Proposal of work

MIGRATORY FLOWS

International ONG Consultants

Consultants
Diego Marcelo Ledesma
Aylin Veronica Sequera Celis
Mónica virginia Ramos Dávila
Nelson Alejandro Castro Andrews
Daniela Berenice Contreras Villafuerte

INDEX

ACTUAL SITUATION

The current situation of migration and emigrants from 1990 to 2020 is complex and has evolved over the years. In general, there has been an increase in the number of people migrating from one country to another, whether for economic, political, or security reasons. However, the COVID-19 pandemic has significantly affected international mobility and has led to a decrease in the number of migrants.

Politically, many countries have implemented more restrictive government measures to control the entry and exit of people, which has made migration more difficult. In addition, increasing political polarization and anti-immigration discourse in many parts of the world have led to increased discrimination and exclusion of migrants.

Regarding the economic situation, globalization and economic growth in some countries have led to increased labor mobility, but the economic crisis caused by the pandemic has led to increased unemployment and decreased demand for migrant workers.

The current situation of migration and migrants is complex and constantly evolving, requiring a comprehensive and equitable solution that takes into account the rights and needs of all involved. The vast majority of people continue to live in the countries where they were born: according to a recent study, one in 30 people.

Generally, when addressing the issue of migration, the starting point is numbers. If we understand the changes in scale, the emerging trends and the evolution of demographic variables that bring about the social and economic transformations in the world caused by phenomena such as migration, we will better understand the changes in the world we live in and be able to better plan for the future. According to the latest estimates, in 2020 there will be approximately 281 million international migrants in the world, a figure equivalent to 3.6% of the world's population.

Globally, the estimated number of international migrants has increased over the past five decades. The estimated total of 281 million people living in a country other than their country of birth in 2020 is 128 million higher than the 1990 figure. Data on migration flows capture the number of migrants entering and leaving (inflows and outflows) a country during a specific period, e.g., a year; in our case, we capture part

of the current situation with data that come from 1990. These data are essential for understanding global migration patterns and how different factors and policies in countries of origin and destination may be related to these migration flows. There are really many causes and we understand that the context of the migratory flow is from the beginning of civilization, so it is complicated to understand the innumerable causes, we will focus more on establishing the existing correlations in each of the data that are the subject of this scientific analysis.

GENERAL GOAL

Therefore, the specific goal is to study and analyze migratory flows to take actions and opportunities to offer, with the staging of data science and thus fill a gap with the implementation of an API based on Machine Learning for both Institutions, organizations, and P2P.

SPECIFIC GOALS (KPIs)

Since data on migration flows are often incomplete and not comparable across countries, our main objective is to study and analyze the number of movements by linking changes in the data on the number of migrants over time.

Using Mathematical methods and several of its branches such as statistics, Computer Science IA which are fairly new fields such as Data Science, and its various methods, technologies procedures, implementation, and development, for which in this objective task we estimate the flows of migrants year by year, even in decades, which are required to address the differences in the migrant population totals.

For example, if the number of foreign-born in a region increases between two time periods, we estimate the minimum migration flows between that region. in all regions and all other countries of the world that are required to deal with this increase.

For each country, we estimate the minimum amount of migration flows needed to equalize the differences in stocks, assuming that people are more likely to stay than to move, thus making this information available on the web.

This estimation procedure is replicated simultaneously for all 196 countries. To estimate first the SWOT and each of the site-specific flow tables studied.

Resulting in a comparable set of global migration flows. Migrant population counts are modified to control births and deaths during the period and based on this we will publish in our web architecture an interactive Dashboard with state-of-the-art visualization protocols so that our clients and users, in general, have the best and most timely information.

This allows our country-specific net migration flows to closely match the net migration estimates published by the United Nations (as one of the sources of the authority of the chosen Datasets).

Also our API in the cutting edge Machine Learning field specifically by making use of AI we will be providing a high-level service.

KPIs

The term KPI, which stands for Key Performance Indicator, refers to a series of metrics used to synthesize information on the effectiveness and productivity of the actions carried out in a business in order to be able to make decisions and determine which have been most effective in meeting the objectives set in a specific process or project.

