



# ANÁLISIS DE PATRONES MIGRATORIOS A NIVEL MUNDIAL

*DEPARTAMENTO DE DATOS DE ANALYTIC HOUND | FEBRERO 2023*

## **DISCLAIMER:**

**Analytic Hound** ® es una consultora con base en Buenos Aires (Argentina) sin ningún tipo de conflictos de interés sociopolítico o económico relacionados a sus diferentes objetos de estudio. La siguiente documentación ha sido publicada en buena fe, dejando en claro que la información aquí proporcionada es sólo de carácter general. No ofrecemos ninguna garantía sobre la integridad, fiabilidad y precisión de esta información. Cualquier acción que pudiera ser tomada basada en la información aquí presente es estrictamente responsabilidad del lector. No somos responsables de ninguna pérdida o daño en relación al uso de este documento.

Las opiniones que puedan expresarse a lo largo de esta documentación son de estricta responsabilidad de los autores. Se deja constancia que la empresa Henry ®, el “project owner” o cualquier otra entidad que pudiera ser vinculada al análisis aquí citado no presenta relación real de ningún tipo.

*El equipo de Analytic Hound* ®

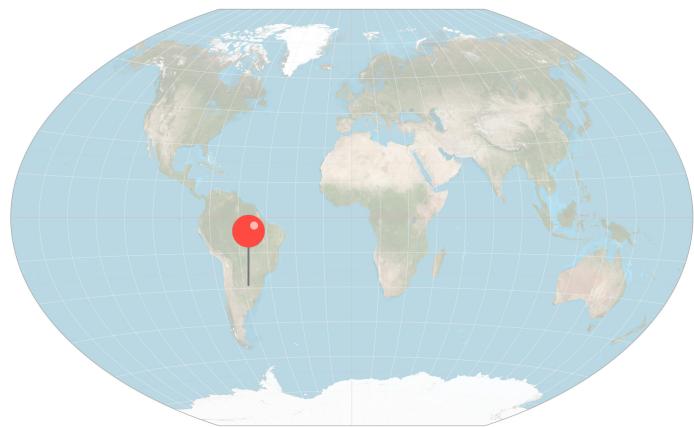
## **EQUIPO:**

MYSLER, ALAN	ZAPATA, BELEN	RODRIGUEZ, LUCAS	BELL, EUGENIA	PINI, MAURO
CIENTÍFICO DE DATOS	ANALISTA DE DATOS	INGENIERO DE DATOS	INGENIERO DE DATOS	ANALISTA FUNCIONAL



## ACERCA DE NOSOTROS:

**Analytic Hound** ® es una consultora tradicional altamente especializada con sede en Buenos Aires, Argentina. Nuestra misión es mejorar la sociedad en la que vivimos a través de la consultoría y, para conseguirlo, estudiamos, diseñamos, ejecutamos y evaluamos proyectos y acciones que pueden mejorar la sociedad tal y como la conocemos. Este objetivo se logra ya sea por proyectos promovidos por el sector público (ONG, gobiernos, etc.) como así también del sector privado.

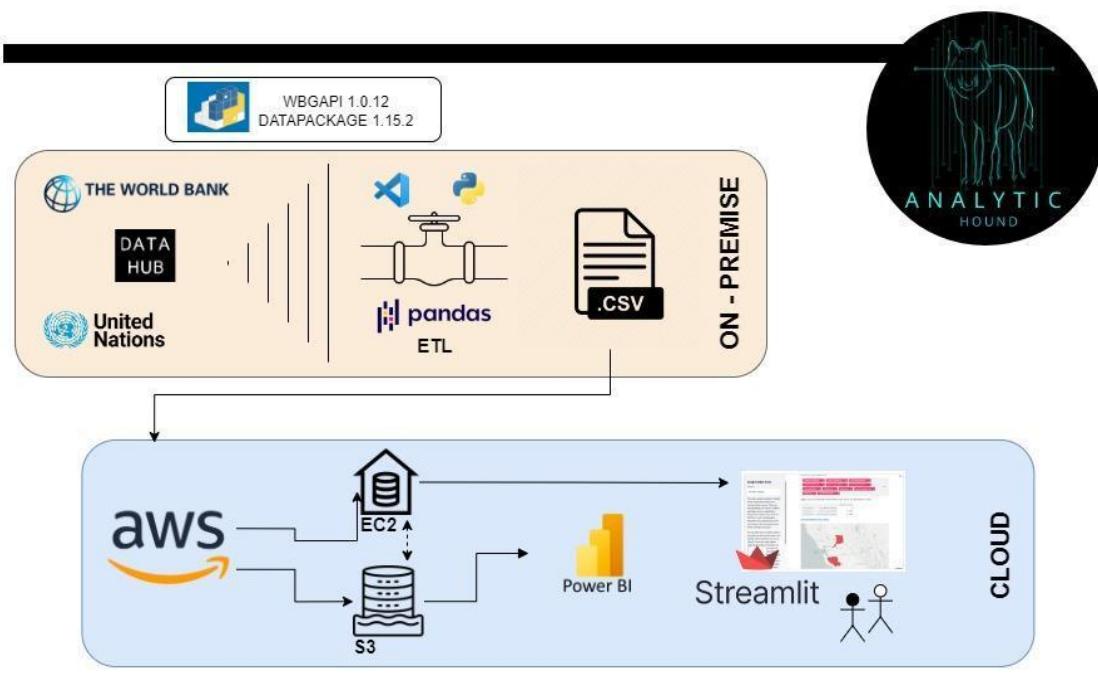


## RESUMEN:

A lo largo de la historia, los flujos migratorios a escala mundial han sufrido múltiples cambios. Estos no son más que la traducción de las fluctuaciones vinculadas a los diferentes aspectos sociopolíticos, demográficos, ambientales y económicos entre otros. El equipo consultor de **Analytic Hound** ®, con el aporte de datos proporcionados por **Data Hub**, **World bank group** y **United nations**, ha desarrollado una plataforma interactiva para el análisis dinámico del movimiento humano alrededor del mundo a partir del año 2000.

El siguiente gráfico refleja a modo esquemático la síntesis de la totalidad del flujo de trabajo vinculado al proyecto “**análisis de patrones migratorios a nivel mundial**”. Entiéndase al mismo como un mapa de referencia, el cual es posible consultar a lo largo de toda esta documentación. Su fin es acercar al lector una hoja de ruta que facilite la comprensión del correcto orden y ensamblaje de los distintos engranajes que entran en juego a la hora de su desarrollo.

El detalle y descripción de cada etapa será abordada en el apartado correspondiente a cada una de ellas. Se invita al lector a hacer uso del índice de este instrumento para ubicar de manera expedita las dudas puntuales que pudieran acontecer.



## ÍNDICE

### **1. Introducción**

- 1.1 Metodología de trabajo
- 1.2. Entendimiento de la situación actual
- 1.3. Objetivos
- 1.4. Alcance y fuera de alcance
- 1.5. KPIs (Key Performance Indicators)
- 1.6. Stack tecnológico

### **2. Procesamiento de datos**

- 2.1. Fuentes
- 2.2. Carga incremental
- 2.3. Informe y pipeline | ETL Completo
- 2.4. Diccionario de variables (features)
- 2.5. Estructura de datos (Data Warehouse, Data Lake)
- 2.6. Implementación
- 2.7. Solución de automatización del DW

### **3. Aplicación de aprendizaje automático**

- 3.1. Machine Learning - Regresión lineal
- 3.2. Hipótesis de trabajo
- 3.3. Selección de features
- 3.4. Implementación

### **4. Visualizaciones**

- 4.1. Power BI
- 4.2. Streamlit

### **5. Insights y Conclusiones**

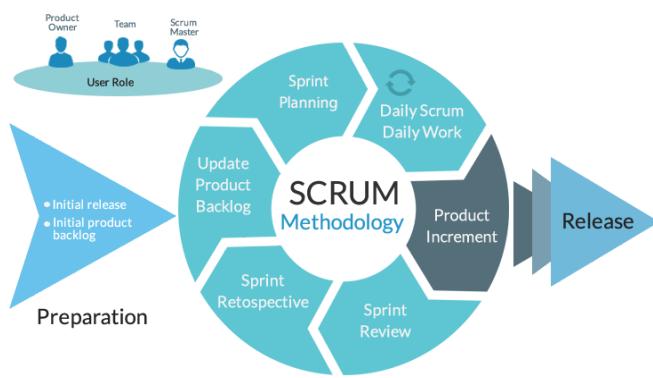
ANEXOS:

- 1. Repositorio
- 2. Fuentes bibliográficas
- 3. Agradecimientos

# 1. INTRODUCCIÓN

## 1.1 METODOLOGÍA DE TRABAJO

El equipo de ***Analytics Hound*** ® cuenta con capacitación y entrenamiento en la metodología de trabajo conocida popularmente como **SCRUM**. La utilización de las herramientas que hacen a esta modalidad servirán de marco de trabajo para coordinar y agilizar en forma efectiva los distintos aspectos del desarrollo de este proyecto.

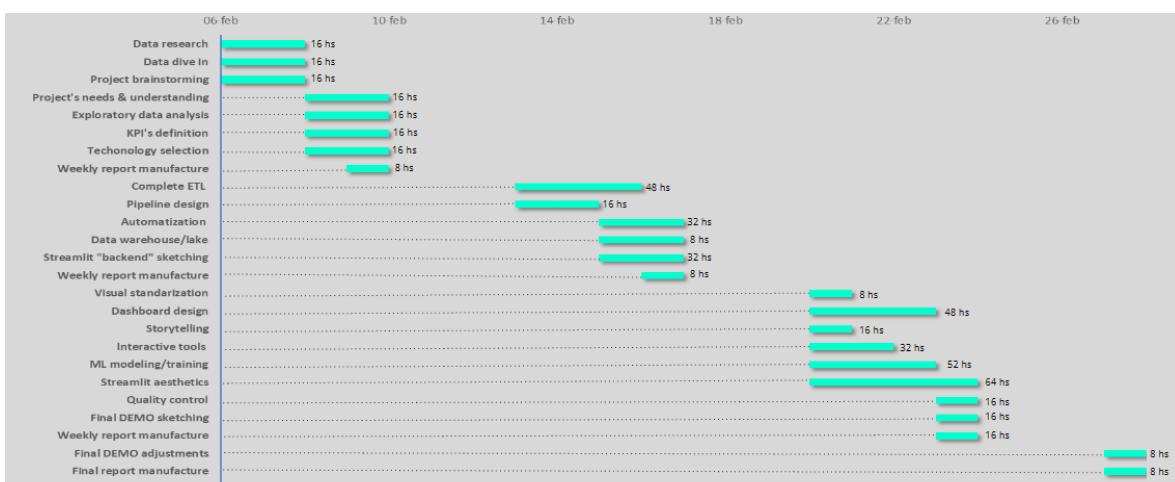


Primariamente, el foco estará puesto en:

- ❖ **Reunión de planificación tipo Sprint:** Al comienzo de cada *Sprint* semanal con todo el equipo, se realizará una inspección minuciosa de la lista de entregables prioritarios (tareas, requisitos y funcionalidades que requiere el proyecto para esa semana específica), seleccionando los elementos correspondientes sobre los cuales se requerirá trabajar durante el *Sprint* semanal.
- ❖ **Scrum diario:** Diaria y matutinamente, previo al encuentro con nuestro asesor externo *Gonzalo Posse*, el equipo se reunirá en este espacio. Se destinará un máximo de 15 minutos a realizar un seguimiento del proyecto en donde se controlará el cumplimiento de las tareas asumidas. Además, se acordarán los objetivos de trabajo diario y se analizarán los posibles problemas que directamente puedan haber limitado o impedido el cumplimiento de los objetivos.
- ❖ **Revisión tipo Sprint:** Al finalizar el *Sprint* semanal, posterior a la reunión sostenida junto al propietario del proyecto (*Product owner*) *Jonathan Deiloff*, el equipo se

reunirá para evaluar retrospectivamente el desempeño. El objetivo de esta etapa consiste en realizar un análisis del cumplimiento del *Sprint* semanal e identificar oportunidades de mejora para el siguiente.

Para obtener una idea global de la distribución y organización de tiempo y esfuerzo de trabajo de las diferentes tareas, se desarrolló un **gráfico de gantt**. El requerimiento de este último fue explicitado por el cliente para lograr un correcto entendimiento de la gestión de tiempos y tareas vinculadas a la totalidad del proyecto.



## 1.2 ENTENDIMIENTO DE LA SITUACIÓN ACTUAL

En esta oportunidad, el informe solicitado por nuestro cliente abarca la temática relacionada a los Flujos Migratorios (FM). Los FM hacen referencia al número de migrantes que llegan a una zona o parten de una zona en el transcurso de un período específico de tiempo. Dichas zonas pueden ser países, regiones, ciudades o continentes. El término surge en reemplazo a otros términos que poseen connotación más despectiva, tales como “masa” u “oleada” de inmigrantes.

