiantes de pregrado en ingeniería, de los cuales 665 registros válidos fueron analizados. Los datos fueron recolectados mediante un cuestionario estructurado, compuesto por 97 variables, 67 de ellas con escalas tipo Likert asociadas a constructos socio cognitivos relacionados con la percepción de estudiantes respecto al área STEM.*

*Se adoptó la metodología CRISP-DM para estructurar el proceso de análisis, incluyendo la limpieza de datos, validación de consistencia interna (alfa de Cronbach), reducción de dimensionalidad con análisis de componentes principales (PCA) y modelado mediante algoritmos de aprendizaje automático no supervisado (K-means y DBSCAN). La evaluación de los modelos se realizó mediante el índice de silueta.*

*Los resultados revelaron que los modelos no lograron identificar grupos claramente diferenciados, dado que los datos tendieron a agruparse en un solo clúster con una forma semi-elíptica, y bajos valores de silueta ( $<0.5$ ), lo que sugiere una alta homogeneidad perceptiva entre los estudiantes. Sin embargo, el análisis inferencial por diferencia de medias identificó diferencias estadísticamente significativas entre hombres y mujeres en 9 de los 14 constructos, destacando dimensiones como autoeficacia, autorregulación y sentido de pertenencia, donde las mujeres reportaron en algunos casos percepciones más bajas.*

*Se concluye que, aunque los algoritmos de agrupamiento no revelaron subgrupos latentes, el estudio sí evidencia divergencias perceptuales por género. La importancia de estos hallazgos radica en que permiten avanzar hacia políticas educativas más equitativas y fundamentadas en evidencia empírica. Además, sugiere que no existe una desventaja homogénea femenina, sino diferencias específicas que pueden abordarse con intervenciones diferenciadas por constructo.*

**Palabras clave-- STEM, Constructos Sociocognitivos, Género, Brecha, Aprendizaje Automático.**

## I. INTRODUCCIÓN

La participación de las mujeres en programas STEM es consistentemente más baja que la de los hombres. A nivel internacional, haciendo alusión al Reporte global de brechas de género 2024 se puede afirmar que a pesar del incremento continuo de participación desde el año 2016, al comparar la presencia de mujeres en áreas de Ciencia, tecnología, ingeniería y matemáticas STEM– con un 28,2% está aún por debajo de la participación femenina en otros sectores con 47,3% [1]. A nivel nacional, de acuerdo con los datos de estudiantes matriculados 2023 del Sistema Nacional de la Información de la Educación Superior - SNIES se puede apreciar que la proporción de las mujeres que ingresaron a programas STEM fue de tan solo 32,29% en el año. Lo anterior nos alerta sobre la necesidad de seguir analizando las fuentes de esta brecha.

Elementos culturales y sociales inducen de una manera tácita a la población femenina a alejarse de programas relacionados con STEM. En esta línea, algunos autores Identifican tras el análisis de artículos construidos a lo largo de treinta años, causas para este fenómeno: Prejuicios basados en género, creencias sobre lo que se debería ser y tener en el sector productivo, balance familiar y valores de vida diferenciados, inclinaciones y deseos profesionales, debilidades y fortalezas tanto en el cerebro de la mujer como del hombre [2]. Por otra parte, Costa-Lizama y otros investigadores, indican que persisten creencias sociales hacia definir los trabajos en STEM como tema de hombres [3]. Desde esta perspectiva, en el presente estudio se recopiló información sobre la percepción de una muestra de estudiantes de áreas STEM (mujeres y hombres) en una universidad privada de Medellín-Colombia. El instrumento fue construido en torno a constructos socio cognitivos como la autoeficacia, la autorregulación, la motivación, la motivación intrínseca y extrínseca, el sentido de pertenencia.

En este artículo se plantea un estudio con enfoque cuantitativo y se inscribe en un diseño de investigación no experimental de corte transversal analizando 14 constructos socio cognitivos que provienen de la aplicación de 1007 encuestas en una Universidad privada de Medellín a una muestra que incluyó

tanto mujeres como hombres y a partir de los resultados del análisis, derivar algunas recomendaciones para mejorar en relación con la percepción de los constructos que se están estudiando. La pregunta a resolver en el estudio consiste en: al explorar los datos de las encuestas realizadas. ¿Se puede afirmar que hay diferencias significativas entre hombres y mujeres al responder las preguntas del cuestionario asociadas a los constructos de autoeficacia, autorregulación, motivación intrínseca y extrínseca, sentido de pertenencia al programa, el sentido de pertenencia al campo STEM?

Inicialmente se hará la fundamentación y definición de los términos clave necesarios para poder orientarse a lo largo de la lectura del artículo, para luego describir el método bajo el que se desarrolló el estudio incluyendo la obtención de la información por parte de los encuestados. La organización y transformación de la data recolectada y los análisis cuantitativos realizados para extraer el conocimiento contenido en los datos, mediante métodos estadísticos del p-value para demostrar diferencia estadística de las respuestas por género. y técnicas de aprendizaje automáticos no supervisados como los son el análisis de componentes principales y técnicas de agrupamiento como el K-means y DBSCAN. Con esta descripción, se procede a la publicación de los resultados y hallazgos obtenidos del ejercicio de investigación. Posteriormente se emiten las conclusiones y perspectivas relacionadas con el fenómeno bajo estudio. Será valioso para llegar a proponer recomendaciones y acciones pertinentes que ayuden a seguir cerrando la brecha de género.

## II. MARCO CONCEPTUAL

A continuación, se enuncian los conceptos de base para lograr el entendimiento de todo el análisis hecho en este artículo dividido en dos fases: una relacionada con los aspectos sociométricos que puntualmente se miden en el estudio y otra dedicada a las técnicas cuantitativas que se utilizaron en este estudio.

En primer lugar, se hace referencia a la historia y significado del término STEM que según Donahoe, se popularizó gracias a Judith A Ramaley exdirectora del *National Science Foundation* en el año 2001, sustituyendo al poco atractivo acrónimo SMET. [4] En líneas generales el término hace referencia a áreas de ocupación y de conocimiento relacionados con: ciencia (science), tecnología (technology), ingeniería (engineering) y matemáticas (mathematics).

Un segundo concepto importante es el de constructos sociocognitivos, cuya utilidad radica en entender cómo los factores individuales y sociales están involucrados y afectan el comportamiento, el aprendizaje y la toma de decisiones académicas. Según Bandura, el concepto hace referencia a elementos mediadores entre los estímulos del entorno y las respuestas de los individuos [5]. Lo anterior incluye

dimensiones como creencias, actitudes, expectativas y capacidades autorreguladoras [6]. En educación, estos constructos están relacionados con el rendimiento académico, la persistencia, la elección de carrera y la adaptación al entorno universitario [7].

Un constructo sociocognitivo relevante es la autoeficacia, que hace referencia a la creencia de un individuo sobre su capacidad de realización y ejecución de acciones necesarias para alcanzar logros [6]. En áreas STEM, la autoeficacia influye en la elección de carrera, el rendimiento académico y la disposición a enfrentar desafíos. Algunas investigaciones muestran diferencias de género en los niveles de autoeficacia, teniendo en cuenta que STEM tiene áreas tradicionalmente masculinizadas [8].

Otro constructo clave es la **autorregulación**, que significa la habilidad del estudiante para planificar, monitorear y evaluar su propio aprendizaje. Para operativizar es necesario establecer las metas, controlar el esfuerzo, gestionar el tiempo y usar estrategias cognitivas para obtener mejores resultados [9]. Los estudiantes autorregulados son más autónomos, persistentes y normalmente tienen mejor desempeño.

La motivación, entendida como el impulso de una persona para actuar de determinada manera, tiene dos tipos: motivación intrínseca y motivación extrínseca. La primera se refiere al deseo de realizar una actividad por el interés o la satisfacción que esta genera en sí misma; la segunda, induce a la acción acorde a recompensas externas o reconocimiento [10]. Por último, está el sentido de pertenencia, definido como la percepción del estudiante de ser aceptado, valorado e incluido en un grupo o entorno determinado. En programas STEM, este constructo aplicado al programa de pregrado o al campo disciplinar puede determinar la identidad profesional en formación, la continuidad en la carrera y la satisfacción académica [11]. Algunos autores coinciden en que donde se evidencian brechas de género o una cultura institucional excluyente, el sentido de pertenencia puede verse afectado [12], especialmente para las mujeres y minorías subrepresentadas [13].

En resumen, los constructos sociocognitivos antes mencionados no actúan de forma aislada, sino que se interrelacionan en complejas dinámicas que configuran la experiencia educativa de los estudiantes. Estudiar estas variables con enfoque de género permite identificar barreras estructurales y percepciones subjetivas que afectan la equidad y el acceso efectivo a oportunidades en campos STEM.

Ahora es necesario incorporar algunos conceptos de corte más técnico que serán usados en las secciones siguientes para la metodología y los hallazgos que se presentan en este proyecto.

El aprendizaje automático, también conocido como aprendizaje de máquinas, computacional o automatizado (Machine Learning, ML), es una disciplina dentro de las ciencias de la computación y una rama de la inteligencia artificial. Su propósito principal es diseñar métodos matemáticos-computacionales que permitan a los sistemas adquirir conocimientos a partir de datos. Un sistema puede considerar que está aprendiendo cuando, al enfrentar una tarea específica (T) y obtener experiencia (E), su desempeño medido a través de una métrica (P) mejora conforme acumula dicha experiencia [14].

En las técnicas utilizadas de aprendizaje automático se tienen dos grandes grupos: el aprendizaje supervisado y el aprendizaje no supervisado.[15] En el primer caso el objetivo es predecir con la mayor precisión posible el comportamiento de una variable dependiente que se denomina como salida (meta, clase o etiqueta) usando un arreglo de variables independientes, conocidas como entradas, características o atributos. Como ejemplo ampliamente reconocido se menciona la regresión lineal. Esta técnica toma una sola variable dependiente continua y se esfuerza por encontrar la línea recta que mejor la relaciona con las varias entradas  $x_i$  que se le han proporcionado, de modo que las diferencias entre los valores reales y los pronosticados sean las más pequeñas posibles.

En este estudio, el tipo de técnica de aprendizajes a utilizar pertenece al conjunto complemento de técnicas no supervisadas, para este no se requiere la variable de respuesta y de salida. Recibe el nombre de no supervisado porque este tipo de algoritmos permite que la máquina de cómputo identifique relaciones y patrones complejos sin la necesidad de especificaciones puntuales por parte del usuario. [16]

En lo relativo a la parte técnica, es necesario iniciar con la metodología que enmarca la manera en que se realizaron los diversos procesos de interacción con los datos recolectados durante el estudio, la cual corresponde a la denominada *Cross Industry Standard Process for Data Mining* (CRISP-DM). La cual consiste en un proceso lógico que permite extraer conocimiento a partir de una colección de datos sin estar limitada a un área de negocio o área de conocimiento determinada con una serie de pasos estandarizados que promueven la rapidez, estandarización, ahorro en costos y la escalabilidad tanto en los estudios con conjunto de datos de pequeña cantidad como en los conjuntos de datos muy grandes.[17] (ver Fig. 1)

Teniendo en cuenta la pregunta que se desea resolver en la investigación, se utilizaron técnicas de aprendizaje automático no supervisados de agrupamiento o clustering, que se pueden definir a grandes rasgos como la técnica de extracción de conocimiento a partir de datos que agrupa a las observaciones en categorías que tienen características o rasgos que son similares entre sí, pero diferentes entre grupos. Este tipo de

análisis se puede trabajar en casos que involucren muestras de: personas, bienes, transacciones comerciales. [18] Dentro de la diversidad de técnicas que pertenecen a este conjunto de algoritmos de aprendizaje automático se seleccionaron dos: Técnica de clusterización con K-means y el algoritmo de agrupamiento DBSCAN

El algoritmo K-means tiene como finalidad dividir un conjunto de datos en K grupos distintos, tomando como criterio principal la similitud entre los puntos de datos. Su objetivo esencial es minimizar la suma de las distancias al cuadrado entre cada punto y el centroide del clúster al que pertenece.

