SEGMENTACIÓN DE LOS FLUJOS MIGRATORIOS EN COLOMBIA: IDENTIFICACIÓN DE SUBGRUPOS Y CARACTERÍSTICAS COMUNES

CINDY VANESSA AGUIRRE MARÍN

Trabajo de grado

Asesor

JUAN DAVID MARTINEZ VARGAS

Co-asesor

LINA MARIA SEPÚLVEDA CANO

UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
MAESTRÍA EN CIENCIAS DE LOS DATOS Y ANALÍTICA
MEDELLÍN
2024

Resumen

El aumento de la migración global ha intensificado los flujos migratorios, destacándose como un fenómeno relevante para las políticas globales, regionales y nacionales. En Colombia, a partir de 2015, la migración venezolana ha llevado a una mayor atención a los flujos migratorios. Este estudio analiza los flujos migratorios hacia Colombia en 2023 utilizando técnicas de aprendizaje automático. Con datos de Migración Colombia, se aplicó *K-Means* para segmentar los flujos migratorios y *UMAP* para reducir la dimensionalidad de los datos. Los resultados revelan cuatro clusters principales, definidos por la región de origen, motivo del viaje, región de hospedaje y mes de llegada. La mayoría de los flujos corresponden a turistas, lo que sugiere que los datos de puntos migratorios oficiales reflejan principalmente movimientos turísticos y no necesariamente otros tipos de migración. Las técnicas de *Machine Learning* demostraron ser efectivas para descubrir patrones complejos en los datos categóricos, y la interpretación mediante *SmartExplainer* de *SHAPash* facilitó la comprensión de estos patrones. Este estudio no solo segmentó adecuadamente los flujos migratorios, sino que también proporcionó herramientas interpretativas para futuros análisis de datos categóricos.
Palabras clave: *Flujos migratorios, Clustering, Machine Learning, K-Means, UMAP, SHAP.*

Abstract

The increase in global migration has intensified migratory flows, emerging as a relevant phenomenon for global, regional, and national policies. In Colombia, since 2015, Venezuelan migration has sparked interest in migratory flows. This study analyzes migratory flows to Colombia in 2023 using *Machine Learning* techniques. *K-Means* was applied in order to segment data from Migration Colombia, while *UMAP* was used to reduce the dimensionality of the data itself. The results reveal four main *clusters*, defined by the region of origin, reason for travel, host region, and month of arrival. Most flows correspond to tourists, suggesting that the data from official migration points primarily reflect tourist movements and not necessarily other types of migration. *Machine Learning* techniques proved effective in uncovering complex patterns in categorical data, and interpretation using *SmartExplainer* by *SHAPash* facilitated the understanding of these patterns. This study not only adequately segmented migratory flows but also provided interpretative tools for future analyses of categorical data.
Keywords: *migratory flows, Clustering, Machine Learning, K-Means, UMAP, SHAP.*

0. INTRODUCCIÓN

El aumento de la migración en el mundo ha traído consigo una intensificación de los flujos migratorios, convirtiéndose en un fenómeno de creciente relevancia global. Estos flujos incluyen no solo la migración tradicional, sino también el turismo, el comercio y la migración laboral, entre otros. La movilidad humana, que alguna vez fue vista principalmente desde una perspectiva estadística, ahora es una prioridad para quienes formulan políticas a nivel global, regional y nacional. La Agenda 2030 para el Desarrollo Sostenible, en particular, subraya la importancia de recopilar datos precisos sobre estos flujos para facilitar una migración segura y ordenada, como se destaca en la meta 10.7 de los Objetivos de Desarrollo Sostenible —ODS— y en otras metas relacionadas (Global Migration Group, 2017).

Colombia, históricamente caracterizada por una movilidad significativa de su población hacia el exterior (Cancillería de Colombia, s.f.), experimentó un cambio notable a partir de 2015, cuando el flujo de migrantes se intensificó, principalmente debido a la migración de personas de nacionalidad venezolana (Departamento Administrativo Nacional de Estadística [DANE], 2022). Este cambio demográfico ha provocado ajustes en las políticas económicas y sociales del país, resaltando la importancia de contar con datos precisos y segmentados para una gestión efectiva.

1. PLANTEAMIENTO DEL PROBLEMA

El estudio aborda el creciente interés en los últimos años sobre el contexto migratorio colombiano, que empezó a tener mayor relevancia desde el incremento de la migración venezolana a Colombia (Departamento Administrativo Nacional de Estadística [DANE], 2022), destacando la necesidad de estudiar los flujos migratorios crecientes hacia el país. Para ello, Migración Colombia, dispone de varios productos e instrumentos de difusión para analizar el flujo migratorio de las personas que ingresan y salen al país a través de puestos migratorios oficiales, tales como: boletines migratorios con periodicidad anual, tableros de visualización estadística y la base de datos abierta de entradas de extranjeros a Colombia y las salidas de colombianos desde el 2012 (Migración Colombia, 2022b).

Dado lo anterior, se utilizará la base de datos de entradas y salidas de personas proporcionada por Migración Colombia para el año 2023, aplicando algoritmos de agrupamiento para su segmentación. Este método tiene el potencial de proporcionar un análisis más detallado de los flujos migratorios. Sin embargo, la naturaleza 100% categórica de los datos presenta desafíos significativos. Estos datos

requieren un preprocesamiento meticuloso para convertir las variables en formatos adecuados para los algoritmos de *Machine Learning,* así como métodos robustos para su interpretación y análisis. Los conjuntos de datos de alta dimensionalidad corren el riesgo de ser muy dispersos, y cuantas más dimensiones tenga el conjunto de entrenamiento, mayor será la probabilidad de sobreajuste (Géron, 2022). Además, la naturaleza ordinal y/o nominal de los datos puede complicar la interpretación de los modelos.

Para abordar estos desafíos, se emplearán algoritmos de clusterización. Los algoritmos de clusterización son técnicas de *Machine Learning* que agrupan los objetos de un conjunto de datos según su similitud, de manera que los objetos dentro de un grupo —clúster— sean más similares entre sí que aquellos en grupos distintos. Aunque los humanos tienden a agrupar por similitud de forma natural, la definición precisa de un clúster en un contexto algorítmico es compleja, lo que ha dado lugar a una amplia variedad de algoritmos de *clustering* (Sancho, 2023). Con este proceso se busca agrupar a los individuos que ingresan al país en clústeres homogéneos según sus características similares, facilitando así una comprensión más detallada de los diferentes tipos de flujos migratorios, como el turismo, la migración por razones laborales, educativas, comerciales, entre otros.

En resumen, el desafío radica en encontrar métodos efectivos para agrupar e interpretar datos completamente categóricos, lo que permitirá aprovechar al máximo la riqueza de los datos disponibles.

2. JUSTIFICACIÓN

La migración es parte de la historia de los pueblos. La movilidad, por su parte, es un derecho de las personas en todo el mundo, que tiene un impacto económico, social y político en los países de origen y en los de destino; por ello, es importante promover la migración ordenada y regulada, para lograr la planeación de servicios y el acceso a oportunidades para los migrantes. Ante este escenario, surge la necesidad urgente de que los países desarrollen mecanismos efectivos para cuantificar y analizar estos flujos migratorios. La producción de información estadística se fundamenta en los principios de las Naciones Unidas sobre las estadísticas oficiales, las recomendaciones de la Organización Internacional para las Migraciones —OIM— y las buenas prácticas establecidas por entidades como el DANE en Colombia. Este enfoque detallado permite la compilación de estadísticas sobre flujos migratorios, proporcionando definiciones estandarizadas y lineamientos para la recopilación y tabulación de datos,

los cuales pueden ser usados por entidades públicas, privadas o con fines académicos para analizar diferentes patrones y cubrir las necesidades de información específicas.

La presente investigación propone contribuir mediante el uso de técnicas de aprendizaje automático para el procesamiento y análisis de información estadística obtenida de fuentes oficiales, como la base de datos de Migración Colombia. Este enfoque se centra en segmentar a la población que ingresa a Colombia, con el objetivo de avanzar en la comprensión de los flujos migratorios en el país. Al utilizar herramientas avanzadas de *Machine Learning*, se busca identificar patrones ocultos y presentar hallazgos basados en datos.

3. OBJETIVOS

3.1 GENERAL

Analizar los patrones y características comunes de la población que ingresa a Colombia en 2023 utilizando técnicas avanzadas de segmentación, con el objetivo de generar conocimiento sobre las particularidades de esta población en ese período específico.

3.2 ESPECÍFICOS

- Identificar las características relevantes para la segmentación de la población que ingresa a Colombia en el 2023
- Formular un modelo de segmentación mediante la aplicación de técnicas de aprendizaje automático.
- Interpretar los resultados del modelo para comprender los segmentos identificados en el conjunto de datos.

4. MARCO TEÓRICO

4.1 Flujos migratorios en Colombia

Migración Colombia, como autoridad de vigilancia y control migratorio del Estado Colombiano, tiene la misión de ejercer control migratorio eficaz. La necesidad de procesar los datos obtenidos surge de la importancia de contar con información precisa y oportuna para la toma de decisiones y la formulación de políticas migratorias. Esta información es recopilada a través de 46 Puestos de Control Migratorio, lo cual permite capturar, registrar, procesar y analizar los datos migratorios de manera eficiente. La base de

datos de "entradas y salidas de personas del país" es resultado de estos esfuerzos y se estructura para proporcionar cifras exactas y oportunas que faciliten la toma de decisiones tanto dentro de la entidad como a nivel gubernamental y a entidades externas. Esta operación estadística se implementa mediante metodologías claras para la recolección, procesamiento, análisis y divulgación de información, garantizando así la calidad y la utilidad de los datos en la producción de información estadística migratoria (Migración Colombia, 2022a).

Es importante destacar que la información producida por Migración Colombia es útil para la formulación de políticas públicas y es una de las pocas fuentes de información sobre movimientos internacionales de personas en el país (Migración Colombia, 2022a). A partir de lo anterior, se crean instrumentos de difusión para analizar la información construida, tales como los informes, reportes y gráficas, que se divulgan a través de la página web oficial de Migración Colombia.

La información estadística sobre los flujos migratorios se presenta en varios formatos, entre ellos, el Boletín Migratorio, un documento analítico elaborado anualmente sobre los eventos migratorios más relevantes del año anterior, utilizando herramientas de estadística descriptiva como comparativos e identificación de tendencias. Además, se publican tableros de visualización estadística, que son reportes interactivos con gráficos, tablas y mapas de georreferenciación que muestran la información sobre flujos migratorios (Migración Colombia, 2022b).

La investigación llevada a cabo por Recaño-Valverde et al. (2012) utilizaron la base de datos de entrada y salida de personas al país para construir un modelo comparativo que permite una estimación indirecta de los flujos migratorios. Este estudio se centró en la comparación de los flujos migratorios entre Colombia y España, revelando hallazgos significativos. La metodología implementada permitió identificar que un 92,2% de los inmigrantes colombianos en España, según la estadística de variaciones residenciales —EVR— española, correspondían a los flujos estimados por la base de datos colombiana. Además, se descubrió que la estructura demográfica de los migrantes, en términos de edad y sexo, era coherente entre ambas bases de datos, validando así la robustez del modelo colombiano. Estos resultados son esenciales para comprender los patrones migratorios y mejorar las políticas públicas relacionadas con la migración entre Colombia y España (Recaño-Valverde et al., 2012).

4.2 *Machine Learning* aplicado a flujos migratorios

Hoffmann y Luengo-Oroz (2023) desarrollan un marco computacional para analizar flujos migratorios, abordando la necesidad de entes gubernamentales y no gubernamentales de tomar decisiones estratégicas informadas por datos. Debido a la falta de guías precisas para manejar estos fenómenos sociales, se

investigan herramientas computacionales y análisis predictivos, como aprendizaje automático, inteligencia artificial, simulaciones y pronósticos estadísticos, para mejorar la precisión de las predicciones. El marco descrito en su investigación incluye la recolección de datos de diversas fuentes (celulares, redes sociales, datos económicos y medioambientales, teledetección, datos geográficos, etcétera, lo cual también es mencionado por Polimis y Zagheni (2020); Fiorio et al. (2021); Zagheni, Weber y Gummadi (2017), como se citó en Molina et. al (2022), y la incorporación de características clave de la región y sus vecinos para comprender mejor las razones detrás de los flujos migratorios. Para manejar la gran cantidad de información y variables que puedan surgir, Hoffmann y Luengo-Oroz (2023) proponen simplificar el modelo mediante algoritmos de selección automatizada de características, como los métodos Forward o Backward, o mediante el análisis de componentes principales —*PCA*—. Además, destacan técnicas de aprendizaje automático que pueden ayudar en la selección de características, como *Random Forest*, que permite evaluar la importancia de las características durante la construcción del modelo.

La elección del modelo dependerá de si se trata de un problema de clasificación o regresión y Hoffmann y Luengo-Oroz (2023) proponen en su estudio el uso de una variedad de algoritmos de aprendizaje supervisado para abordar estos problemas. Entre los algoritmos destacados se encuentran:

- *Lasso y Ridge Regressions:* estos modelos de regresión lineal incorporan una penalización de regularización en los pesos de los coeficientes del modelo, lo que permite considerar un gran número de variables mientras se restringe el sobreajuste.
- *Support Vector Machine —SVM—*: esta técnica de clasificación busca separar las observaciones utilizando un hiperplano. El algoritmo penaliza las observaciones que caen del lado incorrecto del hiperplano, y la penalización aumenta con la distancia desde este.
- *Random Forest*s: los modelos de bosques aleatorios se construyen a partir de un conjunto de muchos árboles de decisión. Para evitar el sobreajuste, el modelo selecciona una nueva muestra de observaciones para cada árbol y un nuevo subconjunto de variables para cada división en el árbol. Las predicciones de los árboles individuales se promedian en un conjunto.
- *Boosting Algorithms —e.g., AdaBoost, XGBoost—*: estos modelos de conjunto se construyen secuencialmente, donde cada modelo intenta corregir los errores del anterior.

Hoffmann y Luengo-Oroz (2023) destacan que, en la práctica, es común probar muchos modelos diferentes debido a la falta de un criterio previo fuerte sobre cuál funcionará mejor para el problema en cuestión. Este enfoque es adoptado, por ejemplo, por Huynh y Basu (2020) en su modelo de

desplazamiento interno y Nair et al. (2020) en su modelo de migración mixta, como se citó en Hoffmann y Luengo-Oroz (2023). A menudo, los modelos sofisticados como *Random Forest*s, *Adaboost y XGBoost* son los que mejor desempeño tienen debido a su flexibilidad y capacidad para ajustarse bien a los datos.

Por ejemplo, Molina et al. (2022) usaron un modelo de *Random Forest* para estudiar cómo el cambio climático influye en las decisiones de migración entre México y Estados Unidos. Este enfoque de *Machine Learning* permitió predecir si una persona migrase o no, capturando relaciones complejas entre variables climáticas, sociales, económicas y demográficas. Sin embargo, lo modelos de aprendizaje automático a menudo producen resultados "caja negra" dado que no especifican claramente cómo se vinculan las variables predictoras con el resultado —como si lo hace un modelo paramétrico a través de la estimación de sus coeficientes—. Para abordar esta limitación, el estudio identificó los predictores climáticos más importantes, con el fin de incluirlos en un modelo más simple y entendible que permita estudiar detalladamente su relación con la migración y otros factores influyentes.

De manera similar, el análisis de clúster jerárquico se ha utilizado para investigar los patrones de migración interregional en Turquía durante el período de 2008 a 2010 (Akin y Dökmeci, 2014). Este método agrupa regiones según similitudes o diferencias en los flujos de inmigración y emigración. Al emplear diversas técnicas de enlace, como el enlace simple, completo y promedio, se determinó la distancia óptima entre clústeres utilizando la correlación cofenética. Los resultados mostraron que Estambul domina el clúster más grande del país, destacando su influencia en los movimientos migratorios. Este enfoque permitió identificar las regiones que generan o atraen mayores flujos migratorios y entender si estos flujos reflejan desigualdades regionales. La comprensión de estas dinámicas es crucial para superar dichas desigualdades y puede informar estrategias de desarrollo nacional, dirigiendo inversiones públicas y fomentando inversiones privadas en regiones subdesarrolladas mediante subsidios económicos (Akin y Dökmeci, 2014).