The key KPIs covered by our company are as follows:

- Total World Population by country between 1990 and 2020. **Metric:** Annual Population Ratio.
- Total World Population by region between 1990 and 2020. **Metric:** Annual Population Ratio.
- Percentage change in migrant population by country between 1990 and 2020. **Metric:** Migration Ratio.

- Behavior of net migration versus population growth rate, stagnation, or decline. **Metric:** net migration/population growth rate.
- GDP rate in destination countries with the highest growth per year of migrants. **Metric:** net migration/ GDP per capita growth rate.
- Assess the amount of income that can retain or expel the population of a country or region. **Metric:** Net migration/per capita income.
- Most popular country or region for migrants between 1990 and 2020. **Metric:** Migration Destination.

SCOPE AND ITS LIMITATIONS

SCOPE:

Within the scope under study, this will be: The Study and Analysis of Migratory Movement Flow applying ML within the set belonging to the IA where information management is crucial (Data Analytics, Data Science, Data engineering), Within the temporal scope, the years of analysis will be between 1990 - 2020. This is due to limitations in access to the data.

Make this information available on the web, adding technologies for Framework and Deployment, with sufficient documentation and visualization of the information collected, studied, and analyzed.

We will estimate the number of movements by linking changes in data on the number of migrants over time.

The shared database in the CSV and scraping on the white web and the deep web have data in the range shown. Furthermore, all sources are official. which provides authority and reliability in our study.

Developing an API in the cutting-edge field of Machine Learning, specifically using AI, we will be providing a high-level service.

Within the geographical scope, the behavior of this migratory phenomenon worldwide will be analyzed, and there will be an emphasis on the implementation of learning models that guarantee the accuracy of their answers as well as their classification.

LIMITATIONS:

One of our main limitations is the data on migratory flows, which is often incomplete and not comparable between countries.

It is not intended to create a cataloging or classification of migratory flows.

Lack of time for training the unsupervised model in relation to the delivery date of the developed project.

The years of analysis are between 1990 - 2020. This is due to limitations in access to data.

WORK METHODOLOGY
SCRUM

Scrum is an agile framework for product development that is primarily used in software development, but can also be applied to other projects and products. It is based on a collaborative, empirical approach to project management, where requirements and solutions evolve through dialogue and feedback from the cross-functional team.

Below we list some key elements of the Scrum methodology:

Scrum Team: a multidisciplinary team consisting of developers, product managers, and other stakeholders.

Product Backlog: a list of requirements and desired features for the product.

Sprint: a limited period of time, usually one to four weeks, during which the team works on a set of tasks from the Product Backlog.

Daily Scrum: a daily meeting in which the team meets to review progress and plan the next day's work.

Sprint Review: a meeting at the end of each Sprint in which the team presents what it has completed and receives feedback from the team and stakeholders.

Sprint Retrospective: a meeting at the end of each Sprint in which the team reflects on their process and decides how to improve in the next Sprint.

The goal of Scrum is to deliver value to the customer incrementally and continuously, improving collaboration and communication among team members and adapting to changing requirements and project circumstances.

We have implemented an organizational board in Trello and a workspace in MIRO where we have been iterating our ideas and shaping our proposal.

TECHNOLOGICAL TOOLS

No single piece of software or tool can achieve everything a company needs, which is why company **We flow together** takes advantage of numerous applications and creates a technology stack to achieve its business goals. This collection of tools, platforms, applications, and software pieces allows the company to create its products, carry out its business operations, and monitor performance indicators. All of this is known as a technology stack.

The technology stack of the company **We flow together** consists of the following tools:

- **Agile Methodology and Daily Work:** Trello, Miro, Gant Schedule, Google Docs, Google Meet, and Github.
- **Data Engineering:** Python and some of its libraries: Pandas, Numpy, Scikit-Learn, Geopy; International Business Machines Corporation (IBM), Airflow, MySQL, FastAPI, and Uvicorn, SQL.
- **Data Analysis and Visualization:** Python and some of its libraries: Hvplot, Holoviews, Panel, Seaborn, Matplotlib, NTLK, Scikit-Learn, Scipy, Wordcloud; Power BI.
- **Machine Learning Models:** Programming language Python and some of its libraries: Seaborn, NTLK, Scikit-Learn, Scipy. Supervised and unsupervised learning.
- **Web Environment:** For the frontend: bootstrap, html, css, javascript; and for the backend FastAPI.