El proceso inicia con la selección aleatoria de K puntos en el espacio de características, que actuarán como centroides iniciales. Luego, cada observación del conjunto es asignada al centroide más cercano utilizando una medida de similitud, como la distancia euclidiana. Posteriormente, los centroides se vuelven a calibrar calculando la media de los puntos asignados a cada grupo. Esta asignación y actualización se repite de forma iterativa hasta que los grupos dejan de modificarse significativamente o se alcanza un número máximo de iteraciones previamente definido. [19]

DBSCAN es un algoritmo base para la agrupación basada en densidad. Tiene la capacidad de detectar clústers de diferentes formas y tamaños dentro de grandes conjuntos de datos que pueden contener ruido y valores atípicos (outliers). Sin embargo, una de sus limitaciones es que no logra manejar adecuadamente la variación de densidad local que puede existir dentro de un mismo clúster. El algoritmo emplea dos parámetros clave:  $\epsilon$  (épsilon), que define el radio de vecindad para cada punto, y la cantidad mínima de puntos necesarios dentro de ese radio para considerar que existe una región densa

Funciona iniciando en un punto arbitrario no visitado, explorando su vecindad  $\epsilon$ , y según si contiene suficientes puntos, decide si comienza un nuevo clúster o lo marca como ruido.[20] Una propiedad clave de DBSCAN es que no requiere especificar previamente el número de clústeres, y puede descubrir agrupamientos de forma arbitraria, así como identificar outliers, haciéndolo especialmente útil para conjuntos de datos grandes y no lineales

Para estimar el número óptimo de clústeres y evaluar la pertinencia de los grupos que el algoritmo encuentra Bishop. Explica tres métodos para esta labor, no obstante, en el alcance de este estudio, sólo se utilizarán dos de ellos, que se explican a continuación.

El método del codo, que se basa en graficar la suma de cuadrados intra-clúster (SSW) frente al número de clústeres. El “codo” de la curva —el punto donde la reducción de la SSW se desacelera significativamente— indica el número óptimo de grupos, equilibrando la cohesión interna y la separación

externa, en adición, el silhouette score que evalúa cada observación calculando la diferencia entre la distancia promedio al resto en su mismo clúster y la distancia promedio al clúster más cercano. Un valor alto del índice de silueta señala una buena asignación de los puntos y clústeres bien diferenciados, se maneja como buen valor aquel puntaje que supere el 0.5. [21]

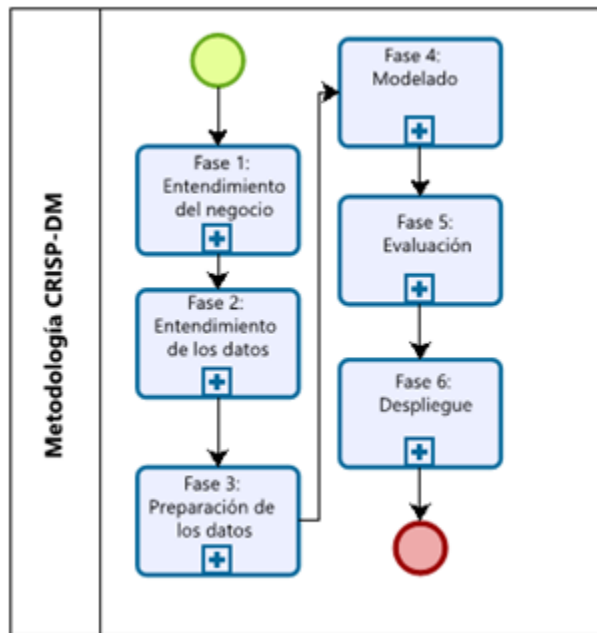


Fig. 1: Pasos de la metodología CRISP-DM

La estrategia para poder medir los constructos en las encuestas realizadas radica en el uso de varias preguntas asociadas a un mismo constructo de manera que permita obtener una respuesta consistente en torno a la percepción de cada uno de los encuestados para cada uno de los constructos, más allá de la intención y de la construcción de la encuesta con el objetivo anteriormente descrito, se precisó en este estudio del uso de un criterio cuantitativo que permita afirmar que las preguntas asociadas a su respectivo constructo son homogéneas y tienden a medir la misma característica. En ese orden de ideas, el criterio seleccionado fue el cálculo del alfa de Cronbach ( $\alpha$ ), índice que permite establecer el grado de homogeneidad entre ítems respecto a la evaluación de una característica. [22]

Para el cálculo de la variable de Cronbach, se utiliza la cantidad de ítems a evaluar y el valor de las varianzas en las respuestas producidas por los sujetos encuestados. [23]

$$\alpha = \frac{K}{K-1} \cdot \left[ 1 - \frac{\sum_{i=1}^K V_i}{V_t} \right] \quad (1)$$

Donde:

$\alpha$ : alfa de Cronbach.

$K$ : número de ítems.

$V_i$ : Varianza de cada ítem.

$V_t$ : Varianza del total.

Los valores que permiten evaluar la consistencia entre reactivos con el alfa de Cronbach se manejan de la siguiente forma Si el resultado es  $\alpha \geq 0.70$  se considera Aceptable, si el dato  $\alpha \geq 0.80$  se considera Bueno y en el caso de  $\alpha \geq 0.90$  se considera Excelente.

En aras de sintetizar la información de los diversos reactivos asociados en cada uno de los constructos socio cognitivos estudiados en este artículo, en lugar de simplemente eliminar variables que representan las preguntas perdiendo información valiosa o calcular la media aritmética ignorando la ponderación de acuerdo a la variabilidad explicada por cada una de las variables. Se tomó la decisión de conseguir esta síntesis deseada por constructo a través del análisis de componentes principales (PCA) dada la capacidad de esta técnica de minería de datos no supervisada de transformar conjuntos de variables posiblemente correlacionadas en una cantidad de menor cardinalidad que el conjunto inicial no correlacionadas llamadas componentes. Mediante estos es posible resumir la información contenida en los datos con un número menor de características que en su conjunto explican la mayor parte de la variabilidad de los datos.[24]

Las características de los datos a que se trabajaron en este artículo contemplaron la agrupación de cierta cantidad de reactivos para consultar a los sujetos encuestados sobre once de los catorce constructos socio cognitivos de este estudio. Para poder resumir las respuestas de las múltiples preguntas (por ejemplo, las relacionadas con autoeficacia que son 9 ítems) en una sola columna capaz de representar la mayor cantidad posible de la variabilidad de las respuestas obtenidas, se tomó la decisión de emplear el análisis de componentes principales (PCA). Esta técnica consiste en buscar un nuevo conjunto de ejes (o "direcciones") en los datos que sean totalmente independientes (normales = 90°) unos de otros. Estos ejes se priorizan por la cantidad de dispersión o variación que representan. La primera componente principal es la que captura la mayor cantidad de variabilidad de los datos muestreados, lo que significa que es la que mejor discrimina entre las distintas opciones. El segundo componente principal incluye dentro de sí el segundo mayor grado de variabilidad y, al ser ortogonal a la primera, capta información que no está ya presente en la primera. Este proceso continúa para las siguientes componentes, cada una aportando nueva información y siendo ortogonal a las anteriores [25]. La técnica tiene una gran ventaja, ya que, como herramienta descriptiva, no requiere de supuestos relacionados con una distribución estadística específica, lo que permite usar esta técnica para exploración de los datos cuantitativos en una amplia gama de contextos [26].

Ahora, posterior al abordaje de algunos conceptos de base para la investigación, en lo que queda de esta sección se presentarán

antecedentes enfocados en estudios que utilizan el análisis de datos y el aprendizaje automático para explorar constructos sociocognitivos en contextos STEM con perspectiva de género.

La integración de técnicas de análisis de datos y aprendizaje automático (ML) en investigaciones sobre género y constructos sociocognitivos en STEM sin duda alguna, ha enriquecido la comprensión de este fenómeno y la detección de patrones complejos y barreras invisibles que posiblemente no son visibles con otro tipo de técnicas y enfoques. A continuación, se presentan algunos estudios interesantes al respecto.

En EE. UU. se han efectuado interesantes estudios con estudiantes de ingeniería, empleando diversos métodos estadísticos y exploratorios. Uno de los hallazgos importantes es que las mujeres tienen niveles más bajos de autoeficacia comparativamente hablando con los hombres, a pesar de que su rendimiento académico es igual o mejor, y esto es más notorio en disciplinas como la física, las matemáticas y la química [27], [28].

En otros estudios se usaron modelos de regresión y path análisis para analizar si el ambiente de aprendizaje modifica las creencias de autoeficacia, interés y sentido de pertenencia. Esto se hizo en cursos introductorios de física. Uno de los hallazgos fue que el reconocimiento percibido amplifica las diferencias de género [29].

Otro estudio de 2025 mostró la relación existente entre ansiedad, autoeficacia y rendimiento académico. El experimento fue ejecutado en cursos de física para biociencias y se abordó desde el análisis predictivo y ML explicativo (por ejemplo, SHAP). Efectivamente, las mujeres experimentaron mayores niveles de ansiedad y menor autoeficacia en evaluaciones de alto impacto.

A nivel de secundaria, también fue posible encontrar investigación en el tema. Mediante técnicas de clustering (TwoStep Cluster Analysis) se logró la identificación de perfiles motivacionales diferenciados por género en STEM. Hubo agrupamientos mixtos combinando niveles intrínsecos y extrínsecos de motivación. Lo anterior generó evidencia de brechas de género en valores percibidos y autoevaluación [27]. Un estudio en Arabia Saudita empleó SEM multigrupo para analizar correlaciones entre autoeficacia, aspiraciones de carrera y autopercepción en 671 mujeres. Confirmó que la “Social Cognitive Career Theory -SCCT” se mantiene en contextos no occidentales, y ratificó que la autoeficacia predice el deseo de seguir en carreras STEM [30].

Otros autores utilizaron codificación cualitativa + ML explicativo para identificar si el reconocimiento recibido en cursos de física está vinculado con la autoeficacia y persistencia de las mujeres, subrayando la importancia de las interacciones instructor estudiante [31].

Algunos elementos comunes en los estudios mencionados son:  
1) Uso de ML y modelos estadísticos avanzados (regresión,

clustering, SEM, interpretabilidad con SHAP). Esto posiblemente se debe a que son enfoques acertados para explorar relaciones no lineales entre género y constructos sociocognitivos; 2) diversidad en áreas de conocimiento (ingeniería, física, biociencias), lo cual facilita la generación de patrones entre disciplinas técnicas STEM; 3) las variables involucradas y que hacen parte de constructos sociocognitivos (autoeficacia, motivación (intrínseca/extrínseca), ansiedad, sentido de pertenencia e identidad de campo); 4) Orientación e interpretación de los agrupamientos generados, que son útiles para identificar mecanismos y grupos vulnerables.

Por otra parte, en el contexto del caso a nivel nacional, el Ministerio de Educación MEN– junto con la Organización de Estados Americanos -OEA- y una fundación auspiciada por Siemens, hacen esfuerzos para aumentar la cantidad de mujeres en carreras STEM, mediante la iniciativa STEAM + Género. Ésta propone orientaciones prácticas para que directivos y docentes promuevan aulas inclusivas, reduzcan los estereotipos y faciliten la integración de la perspectiva de género desde la educación inicial [32]. La autora Guevara-Ramírez plantea 10 reglas para empoderar a las mujeres en STEM artículo realizado en el contexto de un país latinoamericano. Las normas sugeridas son: evitar el “efecto Matilda”; empoderar a otras mujeres a través de la solidaridad entre ellas; colaborar con otros grupos de investigación, organizaciones e instituciones. Regla 4: Investigar y publicar. Regla 5: Pide ayuda. Regla 6: Acaba con los prejuicios. No te sientas diferente por ser mujer. Regla 7: Fomentar la educación STEM. Regla 8: Equilibrar el tiempo para desempeñar roles en la sociedad y la familia. Regla 9: Recuerda que nunca es demasiado tarde. Regla 10: Recuerda tus capacidades y la fuerza que tienes. [33]. Iniciativas y propuestas con potencial para continuar en el camino de cerrar la diferencia en la participación de mujeres en el área de conocimiento.

