

Understanding Brain Drain: A Data-Driven Country Recommendation System

Borgia Mattia 853236
Lupinetti Filippo 932984
Monacis Matteo 925759
Paride Russo 853230

Data Science Lab – Academic Year 2024/2025

Abstract

The international migration of highly educated individuals—commonly referred to as *brain drain*—is reshaping labor markets, innovation ecosystems, and economic development worldwide. This creates a dual challenge: while origin countries often experience reduced innovation capacity and economic setbacks (business problem), there is still no data-driven, interactive system to guide individuals and policymakers in understanding or managing these flows (business need).

This project addresses the gap by integrating OECD migration flow data (over 43 million tertiary-educated individuals across 1,166 origin–destination–gender combinations) with standardized well-being metrics and political freedom ratings. Leveraging this enriched dataset, we developed an interactive analytical platform that enables users to:

- Analyze and visualize international migration patterns by origin, destination, and gender.
- Compare countries across socio-economic and political indicators on a normalized scale.
- Identify groups of countries with similar profiles through clustering and dominance analysis.
- Receive personalized migration recommendations tailored to individual priorities.

The methodology combines data fusion, feature engineering, dimensionality reduction, and ranking algorithms into a modular pipeline accessible via a Streamlit web application. The resulting tool supports both policy-oriented structural analysis and individual-level decision-making, offering a reproducible, data-driven perspective on global talent mobility.

Contents

1	Introduction	4
2	Methodology	5
2.1	Workflow Overview	5
2.2	Environment Setup	5
2.3	Migration Data Preparation	5
2.4	Enrichment with Country-Level Indicators	5
2.5	Dataset Fusion and Feature Engineering	6
2.6	Final Dataset	6
3	Exploratory Data Analysis	7
3.1	General Overview	7
3.2	Top Migration Routes	7
3.3	Discussion and Implications	8
4	Country Comparison Interface	9
4.1	Purpose and Applications	9
4.2	Methodological Pipeline	9
4.3	Visualization and Results	10
4.4	Conclusion	10
5	Clustering Analysis	11
5.1	Purpose and Applications	11
5.2	Methodological Pipeline	11
5.3	Visualization and Results	12
5.4	Conclusion	12
6	Dominance Visualization (Hasse Diagram)	13
6.1	Purpose and Applications	13
6.2	Methodological Pipeline	13
6.3	Visualization and Results	14
6.4	Conclusion	15
7	Recommendation System	16
7.1	Purpose and Applications	16
7.2	Methodological Pipeline	16
7.3	User Profile and Results	17
7.4	Conclusion	18
8	Interactive Web Application	19
8.1	User Interface Overview	19
8.2	Core Functionalities	19
8.3	Technological Stack	20
8.4	Deployment and Accessibility	20
8.5	Advantages of the Interactive Approach	20
8.6	Future Developments	20

9	Conclusion	21
9.1	Summary of Contributions	21
9.2	Reflections, Limitations, and Future Development	21
9.3	Final Remarks	22

1 Introduction

The migration of highly educated individuals—commonly referred to as *brain drain*—has become a defining feature of global labor mobility in recent decades. According to OECD data, more than 43 million tertiary-educated people have migrated between 1,166 distinct origin–destination–gender combinations during 2020–2021, reflecting the complex interplay of personal aspirations, socioeconomic conditions, and political climates.

From a structural perspective, the redistribution of human capital creates a critical **business problem**: countries of origin often face long-term developmental setbacks, reduced innovation capacity, and slowed economic growth, while destination countries benefit from productivity gains, expanded knowledge bases, and global competitiveness. This asymmetric exchange raises pressing questions about equity, sustainability, and the long-term societal impact of skilled migration.

At the same time, there is a clear **business need**. Despite the scale and consequences of brain drain, individuals and policymakers still lack accessible, data-driven tools capable of translating multi-dimensional indicators into actionable insights. Existing approaches remain fragmented, focusing on isolated variables or static reports, without supporting personalized decision-making. This project is guided by a central research question:

What factors drive highly educated individuals to leave their home country, and which destinations offer the most attractive opportunities in return?

To address this question, we developed an interactive analytical system that integrates international migration flows with quality-of-life indicators and measures of political freedom. The platform focuses on three core dimensions:

- **Quality of life**: education, employment, housing, healthcare, and safety.
- **Political freedom**: civil liberties and political rights.
- **Socioeconomic context**: income, access to services, and social cohesion.

Beyond the immediate perception of loss, brain drain can in some contexts evolve into a form of *brain gain*. Countries with flexible education and training systems may adapt to talent outflows by upskilling their population, while diaspora networks, return migration, and remittances often contribute to entrepreneurship, knowledge exchange, and long-term development at home.

Our platform therefore serves two complementary purposes: it supports individual decision-making through personalized recommendations, and it provides insights for policymakers seeking to interpret or manage skilled migration trends. Ultimately, this project frames brain drain not merely as a demographic phenomenon, but as a measurable, interpretable decision-making process that can reveal both risks and opportunities for societies.

2 Methodology

The development of our migration analysis and recommendation system followed a structured, multi-phase workflow designed to produce a reliable, consistent, and interpretable dataset. Each stage of the pipeline was guided by two priorities: (i) ensuring analytical rigor through well-motivated methodological choices, and (ii) preserving interpretability to support both exploratory and user-specific analyses.

2.1 Workflow Overview

The complete pipeline comprised the following main stages:

1. **Migration data preparation:** filtering, cleaning, and standardizing raw OECD flows.
2. **Enrichment:** integrating national-level socioeconomic and governance indicators.
3. **Feature engineering:** computing origin–destination differentials (push–pull factors).
4. **Normalization:** rescaling indicators to a common $[0,1]$ range for comparability.
5. **Final dataset construction:** producing a structured table powering all downstream modules.

This harmonized dataset underpins descriptive visualizations, clustering, dominance analysis, and the personalized recommendation engine.

2.2 Environment Setup

The analysis was conducted in Google Colab for its reproducibility, collaboration features, and Google Drive integration. Raw datasets were stored in a shared Drive and copied to the local runtime for faster access. Data processing used `pandas` and `numpy`, visualizations combined `matplotlib` and `seaborn` with interactive `plotly` charts, and analysis employed `scikit-learn` (k-means, PCA, scaling) and `networkx` with `graphviz` for dominance graphs. Interactivity was added via `ipywidgets`, keeping data processing, visualization, and modeling clearly separated.

2.3 Migration Data Preparation

The **OECD Labor Force Surveys** provide microdata on individuals aged 15+, classified by country of birth and destination, sex, education level, and number of migrants. Since the focus is on *brain drain*, we restricted the analysis to individuals with tertiary education (ISCED levels 5–8), capturing the mobility of highly skilled professionals, researchers, and graduates.

