UNIVERSIDAD & SOCIEDAD
APLICACIÓN

Fecha de presentación: abril, 2025
Fecha de aceptación: junio, 2025
Fecha de publicación: julio, 2025

# DE LA CIENCIA DE DATOS PARA PERFILAR Y PREDECIR LOS DIVORCIOS EN ECUADOR

# APPLICATION OF DATA SCIENCE TO PROFILE AND PREDICT DIVORCE IN ECUADOR

Liliana Rodríguez Villacís [1 *]
E-mail: lrodriguez@est.unibe.edu.ec
ORCID: https://orcid.org/0009-0006-9262-3740
Ximena Celi Celi [1]
E-mail: xceli@est.unibe.edu.ec
ORCID: https://orcid.org/0009-0002-9818-2238
Yasmany Fernández Fernández [1,2]
E-mail: yfernandez@doc.unibe.edu.ec, yfernandezf@upec.edu.ec
ORCID: https://orcid.org/0000-0002-9530-4028
[1]Universidad Iberoamericana del Ecuador. Ecuador.
[2]Universidad Politécnica Estatal del Carchi. Ecuador.
*Autor para correspondencia

Cita sugerida (APA, séptima edición)

Rodríguez Villacís, L., Celi Celi, X. & Fernández Fernández, Y. (2025). Aplicación de la Ciencia de Datos para perfilar y predecir los divorcios en Ecuador. *Universidad y Sociedad 17*(4). e5148.

## RESUMEN

Esta investigación combina técnicas de ciencia de datos, como el análisis de clústeres, y diseño de modelos predictivos, para comprender a profundidad los perfiles de divorcio en Ecuador y predecir su ocurrencia dentro de los primeros 15 años de matrimonio. Se utilizaron los datos históricos de divorcios publicados por el Instituto Ecuatoriano de Estadísticas y Censos (INEC), que incluyen variables sociodemográficas como edad al casarse, número de hijos, nivel de instrucción, nacionalidad y años de duración del matrimonio. Se aplicaron modelos supervisados y no supervisados. El primero permitió identificar patrones comunes en las parejas divorciadas, lo cual facultó una mejor comprensión de las variables que influyen en las decisiones de separación mediante el modelo DBSCAN que descubrió perfiles asociados a matrimonios de corta duración, parejas jóvenes y relaciones con baja descendencia. Luego se entrenó el modelo supervisado XGBoost para la predicción del divorcio con ≤ 15 años de matrimonio, el cual predijo que el grupo de mayor riesgo de divorcio se presenta en cónyuges con edades entre 25 y 40 años.

La investigación contribuye de manera significativamente al conocimiento de los divorcios en Ecuador, abordando información valiosa para la toma de decisiones en diversos ámbitos, como el diseño de políticas públicas, la intervención social y la prevención. Además, sirve de apoyo para asesorías personalizadas en instituciones y terapeutas, y orientación efectiva a parejas en riesgo. Finalmente, este artículo sentará las bases para futuros estudios en esta área, fomentando el uso de la ciencia de datos para abordar problemas sociales relevantes.

Palabras clave: Divorcio, Ciencia y tecnología, Análisis de datos, XGBoost, DBSCAN.

## ABSTRACT

This research combines data science techniques, such as cluster analysis, and predictive modeling to deeply understand divorce profiles in Ecuador and predict their occurrence within the first 15 years of marriage. Historical divorce data published by the Ecuadorian Institute of Statistics and Census (INEC) were used, which include sociodemographic variables such as age at marriage, number of children, educational level, nationality, and years of marriage.

Supervised and unsupervised models were applied. The former allowed for the identification of common patterns among divorced couples, which enabled a better understanding of the variables influencing separation decisions. The

DBSCAN model revealed profiles associated with short-term marriages, young couples, and relationships with few children. The supervised XGBoost model was then trained to predict divorce after ≤ 15 years of marriage, which predicted that the highest risk group for divorce is spouses between the ages of 25 and 40. This research contributes significantly to the understanding of divorce in Ecuador, providing valuable information for decision-making in various fields, such as public policy design, social intervention, and prevention. It can also support personalized counseling in institutions and therapists, and provide effective guidance to at-risk couples. Finally, this article will lay the groundwork for future studies in this area, promoting the use of data science to address relevant social problems.

Keywords: Divorce, Science and technology, Survey data, XGBoost, DBSCAN.

## INTRODUCCIÓN

La variación en las tasas de divorcio puede ser significativa en los distintos países y regiones del planeta, mientras que es común que las sociedades occidentales como las de los Estados Unidos o de Europa, presenten unas tasas de divorcio más elevadas que Asia y América Latina (Divorce Rates In The World [Updated 2024], 2024b), así, los países como Canadá, Rusia y Estados Unidos presentan las tasas de divorcio superiores, mientras que los países como Italia o Irlanda, tiene tasas de divorcio más bajas debido a las condiciones culturales y características religiosas (Longobardo, 2024). Al respecto la señala:

Las altas tasas de divorcios en los últimos años se pueden atribuir a una variedad de factores, incluidos los cambios en las normas sociales, las expectativas cambiantes en el matrimonio y las reformas legales que han hecho que el divorcio sea más accesible. Hoy en día, las personas con frecuencia se casan porque creen que han encontrado a su complemento, alguien que las ayudará a crecer personal y emocionalmente. Sin embargo, cuando la relación ya no cumple con esas expectativas, pueden sentirse atrapadas o insatisfechas, lo que conduce al divorcio. (Editorial Team, 2024).

En Ecuador las tasas de divorcio se han incrementado en el territorio ecuatoriano en forma gradual desde el año 1997 (Castelo-Cabay et al., 2021), pero en el 2023 presenta un descenso del 4.23%, lo cual puede indicar un cambio de la tendencia a la baja, o solo representar un periodo atípico al igual que el 2020 y 2021 de la época de pandemia por COVID-19.

