SPRINT #1: PUESTA EN MARCHA DEL PROYECTO Y TRABAJO CON DATOS

En el primer sprint, nos enfocamos en la puesta en marcha del proyecto y en trabajar con los datos. Definimos cuatro KPIs que medirán el impacto de los flujos migratorios, documentamos el alcance del proyecto y realizamos un análisis exploratorio de los datos. Además, establecimos un repositorio en GitHub, implementamos el stack tecnológico elegido y establecimos la metodología de trabajo. Al finalizar este sprint, teniamos un flujo de trabajo establecido y una base sólida para el análisis de datos.

ENTENDIMIENTO DE LA SITUACIÓN ACTUAL

La migración humana es un fenómeno que ha moldeado nuestras sociedades a lo largo de la historia y que sigue siendo relevante en el mundo actual. Las personas se desplazan por diversas razones, como la búsqueda de oportunidades económicas, el escape de situaciones de violencia o la adaptación a condiciones ambientales cambiantes. Estos factores, junto con el desarrollo tecnológico que ha facilitado la comunicación y el transporte, han generado una gran diversidad y complejidad en los flujos migratorios actuales.

En este contexto, el análisis de los datos migratorios se presenta como una herramienta esencial para comprender y abordar los desafíos y oportunidades asociados con la migración en la actualidad. Los datos recopilados sobre los movimientos de población, las rutas migratorias, las características demográficas y los factores impulsadores proporcionan información valiosa que puede guiar la toma de decisiones informada en diversos niveles.

Esto no solo nos permitirá entender la magnitud y dirección de los movimientos de población, sino también explorar las posibles implicaciones socioeconómicas, culturales y políticas de estos flujos.

OBJETIVOS

- Analizar y comprender los flujos migratorios internacionales.
- Recolección y consolidación de datos históricos.
- Identificar países de origen y destino relevantes.
- Reconocer los principales factores para la migración.

ALCANCE

Análisis de Flujos Migratorios y Factores Motivadores: Recopilaremos y analizaremos datos históricos de flujos migratorios, centrándonos en los años comprendidos entre 2000 y 2020. Nos enfocaremos en los 10 países de origen y destino más frecuentes en este período.

Exploraremos las razones detrás de estos movimientos, considerando factores socioeconómicos, políticos y medioambientales.

La elección del período de 2000 al 2020 se basa en la disponibilidad de datos completos y detallados para realizar un análisis sólido, mientras que la inclusión de años anteriores puede estar limitada por la falta de información completa en esos años debido a razones específicas de cada país.

⫸ OBJETIVOS Y KPIS ASOCIADOS (PLANTEO)

Los KPIs que hemos definido se basan en nuestra comprensión de los flujos migratorios y los factores que los impulsan, y están diseñados para generar un impacto positivo en la toma de decisiones y en la sociedad en general.

1. **Objetivo: Reducir la tasa de migración neta para los países origen con mayor flujo migratorio en 1% en el próximo año.**
   - **KPI:** Tasa de Migración Neta disminuida en un 1% en los países origen con los flujos migratorios más significativos.
2. **Objetivo: Reducir la tasa de desempleo en el país de origen en un 3% en los próximos 2 años.**
   - **KPI:** Tasa de Desempleo en el país de origen reducida en un 3% en un período de 2 años.
3. **Objetivo: Reducir la tasa de migración neta para los países destino con mayor flujo migratorio en 1% en el próximo año.**
   - **KPI:** Tasa de Migración Neta disminuida en un 1% en los países destino con los flujos migratorios más significativos.
4. **Objetivo: Reducir la tasa de desempleo en el país de destino en un 3% en los próximos 2 años.**
   - **KPI:** Tasa de Desempleo en el país de destino reducida en un 3% en un período de 2 años.

⫸ SOLUCIÓN PROPUESTA

Nuestra solución se centrará en un enfoque estructurado y colaborativo que nos permitirá alcanzar los objetivos establecidos y crear un producto final integral y significativo. A continuación, detallamos las tareas clave que realizaremos para cumplir con nuestros objetivos y los productos que surgirán de cada etapa:

Metodología y Organización:

Adoptaremos una metodología ágil para el desarrollo de nuestro proyecto, utilizando Scrum como marco de trabajo. Organizaremos nuestro trabajo en sprints de una semana, con reuniones diarias de seguimiento y demos al final de cada sprint. Esto nos permitirá mantener una comunicación constante, abordar desafíos a medida que surjan y adaptarnos rápidamente a cambios.

