**Clasificación demográfica aplicada al análisis de reducción de tasas de natalidad en Colombia**

Lida Vanessa Largo Quintero

Sebastian Valencia Cadena

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

Asesor  
David Villanueva Valdez, M.Sc. Data Scientist

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). *Clasificación demográfica aplicada al análisis de reducción de tasas de natalidad en Colombia*. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVI.

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Encabezado**

**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 Sebastian Valencia Cadena.

**Tabla de contenido**

[Resumen 9](#_3rdcrjn)

[Abstract 10](#_26in1rg)

[1.](#_35nkun2) Descripción del problema 11

[1.1.](#_1ksv4uv) Problema de negocio 11

[1.2.](#_44sinio) Aproximación desde la analítica de datos 11

[1.3.](#_z337ya) Origen de los datos 11

[1.4.](#_3j2qqm3) Métricas de desempeño 11

[2.](#_4i7ojhp) Objetivos 12

[2.1.](#_2xcytpi) Objetivo general 12

[2.2.](#_3whwml4) Objetivos específicos 12

[3.](#_2bn6wsx) Datos 13

[3.1.](#_qsh70q) Datos originales 13

[3.2.](#_3as4poj) Datasets 13

[3.3.](#_1pxezwc) Analítica descriptiva 13

**Lista de tablas**

[**Tabla 1**  Codificación de variables en base de datos original 1](#_2u6wntf)4

[**Tabla 2** Ejemplo de codificación de las variables](#_19c6y18) 14

**Siglas, acrónimos y abreviaturas**

**APA** American Psychological Association

**Cms.** Centímetros

**ERIC** Education Resources Information Center

**Esp.** Especialista

**MP** Magistrado Ponente

**MSc** Magister Scientiae

**Párr.** Párrafo

**PhD** Philosophiae Doctor

**PBQ-SF** Personality Belief Questionnaire Short Form

**PostDoc** PostDoctor

**UdeA** Universidad de Antioquia

# **Resumen**

En los últimos años se ha identificado una tendencia al decrecimiento de la tasa de natalidad en Colombia, como lo expresó el DANE en sus últimos informes (Estadística, 2024), donde se evidenció que entre 2022 y 2023, la cifra disminuyó en un 11%. Este escenario genera expectativa frente a la tendencia en el tiempo y las implicaciones a nivel económico y social.

Este proyecto propone predecir el comportamiento de la tasa de natalidad en Colombia para los próximos cinco años, mediante la analítica de datos y las herramientas que provee el machine learning. Se planea identificar las posibles causas que tienen mayor influencia en la variabilidad de la tasa de natalidad y determinar cuáles podrían influir en su posible disminución.

Lo anterior en conjunto con una segmentación demográfica, que permita la interacción entre la variable de interés y sus características, buscando entender comportamientos de la población relacionados con la natalidad. Los resultados obtenidos se analizarán con base en las tendencias generadas por los modelos. Se espera obtener como resultado una tendencia a la disminución en la tasa de natalidad en Colombia para los próximos cinco años.

Los resultados de este análisis servirán como base para predecir el comportamiento futuro de las tasas de natalidad en Colombia, entendiendo su relación con el sistema pensional; proporcionando así un recurso valioso para la formulación de políticas públicas o estrategias encaminadas a mitigar la influencia negativa en el sistema de pensiones.

Incluye al final de dicha página la dirección de los repositorios GitHub.

*Palabras clave*: *Tasa de natalidad, demografía, sistema pensional, Colombia, analítica, datos, predicción, clasificación.*

# **Abstract**

In recent years, a downward trend in the birth rate has been identified in Colombia, as expressed by DANE in its latest reports (Estadística, 2024), where it was evidenced that between 2022 and 2023, the figure decreased by 11%. This scenario generates expectations in the face of the trend over time and the implications at the economic and social level.

This project proposes to predict the behavior of the birth rate in Colombia for the next five years, through data analytics and the tools provided by machine learning. It is planned to identify the possible causes that have the greatest influence on the variability of the birth rate and to determine which ones could influence its possible decline.

This is done in conjunction with a demographic segmentation, which allows the interaction between the variable of interest and its characteristics, seeking to understand population behaviors related to the birth rate. The results obtained will be analyzed based on the trends generated by the models. The result is expected to be a downward trend in the birth rate in Colombia for the next five years.

The results of this analysis will serve as a basis for predicting the future behavior of birth rates in Colombia, understanding their relationship with the pension system; thus providing a valuable resource for the formulation of public policies or strategies aimed at mitigating the negative influence on the pension system.

