

MÁSTER EN DATA SCIENCE Y BUSINESS
ANALYTICS PRESENCIAL

Predicción del número de
inmigrantes extranjeros en España
mediante modelo de aprendizaje
automático para prever flujos
migratorios internacionales

TFM elaborado por:

Cristian de Andrade Correia
Erick Ernesto Hernández Lara

Tutor de TFM:

Miguel Martín

Madrid, 23 de octubre de 2024

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS	1
ÍNDICE DE FIGURAS	2
RESUMEN.....	5
1. INTRODUCCIÓN	6
2. ANTECEDENTES	9
3. OBJETIVO	11
4. MATERIALES Y MÉTODOS.....	11
4.1 Adquisición de fuentes de datos.....	11
4.2 Descripción de fuentes de datos	12
4.3 Resumen de herramientas e infraestructura empleada	13
4.4 Procedimiento	14
5. RESULTADOS	16
5.1 Etapa 1: Limpieza, preprocesamiento y análisis de datos de inmigración	16
5.2 Etapa 2: Limpieza, preprocesamiento y análisis de variables explicativas	22
5.2.1 Continentes y Subregiones.....	22
5.2.2 Padrón de Inmigrantes Residentes en España	24
5.2.3 Índices de Desarrollo	26
5.2.4 Régimen Político	30
5.2.5 Índices de Democracia	31
5.2.6 Índices de Libertad y otros.....	34
5.2.7 Tasa de Homicidios.....	37
5.2.8 Conflictos Armados.....	39
5.2.9 Turistas Anuales	42
5.3 Etapa 3: Unión y selección de variables	42
5.4 Etapa 4: Prueba y comparación de modelos de <i>machine learning</i>	49
6. CONCLUSIONES	57
6.1 Trabajo Futuro:.....	57
REFERENCIAS BIBLIOGRÁFICAS	58
ANEXOS.....	60

ÍNDICE DE TABLAS

Tabla 1. Conjunto de datos usados para cada dimensión y sus variables principales.	13
Tabla 2. P-valor de las comparaciones pareadas entre grupos de edades mediante la prueba de Dunn	47
Tabla 3. P-valor de las comparaciones pareadas entre grupos de tipo de régimen político mediante la prueba de Dunn.	48
Tabla 4. P-valor de las comparaciones pareadas entre grupos de continentes mediante la prueba de Dunn	48
Tabla 5. P-valor de las comparaciones pareadas entre grupos de subregiones mediante la prueba de Dunn	48
Tabla 6. Métricas en train/test de modelos con mejor rendimiento para el conjunto de datos con agregados de sexo y grupos de edad	50
Tabla 7. Métricas en train/test de modelos con mejor rendimiento para el conjunto de datos sin agregados de sexo y grupos de edad	50
Tabla 8. Métricas en <i>train/test</i> de modelos con peor rendimiento para el conjunto de datos con agregados de sexo y grupos de edad	51
Tabla 9. Métricas en train/test de modelos con peor rendimiento para el conjunto de datos sin agregados de sexo y grupos de edad	51
Tabla 10. Top 10 variables en importancia media relativa por permutaciones	52
Tabla 11. Métricas en <i>test</i> de modelos de red neuronal de una capa después de aplicar normalización para el conjunto de datos con agregados de sexo y grupos de edad.	53
Tabla 12. Métricas en <i>test</i> de modelos de red neuronal de una capa después de aplicar normalización para el conjunto de datos sin agregados de sexo y grupos de edad.	53

ÍNDICE DE FIGURAS

Figura 1. Factores que propician la migración internacional. Factores que propician la migración internacional (Organización Internacional de Migraciones).....	7
Figura 2. Fuentes de información, elaboración propia.....	12
Figura 3. Esquema de las etapas del proyecto	15
Figura 4. Valores que toman las categorías de nacionalidades, sexo y grupos de edad para los <i>datasets</i> de inmigración de 2008-2021 (A) y 2022 (B).	16
Figura 5. Valores las categorías de sexo (izquierda) y grupos de edad (derecha) post-limpieza/estandarización.....	17
Figura 6. Código empleado para encontrar las top nacionalidades del <i>dataset</i> 2008 - 2021 considerando cada año. Se siguió un procedimiento similar para el conjunto de datos de 2022...	17
Figura 7. Top nacionalidades (en códigos ISO) de los conjuntos de datos de inmigración de 2008-2021 y 2022.....	18
Figura 8. <i>Dataframe</i> de la unión de los <i>datasets</i> de datos de inmigración post-limpieza/preprocesamiento	18
Figura 9. Total de inmigrantes en España desde el año 2008 al 2022.....	19
Figura 10. Cantidad de inmigrantes en España por grupo de edad y sexo.....	19
Figura 11. Distribución de inmigrantes en España por año desde el 2008 a 2022.....	20
Figura 12. Top 10 nacionalidades en porcentaje de inmigrantes en España durante el periodo 2008-2022 (A) y el cambio en el número de inmigrantes en ese período para el Top 5 (B).....	21
Figura 13. Cantidad de número de inmigrantes en España por grupo de edad en el año 2022.....	22
Figura 14. Contenientes y subregiones luego de aplicar el filtro del top nacionalidades en inmigración en España.....	23
Figura 15. Cantidad de países por continente (A) y subregión (B) dentro del top nacionalidades en inmigración en España.....	23
Figura 16. Extracto de grupos de países encontrados en padrón de residentes.	24
Figura 17. Extracto de estandarización con ISO 3166.....	24
Figura 18. Suma de residentes por país de origen en el año 2022.	25
Figura 19. Variación anual de los 5 países con mayor cantidad de residentes extranjeros en España.	25
Figura 20. Variación anual de 5 países con menor presencia de residentes extranjeros en España.	26
Figura 21. Tabla de estadísticos descriptivos para las variables de desarrollo.	27
Figura 22. Distribución de las variables de índices de desarrollo y detección de datos atípicos....	28
Figura 23. Contraste entre la tendencia de cambio en el tiempo de los datos de inmigración del Top 4 nacionalidades (A) y algunos índices de desarrollo: (B) "Political and Violence Percentile", (C) "GDP_growth", y (D) "Rule of Law Percentile".....	29

Figura 24. Conteo de países en el top nacionalidades en inmigración por cada tipo de régimen político por año.....	30
Figura 25. Países dentro del top nacionalidades en inmigración con regímenes políticos autocráticos entre el año 2008 y 2022	30
Figura 26. Distribución de las variables de índice de democracia liberal (A) y deliberativa (B), y detección de datos atípicos.....	31
Figura 27. Variación anual del máximo, mínimo, media y mediana para el índice de democracia deliberativa (A) y liberal (B).....	32
Figura 28. Variación anual de la cantidad de inmigrantes en el Top 4 nacionalidades (A) en contraste con la variación de sus índices de democracia liberal (B) y deliberativa (C).....	34
Figura 29. <i>Dataframe</i> antes de preprocesamiento	34
Figura 30. Muestra de indicadores presentes el estudio, suman 172.....	35
Figura 31. <i>Dataframe</i> final con países selectos y <i>pivot</i> aplicado.	35
Figura 32. Promedio anual de todos los países en: (A) Ausencia de corrupción, (B) Frecuencia de sobornos, (C) Grado de respeto a las libertades, (D) Responsabilidad judicial, (E) Igualdad educacional y (F) Igualdad en salubridad.....	36
Figura 33. Distribución de datos de Tasa de Homicidios y detección de datos atípicos.....	37
Figura 34. Distribución de Tasa de Homicidios por año desde el 2008 a 2022 para el top de nacionalidades en inmigración en España.	38
Figura 35. Contraste entre la tendencia de cambio en el tiempo de la Tasa de Homicidios (A) frente a los datos de inmigración en España del Top 4 nacionalidades (B).	39
Figura 36. Distribución las variables de conflictos armados y detección de datos atípicos.....	40
Figura 37. Variación anual de la cantidad de inmigrantes en el Top 4 nacionalidades (A) en contraste con la variación de las muertes por conflictos armados: (B) "One-sided violence_deaths", (C) "Non-state_deaths", (D) "Intrastate_deaths" y (E) "Interstate_deaths".	41
Figura 38. Total de turistas que llegaron a España en avión por año.	42
Figura 39. Código python para unir las variables predictoras al <i>dataset</i> "inmigrantes" con los datos anuales de inmigración de España.	43
Figura 40. Matriz de correlación de Pearson entre las variables cuantitativas.....	44
Figura 41. Distribución del número de inmigrantes mediante <i>boxplots</i> y detección de datos atípicos para los distintos grupos de cada variable categórica: (A) Sexo, (B) Idioma español, (C) Grupo de edad, (D) Régimen político, (E) Continente y (F) Subregión.	45
Figura 42. Resultados de la prueba de Shapiro-Wilk para normalidad. H0: Datos se ajustan a una distribución normal, H1: No se ajustan a una normal. Intervalo de confianza = 0.95.....	46
Figura 43. Resultados de las pruebas de U de Mann-Whitney y Kruskal-Wallis.....	46
Figura 44. <i>Dashboard</i> de predicciones mediante los mejores modelos de red neuronal de una capa para los conjuntos de datos con (izquierda) y sin agregados (derecha) de sexo y grupos de edad. Intervalos de confianza: 90%. También disponible en este link	54

Figura 45. Muestra de predicciones de inmigrantes para Colombia (A) y Brasil (B) para los modelos con (derecha) y sin (izquierda) agregados de sexo y grupos de edad que incluyen los años 2023 y 2024..... 55

RESUMEN

El creciente flujo de inmigración internacional a la Unión Europea (UE) y la compleja red de factores que rigen la dinámica migratoria internacional han generado la necesidad de emplear nuevas estrategias para predecir estos movimientos. Así, este proyecto tuvo como objetivo principal predecir el número de inmigrantes extranjeros que llegan a España, país que ocupa uno de los primeros lugares en inmigración dentro de la UE, mediante un modelo de *machine learning* con un enfoque en variables macro (por encima del individuo), como variables socioeconómicas, políticas y de seguridad. Para ello, se dividió el proyecto en cuatro etapas: i) análisis de datos de inmigración, ii) análisis de variables macro predictoras, iii) unión y reducción de variables, y iv) prueba y comparación de modelos de *machine learning*. El modelo final seleccionado, basado en una red neuronal de una capa con normalización de variables *inputs* y *target* (“Immigrant count”), mostró una buena capacidad predictiva, obteniéndose los siguientes valores de métricas en el conjunto de prueba: R^2 ajustado = 0.975, RSME = 853 y MAE = 293. Nuestro análisis también resalta la relevancia de los grupos de edades y del fenómeno de migración en cadena dentro de la dinámica migratoria internacional en España. Asimismo, debido al enfoque macro de nuestra metodología, se recomienda el uso de este modelo predictivo para estimaciones de 1 o 2 años en el futuro como herramienta de soporte en el proceso de planificación y gestión gubernamental.

Palabras claves: inmigración, España, predicción, *machine learning*, red neuronal

1. INTRODUCCIÓN

Según el Informe de Migraciones de 2022 de la Organización Internacional de Migraciones (OIM)¹, de la Organización de las Naciones Unidas (ONU), la cantidad de migrantes internacionales ha incrementado de 153 millones en 1990 a 281 millones en 2020, acercándose a duplicar dicho valor, y que es, a su vez, más del triple de lo estimado en 1970.

De estos 281 millones de migrantes en el 2020, Europa y Asia acogieron alrededor de 87 y 86 millones respectivamente, que representa el 61% de la población mundial de migrantes, seguidos de Estados Unidos de Norte América con 21%.

Ya en el 2016 el Banco Europeo de Inversiones (BEI) menciona en su análisis de desafíos y oportunidades de la migración en Europa², que la Unión Europea (UE) fue el mayor aportador durante la escalada de inmigrantes y aplicantes de asilo durante los años 2015 y 2016, con una asignación de 10,1 billones de euros en ambos años, indicando que una deficiente integración de los migrantes podría traducirse en una reducción del Producto interno bruto (PBI) per cápita en los años subsiguientes. Sin embargo, mantuvo la visión que, a largo plazo, las migraciones internacionales tendrían un efecto positivo en Europa y resaltaba la necesidad de modificar y mejorar los marcos legislativos en torno a las políticas de integración y optimización del flujo de migrantes.

El éxito de estos esfuerzos recae en el sinergismo de los diversos procesos que regulan el ritmo migratorio y la integración de los inmigrantes, que a su vez están asociados a la capacidad económica, judicial, logística y de defensa de los países receptores para dar respuesta a las exigencias que plantea el continuo aumento de estos movimientos migratorios internacionales.

Por otra parte, desde el punto de vista del individuo, también tenemos numerosos factores que impactan en su decisión por embarcarse en una travesía migratoria. La Organización Internacional para las Migraciones expone un diagrama (Figura 1) que hace un excelente sumario de la relación entre variables “macro”, “meso” y “micro” que afectan la decisión de las personas en migrar, las cuales involucran variables como la seguridad, libertad, empleo, precios, servicios básicos y salud (macro), actuando en conjunto con las condiciones, posibilidades y ventajas personales (micro y meso).

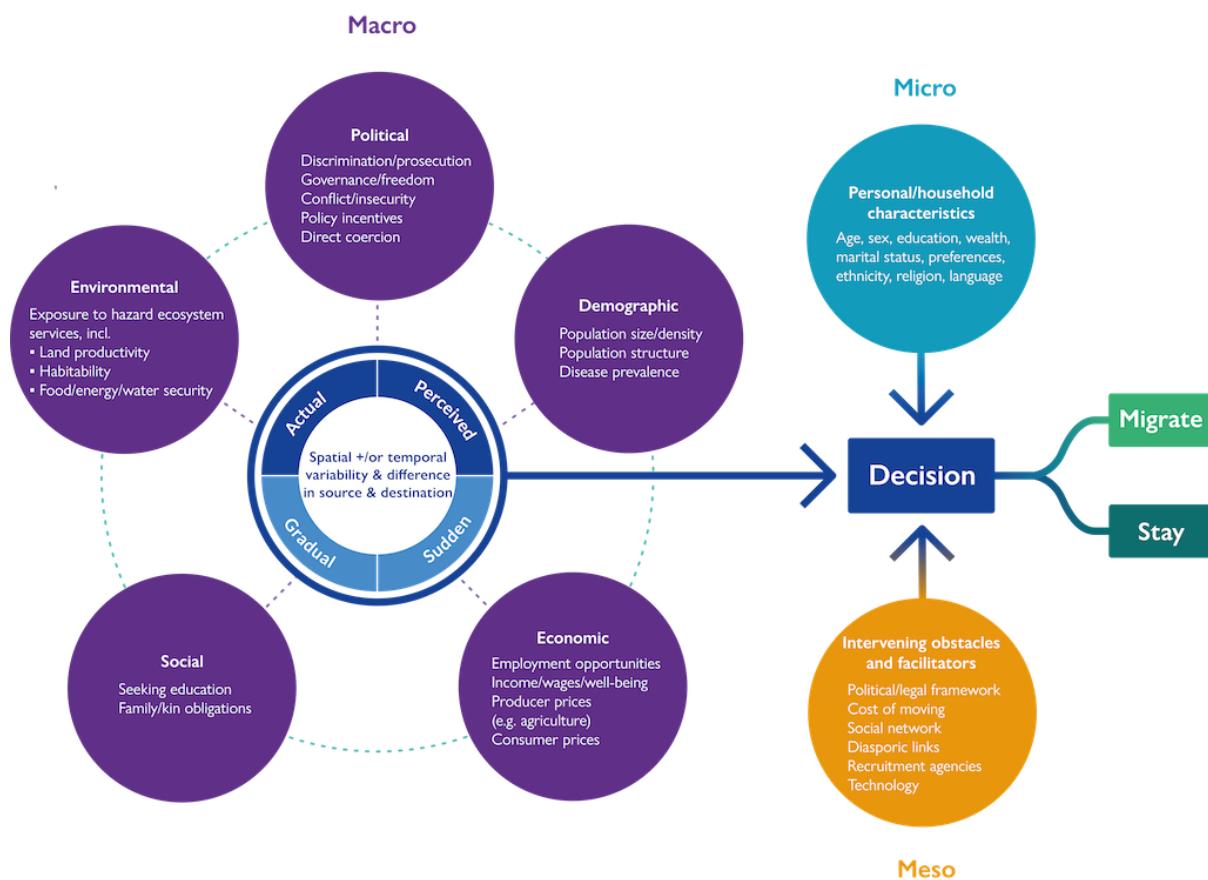


Figura 1. Factores que propician la migración internacional. Factores que propician la migración internacional (Organización Internacional de Migraciones¹).

Así, podemos observar que el estudio del proceso migratorio es un reto complejo debido a la diversidad de factores que lo afectan y el dinamismo de nuestro actual entorno globalizado. Esto se evidencia también en la creciente implementación de políticas en relación con la movilidad humana promovida por la ONU a partir de 2010 que incluye áreas como el cambio climático, desastres y desarrollo sustentable³. Si, además, consideramos eventos de origen bélico, como el actual conflicto entre Ucrania y Rusia, o Israel y el grupo terrorista HAMAS, tendremos una visión general de la complejidad del fenómeno de las migraciones internacionales.

Indudablemente, el mayor esfuerzo y responsabilidad recae sobre los países con mayor recepción de dichos inmigrantes internacionales. De acuerdo con la OIM, tres países de la UE formaban parte del top 10 de países destino para migrantes internacionales a nivel global en el 2020: Alemania, Francia y España, siendo España el país que ocupaba el décimo lugar a nivel global y el tercero de la UE¹, que, además, alcanzó en el 2022 un saldo migratorio de 727.005 personas, el máximo nivel en 10 años⁴.

A razón de lo expuesto, este proyecto busca aprovechar estas tecnologías para estudiar los datos oficiales de inmigrantes en España y desarrollar un modelo predictivo como herramienta de estrategia y planificación que permita estimar el número de inmigrantes extranjeros al territorio español. Y, aunque el fenómeno de la migración también atiende a decisiones personas que pueden

estudiarse desde un punto de vista psicosocial, en este proyecto mantendremos nuestro enfoque en variables macro (por encima del individuo).

2. ANTECEDENTES

Entre la diversidad de aplicaciones en los avances de la inteligencia artificial “clásica” o generativa, tenemos que los desarrollos en Aprendizaje Automático (AA), uno de los subcampos de la Inteligencia Artificial (IA), han permitido abordar problemáticas como la expuesta mediante algoritmos que permitan, por ejemplo, generar predicciones de inmigrantes y grupos específicos⁵, o incluso prever flujos migratorios relacionados con asilo mediante alertas usando datos de eventos y tendencias de Google⁶.

Aydemir et al. (2022)⁵ optaron por emplear un enfoque de predicción directa planteándose dos objetivos: i) predecir el grupo de ingresos (bajo, medio bajo, medio alto y alto) con base al porcentaje de inmigrantes en relación a la población y regiones, en conjunto con datos de variables de desarrollo del Banco Mundial –los primeros en emplear este enfoque– como grupos de edad, nivel de industrialización, áreas de agricultura, tasas de mortalidad, indicadores de servicios, etc., y ii) predecir el stock de inmigrantes de una nacionalidad en base al total a nivel mundial. Sus mejores resultados de tasa de éxito para predecir grupos de ingresos los obtuvieron con Regresión Logística (86.04%) y Máquina de Soporte Vectorial/K-vecinos más cercanos (83.72% para ambos), mientras que para su estudio de regresión, mencionan una tasa de éxito de 98.37% (XGBoost) y 96.42% (*Random Forest*), pero no aclaran la métrica ni mencionan otras; se asume que se trata del R².

Es importante mencionar que Aydemir et al. (2022)⁵ la relevancia de sus modelos considerando que uno de los intereses principales al proyectar migraciones es hacerlo para un amplio período de tiempo (años), sin embargo, su enfoque se ve limitado por la naturaleza de sus datos, los cuales implican el uso de variables de desarrollo, en consecuencia, estimar el futuro estado político, social y económico de las naciones es, en sí mismo, un desafío. En contraste, Carammia et al. (2022)⁷ optaron por enfoque de *rolling window* sobre datos históricos para su estudio predictivo de migraciones relacionadas a pedidos de asilo en la UE, enfocándose en predecir el número de solicitudes con una, dos, tres y cuatro semanas de anticipación, con la finalidad de proveer a las autoridades de una ventana de preparación en términos operacionales. Su estudio usó datos de eventos y búsquedas de internet de los países de origen, detección de inmigrantes en las fronteras de la UE / UE+ (país de origen seguro) y las tasas de aplicación y reconocimiento de asilos en ambos casos, observando que, en casi todas las semanas, sus predicciones se mantuvieron dentro de las bandas de ± 2 errores estándar.

La migración es un fenómeno que sin dudas puede explorarse desde distintos puntos de vista, así como componerlo con diferentes variables que pueden darnos un enfoque variado y concluyente. Tomemos por ejemplo el trabajo de Hosseini y Tarasyev (2018)⁸ quienes observan que las decisiones sobre migrar o no dependen en gran medida de la posibilidad de obtener un salario satisfactorio. Los migrantes, bajo la óptica de las teorías de capital humano, la teoría neoclásica o la teoría estándar de equilibrio, tomarán decisiones racionales con base en la satisfacción salarial,

mismas que pueden considerar la calidad de los mercados laborales en los países de origen y destino, la posible reducción de los salarios de las personas locales en el país destino como resultado de los flujos migratorios descontrolados o la suposición de encontrar un mercado laboral perfecto que supla la diferencia salarial del migrante.

Micevska (2021)⁹ va incluso más allá e introduce una variable como el *internet* como un factor que afecta en gran medida la decisión de migrar, así como ahondar en variables más “tradicionales” como lo son presencia de conflictos en el país de origen y condiciones macroeconómicas generales. Internet se coloca como una variable de peso al influir en la capacidad de solicitar asilo y obtener información detallada antes de migrar. También se especifica que el cambio climático se ha convertido en una razón por sí sola a la hora de tomar la decisión de migrar, siendo básicamente estas personas en “migrantes climáticos”.

El enfoque de nuestro estudio, si bien sigue una lógica empírica en relación con la disponibilidad de variables macroeconómicas, sociales, demográficas y políticas, no nos centraremos en encontrar todas las variables disponibles que influyan en menor o mayor medida al fenómeno de inmigración, sino en explicar la relación que tienen las variables seleccionadas desde un punto de vista más conservador y reducido, haciendo énfasis en los inmigrantes en España de origen extranjero y la situación macroeconómica y socio-política de sus países de origen.

3. OBJETIVO

- Predecir el número de inmigrantes extranjeros en España mediante modelo de aprendizaje automático para prever flujos migratorios internacionales.

Objetivos específicos:

- Listar los países más relevantes en materia de inmigración extranjera en España.
- Obtener un modelo predictivo con coeficiente de determinación cercano o superior al 0.80 (funcional).
- Discutir la importancia de variables.

4. MATERIALES Y MÉTODOS

4.1 Adquisición de fuentes de datos

Nuestro punto de partida fueron los datos de inmigración obtenidos del Instituto Nacional de Estadística de España⁵, los cuales constaban de un grupo de datos anuales de inmigrantes por nacionalidad y variables demográficas (sexo y grupo de edad) desde el año 2008 al 2021 (24287.csv), y otro grupo de datos entre 2021-2022 (61623.csv)¹⁰.

Adicionalmente, con base a nuestro enfoque macro, se realizó un exhaustivo trabajo de investigación durante tres semanas para obtener datos en torno a los factores macroeconómicos, sociales, políticos, libertades, seguridad, entre otros, a considerar para definir nuestras variables predictoras (Figura 2).

El principal desafío durante la investigación fue encontrar bases de datos que abarcaran nuestros años de interés (2008-2022) y que, además, estuviesen completos (ausencia de datos nulos). De manera que, para algunos casos, como para las variables económicas, se usaron varias fuentes en conjunto.

