Logotipo, nombre de la empresa

Descripción generada automáticamente

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/a de TFM:**

Miguel Martín

Madrid, 30 de octubre de 2024

# Índice

[Índice 2](#_Toc1709822763)

[RESUMEN 3](#_Toc927097742)

[1. INTRODUCCIÓN 4](#_Toc61327122)

[2. ANTECEDENTES 7](#_Toc1254347296)

[3. OBJETIVO 9](#_Toc1499882121)

[4. MATERIALES Y MÉTODOS 10](#_Toc871292569)

[4.1 Adquisición de fuentes de datos 10](#_Toc2084564337)

[4.2 Descripción de fuentes de datos 11](#_Toc29696844)

[4.3 Resumen de herramientas e infraestructura empleada 11](#_Toc1966920349)

[4.4 Procedimiento 12](#_Toc1475069877)

[5. RESULTADOS 14](#_Toc1928343364)

[5.1 Etapa 1: Limpieza, preprocesamiento y análisis de datos de inmigración en España 14](#_Toc1132184090)

[5.4 Etapa 2: Limpieza, preprocesamiento y análisis de datos de variables explicativas 22](#_Toc1853550642)

[5.4.1 Continentes y Sub-regiones 22](#_Toc898717322)

[5.4.2 Padrón de Inmigrantes Residentes en España 22](#_Toc439170164)

[5.4.3 Indices de Desarrollo 22](#_Toc1650981470)

[5.4.4 Régimen Político 23](#_Toc125886524)

[5.4.5 Indices de Democracia 23](#_Toc215303433)

[5.4.6 Indices de Libertad y otros 23](#_Toc1039736542)

[5.4.7 Tasa de Homicidios 23](#_Toc1267520712)

[5.4.8 Conflictos Armados 23](#_Toc795103808)

[5.4.9 Turistas Anuales 23](#_Toc1415242180)

[5.5 Etapa 3: Unión y selección de variables 23](#_Toc867466461)

[5.6 Etapa 4: Prueba y comparación de modelos de machine learning 23](#_Toc1084371808)

[Análisis Exploratorio de Datos o EDA 23](#_Toc355544924)

[Conclusiones 28](#_Toc1240685327)

[Referencias bibliográficas 29](#_Toc837951285)

[Anexos 31](#_Toc1435407196)

# RESUMEN

# 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).

A diagram of a decision

Description automatically generated

[**Figura 1**](https://emm.iom.int/es/handbooks/contexto-global-de-la-migracion-internacional/factores-que-propician-la-migracion). 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.

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).

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.

Aydemin 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 SVM/KNN (83.72% para ambos), mientras que para su estudio de regresión, mencionan una tasa de exito 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 Aydemin 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 usado de variables de desarrollo, en consecuencia, estimar el futuro estado politico, 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 Imán H.S y Tarasyev, A (2018)10 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, M. (2021)11 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 Micevska, (2021). 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”, Micevska, (2021).

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

* Lista de nacionalidades mas importantes...
* Obtener un modelo predictivo con coeficiente de determinación cercano o superior al 0.80 (funcional).

# 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*)12.

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 cuáles 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 (en Anexos) 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 la Tabla 2 (ver Anexos) definiciones relevantes de algunas variables que son intuitivas.

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.

**Repositorio**: Se creó un repositorio de GitHubx para los ficheros, notebooks, *exports* resultantes y el escrito final, trabajando progresivamente en el proyecto mediante ramas a nivel local, haciendo *commits* de trabajoy revisiones conjuntas.

**Infraestructura computaciona**l: 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) ~~y tensorflow (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