TEAM ROLES

Having a solid team structure at the organizational level is essential for the success of the project. Therefore, tasks were assigned based on the team members' level of experience and profile, while the rest of the group works as assistants in other areas of the project. It is worth noting that the progress will be supervised by the entire group in order to ensure the correct functioning and progress of the project. For each stage of the project, roles and responsibilities have been defined as follows:

- **Product Manager:** In charge of Paula Villar, some of her responsibilities include: Following up on daily tasks, offering a possible solution when a task is blocked, coordinating meetings with the Product Owner.

- **Data Engineers:** In charge of Nelson Castro, Diego Ledesma and Aylin Sequera, some of their competencies include: Data acquisition, Developing data set processes, Preparing data for predictive and prescriptive models.

- **Data Analytics:** Responsible for Monica Ramos and Daniela Contreras, some of their duties include: Generating reports and evaluations obtained from the data, Developing comprehensive data analyses. Extracting, processing, selecting and grouping data for analysis**.**

- **Machine Learning Engineer:** In charge of Nelson Castro, performing the following activities: Designing and developing machine learning models, Analyzing data to select only valuable data and selecting the best methods to present it. Automating processes and implementing algorithms on machine learning.

GANTT TIMELINE

We rely on this Gantt chart which is a type of bar chart used to represent and plan projects. It shows the duration of individual tasks and their time relationship on a horizontal axis. Each task is represented as a bar indicating the time period in which it is expected to be completed.

The Gantt chart is useful for:

Visualize the project structure: allows you to see the tasks and subtasks and their relationship to time.

Identify dependencies between tasks: allows you to see which tasks must be completed before other tasks can begin.

Plan the use of time: allows estimating the time required to complete each task and the total time of the project.

Progress tracking: allows you to monitor the progress of tasks and adjust your planning accordingly.
-

EDA REPORT

This report aims to communicate the work carried out in the data analysis phase of our project.

In this report, we will establish, for each dataset, first, what is the quality of the data we used. Secondly, the criteria and tools used to overcome the challenges posed by the different data sets.

Datasets

As for the datasets we used, we will mention those from the World Bank (WB) and the UN Department of Economic and Social Affairs (International Migrant Stock 2020).

World Bank

Before we start detailing the work we did in the EDA, we must mention something of utmost importance about the structure of the World Bank datasets. These have values not only for countries but also for various categories of different country groups. Therefore, the original dataset had 266 records (one of which was null), divided as follows: 217 records of national states, 9 of geographical regions, 6 of country groups based on income level, 4 of countries grouped by dividends, 11 by geographic-credit group, 4 of small states, 13 of the "other groups" category, and one record for the "World" category.

The problem with this structure lies in being able to extract the information, either from the countries or from the groups of our interest, without that leading to multiplying the results. Moreover, at the beginning of the project, we did not have data that would direct our course of action. In other words, we still did not know if our work would focus on the migratory analysis of regions or countries.

Also, we understood that these data should either be filtered at some point or discarded. For these reasons, we decided to deal with this matter in a preliminary processing stage, to separate the country data on one side and the data of each of the groups on the other. This way, our original World Bank datasets were divided into seven groups, waiting for the EDA stage.

During this preprocessing stage, we divided the data groups, normalized the column names and text fields (by converting them to lowercase), changed the data types of numeric fields, and dropped columns from years before 1990 and columns containing both the code and the name of the indicator. It's important to note that this did not involve losing information on the indicator in question, as in a previous step, we renamed the year columns to have the

following format: indicator_year. For example, the 2000 column of the "Net migration" dataset became "net_migration_2000".

Once we analyzed the data, the team decided to work on two fronts. The first is to analyze the data from the countries. The second is to analyze the geographical regions' data.