En conclusión, los modelos aprendizaje automático ofrecen herramientas poderosas para comprender y predecir patrones migratorios complejos. Estos métodos permiten capturar relaciones intrincadas entre múltiples variables, proporcionando *insights* valiosos para la formulación de políticas y la planificación estratégica.

4.3 Técnicas de agrupamiento

La agrupación de datos es un problema ampliamente estudiado en minería de datos y aprendizaje automático debido a sus numerosas aplicaciones en el aprendizaje, síntesis, segmentación y marketing. Este proceso consiste en dividir un conjunto de datos en grupos que sean lo más similares posible,

utilizando modelos generativos probabilísticos o enfoques basados en distancia. Este método se aplica en filtrado colaborativo, segmentación de clientes, resumen de datos, detección de tendencias y análisis de datos biológicos, entre otros, proporcionando resúmenes compactos y útiles para diversas aplicaciones (Aggarwal y Reddy, 2015).

Para abordar este problema, se destacan los métodos de selección de características y reducción de dimensionalidad, en los cuales la fase de selección de características es un paso de preprocesamiento clave para mejorar la calidad del agrupamiento, dado que no todas las características del conjunto de datos son relevantes y pueden causar ruido al momento de generar los clústeres. Este enfoque se complementa con las técnicas de reducción de dimensionalidad, que buscan simplificar la representación del conjunto de datos al disminuir el número de variables de entrada o características, conservando la información más importante (Vijay, 2023). En datos de alta dimensionalidad, con cientos o miles de características, la reducción de dimensionalidad ayuda a mejorar la eficiencia computacional, evitar el sobreajuste y facilitar la visualización de los datos sin perder información esencial (Bored ASD, 2021).

Si bien, existen varios métodos para reducir los datos de alta dimensión a un espacio de menor dimensión, para el presente caso de estudio se usará *UMAP —Uniform Manifold Approximation and Projection—*, dado que puede ser utilizado como un paso de preprocesamiento efectivo para mejorar el rendimiento de la agrupación en clústeres (McInnes, Healy y Melville, 2018). *UMAP* es un algoritmo de reducción de dimensiones no lineal que supera algunas limitaciones de *t-SNE*. Es considerablemente más rápido que *t-SNE* y es determinista, lo que significa que siempre produce el mismo resultado con los mismos datos de entrada. Esto permite incorporar *UMAP* en pipelines de *Machine Learning,* ya que puede proyectar nuevos datos en la representación de menor dimensión (Rhys, 2020). *UMAP* también preserva tanto la estructura local como la global, permitiendo interpretar similitudes tanto entre puntos individuales como entre grupos de puntos en dimensiones altas. El algoritmo asume que los datos están distribuidos a lo largo de una variedad, una forma geométrica n-dimensional que localmente parece plana. *UMAP* calcula las distancias entre los puntos a lo largo de esta variedad y optimiza iterativamente una representación de menor dimensión que reproduce estas distancias. Una vez que *UMAP* ha aprendido la variedad de menor dimensión, se pueden proyectar nuevos datos en esta variedad para obtener los valores en los nuevos ejes, ya sea para visualización o como entrada para otro algoritmo de *Machine Learning.* Además, puede realizar reducción de dimensiones supervisada, lo que significa que, dado un conjunto de datos de alta dimensión con etiquetas, puede aprender una variedad que permite clasificar los casos en grupos (Rhys, 2020).

Al momento de ejecutar el código de *UMAP*, se deben tener en cuenta los cuatro hiperparámetros (Rhys, 2020):

- *n_neighbors:* controla el radio de la región de búsqueda difusa. Valores altos incluyen más vecinos y se enfocan en la estructura global, mientras que valores bajos incluyen menos vecinos y se enfocan en la estructura local.
- *min_dist:* define la distancia mínima permitida entre los casos en la representación de menor dimensión. Valores bajos resultan en incrustaciones más "compactas", mientras que valores altos separan más los casos.
- *metric:* define la métrica de distancia que *UMAP* utilizará para medir distancias en la variedad. Por defecto, usa la distancia euclidiana, pero se pueden usar otras métricas como la distancia Manhattan.
- *n_epochs:* define el número de iteraciones del paso de optimización.

4.4 Selección del algoritmo de agrupamiento —*Clustering*—:

Los algoritmos de *clustering* permiten identificar patrones subyacentes en datos sin etiquetar agrupando observaciones en clusters (Kansal et. al, 2018). El objetivo es encontrar agrupaciones óptimas donde las observaciones dentro de cada clúster sean similares entre sí y distintas de otros clusters. A diferencia del análisis de clasificación, el *clustering* no requiere un conocimiento previo del número de grupos ni de sus características (Rencher y Christensen, 2012). Existen diversas metodologías para el *clustering*, como la evaluación de similitudes basadas en medidas de distancia, la elección preliminar de centros de clusters o la comparación de la variabilidad interna y externa de los clusters. También es posible agrupar variables utilizando la correlación como medida de similitud (Rencher y Christensen, 2012). Existen aproximadamente 26 algoritmos de *clustering* de uso común, que se pueden clasificar en 9 categorías (ver Tabla 1).

Tabla 1. Algoritmos tradicionales

Categoría	Algoritmo típico
Algoritmo de agrupamiento basado en la partición	*K-Means, K-medoids, PAM, CLARA, CLARANS*
Algoritmo de agrupamiento basado en la jerarquía	*BIRCH, CURE, ROCK, Chameleon*

Algoritmo de agrupamiento basado en la teoría difusa	*FCM, FCS, MM*
Algoritmo de agrupamiento basado en distribución	*DBCLASD, GMM*
Algoritmo de agrupamiento basado en densidad	*DBSCAN, OPTICS, Mean-shift*
Algoritmo de agrupamiento basado en la teoría de grafos	*CLICK, MST*
Algoritmo de agrupamiento basado en cuadrícula —grid —	*STING, CLIQUE*
Algoritmo de agrupamiento basado en la teoría fractal	*FC*
Algoritmo de agrupamiento basado en el modelo	*COBWEB, GMM, SOM, ART*

Fuente: Tomado de Xu y Thian (2015).

En esta investigación se empleará el algoritmo de *clustering K-Means*, por ser ampliamente utilizado dada su eficacia y simplicidad en la partición de conjuntos de datos. El proceso de *K-Means* implica inicialmente establecer de forma aleatoria las ubicaciones centrales de los grupos o centroides. Luego, se calculan las distancias de los individuos a estos centroides y se agrupa cada uno en el clúster más cercano. Después, se recalculan los centroides y se repite el proceso de agrupación iterativamente hasta que los centroides se estabilizan, ajustándose para minimizar la distancia euclidiana (Martínez, s.f.).

Uno de los desafíos clave en su aplicación es la determinación del número óptimo de clusters, conocido como K. Esta elección se realiza típicamente mediante el método del codo[1], aunque su efectividad puede verse comprometida si los datos no presentan una clara estructura de agrupación (Kansal et al., 2024; Manimozhi, Ruby y Biruntha, 2024). Es esencial considerar que el valor de K influye significativamente en los resultados del *clustering*. Un K más alto puede generar una segmentación más detallada según las características que tenga el conjunto de datos, pero también puede conducir a una sobre-segmentación, dificultando la interpretación y la implementación práctica. Por otro lado, un K más bajo puede

1 Funciona calculando la suma de los errores al cuadrado —*SSE*— de cada punto de datos con su centroide más cercano para diferentes valores de *K* A medida que *K* aumenta, la SSE disminuye, y el valor óptimo de *K* es donde se produce la mayor disminución en la *SSE*, indicando el punto donde se debe dejar de dividir los datos (Kansal et al., 2018).

simplificar la segmentación, y se corre el riesgo de perder sutilezas en los datos subyacentes. Manimozhi, Ruby y Biruntha, 2024).

Una vez generados los clusters, es importante evaluar la calidad y la coherencia de la agrupación resultante. Esto nos permitirá entender si los clusters capturan adecuadamente la estructura subyacente de los datos y si son útiles para nuestros propósitos. Para ello, es común el uso de las siguientes métricas (Gil, 2023):

- El índice de Davies Bouldin se define como la medida de similitud promedio de cada clúster con su clúster más similar, donde la similitud es la proporción de distancias dentro del clúster a las distancias entre clústeres. El valor mínimo del índice DB es 0, mientras que un valor más pequeño —más cercano a 0— representa un mejor modelo que produce mejores clusters.
- El Índice de Calinski-Harabaz se define como la relación entre la suma de la dispersión entre clústeres y la dispersión dentro de los clústeres. Cuanto mayor sea el índice, más separables serán los clusters.
- La puntuación de silueta es una métrica utilizada para calcular la bondad de ajuste de un algoritmo de agrupamiento, pero también puede ser utilizada para determinar un valor óptimo de k. Su valor oscila entre -1 y 1. Un valor de 0 indica que los clusters se superponen y que los datos o el valor de k son incorrectos. Un valor de 1 es el ideal e indica que los clusters son muy densos y están bien separados.

4.5 Interpretabilidad del modelo

En la mayoría de los escenarios, los algoritmos de *Machine Learning —ML—* se consideran de caja negra; es decir, es un sistema cuyo funcionamiento interno está oculto y no se comprende. Un modelo de IA de caja negra toma una entrada, ejecuta uno o varios algoritmos y produce una salida que puede funcionar, pero cuyo proceso interno permanece oculto (Rothman, 202). Por esta razón, es crucial conocer la justificación detrás de las predicciones para disminuir el riesgo de malas predicciones y los sesgos de interpretabilidad.

Para abordar este desafío, se han desarrollado varias técnicas de interpretabilidad. Entre estas técnicas, los métodos basados en la influencia son a menudo utilizados, especialmente en problemas de clasificación. Estos métodos ayudan a comprender el impacto de las características presentes en el conjunto de datos en la toma de decisiones del modelo. Se prefieren en comparación con otros métodos, ya que ayudan a identificar los atributos dominantes del conjunto de datos, tanto estructurados como no estructurados. Uno de los métodos clave para la interpretabilidad de modelos de ML es la importancia

de las características, que evalúa y puntúa las características de entrada según su relevancia para las predicciones. Este enfoque es particularmente efectivo en datos estructurados, permitiendo eliminar características menos relevantes y obtener una visión clara del comportamiento del modelo. Los mecanismos para determinar esta importancia incluyen permutación, correlación estadística y árboles de decisión (Bhattacharya, 2022).

En el ámbito de la Inteligencia Artificial Explicable —*XAI*—, la importancia de las características es crucial para la selección y reducción de características, mejorando así el rendimiento del modelo. Es esencial validar estas características con expertos, ya que la selección puede variar según el mecanismo y el tipo de modelo. Una forma de determinar la importancia de las características es utilizando los valores de *SHAPley*. Por definición, el valor de *SHAPley* es la "contribución marginal media de cada valor de característica en todos los valores posibles en el espacio de características" (Bhattacharya, 2022). *SHAP —SHAPley Additive exPlanations—* utiliza estos valores de *SHAPley* para proporcionar una explicación consistente y justa de la importancia de cada característica en el modelo. *SHAP* también permite visualizar las interacciones entre características mediante el agrupamiento jerárquico, lo cual ayuda a identificar grupos de características que colectivamente impactan el resultado del modelo. Analizando las características más influyentes con *SHAP*, se pueden observar las características importantes en orden descendente de importancia. No obstante, un inconveniente del gráfico de importancia de características de *SHAP* es que solo considera los valores absolutos medios de *SHAPley,* sin mostrar si ciertas características impactan colectivamente el modelo de manera negativa o positiva (Bhattacharya, 2022).

Además de los gráficos de importancia de características, *SHAP* ofrece varias otras visualizaciones útiles para la interpretabilidad del modelo (Bhattacharya, 2022):

- Mapas de calor *SHAP*: muestran cómo cada valor de característica impacta positiva o negativamente el resultado del modelo, aunque pueden ser difíciles de interpretar para usuarios no técnicos.
- Gráficos de resumen *SHAP*: representan tanto las características importantes como el rango de efectos de estas características, usando colores para resaltar impactos positivos y negativos.
- Gráficos de dependencia *SHAP*: revelan la variación del resultado del modelo según características específicas, ayudando a detectar interacciones entre valores de características.
- Gráficos de barras *SHAP* para interpretabilidad local: permiten analizar el impacto positivo y negativo de las características en los datos de inferencia.

- Gráficos en cascada *SHAP*: alternativa visual a los gráficos de barras, mostrando la importancia de las características y sus impactos sin centrarse en cero.
- Diagramas de fuerza *SHAP*: visualizan la predicción del modelo y el impacto de cada característica en comparación con un valor base.
- Gráficos de decisión *SHAP*: comparan los valores de las características de los datos de inferencia con los valores promedio del modelo entrenado, mostrando influencias positivas y negativas.

Para lograr estas visualizaciones, es fundamental elegir el algoritmo explicativo adecuado. *SHAP* proporciona varios algoritmos explicativos que se aplican según el tipo de modelo y datos (Bhattacharya, 2022):

- *TreeExplainer:* para modelos basados en árboles, como *Random Forest*, *XGBoost*, *LightGBM,* y CatBoost. Utiliza una implementación rápida del algoritmo *Tree SHAP* para calcular valores de *SHAPley*.
- *DeepExplainer:* diseñado para modelos de aprendizaje profundo entrenados con datos no estructurados, como imágenes y texto. Se basa en una versión modificada del algoritmo *DeepLIFT.*
- *GradientExplainer:* explica modelos de aprendizaje profundo usando el concepto de gradientes esperados, una extensión de Gradientes Integrados y *SmoothGrad.*
- *KernelExplainer:* proporciona explicabilidad independiente del modelo utilizando un enfoque de regresión lineal local ponderada, similar a *LIME* pero con un enfoque en los valores de *SHAPley*.
- *LinearExplainer:* diseñado para modelos lineales, analiza correlaciones entre características y calcula valores de *SHAPley* eficientes para estos modelos.

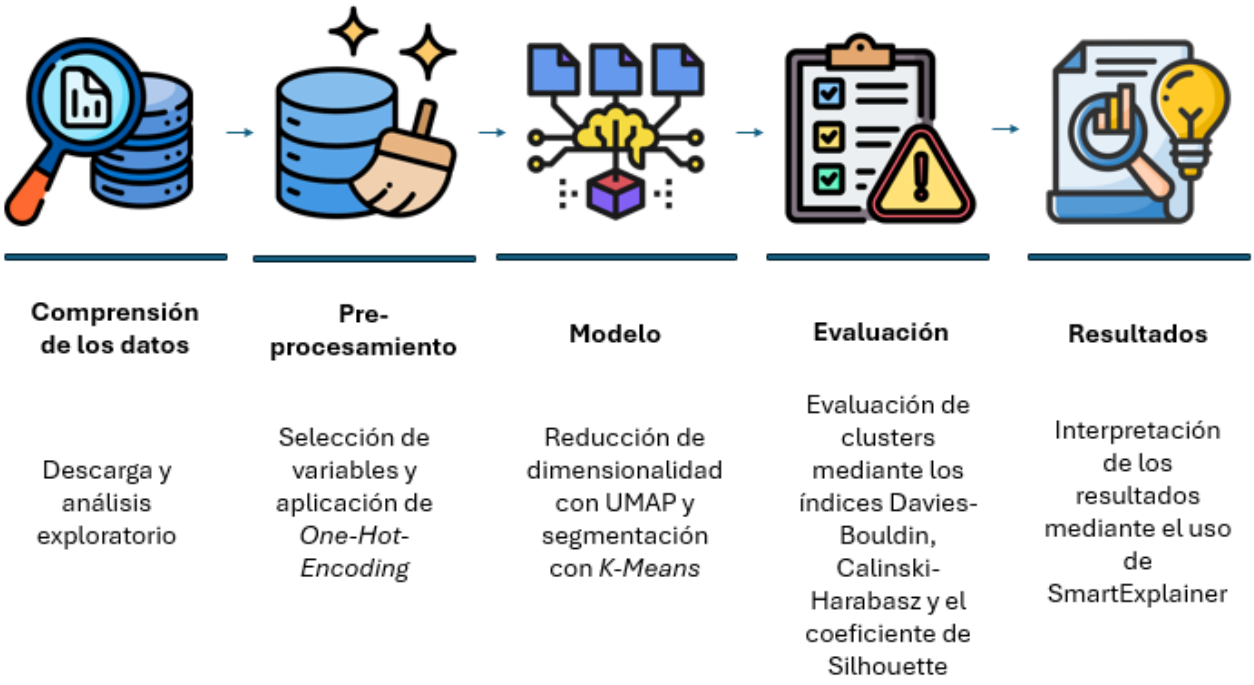
En resumen, los métodos basados en la influencia, junto con herramientas avanzadas como *SHAP* y sus diversos explicadores, mejoran la transparencia y fiabilidad del modelo, permitiendo una interpretación más profunda y aumentando la confianza en las predicciones de *Machine Learning.*

5. METODOLOGÍA

En esta investigación se empleó la metodología *CRISP-DM* para estructurar y guiar el análisis de datos en cada una de las etapas del proyecto. A través de esta metodología, se realizó el preprocesamiento de datos categóricos, se implementaron técnicas de *clustering* como *K-Means* y se aplicaron modelos predictivos utilizando *LightGBM,* los cuales fueron complementados por análisis interpretativos

mediante *SHAP*. El flujo de procesos seguido para desarrollar la metodología se ilustra en el siguiente gráfico:

Grafico 1. Flujo de procesos metodológico



Fuente: elaboración propia.