Actualmente, existen diversas investigaciones que abordan el estudio del divorcio desde diferentes perspectivas. Sin embargo, la aplicación de la ciencia de datos para analizar este fenómeno en Ecuador es aún limitada. A nivel internacional, se han realizado algunos estudios que utilizan herramientas de software avanzadas, técnicas de minería de datos y aprendizaje automático para predecir divorcios o identificar factores de riesgo. No obstante, estos estudios suelen centrarse en contextos culturales y socioeconómicos diferentes, lo que limita su aplicabilidad a la realidad ecuatoriana.

El divorcio es la disolución del vínculo matrimonial y puede fundamentarse en múltiples causas, las cuales pueden ser percibidas como un factor negativo que atenta contra la institucionalidad de la familia, el matrimonio y la misma sociedad (Castrillón, 2021), pero en otros casos también es percibido como el "aumento de la felicidad tanto en hombres como mujeres, especialmente en las etapas posteriores al mismo" (Cavapozzi, 2020).

La ciencia de datos ofrece herramientas y técnicas eficaces para estudiar el divorcio desde diferentes ángulos. Al analizar grandes cantidades de datos y utilizar modelos de aprendizaje automático, se pueden identificar factores de riesgo, predecir divorcios, comprender tendencias y patrones, que conlleven a desarrollar intervenciones y políticas públicas más efectivas.

Desde 1976, el Instituto Nacional de Estadística y Censos – INEC, con la colaboración de las oficinas dependientes de la Dirección General de Registro Civil, viene procesando y publicando de manera anual y continúa la información de matrimonios y divorcios con una cobertura a nivel nacional (Bastidas, 2024). En las décadas anteriores al 2021, el promedio de divorcios aumenta de manera notable, provocando la disolución de hogares, perjuicios en la familia, especialmente en los hijos y en la sociedad en general.

En la siguiente tabla 1 se recopilan los artículos que realizan el análisis del divorcio aplicando análisis de datos y modelos de aprendizaje automático.

Tabla 1: Comparación de Artículos que realizan estudios del divorcio con ciencia de datos.

| Artículo | Año | Autores | Principales hallazgos | Técnicas utilizadas |
|---|---|---|---|---|
| Divorce Prediction with Machine Learning: Insights and LIME Interpretability (Ahsan, 2023) | 2023 | Md Manjurul Ahsan | Evaluó el "Conjunto de datos de predicción del divorcio" (desarrollado a partir de Escala de Predicción del Divorcio de Gottman) utilizando seis algoritmos: SVM, KNN y LDA alcanzaron una precisión del 98.57%. Además, se empleó LIME para explicar detalladamente las probabilidades de predicción, facilitando la diferenciación entre parejas casadas y divorciadas. | Máquinas de vectores de soporte (SVM), K vecinos más cercanos (KNN), análisis discriminante lineal (LDA), Explicaciones Locales Interpretables Independientes del Modelo (LIME) |
| Factores que inciden en los divorcios prematuros en el Ecuador: un modelo de regresión Logística (Castelo-Cabay et al., 2021) | 2021 | Castelo-Cabay, Carrillo-Patarón, & Dávalos-Castelo | Establecer los principales factores por las que los divorcios en el Ecuador se producen en un período menor o igual a 10 años, además de crear y evaluar un modelo de regresión logística que permita predecir si un matrimonio va a tener una duración de al menos una década. El modelo creado para la predicción obtuvo una capacidad de discriminación del 88% para lo que se utilizó la curva ROC como métrica de evaluación. | Modelo de regresión logística para predicción. Curva ROC (Características de funcionamiento del receptor) evalúa gráficamente la capacidad del modelo para discriminar clases binarias. |

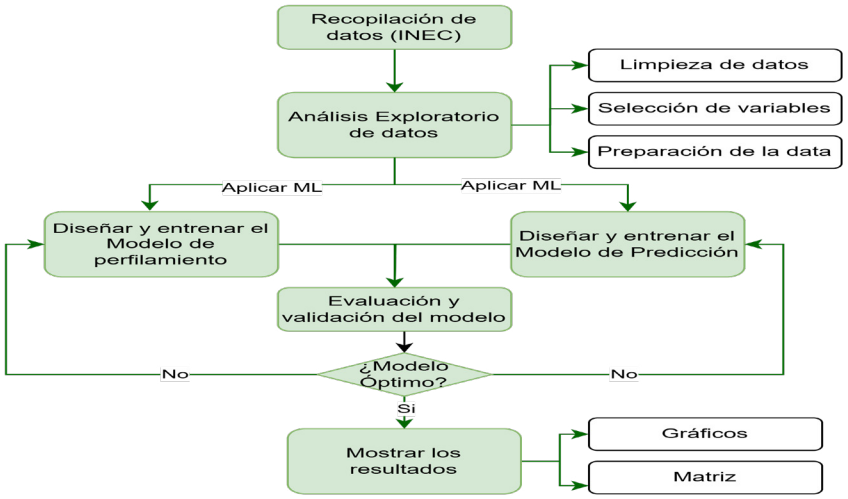
Fuente: Elaboración propia en base a Ahsan (2023); Castelo-Cabay et al. (2021).

Entender los patrones y características de los perfiles de divorcio es importante para desarrollar políticas, estrategias y programas de apoyo adecuados. El objetivo de este estudio es desarrollar un modelo de aprendizaje automático para el perfilamiento y predicción de divorcios en Ecuador, mediante la identificación y análisis de las características de los divorciados junto con la aplicación de modelos se realizan la clusterización (agrupamiento) y predicción de divorcio en base a variables específicas. En el análisis de datos y diseño de modelos de clasificación y predicción se seleccionan las variables clave como edad, nivel educativo, número de hijos, causas, sexo, duración del matrimonio, separación de bienes, y nacionalidad. Se aplican técnicas de análisis de clústeres como DBSCAN para agrupar los datos en perfiles homogéneos, PCA (Análisis de componentes principales) para optimizar la segmentación y gráficos de dispersión y radar para visualizar las diferencias entre los grupos, y XGBoost para la predicción.

