UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista Científica

http://revistas.udistrital.edu.co/ojs/index.php/revcie/index

CIDC

Artículo de investigación científica y tecnológica

Predicción de factores clave en el aumento de la demografía en Colombia a través del ensamble de modelos de Machine Learning

Prediction of Key Factors in Increasing Demographics in Colombia through Ensemble Machine Learning Models

Previsão de fatores-chave para aumentar a demografia na Colômbia através da montagem de modelos de aprendizado de máquina

Hugo-Armando Ordóñez-Erazo[1]
Camilo Ordóñez[2]
Víctor-Andrés Bucheli-Guerrero[3]

Recibido: diciembre de 2021 **Aceptado:** abril de 2022

Para citar este artículo: Ordóñez-Erazo, H. A., Ordóñez, C. y Bucheli-Guerrero, V. A. (2022). Predicción de factores clave en el aumento de la demografía en Colombia a través del ensamble de modelos de Machine Learning. *Revista Científica*, *44*(2), 282-295. https://doi.org/10.14483/23448350.19205

Resumen

El envejecimiento de la población es considerado uno de los fenómenos sociales más significativos que está transformando las economías y las sociedades en todo el mundo. Según la Organización Mundial de la Salud (OMS) el envejecimiento está en aumento. En Colombia, el crecimiento demográfico presenta un incremento natural que muestra una notable diferencia entre las tasas de natalidad y las de mortalidad general. Según el DANE, en Colombia, las tasas de crecimiento natural denotan un vertiginoso declive a lo largo del tiempo. El gobierno central y los gobiernos locales pueden ayudar en la toma de decisiones para establecer políticas de salud sexual y reproductiva. Como herramienta de apoyo aparece el Machine Learning (ML), en el cual existen algoritmos que permiten crear modelos para aprender de los datos e identificar patrones que sirven para apoyar a entes gubernamentales en el proceso de toma de decisiones. Con base en lo anterior, este trabajo propone un método de ensamble de algoritmos de ML que apoye la toma de decisiones respecto al control demográfico enfocado en natalidad. El método de predicción permitió evidenciar que la disminución de nacimientos en Colombia durante los últimos años se debe al cambio en las prioridades de mujeres y hombres. Las mujeres enfrentan discriminación y dificultad en el acceso y la permanencia del empleo a causa de la maternidad. Como consecuencia, se les dificulta articular su vida profesional con el mercado laboral. Las mujeres tienen que asumir una carga desproporcionada de cuidado, por la cual quieren tener menos hijos, es decir uno o máximo dos.

1. Ph. D. Universidad del Cauca, Popayán, Cauca, Colombia. Contacto: hugoordonez@unicauca.edu.co
2. M. Sc. Fundación Universitaria de Popayán, Popayán, Cauca, Colombia. Contacto: camilo.ordonez@docente.fup.edu.co
3. Ph. D. Universidad del Valle, Cali, Valle del Cauca, Colombia. Contacto: victor.bucheli@correounivalle.edu.co

Palabras clave: aprendizaje de maquina; ensamble de modelos; madres; número de hijos; padres; predicciones.

Abstract

Population ageing is considered to be one of the most significant social phenomena that is transforming economies and societies around the world. According to the World Health Organization (WHO), ageing is on the rise. In Colombia, demographic growth exhibits a natural increase, which shows a notable difference between birth and general mortality rates. According to DANE, in Colombia, natural growth rates denote a precipitous decline over time. The Central and local governments can help with decision-making in order to establish sexual and reproductive health policies. Machine Learning (ML) therefore appears as a support tool, in which there are algorithms that allow creating models to learn from data and identify patterns that aid in supporting government entities in the decision-making process. Based on the above, this work proposes a method for ensemble ML algorithms, which supports decision-making regarding demographic control focused on birth. The prediction method made it possible to show that the decrease in births in Colombia in recent years is due to the change in the priorities of women and men. Women face discrimination and difficulty in accessing and staying in employment due to maternity. Consequently, it is difficult for them to articulate their professional life with the job market. Women have to assume a disproportionate burden of care, which is why they want to have fewer children, namely one or two at most.

Keywords: ensemble models; fathers; machine learning; mothers; number of children; predictions.

Resumo

O envelhecimento da população é considerado um dos fenômenos sociais mais significativos que está transformando economias e sociedades em todo o mundo, segundo a Organização Mundial da Saúde (OMS) o envelhecimento está em ascensão. Na Colômbia, o crescimento demográfico mostra um aumento natural que mostra uma notável diferença entre as taxas de natalidade e as taxas de mortalidade geral; De acordo com o DANE na Colômbia, as taxas de crescimento natural denotam um declínio vertiginoso ao longo do tempo. O governo central e os governos locais podem ajudar na tomada de decisões para estabelecer políticas de saúde sexual e reprodutiva. O Machine Learning (ML) surge como uma ferramenta de apoio, na qual existem algoritmos que permitem a criação de modelos para aprender com os dados e identificar padrões que servem para apoiar as entidades governamentais no processo de tomada de decisão. Com base no exposto, este trabalho propõe um método de montagem de algoritmos de ML, que permite apoiar ou auxiliar na tomada de decisão no controle demográfico com foco na natalidade. O método de previsão permitiu mostrar que a diminuição dos nascimentos na Colômbia nos últimos anos se deve à mudança nas prioridades de mulheres e homens. As mulheres enfrentam discriminação e dificuldade de acesso e permanência no emprego devido à maternidade e, consequentemente, é difícil para elas articularem sua vida profissional com o mercado de trabalho. As mulheres têm que assumir uma carga desproporcional de cuidados, por isso querem ter menos filhos, ou seja, um ou dois no máximo.

Palavras-chaves: aprendizado de máquina; mães; montagem do modelo; numero de filhos; pais; previsões.