To explore the following data sets, we decided to use the following graphical methods recurrently:

Since our period covers from 1990 to 2020, we chose to make histograms for the beginning of the period (1990-1992) and the end (2018-2020) to outline the general features of the distribution without having to do so for each year. Secondly, we opted for box plots for the entire period. On the one hand, they deepen the previous information. On the other hand, we can draw ten boxes next to each other, providing a lot of information from one decade in a single plot. Thirdly, we decided on a horizontal bar graph that shows us first, the top 10 countries, and second, the bottom 10. In the case of geographical regions, they are all included in a single graph.

The first dataset we had was called "Net migration" which only had one record and one null column. We decided to eliminate both the row and the field.
We understand net migration as the sum of all people who left a territory and all who entered it during a given period.

countries_net_migration:

After dividing it, we first analyzed the country dataset. It has 217 non-null records and 33 fields. The first two contain the country name and the other its code, according to the Nomenclature used by the WB; the others are the years of the studied period, which goes from 1990 to 2020.

Our exploratory data analysis began by creating a series of histograms, one for the first years of the studied period and the other for the last years. These graphs allowed us to see the distribution of migrations. Thanks to the plot, we noticed, first, that the distribution remained virtually unchanged. Second, the peak in the middle of the distribution looks higher in the latest years than in the first ones. Third, the tails of the histograms indicate that over time there seems to have been an increase in the number of migrants. Fourth, the number of net migrations to the right of the last graph has dramatically decreased. As for the latter, we believe it may be due to a delay in processing the data of the last of the years, something that usually happens in the data series of international organizations due to the large amount of data they process and the time it takes.

The next step we took was to analyze each year with box plots. Except for a single outlier in 1994, we observed a general trend toward increased migration. The values of countries with negative net migration seem to intensify. We also observed some cycles in the outlier values. In other words, we inferred an increase in migration over time, but with ups and downs within an upward trend.
Another analysis tool we used for this dataset was to create two rankings. The first one was a top ten ranking with the highest net migration. That is the ten countries with the highest

migration reception in the world. Here we find the constant presence of the United States as the country with the highest net migration in the world. Other States that appeared continuously in the 90s are the following: Germany, Russia, Arabia, Iran, and Canada. In the late 2010s, Spain, Colombia, and Syria joined this ranking continuously. Russia's presence is also constant.

Finally, in the countries with the highest negative net migration, there aren't many countries that appear constantly. Some of those that led the ranking in the 90s were: Pakistan, China, and Kuwait. Nowadays, the rankings are dominated by Asian countries. Towards the end of the 2010s, we see African countries appearing more consistently. One fact that caught our attention was the appearance of the Republic of Iran on both sides of the migration balance.

region_net_migration:

This dataset has a similar structure to the previous one but for the nine geographical regions of the World Bank Nomenclature. The regions are as follows:

1) East and Southern Africa
2) West and Central Africa
3) East Asia and Pacific
4) Europe and Central Asia
5) Latin America and the Caribbean
6) Middle East and North Africa
7) North America
8) South Asia
9) Sub-Saharan Africa

The quality of this set is undoubted, as it does not present null values. Another thing to mention is that it consists of ten original records, although we removed the one named "World" and the same 33 fields mentioned in the previous dataset.

As we were working with a few registers, the histograms showed that there are regions that are purely a source of migrants, while others are receptors. We observed that some have a very negative net migration while others have a very positive one.
In the histograms of the end of the period, we also noticed a "stretching" of the distribution both to the right and to the left, which speaks of growth in the magnitude of migrations. It turns out that the boxplots reaffirm that fact. The extreme values and the box whiskers far from the mean show that there are regions at the extreme ends of the migration balance.

Finally, we sorted the data in descending order for three periods. First, in the early 90s, second, in the mid-2000s, and third, in the late 2010s. That allowed us to identify the different regions. We observed that except in the 2000s decade, North America was the region with the highest balance. While on the negative side, South Asia, Latin America, Sub-Saharan Africa, and "East Asia, and the Pacific" are the regions that compete for the bottom of the graph.

"Datasets 'Population' & 'Countries_Population'

“countries_population”

This dataset includes information regarding the total population of all countries. It has a similar structure as the first dataset and, just like it, has no missing records.

