NOVA IMS Information Management School

MDDM

Master's Degree Program in

Data-Driven Marketing

BEYOND INTENTIONS: A DATA-DRIVEN MODEL OF BRAZILIAN MIGRATION DESTINATIONS FOR DECISION-MAKING

Predicting individual migration trends to support strategic decisions

Priscila Dias Fenner

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data-Driven Marketing

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

BEYOND INTENTIONS: A DATA-DRIVEN MODEL OF BRAZILIAN MIGRATION DESTINATIONS FOR DECISION-MAKING

Predicting individual migration trends to support strategic decisions

by

Priscila Dias Fenner

Master Thesis presented as partial requirement for obtaining the Master's degree in Data-Driven Marketing, with a specialization in Marketing Intelligence.

Supervised by

Teresa Maria Ferreira Rodrigues, PhD, NOVA Information Management School

Augusto José Rabelo Almeida Santos, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 15 July 2025

Priscila Dias Fenner

DEDICATION

To my husband, Matheus,

Who has been my greatest source of encouragement over the past nine years. From the very beginning, you shared the immigrant journey with me, embracing the unknown, the fears, the highs and the lows, always by my side.

You not only motivated me to pursue my dreams and believed in me during moments when even I had doubts, but you also changed your own plans and dreams so that I could pursue mine. Moving to a new country is never easy, yet you embraced this journey with me, including relocating to Portugal, and supported me with love, patience, and strength every step of the way.

People often say that behind every great man, there is a great woman. I am proud to say that behind this woman, there is also a great man, who dreams with me, plans with me, lives with me, and believes with me.

This achievement is not mine alone. It is ours.

I can’t wait to keep dreaming and building our future together.

ACKNOWLEDGEMENTS

Firstly, I would like to thank God for giving me the strength not to give up on my dream. Studying abroad in Portugal has been a goal of mine since I was 18 years old. I applied for several scholarships during my undergraduate studies, but was not successful. When I discovered the Master's in Data-Driven Marketing and saw that everything I believed in for my professional future was reflected there, I decided to pursue it. However, shortly after enrolling and receiving my acceptance, the pandemic hit, and once again came the frustration of being so close. Then, God gave me a third opportunity, and I embraced it with all my strength.

This journey was not linear. It included challenges, detours, and the experience of living in different countries, each one shaping who I am today and enriching my perspective both personally and professionally.

I am deeply grateful to my family, specially my mother, father, sisters, and my husband, who have always believed in me and have been the hidden force behind every achievement in my life.

I would also like to express my heartfelt gratitude to my grandmother, who sadly passed away while I was studying far from home. I know how proud she was of me and this achievement is also a tribute to her memory.

I would also like to sincerely thank my thesis supervisors, Teresa and Augusto, for believing in the relevance of my research topic and encouraging me to move forward, even with the complexities involved in the data collection process. Their guidance and support were fundamental throughout this journey.

Finally, I would like to thank the Brazilian immigrants who anonymously participated in the data collection. The pride I feel in seeing so many Brazilian communities around the world, and witnessing how much people help one another, just as they helped me reach this research, only deepens my love for being Brazilian. These communities make it possible to feel a little bit of home anywhere in the world.

ABSTRACT

This thesis explores the motivations and challenges faced by Brazilian emigrants, aiming to bridge the gap between descriptive migration studies and predictive modelling. While existing research has predominantly focused on macro-level flows or qualitative analyses, this study adopts a novel approach by building and analysing an original dataset of 173 survey responses from Brazilian migrants to Australia and Ireland. To uncover the patterns influencing destination choices and migration experiences — and to move beyond explanation toward prediction — this study applies a supervised machine learning model (Random Forest), preceded by Principal Component Analysis (PCA) for dimensionality reduction and feature selection. Despite limitations such as a modest sample size, class imbalance, and the presence of noisy self-reported data, the model achieves an overall accuracy of 73% in predicting destination country, correctly identifying 81% of respondents who migrated to Australia and 60% to Ireland. The most influential features are predominantly qualitative, including perceived cost of living, use of social media and messaging apps to access information, and motivations such as language acquisition and job opportunities. These results suggest that behavioural traits and perceptions outweigh sociodemographic characteristics in determining migration outcomes. The findings offer technical insights for institutions such as universities, immigration agencies, and private service providers targeting migrants. The model enables the identification of undecided or influenceable migrant profiles, supports the development of communication strategies tailored to digital information-seeking behaviours, and highlights optimal moments for outreach. Additionally, by introducing a reproducible pipeline for predictive modelling based on primary behavioural data, the study addresses a key methodological gap in current migration research. By shifting the analytical lens from aggregate flows to individual-level decisions, it demonstrates that predictive analytics can generate meaningful insights even within small and complex datasets. This approach contributes a new perspective to migration research and offers institutions a scalable tool to plan strategically and respond proactively to emerging migration trends based on data, rather than assumptions.

KEYWORDS

Brazilian Migration; Predictive Modelling; PCA; Random Forest; Migration Prediction

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
GDP	Gross Domestic Product
KNN	K-Nearest Neighbours
MI	Mutual Information
ML	Machine Learning
PCA	Principal Component Analysis
PLS-SEM	Partial Least Squares Structural Equation Modelling
RF	Random Forest
SDG	Sustainable Development Goal
SEM	Structural Equation Modelling
SN	Subjective Norms
TPB	Theory of Planned Behaviour
UN DESA	United Nations Department of Economic and Social Affairs

1. INTRODUCTION

Migration is a global phenomenon that continues to shape economies, cultures, and societies around the world. The movement of people across borders occurs for various reasons, including economic opportunities, political instability, educational aspirations, and environmental factors. By mid-2024, the number of international migrants worldwide was estimated at 304 million, representing 3.7% of the global population (International Organization for Migration, 2024). This large-scale mobility has profound impacts on both sending and receiving countries, influencing labour markets, social integration policies, and cultural exchanges.

Within this broader context, Brazilian emigration has become an increasingly prominent global trend in recent years. In 2023, an estimated 4.9 million Brazilians were living abroad, a significant increase from 2.7 million in 2015. This represents a growth of 81.4% over the period (Brazilian Ministry of Foreign Affairs, 2024).

This is a growing diaspora that has made a tangible impact on Brazil's economy, society, and culture, as well as, on the countries where they have settled. However, despite the relevance and scale of this movement, there remains a substantial gap in the literature when it comes to understanding its underlying nature, especially the motivations and challenges faced by Brazilian emigrants from a market-oriented perspective.

1.1 RESEARCH MOTIVATION AND PRACTICAL RELEVANCE

Understanding the motivations and challenges faced by Brazilian emigrants provides valuable insights for developing strategies that support their integration and enhance their experiences in host countries.

Despite the scale of this growing diaspora, existing studies on Brazilian migration remain limited and are mostly focused on understanding the phenomenon from descriptive and qualitative perspectives, without offering concrete tools to predict future movements and improve institutional readiness. Current literature generally addresses migration through socio-economic and cultural lenses, without considering a predictive model aimed at strategic decision-making.

This gap is particularly relevant given the growing demand from universities, immigration services, and private organisations for data that can guide more effective outreach, service design, and long-term planning.

1.2 OBJECTIVES AND EXPECTED CONTRIBUTIONS

Unlike previous studies, this research aims not only to understand why Brazilians migrate but also to develop a predictive model capable of anticipating migration trends and providing strategic recommendations for institutions seeking to attract and support these migrants more effectively. By analysing behavioural patterns, sociodemographic factors, and informational needs, the study provides a solid foundation for data-driven strategic decisions. This approach enables institutions to adapt their services and improve their communication strategies, contributing not only to enhancing the experiences of Brazilian migrants but also to shaping institutional strategies and expanding the portfolio of services offered by interested organizations.

The main contribution of this thesis is twofold:

1. The design of a feature-rich dataset derived from a carefully structured survey targeting Brazilian emigrants;
2. The development of a supervised learning framework to predict migration destinations based on key individual attributes (such as age, educational background, and economic status).

In the absence of publicly available datasets or predictive migration at the individual level benchmarks, this study has contributed meaningfully to the literature by building a complete and original pipeline, from survey design to feature engineering and model training. The selection of predictive features was a non-trivial process, supported by Principal Component Analysis (PCA) to isolate the primary drivers influencing destination choice. The final model, based on the Random Forest algorithm, was trained to classify the most likely migration destination. This predictive paradigm, which uses migrant characteristics to anticipate their destination country, remains largely underexplored in migration studies.

By identifying the most informative individual attributes shaping migration decisions, the study delivers a competitive predictive model that can assist institutions in anticipating trends and tailoring their services accordingly. This data-driven approach not only enriches the qualitative understanding of Brazilian migration but also provides a strategic foundation for decision-making across sectors.

2. LITERATURE REVIEW: THE STATE OF THE ART IN MIGRATION STUDIES

2.1 MIGRATION: KEY CONCEPTS AND FRAMEWORKS

Migration is a contested and evolving global process that has influenced human history, economies, and societies from earliest times. While most of the population remains in the same country of birth, the number of people who change residential location at every scale (distinction is made between intra-national, regional and international migration) has increased over the generations. Estimates show that, in mid-2024, there were 304 million international migrants globally, representing 3.7% of the world's population (International Organization for Migration, 2024).

Despite its growing relevance, migration remains one of the most unpredictable aspects of human behaviour. The phenomenon is marked by volatile movements, a multiplicity of causes and motivations, conceptual inconsistencies, data limitations, and major forecasting challenges (Caselli et al., 2004). These difficulties are further exacerbated by inconsistent legal frameworks and divergent criteria for defining and categorising migrants across countries, which hinder both comparative analysis and policy development. Even the basic definition of migration varies between contexts, complicating efforts to ensure reliable data collection and effective governance.

We live in a hyperconnected world where expectations are rising rapidly, and global timeframes often collide, increasing the complexity of migratory dynamics. Expectations are hastily increasing, and the gap between what we want and what, what is happening, and what is offered in services is wide (Rodrigues, 2022). The opportunity to access “too much” information, and the connectivity of more powers of mobility are driving complexity and depth of surging migratory movements. Most countries today are simultaneously both senders and receivers of migrants, regardless of their level of economic and human development, and the underlying causes of migration have become more diverse and multifaceted, although economic motivations continue to prevail, as shown in Figure 1.

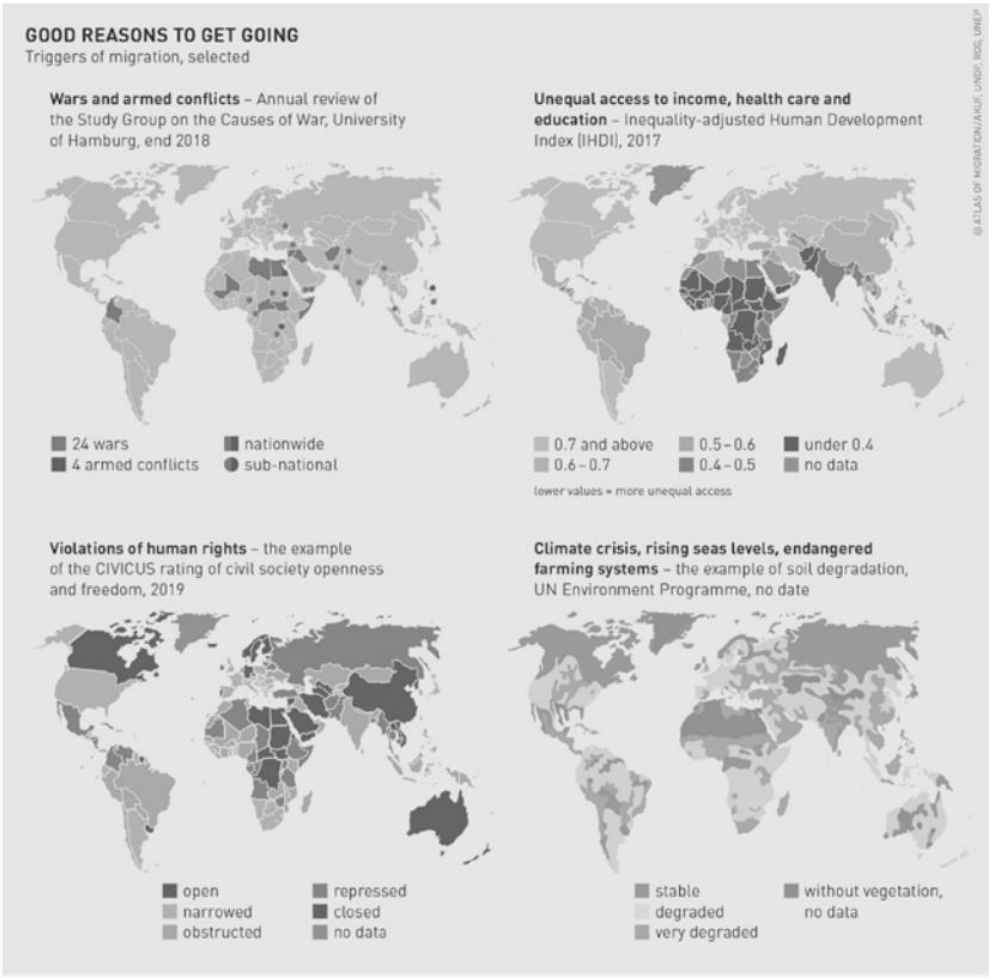

Figure 1 – Regional migration dynamics: Key triggers of migration such as armed conflict, inequality, human rights violations, and environmental degradation (Rosa-Luxemburg-Stiftung, 2019).

To understand how migration is approached in the literature, it is important to distinguish between the concepts, such as *migration* and *migrant.* Migration refers to the process of moving from one location to another. This movement may occur internally, within the borders of a country, or internationally, across national boundaries. On the other hand, the term *migrant* is one that is situational, less absolute, and willing to be many things. The term may emerge as an issue of definition and identity. While it may just refer to someone who has physically moved, it may also refer, in broader sociopolitical context, individuals who may be identified as "migrants" based on their ancestry, nationality, or individual migration status, irrespective of being born and raised in that location. For example, second- or third-generation individuals born in a country may still be categorised as migrants based on their ancestry, and in some cases, those individuals lacking a legal identity and citizenship may be characterised as irregular migrants (McAuliffe, B and Triandafyllidou, A., 2021).

Another key concept in the migration literature refers to the length of stay. International migrants are typically classified into two categories: short-term migrants, defined as individuals who move to a country other than their usual residence for a period of at least three months but less than one year; and long-term migrants, referring to those residing in a different country for one year or longer. While these classifications are essential for demographic analysis and policy-making, their practical application varies. Not all countries adopt these definitions consistently, which complicates international comparability and hinders global statistical harmonisation (UN DESA, 1998; McAuliffe & Triandafyllidou, 2021).

The impact of migration flows extends across economic development, labour markets, and educational systems, particularly in how talent and skills are redistributed globally. Although international migrants represent just 3.6% of the world's population, the vast majority — around 3% — migrate voluntarily, primarily for economic, educational, or family-related reasons. Only about 0.6% migrate involuntarily, often as a result of conflict, persecution, or the absence of fundamental rights (Rodrigues, 2022).

People migrate for a range of reasons, often driven by a combination of economic opportunities, political conditions, educational goals, and environmental factors. Work remains the primary driver of international migration, and migrant workers make up most of the global migrant population. Many of these individuals move to high-income countries (Figure 2), even those located far from their place of origin, reflecting a broader trend toward regions with stronger labour markets and greater economic stability (McAuliffe & Triandafyllidou, 2021).

This pattern has contributed to what is widely known as the "global brain drain", a phenomenon in which highly educated and skilled individuals migrate from lower-income nations to wealthier ones, often resulting in talent shortages in their countries of origin (Docquier & Rapoport, 2012; TheGlobalEconomy.com, 2024). These flows are shaped not only by personal aspirations but also by structural and institutional factors that influence which destinations are perceived as attainable or desirable.

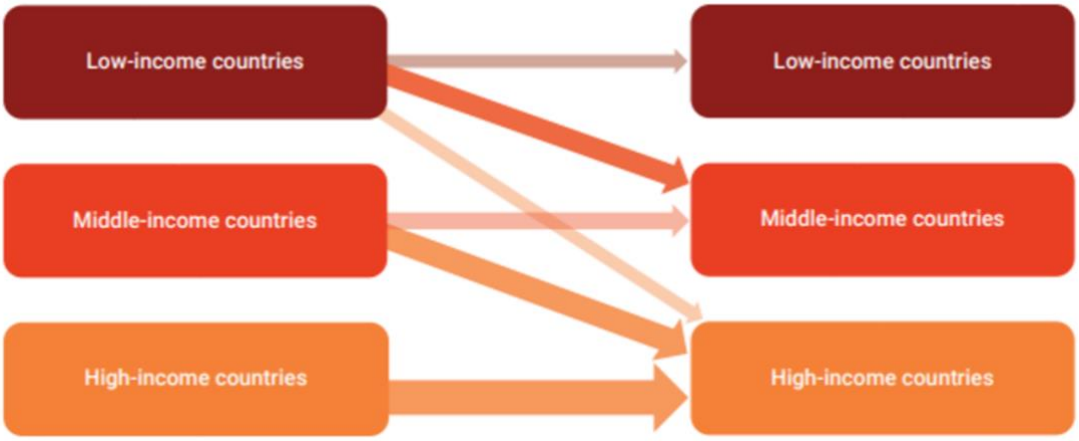

Figure 2 – Migration flows by income level of origin and destination countries. The thickness of each arrow reflects the volume of migration between income groups (World Bank, 2023).

Understanding these motivations requires a comprehensive analysis of the factors that influence such decisions. In the academic literature, migration motivations are often stratified under the push-pull framework. Push factors are the negative conditions that push the migrant away from their home country, such as poverty, violence, and lack of opportunities. In contrast, pull factors are the favourable attributes of destination countries, including job availability, political stability, and quality of life. The pull factors are characteristics of the destination country that attract migrants, including better job prospects, safety, education, and quality of life (Castelli, 2018). While this framework provides a foundational understanding of migration dynamics, it is often insufficient to capture the nuances of decision-making in today's transnational and digitally connected world.

Therefore, migration drivers can also be examined across *macro, meso,* and *micro levels* (Oxford Academic, 2018), which allow us to explore broader systemic, community-based, and individual factors, respectively, that inform migration decisions. Macro-level factors refer to national or global conditions, including immigration policies, economic conditions, and geopolitical events. Meso-level influences involve family ties, diaspora networks, and institutional arrangements. Micro-level factors reflect individual characteristics and perceptions, such as age, education, aspirations, or digital behaviours, that directly inform personal decision-making processes.

This study embraces this multi-layered framework by integrating individual-level data (*micro*), drawing connections to broader institutional policies (*macro*), and accounting for digital and social information flows (*meso*). The following section explores these levels of influence in more detail, drawing on key insights from the literature that inform the development of predictive models capable of forecasting not only *whether* someone will migrate but *where* they are likely to go.

2.2 Understanding Push and Pull Factors

Push factors include economic hardship, political instability, persecution, environmental disasters, and lack of social mobility. Pull factors are positive attributes of destination countries, such as better healthcare, education systems, economic opportunities, and safety (Castelli, 2018; United Nations Network on Migration, 2018). Together, these two types of factors can not only provide a temporal direction to migration movements, but they can shape where migrants choose to go as well.

The macro level factors look at the systemic aspects of migration focusing on factors related to politics, economics, demography, and environment, which are largely outside of people's control (United Nations Network on Migration, 2018). A person's political environment can certainly be a factor, including war, conflict, and persecution, or simply authoritarian governments rendered unsatisfactory to public sentiment, and which naturally force migration by developing unsafe or oppressive conditions. The war in countries experiencing extreme instability, conflict (like Syria) has forced millions of people to be displaced and to seek refuge in other countries (Oxford Academic, 2018).

Economic aspects as a driver for migration at the macro level can be seen as a mix of events on both sides of the equation. Economic crises, high unemployment rates, and poverty can push individuals to migrate to new countries in search of better opportunities. On the other hand, economically stable countries with strong labour markets attract migrants, particularly those seeking to improve their employment conditions and income levels (Castelli, 2018).

Social and demographic aspects are also relevant at the macro level. Rapid population growth, especially in developing regions, has compounded issues related to resource distribution and limited opportunities. These demographic pressures have driven people to migrate to countries with ageing populations and labour shortages (United Nations Network on Migration, 2018).

Environmental changes are another critical macro-level factor. Climate change, natural disasters, and land degradation have increasingly forced people to leave areas that have become uninhabitable. The term "environmental migrants" has emerged to describe those movers who are displaced for several reasons to do with environmental factors (Oxford Academic, 2018). As the world confronts climate change at a steadily increasing pace, the number of people forced to migrate due to environmental conditions is only expected to increase significantly.

The meso-level explores the social, technological, and community-based factors influencing migration. This level highlights the connections between macro-level circumstances and micro-level decisions. It includes the influence of social networks, diaspora communities, communication technologies, and the structure of urban environments (Oxford Academic, 2018).

Diasporas and social networks are essential in shaping migration trends. Migrant communities in place of destination offer migrants the knowledge needed to establish themselves and integrate into society-members of established communities. Members of these established communities frequently have pre-existing ties and support networks, which are particularly valuable in helping new arrivals navigate critical issues such as employment (United Nations Network on Migration, 2018). These diasporas lessen the risks associated with a migration decision and make leaving the home country an easier decision as it is mediated by new, trusted contacts in the country of destination.

Communication technologies, particularly social media, play a significant role in influencing perceptions and decisions about migration. Even though these platforms underscore information without being factual, the cultural pull of seeing a different, and often idealistic quality of life in a developed nation, is important (Castelli, 2018). They create a "pull" effect by portraying these countries as places of abundance and opportunity, encouraging individuals from less prosperous regions to consider migrating. The ability to stay connected with family members abroad also facilitates migration, as individuals feel supported throughout the process (Oxford Academic, 2018).

The urbanization realities and barriers can also impact meso-level factors. Many developing countries continue to urbanize with almost no accompanying economic growth to develop the necessary critical infrastructure to help them cope with needed basic human services, particularly in high density urban environment.

The rapid and unplanned growth of cities in many developing countries has led to increased poverty among households, overcrowding, high unemployment, and a decline in basic living conditions. These conditions have contributed to the emergence of fragile and often unstable urban communities, making migration appear as a more viable alternative. As a result, many urban dwellers perceive fewer advantages in remaining in environments marked by insecurity and limited prospects. In contrast, developed and emerging urban centres often offer more attractive living conditions, reinforcing their appeal as migration destinations (United Nations Network on Migration, 2018).

Finally, the micro-level focuses on the personal and individual characteristics that determine a person’s decision to migrate, and might include education level, personal aspirations, family circumstances, and socio-demographic variables (such as age, gender, and marital status). They shape such outcomes in interaction with macro and meso level contexts, resulting in more or less favourable migration destinations depending on personal backgrounds (Oxford Academic, 2018).

Educated and highly educated people are also more likely to move, particularly when they perceive that they are undervalued in their country of origin. For some, migration is a means of professional progress and improved financial position (United Nations Network on Migration, 2018). Conversely, the very low level of education may also explain worker

migration abroad in response to few local opportunities and for manual work in countries lacking in labour.

Age and gender are also important demographic factors. Younger people, especially single individuals, are more likely to migrate since they have fewer family obligations and are more willing to take risks. Men have generally been more mobile in quest of economic prospects; but the gender disparity in migration is shrinking, and more women are moving on their own in search of work or education. (Castelli, 2018).