Introducción

Según el informe de la Organización Mundial de la Salud (OMS) sobre envejecimiento y salud del 2015 (UNDESA, 2019), el número y la proporción de personas de 60 años o más en la población está aumentando. El envejecimiento de la población es considerado uno de los fenómenos sociales más significativos que está transformando las economías y las sociedades en todo el mundo. Según Watanabe *et al.* (2022), lo anterior se debe a que a nivel mundial, las mujeres tienen menos bebés. La tasa mundial de fecundidad disminuyó de 3,2 nacimientos por mujer en 1990 a 2,5 en 2019. En este sentido, para UNDESA (2020), esto se debe a que en los países desarrollados tienden a tener una tasa de fertilidad más baja por las opciones de

estilo de vida asociadas con la riqueza económica en las que las tasas de mortalidad son bajas, el control a la natalidad es fácilmente accesible y los niños a menudo pueden convertirse en una carga económica causada por el costo de la vivienda, la educación y otros gastos involucrados en su crianza. Frente a lo anterior, en la actualidad aparecen factores claves como la educación superior, las carreras profesionales, el nivel de educación del padre, el estado civil tanto del padre como de la madre, la edad de la madre y del padre, entre otros, que a menudo influyen en que las mujeres tomen la decisión de tener hijos según las características de estos factores (O'Sullivan, 2020).

Otro aspecto notable es que, a partir de 1965, la natalidad disminuyó por la escasez de recursos económicos mezclados con bajos niveles de vida. Adicionalmente, la crisis económica de la década de los ochenta trajo consigo diferentes métodos y técnicas de anticoncepción, cambiando así las características reproductivas en el mundo. De la misma forma, la tasa general de fecundidad de Latinoamérica para el periodo de 1970 a 1975 en promedio se estimó en 5,2 hijos por mujer (Carter 2018), para el periodo de 1995-2000 este promedio descendió a 2,9 hijos por mujer, y se calcula que para el período 2020-2025 disminuirá a 2,2 hijos por mujer.

En Colombia, el crecimiento demográfico demuestra un incremento natural que muestra una notable diferencia entre las tasas de natalidad y las de mortalidad general (Ministerio de Salud y Protección Social, 2020). Según el DANE, en Colombia las tasas de crecimiento natural denotan un vertiginoso declive a lo largo del tiempo, pasando de 22 personas por cada mil habitantes en el periodo de 1985-1990 a una tasa proyectada de 12,1 en el quinquenio 2015-2020 y a una tasa de 0,9 para el quinquenio 2020-2025 (Bailey *et al.*, 2021), representando así una reducción del 45,217 % en la tasa de crecimiento original. En este sentido, según Pérez (2006), en Colombia los niveles de fecundidad están disminuyendo, lo que indica que, si bien el desarrollo económico lograría en algún momento reducciones significativas en los niveles de fecundidad, el gobierno central y los gobiernos locales pueden ayudar en este proceso, en la toma de decisiones para establecer políticas de la salud sexual y reproductiva.

Como herramienta de apoyo a la toma de decisiones aparece Machine Learning (ML), en la cual existen algoritmos que permiten crear modelos para aprender de los datos e identificar patrones que sirven de base en el proceso de toma de decisiones (De la Hoz, Fontalvo y Mendoza, 2020). En este sentido, los algoritmos de ML se han utilizado para apoyar a entes gubernamentales en la toma de decisiones en temas tales como la tendencia de hurtos en Colombia (Ordóñez, Cobos y Bucheli, 2020), para detectar modalidades de secuestro en Colombia (Giraldo Alegría, 2020), también en el órgano legislativo responsable de tomar medidas decisivas, para determinar esquemas relacionados con la tasa de crecimiento social y económico (Sharma y Shekhar, 2020), de la misma forma, para planeación inicial de vacunación con vancomicina, la cual es un antibiótico glicopeptídico de tratamiento primario para las infecciones por Staphylococcus (Matsuzaki *et al.*, 2022), también aplicaciones clínicas en planes de radioterapia para tumores cerebrales (Siciarz *et al.*, 2021).

Con el propósito de ayudar a mejorar el rendimiento de los modelos de ML y optimizar su precisión, en este trabajo se propone un método de ensamble de algoritmos de ML (Gao *et al.*, 2021) que permita dar soporte o ayuda en la toma de decisiones en el control demográfico enfocado en la natalidad. El enfoque propuesto aprovecha las ventajas de los algoritmos de aprendizaje automático, específicamente en árboles de decisión, para abordar las relaciones complejas y no lineales entre la variable dependiente y las variables explicativas de un conjunto de datos. En el enfoque, cada modelo produce una predicción diferente (Islam y Nahiduzzaman, 2022). Las predicciones de los distintos modelos se combinan para obtener una única predicción. La ventaja que se obtiene al armonizar diferentes modelos es que como cada

modelo funciona de forma diferente, sus errores tienden a compensarse (Gao *et al.*, 2021). Esto resulta en un menor error de generalización de la solución. Para ejecutar el modelo se utilizó un *dataset* del Archivo Nacional de Datos de Colombia (ANDA), el cual contiene las estadísticas de nacimientos que se registran a partir de la información proveniente de los certificados de nacido vivo en el territorio nacional.

El método de ensamble definido en este trabajo toma importancia, ya que puede servir de base para la toma de decisiones a los entes gubernamentales en relación con programas de planificación familiar o armonización del control demográfico del país, de la misma forma, sirve de base como referencia a la comunidad científica en trabajos orientados a la analítica de datos, Machine Learning o ensamble de modelos de Machine Learning.

El presente trabajo se encuentra organizado de la siguiente manera, en la siguiente sección se presenta la metodología abordada para la investigación, posteriormente la sección de los resultados y finalmente las conclusiones.

Metodología

Encontrar patrones de información en los datos es el paso fundamental en todo algoritmo de ML. Hoy en día, las técnicas de ML son consideradas herramientas poderosas y pueden ser utilizadas para analizar datos e información de varios típicos. Entre estos tópicos están: fenómenos naturales, enfermedades, aspectos económicos, sociales y naturales. Adicionalmente, pueden ser utilizadas para desarrollar técnicas novedosas para resolver problemas y aportar en la toma de decisiones en relación con los problemas abordados (Chilla *et al.*, 2022). Como las técnicas más poderosas aparecen los métodos de ensamble de algoritmos de ML, los cuales han aumentado drásticamente debido a su alta eficiencia para ayudar a los investigadores a realizar estudios sobre muchos temas en varias ramas del conocimiento. Además, los modelos de ensamble reducen la incertidumbre de cada algoritmo y aumentan su fiabilidad y precisión, lo que permite asegurar y sustentar los resultados de un estudio, ofreciendo un enfoque fiable y eficaz para abordar dificultades complejas en aplicaciones del mundo real (Telikani *et al.*, 2022). En este sentido, a continuación, se describen algunos de los trabajos más representativos en el ensamble de modelos de ML.