Distribución de Tareas y Roles:

- Product Owner: representante.
- Scrum Master: Mentor.
- Equipo de Data Science: Brenda Jaras, Kevin Bonin, Miller Rodríguez y Angie Arango.

Roles y Responsabilidades del equipo:

☑ **Data Engineer: Brenda Jaras**

- **Responsabilidad:** La Data Engineer se encargará de diseñar, construir y mantener la infraestructura de datos, asegurando la disponibilidad, calidad y procesamiento eficiente de los datos.

Tareas Clave:

- Crear y mantener el Data Pipeline: Diseñar, implementar y mantener un flujo de datos automatizado desde múltiples fuentes hasta el Data Warehouse.
- Diseñar y Gestionar la Base de Datos y Data Warehouse: Crear y mantener una estructura de base de datos óptima para el almacenamiento y análisis de datos migratorios.
- Automatizar Procesos de ETL: Utilizar herramientas como AWS Glue para automatizar el procesamiento y la carga de datos en el Data Warehouse.

☑ **Data Scientist: Kevin Bonin y Miller Rodriguez**

- **Responsabilidad:** El Data Scientist se enfocará en analizar los datos, identificar patrones y tendencias, y desarrollar modelos predictivos para comprender mejor los flujos migratorios.

Tareas Clave:

- Análisis Exploratorio de Datos (EDA): Realizar análisis detallados para descubrir relaciones, patrones y características significativas en los datos migratorios.
- Desarrollo de Modelos Predictivos: Utilizar técnicas de Machine Learning para construir modelos que pronostiquen patrones futuros en los flujos migratorios en base a factores históricos y contextuales.

☑ **<u>Data Analyst:</u> Angie Arango**

- **Responsabilidad:** El Data Analyst se encargará de traducir los datos en información comprensible y presentable, destacando los aspectos clave para la toma de decisiones.

Tareas Clave:

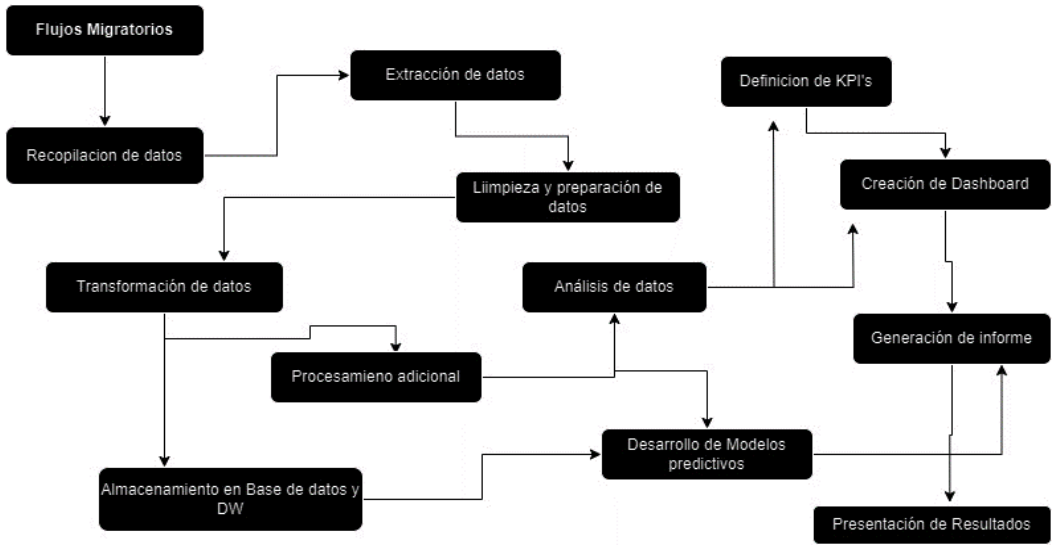
- Análisis Exploratorio de Datos (EDA): Realizar análisis detallados para descubrir relaciones, patrones y características significativas en los datos migratorios.
- Definir KPIs Relevantes: Identificar y definir los KPIs que medirán el impacto de los flujos migratorios y proporcionarán información para la toma de decisiones.
- Creación de Dashboards Interactivos: Diseñar y desarrollar un dashboard interactivo que muestre visualmente los datos migratorios relevantes y los KPIs identificados.
- Generación de Informes: Preparar informes y presentaciones que comuniquen los resultados y hallazgos de manera clara y concisa.
- Colaboración Interdisciplinaria: Trabajar en conjunto con el Data Engineer y el Data Scientist para asegurar la correcta interpretación y visualización de los datos.