A continuación, se describe detalladamente cada una de las etapas:

5.1 Comprensión de los datos:

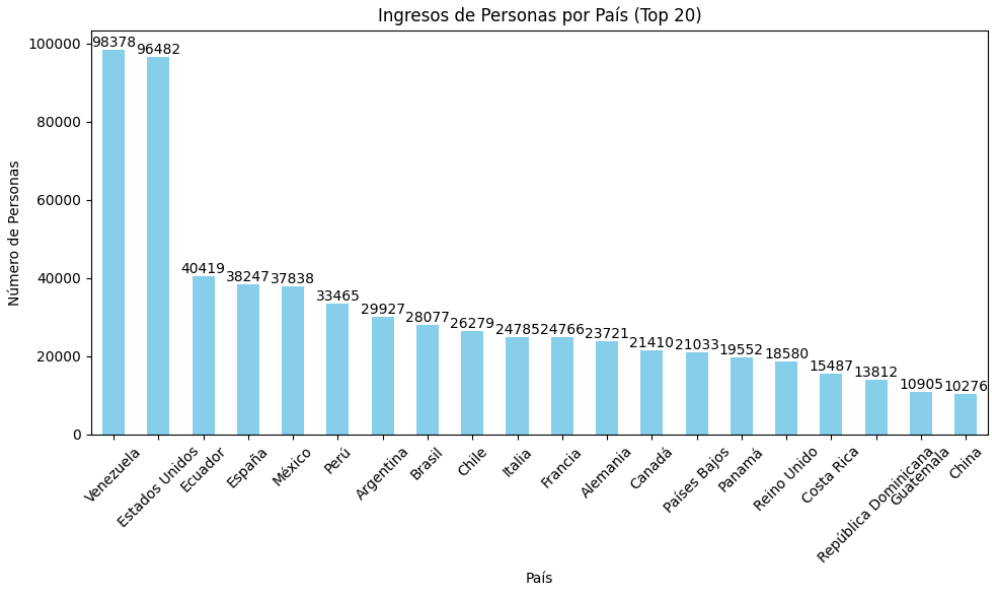
La base de datos utilizada se obtuvo de los tableros de visualización creados por Migración Colombia para analizar los flujos migratorios de personas que ingresan y salen del país. Estos tableros se construyen a partir de la base de datos de "Entrada y Salida de Personas del País", la cual se recolecta mediante la plataforma PLATINUM —Sistema de Plataforma de Información Migratoria— empleada en cada uno de los 46 puestos de control migratorio (Migración Colombia, 2022c). La información disponible abarca desde 2012 hasta 2024; sin embargo, para este análisis, solo se consideraron los datos correspondientes al año 2023, de enero a diciembre.

La información se puede desglosar en extranjeros y colombianos, pero para este análisis, se seleccionaron exclusivamente a los extranjeros que ingresaron a Colombia, excluyendo a los colombianos que reingresaron al país. El conjunto de datos contiene únicamente datos categóricos, con un total de 13

variables y 823,290 filas. No todas las 13 variables fueron incluidas en el análisis; se emplearon las siguientes después del preprocesamiento: mes de llegada a Colombia, región de origen del inmigrante, motivo del viaje a Colombia, región de hospedaje y medio de transporte utilizado para ingresar a Colombia. Cada variable posee subcategorías que fueron estandarizadas para homogenizar los valores y facilitar el análisis. Este proceso se explicará en el apartado de preprocesamiento de datos.

Preliminarmente, en el análisis exploratorio se encontró que las personas de origen venezolano y estadounidense son los que más ingresan a Colombia, siendo el Turismo la razón principal y Bogotá D.C. la ciudad a la que más se dirigen —como se puede ver en el Gráfico 2—.

Gráfico 2. Cantidad de personas que ingresan al país por nacionalidad



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Para obtener un análisis más profundo de las personas que ingresaron al país en 2023, se empleará un modelo de *clustering* —*K-Means*— para detectar patrones y grupos que no son fácilmente visibles mediante un análisis exploratorio. Antes de aplicar el modelo, es necesario realizar el preprocesamiento de los datos.

5.2 Pre-procesamiento de los datos

Las variables del conjunto de datos original eran las siguientes:

Tabla 2. Selección de variables

Variables	Observación
Centro Regional	Eliminada porque no se considera relevante para el análisis
Puesto Migratorio	Eliminada porque no se considera relevante para el análisis
Tipo Transporte	Se mantiene para el análisis. Contiene 5 valores. Sin embargo, fue modificada para reducirla a dos categorías: "terrestre" y "otros"
Ciudad Hospedaje	Eliminada porque se construyó otra variable que agrupara los departamentos y las ciudades: “Región de Hospedaje”
Entrada/Salida	Eliminada porque desde el filtro inicial en el tablero de Migración Colombia se seleccionó solo las Entradas
Meses	Se mantiene sin modificación. Contiene todos los meses del año
Motivo Viaje	Se mantiene para el análisis. Contiene 52 valores. Sin embargo, fue modificada para reducirla a 9 valores. Ver anexo 5.
País Nacionalidad	Eliminada porque se conserva la variable Región Nacionalidad
Rango Edad	Se mantiene. Contiene los siguientes grupos etarios: '0-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70 o más. Sin embargo, los valores se convirtieron en datos ordinales y se eliminaron valores inconsistentes.
Colombiano/Extranjero	Se eliminó porque desde la descarga ya venía predeterminadamente el valor “Extranjero”
Departamento Hospedaje	Eliminada porque se construyó otra variable que agrupara los departamentos y las ciudades: “Región de Hospedaje”
Región Nacionalidad	Se mantiene para el análisis. Contiene 10 valores. Sin embargo, fue modificada para reducirla a 7 valores: América Central y el Caribe, América del Norte, América del Sur, Asia, Europa, Oceanía y Otros
Año.	Se eliminó porque desde la ingesta ya venía predeterminadamente el año 2023

Al agrupar las variables antes de convertirlas a dummies, se reduce la dimensionalidad del conjunto de datos, se simplifica el modelo, se mejora el rendimiento computacional y se disminuye el ruido. Esto facilita la identificación de patrones importantes y evita el sobreajuste, haciendo que el proceso de modelado sea más eficiente y efectivo. A su vez, debido al alto costo computacional de procesar 823,290 filas, se seleccionó una muestra aleatoria significativa de 50,000 filas que permitió hacer una prueba de

concepto para la implementación del modelo. Para ver el preprocesamiento de los datos, consulte el Anexo 5.

El proceso de creación de variables dummies, formalmente conocido como *one-hot encoding*, consiste en transformar las variables categóricas en variables binarias —0 o 1—. Para cada categoría de una variable nominal, se crea una nueva columna en el conjunto de datos. Cada fila tendrá un valor de 1 en la columna correspondiente a su categoría original y 0 en las demás. Este método asegura que el modelo pueda manejar adecuadamente las variables categóricas sin introducir un orden ficticio o jerarquía.

Finalmente, una vez depurada la base de datos de valores nulos e inconsistentes y aplicando el método de *one-hot encoding*, la base de datos con la que se realizará el modelo quedó con 50,000 filas y 35 variables. De estas, 34 fueron nominales y una fue ordinal, correspondiente al rango de edad:

Tabla 3. Variables del modelo

Rango de edad	Mes enero	Mes febrero	Mes marzo
Mes mayo	Mes junio	Mes julio	Mes agosto
Mes septiembre	Mes octubre	Mes noviembre	Mes diciembre
Región de origen: América Central y el Caribe	Región de origen: América del Norte	Región de origen: América del Sur	Región de origen: Asia
Región de origen: Europa	Región de origen: Oceanía	Región de origen: otros	Motivo de viaje: casos específicos y otros
Motivo de viaje: estudios y prácticas	Motivo de viaje: eventos y conferencias	Motivo de viaje: familia y relaciones maritales	Motivo de viaje: negocios y trabajo
Motivo de viaje: protección y situaciones especiales	Motivo de viaje: salud y tratamiento médico	Motivo de viaje: tránsito y movilidad	Motivo de viaje: turismo y viajes
Región de hospedaje: región Amazónica	Región de hospedaje: región Andina	Región de hospedaje: región Caribe	Región de hospedaje: región Orinoquía
Región de hospedaje: región Pacífica	Transporte: otros	Transporte: terrestre	

5.3 Modelo

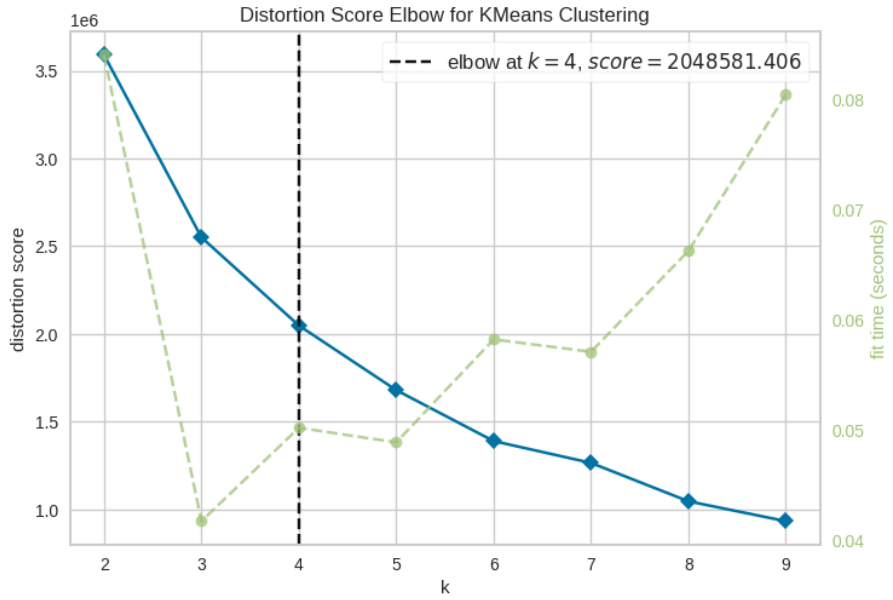
Inicialmente, se procede a reducir la dimensionalidad del conjunto de datos que ha sido transformado mediante *one-hot encoding*, el cual, como se mencionó anteriormente, quedó compuesto por 35 variables y 50 mil filas. Esto facilitará un entrenamiento más eficiente y permitirá encontrar una solución óptima. Para este propósito, se utilizó *UMAP*, el cual es un algoritmo que preserva la estructura global de los datos y tiene un buen rendimiento de tiempo de ejecución, lo que lo hace ideal para identificar las características más relevantes (Holbert, 2022).

El proceso de reducción se realizó configurando *UMAP* para reducir las dimensiones a dos componentes, utilizando la métrica de Manhattan para medir distancias, dado que predeterminadamente se estaba calculando la distancia con la métrica euclidiana y, por la naturaleza de los datos, al ser 100% categóricos, no arrojaba los mejores resultados. Además, se estableció un estado aleatorio fijo para asegurar la reproducibilidad de los resultados. Posteriormente, se aplicó el algoritmo al conjunto de datos, obteniendo un nuevo conjunto de datos con dimensiones reducidas.

A partir del nuevo conjunto de datos, previamente transformado mediante *one-hot encoding* y con dimensiones reducidas, se aplica el modelo de *K-Means* para identificar los clusters. Sin embargo, antes de aplicar el modelo, es crucial determinar el número óptimo de clusters —K— para asegurar una clasificación adecuada de los datos. El proceso se lleva a cabo de la siguiente manera:

- Instanciación del modelo de *clustering* y del visualizador: se instancia el modelo de *K-Means* utilizando la inicialización "*K-Means*++" y se establece un estado aleatorio fijo para garantizar la reproducibilidad de los resultados. Además, se crea un visualizador *KElbowVisualizer* para identificar el número óptimo de clusters K en un rango de 2 a 10.
- Ajuste de los datos al visualizador: el visualizador se ajusta al conjunto de datos reducido para analizar la variación de la inercia en función del número de clusters. Esto permite evaluar cómo cambia la suma de distancias cuadradas dentro de los clusters a medida que se incrementa K.
- Visualización del gráfico: el visualizador genera un gráfico que ayuda a identificar el punto en el cual la inercia deja de disminuir significativamente al aumentar el número de clusters. Este punto óptimo sugiere que el número ideal de clusters K es 4 para el modelo de *K-Means*, como se muestra en el Gráfico 3:

Gráfico 3. Determinación del número óptimo de clusters mediante el método del codo

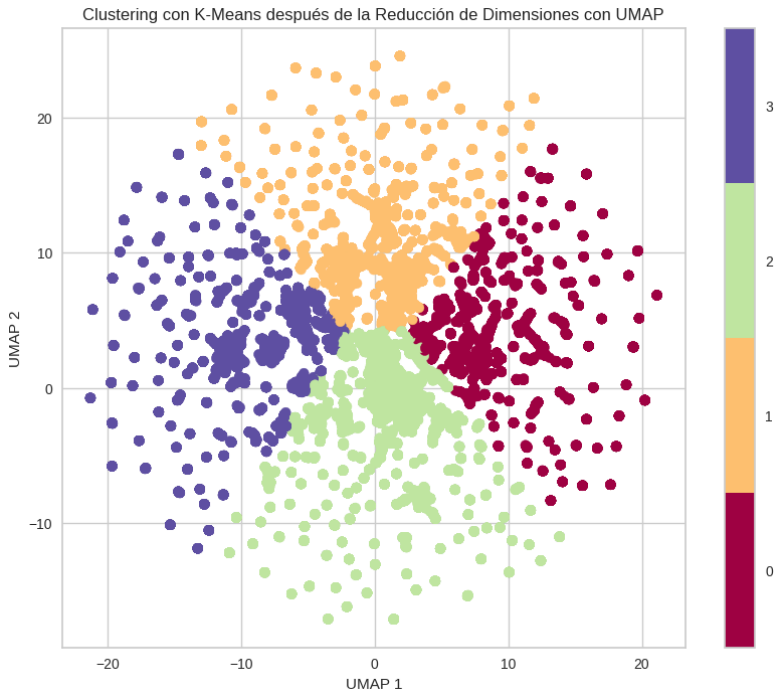


Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Con el número óptimo de clusters identificado como 4, se procede a crear el modelo de *K-Means* configurando varios parámetros importantes. Se establece el número de clusters en 4 y se utiliza la inicialización "*K-Means*++" para mejorar la convergencia. Además, se especifica que el algoritmo se ejecute 10 veces con diferentes centroides iniciales para seleccionar la mejor solución y se fija el número máximo de iteraciones en 100 para asegurar la convergencia. Se establece un estado aleatorio fijo para garantizar la reproducibilidad de los resultados. Posteriormente, se ajusta el modelo de *K-Means* al conjunto de datos reducido con *UMAP* y se predicen los clusters. Este proceso asigna cada punto de datos reducido a uno de los 4 clusters, permitiendo un análisis más detallado y segmentado de los patrones en el conjunto de datos.

