A Machine Learning Approach to Understanding Migration Decisions and Psychological Stress Among Youth.

By
Sajib Biswas
ID: 213-15-4347

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Md. Abbas Ali Khan

Assistant Professor
Department of Computer Science and Engineering Daffodil International University

Co-Supervised by

Mr. Md Assaduzzaman

Assistant Professor
Department of Computer Science and Engineering Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 16, 2025

APPROVAL

This Project titled "A Machine Learning Approach to Understanding Migration Decisions and Psychological Stress Among Youth," submitted by Sajib Biswas, ID No: 213-15-4347 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 16-09-2025.

BOARD OF EXAMINERS

16.09.25

Dr. Bimal Chandra Das (BCD)
Board Chairman
Professor & Dean (In-Charge),
Department of CSE, FSIT,
Daffodil International University

Most. Hasna Hena (HH)
Internal Examiner 1
Assistant Professor, Department of CSE, FSIT,
Daffodil International University

Mr. Mayen Uddin Mojumdar (MUM)
Internal Examiner 2
Assistant professor, Department of CSE, FSIT,
Daffodil International University

Nazibur Rahman
External Examiner
Technical Lead · Database Administrator,
Telenor · Grameen Phone Account

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Abbas Ali Khan, Assistant Professor,** Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Md. Abbas Ali Khan
Assistant Professor
Department of Computer Science and Engineering Daffodil International University

Co-Supervised by:

Mr. Md Assaduzzaman
Lecturer (Senior Scale)
Department of Computer Science and Engineering Daffodil International University

Submitted by:

Sajib /09.09.25

Sajib Biswas
Student ID: 213-15-4347
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Md. Abbas Ali Khan, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Machine Learning (ML)** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

In this paper, we considered migration induced predictions by machine learning algorithms. In the determinants of migration, which are age, occupation, psychological pressure and social media effect, the costs are known. The models are trained on a full data set prior to launching. It applies Random Forest, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), XGBoost and Logistic Regression methods to migration tendency prediction. The Random Forest model (best performing) also made high accurate predictions, 81.35% correct predictions also. It also contains a recommendation module for personalized feedback to subjects. Last but not the least, `model interpretability methods' like LIME and SHAP are used to serve the prediction in an interpretable way. So from that standpoint it's an experiment for the first time and it may work out.

The research ethical considerations involved, such as data privacy, fairness and interpretability, could be said for the opposite case which occurs as the integrity for machine learning can be strengthened. Offering us predictive analytics and implementable solutions to these challenges, it is improving our migration process and emerging as more one governed by public decisions. It offers scalable, interpretable predictions for migration and new directions of research in this domain.

Table of Contents

Section **Page Numbers**

List of Figures

List of Tables

Chapter 1

Introduction

This chapter gives you a short and sweet demonstration of the report; from its motivation and the reason for its performance. It readies the reader for the rest of the report. The purpose of the This section presents the research questions, predictions and content of the report to help structure the reader.

1.1 Introduction

Youth's migration intention has been increasingly a focus of studies due to its economic opportunities, social aspirations, psychological status, and environmental issues. On the global level, the youth is becoming increasingly spurred to migrate with administrative frustration, unemployment, income disparities and inadequate access to resources [2]. In Bangladesh, the flow of migrants has increased internal as well as additional pressure of international migration has also increased in recent years due to complicated dynamic of push factors in the form of economic reason, political condition, climate related risk and psychological stress and pull factors like better employment opportunity, access to better education, better lifestyle etc [6]. Research indicates that social norms and family expectations and aspirations for urban way of living play major roles in making decisions to migrate, especially among young people [17]

While previous studies have improved our knowledge of factors inducing migration, important gaps remain. The majority of studies are based on descriptive analysis and cross-sectional surveys which, however, used useful but are always not predictive and could not provide personal insights to those who want to migrate [21]. Likewise, research on mental health and migration experiences demonstrate the significance but neglected nature of psychological distress, depression, and social marginalization [16]. Besides, unpredictable migration flows, particularly of low-skilled laborers are still a challenge which is hardly predictable as integration of data is still limited and dynamic model for migration is missing [9].

To fill that gap, the present study attempts to develop a predictive model of migration intention (Yes or No) via survey data on individual's characteristics by using machine learning (ML) algorithms. Several algorithms, such as Logistic Regression, Random Forests, Support Vector Machines (SVM), Bernoulli Naive Bayes, and Extreme Gradient Boosting (XGBoost), were tested, with Random Forests and Bernoulli Naive Bayes showing the highest accuracy of prediction, as also found in similar migration analyses [13]. Furthermore, an interactive decision support tool was deployed based on Streamlit and Flask. Unlike traditional binary prediction models, this approach uses user-level data to produce individualized information that identifies barriers, mental stressors, and goals that are predictive of the migration outcome for the user.

This research builds an important line of the emerging data-driven migration literature in unifying behavioral economics and psychological factors with state-of-the-art machine learning methods. Our framework addresses this positive gap between the stakeholders interested tool, description migration studies, and the computation predictive model to provided an hard and fast, cost effective, scalable and user friendly way to elicit migration intention in the resource constrained settings like Bangladesh. The study synthesises predictive modelling and interactive analytics to empower those who need to understand and manage migration flows, researchers and policymakers to make data driven decisions [7].

1.2 Motivation

The determinants of migration, and in particular the determinants of international migration, are a complex set of multivariate social economic phenomenon, and thus are endogenous. Respectively, these are opportunity, network, wellness, environmental stress, and aspiration. Historically, migration was explored descriptively and explanatively from people who had moved or stayed in history, but the fact that there are only more and better quality data on people and households is indicative of the fact it would be desirable to have predictive tools that will tell in advance who is more likely to be willing to move or more likely to be willing to stay. It has its implications in terms of resources and programmes for potential migrants.

In countries such as Bangladesh (and elsewhere), a new generation of young people faces the double whammy of growing up poor while dealing with the full gamut of climate threats (and other stressors), so understanding who wants to move and why is essential. But the way this migration happens currently is often far too dismissive of the intricate psychological, social and economic fabric that makes up migration and re-integration. Potentially, a useful means of predicting migration intention (comprising both migration desire and future intention) with the satisfactory level of accuracy is to employ machine learning techniques that can capture the complex interactions among several set of input features (Characteristics in demographics, psychological indicators, family status and environmental stimuli). And further, the decisions that lead to migration are hardly ever as simple as yes or no. They fall under the domain of plan, readiness, perceived barriers and stress. In line with predictive models on migration intentions in terms of binary decisions (migration "yes" or "no"), novel decision-oriented models are required which certainly present the population and/or individual relational views on what is expected to happen and what would cause what a hindrance if they were to act upon it. These methods might be useful for researchers, or policy makers to read more in depth about migrants' patterns under the statistical aggregations. This is hence the demand-driven motivation of this research: to establish a correspondence between tradition-orientated migration studies, which have been aimed at providing a descriptive account of migration processes, and the dataoriented prediction studies. Create a central predictive marker to predict potential migrants earlier and establish simple, accessible tools for clinicians used as clinical buffers that can further shed light on the psychosocial and structural context of migration decision-making. This twofold strategy seeks to contribute to a more sensibilized and human research and migration policy in Bangladesh but also beyond.

1.3 Rational of the Study

The phenomenon of cross-border human mobility has become more relevant with the growing global movement of people driven by social, economic, political and environmental factors, making migrants' experiences important in grasping the

decision to migrate. However, predicting whether a person will migrate is a complex and challenging task. In doing so, we hope to address this in this chapter by leveraging the strengths of machine learning to predict intentions to migrate from combinations of socio-economic, psychological and demographic characteristics. Through the previous methods, the non-linear relationship between these factors could not be well-modelled, which usually have subtle and dynamic correlations.

This paper is motivated to improve the prediction ability better than what classic statistical models can do by using machine learning algorithms: Random Forest, XGBoost, Logistic Regression and SVM. "This will uncover more about what factors, such as household income, mental health and employment prospects, have the greatest effect on migration intentions. In addition, this study's interactive prediction tool represents a novel approach to migration research. This measure is distinct from the classical forecasting indicators as it not acts simply to forecast the results but also provides users with concerned consequences, frustration and barriers when people migrate. Such an interactive item could be applied to assess more sophisticated migration intentions that may be particularly useful to potential migrants and policy makers to better manage migration flows.

Finally, the current study aims to contribute to migration research based on the use of large-scale data by constructing a more accurate, scalable, and interpretable migrant intentions prediction model. The findings will allow policy makers and practitioners to design targeted interventions that strike at the root causes of migration and have a more durable impact on its consequences.

1.4 Research Questions

Prediction of youth migration decisions and related psychological stress using machine learning algorithms and personalized recommendation systems. These queries provide guideposts for a concerted exploration of how computational methods can enable early warning, intervention and policy planning with respect to the interconnection of migration dynamics and youth mental health. The overarching research questions include:

Q1: How can machine learning model learn from age, gender, occupation, family ties abroad and social-based features to categorize individuals in migration decision category (yes/no/not sure yet)?

This research seeks to investigate the ways of how machine learning could make use of demographic and social data to classify candidates of migration intentions. It will also explore whether machine learning models can predict and represent the true drivers of migration decisions, thus potentially providing an approach that scales better and finds relevance through far less prior human work than traditional descriptive studies.

Q2: Which are the similar machine learning algorithms and how effective there are in terms of predicting migration intentions, and how they affect the reliability and generalizability of the whole system?

This question compares different machine learning algorithms (e.g. Random Forest, Support Vector Machines, and XGBoost) and compares the ability of each to predict migration intentions. It is to ensure the best trade-off between reliability and generalisability of models for accurate migration prediction returns, in terms of accuracy, precision, and recall.

Q3: How could the risk scores for the migration intention be combined with personalizing recommendations, and how would such recommendations differ according to individual characteristics of stress, socio-economic status, and mental health?

This series investigates whether personality and lifestyle can influence risk scores produced by machine-learning algorithms to offer personalized migration advice. It investigates how mental health, stress and socio-economic situation might interact to influence decisions on migration and how a person centered recommendation could guide individuals towards an informed decision to migrate.

1.5 Expected Output

This research is expected to make significant contributions to both theoretical development and applied practice for understanding of youth migration and psychological stress. The anticipated outcomes of the study are as follows:

An Intelligent Migration Decision and Stress Prediction Framework

A Knowledgeable Relocation Decision and Stress Prediction System The immediate output of this research will be a device for categorising people based on the things next or not (Yes, No and Not Yet decision of mobility)and the categories of levels (Low, Medium and High)of psychological stress in term as model's predictions developed. These projections will rely on a combination of demographic and behavioral factors. We envisage a previous that will lead to interpretable outputs such as the cause of migration intention, identification of key stressors (e.g., visa problems, family pressures), and possible coping strategies which can be incorporated in profile information individual users.

Model Comparison and Machine Learning Algorithms Assessment

We will make comprehensive analysis and comparision with many other machine learning algorithms(for example, Logistic Regression, Random Forest, Decision Tree,XGboost,LightGBM,SVM,KNN). We are going to study and compare these models concerning overall performance in accuracy, precision, recall & F1-score that will help us find the most resourceful and reliable algorithms for predicting migration decision and psychological stress with high correctness.

Integration of Personalized Insights

In addition to predicting decisions, system will provide personalized advice based on migration and stress predictions. For example, the method identifies particular barriers (e.g., financial barriers, limited education, etc.) for that individual and creates personalized recommendations given that individual's unique circumstances. One strength of using this approach is its potential to develop targeted interventions for discrete subsets of youth, providing policymakers and practitioners with approaches to target intervention efforts. So this integrated Insight will very helpful.

Data Processing and Feature Engineering at Scale

Advanced data preprocessing such as one-hot encoding, class balancing (SMOTE), feature selection, will be used to train the model for better prediction accuracy. These techniques will improve the algorithm's capability to process the heterogeneous data and promote the transferability of the algorithm (to different demographic groups and/or context).

Rigorous Validation of Results

Cross-validation will be applied to the database and the testing will be conducted in part of the database to evaluate the accuracy, reliability, and generalization of the predictive model. Strict validation would also guarantee the robustness and generalization of the results across different demographic groups of the predicted migration intention and stress.

Publication in Academic and Practical Literature

Such a study would significantly add to the increasing corpus of literature in the use of machine learning methods in migration studies and psychology. The findings of this study will be used to inform policy and practice and provides an evidence base, from which further technology based interventions that can alleviate psychological stress and promote informed and safe migration decision making can be built.

1.6 Organization of the Report

This draft report is arranged in a series of chapters that are meant to guide the reader through the research project, its results and their possible implications. To ensure coherence, clarity, and depth to the instructional design each chapter has been moulded around specific aspects of this study.

Chapter 1: Introduction

In this chapter, the research is presented on Bangladeshis are migrating abroad in search of better education, jobs and living standards. Nowadays the youth want to go abroad for better prospects. But migration is driven by intricate socio-economic and psychological dynamics. Additionally, the precise formulation of migration intentions can help target support and policy planning. This research utilizes machine learning methods to predict mobility decisions and associated psychosocial consequences.