**Figura 2.** Fuentes de información, elaboración propia

A grandes rasgos, las fuentes de los datos consultadas y empleadas comprenden organismos no gubernamentales, estudios y trabajos de investigación, reportes periodísticos y organismos multinacionales, para los cuales haremos referencia en el siguiente apartado.

4.2 Descripción de fuentes de datos

Adicional a los datos centrales de inmigración de España, nuestros datos comprenden una amplia diversidad de variables. Contamos con, por ejemplo, información poblacional de número de residentes, valores booleanos para representar ausencia o presencia de alguna característica, índices económicos, índices de libertad/democracia y también valores estandarizados de 0 a 1 en algunas variables para representar su grado de solidez relacionada al país.

La Tabla 1 muestra el resumen de los distintos conjuntos de datos utilizados, la dimensión que abarcan, su descripción y/o justificación de su uso, variables que incluyen y enlaces web de referencia (si aplica). Adicionalmente, se agrega en el Anexo 1 las definiciones relevantes de algunas variables que son intuitivas.

Tabla 1. Conjunto de datos usados para cada dimensión y sus variables principales.

Dimensión	Conjunto de datos	Comentario / Descripción	Variable(s) Principal(es)	Referencia
Poblacional	Padrón	Cantidad de residentes extranjeros en España por nacionalidad entre 1998 y 2022	1) Número de Residentes	[13]
Socio-Económica	Índices de Desarrollo	Índices socioeconómicos anuales por país con los que se busca considerar y analizar el impacto de la condición de vida de cada uno sobre el incremento o disminución de inmigrantes hacia España.	1) Porcentaje de Desempleo 2) Percentil - Política y Violencia 3) Probabilidad de morir joven 4) Percentil - Calidad Regulatoria 5) Percentil - Estado de Derecho 6) Percentil - Representatividad y Participación Ciudadana 7) Porcentaje de trabajadores asalariados 8) Crecimiento de PBI (anual) 9) Inflación - deflactor del PBI (anual)	[14] [15] [16]
	Índices de Democracia	Índices de democracia anuales por país basado en las estimaciones de expertos y el índice de V-Dem. Con ellas se busca capturar el efecto de depreciación o mejoría de los derechos democráticos en los países sobre la inmigración.	1) Índice de democracia liberal 2) Índice de democracia deliberativa	[18]
Socio-Política	Índices GSOD	Indicadores (<i>Global State of Democracy Indices - GSOD Indices</i>) anuales por país que miden la tendencia democrática en base a distintas categorías	1) Ausencia de Corrupción 2) Libertades Civiles 3) Derechos civiles 4) Calidad de educación 5) Igualdad de oportunidades 6) Equidad en salud (acceso) 7) Rendición de cuentas judicial 8) Intercambios corruptos en el sector público	[19] [20]
Política	Régimen Político	Identifica el tipo de régimen político de los países por año y se distinguen entre: autocracias no electorales (0), autocracias unipartidistas (1), autocracias multipartidistas sin ejecutivo electo (2), autocracias multipartidistas (3), democracias excluyentes (4), democracias masculinas (5), democracias electorales (6) y políarchías (7)	Régimen Político	[17]
	Tasa de Homicidios	Tasa de homicidios por cada 100,000 habitantes por países/año.	Tasa de homicidios por cada 100,000 habitantes	[21] [22]
Seguridad	Conflictos Armados	Los conflictos armados reflejan la cantidad de muertes anuales por país según cuatro tipo de conflictos: <i>One-sided violence</i> , <i>Non-state</i> , <i>Intrastate</i> e <i>Interstate</i> . Con estas variables se incorporar efecto de los conflictos de tipo bélico en el incremento de movimientos migratorios	1) One-sided violence_deaths 2) Non-state_deaths 3) Intrastate_deaths 4) Interstate_deaths	[23]
Geográfica	Continentes y Subregiones	Agrupación oficial de países usada por el Instituto Nacional de Estadística (INE) de España y la Eurostat	1) Continente 2) Subregión (ej: A. del Sur, UE, Caribe)	[12]
Socio-cultural	Idioma Castellano	Variable categórica (1/0) con los que se especifica si el idioma castellano es o no el idioma oficial del país. Con esta variable se busca incluir el efecto del dominio del mismo idioma hablado en España sobre la decisión preferente de este destino por los inmigrantes.	Idioma Castellano	Elaboración propia
	Pandemia	Variable categórica (1/0) para la presencia de restricciones sanitarias debido a la pandemia de COVID 19 (2020-2021) Con esta variable se busca incluir el efecto de las presencia de restricciones sanitarias en la variación del número de inmigrantes	Restricciones por Pandemia	Elaboración propia
Salud	Post-Pandemia	Variable categórica (1/0) para indicar el primer año después de las restricciones sanitarias del COVID (2022) Con esta variable se busca considerar la acumulación de inmigrantes durante las restricciones sanitarias entre 2020-2021, los cuáles no lograron viajar a España, y así incluir la particularidad del año 2022 al ser el primer año después de la flexibilización casi total de las restricciones.	Año Post-Pandemia	Elaboración propia
Turismo	Turistas Anuales	Cantidad de Turistas extranjeros que llegaron a España en Avión por año. Con esta variable buscamos hacer una relación de inmigrantes-por-turista para proveerle al modelo una variable de magnitud para el efecto de la pandemia	Turistas Anuales	[24]

Según cada caso, los conjuntos de datos fueron sujetos a un preprocesamiento, análisis exploratorio y reducción dimensional previo a su uso para el entrenamiento de modelos de aprendizaje automático.

4.3 Resumen de herramientas e infraestructura empleada

Base de datos: Los datos recopilados y usados en este proyecto fueron datos estructurados/relacionales y, siendo un estudio temporal de inmigración, las claves centrales con los que se relacionaron los distintos ficheros empleados se basan principalmente en: año y nombre del país/código, pero pueden variar según los datos que aporta cada fichero.

Repository: Se creó un repositorio de [GitHub](#) para los ficheros, *notebooks*, *exports* resultantes y el escrito final, trabajando progresivamente en el proyecto mediante ramas a nivel local, haciendo *commits* de trabajo y revisiones conjuntas.

Infraestructura computacional: Se trabajo en local mediante equipos personales y software de uso libre y/o estándar: VS Code, Git/GitHub, Jupyter Notebooks, MS Excel, MS Word y MS PowerPoint (diagramas). Adicionalmente, las librerías principales que se usaron en lenguaje python durante el proyecto fueron: pandas, numpy, matplotlib, altair y scikit-learn (*machine learning* y redes neuronales). Las versiones usadas fueron las siguientes:

Visual Studio Code 1.93.1

Jupyter 3.6.0

Python 3.11.5

Pandas 2.0.3

Numpy 1.24.3

Altair 5.0.1

Matplotlib 3.7.2

Seaborn 0.12.2

Plotly 5.9.0

Ipywidgets 8.0.4

Scipy 1.11.1

Scikit-posthocs 0.9.0

Scikit-learn 1.2.2

Xgboost 2.0.3

Mapie 0.9.1

Joblib 1.2.0

4.4 Procedimiento

Teniendo en cuenta la variedad de conjuntos de datos en nuestro estudio y la necesidad particular de limpieza/preprocesamiento de cada uno, se decidió dividir el proyecto en 4 fases (Figura 3) que proporcionaran una secuencia ordenada de limpieza, análisis y prueba de algoritmos de *machine learning*:

- **Etapa 1:** Centrada en el trabajo con los datos de inmigración de España segregados por variables demográficas.
- **Etapa 2:** Enfocada en las variables predictoras.
- **Etapa 3:** Unión de todos los conjuntos de datos para análisis de selección de variables.
- **Etapa 4:** *Machine learning*.

En la Etapa 1 (E1), se realizó la limpieza, preprocesamiento y exploración inicial de los conjuntos de datos de inmigración, obteniendo los primeros *insights* y el top países en número de inmigrantes a incluir en el modelo. Luego, a partir de las observaciones y top países obtenidos en la Etapa 1, se inició la limpieza, preprocesamiento y análisis de los distintos conjuntos de datos de las variables predictoras (Etapa 2 - E2), finalizando con la exportación de los conjuntos de datos preprocesados de ambas etapas.

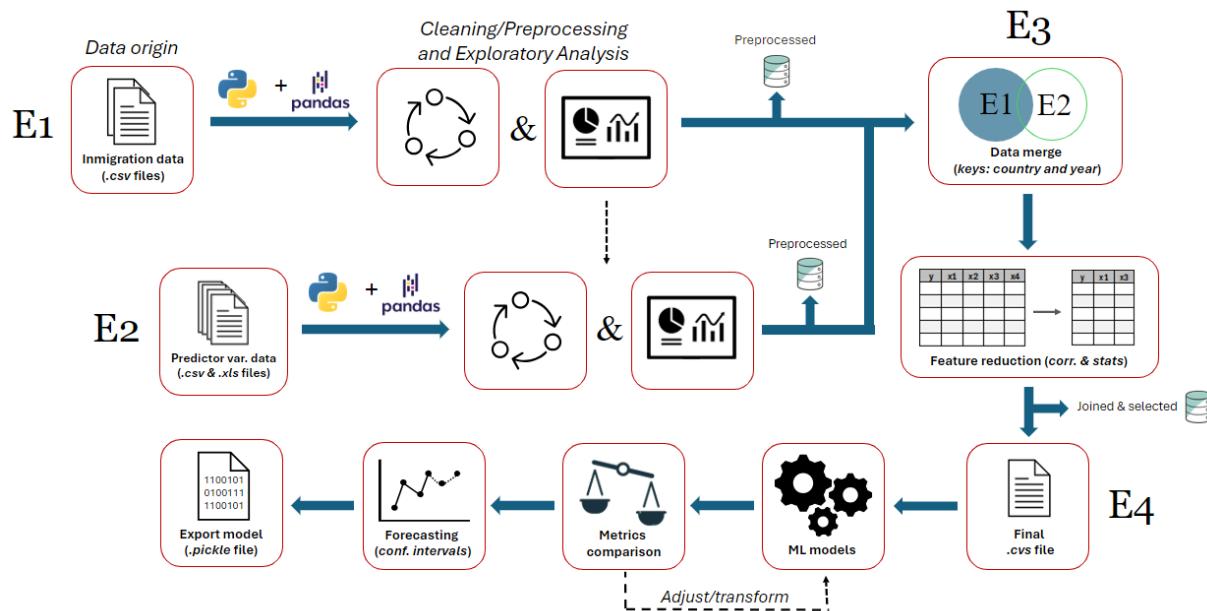


Figura 3. Esquema de las etapas del proyecto, teniendo: i) Etapa 1 (E1) para la limpieza, preprocesamiento y análisis de los datos de inmigración de España, ii) Etapa 2 (E2) orientada a una limpieza, preprocesamiento y análisis orientado según los *insights* de la etapa 1, iii) Etapa 3 (E3) de unión y selección de variables y iv) Etapa 4 (E4) para la prueba y comparación de modelos de *machine learning* con nuestro conjunto de datos final para realizar predicciones.

Posteriormente, se procedió con la unión de todas las variables predictoras al conjunto de datos central de datos de inmigración (Etapa 3 - E3), la inclusión de tres variables categóricas adicionales relacionadas a identificar los países de habla hispana y el período de pandemia/postpandemia (2020-2021 y 2022, respectivamente), así como la reducción de variables mediante el estudio de correlaciones y contrastes de hipótesis.

Finalmente, con los datos exportados en la Etapa 3, se probó algoritmos basados en modelos lineales, de árboles y redes neuronales de una o dos capas (Etapa 4 - E4), comparando distintas métricas para observar el rendimiento de los modelos, identificar los mejores, realizar ajustes de parámetros y transformación de distribución (a normal) necesarias para disminuir el error hasta obtener aquel con el mejor rendimiento posible. Luego se procedió a exportarlo junto con los demás archivos que aplicasen (escalado y/o transformación) para realizar predicciones, incluyendo la estimación de un intervalo de confianza del 90%.

Cada etapa se desarrolló por separado en un *jupyter notebook*, los cuales están disponibles en la carpeta “Notebooks” asociada con este proyecto.

5. RESULTADOS

5.1 Etapa 1: Limpieza, preprocesamiento y análisis de datos de inmigración

Inicialmente, como ambos conjuntos de datos de inmigración (2008 - 2021 y 2021-2022) presentaban información del año 2021, se filtró el segundo de los conjuntos de datos de inmigración para tener únicamente los datos que corresponden al año 2022 (referirse al *jupyter notebook* "Etapa 1 y Etapa 2" para mayores detalles). Luego, se compararon los valores que tomaban las variables categóricas (Figura 4), observando que:

- El conjunto de 2008-2021 estaba en idioma inglés, mientras que el de 2022 estaba en español.
- Había una cantidad muy grande de nacionalidades, y también observamos agrupaciones de regiones/continentes como: "UE27_2020 sin España", "Otro país de Asia", "América del Norte", entre otros.
- Los grupos de edades son diferentes entre los *datasets*, en los datos de 2008 – 2021 hay rangos de edades que aumentan intervalos de cinco años ("From 0 to 4 years old", "From 5 to 9 years", ...), mientras que en el segundo los rangos son más amplios ("De 0 a 15 años", "De 16 a 24 años", ...). Y también se observó que algunas de las categorías del 2008-2021 tenían espacios en blanco.

A partir de estas observaciones iniciales, se unificaron las categorías de sexo en inglés y se estandarizaron los grupos de edades (Figura 5).

```

----- esc[1m DATOS 2008 - 2021 esc[0m -----
esc[1m Nationality: esc[0m 71 valores
['Total' 'Spanish' 'País de la UE27_2020 sin España'
'País de la UE28 sin España' 'Belgium' 'Bulgaria' 'Denmark' 'Finland'
'France' 'Ireland' 'Italy' 'Netherlands' 'Poland' 'Portugal' 'Germany'
'Romania' 'Sweden' 'Lithuania' 'Other countries of the European Union'
'País de Europa menos UE27_2020' 'United Kingdom'
'País de Europa menos UE28' 'Norway' 'Switzerland' 'Ukraine' 'Moldova'
'Russia' 'Other European Union countries' 'De Africa' 'Algeria'
'The Gambia' 'GHANA' 'Guinea' 'Equatorial Guinea' 'Mali' 'Morocco'
'Mauritania' 'Nigeria' 'Senegal' 'Other African countries'
'De América del Norte' 'United States of America' 'Mexico' 'Canada'
'De Centro América y Caribe' 'Cuba' 'Honduras' 'Nicaragua'
'Dominican Republic' 'Other Central American and Caribbean countries'
'De Sudamérica' 'Argentina' 'Bolivia' 'Brazil' 'Colombia' 'Chile'
'Ecuador' 'Paraguay' 'Peru' 'Uruguay' 'Venezuela'
'Other South American countries' 'De Asia' 'Bangladesh' 'China'
'Philippines' 'India' 'Pakistan' 'Other Asian countries' 'De Oceanía'
'Stateless persons']

esc[1m Sex: esc[0m 3 valores
['Both sexes' 'Males' 'Females']

esc[1m Age group: esc[0m 20 valores
['Total' 'From 0 to 4 years old' 'From 5 to 9 years'
'From 10 to 14 years' 'From 15 to 19 years' 'From 20 to 24 years'
'From 25 to 29 years' 'From 30 to 34 years old'
'From 35 to 39 years old' 'From 40 to 44 years'
'From 45 to 49 years old' 'From 50 to 54 years old'
'From 55 to 59 years old' 'From 60 to 64 years old'
'From 65 to 69 years old' 'From 70 to 74 years' 'From 75 to 79 years'
'From 80 to 84 years' 'From 85 to 89 years' '90 years old and over']

----- esc[1m DATOS 2022 esc[0m -----
esc[1m Nationality: esc[0m 69 valores
['Total' 'Española' 'UE27_2020 sin España' 'Bélgica' 'Bulgaria'
'Dinamarca' 'Finlandia' 'Francia' 'Irlanda' 'Italia' 'Países Bajos'
'Polonia' 'Portugal' 'Alemania' 'Rumanía' 'Suecia' 'Lituania'
'Otro país de la Unión Europea sin España' 'Europa menos UE27_2020'
'Noruega' 'Reino Unido' 'Suiza' 'Ucrania' 'Moldavia' 'Rusia'
'Otro país del resto de Europa' 'De Africa' 'Argelia' 'Gambia' 'Ghana'
'Guinea' 'Guinea Ecuatorial' 'Mali' 'Marruecos' 'Mauritania' 'Nigeria'
'Senegal' 'Otro país de África' 'América del Norte'
'Estados Unidos de América' 'México' 'Canadá'
'De Centro América y Caribe' 'Cuba' 'Honduras' 'Nicaragua'
'República Dominicana' 'Otro país de Centro América y Caribe'
'De Sudamérica' 'Argentina' 'Bolivia' 'Brasil' 'Colombia' 'Chile'
'Ecuador' 'Paraguay' 'Perú' 'Uruguay' 'Venezuela'
'Otro país de Sudamérica' 'De Asia' 'Bangladesh' 'China' 'Filipinas'
'India' 'Pakistán' 'Otro país de Asia' 'De Oceanía' 'Apátridas']

esc[1m Sex: esc[0m 3 valores
['Ambos Sexos' 'Hombre' 'Mujer']

esc[1m Age group: esc[0m 8 valores
['Total' 'De 0 a 15 años' 'De 16 a 24 años' 'De 25 a 34 años'
'De 35 a 44 años' 'De 45 a 54 años' 'De 55 a 64 años' 'De 65 y más años']

```

Figura 4. Valores que toman las categorías de nacionalidades, sexo y grupos de edad para los *datasets* de inmigración de 2008-2021 (A) y 2022 (B).

----- Variable "Sex" -----	----- Variable "Age group" -----
Datos 2008 - 2021: 3 valores ['Both' 'Females' 'Males']	DATOS 2008 - 2021: 8 grupos ['All' '0 - 14' '15 - 24' '25 - 34' '35 - 44' '45 - 54' '55 - 64' '65+']
Datos 2022: 3 categorías ['Both' 'Males' 'Females']	DATOS 2022: 8 grupos ['All' '0 - 14' '15 - 24' '25 - 34' '35 - 44' '45 - 54' '55 - 64' '65+']

Figura 5. Valores las categorías de sexo (izquierda) y grupos de edad (derecha) post-limpieza/estandarización.

En cuanto a las nacionalidades, se removieron aquellas categorías relacionadas a regiones/continentes y se hizo una reducción del número de nacionalidades. Para ello, se consideraron como “top nacionalidades” a aquellas que, en conjunto y de forma decreciente, englobasen alrededor del 75% de los inmigrantes en cada año (Figura 6), seleccionando los valores únicos en una lista final para filtrar cada *dataset* con sus top, obteniendo un total de 27 nacionalidades (Figura 7).

```
# Lista vacía
top_nacio = []
years = grouped_0821.Year.unique().tolist()

# Bucle para extraer los top países por año en datos de 2008-2021
for year in sorted(years):
    # Filtrar datos, agrupar por nacionalidad, sumar inmigrantes y convertir a dataframe
    sum_year = grouped_0821[(grouped_0821['Year'] == year) &
                           (grouped_0821['Sex'] == 'Both') &
                           (grouped_0821['Age group'] == 'All')]\n                           .groupby(['Nationality'])\n                           ['Immigrant count'].sum().sort_values(ascending = False)\n                           .to_frame().reset_index()

    # Obtener porcentaje en relación al total de inmigrantes en el año
    sum_year['% of total'] = sum_year['Immigrant count'] / sum_year['Immigrant count'].sum() * 100
    # Obtener suma acumulada de los porcentajes
    sum_year['% acumulado'] = sum_year['% of total'].cumsum()
    # Filtrar top nacionalidades con suma acumulada menor a 76% (punto de corte 75%)
    sum_year = sum_year[sum_year['% acumulado'] <= 76]
    # Crear una lista con las nacionalidades del top
    top = sum_year['Nationality'].tolist()
    # Agregar en lista vacía el top nacionalidades de cada año
    top_nacio = top_nacio + top

top_nacio

# Lista vacia para valores únicos
top_unicos = []

# Recorrer todos los elementos de la lista "top_paises"
for nacionalidad in top_nacio:
    # Revisar si existe en "top_unicos" o no, y agregar a la lista si no está
    if nacionalidad not in top_unicos:
        top_unicos.append(nacionalidad)

top_unicos
```

Figura 6. Código empleado para encontrar las top nacionalidades del *dataset* 2008 - 2021 considerando cada año. Se siguió un procedimiento similar para el conjunto de datos de 2022.

Después de haber filtrado los datos de inmigración con el top nacionalidades, se estandarizó el formato de estas en una nueva columna “Nationality code” usando la nomenclatura internacional de tres códigos (ALPHA-3), descrita en la ISO 3166¹¹, para luego proceder a juntar ambos conjuntos de datos en un *dataframe* (Figura 8).

----- Variable "Nationality code" -----	
DATOS 2008 - 2021: 27 grupos	
['DZA' 'ARG' 'BRA' 'BGR' 'CHN' 'COL' 'CUB' 'DOM' 'ECU' 'FRA' 'DEU' 'HND' 'ITA' 'MAR' 'NIC' 'PAK' 'PRY' 'PER' 'PRT' 'ROU' 'RUS' 'SEN' 'ESP' 'UKR' 'GBR' 'USA' 'VEN']	
DATOS 2022: 27 grupos	
['ESP' 'BGR' 'FRA' 'ITA' 'PRT' 'DEU' 'ROU' 'GBR' 'UKR' 'RUS' 'DZA' 'MAR' 'SEN' 'USA' 'CUB' 'HND' 'NIC' 'DOM' 'ARG' 'BRA' 'COL' 'ECU' 'PRY' 'PER' 'VEN' 'CHN' 'PAK']	

Figura 7. Top nacionalidades (en códigos ISO) de los conjuntos de datos de inmigración de 2008-2021 y 2022.

<class 'pandas.core.frame.DataFrame'>				
RangeIndex: 9720 entries, 0 to 9719				
Data columns (total 5 columns):				
#	Column	Non-Null Count	Dtype	
0	Year	9720 non-null	int64	
1	Nationality code	9720 non-null	object	
2	Sex	9720 non-null	object	
3	Age group	9720 non-null	object	
4	Immigrant count	9720 non-null	int64	
dtypes: int64(2), object(3)				
memory usage: 379.8+ KB				

Year	Nationality code	Sex	Age group	Immigrant count
0 2008	DZA	Both	0 - 14	759
1 2008	PER	Males	35 - 44	2938
2 2008	PER	Males	45 - 54	1128
3 2008	PER	Males	55 - 64	265
4 2008	PER	Males	65+	156
...
9715 2022	PAK	Males	55 - 64	330
9716 2022	PAK	Females	55 - 64	146
9717 2022	PAK	Both	65+	169
9718 2022	PAK	Males	65+	99
9719 2022	PAK	Females	65+	70

Figura 8. Dataframe de la unión de los datasets de datos de inmigración post-limpieza/preprocesamiento

Finalmente, continuando con los análisis de los valores totales de inmigración con relación a su evolución en los años (Figura 9) y por grupo de edad/sex (Figura 10), se observó:

- Entre 2008-2013 hay una caída de número de inmigrantes y a partir del 2014 aumenta progresivamente hasta alcanzar un pico de 750,480 inmigrantes en 2019, para luego disminuir en el 2020 y 2021 (período de pandemia y restricciones sanitarias relacionadas al

COVID 19) y mostrar un gran salto hasta ≈ 1.25 millones de inmigrantes en el 2022 (postpandemia), superando al 2019 en aproximadamente 500,000 inmigrantes.