Trabajos relacionados

Islam y Nahiduzzaman (2022) proponen un método de ensamble para la extracción de características basado de un modelo de aprendizaje profundo para la detección de COVID19 a partir de imágenes de tomografía computarizada. El método es propuesto para analizar un problema de salud pública en más de 200 países del mundo. Dado que la detección de COVID19 mediante la reacción en cadena de la polimerasa con transcripción inversa (RT-PCR) requiere mucho tiempo y es propensa a errores, la solución alternativa de detección son las imágenes de tomografía computarizada (TC). En el método se implementaron varios algoritmos de aprendizaje automático: Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR) y Random Forest (RF). El proceso de evaluación demostró que el método logró muy buenos resultados en relación con la precisión. De la misma forma, en Huang *et al.* (2022) se propone un método de ensamble para la predicción del riesgo cardiovascular sobre un conjunto de factores de estilo de vida. Este estudio analizó el estilo de vida y el control continuo de la presión arterial. La puntuación de riesgo convencional de referencia comparada fue la puntuación de riesgo de Framingham (FRS). Las variables de resultado fueron de bajo o alto riesgo según la puntuación de calcio 0 o la puntuación de calcio 100 y superior. El método de ensamble se construyó con base en: Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), Decision Tree (DT). Combinando todos los grupos de factores de riesgo (cuestionarios de encuestas de estilo de vida,

análisis de sangre clínicos, presión arterial ambulatoria de 24 horas y control de la frecuencia cardíaca) junto con la selección de características, la predicción de grupos de riesgo de ECV bajo y alto se mejoró aún más a 0,791 y 0,790. En la misma línea de la salud, Chilla *et al.* (2022) describen un método de ensamble para la clasificación de pacientes con esquizofrenia y controles sanos utilizando diversos marcadores neuroanatómicos. El estudio buscó clasificar pacientes con esquizofrenia utilizando un conjunto diverso de medidas neuroanatómicas (volúmenes corticales y subcorticales, áreas y grosor corticales, curvatura media cortical); además, correlacionaron dichas características neuroanatómicas con las puntuaciones de evaluación de la calidad de vida (QoL) dentro de la esquizofrenia. En el proceso de evaluación el método de ensamble logró precisiones de clasificación que oscilan entre el 83 y el 87 %. Además de las puntuaciones de calidad de vida más bajas dentro de la cohorte de esquizofrenia, se encontraron correlaciones significativas entre las medidas neuroanatómicas específicas y la salud psicológica.

Por otra parte, los métodos de ensamble se han aplicado a temas como la predicción de ataques terroristas tal como en Olabanjo *et al.* (2021) se propone un método de ensamble para la predicción de zonas de peligro en lucha contra el terrorismo global. El objetivo de este trabajo fue desarrollar un modelo de aprendizaje automático que combinó Support Vector Machine y K Nearest Neighbours para la predicción de continentes susceptibles al terrorismo. Los datos se obtuvieron de Global Terrorism Database. La evaluación del modelo permitió obtener ganancia de información y las funciones basadas en ensamble produjeron una precisión del 94,17 %, 97,34 % y 97,81 %, respectivamente, en la predicción de zonas de peligro. De la misma forma, en Sakhnini *et al.* (2021) se presenta un método de ensamble para identificación y localización de ataques en la capa física en la red, fundamentado en que los problemas de seguridad cibernética se han investigado ampliamente; para esto proponen un modelo inteligente de detección e identificación de ataques capaz de clasificar el tipo de ataque en la capa física basado en un conjunto de métodos de ML. El modelo propuesto fue evaluado con un conjunto de datos de redes inteligentes simulado por los Laboratorios Nacionales de Oak Ridge y se compara con clasificadores de aprendizaje automático tradicionales. La localización de ataques y fallas se prueba dividiendo los datos y midiendo la correlación de las métricas de localización producidas por el método propuesto. Los resultados demostraron la efectividad del método propuesto para clasificar y localizar ataques en comparación con los enfoques de pares.

Como se puede apreciar en relación con los trabajos analizados, el ensamble de algoritmos de ML aporta significativamente en desarrollo de métodos de predicción o clasificación;, en ese sentido, toma pertinencia el desarrollo de la presente investigación, ya que por medio la definición de un método ensamble de algoritmos de ML se pueden analizar y predecir las tendencias de natalidad o fecundidad, aspecto que impacta directamente en la economía y la sociedad en Colombia. Es importante mencionar que el método de ensamble de algoritmos de ML también puede servir de base en el apoyo en la toma de decisiones al gobierno nacional para la generación de políticas sociales o económicas derivadas del control de la natalidad

Modelo propuesto

Los datos

El *dataset* fue tomado de las estadísticas de vitales EEVV del ANDA, sistema nacional de datos abiertos. El *dataset* registra las estadísticas de nacimientos que se producen a partir de la información proveniente de los certificados de nacido vivo en Colombia; la información según el ANDA se establece como fuente primordial para estimadores tales como como tasa bruta de natalidad y tasas de fecundidad. Información que sirve de base para diseñar planes en salud y política social. El *dataset*

contiene los datos de la natalidad en Colombia de los años 2009 a 2019, está conformado por 23 columnas y 642.657 registros. La Tabla 1 presenta la descripción del *dataset*.

Tratamiento de datos

Este se llevó a cabo siguiendo la metodología CRIPS-DM e inició con un análisis exploratorio de datos con el fin de obtener un entendimiento de los datos; en este proceso se removieron columnas o variables que no aportaban a la solución, se eliminaron datos duplicados, los valores faltantes fueron completados con el promedio entre el valor anterior y el siguiente en cada columna; después se eliminaron aquellos registros que aún contenían valores nulos; por último, en el caso de la regresión, los datos originales se normalizaron con el método Min-Max, transformando los valores en un rango entre cero y uno. Se analizaron las distribuciones de las variables, los patrones que presentaban, y se identificó como se relacionaban las variables entre sí. Seguidamente, se eliminaron algunas variables tales como: AREANAC, SIT_PARTO, AREA_RES, SEG_SOCIAL, IDPERTET.