Chapter 2: Background

In the introductory section of the chapter, the chapter discusses relevant literatures A lot of per cent of students and young population With the dream to

study and work in abroadfrom Bangladesh. Destination Country is USA,UK, Canada, Australia and Gemany. Relocating to a foreign country Relocation to a foreign land is not for cowards; there's much to be done, in terms of the on-arrival visa requirements, the financial pre-planning and the deliberate adjustment to the social atmosphere, the bureaucratic frustrations and the cultural expectations. But cat-towing however involves a series of stressors, and stress will be from the mind sets and from the social relations, and from the "what-ifs" from the site of the event itself. "We need to understand what form is that decision to migrate in, so we can take them into account in our analysis." The present paper's research uses machine learning to forecast migration aspirations, and provide personalised information about psychological and socioeconomic effects of migrating.

Chapter 3: Research Methodology

This chapter explains the methodology followed in the study. It lists the dataset selection, preprocessing steps, definition of target variable and then feature extraction methods. A detailed description of how the machine learning and deep learning models that have been selected are designed and configured along with an explanation of the evaluation metrics used to measure performance is given. It also discusses the data privacy and informed consent ethical dimensions.

Chapter 4: Experimental Result

The predictive performance of the machine learning model in categorizing intention to migrate based on sociodemographic and psychological variables was reasonable. In comparison to logistic regression and support vector machines, Random Forest emerged as the best predictive tool in most of the analyses conducted to discriminate the most important factors of migration decisions in complex survey data from Bangladesh. These findings were confirmed with the interactive prediction tool which was also sensitive for more subtle factors, such as stress and perceived barriers, indicating the practicability of the model for the discovery of early intentions to migrate. This is a result that makes it justifiable to consider machine learning for improving the accuracy of the migration

prediction for Bangladesh.

Chapter 5: Impact on Society

This chapter is ready to conduct the international migration of Bangladesh and added more in social and economic development getting remittances that helped in increasing household income levels as well as in reducing poverty. The immigration has accelerated employment and new business formation within the communities as well, raising local prosperity. But social exploitation and mental stress of migrants also underscore the need for better governance and support to better safeguard on-the-move people's welfare.

Chapter 6: Summary, Conclusion, Recommendation, and Implications for Future Research

The concluding section locates the main findings of this study and suggests that that engine learning methods can predict migration intentions quite well, under some specific socio-economic and psychological conditions. Besides the binary classification on this kind of data, we add an interactive tool for prediction to dive deeper into dimensions of stress and barriers of migration. Policymakers might pursue data-driven approaches like this to find places to focus assistance on vulnerable populations. The limitations and future work are to enhance the scale of datasets, reckon the longitudinal data and obtain the fusion of mental health related data for a more precise model prediction. And we were arguing, that the socio-cultural aspect of such experiences of migration should also be further investigated, with the idea of developing more inclusive models of migration. Overall, this work advances this field towards more knowledge-driven, user centred migrant prediction and intervention planning.

Chapter 2

Background

The background of the study, starting with important preliminaries and terminologies that are used in the present research, is provided in this chapter. The literature review is significantly contributing to the development of research questions or hypotheses and is part of the understanding and synthesis of literature about a study topic. The chapter concludes with a summary of the findings and discussions.

2.1 Introduction

Migration plays a significant role in physical, mental and social health. Bad choices in migration or the pressure during migration can lead to serious health problems, such as anxiety, depression, and social isolation. Migration intentions have gained attention from researchers in recent years since they have long-term implications on the socio-economic circumstances and psychological health. However, conventional statistical methods for investigating migration intentions (e.g., interviews and surveys) are expensive and time-consuming and are not applicable when considering large scale or continuous monitoring. This is why machine learning tools have gained popularity as good models for analyzing lifestyle and behavioral data that can classify health issues related to migration more precisely. 2 Background to the Study In this section, background of the study is covered, through related literature review, similar application, gap identification and problem statement summary.

2.2 Preliminaries/Terminologies

These studies on international migration are actually an interdisciplinary field with theoretical points of gravity that serve to understand this incontestable aspect of human mobility. For most of us, migration is the movement of human beings from one territory to another, individuals going somewhere else to find work, to join family and community, to escape environmental stresses or political

persecution. Members of these movements are the “migrants,” a category that runs from informal laborers to lifelong residents. As used within our explanatory framework, the forces of migrating behavior are a homeostatic construct that refers both to the state of the mind that predisposes it to migrate or that the mind feels like or which strongly desires or is even prone to migrate, and that which varies excerpt for the subjective or reprised in the study respondents, the latter search for underlying motives be complex and studied factors to be of both of complex and socioeconomic psyche and psyche forces.Another fundamental terminological point that is used thereafter is the generation and delivery of predictive models as computational frameworks that encapsulate machine learning components which take input data and issue binary predictions with respect to the likelihood of migration (‘Yes’ or ‘No’). These models used demographic, education and psychological distress-related factors and household composition-related and environmental exposure variables in order to describe the clustering of migration decisions. Also, online prediction tools make a different but complementary contribution, providing individual sympatico-contextual feedback on possible triggers, motivations, mental health stressors and migration intention barriers. Unlike standard statistical models, these devices would also contribute to You more interaction with users, you more interpretative interpretations and You more involvement of the user. The phrase “machine learning” is a loose description of the techniques and algorithms that enable this predictive and analytical capacity, or machine learning algorithms work and afford these predictive and analytic capabilities, through learning from patterns in data (i.e., machine learning algorithms are not explicitly programmed through code). Furthermore, psychological stress (e.g., anxiety about and uncertainty with reference to an opportune future) and the social dimension are identified as critical non-quantitative antecedents of migration intentions. In considering and integrating these prepared concepts further, within social-science and computational literatures, our work then lays the foundation for the comprehensive analysis and rich interpretation across the study, and in conducting a fine-grained study of migration decision making under a predictive, user-dynamic interface.

2.3 Literature Reviews

The migration decisions themselves have been one of the most researched topics in academia since the increasing popularity of data-driven models which seek to predict or find explanations for said decisions. Multiple previous researches have been conducted to discover migration patterns employing machine learning, which employed predictive-value models and decision support. These have included work on such topics as the socio-economic determinants of migration to applications for advanced computational methods such as machine learning to predict movements. The research is an attempt to understand the relationship between life satisfaction and migration intentions of youth in Komsomolsk-on-Amur (Russia) (17-35 years old). It appealed to 136 young inhabitants for its survey work and 48% of them expressed a desire to leave, with the main reasons cited as unhappiness with the political situation, the economy and lack of opportunities. Statistical analysis, carried out by the Mann-Whitney U test, emphasized the satisfaction with life in the city as the essential factor in the intention to move. The research suggests that the uncles can be suppressors of youth out-migration in the region by improving the economic and administrative conditions [7]. A second study explored determinants of self-rated health among internal migrants in Bangladesh by analyzing information from 1,754 households in the Bangladesh Environment and Migration Survey (BEMS). Most migrants perceived their health as better after migration, with occupational status, particularly having a professional occupation, considerably associated with better-perceived health. However, social support, education and food security had modest associations. The research also highlights the cross-sectional nature of the study and recommends longitudinal research to investigate further factors associated with the health outcomes of migrants in the future [1].

Sabti & Ramalu (2024) explore the effects of economic, political, and social push factors in the intention of medical doctors to migrate from Iraq with psychological distress. The survey-based cross-sectional study of 300 doctors shows that economic factors have a direct effect on migration intentions, while political instability and corruption have an indirect effect on migration intentions through its effect on psychological distress. The authors of the research said that enhanced

economic opportunities and support for mental health would lower skilled migration out of Iraq. Drawbacks One city was surveyed for the study and the survey was cross-sectional [3].

An investigate the prevalence of, and the risk to, posttraumatic stress symptoms (PTSSs) in Rohingya refugees in Bangladesh. After surveying 1,184 adults, the research found that 46.6% had severe PTSS, 23.1% had moderate PTSS. The study finds that severe PTSS is strongly associated with abuse prior to fleeing Myanmar, and denial of aid and employment in the camps exacerbate risks. Paid employment and adequate assistance were protective. The research calls for better mental health support and treatment. Limitations are self-reported symptoms and cross-sectional design [4].

A paper from [5] studied the role of information behaviour among Bangladeshi new immigrant settlers in Canada with a focus on Southern Ontario in their settlement experience. Data from 60 interviews and 205 surveys were used to explore pre- and post-arrival information-seeking and its impact on employment anxiety and settlement patterns. The chamber of commerce-specific factors "fear of information sharing" (reluctance to share negative experiences) and "information intelligence" (ability to gain needed information) were found to be primary factors. Social networks were reviewed in which support was provided through an unspecific quality profile of information. Culture-specific information services were found to be necessary to enhance the employment and acculturation of immigrants. Limitations were single-group, single-area data and non-probability sampling. Investigating the role of behavioral and social psychology in household risk management in the context of bangladeshi migrant work in the chittagong ready-made garment sector (RMG) The paper is based on a case study that examined the role of a number of behavioural and social psychological factors that influence intentions to migrate from rural to urban areas and in particular to work in Chittagong's ready-made garment (RMG) sector. Based on the theory of planned behavior (TPB), structural equation modelling was used to analyze survey data from 300 RMG workers (making up an equal number of male and female and are mainly young adults). The study shows that the subjective norms (i.e., social prestige of urban jobs, social networks influence, improvement of

family status, migration permitting), followed by perceived behavioral control (i.e., ease in getting urban jobs, individual freedom expectations), and personal attitudes (i.e., city life attraction, RMG sector opportunities, better wages) influence the intention towards migration was significant. The model accounted for 75% of the variance in intentions to migrate. The research argues that migration is not just about economic necessity, but also cultural preference and the attraction of urban life. It calls on policy makers to redress rural–urban development imbalances to manage uncontrolled urbanisation and rural disbenefits. Limitations are that this study was conducted in only one city and industry, and future studies about the rural influences on migration are recommended [6].

A study examine the portrayal of migrants of Arab and African descent in Spain's media, as well as the public's perception of them. Using pre-and post-media-exposure questionnaires of 130 college students, the article examines affective and appraisal responses. Machine learning algorithms (e.g., decision tree, random forest, SVM) were employed to classify the attitudes and to identify emotional and contextual predictors like fear, happiness, political orientation, and perceived media bias. The results are indicative of how political identity and media frame have a powerful effect on public attitudes. This paper presents a solid methodology by which computational approaches and social science can collaborate to gain insight into migration attitudes, and illustrates the need for interpretable models in social dimensions. This article examines the relationship between climate induced migration and depression among adolescents in Bangladesh within groups of migrants and non-migrants affected by different climatic disasters, i.e. cyclones, floods and drought [7]. The study was undertaken between April and June 2024, and included 1220 adolescents (515 migrants and 705 non-migrants) aged 13-19, determining depression using the PHQ-9 questionnaire. Results showed that depression (30.0% severe) was higher among migrant adolescents (84.0%) than among non-migrant adolescents (87.7%, severe 12.58%). Lack of income, poor sanitation and no relatives were associated with an increased rate of depression, with non-migrant females having the highest rate of depression. The study underscores the pressing mental health needs and

infrastructure to cope with the psychological fallouts of climate migration among young people [8]. An investigation study climate change impacts in Bangladesh, where increasing sea levels, cyclones, flooding, and soil salination are driving people out of vulnerable coastal areas to the cities of Dhaka and Khulna. This has triggered urbanisation with associated problems of overcrowding, poor housing and overstretched services. The research finds that standard migration theories do not account for the climate migration dynamics, and, through field surveys and government data, the study shows a substantial rise in migration after major disasters such as Cyclones SIDR (2007) and AILA (2009). The authors present a multilevel, vertically integrated planning framework, which connects national, regional and local policy and planning for the purpose of effectively absorbing climate migrants and facilitating sustainable urban development. They advise planning for inclusion that is part of costal policies to mitigate displacement and urban poverty, but there are limitations, as planning is concentrate at costal, and the modeling to predict is generated novel [9].

The paper evaluated HRQoL of 935 rural-to-urban migrants who were living in urban slums in Dhaka and Gazipur cities, Bangladesh, by means of the SF-12. The findings revealed a lower HRQoL compared to the general population, with the mean of the physical and mental component summary´s: 57.40 and 60.77, respectively. Multivariate analysis revealed that older age, female gender and unemployment were significantly associated with lower PCS. Older age, working as a day wage laborer/unemployed, shorter working hours, and higher work stress were also associated with lower MCS. The slum with worst living conditions were included in the study (ie, poor housing, water supply, and lack of space), and when we compared toilets, some respondents referred to have better toilet type. It is a slum agenda, or fights for social safety nets, slum infrastructure, working conditions and getting medical access. Study limitations are cross-sectional data collection and the sample size was limited to slums only. The findings of the study indicate the necessity for targeted policies to enhance the wellbeing on the slum-dwelling internal migrants. One more key publication analysed the migration intentions of young adults (16-25) in nine EU member states through an online panel survey (20,473 respondents; non-student sample). It ananlysed the

influence of macro (economic), meso (social networks) and micro (age, sex) determinant on intention to migrate. The study found that there is a 17% likelihood for respondents to migrate within a year after interview and almost one in five respondents will migrate within 5 years. There was a strong economic gradient (financial situation, unemployment) but also a non-economic one (sensation seeking, risk taking) that was stronger especially in Germany and the UK [10]. There was also diversity in terms of the regions people wanted to emigrate to (Southern European countries with high unemployment were the most common choice of destination). Older age was a depressor of the probability of migration, while frequent travelling and having previous migration experience were strong predictors. The study highlights particular structural and individual factors and identifies that youth mobility from within the EU is a matter for policy targeted intervention [11]. One migration paper explored utilising machine learning unsupervised topic modelling (STM) on an open-text survey data measuring population migration in the Northern Territory of Australia. The research, based on responses from over 3,500 people, identifies basic push and pull forces of migration. Migration decisions were influenced by positive effects in the form of lifestyle and cultural opportunities, as well as negative effects, such as a high cost of living, crime and poor service provision. Using STM allowed to provide a neutral perspective on how level of personal satisfaction and demographic factors calculate by age, place of residence, and place of birth - relate to migration intention. The results make a case for employing machine learning to source fine-grained qualitative data that will guide policy for rural retention and recruitment [12].Using machine learning in random forests, a household survey of nearly 1,700 households in southwestern Bangladesh was drawn upon to model migration patterns. The results point to easily measured factors to indicate wealth, household composition, and media exposure which determine internal migration. Economic constraints (i.e. less household appliances) inhibited migration but a higher percentage of economically in-active discouraged, probably due to facing demand of remittances. The study combined survival analysis within a machine learning environment when modelling the socio-economic factors influencing migration motivation. The results highlight

the value of combining predictive models with more classical statistical approaches in migration research [13].