Para visualizar los resultados del *clustering*, se generó un gráfico de dispersión que muestra cómo se agrupan los puntos en el espacio reducido. En el Gráfico 4, se puede ver como cada punto representa una observación y su color indica el clúster asignado. Utilizando las coordenadas de *UMAP*, se aprecia cómo las observaciones se distribuyen y agrupan en el espacio bidimensional. Este gráfico facilita la interpretación visual de los clusters formados, destacando la efectividad del proceso de reducción de dimensionalidad y el agrupamiento realizado.

Gráfico 4. Visualización de clusters con *K-Means* tras Reducción de Dimensiones con *UMAP*



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

5.4 Evaluación del modelo de *clustering*

Para evaluar la calidad del modelo de *clustering*, se utilizaron varias métricas: el índice Davies-Bouldin, el índice Calinski-Harabasz y el coeficiente de Silhouette. Los resultados obtenidos son los siguientes:

- El índice Davies-Bouldin arrojó un resultado de 0.93. Dado que este índice mide la compacidad y separación de los clusters, un valor más bajo indica clusters más compactos y mejor separados. Este valor sugiere que los clusters están bien separados y son relativamente compactos, lo cual es una indicación positiva de la calidad del agrupamiento.
- El índice Calinski-Harabasz evalúa la dispersión de los clusters, donde un valor más alto indica una mejor definición de los clusters. Como se obtuvo un valor de 29734.13 indica una definición clara de los clusters, sugiriendo que los datos están bien agrupados. Este alto valor demuestra que los clusters son internamente coherentes y externamente separados.
- El coeficiente de *Silhouette* calcula qué tan similares son los objetos dentro de los mismos clusters en comparación con los objetos de otros clusters. Un valor cercano a 1 indica clusters bien diferenciados, mientras que un valor cercano a -1 indica que los puntos pueden haber sido

asignados al cluster incorrecto. El resultado que se obtuvo al calcularlo fue de 0.33, lo cual sugiere que los clusters están razonablemente definidos, aunque hay margen de mejora en la separación y compacidad de los clusters. Este valor intermedio indica que, aunque los clusters son útiles, podría haber superposición entre algunos de ellos.

En resumen, las métricas de evaluación sugieren que los clusters formados están bien definidos y separados. Estos resultados proporcionan una base sólida para la interpretación adicional utilizando métodos *SHAP*.

5.5 Interpretación de los resultados con *SmartExplainer* de *SHAPash*

Para interpretar los clusters formados, se entrenó un clasificador *LightGBM* utilizando el conjunto de datos previamente transformado mediante *one-hot encoding* y con dimensiones reducidas, utilizando las etiquetas de los clusters generadas por *K-Means* como las etiquetas objetivo. El modelo *LightGBM* es bastante potente y funciona bien con variables categóricas y numéricas (Gil, 2023). El contenido de cada clúster fue posible visualizarlo gracias al uso de *SmartExplainer* de *SHAPash*, el cual consiste en reagrupar las características que comparten similitudes para identificar qué tema es importante y lo hace utilizando el *backend* de *SHAP* para calcular las contribuciones (*SHAPash*, s.f.). Este enfoque permitió evaluar la importancia de las características y cómo contribuyen a la formación de cada clúster, además es especialmente útil para este tipo de conjunto de datos que son completamente categóricos y poseen muchas características. A continuación, se describe el proceso y los resultados obtenidos:

- Se procedió a entrenar un clasificador *LightGBM* —LGBMClassifier— utilizando el conjunto de datos preprocesado, que incluye las características transformadas mediante *one-hot encoding* y reducidas dimensionalmente mediante *UMAP*. Las etiquetas objetivo fueron los clusters identificados por el modelo *K-Means*. El propósito del entrenamiento fue evaluar la importancia relativa de las características en la formación de cada cluster. Durante el entrenamiento, *LightGBM* utilizó la configuración automática de multi-threading y procesó un total de 74 bins con 50,000 puntos de datos y 35 características utilizadas.
- Utilizando *SHAPash*, se creó un *SmartExplainer* que toma como entrada el clasificador *LightGBM* previamente entrenado. Para facilitar el análisis de las características y mejorar la interpretación de los resultados, se realizó un reagrupamiento de las variables en categorías temáticas. Este proceso se llevó a cabo de la siguiente manera:
    - ✓ Agrupación de características: las variables se agruparon en conjuntos temáticos basados en su naturaleza y relevancia común. Por ejemplo, todas las variables relacionadas con

los meses del año se agruparon bajo la categoría "Mes", mientras que las variables que indicaban la región de origen de los individuos se agruparon bajo "Región de origen". Otros grupos definidos incluyeron "Motivo del viaje", "Región de hospedaje" y "Medio de transporte".

- ✓ Etiquetado de los grupos de características: a cada grupo de características se le asignó una etiqueta descriptiva que refleja su contenido temático. Esto incluye etiquetas como "Mes de viaje", "Región de origen", "Motivo del viaje", "Región de hospedaje" y "Medio de transporte". Estas etiquetas facilitan la interpretación y el análisis de los resultados.
- ✓ Creación del *SmartExplainer*: una vez definidos los grupos de características y sus respectivas etiquetas, se procedió a crear un *SmartExplainer* con *SHAPash*. Este paso involucró la inicialización del *SmartExplainer* con el modelo de *LightGBM* y la estructura de agrupación de características previamente definida.
- ✓ Compilación del *SmartExplainer*: el *SmartExplainer* se compiló utilizando el conjunto de datos que había sido transformado mediante *one-hot-encoding* y reducido dimensionalmente, junto con las predicciones de los clusters generadas por el modelo *K-Means*. Este proceso permite que *SHAPash* comprenda cómo cada característica del conjunto de datos contribuye a la formación de los clusters. Los resultados se entregan con la reagrupación de las características en categorías temáticas, facilitando así la interpretación y el análisis detallado.

5.6 Resultados

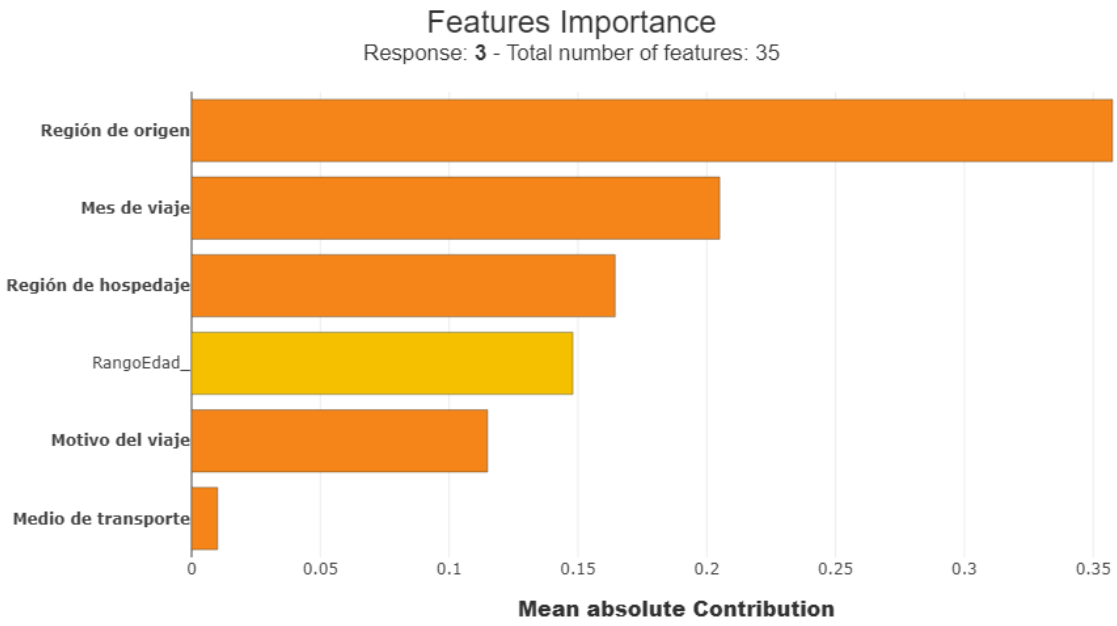
Gracias a los *SmartExplainer* creados, se obtuvieron los resultados correspondientes a cada clúster, entregados con la reagrupación de características en categorías temáticas. Cabe destacar que, por defecto, *SmartExplainer* grafica el último clúster —el clúster 3—, por lo que fue necesario asignar temporalmente el nombre "clúster 3" a cada uno de los clústeres de manera secuencial para generar los gráficos correspondientes a cada uno de ellos.

5.6.1 Análisis del clúster 0

Este clúster se distingue principalmente por la región de origen, el mes de viaje y la región de hospedaje. La mayoría de los viajeros provienen de América del Sur, indicando la importancia de la proximidad geográfica y las relaciones sociopolíticas en los patrones de los flujos migratorios. Los picos migratorios en enero y agosto sugieren una fuerte estacionalidad, vinculada posiblemente a vacaciones y recesos

laborales. Las personas que pertenecen a este clúster prefieren alojarse en la Región Andina, seguida por la Región Caribe. Ver Gráfico 5:

Gráfico 5. Clúster 0



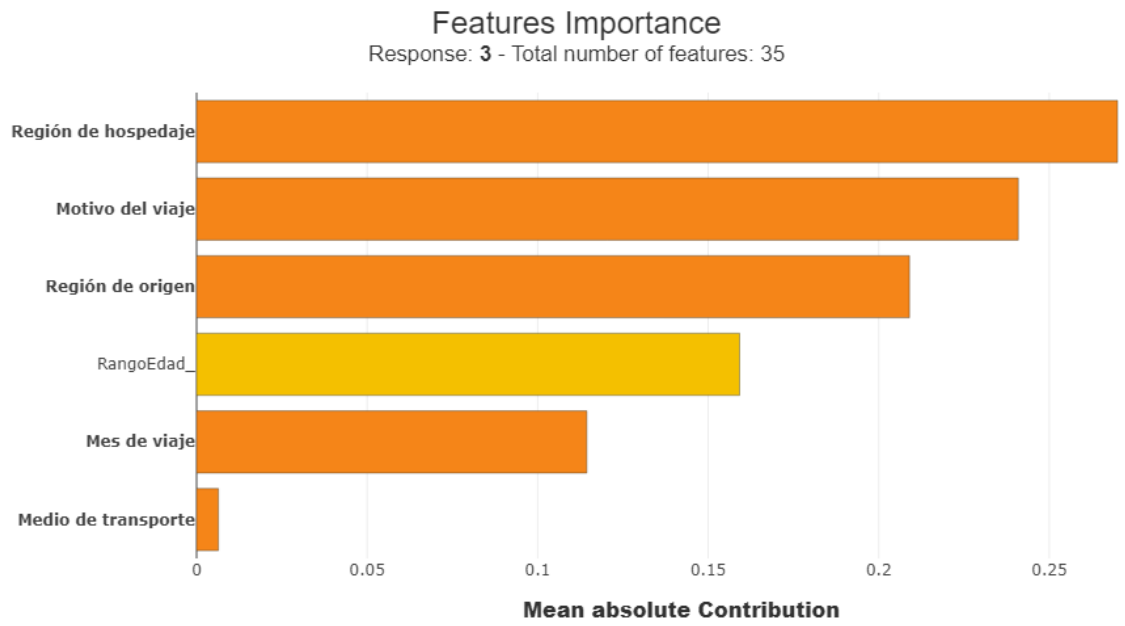
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Nota: el eje horizontal muestra la contribución media absoluta de cada característica a la predicción del modelo. Cuanto mayor es el valor, más importante es la característica para el modelo. (Para conocer cada gráfico de la importancia de las características de cada grupo consultar el Anexo 2).

5.6.2 Análisis del clúster 1

En el clúster 1, la región de hospedaje emerge como la variable más influyente, con una marcada preferencia por la Región Pacífica y siendo el turismo el principal motivo de viaje, recibiendo a más personas de origen norteamericano. Sin embargo, también se puede observar que los negocios y el trabajo influyen de alguna manera en la segmentación de esta población demográfica. Ver Gráfico 6:

Gráfico 6. Clúster 1



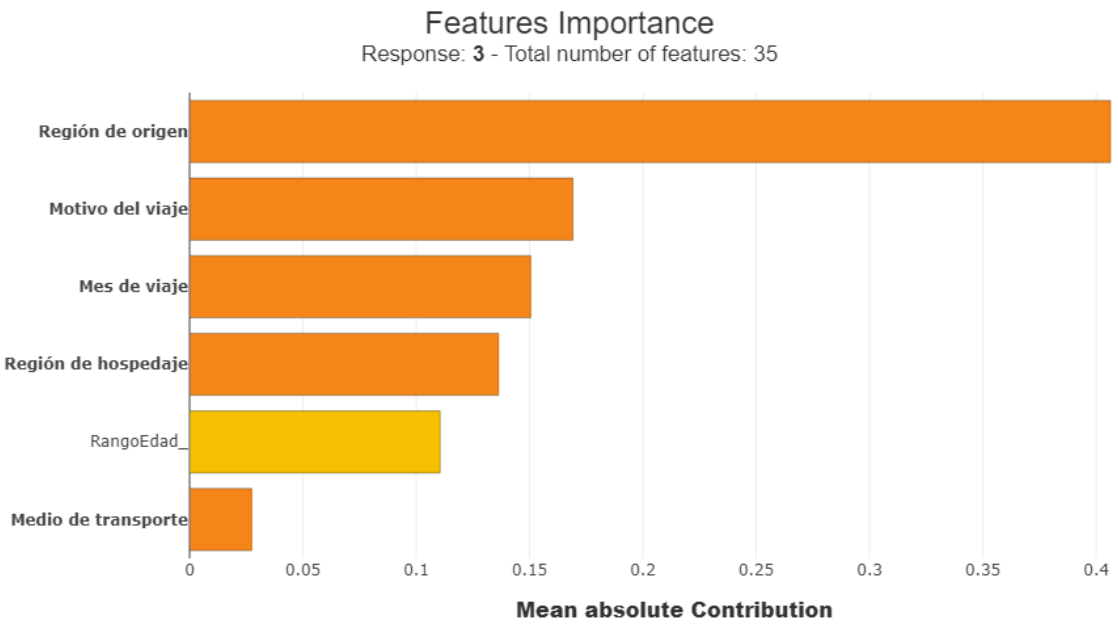
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Nota: el eje horizontal muestra la contribución media absoluta de cada característica a la predicción del modelo. Cuanto mayor es el valor, más importante es la característica para el modelo. (Para conocer cada gráfico de la importancia de las características de cada grupo consultar el Anexo 3).

5.6.3. Análisis del clúster 2

La región de origen y el motivo del viaje son las variables más significativas en este clúster. Predominan las personas de América del Sur, seguidos por América del Norte y América Central y el Caribe, reafirmando la proximidad geográfica como un factor clave. El turismo y los viajes son los principales motivos del flujo migratorio, con una contribución notable de negocios y trabajo. Los meses de noviembre y diciembre muestran picos de afluencia, probablemente relacionados con festividades y vacaciones. Ver Gráfico 7:

Gráfico 7. Clúster 2

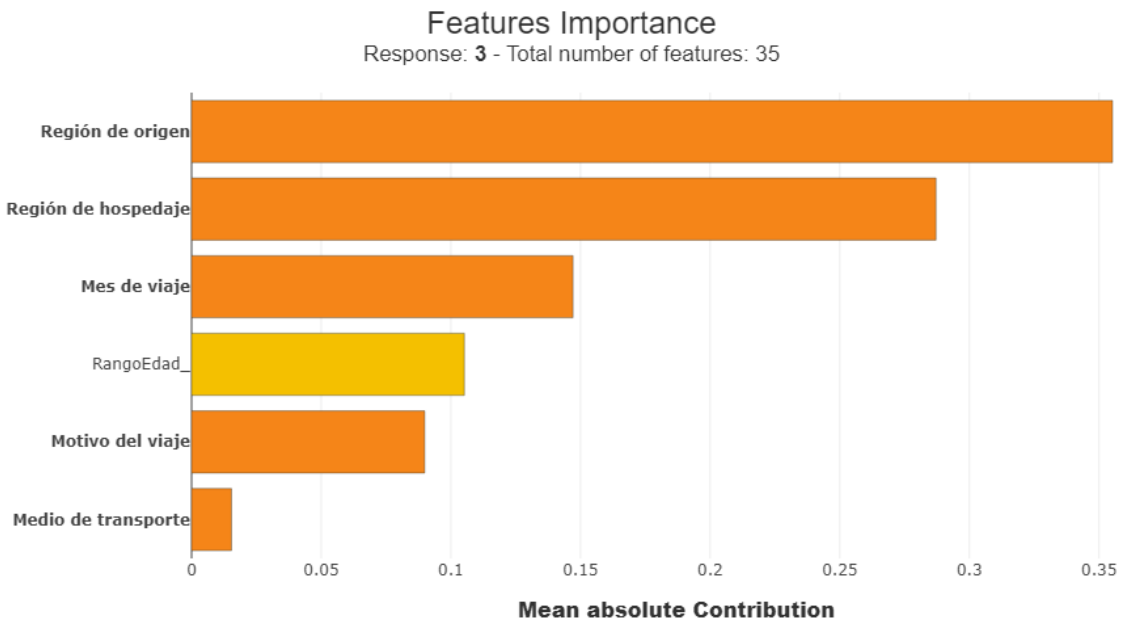


Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación. *Nota:* el eje horizontal muestra la contribución media absoluta de cada característica a la predicción del modelo. Cuanto mayor es el valor, más importante es la característica para el modelo. (Para conocer cada gráfico de la importancia de las características de cada grupo consultar el Anexo 4).