### III. METODOLOGÍA

Para maximizar la confiabilidad y pertinencia de los resultados conseguidos en el presente estudio se aplicarán buenas prácticas de extracción de conocimiento a partir de los datos, mediante el desarrollo de la metodología CRISP DM

Esta se toma como base debido a su madurez y uso masivo en el sector productivo, CRISP-DM permite la extracción de conocimiento del negocio y de los datos, ofrece una guía para la preparación de los datos, el modelado y la validación de los modelos lo que resulta en la creación de un proceso reproducible y escalable para recolectar y analizar datos, teniendo en cuenta nuevas relaciones entre variables y mejoras en la forma de experimentar con ellas, con retroalimentaciones constantes de cada una de las fases del ejercicio.

La metodología CRISP-DM está conformada por seis pasos iterativos y con posibilidad de retornar de manera cíclica a



pasos anteriores de acuerdo con las necesidades del proyecto en el que se esté trabajando.[34]

TABLA I  
DESCRIPCIÓN DE PASOS DEL MODELO CRISP-DM [34]

PASOS	DESCRIPCIÓN
Entendimiento del negocio	En este paso inicial se busca entender el contexto en el que se va a trabajar y establecer cuál será el objetivo del proyecto, definiendo el tipo de análisis que se requiere y estableciendo los criterios de éxito esperados, acorde a la situación o negocio estudiado.
Entendimiento de los datos	En este paso se recopilan los datos disponibles y se analizan para conocer su estructura y su calidad. Es primordial explorar las variables, las observaciones, su comportamiento y cómo se relacionan entre sí, utilizando para esta labor herramientas estadísticas.
Preparación de los datos	Se lleva el conjunto de datos “crudo” a una disposición de los datos que puedan ser utilizados en los análisis requeridos, esto implica la selección de las variables relevantes de los criterios de selección que deben cumplir las muestras para pertenecer al conjunto de datos a procesar, y aplicar los métodos necesarios (escalamiento, OneHotEncoding, Labeled Encoding, bins) en función de las técnicas para hacer el análisis.
Modelado	Se elige(n) la(s) técnica(s) de modelado más adecuada(s) al problema de negocio y a los datos disponibles. Se construye y prueba el modelo, haciendo la sintonización de parámetros y seleccionando el mejor resultado según criterios de evaluación.
Evaluación del modelo	Los resultados del modelo se comparan con los objetivos planteados en la primera fase. Se analiza si las metas se alcanzaron y si el modelo es útil para tomar decisiones, además de revisar el proceso completo.
Despliegue del modelo	Finalmente, se pone en marcha la solución. Esto puede implicar la generación de un informe o la implementación de un componente de software. También se podrían contemplar tareas de monitoreo y mantenimiento.

A. Entendimiento del Contexto.

Este estudio fue realizado en la Universidad EAFIT, en Medellín, Colombia bajo el proyecto de la alianza 4U “STEM sin fronteras de género: Estrategias para una Educación Inclusiva” llevada a cabo el *Institute for the future of Education* del Tecnológico de monterrey, México junto con las universidades EAFIT, ICESI y la Universidad del Norte. El macro proyecto en el que el presente estudio se anida, tiene como objetivo definir recomendaciones para la enseñanza-aprendizaje, las políticas institucionales y de la industria mediante un análisis estructurado para la creación de entornos de aprendizaje equitativos e inclusivos con perspectiva de género.

La menor participación de mujeres en áreas relacionadas con STEM es una preocupación de orden mundial que tiene implicaciones económicas, sociales y políticas, y se acentúa en las áreas específicas de ciencias computacionales, ingeniería y física. A su vez, este fenómeno ocurre en los países sin importar

el nivel de equidad social, como ejemplo se cita el caso de Finlandia, país que lidera el ranking de indicador de equidad de género del Foro Económico Mundial, pero que paradójicamente tiene una de las mayores brechas de género en los títulos universitarios en los campos STEM. Argumentos numéricos que avalan las afirmaciones realizadas, se presentan en los Estados Unidos de América en el que la fundación nacional de ciencias reporta que aunque las mujeres obtienen el 57% de todos los títulos en EEUU relacionados con biología, química y matemáticas desde finales de los 90’s, ellas no representan más del 20% de los títulos obtenidos en áreas de la ciencia computacional, ingeniería y matemáticas aplicadas, en adición, sólo una cuarta parte de doctorados en matemáticas y estadísticas con obtenidos por mujeres. [35]

En la misma línea de argumentos numéricos en Colombia de acuerdo con los datos a corte 2023 del SNIES. Se puede apreciar que la proporción de las mujeres que ingresaron a programas STEM del sistema de educación superior colombiano fue de tan solo 32,29% en ese año.

Por otra parte, los datos consultados en SNIES en términos de estudiantes que se gradúan del sistema de educación superior colombiano en programas de áreas CINE de campo amplio: 1. Ciencias naturales, matemáticas y estadística, 2. Ingeniería, Industria y construcción. y 3. Tecnología de la información y comunicación (TIC). Que se tomaron como programas relacionados con STEM, se evidencia que Colombia no es ajeno a la infra participación de las mujeres en estos programas a lo que Cherney describió en su estudio de 2023.

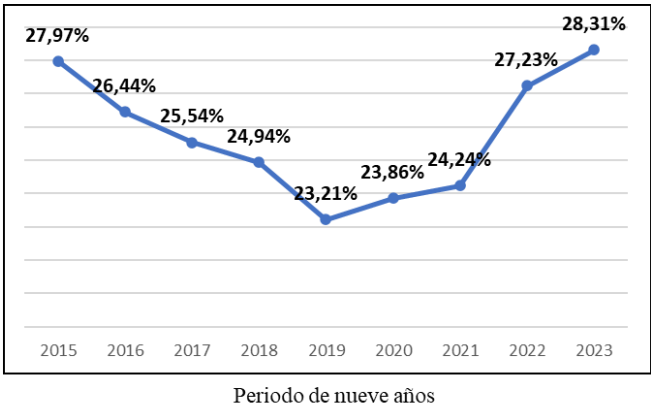


Fig. 2: Porcentaje de graduadas femeninas en programas STEM respecto al total de graduados en Colombia periodos 2015 al 2023. [36]

En los últimos 9 años con data disponible de los graduados en Colombia de programas STEM pese a que la participación de mujeres va en aumento en los últimos años, se evidencia que nunca ha alcanzado el 30% de participación respecto del total.

El caso más dramático se aprecia en las carreras del área CINE asociados a programas de Tecnología de la información y comunicación (TIC) en el que a lo largo de los nueve años del registro en los que en los que a duras penas ha superado el 20% de participación y de hecho en el año más reciente (2023) las

mujeres fueron sólo el 17% de los 4642 graduados en esa área, tasa muy baja de mujeres aún. [36]

En la literatura consultada se encontraron varios factores que pueden ser causas de este efecto presentado con anterioridad, en este artículo se trabaja sobre la tendencia de asociada a las personas en el momento que toman decisiones profesionales basándose en parte en sus fortalezas personales en lo que la literatura consultada denominada como la teoría experiencia - valor del neurocientífico John Eccles, y tras un experimento realizado para ver el impacto en el área de las matemáticas, se observaron diferencias de género en cuanto a las creencias y el logro profesional, siendo los estudiantes varones más propensos que las mujeres a seguir y lograr carreras relacionadas con las matemáticas. Aunque las diferencias mencionadas de género no se pudieron explicar por las diferencias en las creencias sobre las matemáticas como materia académica. [37]

Los resultados presentados en este artículo hacen parte de la primera etapa del proyecto que es el marco referencial de este artículo, cuyo objetivo es realizar un estudio cuantitativo de catorce (14) constructos socio cognitivos entre estudiantes de carreras STEM de la Universidad EAFIT mediante técnicas de análisis cuantitativo y aprendizaje automático.

Para la realización del estudio se cuenta con recursos aprobados en uno de los centros de costos de la universidad EAFIT, un equipo de investigadoras y estudiantes de maestría entre otros colaboradores. Dentro de los riesgos contemplados a la hora de ejecutar las actividades era no obtener la colaboración de los colegas profesores para tomar tiempo de la clase que permitiera a los estudiantes diligenciar las encuestas, a su vez, que un porcentaje alto, superior al 50% de los encuestados tuvieran la condición de ser menores de edad, lo que provocaría por el código de ética del estudio, no proceder con la recolección, uso y publicación de sus datos incluso siendo anonimizados. Otro riesgo durante el estudio es que además de los menores de edad, hubiera contaminación de datos de personas que no fueran de programas STEM o que un porcentaje alto de encuestados abandonaran el instrumento por no sostener entre 12 y 14 minutos que requerían para completar todas las preguntas.

Para minimizar la materialización de riesgos de respuestas nulas e incompletas se decidió escoger cursos de asignaturas de todos los semestres de los distintos programas seleccionados, sin hacer un sesgo hacia primer semestre, zona en donde se esperaba tener una mayor concentración de estudiantes menores de edad. En adición se hizo la inversión en incentivos para otorgar a los estudiantes previa comprobación de encuesta totalmente diligenciada en la pantalla de los dispositivos en que los estudiantes realizaron el ejercicio. Respecto a la mitigación de riesgo de contaminación de los datos, las visitas a salones se hacían en función de la programación de las aulas que tuvieran asignaturas pertenecientes a los programas STEM, y en los casos de que fuese alguna asignatura transversal o en la que pudieran inscribirse otros programas, el encargado de recoger las respuestas exigía conocer en qué programas estaban

matriculados los estudiantes indicando en función de sus respuesta cuáles de los alumnos eran los que debían diligenciar el instrumento.

El resultado deseado de este ejercicio de minería de datos es poder visualizar descubrimientos acerca de cómo es el comportamiento de la percepción de los constructos asociados al área STEM, en los en los estudiantes de este tipo de programas en la Universidad EAFIT y analizar si existen diferencias significativas del comportamiento de las variables entre los géneros (entre mujeres y hombres). El cumplimiento de este objetivo permitirá contar con evidencia y con un punto de partida para análisis posteriores relacionados con la construcción de recomendaciones y proposición de políticas basadas en argumentos objetivos que en el plazo más corto posible permita la reducción de la brecha de género en los estudiantes de programas STEM.

## *B. Entendimiento de los datos.*

Para el desarrollo de este artículo se recibe el conjunto de datos producto de la aplicación del instrumento de recolección de datos (encuesta), titulado *“Desde STEM hasta mi programa académico: una mirada personal”*. Aplicado a estudiantes de la Universidad EAFIT en su sede de Medellín de los programas: Biología, Diseño Urbano y Gestión del Hábitat, Geología, Ingeniería Agronómica, Ingeniería Civil, Ingeniería de Diseño de Producto, Ingeniería de Procesos, Ingeniería de Producción, Ingeniería de Sistemas, Ingeniería Física, Ingeniería Industrial, Ingeniería Matemática, Ingeniería Mecánica.

La encuesta señaló a los sujetos preguntados que su objetivo consistía en explorar las perspectivas de los estudiantes sobre su desarrollo personal y su experiencia en los programas académicos de las áreas de ciencias, ingeniería, tecnología y matemáticas (STEM, por sus siglas en inglés).

### *B.1. Descripción del cuestionario*

El cuestionario fue aplicado en una herramienta digital con tabulación automática al momento de que el sujeto encuestado respondía las preguntas y accionaba un botón indicando el cierre de su intento, el cuestionario estaba para cumplir con el compromiso de no trabajar con datos de menores de edad, así que si un estudiante digitaba su edad menor que 18 años el instrumento le agradecía su participación y le llevaba por fuera del cuestionario.