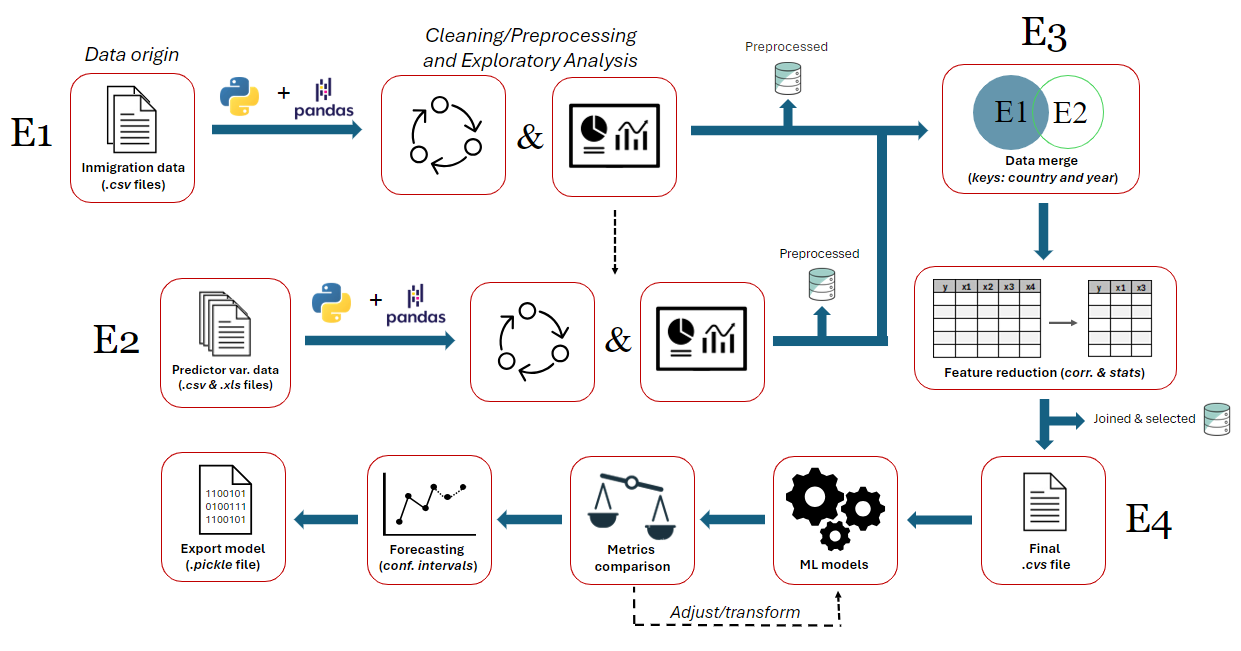
Tensorflow 2.16.1 ~~/ Keras 3.3.3~~

### 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 y de árboles (Etapa 4 - E4), comparando distintas métricas para observar el rendimiento de los modelos, identificar los mejores y realizar ajustes/transformaciones de éstos para disminuir el error hasta obtener el modelo con el mejor rendimiento posible. Luego se procedió a exportarlo y 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.

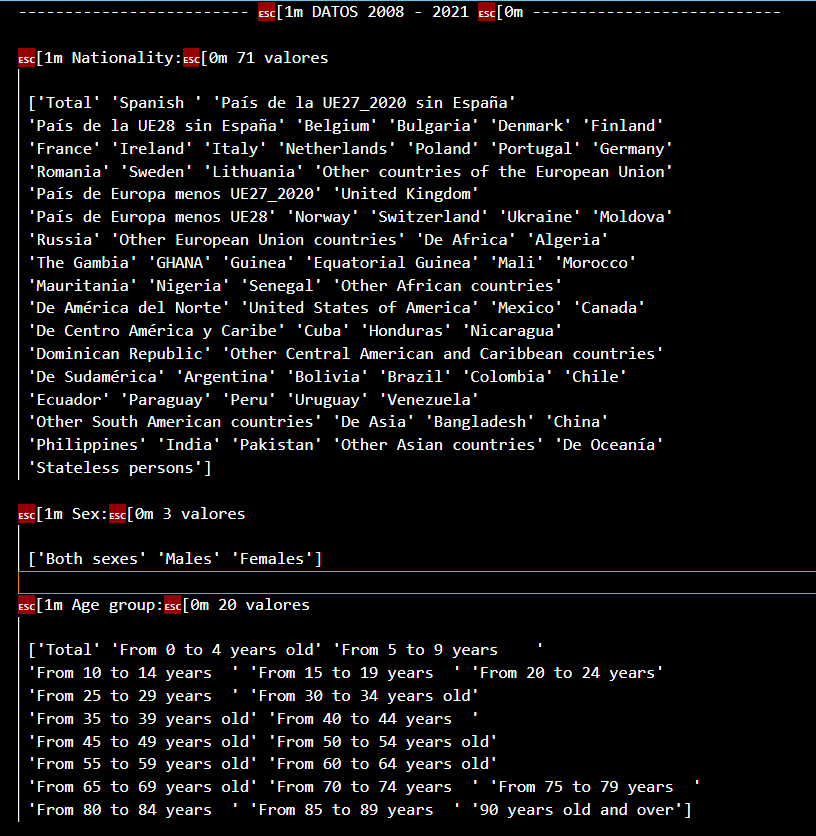
# 5. RESULTADOS

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

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” para mayores detalles de esta etapa). Luego, se compararon los valores que tomaban las variables categóricas (Figura 4), observando que:

* El conjunto de datos para los años 2008-2021 estaba en idioma inglés, mientras que el de 2022 estaba en español. Procedimos a estandarizar los nombres de los países en inglés, así como usar un estándar ISO3 con códigos de países.
* 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)**

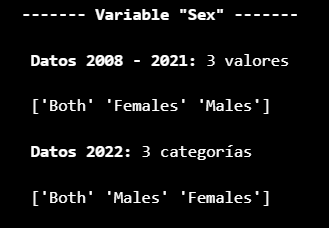
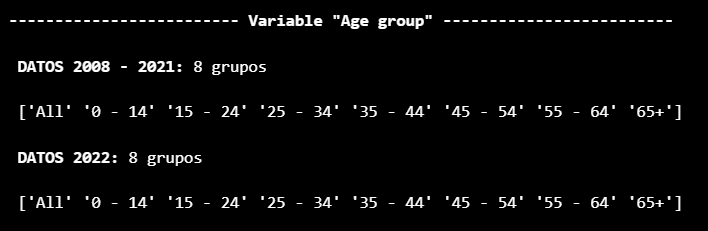


**(B)**



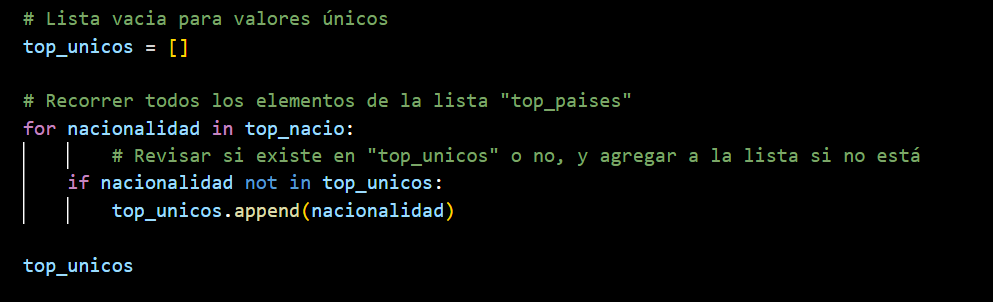
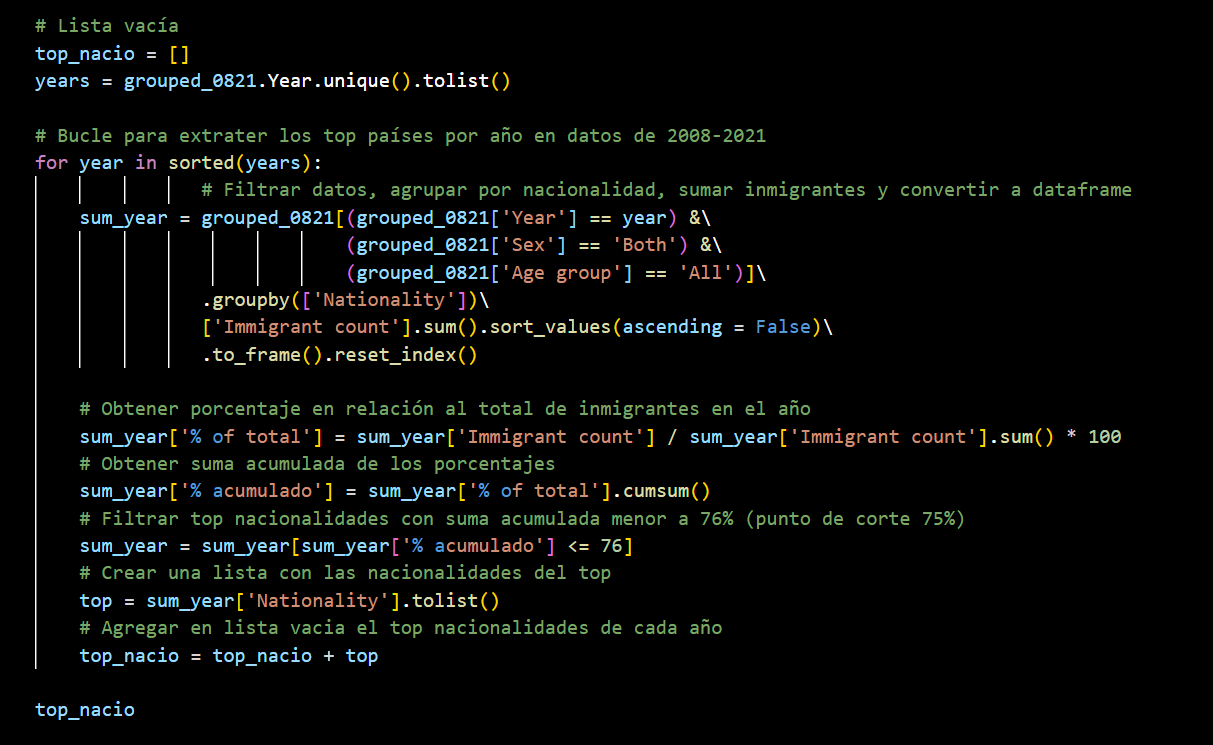
**Figura 4.** Valores que toman las categorías de nacionalidades, sexo y grupos de edad para los *datasets* de inmigración de 2008-2021 (A) y 2022 (B).