5.6.4. Análisis del clúster 3.

Este clúster se caracteriza por la región de origen, la región de hospedaje y el mes de viaje. La mayoría de las personas que ingresan al país provienen de América Central y el Caribe y América del Sur. La Región Caribe es el destino preferido, seguida por la Región Pacífica. Los picos de los flujos migratorios en agosto y diciembre podrían estar vinculados con vacaciones y festividades de fin de año. Ver Gráfico 8:

Gráfico 8. Clúster 3



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación. *Nota:* el eje horizontal muestra la contribución media absoluta de cada característica a la predicción del modelo. Cuanto mayor es el valor, más importante es la característica para el modelo. (Para conocer cada gráfico de la importancia de las características de cada grupo consultar el Anexo 5).

5.6.5. Características comunes y diferenciadoras

Una de las características más consistentes en todos los clústeres es el motivo del viaje, siendo "Turismo y viajes" la razón predominante para los desplazamientos en todos los segmentos de la población analizada. Asimismo, el medio de transporte terrestre es el más común en todos los clústeres. Las características de mes de viaje y región de hospedaje también son recurrentes, subrayando su importancia en la caracterización de los grupos.

En contraste, las diferencias más notables entre los clústeres se observan en las regiones de origen y hospedaje. Estas diferencias son clave para entender las particularidades de cada clúster.

Región de origen:

- Clúster 0 y 2: predominan las personas que vienen de América del Sur.
- Clúster 1: predomina América del Norte como región de origen.
- Clúster 3: América Central y el Caribe son las regiones más comunes.

Región de hospedaje:

- Clúster 0: la región Andina es la principal área de hospedaje.
- Clúster 1 y 2: la región Pacífica destaca como destino principal.
- Clúster 3: la región Caribe emerge como el destino predominante.

Si bien, aunque hay características comunes como el motivo del viaje y el medio de transporte que influyen uniformemente en todos los clústeres, las regiones de origen y hospedaje son las variables que más varían entre ellos. Estas diferencias proporcionan una comprensión matizada de los patrones de movilidad y preferencias de las personas que entran a Colombia.

6. CONCLUSIONES

En esta investigación se consiguió una segmentación exitosa de los flujos migratorios en Colombia, término utilizado por Migración Colombia para describir el conjunto de datos analizado, revelando patrones y características distintivas de las personas que ingresaron al país en 2023. Utilizando la metodología *CRISP-DM*, se realizó un análisis detallado desde el preprocesamiento de datos hasta la aplicación de técnicas de *clustering* como *K-Means*, y la interpretación de resultados mediante *LightGBM* y *SmartExplainer* de *SHAPash*. Los resultados identificaron cuatro clusters principales, cada uno con características únicas en términos de región de origen, motivos de viaje y preferencias de hospedaje, destacando la importancia de factores como la proximidad geográfica y la estacionalidad.

Las variables más influyentes en la segmentación fueron la región de origen, el motivo del viaje, la región de hospedaje y el mes de llegada. Dado que el motivo de viaje más común fue “turismo y viajes”, no se pueden considerar que son inmigrantes dado que tienen una instancia corta en el país — suponiendo que la declaración del motivo de viaje sea verídica al momento de entrar al país—. Esto lleva a concluir que, para analizar otro tipo de migraciones, la información que es recolectada en puntos migratorios oficiales captura en su mayoría a personas turistas y que no se está teniendo en cuenta los flujos migratorios por puntos migratorios no autorizados, dado que allí no es posible tener el control migratorio.

La aplicación de *one-hot-encoding* para transformar las variables categóricas en variables binarias, la reducción de dimensionalidad mediante *UMAP* y la aplicación de *K-Means* demostraron ser efectivas para descubrir patrones complejos en los datos categóricos. Las métricas de evaluación indicaron que los clusters están bien definidos y separados, proporcionando una base sólida para la interpretación adicional y confirmando la efectividad del modelo utilizado.

Por otro lado, el uso de *SmartExplainer* de *SHAPash* junto con *LightGBM* permitió una interpretación detallada de las características influyentes en cada clúster, facilitando la comprensión de los patrones en los flujos migratorios. Además, los resultados se entregaron con la reagrupación de características en categorías temáticas, lo que mejoró aún más la interpretación. En resumen, esta investigación no solo segmentó adecuadamente los flujos migratorios en Colombia, sino que también proporcionó insumos para interpretar datos 100% categóricos utilizando técnicas de *Machine Learning.*

7. RECOMENDACIONES

Con base en los hallazgos de esta investigación, se sugiere ampliar el análisis a períodos de tiempo más largos para futuras investigaciones sobre los flujos migratorios en Colombia. Esto permitiría observar tendencias a lo largo del tiempo y comprender mejor cómo evolucionan los patrones migratorios. Analizar datos de varios años proporcionará una visión más completa y permitirá identificar variaciones significativas entre diferentes años. Además, se recomienda incluir variables numéricas en los conjuntos de datos para facilitar la creación de visualizaciones convencionales de *SHAP*, lo que mejorará la interpretación de los modelos de *Machine Learning* y proporcionará una comprensión más clara de las características que influyen en los flujos migratorios.

Para profundizar en otros tipos de migración y obtener una comprensión más detallada, se recomienda excluir del análisis el motivo de viaje "Turismo y viajes". Esto permitirá concentrarse en los flujos migratorios relacionados con la migración laboral, familiar, educativa y otras razones. Al eliminar los datos de turistas, se podrán identificar y analizar patrones específicos de los migrantes que permanecen en el país por períodos más prolongados y comprender mejor sus características.

REFERENCIAS

1. Aggarwal, C. C. y Reddy, C. K. (Ed). (2015). *Data Clustering: Algorithms and Applications.* Chapman and Hall/CRC.
2. Akın, D. y Dökmeci, V. (2014). Cluster Analysis of Interregional Migration in Turkey. *Journal of Urban Planning and Development, 141*(3). https://doi.org/10.1061/(ASCE)UP.1943-5444.0000223
3. Bhattacharya, A. (2022). *Applied Machine Learning Explainability Techniques.* Packt Publishing.
4. Bored ASD. (15 de febrero de 2021). Técnicas de aprendizaje no supervisado. *Medium.* https://parnerdat.medium.com/tecnicas-de-aprendizaje-no-supervisado-a2a7b2809156
5. Cancillería de Colombia. (s.f.). *Antecedentes históricos y causas de la migración.* https://www.cancilleria.gov.co/colombia/migracion/historia#:~:text=Colombia%2C%20hist%C3%B3ricamente%2C%20se%20ha%20caracterizado,la%20d%C3%A9cada%20de%20los%20a%C3%B1os
6. Departamento Administrativo Nacional de Estadística (DANE). (2022). *Reporte estadístico de Migración.* https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/estadisticas-de-migracion
7. Global Migration Group. (2017). Handbook for Improving the Production and Use of Migration Data for Development. KNOMAD.
8. Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition.* O'Reilly Media, Inc.
9. Hoffman P. K, y Luengo-Oroz, M. (2023). Predictive Modelling Of Movements of Refugees and Internally Displaced People: Towards A Computational Framework. *Journal of Ethnic and Migration Studies, 49*(2), 408-444. https://doi.org/10.1080/1369183X.2022.2100546
10. Holbert, C. (05 de junio de 2022). PCA, *t-SNE*, and *UMAP* Classification of Vegetable Oils. Charles Holbert Blog. https://www.cfholbert.com/blog/pca-tsne-*UMAP*/
11. Manimozhi, S., Ruby, D. y Biruntha, K. (2024). *An Elegant Evaluation: Triangulating Clustering Methods for Customer Segmentation.* 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE) (pp. 2-6).
12. Kansal, T., Bahuguna, S., Singh, V. y Choudhury, T. (2018*). Customer Segmentation Using K-Means Clustering.* International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) (pp. 135-139).

13. Martínez, I. (s.f.). *Tema 4. Agrupación o clustering.* Universidad de Valencia. https://www.uv.es/mlejarza/datamine/T4.pdf
14. McInnes, L., Healy, J. y Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* https://*UMAP*-learn.readthedocs.io/en/latest/*clustering*.html#using-*UMAP*-for-*clustering*
15. Migración Colombia. (2022a). *Así se construyen nuestros datos estadísticos: flujos migratorios.*
16. Migración Colombia (2022b). *Boletín estadístico anual de flujos migratorios 2022.*
17. Migración Colombia. (2022c). *Ficha metodológica. Registros administrativos. operación estadística: "entrada y salida de personas del país".*
18. Molina, M. D., Chau, N., Rodewald, A. D. y Garip, F. (2022). How to Model The Weather-Migration Link: A Machine-Learning Approach to Variable Selection in The Mexico-U.S. Context. *Journal of Ethnic and Migration Studies, 49*(2), 465–491. https://doi.org/10.1080/1369183X.2022.2100549
19. Recaño-Valverde. J., Sánchez-Barriga, C., Martínez-García, J. y Rivera-Sepúlveda, V. N. (2012). *Una nueva base de datos para la estimación de los flujos migratorios internacionales de Colombia: Metodología y resultados comparativos.* Departamento Administrativo Nacional de Estadística (DANE).
20. Rencher, A. C. y Christensen, W. F. (2012). *Methods of Multivariate Analysis.* John Wiley & Sons.
21. Rothman, D. (2020). *Hands-On Explainable AI (XAI) with Python.* Packt Publishing.
22. Rhys, H. (2020). *Machine Learning with R, the Tidyverse, and Mlr.* Manning Publications. https://learning-oreilly-com.ezproxy.eafit.edu.co/library/view/machine-learning-with/9781617296574/OEBPS/Text/kindle_split_026.html#ch14lev1sec2
23. Sancho, F. (15 de diciembre de 2023). Algoritmos de *Clustering*. *MatematIA. https://www.cs.us.es/~fsancho/Blog/posts/Clustering/*
24. *SHAPash*. (s.f.). Groups of Features. https://*SHAPash*.readthedocs.io/en/latest/overview.html
25. Vijay, H. (18 de mayo de 2023). Dimensionality Reduction: PCA, tSNE, *UMAP*. *Auriga.* https://aurigait.com/blog/blog-easy-explanation-of-dimensionality-reduction-and-techniques/
26. Xu, D. y Tian Y. (2015). A Comprehensive Survey of *Clustering* Algorithms. *Annals of Data Science,* 2, 165-193. https://link.springer.com/article/10.1007/s40745-015-0040-1