Colaboración y Comunicación:

Los tres roles trabajarán en estrecha colaboración, compartiendo sus hallazgos y contribuciones en reuniones regulares de equipo.

⏩ **FLUJO DE TRABAJO**

En esta sección, presentamos el diseño detallado del flujo de trabajo que seguiremos en nuestro proyecto, abarcando desde la recopilación de datos hasta la presentación de resultados. Cada etapa es crucial para alcanzar nuestros objetivos y crear un análisis completo de los flujos migratorios.



STACK TECNOLÓGICO:

- **Lenguaje de Programación: Python**

Utilizaremos Python como lenguaje principal para el desarrollo de scripts y código necesario en todas las etapas del proyecto. Python es ampliamente conocido en la comunidad de análisis de datos y es compatible con una variedad de bibliotecas y frameworks para análisis y modelado.

- **Extracción y Limpieza de Datos:**
  - **Bibliotecas: pandas, NumPy**

Utilizaremos la biblioteca pandas para la manipulación y limpieza de datos. pandas nos permitirá cargar, transformar y limpiar los datos de manera eficiente. NumPy se usará para operaciones numéricas y cálculos matriciales.

- **Visualización de Datos:**
  - **Bibliotecas: Matplotlib, Seaborn, geopandas**

Para crear visualizaciones claras y efectivas de los datos. geopandas nos permitirá trabajar con datos geoespaciales.

- **Base de Datos:**
  - **Base de Datos: MySQL**

Lo utilizaremos como sistema de gestión de base de datos relacional para almacenar datos estructurados.

- **Ingesta de Datos:**
  - **Servicio: AWS Lambda**

Para la ingesta de datos, permitiendo el procesamiento de datos en tiempo real.

- **Almacenamiento en Data Lake:**
  - **Servicio: AWS S3**

Para almacenar y gestionar grandes volúmenes de datos de manera escalable y segura.

- **Transformación de Datos:**
  - **Servicios: AWS Lambda y AWS Glue**

Utilizaremos AWS Lambda para transformaciones de datos específicas y AWS Glue para procesos de ETL más complejos y automatizados.

- **Creación de Dashboards Interactivos:**
  - **Herramienta: Power BI**

Para diseñar y desarrollar dashboards interactivos que visualicen de manera efectiva los resultados del análisis de datos.

- **Modelo de Machine Learning:**
  - **Framework: FastAPI**

Para desarrollar y exponer el modelo de machine learning como una API web rápida y eficiente.

- **Almacenamiento del Modelo:**
  - **Módulo: Pickle**

Se utiliza para guardar y cargar el modelo de machine learning de manera eficiente.

- **Despliegue de la API:**
  - **Plataforma: Render**

Se utiliza para el despliegue de la API de FastAPI, lo que permite el acceso público a la funcionalidad del modelo de machine learning.

- **Interfaz de Usuario Mejorada:**
  - **Herramienta: Streamlit**

Se emplea para mejorar la interfaz de usuario y proporcionar una experiencia de usuario más sencilla e intuitiva.

Este stack tecnológico integral permitirá llevar a cabo las diferentes etapas del proyecto, desde la recopilación y transformación de datos hasta la visualización de resultados, siguiendo una metodología ágil y colaborativa.

⫸ GESTIÓN DE PROYECTOS Y COLABORACIÓN

- **Herramientas de Gestión de Tareas**

Trello: Lo usamos como nuestra herramienta principal para la gestión de tareas. Trello nos permite crear tableros, listas y tarjetas para organizar y priorizar entregables, hitos y otros aspectos del proyecto.

