Análisis de Emigración Amazónica en Ecuador: Clasificación, Clustering y Reglas de Asociación

Fabian Ledesma
Universidad de Cuenca
Cuenca, Ecuador
andres.ledesma@ucuenca.edu.ec

Bryan Mendoza
Universidad de Cuenca
Cuenca, Ecuador
bsteven.mendoza7@ucuenca.edu.ec

***Abstract*—La emigración segregada por regiones en Ecuador, específicamente en la Amazonía, es un fenómeno poco estudiado en cuanto a sus características y factores subyacentes. Este estudio utiliza datos del Censo 2022 de Ecuador para analizar y caracterizar el comportamiento migratorio de los habitantes de la región amazónica. Se aplican técnicas de inteligencia artificial, como KNN y MLP, para identificar las zonas de residencia de los emigrantes antes de su salida del país con altas precisiones de clasificación, así como K-means para identificar grupos representativos de emigrantes amazónicos. Finalmente, se han identificado patrones clave en la relación entre variables sociodemográficas de los emigrantes.**

***Keywords*—Emigración Amazonía, Clasificación, Clustering, Reglas de asociación, Censo Ecuador 2022.**

I. Introducción

La migración es un fenómeno global con múltiples implicaciones económicas, sociales y culturales. En Ecuador, la emigración ha ido en aumento en las últimas décadas debido a diversos factores, como problemas sociales, económicos y de salud. Sin embargo, los patrones migratorios específicos de la población amazónica han sido poco estudiados, a pesar de su relevancia para el desarrollo de políticas públicas más efectivas [1].

Este estudio tiene como objetivo analizar y caracterizar el comportamiento migratorio de los habitantes de la Amazonía ecuatoriana utilizando datos del Censo 2022 y técnicas de inteligencia artificial. En particular, se busca identificar patrones migratorios, segmentar a los emigrantes en grupos representativos y analizar la relación entre variables sociodemográficas y la decisión de emigrar. Para ello, se emplearon métodos de aprendizaje supervisado, como K-Nearest Neighbors (KNN) y Multilayer Perceptron (MLP) para la clasificación, así como técnicas de aprendizaje no supervisado, como K-Means para clustering y reglas de asociación para descubrir patrones en los datos.

El resto de este artículo se organiza de la siguiente manera: la Sección 2 describe brevemente el conjunto de datos utilizado. La Sección 3 describe la metodología empleada para el preprocesamiento de los datos, así como para las tareas de clasificación, clustering y extracción de reglas de asociación. A continuación, la Sección 4 analiza los resultados obtenidos. Finalmente, la Sección 5 expone las conclusiones de esta investigación.

II. Descripción del conjunto de datos

El conjunto de datos utilizado fue extraído de la página oficial del Instituto Nacional de Estadística y Censos (INEC) correspondiente al último censo de población y vivienda de Ecuador, realizado en el año 2022, con segregación cantonal[2]. Este conjunto de datos incluye los siguientes archivos: el conjunto de datos de población, el de mortalidad, el de emigración, el de hogar y el de vivienda. En la Tabla I se presenta la información general de cada uno de los conjuntos de datos mencionados.

Dataset	Núm variables	Tipo variables	Núm registros
Población	93	Categóricas, enteras	16.938.986
Emigración	14	Categóricas, enteras	124.992
Mortalidad	16	Categóricas, enteras	250.746
Hogar	44	Categóricas, enteras	5.193.548
Vivienda	32	Categóricas, enteras	6.611.555

TABLE I: Descripción de los conjuntos de datos.

- **Conjunto de datos emigración**
  Está compuesto por información relacionada con los emigrantes, como el año de salida (”E01”), el sexo (”E02”) y el país de residencia actual (”E04”), entre otras variables incluidas en cada instancia del conjunto de datos. Además, se incorporan variables destinadas a la vinculación con otros conjuntos de datos, como el identificador del hogar (”ID_HOG”) y el identificador de la vivienda (”ID_VIV”).

- **Conjunto de datos vivienda**
  Está conformado principalmente por información sobre los aspectos físicos y la ubicación de este espacio. Entre las variables incluidas se encuentran el número de cuartos (”V15”), el estado del piso (”V08”), el material predominante del piso (”V07”) y la provincia en la que se ubica la vivienda (”I01”). Además, se incorporan variables destinadas a la vinculación con otros conjuntos de datos, como el identificador de la vivienda (”ID_VIV”).

- **Conjunto de datos hogar**
  Está compuesto por diversas características, como la disponibilidad de servicios (”H03”), el número de dormitorios (”H01”), el tipo de agua de consumo del hogar (”H06”) y la disponibilidad de internet (”H1004”), entre otras variables que describen las condiciones del hogar. Además, se incluyen variables destinadas a la vinculación con otros conjuntos de datos, como el identificador del hogar (”ID_HOG”) y el identificador de la vivienda (”ID_VIV”).

- **Conjunto de datos mortalidad**
  Describe instancias de personas fallecidas en Ecuador desde el año 2020 hasta la fecha del censo. Entre las características recopiladas se encuentran el mes de fallecimiento (”M0201”), el año de fallecimiento (”M0202”) y la edad al fallecer (”M03”), entre otras variables registradas en este conjunto de datos. Además, se incluyen variables destinadas a la vinculación con otros conjuntos de datos, como el identificador del hogar (”ID_HOG”) y el identificador de la vivienda (”ID_VIV”).