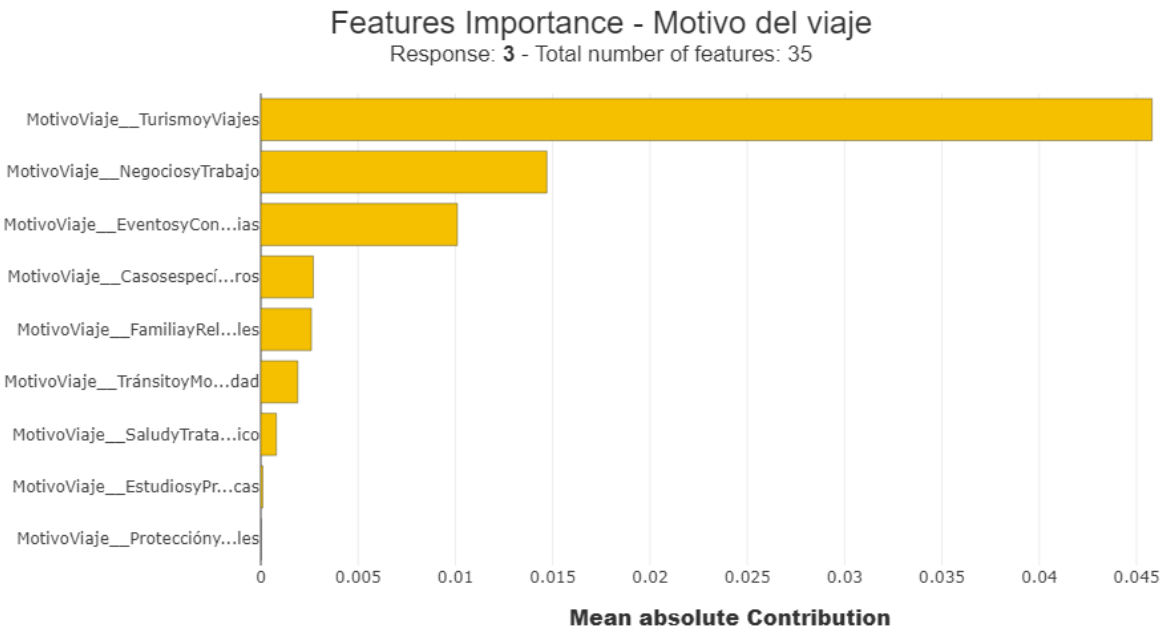
ANEXOS

Anexo 1. Notebooks

Enlace: https://github.com/Vanessaaguim/Trabajo-de-grado

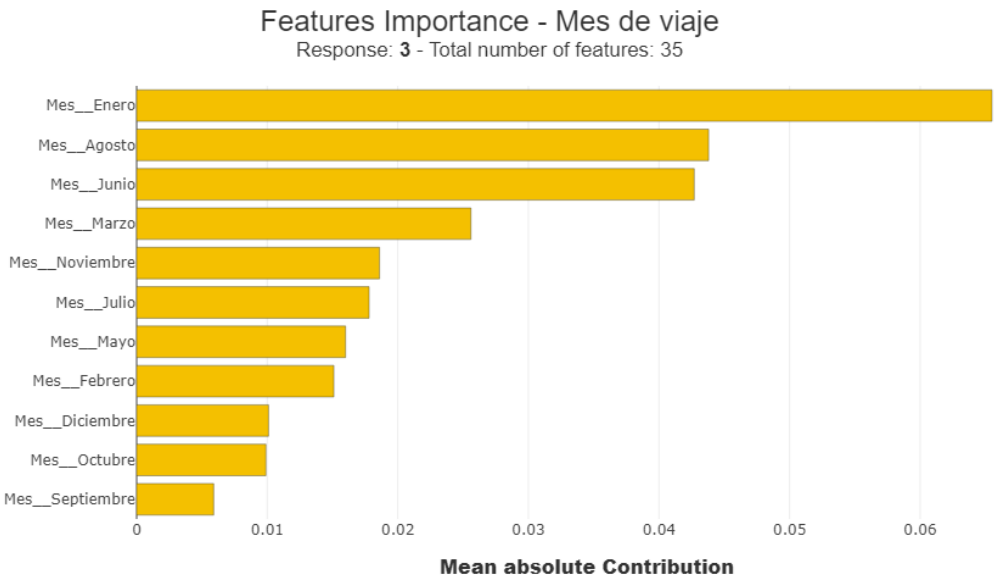
Anexo 2. Desagregaciones del clúster 0

Gráfico 9. Clúster 0: motivo de viaje



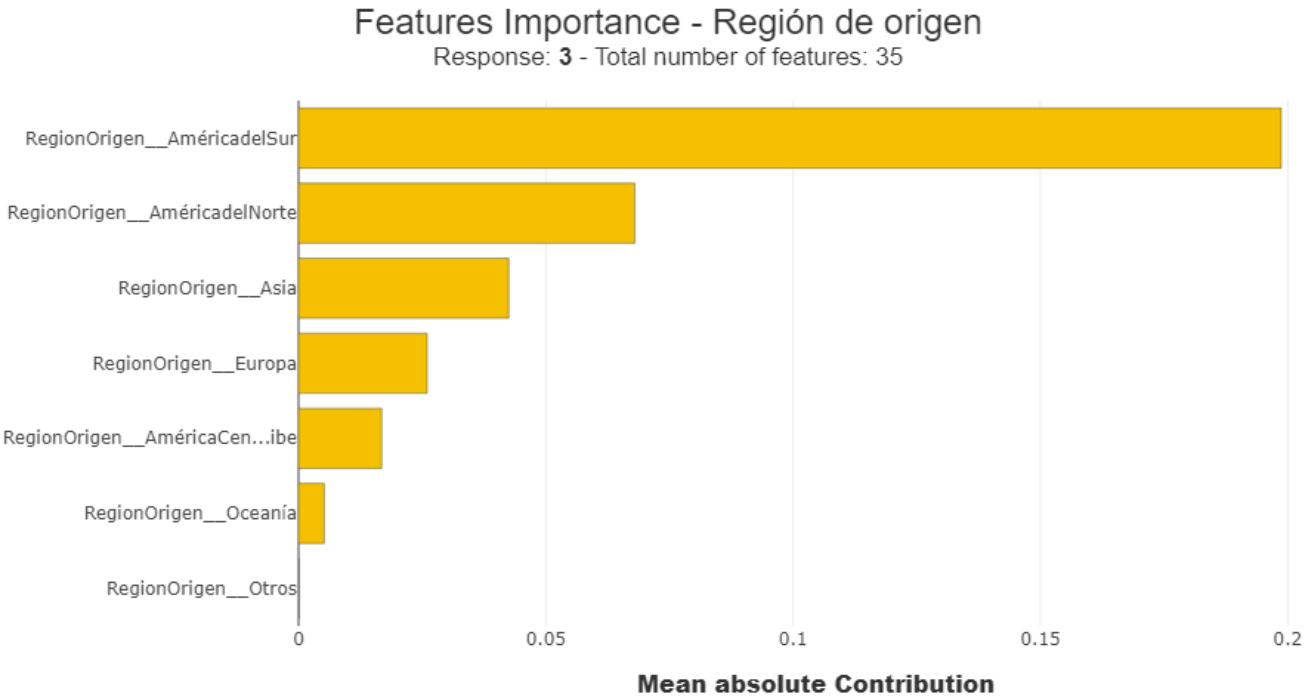
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 10. Clúster 0: mes del viaje



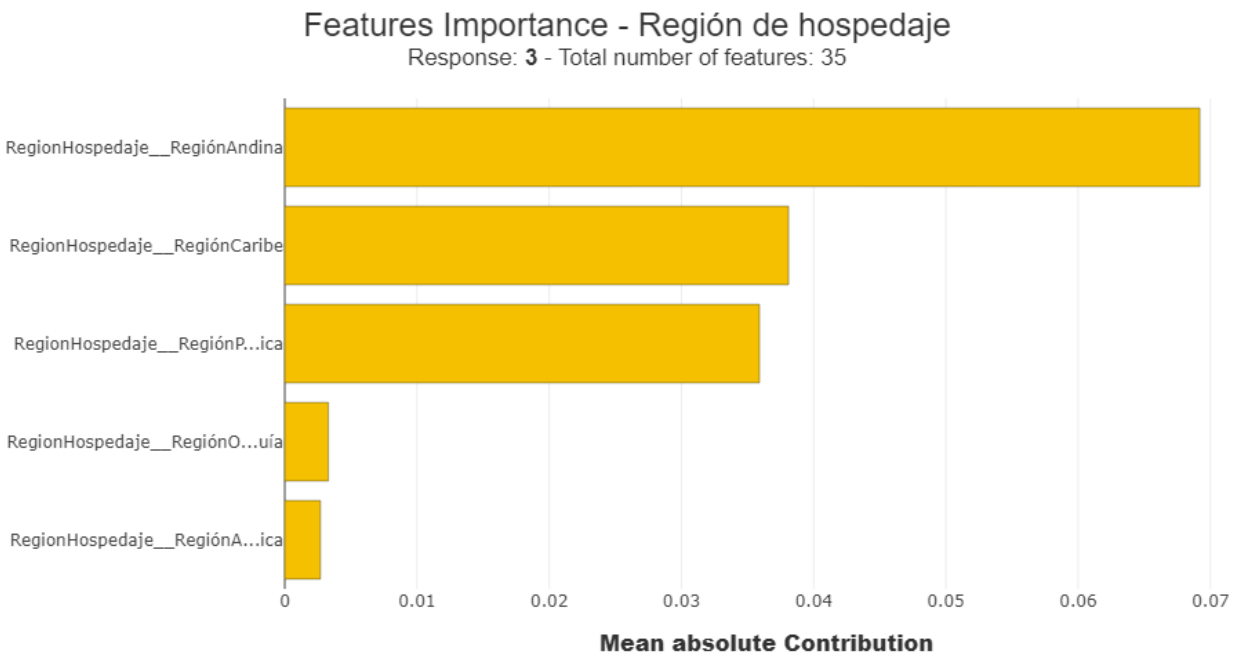
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 11. Clúster 0: región de origen



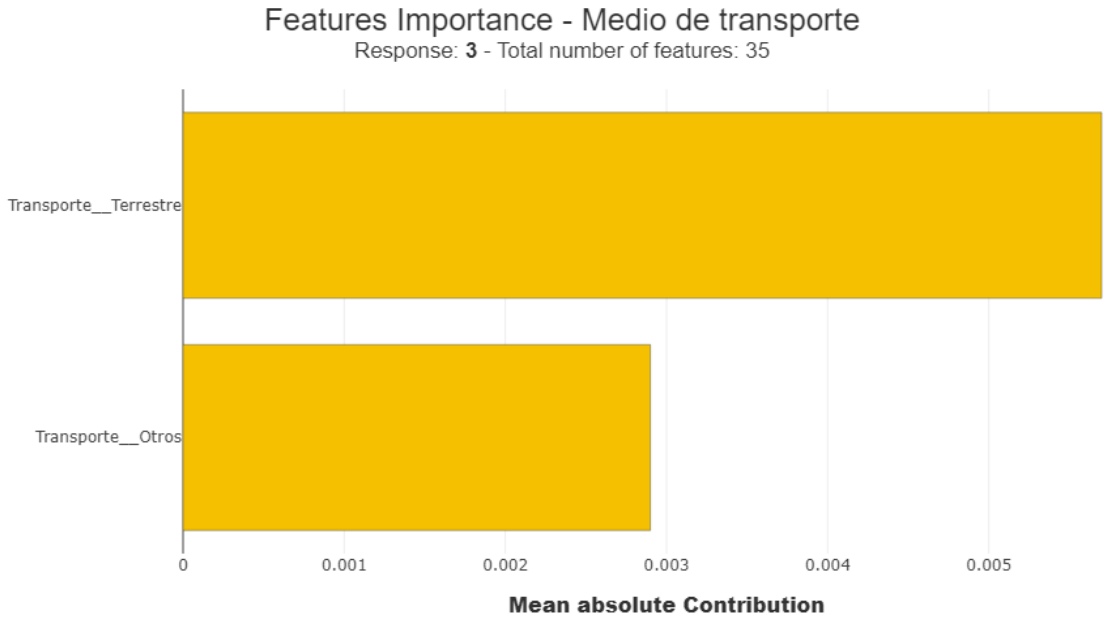
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 12. Clúster 0: región de hospedaje



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

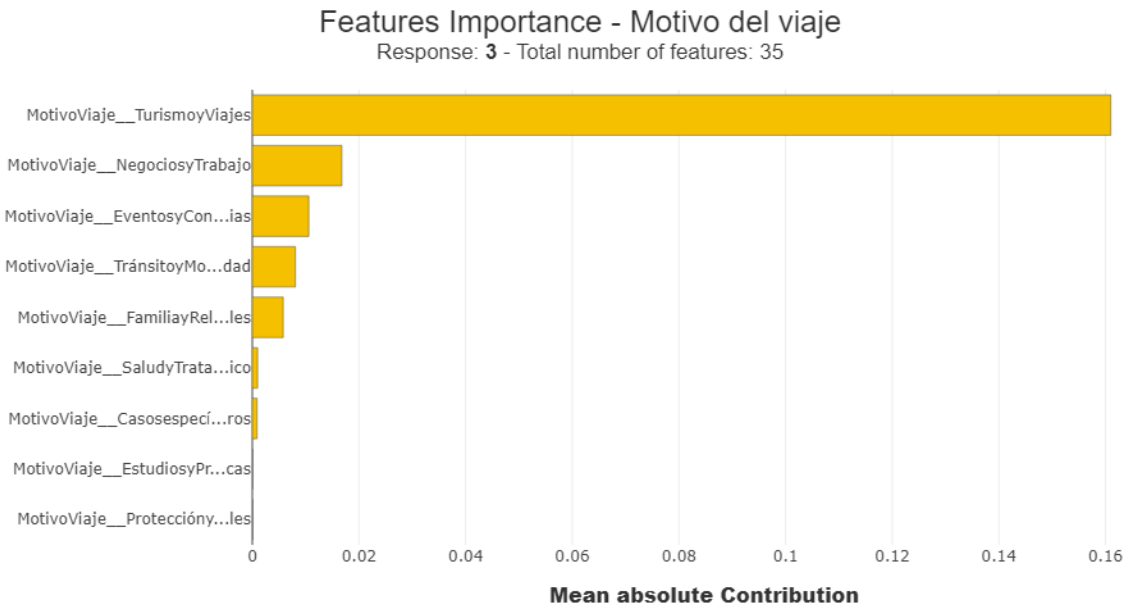
Gráfico 13. Clúster 0: medio de transporte



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

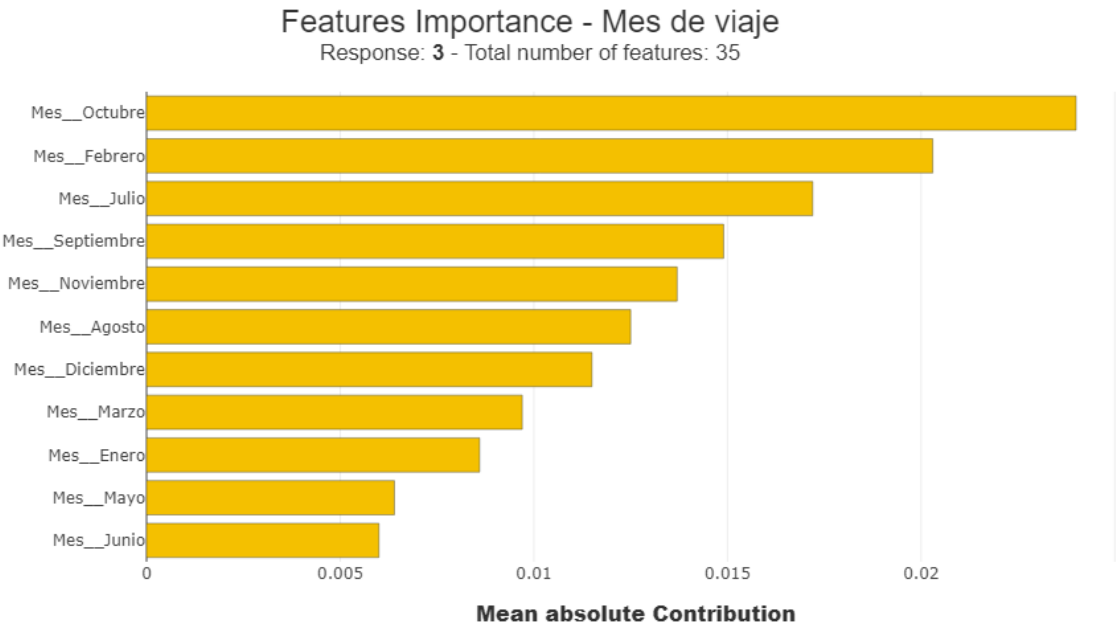
Anexo 3. Desagregaciones del clúster 1

Gráfico 14. Clúster 1: motivo del viaje



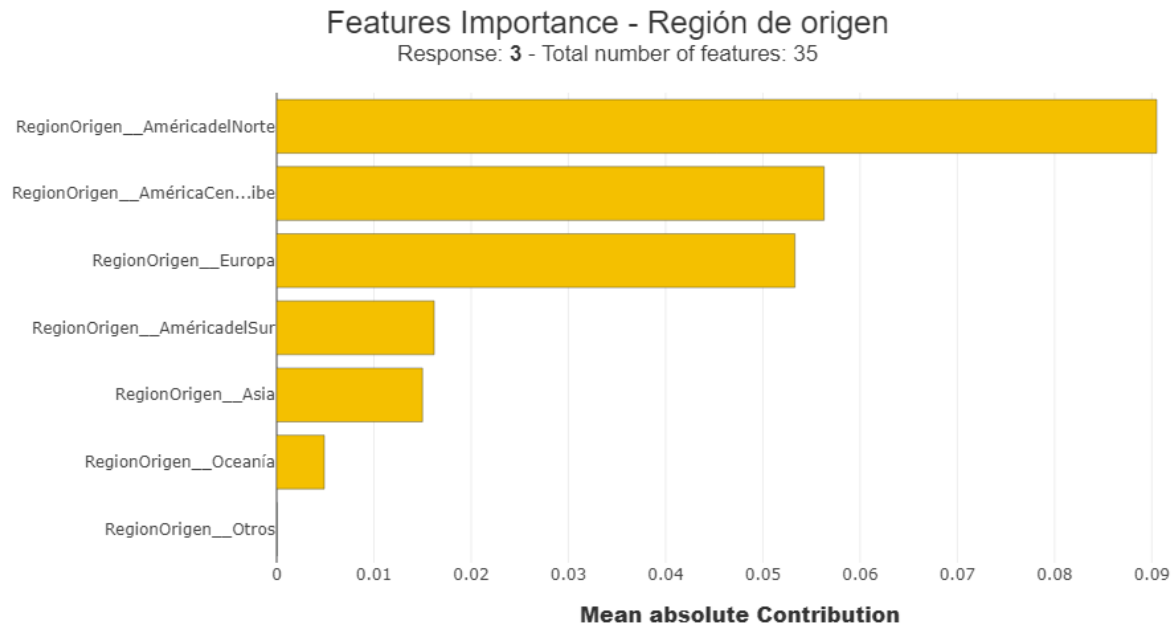
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 15. Clúster 1: mes de viaje



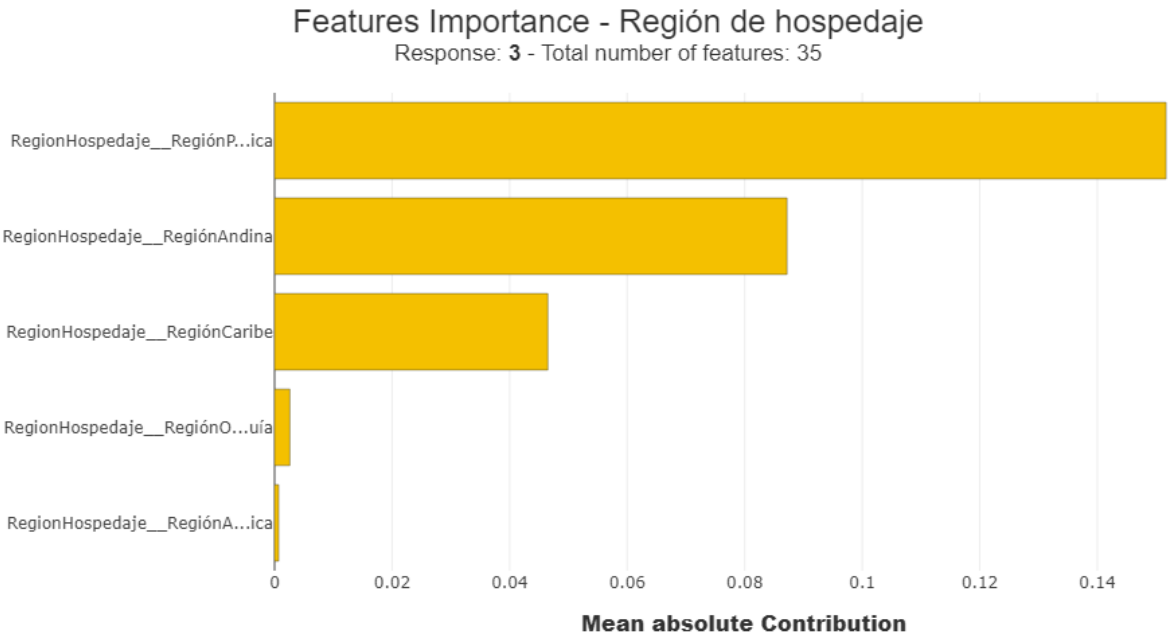
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 16. Clúster 1: región de origen