We assembled the dataset by merging five OECD tables (T1, T3, T4, T5, T6) using the triplet (`origin`, `destination`, `sex`) as a primary key. Country identifiers were standardized to ISO 3166-1 alpha-3 codes (e.g., ITA, FRA) to ensure compatibility with external indicators. Flows with missing identifiers or zero migrant counts were discarded, prioritizing data quality over absolute coverage. The result, stored as `df_spostamenti`, represents the cleaned and standardized base for subsequent enrichment.

2.4 Enrichment with Country-Level Indicators

To provide context beyond raw migration counts, we enriched the dataset with socioeconomic and political indicators.

The **OECD Well-being Framework** measures national performance across eleven dimensions—Education, Jobs, Income, Health, Housing, Safety, Civic Engagement, Environment, Community, Life Satisfaction, and Accessibility to Services—originally on a 0–10 scale. These

values were later rescaled to the $[0,1]$ range for analytical consistency, and missing national values were imputed from regional averages where available.

Political freedom was measured using the **Freedom House Reports**, which score Political Rights (PR) and Civil Liberties (CL) from 1 (most free) to 7 (least free). To align these with the well-being indicators, we inverted and normalized the scales so that higher values consistently represent better conditions.

This produced `df_nazioni`, where each row corresponds to one country with its standardized identifiers and normalized indicators.

2.5 Dataset Fusion and Feature Engineering

To combine the datasets, we duplicated `df_nazioni` into two copies: one for origin countries (`df_origin`) and one for destination countries (`df_dest`), renaming all fields to indicate their role (e.g., `Health_origin`, `Health_dest`). The cleaned migration records were merged first with the origin indicators and then with the destination indicators.

After merging, intra-country flows were removed, as they do not represent international migration. All continuous variables were normalized to a $[0,1]$ range using min-max scaling:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This choice preserves relative distances between countries while ensuring that each indicator contributes equally to aggregated metrics, an essential condition for fair comparisons in clustering and scoring.

A central feature engineering step was the computation of indicator differentials (Δ_i) between destination and origin:

$$\Delta_i = \text{Indicator}_{\text{dest},i} - \text{Indicator}_{\text{origin},i}$$

These deltas quantify the potential gain (positive) or loss (negative) in each dimension when moving from one country to another, operationalizing the push-pull factors that underpin our recommendation logic.

2.6 Final Dataset

The resulting `df_final` dataset integrates all preprocessing, enrichment, and feature engineering steps into a single, structured table. Each record corresponds to a unique migration flow of tertiary-educated individuals and contains:

- **Origin and destination identifiers:** ISO 3166-1 alpha-3 codes and full country names.
- **Sex of migrants:** Male or Female.
- **Migrant count:** total number of individuals in the flow.
- **Normalized indicators (origin and destination):** socioeconomic and political scores rescaled to $[0,1]$.
- **Indicator differentials (Δ_i):** destination value minus origin value for each indicator.

In total, the dataset includes **1,166 valid origin-destination-sex combinations** and more than thirty numerical features per record. This multidimensional structure supports all downstream analytical tasks, including descriptive visualizations, clustering, dominance analysis, and personalized recommendations.

3 Exploratory Data Analysis

Before introducing algorithmic modules or interactive visualizations, we conducted an exploratory data analysis (EDA) to identify structural patterns in the migration dataset. This phase served two purposes: validating the integrity of the preprocessed data and establishing a baseline for interpreting global migration trends.

Using descriptive statistics, aggregated summaries, and visual inspection, we examined the distribution of migration flows, demographic composition, and the relative performance of countries across quality-of-life and political indicators. These insights informed the feature selection and design choices for the clustering, dominance, and recommendation modules.

3.1 General Overview

The final dataset comprises **1,166 distinct migration flows**—unique combinations of origin, destination, and sex—representing more than **43 million tertiary-educated individuals**. This scope offers a representative snapshot of global talent mobility.

Figure 1 shows the distribution of migrants by sex, which is slightly skewed toward female movers. This aligns with recent trends showing greater female participation in tertiary education and international labor mobility, confirming that brain drain is not gender-exclusive.

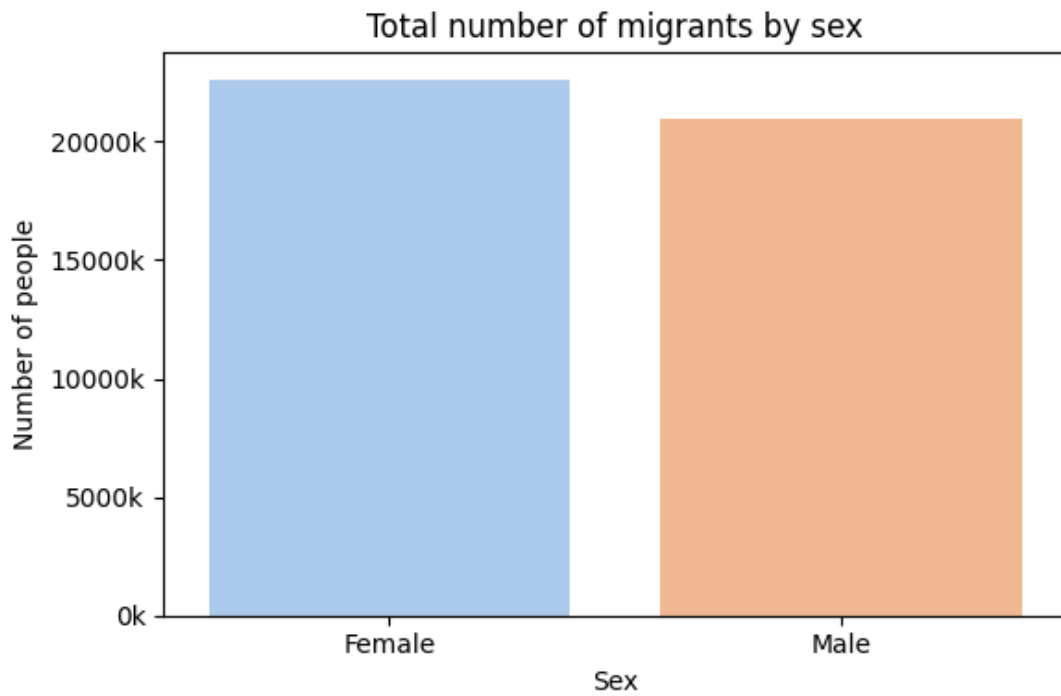


Figure 1: Total number of migrants by sex.

3.2 Top Migration Routes

To identify dominant patterns of brain drain and gain, we aggregated flows by country, ranking nations by total incoming and outgoing tertiary-educated migrants (Figures 2 and 3).