- **Conjunto de datos población**
  Describe información sobre las personas residentes en Ecuador. Entre las características registradas se incluyen el sexo al nacer (”P02”), la posesión de un documento de identidad (”P0602”) y la provincia de nacimiento (”P08”), entre otras variables. Este conjunto de datos es el más extenso en términos de variables, con un total de 93 en la composición de cada instancia. Además, se incorporan variables destinadas a la vinculación con otros conjuntos de datos, como el identificador del hogar (”ID_HOG”) y el identificador de la vivienda (”ID_VIV”).

Una vez descritos cada uno de los conjuntos de datos, los conjuntos requeridos para este estudio son los de vivienda, hogar y emigración. En contraste, los conjuntos de datos de población y mortalidad no resultan adecuados para establecer una correspondencia con los previamente mencionados. El conjunto de datos de mortalidad identifica a personas fallecidas, mientras que el de población hace referencia a personas residentes en Ecuador.

III. Metodología

A. Preprocesamiento de datos

1) **Segregación de instancias**
   Se seleccionó únicamente las instancias correspondientes a las provincias de la Amazonía, utilizando como criterio el identificador ”I01 (código de provincia)”, una variable presente en los tres conjuntos de datos (emigración, vivienda y hogar).
   Posteriormente, se realizó la asociación entre los conjuntos de datos. En primer lugar, se estableció la relación entre los conjuntos de datos de hogar y vivienda a través de la variable ”ID_VIV” (identificador de vivienda), lo que permitió asociar cada hogar a su respectivo espacio físico, es decir, una vivienda. En segundo lugar, se efectuó la asociación entre el conjunto de datos resultante y el conjunto de datos de emigración mediante la variable ”ID_HOG”, lo que permitió atribuir a cada instancia correspondiente a un emigrante las características del hogar al que pertenece.

2) **Eliminar columnas repetidas**
   Se eliminó las columnas repetidas, ya que, al asociar los conjuntos de datos, se identificaron variables duplicadas. Un ejemplo de ello es la variable ”I01”, presente en cada uno de los conjuntos de datos. Como siguiente paso, se eliminó las variables que correspondían a la concatenación de otras variables ya existentes. Por ejemplo, la variable ”CANTON” resultaba de la concatenación de las variables ”I01” e ”I02”; la variable ”ID_VIV” de la combinación de ”I01”, ”I02” e ”I10”; la variable ”ID_HOG” de ”I01”, ”I02”, ”I10” e ”INH”; y la variable ”ID_EMI” de ”I01”, ”I02”, ”I10”, ”INH” y ”E00”. Adicionalmente, se eliminó la variable ”V0202”, debido a que contenía exclusivamente valores faltantes, y la variable ”H1101” (Número de personas fallecidas por hogar), que presentaba un **94.23%** de valores faltantes.

3) **Tratamiento de ruido**
   Se llevó a cabo un tratamiento de ruido en las siguientes variables: ”H11” (Presencia de una persona fallecida a partir de enero de 2020), ”H12” (Alguna persona viajó a otro país y no regresó a partir de noviembre de 2010), ”H15” (Presencia de una persona no mencionada en el hogar), ”E01” (Año de salida), ”E03” (Edad al salir) y ”E04” (País actual de residencia). Estas variables comparten la particularidad de admitir respuestas como 9, 999 o 9999, las cuales, según el diccionario de variables [3], indican que la respuesta se ignora. Estos valores fueron reemplazados por valores faltantes y, posteriormente, se aplicó el imputador ”KNNImputer” con una configuración de $k\_vecinos\_cercanos = 5$.

4) **Tratamiento de valores faltantes**
   Para este proceso, se utilizó la biblioteca sklearn, empleando el imputador ”KNNImputer” con una configuración de $k\_vecinos\_cercanos = 5$. Las variables tratadas fueron ”H0801N” (Número de perros por hogar), que presentaba un 29.62% de valores faltantes, y ”H0802N” (Número de gatos por hogar), con un 60.02% de valores faltantes.

5) **Tratamiento de datos atípicos**
   Por último, las variables categóricas no fueron modificadas. En el caso de las variables de tipo entero, se identificaron valores atípicos, como se muestra en la Imagen 1; sin embargo, no se aplicó ninguna corrección, ya que el diccionario de variables del Censo Ecuador 2022 [3] establece rangos numéricos dentro de los cuales deben situarse los valores de dichas variables. Por ejemplo,

la variable ”E03”, que describe la edad de salida del emigrante, admite valores entre 0 y 98 años, además del valor 999, que indica que el dato es desconocido. Por lo tanto, los valores observados se encontraban dentro de los rangos especificados.

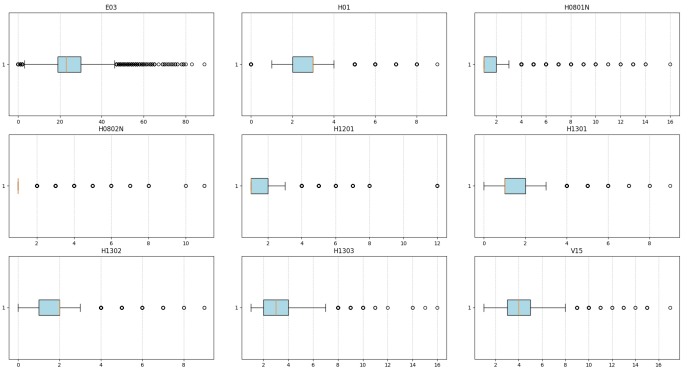

Fig. 1: Outliers presentes en las variables.

