Análisis de reducción de tasas de natalidad en Colombia: un enfoque con técnicas de aprendizaje automático

Lida Vanessa Largo Quintero
Sebastián Valencia Cadena

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
María Bernarda Salazar Sánchez, Doctora (PhD) en Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(Largo Quintero & Valencia Cadena, 2024)
Referencia **Estilo APA 7 (2020)**	Largo Quintero L.V. & Valencia Cadena S. (2024). *Análisis de reducción de tasas de natalidad en Colombia: un enfoque con técnicas de aprendizaje automático*. [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VII.

Grupo de Investigación Intelligent Information Systems Lab – In2Lab.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

Dedicamos este trabajo de grado a nuestros seres queridos, quienes nos han brindado su amor, apoyo y paciencia a lo largo de este arduo camino. A nuestros padres y amigos, gracias por creer en nosotros y motivarnos a alcanzar nuestros sueños. Este logro es también suyo. Con cariño, Lida Vanessa Largo Quintero y Sebastián Valencia Cadena.

Agradecimientos

Agradecemos profundamente a nuestra asesora, la profesora María Bernarda Salazar Sánchez, por su invaluable guía, dedicación y apoyo constante durante el desarrollo de esta investigación. Su experiencia, orientación académica y compromiso fueron fundamentales para la culminación exitosa de este trabajo. Sus conocimientos y retroalimentación fueron determinantes para enriquecer la calidad científica de nuestra investigación.

Asimismo, expresamos nuestra gratitud a la Universidad de Antioquia por brindarnos los espacios y recursos necesarios para llevar a cabo este proyecto de investigación.

.

Tabla de contenido

Lista de Figuras

Lista de Tablas

Análisis de reducción de tasas de natalidad en Colombia: un enfoque con técnicas de aprendizaje automático

***Resumen*— Actualmente, la tasa de natalidad y mortalidad a nivel global muestra un alto dinamismo con una tendencia marcada hacia la disminución, lo que refleja cambios demográficos significativos, concentrados en diversas regiones [1]. En este contexto y desde una perspectiva nacional, las pirámides poblacionales en Colombia ya evidencian modificaciones notables en su estructura, mostrando una mayor concentración de población joven para el año 2023 y una proyección de aumento en la proporción de habitantes en edades avanzadas en los próximos años [2]. Esta evolución, es una clara consecuencia de la disminución en las tasas de natalidad y mortalidad en el país.**

En este sentido y de acuerdo con el último boletín técnico reportado por el DANE, sobre estadísticas vitales en Colombia se ha evidenciado una marcada disminución en la cantidad de nacimientos de los años 2019 al 2023, lo cual significa en cifras un decrecimiento porcentual del 9,6% del año 2022 al 2023 y del 15,2% del 2023 a lo corrido del 2024 [3]. Este declive, genera gran expectativa sobre el comportamiento de esta variable a futuro y el estudio de sus implicaciones económicas y sociales.

Este proyecto se enfoca en analizar la reducción de la tasa de natalidad en Colombia, a partir de los microdatos estadísticos entregados por el DANE para los años 2019 a 2022 [4], utilizando técnicas de aprendizaje automático. Incluyendo dentro del ejercicio, las características demográficas asociadas a la población muestreada y sus efectos políticos, sociales, económicos o culturales, con base en la evidencia científica y literaria disponible en las bases de datos actuales.

Se realizó una predicción de la tasa de natalidad utilizando los modelos de machine learning Random Forest, Árbol de Decisión y XGBoost, encontrando resultados muy positivos en cuanto al ajuste de los datos predichos en comparación con las muestras utilizadas para entrenar los modelos. Se identificó al modelo XGBoost como el que presentó las mejores métricas de desempeño, evidenciando una mayor precisión en la predicción de la tasa de natalidad y permitiendo analizar su relación con las características demográficas asociadas. Se espera que los resultados obtenidos permitan identificar tendencias futuras y ofrezcan insumos para la formulación de políticas públicas, con el fin de mitigar posibles impactos negativos en el comportamiento de esta variable.

***Palabras claves* —** Tasa de natalidad, demografía, Colombia, analítica, datos, predicción, clasificación, XG Boost, Random Forest, Árbol de decisión.

I. Introducción

Las poblaciones están sujetas a cambios continuos, estos pueden ser de estado (tamaño, sexo, edad, ubicación geográfica, densidad poblacional, etc) o por movimientos (tasas de: natalidad, mortalidad, crecimiento, fecundidad, entre otros). Dichos cambios, se generan por procesos de entrada (nacimientos, inmigraciones) y salida (muertes, emigraciones) de los individuos, influenciados históricamente por las guerras, la ubicación geográfica, la cultura, el acceso a los recursos básicos, políticas de gobierno, economía, entre otros [5]. En este sentido, analizar las dinámicas poblacionales desde la demografía es fundamental porque permite entender la estructura, dinámica y evolución de las comunidades humanas, lo cual es clave para la planificación y la sostenibilidad de las mismas, en el tiempo.

En el panorama global actual se evidencia una tendencia marcada hacia la disminución de la natalidad y una mayor concentración de la población en edades más longevas. Según el Fondo de Población de las Naciones Unidas (UNFPA) en la actualidad se calcula que dos terceras partes de la población del planeta residen en países o zonas donde las tasas de natalidad son bajas y donde la fecundidad no llega al umbral de reemplazo, este indicador refleja el nivel de fecundidad necesario para que una generación sustituya a otra en la reproducción de las poblaciones humanas a largo plazo, garantizando la conservación de la misma en el tiempo. Para lograrlo, cada mujer en edad fértil debe tener, en promedio, 2,1 hijos a lo largo de su vida [6]. Adicionalmente, un estudio de la universidad de Washington expone que la Tasa de Fertilidad Total (TFT) mundial ha caído hasta menos de la mitad en los últimos 70 años, desde aproximadamente cinco hijos por cada mujer en 1950 hasta 2.2 hijos en 2021, con más de la mitad de todos los países y territorios (110 de 204) por debajo del nivel de reemplazo poblacional [1].

De acuerdo con el estudio realizado por la Web World Expectancy, la pirámide poblacional global muestra una

marcada tendencia a la disminución de los nuevos nacimientos a menos del 5% y una concentración de la población en edades entre los 25 y los 34 años, para el año 2020. Con esta tendencia para el año 2050 la natalidad en el mundo será inferior al 2.5% y la población se concentra en edades jóvenes y adultas entre los 15 y 44 años [7].

En el caso de Colombia, se evidencia una leve disminución en el porcentaje de nacimientos entre el año 2015 al 2020 con un valor cercano a los 3.5% y una concentración de la población entre los 15 y los 29 años. Proyectándose al 2050 con una tasa de natalidad inferior al 2.5% y una concentración de la población entre los 45 y 49 años. Esto indica un notable envejecimiento de la población, seguido por un desequilibrio en el umbral de reemplazo [6].

En este contexto, la tasa de natalidad se presenta como un indicador demográfico esencial para comprender los cambios poblacionales relacionados con los movimientos, generando repercusiones significativas en el desarrollo social y económico de un país.

En Colombia, este indicador ha experimentado una notable disminución en las últimas décadas, según el Departamento Administrativo Nacional de Estadística [3], la tasa de fecundidad ha bajado de 3.1 hijos por mujer en 2010 a aproximadamente 2.3 en 2022, lo que plantea interrogantes sobre las causas subyacentes de esta reducción.

Entender las razones detrás de esta disminución es fundamental para las políticas públicas relacionadas con la salud reproductiva, la educación y el bienestar social. Investigaciones previas sugieren que factores como el acceso a métodos anticonceptivos, la educación de la mujer [8], las condiciones económicas [9] y los cambios en las estructuras familiares [10], se han propuesto como influencias clave. Sin embargo, la complejidad de estos factores y su interacción hace que su análisis sea un reto.

