Modelado de desigualdad de género en carreras TIC argentinas mediante regresión múltiple

Guillermo Rodriguez[1], Gabriela Espinoza Picado[2]

guillermo.rodriguez@isistan.unicen.edu.ar, gespinozapicado@gmail.com

[1]ISISTAN, CONICET –UNICEN, Tandil, Buenos Aires, Argentina

[2]Universidad de Palermo, Buenos Aires, Argentina

Resumen. Este trabajo presenta un análisis cuantitativo de la brecha de género en carreras de Computación, Sistemas e Informática (CSI) en universidades argentinas, en el período 2010-2015. A partir de un enfoque de Ciencia de Datos, se empleó aprendizaje automático supervisado, utilizando un modelo de regresión lineal múltiple, para identificar los principales factores que afectan la graduación de estudiantes. Los resultados muestran que ser estudiante mujer disminuye significativamente la probabilidad de egreso, mientras que estudiar en universidades privadas o en la provincia de Buenos Aires incrementa dicha probabilidad. Este estudio proporciona evidencia empírica relevante para el diseño de políticas públicas orientadas a reducir las desigualdades de género en carreras STEM. Como trabajo futuro, se propone analizar esta problemática incorporando técnicas de machine learning más avanzadas y extendiendo el análisis a otras carreras STEAM.

Palabras Claves: Desigualdad de género, Ciencias de la Computación, Informática, Carreras universitarias, Análisis de datos.

Modeling Gender Inequality in Argentine ICT Degrees through Multiple Regression

Abstract. This paper presents a quantitative analysis of the gender gap in Computer Science, Information Systems, and Informatics (CSI) degree programs at Argentine universities during the 2010–2015 period. Using a Data Science approach, supervised machine learning was applied—specifically, a multiple linear regression model—to identify the main factors influencing student graduation. The results show that being a female student significantly decreases the probability of graduation, while studying at private universities or in the province of Buenos Aires increases this probability. This study provides relevant empirical evidence for the design of public policies aimed at reducing gender inequality in STEM fields. As future work, the

analysis will be extended by incorporating more advanced machine learning techniques and including other STEAM-related degrees.

Keywords: Gender inequality, Computer Science, Informatics, University degrees, Data analysis.

1. Introducción

Al igual que en otras áreas, STEM (Ciencia, Tecnología, Ingeniería y Matemática con sus siglas en inglés) no se encuentra alejada de las dificultades dispares entre los diferentes géneros. Según datos del Instituto de Estadística de la UNESCO, en el sector empresarial sólo alrededor del 6% de los países alcanzó la paridad. Incluso, en países como Argentina y Uruguay, donde han alcanzado la paridad de género en el sector público, los hombres se encuentran sobrerrepresentados en el sector privado, donde los salarios suelen ser más elevados.

Por muchos años y en diversas partes del mundo, afirma la UNESCO (UNESCO, 2014) las niñas crecen con la idea y la convicción de que las materias STEM son temas "masculinos" y que la capacidad femenina en STEM es innatamente inferior a la de los hombres. Si bien la exploración sobre integrantes biológicos desmiente cualquier base fáctica de tales creencias, lo cierto es que estas persisten y socavan la confianza, el interés y la voluntad de las niñas para participar en las materias STEM. A pesar de las ventajas y grandes contribuciones que traen las TIC´s, las mujeres aún tienen menos de dos tercios de la oportunidad económica que tienen los hombres, afirma Quoc Hung de ONU Mujeres (Quoc Hung, 2019) en el marco del Día Internacional de las Mujeres y las Niñas en Ciencia y Tecnología, durante la Cuarta Revolución Industrial.

Para la CEPAL (CEPAL, 2016), resulta clave entender cómo es el contexto de la universidad para entender la brecha. América Latina presenta un fuerte rezago en materia de educación universitaria, de posgrado y de producción científica. Hay una alta deserción de mujeres en estas carreras y una inadecuada matrícula universitaria Si las estudiantes no reciben una formación universitaria, se reducen los potenciales estudiantes de posgra-

dos orientados a la Inversión y el Desarrollo (I+D), no se podrá revertir estas brechas porque estas áreas propician las actividades de docencia, ciencias y tecnologías.