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

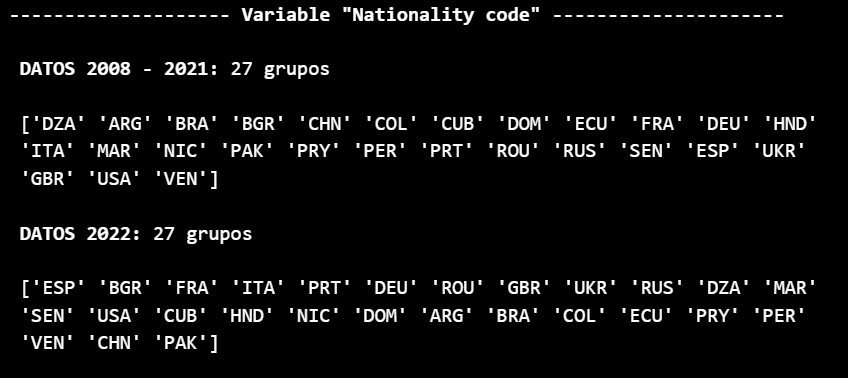
 

**Figura 5**. Valores las categorías de sexo (izquierda) y grupos de edad (derecha) post-limpieza/estandarización.

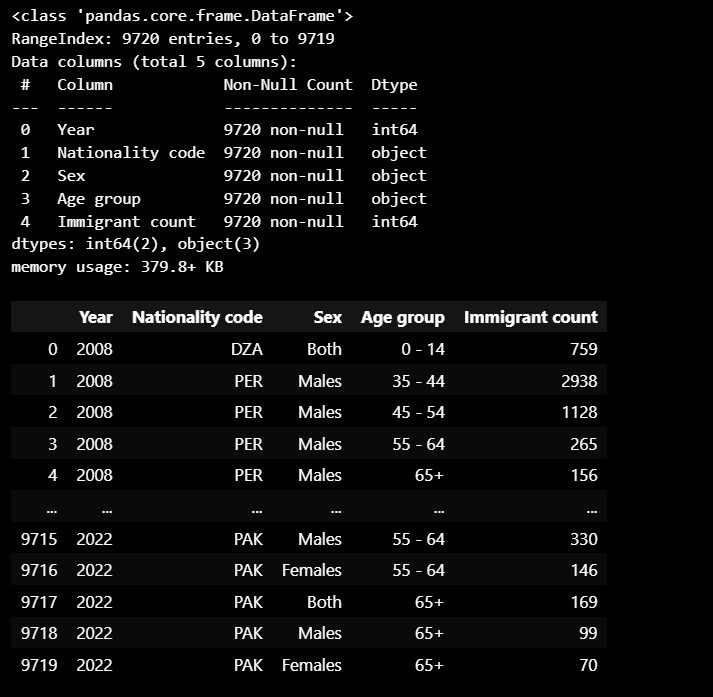
En cuanto a las nacionalidades, se removieron aquellas categorías relacionadas a regiones/continentes y se hizo una reducción del número de nacionalidades. Para ello, se consideraron como “top nacionalidades” a aquellas que, en conjunto y de forma decreciente, englobasen alrededor del 75% de los inmigrantes 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).