Con el avance de la tecnología y el crecimiento del campo de la analítica de datos, las técnicas de aprendizaje automático se presentan como herramientas valiosas para examinar y modelar estos fenómenos [11]. Esta investigación se propone realizar una predicción de la tasa de natalidad en Colombia utilizando métodos de machine learning, con el objetivo de identificar las variables que más influyen en su variabilidad y entender cómo estas podrían incidir en su posible disminución en el futuro.

El presente estudio tiene como objetivo analizar la tasa de natalidad en Colombia en términos de las características demográficas de la población muestreada a través de la aplicación de técnicas avanzadas de analítica de datos y aprendizaje automático. Se espera que los resultados de esta investigación no solo contribuyan al entendimiento académico del fenómeno, sino que también sirvan como base para la formulación de políticas que aborden el dinamismo poblacional del país, con una mayor efectividad.

II. METODOLOGÍA

La metodología propuesta para el análisis exploratorio de datos (EDA) encuentra su fundamento teórico en la obra de John W. Tukey, *Exploratory Data Analysis* (1977) [12], en la cual se establece la importancia de examinar los datos de manera exhaustiva antes de aplicar modelos estadísticos formales. Tukey enfatiza el uso de herramientas gráficas y técnicas descriptivas para identificar patrones, relaciones y posibles inconsistencias en los datos, lo que permite una mejor comprensión y depuración del conjunto inicial. Siguiendo este enfoque, el EDA no solo facilita la reducción de dimensiones y la eliminación de variables redundantes o irrelevantes, sino que también asegura que los datos utilizados se alineen con los objetivos analíticos planteados, optimizando así la precisión y relevancia de los resultados. Esta base conceptual sustenta la estructura metodológica descrita a continuación y guía cada etapa del proceso de análisis (ver
Figura 1.)

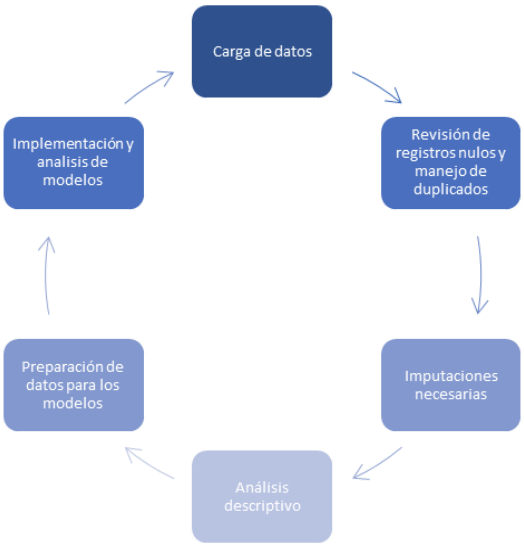

Figura 1. Metodología para el procesamiento de datos.

A. Descripción de la base de datos

El conjunto de datos original se extrae de la plataforma pública de datos abiertos de Colombia, específicamente del apartado de estadísticas de nacimientos publicada por el DANE [3]. Este conjunto de datos está organizado anualmente y contiene diversas variables relevantes para el

análisis de la tasa de natalidad en el país para los años 2019, 2020, 2021 y 2022 (ver Tabla 1).

A continuación, se ilustran las principales variables contenidas en el conjunto de datos (ver Tabla 2), cada una está asociada a una categoría, que a su vez está codificada numéricamente para reducir la robustez del Dataset y facilitar el análisis. Está compuesto por 40 variables, las cuales, tras realizar un análisis exploratorio de datos, se redujo a 27. Esta reducción se obtiene al eliminar aquellas variables que no se alineaban al enfoque de analizar la natalidad en Colombia o cuya información ya estaba contenida/representada en otra variable (ver Tabla 3).

Tabla 1. Número de registros por año

AÑO	TAMAÑO	DIMENSIÓN
2019	62,207 KB	642,661 filas x 39 columnas
2020	59,739 KB	629,403 filas x 39 columnas
2021	97,122 KB	616,925 filas x 39 columnas
2022	90,318 KB	573,626 filas x 39 columnas
TOTAL	310,386 KB	2,462,601 filas x 39 columnas.

Tabla 2. Descripción de las variables de la base de datos. Estadísticas vitales de nacimientos [4].

CAMPO	DESCRIPCIÓN	EJEMPLO
COD_DPTO	Departamento de Nacimiento	17=Caldas
COD_MUNIC	Municipio de Nacimiento	15= Boyacá
AREANAC	Área del Nacimiento	3 = Rural disperso
SIT_PARTO	Sitio de la Parto	1= Institución de salud
OTRO_SIT	Otro sitio, ¿cuál?	2 = Domicilio
SEXO	Sexo del nacido vivo	1= Masculino
PESO_NAC	Peso del nacido vivo, al nacer	1 = Menos de 1.000
TALLA_NAC	Talla del nacido vivo, al nacer	1 = Menos de 20
ANO	Año de la ocurrencia	2022
MES	Mes de la ocurrencia	03 = Marzo
ATEN_PAR	El parto fue atendido por	1 = Médico
T_GES	Tiempo de gestación del nacido vivo	2 = De 22 a 27
T_GES_AGRU_CIE	Tiempo de gestación del nacido vivo ajustado a la agrupación sugerida por la CIE (T_GES_AGRU_CIE)	3 = De 28 a 36
NUMCONSUL	Número de consultas prenatales que tuvo la madre del nacido vivo	00= Ninguna
TIPO_PARTO	Tipo de parto de este nacimiento	2 = Cesárea
MUL_PARTO	Multiplicidad del embarazo	3 = Triple
N_EMB	Número de embarazos, incluido el presente	99 = Sin información
SEG_SOCIAL	Régimen de seguridad social en salud de la madre	2 = Subsidiado
IDCLASADMI	Entidad Administradora en Salud a la que pertenece la madre	1 = Entidad promotora de salud
EDAD_PADRE	Edad del padre en años cumplidos a la fecha del nacimiento de este hijo	999 = Sin información
IDHEMOCLAS	Hemoclasificación del nacido vivo: Grupo Sanguíneo	1 = A
IDFACTORRH	Hemoclasificación del nacido vivo: Factor RH	1 = Positivo
IDPERTET	De acuerdo con la cultura, pueblo o rasgos físicos, el nacido vivo es reconocido por sus padres como	1 = Indígena
EDAD_MADRE	Edad de la madre a la fecha del parto	2 = De 15-19 Años
EST_CIVM	Estado conyugal de la madre	4 = Está viuda
NIV_EDUM	Ultimo nivel de estudio 27que aprobó la ma28dre	2 = Básica primaria
ULTCURMAD	Último año o grado aprobado de la madre	99 = Sin información
CODPRES	País de residencia habitual de la madre en el extranjero	Colombia
CODPTORE	Departamento de residencia habitual de la madre	05 = Antioquia
CODMUNRE	Municipio de residencia habitual de la madre	1 = Cabecera municipal
AREA_RES	Área de residencia habitual de la madre	2 = Centro poblado
N_HIJOSV	Número de hijos nacidos vivos que ha tenido la madre, incluido el presente	3 = 3 hijos
FECHA_NACM	Fecha de nacimiento del anterior hijo nacido vivo	99 = Sin información
APGAR1	Prueba APGAR al minuto del nacido vivo	01 - 10 = Al minuto
APGAR2	Prueba APGAR a los cinco minutos del nacido vivo	01 - 10 = A los cinco minutos
NIV_EDUP	Nivel educativo del padre, último año de estudio que aprobó el padre	5 = Media técnica
ULTCURPAD	Último año o grado aprobado del padre	99 = Sin información
PROFESION	Profesión de quien certifica el nacimiento	2 = Enfermero(a)
TIPOFORMULARIO	Fuente del Certificado	1 = Certificado RUAF-ND
UNNAMED:0	Índice	0 = Registro 0

