

## **Detección automatizada de factores determinantes del embarazo en la adolescencia** **Automated detection of determining factors of pregnancy in adolescence**

Bernardo Céspedes-Panduro<sup>1\*</sup> <https://orcid.org/0000-0002-9606-1478>

Zoraida Judith Huamán-Gutiérrez<sup>2</sup> <https://orcid.org/0000-0002-5533-2454>

<sup>1</sup>Doctor en Estadística Matemática. Docente Auxiliar. Investigador. Departamento de Estadística, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.

<sup>2</sup>Doctor en Estadística Matemática. Docente Principal. Investigador. Departamento de Estadística, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.



\*Autor para la correspondencia. Correo electrónico: [bcspedesp@unmsm.edu.pe](mailto:bcspedesp@unmsm.edu.pe)

### **RESUMEN**

**Introducción:** las adolescentes embarazadas tienen menos probabilidades de construir relaciones sentimentales estables, y más de sufrir trastornos emocionales. También, son más susceptibles a presentar diversas complicaciones durante el embarazo y el parto.

**Objetivo:** evaluar las técnicas de aprendizaje automático para determinar factores de riesgo del embarazo en adolescentes.

**Métodos:** se realizó una investigación con diseño correlacional causal. Los datos se obtuvieron de la Encuesta Demográfica y de Salud Familiar - ENDES 2021 Nacional y Departamental, que abarcó los años 2018 a 2020. En el momento de las entrevistas, su base de datos contenía información de 16 825 mujeres adolescentes peruanas de 12 a 19 años de edad, las cuales constituyeron el universo de estudio. Se implementaron nueve algoritmos: máquina de soporte vectorial, regresión logística binaria, árbol de decisión, refuerzo adaptativo, potenciación del gradiente, refuerzo de gradiente extremo, árboles extremadamente aleatorios, agregación de arranque, y bosque aleatorio. Se consideraron sus métricas como variables a tener en cuenta en la evaluación, la precisión, y el área bajo la curva.

**Resultados:** el algoritmo más preciso fue el bosque aleatorio (0,965825), seguido por la potenciación del gradiente (0,963744), el árbol de decisión, y las máquinas de vectores de soporte (0,963155, ambos).

**Conclusiones:** el bosque aleatorio fue la técnica más precisa; además de la identificación de los factores en cuestión, se distinguieron los tres más importantes. Este estudio es un precedente valioso para la aplicación de las técnicas de aprendizaje automático en la predicción de diversas variables necesarias para mejorar la gestión pública.

**Palabras clave:** automatización; bases de datos estadísticos; embarazo en adolescencia; factores de riesgo.

## ABSTRACT

**Introduction:** pregnant adolescents are less likely to form stable romantic relationships and are more likely to suffer emotional disorders. They are also more susceptible to various complications during pregnancy and childbirth.

**Objective:** to evaluate machine learning techniques to determine risk factors for teenage pregnancy.

**Methods:** a research with a causal correlational design was carried out. The data was obtained from the Demographic and Family Health Survey - ENDES 2021 National and Departmental, which covered the years 2018 to 2020. At the time of the interviews, their database contained information on 16,825 Peruvian adolescent women aged 12 to 19 years, who constituted the study universe. Nine algorithms were implemented: support vector machine, binary logistic regression, decision tree, adaptive boosting (AdaBoost), gradient boosting, extreme gradient boosting (XGBoost), extremely random trees (ExtraTrees), bootstrap aggregation, and random forest. Their metrics were considered as variables to be taken into account in the evaluation, precision, and the area under the curve.

**Results:** The most accurate algorithm was the random forest (0.965825), followed by gradient boosting (0.963744), decision tree, and support vector machines (0.963155, both).

**Conclusions:** the random forest was the most accurate technique; in addition to the identification of the factors in question, the three most important ones were distinguished. This study is a valuable precedent for the application of machine learning techniques in the prediction of various variables necessary to improve public management.

**Keywords:** automation; pregnancy in adolescence; risk factors; statistical databases.

Recibido: 09/08/2024

Aprobado: 23/09/2024

Publicado: 15/01/2025

## INTRODUCCIÓN

En Perú, la proporción de la fecundidad en la adolescencia con respecto a la fecundidad total ha aumentado. Este es el resultado de descensos menos significativos y oscilantes en las cifras de embarazos entre las mujeres de 12 a 17 años; aun cuando se observa un descenso estable de las tasas globales de fecundidad.<sup>(1)</sup>

Según la Encuesta Demográfica y de Salud Familiar - ENDES 2021 Nacional y Departamental,<sup>(1)</sup> entre 2018 y 2020 la fecundidad de las adolescentes peruanas disminuyó 0,5 puntos porcentuales (de 2,8 % a 2,3 %). Esta disminución no es particularmente sustancial, si se tienen en cuenta las consecuencias graves del embarazo en la adolescencia. El cual, desde el punto de vista social, incide en la elevación de los indicadores de abandono escolar en la educación secundaria, aumenta la susceptibilidad a la participación en redes delictivas o de explotación sexual comercial,<sup>(2)</sup> y en última instancia reduce las perspectivas de empleo a largo plazo; con ello, merman los ingresos económicos de la unidad familiar y se fomenta un nivel de pobreza más profundo.<sup>(2,3)</sup> Este fenómeno también genera importantes costes sociales.<sup>(3)</sup>