- Observando los inmigrantes según el grupo de edad, vemos que predomina la inmigración de personas jóvenes de entre 25-34 años, seguidos de jóvenes de entre 15-24. A partir de este último grupo, es notable la disminución progresiva del número de inmigrantes con el incremento de la edad.
- En cuanto al sexo, vemos que la cantidad de mujeres y hombres es similar en todos los grupos, especialmente en los grupos de mayor presencia (15-24 y 25-34 años). Únicamente los grupos de entre 0-14 y 55-64 años muestran una mayor diferencia: el primero hacia los hombres y el segundo hacia las mujeres.

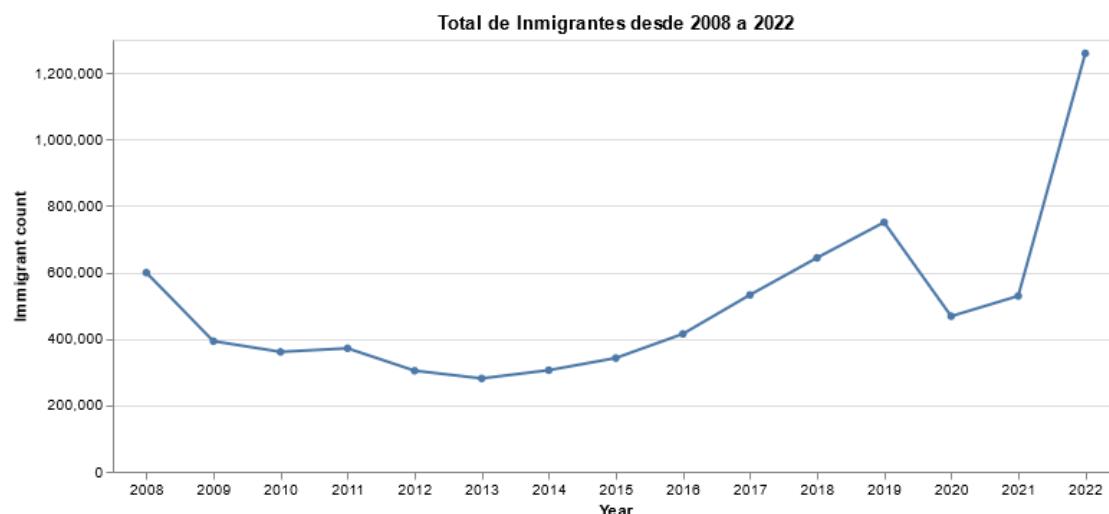


Figura 9. Total de inmigrantes en España desde el año 2008 al 2022.

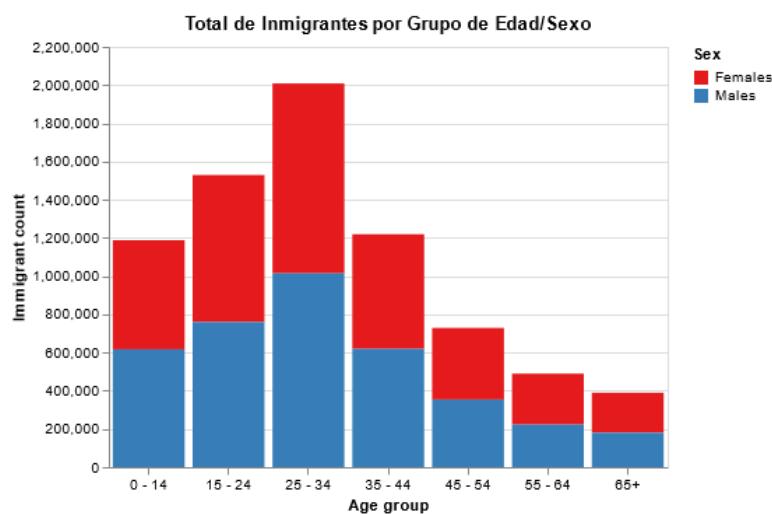


Figura 10. Cantidad de inmigrantes en España por grupo de edad y sexo.

Si observamos la distribución por año en mediante gráficos de cajas (Figura 11), vemos que a medida que hay mayor inmigración, también hay una mayor dispersión de los datos y los valores atípicos son más “extremos”. Esto se debe a que el flujo de inmigrantes por nacionalidad no es igual

y hay algunas que predominan sobre otras, las cuales generan la dispersión observada. Adicionalmente, observando el rango intercuartílico entre 2008-2021, notamos que la mayoría de los datos se concentran en un rango medio/bajo de número de inmigrantes, alrededor o por debajo de 10,000 inmigrantes.

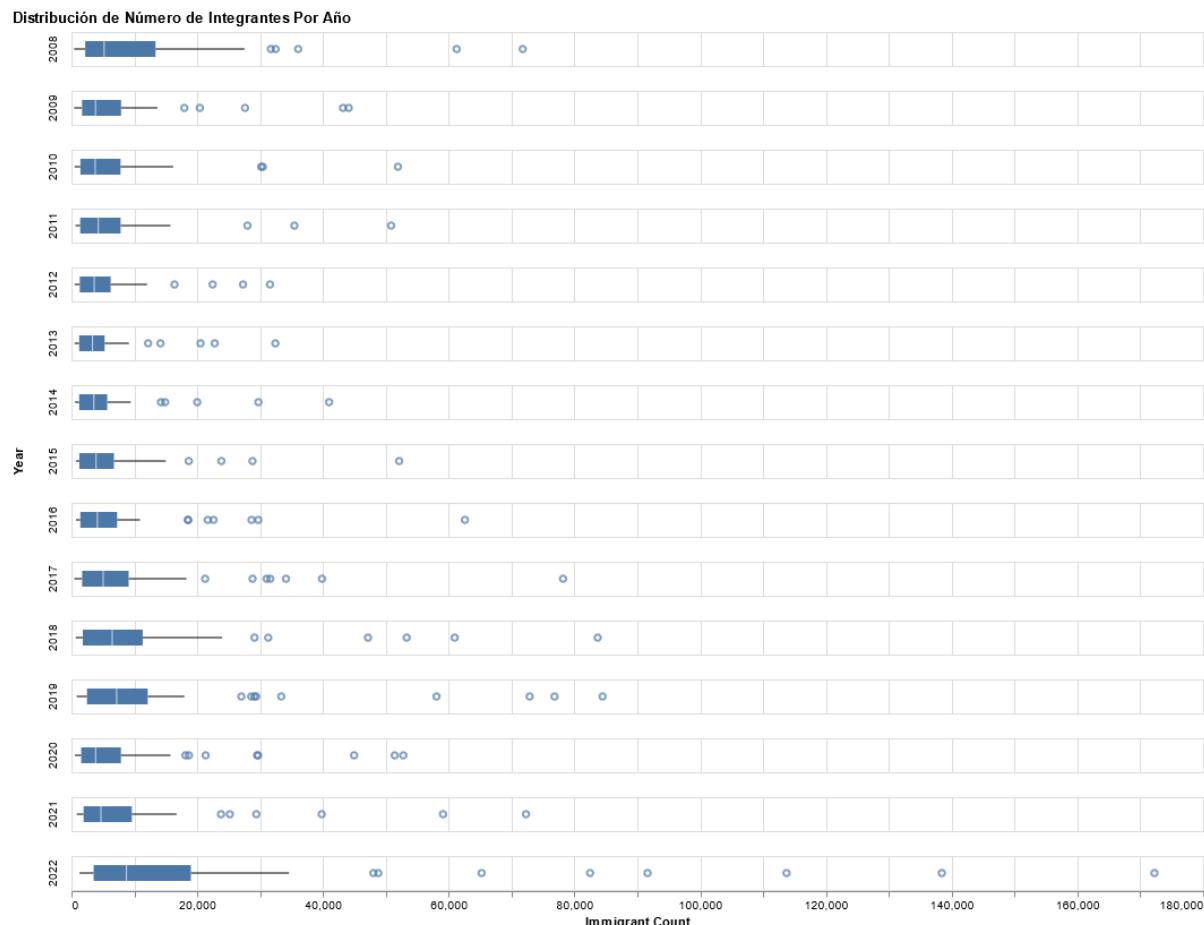
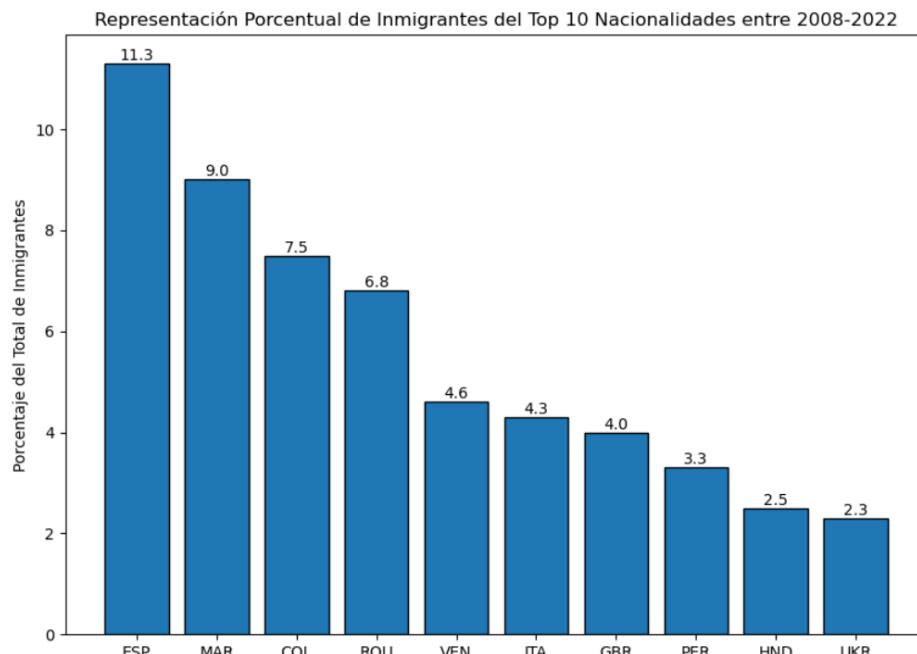


Figura 11. Distribución de inmigrantes en España por año desde el 2008 a 2022.

A partir de estas observaciones, se analizó la cantidad total por nacionalidad, observando que los inmigrantes de nacionalidad española, colombiana, marroquí y rumana poseen la mayor representación porcentual en relación con el total de inmigrantes (Figura 12A). Además, si observamos el número de inmigrantes por año dentro del Top 5 (Figura 12B), notamos que hay diferencias entre los mismo. Por ejemplo, los inmigrantes colombianos y venezolanos están en un rango medio/bajo durante 2008-2015, pero a partir del 2016 su número comienza a aumentar drásticamente. Incluso podemos ver como los colombianos se convierten en la nacionalidad con el mayor número de inmigrantes en el 2022 (postpandemia).

Así mismo, vemos como los rumanos presentan un descenso a partir del 2012 y el número de inmigrantes se estabiliza. Por otro lado, en cuenta a crecimiento, el incremento de españoles es similar a venezolanos y marroquíes a partir del 2012, pero vemos que su número incrementa a un ritmo diferente.

(A)



(B)

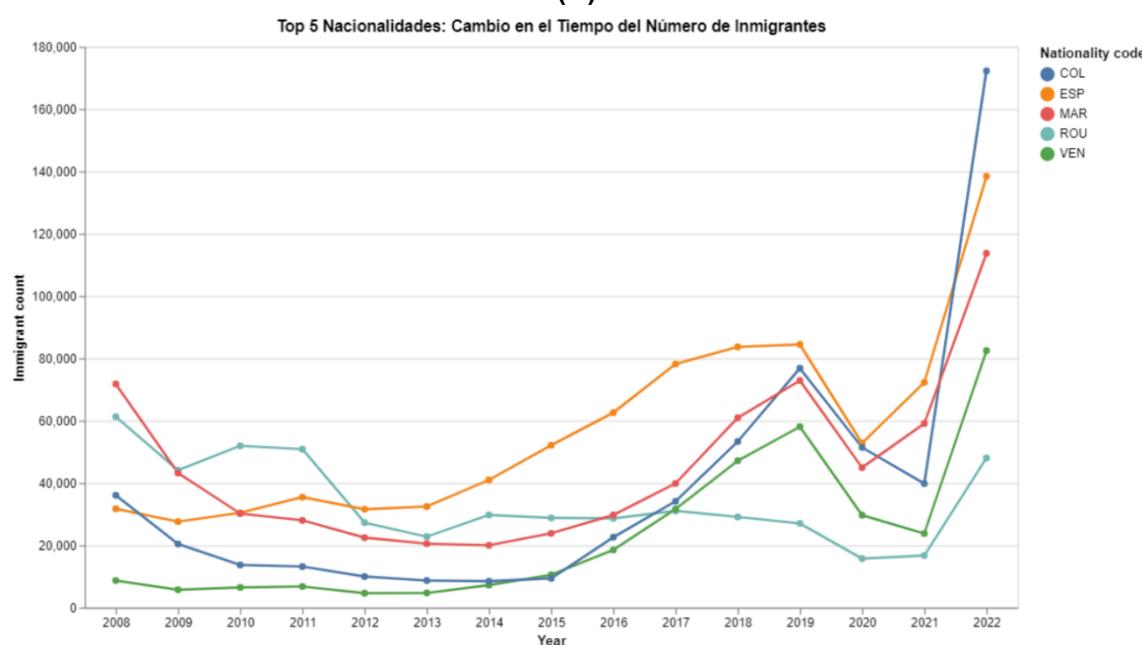


Figura 12. Top 10 nacionalidades en porcentaje de inmigrantes en España durante el periodo 2008-2022 (A) y el cambio en el número de inmigrantes en ese período para el Top 5 (B).

Estos datos indican que las condiciones específicas de cada nación, o nación de origen, del inmigrante tiene un efecto importante a considerar, por lo que, emplear el enfoque macro por país parece ser adecuado.

Otra observación que cabe destacar, relacionada con los grupos de edades, fue que únicamente en el año 2022 se observa un cambio en la tendencia en comparación con la observada previamente en los valores totales de inmigración. Para el 2022, tenemos que el grupo de 15-24

años presenta una cantidad de inmigrantes menor a la del grupo 0-14 años y el grupo de 35-34 supera en cantidad a ambos (Figura 13).

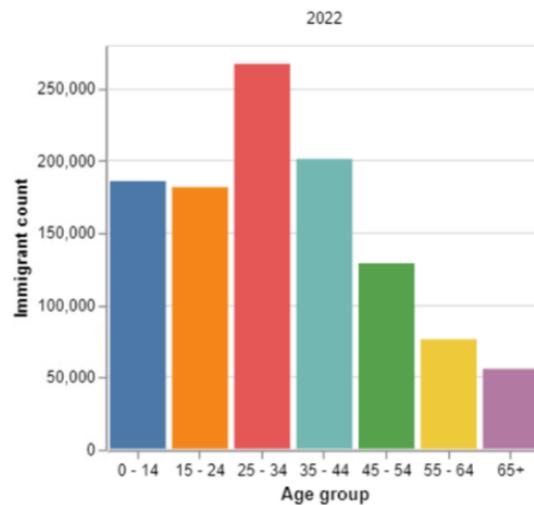


Figura 13. Cantidad de número de inmigrantes en España por grupo de edad en el año 2022.

Es importante destacar que, antes de proceder a exportar el *dataset* con los datos de inmigrantes limpios y preprocesados, se procedió a remover a la nacionalidad española del mismo considerando que nuestro enfoque macro implica que usaremos variables/datos de los países de cada nacionalidad. Por lo que, se estarían asociando datos relacionados con España a los inmigrantes. En consecuencia, los datos de inmigrantes de nacionalidad española requieren un estudio y modelado aparte que no se abarca en este proyecto.

El fichero en formato *top_inmigracion_2008_2022.csv* de los datos de inmigración obtenido de este proceso de limpieza y preprocesamiento, puede en la carpeta “13 - Exports (preprocesamiento)” asociada a este proyecto.

5.2 Etapa 2: Limpieza, preprocesamiento y análisis de variables explicativas

Los resultados se presentan segmentados por variable, enfocados en los aspectos más relevantes, debido a las particularidades que presentaba cada uno de los conjuntos de datos y con la finalidad de seguir el mismo esquema de trabajo empleado. Para mayores detalles del proceso de limpieza y análisis, puede consultarse los *jupyter notebook* que mencionan la “Etapa 2” en la carpeta “Notebooks”, así como los ficheros *.csv* resultantes del preprocesamiento de cada conjunto de datos en la carpeta “13 - Exports (preprocesamiento)”.

5.2.1 Continentes y Subregiones

Los datos de continentes y sub-regiones¹² se filtraron por mapeo con el top nacionalidades obtenidos de la Etapa 1 y se agregaron los códigos de país correspondientes, abarcando cuatro continentes y siete subregiones (Figura 14).

```

Continent: 4 valores
['Europe' 'Africa' 'America' 'Asia']

Sub-region: 7 valores
['European Union' 'Rest of Europe' 'Africa' 'North America'
'Central America and Caribbean' 'South America' 'Asia']

Nationality code: 26 valores
['BGR' 'FRA' 'ITA' 'PRT' 'DEU' 'ROU' 'GBR' 'UKR' 'RUS' 'DZA' 'MAR' 'SEN'
'USA' 'CUB' 'HND' 'NIC' 'DOM' 'ARG' 'BRA' 'COL' 'ECU' 'PRY' 'PER' 'VEN'
'CHN' 'PAK']

```

Figura 14. Contenientes y subregiones luego de aplicar el filtro del top nacionalidades en inmigración en España.

Al graficar la distribución de nacionalidades en el top de inmigración en relación con los contenientes y subregiones (Figura 15), se observó que:

- América es el continente con mayor cantidad de países dentro del top nacionalidades en inmigración ($\approx 45\%$, o 12 de 26), seguido de Europa con $\approx 35\%$ (9 de 26).
- Si se toma en cuenta las subregiones, vemos que la mayor cantidad dentro de América se ubican en América del Sur (7 de 26), y que los de Europa son, mayoritariamente, parte de la Unión Europea, de la cual España también forma parte.

Esto se contrasta con las observaciones hechas en la Figura 12A, donde incluso se ven a países como Colombia, Venezuela, Perú y Honduras dentro del top 10 de nacionalidades en inmigración durante 2008 - 2022.

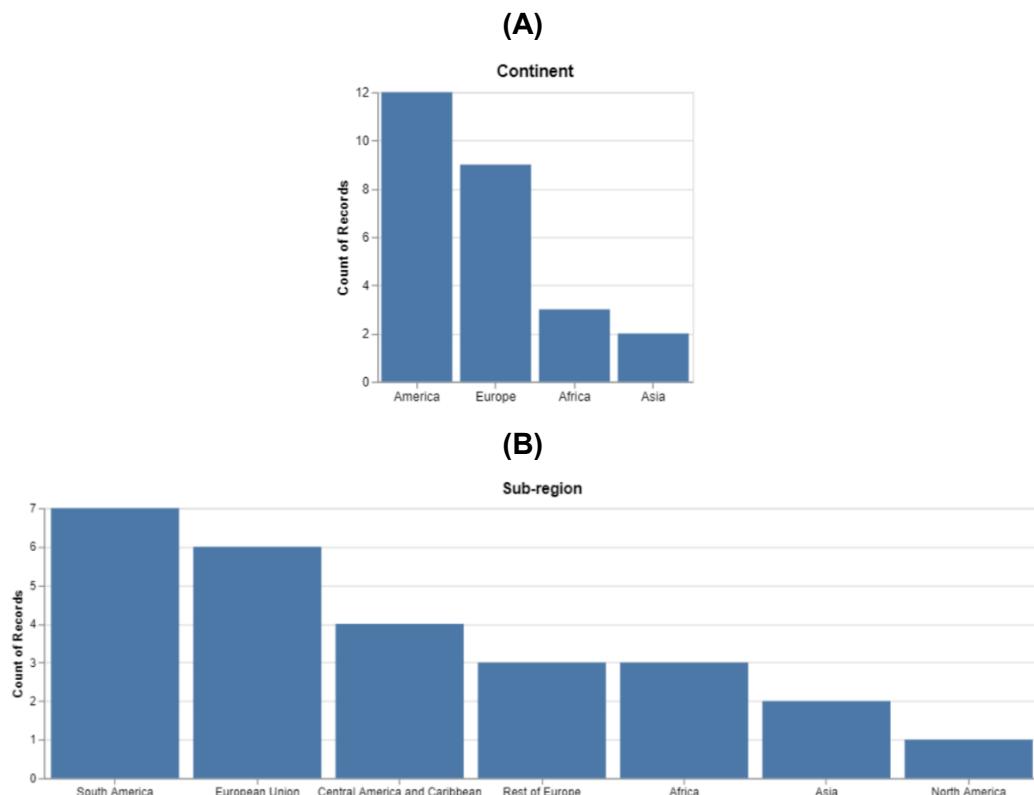


Figura 15. Cantidad de países por continente (A) y subregión (B) dentro del top nacionalidades en inmigración en España.

5.2.2 Padrón de Inmigrantes Residentes en España

Para los datos del padrón de extranjeros residentes en España¹³, encontramos situaciones similares a las encontradas en el conjunto de datos de inmigración de 2008 a 2022, principalmente siendo éstos agrupaciones respecto a los países de origen de los inmigrantes.

```
# Identificamos variables de grupos innecesarias para nuestro análisis
grupos = [
    "PAÍSES AFRICANOS",
    "PAÍSES ASIÁTICOS",
    "PAÍSES DE OCEANÍA",
    "Resto de Países Africanos",
    "Resto de Países de Oceanía",
    "UNIÓN EUROPEA (27_2020)",
    "PAÍSES AMERICANOS",
    "PAÍSES EUROPEOS",
    "PAÍSES EUROPEOS NO UE(27_2020)",
    "Resto de América Central y Caribe",
    "Resto de América del Sur",
```

Figura 16. Extracto de grupos de países encontrados en padrón de residentes.

```
# Definimos diccionario con base en estándar ISO3 para estandarización de datos
codigo = {"Albania": "ALB",
          "Germany": "DEU",
          "Andorra": "AND",
          "Angola": "AGO",
          "Saudi Arabia": "SAU",
          "Algeria": "DZA",
          "Argentina": "ARG",
          "Armenia": "ARM",
          "Australia": "AUS",
```

Figura 17. Extracto de estandarización con ISO 3166.

El resultado sobre el análisis de los datos de los residentes arroja claras tendencias sobre el comportamiento actual del fenómeno de inmigración en España. Es siempre notoria la presencia de ciertos países más que otros en los datos históricos que tomamos como base. También observamos que los niveles de población extranjera cambian a ritmos diferentes conforme pasan los años.

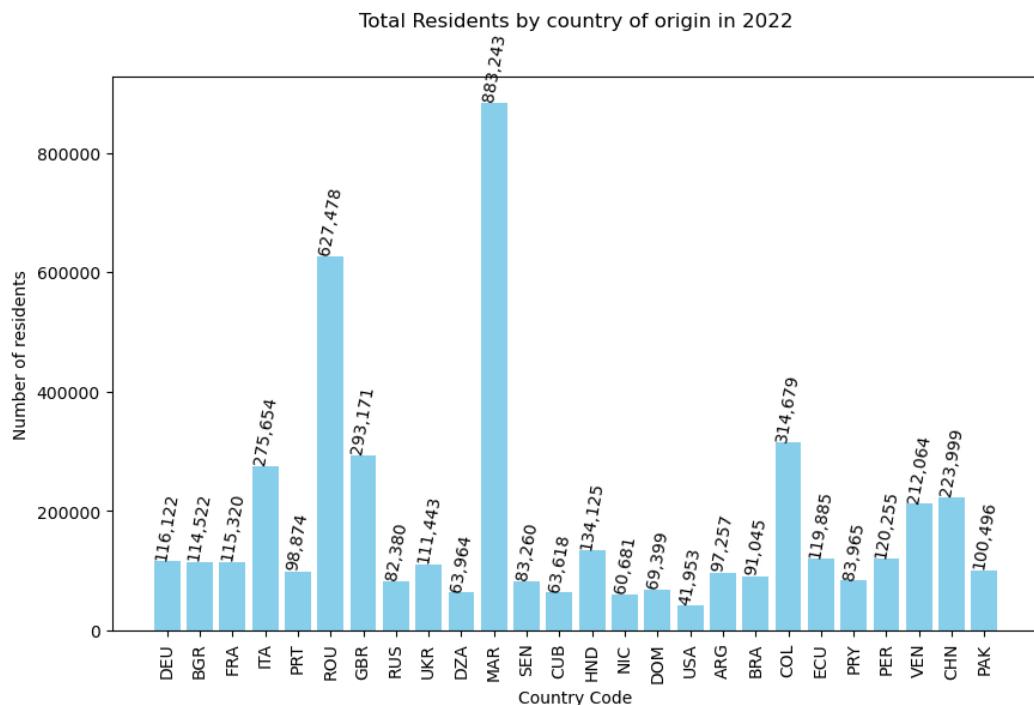


Figura 18. Suma de residentes por país de origen en el año 2022.