B. Clasificación

Antes de aplicar los modelos de clasificación, se establecieron los criterios de selección de variables, la configuración de los algoritmos y la división del conjunto de datos en entrenamiento y prueba. A continuación, se detallan los procedimientos seguidos en esta etapa.

- **Conjunto de datos**
  El conjunto de datos utilizado corresponde al dataset preprocesado en la sección anterior.

- **Etiqueta a predecir**
  La variable a clasificar es ”AUR”, la cual indica la zona de ubicación de la vivienda del emigrante, especificando si esta se encuentra en un área urbana o rural.

- **Algoritmo K Nearest Neighbors(KNN)**
  Para la implementación del modelo KNN, se desarrolló un script en Python utilizando la librería sklearn, específicamente el método KNeighborsClassifier. Se configuró el modelo con k = 1 vecinos cercanos, siguiendo un enfoque basado en la distancia euclidiana para la asignación de clases.

- **Multiple Layer Perceptron(MLP)**
  Se implementó una red neuronal tipo Multi-Layer Perceptron (MLP) empleando la librería sklearn, mediante el método MLPClassifier. La configuración del modelo incluyó dos capas ocultas, cada una con 50 neuronas, y un máximo de 500 iteraciones para la convergencia del algoritmo de optimización.

- **División del conjunto de datos**
  El conjunto de datos se dividió en dos subconjuntos:
  - **Conjunto de entrenamiento:** 80% de las instancias totales.
  - **Conjunto de prueba:** 20% de las instancias totales.

Cada una de las configuraciones mencionadas fue aplicada a tres escenarios distintos:

1) **Sin modificaciones:** Los modelos fueron entrenados directamente, es decir, utilizando los datos sin realizar ninguna transformación previa.

2) **Con escalamiento de datos**
   Se aplicó un proceso de normalización a las variables.

3) **Reducción de dimensionalidad**
   Se realizó un Análisis de Componentes Principales (PCA) para identificar un subconjunto de componentes que explicaran el 90% de la varianza. Como resultado, la dimensionalidad del conjunto de datos se redujo de 69 variables a 42 componentes, tal como se observa en la Figura 2.

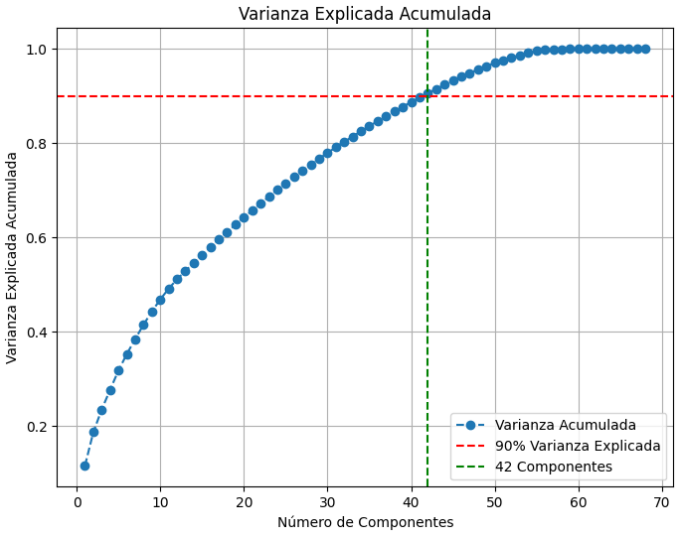

Fig. 2: Varianza explicada por cada número de componentes

Por último, se utilizó la métrica de **accuracy** como métrica global del rendimiento de cada modelo, dado que el dataset, en relación con la variable de clasificación (zona de ubicación de la vivienda del emigrante), esta equilibrado, con un 55.85% de instancias correspondientes a la zona urbana y un 44.15% a la zona rural. Adicionalmente, se emplearon las métricas: precisión, recall y F1 score.

C. Clustering

Se usó un dataset diferente y dos modelos de clustering, los cuales se detallan a continuación.

- **Conjunto de datos**
  El conjunto de datos utilizado corresponde solo al dataset de emigración, el cual fue tratado de la misma manera que se describe en la sección de preprocesamiento.

Este conjunto de datos contiene únicamente información altamente relacionada con la emigración de cada persona, conteniendo 10 variables.

- **K-Means**
  Para este algoritmo, primero se determinó el número óptimo de clusters mediante el cálculo de métricas como el *método del codo* y el *silhouette score*. Una vez determinado el número de clusters (valor de k), se procedió a entrenar el modelo con los datos procesados y normalizados. Los resultados obtenidos (ver Sección IV) permitieron identificar agrupaciones significativas dentro del conjunto de datos, lo que facilitó el análisis de los diferentes perfiles de emigrantes amazónicos y proporcionó información valiosa para la interpretación posterior.