El método de ensamble propuesto

El método usa algoritmos de ML para definir un modelo predictivo. El modelo predictivo está compuesto por los algoritmos de árboles de decisión, específicamente Random Forest, el cual se describe en Hediger, Michel y Näf (2022) y Ordóñez *et al.* (2020). Estos algoritmos son ensamblados mediante una técnica conocida como Bagging, la cual ayuda a mejorar los resultados de la predicción al combinar varios modelos. Este enfoque permite la creación de un mejor rendimiento predictivo en comparación con un solo modelo. La idea básica es aprender de un conjunto de predictores (expertos) y permitirles votar. Bagging disminuye la varianza de una sola estimación, ya que combina

Tabla 1. Descripción del dataset

Variable	Descripción
COD_DPTO	Departamento de nacimiento
AREANAC	Área del nacimiento
SIT_PARTO	Sitio de la parto
SEXO	Sexo del nacido vivo
PESO_NAC	Peso del recién nacido
TALLA_NAC	Talla del nacido vivo
MES	Mes de nacimiento
ATEN_PAR	Número de consultas
NUMCONSUL	El parto fue atendido por (partera, enfermera)
TIPO_PARTO	Tipo de parto de este nacimiento
MUL_PARTO	Multiplicidad de parto
IDHEMOCLAS	Hemoclasificación del nacido vivo: grupo sanguíneo
IDFACTORRH	Hemoclasificación del nacido vivo: factor RH
IDPERTET	De acuerdo con la cultura, pueblo o rasgos físicos, el nacido vivo es reconocido por sus
EDAD_MADRE	Edad de la madre a la fecha del parto
EST_CIVM	Estado conyugal de la madre a la fecha del parto
NIV_EDUM	Nivel educativo de la madre
AREA_RES	Área de residencia habitual de la madre
N_HIJOSV	Número de hijos nacidos vivos que ha tenido la madre, incluido el presente
FECHA_NACM	Fecha de nacimiento del anterior hijo nacido vivo
SEG_SOCIAL	Entidad administradora en salud a la que pertenece la madre
EDAD_PADRE	Edad del padre en años cumplidos a la fecha del nacimiento de este hijo
NIV_EDUP	Nivel educativo del padre

varias estimaciones de diferentes modelos. Así que el resultado puede ser un modelo con mayor estabilidad. Bagging es un modelo homogéneo de oyentes débiles que aprenden unos de otros de forma independiente en paralelo y los combina para determinar el promedio del modelo (Figura 1). A continuación, se describen los pasos de implementación del método de ensamble Bagging.

- Paso 1: se crean múltiples subconjuntos a partir del conjunto de datos original con tuplas iguales, seleccionando observaciones con reemplazo. Entonces, se toman L muestras de arranque (aproximaciones de L conjuntos de datos independientes) de tamaño B (ecuación 1).

$$\{Z_1^1, Z_2^1, Z_B^1\}, \{Z_1^2, Z_2^2, Z_B^2\}, \{Z_1^L, Z_2^L, Z_B^L\}$$

- Paso 2: se crea un modelo base en cada uno de estos subconjuntos (ecuación 2).

$$W_i,\ i = 1 \ldots \ldots L\ , f_1(x) \ldots \ldots f_l(x)$$

- Paso 3: cada modelo aprende en paralelo de cada conjunto de entrenamiento y de forma independiente entre sí (ecuación 3).

$$W_1\ (.), W_2\ (.), W_L\ (.)$$

- Paso 4: las predicciones finales se determinan combinando las predicciones de todos los modelos (ecuación 4). El promedio de fi, para i = 1 hasta L.

$$\bar{f}(x) = \sum_{i=1}^{l} f_i(x)$$

Resultados

Este proceso se realizó para determinar si el modelo realizará una buena tarea de predicción para nuevos y futuros datos que se puedan registrar en el *dataset*. Esto debido a que los nuevos datos pueden tener valores desconocidos. En este sentido, el rendimiento del método de ensamble de algoritmos de ML definido en este trabajo se evaluó mediante las métricas coeficiente de determinación (R^2) y raíz del error cuadrático medio (RMSE) (Bordbar *et al.*, 2022) (ver tabla 2).

Las métricas

El rendimiento del método de ensamble de algoritmos de ML definido en este trabajo se evaluó mediante las métricas coeficiente de determinación (R^2) y raíz del error cuadrático medio (RMSE). La Tabla 2 presenta las ecuaciones, la descripción y el criterio de desempeño de cada métrica de evaluación.

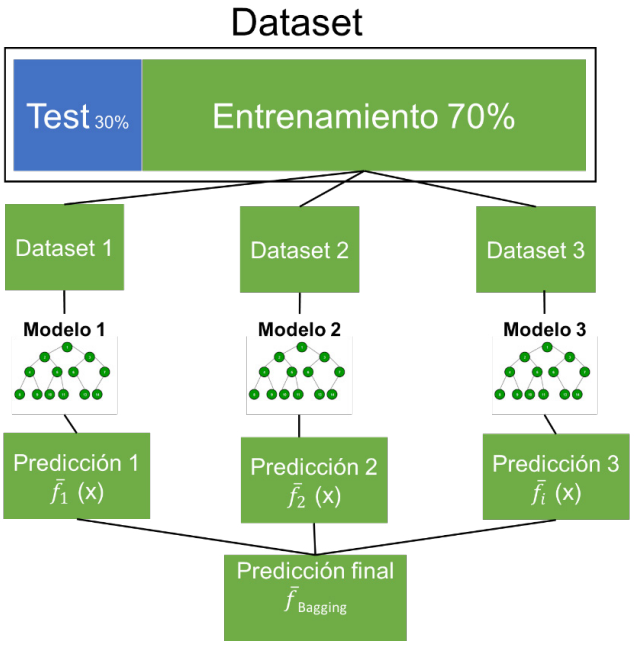

Figura 1. Ensamble Bagging

Las variables a predecir

Según un informe del Ministerio de Salud y Protección Social (2020), la mayoría de las mujeres al momento de tener hijos prestan mayor importancia factores tales como desarrollo profesional, nivel de estudios, proyectos de vida, estado civil atnto del padre como de ellas mismas, lo cual ha venido retrasando la edad para tener hijos y por ende la población en Colombia. Con base en esto, en este trabajo se tomaron como referencias los factores presentados en Ministerio de Salud y Protección Social (2020), los cuales se encuentran representados en el *dataset* por las variables que se muestran en la figura 2. Para medir la importancia

Tabla 2. Métricas de evaluación

Métrica	Ecuación	Descripción	Criterio de desempeño
R^2	$1-\frac{\Sigma(y_i-x_i)^2}{\Sigma(x_i-\underline{x_i})^2}$	Establece qué tanto se aproximan los datos reales a la línea de regresión.	Oscila entre 0 y 1, entre más se aproxime a 1, mejor será el rendimiento del modelo.
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-x_i)^2}$	Representa la diferencia entre los valores reales y los valores estimados por los modelos.	Es un valor positivo y cuanto más se aproxime a 0, mejor será el resultado de la estimación.