Constó de 97 recolectores de características o variables, 13 que tomaban datos de manera automática cómo por ejemplo geolocalización de la persona en el momento de las respuestas, duración de tiempo dentro de la encuesta, ID generado, entre otros las 84 preguntas restantes tomaban los datos producto de la opción señalada o la respuesta del o de la estudiante digitada por teclado. De las cuales 10 se digitaban por teclado y las 74 restantes se obtenían por selección de opción múltiple.



Las respuestas a las preguntas asociadas a los 14 constructos socio cognitivos estaban organizadas en escala de Likert de 5 opciones y de 7 opciones. Fueron en total 67 preguntas de esta clase.

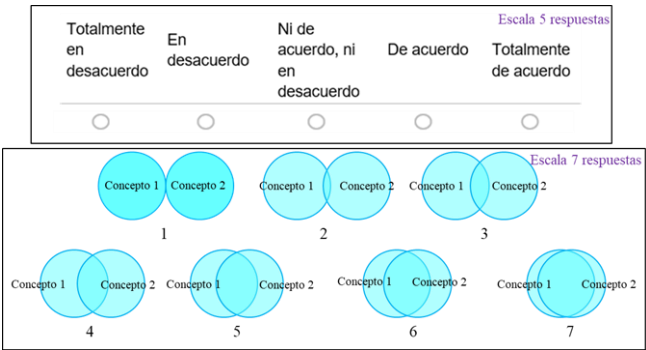


Fig. 3: Tipos de respuesta en escalas de reactivos asociados a la autopercepción de los constructos socio cognitivos en el instrumento.

Preguntas en donde la respuesta esperada era una cadena de caracteres contó con una cantidad de 9. Sólo hubo una (1) pregunta en donde se esperaba un valor numérico. Las otras preguntas que son de opción múltiple pero que no están en escala Likert son en total 7.

### B.2. Visita a salones de asignaturas de programas STEM

Con base en el reporte de la ocupación de los ambientes físicos de aprendizaje de la Universidad EAFIT se procedió a hacer el filtro con los criterios necesarios para que se pudieran visitar salones con asignaturas de los programas STEM de la institución y buscando poder recolectar datos en preferencia de semestres mayores que el primero para reducir la posible respuesta de estudiantes menores de edad que el motor de encuesta en el que se ofrecía el instrumento, descartó de manera automática reduciendo el número de muestras recogidas por visita.

En cada visita se procedía con la solicitud de autorización por parte del profesor encargado, una vez obtenida, se daba el saludo a los estudiantes, se presentaba la persona recolectora de los datos, explicaba de manera general quiénes eran las entidades responsables del estudio, la justificación del estudio, cómo sus respuestas eran valiosos aportes y se preguntaba a los asistentes para control, ¿A qué programa pertenecían? para dar la instrucción a los estudiantes idóneos de que diligenciaran el cuestionario y restringir la participación de estudiantes que no fueran de programa STEM de la Universidad.

Luego de esa introducción se compartía el QR mediante la lectura desde el dispositivo del encuestador o se compartía con ayuda del proyector del salón, en el transcurso de los minutos en que los estudiantes diligenciaron el instrumento, el encuestador permanecía en el aula como facilitador ante cualquier duda del estudiante. Al terminar las preguntas los estudiantes se acercaban al encuestador, mostraban la pantalla indicando la finalización a satisfacción de las preguntas y este procedía a entregar a dicho estudiante el incentivo dispuesto por

ello, la herramienta digital donde se dispuso la encuesta se encargaba de tabular las respuestas de las personas encuestadas.

### B.3. Exploración de los datos resultantes.

Al término del periodo de aplicación del instrumento en los diversos salones de los programas relacionados con STEM de la Universidad EAFIT y luego de la espera para que la universidad representante de la alianza 4U compartiera los datos se realizó la exploración de los datos resultantes del ejercicio.

En la Universidad EAFIT se cuentan con 3163 estudiantes con matrícula activa en los programas relacionados con STEM relacionados en el primer párrafo de esta sección de los cuáles 2132 son Hombres y 1031 son mujeres, a partir de esta población, y luego de aplicar el instrumento a lo largo de dos meses y medio la recolección de muestras sin procesar arrojó la siguiente característica. Hubo en total 1007 sujetos encuestados, de los cuales 493 fueron hombres y 253 fueron mujeres, algo interesante por señalar radica en el hecho de la opción incluida “*Prefiero no contestar*” en la pregunta que indagaba por el sexo biológico y que efectivamente fue utilizada por 6 de los estudiantes encuestados, en adición 255 respuestas fueron de facto valores nulos pues se le dió apertura al instrumento, pero no se pudo diligenciar una sola pregunta, se puede suponer que hubo menores de edad con la intención de entregar su percepción sobre los constructos estudiados pero al colocar la edad el instrumento les dió por terminado el intento, otro hecho a destacar fue la consistencia en el balance de hombres y mujeres tanto en la población como en la muestra con una relación 2:1 en favor de la cantidad de hombres activos en programas STEM de la Universidad EAFIT.

Se destaca que hubo coincidencia en los dos programas que tuvieron la mayor participación tanto de mujeres como hombres y fueron: *Ingeniería de Sistemas e Ingeniería Civil*. con Ingeniería de Sistemas 228 e Ingeniería Civil 114 para los hombres, por su parte, las mujeres con Ingeniería de Sistemas 88 e Ingeniería Civil 46. En términos de la edad de los encuestados se encuentra que la distribución de personas con intención de responder la encuesta tiene un dominio desde los 16 años hasta los 45 años, además se identifica que 86 personas no diligenciaron esta pregunta, se presenta la gráfica sobre los datos brutos recogidos analizando la variable edad para ejemplificar de manera más clara su distribución.

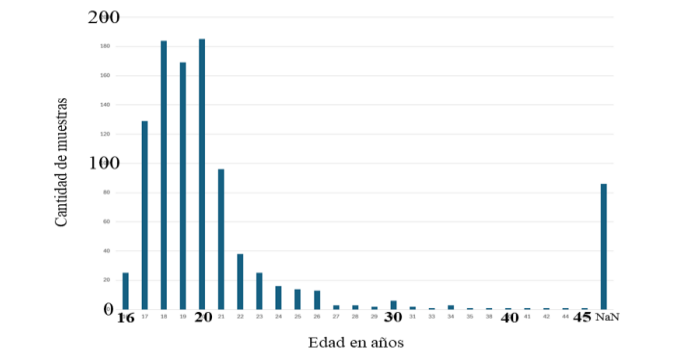


Fig. 4: Distribución de edades en participantes de la encuesta.

Otra característica curiosa se ubica en la pregunta *¿Tienes algún tipo de beca para realizar tus estudios?* en la que los encuestados que respondieron la pregunta y no dejaron el campo vacío respondieron casi en la misma proporción (50-50) tanto los que no tienen beca como los que sí tienen, con un total de 393 muestras indicando que no y 359 muestras indicando que si tienen.

Es pertinente destacar que la tabla en bruto recibida de la institución líder del proyecto 4U tiene unas dimensiones de 97 columnas y 1007 filas de respuestas a la encuesta, a su vez, con presencia de variables que se responden con opción múltiple en las que se esperan dominios discretos en sus respuestas y otras variables fruto de preguntas abiertas en la que los participantes tenían la libertad de colocar cadenas de caracteres libres para desarrollar sus percepciones. Ahora bien, en términos de calidad de los datos y para los próximos pasos del método CRISP-DM, se debe tener en cuenta la presencia de una cantidad considerable de datos faltantes que se podrían considerar en varios subconjuntos: un primer conjunto en el que no se obtuvo ni siquiera una sola respuesta, un segundo conjunto en el que parcialmente se diligenciaron las preguntas de opción múltiple y un tercer grupo que respondió a todas las preguntas de opción múltiple pero dejaron en campo vacío todas las preguntas abiertas.

Tener en cuenta también las preguntas que tienen las siete opciones en escala Likert, respuestas que perfectamente pueden ser de tipo entero pero que están guardadas como punto flotante con decimal igual a cero lo que puede generar valores anómalos al exportarlos en formatos para ser utilizados en distintas herramientas de análisis.

A manera de resumen el conjunto de datos resultante de la encuesta consta de 97 variables, organizadas en diversos tipos de datos que reflejan la complejidad y riqueza del instrumento aplicado. Predominan ampliamente las variables categóricas ordinales (70), Evidenciando la afirmación realizada anteriormente sobre el diseño basado en escalas tipo Likert, orientado a la medición de percepciones, actitudes y creencias. Este núcleo se complementa con variables de tipo cadena de caracteres (10), categóricas nominales (7), además de datos numéricos (decimales y enteros), temporales (DateTime), una variable booleana, y una expresada en porcentaje, lo que evidencia una estructura mixta con potencial tanto para análisis descriptivos como inferenciales.

La dimensión conceptual del instrumento se estructura alrededor de 14 constructos sociocognitivos, entre los que destacan: Autoeficacia (9 ítems), Autorregulación (6), Expectativas de valor percibido en STEM (8), Sentido de pertenencia al programa (8), Expectativas de éxito en STEM (7), y Compromiso académico (5). Estos constructos permiten abordar de manera integral factores psicológicos, motivacionales y contextuales relacionados con la experiencia de los estudiantes de la Universidad EAFIT en programas STEM. La organización del instrumento facilita el análisis y el modelamiento basado en estructuras latentes, constituyéndose

en una base robusta para la interpretación y validación empírica de los resultados.

### *C. Preparación de los datos.*

Una vez entendido el contexto de donde se toma la información, sabiendo que el objetivo es poder visualizar descubrimientos acerca de cómo es el comportamiento de la percepción sobre los constructos socio cognitivos en los estudiantes de programas STEM de la Universidad EAFIT y analizar si existen diferencias significativas del comportamiento de las variables entre los géneros (entre mujeres y hombres) y caracterizando y entendiendo los datos obtenidos producto de la aplicación del instrumento de recolección a 1007 estudiantes de la Universidad, se apertura el proceso de limpieza, ingeniería de características y transformaciones necesarias para poder utilizar los cálculos, modelos y análisis necesarios que permitan extraer efectivamente los insights contenidos en la información recabada.

#### *C.1. Selección de los datos a incluir en los modelos.*

Para seleccionar las muestras que cumplen las condiciones de ser válidas y aportar información relevante para el estudio, los criterios fueron los siguientes: primero, se eliminaron las muestras correspondientes a los menores de edad para cumplir uno de los puntos del código de ética del estudio, en segundo lugar, se eliminan los registros que tienen la variable edad con dato vacío, se decide no hacer imputación de datos en esta situación porque esas muestras, a su vez, carecían de valores en otros campos, en tercer lugar, se tomó la decisión de solamente tener en cuenta las variables asociadas a los constructos con múltiple opción y única selección de respuesta, por lo tanto, así los estudiantes encuestados que no contestaron las preguntas abiertas al final del cuestionario siguen teniendo la información suficiente para utilizar en el presente artículo, para ello en la fase de exploración se evalúan los valores de la variable *Progreso* y se concluye la relación de que valores mayores o iguales que 76 en esta columna tenían la característica de tener todas las preguntas respondidas a excepción de las preguntas abiertas mencionadas anteriormente, así pues, se sustraen del conjunto de datos original las muestras que en la variable *Progreso* eran inferiores a 76.

Por último, se descubre que persisten 4 filas sin las respuestas mínimas necesarias para continuar en el estudio ya sea por datos nulos o faltantes, se individualizaron, por tanto, los registros *ID de respuesta* anónimos de estudiantes que correspondieron a los valores: ('R\_7lSrkdBTUeNyNFv', 'R\_5tLSJarq6XpSzgC', 'R\_6S7SnL86wkW3bUg', 'R\_3RXWSqsPp2FsRvb') y fueron excluidos del conjunto de datos. La tabla original de dimensiones (1007, 97) pasó a ser de (665, 97) dimensiones.