Personal attitudes towards migration, shaped by previous experiences, family influence, and social pressures, also affect behaviour at the micro-level. For instance, members of a migrant family history may be more inclined to migrate themselves, seeing it as a viable and positive opportunity (Oxford Academic, 2018). On the other hand, social norms, including gender roles, can either encourage or restrict migration.

The Figure 3 below summarise the main push and pull factors that influence individuals' decisions to migrate. Arrows connecting the factors to the central concept of *migration* highlight the dynamic interaction between motivations for leaving one place and the perceived benefits of moving to another.

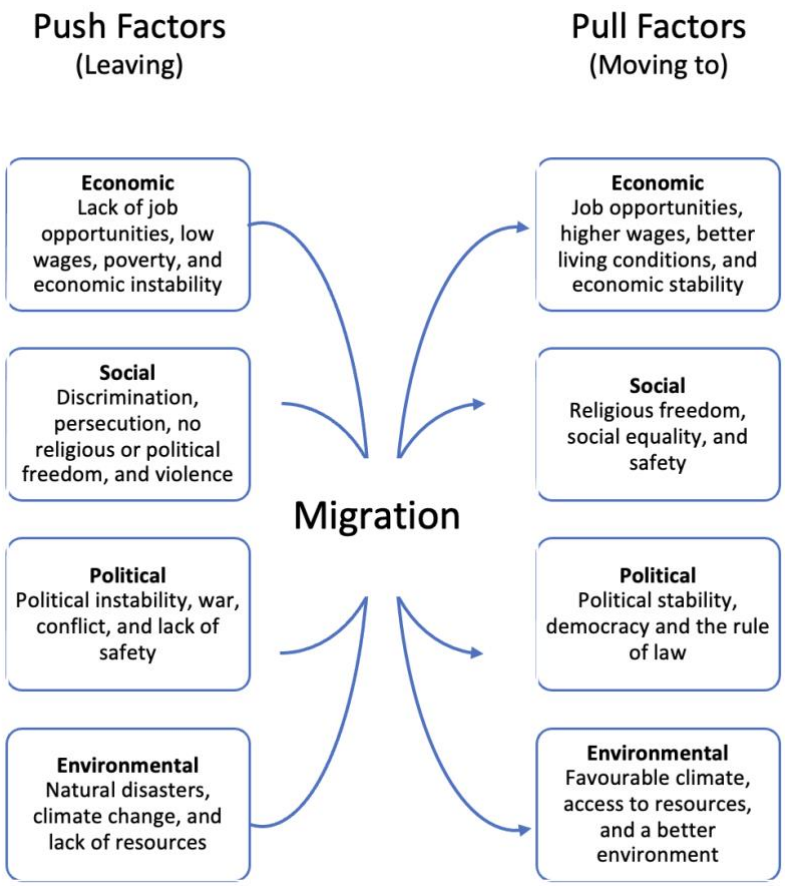

Figure 3 – Main Push and Pull Factors influencing migration (own elaboration).

2.3 MACRO-LEVEL APPROACHES TO MIGRATION: UNDERSTANDING AND PREDICTING MIGRATION FLOWS

In recent years, migration forecasting has increasingly relied on data-oriented approaches, aiming to improve planning, response capabilities, and resource allocation. Macro-level studies examine how structural forces — such as immigration policies, labour markets, and geopolitical instability — influence large-scale migration patterns. Many works focus on estimating migration flows between countries, particularly in contexts of conflict or humanitarian crisis, where accurate forecasting is crucial for economic planning, resource allocation, and institutional response.

This section reviews research that explores both institutional frameworks and data-driven forecasting tools, including machine learning and Bayesian models. While these approaches do not focus on individual motivations (micro-level), they provide essential context for understanding the broader systems in which migration decisions occur.

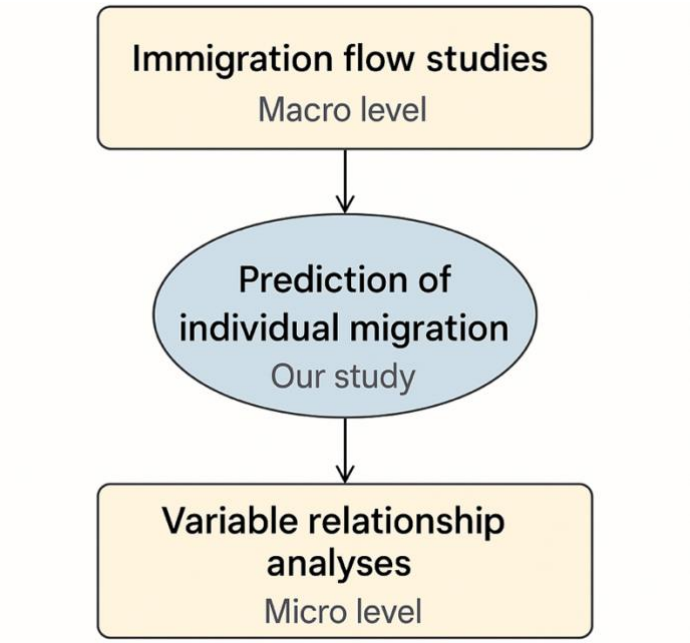

Figure 4 – Positioning of the present study within the literature gap, bridging macro-level immigration flow studies and micro-level variable analyses (own elaboration).

2.3.1 HOW STRUCTURAL FORCES INFLUENCE IMMIGRATION

Migration trends are influenced by structural conditions such as national policies, labour market dynamics, and demographic pressures. As migration forecasting becomes increasingly relevant, predictive models, especially those based on machine learning, are emerging as tools to improve planning and early institutional response. This section is organised in two parts:

the first explores the influence of immigration policies, while the second discusses recent advances in predictive tools used to anticipate migration flows.

Richardson and Lester (2004) consider the immigration policies as one of the significant factors influencing someone to choose to migrate. The authors compare the immigration models and programs in Australia and Canada, specifically historically nations known for attracting skilled migrants. The research discusses the implications of different immigration policies in relation to labour market outcomes, focusing on why Australia tends to achieve better integration of immigrants into the workforce, even though both countries share similar models for immigration tailored to skilled workers.

The study (Richardson & Lester, 2004) adopts a descriptive and interpretative approach, examining how specific policy mechanisms, such as Australia's points-based system and Canada's federal-provincial nominee programmes, translate into measurable labour market outcomes. The analysis places particular emphasis on:

- The role of pre-arrival human capital (education, work experience, language ability);
- The influence of settlement assistance;
- Structural barriers in each country (e.g., credential recognition, employer discrimination).

The main factors identified in the study (Richardson & Lester, 2004) that affect success in the labour market are: a) *Proficiency of English* - Australia's rigorous language testing requirements help ensure that immigrants who arrive in Australia have English language skills to settle quickly into the labour market; b) *Age of migrant*s - Australia attracts younger migrants, particularly those between the ages of 25 and 44, which matches better the processing needs of the labour market and it is also evident that younger immigrants settle more quickly; c) *Pre-assessment* of qualifications - Australia requires immigrants to have their qualifications assessed prior to arriving, thereby facilitating integration into their relevant professions; d) *Access to social benefits* - Australia has a 2-year wait period to obtain social benefits, which motivates faster participation in the labour market, whereas in Canada (and some provinces) immigrants can access welfare benefits immediately on arrival; and e) *Recognition of overseas qualifications* - Canada is a country with decentralization, meaning that there is a major lag time as immigrant's qualifications and recognition are often dealt with gradually. This results in further difficulty in finding suitable immigration roles with professional classes.

The study concludes that both countries have very effective structures for recruiting skilled migrants, however, the more rigorous system of standards for skilled migrants and the structure for labour market integration in Australia lead to a more consistent migratory success. Therefore, elements of Canada's flexible labour market structure implementation could adopt elements of the Australian model including the use of language tests and recognition of qualifications.

2.3.2 THE ROLE OF PREDICTIVE TOOLS IN ANTICIPATING MIGRATION FLOWS

An expanding body of research is turning to predictive modelling to forecast migration flows and improve institutional planning. This section presents recent studies that apply machine learning and statistical tools to anticipate asylum requests, border movements, and population displacements. While these studies typically model aggregate flows, their methodological innovations and forecasting goals are highly relevant to our own work on predicting individual migration decisions.

Forecasting methods aim to generate numerical estimates of future migration flows. In response to the growing demand for more precise projections, these models have become increasingly sophisticated to account for the complex, unstable, and non-linear nature of migration patterns. Academics play a leading role in advancing and applying such forecasting techniques, alongside experts from international organisations (such as Eurostat, the UN Population Division, the World Bank, and the OECD), national statistical agencies, think tanks, and research institutes (Sohst et al., 2020).

While there are several types of forecasting approaches, ranging from argument-based models and intention surveys to time-series extrapolations and machine learning, the core elements remain consistent: each model depends on data inputs (typically from large-scale surveys, census data, or digital traces), statistical techniques, and a set of assumptions regarding how historical patterns can inform future migration outcomes.

Among the more established macro-level approaches are argument-based projections and econometric models. These methods are commonly employed by governmental bodies and international organisations to estimate future migration stock and flow based on historical trends and expert assumptions (Bijak, 2010; Bauer & Zimmermann, 1999). Argument-based projections, for instance, offer low, medium, and high-scenario forecasts, though they lack probabilistic underpinnings. Econometric and gravity models incorporate drivers such as GDP, labour market performance, or geographic distance (Zipf, 1946; Arranz, 2019), assuming that structural conditions largely shape migration behaviour.

Migration intention surveys represent another approach, where individuals' self-reported aspirations to migrate are used to estimate future movements. While useful in contexts with scarce migration data, these surveys rely on the assumption that intentions translate into action, which may not always hold true (Tjaden et al., 2018).

In contrast to these traditional techniques, more recent studies have introduced predictive tools using supervised machine learning, bringing new potential to model non-linear patterns and improve short-term accuracy. While machine learning models have only recently begun to be applied to international migration flows and remain relatively underexplored in the field, early work shows promising results (Robinson and Dilkina, 2018; Bosco et al., 2024). The latter built a forecasting system for European agencies to anticipate asylum applications and irregular border crossings.

While their goal is to anticipate aggregate flows of asylum seekers, our model seeks to forecast individual destination choices among Brazilian migrants, offering predictive insights at the micro level that support universities and mobility services in planning more personalised interventions. In both cases, the common objective is to move beyond reactive strategies by enabling early, data-driven responses to evolving migration dynamics.

To build this system, the authors (Bosco et al., 2024) employed a stacked ensemble combining three supervised machine learning models: Random Forest (RF), Gradient Boosted Decision Trees (GBDT), and Artificial Neural Networks (ANN). Each model was trained independently and then their outputs were aggregated to form a composite prediction, with the goal of leveraging the interpretability of tree-based models and the non-linear learning capacity of neural networks. In terms of performance, each individual model achieved high predictive accuracy:

- The Random Forest performed well in terms of robustness and interpretability, particularly useful when the number of predictors was large and included lagged variables.
- GBDT delivered the best single-model performance, achieving the highest R^2 scores, particularly in forecasting asylum applications, where it explained over 83% of the variance in the validation set.
- The ANN contributed strongly to non-linear interactions and pattern recognition, especially in capturing time-lagged dependencies.

The final stacked ensemble model outperformed all individual models, achieving $R^2 > 0.85$ for irregular border crossings and $R^2 \approx 0.83$ for asylum applications. These R^2 values, commonly used to assess goodness-of-fit in regression, indicated a strong ability of the model to capture the variance in monthly migration patterns, validating its use for operational forecasting.

Like their ensemble, our strategy relies on combining strong learners (e.g., Random Forest) and disciplined feature selection to produce predictions that are not only accurate but also interpretable. Their approach highlighted that rigorous validation (R^2, feature importance, variance explained) and practical utility, rather than model complexity alone, were essential for predictive systems to be used effectively by institutions. This directly informed our thesis, which prioritises transparent, human-readable predictions that can support decision-making by universities, consulates, and migration services.

Similarly, Bayesian hierarchical models have emerged as an alternative forecasting tool, blending time-series data with expert judgement to improve accuracy under uncertainty. These models provide transparency and probabilistic outputs, which are particularly useful in contexts with scarce or unreliable data. Unlike black-box methods, Bayesian models allow integration of prior knowledge, thus enhancing interpretability for decision-makers (Wiśniowski et al., 2013).

These methodological debates also help situate our own study. While many macro-level studies aim to forecast migration flows between countries, few have attempted to predict individual-level destination choices. Our research built on the strengths of predictive tools while shifting the focus from population-level flows to individual decision-making. Similar approaches have employed supervised learning to forecast migration outcomes, though often focused on asylum applications rather than individual preferences (Bosco et al., 2024). This approach contributed to an emerging but still underexplored space in migration research, where micro-level decisions are modelled using predictive analytics.

In doing so, our work fill a notable gap in the literature. While spatial models and time-series extrapolations are useful for broad forecasting, they do not capture the nuanced, feature-level variation that drives personal migration decisions. Behavioural studies using intention surveys or TPB-based approaches, such as Structural Equation Models (SEM), often lack predictive capabilities, serving primarily to explain rather than to forecast. Our approach combines both: we use feature selection and machine learning classification to anticipate migration outcomes at the micro level, offering a more actionable tool for universities, consulates, and mobility services.

Ultimately, we argue that micro-level prediction should not be seen as an alternative to macro-level modelling, but rather as a complementary dimension. By incorporating behavioural, motivational, and informational attributes into predictive frameworks, we contribute a novel, individual-centred perspective to the field of migration forecasting. In this sense, we align with broader forecasting goals while extending their application to the individual scale, a level where practical, data-driven support is still rare but increasingly necessary.

2.3.3 ALGORITHMIC MIGRATION MANAGEMENT AND ETHICAL IMPLICATIONS

A clear critique of algorithmic tools in migration governance has been presented, highlighting not only technical concerns but also the political and ethical consequences of using predictive systems in areas such as visa approval, asylum processes, and migrant risk assessments (Morgenstern & Strijbis, 2024).They raise serious concerns about the lack of transparency in these systems, the risk of built-in bias, and how tools initially developed for helpful purposes can end up serving restrictive or security-driven goals.

To avoid these risks, the authors (Morgenstern & Strijbis, 2024) argue for strong legal protections, open and transparent decision-making processes, and the inclusion of human rights principles in the design of migration-related algorithms. Their main point is that predictive technologies should be not only consistent but also fair and accountable.

Similarly, some reflections were offered on forecasting tools used in the context of forced migration, noting that while expectations for predictive analytics were high, these tools often

failed to deliver in practice due to the complexity of migration systems and limitations in institutional capacity. Nevertheless, governments and agencies continued to invest in such tools because of the urgent need to manage refugee flows and their wide-ranging impacts (Angenendt & Koch, 2024).

Warnings were raised against the use of migration forecasts to justify security-focused or discriminatory actions. Strong ethical safeguards were recommended, including data protection practices that extended beyond individual privacy to consider how entire groups were represented in data. This analysis underlined the importance of ensuring that prediction technologies did not harm vulnerable populations (Angenendt & Koch, 2024).

Although our research focuses on voluntary migration, the ethical concerns raised by both studies remain highly relevant. We handle personal data on people's motivations and migration plans, so we prioritise models that are easy to understand and use feature selection techniques to keep the process transparent. We also take data protection seriously, avoiding shortcuts that could lead to unfair predictions or hidden biases.

Finally, we recognise the limits of what predictive tools can do. Our model is not meant to replace human judgment but to support decision-making by institutions like universities and migration services. By offering useful insights while respecting the privacy and autonomy of individuals, we provide a novel contribution that is both practical and ethically responsible.

2.4 Micro-level approaches to migration: modelling individual motivations and intentions

2.4.1 Factors influencing international students' decisions to stay in Finland: a case study using PLS-SEM

The study on international students in Finland (Nikou & Luukkonen, 2024) investigates the key push and pull factors influencing students' decisions to stay in the host country after graduation. The authors highlight barriers such as language difficulties, challenges in finding employment, and social integration issues, alongside positive influences such as career prospects, recognised educational qualifications, and favourable institutional support.

The research adopts a robust methodological approach by applying Partial Least Squares Structural Equation Modelling (PLS-SEM), well suited for analysing complex linear models with multiple constructs and latent variables. The conceptual model integrates Push-Pull Factor Theory and the Theory of Reasoned Action (TRA), examining how different domains — economic, institutional, environmental, social influence, and personal attitudes — affect the intention to stay in Finland after completing studies.

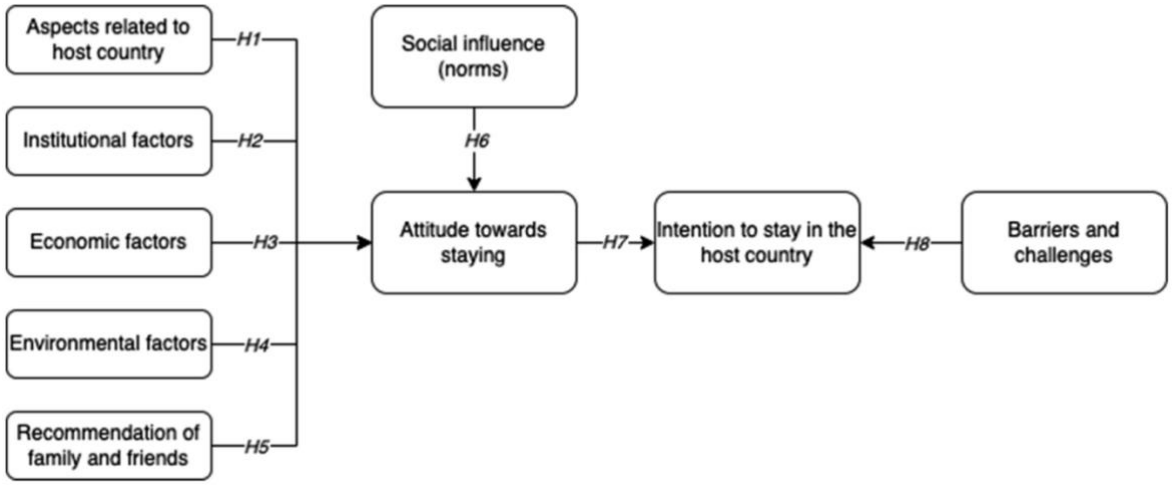

Figure 5 – Conceptual model of migration intentions (Nikou & Luukkonen, 2024)

PLS-SEM was chosen by the researchers for its ability to model both direct and mediated relationships, and several statistical indicators confirmed the model's quality:

- Cronbach's alpha ranged from 0.705 to 0.945, indicating acceptable to excellent internal consistency;
- Composite reliability (CR) and Average Variance Extracted (AVE) values met established thresholds, confirming construct reliability and convergent validity;
- SRMR = 0.084, suggesting good model fit;
- R^2 = 0.69 for "attitude towards staying" and R^2 = 0.39 for "intention to stay", demonstrating moderate to strong explanatory power.

The model tested eight hypotheses. Among the pull factors, the most significant predictors of a positive attitude towards staying were:

- Aspects related to the host country (H1), such as employment and quality of life (β = 0.79, $p < 0.001$),
- Institutional factors (H2), such as the perceived value of Finnish qualifications (β = 0.12, $p < 0.01$),
- Economic factors (H3), including job opportunities (β = 0.10, $p < 0.05$),
- Social influence (H6), reflecting societal norms and support systems (β = 0.11, $p < 0.01$).

In turn, attitude towards staying (H7) had a strong and direct impact on the intention to stay (β = 0.48, $p < 0.001$), and also mediated the relationship between certain pull factors and behavioural intention.

Two pull-related hypotheses did not show significant effects on students' attitudes towards staying, indicating a lower role of geographic proximity or social advice in retention decisions, that were:

- Environmental factors (H4); and
- Family/friend recommendation (H5).

The most significant push factor was barriers and challenges (H8), which had a strong negative impact on the intention to stay ($\beta = -0.30$, $p < 0.001$). Key barriers included difficulty finding work (67%), language obstacles (41%), and lack of social integration (31%), echoing findings from prior literature (Alloh et al., 2018; Khanal & Gaulee, 2019).

These findings affirm that student retention is influenced by a composite of push and pull factors, rather than a single variable. The study concludes that creating positive, inclusive policies, such as improved access to employment, local language training, and social support, is essential to increasing international student retention and addressing labour shortages in Finland.

The insights gained from this research are highly relevant to understanding Brazilian migration dynamics. As with Vietnamese and Finnish cases, career opportunities, institutional recognition, and integration capacity appear to be critical in determining whether international graduates stay or leave, key variables that should also inform our study of Brazilian migrants' motivations and destination preferences.

2.4.2 Factors influencing Vietnamese students' decisions to study abroad and remain in host countries: a PCA approach

A mixed-methods study was conducted with Vietnamese students to understand the motivational factors influencing their migration decisions (Nghia, 2019). The research aimed to explore motivations specifically related to decisions to study outside of Vietnam, as well as intentions to immigrate after completing their education. It followed a sequential exploratory design, beginning with 55 qualitative intercept interviews to identify key motivations, followed by a quantitative survey with 313 respondents to validate and quantify these motivations.

To examine the structure of student motivations, the authors applied Principal Component Analysis (PCA) to reduce dimensionality and uncover underlying factors. The dataset showed acceptable internal consistency ($\alpha = 0.74$) and adequate sampling (KMO = 0.75). Using Varimax rotation, two components were extracted, explaining 49.36% of the total variance, a result considered acceptable for exploratory research in the social sciences.

The first component, interpreted as Pull Factors, showed strong internal consistency (α = 0.83) and accounted for 29.18% of the variance. It included items related to improving international career prospects, gaining international experience, acquiring language proficiency, and seeking high-quality education. Most of its item loadings ranged between 0.65 and 0.82, which are considered strong and well above the standard 0.60 benchmark, indicating reliable grouping of related motivations.

The second component, interpreted as Push Factors, demonstrated acceptable internal consistency (α = 0.70) and explained 20.18% of the variance. It covered motivations such as dissatisfaction with domestic education, limited personal development opportunities, and political or family pressures. Although the total variance explained by this factor was modest, the item loadings (ranging from 0.50 to 0.71) remained within the acceptable range for exploratory models.

Together, these findings suggest that the PCA was statistically sound and aligned with the conceptual framework. The results confirm that pull factors exert a stronger and more consistent influence across the sample, while push factors show more variation, likely due to individual differences in personal, social, or political contexts. The methodology of PCA reinforces the validity of the pull-push distinction in understanding student migration motivations in this cohort.

2.4.3 Decision-making criteria for Taiwanese students: a TPB perspective

To investigate the decision-making criteria of Taiwanese students when choosing an English-speaking host country (the U.S., the U.K., Australia) for their studies, the Theory of Planned Behaviour (TPB) was applied (Gatfield & Chen, 2006). The TPB has been widely used in the social sciences to understand how intention is formed and translated into behaviour, particularly in contexts involving voluntary and future-oriented decisions such as migration and international education.

The TPB suggests that three core constructs influence a person's behavioural intention:

- Attitude towards the behaviour (AB) – the individual's positive or negative evaluation of performing the behaviour;
- Subjective norms (SN) – the perceived social pressure to perform or not perform the behaviour, typically shaped by important referents such as parents, peers, or society at large;
- Perceived behavioural control (PBC) – the perceived ease or difficulty of performing the behaviour, reflecting past experience and anticipated obstacles.

To operationalise these constructs, the authors conducted a four-phase study that combined qualitative and quantitative methods. Following preliminary focus groups and interviews, a

structured TPB-based questionnaire was developed and distributed to a sample of 700 students, from which 518 valid responses were collected for quantitative analysis. A total of 20 decision-related variables were included in the instrument.

