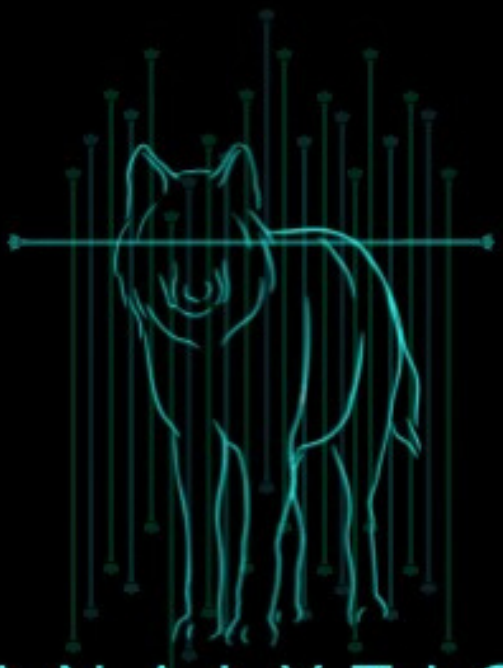


ONG - **Soy Henry**

Reporte Semana #3

DATA ANALYTICS

Indice

Página 3. Machine Learning

Página 8. Visualizaciones:

- Power BI
- Dashboards
- Streamlit

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.

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)
-

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

MACHINE LEARNING - REGRESIÓN LINEAL

Por lo general, es considerado de buena práctica destinar un porcentaje considerable (≈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 el apartado siguiente nos limitaremos a exponer el modelo desarrollado destinado a este proyecto.

IMPLEMENTACIÓN

El equipo de ML plantea como objeto de predicción el “Flujo neto de migrantes”. El deseo y la importancia de conocer las proyecciones de estos flujos futuros radica en la posibilidad que esto brinda a las diferentes naciones para tomar conductas que puedan modificar estos valores según las necesidades internas de los mismos.

Imagínese un país que se encuentra atravesando una crisis social donde no cuenta con la suficiente mano de obra “joven” para el correcto desarrollo de su economía. Conocer el abanico de acciones que podría tomar para modificar esta situación y los diferentes grados de impacto que pudieran tener supone un beneficio de alto valor para la correcta gestión de los recursos de cualquier región.

El enfoque preliminar del desarrollo para el modelo de ML consiste en el siguiente orden de procesos:

- Realizar pruebas con los datos para estudiar las relaciones entre nuestras variables y comparar los resultados.
- Ejecutar un módulo SelectKBest de SciKitLearn. Esta función selecciona las k características principales con las puntuaciones más altas en función de una puntuación que en este caso será "f_regression", que calcula la correlación entre cada característica y la variable de destino y devuelve un valor F y un valor p para cada característica.
- SelectKBest luego selecciona las k principales características con los valores F más altos, lo que indica que tienen la relación lineal más fuerte con la variable de destino.
- El proceso devuelve las siguientes 5 variables... pero ¿cómo se podría comprender mejor el impacto de cada característica en el fenómeno de la migración neta?

IMPLEMENTACIÓN

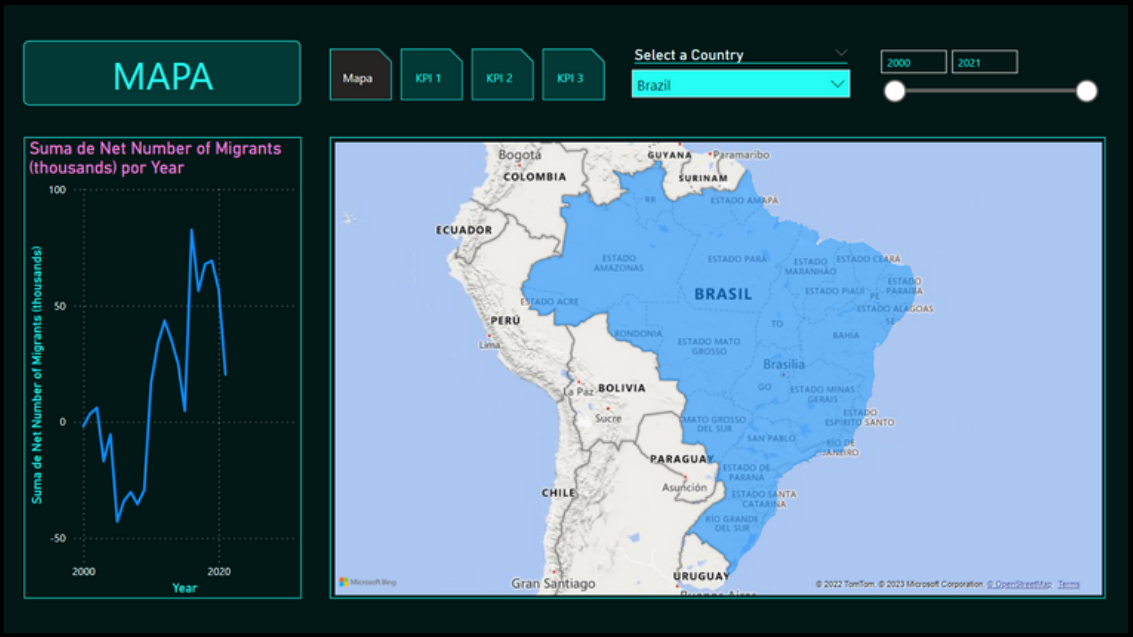
- Utilizamos los coeficientes del modelo de regresión lineal para comprender la dirección y el impacto de cada función en la variable objetivo.
- Luego, se realiza un ordenamiento de los DF por migración neta en orden descendente y se dividen los 192 registros en 5 categorías, del 1 al 5, siendo 1 el grupo de registros con más migración neta. Esto proporcionará una comprensión más profunda de la migración en diferentes escenarios.
- Se hace mención que, en relación a este experimento, tenemos diferentes años de diferentes países en cada grupo. En esta instancia estamos tratando de entender la migración como un fenómeno, independientemente de las fronteras políticas humanas
- Esto es para probar nuestra hipótesis de trabajo la cual sigue el siguiente principio: las variables que explican un flujo neto positivo de migrantes son diferentes a las que explican un flujo negativo.*
- Realizamos un ordenamiento de los datos en estos 5 grupos y luego se ejecuta un SelectKBest al Grupo 1, 3 y 5 para analizar los resultados.
- Ahora, para cada uno de estos grupos, seleccionaremos las K mejores características y calcularemos los coeficientes.