Del mismo modo, en determinadas regiones y grupos sociales la gestación precoz se observa como un fenómeno cultural. Sin embargo, en las grandes ciudades, estos embarazos no suelen ser bien recibidos; solo se dan entre novios que no conviven, o en el seno de parejas consentidas, y suelen culminar con la desprotección de la mujer y el niño, lo que implica la maternidad en solitario. Numerosos embarazos culminan en abortos practicados por personas sin formación profesional adecuada, y en condiciones insalubres, debido a la escasez y altos costes de los servicios médicos especializados y al hecho de que el aborto está prohibido en Perú.<sup>(1)</sup>

Las adolescentes embarazadas tienen menos probabilidades de construir relaciones sentimentales

estables, y más de sufrir trastornos emocionales; entre ellos, depresión, acompañada en ocasiones de ideación suicida,<sup>(3)</sup> y abuso de sustancias.<sup>(2)</sup> Por otra parte, en comparación con las mujeres adultas, desde el punto de vista biológico las adolescentes son más susceptibles a presentar diversas complicaciones durante el embarazo y el parto: preeclampsia,<sup>(4,5)</sup> anemia<sup>(4)</sup>, parto prematuro,<sup>(5,6,3)</sup> neonatos bajos de peso,<sup>(6,5)</sup> y mortalidad materna y perinatal altas.<sup>(5,3)</sup>

En 1956, Davis y Blake<sup>(7)</sup> iniciaron las investigaciones sobre los determinantes de la fecundidad, e introdujeron un modelo de once variables intermedias influyentes, clasificadas en tres grupos. En 1978, Bongaarts<sup>(8)</sup> identificó siete factores que podían influir directamente sobre la probabilidad de fecundidad; los denominó determinantes intermedios de la fecundidad.

Posteriormente, en 2006, Di Cesare y Rodríguez-Vignoli<sup>(9)</sup> propusieron un modelo basado en el de Bongaarts,<sup>(8)</sup> y categorizaron los factores influyentes en dos grupos: variables intermedias (determinantes próximos), que pertenecen a los factores más cercanos al sujeto; y factores subyacentes (determinantes indirectos), asociados a variables socioculturales, ambientales, familiares, e individuales. Este modelo incorporó las variables intermedias posteriores: edad de la primera relación sexual, uso de anticonceptivos, y estado civil. Los determinantes influyentes en la variación de los niveles de fecundidad a través de determinantes próximos de la fecundidad en la adolescencia, se clasifican en tres categorías: socioculturales, familiares, e individuales.<sup>(9)</sup>

En la actualidad, para trabajar con grandes bases de datos se utilizan técnicas de aprendizaje automático mediante procedimientos matemáticos. Los resultados se prueban en una base de datos diferente a la empleada para entrenar el modelo, y se obtienen métricas disímiles a partir de la matriz de clasificación o confusión.

En la presente investigación se analizaron las técnicas de clasificación sobre la predicción existente más utilizadas en el aprendizaje automático; de ese modo, se eligió la que diferencia mejor el embarazo en la adolescencia. Tanto en el sector público como privado de la salud, este conocimiento es imprescindible para la gestión administrativa de recursos y la ejecución de acciones preventivas encaminadas a reducir las tasas de embarazos en la adolescencia y sus repercusiones negativas para la sociedad.

Por ello, el objetivo de los autores fue evaluar las técnicas de aprendizaje automático para determinar factores de riesgo del embarazo en adolescentes.

## MÉTODOS

Se realizó una investigación con diseño correlacional causal<sup>(10)</sup> Los datos se obtuvieron de la *Encuesta Demográfica y de Salud Familiar - ENDES 2021 Nacional y Departamental*,<sup>(1)</sup> que abarcó los años 2018 a 2020. Dirigida por el Instituto Nacional de Estadística e Informática (INEI), en el momento de las entrevistas –el año escolar 2019-2020– su base de datos contenía información de 16 825 mujeres adolescentes peruanas de 12 a 19 años de edad.

Las mujeres que informaron haberse embarazado en edades entre 12 y 19 años se clasificaron como casos de embarazos en la adolescencia.