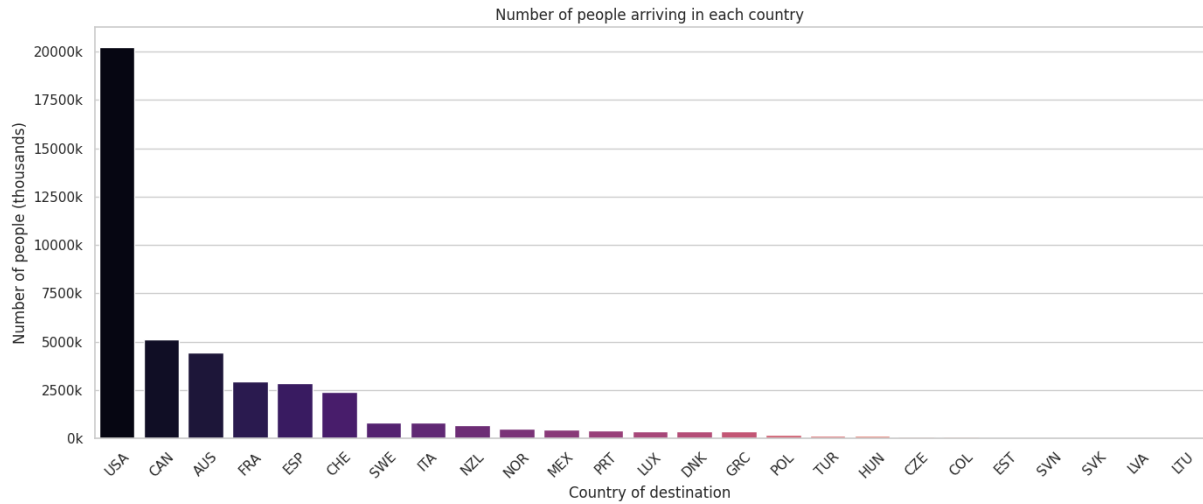


Figure 2: Top destination countries ranked by total incoming tertiary-educated migrants.

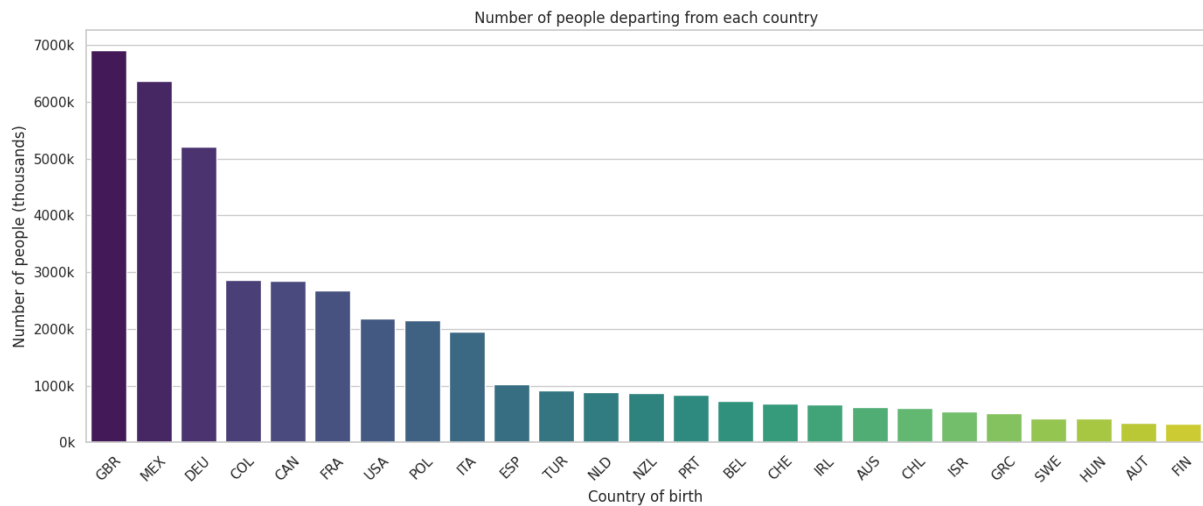


Figure 3: Top origin countries ranked by total outgoing tertiary-educated migrants.

United States, Canada, and Australia lead as destinations, attracting talent through strong labor markets, high living standards, and English-speaking environments. Conversely, **Mexico, the United Kingdom, and Germany** appear among top sources of outbound talent, showing that even advanced economies can experience significant brain drain when local conditions—such as market saturation, high costs, or limited research opportunities—push skilled individuals abroad.

3.3 Discussion and Implications

The exploratory data analysis confirms the robustness and analytical potential of our dataset, revealing key insights. Migration flows follow structured global patterns shaped by persistent socioeconomic disparities and governance differences. No single factor explains these movements: economic opportunities, institutional strength, gender dynamics, cultural proximity, and sector-specific labor demand often interact. Even high-income nations are not immune to *brain drain*; under unfavorable economic or political conditions, they too may face net losses of skilled talent. These findings provide a concise overview of current brain drain dynamics and inform later modules, supporting the integration of multi-dimensional indicators, the inclusion of sex as a variable, and the interpretation of recommendation outputs.

4 Country Comparison Interface

The country comparison module provides a concise yet comprehensive way to evaluate differences between nations across multiple well-being, governance, and economic dimensions. Rather than relying on isolated indicators or static rankings, it offers an integrated visual framework where strengths, weaknesses, and trade-offs become immediately apparent.

By transforming multi-variable datasets into clear comparative graphics, the interface supports both quick overviews and targeted domain-level assessments. This allows users to move beyond intuition or anecdotal impressions, grounding migration-related evaluations in transparent, data-driven evidence.

4.1 Purpose and Applications

The country comparison module addresses a central challenge in migration analysis: evaluating destinations in a way that goes beyond fragmented rankings or anecdotal impressions. By condensing complex multi-variable datasets into transparent visualizations, it highlights both absolute performance and relative trade-offs across well-being and governance dimensions.

This functionality is particularly relevant in the context of brain drain, where origin countries risk losing talent to systematically stronger destinations. At the same time, it provides a benchmarking tool for policymakers and a research instrument for exploring the structural drivers of migration.

Main applications include:

- **Migrants** — compare potential destinations according to personal priorities.
- **Policymakers** — benchmark national performance against peers to identify reforms that strengthen attractiveness.
- **Researchers** — investigate relationships between structural indicators and skilled migration flows.

4.2 Methodological Pipeline

The comparison process is designed to be transparent, replicable, and responsive to user input:

- **Selection of indicators of interest** — The user chooses one or more variables from the normalized dataset, ensuring that all scores are on the $[0, 1]$ scale for consistent cross-country comparability.
- **Definition of the countries to be evaluated** — Two countries are selected through interactive dropdown menus, allowing quick adjustment of the comparison scope.
- **Extraction of relevant data** — The system retrieves the corresponding indicator values from the processed dataset, guaranteeing accuracy and consistency.
- **Generation of the comparative visualization** — A grouped bar chart is created, where each bar represents the score of a specific indicator for one country, making differences immediately visible.
- **Instant update of results upon changes** — Any modification in indicator or country selection triggers an automatic refresh of the chart, enabling iterative exploration without manual re-execution.