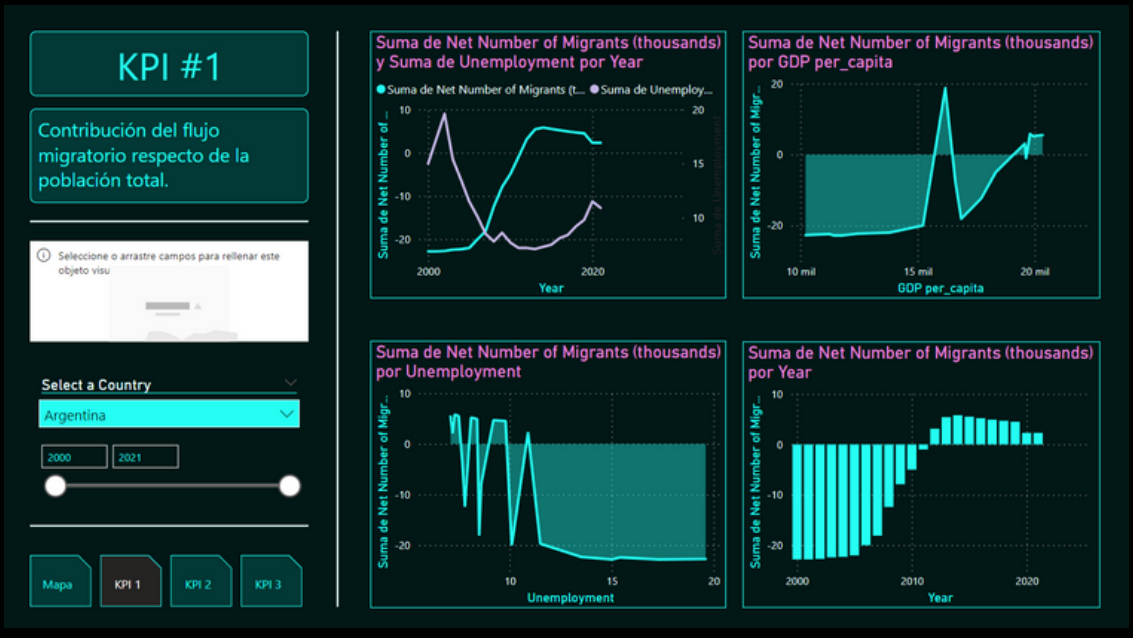
En relación al análisis aquí presentado, se confeccionará en la documentación final un abordaje integral respecto de las conclusiones/insights obtenidas.

VISUALIZACIONES

El uso de herramientas interactivas de visualización permite al usuario navegar de forma íntegra y sencilla los diferentes datos utilizados así como sus diferentes relaciones. 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 ®.

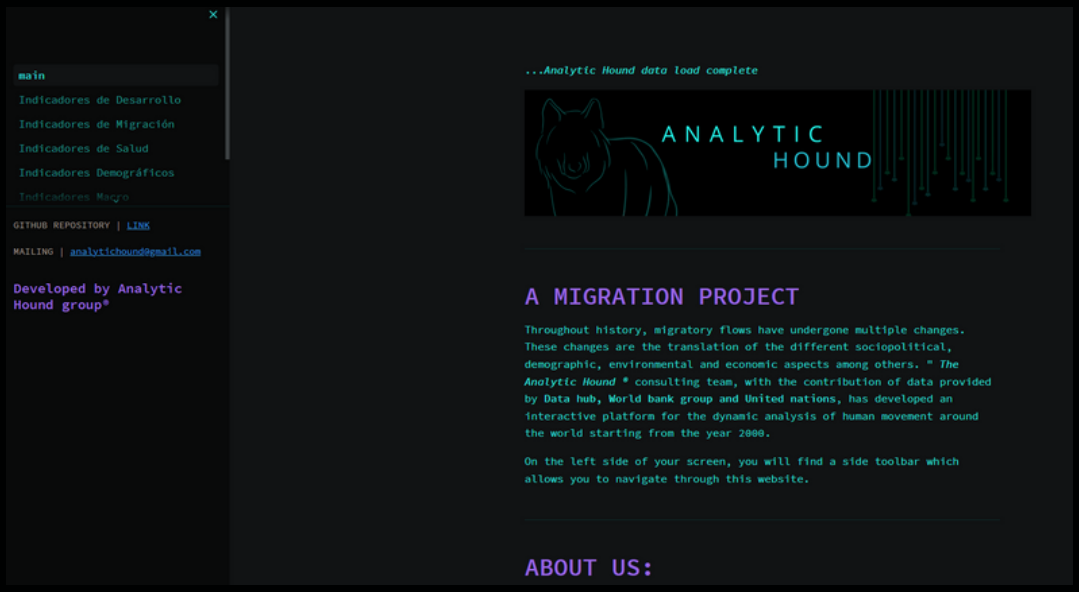
POWER BI:

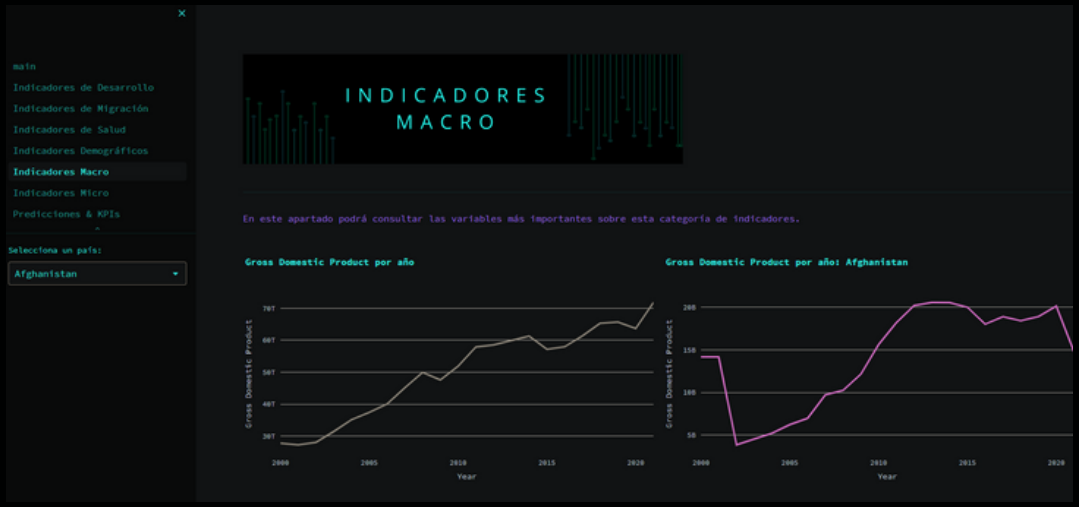


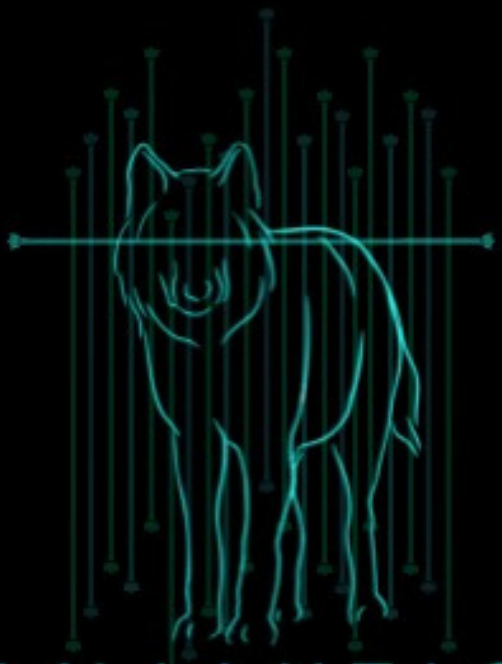


VISUALIZACIONES

STREAMLIT:







ANALYTIC

HOUND