B. Preprocesamiento de los datos

En esta etapa se utilizó la metodología ETL (Extraer, Transformar y Cargar, por sus siglas en inglés) [13], utilizando herramientas de manipulación de datos de alto nivel, como la librería *pandas*, útil para limpieza y agregación de datos [13]. El proceso comenzó con la revisión y análisis de la estructura del conjunto de datos crudos proporcionados por el DANE [4], específicamente las bases de datos de nacimientos de los años 2019 a 2022. Se identificaron las codificaciones definidas para cada campo y su significado, lo que permitió comprender la organización de la información y preparar los datos para su posterior procesamiento. Se inició con el manejo y carga de los datos para obtener la mayor información posible. Posteriormente, se llevó a cabo una revisión exhaustiva para identificar y tratar registros nulos y duplicados, siguiendo las prácticas recomendadas en la literatura [14]. Como resultado, se identificaron y eliminaron un total de

6'085.151 datos nulos, es decir, campos con valores "NaN" o "Null", y se eliminaron 7.004 registros duplicados (ver Tabla 3). Se inició con el manejo y carga de los datos para obtener la mayor información posible. Posteriormente, se llevó a cabo una revisión exhaustiva para identificar y tratar registros nulos y duplicados, siguiendo las prácticas recomendadas en la literatura [14]. Como resultado, se identificaron y eliminaron un total de 6'085.151 datos nulos, es decir, campos con valores "NaN" o "Null", y se eliminaron 7.004 registros duplicados (ver Tabla 3).

Tabla 3. Variables eliminadas del conjunto de datos.

VARIABLE	JUSTIFICACIÓN
IDCLASADMI	No es requerido ya que la entidad prestadora de salud no aporta información importante para el objetivo de análisis.
UNNAMED: 0	Subíndice no requerido ya que al crearse un DataFrame se genera un índice por defecto.
TIPOFORMULARIO	Variable no requerida ya que se trabajará con los datos actuales, independiente del método de obtención de estos.
APGAR1	No requerida de acuerdo con el objetivo de análisis.
APGAR2	No requerida de acuerdo con el objetivo de análisis.
OTRO_SI	No es requerida ya que para ubicación se trabajará con la variable “COD_DPTO”. Además, esta tiene 2457692 registros nulos.
T_GES_AGRU_CIE	El tiempo de gestación no aporta información de acuerdo con el objetivo de análisis propuesto. Además, esta tiene 629402 registros nulos.
FECHA_NACM	La fecha de nacimiento de la madre no se considera relevante para el modelo.
PROFESION	La información académica de los padres se obtendrá de la variable ULT_CURMAD
CODPRES	El tiempo de gestación no aporta información de acuerdo al objetivo de análisis propuesto. Además, esta tiene 2085 registros nulos.
CODPTORE	El tiempo de gestación no aporta información de acuerdo al objetivo de análisis propuesto. Además, esta tiene 36621 registros nulos.
CODMUNRE	El tiempo de gestación no aporta información de acuerdo al objetivo de análisis propuesto. Además, esta tiene 36621 registros nulos.
AREA_RES	El tiempo de gestación no aporta información de acuerdo al objetivo de análisis propuesto. Además, esta tiene 36584 registros nulos.

Posteriormente, se realizaron las imputaciones necesarias para mitigar la influencia de valores faltantes en los análisis subsiguientes. Se utilizó la imputación por mediana, dada su eficacia en la reducción del sesgo causado por valores atípicos en datos demográficos [15]. Una vez completada esta etapa, se llevó a cabo un análisis descriptivo que incluyó gráficos de barras para variables categóricas y una matriz de correlación para variables continuas, permitiendo obtener una visión general, gráfica y detallada de las características principales del conjunto de datos.

Se identificaron variables con valores codificados como 9, 99 o 999, representando ausencia de información. Estos valores fueron tratados como datos nulos, siguiendo las recomendaciones de García et al [15] ya que no aportaban información relevante para los objetivos analíticos del estudio. Las variables NIV_EDUM y NIV_EDUP contenían inicialmente el 19,7% y 14,7% de datos nulos, respectivamente. Estos fueron imputados utilizando información inferida de las variables ULTCURMAD y ULTCURPAD, reduciendo los datos nulos al 10,1% y 3,1%, respectivamente. Posterior a esta imputación, y para evitar redundancias, se eliminaron las variables ULTCURMAD y ULTCURPAD, obteniendo el porcentaje de datos nulos evidenciado en la Figura 2. Este proceso resultó en un conjunto de datos final con 2.462.601 filas y 8 columnas. A pesar de la reducción en el tamaño del conjunto de datos, se mantuvo una representatividad adecuada para los análisis posteriores, garantizando la integridad y validez de los resultados obtenidos. Este proceso resultó en un conjunto de datos final con 2.462.601 filas y 8 columnas.

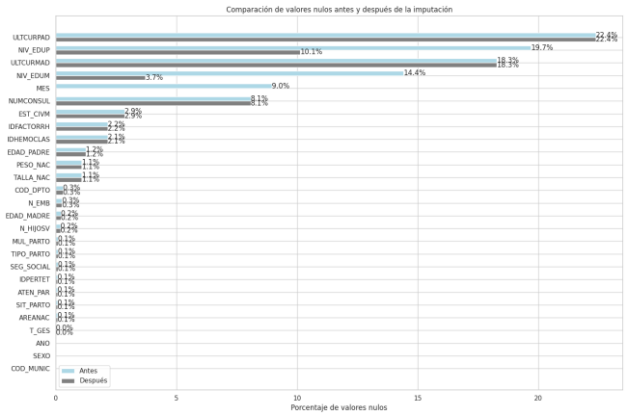

Figura 2**.** Porcentaje de datos nulos en el Dataset.

C. Análisis descriptivo de los datos

Se realiza exploración de las variables categóricas del conjunto de datos, representadas mediante gráficos de barras (ver **¡Error! No se encuentra el origen de la referencia.**), las cuales clasifican los datos en grupos o categorías discretas, aportan información cualitativa crucial para identificar patrones y tendencias en el estudio [16].

Estas variables evidencian una distribución heterogénea de los nacimientos por departamentos (COD_DEPTO), indicando variaciones regionales significativas. Destaca el departamento de Bogotá (Código 11) seguido de Antioquia

(código 5), con un número de nacimientos superior a la media nacional, posiblemente debido a [3]. Se aprecia un balance entre nacimientos de hombres (código 1) y mujeres (código 2), identificados con la variable SEXO. La mayoría de los nacimientos son a término, identificado en la variable T_GES, código 4 (38 a 41 semanas) y se producen por parto natural, identificado en la variable TIPO_PARTO, código 1 (Parto espontáneo).

Las variables relacionadas con el nivel educativo (NIV_EDUM y NIV_EDUP) revelan que tanto las madres como los padres poseen, en su mayoría, educación secundaria (Código 3), lo que sugiere avances en el acceso a la educación y posibles impactos positivos en la salud materno-infantil. Además, una alta proporción de los nacimientos corresponde a madres con acceso a seguridad social (SEG_SOCIAL, código 2) y de estado civil (EST_CIVM) soltero (código 1).

Adicionalmente, los datos muestran una tendencia decreciente anual en el número de nacimientos, reflejando dinámicas demográficas ya documentadas.

Para las variables continuas se implementa una matriz de correlación (ver **¡Error! No se encuentra el origen de la referencia.**). Allí se observa una correlación positiva entre el peso al nacer (PESO_NAC) y la talla al nacer (TALLA_NAC), con un valor de 0.57, lo cual indica que estos factores están relacionados, pero no de manera completamente lineal. También se destacan correlaciones fuertes entre el número de hijos vivos (N_HIJOSV) y el número de embarazos (N_EMB) con valor de 0,92, así como entre la edad de la madre (EDAD_MADRE) y del padre (EDAD_PADRE), con valores de 0.63, lo que sugiere asociaciones esperadas por factores biológicos y sociodemográficos.

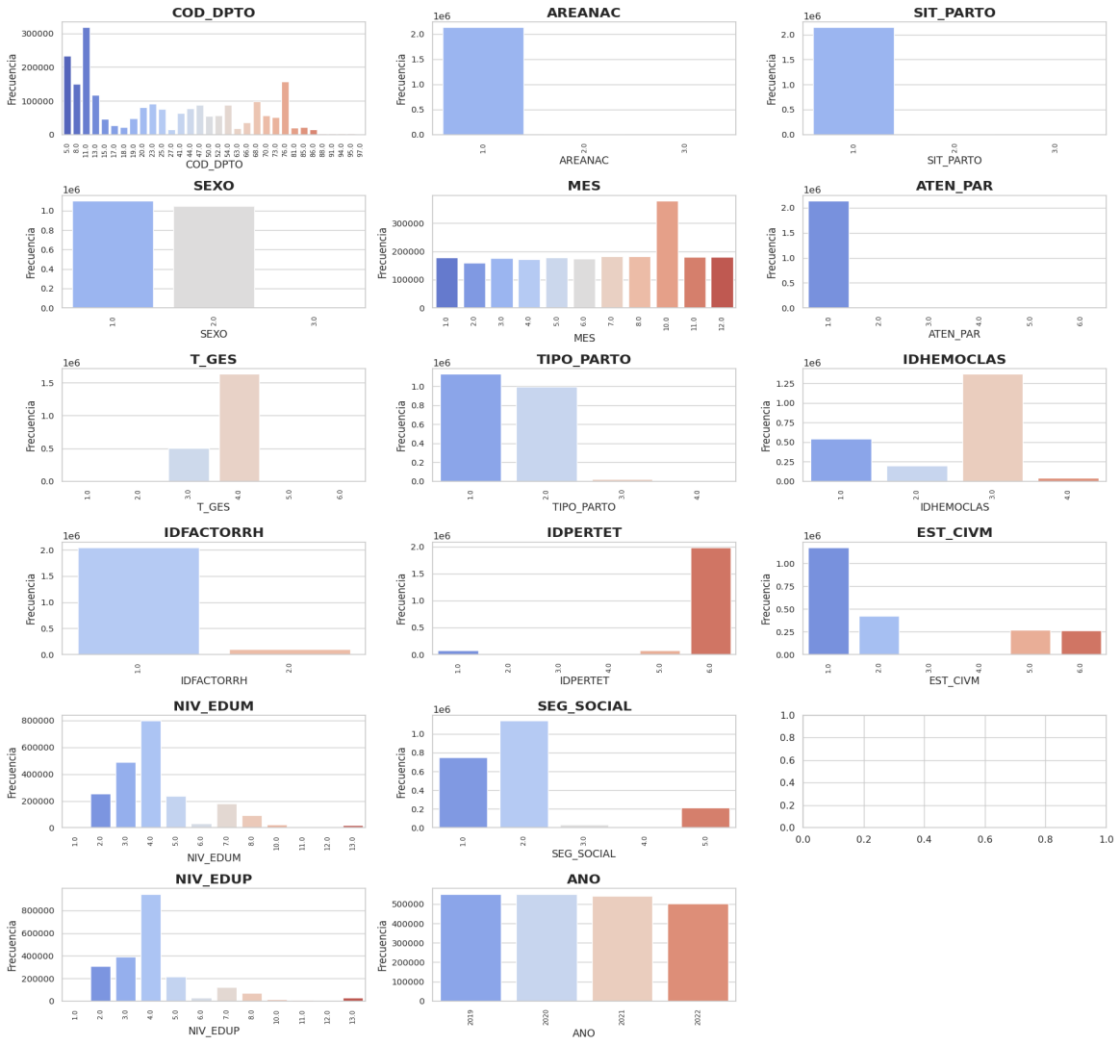

Figura 3. Distribución de variables categóricas del conjunto de datos.

Por otro lado, las correlaciones entre las variables relacionadas con el parto múltiple (MUL_PARTO),

número de consultas prenatales (NUMCONSUL), y las variables como peso y talla al nacer son débiles o nulas, indicando poca relación directa entre estos factores.

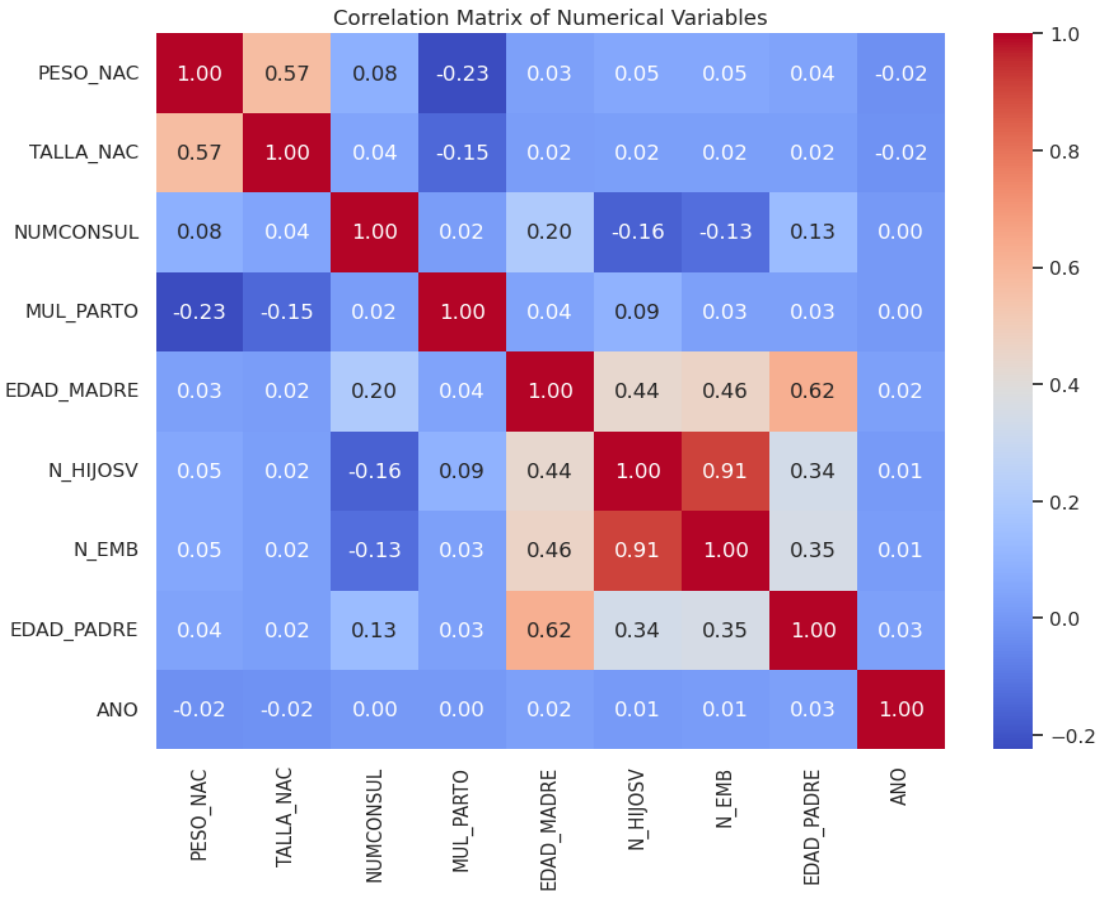

Figura 4. Matriz de correlación para variables continuas del conjunto de datos.

D. Cálculo de nuevas variables

En el marco del análisis de la tasa de natalidad, se procedió a la creación de nuevas variables en el conjunto de datos con el objetivo de calcular este indicador demográfico clave. La tasa de natalidad se define como el número de nacimientos ocurridos en un periodo de tiempo determinado, dividido por la población total de referencia, y multiplicado por 1.000 para expresarla en términos relativos [17]. Matemáticamente, se representa mediante la fórmula:

$$Tasa\ de\ natalidad = \left(\frac{\text{Número de nacimientos}}{\text{Población total}}\right) \times 1,000$$

Para ello, se crearon variables agregadas que contabilizan el número de nacimientos y la población total por departamento y año. Estas variables permiten analizar las tendencias demográficas a nivel regional y temporal, desde la perspectiva del índice de natalidad. Facilitando la identificación de patrones específicos en distintas áreas geográficas.