- **Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)**
  El algoritmo HDBSCAN se basa en la estimación de densidades locales para identificar clústeres y se destaca por su capacidad para manejar eficazmente el ruido en los datos. No obstante, al aplicarlo a este conjunto de datos, no se obtuvieron grupos de emigrantes claramente diferenciados. Se realizaron diversas pruebas ajustando hiperparámetros clave, como el tamaño mínimo de clúster (min cluster size) y el tamaño mínimo de muestra (min samples), con el objetivo de identificar configuraciones que permitieran la formación de clústeres coherentes. A pesar de estos ajustes, HDBSCAN no logró detectar clústeres significativos. Esto se debe a la naturaleza de los datos, cuya distribución de densidad no favorece la creación de agrupaciones definidas, limitando así la efectividad del algoritmo en este contexto.

D. Reglas de asociación

En esta sección, se empleó el software WEKA para la generación de reglas de asociación a partir del conjunto de datos preprocesado anteriormente. A continuación, se detalla la configuración utilizada en el proceso.

- **Conjunto de datos**
  El conjunto de datos utilizado corresponde al dataset preprocesado en la sección anterior.

- **Configuración WEKA**
  En el módulo ”Associate” de WEKA, se seleccionó el método ”Asociador por filtro”, debido a la necesidad de aplicar un filtro de preprocesamiento al conjunto de datos. Este filtro convirtió exclusivamente las variables enteras a nominales, permitiendo la correcta aplicación de los algoritmos de asociación.
  Se configuraron los siguientes parámetros clave:
  - **Métrica mínima:** Se estableció un umbral de 0.9, lo que significa que la confianza de las reglas generadas debía ser, al menos, del 90%. Con esta configuración se reduce la generación de reglas débiles o con poca utilidad práctica en la interpretación de los datos.
  - **Soporte mínimo:** Se fijó en 0.3 para garantizar que las reglas encontradas sean relevantes dentro del dataset y no sean anomalías o patrones poco representativos.Un soporte del 30% implica que la regla debe aparecer en al menos 1912 instancias, asegurando que los patrones detectados sean comunes y significativos.
  - **Número de reglas generadas:** Se configuró el algoritmo para obtener las 10 mejores reglas basadas en los criterios de evaluación.

- **Métricas para evaluar reglas de asociación** Para la evaluación de las reglas generadas, se utilizó las siguientes métricas:
  - **Confianza (conf):** representa la probabilidad de que el consecuente ocurra dado que el antecedente ha ocurrido.
  - **Lift:** mide la fuerza de la regla en comparación con la ocurrencia independiente de los ítems.
  - **Leverage (lev):** cuantifica la diferencia entre la frecuencia conjunta observada del antecedente y consecuente y la frecuencia esperada si fueran independientes.
  - **Soporte (support):** indica el número absoluto de instancias en las que ocurre la combinación de ítems de la regla.
  - **Convicción (conv):** evalúa qué tan dependiente es la ocurrencia del consecuente respecto al antecedente.

IV. Resultados y Discusión

A. Resultados Clasificación

1) **Algoritmo KNN**

a) **Escenario sin modificaciones**
Como se observa en la Tabla II, la métrica presición es correcta, F1 score de igual forma es una métrica correcta para el modelo, de la misma manera la métrica recall. Por último la métrica **accuracy** 0.695 es una métrica trivial para el modelo.

Tipo	Precisión	Recall	F1 Score
Positivos	0.70	0.73	0.72
Negativos	0.69	0.66	0.67
Accuracy	**0.695**		

TABLE II: Métricas con datos crudos.

b) **Con escalamiento de datos**
se normalizó el dataset, para poder identificar mejoras con respecto al primer análisis, se obtuvo mejores resultados.En la Tabla III se observa las métricas y se evidencia una mejora considerable en los resultados, en la métrica **accuracy** se tiene una mejora del 18.13% con respecto al análisis anterior.

Tipo	Precisión	Recall	F1 Score
Positivos	0.82	0.85	0.83
Negativos	0.83	0.79	0.81
Accuracy			**0.821**

TABLE III: Métricas con datos normalizados.

c) **Reducción de dimensionalidad**
Finalmente, en el análisis de reducción de dimensionalidad con PCA, se utilizó 42 componentes. En este caso, se observó una incremento en rendimiento en varias de las métricas Tabla IV , especialmente en recall de negativos con respecto al modelo con datos normalizados, por ende podemos concluir que este último modelo tiene un incremento en rendimiento de 1.1%.

Tipo	Precisión	Recall	F1 Score
Positivos	0.83	0.85	0.84
Negativos	0.83	0.81	0.82
Accuracy			**0.83**

TABLE IV: Métricas con datos aplicados PCA.

2) **Algoritmo Multilayer Perceptron**

a) **Con escalamiento de datos**
En la Tabla V se pueden evidenciar los resultados obtenidos. En este caso existen mejoras en las métricas y rendimiento comparando a los modelos obtenidos en la sección de KNN. La métrica **accuracy** tiene una mejora de rendimiento de 4.46% respecto a la mejor métrica de la sección anterior.

Tipo	Precisión	Recall	F1 Score
Positivos	0.85	0.90	0.88
Negativos	0.88	0.83	0.86
Accuracy			**0.867**

TABLE V: Métricas con datos normalizados.

b) **Reducción de dimensionalidad**
Los resultados se muestran en la Tabla VI. En este caso, las métricas evidencian que los resultados no superan a los obtenidos previamente. Esto podría explicarse por el hecho de que, aunque PCA es una técnica eficaz para reducir la dimensionalidad, puede conllevar una pérdida de información relevante, lo que afecta el desempeño del modelo. Pero este modelo evidencia un mejor rendimiento que los modelos obtenidos en la sección aplicado KNN.