Los fenómenos de la migración y la movilidad humana son de característica innata al ser humano, de índole milenaria y han afectado casi sin excepción a todas las sociedades del mundo. La migración es un tema complejo, y como tal puede ser distorsionado en grados alarmantes por la desinformación, información errónea y la politización.

El entendimiento de los FM desde la perspectiva que nos provee la ciencia de Datos será el criterio primario de nuestro enfoque, sin caer en sesgos ideológicos ni en opiniones de índole personal <sup>1</sup>.

Confiamos en que explorar y exponer de forma clara, precisa y científica los FM tiene una significativa utilidad para la sociedad en general, las empresas, los policy makers y para todos los agentes vinculables que, de forma directa o indirecta, se ven afectados por este fenómeno.

En una investigación preliminar hemos podido ahondar en los distintos aspectos que afectan la dinámica de los FM. La lista envuelve variables que van desde los distintos lugares, épocas, políticas de los países, aspectos demográficos, sociopolíticos, ambientales y económicos, entre otros.

Nuestra tarea en *Analytic Hound* ® será lograr exponer en detalle el impacto y valor de cada uno de estos aspectos (features), así como su relevancia en relación a la dinámica de los FM.

### **1.3 OBJETIVOS**

En primera instancia insistiremos en explorar las distintas fuentes de datos brindadas por el cliente, así como aquellas que puedan surgir de nuestra investigación preliminar. Con ellas intentaremos revelar a aquellas variables(features) determinantes vinculadas al impacto sobre los FM, así como el correcto entendimiento de las mismas. Alcanzado este punto, procederemos a realizar un análisis de flujos por zonas y continentes, para comprender a grandes rasgos la escala, dirección y frecuencia de los movimientos.

Posteriormente realizaremos un tamizaje del conjunto de datos a nivel países, y haremos un exhaustivo estudio de los rasgos característicos de ellos, clasificándolos en diversas categorías.

---

<sup>1</sup> Para más información consulte el apartado “disclaimer”.

Toda esta información será volcada en una base de datos (BD) creado por nuestro equipo de ingenieros, la cual contendrá en forma tabular las distintas *features* involucradas en este estudio y su relación con los distintos países en función del tiempo. Dicha BD tendrá la característica de actualizar la información de forma automática<sup>2</sup> de forma tal de que la estructura del proyecto tenga una aplicación escalable y nuestra BD se encuentre alineada a las variaciones que las distintas fuentes utilizadas puedan sufrir en el futuro.

Accesoriamente y a modo de predicción, tal cual nos fuera solicitado por el cliente, realizaremos profundas tareas de analítica y aplicaremos modelos de Machine Learning (ML) tratando de responder, entre otras, a las siguientes preguntas:

<i>¿Qué hace que un emigrante se vaya de un país?</i>
<i>¿Cuáles son las características predominantes de los países más elegidos por los inmigrantes?</i>
<i>¿Qué consecuencias generan FM en ambos tipos de países?</i>
<i>¿Cómo podemos medir ese efecto?</i>
<i>¿Cuáles son los rasgos característicos de los países con FM netos positivos y negativos?</i>
<i>¿Qué hace que un país tenga FM netos cercanos a cero?</i>

Por otro lado, desarrollaremos una interfaz tecnológica que permita acceder de forma remota no sólo a la BD, sino también a métricas, estadísticas e indicadores puntuales que persigamos a raíz de la investigación y compresión de los datos.

En última instancia, se confeccionará un informe detallado y completo vinculado al desarrollo de este proyecto. El mismo contará con los distintos descubrimientos por parte del equipo de analistas, con las respectivas conclusiones e insights que los datos pudieran proveer.

#### **1.4 ALCANCE Y FUERA DE ALCANCE:**

En concordancia con la fecha de entrega pactada con el cliente para el 28 de febrero de 2023, se plantea los siguientes alcances del proyecto:

---

<sup>2</sup> Para más información consulte el apartado “carga incremental de datos”

- ❖ Informe completo del análisis, desarrollo y entregables del trabajo realizado.
- ❖ Link de acceso a repositorio privado con la totalidad de los archivos y documentos vinculados al proyecto.
- ❖ KPIs (Key performance indicators) propuestos por el equipo de análisis con su respectiva explicación, información de uso e interpretación
- ❖ Dashboard Interactivo con diferentes esquemas y filtros que faciliten la correcta comprensión de los datos a través de métricas, estadísticas y KPIs
- ❖ Modelo completo de Machine learning (Diseño, entrenamiento y puesta en acción) con acceso al código e instructivo del proceso.
- ❖ Presentación final al cliente (01 de marzo de 2023) junto a todo el equipo de Analytic Hound ® . Muestra integral y guiada por los profesionales involucrados. Se trata de una presentación “step by step” de todo el proceso y la exposición del producto final.

Se deja constancia que, en previo acuerdo con el cliente, el desarrollo y gestión de este proyecto no contempla los siguientes:

- ❖ Previsión rigurosa de los flujos futuros de FM. Bajo ninguna circunstancia el equipo de Analytic Hound ® pretende generar predicciones que sean utilizadas para la toma de decisión en los distintos ámbitos de aplicabilidad.
- ❖ Modificaciones en los requerimientos realizados más allá de la segunda semana de trabajo.
- ❖ Presentar información a refugiados provenientes de distintos países, solicitantes de asilo y desplazados por conflictos políticos y bélicos respecto de cursos de acción ante su situación personal. El equipo de Analytic Hound ® aprovecha esta instancia para manifestar su apoyo a los derechos humanos y la necesidad de su extensión a todas las personas independientemente de ideales, religiones u orientaciones de cualquier índole.

## **1.5 KPIS (KEY PERFORMANCE INDICATORS)**

Los indicadores clave de rendimiento o, más comúnmente conocidos **KPIs**, son indicadores clave de rendimiento que se utilizan para evaluar el éxito de las acciones y/o procesos en la medida en que estos contribuyen a la obtención de los objetivos, para determinar en función de estos si las acciones dan los frutos esperados o si en cambio es necesario realizar correcciones.

La importancia de los KPIs radica en su característica clave de permitir medir en tiempo real el funcionamiento de las estrategias de negocios, marketing o ventas, brindando una información valiosa para la toma de decisiones estratégicas.

Los KPIs también desempeñan una importante función comunicativa ya que informan a directivos, empleados e inversores sobre la evolución de la empresa respecto a los objetivos establecidos, para que todos puedan trabajar con una visión y una meta común.

A partir de la exploración en profundidad de las distintas variables y datos obtenidos durante la primera etapa del proyecto, el equipo de analistas propone la formulación de los KPIs aquí descritos.

La metodología utilizada para la definición de los mismos, queda representada por el modelo de trabajo reconocido como objetivos **SMART**. El acrónimo hace referencia a “**S**pecific” (específico), “**M**easurable” (mensurable), “**A**ttainable” (alcanzable), “**R**elevant” (relevante) y “**T**ime-bound” (con límite de tiempo).

En los siguientes párrafos nos dedicaremos a la explicación y análisis de los KPI seleccionados. Además, proporcionamos el diseño del código de cada KPI en lenguaje de python.

---

### **KPI 1**

**Expresa la incidencia del flujo migratorio sobre la población total del país elegido.**

*Ej. Un valor de 0,2 implica que el flujo migratorio contribuyó en un 0,2% al aumento de la población. En cambio, números negativos como -0,6 implican que el flujo migratorio contribuyó a una disminución del 0,6% de la población total.*

---

### **KPI 2**

**Representa la variación del flujo neto de migrantes respecto al año previo.**

- ❖ Valores **positivos**: Más personas inmigran o menos emigran
- ❖ Valores **negativos**: Más personas emigran o menos inmigran
- ❖ **Cero**: no se registran cambios respecto al flujo anual (Ej. Migración neta se mantiene igual respecto al año previo)

```
# Group the data by country
grouped = df.groupby('Country Name')

# Calculate the difference in 'Net Number of Migrants (thousands)' between consecutive years for each country
diff = grouped['Net Number of Migrants (thousands)'].diff()

# Add the difference as a new column to the original DataFrame
df['KPI_2'] = round(diff,2)

# Replace the NaN values with 0 (since the first row for each country will have a NaN value)
df['KPI_2'].fillna(0, inplace=True)
```

---

### **KPI 3**

**Expresa la variación porcentual del flujo neto de migrantes respecto al año anterior.**

*Ej. Un valor **positivo** de 20.36 implica un aumento del 20.36% en la migración de ese año respecto al previo.*

```
# Create a new column 'KPI_3'  
  
df['KPI_3'] = round((np.sign(df['KPI_2']) * abs(df['KPI_2']) / df['Net Number of Migrants (thousands)'].shift(1) * 100), 2)  
  
# Replace the NaN values with 0 (since the first row for each country will have a NaN value)  
  
df['KPI_3'].fillna(0, inplace=True)
```

---

Los siguientes KPIs toman para su medición cierta features específicas que surgen como resultado de la implementación de los modelos ML<sup>3</sup>. Para la construcción de los KPIs 4 y 5, se utilizarán las "m = 5" características más importantes. De esta forma, las autoridades y los hacedores de políticas sabrán de manera concisa a qué variables dirigir sus acciones para maximizar el resultado sobre la variable target.

Algunas de las características de nuestra regresión presentan *coeficientes negativos*, lo cual implica una relación inversa con la variable target. De esta manera, a medida que aumentan los valores de dichas features, la variable objetivo disminuye, y viceversa. Esta situación se ha tenido en cuenta, y puede ser manejada de formas diferentes. Para tratar con esta problemática, hemos elegido un criterio que proporciona al cliente una interpretación final simple e intuitiva del KPI, sin necesidad de explicar en detalle ciertos tecnicismos.

Existen múltiples consideraciones en relación a la especificidad del desarrollo de un KPI, y muchas métricas aún más específicas podrían ser pensadas basándose en las futuras necesidades del cliente. Este KPI proporciona un enfoque global y general, cuya intención es mostrar al cliente que la información se puede gestionar y procesar de diferentes formas.

Estos KPIs se adaptan con precisión a las características de cada grupo. Su utilidad y funcionalidad se maximiza mediante el seguimiento de los pasos indicados a continuación.

---

<sup>3</sup> Para más información, dirigirse al apartado 3.3 “Selección de features”

El KPI 4 será más útil para países que tengan características similares a los países del Grupo 1 (Flujo Migratorio Neto grande y positivo).

**KPI 5** resulta más útil para países pertenecientes al grupo 5 (flujo negativo de inmigrantes)

Más adelante se indicará el método utilizado para obtener de cada país una estimación puntual de a qué grupo pertenece.

**KPI 4:**

Este índice es un conjunto de las características consideradas de mayor relevancia según nuestro modelo ML para explicar la recepción de migrantes en los países y años que reciben la mayor cantidad de migración neta.

Este KPI es la media ponderada de los "m" indicadores elegidos, pero normalizados. Hacemos esta transformación para evitar el sesgo de magnitud, mejorar la interpretación del KPI, mejorar el rendimiento del modelo y disminuir el impacto de los valores atípicos en los datos.

Este KPI proporciona una estimación de la medida del atractivo del país para recibir migrantes. Los niveles más altos de KPI 4 están relacionados con mayores flujos migratorios positivos.

Los features más importantes que arrojó el modelo de ML para los países del Group 1 son:

- ❖ Exports of goods and services (PCT of GDP)
- ❖ Gross Domestic Product
- ❖ Refugee population by country or territory of asylum
- ❖ Refugee population by country or territory of origin
- ❖ Infant Mortality Rate (infant deaths per 1,000 live births)

El código para este KPI es el siguiente:

```
# List of the selected features
SK_4 = ['Exports of goods and services (PCT of GDP)',
        'Gross Domestic Product',
        'Refugee population by country or territory of asylum',
        'Refugee population by country or territory of origin',
        'Infant Mortality Rate (infant deaths per 1,000 live births)']

# Standardize the columns
df[SK_4] = (df[SK_4] - df[SK_4].mean()) / df[SK_4].std()

# Assign a negative sign for the features with negative coefficient
df['Exports of goods and services (PCT of GDP)'] = -df['Exports of goods and services (PCT of GDP)']
df['Infant Mortality Rate (infant deaths per 1,000 live births)'] = -df['Infant Mortality Rate (infant deaths per 1,000 live births)']

# Create the KPI column
df['KPI_4'] = df[SK_4].sum(axis=1)/5
```

Más allá de la utilidad del KPI 4, el desempeño y la efectividad de las políticas implementadas en un determinado país también pueden medirse evaluando el desempeño del indicador en varios períodos y su evolución en el tiempo.