Por su parte, la UNESCO (UNESCO, 2019), en su reporte sobre las *Women in Science*, señala que entre las razones por las que continúan siendo relevantes estos temas dentro del este ámbito es que las universidades tampoco están tomando las acciones correspondientes para cerrar esta brecha, por ejemplo, poner incentivos como becas, hacer un llamado inicial a mujeres para recomendaciones de trabajos, más profesoras mujeres para normalizar y ver profesionales del mismo género, galardonar trabajos académicos creados por mujeres y publicarlos, estar conscientes de la desigualdad de condiciones en que las mujeres viven, etc. Estas y muchas otras medidas que se pueden tomar para poder incrementar el porcentaje de mujeres en carreras TIC's y fomentar adicionalmente mayor número de profesionales investigadoras mujeres.

Este trabajo tiene como objetivo comparar el total de egresados de carreras asociadas a la programación que permite la comparación entre medias de dos poblaciones independientes, por ejemplo, comparar la matrícula y el egreso de estudiantes hombres y mujeres en carreras relacionas a la programación. Asimismo, el objetivo principal de la investigación será determinar las variables que han afectado la totalidad de los egresados en CSI, de los estudiantes en toda la Argentina, durante los años 2010-2015, considerando los siguientes macro factores; género, tipo de universidad o instituto de origen, ubicación geográfica, re cursantes, nuevos ingresos, éstas entre otras variables, mediante el uso de *machine learning*, utilizando específicamente aprendizaje supervisado para realizar el análisis con el método de regresión lineal múltiple.

El resto del trabajo se organiza de la siguiente manera: la Sección 2 plantea la hipótesis de investigación. La Sección 3 describe el caso de estudio realizado. La Sección 4 reporta los resultados obtenidos. Finalmente, la Sección 5 concluye el trabajo y plantea futuras líneas de investigación.

2. Hipótesis de la investigación

Existen factores que atentan contra el número de matriculadas y egresadas en carreras de TIC's, en la Argentina en el período en estudio. *Machine learning* estudia algoritmos informáticos según Lee, Chen, y Lee (Lee y otros, 2019) para aprender a realizar tareas específicas. El aprendizaje que se realiza siempre se basa en algún tipo de observaciones o datos, para una directa experiencia.

El objetivo principal es que los algoritmos deberían ser eficientes para maximizar el tiempo y el espacio, ya que esto se considera muy valioso por la cantidad de datos que producen en la actualidad, buscando además algoritmos que se pueden aplicar fácilmente a una amplia clase de problemas de aprendizaje.

Esto se no limita únicamente al campo de las TIC's, sino que se vuelve interdisciplinario, para hacer políticas públicas de la manera adecuada, mejorar la salud, también en el campo social, educativo, biológico, etc, puede ser aplicado a cualquier ámbito y visibilizar lo que antes era inmedible y menos con la precisión requerida para así y mejorar la vida de los seres humanos y del mundo.

La regresión es un *machine learning* automático supervisado, como lo explica Schapire (Schapire, 2019) lo que significa que debe proporcionar un conjunto de datos de entrenamiento etiquetado que tenga algunas variables de características independientes y una variable dependiente. El algoritmo de regresión identifica las relaciones entre las variables de características y la variable dependiente.

En esta investigación se compara el total de egresados de carreras asociadas a la programación que permite la comparación entre medias de dos poblaciones independientes, por ejemplo, comparar la matrícula y el egreso de estudiantes hombres y mujeres en carreras relacionas a la programación. El análisis se aplica bajo las siguientes hipótesis:

$H_0 =$ Existen factores que atentan contra el número de matriculadas y egresadas en carreras de TIC's, en la Argentina en el período en estudio.

H_1: No existen factores que atentan contra el número de matriculadas y egresadas en carreras de TIC's, en la Argentina en el período en estudio.

Con un nivel de significancia $\alpha = 0{,}05$.

3. Caracterización descriptiva de la población de estudio

Es importante entender la realidad actual de carreras que tienen que ver con programación en Argentina para los años 2010-2015. Chicas en Tecnología (en adelante, "CET") y Medallia comparten la base de datos del primer relevamiento cuantitativo de mujeres en programación de la Argentina. Se analizan 73 carreras relacionadas con programación de más de 80 universidades e institutos universitarios de todo el país.

