APLICACIÓN DE PROCESAMIENTO DE LENGUAJE NATURAL EN ENSAYOS EN PRIMERA PERSONA: EXPLORACIÓN DE LIBROS “MATILDA Y LAS MUJERES EN INGENIERÍA EN AMÉRICA LATINA”

Natural Language Processing applied in essays in the first person: Exploration of Books "Matilda and Women in Engineering in Latin America."

Guadalupe Pascal [1], Soledad Bernachea [1], Lucía Palavecino [1], Milagros Tevez Sauco [1]

gpascal@ingenieria.unlz.edu.ar, sbernachea@ingenieria.unlz.edu.ar, lpalavecino@ingenieria.unlz.edu.ar, mtevezsauco@ingenieria.unlz.edu.ar

[1] Facultad de Ingeniería, Universidad Nacional de Lomas de Zamora, 1832, Lomas de Zamora, Argentina.

Recibido 20/10/2021; Aceptado: 29/12/2021

Resumen: Los niveles de participación de las mujeres en las áreas de Ciencias, Tecnologías, Ingenierías y Matemáticas en la región de América Latina representan y responden a una problemática estructural. En particular, se entiende que los roles que ocuparon y ocupan las mujeres en el desarrollo de la historia de la ingeniería, se encuentran estrictamente relacionados con la estructura desigual de géneros en las sociedades pasadas y presentes, cuya evidencia y magnitud puede captarse a través de las brechas de género existentes. En este contexto, resulta imprescindible reconocer la responsabilidad que tienen las instituciones de educación superior de las carreras afines sobre la problematización de las brechas de género en los propios espacios de formación y desempeño profesional con el objetivo de transformar la situación actual. En la presente investigación se aplican técnicas de procesamiento de lenguaje natural y text mining sobre ensayos en primera persona de diversas profesionales mujeres del campo de las Ingenierías con el objetivo de conocer las principales representaciones generadas y publicados en los Libros “Matilda y las mujeres en Ingeniería en América Latina” durante el 2019 y el 2021.

Palabras-clave: Brecha; Género; Matilda; STEM; Ingeniería

Abstract: The levels of participation of women in the areas of Science, Technology, Engineering, and Mathematics in the Latin American region represent and respond to a structural problem. In particular, in the development of engineering history, the women's role is strictly related to the unequal gender structure in the past and present societies, whose evidence and magnitude can be explained through existing gender gaps. In this context, it is essential to admit the responsibility that higher education institutions of related careers have to problematize gender gaps in their training and

professional performance spaces to transform the current situation. In the present research, we applied natural language processing and text mining techniques to first-person essays of diverse professional women in the field of Engineering. The main aim is to know the foremost representations produced and published in the Books "Matilda and women in Engineering in Latin America" during 2019 and 2021.

***Keywords**:* Gap; Gender; Matilda; STEM; Engineering

1. Introducción

El problema de las brechas de género en las profesiones de las Ciencias, las Tecnologías, las Ingenierías, y las Matemáticas (en adelante “STEM” por sus siglas en inglés) sigue vigente en todo el mundo. Particularmente en América Latina, este problema es especialmente grave debido a su intrínseca relación con el desarrollo socio-económico de la región. (García-Holgado, Díaz & García Peñalvo, 2019)

Un abordaje desde las causas permite reconocer que pueden ser variadas las razones que conllevan a la problemática. Morales Inga S., y Morales Tristán (2020) en su artículo “¿Por qué hay pocas mujeres científicas? Una revisión de literatura sobre la brecha de género en carreras STEM” proponen una tipificación que organiza los estudios sobre la brecha de género en carreras STEM según los tres tipos de explicación que obtuvieron de su relevamiento bibliográfico. Por un lado, una explicación psicológica, que enfatiza en la instancia del individuo y explica la brecha en términos de autoconcepto, autoconfianza y autoeficacia, creencias y percepciones, y diferencias en intereses y preferencias. Por otro lado, una explicación sociocultural, que enfatiza la importancia de la cultura y explica la brecha por efecto de influencia parental y socialización, discriminación y sesgos, estereotipos y roles de género; y por último, una explicación biológica, que enfatiza en los rasgos del sexo anatómico y explica la brecha por su impacto en las diferencias de género.