De esta forma, VAR\_KPI\_4 nos muestra la variación porcentual de KPI 4 entre el periodo n y el periodo n-1. Proporciona una estimación efectiva de la variación en el grado de deseabilidad del país para recibir FM entre el período actual y el siguiente.

En términos prácticos, es un indicador rápido y fácil de entender que nos permite evaluar de un vistazo el grado de efectividad de las políticas implementadas.

Ej: VAR\_KPI\_4 = 0.12 implica que el incremento porcentual interanual de KPI\_4 ha sido del 12%. Este KPI que otorga la data debe ser comparado con los objetivos que se determinen por parte de los hacedores de política.

```
#Formula of VAR_KPI_4

df['VAR_KPI_4'] = round(df['KPI_4'].pct_change()*100,2)

# Replace the NaN values with 0 (since the first row for each country will have a NaN value)

df['VAR_KPI_4'].fillna(0, inplace=True)
```

### **KPI 5:**

La lógica es similar a la de KPI 4, excepto que esta canasta mide las características que más explican los flujos migratorios en el país donde la migración neta es negativa, particularmente aquellos registros que pertenecen al grupo 5.

Como antes, es deseable un KPI creciente a lo largo del tiempo porque indica que las condiciones adversas son menores y la migración neta está aumentando (o tiene un número menos negativo).

De acuerdo con nuestro modelo ML, las características principales para el Group 5 son:

- ❖ Imports of goods and services (PCT of GDP)
- ❖ Population Density
- ❖ Refugee population by country or territory of asylum
- ❖ Labour force, total
- ❖ Infant Deaths, under age 1 (thousands)

El código correspondiente a este KPI es el siguiente:

```
# List of the selected features
SK_5 = ['Imports of goods and services (PCT of GDP)',
        'Population Density',
        'Refugee population by country or territory of asylum',
        'Labour force, total',
        'Infant Deaths, under age 1 (thousands)']

# Standardize the columns
df[SK_5] = (df[SK_5] - df[SK_5].mean()) / df[SK_5].std()

# Assign a negative sign for the features with negative coefficient
df['Imports of goods and services (PCT of GDP)'] = -df['Imports of goods and services (PCT of GDP)']

# Create the KPI column
df['KPI_5'] = -df[SK_5].sum(axis=1)/5
```

Como vimos en el apartado anterior, en este caso presentamos VAR\_KPI\_5, que nos muestra la variación porcentual de KPI 5 entre el periodo n y el periodo n-1.

Ej. VAR\_KPI\_5 = 0,08 implica que el objetivo a conseguir es que haya un incremento porcentual del 8% interanual del KPI 5. El cliente puede comparar el valor del KPI que quiere conseguir con los datos que le proporciona la tabla.

```
# Formula of VAR_KPI_4  
  
df['VAR_KPI_5'] = round(df['KPI_5'].pct_change()*100,2)  
  
# Replace the NaN values with 0 (since the first row for each country will have a NaN value)  
  
df['VAR_KPI_5'].fillna(0, inplace=True)
```

## 1.6 STACK TECNOLÓGICO



**Python:** Lenguaje de alto nivel de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código, se utiliza para desarrollar aplicaciones de todo tipo. Lenguaje de programación multiparadigma. Soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma. Administrado por Python Software Foundation, posee una licencia de código abierto, denominada Python Software Foundation License. Se clasifica constantemente como uno de los lenguajes de programación más populares.

**Visual studio code (VSC):** Editor de código fuente desarrollado por Microsoft para Windows, Linux, macOS y Web. Incluye soporte para la depuración, control integrado de Git, resultado de sintaxis, finalización inteligente de código, entre otras funciones idóneas para la construcción de este proyecto. Librerías dentro de VSC:

- ❖ **Pandas:** librería escrita para el lenguaje de programación de python diseñada con el fin de permitir la manipulación y análisis de datos. Ofrece estructuras de datos y operaciones para la manipulación de tablas numéricas y series temporales.
- ❖ Las librerías **WBGAPI** 1.0.12 y **DATAPACKAGE** 1.15.2 se describen en el apartado “Carga incremental de datos”.

**Amazon Web Services (AWS):** Colección de servicios de computación en la nube pública (también llamados servicios web) que en conjunto forman una plataforma de computación en la nube, ofrecidas a través de Internet por Amazon.com.

**Microsoft Power BI:** Software de visualización de datos interactivo desarrollado por Microsoft con un enfoque principal en la inteligencia empresarial .

**Streamlit:** Biblioteca Python de código abierto que facilita la creación de aplicaciones web personalizadas para el aprendizaje automático y la ciencia de datos

## 2. PROCESAMIENTO DE DATOS

### 2.1. FUENTES

Durante la instancia de búsqueda de distintas fuentes datos que aportaran a la elaboración de este proyecto, se realizó en forma preliminar una minuciosa examinación de aquellos datos propuestos por el cliente. Frente al hallazgo de datos de mayor calidad, variedad y frecuencia temporal en comparación con los indicados por el cliente, el equipo de *Analytic Hound* ® optó por la utilización de sus propias fuentes. Las mismas se enlistan a continuación con los respectivos links enlazados a la página oficial de cada una de ellas:

- ❖ **World Bank:** <https://data.worldbank.org/>
- ❖ **United Nations:** <https://www.un.org/development/desa/pd/data-landing-page>
- ❖ **Accesorio de data de UN:** <https://data.un.org/>



### 2.2. CARGA INCREMENTAL

El aplicativo de la carga incremental de datos hace referencia a la actualización de los mismos en la medida que las fuentes principales sean modificadas a lo largo del tiempo.

En el contexto del análisis migratorio del proyecto en cuestión, dicha actualización es esperable que suceda con una periodicidad promedio de tipo semestral/anual.

Para la correcta implementación de la carga incremental, se remitió a la documentación oficial destinadas a desarrolladores dentro de las bases de datos seleccionadas. Con el soporte de las mismas, se establece el pipeline correspondiente a dicha tarea.

Como puede observarse en los links adjuntos, las siguientes librerías de python fueron utilizadas para la extracción directa de datos de las distintas fuentes:

- ❖ **WBGAPI 1.0.12** | <https://pypi.org/project/wbgapi/>
- ❖ **DATAPACKAGE 1.15.2<sup>4</sup>** | <https://pypi.org/project/datapackage/>

En resumen, el código utilizado para esta tarea presenta el siguiente orden de ejecución:

1. Importación de librerías
2. Extracción de datos en forma directa mediante librerías de python.
3. Guardado de las mismas en variables definidas que luego son materia prima del proceso de ETL que sucederá inmediatamente luego de finalizada la carga.

### **2.3. INFORME Y PIPELINE | ETL COMPLETO**

El análisis e informe correspondiente al proceso de extracción, transformación y carga (ETL) de datos con su respectivo pipeline puede ser consultado directamente en el archivo “*ETL.ipynb*”, alojado en el github oficial. Se facilita el acceso al mismo mediante el siguiente link de carácter público:

<https://github.com/Analytic-Hound-Consulting/ONG-Henry>

### **ORIGEN DE DATOS**

Los datos que se utilizan en este proyecto provienen de múltiples archivos CSV que contienen información sobre diversas variables de todos los países y regiones. Dichas variables son estudiadas según su influencia en la migración positiva y negativa de cada país. Los datos fueron extraídos de sitios web mundiales, los cuales ponen a disposición dicha información.

### **PROCESO DE TRANSFORMACIÓN**

Una vez que se importaron los datos del archivo CSV, se realizaron varias transformaciones para prepararlos para su uso. Estas incluyeron:

---

<sup>4</sup> Nota: datahub presenta un funcionamiento con datos de tipo “paquete” permitiendo su manejo con la librería mencionada.

<i>Renombrar columnas para mayor claridad.</i>
<i>Completar los datos faltantes con criterio.</i>
<i>Revisar y eliminar filas duplicadas.</i>
<i>Eliminar columnas innecesarias.</i>
<i>Eliminar datos correspondientes a años previos al 2000.</i>
<i>Reorganización de las columnas para uniformidad.</i>

Para llevar a cabo estas transformaciones, se diseñaron dos pipelines que agrupan las diversas transformaciones que se deben aplicar a los distintos datasets. El lenguaje utilizado es Python, y las bibliotecas principales fueron *Pandas*, *wbgapi*, *datapackage*, y *scikit learn*.

## **DESTINO FINAL**

Los datos transformados se exportaron a un archivo CSV para su posterior utilización y evaluación para los modelos de aprendizaje. Todo el pipeline fue entregado al equipo de ingeniería para que puedan levantarla en la nube.

## **CONCLUSIONES**

El proceso de ETL fue exitoso, tanto en la recolección como en la transformación de los datos provenientes de diferentes fuentes. Las transformaciones realizadas permiten que los datos se analicen de manera efectiva para obtener información valiosa sobre el efecto que distintas variables tienen sobre la migración.

## 2.4. DICCIONARIO DE VARIABLES (FEATURES)

CATEGORÍA	VARIABLE	DEFINICIÓN	MEDIDA
MACROECONOMÍA	<b>Producto bruto interno (PBI)</b>	Valor agregado por todos los residentes productores en la economía de cada país, en U\$D	(U\$D)
	<b>Crecimiento del PBI:</b>	Tasa de crecimiento porcentual anual del PBI	(Anual %)
	<b>Gasto final en consumo:</b>	Sumatoria del gasto de consumo final de hogares (consumo privado) y gasto de consumo final gubernamental	(% de PBI)
	<b>PBI per cápita:</b>	Valor agregado por un único producto dentro de la economía de un país	(US\$)
	<b>INB per cápita (Método "atlas"):</b>	Ingreso nacional bruto (conversión a dólares utilizando el método del banco atlas), dividido por la población a mitad del año	(US\$)
	<b>Ahorros (bruto)</b>	Calculado como ingreso nacional bruto (INB) - consumo total.	(% of PBI)
	<b>Índice de precios al consumidor:</b>	Costo del consumidor promedio en adquirir una canasta de bienes y servicios (Este puede ser fijo o cambiar a intervalos específicos, como anualmente)	(2010 = 100)
	<b>Desempleo (total)</b>	Porcentaje de la fuerza laboral que no cuenta con trabajo pero que se encuentra disponible y en la búsqueda de empleo	(% de la fuerza laboral total)
	<b>Fuerza laboral</b>	Personas de 15 años o más que aportan mano de obra para la producción de bienes y servicios durante un período determinado. Incluye a las personas que actualmente están empleadas y las personas que están desempleadas pero buscan trabajo	(Total)

	<b>Reservas (total)</b>	Comprende las tenencias de oro monetario, los derechos especiales de giro, las reservas de los miembros del FMI en poder del FMI y las tenencias de divisas bajo el control de las autoridades monetarias.	(Oro + US\$)
	<b>Exportación de bienes y servicios</b>	Valor de todos los bienes y otros servicios de mercado proporcionados al resto del mundo. (Excluye: compensación de empleados e ingresos por inversiones y pagos de transferencias)	(% of PBI)
	<b>Importación de bienes y servicios</b>	Valor de todos los bienes y otros servicios de mercado recibidos del resto del mundo.	(% of PBI)
<b>MICROECONOMÍA</b>	<b>Inversión extranjera directa, entradas netas</b>	Entradas netas de inversión para adquirir una participación de gestión duradera (10% o más de las acciones con derecho a voto) en una empresa que opera en una economía distinta a la del inversor. Es la suma del capital social, la reinversión de utilidades, otro capital a largo plazo y el capital a corto plazo que se muestra en la balanza de pagos. Esta serie muestra las <i>entradas netas</i> (entradas de nuevas inversiones menos desinversión) en la economía declarante de inversores extranjeros, y se divide por el PIB	(% de PBI)
	<b>Impuesto total y tasa de contribución</b>	Monto de los impuestos y contribuciones obligatorias y exenciones como parte de las utilidades comerciales. Se excluyen los impuestos retenidos (como el impuesto sobre la renta de las personas físicas) o recaudados y remitidos a las autoridades fiscales (como los impuestos sobre el valor añadido, los impuestos sobre las	(% de ganancias)