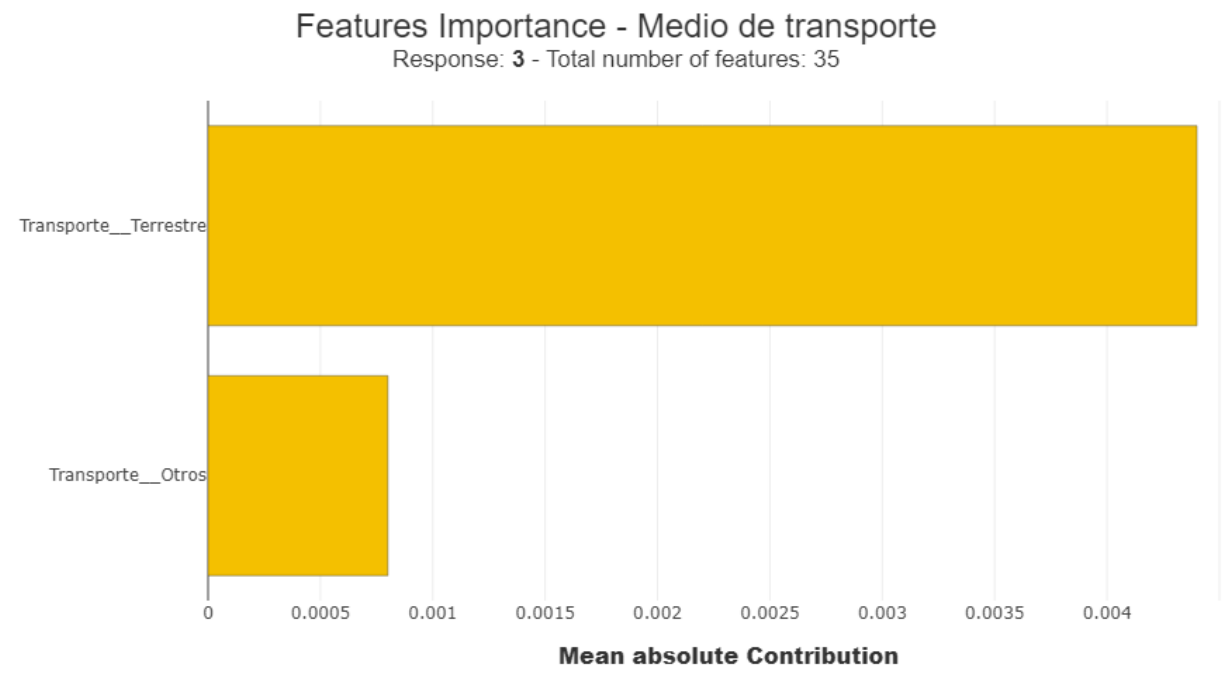
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 17. Clúster 1: región de hospedaje



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

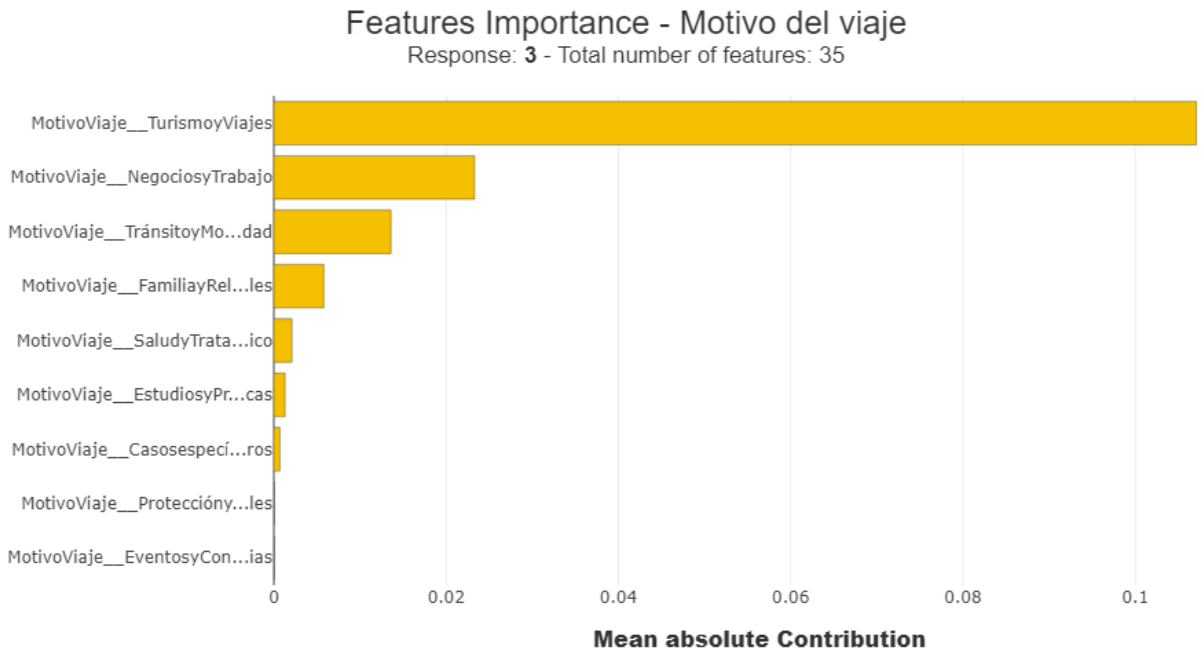
Gráfico 18. Clúster 1: medio de transporte



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

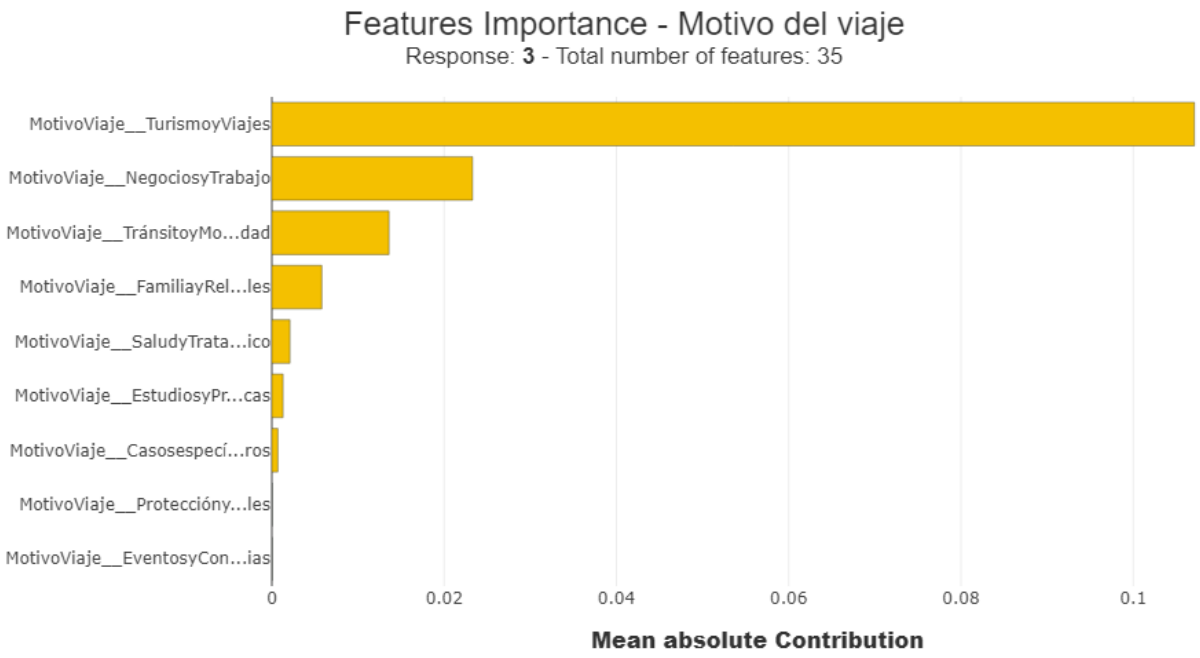
Anexo 4. Desagregaciones del clúster 2

Gráfico 19. Clúster 2: motivo del viaje



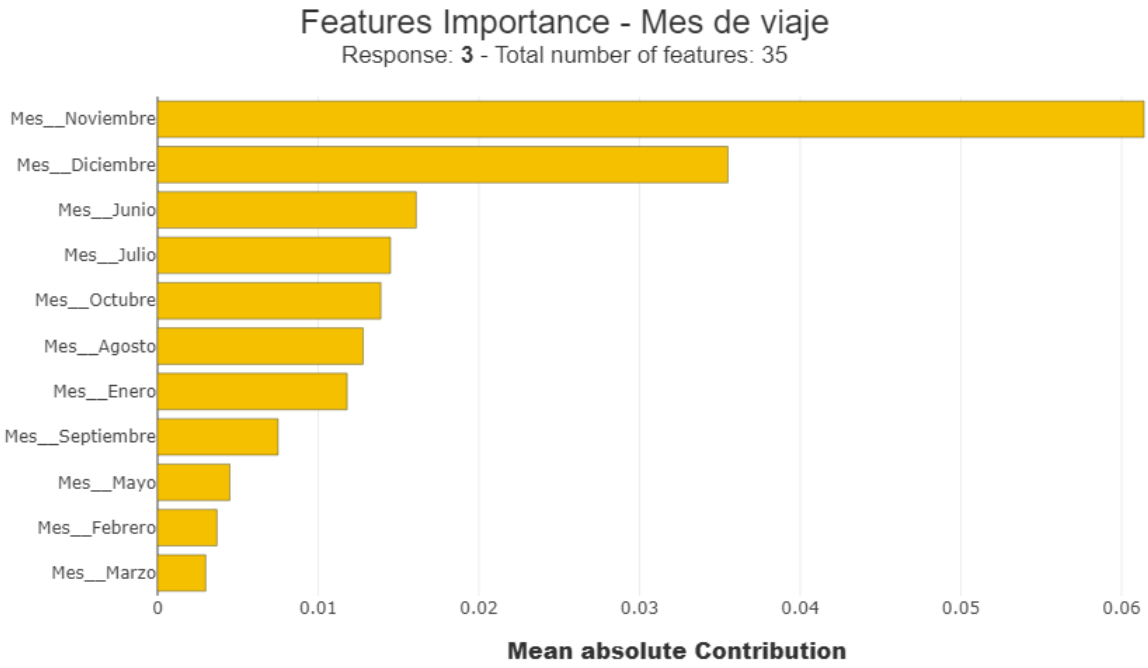
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 20. Clúster 2: motivo del viaje



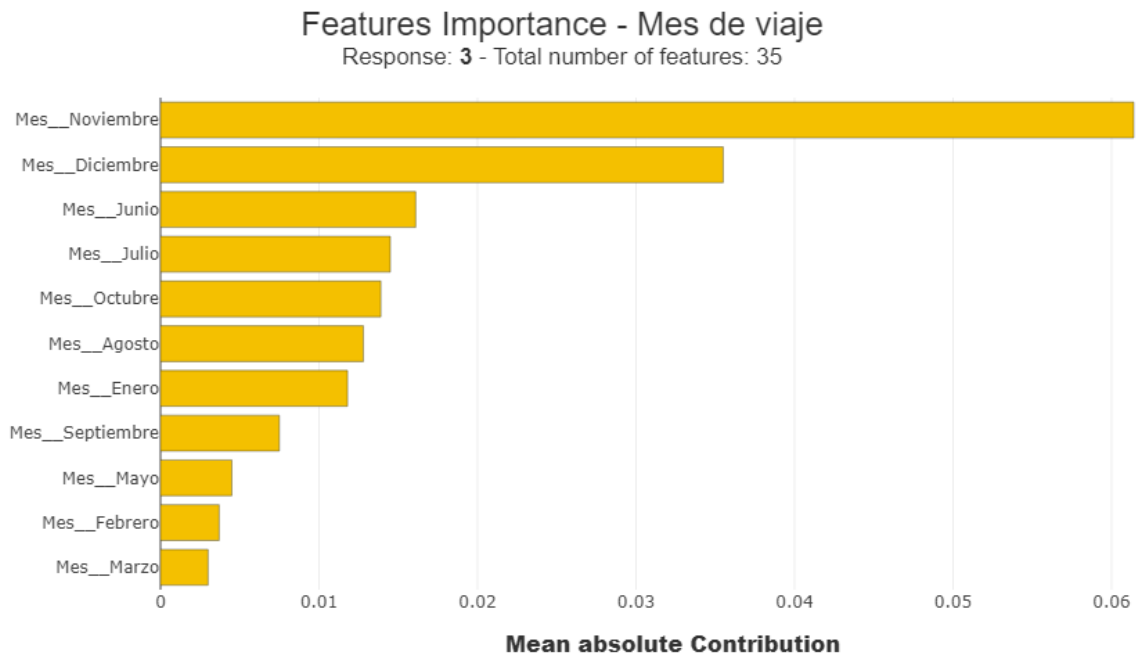
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 21. Clúster 2: mes de viaje



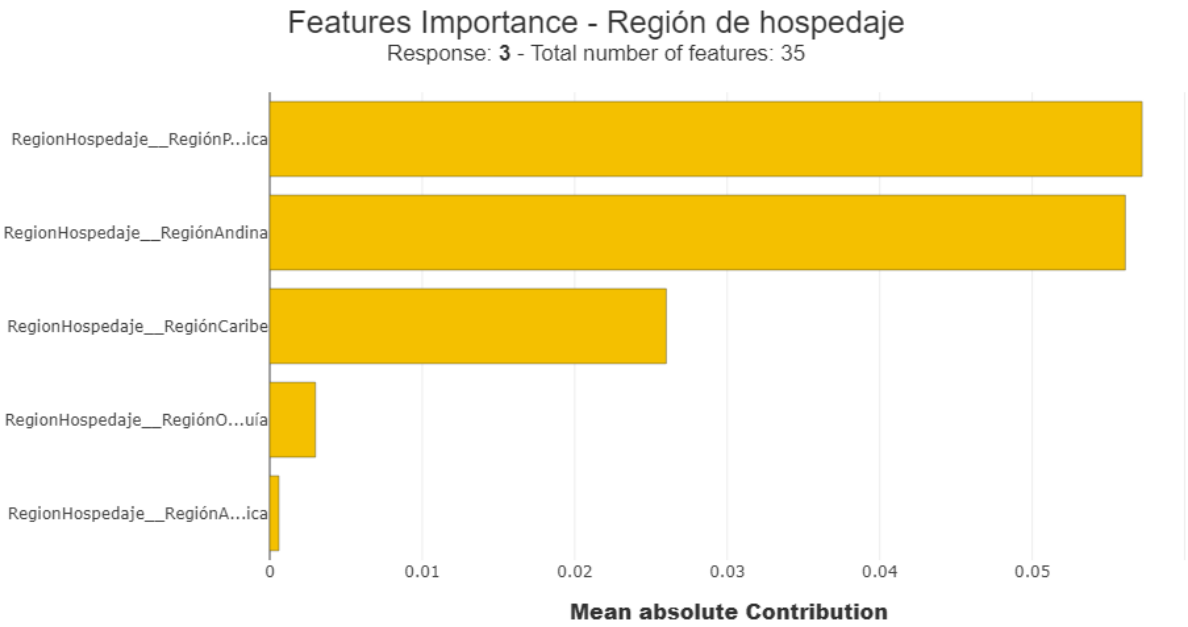
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 22. Clúster 2: mes de viaje



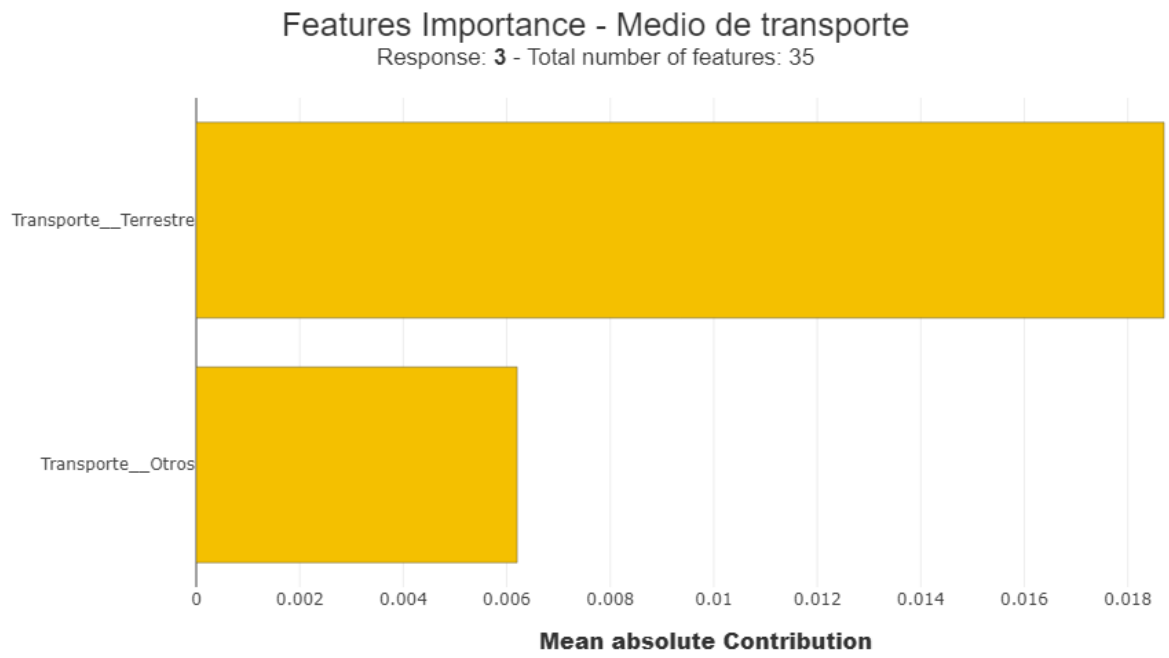
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 23. Clúster 2: región de hospedaje



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

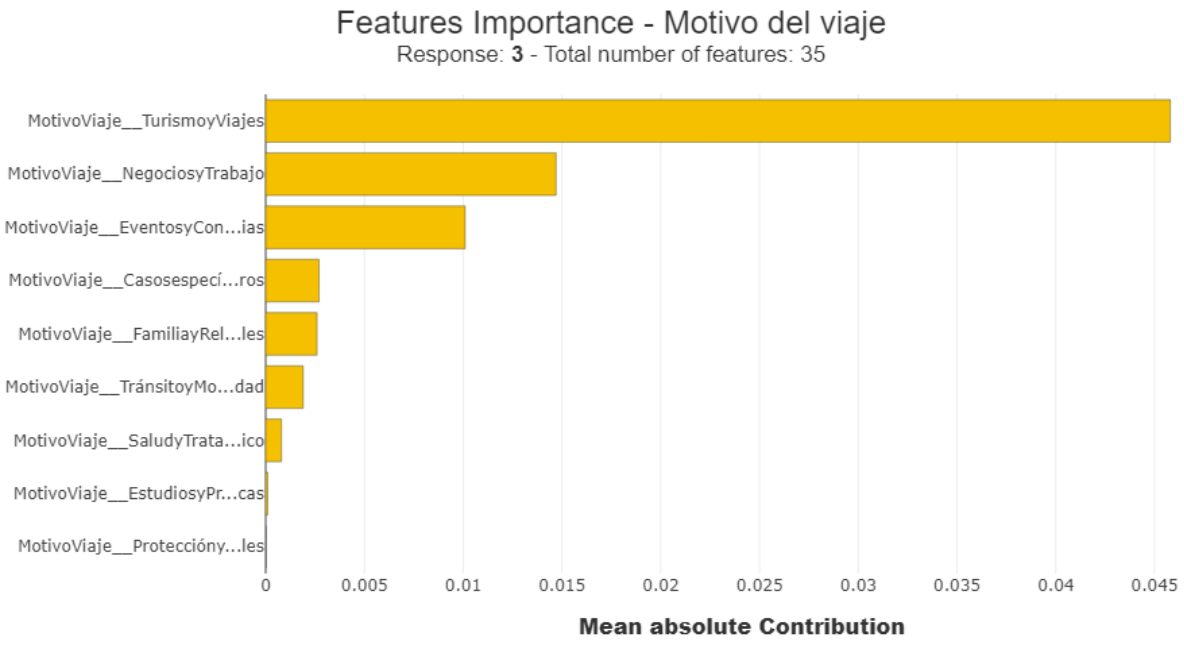
Gráfico 24. Clúster 2: medio de transporte



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

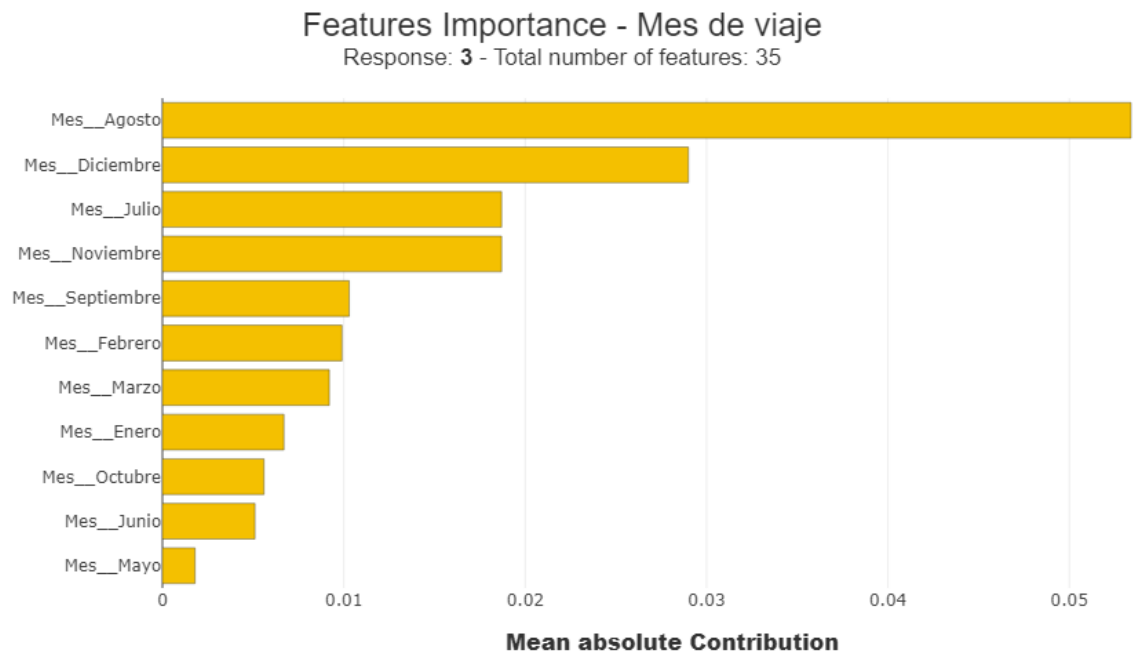
Anexo 5. Desagregaciones clúster 3

Gráfico 25. Clúster 3: motivo del viaje



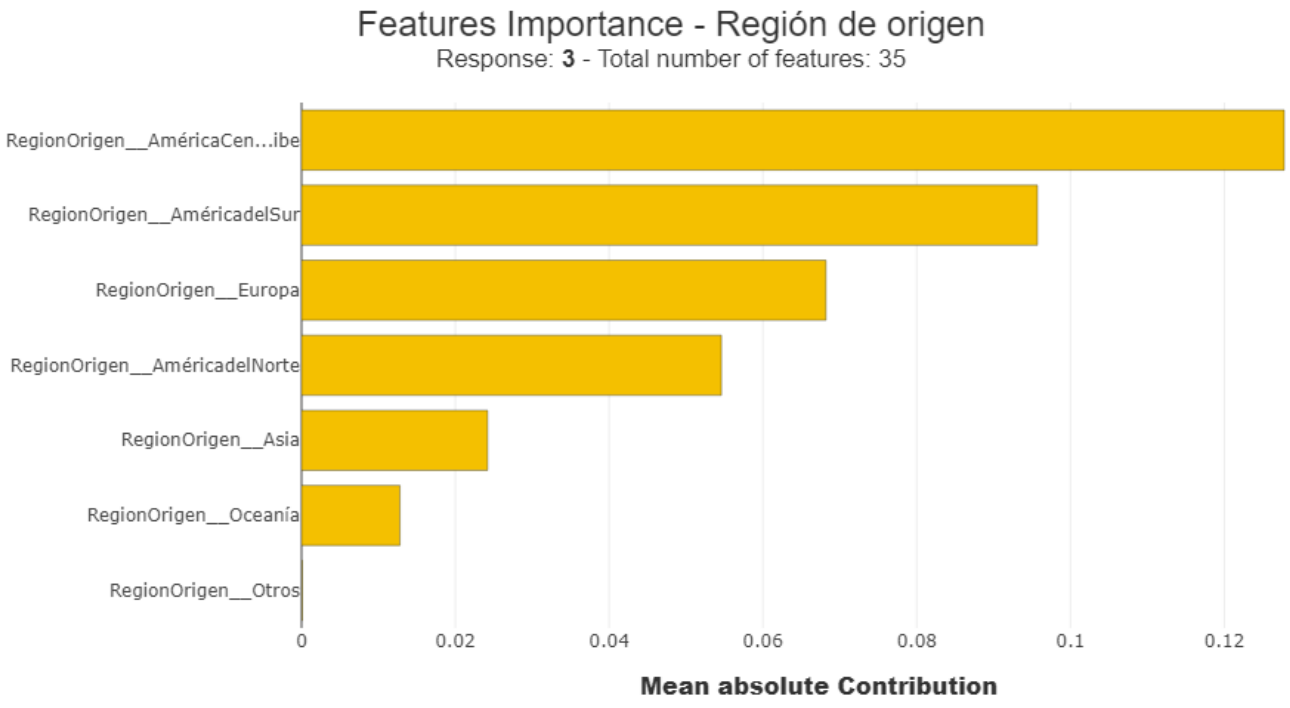
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 26. Clúster 3: mes de viaje



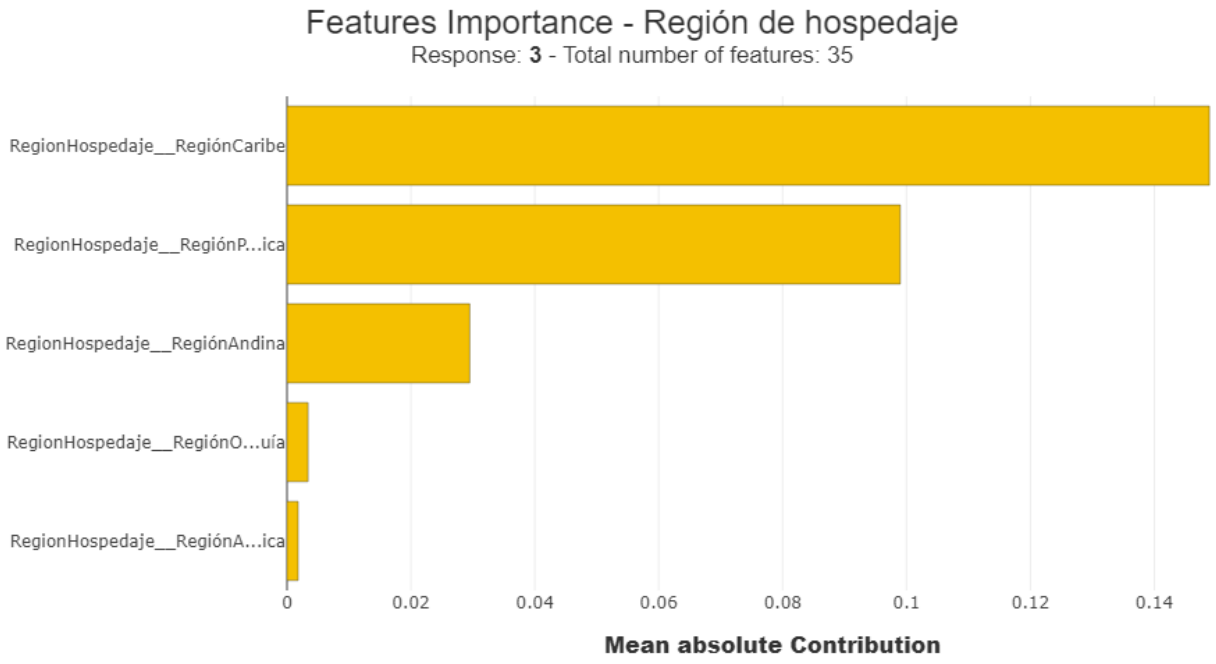
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 27. Clúster 3: región de origen



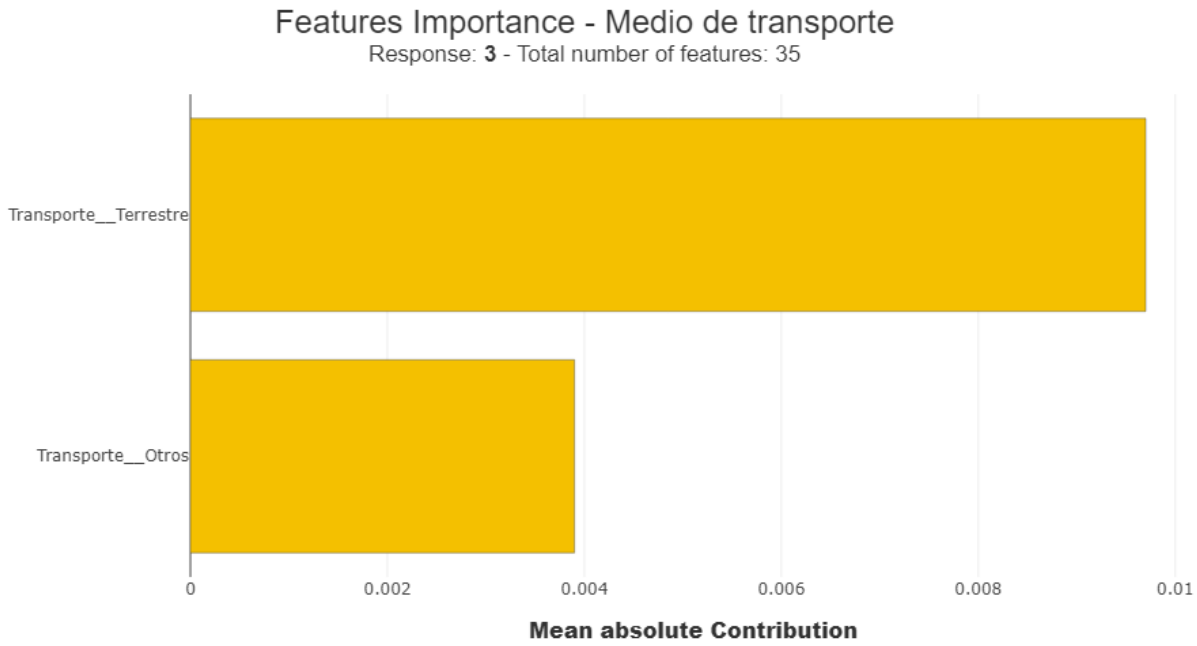
Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 28. Clúster 3: región de hospedaje



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.

Gráfico 29. Clúster 3: medio de transporte



Fuente: elaboración propia a través de Python con el conjunto de datos obtenido para esta investigación.