Año	Estudiantes Varones	Estudiantes Mujeres	Egresados Varones	Egresados Mujeres	Porcentaje de graduación Varones	Porcentaje de graduación Mujeres
2010	60.077	14.559	2.818	814	4,7%	5,6%
2011	72.128	16.055	3.671	1.115	5,1%	6,9%
2012	66.329	14.437	3.233	830	4,9%	5,7%
2013	61.524	12.688	2.959	730	4,8%	5,8%
2014	63.683	12.648	3.078	654	4,8%	5,2%
2015	66.638	12.801	3.073	718	4,6%	5,6%
Total	390.379	83.188	18.832	4.861	4,8%	5,8%

Fig. 1. Datos de la población a estudiar

Como se mencionó anteriormente, se repite el patrón de poca participación de mujeres en carreras con materias de programación (ver Fig. 1, Fig. 2 y Fig. 3). Se observa que, de 473.567 estudiantes en este tipo de carreras, sólo 18.832 fueron estudiantes mujeres. Sin embargo, la tasa de egresadas es mayor que la de sus pares hombres para todos los años, resultado semejante a estudios antes mencionados.

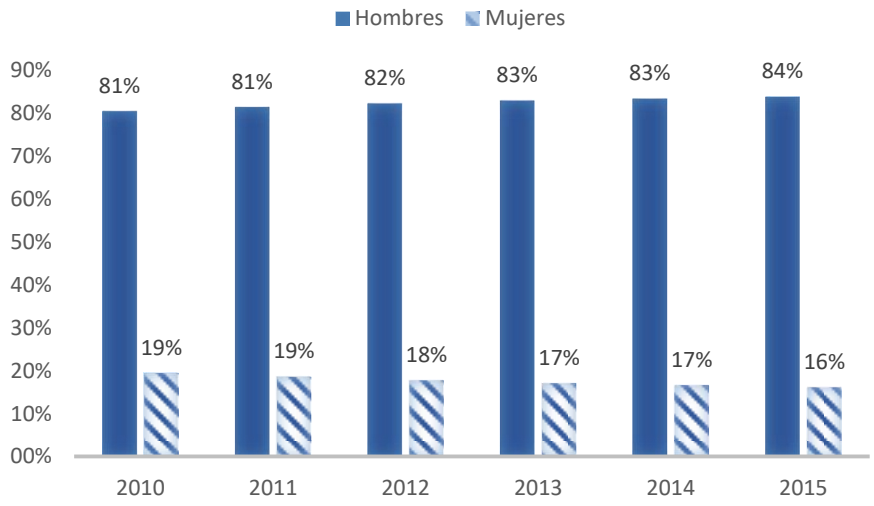

Fig. 2. Gráfico tasa de matrícula.

Además, la mayor concentración de estudiantes se encuentra en la provincia de Buenos Aires.

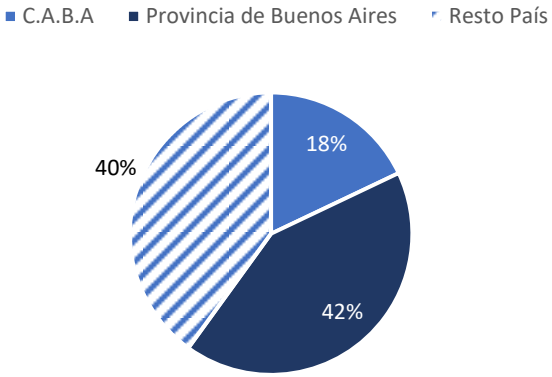

Fig. 3. Porcentaje de estudiantes por área.

4. Evaluación del modelo propuesto

En esta sección se reportan los resultados obtenidos. Mediante el lenguaje R, programo la regresión lineal múltiple. Primeramente, se realizó una regresión con 16 variables independientes. Se concluyó que hay un efecto de enmascaramiento entre las variables ya hay variables marginalmente significativas y que hacen ruido en el modelo, por lo tanto, para mejorar la selección de las variables y calibrar mejorar el modelo se usa la función stepAIC backward, utilizando la función stepAIC ( ) del paquete MASS de R. Los criterios de información de Akaike, esa una medida de ajuste del modelo que tiene en cuenta la complejidad de este, toma en cuenta el número de parámetros y corrige posibles problemas de *overfitting* en el modelo, (Consuelo, Robert Suchting, & Olvera, 2018).