4.3 Visualization and Results

As an example, we compare Italy and France across seven user-selected indicators, with all values rescaled to the $[0, 1]$ range for direct comparability (Figure 4).

France shows higher scores in most domains, particularly *Education*, *Jobs*, *Income*, *Housing*, and *Access to Services*. Italy matches France closely in *Safety*—even slightly exceeding it—and remains competitive in *Life Satisfaction*, despite lower performance in economic and infrastructure-related indicators.

This comparison illustrates how the tool reveals both clear disparities and nuanced strengths, supporting balanced, evidence-based migration decisions.

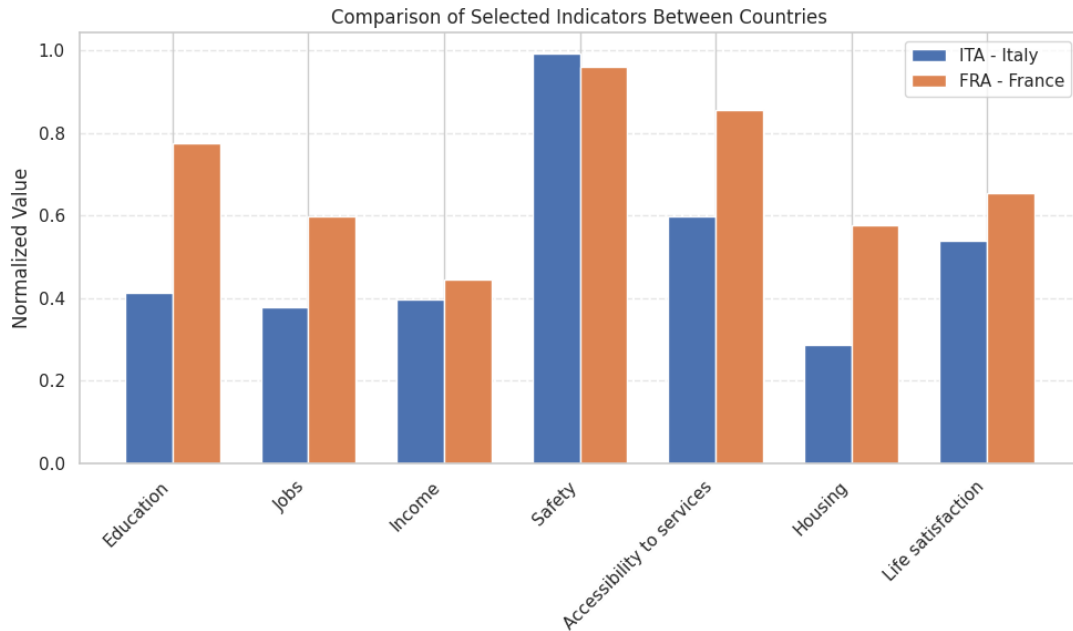


Figure 4: Comparison of selected indicators between Italy and France (normalized scores).

4.4 Conclusion

The country comparison interface transforms complex, multi-dimensional datasets into clear, accessible visual insights that make cross-country differences immediately understandable. By enabling fully customizable, indicator-level evaluations, it allows users to highlight specific domains of interest, compare structural strengths and weaknesses, and assess trade-offs between potential destinations. This functionality not only supports more informed migration-related decisions but also integrates seamlessly with the clustering, dominance, and recommendation modules, ensuring that descriptive comparisons feed directly into the platform’s broader analytical and decision-support framework.

5 Clustering Analysis

To go beyond individual country comparisons, we developed a clustering module that groups nations according to their structural similarities in well-being and governance indicators. Unlike single-variable rankings, clustering considers multiple dimensions simultaneously, revealing *latent patterns* in the data—countries that may be geographically distant or culturally distinct but share comparable living conditions and political frameworks.

By uncovering these relationships, clustering serves both descriptive and exploratory purposes. It helps interpret migration flows in terms of groups of structurally similar destinations and offers a richer context for identifying potential migration alternatives beyond the most obvious choices.

5.1 Purpose and Applications

The clustering module creates structural “neighborhoods” of countries, grouping them by multi-dimensional similarities in well-being and governance. This approach goes beyond single-country comparisons, revealing alternative destinations that may not be obvious but share comparable socio-economic profiles.

In the brain drain debate, clustering helps identify not only the top destinations that attract skilled workers, but also viable second-tier options. For policymakers, it highlights peer groups with similar challenges and opportunities, enabling coordinated strategies to retain or attract talent.

Applications include:

- **Migrants** — discover alternative destinations that align with their preferences.
- **Policymakers** — situate their country within structural clusters and design peer-based reforms.
- **Researchers** — explore links between cluster membership, migration flows, and long-term development.

5.2 Methodological Pipeline

The clustering procedure follows a structured, data-driven sequence to ensure both interpretability and analytical rigor:

- **Identification of relevant dimensions** — The user selects at least two indicators from the normalized dataset ($[0, 1]$ scale), ensuring comparability across all variables considered.
- **Preparation of variables for analysis** — Selected indicators are standardized to have zero mean and unit variance, preventing scale differences from biasing the clustering process.
- **Reduction of dimensionality when necessary** — If more than two indicators are chosen, Principal Component Analysis (PCA) projects the data into two components, retaining maximum variance while simplifying visualization.
- **Application of clustering algorithm** — The k-means method partitions countries into groups, with the optimal number of clusters determined by maximizing the silhouette score within the range $k \in [2, 9]$.
- **Profiling and qualitative interpretation** — For each resulting cluster, average indicator scores are calculated and classified into performance categories (High ≥ 0.7 , Medium $0.4\text{--}0.7$, Low < 0.4) to facilitate interpretation of structural characteristics.

5.3 Visualization and Results

The module outputs a two-dimensional scatter plot where each point represents a country, labeled by ISO code and color-coded by cluster membership (Figure 5). The axes correspond to the first two principal components from PCA, combining the selected indicators into orthogonal dimensions of maximum variance.