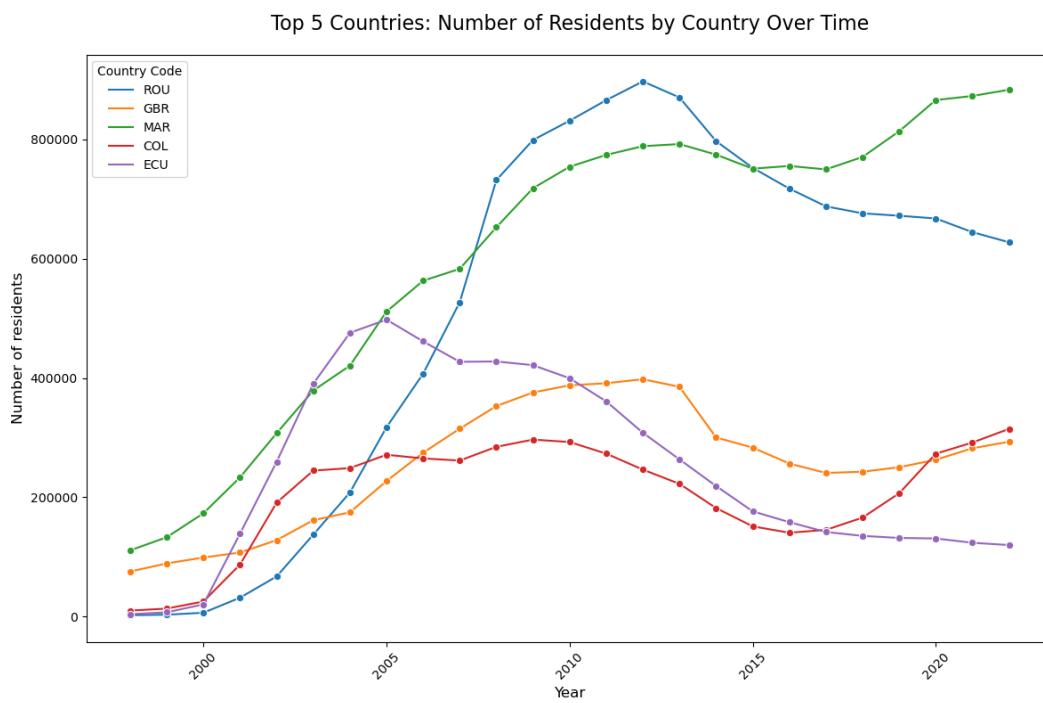


Figura 19. Variación anual de los 5 países con mayor cantidad de residentes extranjeros en España.

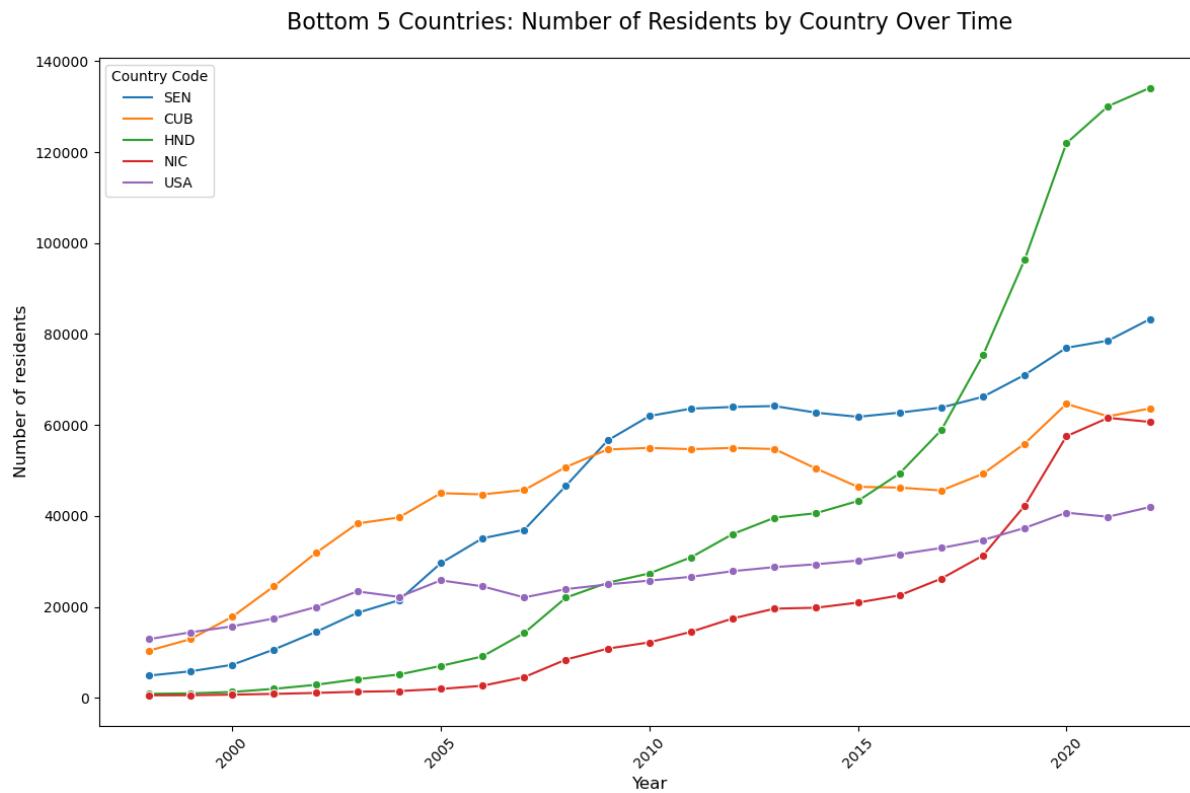


Figura 20. Variación anual de 5 países con menor presencia de residentes extranjeros en España.

Como observaciones generales, tenemos que:

- Marruecos y Rumanía se colocan como los principales países de origen de inmigrantes.
- Se observan periodos de disminución de residentes por emigración, por lo que es importante considerar el equilibrio entre inmigración/emigración.
- Hay caso de crecimiento abrupto de residentes extranjeros, como se observa para los ecuatorianos, marroquíes y rumanos a partir del año 2000, o hondureños a partir del 2016.
- Las tendencias parecen no tener relación directa con idioma o históricos, por lo que estas variables aisladas no parecen ser significativas de forma aislada, y habrá que considerar su interacción con otras, como regiones geográficas, condiciones socioeconómicas, políticas y otras.

5.2.3 Índices de Desarrollo

El estudio de los diversos índices de desarrollo socioeconómico requirió el uso de diversas fuentes^{14, 15, 16}, incluyendo su adecuada transformación, para completar la información faltante de algunos dentro del período 2008 – 2022, siendo necesario imputar con la media un solo valor de la variable “Salaried workers %” para Ucrania.

La descripción inicial de los datos de los índices de desarrollo con algunos estadísticos descriptivos (Figura 21) mostró todas las variables presentan alta dispersión (observa la relación de

la media/desviación estándar y mínimo/máximo). Lo que se aprecia en mayor magnitud para la variable económica “Inflation_anual”, que presenta una dispersión muy alta.

Esto nos dice que tenemos países con índices en ambos extremos, tanto niveles bajos como niveles muy altos.

	Unemployment %	Political and Violence Percentile	Probability of dying young	Regulatory Quality Percentile	Rule of Law Percentile	Voice and Accountability Percentile	Salaried workers %	GDP_growth	Inflation_anual
count	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000	405.000000
mean	7.320049	40.355729	4.977778	51.062514	46.169861	50.855337	67.271618	1.915147	232.486373
std	3.952777	20.396348	3.071540	26.643003	26.585179	26.370549	17.865015	5.133622	3395.366879
min	0.420000	0.473934	1.300000	0.000000	0.000000	3.846154	27.506375	-30.000000	-11.161615
25%	4.590000	25.000000	2.700000	32.227489	25.118483	28.078817	51.437373	0.736267	1.503739
50%	6.760000	41.981133	4.500000	48.095238	41.428570	52.112675	68.668624	2.694907	3.579787
75%	9.250000	57.075470	6.800000	73.684212	62.857143	74.647888	84.020816	4.410481	7.667622
max	33.300000	89.150940	18.100000	98.571426	95.734596	96.059113	94.135947	13.417530	65374.081840

Figura 21. Tabla de estadísticos descriptivos para las variables de desarrollo.

Al observar la distribución de estas variables y los datos atípicos (Figura 22), resaltaron los siguientes aspectos:

- Los datos de las variables “GDP_Growth” (Figura 22B) y “Unemployment %” (Figura 22I) son las que se asemejan más a una distribución normal, pero ambas presentan valores extremos atípicos.
- La inflación anual (Figura 22E) presenta datos tan dispersos que un histograma sólo puede agruparlos en valores de inflación negativa (desinflación) y valores positivos (aumento de inflación). Nótese que en los gráficos de cajas solo se distinguen los valores atípicos debido al amplio rango en los valores.
- Para “Probability of dying young” (Figura 22C), la mayoría de los valores están en un rango de probabilidad baja, con algunos datos que muestran una mayor probabilidad, mientras que para “Salaried workers %” (Figura 22H) predomina un porcentaje medio/alto.
- El resto de las variables no poseen una tendencia clara, sino que presentan grupos con menor o mayor valores a lo largo del rango de los datos.

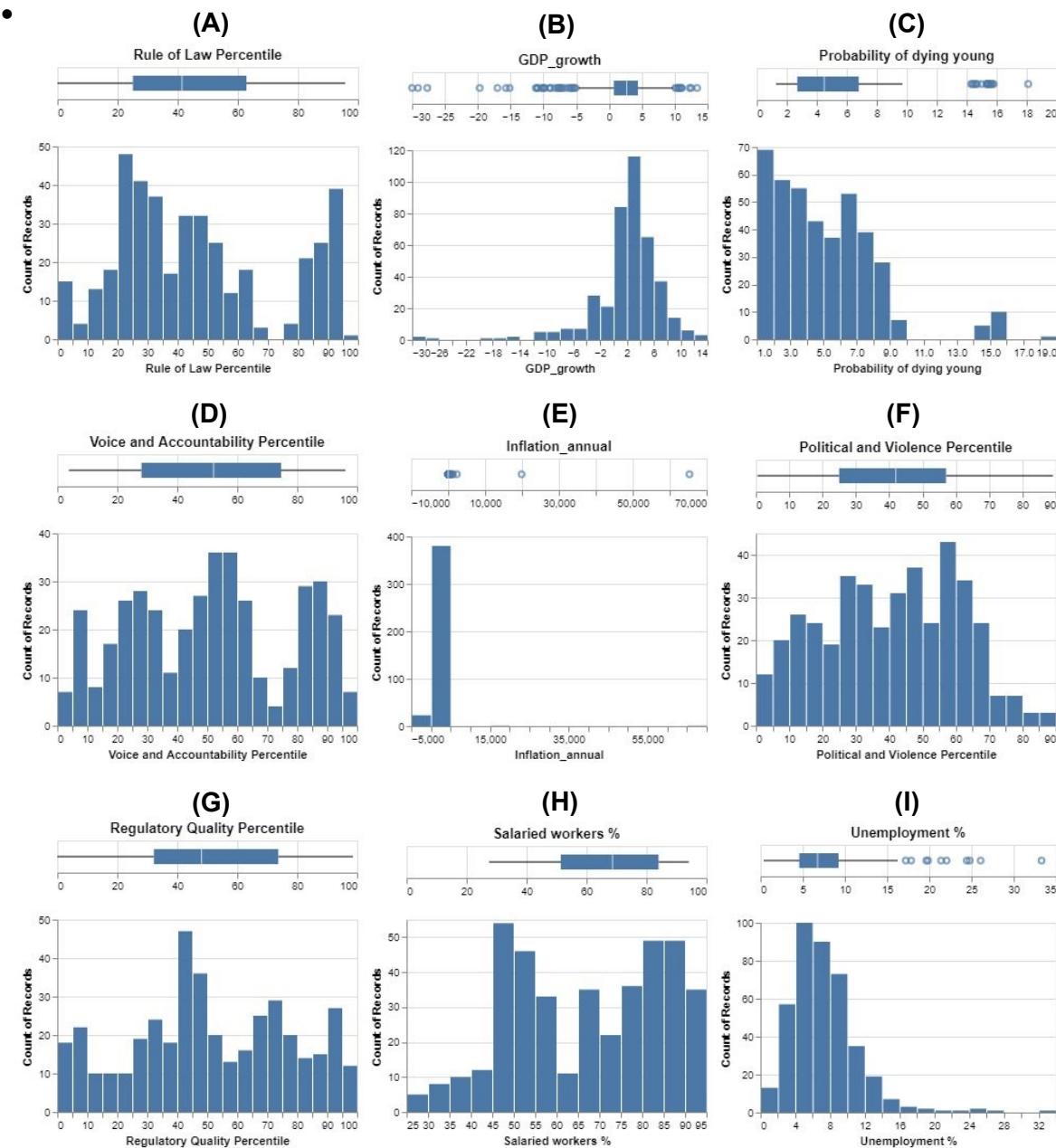


Figura 22. Distribución de las variables de índices de desarrollo y detección de datos atípicos. (A) “Rule of Law Percentile”, (B) “GDP_growth”, (C) “Probability of dying young”, (D) “Voice and Accountability Percentile”, (E) “Inflation_annual”, (F) “Political and Violence Percentile”, (G) “Regulatory Quality Percentile”, (H) “Salaried workers %” y (I) “Unemployment %”.

Luego, al comparar el cambio en el tiempo de alguna de estas variables en contraste con la tendencia de los datos de inmigración del Top 4 –Colombia, Marruecos, Rumanía y Venezuela– entre las nacionalidades (Figura 23), se observó que:

- Para Venezuela, observando el rango de 2013 a 2019, es evidente la correlación inversa con la inmigración en los datos de PBI (Figura 23A vs Figura 23C). Sin embargo, dicha relación es de menor magnitud en los otros top países.

- También observamos como, en relación con los percentiles de estado de derecho y estabilidad política/violencia (Figura 23B), los países del Top 4 están en los percentiles medios/bajos.
- Es interesante ver que los dos países de América del Sur del Top 4 (Colombia y Venezuela) están en los percentiles más bajos, particularmente a partir del 2017, siendo Venezuela el caso más extremo.
- Nótese también que Rumanía mantiene un mejor percentil en el tiempo y es, además, la nacionalidad que muestra la menor variación en el número de inmigrantes por año. Caso contrario para Colombia, Venezuela y Marruecos cuyo número de inmigrantes incrementa significativamente y son, además, los países peor ubicados en los percentiles.

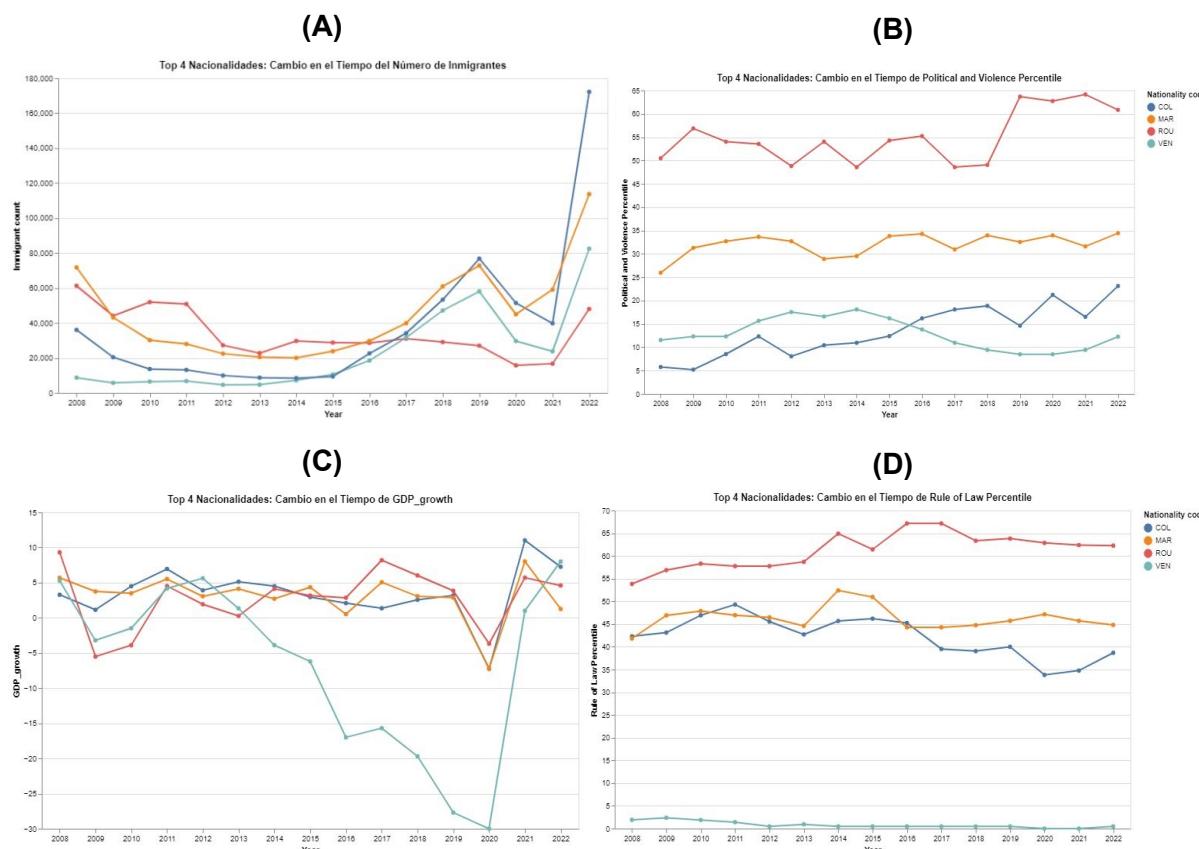


Figura 23. Contraste entre la tendencia de cambio en el tiempo de los datos de inmigración del Top 4 nacionalidades (A) y algunos índices de desarrollo: (B) "Political and Violence Percentile", (C) "GDP_growth", y (D) "Rule of Law Percentile".

Esto nos muestra que cada país/variable se comporta y relaciona de forma diferente con la inmigración. Además, esto también deja en claro la necesidad de incluir más variables que describan otros tipos características y escenarios.

5.2.4 Régimen Político

Los datos de regímenes políticos¹⁷ están en una escala discreta del cero al siete, yendo desde los tipos de regímenes más autocráticos hasta los más democráticos (referirse al Anexo 1 para las definiciones correspondientes).

Luego de filtrar las 26 nacionalidades (top nacionalidades) entre 2008 - 2022 y graficar la cantidad de cada tipo de régimen por año (Figura 24), se observó que predominan regímenes políticos del tipo "7" (poliarquías), seguido de tipo "6" (democracias electorales). También vemos la ausencia de regímenes "4" y "5" (democracia excluyente y democracia masculina), y luego tenemos un grupo que son menores o iguales 3 (autocracias). Además, se observa que los conteos se mantienen hasta el 2015 y, a partir de allí, hay algunas variaciones, de las cuales resalta el incremento de la cantidad de regímenes tipo "1" (autocracia unipartidista), hasta llegar a estar presente en 4 países para el año 2022, y la disminución de regímenes tipo "7" e incremento de tipo "6" hasta alcanzar una cantidad similar en el 2022.

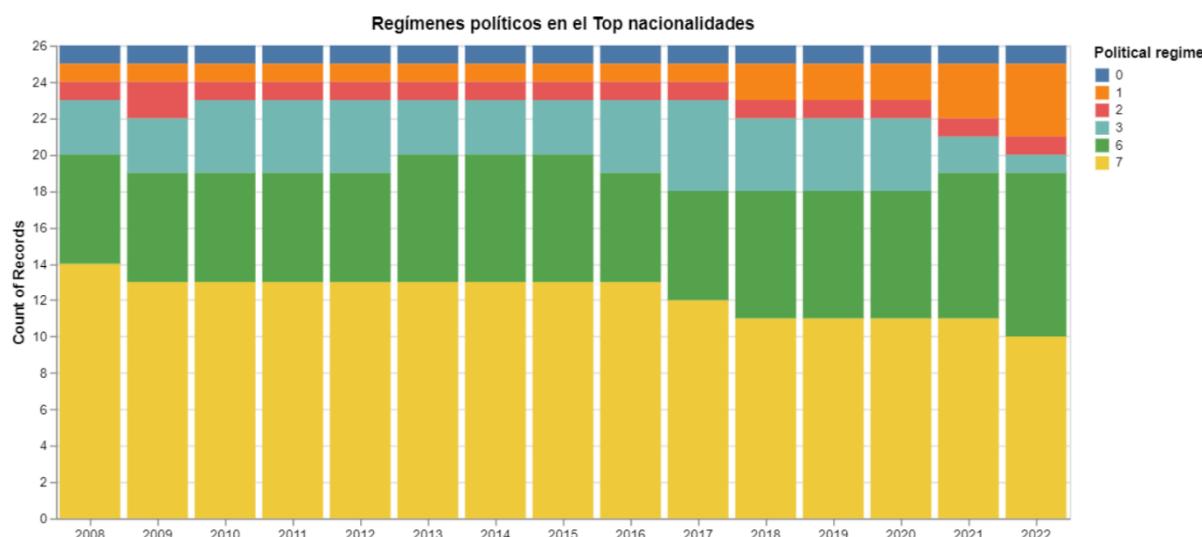


Figura 24. Conteo de países en el top nacionalidades en inmigración por cada tipo de régimen político por año.

Adicionalmente, entre los países con algún tipo de régimen autocrático (0, 1, 2 o 3) entre 2008 – 2022, se identificaron tres países dentro del Top 10: Marruecos, Venezuela y Honduras (Figura 25).

----- Países con regímenes autocráticos entre 2008 - 2022 -----	
['DZA' 'CHN' 'CUB' 'HND' 'MAR' 'NIC' 'RUS' 'VEN']	

Figura 25. Países dentro del top nacionalidades en inmigración con regímenes políticos autocráticos entre el año 2008 y 2022.

5.2.5 Índices de Democracia

Los datos de índice de democracia liberal y deliberativa basado en las estimaciones de expertos y el índice de V-Dem¹⁸ muestran una tendencia similar, en cuanto su distribución, con algunas de las variables de índices de desarrollo observadas en el punto 5.2.3 (Figura 26), pudiéndose destacar que:

- El índice de democracia liberal presenta tres grupos de concentración de datos entre nuestros top nacionalidades en inmigración: aquellos con índice bajo (< 0.35), otros en la zona media/media-alta y otro con índices de democracia alto (> 0.75).
- Vemos que para el índice de democracia deliberativa no se distinguen tres grupos como en el anterior, sino que observan algunos picos en las tres zonas, como entre $0.25 - 0.35$, $0.55 - 0.65$ y $0.80 - 0.85$.