The authors applied exploratory factor analysis (EFA) using principal axis factoring with oblique (Oblimin) rotation to validate the three-factor TPB structure. The dataset demonstrated excellent sampling adequacy, with a Kaiser-Meyer-Olkin (KMO) value of 0.893, and Bartlett's test of sphericity was statistically significant ($p < 0.01$), confirming the suitability of the data for factor analysis. The analysis yielded a three-factor solution explaining 44% of the total variance, which aligned well with the TPB framework and supported the structural validity of the model.

Subsequently, multiple regression analysis was conducted to evaluate the predictive power of the three TPB components. The model produced an R^2 value of 0.298, indicating that the three constructs jointly explained approximately 30% of the variance in students' intention to study abroad. All three predictors were statistically significant, with attitude towards the behaviour ($\beta = 0.285$) emerging as the strongest predictor, followed by subjective norms ($\beta = 0.239$) and perceived behavioural control ($\beta = 0.185$).

Regarding the factor structure, the most influential variables under attitude (AB) had loadings ranging from 0.622 to 0.779. These included the economic performance of the destination country, career advancement opportunities, academic reputation, and the perceived value of foreign degrees in Taiwan's labour market, all of which underscore the career-oriented and prestige-driven motivations of students.

For subjective norms (SN), variables loaded between 0.704 and 0.813, with the influence of family members being most prominent, followed by friends and teachers. These findings highlight the strong role of social expectations and interpersonal influence in shaping students' overseas education plans.

Though perceived behavioural control (PBC) was the weakest among the three constructs, it still contributed significantly to students' evaluations, with factor loadings between 0.467 and 0.741. Key concerns included the cost of living, tuition fees, and the difficulty of gaining admission, reflecting the practical constraints students considered when determining the feasibility of studying abroad.

Overall, the study provides empirical support for the Theory of Planned Behaviour in the context of international education decision-making. It demonstrates that Taiwanese students' intentions to study abroad are shaped by a combination of personal attitudes, social influences, and perceived constraints. These findings reinforce the multidimensional nature of student mobility and suggest that policy interventions or marketing strategies targeting international students should address not only institutional prestige and employability outcomes, but also family expectations and affordability.

In addition to examining general decision-making criteria, one study also explored how students' preferences varied by destination country (Australia, the United States, and the United Kingdom). Each country was associated with distinct motivational patterns that aligned with the TPB components (Gatfield & Chen, 2006).

Australia was primarily chosen for its affordability and proximity, with lower costs of living and tuition (loading = 0.699) and geographic closeness to Taiwan perceived as major advantages. These factors contributed to students' perceived behavioural control, as they viewed Australia as a more accessible and practical option. Marketing efforts also reinforced Australia's appeal by highlighting cultural familiarity, shorter travel time, and post-study work opportunities.

In contrast, the United States attracted students for its academic prestige (loading = 0.622) and career development potential, aligning closely with attitudes towards the behaviour. Students viewed the U.S. as offering elite education and stronger professional networks, particularly for those prioritising long-term employability.

The United Kingdom was favoured for its shorter programme durations (e.g., one-year Master's degrees; loading = 0.539) and its global academic reputation, appealing to students seeking quicker returns on their educational investment and internationally recognised qualifications. This preference again reflects strong attitudinal motivations, combined with practical considerations under PBC.

The authors noted the role of national marketing strategies in shaping students' perceptions. Countries such as Australia, the US, and the UK actively promote their institutions through campaigns emphasising academic excellence, professional pathways, and international alumni success stories. These narratives align strategically with students' existing beliefs, social norms, and expectations, reinforcing the TPB dimensions and influencing both intentions and choices.

Ultimately, the findings underscore how economic, academic, and social considerations interact differently across destination countries. For Taiwanese students, decisions were guided by a blend of career aspirations, financial feasibility, and social influence, highlighting the importance of destination-specific messaging that resonates with the full spectrum of motivational drivers.

2.4.4 EARLY EXPERIENCES AND MOTIVATIONS OF INTERNATIONAL STUDENTS IN THE UK

Qualitative and quantitative methods (interviews and questionnaires) were used to gain insight into the commencement experiences of first-year international students at a UK university (Cowley & Hyams-Ssekasi, 2018). The study emerged from an interest in exploring the motivations for studying abroad in addition to considering the role of the university's induction programs and any challenges faced by students upon their arrival. A mixed-methods approach was adopted, and the quantitative data generated from questionnaires

complemented the qualitative data generated from interviews conducted with 20 international business students from China, Nigeria and India, which further enhanced the understanding of students' academic and social transition.

Rather than applying inferential statistical models or formal theoretical frameworks such as TPB or PCA, the authors used thematic content analysis to identify recurring patterns across responses. While the study referenced concepts from push-pull theory, it was not deployed as a formal analytical model. Instead, the research relied on triangulation between data sources (survey and interview) to enhance validity and depth of insight.

Key findings indicated that students were motivated primarily by a) the UK's reputation for quality education; b) career advancement opportunities; and c) the prospect of gaining international experience. However, they faced significant challenges, including d) language barriers; e) adapting to academic expectations; and f) adjusting to new social norms. Generally, the induction processes were useful, but some gaps were identified related to late arrivals and participants not being aware of the UK educational systems. Moreover, participants highlighted their struggles with non-academic challenges including homesickness, money concerns and housing, which indicate that this type of support should not be limited to student's initial arrival/induction period.

The study further acknowledged the influence of marketing and recruitment communications in shaping students' decisions. Marketing campaigns will provide "pull" factors by promoting the UK's education brand, career potential and support networks for international students. Promotional messages focusing on the potential, academic reputation and support services that would resonate with student's primary motivations, and simultaneously reinforcing the idea that they would receive a well-engineered structure and support throughout their studies. In discussing marketing campaigns, focus will often be placed on indicating the support offered in relation to orientation, language support and housing, all of which are frequently cited as expectations and concerns in the present study. Moreover, for students from countries where financial costs are a primary consideration, the presence of financial incentives which may appeal to students, such as scholarships or flexible payment options are a critical issue.

2.4.5 KEY PREDICTORS OF MIGRATION INTENTIONS

Understanding the predictors of migration intentions is central to building robust and meaningful models. In order to adequately predict migration choices, it is important to recognize, understand and define the factors having consistent effects on future intentions to migrate. A wide body of literature has explored these motivations through a variety of methodological lenses, ranging from behavioural theory and regression analysis to exploratory mixed methods and comparative policy evaluation. While some studies use modelling techniques such as Structural Equation Modelling (SEM) to estimate the strength of

relationships between variables, others adopt alternative approaches, highlighting the diversity of tools available to understand and anticipate migration-related decisions.

Indeed, predictions in the migration literature have not been based exclusively on SEM. Partial Least Squares SEM (PLS-SEM) has been used to evaluate how institutional, economic, and social factors shape international students' intentions to remain in the host country (Nikou & Luukkonen, 2024). The Theory of Planned Behaviour (TPB) has also been applied alongside multiple regression to predict overseas study decisions (Gatfield & Chen, 2006). Additionally, Principal Component Analysis (PCA) has been used to reduce motivational variables into latent factors, which are then analysed in relation to study and stay intentions (Nghia, 2019).

Together, these methodologies provide important insights into the factors associated with migration behaviour, but they often prioritise explanation over prediction. That is, while they clarify which variables are meaningful, they do not always attempt to forecast migration intentions or outcomes in a way that could be generalised to new populations.

Our approach builds on these theoretical and empirical foundations but shifts toward prediction. This thesis builds on those foundations by applying a supervised machine learning framework to model and predict migration decisions. Rather than relying solely on pre-defined hypotheses, this approach allows the data to reveal non-linear interactions and variable importance with greater flexibility. The goal is not only to understand migration motivations but to develop a model that can predict migration patterns in a data-centric and scalable way.

To frame this contribution, the following sections synthesise the main variables identified in the literature as predictors of migration intention. These are grouped across four categories: socio-demographic characteristics, perceptions of host countries, contextual and relational factors, and behavioural and cognitive dimensions.

Socio-demographic characteristics

Numerous socio-demographic characteristics have emerged as notable indicators of migration inclination.

Age is a constant predictor of migration potential. Younger individuals are generally more accessible both physically and socially, and tend to seek international educational and occupational opportunities (Castelli, 2018). Youth tends to imply fewer obligations and greater elasticity that gives relocation more viability.

Gender also plays an important role in migration pathways, as historically men have migrated for work, while recent years have demonstrated an upward trend in female migration for educational and employment purposes. Women are increasingly migrating of their own accord with the ambitions of personal growth and professional opportunities (United Nations Network on Migration, 2018).

Education level as well is a strong predictor of migration stream. Higher education, as seen in the past, strongly predicts international migration, especially when an individual is pursuing an international work opportunity which reflects their level of education (Oxford Academic, 2018). Education typically provides individuals with a greater scope of awareness regarding global options, as well as a greater likelihood of connecting with the local labour market once they migrate.

Similarly, marital status impacts migration potential; single individuals, or individuals with fewer familial ties, are more manageable in terms of migratory angle (Oxford Academic, 2018). Conversely, families with dependents or members with complicated ties experience too many obligations and the possibility of leaving loved ones in an unfamiliar environment becomes more difficult.

More than likely, region of origin matters significantly. People from conflict motivated regions or regions that are facing extreme economic incidence should look to greater safety and prosperity abroad. This simply gives credence to the structural conditions of macro-level origin that impact intention in migration (United Nations Network on Migration, 2018).

Household size can also indirectly discriminate against migration commitment decisions. Larger households inherently increase financial strain on the individual and in turn that could lead the individual to migrate to improve income potential and generate support towards a family (Oxford Academic, 2018).

Finally, employment status is a major driver. Unemployment or underemployment pushes individuals to seek more stable opportunities abroad, making labour market instability a prominent push factor in migration studies (Castelli, 2018).

Perceptions of the host country

In addition to individual characteristics, perceptions of the destination country have been identified as significant influencers of migration intent.

Perceived characteristics of the host country, such as job opportunities, living standards, and quality of life, have been identified as some of the most compelling pull factors (H1) (Nikou & Luukkonen, 2024). Positive perceptions of the destination country may be more salient than conditions in the country of origin and can serve as a driving force for purposeful destination selection.

Institutional factors (H2) such as the reputation and credibility of educational institutions in the destination country, access to support services for academics, and availability of resources are also perceived to significantly impact intention to migrate early in education and career development, especially individuals who are students and young professionals.

Additionally, economic factors (H3), such as anticipated salary, anticipated employment opportunities, and general economic stability, are important factors. Countries that are

perceived to be economically secure and prosperous tend to attract a higher number of skilled migrants than countries that are economically less secure.

Social influence (H6) is defined as the role of social norms, expectations of peers, and support systems in the destination country which has been established as a prominent variable. Support systems facilitate the migration decision and make the post-migration transition period manageable and easy.

Moreover, positive attitudes towards non-return (H7) are associated with the intention to remain in the destination country after arrival. This notion encompasses both rational and emotional experiences in relation to the situational factors of living in the destination country.

However, barriers and challenges of migration (H8) such as language barriers, cultural distance, lived experiences of discrimination, and high cost of living, are considered push-back factors of migration. When barriers are experienced at a magnitude greater than perceived pull factors, the individual may choose not to migrate, or select an alternate destination.

Contextual and relational factors

Some studies further refine our understanding by introducing variables that combine personal, social, and contextual dimensions (Nghia, 2019).

Gender differences are again emphasised, with findings indicating that men are generally more influenced by push factors (Nikou & Luukkonen, 2024), such as unemployment, poverty, or political instability in their country of origin. In contrast, women tend to respond more strongly to pull factors, including the availability of support networks, better quality of life, and improved opportunities for education or employment in the destination country.

Study status (prospective vs. current students) influences migration expectations, with those already studying abroad more likely to develop intentions to remain. This suggests that experience in the host country alters perceptions and increases the desire to stay.

Field of study and language proficiency have also emerged as relevant predictors. Students in fields with global demand (e.g., STEM, business) are more likely to consider international careers, while higher proficiency in the host country's language significantly enhances integration potential and retention.

Family influence and financial support are strong social and economic anchors. Support from family can encourage migration, but financial limitations may also restrict access to international mobility.

Cultural familiarity, social networks, and prior international experience shape confidence and perceived feasibility of migration. Individuals with exposure to international contexts or existing connections abroad are more inclined to pursue migration paths.

Behavioural and cognitive dimensions

Building on the Theory of Planned Behaviour, one study examined how attitudes, norms, and perceived control inform students' decisions to study abroad (Gatfield & Chen, 2006).

Attitudes toward behaviour (AB), including the belief that studying abroad improves job prospects, is aligned with economic performance of the destination country, and enhances the value of degrees, were found to be the strongest predictors of migration intention.

Subjective norms (SN), defined as the influence of family, friends, teachers, and other reference groups, shape not only initial interest in migration but also the choice of destination.

Perceived behavioural control (PBC) encompasses factors such as cost of tuition, living expenses, course length, and the perceived difficulty of being accepted into foreign institutions. These perceptions can either facilitate or hinder migration planning.

Implications for predictive modelling

The identification of these variables is essential for building robust predictive models of migration. Grounded in empirical literature, the selected predictors — spanning sociodemographic characteristics, perceptions of destination countries, and behavioural intentions — provide a strong foundation for modelling the complexity of individual migration choices. Such models are not only useful for analysing past trends, but also for anticipating future flows and supporting institutional planning.

This thesis adopts a data-driven methodology that, while focused on individual-level data, is oriented toward prediction rather than explanation. By combining supervised learning techniques with practical dimensionality reduction through PCA, the research seeks to forecast real-world destination choices in a scalable and interpretable way. The following chapters will detail the methodological pipeline developed to implement this predictive approach.

2.5 BRAZILIAN MIGRATION: CONTEXT AND CHARACTERISTICS

Recent estimates indicate that over 4.9 million Brazilians were living abroad in 2023, with the primary countries of residence being the United States, Japan, Portugal, and Spain. Approximately 557,000 Brazilians reside in Europe, largely due to historical and linguistic ties with Portugal. North America continues to attract individuals in search of economic opportunities, while in Asia, Japan remains a significant destination, influenced by labour market demand and long-standing cultural connections rooted in Brazil's own immigration history. The global distribution of Brazilian emigrants is illustrated in Figure 6.

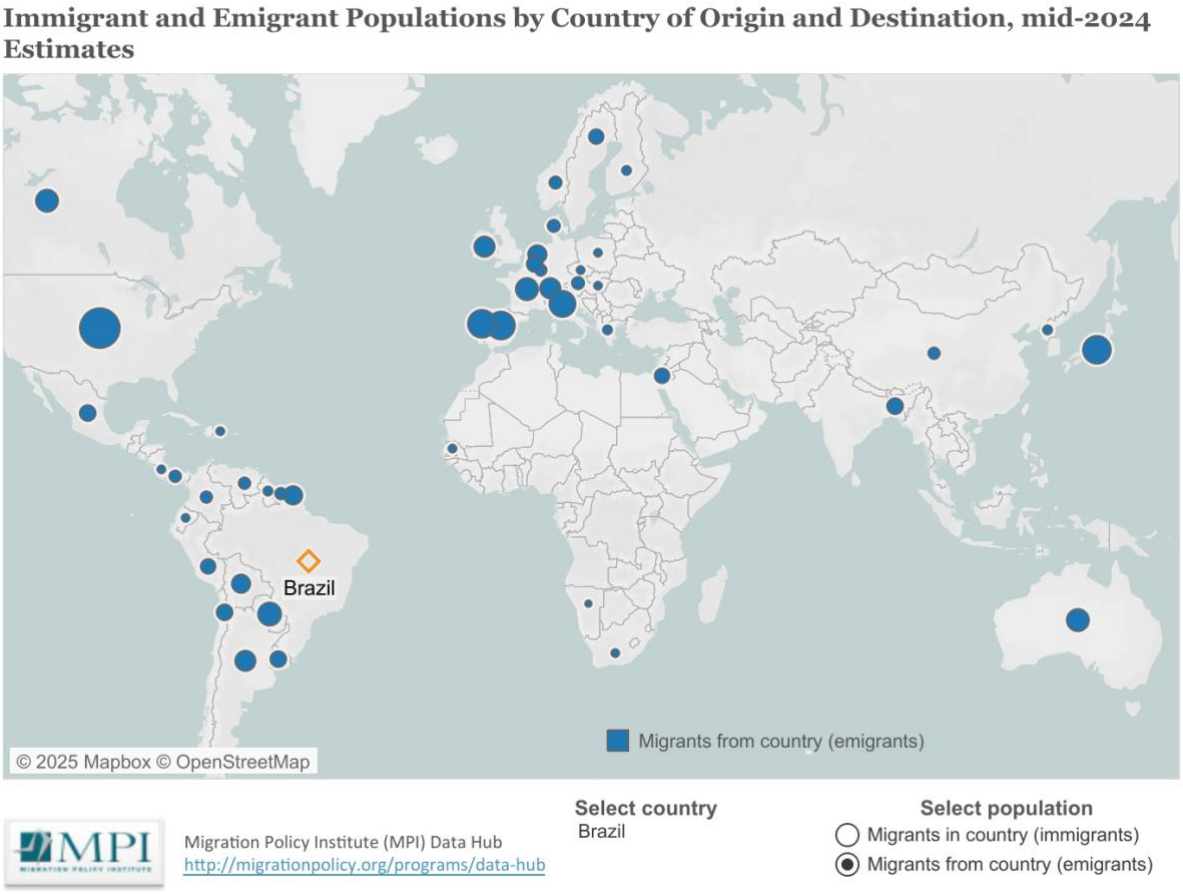

Figure 6 – Main countries of destination for Brazilian emigrants (mid-2024 estimates)

Although this thesis is directed toward understanding the rationales for Brazilian expatriates in Australia and Ireland, other studies, like the Expat Insider 2019 survey, capture the reverse perspective from what it is like for immigrants in Brazil. For Brazil, which ranks 61 out of 64 destinations in the Expat Insider survey, Brazil is one of the ten major countries having negative expat experiences. Brazil is known for its kind hospitality and citizenship integration, however it is crucial to recognize the fact that many immigrants throughout Brazil face substantial issues related to their safety, political stability, and life quality, which rank towards the bottom of the global ratings (International, 2019).

The main push factors identified include:

a) Safety & Security: Brazil scores last in the Safety & Security category, with a significant 61% of expats expressing concerns about personal safety, compared to just 9% globally. Additionally, 53% of respondents are dissatisfied with political stability, far above the global average of 17% (InterNations, 2019).

b) Quality of Life & Family Concerns: Brazil has low marks in the Family Life Index and is ranked second to last (35th of 36 countries). Expat parents raising kids in Brazil are especially concerned with safety issues. Approximately 54% of parents indicated they

were worried (the global average was only 9%). A significant portion of respondents (38%) rated education in Brazil poorly, more than double the global average (16%) (InterNations, 2019).

c) Ease of Settling In: Despite these challenges, Brazil fares better in social integration, ranking 37th for ease of settling in. Over 80% of expats find Brazilians friendly, surpassing the global average of 68%. However, language remains a significant barrier, as nearly 74% of expats find it challenging to live in Brazil without speaking Portuguese (InterNations, 2019).

In contrast, these unfavourable views of foreigners living in Brazil can also illuminate some of the reasons why Brazilians may choose to leave Brazil to seek safety, improved living conditions, and socio-economic stability. Understanding some of the motivations that lead foreigners to come to Brazil can help contextualize why many Brazilians consider leaving Brazil as a viable alternative.

This way of thinking can help to examine how the push factors that arise for foreigners in Brazil (such as insecurity and instability) overlap with motivations for leaving Brazil. Therefore, we can gain a more complete understanding of migration processes by addressing the push factors that cause Brazilians to leave, and the pull factors that foreigners in Brazil might consider when pursuing life in Brazil.

3. METHODOLOGY

This chapter outlines the methodology employed in the research that addressed the motivations and challenges of Brazilian emigrants from a data-centric perspective. The ultimate goal was to create a predictive model to forecast migration trends and inform decisions made by public and private institutions in the future.

The methodology is presented in two sections, namely: the creation of the dataset and the predictive modelling process. Unlike major works in the literature that typically rely on structural equation modelling (a linear approach) to assess the strength of relationships between hypotheses and the target variable, and owing to the complex, non-linear nature of the immigration phenomenon, this study adopted a supervised learning approach to generate predictions automatically.

This chapter details each phase of the methodology, including data collection, feature engineering, data preparation, and model training.

3.1 RESEARCH DESIGN

3.1.1 IDENTIFICATION OF KEY VARIABLES

The modality of the first phasing in this research objectives was centred on establishing the most likely variables that would significantly affect participant migration decisions. The research journey commenced with an extensive Review of Literature.

The literature review identified a wide range of predictors in the political, socio-economic, and cultural dimensions as predictors. The aim was to create an all-encompassing set of variables to inform survey design and serve as the basis for subsequent analysis. The literature consistently highlighted the relevance of age, gender, educational qualifications, marital status, household dimensions, and the region of origin, to be significant individual features. In addition, perceptions related to the destination country, its institutional prestige, visa policy, and social influence were all recognised as implicit and explicit drivers.

3.1.2 RESEARCH GAP AND OBJECTIVES

Despite the growing relevance of Brazilian migration, there is a notable lack of research regarding the formal understanding of the underlying motivations behind this movement, particularly from a market-oriented perspective. Most existing studies rely on descriptive analyses and do not incorporate predictive modelling techniques that could enable institutions to attract and better serve Brazilian emigrants. This paper aims to identify elements that drive migration decisions, their cause and effect relationships, and predictive

models that could be practically useful for facilitating institutions and decision making on behalf of Brazilian emigrants.

3.1.3 Push and Pull Factors Framework

To guide the design of the questionnaire and ensure theoretical consistency, the study adopted a multi-level analytical approach, rooted in established migration literature and adapted to support predictive modelling. This framework helped structure both the survey instrument and the variable selection process by mapping migration drivers across three analytical layers:

- *Macro-Level:* Political, economic, and environmental conditions in sending and receiving countries, such as economic instability, crime, and visa policies.
- *Meso-Level:* The influence of social networks, diaspora communities, and digital platforms that facilitate or constrain access to information and support during the decision-making process.
- *Micro-Level:* Individual motivations and aspirations, including socio-demographic characteristics, educational and professional background, and personal goals.

This layered structure informed the mapping of survey items to model features, supporting the construction of variables that reflect not only personal attributes but also contextual and informational factors. For example, questions about safety perceptions, access to information, or institutional trust were explicitly derived from the meso-level perspective, while items capturing motivations and family structure aligned with the micro level.

By incorporating this framework into the research design, we ensured that the dataset captured a wide spectrum of potential drivers of migration behaviour, improving the model's capacity to detect relevant patterns across different layers of influence. The multi-level approach also aligns with the literature reviewed in Chapter 2, offering a conceptual bridge between theory and methodology that reinforces the validity of our modelling strategy.

3.2 Survey Methodology

3.2.1 Questionnaire Design and Variable Mapping

The data collection stage was designed to capture both quantitative and qualitative data through an online survey, aimed at identifying behavioural patterns and informational needs relevant to the migration decision-making process. The survey explored where Brazilian migrants seek information, the factors that influence their decisions, and how prepared services are to receive them. The questionnaire was divided into two main thematic areas:

1. *Pre-Migration Factors:* This section covered respondents' circumstances prior to migrating, including socio-demographic characteristics (e.g., age, education level), professional background, financial resources, family structure, and sources of information used during the decision-making process.
2. *Post-Migration Factors:* This section assessed respondents' experiences and perceptions in the host country, including feelings of safety, economic opportunities, integration, and future intentions.