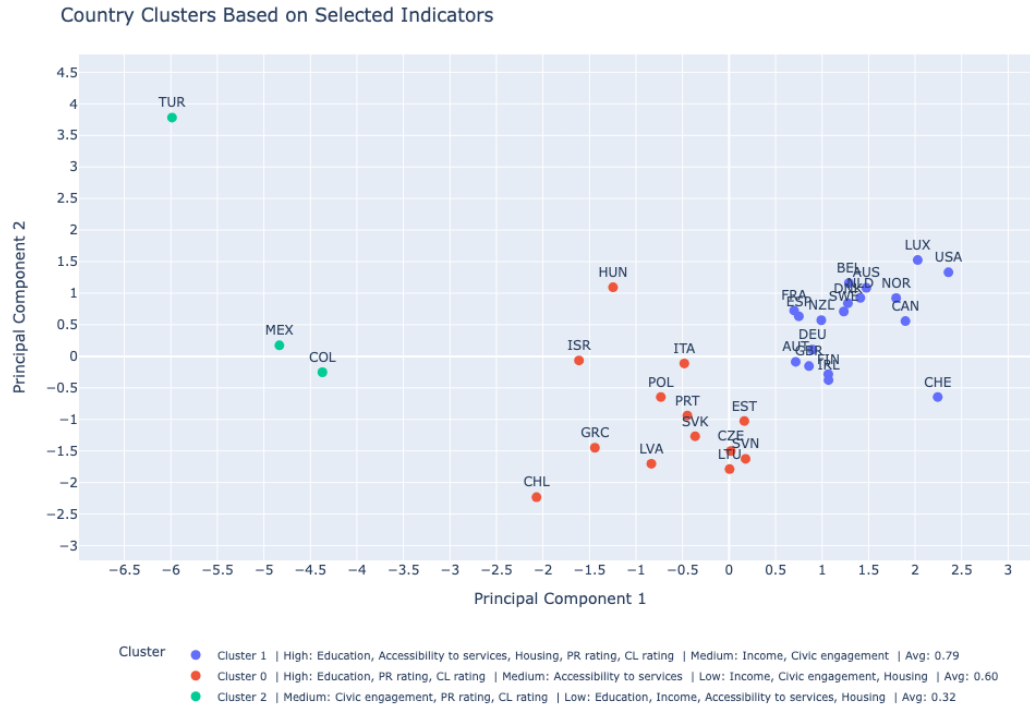


Figure 5: Country clusters based on selected well-being and governance indicators. Each color denotes a distinct cluster, derived from k-means optimization.

In the current configuration, three distinct clusters emerge:

- **Cluster 1 (blue)** — High in *Education*, *Access to Services*, *Housing*, *Political Rights*, and *Civil Liberties*; Medium in *Income* and *Civic Engagement*. Includes most Western European countries, North America, and high-performing smaller states such as Luxembourg and Switzerland.
- **Cluster 0 (red)** — High in *Education*, *Political Rights*, and *Civil Liberties*; Medium in *Access to Services*; Low in *Income*, *Civic Engagement*, and *Housing*. Includes several Eastern and Southern European countries and some Mediterranean nations.
- **Cluster 2 (green)** — Medium in *Civic Engagement*, *Political Rights*, and *Civil Liberties*; Low in *Education*, *Income*, *Access to Services*, and *Housing*. Includes countries such as Turkey, Mexico, and Colombia.

5.4 Conclusion

By condensing multi-dimensional indicators into interpretable clusters, this module bridges the gap between raw statistics and actionable insight. It reveals structural “neighborhoods” of countries—helping individuals discover lesser-known but promising destinations and enabling policymakers to situate their nations within a network of comparable peers. Within the broader system, clustering complements the country comparison and recommendation tools, providing a mid-level lens that links granular metrics to global migration patterns.

6 Dominance Visualization (Hasse Diagram)

While clustering groups countries by similarity, it does not establish clear *hierarchies* or reveal explicit inequalities between them. To address this gap, we implemented a module based on dominance relationships, visualized through a *Hasse diagram*. Rooted in partial order theory, this approach defines when one country can be considered strictly more attractive than another across multiple well-being indicators, enabling a principled evaluation of potential migration destinations.

6.1 Purpose and Applications

In this framework, a country A is said to *dominate* a country B if:

$$A_i \geq B_i \quad \text{for all selected indicators } i, \quad \text{and} \quad A_j > B_j \quad \text{for at least one indicator } j.$$

This ensures that A is at least as good as B across all dimensions considered and strictly better in at least one. Dominance differs from similarity in two important ways: it is *asymmetric* (a country can dominate another without being similar to it) and *directional* (it clearly implies a “better–worse” relationship).

Such a perspective is directly relevant to the issue of *brain drain*: dominant countries, being systematically stronger in terms of well-being and governance, tend to attract highly skilled individuals, while dominated countries risk losing them.

The main applications of the dominance module are:

- **Migrants** — identify destinations that guarantee structural improvements across all selected dimensions.
- **Policymakers** — benchmark against systematically stronger countries and adopt targeted reforms, turning potential human capital losses into opportunities for *brain gain*.
- **Researchers** — analyze asymmetrical relationships in multi-indicator comparisons and link them to migration flows, validating assumptions about the mechanisms of *brain drain*.

6.2 Methodological Pipeline

The dominance module follows a sequence of steps that transform user-selected criteria into an interpretable hierarchy of countries. The process begins with the selection of up to three well-being or governance indicators—such as *Jobs*, *Income*, or *Health*—which define the dominance comparison. An optional fourth indicator (e.g., *Life Satisfaction*) can be chosen to determine node color in the visualization.

Once the indicators are set, the module proceeds as follows:

- **Data Retrieval and Normalization** — Extract normalized values of the selected indicators from the final dataset, scaled to the $[0, 1]$ range for comparability.
- **Dominance Graph Construction** — Create a directed edge from A to B if A performs at least as well as B in all selected indicators and strictly better in at least one. Apply a transitive reduction to remove redundant edges while preserving the core hierarchy.
- **Graph Layout and Visualization** — Display the resulting Directed Acyclic Graph (DAG) in a layered layout, with node size proportional to the number of countries dominated and node color representing the qualitative variable.
- **Summary Metrics** — For each country, calculate the number of countries it dominates and the number of countries by which it is dominated.

6.3 Visualization and Results

Figure 6 shows a Hasse diagram constructed using *Jobs*, *Income*, and *Health* as dominance variables, with node color representing *Life Satisfaction*.