- **Comunicación y Colaboración**

WhatsApp: Lo utilizamos para la comunicación adicional y la coordinación fuera del entorno de trabajo.

Discord: Para llevar a cabo reuniones de equipo, facilitar la comunicación en tiempo real y realizar sesiones de voz o chat en vivo cuando es necesario.

- **Control de Versiones**

GitHub: GitHub es nuestra plataforma de control de versiones, donde almacenamos y gestionamos el código fuente y los archivos del proyecto.

⫸ ANÁLISIS SOBRE LOS DATOS

Entendemos la importancia de trabajar con datos de alta calidad y, por lo tanto, realizamos un análisis detallado de los metadatos de los datos con los que vamos a trabajar en nuestro proyecto de análisis de flujos migratorios internacionales. A continuación, proporcionamos una descripción de los metadatos relevantes:

Fuente de Datos:

Los datos se obtuvieron de fuentes confiables, que incluyen organismos internacionales como la Organización de las Naciones Unidas (ONU), la Organización Internacional para las Migraciones (OIM), el Banco Mundial, entre otras.

Confiabilidad de las Fuentes:

Las fuentes utilizadas son ampliamente reconocidas y utilizadas en estudios de migración. Son confiables y proporcionan datos actualizados y precisos sobre flujos migratorios a nivel global.

Descripción de los Datos:

Los datos contienen información sobre flujos migratorios internacionales, incluyendo:

- País de origen de los migrantes.
- País de destino de los migrantes.
- Porcentaje de mujeres, en país destino, y país origen.
- Factores de migración: muertes por conflictos, esperanza de vida, nivel de corrupción, inflación del deflactor del PIB, tasa de desempleo, crecimiento del PIB, homicidios intencionales, emisiones de CO2, migración neta y población neta.
- Información geográfica nombre del País, Código del país.
- Línea de tiempo de análisis, Año, entre el 2000 y 2020.
- Población de refugiados por país de origen.
- Población de refugiados por país de asilo.

Tipos de Datos:

Los tipos de datos incluyen números enteros para datos demográficos, números decimales para indicadores económicos, porcentajes, texto para descripciones y categorías.

Método de Adquisición:

Acceso a Datos Gubernamentales a través de APIs

Fuentes de Datos Gubernamentales: Lista de las fuentes de datos gubernamentales que se utilizaron

APIs Utilizadas: Especificación de las APIs que se accedieron

Documentación de API: Enlace a la documentación oficial de cada API

Autenticación y Autorización: Descripción de cómo se gestionó la autenticación y la autorización para acceder a las APIs

Proceso de Adquisición de Datos: Descripción de cómo se recopilaron, procesaron y almacenaron los datos obtenidos de las APIs.

Fechas de Adquisición y Última Actualización: Incluir fechas.

SPRINT #2: DATA ENGINEERING

En el segundo sprint, nos enfocamos en configurar pipelines y automatización, lo que nos permitió agilizar el proceso de adquisición, transformación y carga de datos. También finalizamos el análisis exploratorio de datos (EDA), desarrollamos una prueba de concepto del dashboard, realizamos una prueba de concepto del modelo de machine learning y creamos el diccionario de datos junto con el modelo Entidad-Relación (ER). Además, diseñamos un workflow que detalla las tecnologías específicas usadas, lo que nos permitió gestionar de manera eficiente todas las etapas del proyecto.

Con el objetivo de lograr esto, se realizaron las siguientes acciones y se utilizaron las herramientas y servicios correspondientes:

- Para iniciar nuestro pipeline de datos, configuramos AWS Lambda como servicio para la ingesta de datos en tiempo real, asegurando que los datos estén siempre actualizados y disponibles para su procesamiento.
- Implementamos AWS S3 como nuestro servicio de almacenamiento inicial para gestionar grandes volúmenes de datos de manera escalable y segura. Esto forma parte de la infraestructura inicial del pipeline.

- Utilizamos AWS RDS (Amazon Relational Database Service) como un paso intermedio antes de MySQL, para la gestión y estructuración de los datos procesados por el pipeline. AWS RDS proporciona una base sólida para la transformación y preparación de los datos antes de almacenarlos finalmente en MySQL.
- Finalmente, implementamos MySQL como nuestro sistema de gestión de base de datos relacional, aprovechando su capacidad para almacenar datos estructurados de manera eficiente. Esto es esencial para nuestras necesidades de almacenamiento y análisis de datos, y MySQL se integra en nuestro pipeline después de pasar por AWS RDS.