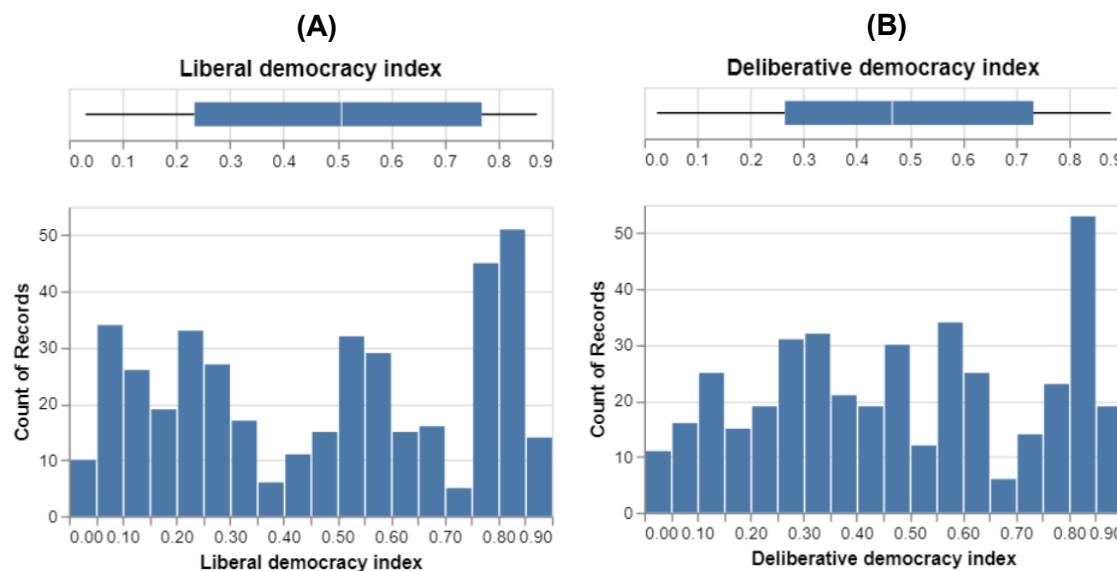
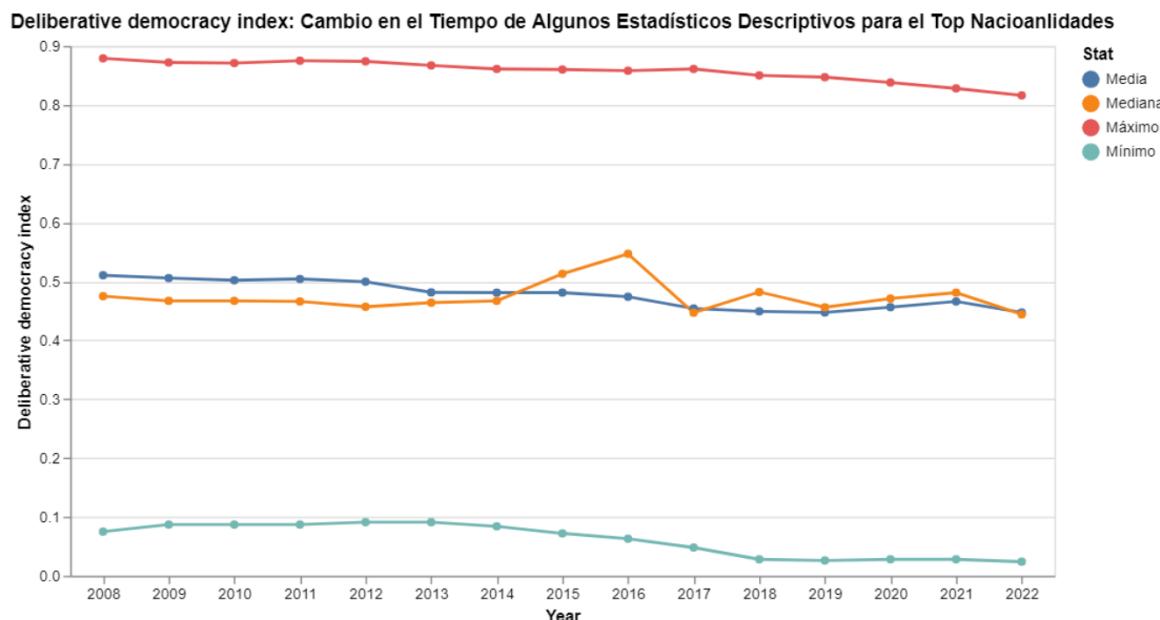


Figura 26. Distribución de las variables de índice de democracia liberal (A) y deliberativa (B), y detección de datos atípicos.

Además, luego de estimar y graficar algunos estadísticos descriptivos por año para ambos índices (Figura 27), se observó que:

- Considerando las observaciones en cuanto a la distribución, se esperaba, observar el mínimo en torno a 0 y máximo próximo a 1, pero la similitud entre la media y mediana nos dice que la relación entre los índices de los países en el top de nacionalidades cambia en una forma similar a través del período 2008 - 2022.
- A partir de 2016 hay una leve disminución de los índices de democracia, especialmente notable en el índice de democracia deliberativa.

(A)



(B)

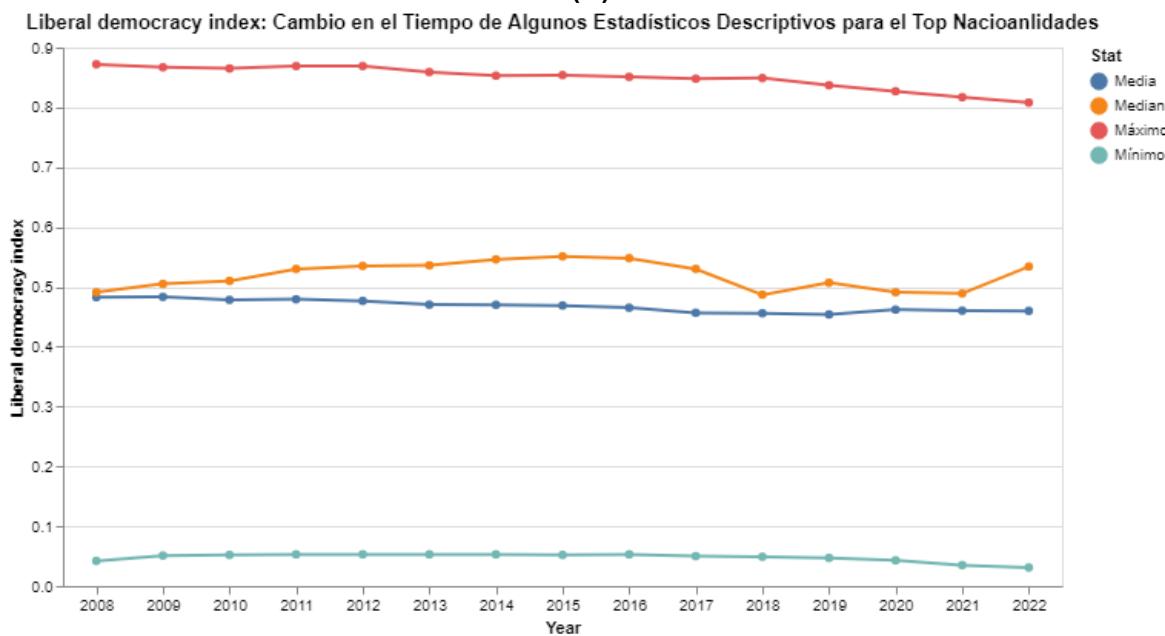


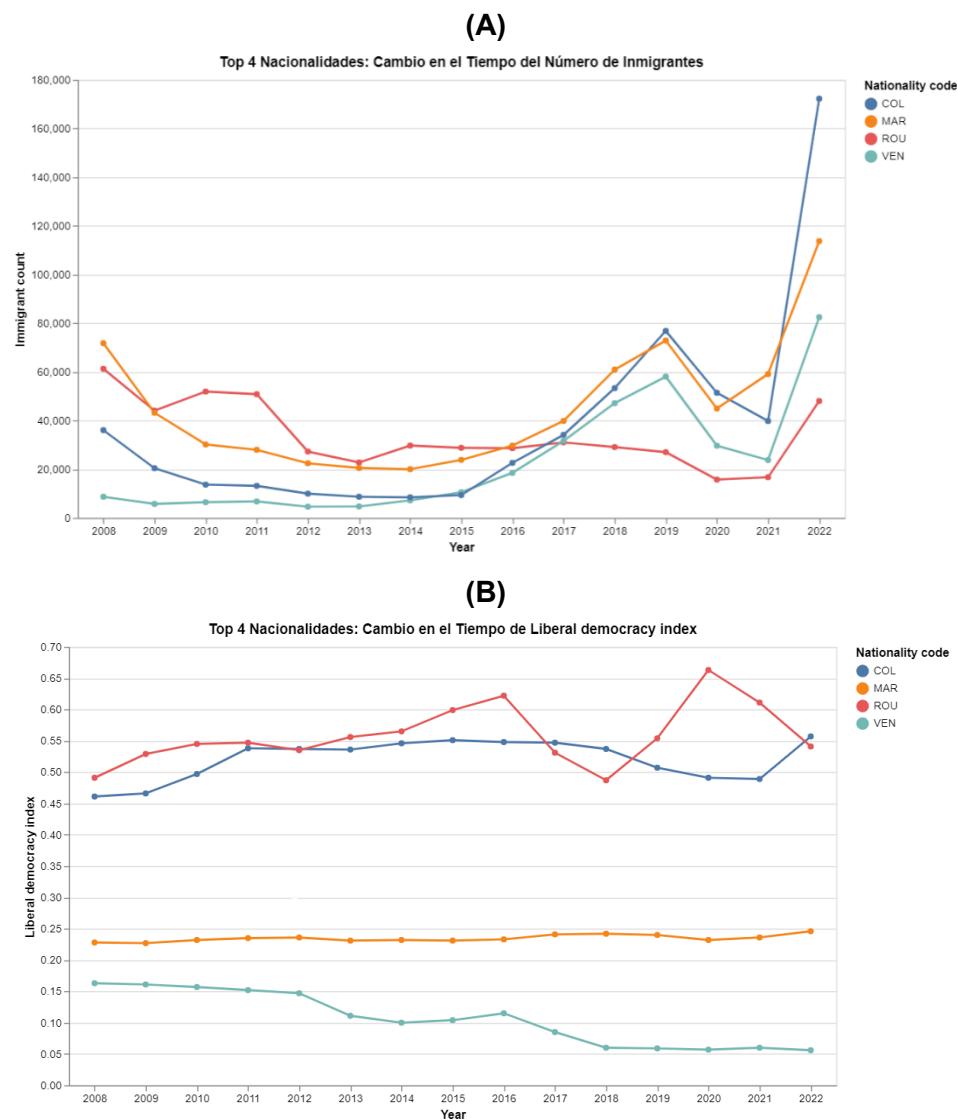
Figura 27. Variación anual del máximo, mínimo, media y mediana para el índice de democracia deliberativa (A) y liberal (B).

Finalmente, al contrastar el cambio en los años de la cantidad de inmigrantes (Figura 28A) con los índices de las Top 4 nacionalidades en inmigración (Figura 28B, 28C) vemos que:

- Hay mayor variación en el tiempo para el índice de democracia deliberativa.
- Los índices de Colombia y Rumanía oscilan en la zona media, mientras que Marruecos y Venezuela se mantienen en la zona baja.
- Previamente, en el análisis de los índices de desarrollo, se observó más similitudes entre Rumanía/Marruecos y Colombia/Venezuela (ambos de América del Sur), pero ahora vemos más cercanía entre Venezuela/Marruecos y Rumanía/Colombia, por lo tanto deben

considerarse el régimen político en el período 2008 – 2022. Así, tenemos que Colombia y Rumania tienen democracias electorales, mientras que Marruecos y Venezuela presentan regímenes autocráticos.

- Para Marruecos, ambos índices se mantienen alrededor de 0.25 en el tiempo, mientras que en Venezuela se observa una disminución progresiva a partir del año 2013, punto en el cual también empieza a aumentar la inmigración de venezolanos a España.
- Vemos que para Colombia hay un incremento y estabilización entre 2009 - 2017, seguido por una caída que puede relacionarse con incremento de inmigrantes colombianos. Para Rumania, vemos un incremento en los índices entre 2008 - 2016, período en el cual disminuye la inmigración progresivamente hasta estabilizarse.



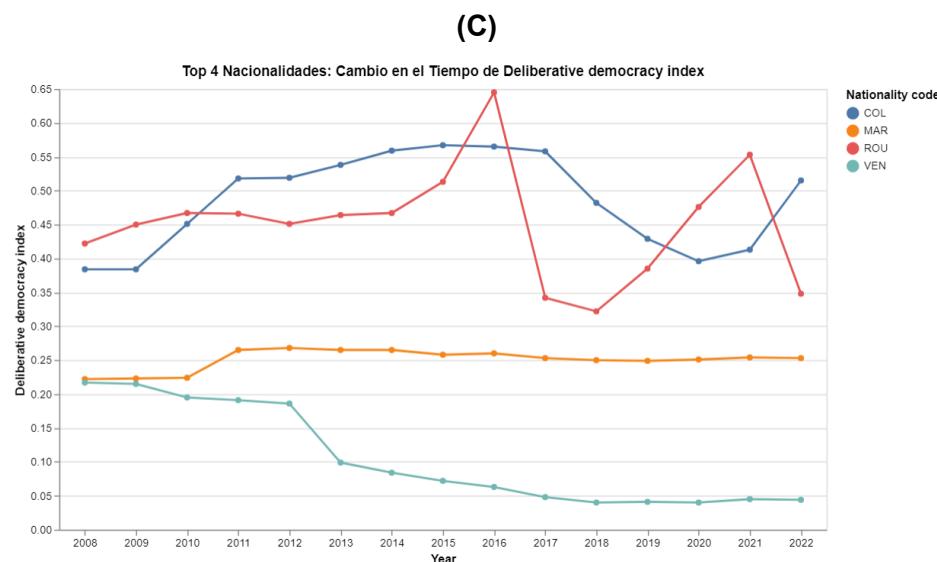


Figura 28. Variación anual de la cantidad de inmigrantes en el Top 4 nacionalidades (A) en contraste con la variación de sus índices de democracia liberal (B) y deliberativa (C).

5.2.6 Índices de Libertad y otros

Nuestro archivo de trabajo se basa en las investigaciones de *International IDEA*, mismas que se vierten en el compendio de índices *Global State of Democracy* (GSOD)¹⁹. En este archivo, venían muchos índices que no nos eran de utilidad todos, por lo que procedimos a realizar un limpiado y estandarización de datos.

Para este *dataframe*, realizamos un *pivot*, eliminamos cadenas de texto innecesarias en la escala de los índices (informativamente, todos los índices traían la leyenda “highest score=1”).

	Economy ISO3	Economy Name	Indicator ID	Indicator	2008	2009	2010	2011	2012	2013	2014	2015	2016
0	AFG	Afghanistan	IDEA.GSOD.abs_corrupt_est	Absence of Corruption (highest score=1)	0.20	0.20	0.16	0.16	0.18	0.21	0.22	0.24	0.24
1	AFG	Afghanistan	IDEA.GSOD.access_just_est	Access to Justice (highest score=1)	0.25	0.25	0.25	0.25	0.24	0.23	0.23	0.23	0.23
2	AFG	Afghanistan	IDEA.GSOD.basic_welf_est	Basic Welfare (highest score=1)	0.31	0.31	0.34	0.34	0.34	0.34	0.37	0.37	0.36

Figura 29. Dataframe antes de preprocesamiento

```
# Inspeccionamos nombre de variables
df2["Indicator"].unique()

array(['Absence of Corruption', 'Access to Justice', 'Basic Welfare',
       'Civic Engagement', 'Civil Liberties', 'Credible Elections',
       'Direct Democracy', 'Effective Parliament',
       'Electoral Participation', 'Elected Government',
       'Freedom of Expression', 'Freedom of Movement',
       'Free Political Parties', 'Freedom of the Press',
       'Freedom of Religion', 'Gender Equality', 'Inclusive Suffrage',
       'Judicial Independence', 'Local Democracy', 'Participation',
       'Personal Integrity and Security', 'Political Equality',
       'Predictable Enforcement', 'Representation', 'Rights',
       'Rule of Law', 'Social Group Equality', 'Direct democracy',
       'EMB autonomy', 'EMB capacity'],
      dtype='object')
```

Figura 30. Muestra de indicadores presentes el estudio, suman 172.

	Country code	Indicator	Year	Value
696	ARG	Absence of Corruption	2008	0.51
2396	BGR	Absence of Corruption	2008	0.53
3264	BRA	Absence of Corruption	2008	0.51
4604	CHN	Absence of Corruption	2008	0.43
5471	COL	Absence of Corruption	2008	0.51

Figura 31. Dataframe final con países selectos y pivot aplicado.

Los índices que usados fueron: "Absence of Corruption", "Public sector corrupt exchanges", "Civil Liberties", "Judicial accountability", "Educational equality" y "Health equality".

En general, los países tienen datos que varían de acuerdo con su composición política y de acceso a servicios, teniendo la mayoría de ellos tiene un rendimiento por indicador cercano al 0.5 (1 es la calificación más alta).



Figura 32. Promedio anual de todos los países en: (A) Ausencia de corrupción, (B) Frecuencia de sobornos, (C) Grado de respeto a las libertades, (D) Responsabilidad judicial, (E) Igualdad educacional y (F) Igualdad en salubridad.

Los países en general más desarrollados como Gran Bretaña logran calificaciones que rondan los 0.8 puntos, mientras que otros como Marruecos promedian el 0.4 de eficiencia. ¿Puede ser esto indicio de que los países más desarrollados emigren a España por cuestiones de placer o negocios mientras que los países con menor desarrollo lo hagan por cuestiones humanitarias o por persecución?

5.4.7 Tasa de Homicidios

Al igual que en las variables anteriores, los datos de tasa de homicidio por cada 100,000 habitantes por país y año^{20, 21} se limpiaron y filtraron para el top nacionalidades y el período 2008 – 2022. Cabe destacar que los datos de Reino Unido estaban informados en tres regiones (Irlanda del Norte, Escocia e Inglaterra-Gales), por lo que el valor final para Reino Unido fue la media de las tres por año.

Posteriormente, se observó mediante un histograma que alrededor del 80% de los datos es inferior a 10 homicidios por 100.000 habitantes (Figura 33), y en el 20% restante hay una gran variabilidad. Además, los datos atípicos pertenecen a algunos países de América del Sur, especialmente Venezuela y Honduras.

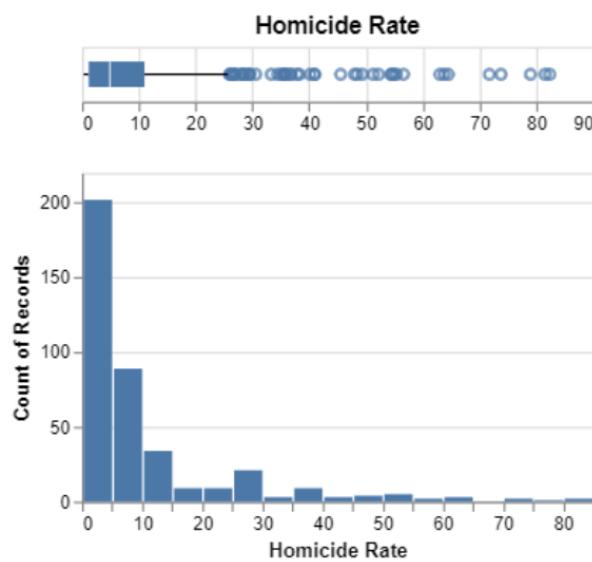


Figura 33. Distribución de datos de Tasa de Homicidios y detección de datos atípicos.

Al analizar la distribución por año, se observó que la distribución de la tasa de homicidios es muy similar entre los años 2008 – 2011 (Figura 34) y que partir del 2012 tiende a bajar hasta alcanzar su menor rango en el 2020, año del inicio de la pandemia del COVID-19. Asimismo, se identifica con mayor facilidad que Venezuela, Colombia, y Honduras permanecen con valores tasa de homicidios atípicamente elevados en todo el período (no apreciable en la figura), añadiéndose Brasil a partir del 2012.

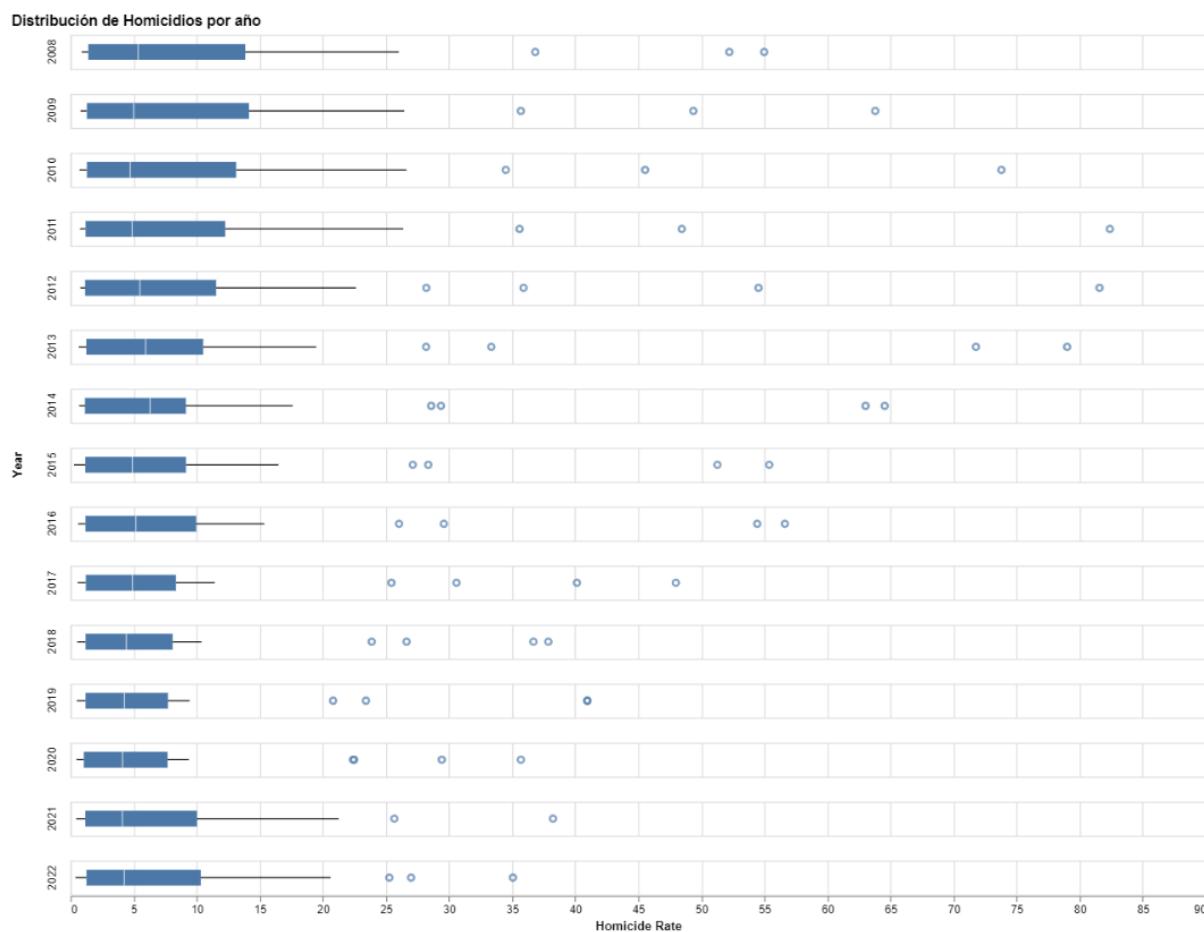


Figura 34. Distribución de Tasa de Homicidios por año desde el 2008 a 2022 para el top de nacionalidades en inmigración en España.