		ventas o los impuestos sobre bienes y servicios).	
	<b>Tiempo requerido para iniciar un negocio</b>	Número de días calendario necesarios para completar los trámites para operar legalmente un negocio (En caso de poder agilizar el proceso con un costo adicional, se elige el trámite más rápido)	(Dias)
	<b>Gastos en investigación y desarrollo</b>	Gasto bruto en investigación y desarrollo. Incluye el gasto en 4 sectores principales: empresa comercial, gobierno, educación superior y privado sin fines de lucro. Abarca la investigación básica, la investigación aplicada y el desarrollo experimental.	(% de PBI)
<b>DESARROLLO SOCIAL</b>	<b>Gasto público en educación:</b>	Gasto público en educación expresado en porcentaje de PBI	(% de PBI)
	<b>Tasa de finalización de escuela primaria</b>	Número de nuevos ingresos (matrículas menos repetidores) en el último grado de educación primaria, independientemente de la edad, dividido por la población en edad de ingreso al último grado de educación primaria	(% del grupo de edad relevante)
	<b>Homicidios intencionales</b>	Homicidios ilegales infligidos deliberadamente como resultado de disputas domésticas, violencia interpersonal, conflictos violentos por los recursos de la tierra, violencia entre pandillas por el territorio o el control, y violencia depredadora y asesinatos por parte de grupos armados.	(Por cada 100,000 personas)
	<b>Acceso a electricidad</b>	Porcentaje de población con acceso a electricidad. Los datos de electrificación se recopilan de la	(% de población)

		industria, encuestas nacionales y fuentes internacionales.	
	<b>Personas que utilizan al menos servicios básicos de saneamiento en áreas urbanas:</b>	Instalaciones de saneamiento mejoradas que no se comparten con otros hogares. Este indicador abarca tanto a las personas que utilizan servicios básicos de saneamiento como a las que utilizan servicios de saneamiento gestionados de forma segura.	(% de población)
	<b>Suscripciones celulares móviles:</b>	Suscripciones a un servicio público de telefonía móvil que dan acceso a la <i>red telefónica pública conmutada</i> mediante tecnología celular. El indicador incluye (y se divide en) el número de suscripciones de pospago y el número de cuentas de prepago activas. El indicador se aplica a todas las suscripciones celulares móviles que ofrecen comunicaciones de voz.	(Por cada 100 personas)
<b>DEMOGRAFÍA</b>	<b>Población total:</b>	Indicador basado en la definición de “población”, que considera a todos los residentes independientemente de su estatus legal o ciudadanía.	(Total)
	<b>Densidad poblacional:</b>	Cantidad de personas por metro cuadrado de superficie	(Personas por metro cuadrado de superficie)
	<b>Transición demográfica:</b>	Calculado como: Nacimientos - muertes	(Miles)
	<b>Tasa de transición demográfica:</b>	Tasa de natalidad menos la tasa de mortalidad de una población particular, durante un período de tiempo particular	(Por cada 1,000 personas)
	<b>Tasa de crecimiento poblacional:</b>	Tasa exponencial de crecimiento de la población a mitad de año desde el año $t-1$ hasta el $t$ . (Ver definición de “Población total”)	(% Anual)

<b>SALUD</b>	<b>Tasa bruta de natalidad</b>	Número de nacidos vivos ocurridos durante el año.	(Por cada 1,000 personas)
	<b>Edad media (al 1 de julio)</b>	Edad que divide a la población en dos partes de igual tamaño, es decir, hay tantas personas con edades por encima de la mediana como con edades por debajo de la mediana.	(Años)
	<b>Esperanza de vida al nacer (ambos sexos)</b>	Número de años que viviría un recién nacido si los patrones de mortalidad prevalecientes en el momento de su nacimiento se mantuvieran iguales a lo largo de su vida	(Años)
	<b>Tasa de mortalidad infantil</b>	Número de muertes infantiles por cada 1.000 nacidos vivos	(Por cada 1,000 nacidos vivos)
	<b>Muertes infantiles (menores de 1 año)</b>	Muerte de un bebé antes de su primer cumpleaños	(Miles)
	<b>Tasa de mortalidad materna:</b>	Número de mujeres que mueren por causas relacionadas con el embarazo durante el embarazo o dentro de los 42 días posteriores a la terminación del embarazo. (Datos se estiman con un modelo de regresión utilizando información sobre la proporción de muertes maternas entre las muertes no relacionadas con el SIDA en mujeres de 15 a 49 años, la fecundidad, las parteras y el PBI)	(Por cada 100.000 nacidos vivos.)
<b>MIGRACIÓN</b>	<b>Número neto de migrantes</b>	Total neto de migrantes durante un período determinado. Calculado como número de inmigrantes menos el número de emigrantes (ciudadanos como no ciudadanos)	(Miles)
	<b>Tasa neta de migración</b>	Diferencia entre el número de inmigrantes y el número de	(Por cada 1,000 personas)

		emigrantes (personas que salen de un área) a lo largo del año	
	<b>Población de refugiados por país o territorio de origen</b>	Los refugiados son personas reconocidas como refugiados en virtud de la Convención sobre el Estatuto de los Refugiados de 1951 o su Protocolo de 1967, la Convención de la Organización de la Unidad Africana de 1969 que rige los aspectos específicos de los problemas de los refugiados en África, personas reconocidas como refugiados de conformidad con el estatuto del UNHCR ( <i>United Nations High Commissioner for Refugees</i> ), o personas a las cuales se les otorgó el estatus humanitario de refugiado y se les proporcionó protección temporal. Los solicitantes de asilo, personas que han solicitado asilo o estatus de refugiado y que aún no han recibido una decisión o que están registrados como solicitantes de asilo, están excluidos. Los refugiados palestinos (y sus descendientes) son personas cuya residencia fue Palestina entre junio de 1946 y mayo de 1948 y que perdieron sus hogares y medios de vida como resultado del conflicto árabe-israelí de 1948. El país de asilo es el país donde se presentó y concedió una solicitud de asilo.	(Anual, sumatoria)
	<b>Población de refugiados por país o territorio de asilo</b>	La definición de refugiado se encuentra en el cuadro superior. Se entiende por país de asilo a aquel donde se presentó y concedió una solicitud de asilo.	(Anual, sumatoria)

## 2.5. ESTRUCTURA DE DATOS (DW, DL)

En lineamiento con los requerimientos del cliente, y optando por la alternativa más eficaz, el departamento de engineering determinó montar un **Data Lake** (DL) y un **Data Warehouse** (DW) en los servicios Cloud de *Amazon Web Services*.

Las principales razones sobre las cuales se sustenta esta elección se relacionan al hecho de que en la misma plataforma de AWS se ofrecen ambos servicios de DL (llamado **S3**) y DW (**EC2**)

En primer lugar, resulta primordial entender que ambos conceptos (*DL* y *DW*) se refieren a herramientas utilizadas para almacenar grandes cantidades de datos de manera organizada. Sin embargo, existen algunas diferencias clave entre ellos.

Un **Data Lake** es un repositorio de datos que se utiliza para almacenar grandes cantidades de información, *sin importar su formato o estructura*. En un Data Lake, se pueden almacenar datos estructurados, semiestructurados y no estructurados, lo que permite a los analistas explorar y analizar los datos en su forma original.

Por otro lado, un **Data Warehouse** es un repositorio de datos que se utiliza para almacenar datos *estructurados y procesados*, con el objetivo de facilitar su análisis y consulta. Los datos almacenados en un Data Warehouse se han organizado y estructurado previamente para que puedan ser utilizados de manera más eficiente.

Ahora bien, es importante tener en cuenta que ambas herramientas pueden ser utilizadas de manera *complementaria*. Es decir, se puede utilizar un Data Lake para almacenar los datos en su forma original y luego procesarlos y transformarlos para cargarlos en un Data Warehouse, donde se pueden analizar de manera más eficiente. Este último formato de uso en forma complementaria responde a cómo fueron utilizados los mismos para el desarrollo de este proyecto en particular.

Algunas de las ventajas de utilizar un **Data Lake** son:

- ❖ *Almacenamiento de datos en su forma original:* Permite almacenar grandes cantidades de datos sin tener que estructurarlos previamente, lo que puede ser útil para el análisis exploratorio y la minería de datos.
- ❖ *Escalabilidad:* Los Data Lakes son altamente escalables y pueden crecer fácilmente a medida que se agregan más datos.
- ❖ *Flexibilidad:* Permite almacenar datos de diferentes formatos y estructuras, lo que lo hace ideal para almacenar datos no estructurados o semiestructurados.

Por otro lado, algunas de las ventajas de utilizar un **Data Warehouse** son:

- ❖ *Facilidad de análisis:* Los datos se estructuran previamente para que puedan ser analizados y consultados de manera más eficiente.
- ❖ *Integración con herramientas de análisis:* Los Data Warehouses suelen integrarse con herramientas de análisis de datos, lo que permite a los usuarios analizar los datos de manera más eficiente.
- ❖ *Mayor seguridad:* Los datos en un Data Warehouse suelen estar más seguros que en un Data Lake, ya que se han tomado medidas para garantizar la integridad y confidencialidad de los datos.

En resumen, tanto el *DL* como el *DW* son herramientas importantes para el almacenamiento y análisis de datos. La elección de cuál y cómo utilizarlos dependerá de las necesidades específicas de cada proyecto en particular y los objetivos del análisis de datos.

Respecto al **Amazon Elastic Compute Cloud (EC2)**, el mismo es un servicio de AWS que proporciona una capacidad de cómputo escalable en la nube. Permite a los usuarios lanzar y administrar instancias de servidores virtuales en la nube, permitiendo un control completo sobre su entorno de cómputo.

Algunas de las características y beneficios de Amazon **EC2** son:

- ❖ *Escalabilidad:* Una de las principales ventajas de EC2 es que es altamente escalable. Los usuarios pueden aumentar o disminuir la capacidad de cómputo en función de sus necesidades en cualquier momento, lo que les permite ajustar su capacidad de cómputo para satisfacer la demanda de su aplicación o servicio.
- ❖ *Diversidad de instancias:* Amazon EC2 ofrece una amplia variedad de tipos de instancias, desde instancias generales hasta instancias especializadas, cada una diseñada para satisfacer necesidades específicas. Esto permite a los usuarios elegir la instancia que mejor se adapte a sus necesidades de cómputo.
- ❖ *Flexibilidad:* Los usuarios tienen el control total sobre su entorno de cómputo. Pueden seleccionar el sistema operativo, la configuración de la red y la cantidad de capacidad de almacenamiento que necesitan para su instancia.
- ❖ *Pago por uso:* EC2 utiliza un modelo de pago por uso, lo que significa que los usuarios solo pagan por la capacidad de cómputo que utilizan. Esto les permite reducir sus costos y optimizar su presupuesto de TI.
- ❖ *Integración con otros servicios de AWS:* Amazon EC2 se integra con otros servicios de AWS, como Amazon Simple Storage Service (S3), Amazon Relational Database Service (RDS) y Amazon Elastic Load Balancing, lo que permite a los usuarios crear soluciones de TI más completas y complejas.
- ❖ *Seguridad:* Amazon EC2 se integra con otras herramientas de seguridad de AWS, como AWS Identity and Access Management (IAM) y Amazon Virtual Private Cloud (VPC), lo que ayuda a los usuarios a proteger su entorno de cómputo y sus datos.

En resumen, Amazon EC2 es una herramienta poderosa que permite a los usuarios lanzar y administrar servidores virtuales en la nube de manera escalable y flexible. Su modelo de pago por uso, su amplia variedad de tipos de instancias y su integración con otros servicios de AWS lo convierten en una solución ideal para este proyecto gracias a su capacidad de cómputo eficiente y escalable. En este caso, dicho servicio se utilizará para el alojamiento (hosting) de **streamlit**.

Por otra parte, **Amazon Simple Storage Service (S3)** es un servicio de almacenamiento de objetos en la nube de AWS. S3 proporciona una solución de almacenamiento altamente

escalable, segura y duradera que permite a los usuarios almacenar y recuperar grandes cantidades de datos desde cualquier lugar del mundo.

Algunas de las características y beneficios de Amazon S3:

- ❖ *Escalabilidad*: Amazon S3 es altamente escalable y permite a los usuarios almacenar y recuperar grandes cantidades de datos en la nube sin preocuparse por la capacidad de almacenamiento. Los usuarios pueden escalar su capacidad de almacenamiento según sea necesario y pagar solo por la capacidad que usan.
- ❖ *Durabilidad y disponibilidad*: S3 es altamente durable y está diseñado para proporcionar una disponibilidad del 99,99%. Esto significa que los usuarios pueden estar seguros de que sus datos estarán disponibles en todo momento y estarán protegidos contra la pérdida de datos.
- ❖ *Seguridad*: Amazon S3 utiliza cifrado SSL/TLS para proteger los datos en tránsito y cifrado AES-256 para proteger los datos en reposo. Los usuarios pueden configurar políticas de acceso y control de acceso basado en roles para proteger sus datos y mantenerlos seguros.
- ❖ *Flexibilidad*: S3 es compatible con una amplia variedad de tipos de archivos y formatos, lo que lo hace ideal para una variedad de casos de uso, desde almacenamiento de archivos multimedia hasta copias de seguridad de bases de datos y registros de aplicaciones.
- ❖ *Integración*: Amazon S3 se integra con otros servicios de AWS, como Amazon EC2, Amazon Lambda y Amazon CloudFront, lo que permite a los usuarios crear soluciones de TI más completas y complejas.
- ❖ *Analytics*: Amazon S3 ofrece una variedad de herramientas de análisis que permiten a los usuarios comprender mejor sus datos y obtener información valiosa. Por ejemplo, S3 puede utilizarse con Amazon Athena para realizar consultas ad-hoc en datos almacenados en S3.