⫸ ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Finalizamos el proceso de análisis exploratorio de datos (EDA) que se inició en el primer sprint. En esta etapa, llevamos a cabo un análisis exhaustivo en una muestra representativa de los datos. Esto incluyó la identificación de valores atípicos, la exploración de las distribuciones de variables y la realización de correlaciones preliminares.

El EDA en el primer sprint nos permitió adentrarnos en los datos y comenzar a comprender sus características. Durante el segundo sprint, completamos este proceso, lo que nos proporcionó una comprensión más profunda de la calidad de los datos, patrones emergentes y áreas de interés clave para el análisis posterior.

Este enfoque secuencial en el EDA nos permitió abordar de manera más efectiva la preparación de los datos y la formulación de preguntas críticas para el análisis de flujos migratorios internacionales.

⫸ PRUEBA DE CONCEPTO DEL DASHBOARD:

Durante este sprint, creamos una versión simplificada del dashboard que tenemos previsto implementar en etapas posteriores del proyecto. Esta versión de prueba incluyó la conexión con nuestra base de datos MySQL y la incorporación de algunas visualizaciones preliminares junto con datos de muestra.

El objetivo principal de esta prueba de concepto fue permitirnos explorar la interacción y la estructura del dashboard antes de trabajar con datos completos. Al utilizar datos de muestra, pudimos evaluar la usabilidad y la funcionalidad del dashboard, identificar posibles problemas y realizar mejoras tempranas en el diseño y la experiencia del usuario.

PRUEBA DE CONCEPTO DEL MODELO DE MACHINE LEARNING:

Además de trabajar en el dashboard, también llevamos a cabo una prueba de concepto para nuestro modelo de machine learning diseñado para predecir flujos migratorios. En esta fase, creamos versiones de prueba del modelo utilizando datos de ejemplo.

La finalidad de esta prueba de concepto fue identificar posibles problemas de usabilidad y funcionalidad en los productos de machine learning antes de presentar la versión final. Al utilizar datos de ejemplo, pudimos evaluar el rendimiento del modelo, ajustar parámetros y comprender cómo se comporta antes de aplicarlo a los datos completos.

DICCIONARIO DE DATOS:

Definimos un diccionario de datos completo que proporciona una descripción detallada de todas las variables, atributos y tablas utilizadas en nuestra base de datos. Este diccionario de datos es una herramienta esencial para comprender y gestionar la estructura de nuestros datos, lo que facilita la comunicación entre los miembros del equipo y garantiza la coherencia en la interpretación de los datos.

MODELO ENTIDAD-RELACIÓN (ER):

Diseñamos un modelo Entidad-Relación (ER) que representa la estructura y las relaciones entre las entidades en nuestra base de datos. El modelo ER nos proporciona una representación gráfica de cómo se organizan los datos y cómo interactúan las diferentes entidades. Esto es fundamental para el diseño y la gestión efectiva de nuestra base de datos relacional.

SPRINT #3: DATA ANALYTICS + ML

En el tercer sprint, creamos un dashboard interactivo utilizando Power BI para visualizar y explorar los resultados del análisis de flujos migratorios internacionales, lo que facilitó la comprensión de los datos y la toma de decisiones. También realizamos un análisis en profundidad de los KPIs definidos previamente, lo que nos proporcionó insights detallados sobre patrones y tendencias en los flujos migratorios. Además, continuamos mejorando nuestro modelo de machine learning para predecir flujos migratorios, refinando algoritmos y ajustando parámetros. Finalmente, trabajamos en la automatización para mantener actualizado el dashboard y el modelo, garantizando la disponibilidad de datos y la relevancia del análisis.