Esta tasa de natalidad (IN_NATALIDAD) es calculada de acuerdo con los reportes del número de nacimientos (NACIMIENTOS) y población total (PTOTAL) de cada departamento (N_DTPO), actualizado acorde para cada mes (MES) y año analizado (ANO) en el conjunto de datos.

E. Desarrollo de modelos

Después de finalizar la fase de preparación de los datos, se inició la aplicación y evaluación de los modelos predictivos de natalidad para cerrar el ciclo metodológico y cumplir con los objetivos analíticos planteados. Los modelos se evaluaron mediante métricas de desempeño, como el error cuadrático medio y el coeficiente de determinación, que

permiten identificar su efectividad y aportar retroalimentación al proceso. Para la predicción de la tasa de natalidad desde la perspectiva sociodemográfica, se propusieron tres modelos de aprendizaje automático: árbol de decisión, bosque aleatorio y XGBoost. Estos modelos fueron seleccionados debido a su eficacia comprobada en estudios demográficos y su capacidad para manejar variables tanto categóricas como continuas

Árbol de Decisión (Decision Tree): Un árbol de decisión es una estructura jerárquica que segmenta iterativamente un conjunto de datos en subconjuntos homogéneos basándose en valores de variables predictoras. Cada nodo interno representa una prueba sobre una variable, cada rama denota el resultado de la prueba y cada hoja representa una predicción del valor de la variable objetivo. En el contexto de la predicción de la tasa de natalidad, los árboles de decisión permiten identificar patrones y relaciones entre variables sociodemográficas y económicas que influyen en las tasas de natalidad mensuales y anuales [18]

Bosque Aleatorio (Random Forest): El bosque aleatorio es un conjunto de árboles de decisión que operan de manera conjunta para mejorar la precisión de las predicciones y controlar el sobreajuste. Cada árbol en el bosque se construye a partir de una muestra aleatoria del conjunto de datos y considera un subconjunto aleatorio de variables en cada división, lo que introduce diversidad y reduce la correlación entre los árboles. En la predicción de la tasa de natalidad, los bosques aleatorios pueden manejar grandes conjuntos de datos con múltiples variables predictoras, proporcionando estimaciones robustas y precisas [19].

Para implementar este modelo se seleccionaron 50 estimadores o árboles, con el fin de equilibrar el desempeño del modelo y el tiempo de entrenamiento. Si bien un mayor número de árboles puede mejorar la estabilidad y precisión del modelo, se optó por 50 debido a que en pruebas iniciales se observó que el desempeño no mejoraba significativamente al aumentar este valor, mientras que el costo computacional sí incrementaba.

XG Boost (Extreme Gradient Boosting): XG Boost es una implementación optimizada del algoritmo de boosting de gradiente, que construye modelos predictivos de manera secuencial, donde cada nuevo modelo corrige los errores de los modelos anteriores. Este enfoque permite manejar datos con alta dimensionalidad y complejidad, y es conocido por su eficiencia computacional y alto rendimiento en tareas de predicción. En el análisis de la tasa de natalidad, XG Boost puede capturar interacciones complejas entre variables y proporcionar predicciones precisas al considerar factores temporales y espaciales [20].

El detalle de la fase de análisis y desarrollo de modelos fue implementada en lenguaje Python y se encuentra disponible en el repositorio de GitHub de los autores: [Enlace disponible en: https://github.com/Sebastianvc16/Monografia_UdeA_SV-VL/blob/main/C%C3%B3digo_monografia_VanessaLargoQuintero_SebastianValenciaCadena.ipynb]".

F. Métricas de evaluación

Para evaluar el rendimiento de los modelos predictivos de la tasa de natalidad, se implementaron cuatro métricas estadísticas fundamentales que proporcionan una evaluación comprehensiva del desempeño de los modelos:

- **Error Cuadrático Medio (Mean Squared Error, MSE)**: Desarrollado originalmente por Gauss en sus trabajos de estadística [21], el MSE cuantifica el promedio de los errores al cuadrado entre los valores predichos y los valores reales. Matemáticamente, eleva al cuadrado la diferencia entre valores predichos y reales, lo que amplifica las desviaciones más significativas [22].

- **Raíz del Error Cuadrático Medio (Root Mean Squared Error, RMSE)**: Como derivación directa del MSE, el RMSE fue introducido por pioneros de la estadística como Fisher [23] para proporcionar una medida de dispersión de los errores en las mismas unidades de la variable objetivo. Un RMSE más cercano a cero indica mejor precisión del modelo, siendo especialmente útil en contextos donde la magnitud absoluta del error es relevante [24].

- **Error Absoluto Medio (Mean Absolute Error, MAE)**: Propuesto por investigadores en estadística aplicada [25], el MAE representa el promedio de los valores absolutos de los errores entre predicciones y valores reales. A diferencia del MSE, no eleva al cuadrado los errores, lo que lo hace menos sensible a valores atípicos. Esta característica lo convierte en una métrica más robusta cuando el conjunto de datos presenta valores atípicos significativos. En términos de interpretación, representa el error promedio en las mismas unidades de la variable analizada [26].

- **Coeficiente de Determinación (R-squared, R^2)**: Originalmente desarrollado por Ronald Fisher en sus trabajos sobre correlación, el R^2 cuantifica la

proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes. R^2 cercano a 1 indica que el modelo explica casi toda la variabilidad de los datos [27].

Estas métricas permiten una evaluación multidimensional del rendimiento de los modelos de predicción de la tasa de natalidad, proporcionando un mejor entendimiento de la precisión, la capacidad predictiva y la bondad de ajuste de cada modelo.

III. Resultados y Discusión

Se implementaron tres modelos de aprendizaje automático: Random forest (RF), XGboost y árbol de decisión. Cada uno se entrena con las mismas variables predictoras: código de departamento (COD_DPTO), año (ANO), mes (MES), área de nacimiento (AREANAC), nivel educativo de la madre (NIV_EDUM), nivel educativo del padre (NIV_EDUP) y código de municipio (COD_MUNIC).

Los resultados después de aplicar la técnica de RF (ver **¡Error! No se encuentra el origen de la referencia.**y **¡Error! No se encuentra el origen de la referencia.**), evidencia una proximidad de los puntos correspondientes a los valores predichos en relación con los valores reales, eso evidencia un buen desempeño del modelo donde los principales errores se concentran en los valores extremos. En la Figura 5 se destaca que el modelo de RF replica de manera consistente los patrones observados.

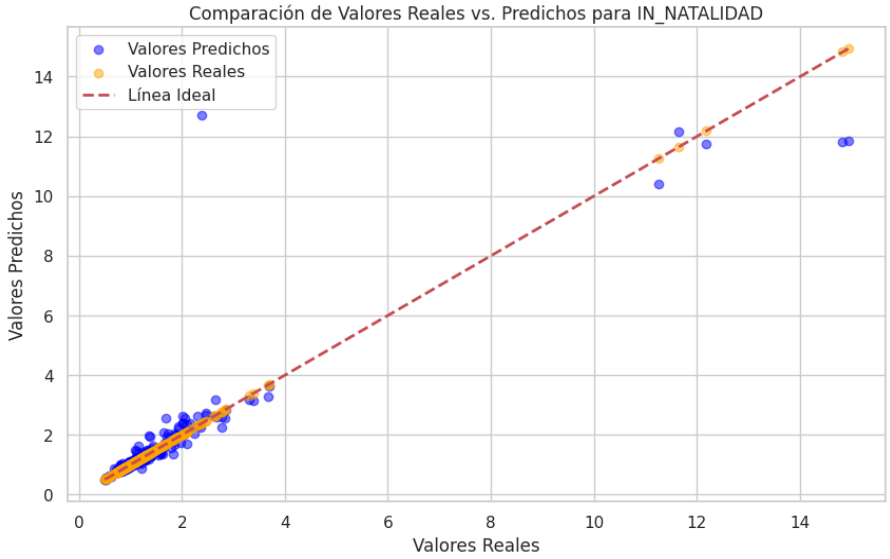

Figura 5. Índice de natalidad con modelo de Random Forest.