En resumen, Amazon S3 es un servicio de almacenamiento de objetos altamente escalable, seguro y duradero que proporciona a los usuarios la flexibilidad para almacenar

y recuperar grandes cantidades de datos en la nube. Sus características de durabilidad, disponibilidad y seguridad, su integración con otros servicios de AWS y sus herramientas de análisis lo convierten en una solución ideal para empresas y desarrolladores que buscan una forma segura y escalable de almacenar y analizar grandes cantidades de datos en la nube.

## 2.6 IMPLEMENTACIÓN

El procedimiento establecido para la estructuración del servicio en la nube fue el siguiente.

En primer lugar, se decidió crear una instancia en *EC2* y realizar una carga manual de archivos desde Visual Studio Code. Esto se efectuó con el objetivo de lograr un mejor entendimiento del comportamiento de dicha instancia. Posteriormente se procedió a la creación de un depósito en *S3* para almacenar la información, y realizar luego una prueba manual de carga de archivos desde AWS.

Siendo necesario contar con la movilización de información de uno a otro, se dispuso de un rol para la conexión entre ambos (*EC2* y *S3*). Es importante aclarar que la misma no impide el acceso individual a ninguno de ellos, ni tampoco el hecho de que uno pueda borrar datos sin afectar al otro.

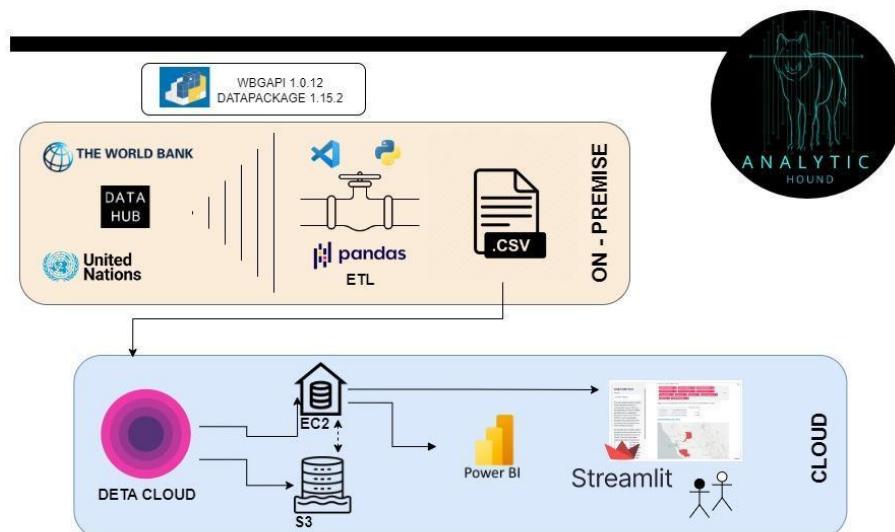
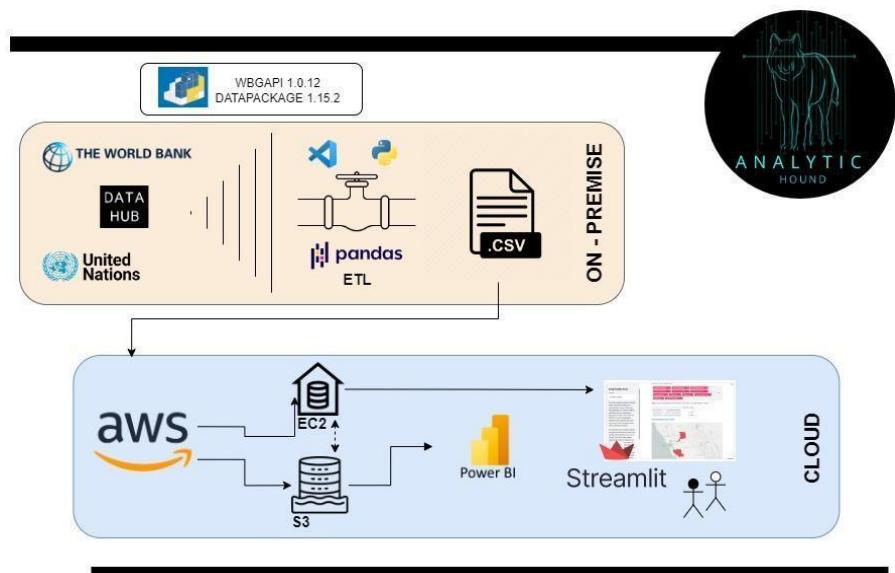
Inicialmente, la idea original se orientó a automatizar el vínculo entre GitHub y el depósito. Esta fue abandonada a favor de una opción más rápida y sencilla. Se realizó la modificación del *pipeline* del ETL para que el mismo, una vez finalizado, ejecute directamente el envío del archivo csv final con todos los datos procesados al depósito, para luego realizar su carga automática al almacén de datos.

Por último, la implementación de *CRON* permitirá la ejecución de un comando responsable de la actualización del archivo continente de los datos sobre los cuales se realiza el análisis de este proyecto. De esta forma contaremos con la totalidad del proceso automatizado, ejecutándose el mismo una vez por mes.

## 2.6. SOLUCIÓN DE AUTOMATIZACIÓN DEL DW

En el siguiente diagrama detallamos el flujo de trabajo que satisface el requerimiento realizado por el *product owner* al final del *Sprint #1*. Dicha modificación de los entregables consistió en considerar un plan alternativo para el cumplimiento de los requisitos en caso de que ante alguna eventualidad no se lograra implementar la opción primaria.

En esta alternativa, decidimos la utilización del servicio en la nube **DETA cloud service** para el almacenamiento de datos y la implementación de Streamlit. Esta versión funciona a modo de “PLAN B”. La misma ya ha sido testeada e implementada para ser utilizada en caso de necesidad.



### 3. APPLICACIÓN DE APRENDIZAJE AUTOMATIZADO

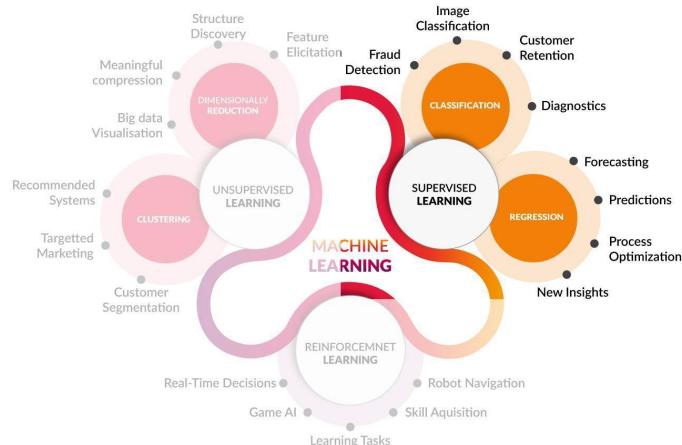
#### 3.1. MACHINE LEARNING - REGRESIÓN LINEAL

Partiendo de la definición técnica, el concepto de **Machine Learning (ML)** se centra en el uso y desarrollo de algoritmos que utilizan datos para imitar la forma en la que los humanos aprenden, mejorando gradualmente su precisión. Considerando que a los fines de este documento sería excesivo e insuficiente el desarrollo de esta temática, se sugiere un material de lectura complementaria para aquellos deseosos de conocer más acerca de este maravilloso mundo.<sup>5</sup>

A modo esquemático y sin intenciones de profundizar en lo que la temática de machine learning abarca, así como sus aplicaciones por las cuales ha logrado revolucionar la forma de desarrollar tecnología en los últimos años, se presenta aquí una breve introducción al modelo utilizado para el proyecto de “*Análisis de patrones migratorios a nivel mundial*” desarrollado por *Analytic Hound* ®

De acuerdo con el tipo de **etiqueta o variable de salida** (comúnmente llamada “y”), los modelos de **aprendizaje supervisados** pueden ser clasificados en:

- ❖ **Clasificación:** en estos modelos la etiqueta es un tipo de categorías (Ej. enfermo/sano, gato/perro/pájaro, spam/no spam)
- ❖ **Regresión:** la variable de salida es un valor numérico. (Ej. precio, cantidad, temperatura)



<sup>5</sup> Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems - <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>

En forma resumida, los modelos de aprendizaje supervisado utilizan las diferentes **variables** o **features** (únicamente tolerados en formato numérico) para determinar los patrones (**algoritmos**) que permiten predecir una variable de salida o etiqueta seleccionada.

Por lo general, es considerado de buena práctica destinar un porcentaje considerable ( $\approx 80\%$ ) de los datos al **entrenamiento** del modelo y luego el restante para el **testeo** del mismo. Todo esto sucede bajo la premisa de “A mayor cantidad de datos de entrenamiento, mejor predicción”.

Finalmente, una vez desarrollado el modelo, el mismo puede ser evaluado mediante diferentes fórmulas de cuantificación del error. Entre ellos se encuentran el **MAE** (error absoluto medio), **MSE** (error cuadrático medio), **RMSE** (raíz cuadrada del error cuadrático medio) y **R cuadrado** (coeficiente de determinación) entre otros.

Habiendo comentado brevemente los pormenores que hacen a los modelos de ML supervisados, en los apartados siguientes nos limitaremos a exponer el modelo desarrollado destinado a este proyecto.

### 3.2. HIPÓTESIS DE TRABAJO

El interés del uso de herramientas de aprendizaje automatizado se encuentra orientado al descubrimiento de aquellas características (*features*) que presentan mayor importancia a la hora de explicar los fenómenos migratorios. Esto sucede con el fin de proporcionar diferentes KPIs a gobernantes y otras entidades que pudieran presentar interés, de forma tal que se utilicen como instrumentos de monitoreo y medición de la situación migratoria dentro de sus fronteras. Asimismo, los KPIs desarrollados resultaran ser poderosas herramientas para el desarrollo de políticas de gobierno y toma de decisiones en la materia.

La **hipótesis** de trabajo se basa en la presunción de que aquellas características que mejor explican los flujos migratorios netos son diferentes según se trate de un país con migración neta positiva o negativa. Con el objetivo de dar respaldo a esta hipótesis, ejecutaremos un

modelo de ML utilizando las características de **SelectKBest** para diferentes registros en función del año y país al cual corresponde el flujo migratorio.

Se realizaron una serie de pruebas para conocer las relaciones entre nuestras variables y comparar los resultados.

En primer lugar, ejecutaremos el módulo **SelectKBest** de **SciKitLearn** para realizar una selección de las “*k*” características principales con las puntuaciones más altas obtenida de una función de puntuación llamada "**f\_regression**". Dicha función realizará el cálculo de la correlación existente entre cada característica y la variable de destino, devolviendo el coeficiente de regresión y un **valor-p** para cada característica. En nuestro abordaje, consideraremos a la variable target como el '**Número neto de migrantes (miles)**'.

*SelectKBest* luego seleccionará las *k* principales características con los *valor-F* más altos, indicando la relación lineal más fuerte vinculada a la variable de destino. Resumidamente, las características seleccionadas serán entonces aquellas con mayor injerencia a la hora de explicar los movimientos en las cifras migratorias.

Para lograr una mejor comprensión acerca del potencial efecto de cada feature sobre el fenómeno de la migración neta, se utilizan los coeficientes del modelo de regresión lineal para dar cuenta de la dirección así como el impacto de cada característica en relación a la variable objetivo.

Se aplica esta metodología a todos los países, para obtener una visión general de esta idea.

En primera instancia, se seleccionan las 5 mejores características para explicar los flujos migratorios a nivel mundial:

- ❖ Population Total
- ❖ Natural Change, Births minus Deaths (thousands)
- ❖ Population Growth Rate (percentage)
- ❖ Net Migration Rate (per 1,000 population)
- ❖ Infant Deaths, under age 1 (thousands)

Estas 5 variables y sus respectivos coeficientes implican que, considerando todos los países, la Migración Neta tiende a explicarse globalmente por dichas características.

Observando las features aquí enlistadas puede resultar intuitivo a simple vista considerar que dichas variables tengan implicancias en los flujos migratorios. Profundizaremos en el análisis más detallado una vez obtenidos los datos completos para comparar, inferir patrones y sacar conclusiones.

Un *valor-p* es una medida estadística de la evidencia contra la hipótesis nula de que el coeficiente de una determinada característica es cero. En forma práctica, un *valor-p* inferior o igual al nivel de significancia del 5 % sugiere que el coeficiente de dicha característica resulta estadísticamente significativo con un nivel de confianza del 95 %.