Otro enfoque plausible consiste en observar las brechas de género en STEM a través de emergentes empíricos en los diferentes estadios en la formación y el desarrollo profesional de las personas: la educación inicial y la escolarización; la formación superior; y el desempeño profesional en el ámbito laboral. Estos emergentes responden a la división de los recursos y las oportunidades entre los géneros a partir de un sistema de símbolos y representaciones creados arbitrariamente por las sociedades; las cuales se aprenden desde los primeros años de vida direccionando las elecciones humanas.

Tanto para el enfoque causal como para el de los emergentes empíricos, resulta sustancial comprender la importancia que alcanzan las representaciones sociales que configuran la sociedad, entendiendo por representación social al producto de las prácticas resultantes de la comunicación social o de experiencias grupales, las cuales se producen en situaciones colectivas otorgándoles un sentido. (Castorina y Barreiro, 2012). Desde este marco, intentar responder la pregunta ¿Qué representaciones surgen de las profesionales mujeres de América Latina en materia de género y STEM, a partir de sus experiencias? resulta sumamente enriquecedor para abordar aproximaciones en torno a las brechas de género en dicho campo de aplicación.

El objetivo principal de la investigación es recopilar y analizar ensayos en primera persona a partir de los libros “Matilda y las mujeres en Ingeniería en América Latina” producidas por la Cátedra Abierta Latinoamericana Matilda y Las Mujeres en

Ingeniería (CAL-Matilda) en pos de explorar las representaciones sociales construidas y divulgadas sobre el rol de las mujeres en las Ingenierías en la región en el período 2019-2021.

2. Matilda y Las Mujeres en Ingeniería en América Latina

La CAL-Matilda, es una iniciativa interinstitucional de la Asociación Colombiana de Facultades de Ingeniería (ACOFI), el Consejo Federal de Decanos de Ingeniería de Argentina (CONFEDI) y la Latin American and Caribbean Consortium of Engineering Institutions (LACCEI). Su misión es “consolidarse como un espacio académico para el debate, la reflexión, la construcción colectiva de la docencia e investigación y la realización de actividades dinamizadoras de la igualdad de derechos, oportunidades y espacios para las mujeres en el ámbito académico y profesional y para el fomento de las vocaciones por la ingeniería en niñas y jóvenes en América Latina y el Caribe” (Páez Pino, 2020).

Su nombre Matilda refiere al fenómeno homónimo, “efecto matilda”, el cual documenta que existe un prejuicio en contra de reconocer los logros de las mujeres científicas, cuyo trabajo a menudo se atribuye a sus colegas de género masculino. Este fenómeno fue descrito por primera vez por Matilda Joslyn Gage (1883) en su ensayo «La mujer como inventora», sin embargo no fue hasta una década después cuando la historiadora de la ciencia Margaret W. Rossiter (1993) dió entidad al término e ilustró diversas mujeres que habían sufrido el efecto Matilda, entre las que se incluyen Marie Curie, Lise Meitner, Marietta Blau, Rosalind Franklin y Jocelyn Bell Burnell. (Rafael, 2019)

En este sentido, la colección de libros “Matilda y las mujeres en Ingeniería en América Latina” fue elaborada con la finalidad de visibilizar diversas mujeres profesionales a través de relatos en primera persona sobre su experiencia formativa, profesional y personal en el ámbito de las Ingenierías.

4. Materiales y Métodos

4.1. Corpus de trabajo

La colección de libros “Matilda y las mujeres en Ingeniería en América Latina” se organiza en tres tomos (Figura 1). Dichas publicaciones corresponden a ensayos en primera persona de más de 100 mujeres ingenieras y profesionales afines, publicados entre el 2019 y 2021, con el objetivo de visibilizar diferentes experiencias de profesionales en el campo de las Ingenierías. Entre las autoras, se encuentran mujeres que se desempeñan en los sectores productivos, académicas, investigadoras, estudiantes de grado y posgrado, mentoras y referentes. Uno de los impactos esperados de la colección de libros de la CAL-Matilda fue generar un conjunto de referentes que permita motivar y generar representaciones en niñas y jóvenes que deseen estudiar carreras STEM. Sin embargo, las más de 450 páginas resultan una oportunidad única para obtener información reveladora que sobrepasa sus motivaciones iniciales.