Luego de eliminadas las muestras no relevantes para el estudio, llegó el turno de seleccionar las variables-columnas que no continuarían siendo parte del dataset, el criterio para la

escogencia de variables fue eliminar las que tenían relación con el acto de diligenciar la encuesta, de esta manera, variables como: *Fecha de inicio*, *Fecha de finalización*, *Tipo de respuesta*, *Dirección IP*, *Progreso*, *Duración (en segundos)*, *Finalizado*, *Fecha registrada*, *Latitud de la ubicación*, *Longitud de la ubicación*, *Canal de la distribución*, *Idioma del usuario*, entre otras, para complementar el ejercicio, se eliminaron las columnas de las preguntas abiertas, lo que permitió tener a la altura de esta fase del estudio un dataset con dimensiones de (665, 75)

El siguiente paso en el procesamiento consistió en la transformación de los datos en escala Likert, las variables categóricas ordinales y las categóricas nominales. En variables representadas por números, que son con los que puede trabajar la máquina a la hora de uso de los modelos.

Después se hizo el cambio de nombres de las columnas, debido a que la denominación de las variables coincidía con el texto original de la encuesta, por esto, muchos nombres de variables constaban con más de cien caracteres, se les colocó nombres codificados, con menos caracteres pero con información suficiente para conocer de qué trata cada una de ellas, en el caso de las columnas asociadas a constructos se les cambió de una pregunta extensa, tal y como aparecía en la encuesta original a un formato, XXXXX\_# en donde las equis hacen referencia al nombre del constructo y el numeral a la posición en que iba apareciendo en el dataset.

Una vez ejecutada esta transformación se procede al cálculo del alfa de Cronbach para validar de manera cuantitativa cuáles de los conjuntos de datos asociados a los constructos sociocognitivos se podían seguir utilizando en el estudio, el siguiente paso fue aplicar análisis de componentes principales a cada una de las “*sub-matrices*” correspondientes a cada uno de los constructos estudiados para generar un vector que represente a todo el constructo que resuma la percepción a cada uno de ellos, es pertinente resaltar que el alfa de Cronbach y la reducción de dimensiones con PCA sólo se aplicó a los constructos con más de una pregunta. Luego de aplicar este proceso las dimensiones del dataset quedaron en (665, 33).

#### D. Modelado

Una vez finalizada la preparación de los datos, se procedió al modelado mediante técnicas de agrupamiento no supervisado, seleccionando K-means, DBSCAN como algoritmos principales. La decisión de aplicar técnicas de clustering se sustentó en la necesidad de explorar posibles patrones ocultos y estructuras latentes entre los estudiantes encuestados, sin imponer supuestos previos sobre la distribución o categorías de los datos.

Antes del entrenamiento de los modelos, se realizó un escalamiento a los datos para que sus valores originales no alteraran los resultados, dada la sensibilidad de K-means a datos

con diferentes escalas, para ello se transformaron a valor estándar Z y con ello poder etiquetar de una manera más eficaz a que clúster asignó cada uno de los datos, para su reducción de dimensionalidad utilizando Análisis de Componentes Principales (PCA), con el fin de simplificar el espacio de características y mejorar la interpretabilidad al permitir graficar en dos dimensiones conocer la estructura visual de los datos.

#### E. Evaluación del Modelo

Para la evaluación de los resultados de los modelos de clustering, se emplearon índices de validación interna, en particular el índice de silueta, que permite cuantificar cuán bien están definidos los clústeres generados. Los valores obtenidos en todos los enfoques fueron bajos (menores a 0.5), lo cual indicó una estructura de agrupamiento débil o difusa. En la visualización mediante técnicas como PCA, los datos mostraron una distribución sin evidencias claras de subgrupos internamente diferenciados, lo cual sugiere una alta similitud entre las respuestas de los estudiantes.

Tanto en los modelos con constructos resumidos como en la matriz completa, se observó una fuerte tendencia a la concentración de los datos en un único grupo, con la aparición de valores periféricos que fueron identificados como posibles outliers. Este comportamiento podría explicarse por una correlación significativa entre los constructos indagados, o por la naturaleza homogénea de las percepciones entre los estudiantes de programas STEM en la institución, pero en esta investigación no es el alcance deseado.

#### F. Despliegue del Modelo

En coherencia con el enfoque de la metodología CRISP-DM, para este estudio la etapa de despliegue no contempla la integración de los modelos en un sistema automatizado, sino la generación de un informe detallado que sintetice los hallazgos obtenidos. Este informe constituye el principal producto del estudio y está orientado a facilitar la extracción de insights relevantes que apoyen la toma de decisiones.

El informe mencionado consiste en el presente artículo en que se reúnen visualizaciones gráficas, análisis descriptivos y resultados del modelado, con énfasis en la interpretación de patrones generales y diferencias significativas por género. La información contenida en el reporte tiene el potencial para alimentar futuras investigaciones, incluyendo la formulación de estrategias pedagógicas e institucionales para el cierre de brechas de género en educación superior STEM.

### IV. RESULTADOS DE DATOS CUANTITATIVOS

La presente sección expone los hallazgos derivados del análisis estadístico y computacional aplicado a las encuestas

realizadas a estudiantes de programas STEM de la Universidad EAFIT de la ciudad de Medellín Primero, se presenta el conjunto de datos resultante de los procesos ejecutados en las fases de entendimiento y procesamiento de los datos, a su vez, la caracterización de los estudiantes encuestados. Posteriormente, se presentan las diferencias de género en los 14 constructos sociocognitivos. Además de los valores promedio en cada constructo para cada género, se presentan las distribuciones de cada género.

Una vez expuesto lo anterior se hará visible los resultados de la aplicación de técnicas de machine learning no supervisadas de agrupamiento para explorar la constitución subyacente de las personas encuestadas en el estudio y saber de la existencia de subconjuntos presentes tácitamente entre ellos.

#### A. Caracterización de los estudiantes encuestados.

En cuanto a las características sociodemográficas de la muestra que se seleccionó para el estudio ( $n=665$ ), se observa una predominancia de estudiantes que se identifican como hombres (65.6%), seguidos de mujeres (33.7%) y un pequeño grupo que prefirió no declarar su sexo biológico (0.8%). Al analizar la variable edad, los hombres presentaron una media de 20.4 años, mientras que las mujeres reportaron una media ligeramente menor de 19.8 años, con una menor dispersión. En términos de acceso a becas, el 55.4% de las mujeres manifestó contar con algún tipo de apoyo económico, frente al 43.3% de los hombres, lo cual podría reflejar políticas institucionales de equidad o factores contextuales asociados. Respecto al capital educativo familiar, los padres y madres de hombres tienden a tener una mayor representación en niveles profesionales, mientras que en el caso de las mujeres se observa una mayor frecuencia en niveles técnico-tecnológicos. Estos datos permiten trazar un perfil preliminar de los sujetos encuestados, donde se vislumbran posibles diferencias estructurales por género que podrían estar asociadas con su permanencia o percepción del entorno STEM.

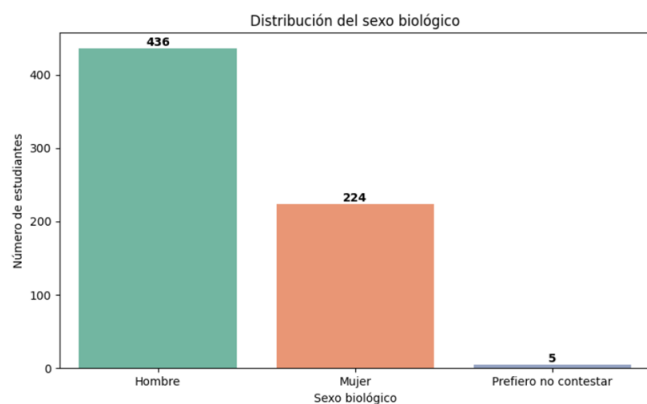


Fig. 5: Cantidad de mujeres, hombres y personas que decidieron no contestar la pregunta asociada al sexo biológico.

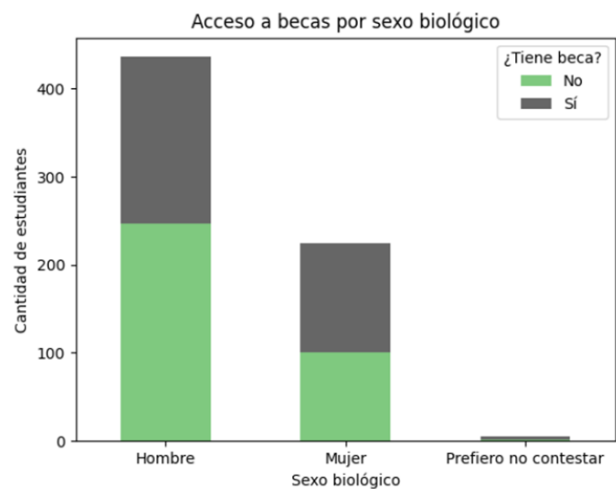


Fig. 6: Acceso a algún tipo de beca en encuestados estudiantes STEM en Universidad EAFIT sede Medellín

A su vez, respecto a la distribución de estudiantes por programa académico relacionado al área STEM matriculados se encontró que la mayoría de las personas encuestadas pertenecen a Ingeniería de Sistemas con 266 personas.

#### B. Resultados de limpieza e ingeniería de características

Luego de aplicar los pasos señalados en la metodología el producto obtenido fue un conjunto de datos de dimensiones (665, 33) seiscientos sesenta y cinco muestras y treinta y tres variables, convertidas en valores numéricos, preparados para su inclusión dentro de los algoritmos de machine learning no supervisados de clusterización. En donde cada uno de los constructos socio cognitivos a estudiar aparece resumido en una sola columna, mediante el uso del análisis de componentes principales.

Las variables de categóricas ordinales se transformaron con etiquetado en código asignando un número a cada valor observado de la variable, esta técnica se usó para las variables del grado de escolaridad del padre o tutor y madre o tutora y la variable del semestre cursado del estudiante. Por su parte, la variable de programa académico al no tener de manera intrínseca un orden establecido se transformó en valores numéricos utilizando la técnica de one hot encoding.

Dentro de la limpieza y la transformación de valores de las variables objeto de estudio, se realizó a todos los subconjuntos de reactivos que evalúan a 11 de los 14 constructos, debido a que tres de ellos correspondientes a: “Compatibilidad género programa, Compatibilidad el propio estudiante con el área STEM y Compatibilidad entre el género y el área STEM, tienen la particularidad de que sólo un reactivo es el que mide la percepción por lo tanto no fueron resumidos. A los demás se les aplicó la prueba del alfa de Cronbach, entregando resultados superiores a 0.7, evidencia que permitió hacer uso del resumen de los conjuntos de preguntas que midieron los constructos socio cognitivos.

Alfa de Cronbach para AUTOEFICACIA: 0.9286  
 Alfa de Cronbach para AUTORREGULACION: 0.7694  
 Alfa de Cronbach para MOTIVACION INTRINSECA: 0.7143  
 Alfa de Cronbach para MOTIVACION EXTRINSECA: 0.7938  
 Alfa de Cronbach para SENTIDO DE PERTENENCIA AL PROGRAMA: 0.8919  
 Alfa de Cronbach para SENTIDO DE PERTENENCIA A STEM: 0.9139  
 Alfa de Cronbach para COMPROMISO ACADÉMICO: 0.7872  
 Alfa de Cronbach para EXPECTATIVAS DE EMPLEO: 0.9412  
 Alfa de Cronbach para EXPECTATIVAS DE VALOR TENER ÉXITO EN STEM: 0.9428  
 Alfa de Cronbach para EXPECTATIVAS DEL VALOR PERCIBIDO EN STEM: 0.9347  
 Alfa de Cronbach para INTENCIÓN DE ABANDONAR: 0.9258

Fig. 7: Evidencia de valores obtenidos índice alfa de Cronbach para afirmar consistencia entre los reactivos que miden cada constructo.