## MATERIALES Y MÉTODOS

Se ha realizado un estudio de carácter exploratorio y cuantitativo, con un enfoque descriptivo-predictivo para entender la participación de las distintas variables del divorcio y ejecutar el pronóstico de divorciados. Con los datos del divorcio del INEC se analizaron las variables sociodemográficas como duración del matrimonio, edad, número de hijos, sexo, separación de bienes, nacionalidad y nivel de instrucción de los cónyuges. En la Figura. 1 representa la metodología general de investigación aplicada en este estudio.

Fig. 1: Metodología.




Fuente: Elaboración propia.

a) Recopilación de datos

Los datos provienen de los datos abiertos del portal del INEC, y corresponden a los registros de divorcios de los años 2021, 2022 y 2023. Se descartan los periodos anteriores, ya que en el 2020 hubo la pandemia de COVID-19 cuyas medidas de confinamiento modifican el comportamiento humano generando sesgos, anomalías y valores atípicos lo que produce resultados alejados de la realidad actual.

b) Análisis exploratorio

A la data de 50 variables y 24595 registros se realiza la limpieza y pre procesamiento de datos ejecutando tareas de transformación y formateo de variables (edad, nivel educativo, estado civil, fechas, etc.), manejo de valores nulos y atípicos, imputaciones, cálculo a través de las fechas, totales, categorización de datos cualitativos y una pre-selección de variables numéricas importantes. En el análisis exploratorio se utilizaron la diferentes funciones y librerías de Python, mapas de calor y gráficos estadísticos que permitieron observar el comportamiento entre las diferentes variables participantes.

Finalmente, la data preparada para el análisis consta de 22381 registros y 12 variables. El diccionario de datos de estas variables se detalla a continuación:

En las siguientes variables el sufijo "_1" refiere al género masculino y "_2" al femenino.

- 'edad_1', 'edad_2': Edad del divorciado o divorciada a la fecha de inscripción.
- 'hijos_1', 'hijos_2': Números de hijos a cargo del divorciado o divorciada en la inscripción.
- 'total_hijos': Variable del cálculo de hijos_1 e hijos_2.
- 'sexo_1', 'sexo_2': Sexo del divorciado o divorciada.
- 'separacion_bienes': Indica si registra capitulación de bienes.
- 'nacion_1', 'nacion_2': Indica si la nacionalidad es ecuatoriana o no.
- 'nivel_inst1', 'nivel_inst2': Nivel de instrucción del divorciado o divorciada.

c) Diseñar y entrenar modelos de aprendizaje automático

Con el fin de entrenar y probar el modelo de aprendizaje automático, se utilizó el 80% del conjunto de datos para el entrenamiento y el 20% del otro conjunto de datos para las pruebas. Se aplicaron varios modelos de clasificación y clusterización:

- Perfilamiento (Agrupamiento): K-Medias, DBSCAN (Agrupamiento espacial basado en densidad de aplicaciones con ruido) y GMM (Modelo de Mezcla Gaussiana), para lo cual se aplicó previamente la estandarización y la reducción de variables mediante el algoritmo PCA (Análisis de Componentes Principales). Con el objetivo de realizar un perfilamiento no supervisado de los matrimonios disueltos, se aplicaron estos algoritmos de clustering, utilizando variables numéricas y binarias relacionadas con edad, número total de hijos, duración del matrimonio y nacionalidad de los cónyuges.

- Predicción (Clasificación): Bosques aleatorios y XGBoost (Refuerzo de Gradientes Extremo), se agregó la variable de supervisión "duración" a todos los registros que tengan una duración de matrimonio igual o menor a 15 años se les dio el valor de 1 y a los matrimonios mayores a ese tiempo se les dio el valor de 0.

d) Evaluar y validar los modelos

Por cada uno de los modelos se aplicaron las métricas de evaluación válidas, generando los resultados mostrados en la Tabla 2. Hubo múltiples ejecuciones por modelo antes de obtener estos resultados mientras se recalibraba la selección de variables hasta obtener métricas adecuadas. La clasificación, se trabajó con las siguientes variables: 'edad_1', 'edad_2', 'hijos_1', 'hijos_2', 'total_hijos', 'nación_1', 'nación_2', 'nivel_edu1', 'nivel_edu2', 'divorcio', y para la agrupación: 'edad_1', 'edad_2', 'total_hijos' y 'nación_1', 'nación_2'.

Tabla 2: Métricas de modelos aplicados.

| Perfilamiento (Agrupamiento) | | | | Predicción (Clasificación) | |
|---|---|---|---|---|---|
| **Modelo** | **Puntuación de Silueta** | **Calinski-Harabasz** | **Davies-Bouldin** | **Modelo** | **Exactitud (Accurancy)** |
| K-Medias | 0.33 | 7227 | 1.31 | Bosques aleatorios | 0.82 |
| DBScan | 0.75 | 2745 | 1.20 | XGBoost | 0.83 |
| GMM | 0.74 | 4758 | 0.57 | | |

Fuente: Elaboración propia.

Los modelos DBSCAN y XGBoost generaron los mejores resultados para la clusterización (agrupamiento) y clasificación respectivamente.

**DBSCAN** (Agrupamiento espacial basado en densidad de aplicaciones con ruido), es un modelo automático no supervisado que agrupa instancias similares y resuelve problemas complejos para identificar grupos naturales (clústeres) basado en la densidad de los divorcios con características similares, sin necesidad de especificar previamente el número de clústeres, y que además identifica eficazmente los valores atípicos en un conjunto de datos.

Las variables fueron pre-escaladas y se empleó un PCA para la reducción de dimensión y una mejor separación de los grupos. Los parámetros particulares del algoritmo DBSCAN (eps=1.5, min_muestras=10) llevaron al algoritmo a crear grupos basados en densidad, con grandes requisitos de vecindario (1.5) así como al menos un número grande de puntos cercanos (min_muestras=10) para definir puntos núcleo del grupo. Los puntos que no cumplieron con estos criterios fueron considerados valores atípicos.