Figura 1: Portada Libros Matilda 1, 2 y 3, disponible en: https://catedramatilda.org

4.2. Aplicación de técnicas de la ciencia de datos en grandes volúmenes de texto.

Para el abordaje del análisis de datos se aplicarán técnicas de Data Mining, también conocido como Descubrimiento del Conocimiento en Bases de Datos (Knowledge Discovery in Databases KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Esta metodología es el resultado de la intersección entre diversas técnicas de estadística y algoritmos para el análisis de datos con herramientas informáticas y computacionales (Witten, Frank, & Hall, 2005).

La Figura 2 muestra la metodología KDD, la cual inicia con el relevamiento de los conocimientos previos y la identificación de los objetivos. A continuación, se realiza la recolección y estructuración de datos para luego realizar una limpieza y depuración de los mismos. Seguidamente, se realiza una transformación de los datos tanto en formato como en escala; para luego realizar el análisis propiamente dicho y generar nuevos conocimientos, finalidad principal de la minería de datos (Fayyad et al., 1996).

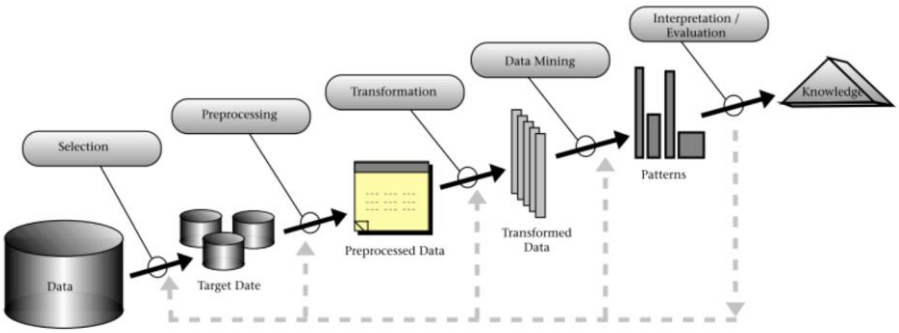


Figura 2: El proceso de generación de conocimiento en bases de datos. Fuente: Introduction to data mining. Fayyad, Usama, et. al. (1996).

En síntesis, las técnicas de data mining están destinadas a la recolección, depuración y análisis de datos con la finalidad de generar conocimiento. En el mismo sentido, el proceso de text mining es similar al de data mining, pero en este caso se obtiene información con alto valor agregado a partir de grandes cantidades de texto, con información no estructurada. (Tan, Steinbach & Kumar, 2006).

5. Análisis de resultados

5.1. Preprocesamiento y transformación

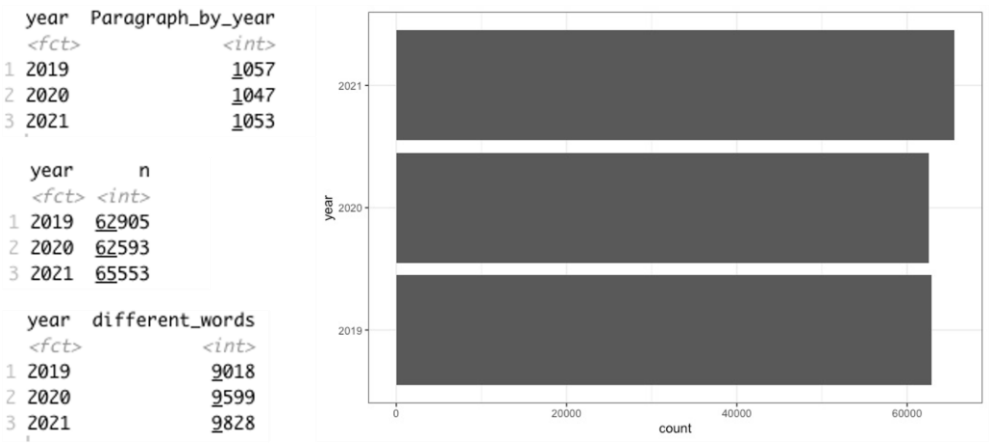
El primer paso consistió en la creación del corpus de trabajo tabulado a partir de los tres libros “Matilda y las mujeres en Ingeniería en América Latina” no estructurados disponibles en sus sitios web. Seguidamente se aplicó un proceso limpieza de texto, el cual consiste consiste en eliminar del texto todo aquello que no aporte información sobre su temática, estructura o contenido; en este caso se eliminaron patrones no informativos (como el uso de urls de páginas web), signos de puntuación, etiquetas HTML, caracteres sueltos y números. Finalmente, se procedió a tokenizar el texto, es decir, dividirlo en las unidades mínimas o elementos más sencillos que lo conforman, en este caso las palabras.