DASHBOARD

El dashboard que hemos creado se compone de varias páginas que ofrecen una visión global de los flujos migratorios y sus factores determinantes:

- **Portada: Flujos Migratorios:** Esta es la página principal del dashboard, donde se puede ver una panorámica de los flujos migratorios y acceder al resto de páginas.
- **Página 1: Factores Migratorios:** En esta página, se muestran gráficos de algunos factores migratorios relevantes, como el crecimiento del PIB, el control de la corrupción, la tasa de desempleo, la migración neta y la población total de cada país. Estos factores son fundamentales para entender las tendencias migratorias.
- **Página 2 - Origen y Destino / Refugiados:** En esta página, se presenta una comparación de la población total, distinguiendo entre hombres y mujeres, tanto en los países de origen como en los países de destino a nivel mundial. El objetivo es analizar la equidad de género en la migración en general. Además, se proporciona información sobre la cantidad de personas en países de acogida en relación con las que salen de sus países de origen en busca de refugio, lo que arroja luz sobre la crisis de refugiados.
- **Página 3: KPIs de Países de Origen:** Esta página presenta los KPIs (Indicadores Clave de Desempeño) de los países de origen, que consisten en la reducción de la tasa de desempleo en un 3% en los próximos 2 años, comparando los períodos 2016-2017 y 2018-2019. También se evalúa la reducción de la tasa de migración neta en un 1% en el próximo año, comparando 2018 y 2019, para establecer metas y medir su cumplimiento en los 10 países de origen más frecuentes de emigración.
- **Página 4: KPIs de Países Destino:** De forma similar a la página anterior, esta página muestra los KPIs de los países de destino, enfocándose en la reducción de la tasa de desempleo y la migración neta en los 10 países de destino más habituales para la inmigración.
- **Página 5: Modelo (Aún en Proceso):** Esta página se destina a analizar el modelo de predicción, que permitirá obtener conclusiones relevantes relacionadas con los flujos migratorios. Aunque esta sección aún está en desarrollo, será esencial para comprender mejor las proyecciones.

MODELO DE PREDICCIÓN DE FLUJOS MIGRATORIOS

En la fase de preprocesamiento de datos, hemos construido un DataFrame utilizando la base de datos MySQL, que previamente configuramos mediante la automatización de pipelines. Además, creamos un archivo CSV diseñado para simplificar la carga de datos en la API. Como parte de este proceso, también llevamos a cabo un análisis de correlación para evaluar las relaciones entre las variables específicas de nuestro conjunto de datos y nuestra variable objetivo, que en este contexto sería la migración neta. . Este conjunto de tareas nos

proporcionó una base sólida para nuestro posterior trabajo de modelado y análisis de flujos migratorios. La comprensión y anticipación de los cambios en la migración son de gran importancia para la toma de decisiones en los ámbitos económico y social.

Variables Utilizadas:

- **Crecimiento_PIB:** Crecimiento anual del Producto Interno Bruto (%).
- **Tasa_desempleo:** Tasa de desempleo (%).
- **Muertes_Conflicto:** Número de muertes debidas a conflictos.
- **Control_Corrupcion:** Control de Corrupción (rango percentil).
- **Migración_neta:** Migración neta (número de personas).

Modelo de Machine Learning:

- **Selección del Modelo:** Utilizamos un modelo de regresión lineal para predecir la migración neta.
- **Entrenamiento del Modelo:** Entrenamos el modelo con datos del período 2000-2020 para aprender patrones migratorios.
- **Métricas de Evaluación:** Utilizamos el Error Cuadrático Medio (MSE), obteniendo un valor de 29.8136, y el coeficiente R^2 de 0.1280. Esto indica que el modelo tiene margen de mejora.
- **Predicciones Futuras:** Planeamos utilizar el modelo para realizar predicciones de migración neta en años posteriores, lo que nos ayudará a comprender mejor los flujos migratorios.

AUTOMATIZACIÓN PARA MANTENER EL DASHBOARD Y EL MODELO ACTUALIZADO:

Implementamos procesos automatizados que aseguran que tanto el dashboard interactivo como el modelo de predicción se mantengan actualizados de manera regular. Esta automatización incluye la actualización de datos desde la fuente, la recopilación de información más reciente y la incorporación de nuevos datos a medida que están disponibles. Esto garantiza que los usuarios siempre tengan acceso a información actualizada y relevante, lo que es esencial para la toma de decisiones informadas y la continuidad del análisis a lo largo del tiempo.