ANALYTIC HOUND

ONG – **Soy Henry**

Report Week #3

DATA ANALYTICS

Milestones - Index

MACHINE LEARNING - LINEAR REGRESSION

Starting from the technical definition, the concept of Machine Learning (ML) focuses on the use and development of algorithms that use data to mimic the way humans learn, gradually improving their accuracy. Considering that for the purposes of this document, the development of this subject would be excessive or insufficient, a complementary reading material is suggested for those wishing to know more about this wonderful world.

In a schematic way and without intending to delve into what the subject of machine learning covers, as well as its applications by which it has managed to revolutionize the way of developing technology in recent years, a brief introduction to the model used for the project is presented here. of "Analysis of migration patterns worldwide" developed by Analytic Hound ®.

According to the type of label or output variable (commonly called "and"), supervised learning models can be classified into:

- Classification: in these models the label is a type of categories (eg sick/healthy, cat/dog/bird, spam/not spam)
- Regression: The output variable is a numeric value. (eg price, quantity, temperature)

In summary, supervised learning models use the different variables or features (only tolerated in numerical format) to determine the patterns (algorithms) that allow predicting a selected output variable or label.

MACHINE LEARNING - LINEAR REGRESSION

In general, it is considered good practice to allocate a considerable percentage (≈80%) of the data to model training and then the rest for model testing. All this happens under the premise of “The more training data, the better prediction”.

Finally, once the model has been developed, it can be evaluated using different error quantification formulas. Among them are the MAE (mean absolute error), MSE (mean square error), RMSE (square root mean square error) and R square (coefficient of determination) among others.

Having briefly commented on the details that make up the supervised ML models, in the following section we will limit ourselves to exposing the model developed for this project.

IMPLEMENTATION

The ML team proposes the "Net flow of migrants" as a prediction object. The desire and importance of knowing the projections of these future flows lies in the possibility that this offers different nations to adopt behaviors that can modify these values according to their internal needs.

Imagine a country that is going through a social crisis where it does not have enough "young" labor for the correct development of its economy. Knowing the range of actions that could be taken to modify this situation and the different degrees of impact that they could have is a highly valuable benefit for the correct management of resources in any region.

The preliminary development approach for the ML model consists of the following order of processes:

- Carry out tests with the data to study the relationships between our variables and compare the results.
- Run a SciKitLearn SelectKBest module. This function selects the top k features with the highest scores based on a scoring function which in this case will be "f_regression", which calculates the correlation between each feature and the target variable and returns an F-value and a p-value for each feature.
- SelectKBest then selects the top k features with the highest F values, indicating that they have the strongest linear relationship with the target variable.
- The process returns the following 5 variables... but how could one better understand the impact of each characteristic on the phenomenon of net migration?