Finalmente, después de comparar las tasas de homicidios del Top 4 en inmigración (Figura 35A) en contraste con la cantidad de inmigrantes (Figuras 35B) se observó que:

- No parece haber una relación clara en cuanto a tendencia entre la inmigración y homicidios para las Top 4 nacionalidades; pero sí notamos que Colombia y Venezuela, ambos de América del Sur, tienen tasas de homicidios elevadas durante amplios períodos de tiempo, especialmente entre 2008 - 2013.
- Marruecos y Rumanía, ambos países mucho más próximos geográficamente a España (siendo Rumanía parte de la Unión Europea), tienen una tasa de homicidios baja y muy similar entre ellos.

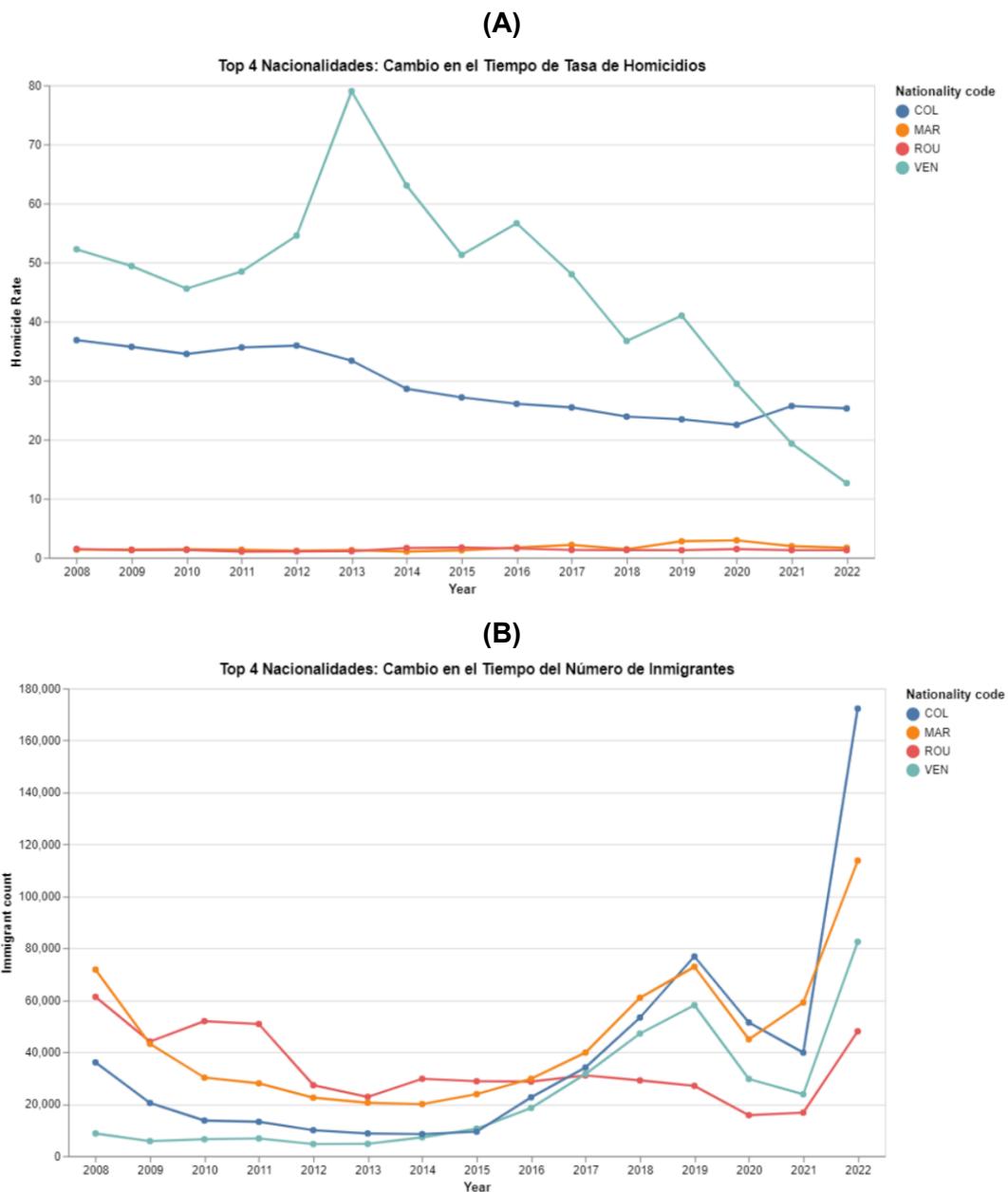


Figura 35. Contraste entre la tendencia de cambio en el tiempo de la Tasa de Homicidios (A) frente a los datos de inmigración en España del Top 4 nacionalidades (B).

A partir del análisis, puede inferirse que la tasa de homicidios, en vez de tener un efecto inmediato, presenta un efecto acumulativo en el tiempo, en cuyo caso, será importante que los modelos de *machine learning* logren identificar relaciones entre variables, además de considerar las condiciones particulares de cada país bajo este complejo esquema de enfoque macro.

5.2.8 Conflictos Armados

Continuando con las variables relacionadas con la inseguridad, se consideró también incluir y analizar las muertes debidas a cuatro tipos de conflictos armados²² (referirse al Anexo 1 para las definiciones correspondientes).

Después de filtrar los datos de nuestro interés, se observó que predomina la baja cantidad o ausencia de muertes dentro del top países en inmigración a España (Figura 36), lo que explica porque los gráficos de caja no son visibles y considera como datos atípicos los registros de muertes para Ucrania, Pakistán, Colombia, Brasil y Venezuela (no apreciable en la figura), que son los países que se repiten dentro de los valores atípicos.

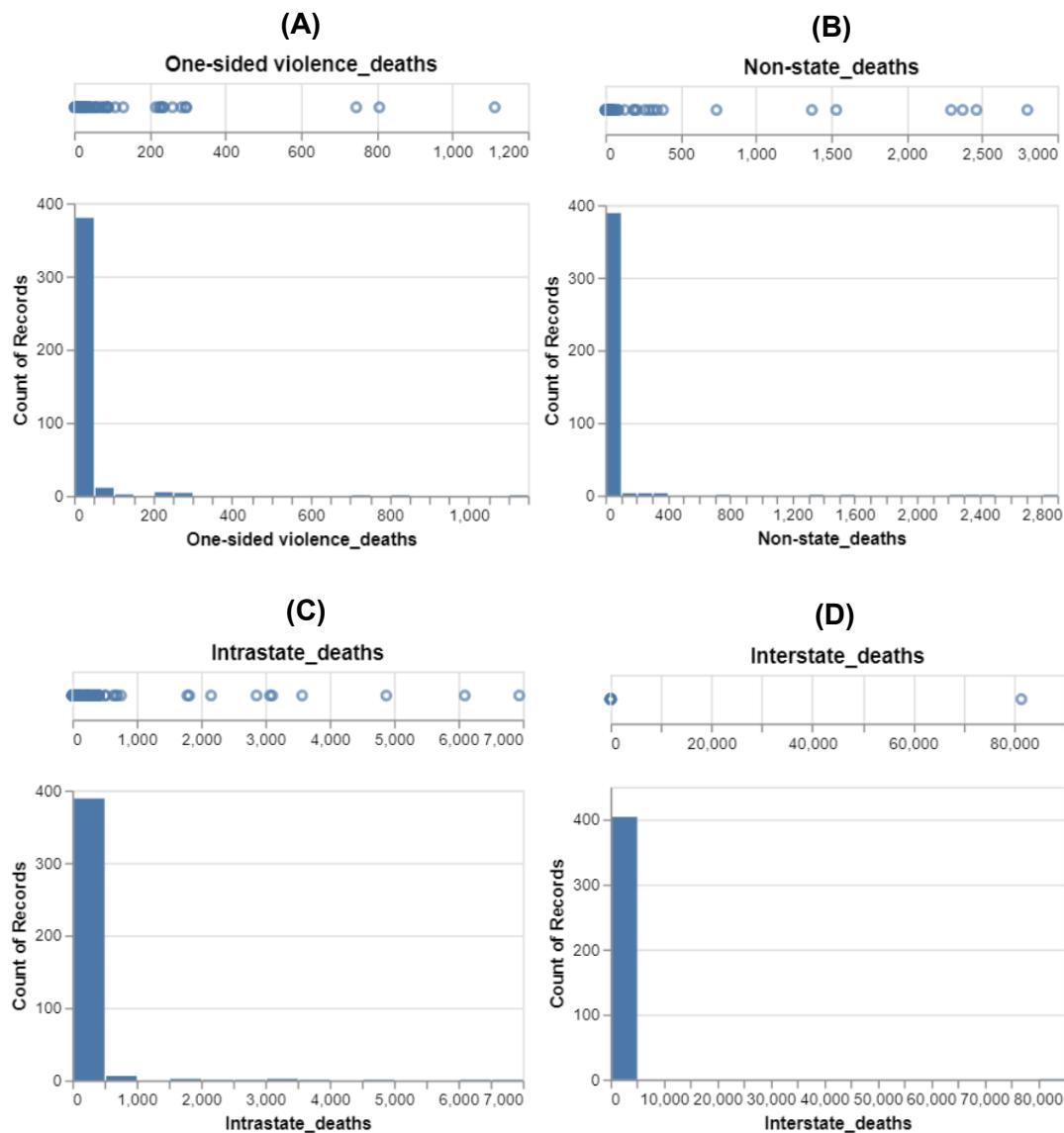


Figura 36. Distribución las variables de conflictos armados y detección de datos atípicos. (A) "One-sided violence_deaths", (B) "Non-state_deaths", (C) "Intrastate_deaths" y (D) "Interstate_deaths".

Debido a la dificultad para observar tendencias totales por la predominancia de algunos países, se filtraron los países más relevantes (Brasil, Colombia, Venezuela y Ucrania) para, nuevamente, contrastar la variación de los datos de inmigración con las muertes por estos conflictos (Figura 37), observando que:

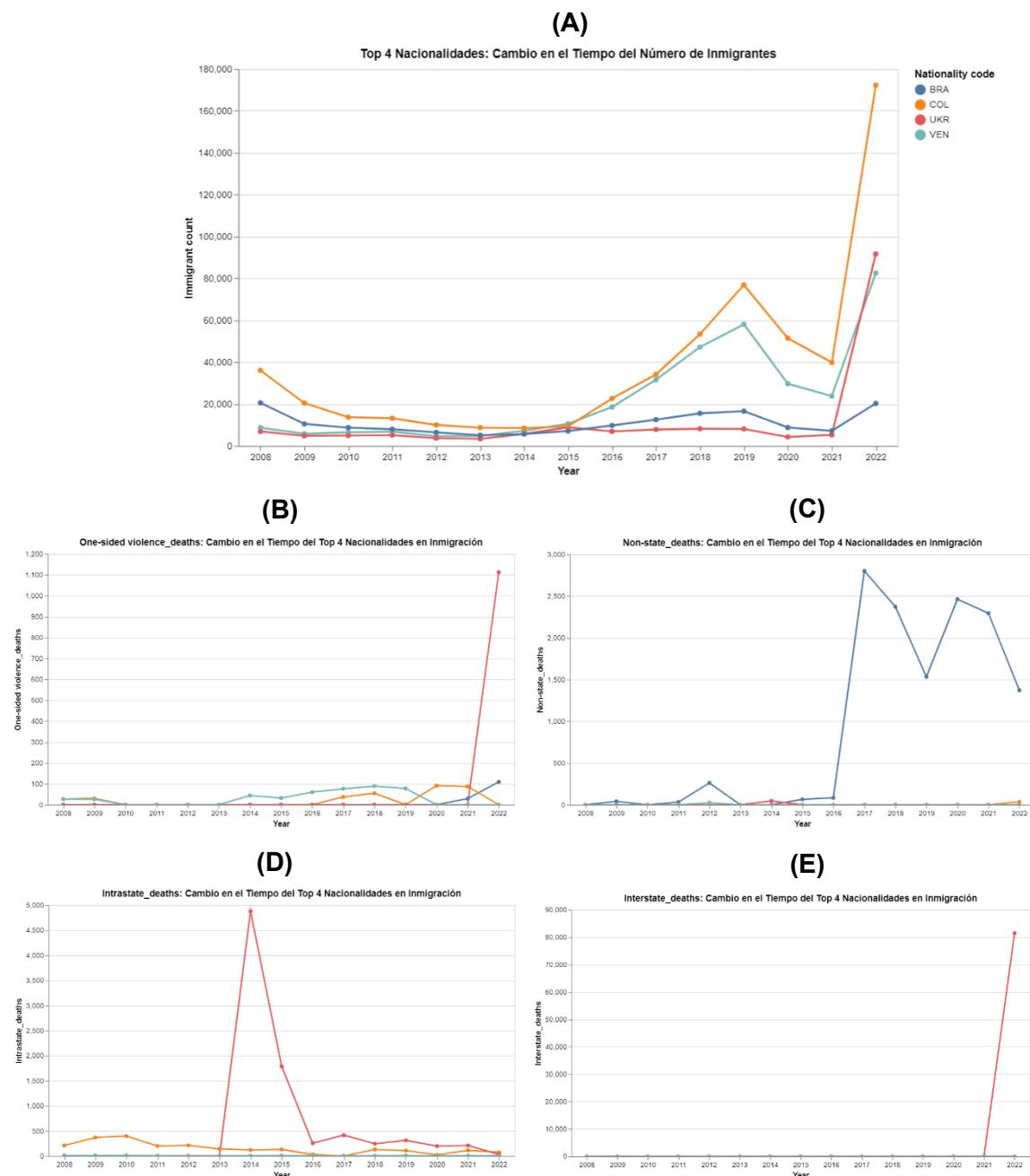


Figura 37. Variación anual de la cantidad de inmigrantes en el Top 4 nacionalidades (A) en contraste con la variación de las muertes por conflictos armados: (B) "One-sided violence_deaths", (C) "Non-state_deaths", (D) "Intrastate_deaths" y (E) "Interstate_deaths".

- Para Ucrania, hay una correlación en el pico de muertes por "One-sided violence_deaths" (Figura 37B) y "Interstate_deaths" (Figura 37E) en 2022, debido al conflicto con Rusia, y el incremento de inmigrantes.
- También vemos que, entre 2014 y 2019, los valores de "One-sided violence_deaths" tienen un incremento para Colombia y Venezuela, al igual que hay un aumento progresivo de inmigración en ese mismo intervalo. Además, se aprecia algo similar para Brasil con relación a "Non-state_deaths" (Figura 37C) e inmigración a partir del 2016 (Figura 37A).

Lo observado con estas cuatro nacionalidades, parece soportar la hipótesis que mientras mayores son las consecuencias de estos conflictos, mayor es su efecto en los desplazamientos migratorios, ya sea por razones humanitarias o nacionales buscando mejores condiciones de seguridad.

5.2.9 Turistas Anuales

En comparación con los *dataframes* anteriores, el conjunto de datos de turistas tiene un volumen menor de datos. Estos datos representan un volumen de personas que llegan a España por medio de aeropuertos²³.

La finalidad de incluir estos datos es para introducir una variable explicativa que pudiera darnos un grado de magnitud a la hora de modelar.

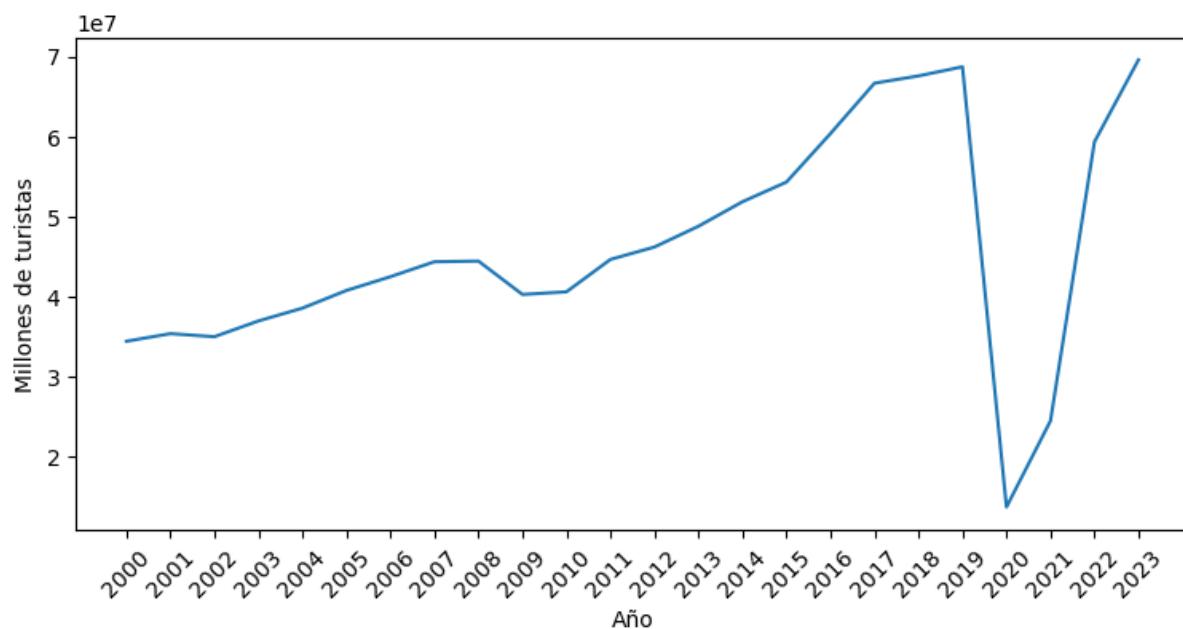


Figura 38. Total de turistas que llegaron a España en avión por año.

Como otros indicadores, la presencia de turistas se ve disminuida en el año de 2020, misma que se recupera a partir de 2022.

5.3 Etapa 3: Unión y selección de variables

Las variables predictivas (Etapa 2) se añadieron al conjunto de datos de inmigrantes (Etapa 1) con un *merge* (Figura 40), teniendo únicamente datos nulos para Senegal en la variable “Homicide Rate” ya que solo se encontraron datos para el año 2015.

Adicionalmente, se agregaron tres variables categóricas: i) una variable de idioma castellano para identificar los países que tienen al castellano como idioma oficial ('Spanish language' = 1), ii) otra para etiquetar a los años 2020 y 2021 por ser los años donde hubo fuertes restricciones a nivel global debido a la pandemia del COVID 19 ("Restricciones_pandemia" = 1) y iii) otra para etiquetar el año 2022 como el primer año con flexibilización de restricciones ("Año post_pandemia" = 1).

```
# Renombrar columna de código de país como el df "inmigrantes"
residentes.rename({'Country code' : 'Nationality code'},
| | | | | inplace = True,
| | | | | axis = 1)

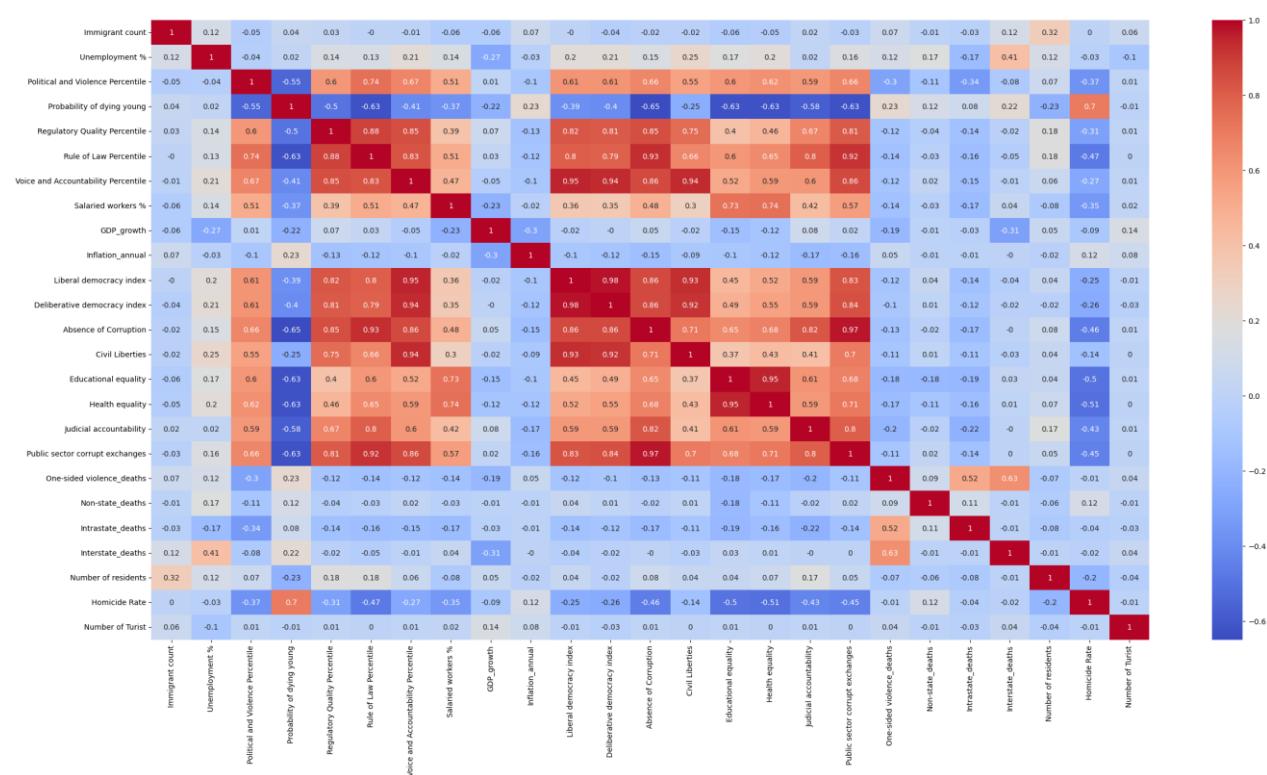
libertad.rename({'Country code' : 'Nationality code'},
| | | | | inplace = True,
| | | | | axis = 1)

# Merge cada df en "inmigrantes"
inmigrantes_merge = inmigrantes.merge(indices_desarrollo, on = ['Year', 'Nationality code'], how = 'left')\
| | | | | .merge(indices_democracia, on = ['Year', 'Nationality code'], how = 'left')\
| | | | | .merge(regiones, on = ['Nationality code'], how = 'left')\
| | | | | .merge(libertad, on = ['Year', 'Nationality code'], how = 'left')\
| | | | | .merge(conflictos_armados, on = ['Year', 'Nationality code'], how = 'left')\
| | | | | .merge(residentes, on = ['Year', 'Nationality code', 'Sex'], how = 'left')\
| | | | | .merge(regimen_politico, on = ['Year', 'Nationality code'], how = 'left')\
| | | | | .merge(homicidios, on = ['Year', 'Nationality code'], how = 'left')\
| | | | | .merge(turismo, on = ['Year'], how = 'left')\

inmigrantes_merge.info()
inmigrantes_merge.iloc[:, :15]
```

Figura 39. Código python para unir las variables predictoras al dataset “inmigrantes” con los datos anuales de inmigración de España.

A partir de estos datos, se observó altas correlaciones entre las variables de desarrollo asociados a estado de derecho “Regulatory Quality Percentil”, “Rule of Law Percentil” y “Voice and Accountability Percentil”), los índices de democracia y las variables del conjunto de datos relacionados a libertad (Figura 41, observar la zona cálida en la matriz de correlación). En particular, si consideramos el valor 0.85 como valor de referencia para alta correlación de Pearson, vemos:

**Figura 40.** Matriz de correlación de Pearson entre las variables cuantitativas.

- Alta correlación entre los índices de democracia, "Absence of Corruption", "Voice and Accountability Percentile" y "Civil Liberties". En consecuencia, se removió: "Deliberative democracy index", "Absence of Corruption", "Voice and Accountability Percentile" y "Civil Liberties".
- Alta correlación entre "Health equality" y "Educational equality"; se removió "Educational equality".
- Alta correlación entre "Regulatory Quality Percentile", "Public sector corrupt exchanges" y "Rule of Law", relacionadas con el sector público y la ley. Se removió "Regulatory Quality Percentile" y "Public sector corrupt exchanges".
- No hay correlación lineal con la variable de número de inmigrantes.