### C. Resultados de la aplicación de estadística inferencial

En el orden de obtener una perspectiva amplia del caso de STEM enfocado en género, además de los modelos de aprendizaje automático no supervisados, también se hizo uso de técnicas de estadística inferencial para evaluar si las personas encuestadas de acuerdo con la información captada al diligenciar el instrumento, presentan con un nivel suficiente de confianza la evidencia estadística suficiente para afirmar que existen diferencias en las percepciones de acuerdo con el sexo biológico.

Durante el tratamiento de los datos para responder la pregunta de investigación mediante este enfoque se pudo conocer otro conjunto de características sobre la muestra encuestada. Por ejemplo, se encontró que las 665 muestras la mayoría pertenecen al programa de ingeniería de sistemas con 266 estudiantes, le sigue ingeniería civil con 146 estudiantes. Sólo estos dos programas aportaron cerca del 50% de los sujetos encuestados.

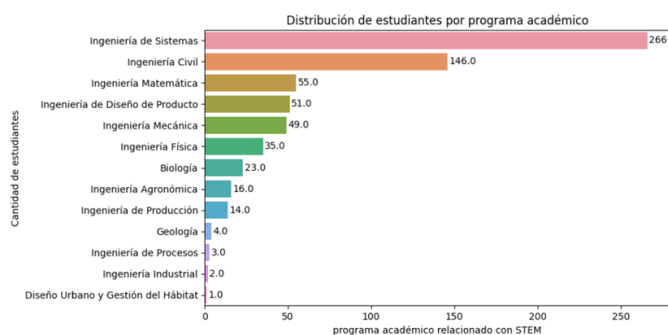


Fig. 8: Distribución de estudiantes por programa académico

A continuación, se muestra de manera gráfica la participación tanto de mujeres como hombres en los dos programas que más aportaron muestras para este estudio. Ingeniería de Sistemas con (H=72,6%, M=22,4%) e Ingeniería Civil con (H=68,5%, M=30,1%, PN=1,4%)

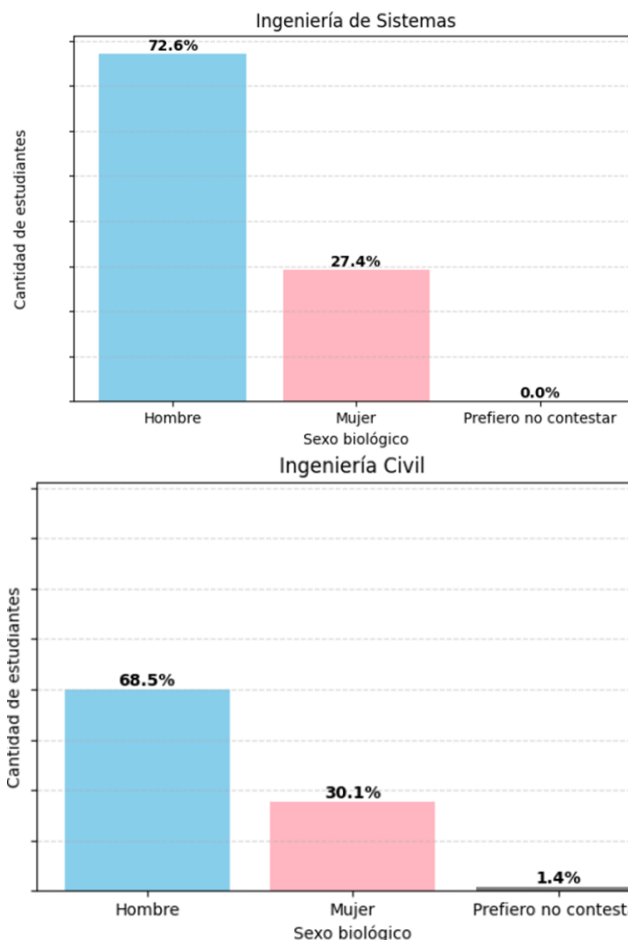


Fig. 9: Distribución de mujeres y hombres en los encuestados que pertenecen a Ingeniería de Sistemas e Ingeniería Civil

Luego del mencionado tratamiento de datos se hizo el procedimiento de cálculo de los parámetros necesarios para hacer la prueba de diferencia de medias entre las mujeres y hombres dentro del estudio, con ayuda del *p-value* cuyo valor fue menor o igual a 0.05 se pudo identificar cuáles de los constructos sociocognitivos tienen la evidencia estadística suficiente para afirmar que hay diferencias significativas entre hombres y mujeres al responder el instrumento de evaluación asociadas a ciertos constructos sociocognitivos indagados en esta encuesta. Para exponer el hallazgo bajo el enfoque de la estadística inferencial se presenta la siguiente tabla.

**TABLA II**  
**ANÁLISIS DIFERENCIA DE MEDIAS MUJERES-HOMBRES POR CONSTRUCTO**

Constructos	P-value	Promedio			Desviación estándar		
		Total	H	M	Total	H	M
Edad (años)	0,06	20,24	20,43	19,84	2,68	3	1,85
Sexo biológico	0	-	-	-	-	-	-
Semestre	0,41	5	5,05	4,91	2,43	2,43	2,44
Beca	0	0,48	0,43	0,55	0,5	0,5	0,5
Autoeficacia	0	4,21	4,26	4,09	0,69	0,67	0,7
Autorregulación	0	3,83	3,78	3,92	0,68	0,69	0,65
Motivación intrínseca	0,03	3,99	4,03	3,93	0,67	0,68	0,64
Motivación extrínseca	0,03	3,78	3,73	3,87	0,89	0,87	0,92
Sentido de pertenencia al programa	0	4,19	4,23	4,12	0,68	0,7	0,63
Sentido de pertenencia a STEM	0,81	3,84	3,84	3,84	0,87	0,89	0,84
Compromiso Académico	0,1	4,55	4,52	4,62	0,6	0,64	0,5
Expectativa de empleo STEM	0,91	4,22	4,23	4,2	0,84	0,82	0,87
Expectativa de éxito en STEM	0,05	4,12	4,15	4,04	0,76	0,76	0,75
Valor Percibido	0,68	4,12	4,11	4,13	0,76	0,75	0,77
Intención Abandonar	0,43	1,94	1,99	1,83	1,07	1,13	0,92
Genero/Programa	0	4,95	5,44	3,98	2,02	1,97	1,77
STEM y el estudiante	0	5,15	5,41	4,66	1,63	1,57	1,64
Genero y STEM	0	5	5,43	4,16	1,89	1,87	1,66

Se evidenció que en 9 de los 14 constructos socio cognitivos estudiados hay evidencia estadística suficiente para afirmar que hay diferencias en sus medias por su pertenencia a cada género biológico tal y como se aprecia en la tabla los constructos que tienen esta condición son: Autoeficacia, Autorregulación, Motivación intrínseca, Motivación Extrínseca, Sentido de pertenencia al programa, Expectativa de éxito en STEM, Compatibilidad entre el género y el programa que está cursando el estudiante, compatibilidad entre el estudiante y STEM, compatibilidad entre el género y STEM. A continuación, se expone de manera gráfica el comportamiento de las distintas variables asociadas a los constructos por género de las 9 variables mencionadas.

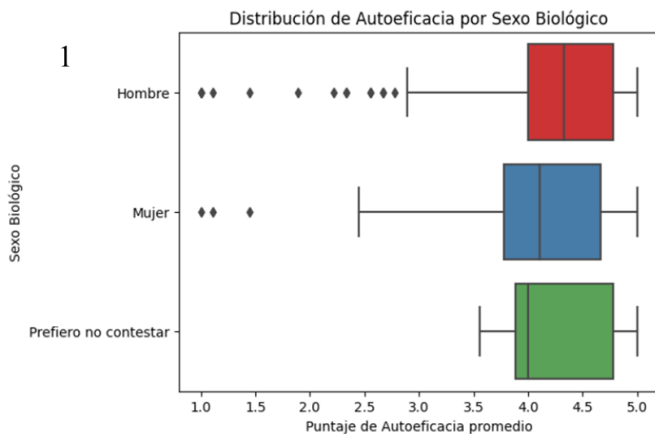


Fig. 10: Distribución de constructo autoeficacia por género

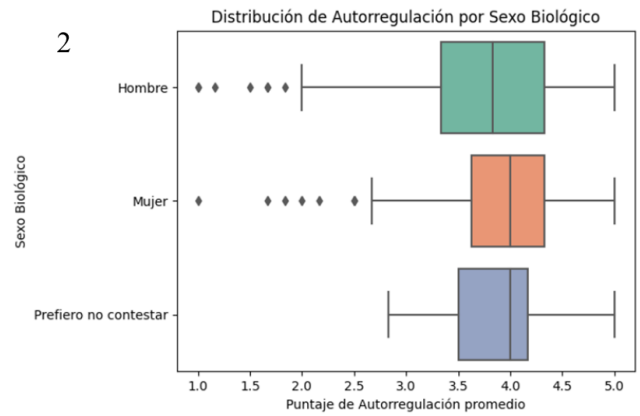


Fig. 11: Distribución de constructo autorregulación por género

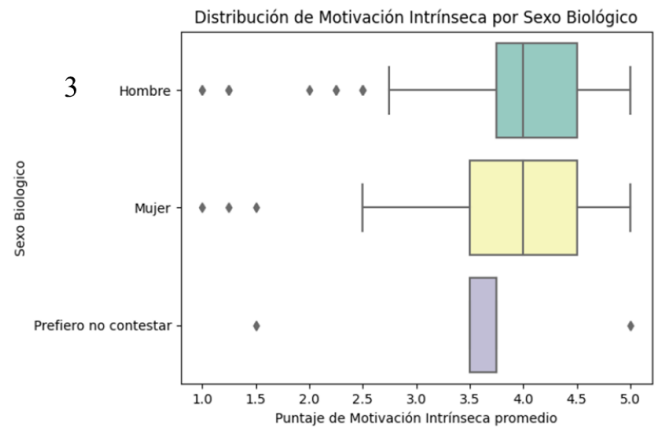


Fig. 12: Distribución de constructo motivación intrínseca por género

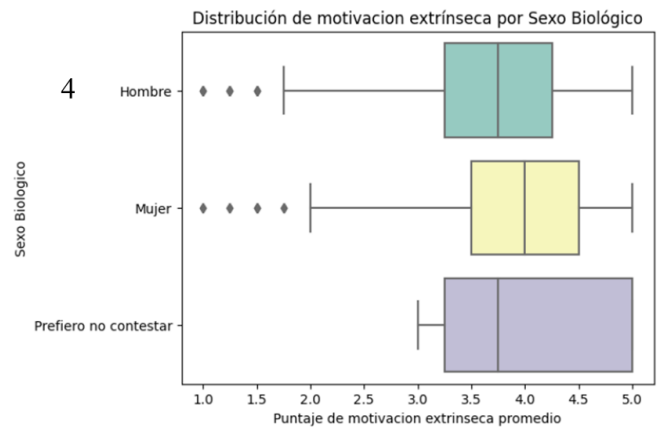


Fig. 13: Distribución de constructo motivación extrínseca por género



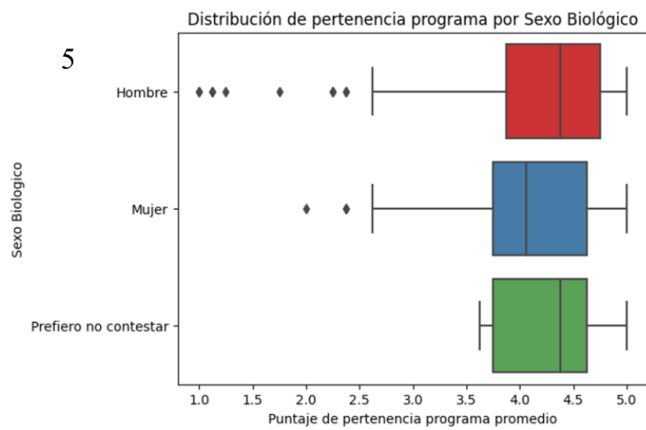


Fig. 14: Distribución de constructo pertenencia al programa por género

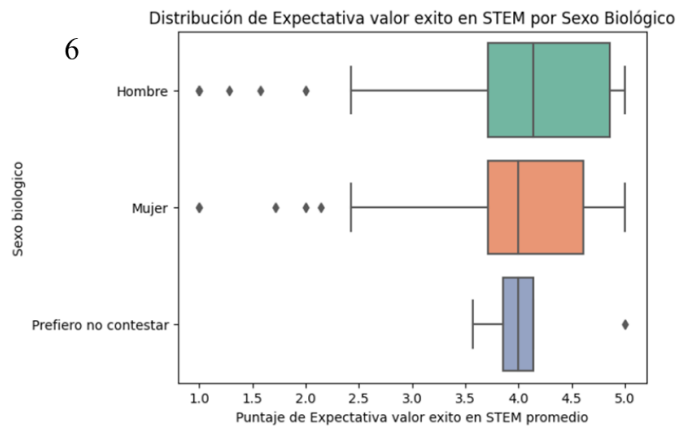


Fig. 15: Distribución de constructo expectativa de valor de éxito en el área STEM por género