	EDAD_MADRE	EST_CIVM	NIV_EDUM	N_HIJOSV	EDAD_PADRE	NIV_EDUP
EDAD_MADRE	1.000000	0.082801	0.052226	0.380975	0.074163	-0.028138
EST_CIVM	0.082801	1.000000	0.313823	-0.077800	0.195487	0.230429
NIV_EDUM	0.052226	0.313823	1.000000	0.013120	0.337082	0.537084
N_HIJOSV	0.380975	-0.077800	0.013120	1.000000	0.014422	0.024870
EDAD_PADRE	0.074163	0.195487	0.337082	0.014422	1.000000	0.393974
NIV_EDUP	-0.028138	0.230429	0.537084	0.024870	0.393974	1.000000

Figura 2. Matriz de correlación de las variables

de cada una de estas variables, se analizó la correlación lineal de importancia a cada par de variables. Como muestra la imagen, existe una correlación positiva entre el nivel de educación del padre con el nivel de educación de la madre, de la misma forma, la edad de la madre con el número de hijos, el nivel de educación del padre con la edad de la madre, estas correlaciones positivas demuestran que de alguna forma las mujeres tienen presente estos factores al momento de planear tener hijos.

Resultados de la predicción

Una vez analizadas las relaciones entre las variables, como primera instancia, se evaluó cuál de los algoritmos de ML presenta mejores resultados para definir el método de ensamble. Para esto se evaluaron los algoritmos de regresión lineal estándar (línea naranja), regresión k-NN (línea roja) y árboles de decisión (línea verde).

La figura 3a muestra la predicción de formación de la madre, en esto se puede apreciar que la tendencia en los últimos años ha ido cambiando, se pasa de una formación media técnica (2005) a una formación de alto nivel como lo es especialización o maestría a partir de 2019 y en adelante. De la misma forma, el nivel de educación del padre (ver figura 3b) ha pasado de una técnica profesional desde 2005 a niveles de profesional a partir del año 2012. Como se puede observar, a medida que pasa el tiempo, tanto mujeres como hombres se preocupan más por su preparación académica y su perfil profesional, factor por el cual la intención de tener hijos se ve aplazada hasta completar su formación.

En relación con la evaluación del rendimiento de los algoritmos en la Tabla 3, se observa que para la variable nivel de educación de la madre los mejores resultados los obtiene Decision Tree (árboles de decisión) tanto para RMSE como para R^2, Esto demuestra que las predicciones realizadas están acordes, es decir que a nivel que el tiempo va

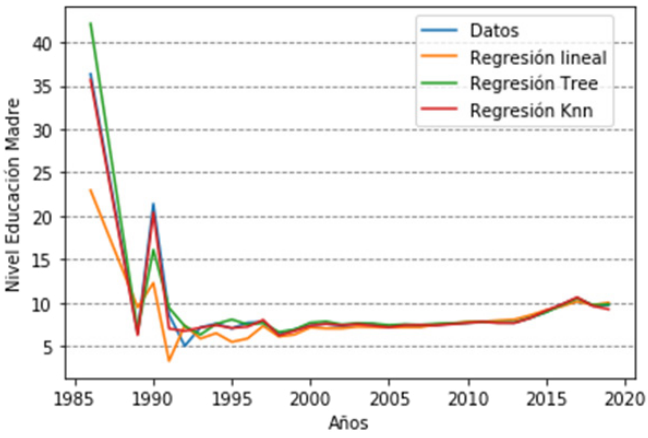

Figura 3a. Nivel formación madre

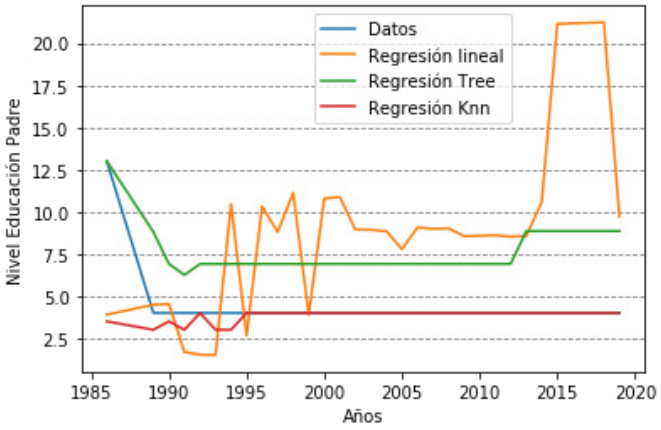

Figura 3b. Nivel formación padre

Valores por categoría.
1 - Preescolar, 2 - Básica primaria, 3 - Básica secundaria, 4 - Media académica o clásica, 5 - Media técnica, 6 - Normalista, 7 - Técnica profesional, 8 - Tecnológica, 9 - Profesional, 10 - Especialización, 11 - Maestría, 12 - Doctorado, 13 - Ninguno, 14 - Sin información

Tabla 3. Evaluación variable nivel de educación madre y padre

Algoritmo	RMSE		R^2	
	Madre	Padre	Madre	Padre
Regresión lineal	7,1008	12,5907	0,3526	0,3519
Decision Tree (árboles de decisión)	4,2880	11,0368	0,5495	0,40330
Regresión k-NN	4,6151	11,9661	0,3536	0,11630