Esto se traduce en una fuerte evidencia vinculada a la relación entre cada una de las variables y el "Número Neto de Migrantes (miles)" en la determinada población de la cual se extrajo la muestra utilizada. Significaria que existe menos del 5% de probabilidad de que la relación observada entre el predictor y la variable de respuesta se deba al azar.

Según la información obtenida a raíz de las pruebas realizadas, los *valores-p* parecerían ser muy bajos (mucho menores a 0,05). Esto indicaría que los coeficientes correspondientes son estadísticamente significativos con un nivel de confianza del 95 %. Por lo tanto, es posible afirmar con alto grado de certeza que las características presentan un impacto significativo en la variable objetivo.

### 3.3. SELECCIÓN DE FEATURES

La multicolinealidad se refiere a una situación en la cual dos o más variables independientes en un modelo de regresión presentan alto grado de correlación. Como resultado, los coeficientes de las variables del modelo pueden volverse inestables y difíciles de interpretar, y el rendimiento general del modelo puede verse afectado. Este efecto puede ser un problema por varias razones que exceden al desarrollo de esta documentación.

Una forma de verificar la multicolinealidad se basa en la utilización del **factor de inflación de varianza (VIF)** para cada característica. Calcular el *VIF* permite medir cuánto aumenta la varianza del coeficiente de regresión estimado para una característica dada debido a la multicolinealidad con las otras características. Un valor *VIF* de 1 indica ausencia de multicolinealidad, mientras que los valores más altos se relacionan a niveles crecientes.

Realizados los cálculos, las siguientes variables presentan un *VIF*> 5:

- ❖ Access Elect.
- ❖ Crude Birth Rate (births per 1,000 population)
- ❖ Natural Change, Births minus Deaths (thousands)
- ❖ Population Growth Rate (percentage)
- ❖ Population Total
- ❖ Rate of Natural Change (per 1,000 population)

En continuidad con los lineamientos expuestos previamente en relación al concepto de *VIF*, estas features serán eliminadas y no tenidas en cuenta para el análisis aquí desarrollado.

Finalmente, el resto de las variables serán utilizadas para trabajar con el modelo de ML.

### 3.4. IMPLEMENTACIÓN

Para probar la hipótesis propuesta, se procede a dividir y clasificar los datos en función de los flujos migratorios netos.

Se realiza un ordenamiento de los mismos por migración neta en orden descendente y una clasificación de todos los registros en 5 categorías. Se incluirá en el grupo 1 aquellos registros con mayor Migración Neta positiva y en el grupo 5 a los que presenten para la misma variable objetivo un mayor valor negativo. De esta manera se obtiene una

comprensión de mayor profundidad en lo que respecta a la migración en diferentes escenarios y tiempos.

Es de destacar que en esta experimentación, los grupos quedarán formados por diferentes años y diferentes países . Esta situación obedece a que no estaremos estudiando el comportamiento de los flujos migratorios para un país en particular. Sino que intentaremos entender la migración como un fenómeno social inherente a la naturaleza humana, independientemente de las fronteras políticas preestablecidas, las cuales se encuentran sujetas a constantes modificaciones.

Una vez divididos los datos, se agrupan aquellos con valores 1, 3 y 5. De esta forma contaremos con registros con alta migración neta positiva (Grupo 1), registros con migración neta “con tendencia a valores neutros” (Grupo 3), y un último grupo con migración neta negativa (Grupo 5).

Para cada uno de estos grupos, se realiza una selección de las 5 características de mayor peso y se calculan los coeficientes correspondientes. Se verifican los *valores-p* y algunos datos estadísticos de relevancia.

A los fines de este desarrollo, se prescinde de aquellas características no numéricas, la característica target 'Número neto de migrantes (miles)' y 'Tasa neta de migración (por 1.000 habitantes)'.

Esta última variable se encuentra altamente correlacionada con el target. La misma se elimina de la muestra considerando que podría traer al modelo problemas de endogeneidad y autocorrelación.

No serán consideradas las variables:

- ❖ Country Code
- ❖ Country Name
- ❖ Year
- ❖ Net Number of Migrants (thousands)
- ❖ Net Migration Rate (per 1,000 population)

- ❖ Access Elect.
- ❖ Crude Birth Rate (births per 1,000 population)
- ❖ Natural Change, Births minus Deaths (thousands)
- ❖ Population Growth Rate (percentage)
- ❖ Population Total
- ❖ Rate of Natural Change (per 1,000 population)

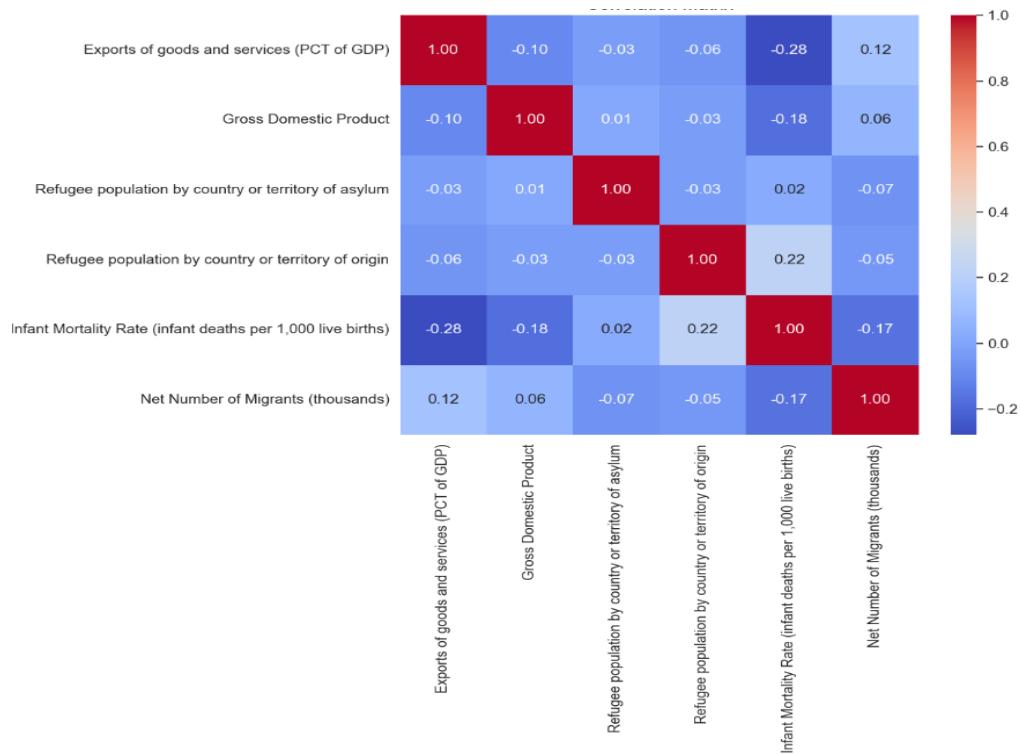
### **GRUPO 1 (Migración neta positiva)**

Las 5 características de mayor peso correspondientes a este grupo son las siguientes:

- ❖ Exports of goods and services (PCT of GDP)
- ❖ Gross Domestic Product
- ❖ Refugee population by country or territory of asylum
- ❖ Refugee population by country or territory of origin
- ❖ Infant Mortality Rate (infant deaths per 1,000 live births)

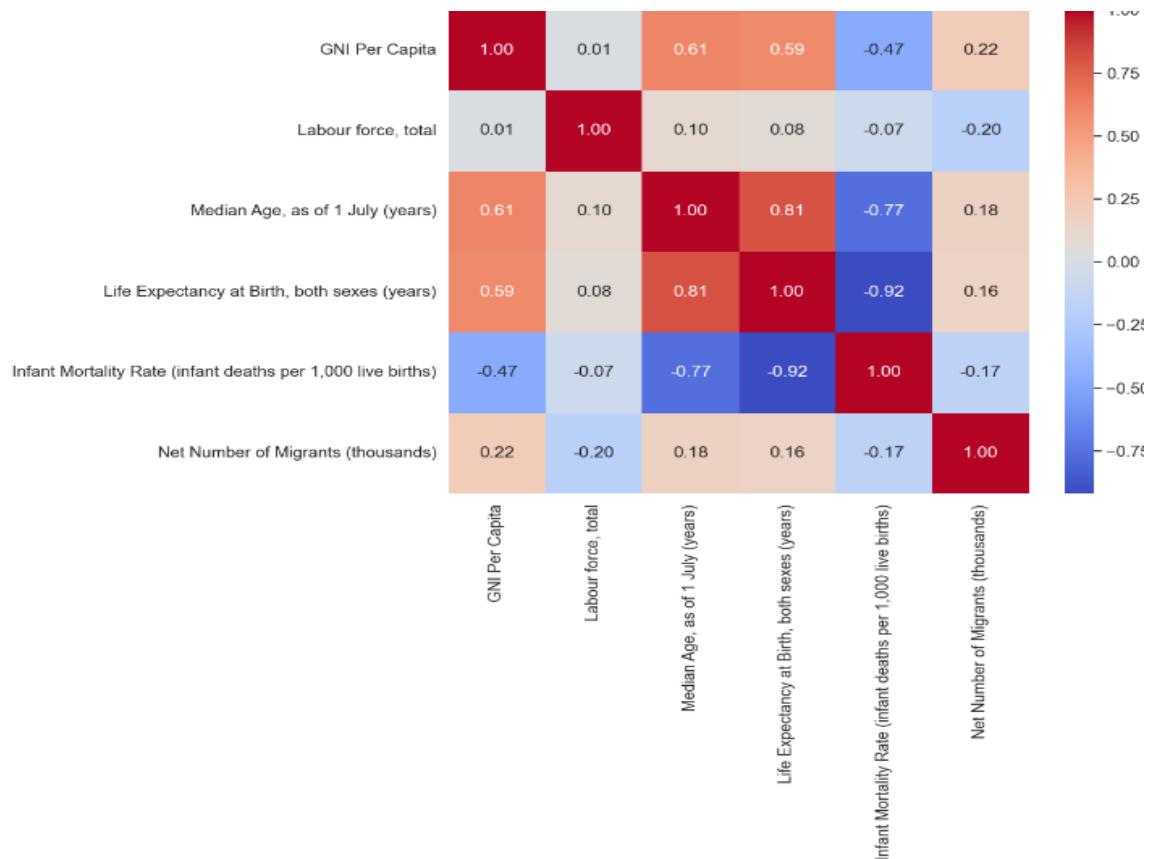
Se adjunta una matriz de correlación de las features en cuestión

### **GRUPO 3 (Migración neta “con tendencia al neutro”)**



Las 5 características de mayor peso correspondientes a este grupo son las siguientes:

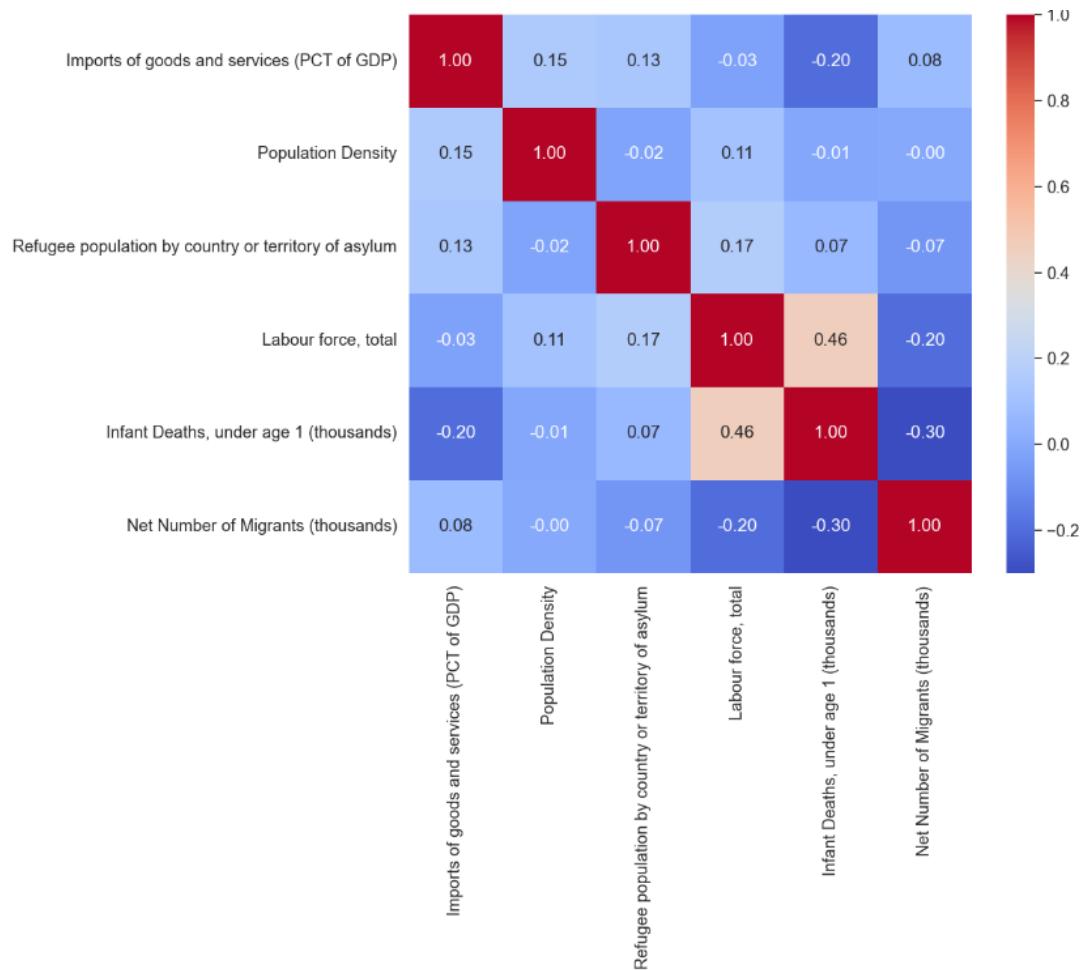
- ❖ GNI Per Capita
- ❖ Labour force, total
- ❖ Median Age, as of 1 July (years)
- ❖ Life Expectancy at Birth, both sexes (years)
- ❖ Infant Mortality Rate (infant deaths per 1,000 live births)



