Segmentación de clientes aplicado al análisis de reducción de tasas de natalidad en Colombia

Seminario de Analítica y Ciencia de Datos

# Resumen Descriptivo del Proyecto

En los últimos años se ha identificado una tendencia al decrecimiento de la tasa de natalidad en Colombia, como lo expresó el Departamento Administrativo Nacional de Estadística (DANE) en sus últimos informes (Citar), donde se registró una cifra de disminución en la natalidad del 11% entre el 2022 y el 2023 [1]. Este escenario genera expectativa frente a su comportamiento en el tiempo y sus implicaciones a nivel económico y social, dado el impacto que genera en el sistema pensional colombiano.

Este proyecto propone realizar la predicción en el comportamiento 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 la variabilidad de la misma, determinando cuáles de ellas influyen en su posible disminución. Todo esto, en conjunto con una segmentación demográfica de clientes que permita la interacción entre la variable de interés, sus causas y características. Los resultados obtenidos serán analizados de acuerdo a las tendencias generadas por los modelos programados.

Teniendo en cuenta las estadísticas presentadas por el DANE en sus últimos informes [1], se espera entonces obtener como resultado una tendencia a la baja, es decir, una disminución en la tasa de natalidad en Colombia para los próximos cinco años. Adicionalmente, se propone también establecer posibles características y comportamientos de la población relacionados con la natalidad, con la intención de determinar las causas principales de este declive.

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

1. Problema.
   1. ¿Cuál es mi problema alrededor de mi proyecto?

El decrecimiento de la tasa de natalidad en Colombia y como este impacta de manera negativa el sistema pensional del país, tal como se encuentra definido actualmente.

* 1. ¿Por qué se manifiesta el problema?

Podemos visualizar la manifestación de este problema a partir de los datos entregados por el DANE en comunicado de prensa del 24 abril del 2023, en los últimos 10 años la tasa de natalidad ha decrecido. Con un descenso sostenido que va de -2.1% a -1.0% entre los años 2018 y 2022. Este último, con una disminución histórica del -7.7% [1], si este comportamiento se sostiene a la baja en los años venideros, acarreará un sinnúmero de consecuencias negativas para el sistema pensional colombiano.

* 1. ¿De quién es el problema?

A este punto, se podría derivar la tenencia del problema en tres sujetos que exponen las siguientes perspectivas:

Sujeto A. Población colombiana en etapa productiva: Hace referencia a quienes se encuentran cotizando o aportando un porcentaje de su salario para disfrutar de su pensión en un futuro, este sujeto forma parte fundamental en el bono demográfico bajo el cual se concibe el sistema pensional colombiano en la actualidad [2]. La variabilidad en la tasa de natalidad supone riesgos para la obtención de una futura pensión o quizás un posible aumento en su aporte económico por un desbalance en el bono mencionado.

Sujeto B. Entes territoriales o instituciones de orden gubernamental: Se refiere aquellas entidades que tienen la responsabilidad de caracterizar y modificar el sistema pensional colombiano o de tomar decisiones encaminadas a la destinación de los recursos de pensión [3].

Sujeto C. Estamentos académicos o del sector empresarial: Se trata de aquellas personas interesadas en conocer el comportamiento de la tasa poblacional en el tiempo y que identifiquen un reto en el tratamiento estadístico de los datos entregados por el DANE, con el fin de utilizar metodologías de la analítica de datos aplicada a realizar predicciones y/o análisis estadísticos con diferentes enfoques.

1. Datos.
   1. Describa la fuente de datos (Sea más específico que en documento pasado):

La base de datos es proporcionada por la plataforma de datos abiertos Colombia [4], corresponde 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” [5].

Las bases de datos se encuentran publicadas 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, podemos encontrar en total 39 variables, entre las que se encuentran: Departamento de Nacimiento, Municipio de Nacimiento, Sitio de la Parto, Sexo del nacido vivo, Edad de la madre a la fecha del parto, Régimen de seguridad social en salud de la madre, Ultimo nivel de estudio que aprobó la madre, entre otras que serán tomadas en cuenta o descartadas para el desarrollo de este proyecto [5].

* 1. ¿Cuál es la calidad que posee dicha fuente?

Para el análisis de la calidad de la fuente de los datos se sometió la misma al checklist propuesto por la universidad de valencia para la evaluación de una fuente de información, quienes a su vez se basan en el libro *“Búsqueda y recuperación de la información”* [1][2]*.*

|  |  |  |
| --- | --- | --- |
| ***Ítem a evaluar*** | ***Cumple(Si/No)*** | ***Justificación*** |
| Relevancia | Si | Los datos aportan la información necesaria para hacer el análisis requerido. |
| Autoridad | Si | La fuente es confiable y está avalada por el gobierno nacional. |
| Actualidad | Si | La información se actualiza de forma anual y se tienen datos hasta el año 2022, estaría pendiente el 2023. |
| Exactitud | Si | En la descripción del DataSet está especificado el origen y la forma en que fue recolectada la información. |
| Consistencia | Si | Los datos son consistentes a lo largo del tiempo y entre diferentes conjuntos de datos del mismo tipo. |
| Objetividad | Si | Dado que son estadísticas de orden nacional, no se presenta ningún tipo de sesgo. |
| Accesibilidad | Si | El acceso es fácil y gratis. |
| Cobertura | Si | La fuente cubre todos los aspectos fundamentales para el análisis. |
| Documentación y soporte | Si | La documentación está detallada y con una amplia descripción. |
| Licenciamiento y restricciones de uso | No | No tiene restricciones de uso. |

***Tabla 1.*** *Criterios de evaluación de la calidad de la fuente de información.*

De acuerdo a lo anterior, se puede concluir que la fuente de datos es confiable. Por tanto, es pertinente para su uso en el proyecto planteado.

* 1. ¿Es accesible y disponible?

Si, se encuentra disponible y accesible en las bases de datos abiertas del gobierno nacional de Colombia.

1. Hipótesis
   1. ¿Qué es lo que pretende probar?

Que la tasa de natalidad seguirá una tendencia decreciente en los próximos cinco años y esto tendrá consecuencias negativas para el sistema pensional colombiano. Afectando el acceso al bono demográfico que subsidia gran parte de los grupos poblacionales que acceden a este beneficio en la actualidad [2]. Esto último, sujeto a interpretación en caso de que tal hipótesis sea comprobada.

* 1. ¿Cuáles pueden ser las respuestas esperadas a lo que pretende probar?

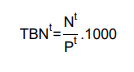
Se esperaría que se compruebe la hipótesis inicial, es decir, que la natalidad en Colombia disminuya. Sin embargo, es necesario considerar el escenario en el que por el contrario esta tasa, crezca.

En ambos casos valdría la pena analizar cómo esta predicción impactaría de manera positiva o negativa el sistema pensional colombiano. Todo esto, basado en la identificación previa de las causas y el análisis de las interacciones entre ellas y las características demográficas de la población, buscando ofrecer un soporte claro que permita focalizar las principales áreas de intervención.

* 1. ¿Qué acciones debo tomar para cada respuesta anterior?

Con el fin de probar si la hipótesis planteada se cumple, se obtendrá una muestra a partir de los datos entregados por el DANE para los últimos 5 años, donde se cuantifique el número de nacimientos y las variables que afectan o generan variabilidad en los mismos. Con esta información se buscará entrenar un modelo que permita predecir el comportamiento de los nacimientos en Colombia y cuáles son las características más relevantes para este análisis. Teniendo como supuesto o variable predictora, la disminución de los nacimientos en Colombia.

Posteriormente, esta información se deberá entregar en términos de tasa de natalidad, entiéndase este dato como “*el total de nacimientos de madre perteneciente a un determinado ámbito en el año t por cada 1.000 habitantes*” [2]. Utilizando la fórmula:



Con este cálculo se espera obtener información suficiente sobre la tendencia que tomaría la TBN. Obteniendo así, el punto de partida para el análisis investigativo sobre el impacto de este comportamiento en la estructura del sistema pensional de Colombia.

Adicionalmente, se aplicará la metodología de segmentación demográfica de clientes sobre la población objeto de estudio, con el fin de caracterizarla y realizar una identificación de las principales causas que ayudan a explicar la probable caída en la natalidad pronosticada mediante el modelo planteado.

1. Solución.
   1. Tipo. (¿Qué tipo de algoritmo de ML es? ¿Qué técnica es?

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 [3]. Específicamente, algoritmos de **regresión lineal**, utilizando la base de datos previamente obtenida desde el DANE, permita entrenar un modelo que se ajuste a dichos datos, prediciendo la tasa de natalidad en Colombia, desde sus causas.

En paralelo, se implementará un algoritmo de aprendizaje no supervisado, que permita **segmentar las poblaciones**, entiéndase como clientes para el algoritmo, con base en sus características demográficas [4]. Se espera encontrar algún tipo de relación con 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 sobre los posibles focos de intervención a futuro, en relación a su impacto en el sistema pensional colombiano.

* 1. ¿Cómo debe ser implementado?

Para el algoritmo de machine learning supervisado (**regresión lineal**), se seguirá la siguiente ruta de implementación:

* Definición y declaración de librerías de interés.
* Carga de la base de datos seleccionada.
* Limpieza de datos:
  + Se eliminan columnas que no son de interés para el análisis.
  + Se eliminan valores nulos, valores duplicados, se identifican variables categóricas no requeridas, entre otros.
* Visualización de datos:
  + Identificación de variables numéricas y categóricas e identificación de su aporte al análisis a partir de técnicas gráficas, tales como: histogramas, diagramas de cajas y bigotes, matrices y graficas de correlación de variables, entre otros.
  + Identificación de percentiles, media, mediana, moda y demás variables estadísticas de interés para el análisis.
* Preparación de datos:
  + Se identifican valores atípicos eliminando, imputando o modificando de acuerdo a su cantidad e influencia en el modelo.
  + Se analizan y realizan las imputaciones requeridas.
* Se crea y entrena el modelo a partir de la exploración de datos.
* Se realiza un diagnóstico, se prueba y se evalúan los resultados.

Para el algoritmo de machine learning supervisado (**segmentación demográfica de clientes**), se seguirá la siguiente ruta de implementación:

* Definición y declaración de librerías de interés.
* Carga de la base de datos seleccionada.
* Limpieza de datos:
  + Se eliminan columnas que no son de interés para el análisis.
  + Se eliminan valores nulos, valores duplicados, se identifican variables categóricas no requeridas, entre otros.
* Visualización de datos:
  + Identificación de variables numéricas y categóricas e identificación de su aporte al análisis a partir de técnicas gráficas, tales como: histogramas, diagramas de cajas y bigotes, matrices y graficas de correlación de variables, entre otros.
* Se analizan y realizan las imputaciones requeridas.
* Se entrena el modelo.
* Se obtención de los centroides y asignación de clusters
* Se realiza evaluación del modelo.
* Se realiza visualización de los clusters.
* Se Interpretación de los clusters.

Es de anotar, que estas metodologías de implementación están sujetas a cambios si en el desarrollo del presente proyecto se identifican técnicas o tareas que logren un mejor resultado o ajuste de los datos a los modelos planteados.

* 1. ¿Qué resultados se esperan?

A partir de los resultados cuantitativos obtenidos en el modelo de regresión lineal, se espera comprobar la hipótesis inicial, es decir, que la tasa de natalidad en Colombia tiende a disminuir en los próximos cinco años. Esto, acompañado del análisis de las causas que generan mayor cambio en dicha variable, con el fin de interpretarlas y analizarlas a la luz de su relación con el sistema pensional colombiano.

Del análisis de segmentación se clientes se espera generar agrupaciones poblacionales a partir de la relación entre sus características demográficas y el decrecimiento en la tasa de natalidad.

En resumen, al analizar los resultados derivados de la implementación de ambas técnicas de machine learning se espera comprobar tres supuestos:

* El comportamiento de las tasas de natalidad en Colombia en los próximos 5 años tiende a decrecer.
* Se generan al menos tres grupos, cada uno con una característica demográfica común, la cual se relaciona con una disminución en la tasa de natalidad en Colombia.
* Al menos dos de las variables independientes del con junto de datos se constituyen en las causas del decrecimiento de las tasas de natalidad en Colombia. Esto se identificará a partir de su influencia en la variabilidad del modelo de regresión lineal planteado.

1. KPI’s
   1. ¿Cómo espera evaluar el modelo?

La evaluación del modelo de regresión, implica examinar en primer lugar la precisión de las predicciones y su confiabilidad, esto es especialmente importante en contextos de planificación, donde las decisiones basadas en predicciones incorrectas pueden tener consecuencias significativas.

En segundo lugar, se examinará la relevancia de las variables, al evaluar la significancia estadística de cada coeficiente, se pueden determinar cuáles variables tienen un impacto real en la variable dependiente (tasas de natalidad en este caso). Esto ayudará a evitar que se tomen decisiones basadas en factores que en realidad no afectan el fenómeno estudiado.

En tercer lugar, se buscará identificar problemas de sobreajuste, para evitar que su rendimiento decaiga en caso de enfrentarse a nuevos datos. Por último, también se incluirá la validación de la robustez del modelo con el fin de comprobar si funciona en otros subconjuntos de datos y así asegurar que el modelo es generalizable y que no solo se ajusta para un subconjunto específico de datos [10].

* 1. ¿Qué métricas debe usar?

Se plantea de manera preliminar la implementación de las siguientes herramientas de análisis estadístico:

* Revisión de la calidad del ajuste: R-cuadrado (R²) y R-cuadrado ajustado.
* Análisis de residuos: normalidad de los residuos, homocedasticidad e independencia de los residuos.
* Verificación de la significancia de los coeficientes: pruebas t para coeficientes e intervalos de confianza para los coeficientes.
* Errores y métricas de precisión: MSE, RMSE y MAE.
* Validación Cruzada: K-fold cross-validation.
* Análisis de sensibilidad y robustez: Influencia de outliers y poder predictivo con diferentes subconjuntos de datos.

Es de aclarar que en el desarrollo del proyecto planteado podrán agregarse otro tipo de métricas como resultado del proceso investigativo.

1. Actores.
   1. ¿Quién es su cliente?

Entidades públicas, tales como, banco de la república, gobierno nacional, fondos de pensiones, academia y demás interesados en medir el comportamiento de la natalidad en Colombia con enfoques en su impacto en el modelo pensional colombiano.

* 1. ¿Quiénes son los interesados (Stakeholders)?

Entre ellos podemos encontrar, estudiantes, investigadores, analistas de mercado, analistas de datos de entidades públicas que deseen cimentar la toma de decisiones en relación al sistema pensional con un análisis confiable, metódico y bien estructurado basado en las tasas de natalidad en Colombia y la predicción de su comportamiento a futuro. También, estudiantes que consideren mejoras para los modelos planteados.

* 1. ¿Quién usará la solución?

Los sujetos descritos en el ítem 1. Es decir, quienes tienen el problema.

* 1. ¿A quién impactará?

De forma directa a las entidades encargadas de la distribución de recursos, reformas políticas y evaluación del sistema pensional colombiano. Toda vez que podrán contar con una base de información medible, analizada y cuantificada que apoye en la toma de decisiones relacionadas con el sistema pensional de país.

De forma indirecta a los ciudadanos de a pie que hacen parte del bono demográfico y que estén interesados en conocer el comportamiento del sistema pensional a partir de las tasas de natalidad, generando conciencia sobre el futuro de la sociedad siempre que su interpretación pueda enmarcarse alrededor de esta variable.

Bibliografía

|  |  |
| --- | --- |
| [1] | Departamento Administrativo Nacional de Estadística (DANE), «Principales resultados de las Estadísticas Vitales en nacimientos y defunciones para el cuarto trimestre de 2023pr año acumulado 2023pr (1 de enero al 31 de diciembre de 2023) y año corrido 2024pr (1 al 31 de enero de 2024),» DIAN , Bogotá , 2024. |
| [2] | Universidad externado de Colombia , «derlaboral.uexternado.edu.co,» Universidad externado de Colombia , 08 09 2020. [En línea]. Available: https://derlaboral.uexternado.edu.co/aspectos-teoricos-de-la-reforma-pensional/como-afectan-la-modificacion-demografica-y-el-mercado-laboral-al-sistema-pensional/. [Último acceso: 2024 03 02]. |
| [3] | O. A. O. Maldonado, «Administración del sistema colombiano de pensiones obligatorias,» *Studiositas,* nº ISSN 2215-728X, pp. 71-82, 2010. |
| [4] | Datos abiertos, «datos.gov.co,» [En línea]. Available: https://www.datos.gov.co/. [Último acceso: 02 04 2024]. |
| [5] | D. A. N. d. E. (DANE), «datos.gov.co,» DANE , 2019,2020,2021,2022,2023. [En línea]. Available: https://www.datos.gov.co/Estad-sticas-Nacionales/Estad-sticas-Vitales-EEVV/kk5w-ugzm/about\_data. [Último acceso: 24 04 2024]. |
| [6] | U. d. Valencia, «UV,» Secció d'Informació Bibliogràfica. Biblioteca Ciències Socials "Gregori Maians", [En línea]. Available: https://www.uv.es/cibisoc/tutoriales/trabajo\_social/431\_criterios\_de\_evaluacin.html. [Último acceso: 2024 04 24]. |
| [7] | N. P. M. M. Ferran Ferrer, «Búsqueda y recuperación de la información,» de *Búsqueda y recuperación de la información*, Barcelona , UOC, 2009. |
| [8] | IBM, «www.ibm.com/,» NA , [En línea]. Available: https://www.ibm.com/es-es/topics/unsupervised-learning. [Último acceso: 25 04 2024]. |
| [9] | IBM, «www.ibm.com,» NA, [En línea]. Available: https://www.ibm.com/es-es/topics/supervised-learning#:~:text=El%20aprendizaje%20supervisado%2C%20tambi%C3%A9n%20conocido,o%20prev%C3%A9n%20resultados%20con%20precisi%C3%B3n.. [Último acceso: 2024 04 24]. |
| [10] | T. I. b. t. A. o. L. R. Algorithm, «towardsdatascience.com,» MEDIUM, 05 07 2021. [En línea]. Available: https://towardsdatascience.com/linear-regression-assumptions-why-is-it-important-af28438a44a1. [Último acceso: 23 04 2024]. |

# Rúbricas

R01: Completitud y consistencia: el estudiante realiza una entrega del momento evaluativo con todos los elementos desarrollados, manteniendo una relación coherente entre los elementos solicitados. El estudiante usa una consistencia argumentativa que incluye el uso de referencias cuando hay mérito de autoría entre otros.

R02: Pertinencia: el proyecto del estudiante está dentro del ámbito de la analítica y la ciencia de datos y es evidente el desarrollo de cada uno de los elementos del Data Product Canvas

R03: Solución: el estudiante es capaz de responder a las preguntas del Data Product Canvas con la consistencia necesaria para relacionarla con el proyecto que desarrolla. Hay una relación evidente entre ellos. Las herramientas de MAchine Learning son una solución viable dentro de la descripción de la necesidad y objetivos del proyecto.

R04: Datos: los datos sugeridos son pertinentes al proyecto y al alcance delimitado en el momento evaluativo anterior.