Tipo	Precisión	Recall	F1 Score
Positivos	0.85	0.86	0.85
Negativos	0.84	0.83	0.83
Accuracy			**0.842**

TABLE VI: Métricas con datos aplicados PCA.

En conclusión el modelo con mejor rendimiento y métricas; el algoritmo Multilayer Perceptron con datos normalizados y una configuración de dos capas ocultas de 500 neuronas cada una y un máximo de 500 iteraciones.

B. Resultados Clustering

En esta sección se presentan los resultados obtenidos al aplicar el modelo de clustering llamado K-Means al conjunto de datos. El objetivo fue identificar agrupaciones significativas que proporcionen una mejor comprensión del contenido del dataset.
El algoritmo K-Means fue seleccionado para este estudio debido a su simplicidad y eficacia al trabajar con datasets de alta dimensionalidad. No obstante, este modelo requiere conocer el número óptimo de clusters (K) a priori. Para determinar este valor, se utilizaron métricas como el método del codo y el coeficiente de silueta.
Analizando la Figura 3, el método del codo muestra un punto de inflexión (similar a un ”codo”) entre 6 y 7 clusters, donde la disminución en la suma de los errores cuadráticos se vuelve menos pronunciada. Este comportamiento sugiere que el número óptimo de clusters podría ser 6.

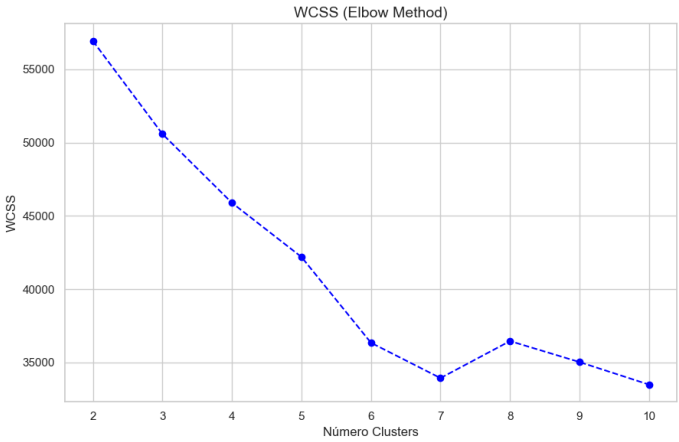

Fig. 3: Método del codo

Por su parte, el coeficiente de silueta (Figura 4) mide la calidad de los clusters evaluando la cohesión (proximidad de los puntos dentro de un mismo cluster) y la separación (distancia entre clusters), también indicó que 6 clusters es la mejor opción, ya que este número presentó el valor más alto del coeficiente de silueta.
En conclusión, se determinó que el número óptimo de clusters para este estudio es 6, dado que ambos métodos de evaluación convergieron en esta recomendación.

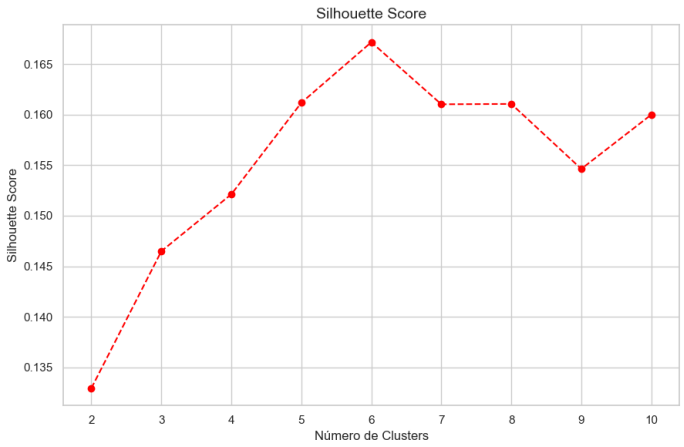

Fig. 4: Silhouette score

Tras entrenar el algoritmo K-Means con el dataset de emigración, filtrado exclusivamente para los habitantes de la región Amazónica, se obtuvieron seis clusters, los cuales se visualizan en la Figura 5 en dos dimensiones, utilizando Análisis de Componentes Principales (PCA).

En la Figura 5, no se aprecia claramente una separación definida entre los clusters. Esto podría deberse a que los datos no presentan una estructura lineal. Otra posible razón es que, dado que PCA busca maximizar la varianza en las primeras componentes principales al reducir la dimensionalidad, este proceso podría no preservar la estructura de separación de los clusters identificada por K-Means en las dimensiones originales del dataset.

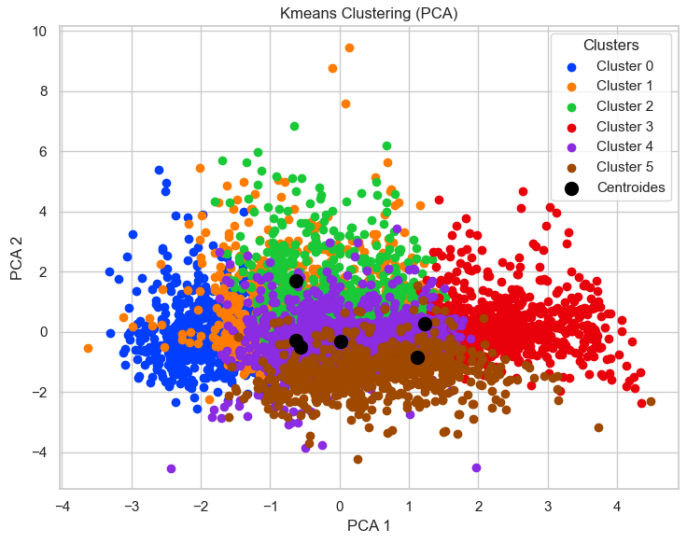

Fig. 5: Clusters formados con k-means.