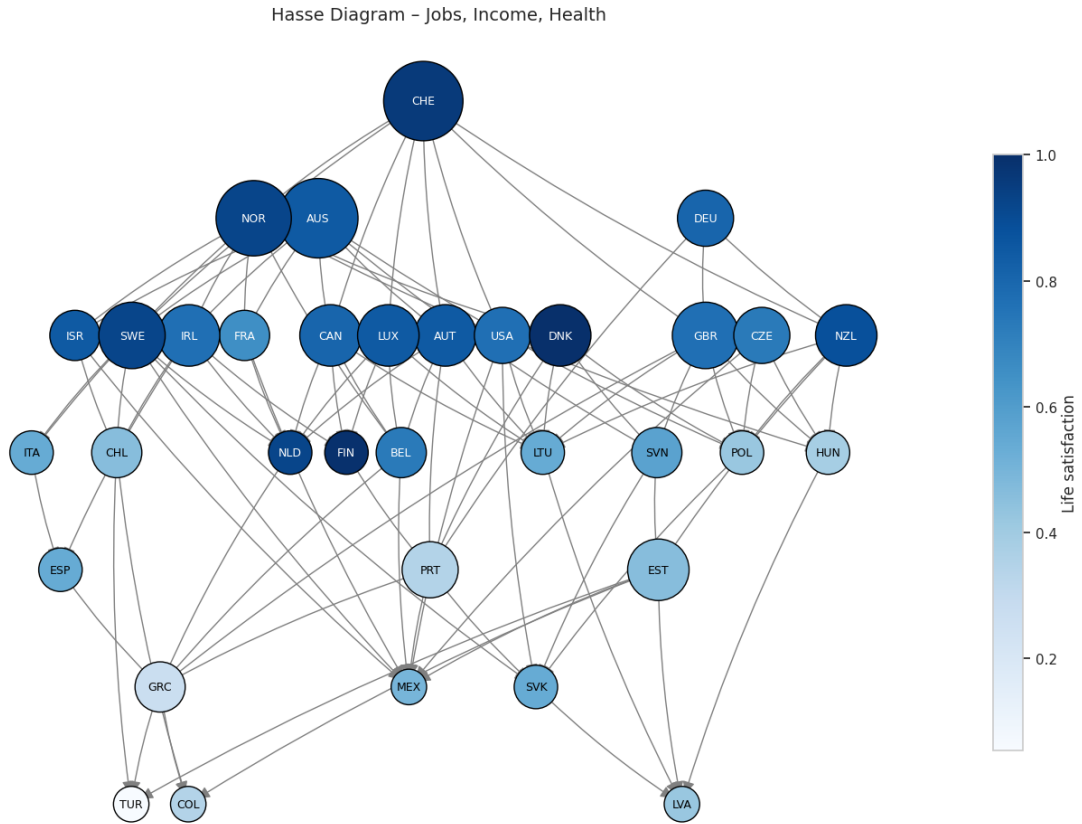


Figure 6: Hasse diagram with dominance variables *Jobs*, *Income*, and *Health*, node color scaled by *Life Satisfaction*. Larger nodes dominate more countries.

The diagram reveals a clear hierarchy:

- **Top-tier countries** — Switzerland (CHE), Norway (NOR), and Australia (AUS) dominate a wide range of other nations, combining strong labor markets, high incomes, and excellent health systems with high life satisfaction.
- **Upper-middle layer** — France (FRA), Canada (CAN), Luxembourg (LUX), and the USA (USA) dominate numerous peers but are themselves dominated by one or two top performers.
- **Intermediate performers** — Italy (ITA), Chile (CHL), and Lithuania (LTU) dominate some lower-tier nations but are frequently outperformed by those above them in the hierarchy.
- **Lower tier** — Turkey (TUR), Colombia (COL), and Latvia (LVA) appear at the bottom, dominated by many others and with comparatively weaker scores in the selected dimensions.

6.4 Conclusion

The Hasse diagram complements similarity-based analyses by introducing explicit hierarchies grounded in multidimensional performance. It answers the question “which countries are *truly* better according to these indicators?” rather than merely grouping nations by likeness. Within the broader analytical system, this module offers a powerful visual and conceptual tool for prioritizing migration destinations, benchmarking national performance, and interpreting competitive advantages in a structured, evidence-based manner.

7 Recommendation System

The final and most important component of our analytical framework is a personalized *recommendation engine*, designed to simulate how an individual might evaluate migration destinations by integrating both subjective preferences and objective country-level indicators. This module bridges the gap between personal aspirations and structural realities, generating tailored migration advice grounded in measurable data.

7.1 Purpose and Applications

The recommendation engine connects individual aspirations with measurable country-level conditions, combining push factors (dissatisfaction with the origin) and pull factors (desired features in the destination). This hybrid approach transforms subjective preferences into data-driven recommendations.

Within the brain drain vs brain gain debate, the module has a dual role: it helps individuals identify destinations aligned with their priorities, and it provides policymakers with insight into the factors that make certain countries more attractive than others.

Applications include:

- **Migrants** — receive personalized recommendations that balance improvement over the origin with preference alignment.
- **Policymakers** — understand drivers of national attractiveness and design reforms to retain or attract talent.
- **Researchers** — study how push–pull dynamics shape migration choices and their wider economic implications.

7.2 Methodological Pipeline

The recommendation algorithm evaluates each potential destination through a structured six-step process, designed to combine *relative improvement* over the origin country, *absolute proximity* to the user’s ideal conditions, and empirically observed migration flows.

Step 1 — Push Score. For each push factor (representing an aspect the user wishes to improve compared to their country of origin), the system measures how much better the destination performs. The size of this improvement is multiplied by the personal weight the user assigned to that factor, ensuring that more important aspects have a stronger influence on the final score.

Step 2 — Pull Score. The system evaluates the destination in absolute terms, regardless of the origin country, by checking the performance of the selected pull factors (qualities the user is looking for in a destination) and multiplying each by its importance weight. This captures how well the destination matches the desired living conditions.

Step 3 — Raw Composite Score. Push and Pull Scores are summed to produce a preliminary overall value for the destination, reflecting both improvement over the origin and intrinsic appeal.

Step 4 — Cosine Similarity. To assess how close the destination’s full profile is to the user’s “ideal profile” across all selected indicators, the algorithm calculates a cosine similarity. This geometric measure evaluates the angle between two vectors — the user’s ideal and the destination’s actual indicators. Values closer to 1 indicate stronger alignment.

Step 5 — Migration Flows. To capture revealed preferences and network effects, the algorithm incorporates the number of migrants moving from the origin to each destination (filtered by sex, log-transformed and normalized), serving as a proxy for destination popularity and existing diaspora communities.

Step 6 — Final Score. The Push–Pull Composite Score (`score_norm`), the cosine similarity measure, and the normalized migration flows (`number_norm`) are combined, for each destination country, into a single weighted index:

$$\text{Final Score} = 0.4 \cdot \text{score_norm} + 0.4 \cdot \text{similarity} + 0.2 \cdot \text{number_norm}.$$

This ensures that the recommendations reflect both subjective preferences and structural realities, while also aligning with empirically observed migration dynamics.

Since all three components are normalized within the $[0,1]$ interval, the Final Score also ranges between 0 and 1. Higher values indicate stronger alignment with both user preferences and observed migration dynamics, making destinations directly comparable on a unified scale.

This hybrid scoring strategy combines three complementary perspectives:

- **Directionality** — ensuring that the suggested destinations offer genuine improvements compared to the origin.
- **Profile Alignment** — ensuring that the country’s overall characteristics match the user’s priorities.
- **Empirical Validation** — grounding the recommendations in real migration flows to account for popularity and network effects.

By blending these perspectives, the system generates recommendations that are both analytically robust and personally relevant.

7.3 User Profile and Results

For the scenario analyzed, the user is a male from Italy (ITA) dissatisfied with several domestic conditions. He assigns importance scores (1–10 scale) to aspects he wishes to improve (push factors): Jobs (7), Income (8), Safety (5), Health (4), Accessibility to Services (7), Life Satisfaction (4). He also specifies desired conditions in the destination country (pull factors): Jobs (9), Income (10), Safety (8), Health (7), Life Satisfaction (9), Political Rights (PR) (10), and Civil Liberties (CL) (10). The emphasis on PR and CL underscores the importance of political freedom and governance quality in his decision-making.

Applying the methodology to this profile yields the following top 5 destinations:

1. **Switzerland (CHE)** — 0.9980
2. **Luxembourg (LUX)** — 0.9104
3. **Norway (NOR)** — 0.9007
4. **Australia (AUS)** — 0.8908
5. **Denmark (DNK)** — 0.8747

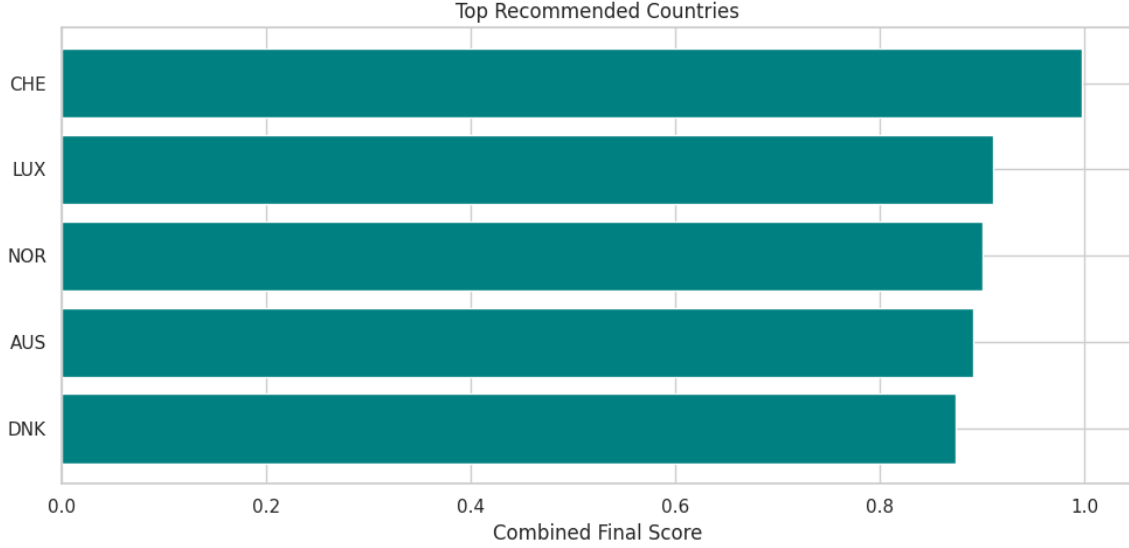


Figure 7: Top 5 recommended countries for the specified user profile.

To enhance transparency, the system provides a breakdown of main improvements contributing to each country’s score:

- **CHE** — Major gains in *Income* (+0.60), *Jobs* (+0.52), *Life satisfaction* (+0.42), with notable improvements in *Accessibility to services* (+0.38) and *Health* (+0.06).
- **LUX** — Strong increases in *Jobs* (+0.36), *Income* (+0.32), and *Life satisfaction* (+0.31), plus better *Accessibility to services* (+0.28) and slight improvement in *Safety* (+0.01).
- **NOR** — Large boosts in *Jobs* (+0.50), *Life satisfaction* (+0.38), and *Accessibility to services* (+0.28), alongside gains in *Income* (+0.09) and *Health* (+0.03).
- **AUS** — Marked improvements in *Jobs* (+0.45), *Life satisfaction* (+0.31), and moderate gains in *Income* (+0.21) and *Health* (+0.09).
- **DNK** — Balanced profile with increases in *Jobs* (+0.46), *Life satisfaction* (+0.46), and *Accessibility to services* (+0.30).

The reported values (e.g., *Income* +0.60) indicate the normalized improvement of each indicator in the destination compared to the origin country, weighted by the importance assigned by the user. Such explanations reinforce user trust by making the recommendation process transparent and grounded in quantifiable evidence.

7.4 Conclusion

The recommendation engine personalizes migration analysis, transforming raw multi-dimensional data into actionable, user-specific insights. By integrating push and pull dynamics with profile–destination similarity, it supports migration decisions that are both data-driven and aligned with personal priorities. Within the broader system, it functions as the final decision-support layer, complementing the exploratory, comparative, and hierarchical analyses of previous modules.

8 Interactive Web Application

To make the analytical framework accessible to a wider audience, we developed a fully functional **interactive web application** using the **Streamlit** framework. This application integrates all analytical modules described in this report into a dynamic, user-friendly environment where visitors can explore data, compare countries, and receive personalized migration recommendations without requiring programming skills or complex installations.

8.1 User Interface Overview

The application is structured into distinct sections, each corresponding to a specific analytical task. The homepage presents a navigation sidebar, a concise description of the project's objectives, and immediate access to all interactive modules.

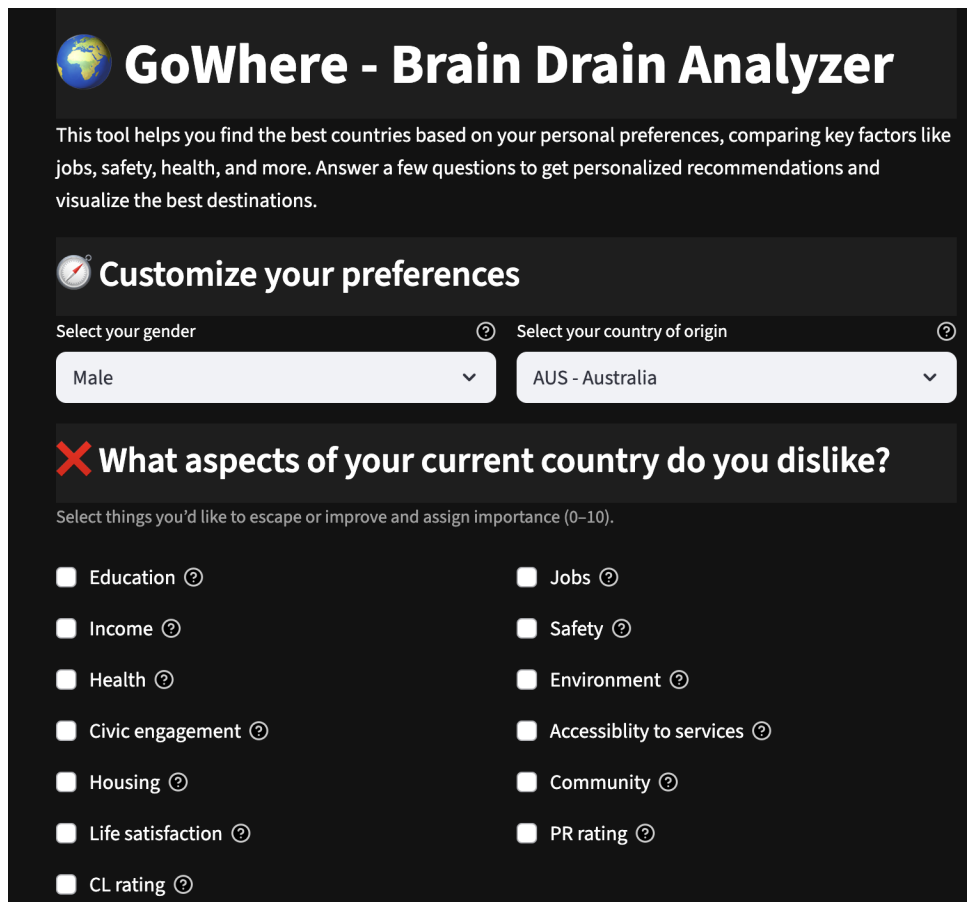


Figure 8: Homepage of the web application with sidebar navigation and project overview.

8.2 Core Functionalities

The platform offers several interactive tools that address different analytical needs:

- **Personalized Recommendation Engine** — Definition of push and pull factors, followed by generation of a ranked list of countries tailored to the user's stated preferences.
- **Country Comparison** — Side-by-side evaluation of two countries across selected well-being and governance indicators, using normalized values for direct comparability.
- **Clustering Interface** — Grouping of countries based on similarity in quality-of-life attributes, enabling discovery of non-obvious but structurally comparable destinations.

8.3 Technological Stack

The application was developed in **Python** and deployed via **Streamlit Cloud**, with the source code hosted on **GitHub**. Streamlit retrieves the Python scripts from the linked repository, executes them in a controlled environment, and exposes the interactive interface as a web application. The architecture is modular and relies on the following main libraries:

- **Streamlit** — Web interface and interactive components.
- **Plotly** — Dynamic charts and scatterplots for clustering results.
- **NetworkX** + **Graphviz** — Construction and rendering of Hasse diagrams.
- **scikit-learn** — Clustering algorithms, dimensionality reduction (PCA), and similarity computations.
- **pandas** + **numpy** — Data manipulation and backend logic.

8.4 Deployment and Accessibility

The application is publicly available online at:

<https://braindrain-4zkeytvdhytywh6uhknkap.streamlit.app/>

All modules can be accessed directly through a browser with no need for local setup. Users can modify migration scenarios, adjust preferences, and view results in real time, making the platform suitable for both casual exploration and structured policy analysis.

8.5 Advantages of the Interactive Approach

Deploying the system as a web application greatly increases accessibility, removing technical barriers and allowing users to engage directly with the analytical tools. Real-time interaction encourages experimentation, while the versatility of the platform makes it equally useful for public presentations, workshops, academic research, and individual decision-making. Users can iteratively refine their migration profile and immediately observe how recommendations and rankings evolve.

8.6 Future Developments

While the current version is fully functional, several enhancements are planned to further improve usability and analytical scope:

- Export functionality for personalized reports and country comparisons in PDF format.
- UX refinements such as contextual tooltips, pre-filled example profiles, and interactive maps for geographic exploration.

These improvements will enhance the platform’s value as a decision-support tool for both individual users and institutional stakeholders.

9 Conclusion

This work presents a comprehensive and interactive framework for analyzing the global phenomenon of *brain drain*—the migration of highly educated individuals in search of better opportunities abroad. By combining rigorous data analysis with a personalized decision-support system, it addresses both the *business problem* of talent loss in origin countries and the *business need* for tools that can guide individual choices and inform policy. The result is a dual-purpose platform: a structural perspective on global migration patterns and an individual-level tool for evaluating and simulating relocation choices.

Our approach integrates multiple real-world data sources: detailed OECD migration flows, standardized well-being indicators from the OECD Better Life framework, and political freedom ratings from Freedom House. Through this integration, we have built a modular and extensible system capable of descriptive analysis, cross-country comparison, clustering, dominance assessment, and personalized recommendations.

9.1 Summary of Contributions

Over the course of this project, we designed and implemented an integrated analytical framework that brings together data collection, processing, visualization, and decision support. The main contributions can be summarized as follows:

- **Unified dataset** — A harmonized integration of OECD migration data, well-being indicators, and governance metrics, normalized to ensure comparability across countries and dimensions.
- **Exploratory migration dashboard** — An interactive environment to analyze volumes, demographic breakdowns, and net migration balances, supporting both macro-level and granular insights.
- **Interactive country comparison** — A flexible interface for evaluating trade-offs between nations across well-being and governance indicators, enabling targeted comparisons.
- **Clustering and dominance analysis** — Two complementary approaches: clustering uncovers latent patterns of structural similarity, while dominance establishes explicit hierarchies across multiple dimensions.
- **Personalized recommendation engine** — A hybrid scoring model that combines push–pull factors with profile–destination similarity, generating transparent and data-driven migration suggestions.

Together, these components form a coherent, modular platform that can both describe existing migration patterns and support individual decision-making, bridging the gap between statistical evidence and personal relocation choices.

9.2 Reflections, Limitations, and Future Development

The project highlights the value of data fusion and interactive analytics for addressing complex global challenges. However, several limitations remain: the current focus on OECD countries excludes many emerging economies relevant to migration; static indicators do not capture dynamic shocks or rapid changes; and user profiles could be enriched (e.g., with age, occupation, family status, migration costs) to simulate more realistic scenarios.

Future developments could address these gaps by extending coverage to non-OECD countries, integrating geospatial visualizations and exportable reports, and applying machine learning to improve prediction accuracy, cluster detection, and preference inference.

9.3 Final Remarks

Migration is not only a matter of statistics or policy frameworks—it concerns individuals seeking to improve their quality of life. Importantly, brain drain is not always a net loss: under the right conditions, countries can experience *brain gain* through upskilling, diaspora networks, return migration, and remittances. Our platform thus supports both perspectives: it empowers individuals in making informed, data-backed decisions, while providing policymakers with insights to anticipate or manage skilled migration flows.

In a world where human capital is increasingly mobile and global inequality remains a pressing issue, tools like ours can foster a more transparent, evidence-based dialogue on international migration—helping societies navigate the thin line between brain drain and brain gain.

References

- [1] OECD. *Database on Immigrants in OECD and Non-OECD Countries*. Available at: <https://www.oecd.org/en/data/datasets/database-on-immigrants-in-oecd-and-non-oecd-countries.html>
- [2] Freedom House. *Freedom in the World*. Available at: <https://freedomhouse.org/report/freedom-world>
- [3] OECD. *OECD Regional Well-Being Tool*. Available at: <https://www.oecd.org/en/data/tools/oecd-regional-well-being.html>
- [4] Yale Economic Growth Center. *Brain Drain or Brain Gain?*. Available at: <https://egc.yale.edu/brain-drain-or-brain-gain>