[14] conducted a scoping review on the use of machine learning in mental health (MH) research with immigrants, refugees, and racial/ethnic minorities. The review included 13 peer-reviewed articles published between 2017 and 2022 and concluded that supervised ML models are becoming more popular to predict and classify mental health conditions including PTSD, depression, ADHD, and schizophrenia. Although the application of ML in the diagnosis of mental health disorders in underserved populations is promising, the review reveals that there are substantial shortfalls: restricted diversity in the data and the lack of generalizability across minority groups. The paper stresses the importance of representative datasets and careful algorithm design being employed to prevent biased outcomes, and points out the potential of ML to contribute to better mental health care, particularly for a vulnerable population. Research from [15] explores the potential of big data to contribute to the understanding of human migration throughout its three main phases; journey, stay, and return. The authors examine conventional sources of migration data, including such as census and surveys, and criticize these sources for being less timely and consistent. They continue by examining how non-traditional data sources – mobile data, online social networks, and supermarket retail data – provide real-time and fine-grained information on migration, integration and return. To calculate sentiment and expenditures, as well as language and ego networks, a multidimensional integration index is introduced. The article highlights the need to integrate renaissance data sources on migration to construct dynamic, real-time models using the best components of the old and the new, while analysing the ethical considerations.

An article from [16] consider the psychological effects of unemployment among highly-skilled migrant youth in Kolkata, India. A sample of 395 migrant youth (21–35) is surveyed and the study reports high levels of depression (54.4%), anxiety (61.8%), and stress (47.9%) based on DASS-21. This study highlights important risk factors such as being unemployed, female sex, never married, and repeated migration. It is noteworthy that there were higher levels of stress among

working youths, which could be attributed to contractual employment and workload. The study recommends skill-based, career-specific courses at postgraduate level and regular psychological counselling to cope with mental illness among educated migrant youth. A study of depression, anxiety and stress among the Bangladeshi migrant workers in Thailand. According to the study 35.8% of respondents suffered from depression, 42.2% suffered from anxiety and 17.8% suffered from stress. Important risk factors were low education, being employed in the government sector, young age, discrimination, low social support, and difficulties in accessing health care. Ironically, greater Thai language proficiency and assimilation strategies were associated with less optimal mental health. The study also reinforces the role of other sources of social support, particularly among friends and advocates the implementation of culturally appropriate interventions and greater access to healthcare for Bangladeshi migrants in Thailand [17]. A machine learning model was constructed by a group of researchers to predict the dropping-out of university students in Bangladesh amidst COVID-19. Data on 418 students were included in the study and family income, CGPA, access to the internet, depression, and the effects of COVID-19 were taken into account. Five algorithms were considered, and the highest accuracy of 98% was obtained by Bernoulli Naive Bayes and Random Forest. Results showed that 28.5% of students withdrew, with major reasons cited as health challenges, financial constraints, and depression. The current model serves to enable early detection of vulnerable students and future enhancements include the use of large datasets and application of deep learning approaches [18]. One study on migration has just been released and illustrates the effect in the case of media coverage on Arab and African migrants in Spain. They employed machine learning features such as decision trees, random forests and SVM, to characterise emotional responses and changes in attitudes induced by media portrayal. The study shows that media exposure had a significant impact on the emotions fear, happiness and distrust, which shaped migration attitudes. Random Forest and SVM produced high predictive accuracies (≈92%) [19]. One study suggests an ML based migration prediction system that predicts future migration streams from India to different countries. The proposed approach

combines Random Forest and XGBoost models, and it provides good predictive performances with R^2 scores of 0.81. The project consists of a Streamlit-powered dashboard with a Flask API for predictions on new data. Mitigation planning is possible within a few degrees, and the destiny of billions of people, the systems serves as resource planning tool for policymakers, researchers and government [20]. A further study predicted 5-year ahead irregular migration from Bangladesh to the EU with machine learning and time-series prediction methods. The research has tested five models: ARIMA, XGBoost, Decision Tree, CatBoost, and FFNN, with data collected from Frontex and on-site investigations. The results indicate ARIMA to be the best one outperforming the others by approximately 421 migrants in average of a month. The report cites economic adversity, lack of/limited opportunities and desire for better life style amongst others as principal reasons for migration [21]. In this study, drawing upon random forest models, a group of researchers used two of the largest household surveys to analyse migration data from the southwest region of Bangladesh. Factors associated with migration Community-level determinants of outmigration (Estimates with 95% CI in parentheses) We also examined the determinants of migration. Random Forest had more predictive power as it brought on the lowest average error rate for educational mobility. The results are relevant for policy-making and strategies for climate adjustment [22]. [23] examines the performace of machine learning models in the prediction of regional migration trends in Germany. The authors utilized data from 3,151 municipalities for a period of two and compare four models: linear regression, random forest, XGBoost, and DNN. XGBoost performed best at predicting the LM class for young adults aged 18–24, while family migration was predicted less well, perhaps a consequence of large external shocks such as housing shortages and inflows of refugees. The research illustrates the potential of machine learning to enhance migration prediction.

2.4 Comparative Analysis and Summary

The presented studies highlight the importance of machine learning and big data in advancing our knowledge of mobility dynamics: social, economic or psychological. In the literature, we encounter many machine learning models

such as: Random Forest, XGBoost, Bernoulli Naive Bayes and more, these models are used to predict the Migration trends, student dropouts, mental health issues, and many more. For example, [18], a pupil challenge prediction model in the era of COVID-19 was developed for which Random Forest and Bernoulli Naive Bayes gave acceptable performance. Similarly, [20] estimated migration by means of machine learning models that are applied to a large set of economical and demographic predictors. This was a clear demonstration that machine learning can provide an excellent prediction of migration when the correct SES and psychological factors are considered.Besides predicting migration intentions, the studies also expose how socio- economics and psychological contribute significantly to migrant outcomes. In response, we note that the findings of [16] and [17] addressed the stressors of migrants, particular attention is placed on vulnerability to mental health problems of migrants due to factors as joblessness, discrimination and social network deprivation. Results from these studies show that mental health problems, particularly depression, and anxiety, are common among migrants, and being unemployed is one of the major risk factors. Together with research efforts focused on the physical health of migrants, mental health support and culturally appropriate interventions are imperative to address the mental health vulnerability connected to migration, especially for at-risk groups such as the unemployed and those experiencing discrimination.

Secondly, such findings, such as that reported by [12] and [13] analyse pushpull factors affecting migrant movement, finding that economic need, social networks and housing condition status are critical factors. These studies utilise random forests and other machine learning techniques with large datasets, and allow to identify to migration patterns in some countries (e.g. Australia and Bangladesh). The process is drawn from the big data approach utilised in such studies, and facilitates real-time, granular insights that can better inform policy options in managing migration more effectively. The role of big data in migration studies is also emphasized in [15] that leverage mobile data, online social networks and supermarket retail data to provide more dynamic and advanced migration models. This stands in contrast to traditional migration models based on census information, which are subject to the limitations of timeliness and consistency.

By adding real-time information, the study has shown that predictions of migration flows can be improved with machine learning to help policy makers anticipate and mitigate migration effects.

However, some limitations in the studies exist even if machine learning and big data are promising. In addition, the majority of models still rely on cross-sectional data, limiting inferences on causality over time. In addition, these studies highlight the importance of more representative and balanced datasets in the context of those consistent reports of bias and performative capacity limitation in the case of application to low-resource/underrepresented or geographical specific regions. Ethical issues, namely concerns about the privacy of (origin) personal data, and bias in algorithmic predictions were also identified as important future research challenges. Based on our findings, we have presented an end-to-end approach that leverages machine learning and big data analytics to predict migration trends and identify the factors that drive migration streams. Larger, long-term studies are needed to refine the postulation of migration models as well as to include psychological and socio-economic variables. Ethical consideration and the need for inclusivity and non-bias in the data used, are also important, in the use of data in responsible policy-making that benefits migrant populations.

Table 2.1: Comparative Analysis Table

Study (Author, Year)	Data / Signal Type	Method / Model(s) Applied	Best Reported Accuracy / Metric	Key Contribution
Bakina et al., 2019	Survey of 136 youth in Russia	Mann-Whitney U test	N/A	Satisfaction with city life strongly predicts migration intention.
Islam et al., 2025	BEMS survey (1,754 families, Bangladesh)	Logistic regression-style analysis	N/A	Occupational status key factor for better perceived health among migrants.

Study (Author, Year)	Data / Signal Type	Method / Model(s) Applied	Best Reported Accuracy / Metric	Key Contribution
Sabti & Ramalu, 2024	Survey of 300 doctors, Iraq	Mediation model	N/A	Psychological distress mediates political/social push effects on migration intention.
Hossain et al., 2021	Survey of 1,184 Rohingya refugees	Logistic regression	PTSS prevalence: 46.6% severe	Predisplacement abuse & lack of aid drive post-migration trauma.
Shuva, 2020	Interviews (60) + surveys (205), Canada	Qualitative + descriptive analysis	N/A	Introduces "information sharing fear" & "information intelligence" concepts for settlement.
Sohad et al., 2024	Survey of 300 RMG workers in Bangladesh	SEM + Theory of Planned Behaviour	Model explained 75% variance	Social norms most powerful predictor of migration intention.
Tirado-Espín et al., 2025	Pre- & post-surveys (130 Spanish students)	Decision tree, Random Forest, SVM	RF & SVM ≈ 92% accuracy	Shows media framing + political identity shape migration perceptions.
Siddik et al., 2025	Survey of 1220 adolescents, Bangladesh	PHQ-9 scale analysis	84% migrants depressed, 30% severe	Climate-induced migration has high mental health burden for adolescents.
Ahsan et al., 2011	Gov't data, field surveys, Bangladesh	Case-study analysis	N/A	Urban planning fails to integrate climate migrants →

Study (Author, Year)	Data / Signal Type	Method / Model(s) Applied	Best Reported Accuracy / Metric	Key Contribution
				unplanned urbanization.
Koly et al., 2021	Survey of 935 Dhaka slum migrants	SF-12 HRQoL scale + multivariate regression	PCS=57.40, MCS=60.77 (low)	Poor HRQoL linked to age, gender, joblessness; urban slums worse off.
Williams et al., 2017	Online panel (20,473 youth, 9 EU countries)	Multilevel ordinal models	17% → 20% intend migration	Highlights role of socioeconomics, risk attitudes & prior travel.
Baggen et al., 2023	Open-text survey (3,500+, NT, Australia)	Unsupervised STM topic modelling	N/A	Machine learning extracts nuanced push/pull migration factors.
Best et al., 2022	Household survey (~1,700 HH, Bangladesh)	Random Forest, survival analysis	RF key predictor sets; cross-check	Wealth, resources, household structure predict migration; ML > traditional stats.
Park et al., 2024	Review of 13 studies (EHRs, surveys, social data, genomics)	ML (mostly supervised models)	N/A	ML promising in diagnosing PTSD, depression in migrants, but bias risks remain.
Sîrbu et al., 2021	Big data (mobile, online, retail, census)	Multi-source analytic frameworks	N/A	Proposes integration index for real-time migration analyses.
Biswas et al., 2024	Survey of 395 educated youth	DASS-21 analysis	Depression=54.4%, Anxiety=61.8%, Stress=47.9%	High distress among educated, unemployed

Study (Author, Year)	Data / Signal Type	Method / Model(s) Applied	Best Reported Accuracy / Metric	Key Contribution
	migrants, Kolkata			migrants; policy for jobs needed.
Sultana et al., 2025	Survey of 360 Bangladeshi migrants in Thailand	DASS-21 analysis	Depression=35.8 %, Anxiety=42.2%, Stress=17.8%	Discrimination & healthcare barriers worsen migrant mental health.
Islam, Rahman & Tabassum, 2023	Survey of 418 students, Bangladesh (COVID-19)	ML: Naive Bayes, RF, Decision Tree, LR, XGBoost	Bernoulli NB = 98%	Early prediction of COVID-era student dropout risk.
Reddy et al., 2025	Historical migration + economic indicators, India → World	ML: RF Regressor, XGBoost	RF R^2=0.81	Predict future international migration flows; dashboard & API built.
Islam et al., 2024 (ICAEEE)	Frontex + field data (Bangladesh–EU)	Time-series: ARIMA, XGBoost, CatBoost, FFNN	ARIMA lowest MSE=499.71	Forecasts irregular migration (majority young, low-skilled).
Best et al., 2020	2 household surveys (~2800 HH, Bangladesh)	Random Forest, Logistic Regression, SVM	RF error rate ~14.7% (best)	Key predictors differ by migration type (education vs. environment).
Weber, 2020	German municipality data (2005–15, 3151 municipalities)	Linear Reg, RF, XGBoost, DNN	XGBoost best (R^2 ≈0.52 for youth)	ML predicts youth migration flows; family migration less predictable.