The histograms from the early 1990s exhibit a skewed distribution to the right. Also, there are nearly 200 countries located on the left side of the graph, with two outliers on the right side and the rest of the countries being positioned between the lower population range and 300,000 residents. By the end of the analyzed period, the histograms appeared almost identical with the only difference being the overall increase in population.

The box plots confirmed that the entire period showed a similar trend. Another finding from this analysis was that the second outlier was trending towards the value of the first outlier, meaning that the second most populous country was growing faster than the first.

From the bar graphs, we identified that the two exceptional values correspond to China and India. Meanwhile, the United States has sustained its position as the third most populous country. Additionally, Russia is losing its position as one of the most populous nations globally, just like Brazil, which dropped one position in the ranking to Pakistan. Furthermore, Nigeria emerged as the first African country in this classification. To sum up, the demographic ranking is dominated by Asian countries."

"region_population"

This dataset comes from the same original dataset as the previous one but records geographical regions. It has the same structure as all the BM's geographical region sets (9x33 after removing the "world" row).
The histograms allowed individualization of each region, although we couldn't identify them. Perhaps with so few records, this is not the best choice as the information it contains is redundant with the ranking, and there are not enough records to analyze the distribution.

On the other hand, the boxplots have shown great information. Firstly, in the boxplots of the 1990s, we see a single outlier and a tendency for the whiskers to lengthen what speaks of acceleration. Towards the end of the 2000s decade, this lengthened whisker gave rise to an outlier, which then reversed in the mid-2010s due to the growth of central values. That could speak of a sharp increase in the total population. However, we do not rule out the possibility of a plotting error since the central measures do not grow for two decades. We will analyze that in a deeper analysis stage.

As for the ranking by largest population, as we saw above, Asian countries have the largest population. For this reason, it is not surprising that the three regions with the largest world population are "East Asia and the Pacific," "South Asia," and "Central and Asia Europe," in that order; however, the latter has lost the third place to the Sub-Saharan Africa region. Latin America and the Caribbean, which used to be the fifth most populated region in the 90s,

have been surpassed by the Eastern and Southern African region. North America, which once was the third least populated region in the 90s, has fallen to the last place after being surpassed by the "Middle East and North Africa" and "West and Central Africa" regions.

"GDP per capita (constant 2015) Datasets"

This dataset contains data about the income measured in US dollars per capita of a country or region.

Dataset "countries_gdp_cap"

So far, this was the dataset with the lowest quality of data, presenting missing values in all year columns, with a maximum of 23% in 1990 and a minimum of 3% in 2015. Our first option is not to delete the missing values during the EDA phase. We are waiting to obtain data from reliable sources that allow us to replace the missing values. Otherwise, we will impute values in a later stage.

The histogram graphs showed that in the 90s decade, distributions skewed to the right, with a cluster of about 120 countries in the lowest per capita income zone in the world some countries, we estimate about 35, in a second zone that does not reach 30,000 dollars per capita, then about 25 in the 30,000-40,000 USD per capita zone, and a few more to the right, and with the highest per capita income exceeding 100,000 USD.
In the last years of our period, we noticed that some countries joined the lowest income zone. Also, we observed that others have moved to the right. That is, they have increased their income per capita. The highest values are nearly 200,000 dollars.

As for the box plots, they do not yield much information about what is happening in the central zones of the distribution. However, we can observe that the box sizes have grown over the years. That may be due to either some countries moving from a medium-income zone to a high-income or perhaps the appearance of more outliers and the growth of existing ones pulling the central tendency measures. What is completely clear is that the per capita income of some countries has skyrocketed.

Bar charts provided a lot of information as they showed the names of countries with the highest per capita income. At the beginning of the 90s, we could see that many states leading these rankings were small. The leader was Monaco, Bermuda second, and Switzerland and Luxembourg shared third place. Other countries that constantly appeared were: Australia, the United States, Norway, Denmark, and Brunei.

Towards the end of the 2010s, Monaco, Bermuda, Luxembourg, and Switzerland continued to appear in the same places. But we also witnessed the rise of a new wave of wealthy small states, such as Qatar, the Cayman Islands, the Isle of Man, Macao, and Singapore. The United States only appeared in 2020 in tenth place. In addition, the presence of Ireland in the chart was a constant in recent years.