## GRUPO 5 (Migración neta negativa)

Las 5 características de mayor peso correspondientes a este grupo son las siguientes:

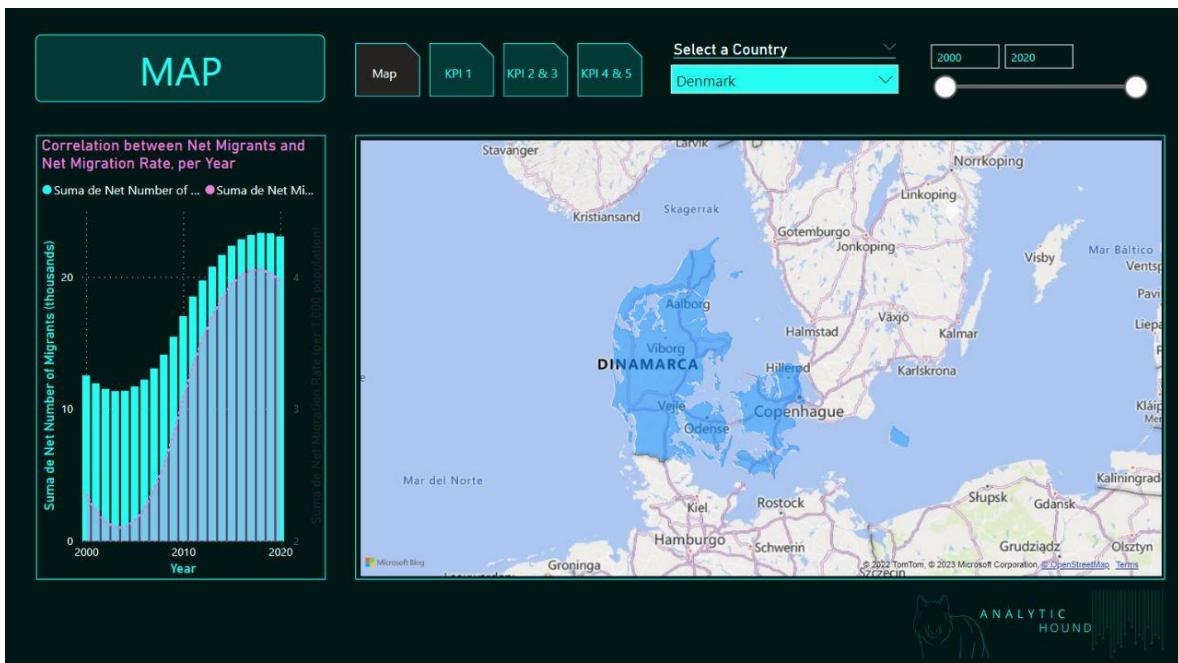
- ❖ Imports of goods and services (PCT of GDP)
- ❖ Population Density
- ❖ Refugee population by country or territory of asylum
- ❖ Labour force, total
- ❖ Infant Deaths, under age 1 (thousands)

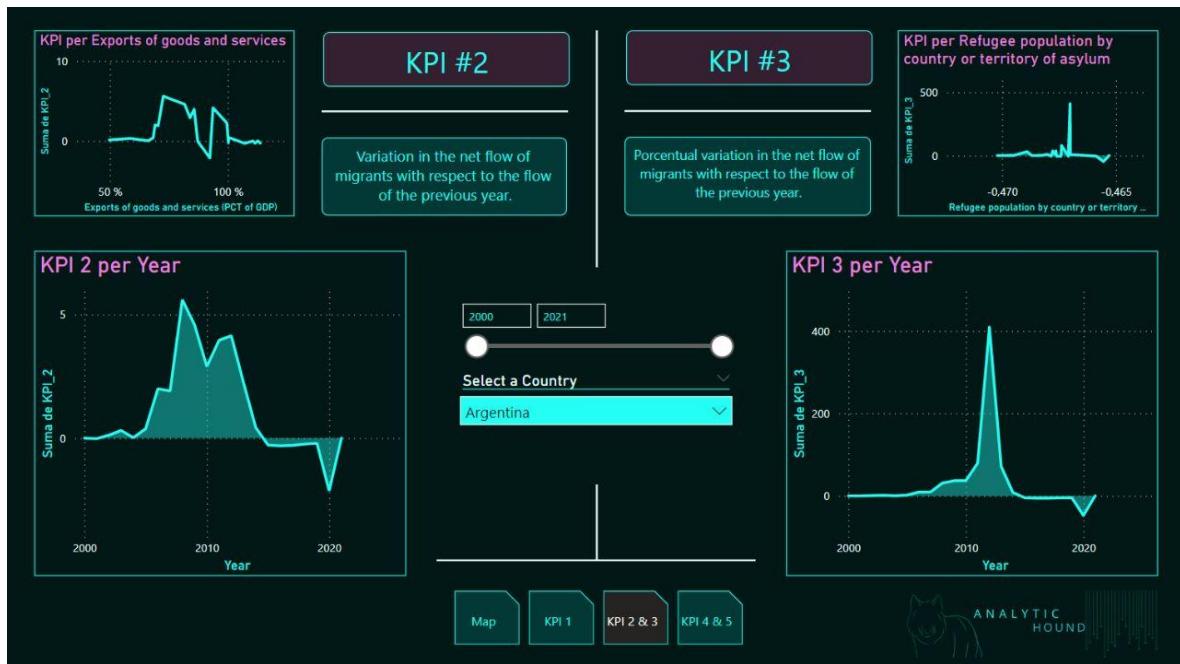


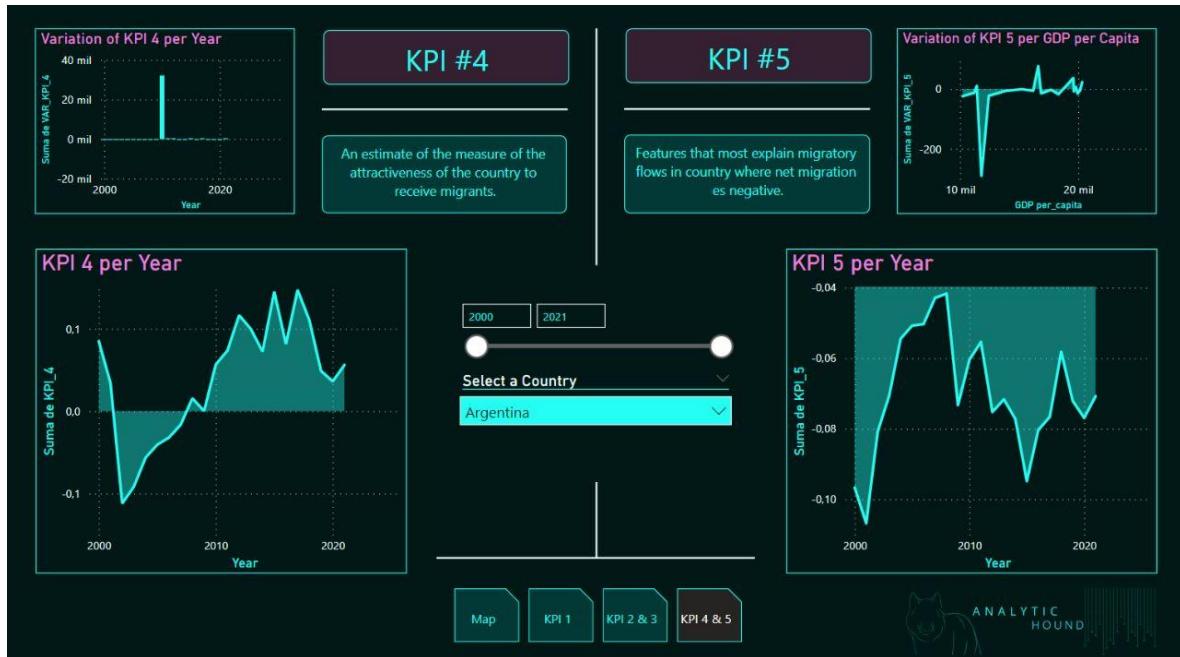
## 4. VISUALIZACIONES

El uso de herramientas interactivas de visualización permite al usuario navegar de forma global y sencilla los diferentes datos utilizados así como sus interrelaciones. La manera en que estos datos se vinculan y las conclusiones derivadas de ellos, son la *llave maestra* de nuestro proyecto. A continuación se presentan en forma esquemática algunas imágenes correspondientes a las visualizaciones desarrolladas por el equipo de analistas de **Analytic Hound ®**.

### 4.1 POWER BI







## 4.2 STREAMLIT (LINK DE ACCESO: <http://107.21.7.155:8501/>)

Home

Demographic Indicators

GITHUB REPOSITORY | [LINK](#)

MAILING | [analytichound@gmail.com](mailto:analytichound@gmail.com)

Developed by Analytic Hound Group®

...Analytic Hound data load complete

ANALYTIC HOUND

A MIGRATION PROJECT

Throughout history, migratory flows have undergone multiple changes. These changes are the translation of the different sociopolitical, demographic, environmental and economic aspects among others. "The Analytic Hound"

Home

**Demographic Indicators**

Development Indicators

Health Indicators

Macro Indicators

Micro Indicators

Migration Indicators

Study of Features with ML

KPIs

Select a country:

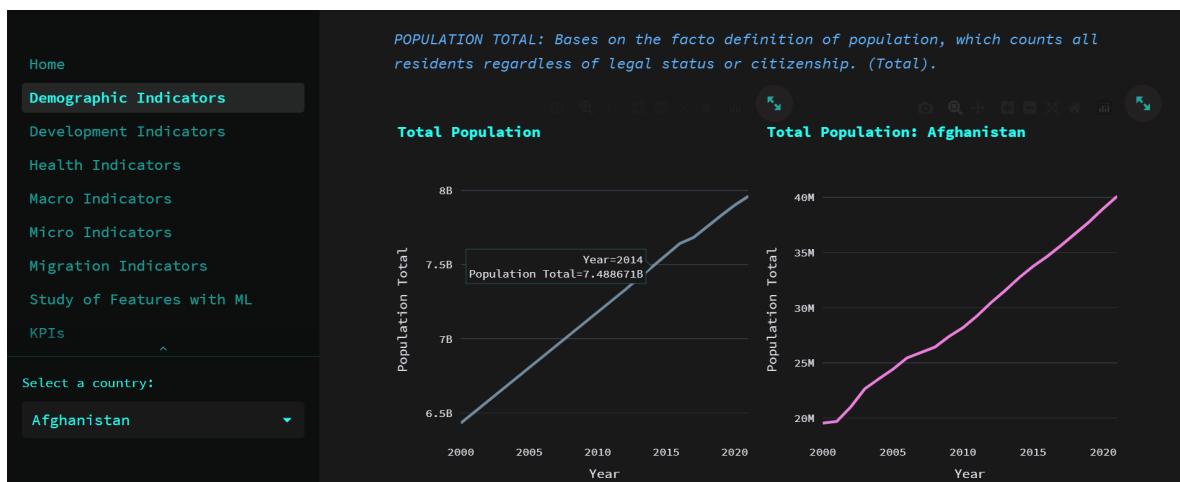
Afghanistan



IN THIS SECTION YOU CAN CONSULT THE MOST IMPORTANT VARIABLES ON THIS CATEGORY OF INDICATORS.