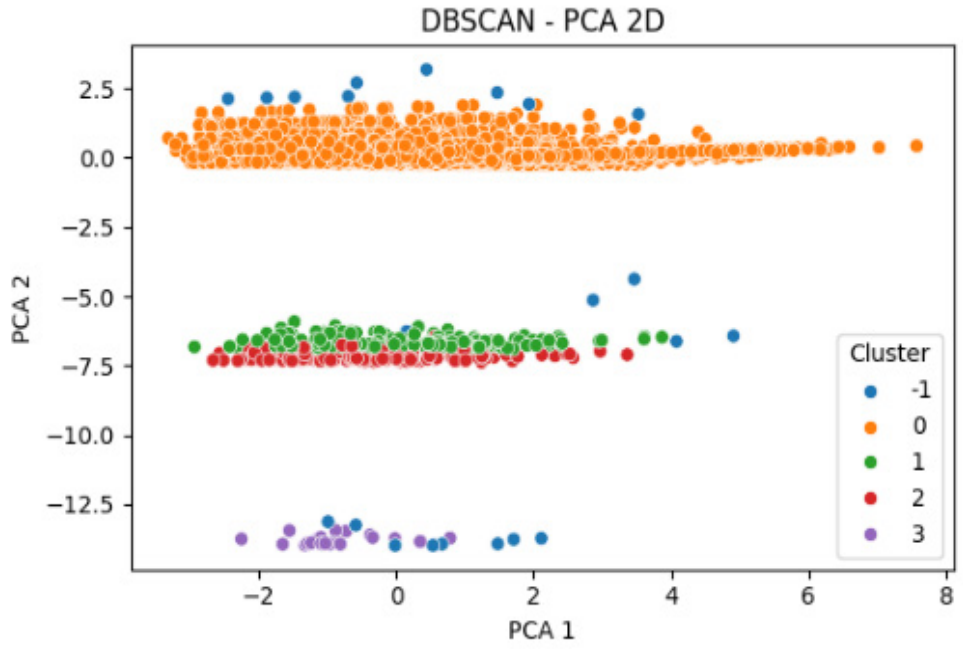
**XGBoost** (eXtreme Gradient Boosting), es un modelo de aprendizaje automático supervisado de gran eficiencia, precisión y escalabilidad que se compone de una secuencia de árboles de decisión (Zúñiga, 2020). Por lo cual este modelo permitió obtener las mejores métricas para predecir el divorcio dentro de los primeros 15 años utilizando las variables que más aportaban al modelo permitiendo cuantificar la importancia de cada una de las variables mediante la combinación jerárquica de árboles.

Para tener control sobre el entrenamiento del modelo XGBoost se desactivó el codificador automático de etiquetas asignando el valor de False a la variable use_label_encoder (use_label_encoder=False), y se configuró como métrica de rendimiento del modelo (eval_metric), la pérdida logarítmica (eval_metric='logloss'). Al ser un modelo de clasificación binaria la configuración, favoreció la predicción de la probabilidad de que la muestra pertenezca a la clase positiva (clase 1: divorcios dentro de los 15 años), y mide que tan lejos están esas probabilidades predichas del valor real (0 o 1). La métrica Logloss mide la diferencia entre las probabilidades predichas y las etiquetas reales. El modelo permitió obtener una precisión de 0.83 que significa que hace una buena predicción. En el artículo realizado por Castelo-Cabay et al. (2021), encontraron que había una gran probabilidad de divorcio antes de los 10 años, sin embargo, al realizar el presente análisis con los datos de los años 2021, 2022 y 2023 se encontró que hay una gran probabilidad de divorcio dentro de los 15 primeros años de matrimonio (0.83%).

e) Mostrar los resultados

La mejor forma de mostrar los resultados es a través de gráficos, para la clusterización se utilizaron gráficos de dispersión que permiten visualizar el agrupamiento de los clústeres formados después de la reducción a 2 dimensiones mediante PCA, en la Figura. 2 se pueden visualizar los cuatro clústeres generados y también el grupo con datos atípicos representados con el número -1.

Fig. 2: Gráfico de dispersión de Clústeres DBScan.

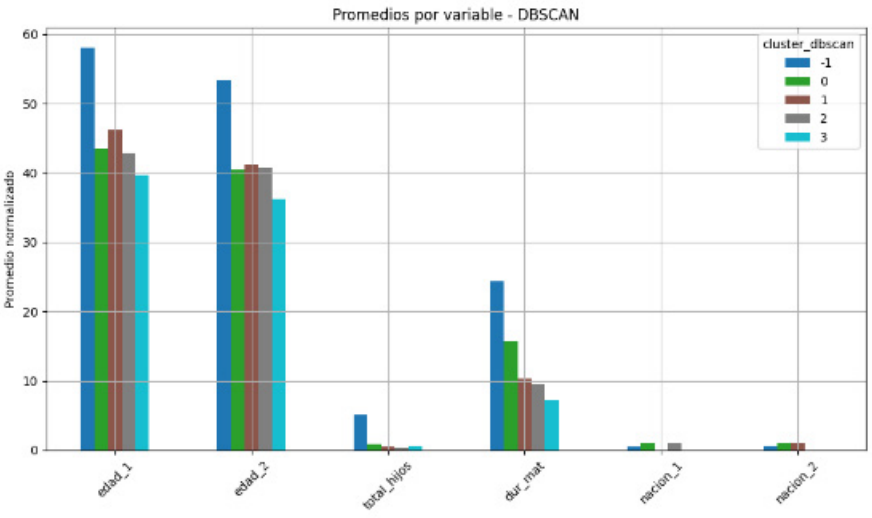



Fuente: Elaboración propia con librerías de Python.

También se realizaron gráficos de barra para la comparación del impacto de las variables en cada clúster, como se puede apreciar en el gráfico de la Figura. 3 en el eje de las X se tienen cada una de las variables que conforman el clúster y en el eje de las Y los promedios de valores de cada variable, como se muestra en la Figura. 3 cada grupo o clúster se representa de un color diferente.