As for the countries with the lowest per capita income, we can mention the following in the 90s: Myanmar, Mozambique, Ethiopia, Malawi, Burkina Faso, Rwanda, Uganda, Niger,

Nepal, Equatorial Guinea, and Burundi. As we can see, there were mainly African countries added to some Asian ones.
The list of countries at the end of the 2010s was the following: Burundi, Central African Republic, Malawi, Somalia, Madagascar, Democratic Republic of the Congo, Niger, Afghanistan, Togo, Mozambique, and Sierra Leone. As we detected, some names changed, but Africa continued to have the poorest countries in the world, and in this case, we also observed the presence of some Asian states.

UN International Migrant Stock 2020

The UN dataset has a similar structure to the one of the World Bank. It is divided by continents, geographical regions, economic regions, and countries.
It contains information about the migration flow between 234 countries, 6 continents, 10 economic regions, and 23 geographical regions. That flow is analyzed between the years 1990 and 2000, divided by five-year periods.
In order to clean the dataset we had to normalize it, which included reorganizing and renaming the columns, which were disorganized by the process of importing the data. We also had to reorder the values within the columns, save null values, and remove duplicates. The columns 'Index', 'Destination Notes', and 'Destination_Type_of_Data' were removed as well since they didn't bring anything useful to the analysis for being repetitive or containing a majority of null values. During this process, we were able to preserve the integrity of all useful data.

Due to the number of registered countries, we've decided to take the ones we consider most relevant in order to visualize the data and get a preliminary idea of the tendencies. Throughout this project, we will deepen the analysis of those countries. We did the same for the geographical regions.

We decided to divide the UN dataset into 7 subsets in order to make a better analysis of it. It is now as follows:

1) Countries
2) Continent of origin
3) Continent of destination
4) Economic region of origin
5) Economic region of destination
6) Geographical region of origin
7) Geographical region of destination.

We used bar charts to analyze the countries and the geographical regions, and line charts for the continents and economic regions. We chose those visualization methods for their clarity and simplicity. They allowed us to see at first glance the changes in migration flows between different places and through time. This showed a noticeable growth in the migrant population over time.

Countries

For the first visualization of the data, we decided to take Rwanda, Israel, Spain, and Mexico as countries of origin. In those cases, we observed a tendency that will be corroborated by the analysis of the other datasets: the lesser economically developed countries are the ones with the greater number of emigrants. It is very noticeable the growth of emigration from Rwanda and the decrease of it in Spain and Israel. This, however, doesn't affect the general tendency. México remains the country with the greatest number of emigrants.

Continents

In this case, we can see the growth of North America as a receptor of immigrants. Europe, Africa, Oceania, and Latin America and the Caribbean remain fairly constant. Asia is the continent that fluctuates the most in the timeline. It also has the greatest number of emigrants, followed by Europe. According to the available documentation, this may be due to the fact they both experience a significant intraregional migration, as does Africa, which holds third place in this ranking. In this dataset, we can also see that as the number of migrants grows globally, so does the absolute number of migrants from each of these continents, but the tendency and the proportion stay the same.

Economic regions

This dataset is particularly interesting, for it reveals a completely different perspective compared to the others.
The phenomenon registered here is this: the regions with less economic power are the ones that have a greater number of immigrants. This could be caused by intraregional migration.

Another interesting phenomenon is the proportional decrease in migration to most developed regions between 1990 and 1995, a period in which the number of migrants worldwide almost doubled. These new waves of migrants seem to choose lesser-developed regions as their destination.

Geographical regions

As we mentioned before, geographical regions reflect the behavior of the continents, with Europe and North America being the ones that receive the largest number of people, and Latin America and the Caribbean the ones that receive the fewest.

On the other hand, South East Asia doesn't experience noticeable changes, due perhaps to the restrictive immigration policies of most of the countries that conform it. But it does show an increase in the number of emigrants. Europe and the United States are the ones with the greatest number of emigrants, possibly due to Europe's internal migration.

This is a preliminary analysis and its conclusions need to be revised by further study of the available data.