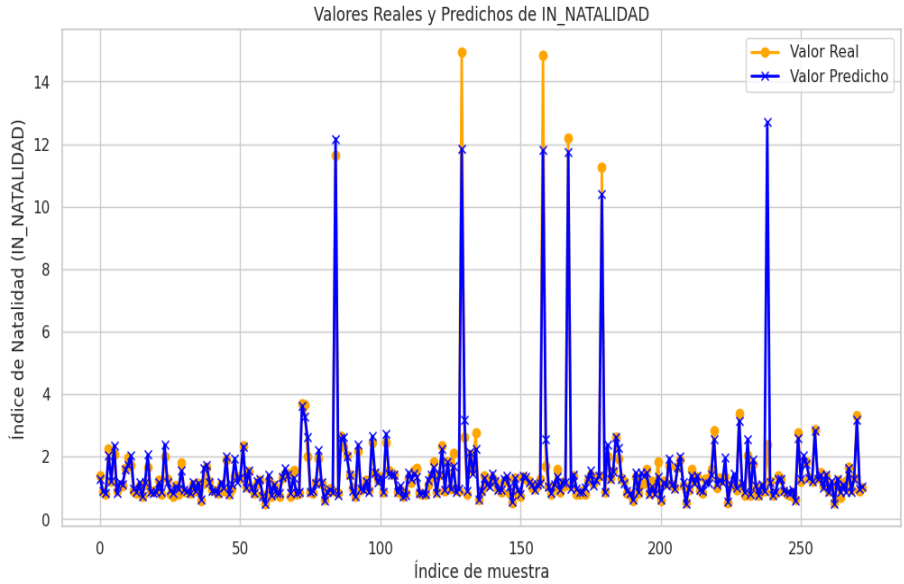

Figura 6. Índice de natalidad: Predicción vs entrenamiento.

Algunos puntos con índices de natalidad reales altos (mayores a 10) muestran predicciones del modelo que están notablemente por encima de los valores reales, estos valores atípicos se deben a la presencia de características únicas en ciertos registros que no fueron representadas durante el entrenamiento del modelo.

Por otra parte, al aplicar la técnica de XGboost (ver **¡Error! No se encuentra el origen de la referencia.** y **¡Error! No se encuentra el origen de la referencia.**) se observa una tendencia general de alineación entre los valores reales del índice de natalidad y las predicciones realizadas por el modelo. Esta alineación está indicada por la cercanía de los puntos a la línea roja discontinua, que representa la relación ideal entre ambos. Sin embargo, también se identifican valores atípicos, definidos como aquellos puntos que se alejan considerablemente de dicha línea. Estos valores representan casos en los que el modelo de predicción subestimó o sobreestimó el índice de natalidad real de manera significativa.

Finalmente, en la **¡Error! No se encuentra el origen de la referencia.**, se ilustra de manera general las características más relevantes de la técnica de árbol de decisión en relación a la predicción de la tasa de natalidad en Colombia. El modelo refleja que las variables geográficas (COD_DPTO) y temporales (MES, ANO) son las principales determinantes del índice de natalidad, ya que la división inicial basada en departamentos (COD DPTO) destaca una segmentación del 95% de las muestras para el umbral relacionado en el árbol. Posteriormente, los meses (MES 10.0, MES 3.0, MES 6.0) y el año (AÑO 2022) influyen de forma significativa en las predicciones, tal como se describen en la Tabla 4.

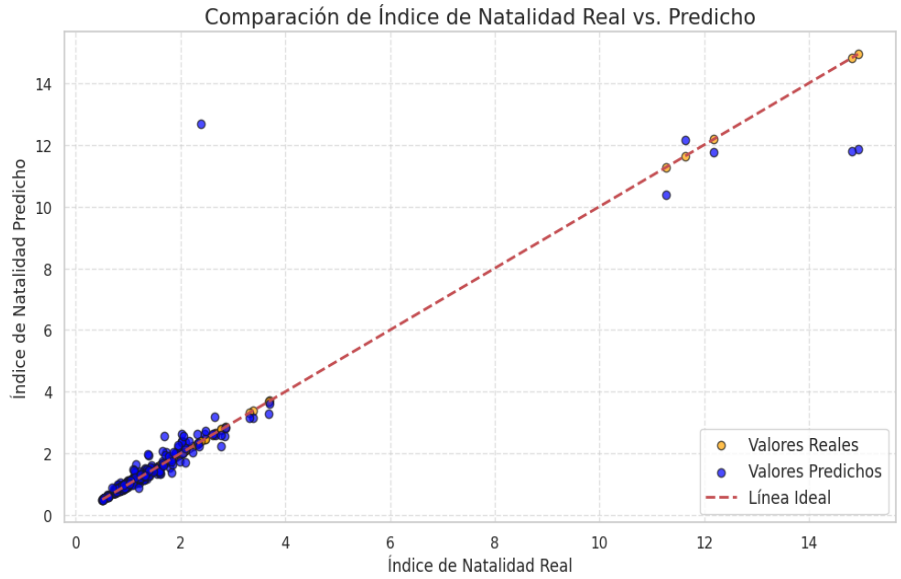

Figura 7. Gráfico de regresión: Índice de natalidad predicho vs real, utilizando XGBoost.

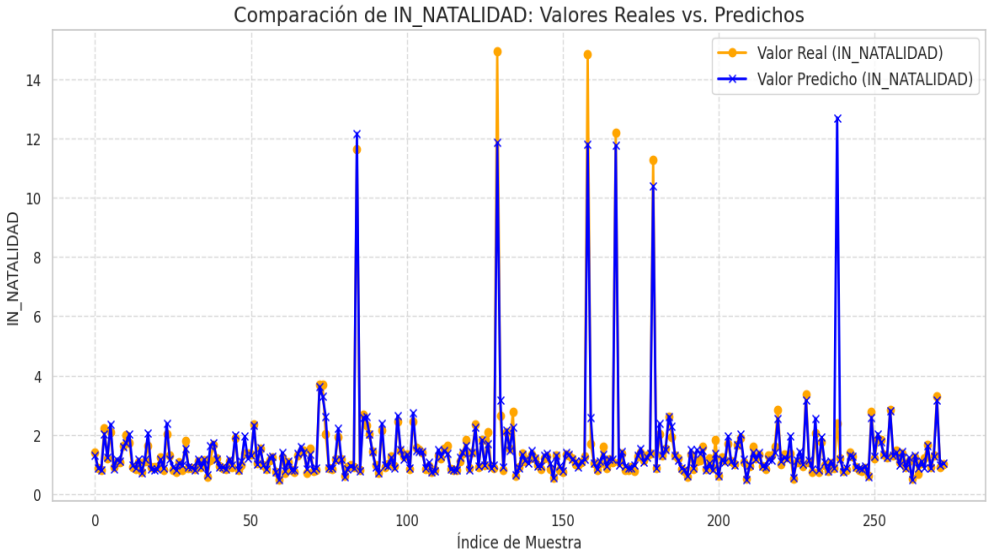

Figura 8. Índice de natalidad predicho vs real, utilizando XGBoost.