$$Total\ Egresados = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i}$$

$$\begin{aligned} Total\ Egresados &= 1{,}33 + 6{,}69 Año + 5{,}83\ TipoInstitucion + 9{,}79\ \text{BuenosAires} \\ &+ 8{,}94\ \text{UniversidadPrivada} + 7{,}90\ \text{Grado} \\ &+ 4{,}80\ EstudiantesVarones - 3{,}93\ NuevasMujeres \end{aligned}$$

4.1 Evaluación de la calidad del modelo

Para para verificar que efectivamente se puede aplicar regresión lineal en este caso de estudio, existen algunos supuestos estadísticos que el modelo debe cumplir. Estos son algunos:

- **Linealidad**

El R^2 es de 0,8188, es decir las variables elegidas explica en 82% el modelo. Este coeficiente también conocido como coeficiente de determinación se utiliza para ver el grado efectividad que tienen las variables independientes en explicar la variable dependiente.

Existe linealidad si se presenta una relación significativa entre la variable que se quiere predecir y las otras variables. Puede usarse el coeficiente R cuadrado ajustado, para saber si existe linealidad, debe tener un valor mayor o igual a 0,7.

Además, se contrasta con el estadístico F, el cual se usa para determinar si de entre un grupo de variables independientes, existe al menos que explica una parte significativa de

la variación de la variable dependiente F-statistic: 35.44, con *p-value 0*,002, por lo tanto, el modelo es globalmente significativo.

- **Normalidad**

  Los residuos deben presentar una distribución normal, y la ausencia de normalidad supone poca precisión en los intervalos de confianza creados por el modelo. Los residuos se deben distribuir de forma normal con media cero. La mayor cantidad de errores cercanos al cero por lo que se puede decir que es normal. Otra forma de probar la normalidad, es usando el test Shapiro-Wilk (W=0.751, p-valor=2.2e-16). Se comprueba el contraste de normalidad para los residuos estandarizados del modelo ajustado.

- **Homocedasticidad**

  Este supuesto asume que los residuos en las predicciones son constantes en cada predicción (es decir, varianza constante). Este supuesto validó que los residuos no aumentan ni disminuye cuando se predicen valores cada vez más altos o más pequeños.

  La varianza de los residuos debe de ser constante en todo el rango de observaciones. Otra forma de comprobarlo es utilizando la prueba Breusch-Pagan, con un *p-value* < 0,05 se comprueba que no hay evidencia de que la varianza de los residuos es homocedástica (BP=138.69, p-valor=2.2e-16).

- **Autocorrelación de los errores**

  Este supuesto asume que los residuos no están auto-correlacionados, por lo cual son independientes. La autocorrelación es cuando el residuo en la predicción de un valor es afectado por el residuo en la predicción del valor más cercano. Utilizando la prueba Durbin-Watson (DW=1.84, p-valor=0.011) puede concluirse que no existe dependencia de los residuos.

- **Multicolinealidad**

  Por último, hay que hacer pruebas de multicolinealidad que se refiere a una o varias variables explicativas son una combinación lineal de otra(s).

Resumiendo, el modelo propuesto cumple todos los supuestos para que el modelo regresivo explique el fenómeno, a continuación, los resultados del modelo, una vez confirmado que el modelo es lo suficientemente robusto, para explicar el fenómeno.

4.2 Análisis de resultados

En esta sub-sección se analizan los resultados obtenidos luego de obtener y evaluar el modelo de regresión lineal logrado.

	β	Pr(>|t|)	
(Intercept)	1,33	0,1	
Año	-6,69	0,096	
Tipo Institución	5,83	0,187*	.
Buenos Aires	9,79	6,99e-11	***
Universidad Privada	8,94	2,80e-07	***
Grado	7,91	6,83e-06	***
Estudiantes Varones	4,81	< 2.e-16	***
Estudiantes Mujeres	-3,93	2,61e-11	***

Fig. 4. Resultados del modelo.

Los "***" que se denotan en la Fig. 4, significa que con un 95% de confianza se rechaza la hipótesis nula y demuestra que la variable independiente tiene influencia sobre la variable dependiente.

Por cada estudiante graduado de las carreras de CSI, haber estudiado en la provincia de Buenos Aires, aumenta su probabilidad de egreso en 9,79 veces con respecto a sus pares de otras regiones de Argentina. Además, si una persona estudia un grado y no un pregrado, es decir una carrera universitaria completa tiene más oportunidades de egresarse en 7,91 veces, que quien estudia un pregrado para los estudiantes del período de estudio.