**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 las mismas en una nueva columna “Nationality code” usando la nomenclatura internacional de tres códigos (ALPHA-3), descrita en la ISO 316613,para luego proceder a juntar ambos conjuntos de datos en un *dataframe* (Figura 8).



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



**Figura 8**. *Dataframe* de la unión de los *datasets* de datos de inmigración post-limpieza/pre-procesamiento

Finalmente, continuando con los análisis de los valores totales de inmigración en 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.

A graph with a line and a blue line

Description automatically generated

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

A graph of a number of people

Description automatically generated

**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.

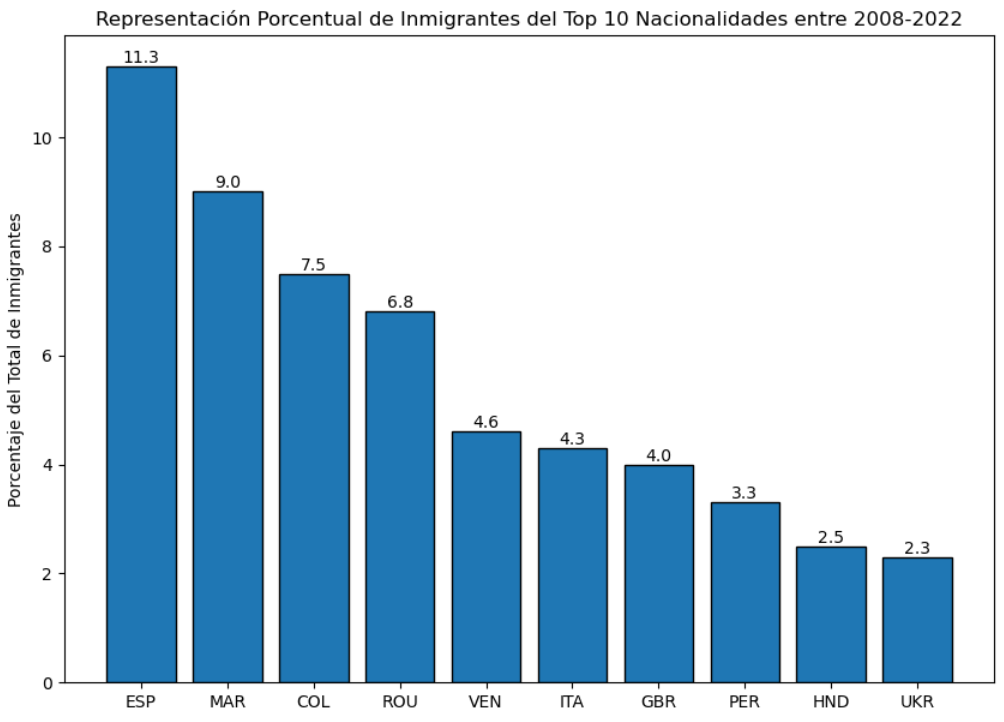
A screenshot of a computer

Description automatically generated**Figura 11.** Distribución poblacional de inmigrantes en España por año de 2008 a 2022.