IMPLEMENTACIÓN

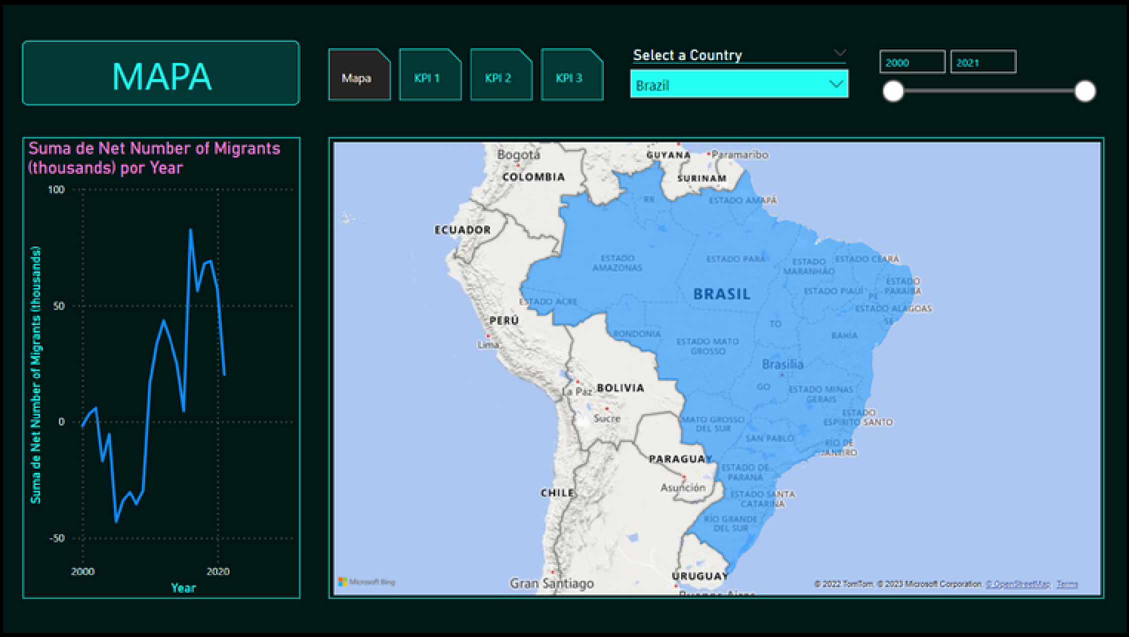
- We use the coefficients from the linear regression model to understand the direction and impact of each function on the target variable.
- Then, the DFs are ordered by net migration in descending order and the 192 records are divided into 5 categories, from 1 to 5, with 1 being the group of records with the most net migration. This will provide a deeper understanding of migration in different scenarios.
- Mention is made that, in relation to this experiment, we have different years from different countries in each group. In this instance we are trying to understand migration as a phenomenon, regardless of human political borders.
- This is to test our working hypothesis which follows the following principle: the variables that explain a positive net flow of migrants are different from those that explain a negative flow.*
- We sort the data into these 5 groups and then run a SelectKBest on Group 1, 3 and 5 to analyze the results.
- Now, for each of these groups, we will select the K best features and calculate the coefficients.

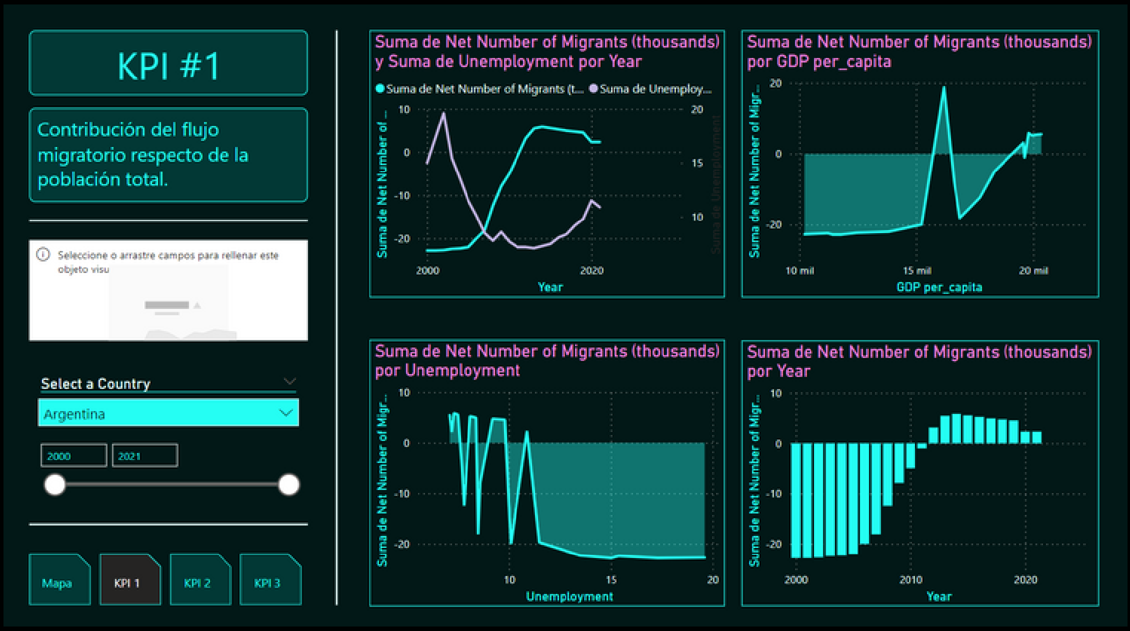
In relation to the analysis presented here, a comprehensive approach to the conclusions/insights obtained will be prepared in the final documentation.