C. Interpretación de Clusters

La interpretación de los clusters generados fue un paso crucial para dar sentido a las agrupaciones y comprender los diferentes grupos de emigrantes de la Amazonía representados en cada cluster. Para interpretar y caracterizar los clusters, se analizaron la moda y las distribuciones de frecuencia de las variables categóricas más relevantes del dataset.

En la Figura 6 se visualizan las frecuencias de cada categoría correspondiente al área de residencia previa de los emigrantes, ya sea urbana (1) o rural (2), para cada uno de los clusters. Este diagrama permite identificar patrones claros en las agrupaciones, como que en el cluster 1 el 99% de los emigrantes provienen de la zona rural, mientras que el cluster 2 está compuesto exclusivamente por personas que salieron de la zona urbana. De manera similar, se observan características distintivas en los demás clusters, lo que facilita su interpretación y análisis.

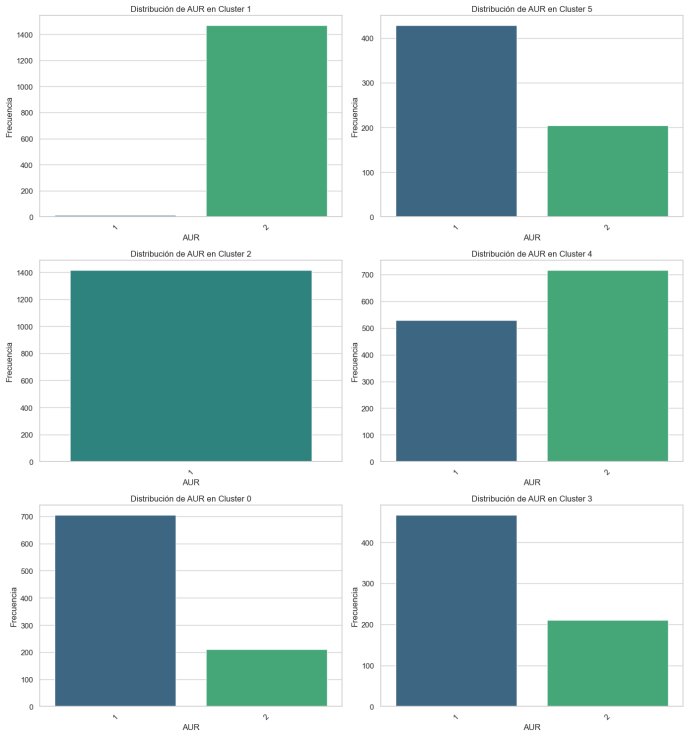

Fig. 6: Distribución de área urbana o rural de cada cluster

Para los datos numéricos, como la edad de los emigrantes al momento de salir del país, se analizaron diversas medidas estadísticas, incluyendo la media, la mediana, los cuartiles, y los valores máximos y mínimos. Con el fin de obtener una representación más clara y facilitar la interpretación de los datos, se elaboraron diagramas de caja para cada clúster.

A través de la visualización y el análisis de estos diagramas, se identificó que, en el último clúster, la mayoría de los emigrantes tenían entre 40 y 52 años, siendo la persona más joven de este grupo de 32 años. Este patrón sugiere que dicho clúster está compuesto por emigrantes adultos, todos mayores de 31 años, con la mayoría de ellos teniendo entre 40 y 52 años al momento de emigrar.

Luego de realizar el análisis anteriormente descrito por cada variable significativa como lo es el año de salida del emigrante, el sexo, a que país emigró, en que zona (urbana o rural) habitaba el emigrante en el país, y la edad que tenía al momento de salir de país. Se logró caracterizar los clusters,

los cuales se describen a continuación.

- **Primer cluster (Cluster 0):** Este grupo de emigrantes, tiene como destino principal a Estados Unidos. Se registró un pico migratorio en 2021, cuando más de 200 personas dejaron el país, mientras que en 2020 y 2022 la cifra alcanzó aproximadamente 150 personas por año. Este cluster presenta una notable diferencia de género, con cerca de 580 hombres frente a 330 mujeres. La mayoría de los emigrantes en este grupo residía en zonas urbanas y tenía entre 20 y 29 años al momento de salir del país, lo que podría reflejar un perfil de jóvenes en búsqueda de oportunidades laborales o educativas en el extranjero.

- **Segundo cluster (Cluster 1):** Los emigrantes de este cluster provienen principalmente de zonas rurales de Ecuador y tienen como destino principal a Estados Unidos, aunque una pequeña proporción emigró a España. Los años con mayor actividad migratoria fueron 2021 y 2022, con algunos movimientos menores registrados en 2019 y 2020. Este grupo está compuesto principalmente por jóvenes de entre 19 y 27 años, lo que podría indicar una migración temprana motivada por la falta de oportunidades económicas en las zonas rurales o falta de interés en seguir las actividades tradicionales de sus culturas.