*Keywords: Birth rate, demographics, pension system, Colombia, analytics, data, prediction, classification.*

# **Descripción del problema**

## **Problema de negocio**

La tasa de natalidad en Colombia ha tenido una tendencia a la disminución que ha oscilado entre el -1.0 % al -2.1 % ((DANE), 2022), en el último decenio. Este comportamiento tuvo un cambio significativo entre los años 2021 y 2023, con un descenso del -7.0 % y -11.0% (Estadística, 2024), respectivamente. Por lo tanto, se busca predecir el comportamiento de la tasa de natalidad en Colombia y las posibles afectaciones que podría generar en el contexto del sistema pensional del país, tal como se encuentra definido actualmente. Para este fin, se analizarán las bases de datos proporcionadas por el DANE sobre las estadísticas de nacimientos en Colombia en los últimos 5 años y las características demográficas de la población. Las métricas de desempeño podrían incluir la predicción de la tasa de natalidad para Colombia y la caracterización de la población mediante el uso de algoritmos de clasificación.

## **Aproximación desde la analítica de datos**

Con el fin de predecir el comportamiento de la variable de interés (tasa de natalidad en Colombia) se hará uso de técnicas de aprendizaje supervisado (IBM, 2024). Específicamente, algoritmos de regresión, utilizando la base de datos obtenida desde el DANE, con el fin de entrenar un modelo o modelos que se ajusten a las variables contenidas en este conjunto de datos. Con esta información, se realizará la predicción de la tasa de natalidad en Colombia para los próximos 5 años, teniendo en cuenta sus causas.

En paralelo, se implementará un algoritmo de aprendizaje no supervisado que permita clasificar las poblaciones con base en sus características demográficas (DataCamp, 2024). Se espera encontrar algún tipo de relación entre las causas de mayor influencia en la predicción del comportamiento a futuro de la tasa de natalidad en Colombia y así fortalecer el análisis predictivo, aportando información útil sobre los posibles focos de intervención a futuro, en relación con la influencia en el sistema pensional colombiano.

## **Origen de los datos**

Los datos utilizados son proporcionados por la plataforma de datos abiertos Colombia (Colombia, 2024), y corresponden a las estadísticas de orden nacional sobre los nacimientos, desde el año 2019. De manera específica la base de datos a analizar se denomina “Estadísticas Vitales – EEVV” (DANE, 2024), disponibles en formato csv. Esta información es publicada de forma anual y para el desarrollo del presente proyecto se tomarán 5 bases de datos, correspondientes a 5 años (2023, 2022, 2021, 2020, 2019). En ellas, se podrá encontrar en total 39 variables (DANE, microdatos.dane.gov.co, 2024).

## **Métricas de desempeño**

Se plantea de manera preliminar la implementación de las siguientes herramientas de análisis estadístico tanto para el modelo de clasificación como para el modelo de regresión:

En el análisis de un modelo de regresión, se realizará una revisión exhaustiva de la calidad del ajuste, utilizando el coeficiente de determinación, tanto R-cuadrado (R²) como R-cuadrado ajustado, esto servirá para evaluar el porcentaje de varianza explicada por el modelo. Además, se llevará a cabo un análisis de residuos, comprobando la normalidad, la homocedasticidad y la independencia de los residuos, lo que ayudará a verificar que los supuestos del modelo se cumplan.

Para verificar la significancia de los coeficientes, se plantea emplear pruebas t y se calcularan intervalos de confianza para cada coeficiente, con el fin de analizar si los parámetros estimados son significativamente diferentes de cero. Usando las métricas de error y precisión como el error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE) evaluaremos cuantitativamente la precisión del modelo. Además, por medio de la validación cruzada, específicamente K-fold cross-validation, se evaluará la capacidad de generalización del modelo. Finalmente, se realizará un análisis de sensibilidad y robustez para examinar la influencia de outliers y evaluar el poder predictivo del modelo utilizando diferentes subconjuntos de datos, asegurando así la estabilidad y fiabilidad del modelo en diversas condiciones.

Para el modelo de clasificación se realizarán evaluaciones de rendimiento que comprenderán la exactitud del modelo, junto con un análisis detallado de clasificación que abarca la precisión, la exhaustividad y la puntuación F1 score. Además, se empleará una representación gráfica de la matriz de confusión para ofrecer una visualización clara de cómo el modelo clasifica las poblaciones.

Cabe aclarar que en el desarrollo del proyecto planteado podrán agregarse o retirarse ciertas métricas como resultado del proceso investigativo.

# **Objetivos**

## **Objetivo general**

Realizar la predicción de la tasa de natalidad en Colombia para los próximos cinco años, por medio de la analítica de datos y las herramientas que provee el machine learning. Se planea identificar las causas que tienen mayor influencia en su variabilidad, determinando cuáles de ellas inciden en su posible disminución.