Se utilizaron ocho variables independientes, agrupadas de la siguiente manera: variables intermedias (edad en la primera relación sexual, ¿ha tenido una primera unión?, ¿ha estado casada o ha convivido?, ¿alguna vez usó cualquier método anticonceptivo o usa alguno actualmente); factores socioculturales y ambientales (etnia, área de residencia, y región de residencia); factores familiares (número de miembros del hogar, e índice de riqueza); factores individuales (migración, conocimiento del ciclo ovulatorio, ¿ha escuchado hablar de planificación familiar por medios de comunicación?, número deseado de hijos, edad –en años cumplidos–, empleo en los últimos 12 meses, educación –en años académicos–, ¿actualmente asiste a la escuela, colegio, instituto, o universidad?

Para el análisis, se revisaron las variables en el conjunto de datos de ENDES y se eliminaron las que no se consideraron significativas: las relacionadas con la fecha de la entrevista, y la identificación (ID) del caso y el conglomerado.

Se identificaron las diferentes categorías que debe incluir cada variable; los datos faltantes se imputaron mediante la función *missForest* de la librería *missForest* del programa estadístico R. Se ejecutó el análisis exploratorio de datos (EDA, por sus siglas en inglés) univariado y bivariado con el programa estadístico *Python*. Se convirtieron los datos categóricos en números, mediante la codificación uno caliente (*one hot*) con el enfoque de codificación de etiquetas. Los datos se partitionaron en conjuntos de entrenamiento y prueba; y a continuación se escalaron o normalizaron.

Se implementaron nueve algoritmos: máquina de soporte vectorial, regresión logística binaria, árbol de decisión, refuerzo adaptativo (*AdaBoost*), potenciación del gradiente (*gradient boosting*), refuerzo de gradiente extremo (*XGBoost*), árboles extremadamente aleatorios (*ExtraTrees*), agregación de arranque

(*bagging*), y bosque aleatorio (*random forest*). Estos modelos se seleccionaron por su uso recomendado y sostenido en el tiempo, tanto para el aprendizaje automático de tareas de clasificación<sup>(11,12)</sup> como en la salud pública y la investigación en ciencias del comportamiento.<sup>(13-16)</sup>

Para probar el rendimiento de los modelos, se utilizó la matriz de confusión, en busca de un indicador de precisión y error. Se calculó el área bajo la curva característica operativa del receptor (ROC (AUC), por sus siglas en inglés), utilizada como indicador de precisión.

La curva ROC es un gráfico de la tasa de positivos verdaderos de la prueba (eje Y), frente a la correspondiente tasa de positivos falsos (eje X); es decir, sensibilidad frente a especificidad. De manera simplista, AUC es la posibilidad de que una observación seleccionada al azar de que una adolescente tuvo un embarazo se clasifique como esa.<sup>(17)</sup>

Se determinó la importancia de los predictores en el modelo mejor de los encontrados. Para analizar y procesar los datos se utilizaron los programas para IBM SPSS 26.0, R) con su interfaz *R Studio*), y *Python* con su interfaz (la aplicación *Google Colaboratory* con las “bibliotecas” necesarias: *pandas*, *SciPy*, *keras*, *NumPy*, *scikitlearn*, *TensorFlow*).

La investigación se apejó a los aspectos éticos establecidos en la *Declaración de Helsinki*.<sup>(18)</sup> Se siguieron las normas de confidencialidad establecidas para garantizar el uso de la información solo con fines científicos, sin divulgar datos específicos de los participantes en el estudio.

## RESULTADOS

Los embarazos ocurrieron en 1 830 (10,90 %) de las 16 825 adolescentes entre 12 y 19 años de edad estudiadas (Tabla 1).

**Tabla 1** - Distribución de la variable embarazo en la adolescencia

Embarazo en la adolescencia	Frecuencia	Porcentaje
No	14995	89,10
Sí	1830	10,90
<b>Total</b>	16825	100,00

Fuente: INEI – ENDES 2019-2020.

La mayoría no ha tenido una primera unión (90,60 %), no ha estado casada o convivido (98,60 %), no ha usado alguna vez algún método anticonceptivo (80,50 %), y tampoco lo usaba al momento del

estudio (89,60%). Su lengua materna es el español u otra extranjera (87,90%), no tiene conocimiento del ciclo ovulatorio (87,00%), no ha escuchado hablar sobre planificación familiar por medios de comunicación (69,50%). Ha estado desempleada los últimos 12 meses (75,40%), y asiste a la escuela, colegio, instituto o universidad (78,90%).

Por otro lado, 87,10% de las adolescentes que han tenido una primera unión se han embarazado una vez, en comparación con las que no (3,00%). De las que han estado casadas o convivido, 84,60% han tenido un embarazo, a diferencia de 9,80% que no lo han estado. De aquellas que usaron alguna vez cualquier método anticonceptivo, 51,10% han tenido un embarazo en comparación con 1,10% de las que no lo usaron. De las que al momento del estudio usaban métodos anticonceptivos, 61,30% ya se habían embarazado, en comparación con las que no los usaban (5,00%).

Por último, 16,60% de las adolescentes cuya lengua materna es otra (nativa u originaria) ya han tenido un embarazo (Tabla 2).

**Tabla 2-** Características de las adolescentes según embarazos

Variable	Adolescentes	Porcentaje	Embarazo adolescente	Porcentaje de embarazo adolescente
<b>Ha tenido una primera unión</b>				
No	15240	90,60	450	3,00
Sí	1585	9,40	1380	87,10
<b>Ha estado casada o ha convivido</b>				
No	16584	98,60	1626	9,80
Sí	241	1,40	204	84,60
<b>Alguna vez usó cualquier método</b>				
No	13538	80,50	151	1,10
Sí	3287	19,50	1679	51,10
<b>Uso actual de métodos anticonceptivos</b>				
No	15075	89,60	758	5,00
Sí	1750	10,40	1072	61,30
<b>Etnia</b>				
Castellano u otra lengua extranjera	14787	87,90	1598	10,80
Otra lengua nativa u originaria	391	2,30	65	16,60
Quechua	1647	9,80	167	10,10
<b>Área de residencia</b>				
Rural	6124	36,40	814	13,30
Urbana	10701	63,60	1016	9,50
<b>Región de residencia</b>				
Lima Metropolitana	1800	10,70	143	7,90
Resto Costa	4481	26,60	479	10,70
Selva	4445	26,40	617	13,90
Sierra	6099	36,20	591	9,7
<b>Índice de riqueza</b>				
Muy rico	1414	8,40	58	4,10
Rico	2184	13,00	142	6,50

Clase media	3019	17,90	275	9,10
Pobre	4342	25,80	534	12,30
Pobre extremo	5866	34,90	821	14,00
<b>Desplazamiento (migración)</b>				
No	12239	72,70	978	8,00
Sí	4586	27,30	852	18,60
<b>Conocimiento del ciclo ovulatorio</b>				
No	14641	87,00	1538	10,50
Sí	2184	13,00	292	13,40
<b>Ha conocido sobre planificación familiar por medios de comunicación</b>				
No	10567	62,80	1242	11,80
Sí	6258	37,20	588	9,40
<b>Empleo en los últimos 12 meses</b>				
No	12693	75,40	914	7,20
Sí	4132	24,60	916	22,20
<b>Actualmente asiste a la escuela, colegio, institución o universidad</b>				
No	3554	21,10	1482	41,70
Sí	13271	78,90	348	2,60

Fuente: INEI – ENDES 2019-2020.

La edad promedio de las adolescentes al momento de la primera relación sexual fue 14,64 años (mediana de 14 años); es decir, 50% tenían alrededor de 14 años, por lo cual se supone que este dato tiende a una distribución simétrica. El número promedio de miembros del hogar fue de 5,49 (mediana de cinco miembros), con tendencia a una distribución asimétrica. El número deseado promedio de hijos fue de 1,9, y el de años de educación 8,46 (mediana de 8 años) (Tabla 3).

**Tabla 3** -Estadísticas de las variables cuantitativas de las adolescentes

Variables independientes	Media	Desviación estándar <sup>1/</sup>	Mínimo	Mediana	Máximo
Edad en la primera relación sexual	14,64	1,218	9	14	19
Número de miembros del hogar	5,49	2,082	1	5	24
Número deseado de hijos	1,9	0,624	0	2	8
Edad (en años cumplidos)	15,22	2,258	12	15	19
Educación (en años)	8,46	2,226	0	8	15

Fuente: INEI – ENDES 2019-2020.

1/ Desviación estándar.

Para el entrenamiento se reservaron 80% de los datos, y 20% para evaluar el rendimiento del modelo. Su precisión se estableció según la proporción de predicciones correctas hechas por el modelo sobre todos los tipos de predicciones realizadas (Tabla 4). El bosque aleatorio resultó el algoritmo con mejor

precisión (0,965825), calculado a partir de las matrices de confusión. Le siguieron el impulso por gradiente (0,963744), el árbol de decisión (0,963155), y la máquina de vectores de soporte (0,963155).

**Tabla4** - Algoritmos de aprendizaje automático según su precisión

Modelo	Precisión
Bosque aleatorio	0,965825
Potenciación del gradiente	0,963744
Árbol de decisión	0,963150
Máquina de soporte vectorial	0,963150
Árboles extremadamente aleatorios	0,961664
Refuerzo adaptativo	0,957801
Regresión logística binaria	0,957504
Agregación de arranque	0,956909
Refuerzo de gradiente extremo	0,938782

Fuente: INEI – ENDES 2019-2020.

Los nueve algoritmos propuestos obtuvieron indicadores AUC superiores a 0,90. De acuerdo con los intervalos propuestos para medir la calidad de predicción del modelo a partir de la curva ROC,<sup>(19)</sup> los nueve se evaluaron de muy buenos a excelentes.

En cuanto al área bajo la curva, los algoritmos con mejores métricas (AUC de 0,99 para cada uno) fueron el bosque aleatorio, refuerzo de gradiente extremo, y regresión logística binaria. La máquina de soporte vectorial presentó 0,95; y los restantes entre 0,96 y 0,98 (Tabla 5).

**Tabla 5** - Comparación de los algoritmos según sus métricas del área bajo la curva

Modelo	Área bajo la curva
Bosque aleatorio	0,99
Potenciación del gradiente	0,99
Regresión logística binaria	0,99
Árboles extremadamente aleatorios	0,98
Refuerzo de gradiente extremo	0,98
Refuerzo adaptativo	0,97
Agregación de arranque	0,97
Árbol de decisión	0,96
Máquina de soporte vectorial	0,95

Fuente: INEI – ENDES 2019-2020.

El bosque aleatorio posibilita calcular la importancia de las variables. Mediante este algoritmo se determinó que, en concordancia con los datos de entrenamiento, los tres factores predictores más influyentes (en orden de importancia) para predecir el embarazo en la adolescencia son: ha tenido una primera unión (30,90 %), alguna vez usó cualquier método anticonceptivo (15,23 %), y actualmente asiste a la escuela, colegio, instituto o universidad (7,47%) (Tabla 6).

**Tabla 6** – Orden de importancia de las variables (factores predictores de embarazo en la adolescencia), según el algoritmo del bosque aleatorio

Variable	Descripción	Importancia (%)
Unión	Ha tenido una primera unión	30,9002
Anticonceptivo	Alguna vez usó cualquier método	15,2260
Asiste	Actualmente asiste a la escuela, colegio, instituto, o universidad	7,4697
Uso	Uso actual de métodos anticonceptivos	7,4152
Primera	Edad en la primera relación sexual	6,9659
Edad	Edad (en años cumplidos)	5,7233
Miembros	Número de miembros del hogar	5,5032
Educación	Educación (en años)	4,4718
Región	Región de residencia	2,6099
Riqueza	Índice de riqueza	2,5800
Deseado	Número deseado de hijos	2,2696
Empleo	Empleo en los últimos 12 meses	1,6407
Casada	Ha estado casada o ha convivido	1,6308
Migración	Migración	1,3363
Planificación	Han escuchado hablar de planificación familiar por medios de comunicación	1,3359
Área	Área de residencia	1,0519
Ciclo	Conocimiento del ciclo ovulatorio	1,0066
Etnia	Etnia	0,8630

Fuente: INEI – ENDES 2019-2020.

## DISCUSIÓN

En la presente investigación se constató la utilidad de las técnicas de aprendizaje automático para clasificar y predecir factores de riesgo del embarazo en la adolescencia. Los estudios predictivos sobre

este tema son de interés, debido a que con frecuencia las adolescentes embarazadas postergan su desempeño académico y laboral.

Las variables más importantes en cuanto factores predictores de embarazos en adolescentes fueron: ha tenido una primera unión, alguna vez usó cualquier método anticonceptivo, y actualmente asiste a la escuela, colegio, instituto, o universidad. Quezada y cols.,<sup>(20)</sup> en Colombia, determinaron variables similares: actualmente tiene pareja sexual, usa métodos anticonceptivos prescritos por una entidad promotora de salud, el individuo está matriculado en la escuela, inició la actividad sexual al menos a los quince años de edad, planifica sus relaciones y forma de educación; para ello, estos autores utilizaron el algoritmo J48 de aprendizaje automático.<sup>(21)</sup>

Rosales-López<sup>(22)</sup> usó el modelo de fecundidad de Bongaarts,<sup>(8)</sup> y también encontró que la variable más importante es la nupcialidad o unión temprana. Sin embargo, el metaanálisis de Fasula y cols.,<sup>(23)</sup> expuso como factores de riesgo principales para el embarazo en la adolescencia: haber sufrido abuso infantil, participar en conductas delictivas graves o repetidas, abusar de sustancias, tener un padre adolescente; disfunción familiar grave, no convivir con ninguno de los padres, y ser de etnia hispana.

Garza-Reyna y cols.,<sup>(24)</sup> detectaron otros factores: uso deficiente de los métodos anticonceptivos (a pesar de conocerlos), ser hija de padres adolescentes, haber tenido varias parejas sexuales, pertenecer a los estratos socioeconómicos medibajo y bajo, ser dependiente económicamente de los padres.

Asare y cols.<sup>(25)</sup> mediante regresión logística determinaron que las adolescentes residentes en áreas rurales, desempleadas, y de nivel económico bajo, tienen mayores probabilidades de quedar embarazadas en esa etapa de sus vidas, en comparación aquellas de áreas urbanas, escolarizadas, y nivel económico alto. Por otra parte, D'Añari-Cabrera<sup>(26)</sup> encontró que la edad de inicio de las relaciones sexuales y el nivel de instrucción se asocian con el embarazo en la adolescencia, de forma estadísticamente significativa.

En la presente investigación, el uso del aprendizaje automático posibilitó identificar la técnica mejor para predecir el embarazo en la adolescencia. El bosque aleatorio presentó métricas óptimas para la predicción, la cual fue excelente. Rawat y cols.,<sup>(27)</sup> a partir de datos de seguros médicos y de salud para la presentación y aceptación de reclamos, también identificaron esta técnica como la mejor para sus propósitos.

No obstante, este resultado diverge de los hallazgos de otras investigaciones. Oermann y cols.<sup>(28)</sup> identificaron el modelo de regresión logística como el más eficaz, con 74 % de precisión para predecir

la mortalidad de pacientes con malformaciones arteriovenosas cerebrales, intervenidos quirúrgicamente mediante radioterapia. En el estudio de Asadi y cols.<sup>(29)</sup> las redes neuronales fueron el modelo más eficaz ( $97,5\% \pm 2,5\%$ ). Raj y cols.<sup>(30)</sup> también utilizaron este último ( $AUC=0,90$ ) para la predicción del matrimonio infantil en la India. Es importante puntualizar que en la presente investigación esta técnica no se tuvo en cuenta.

Este estudio tuvo limitaciones. La escasez de fuentes teóricas nacionales, relacionadas con el empleo de las técnicas de aprendizaje automático para investigar el embarazo en la adolescencia, fue un obstáculo para contrastar los resultados de forma más amplia. La muestra se seleccionó a partir de la información aportada por el Instituto Nacional de Estadística e Informática con su Encuesta Demográfica y de Salud Familiar, la cual desarrolla con la finalidad de calcular indicadores de desnutrición y mortalidad en la niñez; por tanto, no tiene en cuenta la variable dependiente embarazo en la adolescencia.

## CONCLUSIONES

Se evaluaron las técnicas de aprendizaje automático para determinar factores de riesgo del embarazo en adolescentes peruanas. El bosque aleatorio resultó la mejor por su precisión; además de la identificación de los factores en cuestión, se distinguieron los tres más importantes. Este estudio es un precedente valioso para la aplicación de las técnicas de aprendizaje automático en la predicción de diversas variables necesarias para mejorar la gestión pública. Más cuanto que en este campo se utilizan bases de datos grandes, como las desarrolladas por el Instituto Nacional de Estadística e Informática, accesibles al público en general.

## REFERENCIAS BIBLIOGRÁFICAS

1. Instituto Nacional de Estadística e Informática. Encuesta Demográfica y de Salud Familiar - ENDES 2021 Nacional y Departamental [Internet]. Lima: INEI; 2022 [citado 2 Sep 2024]. Disponible en: [https://www.inei.gob.pe/media/MenuRecursivo/publicaciones\\_digitales/Est/Lib1838/pdf/Libro.pdf](https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1838/pdf/Libro.pdf)
2. Buitrago-Ramírez F, Ciurana-Misol R, Fernández-Alonso MC, Tizón JL; Miembros del Grupo de Salud Mental. Prevención de los trastornos de la salud mental. Embarazo en la adolescencia. Aten

Primaria [Internet]. Oct 2022 [citado 2 Sep 2024];54 Supl 1: 102494. Disponible en: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9705218/pdf/main.pdf>

3. Hernández-Cordero AL, Gentile A, Santos-Díaz E. Perspectivas teóricas para el análisis de la maternidad adolescente. Barataria [Internet]. 2019 [citado 2 Sep 2024];26:135-54. Disponible en: <https://revistabarataria.es/web/index.php/rb/article/download/399/710/1546>

4. Ranjbar A, Jahromi MS, Boujarzadeh B, Roozbeh N, Mehrnoush V, Darsareh F. Pregnancy, childbirth and neonatal outcomes associated with adolescent pregnancy. Gynecol Obstetr Clin Med [Internet]. Jun 2023 [citado 2 Sep 2024];3(2):100-5. Disponible en: <https://www.sciencedirect.com/science/article/pii/S2667164623000131>

5. Azimirad A. Pregnancy in adolescence: It is time to get ready for generations Z and Alpha. Gynecol Obstetr Clin Med [Internet]. Jun 2023 [citado 2 Sep 2024];3(2):71-5. Disponible en: <https://www.sciencedirect.com/science/article/pii/S2667164623000374>

6. Eliner Y, Gulersen M, Kasar A, Lenchner E, Grünebaum A, Chervenak FA, et al. Maternal and neonatal complications in teen pregnancies: a comprehensive study of 661,062 patients. J Adolesc Health [Internet]. Jun 2022 [citado 2 Sep 2024];70(6):922-7. Disponible en: <https://pubmed.ncbi.nlm.nih.gov/35165030/>

7. Davis K, Blake J. Social structure and fertility: an analytic framework. Econ Dev Cult Change [Internet]. Abr 1956 [citado 2 Sep 2024];4(3):211-35. Disponible en: [https://u.demog.berkeley.edu/~jrw/Biblio/Eprints/%20D-F/davis.blake.1956\\_intermediate\\_variables.pdf](https://u.demog.berkeley.edu/~jrw/Biblio/Eprints/%20D-F/davis.blake.1956_intermediate_variables.pdf)

8. Bongaarts J. A framework for analyzing the proximate determinants of fertility. Popul Dev Rev. 1978;4(1):105-32.

9. Di Cesare M, Rodríguez-Vignoli J. Análisis micro de los determinantes de la fecundidad adolescente en Brasil y Colombia. Pap. poblac [Internet]. Jun 2006 [citado 2 Sep 2024];12(48):107-40. Disponible en: <https://www.redalyc.org/pdf/112/11204806.pdf>

10. Hernández-Sampieri R, Mendoza-Torres CP. Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta. 2da ed. Ciudad de México: Editorial Mc Graw Hill Education; 2023.

11. Huang B, Zhu Y, Usman M, Chen H. Semi-supervised learning with missing values imputation. J Knowledge-Based Sys [Internet]. Ene 2024 [citado 2 Sep 2024];284:111171. Disponible en: <https://arxiv.org/pdf/2106.01708>

12. Shaon SH, Karim T, Shakil S, Hasan Z. A comparative study of machine learning models with

LASSO and SHAP feature selection for breast cancer prediction. Healthcare Analytics [Internet]. Dic 2024 [citado 2 Sep 2024];6:100353. Disponible en: [https://www.researchgate.net/profile/Md-Shazzad-Hossain-Shaon/publication/381772917\\_A\\_comparative\\_study\\_of\\_machine\\_learning\\_models\\_with\\_LASSO\\_and\\_SHAP\\_feature\\_selection\\_for\\_breast\\_cancer\\_prediction/links/667e41caf3b61c4e2c94833f/A-Comparative-Study-of-Machine-Learning-Models-with-LASSO-and-SHAP-Feature-Selection-for-Breast-Cancer-Prediction.pdf](https://www.researchgate.net/profile/Md-Shazzad-Hossain-Shaon/publication/381772917_A_comparative_study_of_machine_learning_models_with_LASSO_and_SHAP_feature_selection_for_breast_cancer_prediction/links/667e41caf3b61c4e2c94833f/A-Comparative-Study-of-Machine-Learning-Models-with-LASSO-and-SHAP-Feature-Selection-for-Breast-Cancer-Prediction.pdf)

13. Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. Lancet Oncol [Internet]. May 2019 [citado 2 Sep 2024];20(5):e262-73. Disponible en: <https://www.sciencedirect.com/science/article/pii/S1470204519301494?via%3Dihub>

14. Parzinger M, Hanfstaengl L, Sigg F, Spindler U, Wellisch U, Wirnsberger M. Comparison of different training data sets from simulation and experimental measurement with artificial users for occupancy detection — Using machine learning methods Random Forest and LASSO. Build Environ [Internet]. Sep 2022 [citado 2 Sep 2024];223:109313. Disponible en: <https://www.sciencedirect.com/science/article/pii/S0360132322005352>

15. Eddie D, Prindle J, Somodi P, Gerstmann I, Dilkina B, Saba SK, et al. Exploring predictors of substance use disorder treatment engagement with machine learning: The impact of social determinants of health in the therapeutic landscape. J Subst Use Addic Treat [Internet]. Sep 2024 [citado 2 Sep 2024];164:209435. Disponible en: <https://www.sciencedirect.com/science/article/abs/pii/S2949875924001474>

16. Obaido G, Mienye ID, Egbelowo OF, Emmanuel ID, Ogunleye A, Ogbuokiri B, et al. Supervised machine learning in drug discovery and development: algorithms, applications, challenges, and prospects. Mach Learn Applic [Internet]. Sep 2024 [citado 2 Sep 2024];17:100576. Disponible en: <https://www.sciencedirect.com/science/article/pii/S2666827024000525>

17. Richardson E, Trevizani R, Greenbaum JA, Carter H, Nielsen M, Peters B. The receiver operating characteristic curve accurately assesses imbalanced datasets. Patterns [Internet]. 2024 [citado 4 Ene 2024];5(6):100994. Disponible en: <https://www.cell.com/action/showPdf?pii=S2666-3899%2824%2900109-0>

18. Asociación Médica Mundial. Declaración de Helsinki de la AMM. Principios éticos para las investigaciones médicas en seres humanos. Ratificada en la 64ª Asamblea General, Fortaleza, Brasil,

octubre 2013. Helsinki: 18ª Asamblea Mundial; 1964 [citado 4 Ene 2024]. Disponible en: [http://www.anmat.gov.ar/comunicados/HELSINSKI\\_2013.pdf](http://www.anmat.gov.ar/comunicados/HELSINSKI_2013.pdf)

19. Martínez-Pérez JA, Pérez-Martín PS. La curva ROC. SEMERGEN [Internet]. Feb 2023 [citado 4 Ene 2024];49(1):e101821. Disponible en: <https://static.elsevier.es/ficheros/7.pdf>

20. Quezada MA, Tobón-Rivera A, Castrillón-Gómez OD. Minería de datos: una aplicación para determinar cuáles factores socio-económicos influyen en el embarazo adolescente. Inform Tecnol [Internet]. 2020 [citado 4 Ene 2024];31(6):53-60. Disponible en: <https://scielo.conicyt.cl/pdf/infotec/v31n6/0718-0764-infotec-31-06-53.pdf>

21. Avelar-Jaime D, López-Ramírez M, Rivera-Romero CA, Guzmán-Cabrera R. Clasificación del Corpus BBC News Summary utilizando J48 en Weka. Jóvenes Cienc [Internet]. 2023 [citado 4 Ene 2024];25:[aprox. 6 p.]. Disponible en: <https://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/download/4212/3692/13737>

22. Rosales-López JY. Determinantes próximos de la fecundidad adolescente en Honduras periodo 2011-2012 [Internet]. Tegucigalpa: Universidad Nacional Autónoma de Honduras; 2019 [citado 4 Ene 2024]. Disponible en: <https://tzibalnaah.unah.edu.hn/bitstream/handle/123456789/11522/Determinantes%20pr%e3%b3ximos%20de%20la%20fecundidad%20adolescente%20en%20Honduras%20periodo%202011-2012.pdf?sequence=2&isAllowed=y>

23. Fasula AM, Chia V, Murray CC, Brittain A, Tevendale H, Koumans EH. Socioecological risk factors associated with teen pregnancy or birth for young men: a scoping review. J Adolesc [Internet]. Jul 2019 [citado 4 Ene 2024];74(1):130-45. Disponible en: <https://www.sciencedirect.com/science/article/abs/pii/S014019711930096X?via%3Dihub>

24. Garza-Reyna D, Cruz-Villareal M, Alanís-Cruz A, Flores-Acosta CC, Ramírez-Colunga C, Soria-López J, et al. 120. Sociodemographic and psychosocial factors associated with adolescent pregnancy. J. Pediatr. Adolesc. Gynecol. [Internet]. Abr 2024 [citado 4 Ene 2024];37(2):297-98. Disponible en: <https://www.sciencedirect.com/science/article/abs/pii/S1083318824001426?via%3Dihub>

25. Asare BYA, Baafi D, Dwumfour-Asare B, Adam AR. Factors associated with adolescent pregnancy in the Sunyani Municipality of Ghana. Int. J. Afr. Nurs. Sci. [Internet]. 2019 [citado 4 Ene 2024];10:87-91. Disponible en: <https://www.sciencedirect.com/science/article/pii/S2214139118300817>

26. D'Añari-Cabrera JR. Factores biosociodemográficos asociados al embarazo precoz en adolescentes gestantes atendidas en el HRHDE, abril–mayo 2019 [Internet]. Arequipa: Universidad Nacional de San Agustín de Arequipa; 2019 [citado 4 Ene 2024]. Disponible en: <https://repositorio.unsa.edu.pe/bitstreams/c66a10e4-9339-419d-830a-72810f74cde9/download>
27. Rawat S, Rawat A, Kumar D, Sabitha A. Application of machine learning and data visualization techniques for decision support in the insurance sector. Int. J. Inf. Manag. Data Insights [Internet]. Nov 2021 [citado 4 Ene 2024];1(2):100012. Disponible en: <https://www.sciencedirect.com/science/article/pii/S2667096821000057>
28. Oermann EK, Rubinsteyn A, Ding D, Mascitelli J, Starke RM, Bederson JB, et al. Using a machine learning approach to predict outcomes after radiosurgery for central arteriovenous malformations. Scientific Reports [Internet]. 2016 [citado 4 Ene 2024];6:21161. Disponible en: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4746661/pdf/srep21161.pdf>
29. Asadi H, Kok HK, Looby S, Brennan P, O'Hare A, Thornton J. Outcomes and complications following endovascular treatment of brain arteriovenous malformations-a prognostication attempt using artificial intelligence. World Neurosurg [Internet]. Dic 2016 [citado 4 Ene 2024];96:562-9. Disponible en: <https://www.sciencedirect.com/science/article/abs/pii/S1878875016309160?via%3Dihub>
30. Raj A, Dehingia N, Singh A, McDougal L, McAuley J. Application of machine learning to understand child marriage in India. 2020. SSM Popul Health [Internet]. 2020 [citado 4 Ene 2024];12:100687. Disponible en: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7732880/pdf/main.pdf>

### Conflictos de intereses

Los autores declaran que no existen conflictos de intereses.

### Contribuciones de los autores

Bernardo Céspedes-Panduro: conceptualización, curación de datos, metodología, administración del proyecto, *software*, validación, visualización, y redacción del borrador original.

Zoraida Judith Huamán-Gutiérrez: análisis formal, adquisición de fondos, investigación, recursos, supervisión, redacción, revisión, y edición.



---

### **Financiación**

Universidad Nacional Mayor de San Marcos, Lima, Perú.