- **Tercer cluster (Cluster 2):** Este cluster está compuesto exclusivamente por emigrantes provenientes de zonas urbanas de la región amazónica de Ecuador. Estados Unidos se mantiene como el destino principal, seguido por España y, en menor medida, Argentina. La proporción de género es equilibrada, lo que sugiere una participación similar de hombres y mujeres en este flujo migratorio. La mayoría de los emigrantes tenía entre 19 y 27 años de edad al momento de emigrar, reflejando un perfil demográfico joven y urbano característico de esta región.

- **Cuarto cluster (Cluster 3):** Los emigrantes de este cluster destacan por su diversidad de destinos. Aunque Estados Unidos sigue siendo el principal, países como Colombia, España, Rusia, Argentina, y, en menor medida, Chile y Perú también son relevantes. La migración de este grupo ha mostrado un crecimiento constante desde 2015, alcanzando su punto máximo en 2022. Existe un equilibrio notable entre hombres y mujeres, con una ligera mayoría masculina. La mayoría de los emigrantes eran jóvenes de entre 19 y 29 años, provenientes principalmente de zonas urbanas, aunque hubo algunas excepciones provenientes de áreas rurales. Esta diversidad de destinos podría estar relacionada con redes familiares o laborales ya establecidas en múltiples países.

- **Quinto cluster (Cluster 4):** Este grupo, muestra una concentración migratoria significativa en los años 2021 y 2022, con un flujo casi equitativo entre ambos años. Más del 70% de los emigrantes eran hombres jóvenes, con edades entre 19 y 27 años. Aproximadamente el 60% de estas personas provenían de zonas rurales, reflejando la continua migración del campo a países como Estados Unidos, que fue el principal destino, seguido por España, y en menor medida, Colombia y Argentina.

- **Sexto cluster (Cluster 5):** Este cluster destaca por estar compuesto por emigrantes de mayor edad en comparación con los demás grupos, con rangos que van de los 40 a los 52 años. La mayoría provenía de zonas urbanas y mostraba un equilibrio de género, con una participación similar de hombres y mujeres. Estados Unidos y España se posicionaron como los destinos principales. Este perfil sugiere una posible migración de adultos en busca de estabilidad económica o reunificación familiar.

D. Resultados reglas de asociación

En la Tabla VII, se puede evidenciar las diez mejores reglas de asociación, cada una de estas cumple con las métricas para atribuir una relación fuerte entre las mismas. En la métrica **conf**, si conf ≈ 1 se considera una métrica máxima, muchas de ellas cumplen con el enunciado. En la métrica **lift** todas las reglas cumple con $fift > 1$, por ende se puede concluir que la relación por parte de antecedente y consecuente en la reglas son realmente fuertes. En la métrica **lev** $lev > 0$, indica que existe una fuerte dependencia entre el antecedente y consecuente en todas las reglas. En la métrica **conv** $con > 1$, indica que todas las reglas son significativas. Ahora bien indicaremos el significado de las variables involucradas en cada una de las reglas.

Regla	conf	lift	lev	support	conv
V04=1 V07=2 → DEF_HAB=1	1	2.05	0.16	998	333.52
V04=1 V06=1 V08=1 → DEF_HAB=1	1	2.04	0.17	1102	123.41
V04=1 V08=1 → DEF_HAB=1	1	2.04	0.19	1215	122.49
V04=1 V06=1 → DEF_HAB=1	0.98	2.01	0.19	1180	26.54
V06=1 V08=1 DEF_HAB=1 → V04=1	0.97	2.14	0.18	1151	19.56
V06=1 DEF_HAB=1 → V04=1	0.97	2.13	0.2	1247	17.4
V04=1 → DEF_HAB=1	0.95	1.95	0.21	1342	10.52
V08=1 DEF_HAB=1 → V04=1	0.94	2.08	0.19	1234	9.45
V04=1 V06=1 DEF_HAB=1 → V08=1	0.92	1.79	0.15	995	6.07
V06=1 DEF_HAB=1 → V08=1	0.92	1.79	0.15	977	5.83

TABLE VII: Reglas de asociación.

1) **V04=1 V07=2 → DEF_HAB=1**
   Esta regla indica, si $V04 = 1$ (el estado del techo de la vivienda es bueno) y $V07 = 2$ (el material predominante del piso es cerámica, baldosa, vinil o porcelanato) entonces $DEF\_HAB = 1$ (el déficit habitacional es aceptable).

2) **V04=1 V06=1 V08=1 → DEF_HAB=1**
   Esta regla indica, si $V04 = 1$ (el estado del techo de la vivienda es bueno), $V06 = 1$ (el estado de las paredes exteriores es bueno) y $V08 = 1$ (el estado del piso es bueno) entonces $DEF\_HAB = 1$ (el déficit habitacional es aceptable).

3) **V04=1 V08=1 → DEF_HAB=1**
   Esta regla indica, si $V04 = 1$ (el estado del techo de la vivienda es bueno) y $V08 = 1$ (el estado del piso es bueno) entonces $DEF\_HAB = 1$ (el déficit habitacional es aceptable).

4) **V04=1 V06=1 → DEF_HAB=1**
   Esta regla indica, si $V04 = 1$ (el estado del techo de la vivienda es bueno) y $V06 = 1$ (el estado de las paredes exteriores es bueno) entonces $DEF\_HAB = 1$ (el déficit habitacional es aceptable).