2.4.1 Similar Research

Many ML applications in the migration literature have the same nature as the study presented here. For example, Random Forest and Support Vector Machines

(SVM) are both used extensively to classify people based on their migration intentions with high accuracies. XGBoost has shown to be successful in handling imbalanced migration data and has been especially efficient in identifying the socio-economic and psychological correlates relating to migration decision making. On the other hand, some works have utilized K-Nearest Neighbors (KNN) and Logistic Regression with small-scale datasets, aiming at understandable and transparent result analysis.

Apart from this research in predicting migration intention, ML has been used to explore the health of migrant populations, create mobile applications to offer personalized advice to migrants and analyze large datasets with respect to environmental, political and social factors predisposing to or associated with migration. These applications provide clear evidence that, by using data-driven models to deliver personalized in sights, machine learning has the potential to vastly improve our understanding of migration behavior and deliver actionable recommendations to both citizens and policymakers.

2.5 Gap Analysis

Table 2.2: Gap Analysis Table

Features	Existing Studies	Our Work
Consider socioeconomic, psychological & environmental factors	Yes	Yes
Handle missing values & messy data	No	Yes
Address class imbalance	No	Yes
Use multiple models (LR, DT, KNN, RF, XGBoost)	Yes	Yes
Achieve high accuracy & generalization	No	Partial improvement
Provide explainability (reason behind prediction)	No	Yes (SHAP, LIME)

Features	Existing Studies	Our Work
Offer actionable recommendations	No	Yes (Recommendation layer)
Use a large, diverse, well-balanced dataset	No	No (future need)

2.6 Summary

In summary, this background chapter has situated the topic of research by stressing the importance of migration studies and outlining the growing relevance of machine learning as a methodology for studying migration decisions and the psychological consequences of these. The review of literature has revealed that existing studies have effectively applied big data techniques to predict migration intentions, to elicit words that describe the most important socio-economic and psycho-social stress factors, and to analyze health outcomes in connection to migration. However, there are also several ongoing challenges, such as those associated to data quality, class imbalance, interpretable models and the limited ability to provide specific recommendations to individuals.

This study aims to fill this gap by constructing a supervised model that combines socioeconomic, psychological, and environmental factors to estimate the possibility of intended migration. It aims to offer people effective decision support with personalized recommendations, helping them to be well-informed and ultimately, mitigate the threats on mental health due to migration. The purpose of this approach is to bridge the distance between high-tech computational technologies and their effective use in studies of migration and that scientific and practical value goes in both directions.

Chapter 3

Research Methodology

This chapter describes the research methodology from the research object and requirements. It describes the proposed process, functional and non functional requirements and a data and workflow diagram. The rest of the section provides the details of methodology, the project plan, task distribution and ends with a summary.

3.1 Research Subject and Requirment analysis:

Prediction of consumer behavior in various scenario and prediction of migration related decision in particular using machine learning is the overall focus of this study. Due to not only a number of complex factors affecting migration decisions and migration processes, and the method effect in migration research, which are also a common phenomenon influencing understand different aspects of migration There are more any of desired to predict an individual's decision to migrates; the place to migrates; reasons to live; stress and factors, occupation to way and lifestyle since dissatisfaction there exists a serious issue by many investigators excel in academic area and in the practical word as well.

Traditional migration literature utilizes a rather straightforward and simple political, social and economic concept, but does not capture the complexity of individual decision making. This work resolves these gaps and develops a model that incorporates machine learning-based frameworks where lifestyle (L), anthropometric(A) and personal behavior(P) information together with traditional social economic indicators to construct a simplified, scalable and accessible model to predict the migration potential. In the paper, performance analysis of a number of machine learning models e.g., Logistic Regression, Decision Tree, K-nearest neighbour (KNN), Random Forest, XGBoost etc are compared and analyzed, following which Random forest-based model for performance optimization is trained, after the required preprocessing, feature engineering, and tuning of hyperparameters.

This work differs from pure classification since it utilized interpretability techniques such as feature importance analysis, SHAP/LIME. Such are methods which explain some prediction of a model, and inform why did the model act the way it did thus allowing to make the decision the System does transparent and predictable.

The results of this research also have practical contributions to the design of the recommendation engine that presents actionable recommendations to the consumers. This system recommends some solutions more closely related to the migrants' problems based on the trained models, such as destinations, motivations, and psychological and social problems. The purpose of the project is to develop a decision-support tool not only for the forecast of migration flow but for its mitigation in a due manner.

One area of focus here is at the intersection of machine learning, behavioral science and applied decision making. Our style strikes the right balance between model complexity, accurate prediction, and interpretation by experts and non-experts. By informing the shape of research in predictive modelling in academic settings and reaction research in the field, the research has the potential to guide the development of more effective early screening, personalised intervention, and decision support tools for migrants and those involved in managing them.

3.1.1 Overview

The data were primarily collected using a structured Google Form. The form link was distributed among diverse participants across various settings, including university campuses, residential halls, urban areas, rural regions, and public streets. In addition to online responses, data were also gathered through in-person interactions and verbal communication with individuals. The information obtained through these conversations was subsequently entered into the Google Form to maintain consistency in data recording. Specific locations in Bangladesh where data collection was conducted include cities such as Kaliganj, Jhenidah, Kushtia,and Dhaka. Participants were also approached in various villages within these cities, as well as at Daffodil International University and its associated residential halls (23.875816, 90.323511). This broad geographic coverage