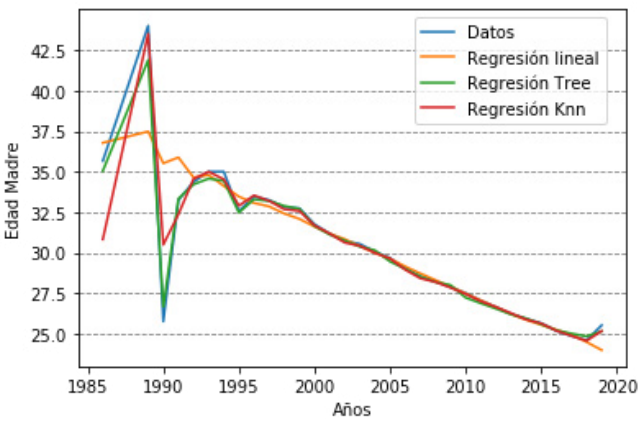

Figura 4a. Edad madre

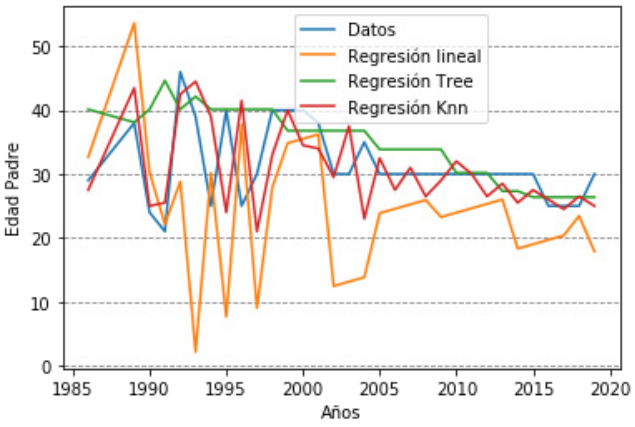

Figura 4b. Edad padre

avanzado, las mujeres se preocupan inicialmente por su proceso de formación profesional. Por otra parte, la variable nivel de educación del padre, los árboles de decisión obtienen los mejores resultados, aspecto que denota que los padres al igual que las madres dan prioridad a su formación o perfil profesional antes de formar una familia o tener más hijos.

En la evaluación de esta variable se tiene que el mejor rendimiento lo obtienen árboles de decisión tanto para RMSE como R^2, estos valores logrados permiten ver que la predicción es coherente con los datos, es notable que la tendencia de padre y madre es tener sus hijos a una edad en que ellos de alguna forma ya tengan estabilidad laboral o económica; en ese sentido, la edad se convierte en un factor fundamental a la hora de planear tener hijos, tanto para el padre como para la madre (Tabla 4).

En relación con la variable de número de hijos nacidos vivos por madre, la Figura 5 muestra que, a medida que ha ido pasando el tiempo, el número de hijos por madre o familia ha ido disminuyendo notablemente; se tiene que de 1990 a 1995 las madres tenían alrededor de tres hijos, pero a partir del año 2000 las cosas van cambiando, ya las mujeres no tienen más de uno o dos hijos, esto se debe a los cambios económicos o sociales. Las mujeres se han preocupado más por su formación profesional y académica, esto debido a que ellas tratan de articular su vida profesional con el mercado laboral, factor que les garantiza estabilidad económica. Debido a lo anterior, las mujeres son más conscientes de los retos asociados y de los costos que implica tener hijos y criarlos, inclusive algunas lo hacen por convencimiento de no querer aumentar o aportar al aumento de la población.

Para la variable número de hijos nacidos vivos, la Tabla 5 muestra los valores logrados obtenidos con base en el rendimiento de la predicción de los algoritmos, en estos se observa que los árboles de decisión logran los mejores valores de RMSE y R^2, seguidos de la regresión k-NN. Ante estos resultados es notable que las predicciones se ajustan a la realidad, ya que en esta época las mujeres no están dispuestas a tener más de dos hijos, esta realidad toma relevancia debido a factores que se han mencionado anteriormente, ya que para tener oportunidades en el mercado laboral que les ofrezca estabilidad económica para poder criar en condiciones óptimas a sus hijos, es necesario contar con un buen perfil profesional o al menos estar en el proceso de formación.

Evaluación método de ensamble

Para esta evaluación se tomaron los algoritmos de regresión lineal, de árboles de decisión, el cual obtuvo los mejores resultados en la fase anterior y se comprobó frente al método propuesto (ver figura 6). En estos se tiene que el método de ensamble propuesto obtiene mayor nivel de similitud y coherencia con los datos, en estos resultados se puede apreciar que la predicción del método propuesto a partir del año 2016 tiende

Tabla 4. Evaluación variable edad madre y padre

ALGORITMO	RMSE		R^2	
	MADRE	PADRE	MADRE	PADRE
Regresión lineal	20,5628	41,4749	-10,319	0,1914
Decision Tree (árboles de decisión)	**3,0286**	**29,7662**	**0,5954**	**0,2802**
Regresión k-NN	3,5351	30,7773	0,4078	-0,0333

Tabla 5. Evaluación variable número de hijos madre. Fuente: elaboración propia

ALGORITMO	RMSE	R^2
Regresión lineal	0,7389	0,1848
Decision Tree (árboles de decisión)	**0,62310**	**0,2639**
Regresión k-NN	0,71304	0,0681

a que las mujeres tienen entre uno y dos hijos como máximo, valores que se mantienen desde el año 2018 y se proyectan pasando el 2020. Estas predicciones se ajustan plenamente a la realidad que se vive en este momento en la sociedad, la situación económica y laboral para las mujeres con poco nivel de escolaridad tiende a ser más complicada, factor por el cual la decisión de tener menos hijos es siempre de mayor importancia.

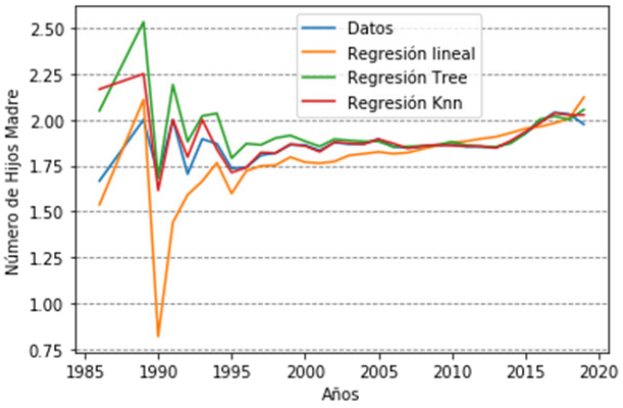

Figura 5. Número de hijos madre

En relación con la evaluación del rendimiento (Tabla 6), el método propuesto obtiene los mejores resultados, para RMSE obtiene un 20 % de mejoría en el rendimiento en relación con los árboles de decisión clásicos y de 30 % para la regresión lineal clásica, esto se debe a que la técnica de Bagging realiza varias predicciones a partir de los subconjuntos de datos creados a partir de los datos de entrenamiento, para posteriormente encontrar un promedio entre estas, de esta forma se reduce la varianza y aumenta la precisión del método predictivo, debido a que se obtienen múltiples muestras de la población y se ajusta un modelo distinto con cada una de ellas. De la misma forma, para R^2 el método propuesto logra mejores resultados en relación con los árboles de decisión y la regresión lineal clásica, los cuales oscilan entre 37 % más que los árboles de decisión y 48 % más para la regresión lineal clásica.