## **Objetivos específicos**

## Implementar modelos de regresión para predecir la tasa de natalidad en Colombia.

## Realizar una clasificación demográfica de la población que permita comprender la interacción entre la variable de interés, sus causas y características en común.

## Identificar las principales causas que influyen en la variabilidad de la tasa de natalidad en Colombia.

* Analizar con base en los resultados obtenidos la influencia en el sistema pensional colombiano.

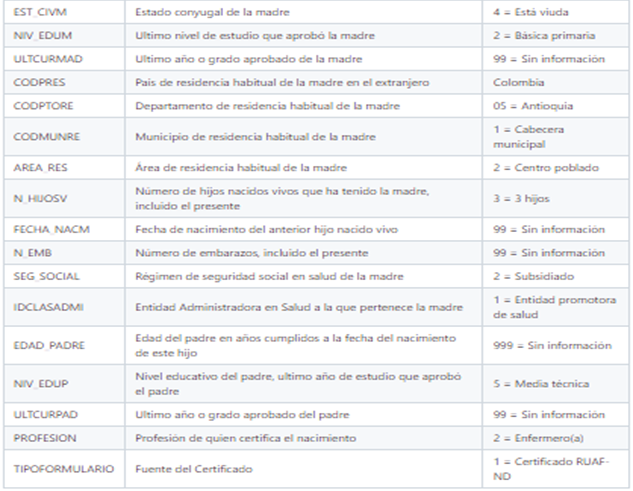
# **Datos**

## **Datos originales**

Como se mencionó previamente, el conjunto de datos original se extraerá de la plataforma pública de datos abiertos Colombia (Colombia, 2024), específicamente en el apartado de estadísticas de nacimientos publicada por el DANE.

A continuación, se ilustran en forma de tabla, las variables contenidas en el conjunto de datos. Además, se dan ejemplos del tipo de variables que contiene:





**Tabla 1.** Codificación de variables en base de datos original

Estas bases de datos se encuentran distribuidas por años, y cada año contiene en sus columnas la información anteriormente descrita. Para cada columna, el DANE estableció una serie de codificaciones con el fin de procesar los datos de forma numérica en los diferentes entornos de desarrollo, ya que muchos de los registros son categóricos de base y esto podría generar inconvenientes a la hora de procesarlos. A continuación, un ejemplo de estas codificaciones:

|  |  |  |
| --- | --- | --- |
| **Nivel educativo del padre** | **Multiplicidad del embarazo** | **Área de residencia habitual de la madre** |
| 1 = Preescolar | 1 = Simple | 1 = Cabecera municipal |
| 2 = Básica primaria | 2 = Doble | 2 = Centro poblado (inspección, corregimiento o caserío) |
| 3 = Básica secundaria | 3 = Triple | 3 = Rural disperso |
| 4 = Media académica o clásica | 4 = Cuádruple o más | 9 = Sin información |
| 5 = Media técnica  … | 9 = Sin información |  |

**Tabla 2.** Ejemplo de codificación de las variables.

Las bases de datos por año se encuentran en formato .csv y tienen el siguiente tamaño:

* **2019:** 62.207 KB
* **2020:** 59.739 KB
* **2021:** 97.122 KB
* **2022:** 90.318 KB
* **2023:** En espera de publicación en las bases de datos abiertas del DANE. Para efectos de cálculo de tamaño total se obtendrá un promedio con los años anteriores aquí descritos (87.347 KB).

En total, el tamaño del conjunto de datos a procesar suma 397.733 KB.

Por otro lado, se tienen la siguiente cantidad de datos:

* **2019:** 642,661 filas x 39 columnas.
* **2020:** 629,403 filas x 39 columnas.
* **2021:** 616,925 filas x 39 columnas.
* **2022:** 573,626 filas x 39 columnas.
* **2023:** 615,653 en promedio. Se está a espera de publicación en las bases de datos abiertas del DANE para tener el dato real.

En total, se tendrían 3.078.268 filas x 39 columnas. Debido al gran volumen de datos e información se estudiará la posibilidad de tomar una muestra representativa de cada año, con el fin de desarrollar el presente proyecto.

## **Datasets**

A partir de los datos originales y descritos previamente se construirán dos Datasets principales: un Dataset de entrenamiento para entrenar el o los modelos con el 70% de los datos originales y un Dataset de prueba que tendrá el 30% de los datos originales, estos serán utilizados para evaluar el modelo.

## **Analítica descriptiva**

Se realizará un análisis descriptivo previo que permita identificar de manera gráfica y estadística el comportamiento de los datos y su distribución, con el fin de evaluarlos en los modelos seleccionados, algunos de estos incluyen:

* Generación de histogramas y matriz de correlación.
* Clusterización por técnica de KNN.
* Generación de diagramas de dispersión.

A lo largo del proceso investigativo se fortalecerán las técnicas a implementar de acuerdo a las métricas planteadas previamente, con enfoque en el cumplimiento de los objetivos propuestos. Es de anotar, que las técnicas enunciadas a continuación ejemplifican el tipo de análisis que se obtendrá, estas son susceptibles de cambio a necesidad.

|  |
| --- |
| # Cargar datos con colab from google.colab import drive import xml.etree.ElementTree as ET from google.colab import auth from pydrive2.auth import GoogleAuth from pydrive2.drive import GoogleDrive from oauth2client.client import GoogleCredentials import os  # Tratamiento de datos import pandas as pd import numpy as np from sklearn.preprocessing import MinMaxScaler  #Gráficos import matplotlib.pyplot as plt from matplotlib import style import seaborn as sns  # Matemáticas y estadísticas import math  # Preparación de datos from sklearn.model\_selection import train\_test\_split from sklearn.preprocessing import normalize from sklearn.cluster import KMeans  # Configuración warnings # ============================================================================== import warnings warnings.filterwarnings('ignore')  #=========================================================================  # Generación de los DataSets  **X = nac\_1['NIV\_EDUM'] y = nac\_1['MUL\_PARTO']   X\_train, X\_test, y\_train, y\_test = train\_test\_split(  X,  y.values.reshape(-1,1),  train\_size = 0.7,  random\_state = 1234,  shuffle = True  )**  #=======================================================================  # Preparación de datos para modelo de regresión  # 8.1 Graficamos la frecuencia mediante un histograma  sns.histplot(x = nac['MUL\_PARTO'], stat = "count", binwidth = 0.6, binrange=(1, 9) ) sns.set\_style("darkgrid") plt.xlabel('Cantidad de nacimientos') plt.ylabel('Frecuencia de nacimientos (2022)') plt.title('Distribución de nacimientos en 2022') plt.show()  # 8.2 Generamos una matriz de correlación de las variables matrizCorr = nac\_1.drop(['NIV\_EDUM', 'MUL\_PARTO'], axis=1) sns.heatmap(matrizCorr.corr(), annot=True, annot\_kws={"size": 5}, cmap='coolwarm') plt.title('Matriz de Correlación') plt.show()  #======================================================================  # Modelo de Clusterización  # Análisis exploratorio de datos sns.scatterplot(x='NIV\_EDUM', y='MUL\_PARTO', data=nac\_1) plt.show()  # Preparación de datos nac\_1\_norm = normalize(nac\_1[['NIV\_EDUM', 'MUL\_PARTO']])  # Creación de datasets kmeans = KMeans(n\_clusters=5, random\_state=42) nac\_1['cluster'] = kmeans.fit\_predict(nac\_1\_norm)  # Análisis descriptivo avanzado sns.scatterplot(x= 'NIV\_EDUM', y='MUL\_PARTO', hue='cluster', data=nac\_1, palette='viridis') plt.show() |

# 

# 

# 

# 

# 

# **Referencias**

(DANE), D. A. (2022). *Principales resultados de Estadísticas Vitales nacimientos y defunciones para el cuarto trimestre de 2022, acumulado del año 2022 y el año corrido de 2023.* Bogotá, Colombia: DANE, Comunicado de prensa.

Colombia, D. a. (20 de 04 de 2024). *datos.gov.co*. Obtenido de datos.gov.co: https://www.datos.gov.co/

DANE. (20 de 05 de 2024). *datos.gov.co*. Obtenido de datos.gov.co: https://www.datos.gov.co/Estad-sticas-Nacionales/Estad-sticas-Vitales-EEVV/kk5w-ugzm/about\_data

DANE. (20 de 05 de 2024). *microdatos.dane.gov.co*. Obtenido de microdatos.dane.gov.co: https://microdatos.dane.gov.co/index.php/catalog/807/data-dictionary

DataCamp. (20 de 05 de 2024). *datacamp.com*. Obtenido de datacamp.com: https://www.datacamp.com/es/tutorial/k-nearest-neighbor-classification-scikit-learn

Estadística, D. A. (22 de 03 de 2024). *Boletín técnico: Estadísticas Vitales (EEVV), Nacimientos en Colombia.* Bogotá, Colombia: DANE. Recuperado el 01 de 05 de 2024, de https://www.dane.gov.co/files/operaciones/EEVV/bol-EEVV-Nacimientos-IVtrim2023.pdf

IBM. (2024). *www.ibm.com*. Obtenido de www.ibm.com: https://www.ibm.com/mx-es/topics/supervised-learning