5) **V06=1 V08=1 DEF_HAB=1 → V04=1**
   Esta regla indica, si $V06 = 1$ (el estado de las paredes exteriores es bueno), $V08 = 1$ (el estado del piso es bueno) y $DEF\_HAB = 1$ (el déficit habitacional es aceptable) entonces $V04 = 1$ (el estado del techo de la vivienda es bueno).

6) **V06=1 DEF_HAB=1 → V04=1**
   Esta regla indica, si $V06 = 1$ (el estado de las paredes exteriores es bueno) y $DEF\_HAB = 1$ (el déficit habitacional es aceptable) entonces $V04 = 1$ (el estado del techo de la vivienda es bueno).
7) **V04=1 → DEF_HAB=1**
   Esta regla indica, si $V04 = 1$ (el estado del techo de la vivienda es bueno) entonces $DEF\_HAB = 1$ (el déficit habitacional es aceptable).

8) **V08=1 DEF_HAB=1 → V04=1**
   Esta regla indica, si $V08 = 1$ (el estado del piso es bueno) y $DEF\_HAB = 1$ (el déficit habitacional es aceptable) entonces $V04 = 1$ (el estado del techo de la vivienda es bueno).

9) **V04=1 V06=1 DEF_HAB=1 → V08=1**
   Esta regla indica, si $V04 = 1$ (el estado del techo de la vivienda es bueno), $V06 = 1$ (el estado de las paredes exteriores es bueno) y $DEF\_HAB = 1$ (el déficit habitacional es aceptable) entonces $V08 = 1$ (el estado del piso es bueno).

10) **V06=1 DEF_HAB=1 → V08=1**
    Esta regla indica, si $V06 = 1$ (el estado de las paredes exteriores es bueno) y $DEF\_HAB = 1$ (el déficit habitacional es aceptable) entonces $V08 = 1$ (el estado del piso es bueno).

Las reglas expuestas presentan principalmente variables relacionadas con la vivienda. En particular, se observó que muchas reglas se asociaban al déficit habitacional, una variable categórica que clasifica las viviendas en tres categorías: déficit aceptable, recuperable e irrecuperable. Los valores indicaron que el déficit habitacional predominante correspondía a la categoría de déficit aceptable.

V. Conclusiones

En este estudio se ha evidenciado que las características de los emigrantes de las provincias de la Amazonía ecuatoriana están estrechamente relacionadas con información relevante sobre sus hogares, particularmente vinculada a las características físicas de sus viviendas. Estos aspectos han demostrado ser determinantes para identificar de manera precisa las zonas de residencia de los emigrantes, sirviendo como base sólida para la construcción de modelos de clasificación eficientes.

El algoritmo Multilayer Perceptron (MLP) ha demostrado un buen desempeño al clasificar las zonas de ubicación de las viviendas de los emigrantes antes de su salida del país, alcanzando un accuracy del 86%. Por su parte, el algortimo K-Nearest Neighbors (KNN) logró un accuracy máximo del 83%. Ambos modelos han demostrado un rendimiento satisfactorio, respaldado por métricas complementarias de evaluación, lo que confirma la eficacia del enfoque propuesto para el análisis del comportamiento migratorio en la región amazónica.

Por otro lado, mediante el modelo de aprendizaje no supervisado K-means, se identificaron seis grupos de emigrantes amazónicos. Entre ellos, destacan jóvenes que emigran buscando oportunidades laborales o educativas, así como aquellos que lo hacen debido a la falta de oportunidades económicas o el desinterés por continuar con las actividades tradicionales de sus culturas. Un grupo relevante también incluye a jóvenes provenientes de zonas urbanas, cuyo destino parece estar relacionado con redes familiares o laborales ya establecidas en diferentes países. Además, se observó un grupo de adultos que emigran en busca de estabilidad económica o reunificación familiar. Es importante señalar que los primeros y terceros clústeres mostraron grandes similitudes, lo que sugiere que podrían ser considerados un solo grupo debido a sus características comunes en cuanto a edad, género y países de destino, con la única diferencia en la zona de residencia.

Finalmente, se puede concluir que, entre los emigrantes de la provincia de la Amazonía, la información sobre los aspectos de vivienda en Ecuador resulta de gran relevancia. Los datos que reflejan el buen estado de los materiales de las viviendas como lo son pares exteriores, techo y piso, así como la predominancia de ciertos materiales de piso específicamente, sugieren un déficit habitacional aceptable dentro de este grupo.

References

[1] L. P. Paredes, "La migración internacional en ecuador: sus causas, consecuencias y situación actual," *Revista de Investigación del Departamento de Humanidades y Ciencias Sociales*, vol. 14, pp. 73–98, 2018. Licencia Creative Commons Atribución-NoComercial-SinDerivar 3.0 Internacional.

[2] Censo Ecuador, "Data - Censo Ecuador," 2022. Accedido el 1 de febrero de 2025.

[3] INEC, "Diccionario de variables," 2022. En línea; consultado en enero de 2025.

[4] INEC, "Guía de usuario," 2024. En línea; consultado en enero de 2025.

[5] OpenClassrooms, "Analyze the results of a k-means clustering," 2020. Accedido: 26-Jan-2025.

[6] S. Urdanegui, "Algoritmos de clustering con scikit-learn en python," 2022. Accedido: 26-Jan-2025.