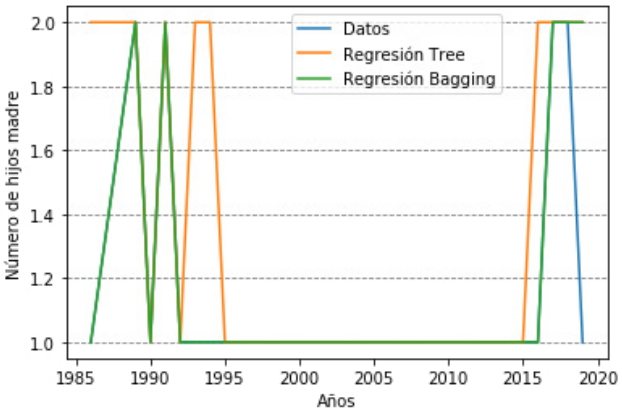

Figura 6. Evaluación del método propuesto

Para corroborar los valores obtenidos en las métricas de rendimiento, se evalúo la influencia de cada uno de los predictores, esto se llevó a cabo a través de la técnica de importancia de permutación (Rengasamy, Rothwell y Figueredo, 2021), la cual identifica la influencia que tiene cada predictor sobre una determinada métrica de evaluación del modelo, esta fue estimada por medio de validación cruzada. El valor asociado con cada predictor se presenta en la tabla 7; se puede observar que para la variable número de hijos vivos madre, el predictor principal (número 1) es la edad de la madre, como se apreció anteriormente este factor es fundamental para una mujer a la hora de tomar la decisión de tener un hijo; otro factor importante es el nivel de educación de la madre (número 2), esto es muy claro, ya que las mujeres tienen como prioridad tener un buen nivel profesional o de formación para poder garantizar la estabilidad de un hijo; otro predictor importante es nivel de educación del padre (número 3), este factor toma importancia para una mujer, ya que cuando el padre también tiene alto nivel de formación o perfil profesional este puede garantizar estabilidad a sus hijos. Como se puede apreciar el predictor de la edad del padre obtiene el último nivel de importancia, esto puede deberse a que en la actualidad en la sociedad prima la competitividad que tienen los padres, tanto madre como padre, es decir, qué tan competitivos son a nivel profesional o laboral, hecho que demanda tiempo y esfuerzo en etapa

Tabla 6. Resultados de rendimiento del método propuesto

Algoritmo	RMSE	R^2
Regresión lineal	0,7466	0,1793
Decision Tree (árboles de decisión)	0,6249	0,2565
Bagging	**0,4252**	**0,6293**

Tabla 7. Importancia de los predictores

Predictor	Importancia
EDAD_MADRE	0,423352
NIV_EDUM	0,348951
NIV_EDUP	0,104343
EST_CIVM	0,041111
EDAD_PADRE	0,033671

de edad temprana, haciendo que la edad para tener hijos se postergue hasta que los padres puedan garantizar estabilidad para los hijos.

Conclusiones

En este trabajo se definió un método de ensamble de algoritmos de ML, para implementar y evaluar el método se utilizó un *dataset* tomado de las estadísticas de vitales EEVV del ANDA, sistema nacional de datos abiertos. Para el tratamiento de los datos se aplicó la metodología CRISP-DM, esta permitió obtener entendimiento del conjunto de datos y conceptualizar en el dominio de estos.

A través de las predicciones del método se pudo llegar a concluir que las nuevas generaciones tienen otros intereses en su vida, algunas mujeres le dan mayor importancia a su proyección profesional, al estudio, a sus proyectos de vida, por los cuales han venido atrasando la edad en la cual tienen hijos. Esto ha llevado a que muchos hombres y mujeres de estas nuevas generaciones decidan que no está en su proyecto de vida tener hijos, además tanto madres como padres son conscientes de los retos asociados y de los costos que implica criarlos, incluso algunos lo hacen por convicción de no querer aumentar o aportar al aumento de la población, factor por el cual la población colombiana ha venido envejeciendo.

Las predicciones muestran que las mujeres a partir del año 2005 pasaron de una formación técnica a tener perfiles en alta formación como es al grado de maestría o especialización del año 2019 en adelante, motivo por el cual la prioridad de tener hijos se atrasa, así mismo, el nivel de educación del padre a partir del año 2005 pasó de una técnica profesional a nivel profesional a partir del año 2012 y en adelante, motivo por el cual también los hombres esperan para tomar la decisión de tener hijos.

En relación con la edad en que las mujeres deciden tener hijos es cuando ya han alcanzado un cierto grado o perfil profesional, lo cual ocurre después de los 25 años, de la misma forma para los padres, estos a los 27 años ya se han formado y tienen alguna estabilidad laboral o económica.

El método de predicción permitió evidenciar que la disminución de nacimientos en Colombia durante los últimos años se debe al cambio en las prioridades de mujeres y hombres. Las mujeres enfrentan los mayores costos asociados a la fecundidad, así mismo enfrentan discriminación

y dificultad en el acceso y la permanencia del empleo a causa de la maternidad, además como consecuencia se les dificulta articular su vida profesional con el mercado laboral. Las mujeres tienen que asumir una carga desproporcionada de cuidado y esta es una de las razones, tal vez una de las más fuertes, por las cuales las mujeres quieren tener menos hijos, es decir uno o dos máximo.

Como trabajos futuros se espera probar en método con otros *datasets*, tales como accesibilidad a la educación en Colombia, índice de habitantes de calle con estado de indigencia o acceso a vivienda propia en Colombia, datos que se encuentran en ANDA, sistema nacional de datos abiertos. Además, complementar el método propuesto con otras técnicas de ensamble tales como Boosting técnica de aprendizaje secuencial, Stacking técnica con procedimiento general para ensamblar modelos base.