Continuando con las variables categóricas, se evaluó su significancia mediante un contraste de hipótesis usando pruebas no paramétricas debido a que ninguna de ellas seguía una distribución normal (Figura 42-43). En consecuencia, se aplicó la prueba de U de Mann-Whitney para dos grupos y Kruskal-Wallis para aquellas variables con múltiples grupos.

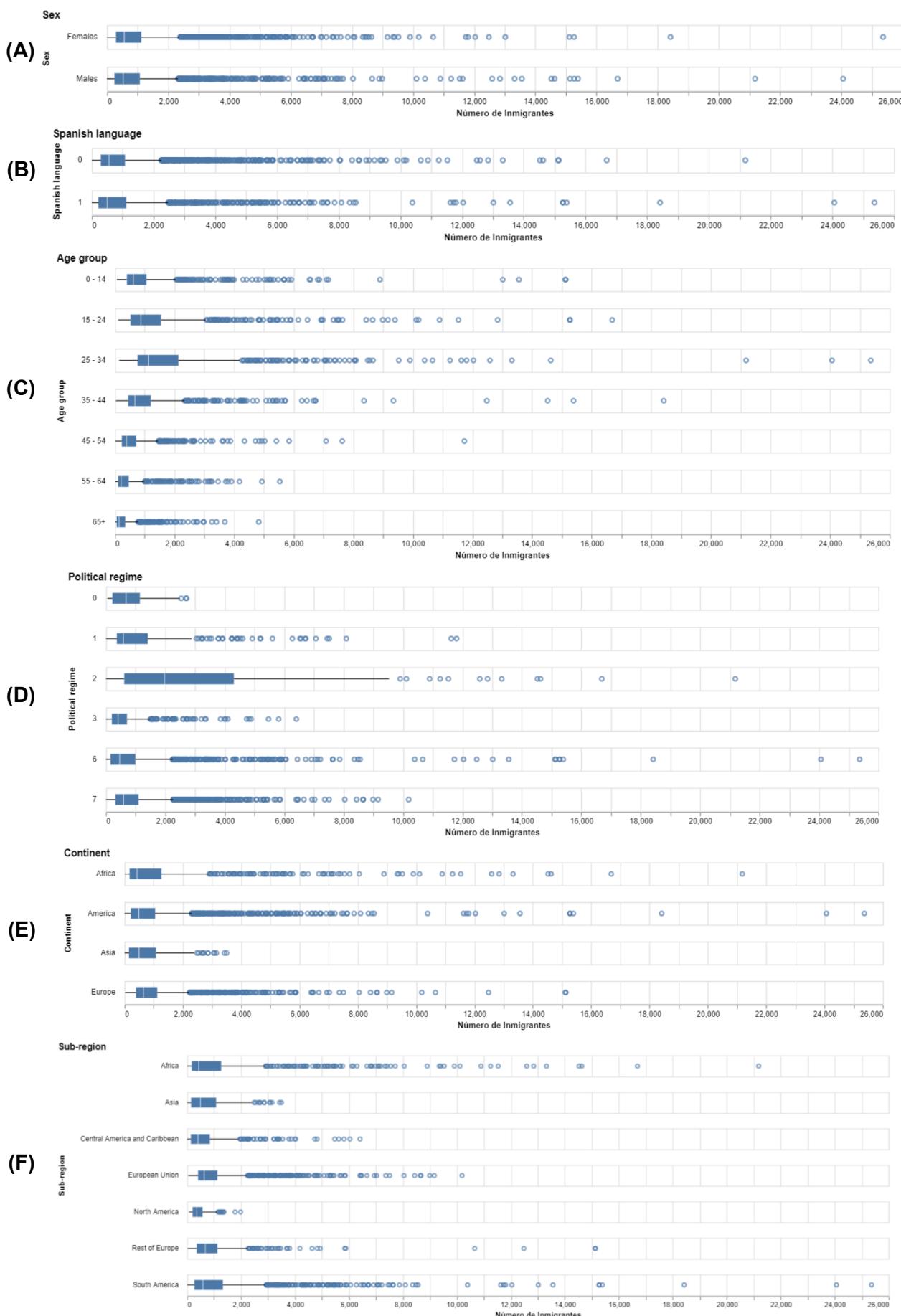


Figura 41. Distribución del número de inmigrantes mediante boxplots y detección de datos atípicos para los distintos grupos de cada variable categórica: (A) Sexo, (B) Idioma español, (C) Grupo de edad, (D) Régimen político, (E) Continente y (F) Subregión.

```

----- Sex -----
Group Females: Statistics=0.5497, p=0.0
Group Males: Statistics=0.5113, p=0.0

----- Age group -----
Group 0 - 14: Statistics=0.51, p=0.0
Group 15 - 24: Statistics=0.5864, p=0.0
Group 25 - 34: Statistics=0.5782, p=0.0
Group 35 - 44: Statistics=0.5288, p=0.0
Group 45 - 54: Statistics=0.5587, p=0.0
Group 55 - 64: Statistics=0.5923, p=0.0
Group 65+: Statistics=0.5947, p=0.0

----- Political regime -----
Group 0: Statistics=0.9071, p=0.0
Group 1: Statistics=0.6338, p=0.0
Group 2: Statistics=0.7825, p=0.0
Group 3: Statistics=0.6483, p=0.0
Group 6: Statistics=0.4623, p=0.0
Group 7: Statistics=0.6851, p=0.0

----- Continent -----
Group Africa: Statistics=0.5627, p=0.0
Group America: Statistics=0.5001, p=0.0
Group Asia: Statistics=0.8624, p=0.0
...
----- Spanish language -----
Group 0: Statistics=0.5476, p=0.0
Group 1: Statistics=0.5098, p=0.0

```

Figura 42. Resultados de la prueba de Shapiro-Wilk para normalidad. H0: Datos se ajustan a una distribución normal, H1: No se ajustan a una normal. Intervalo de confianza = 0.95.

Los resultados de las pruebas no paramétricas indican que hay diferencias significativas en las variables con dos grupos (seco e idioma castellano) y, al menos, diferencias significativas entre dos grupos dentro de las variables con múltiples (Figura 44).

```

----- Variable "Sex" -----
Prueba U de Mann-Whitney: Statistics=4641572.5, p=0.0015

----- Variable "Spanish language" -----
Prueba U de Mann-Whitney: Statistics=10816393.0, p=0.0004

----- Variable "Age group" -----
Prueba Kruskal-Wallis: Statistics=2726.2587, p=0.0

----- Variable "Political regime" -----
Prueba Kruskal-Wallis: Statistics=365.6376, p=0.0

----- Variable "Continent" -----
Prueba Kruskal-Wallis: Statistics=116.0261, p=0.0

----- Variable "Sub-region" -----
Prueba Kruskal-Wallis: Statistics=260.0816, p=0.0

```

Figura 43. Resultados de las pruebas de U de Mann-Whitney y Kruskal-Wallis. H0: No hay diferencia (en términos de tendencia central) entre los grupos, H1: Existe una diferencia (con respecto a la tendencia central) entre los grupos. Intervalo de confianza = 0.95.

En el caso de “Sex”, puede deberse a que se observa una mayor dispersión en los datos de cantidad de inmigrantes hombres, similar a lo que se observa en el grupo de idioma castellano.

Adicionalmente, para las variables con más de dos grupos, se realizó la prueba de Dunn de comparación múltiple no paramétrica para evaluar las diferencias específicas entre grupos. Las observaciones que resaltan son:

- Hay diferencias significativas entre todos los grupos de edades, a excepción de los grupos de 0-14 y 35-44 años (Tabla 2), los cuales, además, son los que mostraban una distribución muy similar en el gráfico de cajas (Figura 42B).
- En cuanto a los tipos de régimen político, resalta el régimen tipo 2 (autocracias multipartidistas sin ejecutivo electo) es significativamente diferente al resto (Tabla 3), resultado que se esperaba con base a lo observado en la Figura 42D. De forma similar, el régimen tipo 3 (autocracias multipartidistas) también presentó diferencias con los demás tipos de regímenes, exceptuando el tipo 6 (democracias electorales).
- El continente europeo fue el único que presentó diferencias significativas con el resto de los continentes (Tabla 4).
- En contraste con los continentes, las subregiones presentan una mayor cantidad de diferencias significativas (Tabla 5). Vemos a América del Norte que se diferencia de todos menos con el Caribe-Centro América. De forma similar, la UE y América del Sur también difieren con todos, exceptuando a “el resto de Europa”.

Tabla 2. P-valor de las comparaciones pareadas entre grupos de edades mediante la prueba de Dunn. H0: No hay diferencias entre los dos grupos, H1: Existe una diferencia entre dos grupos. Intervalo de confianza = 0.95

Grupo 1	Grupo 2	p-valor	¿Diferencias significativas?
0 - 14	15 - 24	1.67E-08	Si
0 - 14	25 - 34	1.31E-29	Si
0 - 14	35 - 44	1.00	No
0 - 14	45 - 54	2.60E-20	Si
0 - 14	55 - 64	1.58E-73	Si
0 - 14	65+	3.22E-110	Si
15 - 24	25 - 34	1.26E-06	Si
15 - 24	35 - 44	3.76E-05	Si
15 - 24	45 - 54	3.13E-54	Si
15 - 24	55 - 64	1.02E-130	Si
15 - 24	65+	1.97E-178	Si
25 - 34	35 - 44	4.42E-23	Si
25 - 34	45 - 54	1.11E-97	Si
25 - 34	55 - 64	1.02E-194	Si
25 - 34	65+	3.72E-252	Si
35 - 44	45 - 54	1.84E-26	Si
35 - 44	55 - 64	7.31E-85	Si
35 - 44	65+	5.20E-124	Si
45 - 54	55 - 64	4.46E-17	Si
45 - 54	65+	1.11E-36	Si
55 - 64	65+	0.000740	Si

Tabla 3. P-valor de las comparaciones pareadas entre grupos de tipo de régimen político mediante la prueba de Dunn. H0: No hay diferencias entre los dos grupos, H1: Existe una diferencia entre dos grupos. Intervalo de confianza = 0.95

Grupo 1	Grupo 2	p-valor	¿Diferencias significativas?
0	1	5.80E-01	No
0	2	0.00	Si
0	3	4.25E-04	Si
0	6	7.59E-02	No
0	7	1.00	No
1	2	6.46E-13	Si
1	3	3.57E-13	Si
1	6	0.00	Si
1	7	3.72E-01	No
2	3	1.94E-51	Si
2	6	0.00	Si
2	7	1.31E-28	Si
3	6	1.23E-01	No
3	7	0.00	Si
6	7	9.47E-14	Si

Tabla 4. P-valor de las comparaciones pareadas entre grupos de continentes mediante la prueba de Dunn. H0: No hay diferencias entre los dos grupos, H1: Existe una diferencia entre dos grupos. Intervalo de confianza = 0.95

Grupo 1	Grupo 2	p-valor	¿Diferencias significativas?
Africa	America	1.00E+00	No
Africa	Asia	1.00	No
Africa	Europe	1.62E-12	Si
America	Asia	1.00E+00	No
America	Europe	6.57E-21	Si
Asia	Europe	1.56E-10	Si

Tabla 5. P-valor de las comparaciones pareadas entre grupos de subregiones mediante la prueba de Dunn. H0: No hay diferencias entre los dos grupos, H1: Existe una diferencia entre dos grupos. Intervalo de confianza = 0.95

Grupo 1	Grupo 2	p-valor	¿Diferencias significativas?
Africa	Asia	1.00	No
Africa	Central America and Caribbean	0.01	Si
Africa	European Union	5.51E-12	Si
Africa	North America	4.34E-03	Si
Africa	Rest of Europe	0.00	Si
Africa	South America	4.32E-05	Si
Asia	Central America and Caribbean	1.34E-01	No
Asia	European Union	0.00	Si
Asia	North America	2.68E-02	Si
Asia	Rest of Europe	9.00E-06	Si
Asia	South America	0.00	Si

Grupo 1	Grupo 2	p-valor	¿Diferencias significativas?
Central America and Caribbean	European Union	6.84E-33	Si
Central America and Caribbean	North America	1.00E+00	No
Central America and Caribbean	Rest of Europe	0.00	Si
Central America and Caribbean	South America	3.06E-20	Si
European Union	North America	4.32E-17	Si
European Union	Rest of Europe	1.00	No
European Union	South America	1.39E-02	Si
North America	Rest of Europe	2.58E-12	Si
North America	South America	0.00	Si
Rest of Europe	South America	1.00	No

A razón estas observaciones, no se removieron ninguna de las variables categóricas del conjunto de datos final para la construcción del modelo de *machine learning* en la Etapa 4.

5.4 Etapa 4: Prueba y comparación de modelos de *machine learning*

Las pruebas con modelos de *machine learning* se hicieron en dos conjuntos de datos, uno con datos de los agregados “Both” y “All” para sexos y grupos de edades, y otro sin los mismo (para mayores detalles, referirse a los *jupyter notebooks* “Etapa 4 – Agregados” y “Etapa 4 – Sin Agregados”). Con este enfoque se buscó evaluar si el error de estimar los agregados es menor al incluirlos en el modelo a que si se obtienen por sumatoria de los grupos.

Además del uso de modelos típicos para *machine learning*, se incluyeron modelos lineales que suelen usarse en presencia de datos atípicos: Huber, Theilsen y RANSAC.

Entre los resultados más resaltantes, tenemos:

- Los modelos de *Hist Gradient Boosting* (HGB), red neuronal de una capa y XGBoost mostraron los mejores resultados en métricas tanto para el modelo con las categorías de agregados como sin ellos (Tabla 6-7). También se observa que, como era de esperarse, las métricas de RSME (Raíz del Error Cuadrático Medio), MAE (Error Medio Absoluto) y MAPE (Error Porcentual Medio Absoluto) son mayores en el conjunto de datos con los agregados de sexo y grupo de edad debido a que, al tener estos valores agregados, la dispersión y magnitud de los datos aumenta.

Tabla 6. Métricas en *train/test* de modelos con mejor rendimiento para el conjunto de datos con agregados de sexo y grupos de edad.

Métrica/conjunto	HGB	Red Neuronal	XGBoost	Random Forest	Decision Tree	KNN
R ² train	0.988	0.976	0.983	0.801	0.838	0.63
R ² test	0.942	0.94	0.884	0.744	0.712	0.633
RSME train	604	866	707	2474	2228	3369
RSME test	1319	1343	1871	2775	2947	3325
MAE train	294	522	302	752	903	1019
MAE test	432	629	572	910	1125	1136
MAPE train	0.41	0.78	0.77	0.85	1.84	1.21
MAPE test	0.5	0.62	0.99	0.81	1.46	1.52

Tabla 7. Métricas en *train/test* de modelos con mejor rendimiento para el conjunto de datos sin agregados de sexo y grupos de edad.

Métrica/conjunto	HGB	Red Neuronal	XGBoost	Random Forest	Decision Tree	KNN
R ² train	0.981	0.968	0.978	0.793	0.69	0.721
R ² test	0.957	0.911	0.91	0.765	0.551	0.697
RSME train	213	290	232	742	908	861
RSME test	321	413	465	672	927	762
MAE train	110	188	119	337	386	376
MAE test	152	242	226	356	437	381
MAPE train	0.23	0.57	0.6	0.74	0.59	0.87
MAPE test	0.29	0.57	0.98	0.79	0.68	0.96

- Las peores métricas se obtuvieron con modelos lineales (clásico, regularizados y orientados a datos atípicos) y de soporte vectorial (SVR) (Tabla 8-9).

Tabla 8. Métricas en *train/test* de modelos con peor rendimiento para el conjunto de datos con agregados de sexo y grupos de edad.

Métrica/conjunto	Lineal	Lineal - Ridge	Lineal - Lasso	Lineal - Theilsen	Lineal - Huber	Lineal - E-Net	Lineal - RANSAC	SVR
R ² train	0.445	-	-	0.372	0.249	-	0.033	0.026
R ² test	0.507	0.507	0.506	0.473	0.277	0.103	0.035	0.022
RSME train	4128	-	-	4389	4800	-	5449	5469
RSME test	3853	3855	3859	3985	4668	5199	5394	5430
MAE train	1773	-	-	1575	1306	-	1653	1659
MAE test	1818	1800	1797	1557	1372	2293	1745	1758
MAPE train	6.5	-	-	4.76	1.39	-	1.6	1.04
MAPE test	4.65	4.56	4.52	3.4	0.94	6.1	0.87	1.01

Tabla 9. Métricas en *train/test* de modelos con peor rendimiento para el conjunto de datos sin agregados de sexo y grupos de edad.

Métrica/conjunto	Lineal	Lineal - Ridge	Lineal - Lasso	Lineal - Theilsen	Lineal - Huber	Lineal - E-Net	Lineal - RANSAC	SVR
R ² train	0.52	-	-	0.242	0.293	-	-16.904	0.055
R ² test	0.513	0.514	0.516	0.251	0.34	0.339	-4.94	0.074
RSME train	1129	-	-	1419	1370		6897	1585
RSME test	966	965	963	1198	1125	1125	3373	1332
MAE train	621			640	526		992	620
MAE test	614	611	608	611	493	625	662	579
MAPE train	3.3			2.82	0.97		0.87	0.89
MAPE test	2.57	2.56	2.53	2.06	0.83	2.37	0.8	0.89

- La importancia media relativa de variables del modelo HGB (por permutaciones), mostró que el número de residentes fue la variable más importante (Tabla 10), seguida del año y el grupo de edad de 25-34 años (grupo predominante en número de inmigrantes). También tenemos variable relacionadas a empleo, número de turistas y otros grupos de edades.

Tabla 10. Top 10 variables en importancia media relativa por permutaciones del modelo HGB en el conjunto sin agregados.

Variable	Media de Importancia	Desviación estand. de Importancia
Number of residents	0.823	0.029
Year	0.202	0.007
Age group_25 - 34	0.186	0.014
Unemployment %	0.185	0.020
Age group_65+	0.177	0.033
Age group_55 - 64	0.138	0.026
Number of Turist	0.098	0.009
Salaried workers %	0.094	0.010
Age group_15 - 24	0.084	0.006
Age group_45 - 54	0.046	0.008

En primer lugar, debido a las tendencias observadas en nuestro análisis inicial, se esperaba observar relevancia con relación a los años y grupos de edades. En segundo lugar, el grupo de edad fue la variable de segregación de los datos de inmigración que mostró la mayor diferencia entre grupos, habiéndose observado la predominancia de los grupos de 15-24 y 25-34 años, y siendo +65 años el grupo más pequeño.

Las variables de empleo nos señalan la relevancia de las condiciones socioeconómicas, pues en economías débiles el sector comercial y empresarial se ve muy impactado, viéndose forzados a cerrar o disminuir su plantilla de empleados, u ofrecer menor cantidad de plazas de trabajo, así mismo, los ciudadanos también tienden a buscar fuentes alternativas de ingresos en la economía informal. Por otro lado, habíamos visto que la tendencia del número de turista se asemeja a la de los inmigrantes, y es notoria que muchos inmigrantes ingresan inicialmente como turista al país, pero sin intención de regresar a sus países de origen.

En cuanto a los números de residentes, este resultado nos dice que el fenómeno de **migración en cadena** está jugando un papel muy importante, especialmente en intervalos donde no hay migración irregular (e.g., debida a causas excepcionales, como conflictos bélicos o pandemia). Este fenómeno ocurre cuando inmigrantes de un país o región siguen a otros que ya han emigrado al mismo destino: debido a que la presencia de una comunidad establecida de la misma actúa como un factor de atracción para más personas del mismo grupo, debido a lazos familiares, culturales o económicos²⁴. Por décadas, uno de los casos más estudiados en España ha sido la inmigración marroquí²⁵, y rumana desde su ingreso a la UE en 2007²⁶: ambos dentro de nuestro top de nacionalidades; Colombia y Venezuela son un caso similar. La inmigración colombiana comenzó a tener su auge a partir del 2000 en la búsqueda de oportunidades y escapar de la violencia y conflictos armados²⁷, mientras que Venezuela es el caso más reciente inmigración en cadena originado por las condiciones sociales, políticas y económicas del país²⁸.

- Comparado con los resultados de las métricas de RSME, R² y MAE, los resultados de métrica de MAPE mostraron los mayores valores error en todos los modelos. Esto ocurre porque el MAPE supone que los errores porcentuales son igualmente importantes en todos los puntos de datos y puede verse afectado de manera desproporcionada por valores atípicos debido a su relación porcentual.
- Considerando el punto anterior, y el valor de la media de ambos conjuntos de datos fueron 2383 (agregados) y 1021 (sin agregados) inmigrantes, el RSME y el MAE parecen ser las métricas más adecuadas para estimar el error de nuestros modelos.

Adicionalmente, tomando en cuenta que uno de los mejores modelos fue el de redes neuronales y que los datos, tanto *inputs* como *target*, presentan *outliers* y no tienen una distribución normal, se realizaron transformaciones de la variable *target* “Immigrant count” para asemejarla a una distribución Gaussiana (normal), y también aplicando dicha transformación tanto en variables *inputs* como en *target*.

Los resultados de métricas mostraron una gran mejoría de los modelos de redes neuronales de una capa en ambos conjuntos (Tabla 11-12), de manera que se eligieron los modelos de red neuronal con transformación de variables *inputs* y *targets* en ambos conjuntos (nota adicional): en relación con el conjunto de datos sin agregados, se eligió el modelo de red neuronal frente al modelo HGB debido a que en gráficas se observó que sus predicciones eran más próximas al valor real en intervalos regulares de inmigración (2008-2019), mientras que el modelo HGB respondió mejor en los año de irregularidades, como 2020 – 2022; gráfico disponible en [este link](#)). Es importante aclarar que se esperaba observar mayores valores de error para el modelo con los agregados de sexo/grupos de edad debido a que, al incluirlos, la dispersión de los datos de inmigración es mayor, sin embargo, fue menor de lo que se pensaba.

Tabla 11. Métricas en *test* de modelos de red neuronal de una capa después de aplicar normalización para el conjunto de datos con agregados de sexo y grupos de edad.

Métricas	RN + <i>target</i> normalizado	RN + <i>inputs/target</i> normalizados
RMSE test	1023	853
MAE test	359	293

Tabla 12. Métricas en *test* de modelos de red neuronal de una capa después de aplicar normalización para el conjunto de datos sin agregados de sexo y grupos de edad.

Métricas	RN + <i>target</i> normalizado	RN + <i>inputs/target</i> normalizados
RMSE test	343	347
MAE test	162	157

Así, con esto modelos se realizó la predicción del conjunto entero de datos, incluyendo intervalos de confianza del 90% mediante *conformance prediction*, para elaborar un *dashboard* comparativo de ambos (Figura 45, disponible también en [este link](#)).