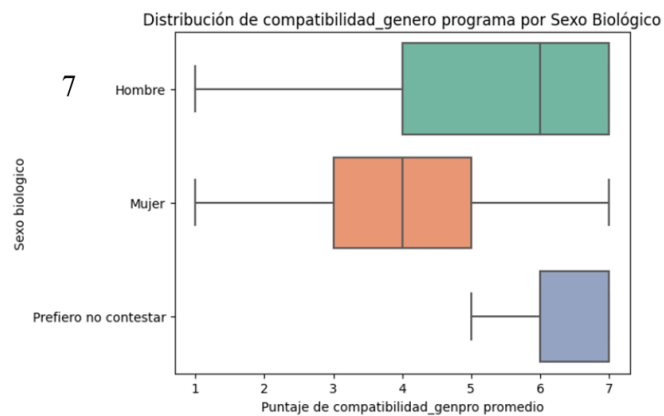


Fig. 16: Distribución de constructo compatibilidad entre el género y programa estudiado por género

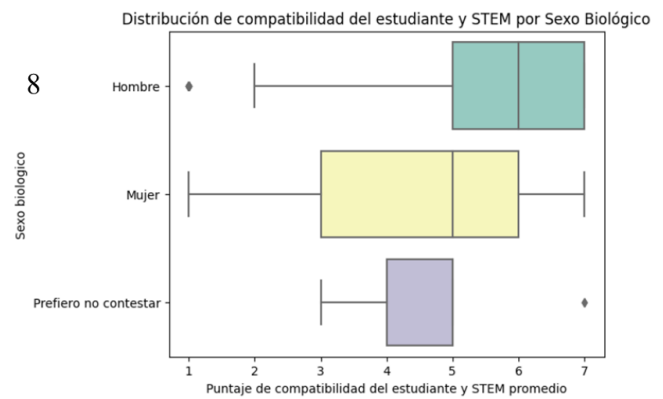


Fig. 17: Distribución de constructo compatibilidad entre el estudiante y el área STEM por género

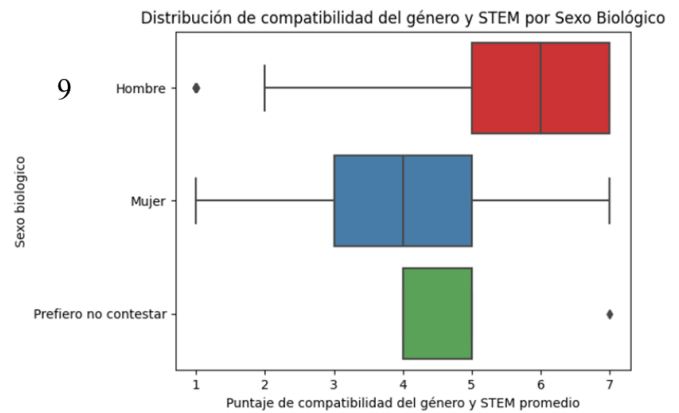


Fig. 18: Distribución de constructo compatibilidad entre el género y el área STEM por género

#### D. Resultados de la aplicación de algoritmos de clustering

Durante el presente estudio se utilizó un enfoque basado en aprendizaje automático no supervisado con los algoritmos de clustering: K-means y DBSCAN, buscando descubrir si existen grupos inherentes a los sujetos encuestados, seguidamente, se presentan los hallazgos para cada una de las técnicas utilizadas.

##### D.1. Hallazgos en el uso de técnica K-means

Se colocó la tabla de datos procesada con todas las variables representadas en un formato numérico, y produjo como resultado el conjunto de datos con una columna adicional en donde le asignó a cada muestra a qué clúster le correspondió.

Para hacer mejor interpretable el hallazgo se muestra la gráfica obtenida para los componentes principales a los que se redujo el dataset procesado por K-means y que muestra la ubicación de cada una de las muestras respecto a esos ejes.

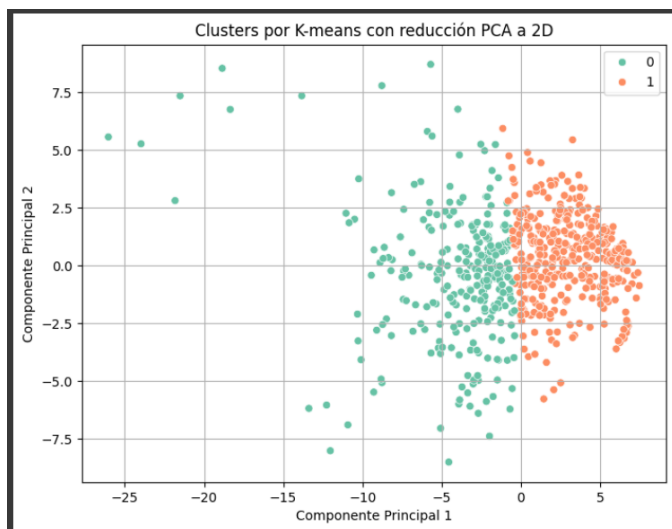


Fig. 19: Distribución de muestras en K-means para un valor de  $K = 2$

Los valores de las métricas para este resultado fueron los siguientes: Silhouette Score: 0.139 (ideal  $\geq 0.5$ ).

#### D.2. Hallazgos en el uso de técnica DBSCAN

En el momento en que se realizó el DBSCAN se hicieron iteraciones para en los parámetros buscando la posibilidad de obtener dos agrupamientos para poder observar si había grupos inmersos dentro de la masa de datos de los sujetos encuestados pese a que inicialmente DBSCAN devolvía sólo un grupo, hasta que el valor de  $\epsilon = 14$  arrojó dos grupos, y se presentó el siguiente comportamiento de manera gráfica.

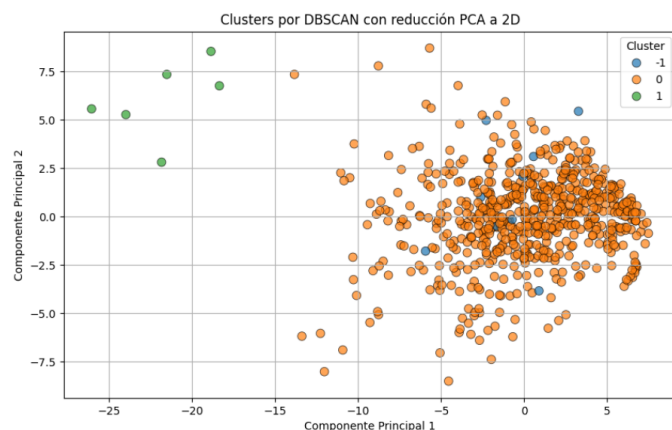


Fig. 20: Distribución de muestras en DBSCAN para un valor de  $\epsilon = 14$

El valor de la métrica del índice de silueta para ver la separación de los grupos no fue satisfactorio para afirmar que hay más de un grupo consistente inherente en los datos con número de clústeres (NO ruido): 2, índice del silhouette: 0.3844 (ideal  $\geq 0.5$ )

## V. CONCLUSIONES Y PERSPECTIVAS

Una conclusión que necesita ser enunciada en primer lugar radica en el hecho de que las afirmaciones realizadas en la introducción y en el entendimiento del contexto respecto a una participación femenina inferior a la cantidad de hombres en programas relacionados con STEM, es consistente con lo que se encontró en la muestra de este estudio con hombres (65.6%), seguidos de mujeres (33.7%). Se refleja una vez más la necesidad de continuar con el proceso de proponer y luego aplicar recomendaciones, estrategias que cierren la brecha entre participación femenina y masculina en las Instituciones de educación superior.

Respecto al enfoque de análisis de los datos en términos de clustering e evidencia una posible ausencia de estructura de clúster clara, pues para las dos técnicas de agrupamiento utilizadas en este artículo agruparon la mayoría de los datos en un sólo grupo de forma semi-elíptica, y con índices de silueta menores que 0.5. Puede ocurrir que no haya un sub agrupamiento interno dentro de las muestras o que las variables indagadas por los reactivos del instrumento presentaron un comportamiento correlacionado entre sí. Y el comportamiento elíptico de las proyecciones en 2D de las muestras indican que la variabilidad de los datos está explicada de mayor manera en una sola dirección.

Este enfoque sugiere que los estudiantes tienen perfiles similares en sus respuestas, y no hay subgrupos naturales diferenciables en las dimensiones abordadas en el estudio, utilizando otras herramientas de análisis no se debe esperar que haya radicales diferencias a nivel general entre los sujetos encuestados.

Efectivamente el análisis ejecutado desde la óptica de evidenciar estadísticamente diferencias de medias a nivel particular estudiando cada uno de los constructos por separado comparando el género biológico (femenino vs masculino) logró descubrir diferencias en 9 de los 14 constructos. Un aspecto a destacar es que no todas las diferencias implicaron una percepción más baja en los 9 constructos, pues de acuerdo con las representaciones gráficas de los resultados en los boxplot (2 y 4) correspondientes a Autorregulación y Motivación extrínseca, el rango Inter cuartil correspondiente a las respuestas de las mujeres muestran concentración de valores más altos de percepción a favor que en las respuestas de los hombres.

Las representaciones gráficas que aparecen en resultados (1, 3, 5 y 6) correspondientes a: Autoeficacia, Motivación intrínseca, Sentido de pertenencia con el programa y Expectativas de valor de éxito en la carrera del estudiante en STEM muestran que en las mujeres tienden a sentir una percepción más baja que los hombres con una diferencia visual sucinta. Sin embargo, en las representaciones gráficas de los constructos (7, 8 y 9) correspondientes a la: compatibilidad entre género y STEM, compatibilidad del estudiante con el área STEM y compatibilidad entre el género y STEM, se revelan

percepciones más marcadas hacia la percepción de que hay una baja *compatibilidad de ellas mismas como personas, que el género biológico al que pertenecen* respecto a los programas que estudian y al área de conocimiento STEM.

El comportamiento presentado por la aplicación de aprendizaje automático sugiere que la población estudiantil presenta un perfil perceptivo homogéneo respecto a los constructos estudiados. El uso de técnicas no supervisadas en este contexto puede no ser suficiente para identificar diferencias sustanciales sin enriquecer previamente los datos, por ejemplo, mediante datos que puedan dar información de los estudiantes respecto a segmentación por contexto, carrera o trayectorias académicas.

A pesar de la homogeneidad general, se identificaron diferencias estadísticamente significativas por género en 9 de los 14 constructos sociocognitivos, de esta manera, los resultados revelan que no existe una desventaja uniforme femenina, sino una percepción diferenciada por tipo de constructo. Esto abre la puerta a un enfoque de intervención más preciso y focalizado, que no se base en estereotipos, sino en evidencia segmentada.