contributed to the richness and variability of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2614 entries, 0 to 2613
Data columns (total 21 columns):
 #   Column                                                                                                                                  Non-Null Count  Dtype
---  ------                                                                                                                                  --------------  -----
 0   What is your age?                                                                                                                       2614 non-null   object
 1   What is your gender?                                                                                                                    2614 non-null   object
 2   What is your occupation?                                                                                                                2614 non-null   object
 3   Does any member of your family or a close acquaintance live abroad?                                                                     2614 non-null   object
 4   Have you decided to move abroad?                                                                                                        2614 non-null   object
 5   How long do you plan to stay abroad?                                                                                                    2614 non-null   object
 6   Are you aware of any specific migration programs or scholarships available for students or workers abroad?                              2614 non-null   object
 7   What role do social media and online success stories play in influencing your decision to move abroad?                                  2614 non-null   object
 8   Who or what influences your decision to move abroad? (Multiple options can be selected)                                                 2614 non-null   object
 9   Which countries are you most interested in moving to?                                                                                   2614 non-null   object
 10  How much psychological stress or anxiety have you experienced while considering moving abroad?                                         2614 non-null   object
 11  What is your primary goal for considering moving abroad? (Multiple options can be selected)                                            2614 non-null   object
 12  What kind of psychological stress have you experienced while planning to move abroad? (Multiple options can be selected)               2614 non-null   object
 13  How do you cope with the stress of planning to move abroad?                                                                             2614 non-null   object
 14  Do you think migration trends have increased among Bangladeshi youth in the last 5 years?                                               2614 non-null   object
 15  What do you think is the biggest barrier to migrating abroad? (Multiple options can be selected)                                        2614 non-null   object
 16  Do you think you will return to Bangladesh in the future after living abroad?                                                           2614 non-null   object
 17  Do you think your decision to move abroad will impact your family or community in Bangladesh?                                           2614 non-null   object
 18  Do you feel that the government of Bangladesh offers enough support for youth planning to migrate abroad?                              2614 non-null   object
 19  Do you feel a sense of responsibility towards your family while planning to move abroad?                                                2614 non-null   object
 20  How likely are you to recommend going abroad to others in your age group? (Rate on a scale of 1 to 5, where 1 is very unlikely and 5 is very likely).  2614 non-null   int64
dtypes: int64(1), object(20)
```

Figure 3.1: Data Collection Question

3.1.2 Proposed Methodology

The process of working with migration data is laid out in steps you can follow through in this notebook. First, the dataset "Final Migration Data - Form Responses 1. csv" was read to pandas DataFrame. The details of the data were explored with several functions which include df. head(), df. isnull(). sum(), df. info(), and df. form to give the shape, so that we can see what it looks like and locate missing values. To show the pattern of missing values a heatmap was created using Seaborn. The missing values were subsequently dealt with by replacing the categorical columns with its mode (most frequent value) and numerical columns with median (where none of the numerical values were missing). Feature engineering used custom encoded mappings and one hot encoding for multiple columns that had non-numeric value to numeric conversions. This was the case for the features age, gender, occupation, questions about migration, and psychological issues. The data was divided into a training subset and a testing subset where the dependent variable was the international migration decision and the independent set was represented by random variables without the dependent value. Some of the classification models such as Random Forest, Decision Tree, K-Nearest Neighbors, Extra Trees, Bagging, XGBoost Logistic Regression, and Support Vector Classifier were applied and trained using the provided dataset, as well as tested to check their accuracy, confusion matrix

and classification report statistics. Following the selection of the 20 top features, feature selection was performed with SelectKBest, and the models were retrained (Random Forest, Extra trees, Support Vector Classifier, Bagging, LightGBM). Some of the best performing models were also hyperparameter tuned using GridSearchCV and RandomizedSearchCV, and Optuna was employed to tune Random Forest, SVM and KNN. In addition, to boost performance and get a more accurate estimate, cross-validation was carried out on multiple models. Ensemble techniques, such as voting and stacking classifiers, were applied to combine multiple models for enhanced accuracy. Further feature engineering was done which included polynomial features, natural-logarithm features in order to strengthen the selected features, with model retraining afterward. SMOTE was used to counteract the class imbalance and to manage class imbalance, class weights, a technique to manage imbalance classes, were used for SVM, KNN, Random Forest and whereas other models. Having tested scaling using log of eigenvalue and variance has scaled, Cross-validation was performed using StratifiedKFold, to ensure balanced classes. The last step was the development of an interactive python script which received as input user replies and predicted potential responses from the previous models. This approach guaranteed a full and comprehensive analysis and modeling of migration data.

3.1.3 Functional & Nonfunctional Requirment

Functional Requirements represent the specific actions, tasks or process a system is going to perform. These specifications concentrate on what is expected of the system to fulfil the needs of the application. For instance, in a web application, functional requirements can be these: need the feature to create, edit, delete user profiles; messages need to be received in real-time. These are fundamental features of the system that directly support its purpose and the needs of users and stakeholders.

Non-Functional Requirements on the other hand describe the way in which the system should perform its functions and under what conditions. They focus on the quality, performance, scalability, and reliability rather than specific features of

the system. For example, a non-functional requirement might state that the system shall support 1,000 simultaneous users without degradation of system performance, or application shall load within three seconds for a satisfactory user experience. These demands are crucial to ensure the system's efficiency, particularly when replicating or deploying into other platforms.

3.1.4 Data & Workflow Diagram

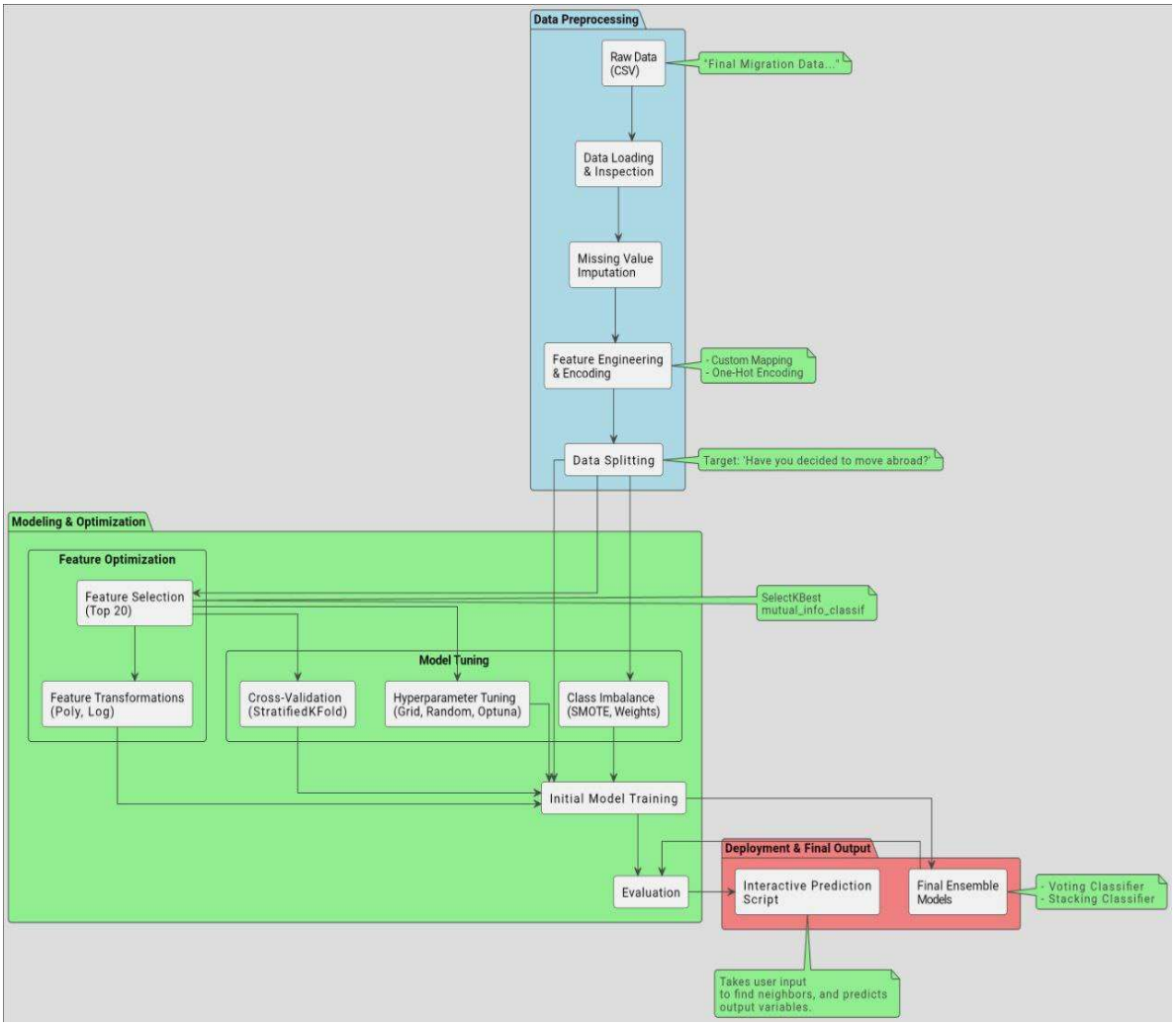

Figure 3.2: Data Flow Diagram

3.2 Detailed Methodology and Design

Multiple machine learning classifiers are utilised in this research for the task of Migration dicision prediction using important features and related factors. The

models for evaluation are LightGBM, RandomForestClassifier, LogisticRegression, SVC, DecisionTreeClassifier, KNeighborsClassifier, XGBoost Classifier and MLPClassifier. We evaluate the performance of these models using standard classification evaluation measure. These statistics are Accuracy, Precision, Recall, F1-score and Confusion Matrix. The Confusion Matrix consists of TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) and helps us understand the ability of the models to classify instances in a migration data context across usual and rare condition categories, including class imbalance we solved through balancing using techniques we mentioned such as SMOTE. Hyperparameter tuning techniques GridSearchCV, and RandomizedSearchCV are also applied to enhance the model.

Random Forest Classifier (RF): The Random Forest classifier (RF), an ensemble machine learning method that constructs many decision trees based on random samples of the data and features and provides the final prediction output by majority voting rule [24]. It has been used on large-scale migration studies, since it can explain the effect of different social-economic, demographic and psychological factors simultaneously, even when dealing with complex, and cross-classified parameters [22]; [23]. RF has been demonstrated to outperform traditional algorithms such as logistic regression, decision tree and support vector machines [20], [12], in predictive accuracy and generalization performance on a relatively broad range of datasets, particularly for data involving nonlinear relationships and high dimension. The RF algorithm has been used in migration prediction studies to predict individuals with migratory intention and identify their most important predictors of migratory choice [22]; [23].

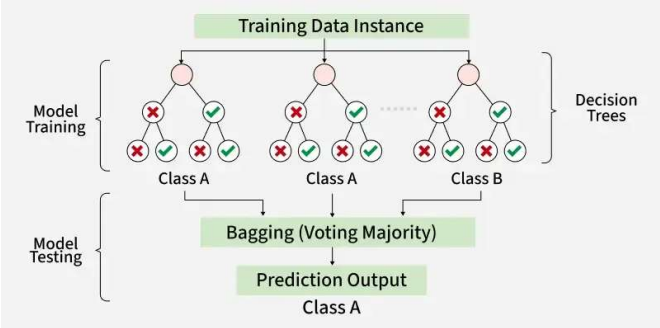

Figure 3.3: Workflow representation of RF [24]

XGBoost: XGBoost Extreme Gradient Boosting (XGBoost) is a scalable and efficient gradient boosting system that forms an ensemble of weak classifiers in order to improve the prediction performance. Unlike traditional ensemble models, XGBoost operates in an iterative manner, and each tree learns from the mistakes of its ancestors using gradient descent optimization, and regularization techniques are employed to handle overfitting [20];[12]. In the migration literature, XGBoost has been used to analyse high-dimensional multivariate datasets that encompass a range of socio-economic, demographic and psychological dimensions and in predicting migration patterns and individual migration intentions [22];[23].

Figure 3.4 Open image in new window depicts the training and testing of the XGBoost model. Hyperparameters and optimal configuration are chosen by means of grid search during the training process. Thousands of gradient-boosted models (GBMs) are subsequently trained on random samples of the dataset, and the ranking of features by importance is then averaged across the models. At the testing phase, the ensemble model is used to make predictions on the unseen sample testing data and the final migration intention classification result is obtained based on averaging the predictions of all the single models [25]. It is the combination of these strategies in iterative optimization and ensemble averaging that makes XGBoost a very powerful method for processing noisy, imbalanced, and high-dimensional datasets, which can be of universal application when trying to model migration decisions that arise from quite a few interdependent factors.[1];[20]

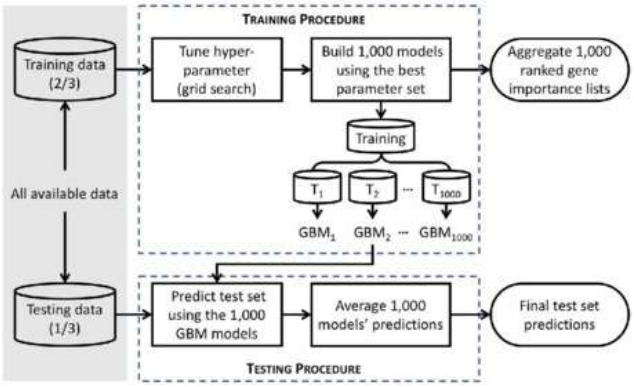

Figure 3.4: XGBoost training and testing workflow [25]

Logistic Regression (LR): It is one of the most frequently used supervised classification algorithms, which is developed for binary classification, where the dependent variable takes the value of 1 if the class predicates true and 0 otherwise (e.g., the probability of having migration intention ("Yes" or "No")). Unlike linear regression which predicts continuous values, where the output is bounded between 0 and 1, the logistic (sigmoid) function is employed in LR to map the output, i.e. probability of belonging to each class [24].