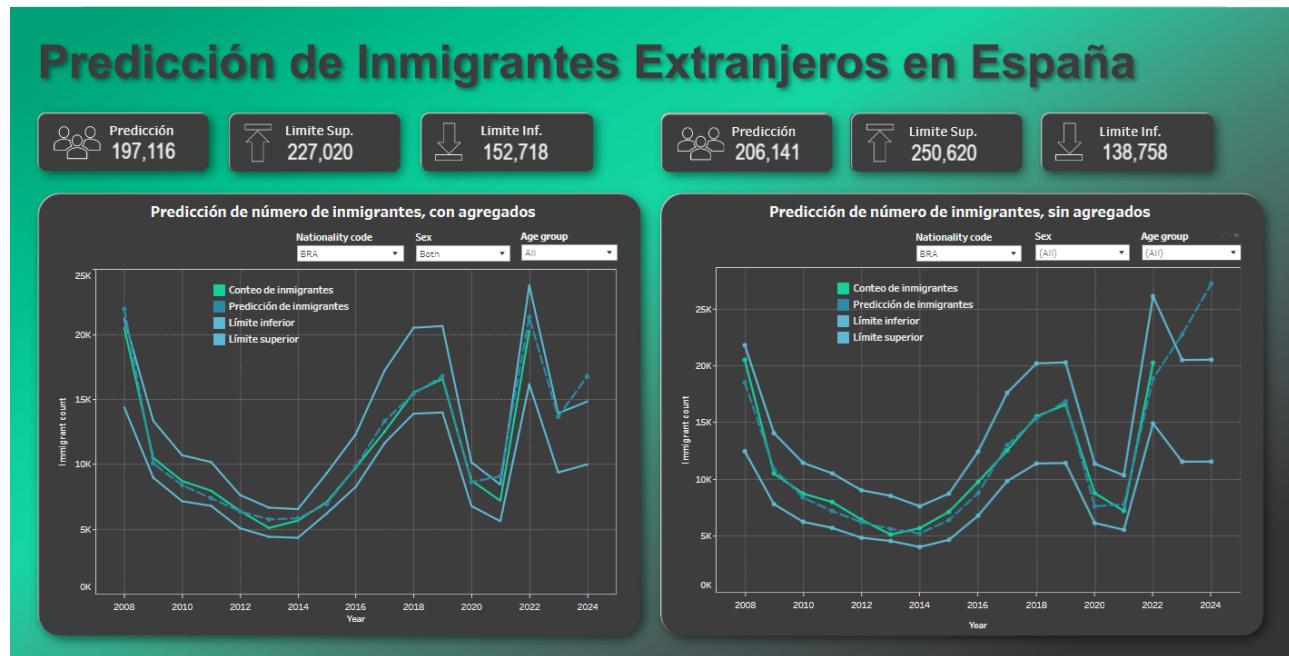


Figura 44. Dashboard de predicciones mediante los mejores modelos de red neuronal de una capa para los conjuntos de datos con (izquierda) y sin agregados (derecha) de sexo y grupos de edad. Intervalos de confianza: 90%. También disponible en [este link](#).

A través de la comparación en *dashboard* de las predicciones a partir de ambos modelos (con agregados y sin agregado), se notó que:

- El modelo sin agregados tiende a tener un menor error de predicción en cambios abruptos de número de inmigrantes en comparación con el modelo con agregados, como en los años 2021 y 2022 durante el efecto pandemia/postpandemia. Sin embargo, el modelo con agregados balancea este punto al presentar intervalos de confianza mucho más acotados.
- El modelo con agregados, además de presentar intervalos más acotado, tiende a predecir con menor error en los casos donde se aplica algunas de las variables de agregación (e.g., ambos sexos o todos los grupos de edades). Esto se debe a que, al haber entrenado con los valores de dichos agregados, logra establecer una mejor distinción, mientras que la predicción del modelo sin agregados es la suma de las predicciones del agregado en cuestión.

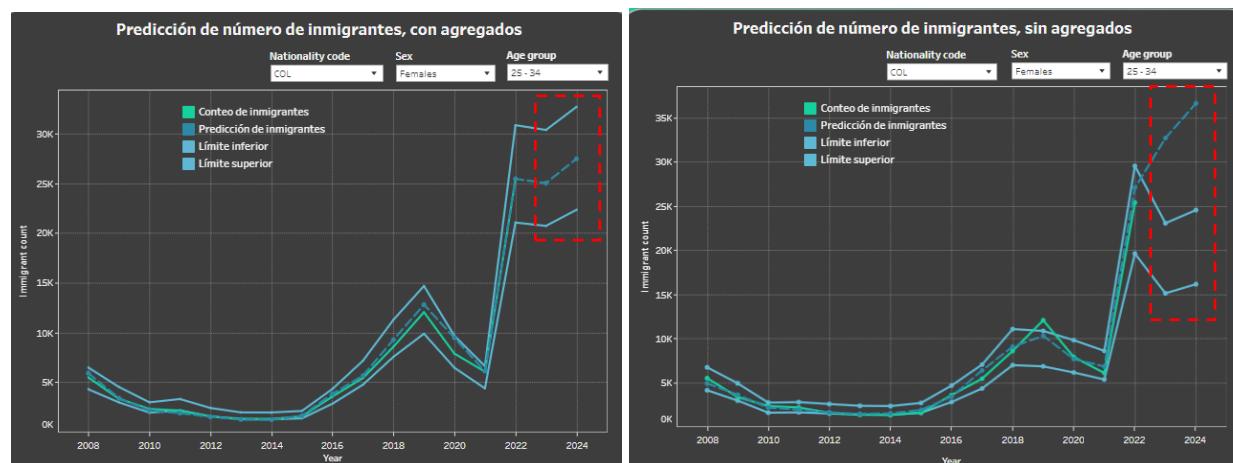
Adicionalmente, a modo de prueba y estudio de los modelos, también se predijo la cantidad de inmigrantes para Colombia y Brasil en los años 2023 y 2024 (Figura 46A y 46B, respectivamente), observando que las predicciones de número de inmigrantes del modelo sin agregados salen del intervalo de confianza y siguen incrementando, continuando la tendencia marcada por el año 2022, mientras que en las predicciones del modelo con agregados hay un “stop” del impulso alcista postpandemia y se mantiene dentro de los intervalos de confianza.

Se pudo constatar lo mencionado debido a que datos preliminares del total de inmigrantes colombianos que llegaron a España en el 2023 indican que fueron 158,600¹⁰, y tenemos que la predicción del modelo con agregados fue 146,985 inmigrantes colombianos, lo que representa un menor error frente a la predicción del modelo sin agregados (252,609 inmigrantes).

Esto puede explicarse debido a que:

- En el 2022 (último punto de tiempo) todavía se está en un contexto irregular (postpandemia) y no se pudo proveer al modelo con datos posteriores para que pudiese aprender el regreso a condiciones “normales”.
- El modelo con agregados hizo un entrenamiento con datos de inmigración con mayor dispersión, por lo que estuvo expuesto a contextos de mayor variabilidad de datos y como se relacionan los grupos dentro de las categorías en ese contexto, por lo que puede presentar una mayor tolerancia al momento de exponerse a años después del 2022.

(A)



(B)

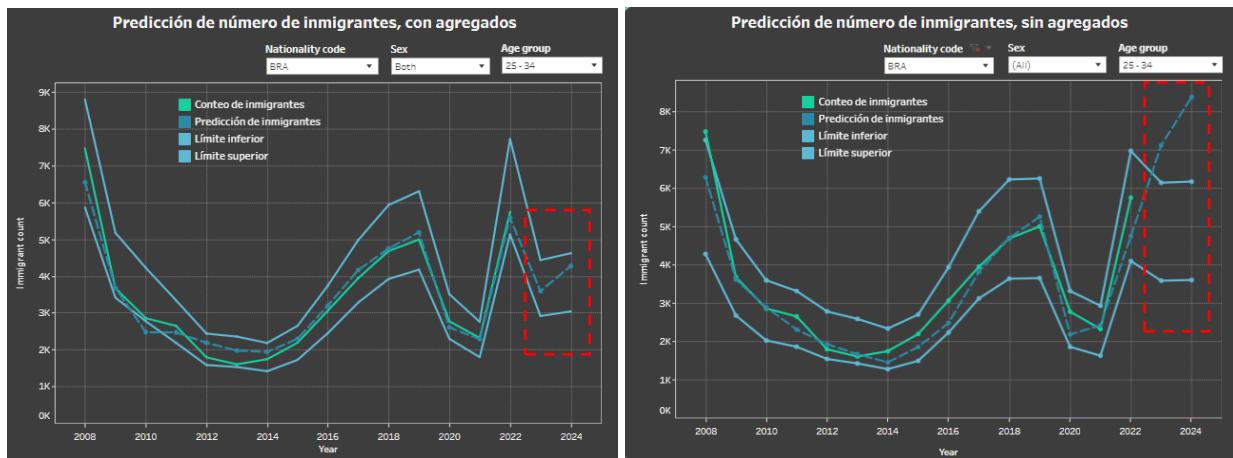


Figura 45. Muestra de predicciones de inmigrantes para Colombia (A) y Brasil (B) para los modelos con (derecha) y sin (izquierda) agregados de sexo y grupos de edad que incluyen los años 2023 y 2024.

A razón de lo expuesto, se considera más adecuado el uso del modelo de red neuronal de una capa que incluyen los agregados debido a que presenta una buena capacidad predictiva, intervalos de confianza más acotados y una mejor respuesta en predicciones a futuro.

Cabe destacar que, debido a la metodología aplicada en nuestro estudio (e.g., enfoque en valores anuales de variables macro pertenecientes a diversos dominios), no es un modelo ideal para extrapolación predicciones a largo plazo. Al ser un modelo predictivo dependiente de valores socioeconómicos, políticos, sociales, etc., no es viable asegurar cómo serán las condiciones de cada país en estas materias en intervalos amplios de tiempo. Es más recomendable implementar el uso de este modelo como una herramienta de estimación de la inmigración a 1 o 2 años de países con alta migración a España como soporte de planificación y gestión gubernamental, y también para evaluar el efecto de posibles escenarios atípicos sobre la inmigración ciudadanos de estos países.

6. CONCLUSIONES

El modelo seleccionado (redes neuronales con normalización de *inputs* y *target*), que incluye categorías de agregados de sexo y grupos de edad, mostró una buena capacidad predictiva para estimar el número de inmigrantes extranjeros, recomendándose su uso para predicciones de 1 o 2 años en el futuro. Se resaltan los siguientes aspectos:

- El modelo superó el objetivo inicial de 0.8 de coeficiente de determinación, mostrando un R^2 ajustado de 0.975 en el conjunto de prueba. Sin embargo, habiendo observado la distribución y dispersión de los datos, además de los datos atípicos, se consideró más apropiado el uso del RSME y MAE como punto comparativo entre modelos para la selección del más indicado, siendo de 853 y 293, respectivamente, para el modelo seleccionado.
- Nacionalidades como Marruecos, Rumania, Venezuela y Colombia dominan el actual escenario de inmigración en España, pero los aspectos que influyen sobre los movimientos migratorios varían entre países debido a sus características diversas. Sin embargo, hay escenarios atípicos pueden generar grandes cambios en la dinámica migratoria de forma imprevista, como se observó durante el periodo de restricciones sanitarias y escenario postpandemia entre 2020 – 2022, o el incremento abrupto de inmigración ucraniana debido al conflicto bélico con Rusia.
- La migración en cadena parece ser el elemento determinante en la selección de España como destino de inmigración para las nacionalidades más relevantes, especialmente dentro de las nacionalidades más importantes en inmigración. Así, el origen del impulso migratorio parece estar vinculado a las condiciones del país de origen (socioeconómicas, políticas, etc.), pero la selección del destino está más vinculada al fenómeno de migración en cadena. Adicionalmente, los grupos de edades son otro elemento importante que considerar debido a la predominancia de inmigración de jóvenes y adultos jóvenes de entre 15-34 años y a la baja inmigración de individuos con +65 años.

6.1 Trabajo Futuro:

- Evaluar el desempeño (métricas) de la media de predicciones de dos de los modelos con mejores resultados, por ejemplo, evaluar las métricas de media de predicciones del modelo de HGB y el modelo de red neuronal de una capa.
- Debido a los escenarios atípicos de pandemia/postpandemia, es importante incorporar al modelo datos oficiales de inmigración de, por lo menos, el año 2023. De esta forma, se puede proveer al modelo de datos que reflejen la nueva tendencia fuera de dicho escenario. Los datos oficiales completos estarían disponibles en el INE¹⁰ a partir de enero de 2025.
- Incluir un estudio de inferencial causal para identificar y cuantificar las relaciones de causa y efecto entre variables.

REFERENCIAS BIBLIOGRÁFICAS

1. International Organization for Migration (2011). *Factores que propician la migración internacional*. Recuperado en Julio de 2024 de <https://emm.iom.int/es/handbooks/contexto-global-de-la-migracion-internacional/factores-que-propician-la-migracion>.
2. Organización Internacional de Migraciones (2022). *Informe sobre las Migraciones en el Mundo 2022*. Geneva. Recuperado en Julio de 2024 de <https://publications.iom.int/books/informe-sobre-las-migraciones-en-el-mundo-2022>
3. Banco Europeo de Inversiones (BEI) (marzo 2016). *Migración y las naciones europeas. Retos, oportunidades y el papel del BEI*. Recuperado en Julio de 2024 de <https://publications.iom.int/books/informe-sobre-las-migraciones-en-el-mundo-2022>
4. CLIMB Database. *Human Mobility in the Context of Disasters, Climate Change and Environmental Degradation Database*. Red de las Naciones Unidas. Recuperado en Julio de 2024 de <https://migrationnetwork.un.org/climb#accordion>
5. Instituto Nacional de Estadística de España (2022). *Estadística de Migraciones y Cambios de Residencia (EMCR)*. Nota de Prensa. Recuperado en Julio de 2024 de https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177098&menu=ultiDatos&idp=1254735573002
6. Aydemir B., Aydin H., Çetinkaya A. y Polat D. Ş. (2022). Predicting the Income Groups and Number of Immigrants by Using Machine Learning (ML). *International Journal of Multidisciplinary Studies and Innovative Technologies*, 6(2), 162-168.
7. Carammia M., Maria Iacus S. y Wilkin T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Nature, Scientific Reports*, 12, 1457.
8. Hosseini, S. I. y Tarasyev, A. (2018). Machine Learning Methods in individual Migration Behavior. *UDC 331.214: 338.27*.
9. Micevska, M. (2021). Revisiting forced migration: A machine learning perspective. *European Journal of Political Economy*, 70, 102044.
10. Instituto Nacional de Estadística de España. *Estadística de Migraciones*. Recuperado en Julio de 2024 de: <https://www.ine.es/dynt3/inebase/index.htm?padre=3694&capsel=1963>
11. International Organization for Standardization (ISO). *Country Codes Collection*. Recuperado en Julio de 2024 de <https://www.iso.org/publication/PUB500001.html>.
12. Instituto Nacional de Estadística de España. *Clasificación de países y territorios*. Recuperado en Julio de 2024 de <https://www.ine.es/daco/daco42/clasificaciones/paisesyterritorios.xls>
13. Instituto Nacional de Estadística de España. *Principales series de Población desde 1998*. Recuperado en Julio de 2024 de <https://www.ine.es/dynt3/inebase/es/index.htm?type=pcaxis&path=/t20/e245/p08/&file=pcaxis&dh=0&capsel=>

14. World Bank Group. *GDP - Inflation - Country Risk*. Recuperado en Julio de 2024 de <https://databank.worldbank.org/GDP---Inflation---Country-Risk/id/ae8fd58a>
15. World Bank. *A Global Database of Inflation*. Recuperado en Julio de 2024 de <https://www.worldbank.org/en/research/brief/inflation-database>
16. International Monetary Fund. *Real GDP growth*. Recuperado en Julio de 2024 de https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/ADVEC/WEOWO_RLD/VEN
17. Skaaning et al. (2023). *Political regime*. Procesado por Our World in Data y recuperado en Agosto de 2024 de <https://ourworldindata.org/grapher/political-regime-lexical?tab=table>
18. The Varieties of Democracy (V-Dem) (2024). *Democracy Data Explorer*. Procesado por Our World in Data y recuperado en Agosto de 2024 de <https://ourworldindata.org/explorers/democracy>
19. World Bank Group. *Global State of Democracy* (GSoD). Recuperado en Julio de 2024 de <https://prosperitydata360.worldbank.org/en/dataset/IDEA+GSOD>
20. International Institute for Democracy and Electoral Assistance (International IDEA). *Democracy Assessment*. Recuperado en Julio de 2024 de <https://www.idea.int/theme/democracy-assessment>
21. United Nations Office on Drugs and Crime. *International Homicide*. Recuperado en Julio de 2024 de <https://dataunodc.un.org/dp-intentional-homicide-victims>
22. Expansión (Datosmacro). *Homicidios Intencionados*. Recuperado en Julio de 2024 de <https://datosmacro.expansion.com/demografia/homicidios>
23. Uppsala Conflict Data Program (2024) y Natural Earth (2022). *War and Peace - Armed Conflict and War*. Adaptado por Our World in Data y recuperado en Julio de 2024 de <https://ourworldindata.org/war-and-peace#explore-data-on-armed-conflict-and-war>
24. Statista. Número anual de turistas internacionales que llegaron a España en avión de 2000 a 2023. Recuperado en Julio de <https://es.statista.com/estadisticas/474883/llegadas-de-turistas-extranjeros-a-espana-en-avion/>
25. Arango, J. (2000). Explaining migration: A critical view. *International Social Science Journal*, 52(165), 283-296.
26. Gozález Pérez, V. (1999). Migraciones africanas a España: El caso de los marroquíes. *Revista Española de Investigaciones Sociológicas (REIS)*, (85), 9-29.
27. Stanek, M. (2011). La migración circular de los rumanos en España: Una respuesta a la crisis económica. *Migraciones*, (30), 147-171.
28. Pedone, C. (2009). La migración colombiana en España: Entre la regularización y la integración. *Revista CIDOB d'Afers Internacionals*, (87), 113-131.
29. González, J. C., & Santacruz, A. (2020). La migración venezolana en España: Redes sociales y factores de atracción. *Migraciones Internacionales*, 11(2), 135-157.

ANEXOS

Anexo 1. Definiciones relevantes de algunas de las variables empleadas.

Variable	Descripción
GDP growth	Tasa de crecimiento porcentual anual del PIB a precios de mercado basados en moneda local constante. Los agregados se basan en dólares estadounidenses constantes de 2010.
Unemployment %	Se refiere a la proporción de la población activa que no tiene trabajo, pero está disponible y busca empleo.
Inflation annual	Medida por la tasa de crecimiento anual del deflactor implícito del PIB, muestra la tasa de variación de los precios en el conjunto de la economía.
Probability of dying young	Probabilidad de morir entre los 20 y los 24 años de edad expresada por cada 1000 jóvenes de 20 años, si están sujetos a las tasas de mortalidad específicas por edad del año especificado.
Political and Violence Percentile	Mide la percepción de la probabilidad de inestabilidad política y/o violencia por motivos políticos, incluido el terrorismo. El rango percentil indica el rango del país entre todos los países cubiertos por el indicador agregado, correspondiendo 0 al rango más bajo y 100 al rango más alto.
Regulatory Quality: Percentile Rank	Capta la percepción de la capacidad del gobierno para formular y aplicar políticas y normativas sólidas que permitan y promuevan el desarrollo del sector privado. El rango percentil indica el rango del país entre todos los países cubiertos por el indicador agregado, correspondiendo 0 al rango más bajo y 100 al rango más alto.
Rule of Law Percentile	Refleja la percepción de hasta qué punto los agentes confían en las reglas de la sociedad y las respetan, y en particular la calidad del cumplimiento de los contratos, los derechos de propiedad, la policía y los tribunales, así como la probabilidad de delincuencia y violencia. El rango percentil indica el rango del país entre todos los países cubiertos por el indicador agregado, correspondiendo 0 al rango más bajo y 100 al rango más alto.
Voice and Accountability Percentile	Recoge la percepción del grado en que los ciudadanos de un país pueden participar en la elección de su gobierno, así como la libertad de expresión, la libertad de asociación y la libertad de los medios de comunicación. El rango percentil indica el rango del país entre todos los países cubiertos por el indicador agregado, correspondiendo 0 al rango más bajo y 100 al rango más alto.
Salaried workers %	Trabajadores que ocupan el tipo de empleos definidos como «empleos asalariados», cuyos titulares tienen contratos de trabajo explícitos (escritos u orales) o implícitos que les otorgan una remuneración básica que no depende directamente de los ingresos de la unidad para la que trabajan.

Homicide Rate Tasa de homicidios por cada 100,00 habitantes

Liberal democracy index	Combina información sobre los derechos de voto, la libertad e imparcialidad de las elecciones, las libertades de asociación y expresión, las libertades civiles y las limitaciones del poder ejecutivo. Va de 0 a 1 (más democrático).
Deliberative democracy index	Combina información sobre los derechos de voto, la libertad e imparcialidad de las elecciones, las libertades de asociación y expresión, así como la medida en que los ciudadanos y los dirigentes debaten opiniones diferentes y buscan el bien público. Va de 0 a 1 (más democrático).
Absence of Corruption	Indica hasta qué punto el ejecutivo y la administración pública, en general, no abusan de su cargo para obtener beneficios personales. de su cargo en beneficio propio; sus rangos van de 0 a 1 (mayor ausencia)
Civil Liberties and Civil Rights	Indica hasta qué punto se respetan los derechos y libertades civiles; su rango va de 0 a 1 (mayor respeto)
Judicial accountability	Conjunto de mecanismos destinados a responsabilizar personal o institucionalmente a los jueces y tribunales por comportamientos y decisiones contrarios a las normas constitucionales o legales; su rango va de 0 a 1 (mayor responsabilidad)
One-sided violence deaths	Incluye las muertes de civiles debidas a conflictos unilaterales que estaban en curso ese año. Se considera un conflicto unilateral a aquel que donde hay uso de la fuerza armada por parte de un grupo armado estatal o no estatal contra civiles que cause al menos 25 muertes de civiles durante un año.
Non-state deaths	Incluye las muertes de combatientes y civiles debidas a conflictos no estatales que estaban en curso ese año. Se considera un conflicto no estatal a aquellos que ocurren entre grupos armados no estatales, como grupos rebeldes, organizaciones delictivas o grupos étnicos, que causa al menos 25 muertes en un año.
Intrastate deaths	Incluye las muertes de combatientes y civiles debidas a combates en conflictos intraestatales que estaban en curso ese año. Se considera un conflicto intraestatal a aquel que ocurre entre un Estado y un grupo armado no estatal dentro del territorio del Estado que causa al menos 25 muertes durante un año
Interstate deaths	Incluye las muertes de combatientes y civiles debidas a combates en conflictos interestatales que estaban en curso ese año. Se considera un conflicto interestatal a aquel que ocurre entre Estados causando al menos 25 muertes al año.

Political
regime

Identifica el tipo de régimen político en los países por año:

- **Autocracia no electoral (0):** los ciudadanos no tienen derecho a elegir al jefe del Ejecutivo ni al Legislativo. Autocracia unipartidista: algunos ciudadanos tienen derecho a elegir al jefe del Ejecutivo o al Legislativo, pero sólo tienen una opción.
 - **Autocracia unipartidista (1):** algunos ciudadanos tienen derecho a elegir al jefe del Ejecutivo o del Legislativo, pero sólo tienen una opción.
 - **Autocracia multipartidista sin ejecutivo elegido (2):** algunos ciudadanos tienen derecho a elegir el poder legislativo y tienen más de una opción, pero el jefe del ejecutivo no es elegido.
 - **Autocracia multipartidista (3):** algunos ciudadanos tienen derecho a elegir al jefe del Ejecutivo y al Legislativo y más de una opción, pero el resultado de las elecciones es seguro.
 - **Democracia excluyente (4):** los ciudadanos tienen derecho a elegir al jefe del ejecutivo y al legislativo en elecciones multipartidistas e inciertas, pero el sufragio está restringido.
 - **Democracia masculina (5):** los ciudadanos tienen derecho a elegir al jefe del ejecutivo y al legislativo en elecciones multipartidistas e inciertas, pero el sufragio está restringido a los hombres.
 - **Democracia electoral (6):** los ciudadanos tienen derecho a elegir al jefe del ejecutivo y al legislativo en elecciones multipartidistas e inciertas. Los ciudadanos tienen derecho a elegir al jefe del ejecutivo y al legislativo en elecciones multipartidistas e inciertas, y gozan de libertades de expresión, reunión y asociación.
 - **Poliarquía (7):** Las poliarquías son democracias electorales que también protegen las libertades de expresión, reunión y asociación.
-