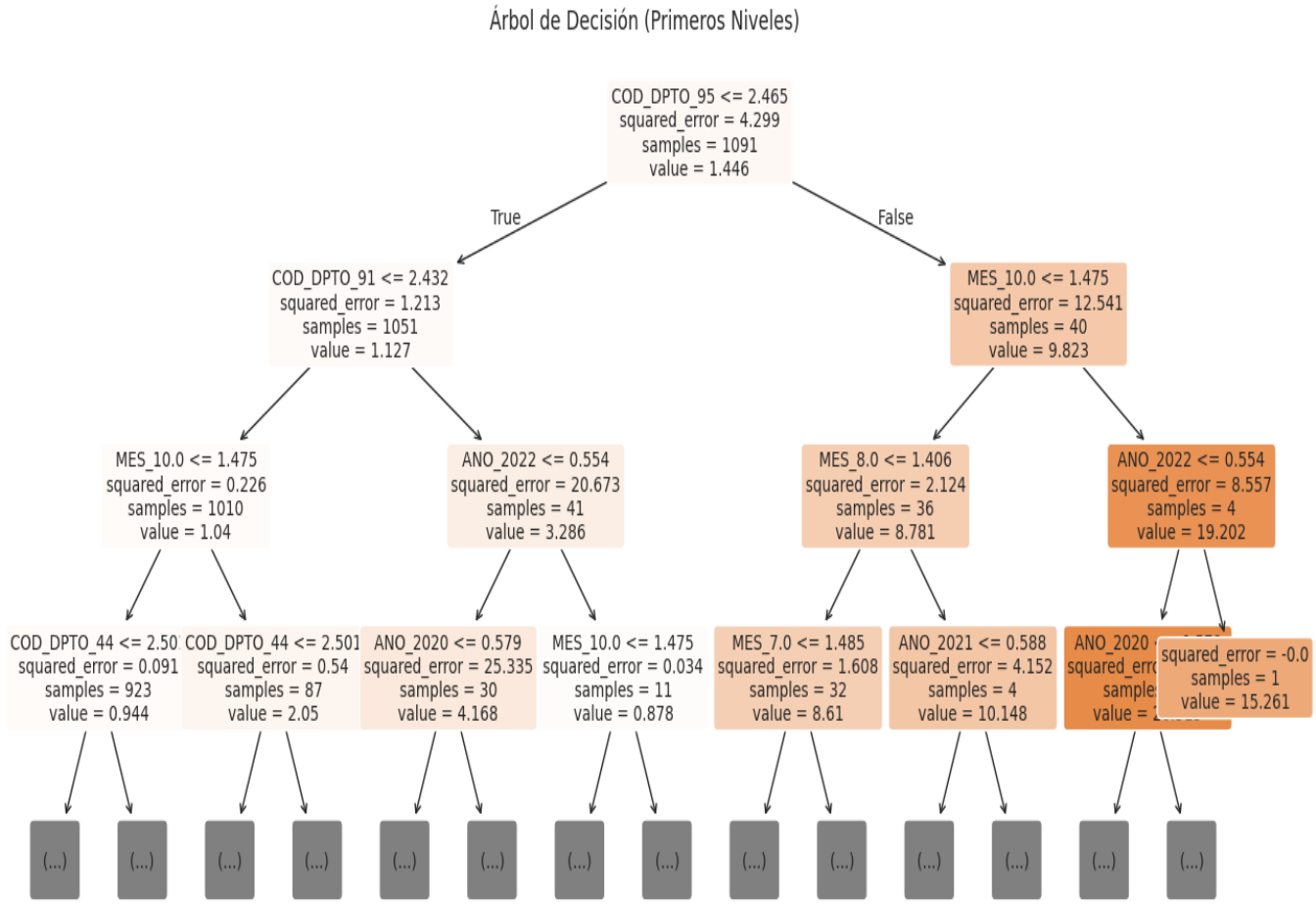

Figura 9. Resultado obtenido con la técnica Árbol de decisión.

Los resultados anteriores evidencian que los tres modelos obtuvieron resultados en términos de métricas de

rendimiento muy similares. En la Tabla 5. Comparación de métricas de desempeño de los modelos de predicción implementados.

, tanto los valores de MSE como el coeficiente R^2 tienen la misma tendencia y son similares (diferencia inferior al 15%). Sin embargo, el mejor desempeño, con un MSE de 0.0778 y un R^2 de 0.9819 en el conjunto de entrenamiento, lo registra el modelo XGBoost, lo que traduce que en un 98.19% es capaz de predecir la variabilidad en los datos de natalidad.

Por otra parte, el Árbol de Decisión también presentó resultados sobresalientes (ver Tabla 4. Interpretación de los resultados obtenidos con la técnica árbol de decisión

), con un MSE de 0.0971 y un R^2 de 0.9424. Este modelo revela que las variables más importantes para predecir la tasa de natalidad son el código de departamento (COD_DPTO), el año (ANO) y el nivel educativo de la madre (NIV_EDUM). Finalmente, el modelo de Random Forest obtuvo un desempeño ligeramente inferior, con un MSE de 0.2600 y un R^2 de 0.8457, pero aun así evidencia una capacidad predictiva sólida.

La consistencia en el desempeño de los tres modelos confirma la robustez de las predicciones puesto que todos los modelos superan un R^2 de 0.84 y los errores cuadráticos medios se mantienen por debajo de 0.2 Además de que las variables predictivas mantienen su importancia relativa a través de los modelos

Tabla 4. Interpretación de los resultados obtenidos con la técnica árbol de decisión

DIVISIÓN/NODO	***VARIABLE DE DECISIÓN***	***UMBRAL***	***DESCRIPCIÓN***
PRIMERA DIVISIÓN	**COD_DPTO_95.0**	**2.465**	*Separa los datos en dos ramas principales según regiones geográficas.*
RAMA IZQUIERDA (TRUE)	**COD_DPTO_91.0**	**2.432**	*Las características geográficas continúan siendo determinantes.*
	MES_10.0	**1.475**	*El mes (probablemente octubre) influye en los patrones de nacimientos.*
	MES_10.0	**1.475**	*El mes sigue siendo relevante, marcando su importancia en esta región.*
RAMA DERECHA (FALSE)	**ANO_2022**	**0.554**	*El año también afecta los patrones de natalidad, dividiendo aún más las ramas.*
	MES_3.0, MES_6.0	**Varios**	*Los meses específicos tienen un impacto clave en las divisiones finales.*
ERRORES CUADRÁTICOS MEDIOS	-	-	*Cada nodo refleja el error cuadrático medio; menores valores indican mayor precisión.*
VALORES PROMEDIOS	-	-	*Representan el índice de natalidad estimado en cada nodo.*

El análisis identifica tres factores principales que influyen en las tasas de natalidad:

- **Factor Geográfico (Código Departamental):** se evidencia variaciones significativas entre departamentos con patrones regionales consistentes. Una disminución del 9.6% de la tasa de natalidad entre 2022-2023 y 15.2% de reducción de 2023 a 2024.
- **Factor Temporal (Año):** Demuestra una tendencia decreciente consistente (2019-2022), con patrones estacionales identificables
- **Factor Educativo (Nivel Educativo Materno):** Se confirma como variable predictiva significativa y muestra correlaciones consistentes con las tasas de natalidad

Tabla 5. Comparación de métricas de desempeño de los modelos de predicción implementados.

Modelo	***Error Cuadrático Medio (MSE)***	***Coeficiente de Determinación (R^2)***

Random Forest	0.2600	0.8357
XG Boost	0.0778	0.9819
Árbol de Decisión	0.0971	0.9424

IV. Limitaciones Metodológicas Identificadas

Las limitaciones metodológicas del presente estudio se concentran, principalmente, en la naturaleza y alcance de los datos analizados. El período de análisis, restringido a los años 2019-2022, representa un intervalo temporal relativamente corto, lo que puede limitar la capacidad de identificar tendencias demográficas a largo plazo. Adicionalmente, se observaron desafíos significativos en la completitud y calidad de los datos, con variaciones importantes en el registro de información entre diferentes departamentos. La ausencia de ciertas variables socioeconómicas fundamentales, como ingreso familiar, acceso a servicios de salud reproductiva o indicadores económicos específicos, representa una limitación que potencialmente restringe la profundidad del análisis. Los registros presentaron heterogeneidad en su calidad, con presencia de datos faltantes en categorías específicas, particularmente en variables educativas y de caracterización demográfica. Estas limitaciones sugieren la necesidad de implementar protocolos más rigurosos de recolección y sistematización de datos, así como de considerar la incorporación de variables adicionales que permitan una comprensión más integral de los factores que influyen en la tasa de natalidad.