Producto de este estudio se pueden dar recomendaciones a las distintas entidades gubernamentales, no gubernamentales, instituciones de educación superior e instituciones con capacidad de intervención en cierre de brechas de género de orientar el diseño de programas de acompañamiento diferenciados por género, no solo para motivar el ingreso de mujeres a programas académicos relacionados con STEM, sino también para apoyar su experiencia académica, especialmente en áreas donde ellas mismas reportan una menor compatibilidad o sentido de pertenencia.

Para futuras investigaciones se propone, revisar las muestras que aparecen como outliers para identificar las condiciones especiales que les clasifican por fuera de la masa de datos, a nivel más general, se podría considerar el uso de otras técnicas de agrupamiento, con un enfoque probabilístico o con cálculos de distancia no lineales proyectando desde espacios vectoriales diferentes.

#### AGRADECIMIENTO/RECONOCIMIENTO

Para la construcción de este artículo hago una mención a la Universidad EAFIT por realizar todas las gestiones administrativas para abrir el espacio para este estudio y por abrir el espacio para recabar la información de los estudiantes matriculados en programas afines a STEM. También un agradecimiento especial las ingenieras Liliana González Palacio y Silvana Montoya Noguera, por sus orientaciones, por la confianza depositada en el autor y por la disposición y dedicación de tiempo para apoyarle en múltiples obstáculos que se presentaron durante la realización del estudio.

#### REFERENCIAS

- [1] Sistema Nacional de la Información de la Educación Superior -SNIES (2023). Estudiantes Matriculados en primer curso 2023 [En línea]. Disponible en: <https://snies.mineducacion.gov.co/portal/ESTADISTICAS/Bases-consolidadas/>
- [2] A. Alam, «Psychological, Sociocultural, and Biological Elucidations for Gender Gap in STEM Education: A Call for Translation of Research into Evidence-Based Interventions», presentado en 2nd International Conference on Sustainability and Equity (ICSE-2021), Bhubaneswar, India, 2022. doi: 10.2991/ahsseh.k.220105.012.
- [3] G. Costa-Lizama, L. San Martín, O. Pinto, y G. Gatica, «Hack4women: un paso hacia la equidad de género», Texto Livre, vol. 15, p. e39348, oct. 2022, doi: 10.35699/1983-3652.2022.39348.
- [4] D. Donahoe, «“The Definition of STEM?”», Today’s Engineer, IEEE-USA, dic. 2013.
- [5] C. J. Heffernan, «Social foundations of thought and action: A social cognitive theory, Albert Bandura Englewood Cliffs, New Jersey: Prentice Hall, 1986, xiii + 617 pp. Hardback. US\$39.50.», Behav. change, vol. 5, n.º 1, pp. 37-38, mar. 1988, doi: 10.1017/S0813483900008238.
- [6] A. Bandura, Self-efficacy: The Exercise of Control. New York, NY, USA: W. H. Freeman/Times Books/Henry Holt & Co., 1997.
- [7] J. E. Ormrod, Human learning, Eighth edition. Hoboken, NJ: Pearson, 2020.
- [8] A. L. Zeldin y F. Pajares, «Against the Odds: Self-Efficacy Beliefs of Women in Mathematical, Scientific, and Technological Careers», American Educational Research Journal, vol. 37, n.º 1, pp. 215-246, mar. 2000, doi: 10.3102/00028312037001215.
- [9] B. J. Zimmerman, «Attaining Self-Regulation», en Handbook of Self-Regulation, Elsevier, 2000, pp. 13-39. doi: 10.1016/B978-012109890-2/50031-7.
- [10] E. L. Deci y R. M. Ryan, Intrinsic Motivation and Self-Determination in Human Behavior. Boston, MA: Springer US, 1985. doi: 10.1007/978-1-4899-2271-7.
- [11] C. Corson y M. G. González-Morales, «Exploring women’s and men’s belonging in STEM», EDI, dic. 2024, doi: 10.1108/EDI-02-2024-0060.
- [12] M. H. Yong, G. Chikwa, y J. Rehman, «Factors affecting new students’ sense of belonging and wellbeing at university», Innovations in Education and Teaching International, pp. 1-14, ene. 2025, doi: 10.1080/14703297.2025.2453104.
- [13] D. R. Johnson et al., «Examining Sense of Belonging Among First-Year Undergraduates From Different Racial/Ethnic Groups», csd, vol. 48, n.º 5, pp. 525-542, sep. 2007, doi: 10.1353/csd.2007.0054.
- [14] T. M. Mitchell, Machine Learning. en McGraw-Hill series in computer science. New York: McGraw-Hill, 1997.
- [15] T. D. Buskirk, A. Kirchner, A. Eck, y C. S. Signorino, «An Introduction to Machine Learning Methods for Survey Researchers», Surv Pract, vol. 11, n.º 1, pp. 1-10, ene. 2018, doi: 10.29115/SP-2018-0004.
- [16] S. Jaggia, A. Kelly, K. Lertwachara, y L. Chen, «Minería de Datos no Supervisada», en Analítica de Negocios Comunicación con Números, 2.ª ed., vol. 1, 1 vols., McGraw Hill, 2023, p. 568. [En línea]. Disponible en: <https://www-ebooks7-24-com.ezproxy.eafit.edu.co/?il=26249&pg=1>
- [17] R. Wirt y J. Hipp, «CRISP-DM: Towards a Standard Process Model for Data Mining», presentado en Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000, pp. 29-39. [En línea]. Disponible en: [https://www.researchgate.net/publication/239585378\\_CRISP-DM\\_Towards\\_a\\_standard\\_process\\_model\\_for\\_data\\_mining](https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining)
- [18] S. Jaggia, A. Kelly, K. Lertwachara, y L. Chen, «Análisis de conglomerados jerárquicos», en Analítica de Negocios Comunicación con Números, 2.ª ed., McGraw Hill, 2023, pp. 538-545. Accedido: 28 de junio de 2025. [En línea]. Disponible en: <https://www-ebooks7-24-com.ezproxy.eafit.edu.co/?il=26249&pg=1>

- [19] C. M. Bishop, Pattern recognition and machine learning. en Information science and statistics. New York: Springer, 2006. p. 424.
- [20] A. Ram, S. Jalal, A. S. Jalal, y M. Kumar, «A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases», IJCA, vol. 3, n.º 6, pp. 1-4, jun. 2010, doi: 10.5120/739-1038.
- [21] C. M. Bishop, Pattern recognition and machine learning. en Information science and statistics. New York: Springer, 2006.
- [22] L. J. Cronbach, «Coefficient Alpha and the Internal Structure of Tests», Psychometrika, vol. 16, n.º 3, pp. 297-334, sep. 1951, doi: 10.1007/BF02310555.
- [23] J. González Alonso y M. Pazmiño Santacruz, «Cálculo e interpretación del Alfa de Cronbach para el caso de validación de la consistencia interna de un cuestionario, con dos posibles escalas tipo Likert», Revista Publicando, vol. 2, n.º 1, pp. 62-67, 2015.
- [24] S. Jaggia, A. Kelly, K. Lertwachara, y L. Chen, «Introducción a la minería de datos», en Analítica de Negocios Comunicación con Números, 2.ª ed., vol. 1, 1 vols., McGraw Hill, 2023, p. 418-425. Accedido: 17 de junio de 2025. [En línea]. Disponible en: <https://www-ebooks7-24-com.ezproxy.eafit.edu.co/?il=26249&pg=1>
- [25] H. Jodlbauer y R. Riedl, «Maintaining Accuracy While Reducing Effort in Online Decision Making: A New Quantitative Approach for Multi-Attribute Decision Problems Based on Principal Component Analysis», JTAER, vol. 19, n.º 4, pp. 2896-2918, oct. 2024, doi: 10.3390/jtaer19040140.
- [26] I. T. Jolliffe y J. Cadima, «Principal component analysis: a review and recent developments», Phil. Trans. R. Soc. A., vol. 374, n.º 2065, p. 20150202, abr. 2016, doi: 10.1098/rsta.2015.0202
- [27] S. Hermans, M. Gijzen, T. Mombaers, y P. Van Petegem, «Gendered patterns in students' motivation profiles regarding iSTEM and STEM test scores: a cluster analysis», IJ STEM Ed, vol. 9, n.º 1, p. 67, oct. 2022, doi: 10.1186/s40594-022-00379-3.
- [28] K. M. Whitcomb, Z. Y. Kalender, T. J. Nokes-Malach, C. D. Schunn, y C. Singh, «A mismatch between self-efficacy and performance: Undergraduate women in engineering tend to have lower self-efficacy despite earning higher grades than men», 12 de marzo de 2020, arXiv: arXiv:2003.06006. doi: 10.48550/arXiv.2003.06006.
- [29] Y. Li, K. M. Whitcomb, y C. Singh, «How learning environment predicts male and female students' physics motivational beliefs in introductory physics courses», en 2020 Physics Education Research Conference Proceedings, Virtual Conference: American Association of Physics Teachers, sep. 2020, pp. 284-290. doi: 10.1119/perc.2020.pr.Li.
- [30] N. Yamani y H. Almazroa, «Exploring career interest and STEM self-efficacy: implications for promoting gender equity», Front. Psychol., vol. 15, p. 1402933, oct. 2024, doi: 10.3389/fpsyg.2024.1402933.
- [31] Y. Li y C. Singh, «The impact of perceived recognition by physics instructors on women's self-efficacy and interest», 2023, doi: 10.48550/ARXIV.2303.07239.
- [32] M. de E. N. MEN, «+ género Una propuesta para fortalecer la educación inicial con equidad». 2021. [En línea]. Disponible en: <https://www.colombiaaprende.edu.co/contenidos/coleccion/guia-steam-genero>
- [33] P. Guevara-Ramírez et al., «Ten simple rules for empowering women in STEM», PLoS Comput Biol, vol. 18, n.º 12, p. e1010731, dic. 2022, doi: 10.1371/journal.pcbi.1010731.
- [34] P. Chapman, «CRISP-DM 1.0: Step-by-step data mining guide», Computer Science, 2000. [En línea]. Disponible en: <https://api.semanticscholar.org/CorpusID:59777418>
- [35] I. D. Cherney, «The STEM paradox: Factors affecting diversity in STEM fields», J. Phys.: Conf. Ser., vol. 2438, n.º 1, p. 012005, feb. 2023, doi: 10.1088/1742-6596/2438/1/012005.
- [36] Sistema Nacional de la Información de la Educación Superior - SNIES (2023) Tableros de Control - Graduados en CINE Campo Amplio 1. Ciencias naturales, matemáticas y estadística, 2. Ingeniería, Industria y construcción. y 3. Tecnología de la información y comunicación (TIC) 2023 [En línea]. Disponible en: <https://hecaa.mineducacion.gov.co/consultaspublicas/tableros/graduados>
- [37] F. Lauermann, Y.-M. Tsai, y J. S. Eccles, «Math-related career aspirations and choices within Eccles et al.'s expectancy-value theory of achievement-related behaviors.», Developmental Psychology, vol. 53, n.º 8, pp. 1540-1559, ago. 2017, doi: 10.1037/dev0000367.

**ANEXO: REPOSITORIO EN GITHUB CON EL REGISTRO DE CODIGO Y DATASET UTILIZADO EN EL ESTUDIO.**

[https://github.com/JGutierrez90/Tesis\\_Juan\\_EAFIT\\_2025/tree/main](https://github.com/JGutierrez90/Tesis_Juan_EAFIT_2025/tree/main)



**Tesis\_Juan\_EAFIT\_2025** Public



JGutierrez90 / Tesis\_Juan\_EAFIT\_2025

[https://github.com/JGutierrez90/Tesis\\_Juan\\_EAFIT\\_2025](https://github.com/JGutierrez90/Tesis_Juan_EAFIT_2025)