Each item in the survey was mapped to a specific feature in the dataset, ensuring alignment with the theoretical constructs outlined in the literature and operationalised through the multi-level framework described earlier. These variables served as inputs for the predictive model.

For full transparency and reproducibility, the complete list of survey items, grouped by thematic section, is provided in Appendix A – Survey Questionnaire.

3.2.2 DATA COLLECTION PROCESS

The survey was disseminated via Facebook groups targeted at Brazilian migrants who had either already emigrated or were in the process of making migration decisions. These groups were selected based on high engagement levels and relevance to the target audience.

Overall, the data collection phase took place from 18 February to 31 March 2025, and resulted in 329 responses. After removing entries completed in under 60 seconds, those that failed the attention check, and those missing either the year of immigration or the motivations section, we were left with 303 valid survey responses. The respondents were located in 22 countries on 5 continents giving the sample rich variation.

However, to ensure uniformity and modelling quality, the final predictive model focused exclusively on responses from Australia and Ireland**,** the two most frequently selected destinations. This filtering allowed for a more robust comparison and reduced issues related to data sparsity**.**

3.2.3 COUNTRY FILTERING AND STRATEGIC FOCUS

Data were collected from around the world, but responses were not equally distributed, as certain countries yielded just a few submissions. Figure 7 presents the number of Brazilian emigrants who selected each country as their destination.

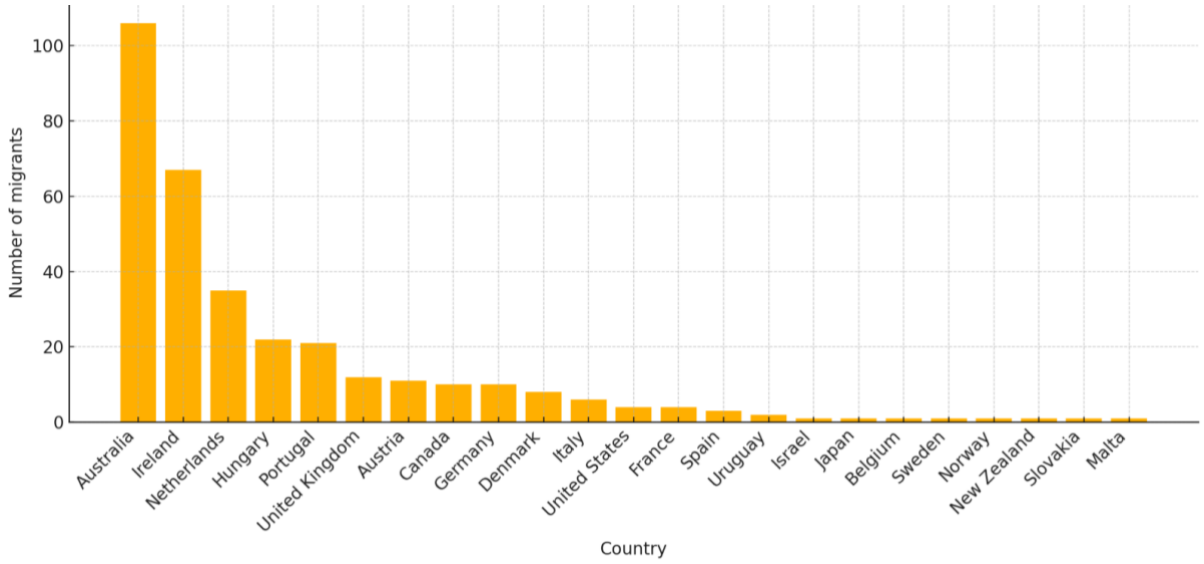

Figure 7 – Distribution of Brazilian migrants by destination country based on survey responses

The top 3 destination countries for Brazilian migrants in this sample are:

- **Australia**: 106 responses (32.22%)
- **Ireland**: 67 responses (20.36%)
- **Netherlands**: 35 responses (10.64%)

These three countries account for nearly two-thirds (63.22%) of all respondents, with the first two (Australia and Ireland) representing the majority at 52.58% of the sample. The basis for successful predictive modelling is, ideally, a critical mass of data per class. For this reason, Australia and Ireland were used in the final modelling stage due to sufficient representation overall and a political will around this topic.

This choice posed a methodological challenge, as both countries share several characteristics, including the English language, structured visa systems for students and skilled workers, and similar economic and lifestyle appeal. Accurately distinguishing between them based on migration motivations and sociodemographic profiles required careful feature selection and modelling strategies.

However, this also represents a key contribution of the study: being able to identify subtle differences using predictive modelling is especially valuable for stakeholders in both the public and private sectors, including universities, recruitment agencies, and immigration services.

3.2.4 SURVEY STRUCTURE AND DATA PRE-PROCESSING

The survey was designed to collect both quantitative and qualitative data. Questions were developed based on insights from the literature review, drawing on previous studies that identified push and pull factors, social influence, and institutional context. Each question was then mapped to different variables with the goal of improving the prediction model accuracy. To ensure quality, multiple validation checks were applied:

- Responses with inconsistent or missing critical data were removed;
- Attention-check questions were embedded in the form;
- Open-ended responses were categorised and subsequently converted into numeric or binary values.

The final dataset was structured, cleaned, and coded to optimise it for machine learning. During the data preparation stage, additional features were engineered to better represent behavioural and motivational aspects relevant to the research.

Missing data were minimal, and no artificial imputation was applied to avoid introducing artificial statistical patterns that could bias the model towards synthetic trends rather than true underlying behaviour. Instead, missing values were treated as a distinct category labelled 'missing' to capture any meaningful signal conveyed by their absence.

After referred pre-processing, the final dataset comprised 55 engineered features per immigrant ready for predictive modelling.

3.3 PREDICTIVE MODELLING APPROACH

This section outlines the methodology employed to construct a model that predicts the destination country of Brazilian emigrants, based on individual and contextual variables gathered via the survey. The modelling pipeline is illustrated in Figure 8, with each step arranged in logical sequence, encompassing data pre-processing, dimensionality reduction, model selection, and evaluation.

The process began with the input of survey data comprising 55 features. These were pre-processed and subjected to dimensionality reduction via *Principal Component Analysis (PCA)*. The selected features were then used to train a supervised machine learning model, specifically, a *Random Forest classifier*, which output the predicted destination classification: Australia or Ireland.

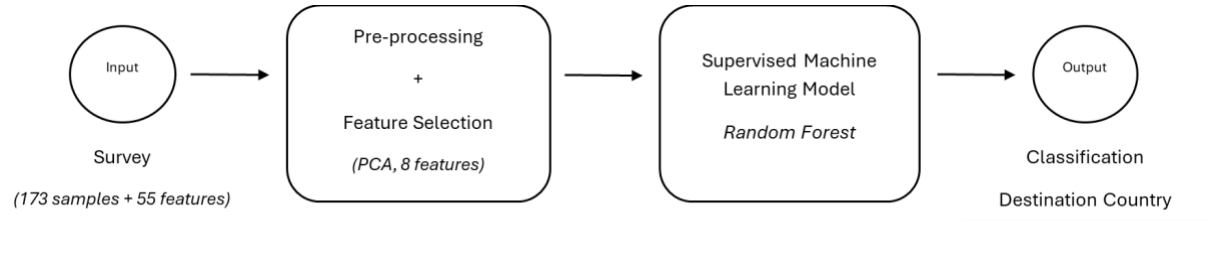

Figure 8 – Predictive modelling pipeline used to classify Brazilian emigrants by destination country.

3.3.1 DATA PRE-PROCESSING AND CLEANING

Given the self-administered nature of the survey, the initial dataset was expected to contain a non-trivial amount of noise. The initial dataset contained 303 responses and 55 variables. Prior to modelling, we restricted the dataset to the two countries with the highest response rates — Australia and Ireland — reducing the sample size to 173. This decision ensured a statistically meaningful amount of data per class, minimised the risk of overfitting, and improved class balance.

Key pre-processing steps included:

- Feature Engineering: All categorical variables were transformed into numeric or binary variables to enable predictive modelling. For multiple-choice questions allowing more than one selection, we applied a one-to-many encoding strategy, creating separate binary features for each option. For single-answer questions, responses were encoded as binary values (yes = 1, no or absence of selection = 0). This transformation ensured consistency and interpretability across all features.
- Missing Values: During feature selection, missing data were not imputed, as imputing them to avoid introducing artificial patterns or distort true relationships in the data. For model training, missing entries were encoded as a distinct category ("missing") to reflect the potential informational value of absence itself, particularly relevant for self-reported survey data. This approach avoided the bias that might arise from using central tendency measures such as the median or mode for imputation.
- Outlier Treatment: The dataset did not contain a significant number of outliers, as most numerical entries originated from categorical or ordinal survey questions and were encoded as binary or ranked variables. Absences were recorded as zeros. The only variable that stood out as a true numerical input was the year of migration. To make it more meaningful and comparable, we engineered a new feature: migration duration.

$$\textit{Years since migration} = 2025 - \textit{Year of migration}$$

- Data Normalisation: Z-score normalisation was applied to all numerical features, including the newly created migration duration, ensuring that variables on different scales contributed equally to the model.

3.3.2 BASELINE MODEL ESTABLISHMENT

In predictive modelling, it is important to establish a reference performance, commonly referred to as a baseline, that allows us to identify how effective any proposed classification model is. In statistical terms, baseline performance is typically represented by a naïve classifier, one that does not use any other information apart from the distribution of classes in the data set.

In our dataset, which comprised Brazilian migrants to either Australia or Ireland, the overall distribution was imbalanced (i.e., a majority class in Australia and a minority class in Ireland). We performed stratified sampling around the target variable to maintain original proportions of each class in the training and test sets.

In the test set, approximately 61% of the observations corresponded to migrants to Australia, and 39% correspond to migrants to Ireland. If we were to think about a naive classifier that predicted what the majority class was (i.e., Australia), our baseline or classification accuracy would be 61%. If the model always predicted the minority class (Ireland), our classification accuracy would be 39%.

These conditional baselines served as reference points for model evaluation and also offered a pragmatic approach for evaluation ensuring predictive algorithms tested against a fair baseline reflecting the class distribution within the evaluation set.

Establishing this baseline was critical because it defined the minimum acceptable performance threshold any sophisticated predictive model must surpass. If a predictive model does not exceed the baseline, it does not add any value to the decision making process more than random guessing along the class distribution in the target variable. This will allow us to objectively differentiate the incremental value added and actual predictive power evolved in creating a classification model using more sophisticated statistical techniques and machine learning.

3.3.3 FEATURE SELECTION TO REDUCE DIMENSIONALITY

It was essential to define an effective feature selection and data preprocessing strategy to optimise performance while ensuring generalisability. OThe initial dataset comprised 303 samples, a relatively small volume for machine learning purposes, and included 55 independent variables, which posed an unfavourable trade-off between dimensionality and sample size.

This limitation was recognised early on, the dataset was filtered before pre-processing. To minimise the risk of overfitting due to class imbalance and ensure statistical reliability, we restricted our training set to the two countries with the highest number of respondents: Ireland and Australia. This filtering step reduced the sample size to 173 entries, providing a more balanced and focused base for training while preserving the dataset's integrity.

Many studies resort to artificial sampling techniques such as SMOTE or other data smoothing methods to compensate for small datasets. However, we deliberately avoided these approaches to preserve the real-world nature and generalisation properties of our ML models. The goal of this research is not only to produce predictive accuracy, but also to ensure that the insights generated can be meaningfully applied to real migration scenarios without artificial bias. Nonetheless, we tested the use of SMOTE on the training set to balance the underrepresented Ireland class and assess whether it could improve overall model accuracy. However, the final results were equivalent to the original baseline, and the SMOTE-based models did not demonstrate stability when comparing validation and test performance, reinforcing our decision to avoid synthetic sampling in the final pipeline.

Given the high dimensionality relative to the sample size, we conducted a series of warm-up experiments to identify which features held the strongest relationship with the target variable. The aim was to reduce dimensionality by selecting the most relevant features (up to 10) to serve as input for the machine learning model. To accomplish this, we tested four complementary feature selection methods:

1. **Mutual Information (MI):** This filter-based method estimates the dependency between each feature and the target variable. As it captures non-linear and non-parametric relationships, MI is particularly useful when dealing with complex interactions. We applied MI in five different experimental configurations to evaluate the consistency and relevance of the top-ranked features. However, MI alone proved insufficient to eliminate data bias, as it consistently selected highly correlated features across different splits. This redundancy could introduce noise and compromise model generalisation.
2. **Correlation analysis:** As a complementary step, we used correlation matrices to detect multicollinearity and eliminate redundant features. This allowed us to reduce dimensionality by eliminating highly correlated features, especially in the early stages of model simplification.
3. **Principal Component Analysis (PCA):** PCA is a transformation-based technique that reduces dimensionality by projecting features onto a new set of orthogonal components that explain the highest variance. It is especially useful for compressing correlated variables. PCA proved to be the most effective method for feature selection in our study, providing better performance results.
4. **Manual feature elimination:** After assessing the outputs of MI and PCA, we manually reviewed the top-ranked features to identify semantic redundancies and

refine the selection. Observing that in linear regression models, highly correlated features may introduce bias and lead to near-singular matrices, which compromise the model's predictive power, we attempted a manual inspection to remove features "by eye", aiming to retain the most distinct and meaningful variables. This step ensured that domain knowledge informed the final decision and improved overall model interpretability.

As a result of this structured selection pipeline, we were able to reduce the initial 55 features to 8 predictors with high informational value using PCA. This dimensionality reduction not only mitigated the risk of overfitting but also enhanced model performance and interpretability.

In the following modelling stage, we trained algorithms such as K-Nearest Neighbours, Logistic Regression, Random Forest, and Neural Networks using the refined dataset. The outcome was a more efficient and generalisable predictive model, supported by a rigorous and iterative selection process that balanced statistical robustness with practical relevance.

3.3.4 PCA FOR FEATURES SELECTION

To support the development of a predictive model that classifies migrants' destination country, we applied Principal Component Analysis (PCA) was applied to the complete set of survey variables. The objective was to identify the features with the highest impact on the target classification "Australia/Ireland," referring specifically to respondents who actually migrated to these countries.

The PCA method identifies linear combinations of variables that explain the maximum variance, revealing underlying patterns without supervision. By focusing on the most explanatory components, we isolated noise and avoid overfitting. Highly correlated variables appear together in the same component, allowing us to group or eliminate redundancies, and the component loadings indicate which features co-occur with the pattern of interest ("Australia/Ireland"), guiding attribute prioritization.

To establish a reference point (baseline), we initially used all features collected in the survey to train the PCA. The idea was to obtain a global decomposition of the variable space without any segment-specific bias, ensuring that the principal components reflected the total data variability.

While mutual information measures the dependency between each feature and the target independently, it cannot account for correlations among features or reveal latent structures in the data. In contrast, PCA:

- Captured joint variability: It found linear combinations of features that explained maximal variance, uncovering multivariate patterns that individual-feature methods miss.

- Reduced multicollinearity: Highly correlated variables load onto the same component, letting us group or eliminate redundant features.
- Avoided discretization pitfalls: Mutual information on continuous or ordinal variables often requires binning, which can introduce arbitrariness and information loss, issues PCA sidesteps by working directly on standardized values.

As part of this study, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the survey dataset and identify variables most relevant for differentiating between Brazilian migrants who chose Australia and those who chose Ireland. Unlike the applications of PCA reviewed in the literature, which often focus on identifying latent structures or linear relationships, here PCA was used pragmatically as a feature selection tool, aimed at isolating the variables that best explain variance associated with the observed outcome, namely the respondent's chosen destination.

We have started the feature selection process of loading our data. The survey results were imported from CSV files into KNIME, a low-code data analytics platform that we used to carry out this research. At this stage, the data included 55 variables, ranging from socio-demographic attributes (e.g., age, income), migration context (e.g., year of migration, decision duration), to motivational and informational factors (e.g., sources of information, perceived barriers). The dataset included both continuous and categorical variables (encoded as binary dummies).

PCA was performed on a dataset of 55 variables using Z-score normalisation (mean = 0, standard deviation = 1) to ensure equal contribution across different scales (e.g., years vs. binary dummies). Since missing values represented less than 2% of the data and were randomly distributed, they were left untreated to avoid introducing artificial structure. The PCA decomposed total variance into orthogonal components, and a scree plot (Figure 9) was used to determine the number of components to retain. We initially considered variables with high absolute loadings across the top components (using |0.3| and |0.5| as thresholds for moderate and strong influence, respectively).

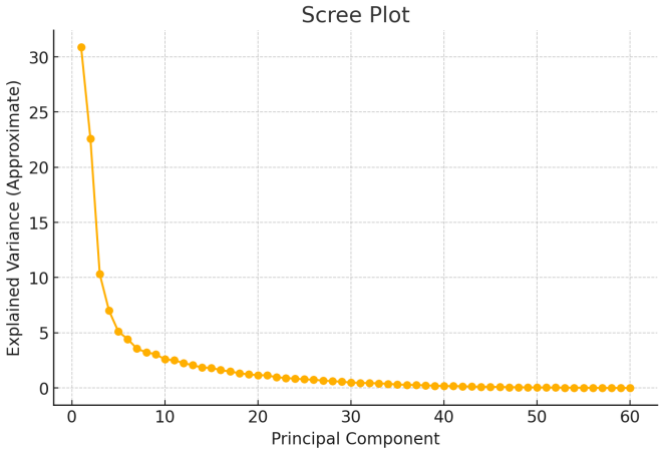

Figure 9 – Scree plot used to guide Principal Component Analysis (PCA) feature selection.

The scree plot displays the explained variance associated with each principal component derived from the original dataset. A sharp drop followed by a gradual flattening (the "elbow") indicates the point at which additional components contribute minimally to the total variance. In this analysis, the plot was used to determine how many components to retain for dimensionality reduction, ensuring that the most meaningful variance in the data was preserved while reducing noise and redundancy.

However, this approach produced marginal gains in model performance. It disproportionately favoured qualitative variables and failed to highlight features that meaningfully contributed to class separation in predictive terms. In response, we refined our strategy.

Given these limitations, we adopted a revised approach. Instead of summing loadings across components, we focused on a single principal component, specifically, the one most strongly aligned with the “Country” variable (Australia vs. Ireland). By ranking features based on their contribution (loading strength) to this single "Country" component, we isolated the eight variables that most clearly captured variance associated with destination choice.

However, when the classifier was trained using the features derived from the global-baseline PCA, its performance dropped significantly, as shown below:

- Overall accuracy: 57%
- Australia accuracy: 81% (above the conditional baseline of 61% for Australia)
- Ireland accuracy: 20% (below the conditional baseline of 39% for Ireland)

This result demonstrated that features maximising total variance do not necessarily discriminate well between Australia and Ireland. Despite strong performance for Australia, the model largely failed to recognise Brazilian migrants to Ireland.

To address this, we refined our methodology by filtering the dataset for Australia and Ireland samples prior to applying PCA. This step removed variance unrelated to our specific classification task and focused the feature selection process on attributes most informative for distinguishing between the two destinations. Z-score standardisation ensured fair weighting across all variables, and the decision to leave missing values untreated avoided the introduction of bias in a low-prevalence context. The resulting set, comprising eight high-contributing variables plus the target, constituted a compact, high-value feature subset optimised for supervised modelling.

This targeted ranking method provided a clearer signal. When used to train classification models, these top eight features led to improved accuracy across both validation and test sets. Notably, they allowed the Random Forest model to achieve higher performance and greater consistency, especially in identifying the minority class (Ireland), validating the effectiveness of this refined PCA-based feature selection strategy.

1. **Barriers to staying – Climate-related barriers**
2. **Future plans**
3. **Migrated alone or with others**
4. **Motivation – To learn/improve the local language**
5. **Barriers to staying – High cost of living**
6. **Access to information – WhatsApp groups**

Based on the final set of variables selected via PCA, we observed that all features selected by PCA are qualitative, relating to perceptions, motivations, or experiences, such as barriers, future plans, or information sources, with no quantitative variable in the first 15 ranked features, missing the objective and sociodemographic aspects of our research. This predominance of qualitative inputs introduces an additional challenge when training the model, as these types of variables are often subjective, encoded in binary or ordinal formats, and may vary considerably between respondents, i.e., technically, they exhibit lower signal-to-noise ration.

At the same time, the outcome revealed a key insight: the sociodemographic profiles of Brazilian migrants to Australia and Ireland are largely similar, and thus do not strongly distinguish between destinations. As such, the choice between the two destinations appears to depend less on structural characteristics like education or age, and more on push and pull factors, such as perceived barriers to settlement, future intentions, or the type of information accessed during the decision-making process.

Given this initial imbalance between qualitative and quantitative features, and after observing their influence on model performance, we chose to manually reintroduce two sociodemographic variables to rebalance the feature set:

7. **Level of education**
8. **How long ago they migrated**

This decision, though empirical, helped to stabilise and improve the model, particularly in identifying patterns linked to time and educational background.

3.3.5 MODEL TRAINING AND VALIDATION STRATEGY

We tested several classification algorithms to determine which would best support the study's objective of predicting migration destinations. The models evaluated included:

- Logistic Regression
- K-Nearest Neighbours (KNN)
- Decision Trees
- Neural Networks
- Random Forest

Each model was trained on the dataset using both the full and PCA-reduced feature sets. Performance was evaluated based on overall accuracy and class-specific accuracy for both Australia and Ireland. Models that failed to outperform the conditional baseline of 61% for Australia and 39% for Ireland were discarded.

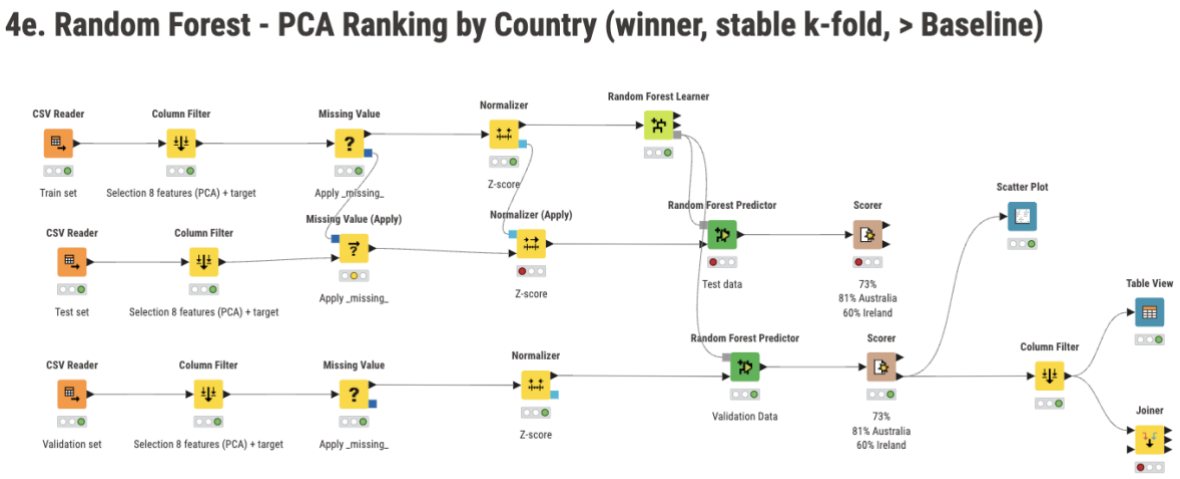

Figure 10 – Final predictive modelling pipeline using Random Forest classifier in KNIME.

