Logotipo, nombre de la empresa

Descripción generada automáticamente

MÁSTER EN DATA SCIENCE Y BUSINESS ANALYTICS PRESENCIAL

Predicción sobre el número de inmigrantes en España mediante modelo de aprendizaje automático para prever flujos migratorios internacionales hacia el país

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](#_Toc462893297)

[RESUMEN 3](#_Toc2125671252)

[1. INTRODUCCIÓN 4](#_Toc1230000241)

[2. ANTECEDENTES 7](#_Toc731643177)

[3. OBJETIVO 8](#_Toc1317909006)

[4. MATERIALES Y MÉTODOS 9](#_Toc1240271292)

[5. RESULTADOS 11](#_Toc1997622069)

[5.1 Adquisición de fuentes de datos 11](#_Toc1679966463)

[5.2 Descripción de fuentes de datos 12](#_Toc1387243860)

[Análisis Exploratorio de Datos o EDA 12](#_Toc1477394137)

[Preprocesado de datos 18](#_Toc1017387196)

[Selección de variables 18](#_Toc1217994107)

[Partición de los datos en conjuntos train y test 18](#_Toc914209585)

[Aplicación de modelos de aprendizaje automático 18](#_Toc1502061584)

[Evaluación y validación de modelos de aprendizaje automático 19](#_Toc1268458081)

[Comparativa de resultados 19](#_Toc751764618)

[Conclusiones 19](#_Toc1874922170)

[Referencias bibliográficas 20](#_Toc1784690071)

[Anexos 21](#_Toc664074316)

# 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) mencionaba, 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 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 los 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.

(International Organization for Migration)

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 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 Unión Europea (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.

Exponer lo que hicieron, como y lo que obtuvieron

Relacionar los hallazgos de los distintos

Buscar un par mas de antecedentes (pueden ser también estrategias mas antiguas que no fuesen de ML, probablemente con metodos estadisticos)

# 3. OBJETIVO

* Predecir el número de inmigrantes en España mediante modelo de aprendizaje automático para prever flujos migratorios internacionales hacia el país.

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

* **Base de datos**: Los datos recopilados y usados en este proyecto fueron datos estructurados/relacionales. Siendo un estudio temporal de inmigración, las claves centrales con los que se relacionaron los distintos ficheros empleados se basaron en: año, nombre del país/código y sexo; las relaciones varían según los datos que aporta cada fichero.
* **Repositorio**: 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 trabajoy revisiones conjuntas.
* **Infraestructura computaciona**l: Se trabajo en local mediante equipos personales y software de uso libre y/o estándares: VS Code, GitBash, Jupyter Notebooks, MS Excel, MS Word y MS PowerPoint (diagramas). Adicionalmente, las librerías principales que se usaron en python durante el proyecto fueron: pandas, numpy, matplotlib, altair, 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

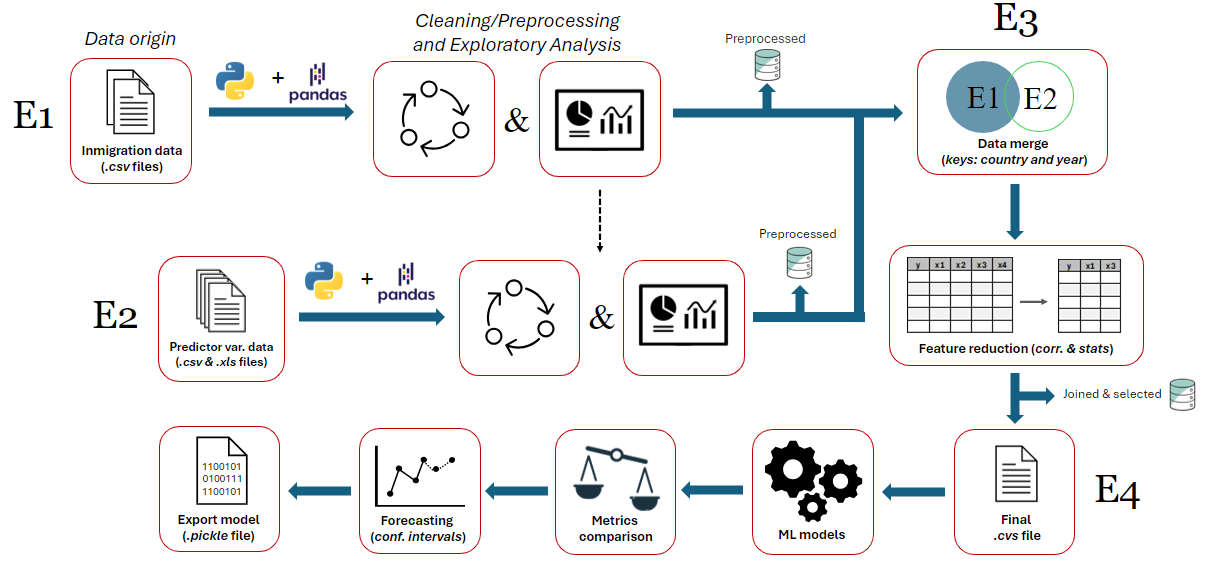
Tensorflow 2.16.1 / Keras 3.3.3

El punto de partida del proyecto fueron los datos de inmigración obtenidos del Instituto Nacional de Estadística de España. A partir de allí, y teniendo en cuenta la variedad de conjuntos de datos a considerar a razón de nuestro enfoque macro del problema, incluyendo la necesidad de limpieza/preprocesamiento que cada uno presentaría, se decidió dividir el proyecto en 4 fases (Figura X) 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.
* **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.

Posteriormente, se procedió con la unión de todas las variables predictoras al conjunto de datos central de datos de inmigración (Etapa 3), 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 seleccionar el mejor modelo, exportarlo y realizar predicciones, añadiendo como un intervalo de confianza del 90%.

Figura X. 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.

# 5. RESULTADOS

### 5.1 Adquisición de fuentes de datos

(pendiente:)

La adquisición de datos originales supuso el planteamiento del objetivo en distintas fases o partes que mutuamente se complementaran. El fenómeno de la migración depende de múltiples factores macroeconómicos, sociales, políticos y también de factores personales, psicológicos o laborales de cada persona sujeto. Los principales factores que identificamos para elaborar nuestros *datasets* originales se dividían en las siguientes categorías:

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

Las fuentes de datos contemplan estudios o pronunciamientos de organismos internacionales, ONG y gobiernos.

### 5.2 Descripción de fuentes de datos

Nuestros datos comprenden de distintos tipos de *datasets* que miden fenómenos que poco tienen que ver entre sí, si los miramos desde un punto de vista pragmático. Contamos con variables que cuentan población en número de residentes, valores booleanos para representar ausencia o presencia de alguna característica como un índice y también valores estandarizados de 0 a 1 para representar el grado de solidez de un indicador.

En el Anexo, presentamos una descripción detallada de todas las variables que usamos en nuestros *datasets*, misma que fue sujeta a un preprocesamiento, análisis exploratorio y reducción dimensional previo al entrenamiento de modelos estadísticos.

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

### 5.4 Etapa 2: Limpieza, preprocesamiento y análisis de datos de variables económicas, sociales y políticas

### 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

A graph of a number of people

Description automatically generated

Figura 2, distribución de la población inmigrante en España por sexo y grupo de edad, 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.

A screenshot of a computer

Description automatically generated

Figura 3, distribución poblacional de inmigrantes en España por cada año de 2008 a 2022, elaboración propia

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. 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.
2. 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>
3. 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>
4. 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>
5. 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>
6. 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.
7. 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.

# Anexos