Agradecimientos

El profesor Hugo Ordóñez agradece por el apoyo y el tiempo asignado para realizar la investigación a la Universidad del Cauca, en la cual labora como profesor asociado; el profesor Camilo Ordóñez agradece a la Fundación Universitaria de Popayán. El profesor Víctor Buchelli agradece a la Facultad de Ingeniería de la Universidad del Valle por el apoyo y el tiempo asignado para realizar la investigación.

Referencias

Bailey, H. H., Janssen, M. F., Varela, R. O., Moreno, J. A. (2021). EQ-5D-5L Population Norms and Health Inequality in Colombia. *Value in Health Regional Issues*, *26*, 24-32. https://doi.org/10.1016/j.vhri.2020.12.002

Bordbar, M., Aghamohammadi, H., Pourghasemi, H. R., Azizi, Z. (2022). Multi-hazard spatial modeling via ensembles of machine learning and meta-heuristic techniques. *Scientific Reports*, *12*. https://doi.org/10.1038/s41598-022-05364-y

Carter, E. D. (2018). Population control, public health, and development in mid twentieth century Latin America. *Journal of Historical Geography*, *62*, 96-105. https://doi.org/10.1016/j.jhg.2018.03.012

Chilla, G. S., Yeow, L. Y., Chew, Q. H., Sim, K., Prakash, K. N. B. (2022). Machine learning classification of schizophrenia patients and healthy controls using diverse neuroanatomical markers and Ensemble methods. *Scientific Reports*, *12*. https://doi.org/10.1038/s41598-022-06651-4

De la Hoz, E. J., Fontalvo, T. J., Mendoza, A. A. (2020). Aprendizaje automático y PYMES: oportunidades para el mejoramiento del proceso de toma de decisiones. *Investigación e Innovación en Ingenierías*, *8*(1), 21-36. https://doi.org/10.17081/invinno.8.1.3506

Gao, K., Yang, Y., Zhang, T., Li, A., Qu, X. (2021). Extrapolation-enhanced model for travel decision making: an ensemble machine learning approach considering behavioral theory. *Knowledge-Based Systems*, *218*, e106882. https://doi.org/10.1016/j.knosys.2021.106882

Giraldo Alegría, S., Ordóñez Palacios, L. E., Bucheli Guerrero V., Ordóñez Erazo, H. (2020). Modelo de redes neuronales para predecir la tendencia de víctimas de secuestro en Colombia. *Investigación e Innovación en Ingenierías*, *8*(3), 38-49.

Hediger, S., Michel, L., Näf, J. (2022). On the use of random forest for two-sample testing. *Computational Statistics and Data Analysis*, *170*, e107435. https://doi.org/10.1016/j.csda.2022.107435

Huang, W., Ying, T. W., Chin, W. L. C., Baskaran, L., Marcus, O. E. H., Yeo, K. K., Kiong, N. S. (2022). Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction. *Scientific Reports*, *12*. https://doi.org/10.1038/s41598-021-04649-y

Islam, R., Nahiduzzaman. (2022). Complex features extraction with Deep learning model for the detection of COVID19 from CT scan images using ensemble based machine learning approach. *Expert Systems with Applications*, *195*, e116554. https://doi.org/10.1016/j.eswa.2022.116554

Matsuzaki, T., Kato, Y., Mizoguchi, H., Yamada, K. (2022). A machine learning model that emulates experts' decision making in vancomycin initial dose planning. *Journal of Pharmacological Sciences*, 148(4), 358-63. https://doi.org/10.1016/j.jphs.2022.02.005

Ministerio de Salud y Protección Social. (2020). *Análisis de Situación de Salud (ASIS): Colombia, 2020*. Bogotá: Ministerio de Salud y Protección Social

O'Sullivan, J. N. (2020). The Social and environmental influences of population growth rate and demographic pressure deserve greater attention in ecological economics. *Ecological Economics*, *172*, e106648. https://doi.org/10.1016/j.ecolecon.2020.106648

Olabanjo, O. A., Aribisala, B. S., Mazzara, M., Wusu, A. S. (2021). An ensemble machine learning model for the prediction of danger zones: Towards a global counter-terrorism. *Soft Computing Letters*, *3*, e100020. https://doi.org/10.1016/j.socl.2021.100020

Ordóñez, H., Cobos, C., Bucheli, V. (2020). Modelo de Machine Learning para la predicción de las tendencias de hurto en Colombia. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, *E29*, 494-506

Pérez, G. J. (2006). *Dinámica demográfica y desarrollo regional en Colombia*. Documentos de Trabajo Sobre Economía Regional, 78. Cartagena: Banco de la República

Rengasamy, D., Rothwell, B. C., Figueredo, G. P. (2021). Towards a more reliable interpretation of machine learning outputs for safety-critical systems using feature importance fusion. *Applied Sciences*, *11*(24), e11854. https://doi.org/10.3390/app112411854

Sakhnini, J., Karimipour, H., Dehghantanha, A., Parizi, R. M. (2021). Physical layer attack identification and localization in cyber-physical grid: An ensemble deep learning based approach. *Physical Communication*, *47*, e101394. https://doi.org/10.1016/j.phycom.2021.101394

Sharma, A., Shekhar, H. (2020). Intelligent Learning based opinion mining model for governmental decision making. *Procedia Computer Science*, *173*, 216-224. https://doi.org/10.1016/j.procs.2020.06.026

Siciarz, P., Alfaifi, S., Van Uytven, E., Rathod, S., Koul, R., McCurdy, B. (2021). Machine Learning for Dose-Volume Histogram Based Clinical Decision-Making Support System in Radiation Therapy Plans for Brain Tumors. *Clinical and Translational Radiation Oncology*, *31*, 50-57. https://doi.org/10.1016/j.ctro.2021.09.001

Telikani, A., Tahmassebi, A., Banzhaf, W., Gandomi, A. H. (2022). Evolutionary machine learning: A survey. *ACM Computing Surveys*, *54*(8). https://doi.org/10.1145/3467477

United Nations Department of Economic and Social Affairs (UNDESA). (2019). *World Population Prospects 2019: Highlights*. New York: United Nations

United Nations Department of Economic and Social Affairs (UNDESA). (2020). *World Fertility and Family Planning 2020: Highlights*. New York: United Nations

Watanabe, J., Kimura, T., Nakamura, T., Suzuki, D., Takemoto, T., Tamakoshi, A. (2022). Associations of Social Capital and Health at a City with High Aging Rate and Low Population Density. *SSM - Population Health*, *17*, e100981. https://doi.org/10.1016/j.ssmph.2021.100981