V. Conclusiones

Los resultados obtenidos son fundamentales para comprender las dinámicas demográficas y poblacionales actuales en Colombia, sirviendo como base de apoyo para diseñar políticas públicas que respondan eficazmente para disminuir la reducción en la tasa de natalidad, entendiendo sus posibles implicaciones económicas y sociales. En este sentido, el desempeño destacado del modelo de XGBoost refuerza su utilidad para este tipo de análisis predictivos en contextos complejos.

Colombia presenta una tendencia descendente en la tasa de natalidad para los años analizados y este ejercicio evidencia la relación directa que ésta tiene con factores demográficos, los cuales son los que más afectan las métricas de análisis de la dinámica poblacional. Esto también puede ser reflejo de cambios significativos en la estructura poblacional del país actualmente y a futuro, con una disminución de los nacimientos y un aumento en la proporción de población en edades avanzadas. Este fenómeno plantea desafíos importantes para la sostenibilidad del umbral de reemplazo poblacional y el equilibrio intergeneracional.

VI. Trabajo futuro

Basado en los hallazgos actuales, se identifican las siguientes áreas para investigación adicional:

- ***Extensiones Analíticas:*** (i) Incorporación de variables económicas cuantitativas; (ii) Ampliación del período de análisis, y (iii) Analizar las interacciones entre variables demográficas.
- **Mejoras Metodológicas:** (i) Desarrollo de modelos específicos por región; (ii) Implementación de análisis de series temporales más detallados, y (iii) Incorporación de técnicas de análisis espacial avanzadas.
- **Validación Adicional:** (i) Comparación con otros modelos predictivos; (ii) Evaluación en diferentes períodos temporales, y (iii) Validación cruzada con datos de otros países.

VII. Referencias

[1] University of Washington School of Medicine, «The lancet,» [En línea]. Available: https://www.healthdata.org/sites/default/files/2024-03/TL%20Capstones%20global%20fertility_ES.pdf. [Último acceso: 05 11 2024].

[2] Programa Nacional de las Naciones Unidas para el Desarrollo, «PNUD,» PNUD, 16 06 2023. [En línea]. Available: https://www.undp.org/es/colombia/discursos/diferencial-demografico-colombia-oportunidad-disminuir-pobreza. [Último acceso: 13 11 2024].

[3] Instituto Nacional de Estadística (DANE), «DANE,» DANE, 2024 09 20. [En línea]. Available: https://www.dane.gov.co/index.php/estadisticas-por-tema?id=34&phpMyAdmin=3om27vamm65hhkhrtgc8rrn2g4. [Último acceso: 2024 11 15].

[4] Instituto Nacional de Estadística (DANE), «Datos abiertos,» DANE, 2024 08 23. [En línea]. Available: https://www.datos.gov.co/Estad-sticas-Nacionales/Estad-sticas-Vitales-EEVV/kk5w-ugzm/about_data. [Último acceso: 01 11 2024].

[5] Universidad de costa rica, «ccp.ucr.ac.cr,» Centro centroamericano de población, 15 08 2023. [En línea]. Available: https://ccp.ucr.ac.cr/cursos/demografia_03/materia/1_demografia.htm. [Último acceso: 10 11 2024].

[6] Fondo de población de las naciones unidas, «www.unfpa.org,» Fondo de población de las naciones unidas, 23 04 2023. [En línea]. Available: https://www.unfpa.org/es/swp2023/too-few. [Último acceso: 07 11 2024].

[7] worldlifeexpectancy, «worldlifeexpectancy,» Fuentes de datos: OMS, CIA, Banco munidal, UNESCO, 15 09 2018. [En línea]. Available: https://www.worldlifeexpectancy.com/es/colombia-population-pyramid. [Último acceso: 10 11 2024].

[8] J. M. R.-G. Teresa González Pérez, «La educación de las mujeres en Iberoamérica. Análisis histórico.,» *Universidad del Atlantico*, vol. XVI, nº 38, pp. 301-306, 2021.

[9] F. A. Cortés Fernando, Tasa de natalidad y variables socio-económicas : una nota metodológica, Santiago, Chile: Comisión Económica para América Latina y el Caribe (CEPAL), 1974.

[10] R. Rodríguez Salón, «uventud, familia y posmodernidad: (des)estructuración familiar en la sociedad contemporánea,» *Fermentum*, vol. 20, nº 57, p. 39 a 55 , 2010.

[11] C. M. y. J. G. . López, «Aplicaciones del aprendizaje automático en la predicción de indicadores sociales,» *ournal of Data Science and Applications*, vol. 10, nº 4, p. 456 a 472, 2023.

[12] J. W. Tukey, Exploratory Data Analysis, Reading, Massachusetts, EE. UU.: Addison-Wesley, 1977.

[13] F. M. D. L. Y. Jorge Kamlofsky, «Tareas ETL más Simples con Pandas: Funciones Útiles Aplicadas sobre Datos Públicos,» Universidad Tecnológica Nacional, Universidad Abierta Interamericana, Buenos Aires, Argentina, 2019.

[14] M. K. y. J. P. Jiawei Han, Minería de datos: conceptos y técnicas, Morgan Kaufmann, 2012.

[15] J. L. ,. F. H. Salvador García, Data Preprocessing in Data Mining, Switzerland: Springer Cham, 2015.

[16] A. Agresti, Categorical Data Analysis, Atlanta, GA: 2014 Springer Science+Business Media Dordrecht, 2022.

[17] P. H. M. G. Samuel H. Preston, Demography: Measuring and Modeling Population Processes, Cowley Road, Oxford: Blackwell publishing, 2001.

[18] J. F. R. O. C. J. S. Leo Breiman, Classification and Regression Trees, New York: Chapman and Hall/CRC, 1984.

[19] G. B. J.-P. V. Erwan Scornet, "Consistency of Random Forests", FR: HAL: European Research Council [SMAC-ERC-280032], 2014.

[20] C. G. Tianqi Chen, XGBoost: A Scalable Tree Boosting System, San Francisco, CA, USA: arXiv, 2016.

[21] C. F. Gauss, Theoria motus corporum coelestium., Hamburgo, Alemania: Sumtibus Frid. Perthes et I. H. Besser, 1809.

[22] D. C. M. y. G. C. Runger, Applied Statistics and Probability for Engineers., Hoboken, Nueva Jersey, EE. UU.: John Wiley & Sons, 2010.

[23] R. A. Fisher, Statistical Methods for Research Workers., Edimburgo, Reino Unido: Oliver and Boy, 1925.

[24] C. J. W. y. K. Matsuura, «Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.,» *Climate Research*, Vols. %1 de %2Vol. 30, No. 1, pp. 79-82, 2005.

[25] P. Lovie, «Mean Absolute Error: A Poor Measure of Average Error.,» *Journal of Experimental Education*, Vols. %1 de %2Vol. 54, No. 3, pp. 213-216, 1986.

[26] D. M. W. Powers, «Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation.,» *Journal of Machine Learning Technologies*, Vols. %1 de %2Vol. 2, No. 1, pp. 37-63, 2011.

[27] R. A. Fisher, «On the mathematical foundations of theoretical statistics.,» *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. Vol. 222, pp. 309-368, 1922.

[28] P. B. G. E. J. G. G. Centro Latinoamericano y Caribeño de Demografía (CELADE), «mérica Latina: avances y desafíos de la implementación del Programa de Acción de El Cairo, con énfasis en el período 2004-2009,» CEPAL, Santiago, chile, 2010.