The Random Forest algorithm emerged as the most robust choice due to its:

- Resistance to overfitting in small datasets;
- Ability to model non-linear interactions;
- Built-in mechanism for handling correlated features.

Model evaluation was conducted using k-fold cross-validation, comparing precision, recall, F1-score, and overall accuracy. The final model was trained using the 8 PCA-selected features and demonstrated superior predictive capacity while preserving generalisation and real-world applicability.

The detailed results of model performance and comparative metrics are discussed in the next chapter.

3.4. METHODOLOGICAL CHALLENGES

Our dataset presented several substantial challenges that required thoughtful methodological choices, significantly influencing our analytical approach.

a) **Noisy data and missing values:**
Given the self-administered nature of the survey, the data collection process inherently introduced noise. Participants entered their responses freely, resulting in highly variable free-text entries, inconsistent coding of categorical answers, and the occasional presence of outliers. Due to these discrepancies, we needed to do a lot of manual data cleaning and validation to ensure the data was consistent and accurate. Also, even if only 2% of the Australia/Ireland subset was missing data and seemed randomly distributed, we decided not to introduce synthetic input data. We did not want to create some false patterns or biases that disrupted the real relationships in the data, as we believed this would impact the integrity of our analyses.

b) **Limited sample size:**
Another critical limitation was our relatively modest sample size, consisting of approximately 303 respondents. Such a limited dataset restricts the complexity and types of models that can be reliably trained, as more sophisticated and flexible algorithms generally require larger amounts of data to capture subtle patterns without overfitting. Thus, our methodological choices had to balance model complexity against the inherent risk of overfitting due to limited data availability.

c) **Country selection constraint:**
At first, our survey attracted responses from participants who relocated to a variety of countries. However, most destination countries had only very few respondents, causing sparse and inadequate data for substantial modelling. Therefore, we took the pragmatic approach to focus solely on Australia and Ireland, which had the most responses. While this was mandatory for our analysis, it also presented a further research problem: Australia and Ireland have significant similarities related to, language, cultural background, exchange-visa programs availability and access, and financial attractiveness (e.g., reputation for high salaries). These similarities make it inherently difficult to distinguish between the two destinations. However, this difficulty also strengthened the study's contribution. Successfully classifying migrants between such similar countries provides valuable insights for institutions involved in international education, mobility services, and public policy.

d) **Class imbalance in filtered data:**
By narrowing our analysis to responses related solely to Australia and Ireland, our usable dataset became even smaller and exhibited pronounced class imbalance, approximately 61% Australia and 39% Ireland. Class imbalance poses a critical issue

because predictive models, if not adequately calibrated, naturally gravitate toward predicting the majority class to maximize accuracy. Consequently, models risked overpredicting migration to Australia at the expense of correctly identifying respondents likely to migrate to Ireland, thereby reducing overall predictive validity, particularly in the minority class.

e) **No synthetic sampling:**
In light of the substantial class imbalance, we chose not to employ synthetic oversampling techniques such as SMOTE. This was contingent on our initial experiments which illustrated that it was possible to deliver acceptable predictive performance without bringing in artificially generated samples. However, this presented additional complexities, as we had to be diligent in scrutinizing and validating our model performance, particularly regarding the accuracy of predictions of the minority class.

f) **Lack of public data and benchmarks:**
Another important challenge was the absence of public datasets or open benchmarks for this specific type of migration prediction problem. To the best of our knowledge, there were no publicly available samples or prior research providing structured, ready-to-use data for supervised learning focused on migration destination prediction at the individual level. As a result, the entire dataset had to be created from scratch, from survey design and data collection to data cleaning, encoding, and preprocessing, and create a complete modelling pipeline tailored to the problem. This meant that each step required careful methodological consideration, without the advantage of relying on existing templates, frameworks, or baselines to guide decision-making or validate our approach.

g) **Predictive versus explanatory objectives:**
Unlike most migration studies, which focus on explanatory models such as Structural Equation Modelling (SEM) to understand relationships between variables, this study was explicitly predictive. The goal was to forecast destination choices, not merely explain them. This difference posed a challenge in comparing our approach to existing research, as few studies had attempted similar predictive tasks using supervised learning techniques.
These cumulative challenges, ranging from noisy data to the absence of predictive benchmarks, shaped the methodological pipeline. To respond to these issues, we adopted a focused approach: restricting the dataset to Australia and Ireland, applying PCA for feature selection, and carefully validating the predictive models. This strategy ensured that the final model remained relevant and robust, despite the constraints of the dataset.

These cumulative challenges — noisy self-reported data, limited and imbalanced samples, and the necessity of predicting outcomes at the individual level rather than aggregate migration flows, as seen in traditional forecasting studies — critically impacted our analytical pipeline. In response, we adopted a targeted strategy: filtering data explicitly to include only respondents migrating to Australia and Ireland, and subsequently applying PCA-based feature selection prior to training predictive models. This methodological pathway ensured that our final models were both relevant and robust within the constraints and complexities of the dataset.

4. RESULTS AND DISCUSSION

This chapter presents the results of the predictive modelling process aimed at classifying the destination country — Australia or Ireland — based on individual and contextual features gathered through the survey. The performance of each algorithm is compared using overall accuracy, class-specific accuracy, and balanced metrics to ensure fairness and robustness in classification.

4.1 DESCRIPTIVE OVERVIEW OF THE DATASET

Before presenting the results of the predictive modelling, this section provides a descriptive overview of the filtered dataset used in the analysis. It focuses on the 173 valid responses from Brazilian emigrants who reported having moved to either Australia or Ireland, the two most frequently selected destinations in the survey (see Figure 7). This focus ensures sufficient sample size for each class, enhancing the reliability of the classification models.

From this subset:

- 106 participants (61%) migrated to Australia
- 67 participants (39%) migrated to Ireland

This distribution served as the foundation for the baseline model. For example, a simple classifier that always predicted “Australia” would reach 61% accuracy. However, this would not provide useful insights, especially in identifying migrants to Ireland. Therefore, more advanced and balanced models were required.

In addition to preparing the data for modelling, this section aimed to explore whether any specific patterns or differences could already be observed across key variables. By examining demographic characteristics, motivations, and access to information, the aim is to detect early signals that might inform prediction or guide feature selection.

At first glance, one might expect clear demographic contrasts between migrants to these two destinations. However, as the descriptive analysis reveals, the overall profiles are remarkably similar. Both countries attract young, educated Brazilians from middle- to upper-middle-income backgrounds. This similarity highlights a central challenge in our study: when conventional variables such as age, education, and income show minimal variation between groups, what then drives the choice of destination?

This finding strengthens the case for adopting dimensionality reduction (via PCA) and supervised machine learning techniques to capture the nuanced, possibly non-linear patterns underpinning individual decision-making. It also reinforces the importance of examining more

qualitative dimensions — such as motivations, perceived barriers, and sources of information — which may operate in combination to explain migration choices.

The following section delves into these socio-demographic variables in greater detail, establishing the foundation for feature selection and modelling.

4.1.1 KEY SOCIO-DEMOGRAPHIC CHARACTERISTICS

Understanding the socio-demographic profile of Brazilian migrants is a critical first step in exploring the underlying factors influencing destination choices. This section examines variables such as age, education, income, region of origin, relationship status, and length of stay to determine whether these structural attributes can help distinguish between migrants who chose Australia and those who chose Ireland. While this descriptive analysis does not yet draw on the predictive model, it helps establish whether observable demographic patterns might serve as early indicators, or whether subtler, behavioural features must be prioritised in the modelling process.

Most Brazilian migrants to both Australia and Ireland reported having migrated between the ages of 25 and 34. However, as shown in Figure 11, proportional differences emerged: Ireland attracted a younger migrant profile, with a higher share of respondents aged 18–29, often at the beginning of their academic or professional careers. Australia, on the other hand, had a larger share of migrants aged 30–39, suggesting a tendency to attract individuals with more established careers or qualifications.

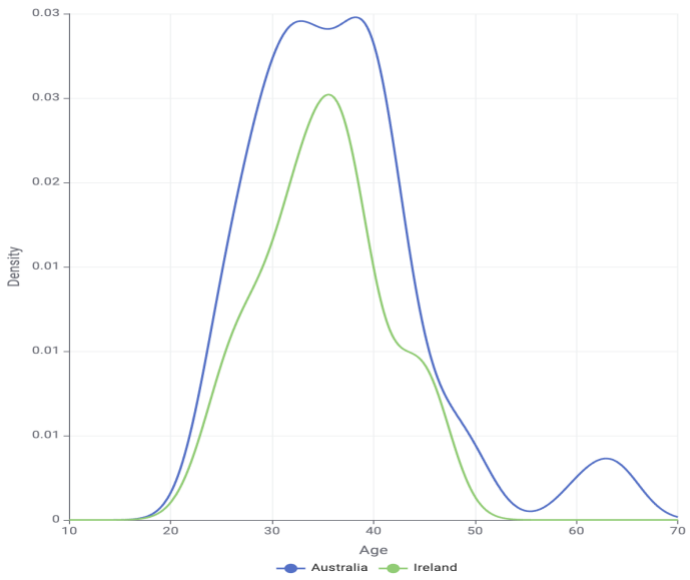

Figure 11 – Kernel density estimation of respondents' age by destination country. This plot illustrates the distribution of age among Brazilian migrants heading to Australia (blue) and Ireland (green).

This distinction indicates that, while both countries appeal to migrants in their early working years, Ireland may be more accessible or appealing to younger migrants, such as students or early-career professionals, whereas Australia may promote more experienced or credentialed candidates, potentially due to its skilled migration system. This trend is further reflected in the educational background of respondents, which reveals differences in the academic qualifications held by migrants to each destination.

Both groups exhibited a high level of education, as shown in Figure 12, with Bachelor's and Master's degrees being the most common. However, when adjusting for the number of respondents, Australia has a higher proportion of highly educated individuals, particularly those holding postgraduate and Master's degrees. This may reflect Australia's emphasis on skilled migration pathways that favour qualified professionals. The broader educational profile observed among migrants to Ireland, including a higher share of individuals with incomplete degrees or technical education, may be partially explained by their younger average age at migration, which often coincides with earlier stages of academic or professional development.

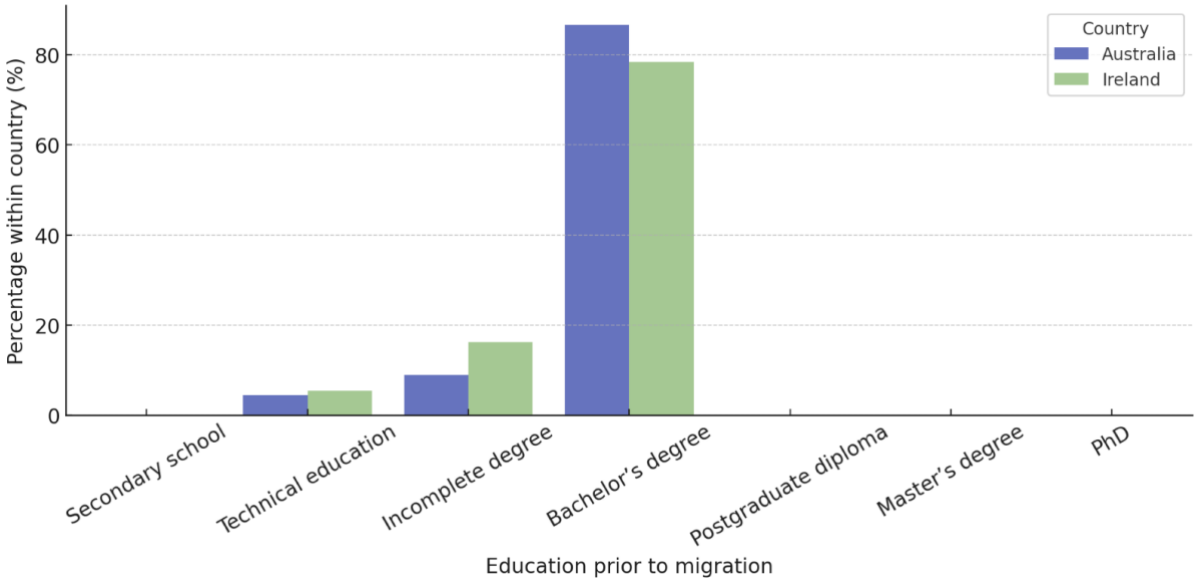

Figure 12 – Educational level prior to migration among Brazilian migrants in Australia and Ireland. Percentages represent the distribution within each country.

The findings on *household income in Brazil before migration* (Figure 13) suggest that Australia proportionally attracted more migrants with higher purchasing power. Several factors may explain this trend:

- Stricter requirements within the Australian visa system, including proof of professional qualifications and financial means;

- Higher perceived costs associated with the immigration process and the cost of living, even at the planning stage (as further illustrated in upcoming charts);
- Greater use of agents and professional consultancies during planning, which demands a higher initial financial investment.

Ireland, on the other hand, with easier entry through short-term student visas and strong informal support networks (such as Facebook groups and friends), may be perceived as a more accessible alternative for those with lower income levels in Brazil.

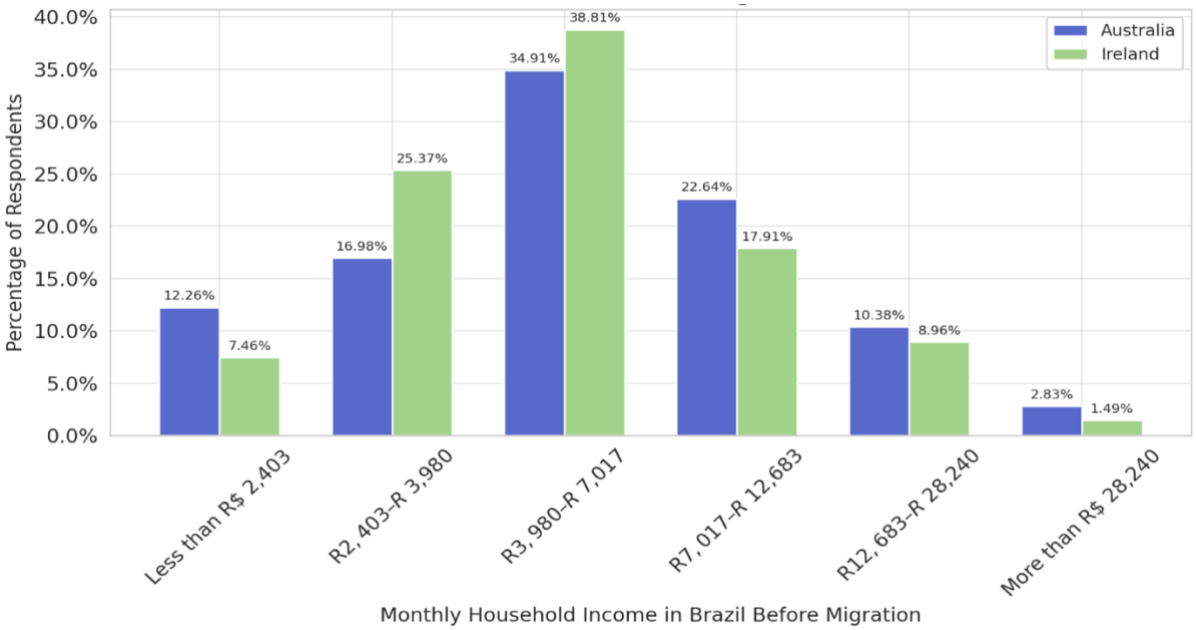

Figure 13 – Monthly household income in Brazil prior to migration among Brazilian migrants to Australia and Ireland.

The socioeconomic distinction is essential for private institutions when defining persona profiles, acquisition channels, pricing strategies, and the services offered in each country.

Given that educational attainment and household income levels are often associated with region of origin within Brazil, we examined whether certain regions were more likely to send migrants to Australia or Ireland, and whether any discernible trends emerged. The bar chart on *region of origin* (Figure 14) illustrates the internal percentage share of each Brazilian region sending migrants to each destination:

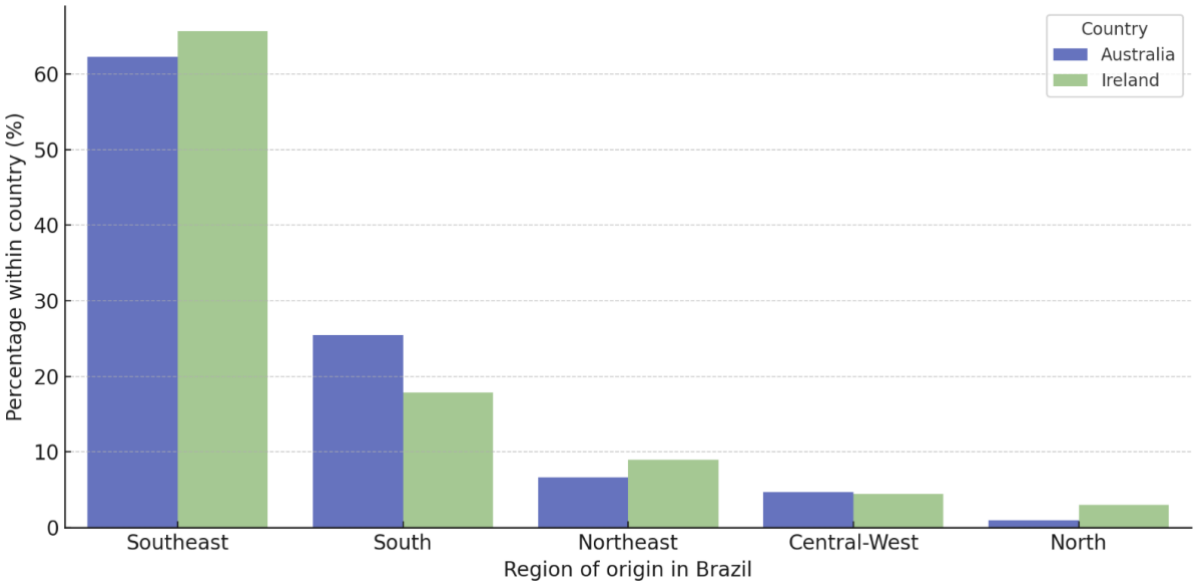

Figure 14 – Region of origin in Brazil among Brazilian migrants to Australia and Ireland. Percentages represent the distribution within each destination country.

- Southern Brazil sends a larger proportion of its migrants to Ireland, indicating that this region may perceive Ireland as a more accessible or suitable destination compared to Australia;
- Southeast Brazil distributes heavily to both countries, but its stronger representation in Australia suggests a more qualified or financially prepared migrant profile;
- Northeast Brazil has a relatively higher share of migrants choosing Ireland, possibly due to lower entry barriers and established support networks;
- For Central-West and North, while the absolute numbers are smaller, Ireland and Australia remain among the top destinations, showing their relevance across all regions.

These regional differences likely reflect a combination of economic preparedness, historical migration patterns, language familiarity, and the influence of diaspora networks that facilitate access to information and shape perceptions of destination accessibility.

Also, a clear distinction emerged between the two destinations in terms of *relationship status* (Figure 15):

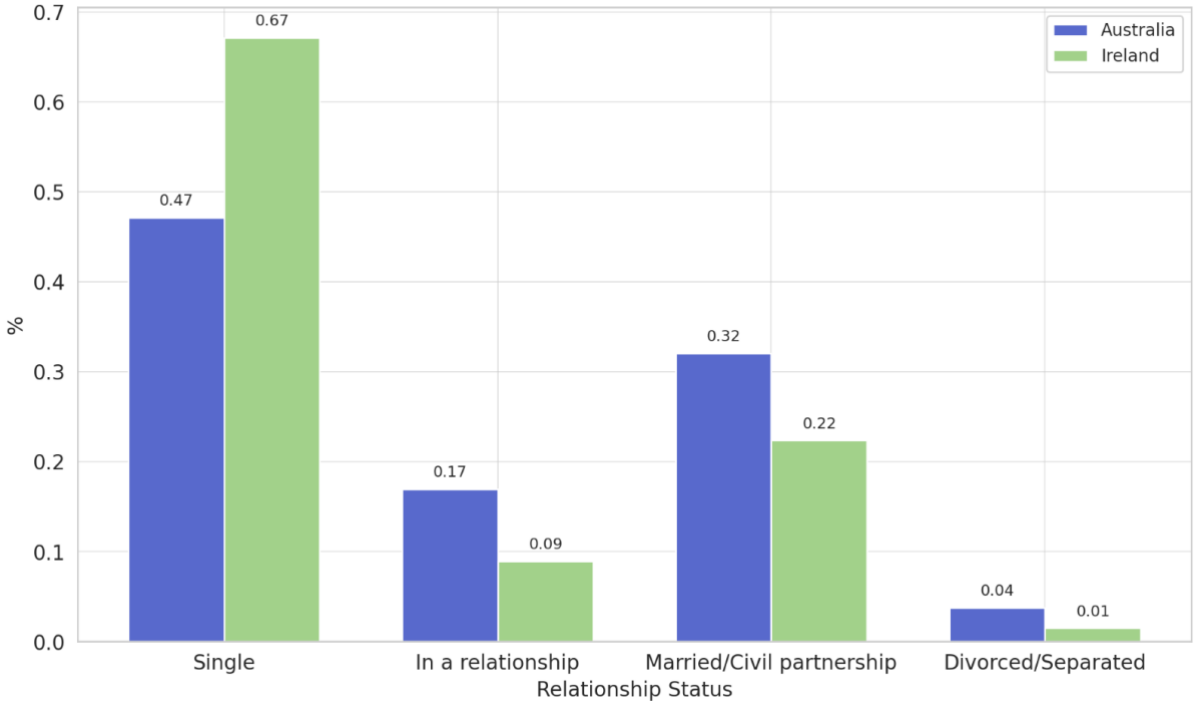

Figure 15 – Relationship status of Brazilian migrants prior to migration to Australia and Ireland. Percentages indicate the distribution within each destination country.

- *Higher proportion of single migrants in Ireland:* Over two-thirds (68.2%) of respondents who migrated to Ireland reported being single, compared to 44.2% in Australia. This suggests that Ireland may be perceived as a more accessible or appealing destination for younger individuals or those without dependants or family obligations.

- *Greater presence of couples and families in Australia:* In contrast, Australia had a significantly higher share of married respondents or those in stable relationships (30.2% vs. 22.7% in Ireland), as well as a larger proportion of individuals in dating relationships (15.9% vs. 9.1%). This pattern may reflect Australia's more structured visa pathways for skilled migrants and family reunification, alongside a greater need for financial planning and long-term commitment prior to migration—consistent with other findings related to higher income and the use of professional planning services among Australian-bound migrants.

- *Lower representation of divorced or separated migrants overall:* Although a small segment, this group was slightly more represented in Australia (4.7%) than in Ireland (1.5%).

These findings provide valuable insights for institutions designing tailored services. Brazil-to-Ireland migrants are more likely to be independent, younger individuals pursuing temporary or exploratory experiences. Conversely, the profile of migrants to Australia is more aligned

with structured, long-term migration involving couples or families with greater financial and professional readiness.

This distinction in migrant profiles is further supported by differences in how long individuals had been living in their host countries. The boxplot in Figure 16 presents the distribution of years since migration among respondents in Australia and Ireland.