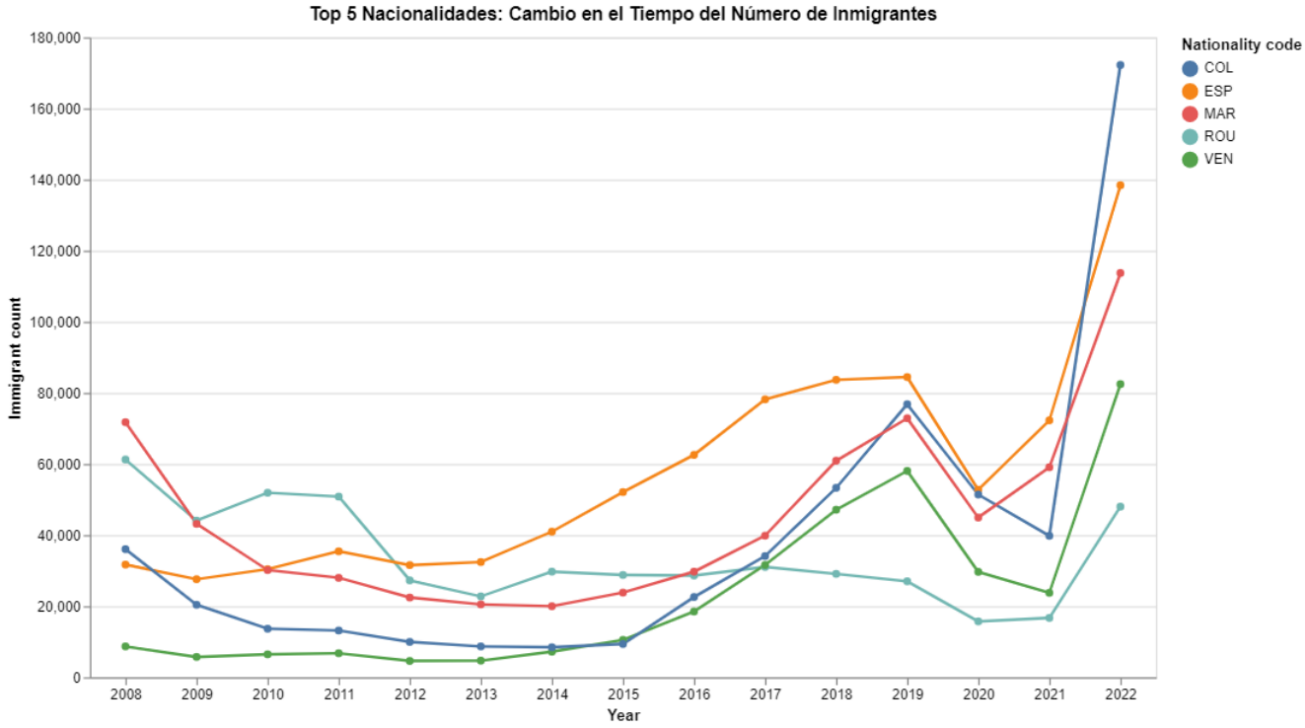
A partir de estas observaciones, se analizó la cantidad total por nacionalidad, observando que los inmigrantes de nacionalidad española, colombiana, marroquí y rumana poseen la mayor representación porcentual en relación al total de inmigrantes (Figura 12A). Además, si observamos el número de inmigrantes por año dentro del Top 5 (Figura 12B), notamos que hay diferencias entre los mismo. Por ejemplo, los inmigrantes colombianos y venezolanos entán 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 cuento 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)**



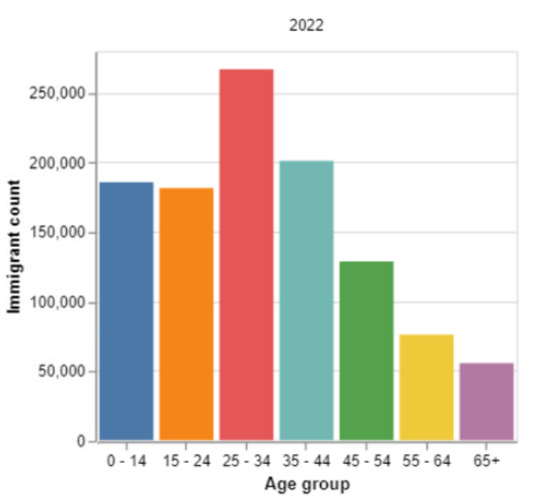
**(B)**



**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 tener en cuenta, 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.

### 5.4 Etapa 2: Limpieza, preprocesamiento y análisis de datos 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. Puede consultarse el *jupyter notebook* “Etapa 2” para mayores detalles del proceso de limpieza y análisis.

#### 5.4.1 Continentes y Sub-regiones

#### 5.4.2 Padrón de Inmigrantes Residentes en España

Para los datos del padrón de inmigrantes residentes en España, 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

A screenshot of a computer program

Description automatically generated

**Figura 14.** Extracto de grupos de países encontrados en padrón de residentes

A screen shot of a computer

Description automatically generated

**Figura 15.** Extracto de estandarización con ISO3

El resultado sobre el análisis de los datos de los residentes arrojan 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 inmigrante van cambiando conforme pasan los años.