Si los egresados, estudiaron en una universidad privada, aumentaron su probabilidad de graduación en 8,94 veces con respecto a estudiantes de universidades públicas.

Si el egresado es hombre, tiene 4,81 veces más posibilidades de graduarse que sus pares mujeres. En cambio, si las estudiantes son mujeres tienen -3,93 veces menos posibilidades de egresarse para las carreras antes mencionadas en Argentina para los años 2010-2015.
En otras palabras un estudiante varón que realizó sus estudios en Buenos Aires, en una universidad privada tiene mayores posibilidades de graduarse.
En conclusión, exceptuando año y tipo de institución, todas las demás variables están relacionadas con la brecha de género en CSI, con un grado de confianza del 95%, es decir con un error estadístico menor al 5%, por lo que se rechaza la hipótesis nula.

5. Conclusiones

En este trabajo se presentó un modelo de regresión lineal que permite explicar la brecha de género que existe en carreras de Ciencias de la Computación e Informática en universidades argentinas. Se comprueba la hipótesis, no sólo hay menos mujeres en carreras relacionas con programación, sino que, sólo el hecho de ser estudiante mujer disminuye la probabilidad de graduarse.

Se recomienda tomar medidas de políticas públicas por parte del Estado, como, por ejemplo; impartir clases de programación que se dicten en nivel primario y secundario. También se apliquen políticas a nivel universitario para aumentar los incentivos para que las mujeres escojan este tipo de carreras. Estos ajustes podrían cambiar el paradigma y estigma acerca de que las mujeres son malas en matemática y lógica.

Por último, es importante realizar el mismo ejercicio en carreras con programación en su plan de estudios para los años venideros y analizar si esta brecha de género se ha disminuido entre hombres y mujeres. También sería importante medir cursos abiertos cortos de programación para estimar si en estos casos la brecha disminuye, ya que cada vez más hay cursos si bien no universitario, pero cursos cortos de ciencia de datos o cursos que las empresas brinden como actualización profesional.

Como trabajo futuro, el plan es realizar el mismo estudio para cada una de las carreras STEAM (disciplinas que combinan Ciencia, Tecnología, Ingeniería, Arte y Mate-

máticas) en Argentina en el mismo período para estudiar el comportamiento de las carreras universitarias que quedaron por fuera de este estudio.

Referencias

A. Powell, A. D. (2012). Gender stereotypes among women engineering and technology students in the UK: lessons from career choice narratives. *European Journal of Engineering Education, Vol. 37, No. 6, 2012, pp. 541-556., 37*(6), 541-556.

Baytiyeh, H. (2013). Are women engineers in Lebanon prepared for the challenges of an engineering profession? *European Journal of Engineering Education, 38*(4), 394-407.

Burke, R., & Mattis, M. (s.f.). Women and minorities in Science, Technology, Engeeniring and Mathematics. *ISBN 978 1 84542 888 4*.

CEPAL, Z. (2015). *La industria del software y los servicios informáticos: un sector de oportunidad para la autonomía económica de las mujeres latinoamericanas.*

Cimpian, J., Kim, T., & McDermott, Z. (2020). Understanding persistent gender gaps in STEM: Does achievement matter differently for men and women? *Science, Vol 368 Issue 6497*.

Consuelo, W.-B., Robert Suchting, R. L., & Olvera, D. E. (2018). Inflammatory markers as predictors of depression and anxiety in adolescents: Statistical model building with component wise gradient boosting. *Journal of Affective Disorders, 234*(276-281).

Díaz, S. (2016). Promoción estudios STEM, Ciencia, Tecnología, Ingeniería y Matemáticas, en Navarra. *Universidad Pública de Navarra*.

E. Hirose, K. M. (2011). Increasing the number of women in engineering at universities and colleges in Japan. *American Society for Engineering Education*.

Eagly, A., & Steffen, V. (1984). Gender stereotypes STEM from the distribution of women and men into social roles. *Journal of Personality and Social Psychology, 46*, 735-754.

Empresarial, U. d. (2016). *Ciencia, tecnología e innovación en la economía digital: Situación de América Latina y el Caribe.* Comisión Económica para América Latina y el Caribe.

Espinoza, G. (2017). *Caracterización de los estudiantes que ingresan a la Universidad de Costa Rica en el período 2010 - 2013*. Universidad de Costa Rica.

