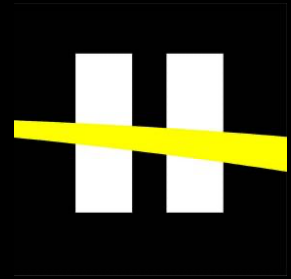


Data Analytics y Modelo de Machine Learning



Para que nuestros usuarios puedan visualizar de forma sencilla e integrada, los diferentes datos utilizados, es que utilizamos 2 herramientas interactivas: Power BI y Streamlit.

Data Analytics

El siguiente análisis de los flujos migratorios en Latinoamérica y Estados Unidos es el resultado del trabajo de las semanas anteriores donde se recopilaron y transformaron datos de gran valor para poder mostrarlos en una visualización sobre los flujos migratorios en Latinoamérica y Estados Unidos, junto con el análisis de las posibles razones que motivan a las personas a migrar, y cómo influyen en esta decisión los factores económicos y la calidad de vida de las personas en su país de origen.

En este análisis trabajaremos los siguientes Indicadores clave de Rendimiento (**KPIs**):

1. Variación de la tasa de migración neta en Estados Unidos por año
2. Variación anual de la Migración neta por país
3. Relación del PIB entre países de Latinoamérica y Estados Unidos, por año
4. Diferencia porcentual entre las Remesas y el PIB per cápita por país y año
5. Variación del % de Desempleo entre países de Latinoamérica y Estados Unidos, por año
6. Variación del índice de felicidad entre países y Estados Unidos en el año 2021

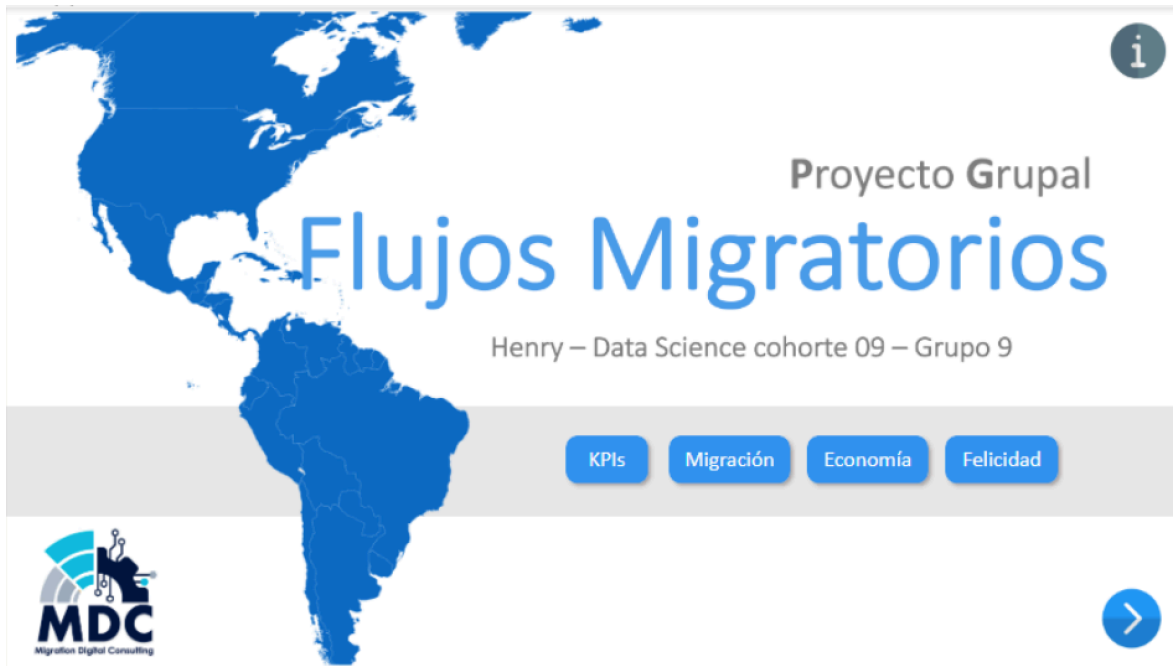
Con el presente dashboard nuestra intención es disponibilizar la información trabajada en este proyecto a nuestros clientes, gobiernos y toda persona que tenga la intención de migrar en Latinoamérica y Estados Unidos.

Proceso preliminar

Una vez los datos fueron procesados y cargados AWS S3 los conectamos con Power Bi para usar las tablas. Se aplicó cierta limpieza a cada tabla para normalizarlas y que su uso fuera más eficiente, quedando los datos divididos en las tablas:

- datos de migración (datos_migración)
- datos de calidad de vida (datos_felicidad)

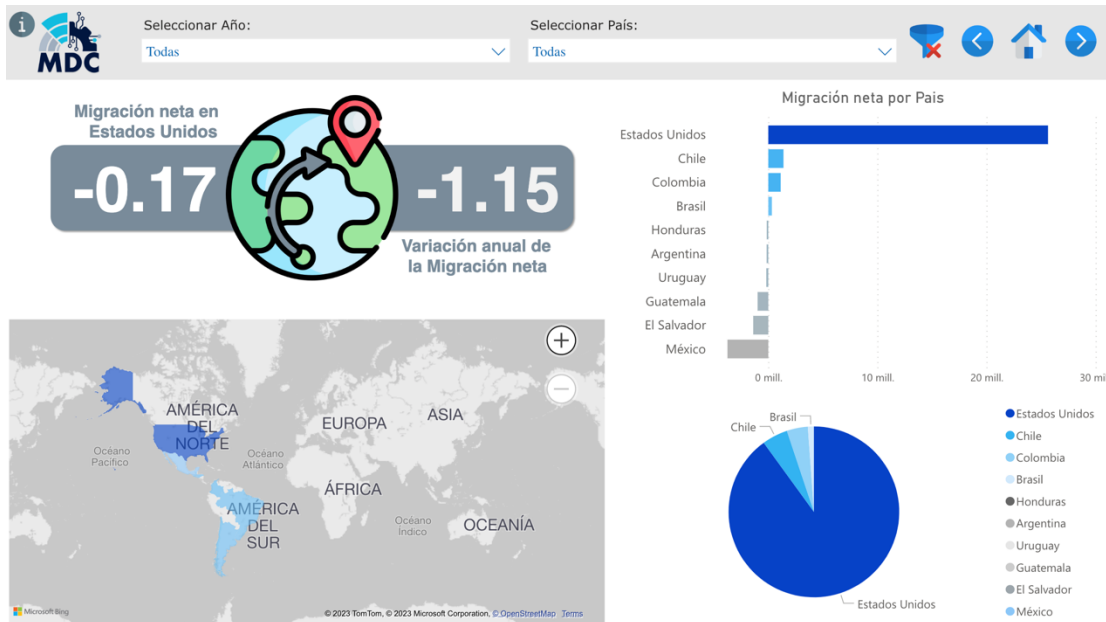
Dashboard



- Estructura del dashboard:
 - 6 páginas: Portada, KPIs, Datos de Migración, Factores decisivos, Indicadores Económicos e Indicadores de Calidad de Vida.
 - Filtro: por año y país
 - Botones y navegabilidad

Para la creación del Dashboard se utilizó el software PowerBI, donde se creó una portada, una página informativa sobre los KPIs abordados en el dashboard y 4 páginas interactivas, enfocadas en los datos de migración, los posibles factores decisivos a la hora de migrar, datos sobre la economía y la calidad de vida.

- **Dashboard Migración:** Esta dashboard interactivo muestra la Migración neta (Inmigración - Emigración) en los países estudiados, se puede filtrar la información por país y año; y se muestran los KPIs 1 y 2 que nos ayudan a comprender la variación en la migración de Estados Unidos y de cada país de Latinoamérica. Los Kpis tanto 1 como el 2, nos son de gran importancia porque partir de un análisis con esos datos nos ayudaría a entender y poder establecer una tendencia de cada país, y que mejor si se lo puede comparar de manera porcentual con el país que adoptamos como base a nuestro estudio, que en este caso es Estados Unidos



- Dashboard Factor decisivo:*** Esta dashboard resume las principales factores que pueden motivar la migración de un país de origen, como son: *la percepción de la corrupción, la libertad de tomar decisiones vitales y el PIB per cápita* de cada país. En este tablero es uno de los más importantes para poder entender el origen de la migración, ya que para empezar a relacionar los datos en sus respectivas gráficas, primero se necesita conocer los datos y que representan para luego recién poder interpretar el valor de las comparaciones. También es para destacar que con la información proporcionada en este tablero se puede sacar algunas conclusiones en respecto del origen de la decisión de los migrante, para ello tenemos que comparar con las teorías de Charles Henry Dow en su teoría de tendencia y la teoría de Vilfredo Pareto con su ley del 70 - 30, e nos cuenta que aproximadamente el 70% de los resultados se deben al 30% de las causas.



- **Dashboard Economía:** Esta dashboard resume , la relevancia queda más notoria luego de poder explicar el tablero anterior podemos destacar a la parte económica como la razón más importante para migrar, y aunque estemos todos de acuerdo que no es la única, si es importante destacar que en gran medida las decisiones pasan por ese lado.



- **Dashboard Calidad de Vida:** Este dashboard expresa la relevancia que posee temas que no están vinculados a la parte económica y que si condiciona el origen de la migración de una manera decisiva. Lo cierto es que podemos tratar de entender las cosas así, como en economía nos hablaban del Ceteris Paribus que se utiliza para analizar algo de manera independiente, dejando todo lo demás constante, y aunque esto puede ser muy útil a la hora de aprender sobre un tema, a la hora de tomar decisiones al partir de eso sería un error muy grande, ya que si se cambia un valor una formula todo lo demás se cambia.



Modelo de Machine Learning

Emprendimos un proyecto de ciencia de datos destinado a analizar y predecir las futuras tasas de desempleo y migración neta en nueve países. Nuestro equipo se encargó de desarrollar un modelo de aprendizaje automático para abordar esta tarea. Empleamos varios modelos de aprendizaje automático, centrándonos en el análisis de series temporales y la previsión. Los modelos que utilizamos en el enfoque de series temporales fueron el regresor XGBoost, la regresión lineal (adaptada a series temporales) y el modelo ARIMA (incluidos sus derivados SARIMA y SARIMAX). Además, para el análisis de regresión, empleamos modelos como XGBoost para la regresión multivariante.

Selección de modelos:

Tras evaluar el rendimiento de los distintos modelos, determinamos que el modelo ARIMA superaba a los demás en términos de rendimiento global. Por lo tanto, seleccionamos el modelo ARIMA como nuestra opción principal para el análisis posterior. A continuación, buscamos conjuntos de datos adecuados entre las opciones disponibles y aplicamos regresión y medias móviles. Como la mayoría de los conjuntos de datos eran no estacionarios, aplicamos la diferenciación de primer orden para lograr la estacionariedad. Posteriormente, empleamos el modelo ARIMA, considerando la mayoría de conjuntos de datos con diferenciación de primer orden o no estacionarios. El modelo ARIMA arrojó unos resultados impresionantes, con unas puntuaciones R-cuadrado que oscilaban entre el 56% y el 93%. Estos resultados son muy prometedores para el análisis de series temporales.

Optimización y despliegue:

Tras la selección y el análisis del modelo, procedimos a automatizar el proceso y lo aplicamos a los nueve conjuntos de datos optimizados para analizar los indicadores de migración neta en los nueve países. Además, optimizamos nueve conjuntos de datos para analizar los indicadores de la tasa de desempleo en estos mismos países. Para automatizar la optimización de los parámetros, utilizamos la biblioteca pmdarima y su módulo autoarima. Este enfoque nos permitió identificar de forma autónoma los mejores parámetros (p, d, q) para los modelos y conjuntos de datos.

Además, aplanamos los 18 modelos utilizando la biblioteca pickle y los cargamos junto con los respectivos archivos CSV que contenían los indicadores de migración neta y tasa de desempleo de los nueve países. Con el fin de proporcionar una interfaz fácil de usar para acceder a los datos y visualizarlos, desplegamos los modelos en Streamlit, donde los usuarios pueden explorar los resultados a través de representaciones gráficas personalizadas creadas con Matplotlib.

Nuestro equipo de ciencia de datos desarrolló con éxito un modelo integral de aprendizaje automático para analizar y predecir las futuras tasas de desempleo y migración neta en nueve países. La selección del modelo ARIMA, junto con la optimización de parámetros y el despliegue automatizado, garantizaron predicciones precisas y eficientes. El proyecto pone de manifiesto nuestra experiencia en ciencia de datos y nuestra capacidad para ofrecer a los clientes ideas y soluciones valiosas. Confiamos en que nuestro enfoque y resultados demuestren el potencial de nuestra consultora en ciencia de datos.

Presentación Demo 3