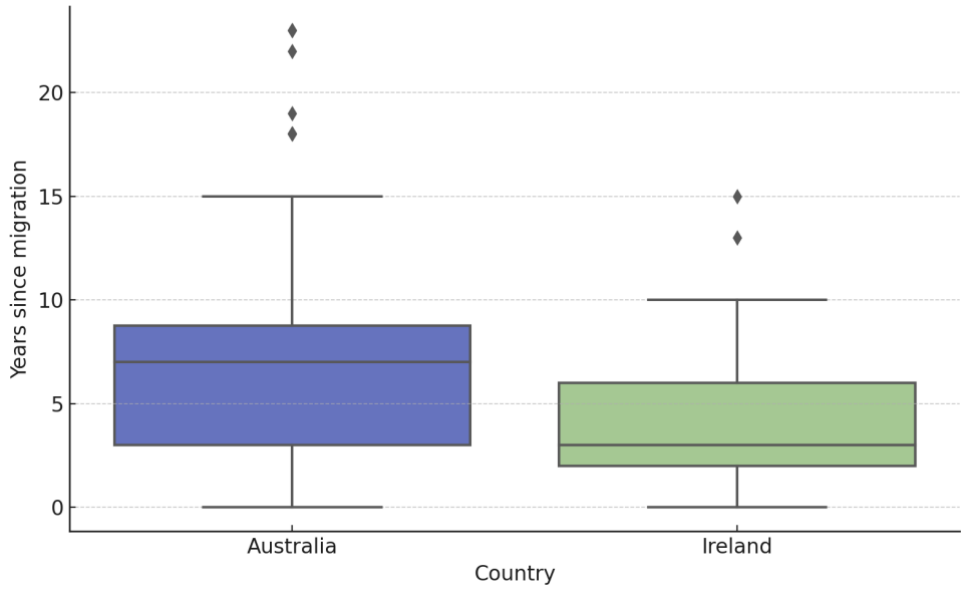

Figure 16 – Years since migration among Brazilian migrants in Australia and Ireland. The boxplot shows the distribution, median, and outliers within each destination country.

Brazilian respondents in Australia had typically been living there for longer periods, with a median between 5 and 10 years. This reflects a more established migration pattern, suggesting that Australia has long served as a destination for skilled or long-term migration. The respondents in Ireland had more recent migration experiences, with a median of just 1 to 2 years. This indicates that Ireland is an emerging destination, likely driven by newer visa opportunities, study programmes, and accessible entry requirements.

The wider range observed in the Australian data points to a mature and diversified migrant population, whereas the clustering of shorter durations in Ireland suggests a wave of recent entrants who may still be in the process of adaptation or transition. These patterns reinforce the interpretation that Australia predominantly attracts long-term settlers, while Ireland currently appeals to individuals seeking temporary or exploratory migration experiences (e.g., study, working holiday, or early-career transitions).

Taken together, the socio-demographic profiles of Brazilian migrants to Australia and Ireland reveal both convergence and divergence across key variables. Although no single

characteristic clearly distinguishes one group from the other, the subtle trends outlined here provide important context for the behavioural and motivational dimensions examined in the following chapter.

4.1.2 MOTIVATIONAL FACTORS FOR MIGRATION

To deepen the understanding of the migration profiles presented earlier, this section explores the underlying motivations that influenced respondents' decision to migrate. Participants were asked to rate the importance of a series of potential motives on a five-point Likert scale, ranging from 1 (not important) to 5 (extremely important). The bar chart on the Figure 17 displays the average score for each motivational factor, comparing Brazilian migrants living in Australia and Ireland.

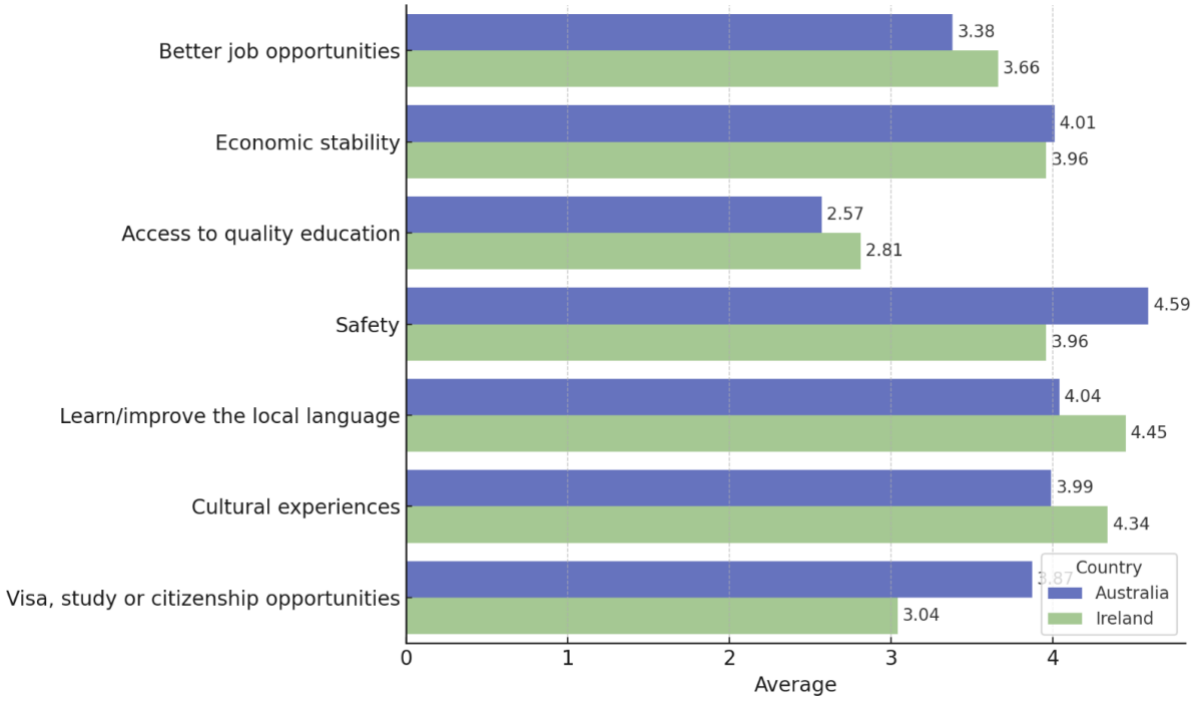

Figure 17 – Average importance of migration motivations among Brazilian migrants in Australia and Ireland.

In general, both groups rated safety, language acquisition, and cultural experiences as highly important motivations for migration, while access to quality education consistently received the lowest average score. Key observations include:

- Safety scored highest among migrants in Australia (M = 4.59), suggesting a greater weight placed on this aspect when choosing that destination;
- Migrants to Ireland showed a stronger inclination toward language learning (M = 4.45) and cultural experiences (M = 4.34) compared to their Australian counterparts;

- Economic stability and visa/citizenship opportunities remained significant for both groups, with a slightly stronger emphasis among those in Australia;
- Job opportunities, while often cited as a general driver of migration, received relatively moderate scores from both cohorts.

In fact, when questioned about *how much the possibilities of future opportunities in the destination country influenced the decision-making process*, approximately 80% of respondents in both Australia and Ireland reported that this factor was either *influential* or *very influential* (Figure 18).

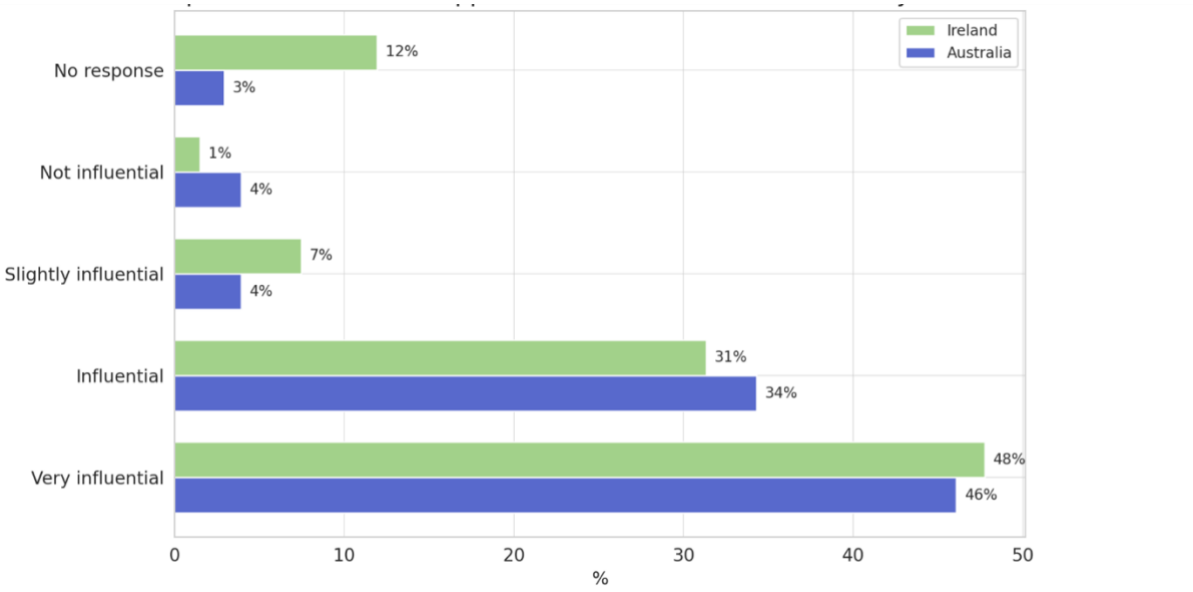

Figure 18 – Perceived influence of future opportunities in the destination country on migration decision-making among Brazilian migrants in Australia and Ireland.

This finding reinforces the notion that Brazilian migrants are strongly forward-looking, particularly motivated by the prospect of career advancement and financial stability. It also supports the interpretation that migration is not merely reactive to current conditions but is often driven by expectations and aspirations related to life improvement abroad.

Participants were also asked to indicate which sources of information they consulted while planning their migration. As shown in Figure 19, the responses reflect distinct strategies between migrants to Australia and Ireland. Many participants have indicated the influence of online sources such as digital influencers, WhatsApp groups, and blogs, reinforcing the increasing role of digital platforms in shaping migration decisions.

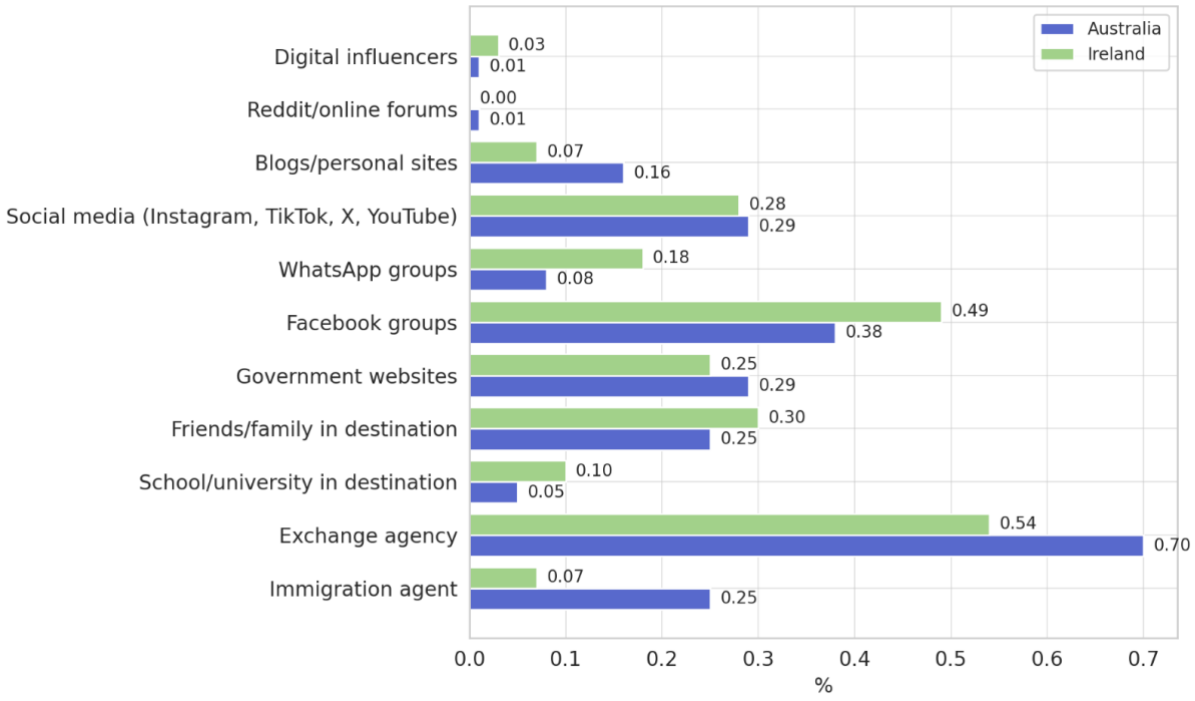

Figure 19 – Sources of information used by Brazilian migrants to prepare for migration to Australia and Ireland.

In Australia, there was a greater reliance on formal and structured channels, such as exchange agencies (70%), immigration agents (25%), and government websites (29%). This profile suggests that migrants heading to Australia tend to seek more technical planning and specialised support from the beginning of their journey. For the market, this reflects a clear demand for more comprehensive and professional services that offer expert consultancy, document preparation, and support with skilled visa applications.

In contrast, Brazilian migrants heading to Ireland reported stronger reliance on informal and community-based sources, such as Facebook groups (49%), WhatsApp groups (18%), and friends or family already living in Ireland (30%). This highlights the role of peer networks and social guidance in shaping migration pathways to Ireland. For service providers, this suggests a greater opportunity to engage through social proof, authentic testimonials, and peer-driven content strategies, including community ambassadors or micro-influencers within migrant groups.

Despite the growing use of social media as an information source in both destinations (29% for Australia / 28% for Ireland), the use of digital influencers and platforms such as Reddit remained limited, suggesting an untapped potential for organisations to build trust and visibility among early-stage decision-makers through diversified and platform-specific digital marketing.

Notably, educational institutions were among the least-used sources (5% for Australia, 10% for Ireland), suggesting that most migrants did not receive direct support or outreach from schools or universities, even when their goals involved formal study.

Overall, these patterns reflect how information-seeking behaviours are shaped not only by individual characteristics, but also by the structural conditions of each destination, including the maturity of local migrant communities, the availability of formal planning tools, and the degree of institutional complexity.

For exchange agencies, universities, migration consultancies, and other private service providers supporting Brazilian migrants, these findings offer valuable strategic insights. In Australia, positioning as a technical and reliable authority may be more effective, while in Ireland, investment in community-building, social validation, and peer-shared experiences appears more aligned with the expectations and behaviours of prospective migrants.

Moreover, these results contribute to the literature by reinforcing that the search for information is not solely a reflection of individual motivations but is also strongly influenced by the structural characteristics and accessibility of the destination country. In this light, understanding how migrants access information provides a foundation for interpreting how they prepare for the journey ahead.

To further contextualise these decision-making processes, Figure 20 presents the main barriers faced during the planning phase. These data help illustrate how initial motivations are often constrained or reshaped by practical limitations, particularly those related to financial and institutional access.

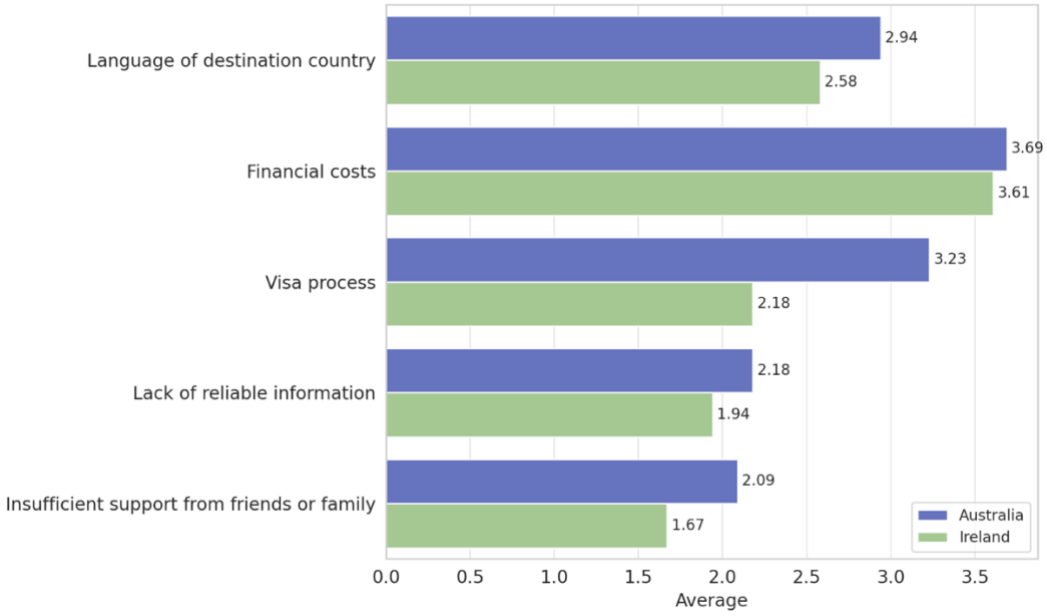

Figure 20 – Main barriers faced by Brazilian migrants during the planning phase of their migration to Australia and Ireland. Values represent the average level of difficulty rated by respondents for each factor.

The most prominent barrier across both destinations was financial cost (M = 3.69 Australia; M = 3.61 Ireland), reinforcing its dual role as both a motivating factor (linked to the search for economic stability) and a limiting one. Migrants were not only driven by the promise of better financial prospects but also hindered by the high cost of reaching them.

Visa-related difficulties were significantly more prominent among migrants to Australia (M = 3.23) than those in Ireland (M = 2.18), reflecting the complexity of Australia's immigration system. This finding supports earlier observations that Australian-bound migrants more frequently rely on formal support structures, such as migration agents and agencies, to navigate visa requirements.

Language barriers were also rated higher among Australian migrants (M = 2.94) than their Irish (M = 2.58), possibly due to stricter English proficiency expectations or more diverse communication demands across Australian states and institutions.

Interestingly, lack of reliable information and limited support from friends or family were less prominent overall, but still more frequently cited by respondents migrating to Australia. This pattern is consistent with previous findings suggesting a more technical, individualised planning approach in Australia, in contrast with Ireland's reliance on community-based support and peer networks.

While motivations for migration tend to focus on long-term goals such as career, safety, or education, the chart above revealed that the path to achieving those goals is shaped by the institutional context of the destination. Migrants to Australia face more structural planning barriers, which likely influence their higher use of professional services. In contrast, those migrating to Ireland navigate fewer formal challenges but tend to rely more on informal preparation and personal networks.

The Figure 21 presents the main barriers to permanence faced after arrival, illustrating how post-migration challenges are experienced differently across the two destinations. Percentages indicate the share of respondents who selected each factor.

The high cost of living stands out as the most significant barrier in both destinations, cited by 50% of respondents in Australia and an even higher 66% in Ireland. This finding echoes previous results from the planning phase, where cost was already identified as a major concern, particularly among those moving to Ireland.

The lack of long-term visa options appears prominently in Australia (39%) and, to a lesser extent, in Ireland (27%). This suggests a potential mismatch between expectations and reality of immigration systems, especially in Australia, where visa pathways are known to be complex and selective. Interestingly, this contrasts with earlier motivation findings, where respondents considering Australia expressed more long-term intentions; yet, structural rigidity appears to limit those aspirations.

Cultural and linguistic barriers are also notably higher in Australia (19% and 23%, respectively), compared to those in Ireland (9% and 6%). While both groups had previously rated “learning or improving the local language” as a key motivation, these figures suggest that the actual linguistic and cultural adaptation process may prove more challenging than initially expected, particularly in Australia.

A unique finding for Ireland is the mention of climatic barriers (5%), a factor not reported by respondents in Australia. This suggests that environmental conditions may also influence the experience of permanence.

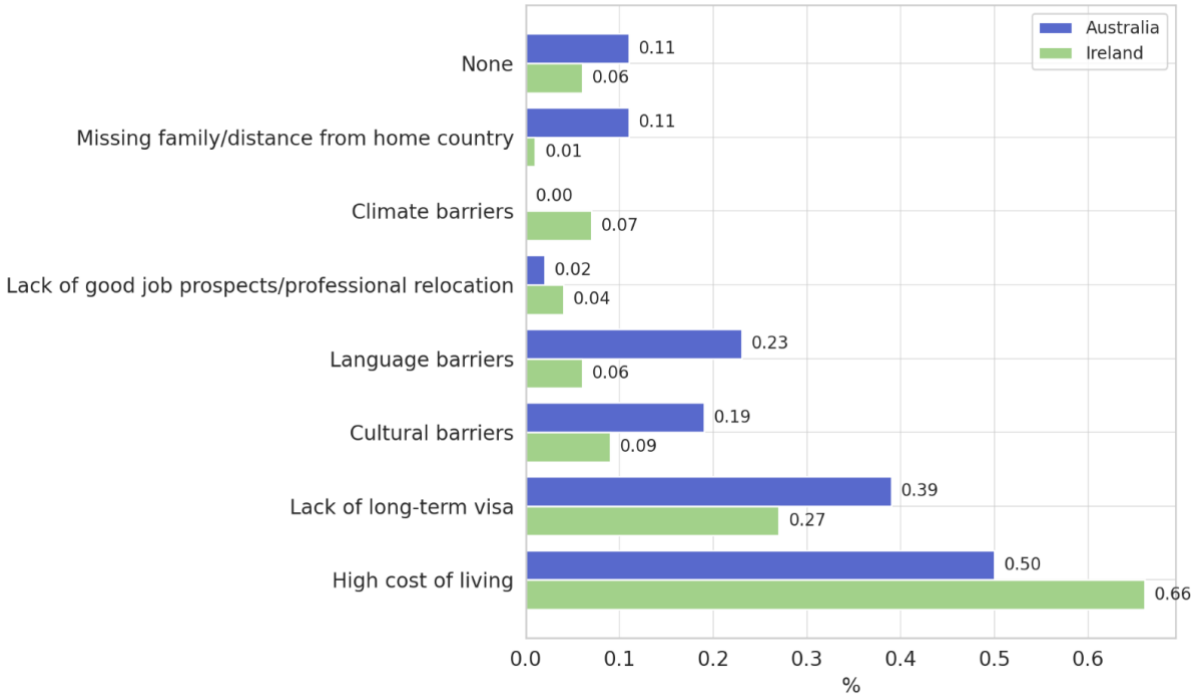

Figure 21 – Main challenges faced by Brazilian migrants when trying to remain in Australia and Ireland.

Interestingly, barriers such as lack of professional opportunities and distance from family were reported less frequently. This may help balance the equation in favour of staying, especially considering that the push and pull factors most cited as motivations, such as financial stability and career opportunities, seem to be partially met. Likewise, homesickness does not emerge as a decisive factor, reducing the emotional pressure to return to the country of origin.

These findings are consistent with migrants stated future intentions. As illustrated in Figure 22, a significant portion of respondents, especially those in Australia, do not intend to return to Brazil. Many have already established their lives abroad or are awaiting permanent residency before making long-term decisions. In Ireland, there is a greater level of uncertainty, with more respondents expressing indecision or the desire to return at an undefined time.