Falgueras, I. (2008). Teoría del Capital Humano: origenes y evolución. Temas Actuales de Economía. *2*, 19-48.

Garbanzo, V. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. Revista Educación. (1), 43-63.

Giménez, G., & Castro, G. (s.f.). ¿Por qué los estudiantes de colegios públicos y privados de Costa Rica obtienen distintos resultados académicos? *Perfiles Latinoamericanos, 49*, 1-37.

Griffith, A. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review, 29*, 911-922.

Jiménez, C., Jones, E., & Vidal, C. (2019). Estudio Exploratorio de factores que influyen en la decisión de la mujer para estudiar ingeniería en Chile. *Información Tecnológica , 30*, 209-219.

K. Beddoes, M. B. (2011). Feminist Theory in Three Engineering Education Journals: 1995-2008. *Journal of Engineering Education, 100*(2), 281-303.

kvochko, E. (2013). *Five ways technology can help the economy.* World Economic Forum.

Lee, C.-F., Chen, H.-Y., & Lee, J. (2019). Mathematics and Statistics: Theory, Method and Application. *Springer*, ISBN 979-1-4939-9427-4.

Lourens, A. S. (2013). The design of a leadership development programme for women engineering students at a Sudafrican university . *120th ASEE Annual conference and exposition.*

Margolis, J., Fisher, A., & Faye, M. (2010). The Anatomy of Interest Women in Undergraduate Computer Science. *Womens Stud. Q. Spec. Issue Women Sci. 28*.

Moguillansky , G. (2015). *La importancia de la tecnología de la información y la comunicación para las industrias de recursos naturales.* Publicación de las Naciones Unidas , UNESCO, CEPAL, Chile.

Moses, L., Hall, C., Wuensch, K., De Urquidi, K., Kauffman, P., Swart, W., & Dixon, G. (2011). Are Math Readiness and Personality Predictive of First-Year Retention in Engineering? *The Journal of Psychology: Interdisciplinary and Applied, 3*, 229-245.

ONESCO. (2016). Inequidad y género en los logros de aprendizaje en educación primaria: ¿Qué nos puede decir TERCE?

Organización de las Naciones Unidas para Educación, la Ciencia y Cultura. (2014). Enseñanza y aprendizaje: lograr la calidad para todos.

Quoc Hung, P. (2019). *Se necesitan cientificas* . Organización de las Naciones Unidas.

Reyes, M. (2011). Unique Challenges for Women of Color in STEM Transferring from Community Colleges to Universities. *Harvard Educational Review, 81*, 241-263.

Rubio, M., & Berlanga Silvente, V. (2012). Cómo aplicar las pruebas paramétricas bivariadas t de Student y ANOVA en SPSS. *Revista d'Innovació n Recerca en Educació(REIRE)*, 83-100.

S. Bucak, N. K. (2011). Influence of gender in choosing a career amongst engineering fields: a survey study from Turkey. *European Journal of Engineering Education, 36*(5), 449-460.

Sanabria, E., Granados, A., & Matamorros, J. (2019). Percepción de estudiantes del Recinto de Paraíso sobre las razones que podrían mediar la brecha de género en el ingreso a la carrera de Informática Empresarial investigación. *Revista Technology Inside, 4*(ISSN: 2215-5392), 11-20.

Schapire , R. (2009). Princeton.

Schapire , R. (2009). *Theoretical Machine Learning.* Princeton.

Shapiro, C., & Sax, L. (2011). Major Selection and Persistence for Women in STEM. *New Directions for Institutional Research , 152*, 5-18.

UNESCO. (2014). Igualdad de género, patrimonio y creatividad. *ISBN 978-92-300008-7*.

UNESCO. (2016). Inequidad de género en los logros de aprendizaje en educación primaria ¿Qué nos puede decir TERCE?

UNESCO. (2017). *Por muchos años y en diversas partes del mundo, las niñas crecen con la idea y la convicción que las materias STEM son temas "masculinos.*

UNESCO. (2020). *Un nuevo informe de la UNESCO pone de relieve las desigualdades de género en la enseñanza de las ciencias, la tecnología, la ingeniería y las matemáticas (STEM).*

United National Educational, S. a. (2019). *Women in Science* (Vol. FS/2019/SCI/55). Fact Sheet No. 55.