**Total Population**

**Total Population: Afghanistan**



## 5. INSIGHTS Y CONCLUSIONES

En los apartados previos, el contenido de este documento ha sido íntegramente abocado a la confección de un manual preciso y detallado vinculado a los distintos procesos y etapas que engloban al procesamiento de los datos.

Resulta imperativo en este punto realizar una aclaración acerca de lo que el equipo de **Analytic Hound** ® considera la *piedra angular* de este proyecto; nos referimos aquí a la “**Transformación de datos en información**”.

Hemos examinado el proceso que sufren los datos desde su obtención a partir de las fuentes, pasando por toda su etapa de transformación, hasta su visualización en forma de gráficos y dashboards, así como su uso en el desarrollo de los distintos modelos de machine learning.

Intuitivamente, resulta esperable cuestionarse el momento preciso en el cual los datos abandonan su lugar como tal para convertirse en valiosa información. Nos referimos a información o **insight** cuando, implementados los diferentes procesos correspondientes, los datos se convierten en generadores de nuevas conclusiones o **conocimiento**, permitiendo su uso primariamente para la toma de decisiones estratégicas.

A continuación, expondremos los diferentes **insights** obtenidos en relación al objeto de estudio de este proyecto.

Remitiéndonos al apartado de machine learning, el resultado del análisis arroja para cada grupo de registros los 5 features más relevantes a la hora de explicar la migración neta. Para el desarrollo de este paper introductorio, por motivos de extensión en la documentación, seleccionaremos únicamente dos de estos grupos: el Grupo 1 y el Grupo 5. Estos últimos aglomeran aquellos registros ubicados en los polos opuestos respecto al flujo migratorio neto. En futuros apéndices de esta documentación, se abordará el análisis de forma más integral y detallada.

En relación al **primer grupo**, donde se nuclean los registros con mayor valor de **migración neta positiva**, podemos encontrar la **exportación de bienes y servicios, el PBI y la**

**mortalidad infantil** como las características más destacadas en relación a la influencia que estas pudieran ejercer sobre una migración positiva.

Así como el PBI (producto bruto interno) es el claro reflejo de la situación económica real de un país, el carácter de “exportador” aportado por la variable correspondiente permite obtener una idea global del grado de industrialización del mismo. Este desarrollo industrial implica necesariamente un correlato laboral, siendo que aquellas regiones con mayor cantidad de exportaciones requieren de una mayor cantidad de mano de obra para la producción de dichos bienes o servicios. Se desprende de esto último que el impulso orientado al desarrollo de la industria, traducido en mayores puestos de trabajo, mayor cantidad de exportaciones y mejoras en la situación del PBI nacional, resultan ser variables que afectan directa e indirectamente sobre la dinámica de los flujos migratorios.

Según la evidencia empírica arrojada por nuestro estudio, existe una relación inversa entre niveles de exportación y flujo migratorio neto. Abordaremos la interpretación específica de esta relación luego de analizar el efecto del feature importación de bienes y servicios sobre el Grupo 5, desarrollando ambos efectos en forma conjunta.

La mortalidad infantil se encuentra íntimamente ligada a las políticas sanitarias y a la calidad de los servicios de salud de cada país. Resulta poco sorpresivo que el coeficiente de regresión de la variable sea negativo, indicando que a mayores niveles de mortalidad infantil, menores flujos migratorios netos habrá. Es sencillo inferir que un valor alto de este indicador resulta en un desincentivo a la hora de definir un proceso migratorio hacia el país en cuestión. La tasa de mortalidad infantil no es otra cosa que el reflejo de las condiciones económicas, sociales y ambientales en la salud de las madres y los bebés, así como el acceso y la efectividad de los sistemas de salud.

Analizando el polo alternativo del flujo migratorio, en aquellos registros pertenecientes al **grupo 5** encontramos agrupados los de mayor **migración neta negativa**. Entre las *features* determinantes y con significancia estadística vinculadas a este conjunto se encuentran la

*importación de bienes y servicios, la densidad poblacional, mortalidad infantil (en menores de 1 año) y población trabajadora.*

En lo que a sistema sanitario refiere, nuevamente surge la importancia de la salud a través de la *mortalidad infantil*. En este caso se observa que el signo del coeficiente de la regresión es negativo, por lo cual se establece que cuanto mayor sea el valor de este indicador, menor su flujo migratorio. Dada la obviedad de la interpretación, no entraremos en detalles en lo que respecta a este indicador.

Recordando la definición brindada en el diccionario de este documento, se entiende por *fuerza laboral* a “Aquellas personas de 15 años o más que aportan mano de obra para la producción de bienes y servicios durante un período determinado. Incluye a las personas que actualmente están empleadas y las personas que están desempleadas pero buscan trabajo”. En este caso, en relación a esta variable, se observa una correlación negativa entre la misma y un flujo migratorio neto negativo. En consecuencia, la existencia de una menor fuerza de trabajo implicaría un mayor número de inmigrantes para un determinado país.

Esta relación es esperable en términos de las bases de las teorías del Comercio Internacional, que consideran la mano de obra (tanto calificada como no calificada) como un factor de producción más, al igual que la materia prima y los insumos. Es de esperar que exista una relación inversa entre mano de obra y migración neta, puesto que al escasear la mano de obra en un país, es natural que se compense con flujos migratorios externos. De forma opuesta, cuando la fuerza laboral crece y tiende al pleno empleo de los factores productivos, los flujos migratorios de mano de obra tenderán a reducirse, entendiendo que serían menos demandados.

Sería interesante indagar en esta relación, ya que ofrece muchas aristas de estudio. Por ejemplo, en el contexto de un mundo globalizado y con el advenimiento de los trabajos “remotos”, será primordial para el correcto desarrollo de la economía contar con un conocimiento preciso de los efectos del mercado laboral para alinear los incentivos de estos trabajadores capacitados, quienes usualmente son los mejores remunerados. Una

buenas políticas al respecto podría generar un impacto potencial en múltiples aspectos de la sociedad generando “derrames” a niveles consumo, recaudación, inversión doméstica, etc.

Interesantemente, del análisis del grupo 5 surge también la variable densidad poblacional. La misma presenta un coeficiente de regresión elevado y negativo, con un p-valor muy significativo determinando su preponderancia en el análisis. En resumidas cuentas, el hallazgo evidencia que los registros con los mayores flujos migratorios netos negativos, se caracterizan por mostrar una relación inversa a la densidad poblacional. Esto se puede interpretarse de varias formas, y sería necesario descomponer el efecto considerando cantidades de inmigrantes y emigrantes, pero concretamente se traduce en que cuanto más densamente poblada sea una determinada zona, menor ingreso de migrantes tendrá o, dicho de otra forma, será mayor la cantidad de gente que egresa respecto a los migrantes que pudieran ingresar.

Es de enfatizar que este efecto demográfico de alto interés presenta una potencialidad de análisis mayor. En posteriores publicaciones, se optará por segmentar dicho indicador en las distintas partes que lo conforman, para lograr un entendimiento más preciso de sus consecuencias.

En relación a las importaciones de bienes y servicios, resulta necesario un mayor caudal de datos para la correcta interpretación de su efecto. Concretamente, el carácter de “importador” se encuentra positivamente relacionado a los flujos migratorios en los países con flujos migratorios netos negativos. Se podría alegar diversos argumentos respecto de las causas detrás de dicho efecto, tales como que la matriz productiva del país es netamente importadora, que la importación es lo único que evita que la sangría migratoria sea mayor, etc. Finalmente, en este caso, sólo es posible aseverar lo postulado previamente como hecho concreto donde se reconoce la relación positiva entre las variables.

En rigor a la exactitud de la interpretación, es donde debemos determinar los alcances de este trabajo. Con lo hasta aquí desarrollado, hemos logrado categorizar por grupos los diferentes países, permitiendo al cliente contar con un intervalo de confianza del 95% de

certeza a la hora de conocer aquellas cuestiones específicas sobre las cuales ha de ponerse especial atención a la hora de tomar medidas que pudieran influir sobre el flujo migratorio del año próximo.

En función de determinar cuestiones específicas sobre un país, grupo o región, se deberá proceder a indagar sobre las métricas y conocer los comportamientos de las variables con mayor grado de detalle (chequear estacionalidades, efectos indirectos, existencia de autocorrelación o rezagos distribuidos,etc)

Ilustraremos este punto de forma general, aclarando la cuestión relacionada a las *Exportaciones* (Grupo 1) e *Importaciones* (Grupo 5).En los países del Grupo 1, al observar una salida de regresión más amplia, observamos que tanto Exportación como Importación están relacionadas negativamente al flujo migratorio neto. A primera vista podría parecer antiintuitivo o inconsistente. Percepción que se podría reforzar al conocer que ambas variables se encuentran relacionadas positivamente a Migración Neta para los países del Grupo 5.

El lector reflexivo entenderá que construir la variable *Exportaciones Netas* podría iluminar esta disyuntiva ya que otorgaría otro tipo de información. Finalmente, la intención en este punto es dejar claro que el tipo de conclusiones puntuales que explican la naturaleza detrás del efecto que percibimos, requieren de una metodología de trabajo alternativa que excede los alcances de este documento.

En versiones posteriores ahondaremos sobre el estudio detallado de estas situaciones surgidas a raíz del análisis. Estudiaremos casos puntuales y específicos que hayan despertado interés, y analizaremos registros y países “outliers”.

En conclusión, como fuera definido inicialmente en esta documentación<sup>6</sup>, hemos cumplimentado la labor mediante la entrega de una poderosa herramienta de política migratoria que otorga información precisa y relevante respecto del efecto per se. Sería posible realizar modificaciones en el modelo para aggiornarlo a las necesidades puntuales de un país, pero nos ha generado mayor interés y nos ha parecido más útil explorar la

---

<sup>6</sup> Para más información, dirigirse al apartado 1.4 “Alcance y fuera de alcance”

naturaleza del efecto y otorgar un instrumento práctico y eficaz para la evaluación y la toma de decisiones.

Ciertamente son muchos los debates centrados en la evolución de los flujos migratorios, muchos los intereses y muchos los interesados en conducir los hilos de este fenómeno netamente humano. Es nuestro mayor deseo haber aportado con esfuerzo y desarrollo una herramienta que permita que las políticas migratorias avancen y se orienten en el futuro hacia una práctica más saludable y controlada, permitiendo en toda su aplicabilidad obtener beneficios para todos los seres humanos de este planeta... a fin de cuentas, todos convivimos dentro de una misma frontera.

## **ANEXOS:**

### **1. REPOSITORIO/REPOSITORY:**

Toda información relacionada al proyecto en cuestión es de público acceso a través del repositorio de github.

<https://github.com/Analytic-Hound-Consulting/ONG-Henry>



EDA-ETL	Add files via upload
Weekly Sprint #1	Add files via upload
Weekly Sprint #2	Add files via upload
Weekly Sprint #3	Add files via upload
graph	Add files via upload
COPYING	Create COPYING
Dashboard.pbix	Add files via upload
Documentacion (ESP).docx	Add files via upload
Documentation (ENG).docx	Add files via upload
README(ENG).md	Update README(ENG).md
README.md	Update README.md

<b>EDA - ETL</b>	Bases de datos (formato .csv) Código y reporte EDA (Análisis Exploratorio de Datos)
<b>Weekly sprint</b>	Bitácora, soporte gráfico, reporte oficial y entregables semanales
<b>Graph</b>	Soporte gráfico
<b>COPYNG</b>	Licencia de uso
<b>Dashboard</b>	Dashboard - Power BI
<b>Documentación</b>	Documentación final (Español e inglés)
<b>Readme</b>	README del repositorio (Español e inglés)

## **2. FUENTES BIBLIOGRÁFICAS**

- ❖ Informe ONU Migración 2022. Organización Internacional para las Migraciones (OIM).
- ❖ Presentación del Dossier estadístico anual sobre inmigración italiana. Publicación Actualidad Internacional Sociolaboral nº 228.
- ❖ eBrain – Using AI for Automatic Assessment at the Hong Kong Immigration Department. American Association for Artificial Intelligence. Wong, R.W.-M. y A.H.W. Chun
- ❖ Econometría, Segunda Edición. Damodar N. Gujarati. Copyright 1986 by McGraw Hill Inc.
- ❖ “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, by Aurélien Géron (O'Reilly). Copyright 2019 Aurélien Géron, 978-1-492-03264-9.”

## AGRADECIMIENTOS:

El equipo de **Analytic Hound** ® agradece la colaboración del *Data World Bank, Data Hub y United Nations* en la provisión de datos necesarios para el desarrollo de este proyecto. El correcto análisis de la situación migratoria mundial, así como el desarrollo de las predicciones realizadas en este documento, no hubiesen sido posible de no contar con el apropiado respaldo de información brindado por dichas fuentes.