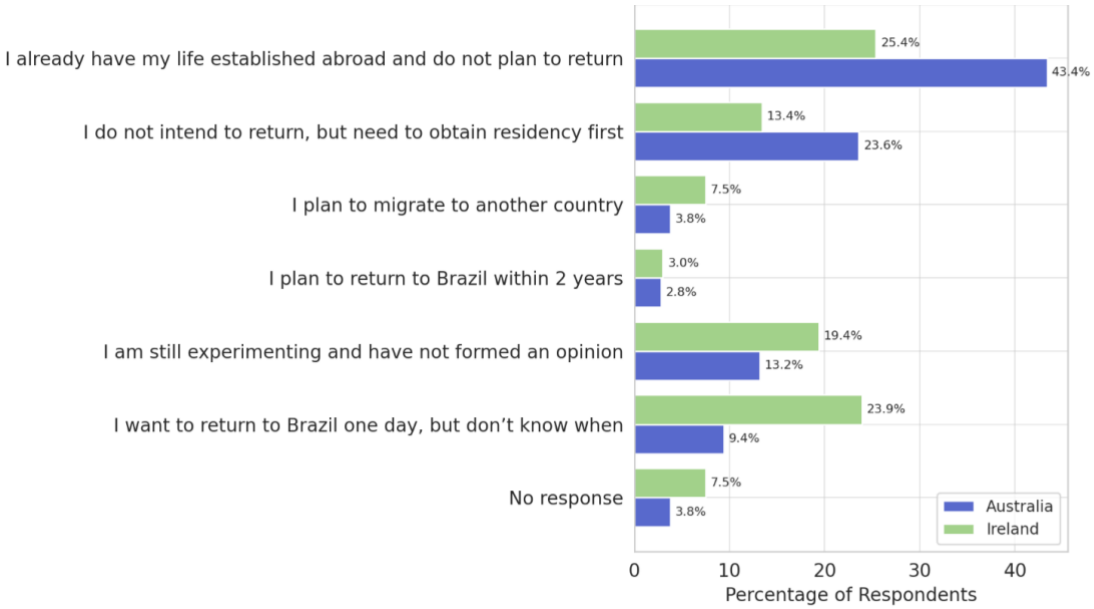

Figure 22 – Future intentions of Brazilian migrants living in Australia and Ireland.

Despite the presence of structural or emotional challenges, such barriers are not strong enough to outweigh the perceived benefits of remaining. On the contrary, many migrants appear to be navigating these difficulties with a forward-looking mindset, reinforcing the continued relevance of their initial motivations, particularly around career and life stability. This observation supports the broader argument that migration decisions reflect a dynamic and cumulative process, rather than a single moment of choice.

Finally, the descriptive findings confirm that Brazilian migrants heading to Ireland and Australia are demographically aligned in many ways. Although there are slight differences in what migrants to Australia and Ireland consider most motivating and challenging, a purely comparative analysis of these averages does not offer a conclusive explanation for country choice. For instance, migrants to Australia may already have a longer-term plan in mind, while those moving to Ireland might be more focused on short-term study or cultural experiences. However, such intentions are often reinforce, or even shaped, by the immigration policies and visa structures of each destination, which either enable or limit certain pathways. This supports the argument that destination choice in this context is influenced less by socio-demographic factors and more by perceived opportunities, personal motivations, and behavioural dynamics.

These insights validate the modelling approach adopted in this thesis: if traditional predictors do not provide sufficient differentiation, then a predictive model must rely on less obvious patterns, those embedded in perceptions and qualitative responses. In this light, the reliance on qualitative variables and PCA-derived features should not be viewed as a limitation but rather as a necessary response to the complexity of real-world migration decisions. This

chapter, therefore, establishes the empirical foundation for the predictive analysis that follows.

When considered alongside the broader literature and the results of the PCA and Random Forest model, this analysis reinforces the view that no single feature explains migration outcomes in isolation. The strength of a model such as Random Forest, especially when combined with PCA for feature selection, lies in its capacity to capture the non-linear and multifactorial nature of these decisions. Ultimately, it is the interaction of multiple elements, rather than any one variable, that drives the final outcome.

4.2 Feature selection results

To reduce dimensionality and improve model interpretability, Principal Component Analysis (PCA) was applied to the filtered dataset containing only Brazilian migrants to Australia and Ireland. Unlike conventional PCA applications, which aim to explain the overall variance in a dataset, this study used PCA to support a supervised learning task: identifying features most associated with the binary classification of destination country.

After applying PCA with Z-score normalisation, we identified a principal component most aligned with the "Country" variable (Figure 23). This component captured the greatest discriminative variance between the two destinations, rather than simply maximising total variance across all respondents.

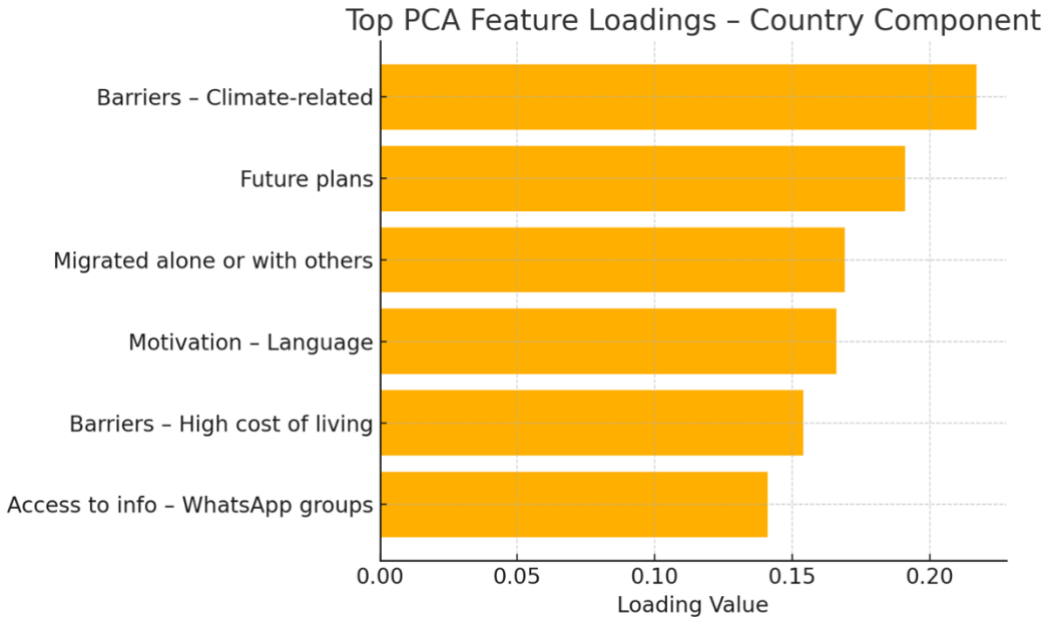

Figure 23 – Top feature loadings for the principal component most associated with the destination country variable, based on PCA with Z-score normalisation.

The six variables with the highest absolute loading values on this component were:

1. **Barriers to staying – Climate-related barriers (loading = 0.217)**
2. **Future plans (loading = 0.191)**
3. **Migrated alone or with others (loading = 0.169)**
4. **Motivation – To learn/improve the local language (loading = 0.166)**
5. **Barriers to staying – High cost of living (loading = 0.154)**
6. **Access to information – WhatsApp groups (loading = 0.141)**

To enrich the interpretability and maintain the contextual relevance of the model, two additional socio-demographic variables were manually reintroduced, as they were underrepresented in the PCA-derived rankings:

7. **Level of education (loading = 0.065)**
8. **How long ago they migrated (loading = -0.29)**

This final set of eight predictors, combined with the target variable "Country", constituted the formed the basis for the supervised learning models. The selected variables offer a balance between subjective perceptions (e.g., motivations and barriers) and structural indicators (e.g., education and migration timing), reflecting both personal decision-making and measurable attributes.

The inclusion of these features improved model performance relative to earlier tests using either the full feature set or raw PCA output alone. Notably, this refined subset enhanced both overall accuracy and stability, particularly in the minority class (Ireland), where many models previously underperformed.

4.3 MODEL PERFORMANCE COMPARISON

This section presents the outcomes of the model training experiments conducted on the selected dataset. The goal was to evaluate predictive performance across different modelling approaches and to assess the impact of dimensionality reduction via PCA.

Five different models were tested: Logistic Regression, K-Nearest Neighbours (KNN), Decision Trees, Neural Networks, and Random Forest. The aim of testing multiple algorithms was motivated by the desire to assess contrasting approaches from interpretable linear models, such as logistic regression, to more complex models capable of capturing non-linear relationships, such as neural networks.

All models were first evaluated using the full set of 55 variables, and then re-evaluated using a reduced set of eight features selected through PCA. To determine whether a model's performance was meaningful, we compared its accuracy against a baseline defined by the most frequent class. In this case, a model needed to outperform the accuracy of simply

predicting the majority class (i.e., always predicting “Australia”) to be considered valid within the scope of our study.

Since the final training dataset was divided between two destination countries — Ireland (39%) and Australia (61%) — we considered a model relevant only if its conditional accuracy for each class exceeded the class’s proportion in the dataset. For instance, a model achieving 80% overall accuracy but only 35% accuracy for Ireland would fall below the 39% baseline and thus be deemed inadequate. Our preference was on models that exhibited good accuracy at both of our key target classes, maximising practical value and avoiding predictive bias due to class imbalance and small sample size.

Initially, Neural Networks were considered because of their potential for modelling complex, non-linear patterns may be helpful for capturing the nuanced motivations behind individual migration choices. We played with simple feed-forward neural-network architectures, but due to our modest sample size there was not enough data for the neural networks to identify stable and consistent patterns of behaviour. The inherent instability of the model led it to a great deal of convergence problems, with performance no better than random guessing. This level of performance was somewhat to be expected, as neural networks generally require a larger dataset to avoid overfitting as well as generate valid predictions.

We then moved on to using the k-Nearest Neighbours (KNN) algorithm because of its simplicity, ease of interpretation, and relatively consistent performance on smaller datasets in our experience. The KNN method classifies unknown samples based upon the majority class of their closest neighbours, and is simple to implement and quickly execute. While KNN returned reasonable overall accuracy compared with the naive baseline, it was significantly poor with our imbalanced class distribution. Specifically, the accuracy of the minority class of Ireland peaked just short of the 40% threshold and only gave marginal increase in performance with our naive baseline of 39%, indicating it has little predictive power to effectively differentiate Irish migrants in respect to our Australian sample.

Next, we used Decision Trees because of their ease of interpretability and ease of implementation. The decision tree classifies unknown samples by sequentially splitting features into branches producing transparent insight into all the factors leading to the sample's migration choice. Despite these merits, decision trees achieved high training accuracy but overfitted, as shown by poorer performance on validation folds. PCA-based selection proved unstable in cross-validation and was deemed unsuccessful. Mutual information feature selection marginally pruned complexity but resulted in substantially lower k-fold validation scores compared to test, with Australian performance dropping below the 65% baseline, indicating poor generalisation. Manual feature sets were inconsistent and thus discarded. This ability to learn did not translate into predictive accuracy on new, unseen data.

We also considered Logistic Regression due to the idea of simplicity, ease of interpretation, and inherent use as a baseline for other models for binary classifiers. Logistic regression seeks

to find linear associations between predictors with probability of the outcome class. Logistic results are easy to interpret. However, decisions about migration classes likely follows complex, nonlinear associations of multiple factors: financial resources, cultural openness, and familial connections. As a result, the accuracy produced from logistic regression was very limited, and neither the baseline nor any feature-selection technique (PCA, Mutual Information or manual) delivered acceptable predictive accuracy. All configurations failed to converge on reliable multiclass discrimination, indicating limited linear separability in the data.

Finally, we settled on using the Random Forest algorithm. We chose Random Forest because it is robust to noisy, imbalanced data, and because it is able to manage complex nonlinear interactions among predictors of the outcome class. Although the random forest produces an aggregate of predictions based on multiple decision trees, each trained on slightly distinct subsets of data and features, the aggregation would take care of overfitting, class imbalance and predictor correlations. Our investigations established that Random Forest provided the best overall performance for balancing bias and variance in prediction levels, offering the best representation of all models we had considered.

Due to previous models returning similar or worse performance levels to our overall naive model, Random Forest was selected for a deeper inquiry into intended applications, as contained in the following sections.

The Random Forest classifier consistently delivered the best balance between sensitivity to both classes and overall performance. Notably, using PCA-reduced features, not only it preserved accuracy but also improved the model's ability to correctly classify migrants. It raised the accuracy for the majority class, Australia, from 61% to 81%, and for the minority class, Ireland, from 39% to 60%, significantly above the baseline.

While some models performed well in terms of overall accuracy, their inability to improve prediction for the minority class (Ireland) highlighted the challenge of imbalanced data. For example, Logistic Regression and Neural Networks hovered near the baseline for Ireland, despite respectable overall results. In contrast, Random Forest demonstrated a consistent advantage, leveraging ensemble learning to generalise well across both classes. By integrating decision boundaries from multiple trees, it avoided the pitfalls of overfitting seen in single-tree models and better managed feature correlations identified during PCA.

4.4 FINAL MODEL: RANDOM FOREST

While simpler models and single decision trees struggled with the minority group, Random Forest gave more reliable predictions for both Australia and Ireland.

Random Forest works by building many decision trees on random slices of the data and then combining their votes for each prediction. This gives it several benefits: it easily handles mixed data types (yes/no, ordered categories and numbers), finds complex patterns automatically, reduces overfitting through averaging, and still works well even when we do not have a large dataset.

To keep things consistent and not too complicated, we set up each forest with 100 trees, each up to six levels deep, a minimum of five samples per leaf, and the Gini index split rule. We also fixed the random seed so our results could be reproduced exactly. With this setup, we evaluated 5 versions of the Random Forest model:

- **All features (baseline):** Confirmed that the forest met our basic accuracy targets, but exhibited considerable fold-to-fold variance.
- **Mutual Information selection:** Delivered 77% overall accuracy on the test split (81% Australia, 70% Ireland) but dropped to 65% in k-fold validation, and 65% for Australia, showing insufficient stability.
- **Manual selection:** Did not improve beyond the training data performance.
- **PCA with summed loadings:** Matched the baseline in both test and cross-validation, offering no real gain.
- **PCA ranked by "Country":** By isolating the eight variables that contributed most strongly to the "Country" principal component, we achieved 73% overall accuracy — 81% for Australia and 60% for Ireland — consistently on both the test and validation score.

In summary, the Random Forest using country-ranked PCA features, balanced bias and variance, most effectively, improved detection of the smaller class, and was the only version to stay consistently strong across all folds.

By comparing these two configurations, we observed that PCA-based dimensionality reduction led to the following improvements:

- **Noise reduction**: By selecting those variables that best capture the variance relevant to our binary classification task (Australia vs. Ireland), we reduced the complexity and redundancy of the model.
- **Efficiency**: As a result, the model became less complex and faster to train, while also improving Ireland-class accuracy by 9 percentage points.

To better understand model performance, we also analysed the confusion matrix on the Figure 24 from the final Random Forest model trained on PCA selected features:

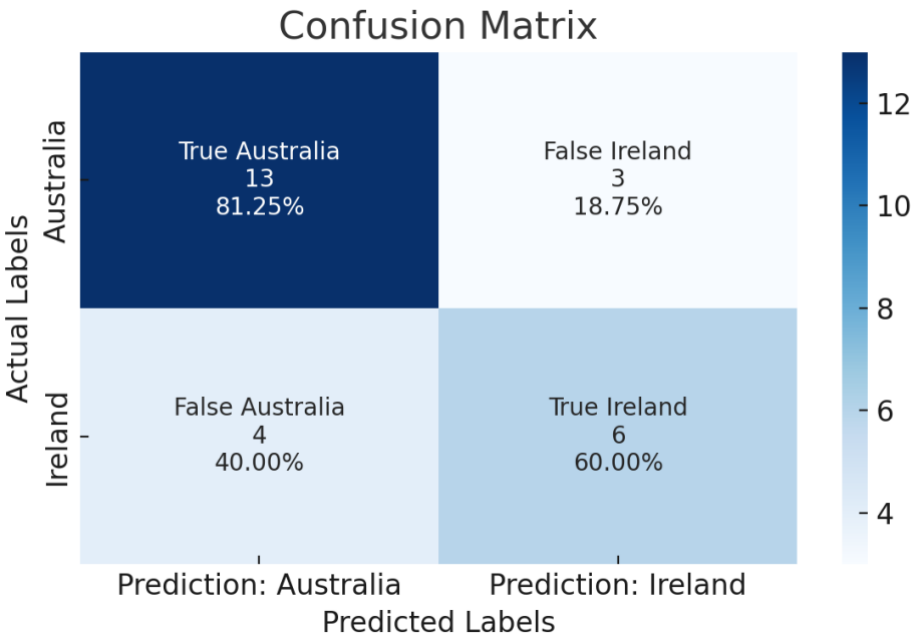

Figure 24 – Confusion Matrix for the final Random Forest classification model predicting the migration destination (Australia or Ireland).

The model achieved 73.08% of overall accuracy, with 81% for Australia and 60% for Ireland, 26.92% of Error Margin and a Cohen's kappa of 0.42%, indicating moderate agreement beyond chance. While some misclassifications remain, especially predicting Australia for true Ireland migrants, the model shows strong discriminatory power, especially given the relatively small and imbalanced dataset. Mean values across classes were also strong: recall and sensitivity both averaged 0.706, precision reached 0.716, and the F1-score stood at 0.710, showing a good balance between precision and recall.

Cross-validation results further confirm that the model maintained solid predictive performance across different subsets of data, underlining its robustness. While classification for Ireland remains more challenging than for Australia, the model still achieved 60% true positives for Ireland (well above the baseline of 39%), representing a 54% relative improvement in prediction accuracy. Precision reached 0.667 and the F1-score was 0.632. Notably, specificity was high at 0.812, meaning the model was more effective in identifying non-Irish cases than Irish ones. These metrics are acceptable considering the smaller number of Irish cases in the dataset and the associated prediction challenges.

For Australia, performance was even stronger, with 81% of true positives compared to a baseline of 61%, resulting in a 33% improvement. Precision was also high at 0.765, and the F1-score reached 0.788, reflecting a strong overall performance in distinguishing Australian respondents. Specificity, however, was slightly lower at 0.600, which means that some non-

Australian cases were mistakenly classified as Australian. These results demonstrate the model's ability to identify relevant patterns and enhance prediction accuracy, even within a relatively small and imbalanced dataset.

4.4.1 Comparative analysis of Random Forest and traditional modelling approaches in migration research

In migration research, a wide range of methodological approaches has been used to understand individual decisions, patterns, and trends. These approaches can generally be divided into two categories: supervised and unsupervised learning models. Supervised models are those that rely on a labelled outcome, typically used for prediction or classification, whereas unsupervised models explore the structure of the data itself, identifying clusters or patterns without a predefined target variable. Most traditional migration studies, however, have relied on explanatory frameworks, particularly Structural Equation Modelling (SEM) and its more flexible variant, Partial Least Squares SEM (PLS-SEM).

PLS-SEM has proven useful in exploratory research aimed at validating theoretical constructs, particularly in studies involving international students. Researchers have applied it to examine how latent factors such as academic prestige, financial cost, and perceived behavioural influence the decision to migrate (Nikou & Luukkonen, 2024; Nghia, 2019). However, despite its strengths in estimating linear relationships, PLS-SEM is limited when applied to more dynamic, non-linear social phenomena like migration. It also requires a predefined structural model, making it less suitable for situations where the relationships between variables are not yet clearly understood or may evolve with context.

More recent migration research, particularly involving forecasting asylum flows or refugee movements, has begun to adopt machine learning models that focus on prediction rather than explanation. One example demonstrated the use of Random Forest to forecast irregular border crossings and asylum requests in Europe with considerable success (Bosco et al., 2024).

While much of the existing literature remains grounded in theoretical validation, either through PLS-SEM, Bayesian modelling, or even institutional simulation-based tools (Angenendt & Koch, 2024), our approach embraces the complexity of migration decisions through a supervised machine learning lens, prioritising accurate prediction over confirmatory explanation. Moreover, we apply this to a voluntary migration context, a less commonly studied area compared to forced migration and asylum flows.

Another distinguishing factor is the type of data employed. Instead of relying on large administrative datasets or macroeconomic indicators, we built an original behavioural dataset through an online survey targeting Brazilians in the process of migration planning. This allowed us to include individual-level variables such as perceptions, motivations, digital behaviours, and country preferences, factors often left out of structural or official data. In

doing so, we aligned our modelling choice (Random Forest) with both the behavioural complexity of the phenomenon and the applied decision-oriented nature of our research goal.

This positions our work at the intersection of macro-level modelling of flows and micro-level behavioural studies. While forecasting studies (Bosco et al., 2024; Wiśniowski et al., 2013) have delivered strong results for aggregate-level predictions, and micro-level studies (Gatfield & Chen, 2006; Nghia, 2019) have clarified behavioural drivers without attempting prediction, few studies have bridged both. Our contribution lies in applying supervised learning to individual-level behavioural data, offering predictive insights typically reserved for macro-level models.

Unlike most existing studies, which either focus on explaining individual migration intentions or forecasting large-scale flows, our research uses supervised machine learning to anticipate destination choices in a voluntary migration context. By working with individual-level behavioural and perceptual data, our model offers a new perspective that complements macro-level approaches with fine-grained, predictive insights that can inform more targeted decision-making

Our study takes a distinct path. First, we focus on voluntary migration, specifically among Brazilian individuals considering relocation for education, work, or lifestyle reasons. This population is underrepresented in predictive research, which still tends to prioritise refugee and asylum contexts. Second, we move away from institutional and macroeconomic datasets by constructing an original dataset composed of behavioural and perceptual data, gathered through online surveys in Brazilian migrant communities. This allows us to model micro-level decision-making, grounded in real-time, with self-reported motivations.

In summary, while traditional models explain *why* people migrate, we explore *who* is likely to migrate, under *which* conditions, and to *which* destination. This predictive shift bridges theoretical insight and practical application, contributing a novel perspective to migration forecasting.

4.4.2 THE SHIFT TO RANDOM FOREST: WHERE OUR RESEARCH STANDS OUT

Random Forest was selected for this study not only for its predictive performance but also for its alignment with the structure and nature of our data. Unlike PLS-SEM, which depends on pre-established theoretical linear relationships and struggles with complex, interactive nonlinear effects, Random Forest is a supervised machine learning algorithm that handles non-linearity, multicollinearity, and missing data with relative ease. It constructs multiple decision trees using bootstrapped samples and random subsets of variables, then it aggregates their outputs, making it well-suited for identifying subtle patterns in large, diverse datasets.

The decision to adopt Random Forest was based on its strong performance in handling heterogeneous, non-linear, and partially incomplete datasets, which are typical of survey data capturing individual perceptions and motivations. Similar applications have used Random Forest to forecast asylum requests and irregular border crossings in Europe with over 80% explained variance (Bosco et al., 2024). In our case, the algorithm proved especially useful for capturing complex interdependencies among economic, social, and demographic variables. In contrast with PLS-SEM, which assumes linear relationships and requires pre-specified model structures, Random Forest identifies patterns and interactions autonomously, offering a model-free framework that aligns more closely with the exploratory and practical goals of our research.

This methodological shift matters. Unlike Bayesian models, which are ideal for probabilistic inference under well-defined structures, or agent-based simulations (Angenendt & Koch, 2024), which demand high computational effort and institutional implementation, Random Forest enables direct, accessible, and easy forecasting using modest technical resources. It bridges the gap between academic rigour and operational application, making it a more agile tool for emerging needs in the migration field.

Our model, therefore, is not only academically relevant but also market oriented. Designed for institutions such as universities, education agencies, and migration consultancies, it provides actionable insights into the profiles most likely to migrate. These insights can inform recruitment, communication strategies, and programme design tailored to the expectations of prospective migrants.