A graph of a number of people

Description automatically generated

**Figura 16.** Suma de residentes por país de origen en el año 2022.

A graph of different colored lines

Description automatically generated

**Figura 17**. Principales 5 países en la inmigración en España, a través del tiempo

Nuestros datos base, cabe destacar, tienen un peso importante en nuestros modelos que más adelante expondremos.

#### 5.4.3 Índices de Desarrollo

Los índices de desarrollo se refieren a un conjunto de variables macroeconómicas que influyen en la vida política de los países y afectan a sus ciudadanos, propiciando una decisión de migración fundamentada. 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)*. 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”).

A screenshot of a computer program

Description automatically generated

**Figura 18**. Dataframe antes de preprocesamiento

A screenshot of a computer

Description automatically generated

**Figura 19**. Muestra de indicadores presentes el estudio, suman 172.

A screenshot of a computer

Description automatically generated

**Figura 20**. Dataframe final con países selectos y pivot aplicado.

Los índices que usamos son: *"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 a su composición política, cultura y economía, aunque en general, la mayoría de ellos tiene un rendimiento por indicador cercano al 0.5 de eficiencia (1 es la calificación más alta).

A graph with a line

Description automatically generated

**Figura 21**. Promedio de todos los países en Ausencia de corrupción.

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?

#### 5.4.4 Régimen Político

#### 5.4.5 Índices de Democracia

#### 5.4.6 Índices de Libertad y otros

#### 5.4.7 Tasa de Homicidios

#### 5.4.8 Conflictos Armados

#### 5.4.9 Turistas Anuales

En comparación con los dataframes anteriores, el dataframe de turistas tiene un volumen menor de datos. Estos datos representan un volumen de personas que llegan a España por medio de aeropuertos, mismos que se pueden representar con un simple dígito por cada año que se estudia. La finalidad de incluir estos datos es para introducir una variable explicativa que pudiera darnos un grado de volumen a la hora de modelar.

A screenshot of a number

Description automatically generated

**Figura 22**. Cuenta de turistas anuales en España

A graph of a number of years

Description automatically generated

**Figura 23**. Suma de visitantes en España 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.

### 5.5 Etapa 3: Unión y selección de variables

### 5.6 Etapa 4: Prueba y comparación de modelos de *machine learning*

### Análisis Exploratorio de Datos o EDA

El análisis exploratorio fue dividido en varios *notebooks* con la finalidad de evitar una confusión al introducir muchas variables en uno solo.

A manera de exposición, exponemos los gráficos más relevantes para comprender el tipo de análisis que ejecutamos, manteniendo el análisis completo en cada *notebook*.

A graph with a line and a blue line

Description automatically generated

Figura 1, total de inmigrantes en España de 2008 a 2022, elaboración propia

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 (post-pandemia), superando al 2019 en aproximadamente 500,000 inmigrantes.

Observando los inmigrantes que llegaron a España durante estos 15 años según la edad, vemos que predomina la inmigración de personas jóvenes de entre 25-34 años, seguidos de jóvenes de entre 15-24. A partir de ese punto, 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.

La distribución por año parece indicar que a medida que hay menor inmigración, la dispersión de los datos es menor y los valores atípicos menos extremos en relación con el rango intercuartílico (caja), mientras que cuando hay aumenta la inmigración en el tiempo, la dispersión también aumenta.

Además, vemos que la mayoría de los datos se concentran en un rango medio/bajo de número de inmigrantes (observar rango intercuartílico entre 2008-2021 que se mantiene alrededor o por debajo de 10,000) y sólo algunas nacionalidades presentan valores "atípicamente" superiores al resto. Y son precisamente estas nacionalidades con grandes valores atípicos los que son de nuestro interés de estudio y modelado, pues en ellas se condensan la mayor masa de inmigrantes hacia España. Más aún, es interesante observar cómo algunas nacionalidades mantienen un número atípicamente alto de inmigración a lo largo de todo el período 2008-2022, como son españoles, colombianos, rumanos y marroquíes. A graph of blue rectangular objects

Description automatically generated with medium confidence

Figura 4, porcentaje por total de inmigrantes en España por país de origen, elaboración propia

Previamente, ya habíamos notado la cantidad de inmigrantes españoles, colombianos, marroquíes y rumanos, y ahora lo observamos también en su distribución porcentual en base al total.

Luego de los top 4, parece variación porcentual menor entre las subsiguientes nacionalidades.

A graph of blue rectangular bars with white text

Description automatically generated

Figura 5, desempeño por indicadores de libertad para Argentina de 2008 a 2022 en promedio, elaboración propia

Para los índices de libertad, obtuvimos gráficas por cada país para conocer el desempeño por cada uno de los indicadores. En el caso de Argentina, observamos que las libertades civiles alcanzan un grado de plenitud (donde 1 significa certeza y 0 ausencia). En general, los países seleccionados obtienen un promedio de desempeño alrededor del 0.5 de todos los indicadores.

A graph with a line

Description automatically generated

Figura 6, desempeño global en indicador de igualdad en salubridad de 2008 a 2022, elaboración propia

Este indicador es particularmente interesante dado el contexto de la pandemia del COVID. El indicador responde a la accesibilidad de los ciudadanos a servicios de salud de alta calidad, para permitirles ejercer sus derechos políticos en la adultez. En franco declive desde 2008, la situación se ve un poco acelerada para 2019, tocando fondo en 2020 y teniendo un rebote. Aunque se aprecia un esfuerzo global por mejorar los servicios de salud en los países, nuevamente se pierde el avance para 2022.

A graph of a graph of a graph

Description automatically generated with medium confidence

Para efectos del estudio, podemos afirmar que los migrantes en España provienen de un trasfondo diverso, muchos marcados por carencias palpables y otros con un nivel de vida alto. Las gráficas anteriores muestran el ejemplo concreto para Pakistán mismo que tiene altibajos de un indicador a otro.

# Conclusiones

# Referencias bibliográficas

1. *Iman, H. S., & Tarasyev, A. (2018). MACHINE LEARNING METHODS IN INDIVIDUAL MIGRATION BEHAVIOR.* Disponible en:<https://elar.urfu.ru/bitstream/10995/68086/1/978-5-8295-0581-3_2018-08.pdf>
2. Micevska, M. (2021). *Revisiting forced migration: A machine learning perspective*. European Journal of Political Economy, 70, 102044. doi:10.1016/j.ejpoleco.2021.102044
3. International Organization for Migration. *Factores que propician la migración internacional*. Año 2011. Recuperado de <https://emm.iom.int/es/handbooks/contexto-global-de-la-migracion-internacional/factores-que-propician-la-migracion> el 19 de septiembre 2024.
4. Organización Internacional de Migraciones (OIM). *Informe sobre las Migraciones en el Mundo 2022*. Geneva. Disponible en: <https://publications.iom.int/books/informe-sobre-las-migraciones-en-el-mundo-2022>
5. Banco Europeo de Inversiones (BEI). *Migración y las naciones europeas. Retos, oportunidades y el papel del BEI*. Marzo 2016. Disponible en: <https://publications.iom.int/books/informe-sobre-las-migraciones-en-el-mundo-2022>
6. CLIMB Database: Human Mobility in the Context of Disasters, Climate Change and Environmental Degradation Database. Red de las Naciones Unidas. Disponible en: <https://migrationnetwork.un.org/climb#accordion>
7. Instituto Nacional de Estadística (INE). *Estadística de Migraciones y Cambios de Residencia (EMCR). Año 2022*. Nota de Prensa. Disponible en: <https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177098&menu=ultiDatos&idp=1254735573002>
8. AYDEMİR Belgin, AYDIN Hakan, ÇETİNKAYA Ali y POLAT Doğan Şafak. *Predicting the Income Groups and Number of Immigrants by Using Machine Learning* *(ML).* International Journal of Multidisciplinary Studies and Innovative Technologies. 2022. 6 (2), 162-168.
9. CARAMMIA Marcello, MARIA IACUS Stefano y WILKIN Teddy*. Forecasting asylum‑related migration flows with machine learning and data at scale*. Nature, Scientific Reports. 2022. 12:1457.
10. Instituto Nacional de Estadística (INE). *Estadística de Migraciones y Cambios de Residencia (EMCR). Año 2022*. Nota de Prensa. Disponible en: <https://www.ine.es/dynt3/inebase/index.htm?padre=3694&capsel=1963>
11. ISO https://www.iso.org/obp/ui/#search

# Anexos