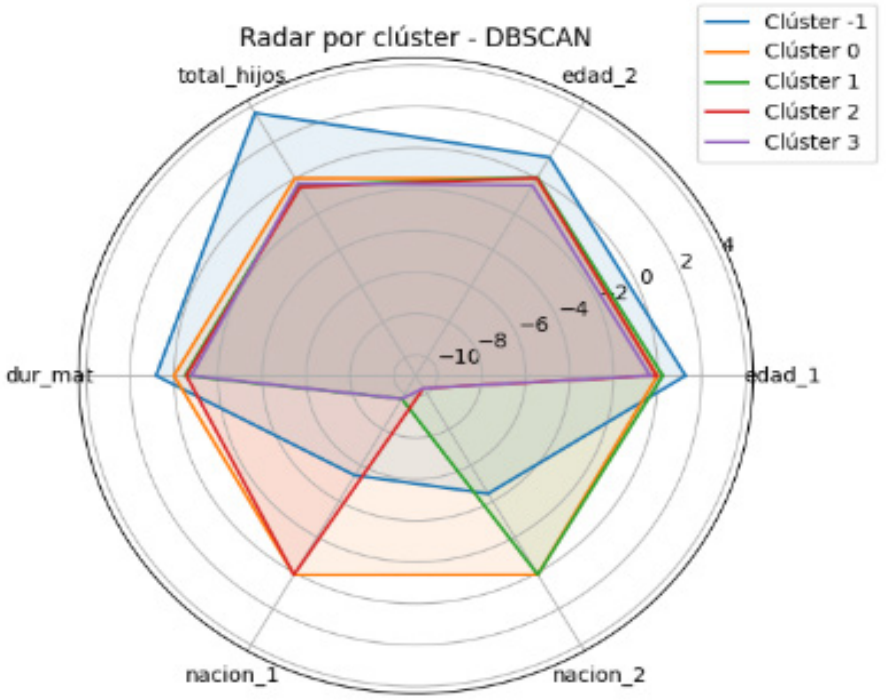
En la Figura. 4, se realizó el gráfico de radar para representar los clústeres sobre múltiples variables que representa la media de diferentes variables para cada clúster generado por DBSCAN, cada línea de color representa un clúster de acuerdo con la leyenda indicada en la parte superior derecha, muestra de manera visual las diferencias entre los clústeres identificados.

Fig. 3: Clústeres DBSCAN, gráfico de barra.




Fuente: Elaboración propia con librerías de Python.

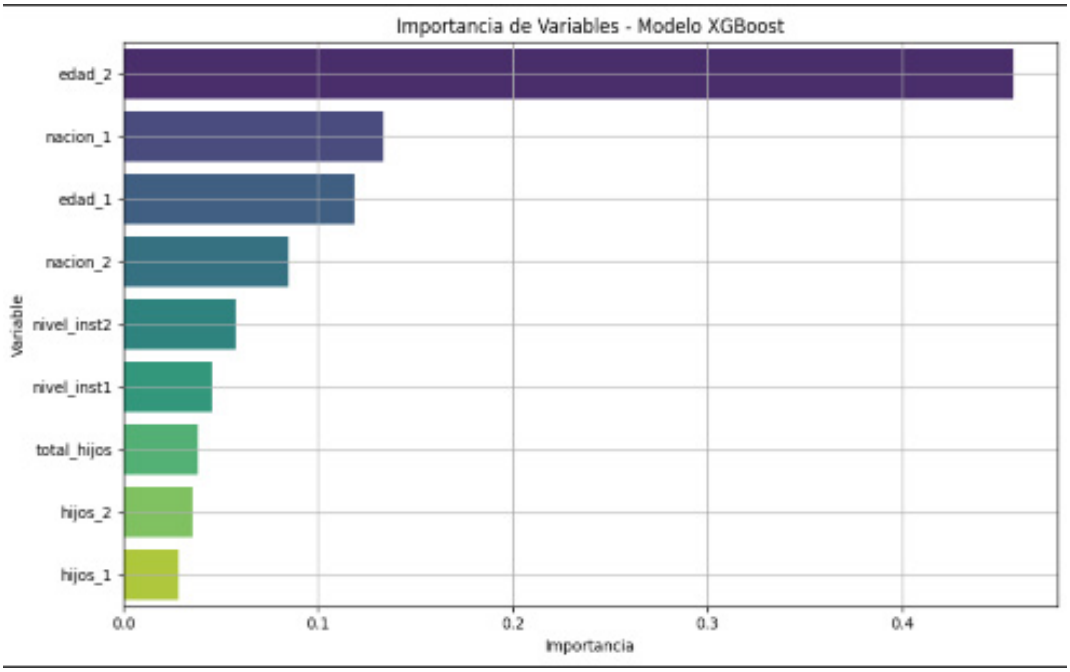
Fig. 4: Clústeres DBSCAN.




Fuente: Elaboración propia con librerías de Python.

Para la clasificación se construyó el gráfico de importancia de características en el proceso de predicción, mediante gráficos de barras. Cada barra muestra a que nivel aporta cada variable al modelo, a simple vista se puede ver que la variable “edad_2” (segundo cónyuge) es la más influyente de acuerdo con la Figura 5, esto quiere decir, que la edad_2 está estrechamente relacionada con la duración del matrimonio.

Fig. 5: Importancia de características, diagrama de barras.




Fuente: Elaboración propia con librerías de Python.

**Resultados del análisis del agrupamiento**

El modelo DBSCAN genera 4 clústeres, con un número de puntos clasificados como ruido (-1), lo que indica que algunos divorcios no se ajustan bien a los patrones identificados. Los clústeres formados se muestran en la Tabla 3.

Tabla 3: Perfilamiento por clúster (promedios).

| clúster | Descripción | edad_1 | edad_2 | total_hijos | dur_mat | nacion_1 | nacion_2 |
|---|---|---|---|---|---|---|---|
| -1 | Ruido | 58.04 | 53.39 | 5 | 24.39 | Extranjero/ Ecuatoriano | Ecuatoriana/ Extranjera |
| 0 | Relaciones duraderas | 43.49 | 40.50 | 1 | 15.67 | Ecuatoriano | Ecuatoriana |
| 1 | Relaciones intermedias | 46.22 | 41.19 | Pocos/Sin hijos | 10.36 | Extranjero | Ecuatoriana |
| 2 | Relaciones breves | 42.89 | 40.72 | Pocos/Sin hijos | 9.62 | Ecuatoriano | Extranjera |
| 3 | Relaciones más jóvenes y sin hijos | 39.62 | 36.29 | Pocos/Sin hijos | 7.24 | Extranjero | Extranjera |

Fuente: Elaboración propia obtenida con DBSCAN en Python.

- **Ruido**: Agrupa los casos que no encajan en ningún patrón denso, son todos los valores que quedaron por fuera de los parámetros establecidos (eps=1.5, min_samples=10), representa relaciones con características muy distintas: cónyuges mayores, más hijos y matrimonios mucho más duraderos, lo que sugiere la existencia de un grupo atípico o inusual en los datos.
- **Relaciones duraderas**: Matrimonios con edades promedio de 58 y 53 años, un alto número de hijos (5) y una larga duración promedio (24.39 años).
- **Relaciones intermedias**: Este clúster se caracteriza por matrimonios con edades de 46 y 41 años, con un número moderado de hijos. La duración promedio es de 10.36 años, indicando que son relaciones más efímeras que las del clúster 0, pero no necesariamente cortas.
- **Relaciones breves**: Agrupa matrimonios con edades de 43 y 40 años, con pocos hijos y una duración de 9.62 años, lo que refleja la tendencia hacia matrimonios más cortos.
- **Relaciones más jóvenes y sin hijos**: Este grupo tiene edades de 39 y 36 años, con un bajo número de hijos y una duración de 7.24 años, indicando matrimonios tempranos y breves.

Este hallazgo de grupos refleja el hecho de que las relaciones con más hijos y edades mayores están asociadas a matrimonios más duraderos, mientras que las relaciones con menos hijos y edades más jóvenes están asociadas con matrimonios más cortos. Esto puede ofrecer un ángulo diferente al abordar las razones y correlaciones del divorcio.

Las métricas de evaluación de la aplicación del modelo son:

- **Silhouette Score**: **0.754,** es un valor alto, que indica que el agrupamiento de DBSCAN ha creado clústeres compactos y bien separados, con buena cohesión interna y buena separación externa.
- **Calinski-Harabasz Score: 2745.38**, este resultado refuerza el resultado anterior de que los clústeres están bien compactos internamente y bien separados entre ellos.
- **Davies-Bouldin Score: 1.20**, indican que los clústeres tienen buena calidad general.

Dado que no se han encontrado estudios que apliquen DBSCAN al análisis de datos de divorcio, existe una oportunidad para las ciencias sociales computacionales. Por ejemplo, se podría aplicar DBSCAN a conjuntos de datos que contengan información sobre factores socioeconómicos, demográficos y conductuales de parejas, con el objetivo de identificar patrones o grupos con mayor propensión al divorcio. Esto podría proporcionar una perspectiva novedosa en la comprensión de las causas y factores asociados al divorcio.

**Resultados del análisis la predicción (clasificación)**

Con respecto a los matrimonios que terminan en divorcio antes de los 15 años de duración, el modelo XGBoost alcanza una tasa de precisión del 88% para la clase de divorcio, demostrando un fuerte poder predictivo del modelo cuando se encontraba un caso de divorcio. Los resultados detallados por clase fueron se resumen en la Tabla 4.

Tabla 4: Métricas de evaluación de la predicción XGBoost.

| Divorcio <=15 años | Precisión (accurancy) | Sensibilidad (recall) | Puntaje F1 (f1-score) | Número muestras (support) |
|---|---|---|---|---|
| 0 (No) | 0.77 | 0.86 | 0.81 | 2844 |
| 1 (Si) | 0.88 | 0.81 | 0.84 | 3871 |

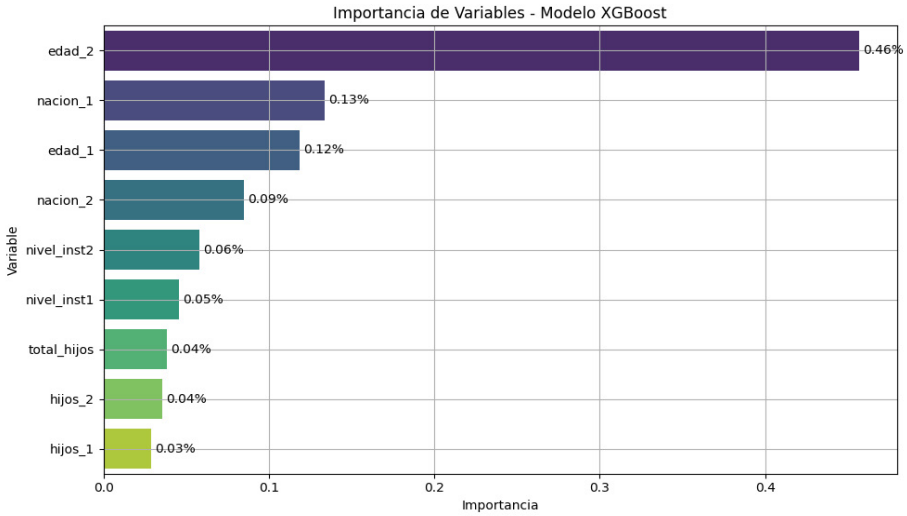
Fuente: Elaboración propia obtenida con XGBoost en Python.

La sensibilidad del 0.81 indica que logra identificar el 81% de los casos de divorcio ocurridos en ≤15 años, y con un Puntaje F1 promedio de 0.84 para los divorcios y 0.81 para la otra clase, lo cual indica un buen equilibrio entre precisión y sensibilidad en ambas clases. El total de hijos: indica que el número de hijos en común está fuertemente relacionado con la duración del matrimonio.

La edad (edad_1, edad_2) de los cónyuges al momento del divorcio. Hay que tener conciencia de que la finalización de los matrimonios puede seguir algún tipo de patrones con algunos puntos conjuntos que señalan que los divorcios a ciertas edades (por encima o por debajo del promedio) tienden a ser más o menos estables. Los niveles educativos de ambos cónyuges (nivel_inst1, nivel_inst2), sugieren que podría existir una asociación entre la educación y el riesgo de estar casado o no. Los resultados no solo proporcionan una predicción eficiente, sino también una explicación sobre los perfiles de estabilidad o disolución temprana de matrimonios.

Al analizar la importancia de las variables para comprender los factores más influyentes en la predicción se identificaron las siguientes: edad_2, nacion1, edad_1, nacion_2, nivel_inst2, nivel_inst1, total_hijos, hijos_1, como se puede visualizar en la Figura. 6.

Fig. 6: Importancia de las variables.



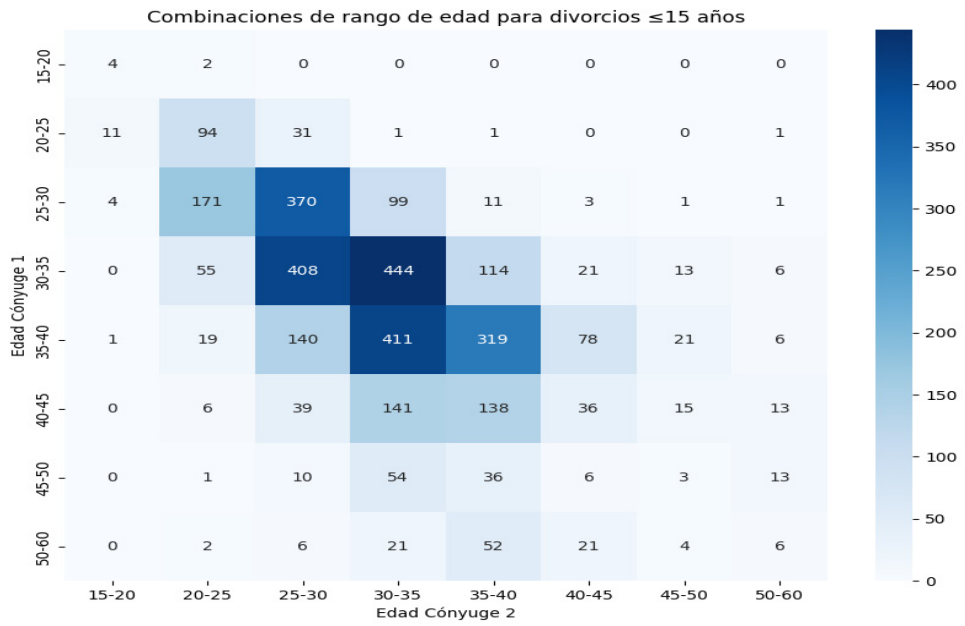

Fuente: Elaboración propia con librerías de Python.

En la Figura. 6 también se puede visualizar que el análisis de importancia de variables en el modelo XGBoost reveló que la edad de la cónyuge (edad_2) fue el predictor más influyente en la clasificación de divorcios ocurridos en los primeros 15 años de matrimonio, representando cerca del 46% de la importancia total del modelo, lo que podría estar relacionado con diferencias de madurez, estabilidad emocional o situación socioeconómica. Fue seguido por, la nacionalidad y la edad del cónyuge en términos de relevancia. Estos hallazgos indican que variables como la edad al momento del matrimonio y la nacionalidad de los cónyuges son macro-indicadores de la duración de la unión. Variables como el nivel de instrucción y la cantidad de hijos presentaron una importancia relativa menor, lo cual indica

que su influencia está mediada por otros factores más estructurales o contextuales, y concuerda con estudios que muestran que un mayor nivel educativo se asocia a una menor tasa de divorcio pues mayor planificación, independencia, estabilidad y tampoco el número de hijos es un factor determinante. A nivel internacional, un análisis de la revista Demographic Research muestra que, en países como Alemania, Noruega y Estados Unidos, las personas con mayor nivel educativo presentan menores tasas de disolución conyugal. Este patrón se ha acentuado en las últimas décadas, indicando una creciente desigualdad en la estabilidad matrimonial según el nivel educativo (Guetto et al., 2022).

En la Figura. 7 se aprecian en un mapa de calor las edades que presentan mayor incidencia en los divorcios dentro de los 15 años de matrimonio, donde las edades con mayor representatividad están entre los 25 y 40 años tanto en hombres como mujeres.

Fig. 7: Mapa de calor de edades en divorcio <= 15 años.




Fuente: Elaboración propia con librerías de Python.

A través de un análisis observatorio se obtiene el conjunto de las edades de ambos cónyuges en los casos de divorcios ocurridos dentro de los 15 años de matrimonio, y se identifican tres grupos de edad que concentraron la mayor proporción de la muestra que se resume en la Tabla 5.

Tabla 5: Rangos de edades la predicción XGBoost.

| Rangos de Edad | Total, de la muestra hombres | Total, de la muestra mujeres |
|---|---|---|
| (25, 30] | 660 | 1005 |
| (30, 35] | 1062 | 1175 |
| (35, 40] | 995 | 676 |

Fuente: Elaboración propia obtenida con XGBoost en Python.

Los resultados obtenidos indican que los divorcios de corta duración se concentran principalmente en individuos que, al momento de la separación, se encuentran en el rango de 25 a 40 años. Esta tendencia puede estar relacionada con factores como la consolidación de la carrera profesional, cambios en las expectativas de vida en pareja, y transiciones familiares, aspectos que futuros estudios cualitativos podrían explorar, ya que pueden aportar una comprensión más profunda sobre los motivos detrás de estas rupturas, explorando cómo la carrera profesional, las transiciones familiares y los cambios en las expectativas influyen en la estabilidad conyugal.

## CONCLUSIONES

Los algoritmos de agrupamiento no supervisado DBSCAN y de aprendizaje supervisado XGBoost se integraron para detectar patrones en la duración de los matrimonios, y se obtuvo una precisión general del 88% para las relaciones

cortas (≤15 años). Los resultados de la clusterización o agrupación revelaron diferentes perfiles de matrimonios, cuyos factores diferenciales son: el número de hijos, la edad de los cónyuges, nivel de instrucción, y duración del matrimonio. Los segmentos o clústeres son óptimos, ya que permiten identificar con mayor claridad los matrimonios que puedan ser sensibles a tener una mayor estabilidad con aquellos que podrían tener una duración más corta.

La aplicación de los modelos supervisado y no supervisado permitieron identificar los factores que afectan en los matrimonios con una duración corta o larga, estas variables son: edad, nivel de instrucción y nacionalidad principalmente. Estos hallazgos facultan la construcción de modelos más especializados y precisos que permitan dar atención a esta temática social y posibilitar la prevención de divorcios. A través del entrenamiento con el modelo de predicción XGBoost se obtuvo una precisión del 88% y una sensibilidad (recall) del 81% para predecir correctamente si un matrimonio termina en divorcio o no dentro de los primeros 15 años, lo que se traduce a un buen desempeño prediciendo relaciones cortas. Estos resultados evidencian que el perfil de divorcio temprano perfilado fue más fácil de identificar. El análisis de la importancia de variables permite interpretar la relación de éstas de manera transparente, destacando que factores como los hijos, la edad y el nivel educativo se vuelven fundamentales en la predicción de matrimonios duraderos. Variables como el sexo y la nacionalidad de los cónyuges estaban menos relacionadas con el modelo, indicando que los factores estructurales (es decir, edad, hijos, educación) ejercen una influencia más fuerte sobre la duración marital.

Los hallazgos del modelo XGBoost se compararon con los realizados usando agrupamiento DBSCAN y se observó un buen acuerdo entre las tendencias de edad en divorcios pre-15 años y subconjuntos de grupos de relaciones emergentes. XGBoost identificó como más representativos los rangos de edad entre 25 y 40 años, siendo más frecuentes entre mujeres. Por otro lado, el análisis de clústeres reveló agrupaciones que también presentan características asociadas a relaciones más frágiles o de menor duración. Particularmente, los clústeres 1, 2 y 3 corresponden a matrimonios con duraciones menores a 11 años, con bajo número de hijos, y edades promedio entre 39 y 46 años, alineándose con el fenómeno de disolución temprana detectado por el modelo supervisado. Este hallazgo refuerza la validez del perfilamiento realizado por XGBoost, al coincidir parcialmente con las tipologías emergentes en el análisis no supervisado.

Los hallazgos obtenidos mediante la aplicación de técnicas de análisis de datos en esta investigación pueden ser una herramienta válida para la planificación y diseño estrategias de prevención y fortalecimiento familiar, además de ser insumo para considerar en de políticas públicas, programas sociales relacionados con la pareja, y por supuesto, pueden ser usados en la predicción temprana del riesgo de divorcio. Las tipologías emergentes propician la identificación de etapas críticas en el ciclo de vida familiar que podrían tratarse mediante consejería profesional, educación emocional y programas segmentados por edad y etapa conyugal. El modelo puede integrarse en estadísticas nacionales para monitorear la salud del vínculo conyugal a lo largo del tiempo, los resultados permiten a gobiernos o ONGs priorizar recursos (psicólogos, talleres, charlas) en ciertos segmentos poblacionales con mayor riesgo de divorcio temprano, así como también promover educación relacional en colegios o universidades, para desarrollar habilidades emocionales y comunicativas en personas jóvenes antes del matrimonio.

## *REFERENCIAS BIBLIOGRÁFICAS*


Ahsan, M. M. (2023, 12 octubre). ***Divorce Prediction with Machine Learning: Insights and LIME Interpretability***. arXiv.org. https://arxiv.org/abs/2310.08620

Bastidas, C. (2024). ***Registro Estadístico de Divorcios-2023 - Datos abiertos Ecuador.*** https://datosabiertos.gob.ec/dataset/registro-estadistico-de-divorcios-2023

Castelo-Cabay, M. J., Carrillo-Patarón, M. E., & Dávalos-Castelo, M. A. (2021). Factores que inciden en los divorcios prematuros en el Ecuador: un modelo de regresión logística***. Polo del Conocimiento***, 6(3), 275–296. https://doi.org/10.23857/pc.v6i3.2464

Castrillón, O. D. (2021). Predicción del divorcio por medio de técnicas inteligentes. ***Información Tecnológica***, ***32***(5), 111-120. https://doi.org/10.4067/s0718-07642021000500111

Cavapozzi, D., Fiore, S., & Pasini, G. (2019). Divorce and well-being. Disentangling the role of stress and socio-economic status. ***The Journal of the Economics of Ageing, 16,*** 100212. https://doi.org/10.1016/j.jeoa.2019.100212

Divorce Rates in the World [Updated 2024]. (2024b, julio 15). https://divorce.com/blog/divorce-rates-in-the-world/

Editorial Team. (2024). ***Surprising Increase in Divorce Rates (1,232% increase)***. The Smart Divorce Coach, Advisor, And Transitionist. https://thesmartdivorce.com/high-divorce-rate#toc_Why_Were_Divorce_Rates_So_Low_in_the

Guetto, R., Bernardi, F., & Zanasi, F. (2022). Parental education, divorce, and children's educational attainment: Evidence from a comparative analysis. ***Demographic Research***, ***46***, 65-96. https://doi.org/10.4054/demres.2022.46.3

Longobardo, J. P. (2024). ***Porcentaje de divorcios en el mundo: cifras definitivas***. Abogado Divorcio Barcelona. https://abogadodivorciobarcelona.cat/porcentaje-de-divorcios-en-el-mundo-cifras-definitivas

Zúñiga, J. J. E. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. ***Ingeniería Investigación y Tecnología***, ***21***(3), 1-16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022