5.2. Exploración de voces

Una vez realizadas las instancias de preprocesamiento y transformación se realizaron diferentes rutinas de minería de textos. En primer lugar, dado que el corpus de trabajo está conformado por tres libros distintos (una publicación por año) resultó interesante explorar no solo la colección en su conjunto sino cada uno de ellos individualmente. Para eso, se caracterizó la escritura de las publicaciones a partir del estudio de las palabras que se emplean, su frecuencia y las relaciones más predominantes.

5.2.1. Cantidad de párrafos y palabras por cada publicación anual

La Figura 3 resume las principales medidas absolutas sobre la cantidad de párrafos y palabras por libro. Se observa que en general las extensiones son comparables, con aproximadamente 1050 párrafos que incluyen más de 60.000 palabras. Al analizar palabras no repetidas, la dispersión se acentúa dado que en el 2019 se han utilizado 9018 palabras y en el 2021 el número asciende a casi 10000.



	year <fct>	Paragraph_by_year <int>
1	2019	1057
2	2020	1047
3	2021	1053

	year <fct>	n <int>
1	2019	62905
2	2020	62593
3	2021	65553

	year <fct>	different_words <int>
1	2019	9018
2	2020	9599
3	2021	9828

Figura 3: Resumen estadístico de cantidad de párrafos y palabras por publicación. Sección izquierda: Para cada año se observa la cantidad de párrafos (paragraph _by_year), la cantidad de palabras (n) y la cantidad de palabras no repetidas (different _ words). Sección derecha: cantidad de palabras por año de publicación.

5.2.2. Palabras más frecuentes por cada publicación anual

En esta sección el interés principal fue reconocer aquellas palabras con mayor presencia en cada una de las publicaciones anuales. Para ello resultó necesario realizar una nueva depuración del corpus orientada a eliminar aquellos términos que no aportan información tales como artículos, preposiciones, pronombres y conectores, entre otros; una instancia por demás habitual en las rutinas de text mining. Dichos términos son llamados “stopword” y en esta oportunidad se trabajó con un set de “stopwords” personalizado que incluye:

- Lista de stopwords por defecto para textos en idioma español
- Lista de stopwords por defecto para textos en idioma inglés
- Lista de stopwords por defecto para textos en idioma portugues
- Lista de stopwords ad hoc con términos como "Lic", "Dr.", "Dra.", "Ing.", entre otros.

Cabe destacar que la definición de los vocablos que deben ser eliminados de un corpus para su posterior análisis refiere a un campo de trabajo en sí mismo.

En la Figura 4 se puede observar que en los tres años de publicación las palabras “mujeres” “ingeniera” “ser” “universidad” tienen un predominio por sobre otro términos, lo cual tiene sentido si lo ponemos en contexto ya que la invitación a la formulación de los ensayos estuvo orientado a visibilizar el rol de las mujeres en las Ingenierías y la convocatoria se dirigió fuertemente hacia mujeres trabajadoras parcial o exclusivamente en Universidades.

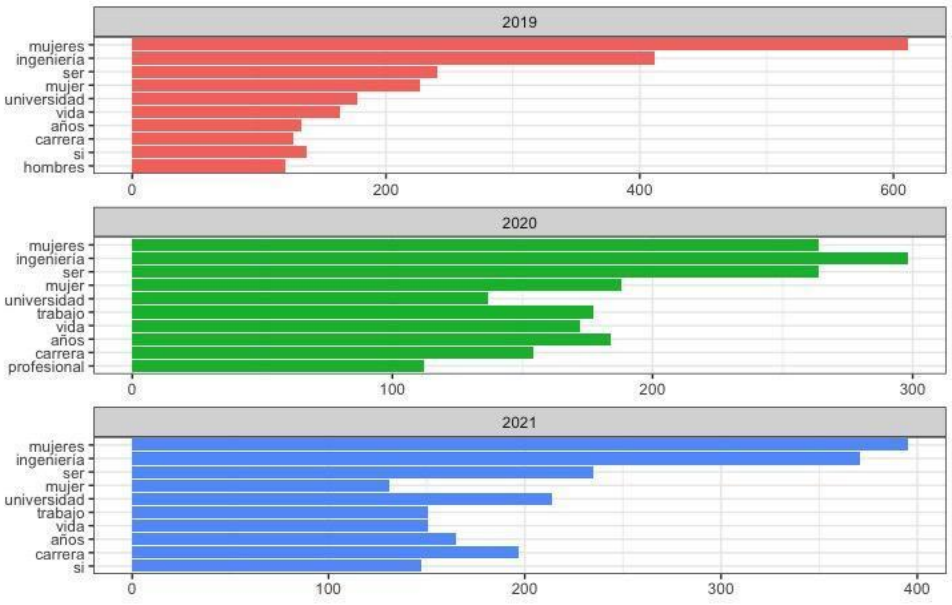


Figura 4: Vista seteada de palabras más frecuentes por año de publicación.

Otra forma de visualizar los términos más frecuentes es mediante nube de palabras (word clouds). En este tipo de representación gráfica, aquellas con mayor cantidad de repetición son presentadas con mayor tamaño de visualización. En la Figura 5 se realizó el análisis en dos sentidos, por un lado una representación amplia con más de 500 de los principales términos recolectados en el corpus completo, y por el otro, el resultado visual de la nube de palabras para cada publicación anual. En los cuatro casos, puede observarse que el contexto universitario sigue siendo fuertemente predominante.

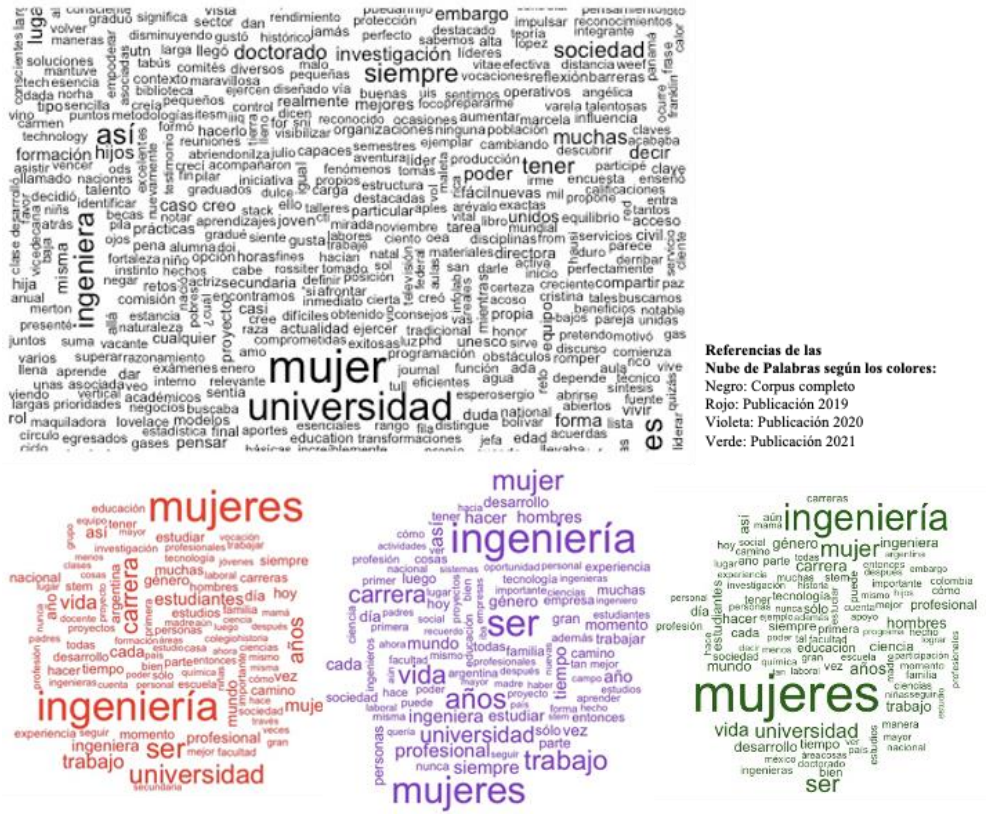


Figura 5: Nube de palabras de corpus completo y por cada publicación anual.

6. Conclusiones preliminares y líneas futuras de trabajo

El análisis de minería de texto realizado sobre la colección de libros “Matilda y las Mujeres en Ingeniería en América Latina” se aúna a los estudios de géneros que buscan problematizar las brechas presentes en diversos campos producto de las desigualdades históricas y estructurales de la sociedad en su conjunto. En particular, dado que al momento no se han registrado estudios precedentes sobre la obtención de conocimiento a partir de la aplicación de técnicas de ciencia de datos en los textos publicados por la CAL-Matilda, se espera que el mismo alcance relevancia y trascendencia en la comunidad afín.

Así mismo, puesto que constituye un capital valioso para avanzar hacia la interpretación de las representaciones sociales generadas en torno a las mujeres académicas y profesionales en STEM en América Latina, se espera que la investigación sirva para motivar futuras investigaciones análogas que superen el alcance de las Ingenierías e indaguen el campo de las Tecnologías, las Ciencias y las Matemáticas.

Sobre los principales resultados obtenidos en el incipiente trabajo de exploración es notable que en las tres ediciones del libro casi no aparecen términos relacionados con “violencia” o “diversidad” y directamente halla nula presencia del término

"trans-". La propagación masiva de "casos de éxitos" de las mujeres trabajadoras en Ingeniería con una mirada sexista y binaria podría fortalecer efectos indeseados como el sesgo de representatividad o el efecto pitufina[1] e impactar negativamente en el espíritu principal de perseguir una educación más justa e igualitaria.

Para trabajos futuros se advierten tres líneas propositivas. La primera de ellas se refiere a optimizar y profundizar el actual análisis de minería de texto realizado sobre los Libros de la CAL-Matilda, concretamente aplicando nuevas rutinas de trabajo exploratorio orientadas a evaluar frecuencias de bigramas o n-gramas y evaluando la sensibilidad del corpus de trabajo con diferentes rutinas de limpieza de stopword. Para la segunda línea futura se propone realizar un análisis de sentimientos, con el objetivo de evaluar el contenido emocional de los textos. Finalmente, la tercera línea sugiere realizar un modelado de tópicos orientado a reconocer cuáles son los grupos naturales de palabras a partir de la aplicación de métodos de clustering no supervisado.

7. Referencias bibliográficas.

CASTORINA, José Antonio; BARREIRO, Alicia. Los usos de las representaciones sociales en la investigación educativa. Educación, Lenguaje y Sociedad, 2012, vol. 9, no 9.

FAYYAD, Usama M., et al. Knowledge Discovery and Data Mining: Towards a Unifying Framework. En KDD. 1996. p. 82-88.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 1996, vol. 39, no 11, p. 27-34.

GARCÍA HOLGADO, Alicia, et al. La brecha de género en el sector STEM en América Latina: Una propuesta europea. 2019.

GARCÍA-HOLGADO, Alicia; DÍAZ, Amparo Camacho; GARCÍA-PEÑALVO, Francisco J. Engaging women into STEM in Latin America: W-STEM project. En Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality. 2019. p. 232-239.

MORALES INGA, Sergio; MORALES TRISTÁN, Oswaldo. ¿ Por qué hay pocas mujeres científicas? Una revisión de literatura sobre la brecha de género en carreras STEM. 2020.

PÁEZ PINO, Adriana. CAL Matilda y las mujeres en ingeniería. Revista de Ingeniería, 2020, vol. 67.

RAFAEL, Crespo García. Género y STEM: una falsa antagonía: Gender and STEM: A false antagonism. Universidad-Verdad, 2019, no 75, p. 61-70.

[1] El "síndrome de Pitufina" es fenómeno acuñado en 1991 por Katha Pollitt, refiere al hecho en el cual se muestra un único personaje femenino rodeado de personajes masculinos, parafraseando aquella historia de la aldea de los pitufos azules en donde solo conviven con una pitufa.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Introduction to data mining. Pearson Education India, 2016.

WITTEN, Ian H., et al. Practical machine learning tools and techniques. En Data Mining. 2005.

8. Agradecimientos

Comité de Investigación y Comité de Mentoreo de la Cátedra Abierta Latinoamericana Matilda.
Equipo de Género de la Facultad de Ingeniería de la Universidad Nacional de Lomas de Zamora.