VISUALIZATIONS

The use of interactive visualization tools allows the user to fully and easily navigate the different data used as well as their different relationships. The way in which these data are linked and the conclusions derived from them are the master key to our project. Below are schematically presented images corresponding to the visualizations developed by the Analytic Hound ® team of analysts.

POWER BI:





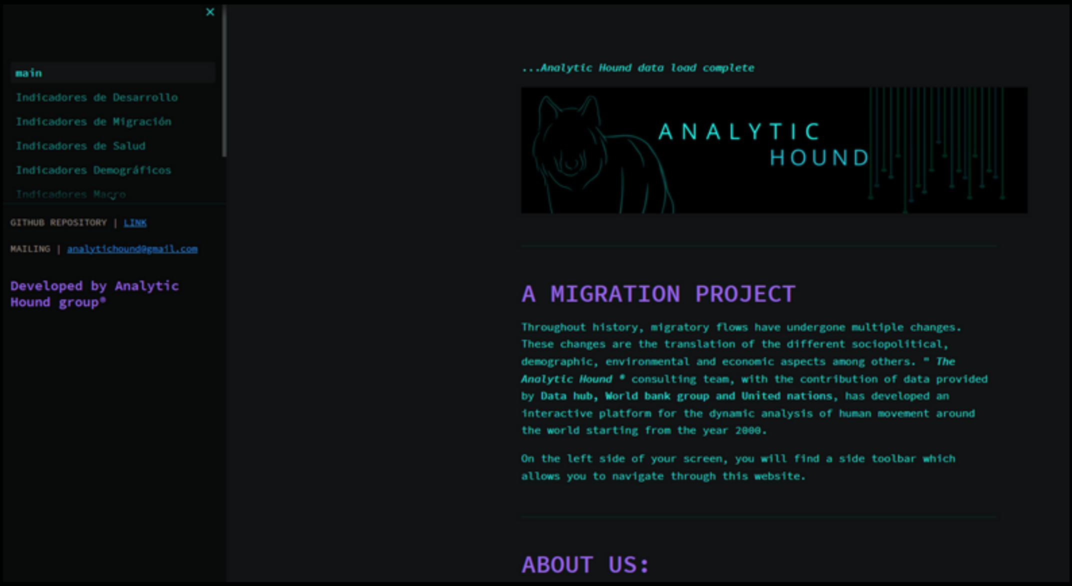
Contribución del flujo migratorio respecto de la población total.

Seleccione o arrastre campos para rellenar este objeto visu

Select a Country
Argentina
2000
2021
Mapa
KPI 1
KPI 2
KPI 3
Suma de Net Number of Migrants (thousands) y Suma de Unemployment por Year
● Suma de Net Number of Migrants (t... ● Suma de Unemploy...
10
0
-10
-20
20
15
10
Suma de Net Number of ...
Suma de Unemployment
2000
2020
Year
Suma de Net Number of Migrants (thousands) por GDP per_capita
20
0
-20
Suma de Net Number of Migr...
10 mil
15 mil
20 mil
GDP per_capita
Suma de Net Number of Migrants (thousands) por Unemployment
10
0
-10
-20
Suma de Net Number of Migr...
10
15
20
Unemployment
Suma de Net Number of Migrants (thousands) por Year
10
0
-10
-20
Suma de Net Number of Migr...
2000
2010
2020
Year

VISUALIZATIONS

STREAMLIT:



GITHUB REPOSITORY | LINK

MAILING | analytichound@gmail.com

Developed by Analytic Hound group®

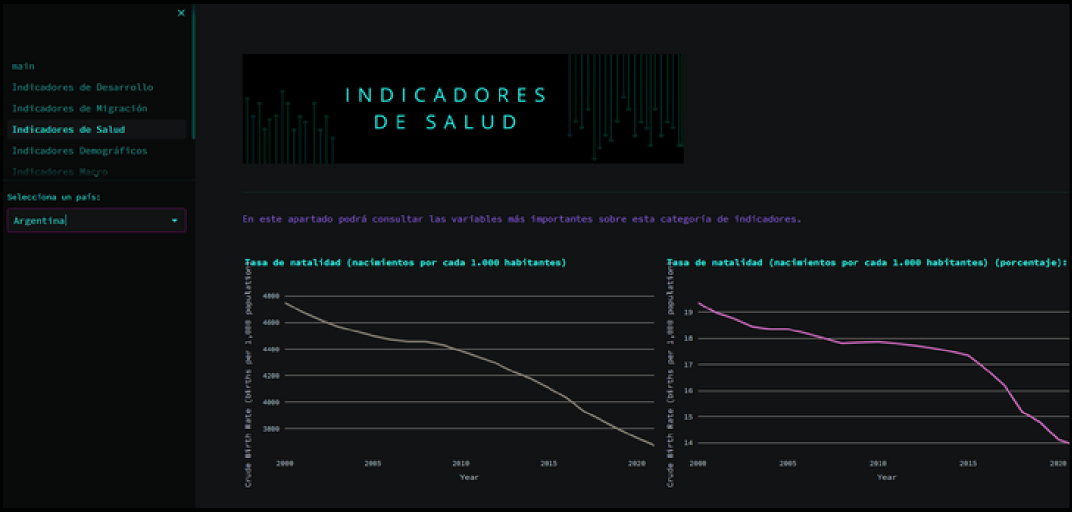
...Analytic Hound data load complete

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* ® consulting team, with the contribution of data provided by Data hub, World bank group and United nations, has developed an interactive platform for the dynamic analysis of human movement around the world starting from the year 2000.

On the left side of your screen, you will find a side toolbar which allows you to navigate through this website.

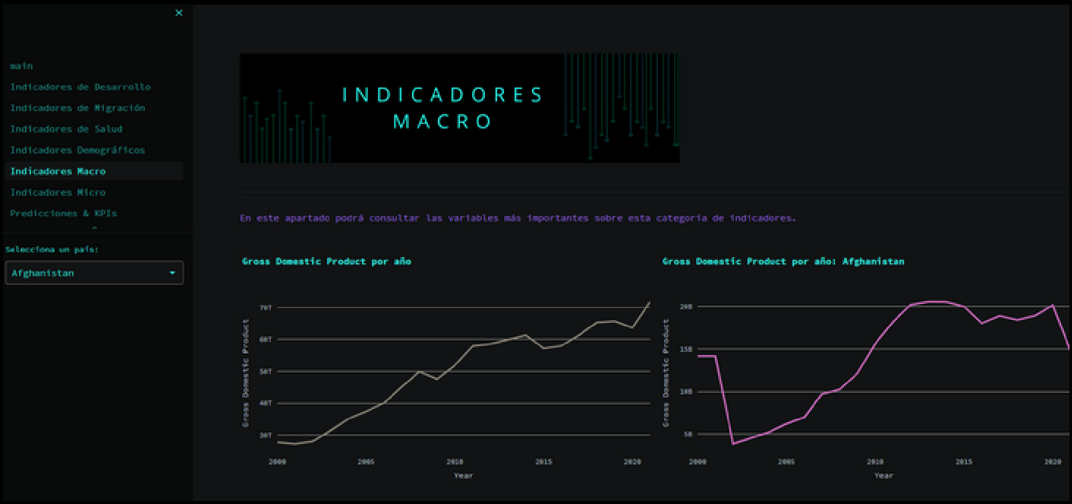
ABOUT US:


main
Indicadores de Desarrollo
Indicadores de Migración
Indicadores de Salud
Indicadores Demográficos
Indicadores Macro
INDICADORES
DE SALUD
Tasa de natalidad (nacimientos por cada 1.000 habitantes)
Crude Birth Rate (births per 1,000 population)
4800
4600
4400
4200
4000
3800
2000
2005
2010
2015
2020
Year
Tasa de natalidad (nacimientos por cada 1.000 habitantes) (porcentaje):
Crude Birth Rate (births per 1,000 population)
19
18
17
16
15
14
2000
2005
2010
2015
2020
Year

Selecciona un país:

Argentina

En este apartado podrá consultar las variables más importantes sobre esta categoría de indicadores.



Selecciona un país:

Afghanistan

En este apartado podrá consultar las variables más importantes sobre esta categoría de indicadores.

ANALYTIC

HOUND