This research reframes migration forecasting by combining interpretability, flexibility, and predictive strength. While acknowledging the value of explanatory frameworks, we advocate for a shift toward prediction, offering a method that is robust, transparent, and adaptable, grounded in behavioural data and aligned with both academic and applied goals.

4.5 Reflections on false positives and false negatives

An important dimension of our model evaluation involves the interpretation of the false positives and false negatives revealed in the confusion matrix. While these misclassifications might initially be seen as shortcomings in model accuracy, they offer insight into a relevant subset of respondents: those situated in a "zone of indecision."

These individuals — those whom the model incorrectly classified — appear to deviate from consistent patterns across the selected features. This may suggest a weaker or more ambiguous migration intention at the time of decision-making, and probably, would have been more susceptible to persuasive action or communication, particularly the marketing and outreach of educational entities, immigration assistance entities, or government programmes.

In contrast, individuals correctly classified with high prediction confidence likely had stronger internal preferences or more established migration plans. These individuals are typically less influenced by messaging around *where* to migrate and more focused on finding support services tailored to a destination they have already chosen.

The “indecisive” group, reflected in the model’s misclassifications, holds significant strategic value. These individuals may be more open to persuasion and responsive to targeted communication or support, particularly in competitive recruitment contexts. As such, the model offers practical value by helping providers identify and engage individuals who are still in the decision-making phase of migration—offering a strategic advantage in international recruitment.

5. CONCLUSIONS AND FUTURE RESEARCH

5.1 WHAT THE RESULTS REVEAL ABOUT BRAZILIAN MIGRANT BEHAVIOUR AND HOW INSTITUTIONS CAN USE THESE INSIGHTS

The predictive analysis of this study indicates broad patterns of structural, motivational, and information-based influences on Brazilian migration choices regarding Australia and Ireland. The most notable predictors of decisions were innate pre-migration status, digital influencers and blogs, and timeframes for making decisions, as revealed by PCA and confirmed through the Random Forest model. This incorporation of technology in migration behaviours aligns with the growing body of literature addressing the relevance of online sources, and peer influence (Dekker & Engbersen, 2014; Komito, 2011).

Key motivations for migrating to Ireland included visa accessibility, safety, and affordability. This shows that institutional and policy-related variables can inform and frame destination choices. However Australia-bound migrants appeared to be indirectly influenced by factors such as long-term employment opportunities, lifestyle, and digital information ecosystems which reflect their ability to visualize and place emphasis on longer term strategic planning and exposure to larger digital narratives.

Additionally, the Random Forest model performed relatively well, especially for correctly classifying destination responses for respondents bound for Ireland (72% accuracy for PCA derived features). Even within a modest sample size, it was still possible to capture behaviourally based patterns that are not surface preferences. In particular, capturing behavioural patterns is useful for identifying latent signals that can help distinguish decisional profiles.

This section discusses the practical and theoretical implications of the study's findings, focusing on three dimensions: (1) what the results suggest about Brazilian migrant behaviour toward Australia and Ireland, (2) how institutions can apply these insights, and (3) the limitations encountered during model development and interpretation.

The insights produced from this research have practical and usable implications for educational institutions, exchange agencies, immigration consultants, and policymakers with experience in engaging with Brazilian migrants. Specifically:

- **Targeting undecided profiles**: Institutions can identify respondents who are less clear in their categorizations, to visually direct campaigns and offers to those who are still considering their migration alternatives. These "indecisive profiles" are more likely to be influenced by outreach strategies and information quality.
- **Tailoring communication**: Since digital influencers and informal networks (WhatsApp groups, blogs) are strong information sources, institutional communication should adapt to these channels, employing influencers or alumni testimonials that resonate with each destination's appeal (e.g., safety for Ireland, lifestyle for Australia).

- **Timing interventions**: As decision-making duration emerged as a key variable, institutions can benefit from understanding when potential migrants are most receptive to messaging, particularly in the early phases when destination indecision is higher.
- **Service adaptation**: Educational institutions can tailor support services in relation to pre-migration status. For instance, persons migrating from unemployment or underemployment positions can require stronger financial planning support, whereas those migrating from institutions through international studies may need help with career or visa counselling, etc.

These strategies can help educational institutions relate better with Brazilian migrant profiles, and improve recruitment quality, user satisfaction and long-term integration works.

5.2 RESEARCH CONTRIBUTION AND ORIGINALITY

To the best of our knowledge, this is the first study to attempt to predict a migrant's destination country using a supervised machine learning model based on individual-level survey data. Unlike most existing research, which relies on public or institutional datasets, every step of this research, from survey design and data collection to feature selection and model training, was carried out independently. Due to the absence of publicly available datasets suitable for individual-level predictive modelling, we developed an original dataset specifically for this research. This lack of open data not only reveals a critical gap in the field but also underscores the importance of this study, capable of generating primary data for predictive purposes.

Because of the limited size of our original dataset, we were unable to explore more complex models such as deep learning, which would be highly prone to overfitting under these conditions. Nevertheless, this constraint reaffirmed the value of interpretable algorithms such as Random Forests, which performed reliably and enabled us to extract meaningful insights even with small data.

This work contributes to bridging the macro-micro divide in migration research: whereas macro-level models typically predict flows between countries and micro-level studies aim to identify associations between variables (e.g., motivations, demographics), our approach takes a step further. It uses individual-level attributes to predict outcomes (namely, country of destination), a shift from explanation to prediction. In this sense, our model complements, rather than replaces, traditional migration research frameworks by offering a new layer of analytical depth.

Notably, many of the micro-level studies we reviewed focused on identifying correlations and the relative importance of individual variables, which was highly valuable during the initial stages of our research, especially to design the survey entailing the descriptive features. However, predictive modelling at this microscopical level remains severely underexplored in

the literature. Our findings suggest that these same variables, when carefully selected and processed, can serve as effective inputs for practical classification models with strong predictive performance.

By combining behavioural theory with machine learning, this research opens a novel pathway for future studies in migration: anticipating individual choices based on data, rather than merely explaining them in retrospect. It also highlights the importance of investing in primary data collection to unlock new analytical possibilities in a field that remains dominated by descriptive approaches.

5.3 LIMITATIONS OF THE MODEL

Despite the encouraging results, several limitations should be noted:

- **Sample size:** The Random Forest model demonstrated strong predictive performance, and its accuracy remained reasonable even under a relatively small sample size (n = 173). However, to enhance generalization and enable the development of more complex models with finer segmentation, such as analysing motivations within each destination, a larger sample size would be necessary.
- **Data bias:** The survey collected self-reported data online, meaning that the sample was likely biased towards migrants with higher digital engagement (e.g. being active on social media). This digital engagement may also have biased the importance of influencers or online communities.
- **Class imbalance**: Although stratified sampling was used, the base distribution (61% Australia vs 39% Ireland) still introduced challenges in achieving equal accuracy across classes. While PCA and Random Forest mitigated this, accuracy for Ireland predictions was still lower in cross-validation (61%) compared to Australia (84%).
- **Exclusion of third countries**: The decision to focus on Australia and Ireland improved model clarity but removed the possibility of analysing wider migration trends. As a result, findings are only valid for comparisons between these two destinations.
- **Interpretability trade-offs**: Although Random Forest outperformed other models, its "black box" nature limits the interpretability of the exact decision paths taken by the model. This may inhibit clarity in the context of policy creation or institutional applications that require clear and defined logic chains.

Future research should look to mixed-methods, move to a larger sample size to promote generalisability across classes, and study more granular predictors such as language skills, financial planning, or emotional readiness, which were difficult to comprehensively study within this data set.

5.4 FUTURE RESEARCH DIRECTIONS

This study opens several avenues for future research, particularly within the growing intersection of behavioural migration studies and predictive modelling.

First, subsequent studies would benefit from an expanded sample size and broader geographic scope. While our model successfully predicted destination choices between Australia and Ireland, its generalisability remains limited by the modest number of observations and the exclusion of alternative destinations. Including additional countries, such as Canada, Portugal, or the United Kingdom, could enable multiclass classification and provide more nuanced insights into the drivers of destination choice.

Second, the model could be improved by incorporating new predictors that were outside the scope of this initial study. Variables related to language proficiency, financial planning, emotional readiness, or the presence of dependents may help refine classification accuracy and further differentiate between types of migration decisions. Such additions could also support the construction of more granular persona profiles for applied use by institutions.

Third, mixed-method approaches may enhance the interpretability and explanatory value of predictive models. Combining supervised machine learning with qualitative interviews or focus groups would allow researchers to validate patterns identified algorithmically and uncover latent or context-specific factors not easily captured through structured survey instruments (Angenendt & Koch, 2024).

Fourth, additional research is needed to explore the longitudinal dimension of migration decisions. Most predictive models, including ours, capture the decision-making process at a single point in time. However, by following migrants over time, future studies could assess how initial motivations and expectations evolve and whether they align with post-migration realities. This time series data would enable the development of models that not only predict destination choice, but also permanence, satisfaction, or the likelihood of return migration.

Lastly, the approach proposed focused primarily on behavioural data and predictive modelling and it should be tested across different migrant populations. Applying similar techniques to other national contexts (e.g., Indian, Filipino, or Nigerian migrants) could help validate the broader applicability of this methodology and support the design of culturally adapted recruitment and support strategies. Successful applications in forecasting forced migration flows (Bosco et al., 2024) suggest there is further potential for using these tools to anticipate patterns in voluntary migration as well.

By addressing these directions, future studies can further bridge the gap between macro-level forecasting and micro-level behavioural insight, building stronger foundations for both academic research and applied migration policy design.

BIBLIOGRAPHICAL REFERENCES

Angenendt, S., & Koch, A. (2024). *Digital tools and data analytics for migration forecasting.* Forced Migration Review, (75), 24–26. https://www.fmreview.org/digital-disruption/angenendt-koch/

Bijak, J. (2010). *Dealing with uncertainty in international migration predictions: From probabilistic forecasting to decision analysis.* EUROSTAT, Work session on demographic projections Lisbon, 28–30 April 2010. Luxembourg: Eurostat.

Bosco, C., Minora, U., Rosińska, A., Teobaldelli, M., & Belmonte, M. (2024). *A machine learning architecture to forecast irregular border crossings and asylum requests for policy support in Europe: A case study.* Data & Policy, 6, e81. https://doi.org/10.1017/dap.2024.48

Caselli, G., et al. (2004). *Démographie: analyse et synthèse*. Vol. V, 330–332 IOM. Glossary on Migrations.

Castelli, F. (2018). *Drivers of migration: Why do people move?* Journal of Travel Medicine, 25(1), tay040. https://doi.org/10.1093/jtm/tay040

Cowley, P., & Hyams–Ssekasi, D. (2018). *Motivation, Induction and Challenge: Examining the Initial Phase of International Students' Educational Sojourn.* Journal of International Students*,* 8(1), 109–130. https://doi.org/10.32674/jis.v8i1.154

Docquier, F., & Rapoport, H. (2012). *Globalization, brain drain, and development*. Journal of Economic Literature, 50(3), 681–730. https://doi.org/10.1257/jel.50.3.681

Gatfield, T., & Chen, C. H. (2006). *Measuring student choice criteria using the Theory of Planned Behaviour: The case of Taiwan, Australia, UK, and USA.* International Journal of Educational Management, 20(7), 544–555. https://doi.org/10.1300/J050v16n01_04

Global Refuge. (2021, July 14). *Why do people immigrate? The different causes of immigration.* Global Refuge. Retrieved from https://www.globalrefuge.org

InterNations. (2019). *Expat Insider 2019: The World Through Expat Eyes.* Retrieved from https://www.internations.org/expat-insider

McAuliffe, M., & Triandafyllidou, A. (2021). *World Migration Report 2022*. International Organization for Migration (IOM). https://publications.iom.int/books/world-migration-report-2022

Ministério das Relações Exteriores do Brasil. (2024). *Comunidades brasileiras no exterior – Estatísticas 2023*. Governo Federal. https://www.gov.br/mre/pt-br/assuntos/portal-consular/comunidades-brasileiras-no-exterior-estatisticas-2023

Morgenstern, S., & Strijbis, O. (2024). *Forecasting migration movements using prediction markets.* Comparative Migration Studies, 12, Article 45. https://doi.org/10.1186/s40878-024-00404-0

Nghia, T. L. H. (2019). *Motivations for Studying Abroad and Immigration Intentions: The Case of Vietnamese Students.* Journal of International Students, 9(3), 758-776. https://doi.org/10.32674/jis.v0i0.731

Nikou, S., & Luukkonen, M. (2024). *The push-pull factor model and its implications for the retention of international students in the host country.* Higher Education, Skills and Work-Based Learning, 14(1), 76–94. https://doi.org/10.1108/HESWBL-04-2023-0084

Oxford Academic. (2018). *Drivers of migration: Why do people move?* Journal of Travel Medicine, 25(1). Retrieved from https://academic.oup.com/jtm

Richardson, S., & Lester, L. H. (2004). *A comparison of Australian and Canadian immigration policies and labour market outcomes* [Report to the Department of Immigration and Multicultural and Indigenous Affairs]. National Institute of Labour Studies, Flinders University. Retrieved from https://www.researchgate.net/publication/252056325_A_Comparison_of_Australian_and_Canadian_Immigration_Policies_and_Labour_Market_Outcomes

Rosa-Luxemburg-Stiftung. (2019). *Atlas of migration: Facts and figures about people on the move* (3rd ed.). https://www.rosalux.de/fileadmin/rls_uploads/pdfs/sonst_publikationen/atlasofmigration2019web1906141.pdf

Robinson, C., & Dilkina, B. (2018). *A machine learning approach to modelling human migration.* In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. ACM. https://dl.acm.org/doi/pdf/10.1145/3209811.3209868

Rodrigues, T. F. (2022). *As migrações europeias numa nova era*. *Relações Internacionais*, 75, 5–11. https://doi.org/10.23906/ri2022.75a01

Sohst, R., Tjaden, J., de Valk, H., & Melde, S. (2020). *The future of migration to Europe: A systematic review of the literature on migration scenarios and forecasts.* International Organization for Migration, Geneva, and Netherlands Interdisciplinary Demographic Institute, The Hague.

TheGlobalEconomy.com. (2024). *Brazil: Human flight and brain drain index.* Retrieved from https://www.theglobaleconomy.com/Brazil/human_flight_brain_drain_index/

Tjaden, J., Auer, D., & Laczko, F. (2018). *Linking migration intentions with flows: Evidence and potential use.* International Migration, 57(1), 36–57.

United Nations Department of Economic and Social Affairs (UN DESA). (1998). *Recommendations on Statistics of International Migration,* Revision 1. United Nations. https://unstats.un.org/unsd/publication/seriesm/seriesm_58rev1e.pdf

United Nations, Department of Economic and Social Affairs, Population Division. (2020). *International Migrant Stock 2020.* (United Nations database, POP/DB/MIG/Stock/Rev.2020). https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2020_international_migrant_stock_documentation.pdf

United Nations Network on Migration. (2018). *Drivers of migration: Why do people move?* United Nations. Retrieved from https://migrationnetwork.un.org

Wiśniowski, A., Bijak, J., Christiansen, S., Forster, J. J., Keilman, N., Raymer, J., & Smith, P. W. F. (2013). *Utilising expert opinion to improve the measurement of international migration in Europe.* Journal of Official Statistics, 29(4), 583–607. https://doi.org/10.2478/jos-2013-0041

World Bank. (2023). *WDR2023 migration database*. Washington, DC. https://www.worldbank.org/wdr2023/data

APPENDIX A – SUVERY QUESTIONNAIRE

Pesquisa sobre motivações e desafios de brasileiros que vivem no exterior

Esta pesquisa faz parte do Master in Data-Driven Marketing da Universidade Nova de Lisboa e tem como objetivo compreender as motivações, desafios e experiências dos brasileiros que vivem no exterior.

Todas as respostas são confidenciais e serão utilizadas exclusivamente para fins acadêmicos. A pesquisa leva aproximadamente 10 minutos para ser concluída.

Agradecemos sua participação!

Seção 1: Fatores socio-demográficos

1. **Qual a sua idade atual?** *
2. **Para qual país você imigrou?** *
    a. Estados Unidos
    b. Canadá
    c. Reino Unido
    d. Irlanda
    e. Austrália
    f. Portugal
    g. Alemanha
    h. Espanha
    i. Itália
    j. França
    k. Outro: [Entrada de texto]

3. **Em que ano você imigrou?**

4. **Durante o processo de decisão, você considerou migrar para outros países?** (Selecione todos que se aplicam) *
    a. Estados Unidos
    b. Canadá
    c. Reino Unido
    d. Irlanda
    e. Austrália
    f. Portugal
    g. Alemanha
    h. Espanha
    i. Itália
    j. França
    k. Outro: [Entrada de texto]
    l. Não considerei outro país para imigrar

5. **Como você se descreve?** *
    a. Masculino
    b. Feminino
    c. Não-binário/Terceiro genêro
    d. Prefere a autodescrição: [Entrada de texto]
    e. Prefere não dizer

6. **Qual era o seu nível de escolaridade antes de migrar?** *
    a. Ensino Fundamental
    b. Ensino Médio
    c. Ensino Técnico
    d. Graduação Incompleta/Em curso
    e. Graduação Completa
    f. Pós-Graduação
    g. Mestrado
    h. Doutorado
    i. Outro: [Entrada de texto]

7. **Qual era o seu estado civil antes de migrar?** *
    a. Solteiro(a)
    b. Namorando
    c. Casado(a)/Em união estável
    d. Divorciado(a)/Separado(a)
    e. Viúvo(a)

8. **De que estado você é do Brasil?**

9. **Com quem você residia antes de migrar?**
    a. Eu morava sozinho(a)
    b. Morava com amigos
    c. Morava com namorado/parceiro
    d. Família nuclear (pais e irmãos/cônjuge e filhos)
    e. Família ampliada (com avós, primos, etc.)

10. **Você imigrou sozinho(a) ou acompanhado(a)?** *
    a. Sozinho(a)
    b. Com cônjuge/companheiro(a)
    c. Com família (filhos, pais, etc.)
    d. Com amigos
    e. Outro: [Entrada de texto]

11. **Qual era sua renda familiar mensal no Brasil antes de migrar?** *
    a. Menos de R$ 2.403
    b. R$ 2.403 – R$ 3.980
    c. R$ 3.980 – R$ 7.017
    d. R$ 7.017 – R$ 12.683
    e. R$ 12.683 – R$ 28.240
    f. Mais de R$ 28.240

12. **Qual era a sua área de atuação antes de migrar?** *
    a. Tecnologia
    b. Saúde
    c. Educação
    d. Engenharia
    e. Comunicação/Marketing
    f. Recursos Humanos
    g. Advocacia
    h. Contabilidade
    i. Vendas/Atendimento ao público
    j. Outro: [Entrada de texto]

Seção 2: Motivações para Migrar

13. **O que te motivou a se mudar para o exterior?** (Classifique na escala em que 1 = Nada importante e 5 = Extremamente importante)
    a. Melhores oportunidades de trabalho
    b. Estabilidade econômica
    c. Acesso à educação de qualidade
    d. Segurança
    e. Aprender/melhorar o idioma local
    f. Experiências culturais
    g. Oportunidades de vistos de trabalho, estudo ou cidadania
    h. Outro: [Entrada de texto]

14. **Como você acessou informações para se preparar para a migração?** (Selecione todas que se aplicam) *
    a. Agente de imigração
    b. Agência de intercâmbio
    c. Escola ou universidade no país de destino
    d. Amigos ou familiares no país de destino
    e. Sites de governo
    f. Grupos no Facebook
    g. Grupos no WhatsApp
    h. Redes Sociais (Instagram, TikTok, YouTube)
    i. Blogs ou sites pessoais
    j. Reddit ou outros fóruns online
    k. Influenciadores digitais. Quem? [Entrada de texto]

l. Outro: [Entrada de texto]

15. **O quanto as possibilidades de oportunidades futuras no país de destino influenciaram sua decisão?** *
    i. Muito influentes
    j. Influentes
    k. Neutras
    l. Pouco influentes
    m. Nada influentes

16. **Quais barreiras você encontrou ao planejar a migração?** (Classifique na escala em que 1 = Nada importante e 5 = Extremamente importante) *****
    a. Língua do país de origem
    b. Custos financeiros
    c. Processo de visto
    d. Falta de informações confiáveis
    e. Apoio insuficiente de amigos ou familiares
    f. Outra: [Entrada de texto]

17. **Como você avaliaria seu conhecimento do idioma local antes de migrar?** *
    a. Fluente
    b. Intermediário
    c. Básico
    d. Nenhum

18. **Você já tinha algum familiar ou amigo próximo morando no país de destino?***
    a. Sim, um familiar próximo
    b. Sim, um amigo próximo
    c. Meu parceiro já morava no país de destino
    d. Apenas conhecidos
    e. Não tinha ninguém

19. **Quanto tempo levou o seu processo de decisão, desde que começou a pesquisar sobre o país de destino até efetivamente se mudar?**
    a. Menos de 1 mês
    b. 1 a 3 meses
    c. 4 a 6 meses
    d. 7 a 12 meses
    e. 1 a 2 anos
    f. 2 anos ou mais

20. **Verificação de Atenção: Selecione "Neutro" para continuar.**
    a. Muito fácil
    b. Fácil
    c. Neutro
    d. Difícil

e. Muito difícil

Seção 3: Após Migrar para o País de Destino

21. **Quanto tempo você está morando no país de destino?** *
    a. Menos de 6 meses
    b. 6 meses a 1 ano
    c. 1–3 anos
    d. 3–5 anos
    e. Mais de 5 anos

22. **Como foi ou está sendo o processo de solicitação de visto para residir no seu país de destino?** *
    a. Extremamente difícil
    b. Parcialmente difícil
    c. Nem fácil nem difícil
    d. Parcialmente fácil
    e. Extremamente fácil
    f. Não precisei de visto, pois possuo cidadania
    g. Não fiz solicitação de visto, estou irregular

23. **Quais fatores foram mais importantes para ajudá-lo(a) a se estabelecer no país de destino?** (Classifique na escala em que 1 = Nada importante e 5 = Extremamente importante) *
    a. Redes sociais /influenciadores digitais
    b. Amigos/familiares já no país de destino
    c. Apoio de comunidades locais
    d. Proficiência no idioma local
    e. Familiaridade com a cultura do país

24. **Quais barreiras você enfrenta para permanecer no país de destino?** (Selecione todas que se aplicam) *
    a. Custos de vida elevados
    b. Falta de visto de longo prazo
    c. Barreiras culturais
    d. Barreiras linguísticas
    e. Outro: [Entrada de texto]

25. **Como você acessou informações para se preparar para a sua migração?** (Selecione todas que se aplicam) *
    a. Agente de imigração
    b. Agência de intercâmbio
    c. Escola ou universidade no país de destino

d. Fóruns online e redes sociais
e. Amigos ou familiares no país de destino
f. Sites de governo
g. Outro: [Entrada de texto]

26. **Quais são seus planos futuros?**
    a. Já tenho minha vida estabelecida fora do Brasil e não planejo voltar
    b. Não pretendo voltar para o Brasil, mas preciso conquistar minha residência primeiro
    c. Planejo voltar para o Brasil em menos de 2 anos
    d. Quero voltar para o Brasil um dia, mas não sei quando
    e. Ainda estou experimentando opções e não tenho uma opinião formada
    f. Planejo imigrar para outro país. Qual?

Encerramento

Obrigado por compartilhar sua experiência! Sua participação é essencial para entendermos melhor as dinâmicas da migração de brasileiros ao redor do mundo.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa