“Predicción trimestral de la tasa de natalidad en el Perú con modelos de aprendizaje automático”

Autores: John Arzapalo, Ariana Burga, Ricardo Lara, Luis Miranda, Alessandro Santé y Juan Zavala.

Resumen – La tasa de natalidad es uno de los principales indicadores que permiten describir la transición demográfica. Durante el último siglo, ha mostrado una tendencia a la baja a nivel global, la cual se replica en Perú. Esta reducción requiere ser analizada para la correcta toma de decisiones y planificación de los distintos niveles de gobierno en los sectores salud, educación y pensionario. Ante esta necesidad, este trabajo plantea la comparación de tres modelos predictivos basados en machine learning para la predicción trimestral de la tasa de natalidad a nivel nacional y departamental en el Perú. Para ello se usó información del Registro de Nacidos Vivos (2015-2025), buscando obtener un modelo con alto desempeño y los factores sociodemográficos más influyentes en la natalidad peruana.

1. Introducción

La transición demográfica que experimenta el Perú se manifiesta en cambios significativos en los patrones de natalidad. A nivel global, este indicador ha mostrado un declive consistente durante el último siglo (The Economist 2023), tendencia que se replica en el contexto peruano con particularidades regionales que demandan un análisis específico a nivel departamental. Esta reducción progresiva genera desafíos asociados al envejecimiento poblacional, donde una base reducida de población joven debe sustentar económicamente a un creciente número de adultos mayores.

La comprensión de la evolución de la tasa de natalidad resulta fundamental para el diseño de políticas públicas descentralizadas en salud, educación y pensiones (Castellares, Camacho y Huarancca 2024: 65). No obstante, los métodos tradicionales de proyección demográfica presentan limitaciones al no capturar la complejidad de las relaciones entre variables sociodemográficas a nivel subnacional.

Frente a estas limitaciones, este estudio propone la implementación comparativa de tres modelos de machine learning para la predicción trimestral a nivel nacional y departamental: Regresión lineal, Regresión KNN y Árboles de regresión. La hipótesis central plantea que entre estos tres enfoques existirá al menos un modelo demostrará desempeño predictivo superior a nivel nacional y departamental, identificando los factores sociodemográficos con mayor poder predictivo para cada contexto geográfico.

La pregunta de investigación que guía este trabajo se centra en determinar:

- ¿Qué modelo de machine learning entre regresión lineal, regresión KNN y árboles de regresión provee las predicciones más precisas de la tasa de natalidad trimestral en Perú a nivel nacional y departamental, y cuáles son las variables que presentan mayor influencia predictiva en cada ámbito territorial?

Para responder esta interrogante, el objetivo general consiste en:

- Desarrollar y comparar un sistema predictivo de la tasa de natalidad trimestral utilizando los tres modelos con datos del Registro de Nacidos Vivos (2015-2025) para los ámbitos ~~nacional y~~ departamental.

Los objetivos específicos incluyen:

- El procesamiento y transformación de los datos del registro con desagregación departamental.
- La implementación y entrenamiento de los tres modelos para ambos niveles geográficos.
- La evaluación comparativa de su desempeño predictivo mediante métricas estandarizadas.
- Evaluación de las variables con mayor influencia en la capacidad predictiva de los modelos a nivel departamental.
- La validación de la capacidad predictiva con los datos más recientes disponibles, contribuyendo así a la mejora de los instrumentos de planificación demográfica nacional y regional.

Comentado [LM2]: Se iria esto

Comentado [LM3]: “La identificación de las variables con mayor poder predictivo en el modelo de mejor desempeño para cada ámbito” .
Lo cambie pues el jp indico que esta parte de la tarea no está directamente vinculada al objetivo general y sugirió que podría eliminarse o reestructurarse para alinearse mejor con el propósito principal del trabajo, evaluar modelos predictivos

Comentado [LM1]: Seria cambiarlo por solo el contexto departamental?

2. Trabajos relacionados

Predicting Future Birth Rates with the Use of an Adaptive Machine Learning Algorithm: A Forecasting Experiment for Scotland
La investigación compara modelos de regresión, entre ellos Regresión Lineal y árboles de decisión, para pronosticar nacimientos mensuales en Escocia. Los resultados evidencian que los métodos basados en árboles superan a los lineales, demostrando su mayor capacidad para representar relaciones no lineales en las tasas de natalidad. (Tzitiridou-Chatzopoulou, Zournatzidou y Kourakos 2024: 4–5; 7-9).

Análisis de reducción de tasas de natalidad en Colombia: un enfoque con técnicas de aprendizaje automático
El estudio utiliza datos del DANE (2019–2022) para predecir la natalidad mediante Árbol de Decisión, Random Forest y XGBoost. Se resalta que el Árbol de Decisión identifica factores determinantes como la educación materna y el año, confirmando su utilidad para explicar patrones demográficos y contrastarlo con enfoques lineales tradicionales. (Largo y Valencia, 2024: 13–17).

Predicción de factores clave en el aumento de la demografía en Colombia a través del ensamble de modelos de Machine Learning
Este trabajo aplica Regresión Lineal, KNN y Árboles de Decisión para analizar la natalidad en Colombia, integrándolos en un modelo de ensamble tipo Bagging. Los hallazgos revelan que el Árbol de Decisión alcanza el mejor rendimiento, validando su eficacia frente a métodos lineales para describir la relación entre variables sociodemográficas y nacimientos (Ordóñez-Erazo, Ordóñez y Bucheli-Guerrero 2022: 285).

Estos tres estudios brindan un marco comparativo útil para nuestro trabajo, al aplicar modelos como Regresión Lineal, KNN, Árboles de Decisión, Random Forest y XGBoost en la predicción de tasas de natalidad y el análisis de factores sociodemográficos. A partir de ellos, nuestro estudio se centra en Regresión Lineal, KNN y Árboles de Decisión, adaptados al contexto peruano y orientados a una predicción trimestral que refleje con mayor detalle la variabilidad temporal del fenómeno.

3. Metodología

Enfoque(s) propuesto:
Este estudio busca predecir la tasa de natalidad trimestral en Perú mediante algoritmos basados en machine learning. Formalmente, sea **X** el conjunto de variables demográficas y temporales, y **Y** la tasa de natalidad en un trimestre determinado para un departamento específico. El problema consiste en encontrar una función **f: X → Y** que aproxime de manera óptima la relación entre las variables de entrada y la natalidad trimestral. Para ello, se evaluarán distintos algoritmos de machine learning, comúnmente aplicados en problemas de predicción temporal para identificar tendencias y patrones recurrentes.

1. Recopilación de los Dataset/s
A través de la Plataforma Nacional de Datos Abiertos del Perú y el Geodir, se obtuvieron dos datasets para este estudio. El primero contiene información individual de cada nacimiento registrado, incluyendo características temporales, demográficas y geográficas. El segundo proporciona el catálogo oficial de ubicaciones geográficas del país, permitiendo la vinculación espacial y agregación departamental de los datos.

- **Registros de Nacidos Vivos en el Perú (2015–2025):**
  Dataset oficial que contiene registros individuales de nacimientos con variables cualitativas y cuantitativas clasificadas por demografía, salud y geografía. Incluye el Código de Ubicación Geográfica (Ubigeo) que permite vincular cada nacimiento con su ubicación específica a nivel de distrito, provincia y departamento.

Tabla 1
Variables del Dataset Datos de Registros de Nacidos Vivos

#	Variable	#	Variable
1	Año del nacimiento	12	Número de embarazos anteriores
2	Mes del nacimiento	13	Número de hijos vivos
3	Peso al nacer	14	Número de hijos fallecidos
4	Tamaño al nacer	15	Número de abortos o nonatos
5	Duración gestacional	16	País de origen de la madre
6	Condición del parto	17	Código de ubicación geográfica (Ubigeo)
7	Sexo del recién nacido	18	Ocupación de la madre
8	Modalidad del parto	19	Código del establecimiento de salud certificador

9	Edad de la madre	20	Lugar de nacimiento
10	Estado civil de la madre	21	Tipo de atención en el parto
11	Nivel educativo de la madre	22	Fuente de financiamiento del parto

Nota. Tomado de Plataforma Nacional de Datos Abiertos: https://datosabiertos.gob.pe/dataset/registros-de-nacidos-vivos-en-el-per%C3%BA-2015%E2%80%932025

- **Catálogo de Ubigeo del INEI (2019):** Listado oficial del Instituto Nacional de Estadística e Informática que detalla cada unidad geográfica del país (ubigeo, departamento, provincia y distrito), así como variables demográficas y geográficas asociadas. Este catálogo sirve como clave de vinculación para la georreferenciación de los nacimientos.

Tabla 2

Variables del Dataset Ubigeo de Perú (INEI)

#	Variable	#	Variable
1	Código de ubicación geográfica (Ubigeo)	5	Población
2	Distrito	6	Superficie
3	Provincia	7	Y
4	Departamento	8	X

Nota. Tomado de Ubigeo de Perú (INEI): https://account.geodir.co/recursos/ubigeo-inei-peru.html

Con esta información y tras análisis podemos señalar que los tres modelos implementados (Regresión Lineal, k-NN y Árboles de Regresión) comparten la misma estructura de entrada y salida, diseñada específicamente para capturar patrones espacio-temporales de natalidad.

Entrada (X):
Los modelos reciben como entrada un conjunto reducido de variables esenciales para la predicción trimestral departamental:

Variables temporales:
- Año del nacimiento (Variable #1, Tabla 1): Permite capturar tendencias de largo plazo
- Mes del nacimiento (Variable #2, Tabla 1): Agregado trimestralmente (Q1: ene-mar, Q2: abr-jun, Q3: jul-sep, Q4: oct-dic)

Con respecto a la dimensión temporal, permiten dividir el dataset siguiendo una secuencia regida por orden cronológico.

Variable de georreferenciación:
- Código de ubicación geográfica/Ubigeo (Variable #17, Tabla 1): Vincula cada nacimiento con su ubicación espacial
- Departamento (Variable #4, Tabla 2): Obtenido al relacionar el Ubigeo de cada nacimiento con el catálogo del INEI, extrayendo los primeros 2 dígitos del código que identifican el departamento
- Población departamental (Variable #5, Tabla 2): Necesaria para estandarizar el número de nacimientos y calcular la tasa por cada 1000 habitantes

El análisis se limita a los departamentos con más de 3 000 registros y menor presencia de valores atípicos, garantizando consistencia estadística.
Esta selección garantiza que el modelo capture de forma integrada las dimensiones temporal (cuándo ocurren los nacimientos), espacial (dónde ocurren, a nivel departamental) y proporcional (en qué magnitud relativa, normalizada por población).

Salida (Y):
El objetivo de los modelos es predecir un único valor numérico continuo: la tasa de natalidad trimestral por departamento. Esta tasa se calcula de la siguiente manera, expresada por cada 1,000 habitantes:
Y = (Total de nacimientos en el departamento D durante el trimestre T / Población del departamento D) × 1000

2. Preprocesamiento y estructuración de Datos
El proceso de preparación de datos fue diseñado para transformar los registros individuales de nacimientos en un formato agregado, adecuado para el modelado de series temporales. El flujo de trabajo consistió en los siguientes pasos clave:

- Limpieza y Validación de Datos Críticos: El primer paso fue una rigurosa limpieza de los datos. Se eliminaron todos los registros que presentaban valores inválidos o faltantes en las variables esenciales para el análisis: año y mes de nacimiento, y el código geográfico (Ubigeo). Esta validación fue fundamental, ya que cualquier registro sin una fecha y ubicación precisas no puede ser utilizado en la agregación espacio-temporal.
- Vinculación y Enriquecimiento: Una vez depurados los datos, se utilizó el código Ubigeo para vincular cada nacimiento con el catálogo geográfico del INEI. Este cruce permitió enriquecer cada registro con dos piezas de información clave: el nombre del departamento y la población total de esa región.
- Agregación Espacio-Temporal: Los datos individuales se agruparon por año, departamento y un nuevo campo de trimestre (derivado del mes de nacimiento). Para cada uno de estos grupos, se contabilizó el número total de nacimientos.
- Cálculo de la Variable Objetivo (Y): La variable a predecir, la tasa de natalidad trimestral, se calculó para cada grupo utilizando la fórmula:

Comentado [LM5]: Añadido, lo que no escucho bien del audio es si 3mil o 2mil, la transcripción lo pone en 3mil

Comentado [LM6]: Lo cambie. Originalmente estaba como:
garantiza que el modelo pueda responder a las preguntas fundamentales de cuándo (dimensión temporal), dónde (dimensión espacial) y en qué proporción (normalización por población) ocurren los nacimientos

Comentado [LM4]: Añadi esto para reincidir en que tenemos en cuenta la división temporal según secuencia (años)

Tasa = (Total de Nacimientos / Población) × 1000. Esta estandarización es crucial para poder comparar la natalidad entre departamentos de diferentes tamaños.

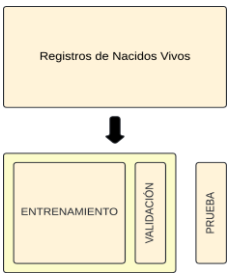
El *dataset* final quedó estructurado con las variables de entrada (X) siendo el año, el trimestre y el departamento, y la variable objetivo (Y) siendo la tasa de natalidad.

El conjunto de datos se dividirá siguiendo la metodología Train-Test Split, asignando un 70% de los datos al conjunto de desarrollo (dev) y un 30% al conjunto de prueba. A su vez, el conjunto de desarrollo se subdividirá en entrenamiento (80%) y validación (20%), permitiendo ajustar y evaluar el modelo antes de su prueba final.

Mi propuesta: Dada la naturaleza secuencial de la información, en este enfoque los datos históricos hasta el presente (2010-2018) serán empleados para entrenar a los modelos, mientras que los más recientes (2020-2025) se emplearán para validar el desempeño de estos. Con eso conseguimos evitar que el modelo entrene con datos futuros, fundamental en problemas de series temporales ya que se debe considerar la dependencia temporal de la información.

El dataset será dividido en conjunto de entrenamiento (2010-2018), validación (2019) y prueba (2020-2025).

Figura 1
División de los Registros de Nacidos Vivos

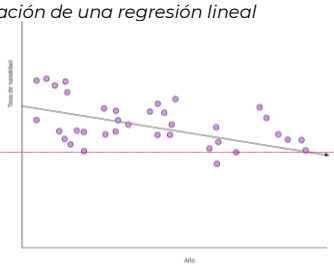


Nota. Elaboración propia: https://drive.google.com/file/d/14si5YAws_pFUK8e78JTK9TuSFQmyF-EJ/view?usp=sharing

3. Modelos y entrenamiento
Para este trabajo se desarrollaron tres modelos predictivos:

Regresión Lineal: Se utilizó como modelo inicial de referencia debido a su capacidad para estimar relaciones directas y proporcionales entre la tasa de natalidad y las variables explicativas. Estudios de modelado demográfico han empleado regresión lineal precisamente por su simplicidad y transparencia en la interpretación de coeficientes estadísticos.

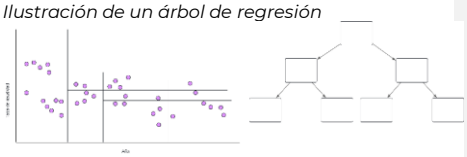
Figura 2
Ilustración de una regresión lineal



Nota. Elaboración propia: https://drive.google.com/file/d/1f5zp-0qv9m1Pgk1BHcNtubw3a-VfKSw5/view?usp=sharing

Árbol de regresión: A diferencia de la regresión lineal, este enfoque puede identificar umbrales críticos. Esta característica lo convierte en un modelo particularmente útil para la detección temprana de cambios estructurales y la planificación diferenciada por departamentos. Esta ventaja de interpretabilidad y detección local ha sido documentada en estudios aplicados en contexto demográficos.

Figura 3
Ilustración de un árbol de regresión



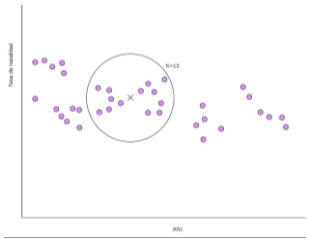
Nota. Elaboración propia: https://drive.google.com/file/d/1tXOWp0L3zpc46VkJeubVonSpbF0tIjqt/view?usp=sharing

Regresión k-NN: Este modelo se incorporó para capturar patrones de similitud temporal en la natalidad. A través del análisis de los trimestres y departamentos, el algoritmo estima la tasa esperada de nuevos periodos según la proximidad a contextos previos. Sus capacidades lo hacen una herramienta adecuada para identificar picos o repeticiones estacionales.

Figura 4
Ilustración de una regresión k-NN

Comentado [LM7]: Verificar los rangos, creo que en la parte de entrenamiento seria mucho si considero de 2010 a 2018 verdad? Perdon estoy un poco perdido con eso

Comentado [LM8]: Tenemos que modificar esto siguiendo las pautas:
- establecer un criterio de división temporal del dataset donde el entrenamiento y la validación estén basados en años anteriores, y la prueba en años más recientes.
- mantener la consistencia temporal en cada ubigeo para evitar fugas de información.



Nota. Elaboración propia: https://drive.google.com/file/d/1Ggd_LdN-45H4yhmKs9ctgY_gB9WFOEir/view?usp=sharing

La comparación entre estos modelos permite evaluar distintos enfoques de predicción, lineal, jerárquico y por similitud, y determinar cuál se ajusta mejor a la naturaleza de los datos demográficos.

Como resultado esperado en la salida, tenemos una predicción de la tasa de natalidad, ya sea a nivel nacional o departamental, lo cual proporciona una herramienta para la planificación.

4. Experimentación y Resultados

- Setup experimental:
  - Describir datos usados (o método para obtenerlos) (si aplica).
  - Describir las métricas de evaluación.
  - Describir los experimentos hechos (qué componentes/parámetros/escenarios se probaron, qué valores, qué estrategia de validación).

  Los experimentos deben ser planificados para poder caracterizar/comparar los enfoques desarrollados. Aquí algunas preguntas que deben responder los experimentos:
  - ¿El enfoque desarrollado resuelve siempre el problema?
  - ¿Qué tan eficientemente lo resuelven?
  - ¿Cuál es el desempeño comparado con otros modelos o técnicas de referencia?
  - ¿Cómo influyen los parámetros del enfoque en su desempeño?

- Resultados y Discusión:
  - Presentar resultados numéricos generados en los experimentos. Hacer un análisis de dichos resultados.

  La presentación de los resultados debe facilitar el entendimiento. En general se deben usar figuras y/o tablas. Recuerde, todos los resultados deben interpretarse. Esforzarse para explicar el formato de las curvas presentadas, dar detalles del tiempo de simulación.

5. Conclusión

Dar las conclusiones principales con base en los resultados obtenidos y a lo que fue planteado en su hipótesis, ¿qué se puede decir del o los enfoques desarrollados y/o del problema abordado?

6. Sugerencias de trabajos futuros

Indicar, por ejemplo: ¿qué cosas se pueden mejorar del enfoque?, ¿qué otros posibles problemas podrían abordarse con el enfoque?

7. Implicancias éticas

Indicar qué implicancias éticas podría generar el trabajo desarrollado de ser escalado (sesgos, posible afectación a la seguridad/privacidad de usuarios, posibilidad de ataques y robo de datos, etc). Sugiera formas de abordar dichos problemas

8. Link del repositorio del trabajo

Puede ser Github, Gitlab, u otro. Dar las credenciales para poder tener acceso.

Comentado [LM9]: Deberiamos actualizar esto? Creo que vole entonces ahi lo dejo nomas lo diiscutimos en conjunto

Comentado [LM10]: Solo departamental

9. Declaración de contribución de cada integrante

Describir los aportes de cada integrante al proyecto.

10. Referencias

[1]. CASTELLARES, Renzo; Diego CAMACHO y Mario HUARANCCA 2024 "Reducción de la tasa de natalidad en el Perú: proyecciones y determinantes". *Moneda*. Lima, número 197, pp. 64-69. Consulta: 29 de setiembre de 2025. https://www.bcrp.gob.pe/docs/Publicaciones/Revista-Moneda/moneda-197/moneda-197-10.pdf

[2]. LARGO, Lida y Sebastián VALENCIA 2024 "Análisis de reducción de tasas de natalidad en Colombia: un enfoque con técnicas de aprendizaje automático". *Universidad de Antioquia*. Medellín. Consulta: 29 de setiembre de 2025. https://hdl.handle.net/10495/44526

[3]. ORDÓÑEZ-ERAZO, Hugo-Armando; Camilo ORDÓÑEZ y Víctor-Andrés BUCHELI-GUERRERO 2022 "Predicción de factores clave en el aumento de la demografía en Colombia a través del ensamble

de modelos de Machine Learning". *Revista científica*. Bogotá, número 44, pp. 282-295. Consulta: 29 de setiembre de 2025. https://doi.org/10.14483/23448350.19205

[4]. THE ECONOMIST
2023 "Global fertility has collapsed, with profound economic consequences". *The Economist*. London, pp. s. p. Consulta: 29 de setiembre de 2025. https://www.economist.com/leaders/2023/06/01/global-fertility-has-collapsed-with-profound-economic-consequences

[5]. TZITIRIDOU-CHATZOPOULOU, Maria; Georgia ZOURNATZIDOU y Michael KOURAKOS
2024 "Predicting Future Birth Rates with the Use of an Adaptive Machine Learning Algorithm: A Forecasting Experiment for Scotland". *International Journal of Environmental Research and Public Health*. Basilea, volumen 21, número 7, pp. 1-11. Consulta: 29 de setiembre de 2025. https://doi.org/10.3390/ijerph21070841

Nota importante:

Figuras y tablas

- Cada figura debe ser leída y comprendida sin necesidad de recurrir a la descripción a lo largo del texto.
- Todos los símbolos usados deben ser explicados en la leyenda.