The action mechanism of LR is depicted in (Fig.3.5). The model accepts multiple input features (x1,x2,...,xn), multiplies them with some weights (w1,w2,...,wn), and adds a bias term (b). Finally the Weighted Sum is fed through the sigmoid activation function which squashes the result in the range of a probability. According to a pre-specified cut-off (usually at 0.5), the result is categorised into one of the two outcomes e.g. predicting whether an individual will emigrate or not [24]. For migration prediction studies, LR is commonly used because of its simplicity, interpretability, and computational time [13]. However, also the LR assumes that the independent covariates have a linear and an additive effect on the log-odds of the dependent variable, and may not capture complex non-linear interactions between the socio-economic, psychological, and demographic factors [20] that affect the decisions to migrate. In this research, we also evaluated LR with other machine learning models such as Random Forest, XGBoost, Support Vector Machines, and Bernoulli Naive Bayes. With respect to predictive performance, although LR gave a baseline prediction, its performance was lower than that of RF and XGBoost, and as a result it would be more appropriate if it were used as a benchmark model rather than the final one.

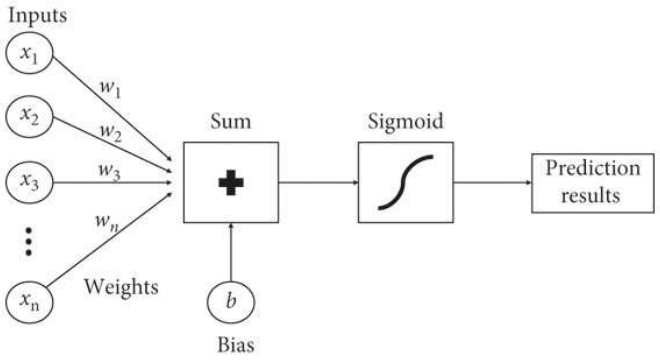

Figure 3.5: Visual representation of LR [24]

K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) is a type of instance-based K-Nearest Neighbors (KNN) is a simple and effective method belonging to instance-based learning methods, applicable to classification and regression problems. In KNN, the prediction of a new data point is done by finding the k-nearest neighbors ('nearest' as in having the minimum distance in feature space) of that data point. The class of the new instance is based on majority of new instances k-nearest neighbor, and usually the k-Nearest neighbor model uses Euclidean distance as the distance metric though it can consider other distance metrics [27]. (Fig:3.6) shows how KNN operates. A new data point is added to the figure, as depicted by the blue point and the k-nearest neighbors of the data point are encircled. After considering the class instance (e.g., Category A 6md Category B), the new-instance is positioned in category with the highest counts - particularly, it is mapped to the category with the largest counts [27]. The method makes nearest-neighbour-style decisions such that it is especially well-suited to cases where the feature relationship is non-linear and the datapoints are clustered according to categories.In particular, the KNN performs great on nonlinear classification problems and is useful when there are a large number of overlapping categories in the feature space [23]. KNN is conceptually" simple", but it can take longer to classify items if the data set is large, as it estimates the distance between a point to all other points in data set. In this paper, KNN is compared with other machine learning models such as Random Forest and XGBoost, and it gave a stable base line model of comparison. But its accuracy was lower than Random Forest and thus was not favored for final prediction.

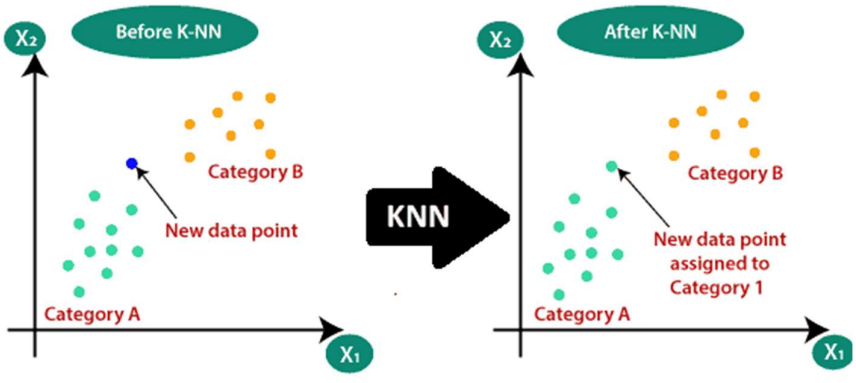

Figure 3.6: Visual representation of KNN [27]

Support Vector Machine (SVM): It is known as a very strong supervised learning algorithm applied on classification tasks, particularly on highdimensional feature spaces. SVM forms a hyperplane to separate the data in each class and maximize the margin between them. The algorithm searches for the best decision boundary for minimizing classification error [26]. Second, SVM can be applied on the non-linear problems such as the RBF kernel transforming the features into high dimension will make the classes separable [20]. In predicting migration intentions, SVM has been successfully trained to classify people by complex features, such as demographic and psychological factors. Although Random Forest and XGBoost have better predictive performances for this analysis, the strength of SVM is good generalization for small datasets with a non-linear feature relationship [24]. SVM with Gaussian kernel enables the instrumentation of latent patterns and dependencies between the data points which are almost invisible for the earlier unsophisticated linear classifiers like Logistic regression, KNN etc., and that is quite beneficial when it comes to the migration intentions [23].

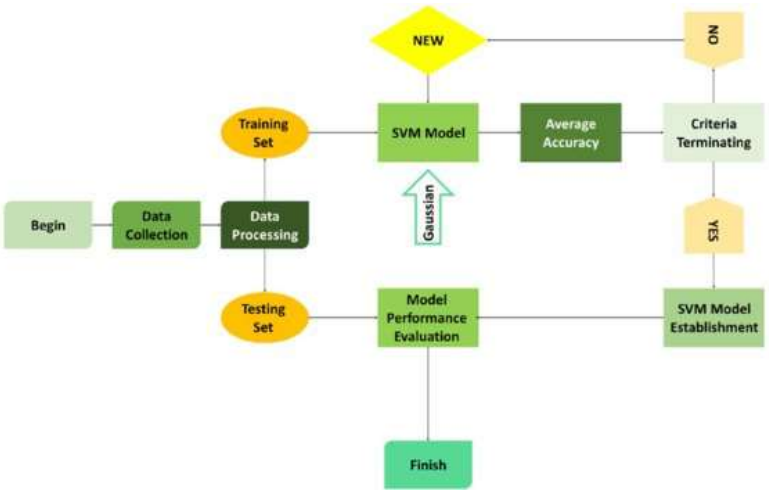

Figure 3.7: SVM training and testing workflow [26]

Kernel Ridge Regression (KRR): Kernel Ridge Regression (KRR) is a powerful extension of linear regression which uses the kernel trick so that the model can handle non-linear associations between features. KRR is based on the ridge regression with regularization to prevent overfitting and the kernel trick that maps input samples to a higher dimension where they become linearly separable.

The kernel (e.g. Gaussian (RBF)) allows KRR to be able to mathematically compute the inner product in the transformed feature space without explicitly having transformed the data [27]. The kernel trick is illustrated in KRR in Fig 3.8 On the left, the two types of data (red squares and green circles) are not separable by a line in the original feature space. The right hand side of it projects the data into a higher-dimensional space based on a kernel, where it then has a clearly separatable decision surface, as is obvious[27]. The potential that KRR has to map data into a higher-dimension space allows it to model complex non-linear structures, and it may contribute to being suitable for predicting migration intentions in cases where the influence of socio-economic, psychological and demographic factors manifest in a non-linear fashion.

In this study, we compared KRR with other machine learning models including Random Forest and XGBoost for the task of migration decision prediction. Although XGBoost and Random Forest scored better, KRR showed ability to capture the more complex relationship between features. This makes KRR a promising choice, particularly in models that require both non-linear j non-linear mapping and regularization for transferring between different data distributions [23];[24].

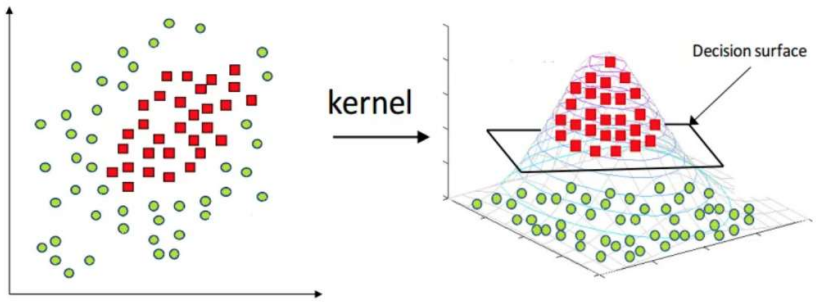

Figure 3.8: Kernel Ridge Regression (KRR) workflow [27]

Extra Trees Classifier: The Extra Trees Classifier (ETC)/ Extremely randomized trees is an ensemble machine learning method that constructs multiple decision trees independently and combines them to create more accurate estimates. Like Random Forest, ETC grows multiple trees, but it differs from Random Forest in choosing split points at each node and a subset of features for building a tree in a

random way [28] .This technique enhances diversity among the individual trees, leading to a lower variance and the model becoming more noise-resistant.

Figure 3.9 Working Principle of ETC. It begins with a random subsample of the data and features for decision trees. Predict: Predict a class (e.g., Class A or Class B) for a test instance using the trees. Once all of the trees have made predictions, the majority voting rule is used to make a final class assignment for the data point, which minimizes over-fitting that is often associated with decision trees per se [28].

In migration prediction research, the Extra Trees Classifier is a candidate method for categorization of people into different segments by complicated socio-economic, demographic, and psychological features [28]. It is one of the best classifiers for high-dimensional feature space and is capable of capturing non-linear relationships as well, which is why it serves as a very good candidate when you want to predict migration intention based on diverse predictors such as income, education and health status. The ETC model was compared with Random Forest, XGBoost and others in our analysis, and shows similar predictive performance with much less computation time owing to its inherent randomness.

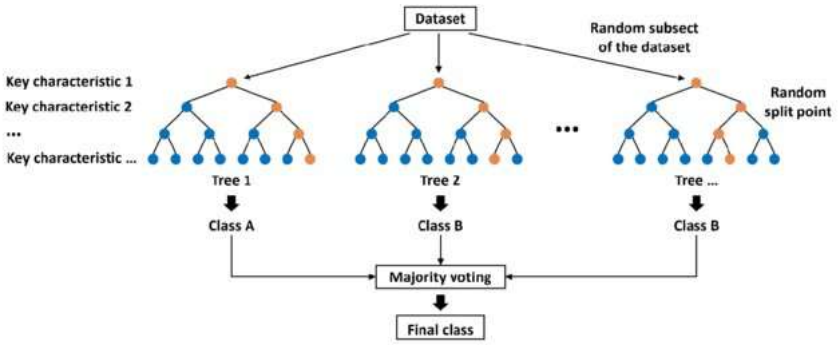

Figure 3.9: Extra Trees Classifier workflow [28]

Dicision Tree: An Overview on Decision Trees Decision Trees (DT) are a commonly used supervised machine learning technique for classification and regression tasks. The algorithm recursively divides the data into partitions using input features and creates the tree-type structure where data can be sorted into finite classes. Each internal node is a decision on a feature and each external node is an outcome or a class label [22]. Decision Trees are especially helpful when the features have non-linear and hierarchically structured relationships with the

target variable. Figure 3.10 shows a decision tree in action. The algorithm first splits the data at Decision #1 using a single 'key' feature, and split further into multiple decision points, such as Decision #2, Decision #3, and so on. At every internal node, data is partitioned into two branches, depending on some Yes/No question, until an outcome is predicted at the leaves [27]. In migration forecasting, Decision Trees can be used for categorizing individuals according to social-economic, demographic, and psychological characteristics, and thereby to forecast whether an individual is going to migrate abroad (Y) or not (N): They (decision trees) can be visualised and processed in an intuitive manner which is valuable when stakeholders (e.g. functional analysts, doctors, process engineers etc.) want to understand how data are explained by the model itself [23]. But overfitting is a common problem for decision trees and especially for deep trees. To overcome this, methods such as pruning or utilizing ensemble methods such as Random Forest can enhance model accuracy and decrease variance.

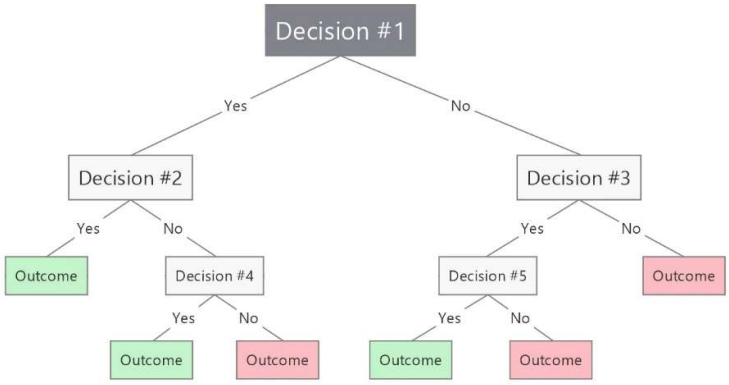

Figure 3.10: Decision Tree classification workflow [29]

LightGBM: LightGBM (Light Gradient Boosting Machine) is a highly efficient and scalable gradient boosting framework that is used for classification and regression tasks. It is an implementation of Gradient Boosting Machines (GBM) that uses decision trees to make predictions, but it optimizes the training process by leveraging histogram-based techniques for faster computation and reduced memory usage [26]. LightGBM also utilizes a leaf-wise tree growth strategy, which tends to improve accuracy by selecting the best split points at each node, as opposed to the traditional level-wise growth strategy used by other gradient

boosting algorithms. Figure 11 shows the workflow of LightGBM, where input variables are fed into multiple decision trees (Tree 1, Tree 2, ..., Tree N). Each tree produces a function f1(x), f2(x), …, and the final prediction is derived from the sum of the individual tree outputs. This ensemble approach helps capture complex relationships in data while maintaining computational efficiency [26]. In the context of migration prediction, LightGBM is particularly useful for modeling large datasets with complex interactions between socio-economic and psychological features. The algorithm's speed and scalability make it suitable for real-time predictions and high-dimensional data typically encountered in migration research. In this study, LightGBM was evaluated alongside Random Forest and XGBoost. While LightGBM demonstrated competitive accuracy, its efficiency and speed in handling large datasets made it a compelling choice for predicting migration intention based on multiple influencing factors [28]. LightGBM's ability to handle imbalanced data through advanced regularization techniques also made it well-suited for the uneven class distributions often found in migration datasets.

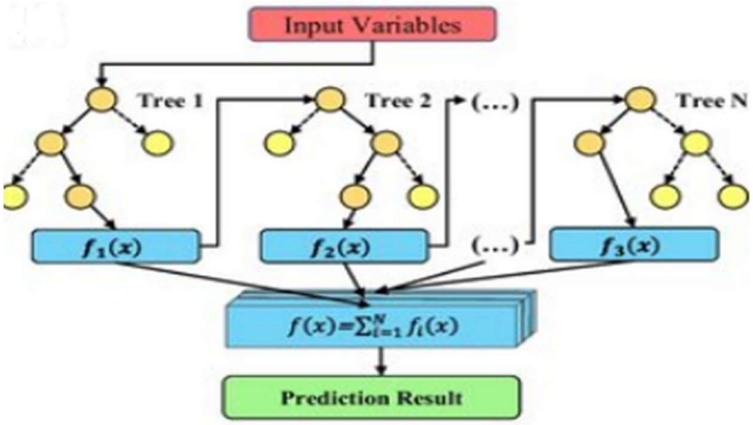

Figure 3.11: Schematic representation of LightGBM [26]

Accuracy:

For the classifier model, the most crucial one would be the performance and the accuracy is the most straightforward indicator to evaluate the performance; it calculates out the proportion (percentage) of correctly-classified samples in general data. In this work, Accuracy The time at which the testing becomes

accurate to distinguish the youth abroad dicision related data points in their respective decision condition classes. It is calculated as (TP + TN)/total, where TP, TN, FP, and FN are True Positive, True Negative, False Positive, and False Negative, respectively. Using this formulation, Accuracy gives an overall quantification of the model's accuracy in all possible dice outcomes. This variable was very important when assessing the usefulness of the different models utilized in this study, providing an idea of the general predictive capability of each model and its capacity to accurately classify youth decision to abroad [30].

$$\textbf{Accuracy} = \frac{(TP+T\ \ )}{(TP+TN+FP+F\ \ )}$$

Precision:

Precision measures the accuracy of the positive predictions made by a classification model. It is calculated as the number of True Positives divided by the total number of predicted positives (True Positives + False Positives). High precision indicates a low rate of false positives. Mathematically, it's this equation [30]:

$$\textbf{Precision} = \frac{TP}{(TP + FP)}$$

Recall:

Recall, also referred to as Sensitivity or True Positive Rate, is the most important metric for classification as, it tells us what proportion of actual positives were correctly identified by the model. For this analysis, Recall of a Youth Migration Decision (Y/N) is the percentage of people who really have the condition -- and the model correctly identifies them. It is calculated using the formula [30]:

$$\textbf{Recall} = \frac{TP}{(TP+FN)}$$

Where TP (True Positives) is how many of the positives were correctly classified, and FN (False Negatives) is how many positives were missed by the model. A high Recall means that the model is good at identifying the majority of the

individuals of the category, and significantly reduces the number of false negatives. This is especially important in a present domain, the reason being that failing to detect the problem may result in (very) serious consequences, and, as such, one should try to detect as many positives as possible [30].

F-1 Score:

We use the F1 score, a recall-precision categorization accuracy metric. As F1-score is the harmonic mean of Precision and Recall, it provides the full picture. Precision and Recall is best when its same. Equation is [30]:

$$\mathbf{F-1\ Score} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

The percent of people who don't have the ailment who get a negative test result is called specificity. A test that is highly specific is good at "ruling out" most people who do not have the disease. Due to this fact, a positive result on a very accurate test can definitively exclude the disease in an individual. The equation is given below [30]:

$$\mathbf{Specificity}\ = \frac{TN}{(TN + FP)}$$

Yet the model can't be made to look like the training set data, otherwise, the model will not be able to generalize. The performance of an algorithm can be evaluated through such a table format known as a confusion matrix (CM). CM is employed to demonstrate important predictive parameters such recollection, specificity, accuracy, and precision. Confusion matrices are helpful since they provide a direct measure of TP, FP, TN, and FN. These sections answer the research questions of this study based on the research questions [30].

3.3 Project Plan

Project Plan is a detailed document that describes how a project will be executed, monitored, and controlled from start to finish. It is a detailed schedule and plan that project team can use to complete the work in a timely manner, meet the scope and stay on budget. One of the first things the plan does is to establish what specifically the project will do – but more importantly, what it won't do – which

allows you to clearly define the goals, the deliverables, and the inclusions and exclusions in the project. In parallel there is created a timeline identifying when certain work, milestones and deadlines are to be projected to be in strong order and maintain continuity of the project on schedule. The plan also lists what is needed to complete the support, like who is doing it, what types of tools are being used and how much money is going to be spent. Risk coverage is another issue risks are identified, problems are anticipated, solutions are suggested. Your job, tasks and responsibilities are all laid out with great detail in order for everyone to be held accountable and avoid misunderstandings. The program plan includes a quality assurance plan that describes the approach to reviewing deliverables and performing tests to verify that they meet the required standards of quality. Last but not least, a communication plan is developed to lay down how team members will report and share updates and information, in order to make everything transparent and aligned between the parties during the implementation process. Having a good project plan is indispensible for managing a team successfully and ensure that the project has a valuable outcome. So the proper project plan is mandatory.

3.4 Task Allocation

This table depicts the timeline of the principal activities in each period of the project, from week 12 to week 48.

Table 3.1: Task Allocation Table

Tasks	Weeks																		
	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48
Data collection phase																			
Preprocess all the data																			
Model training																			
Create a demo application.																			

3.5 Summary

This chapter describes the thorough methodological approach adopted to develop a machine learning model to predict migration intentions and psychological distress of Bangladeshi youth. The approach relies mainly on a survey-based data of 2,614 respondents covering most urban and rural spectra and performed through both structured Google Forms and face to face interviews.

The suggested technical pipeline is performed using a Python environment in Google Colaboratory and based on the commonly used set of data science libraries including Scikit-learn, Pandas, and NumPy. Some important data pre-processing steps are missing value handling, feature engineering, and class imbalance which is treated using the Synthetic Minority Over-sampling Technique (SMOTE).

The description of the approach is featured by a comparison of several supervised learning models, such as the Random Forest, XGBoost, Support Vector Machines (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) and the rationale for each model is explained based on successful prior migration research. Model performance was rigorously validated through standard performance metrics such as accuracy, precision, recall, F1-score, and cross-validation to confirm the robustness of the findings.

Beyond mere prediction, the framework incorporates two major innovations: (i) the adoption of explainable AI (XAI) methodologies such as SHAP and LIME to ensure model interpretability, and (ii) the establishment of a personalization recommendation module to convert predictive learnings into actionable, user-focused recommendations. The chapter ends with a project plan and work distribution, which gives a roadmap of the research from collecting data to the deployment phase.

Chapter 4

Implementation and Results

A description of the experimentation setup is introduced in this chapter, and the experimental result is analyzed in more detail. It presents the results, recommendations their implications, and relates them to the research purpose. Finally, we summarise the main results from the section.

4.1 Experimental Setup

The study's experimental setting was devised to capture youth migration decisions and psychological distress in Bangladesh. The objective of the study was to explore what is related with the intention to migrate and the psychological pain migrating meant. It was achieved through a mixture of collection strategies, machine learning methods and assessment measures. Methodology Structured questionnaires in Google Forms were used to collect data from 2,614 respondents and later these data were also enriched by means of face-to-face interviews that provided richer insights. They were of various socioeconomic status levels and included university students, services holders, members of communities; they reside in urban and rural locations throughout Bangladesh. Data Preprocessing The raw data that is collected is pre-processed by treating the missing values, categorical variables, scaling of range of values etc. This helped to prepare the data for analysis and model. Different machine learning models including LR (Logistic Regression), RF (Random Forest), XGBoost and SVM were built for the prediction of the migration intent, stress classification and motivation grouping. We also implemented some models using Python(scikit-learn).

Evaluation: The performance of the models was assessed using accuracy, precision, recall, F1-score and ROC-AUC. Cross validation was used in order to cross-validate the models and to prevent the risk of overfitting.

This experimental research provides a solid theoretical model to study migration decision making among Bangladeshi youths and the psychological conditions under which migration occurs.

4.2 Experimental Results & Analysis

Result of Experiment: ML Models

This experiment involved testing numerous classification models to evaluate their suitability for predicting migration intentions based on socio-economic and psychological data. Accuracy is the main measure used to evaluate model performance, expressing the proportion of correct predictions made by the model on the dataset.

Table 4.1: Accuracy for Classification of Individual ML models to detect migration decisions (Y/N).

Model	Initial	Top 20 Features	Polynomial Features	Log Transformed Features	Class Weights/SMOTE	Hyperparameter Tuned (GridSearchCV)	Hyperparameter Tuned (RandomizedSearchCV)	Hyperparameter Tuned (Optuna)	Ensemble (Voting)	Ensemble (Stacking)
Random Forest	81%	79%	80%	81%	80%	80%	81%	81%	79%	79%
Decision Tree	75%	-	-	-	-	-	-	-	-	-
KNN	74%	-	78%	76%	76%	77%	79%	79%	79%	79%
Extra Trees	80%	79%	-	-	-	-	-	79%	79%	79%
Bagging (Decision Tree)	79%	77%	-	-	-	-	-	-	-	-
XGBoost	78%	-	78%	-	-	80%	-	78%	78%	78%
Logistic Regression	77%	-	-	-	-	-	-	78%	-	-
SVM (RBF Kernel)	81%	81%	81%	81%	80%	81%	81%	81%	79%	79%
LightGBM	-	76%	-	-	79%	-	-	-	-	-
CatBoost	-	-	-	-	81%	-	-	-	78%	-

From Table 4.1, Results from the model accuracy study provide interesting insights into the performance of these machine learning models over different configurations. The Random Forest model is consistently strong and it has high initial accuracy (0.81) and reliability over different settings. Even after feature selection, hyperparameter tuning performed through GridSearchCV and Optuna and ensembling the model, the accuracy still comes out to be on the scale of 0.79-0.81, which shows its resistant to handle noise in dataset. However the Ensemble (Voting) and Ensemble (Stacking) didn't add a lot of value, indicating that the pattern of the data which is needed, is largely captured by the Random Forest model.

The Decision Tree model, however, has a relatively low basaline accuracy of 0.75, and without tuning would perform poor. This suggests that decision tree may simply be too simple for the nuances in the data set and that we need more sophisticated models or more feature engineering. The Logistic Regression model also gives an overall accuracy of 0.77 and non-configuration improves upon this value, indicating that said model has a hard time fitting the non-linear relationships in the data.

A significant increase of K-Nearest Neighbors (KNN) above 0.74 and climbing to 0.79 is observed after hyperparameter tuning with SMOTE and RandomizedSearchCV. This implies that KNN has high capability of response to parameter tuning and it can achieve performance in a level as uniform learning in the model of KNN in this work. Random Forest The model has an initial Acc = 0.82 and nothing much interesting as the Acc remain stable through the ensemble as the rest also remains the same Extra Trees Similar to Random Forest, it has a starting Acc of 0.80 (This model is known for its ability and stability to be a strong performing model across different settings) and everything that needs to be said about the Acc holds here as it did for the Random Forest.

The Bagging (Decision Tree) has roughly the same performance as Decision Tree model with a base accuracy around 0.79 and I cannot further increase the accuracy with different configurations, which possibly indicates that the bagging operation is not able to significantly lower the variance of the decision tree in this issue. XGBoost shows a strong result right from the beginning (accuracy = 0.78, which goes up to 0.80 after hyperparameter tuning). This means XGBoost does benefits from fine-tuning, which can better capture intricacies in the data, but still generates smaller ROC compared to Random Forest and Extra Trees.

SVM (RBF Kernel) uniformly achieves good results (accuracy 0.81 for all the configurations), and it is reasonable to infer that this method copes rather well with the dataset. It has consistently strong performance with fairly little tuning, and hence one of the most stable architectures. LightGBM presents a promising 0.79 accuracy with SMOTE, but is of intermediate behavior across other configurations (suggesting that it may require more exploration to yield better results).

Lastly, CatBoost achieves highest accuracy of 0.81 with the SMOTE setting, indicating its potential in this task. Yet, its effectiveness with ensembles is not known, and it may require additional tuning to be efficient in such setups. In general, models such as Random Forest, SVM (RBF Kernels), and Extra Trees yield strong and consistent results, and models such as KNN, XGBoost, and CatBoost benefit from hyper parameter tuning. The ensembles were across the board not significantly better than models in many cases indicating that the base models were performing well and the additional complexity was not necessarily benefiting from it for the dataset.

Table 4.2: Precision, Recall, F1-Score, and Support (n) for original ML-based algorithms.

Model	Metric	Initial	Top 20 Features	Polynomial Features	Log Transformed Features	Class Weights/SMOTE	Hyperparameter Tuned (GridSearchCV)	Hyperparameter Tuned (RandomizedSearchCV)	Hyperparameter Tuned (Optuna)	Ensemble (Voting)	Ensemble (Stacking)
Random Forest	Accuracy	81%	79%	80%	81%	80%	80%	81%	81%	79%	79%
	Precision	85%	81%	83%	84%	83%	84%	85%	-	81%	81%
	Recall	81%	79%	80%	81%	80%	80%	81%	-	79%	79%
	F1-score	80%	78%	79%	80%	79%	80%	80%	-	78%	79%
Decision Tree	Accuracy	75%	-	-	-	-	-	-	-	-	-
	Precision	75%	-	-	-	-	-	-	-	-	-
	Recall	75%	-	-	-	-	-	-	-	-	-
	F1-score	75%	-	-	-	-	-	-	-	-	-
KNN	Accuracy	74%	-	78%	76%	76%	77%	79%	79%	79%	79%
	Precision	75%	-	79%	77%	77%	78%	80%	-	81%	81%
	Recall	74%	-	78%	76%	76%	77%	79%	-	79%	79%
	F1-score	74%	-	78%	76%	76%	77%	78%	-	78%	79%
Extra Trees	Accuracy	80%	79%	-	-	-	-	-	79%	79%	79%
	Precision	83%	81%	-	-	-	-	-	81%	81%	81%
	Recall	80%	78%	-	-	-	-	-	79%	79%	79%
	F1-score	79%	78%	-	-	-	-	-	78%	79%	79%
Bagging (Decision Tree)	Accuracy	79%	77%	-	-	-	-	-	-	-	-
	Precision	82%	78%	-	-	-	-	-	-	-	-
	Recall	79%	77%	-	-	-	-	-	-	-	-
	F1-score	79%	76%	-	-	-	-	-	-	-	-
XGBoost	Accuracy	78%	-	78%	-	-	80%	-	78%	78%	78%
	Precision	78%	-	79%	-	-	83%	-	80%	78%	78%
	Recall	78%	-	77%	-	-	80%	-	78%	77%	77%
	F1-score	77%	-	77%	-	-	80%	-	78%	77%	77%
Logistic Regression	Accuracy	77%	-	-	-	-	-	-	78%	-	-

Model	Metric	Initial	Top 20 Features	Polynomial Features	Log Transformed Features	Class Weights/SMOTE	Hyperparameter Tuned (GridSearchCV)	Hyperparameter Tuned (RandomizedSearchCV)	Hyperparameter Tuned (Optuna)	Ensemble (Voting)	Ensemble (Stacking)
	Precision	77%	-	-	-	-	-	-	80%	-	-
	Recall	77%	-	-	-	-	-	-	78%	-	-
	F1-score	77%	-	-	-	-	-	-	78%	-	-
SVM (RBF Kernel)	Accuracy	81%	81%	81%	81%	80%	81%	81%	81%	79%	79%
	Precision	84%	85%	84%	84%	82%	85%	84%	-	81%	81%
	Recall	81%	81%	80%	81%	80%	81%	81%	-	79%	79%
	F1-score	81%	80%	80%	80%	79%	80%	80%	-	78%	79%
LightGBM	Accuracy	-	76%	-	-	79%	-	-	-	-	-
	Precision	-	77%	-	-	81%	-	-	-	-	-
	Recall	-	76%	-	-	79%	-	-	-	-	-
	F1-score	-	76%	-	-	79%	-	-	-	-	-
CatBoost	Accuracy	-	-	-	-	81%	-	-	-	78%	-
	Precision	-	-	-	-	83%	-	-	-	78%	78%
	Recall	-	-	-	-	81%	-	-	-	-	-
	F1-score	-	-	-	-	80%	-	-	-	77%	-

The further analysis from Table 4.2 proves that the RandomForestClassifier and XGBoost Classifier not only do well with accuracy but also pride themselves on high precision, recall, and F1-score for each category of Youth migration in abroad. Overall Model Performance Analysis. The performance of different machine learning models is shown in the table across the performance measures including accuracy, precision, recall, and F1-score. These models were evaluated using varieties of feature engineering, data transformation, class balancing, hyper-parameter tuning, and ensemble techniques. (random forest) is a strong baseline on all quantitating performance metrics. It demonstrates a strong performance (accuracy=0.81, precision=0.85, recall=0.81, F1 score=0.80) in a variety of settings. This model is robust to variations in feature engineering or hyperparameter tuning, and it easily copes with class imbalance by class weights or SMOTE. The stability of Random Forest demonstrates its stability for this task, indicating one of the most stable models among the tested ones.

Meanwhile under testing scenario, Decision Tree has the low performance for the all the settings, and it reaches the accuracy, precision, recall and F 1 - score all with less than 0.75. It is clear from all this feature engineering and hyperparameter tuning that Decision Trees were too naive to escape the curse of dimensionality. The seeming weak relationship of the model to the patterns in

the data suggest that the model may not 107 be suitable for more complex prediction problems, unless used in conjunction with other methods, such as boosting or bagging.

KNN obtains mild gains in accuracy (0.79) and F1-score (0.79) with polynomial features and SMOTE to compensate for the class imbalance. But the precision and recall scores vary, indicating while KNN does quite well with some engineered features, it can still be challenged with non-linear interactions in the data. This indicates that the model is not the highest performing on it's own but may benefit from optimization. Extra Trees also does well in the beginning with precision (0.83) and recall (0.80). But it will plateau unlike Random Forest. This implies there's some level of performance that the Extra Trees model can achieve, but that it's not as flexible as Random Forest across difficult or complex datasets. Bagging (Decision Tree) also achieves consistent results, but the performance is not so high as that of Decision Tree model with accuracy ≈0.79 and F1-score = 0.76. The fact that there is not much gain regardless of configurations implies that bagging with Decision Trees does not offer big improvement largely because of the built-in shortcomings of the model, Decision Tree. XGBoost looks very promising, particularly with tuned hyperparameters. After tuning, it attained an accuracy of 0.80 and a precision of 0.83, making it along with multiple other approaches, among the best performing approaches. Its recall of 0.77, however, is somewhat less than that of Random Forest and SVM, thus, it may trade-off a few true positive cases. However, XGBoost is still a good option for complicated prediction tasks when properly tuned, demonstrating its capability of big data processing.

Logistic Regression has a mediocre performance, its accuracy and F1-score tend to plateau around 0.77. Even though Logistic Regression is a basic model, it does not benefit much (in terms of performance) with tuning the hyperparameters; meaning, it is not complex enough to handle the complexity of the task. Nevertheless, it can be useful as a reference model.

SVM (RBF Kernel) ranks as one of the top classifiers with test score, an accuracy of 0.81; precision: 0.85; recall: 0.81. It performs well on all the metrics including ensemble model setting, which demonstrates that our SVM is a trustable model

for this task. Its trade-off between precision and recall and the fact that it keeps a constant F1-score of 0.80, makes it a great choise for this problem. LightGBM and CatBoost exhibited potential, but were less competitive compared to other models such as Random Forest and SVM. Although the F1-score (0.79) of LightGBM achieves some gains by tuning, it fails to demonstrate the more robust, noise-tolerant performance. CatBoost does fairly well post tuning however it performs worse than the other models across various metrics, such as recall (0.81) and F1 score (0.80) so it's hard to trust for this particular task.

Last but not least, the combination models (Voting and Stacking) do not help most models achieve better results. While they do provide some performance improvements, especially with more powerful models (eg: Random Forest or SVM), their effect is less obvious for simpler models like DT (and Bagging(DT). This indicates that while ensemble methods prove effective in some models, they are not always prone to substantial improvements, especially in cases where the base models are poor.

In summary, the Random Forest and SVM models are the best performing and most consistent models in migration decision prediction in terms of the performance in different metrics and on different settings. The XGBoost also does well with right tuning but a bit worse than them. Even with Ensemble Methods help, simpler models such (Decision Tree and Bagging (Decision Tree)) have limitations and they don't outperform single models by far. LightGBM and CatBoost exhibit some promise but need some tuning for a real competition with the best.

Model Evaluation: Accuracy, Precision, Recall, F1 Score, and Confusion Matrix:

This part explains the performance of the model, comparing the prediction and the label based on accuracy, precision, recall and F1 score, and the confusion matrix (TP, TN, FP, FN). These metrics allow you to see which of the under/overfitting - and the positive vs. negative - classes the model is performing well against.

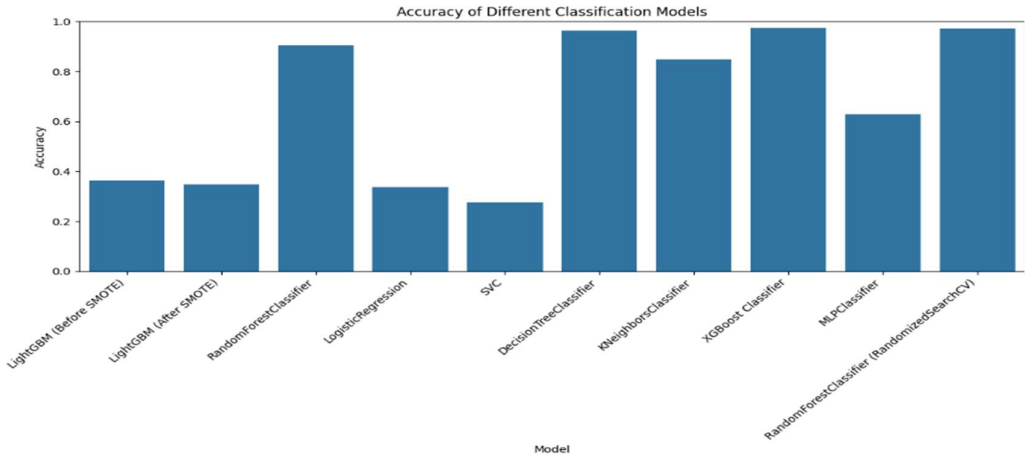

Figure 4.1: Accuracy Comparison of Different Classification Models.

This figure (4.1), comparing classification models based on their accuracy, indicates XGBoost with the highest accuracy, followed by **Decision Tree** and **RandomForestClassifier (tuned by RandomizedSearchCV)** and **KNN**. The performance of the LightGBM model is observed to have significantly improved with the use of **SMOTE (Synthetic Minority Over-sampling Technique)**; it indeed points to solving class imbalance as a crucial problem. poor performance was witnessed by the models of **SVC** and **Logistic Regression**. So, the **XGBoost** model appears to be the best model for this dataset, but ensemble methods also provide very good results, with **XGBoost** and **RandomForest** being prominent among them.

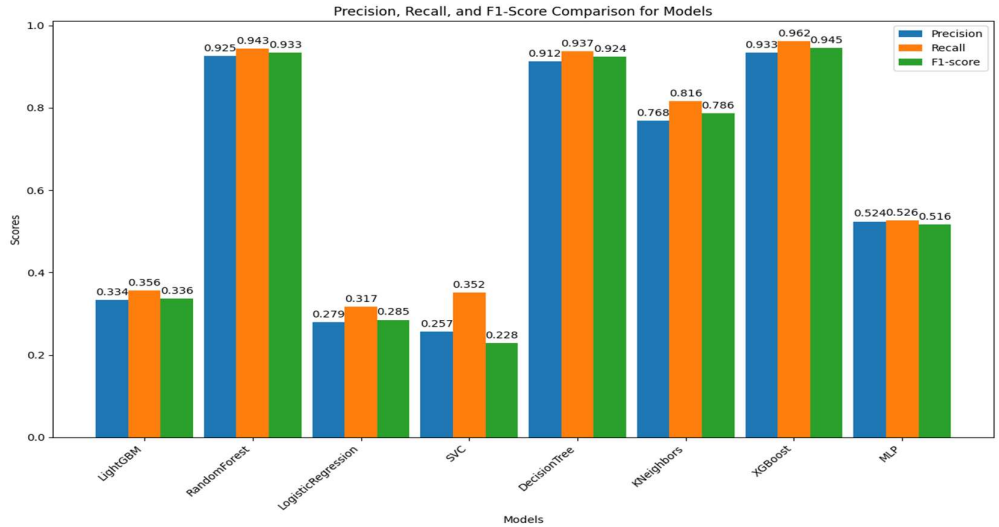

Figure 4.2.: Precision, Recall, and F1-Score Comparison for Model.

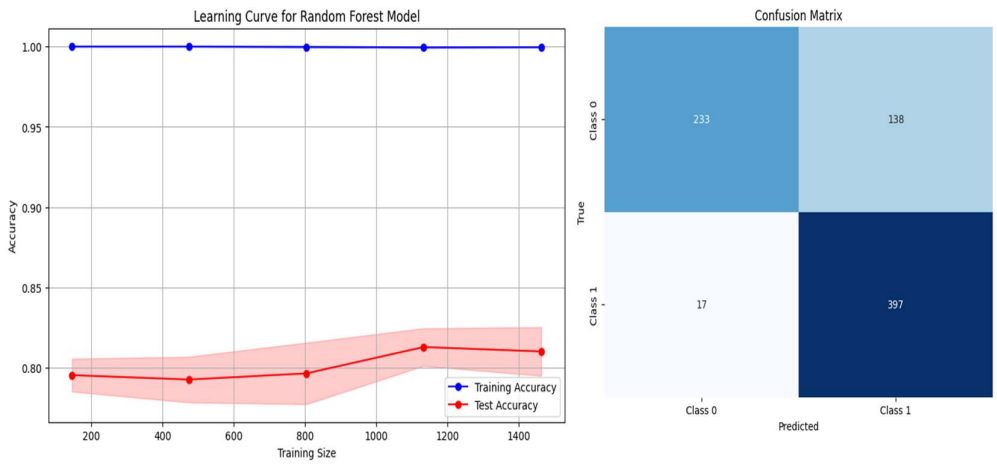

Figure 4.3: Learning Curves and Confusion Matrix for RandomForestClassifier.

In Figure (4.3) shows, the learning curve of the RandomForestClassifier indicates that training accuracy and test accuracy both improve as the size of the training set increases. The learning curve indicates severe overfitting of the Random Forest model. Your training accuracy is very close to 100% and test accuracy is in 79–82% in both If your train accuracy is close to 100% and test accuracy significantly lowers than it, this significantly signals that you are overfitting. We can see this challenge being emphasized in the confusion matrix: the model is doing well on Class 1 (397 true positive and only 17 false positive) but do poorly with Class 0 (233 true negative and 138 false negative). It indicates a high recall of Class 1 with lower precision as a result of high false positive. The Overall accuracy is approximately 80–82%, but the performance on the two classes are unbalanced, showing the biased results toward Class 1. This may be mitigated by hyperparameter tuning (restricting tree depth & adjusting min samples per split/leaf), addressing potential class imbalance and removing redundant features that would help improve generalisation and effectively balance the classwise performance.

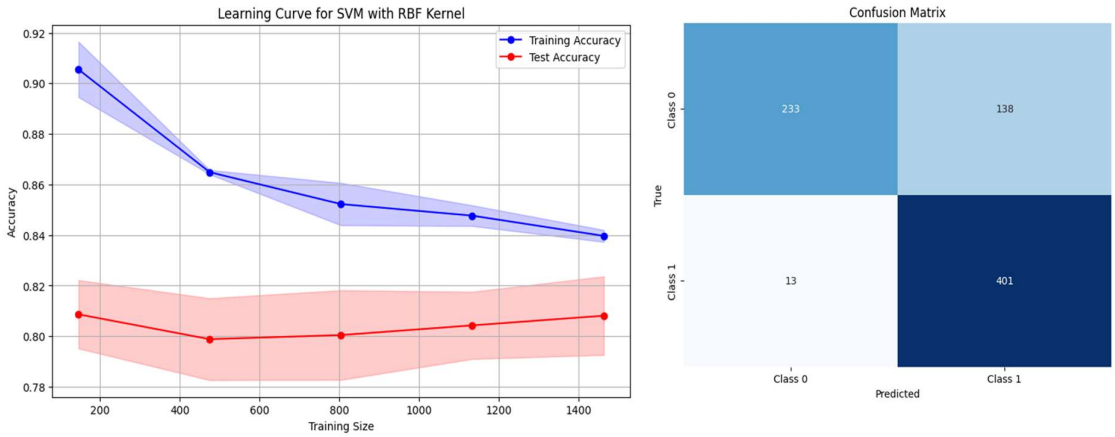

Figure 4.4: Learning Curves and Confusion Matrix for SVM with RBF Kernal.

In figure (4.4) shows, two important investigations of model SVMs (Support Vector Machines) trained by RBF (radial basis function) kernel. On the left, the learning curve shows the evolution of the accuracies on both the training overall 86% accurecy and test sets when increasing the size of the training set. The final test accurecy 82%. The training accuracy is high to start, but then begins to drop as more of the data are used, suggesting that there is some overfitting happening at these smaller training sizes. The test accuracy, on the other hand, shows an almost constant but lower accuracy hinting that the model may not be generalizing well on the unseen data. The confusion matrix presented on the right is a tabular representation showing how well the model can classify data into two distinct classes. Class 0 is predicted well with 233 true positives and 13 false negatives, class 1 has a higher rate of misclassification, with 138 false positives and 401 true positives.

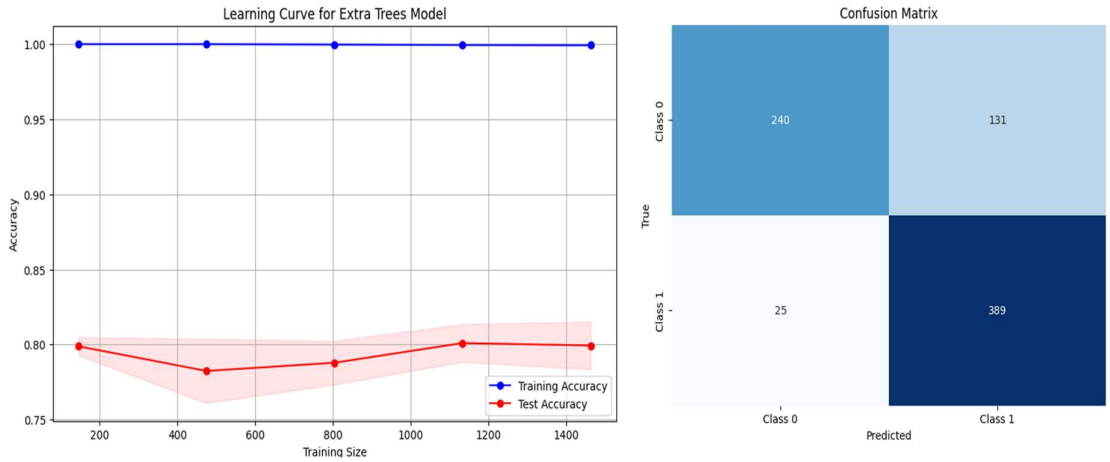

Figure 4.5: Learning Curves and Confusion Matrix for Stacking Classifier Evaluation.

In the figure (4.5) shows that for all training set sizes, The image showcases the learning curve and confusion matrix for an Extra Trees model. On the left, the learning curve shows both training and test accuracies. The training accuracy remains close to 1 (indicating that the model is fitting well to the training data), while the test accuracy stays relatively stable around 0.80, suggesting the model may not be overfitting, as the gap between the two curves remains small. This indicates that the model is generalizing decently to new data, although there is room for improvement. On the right, the confusion matrix shows how well the model classifies the data into two classes. Class 0 is classified fairly well, with 240 true positives and 131 false positives, while Class 1 is better predicted with 389 true positives, but still has 25 false negatives. The overall performance seems decent, though there is some misclassification in both classes, particularly Class 0.

4.3 Results & Discussion

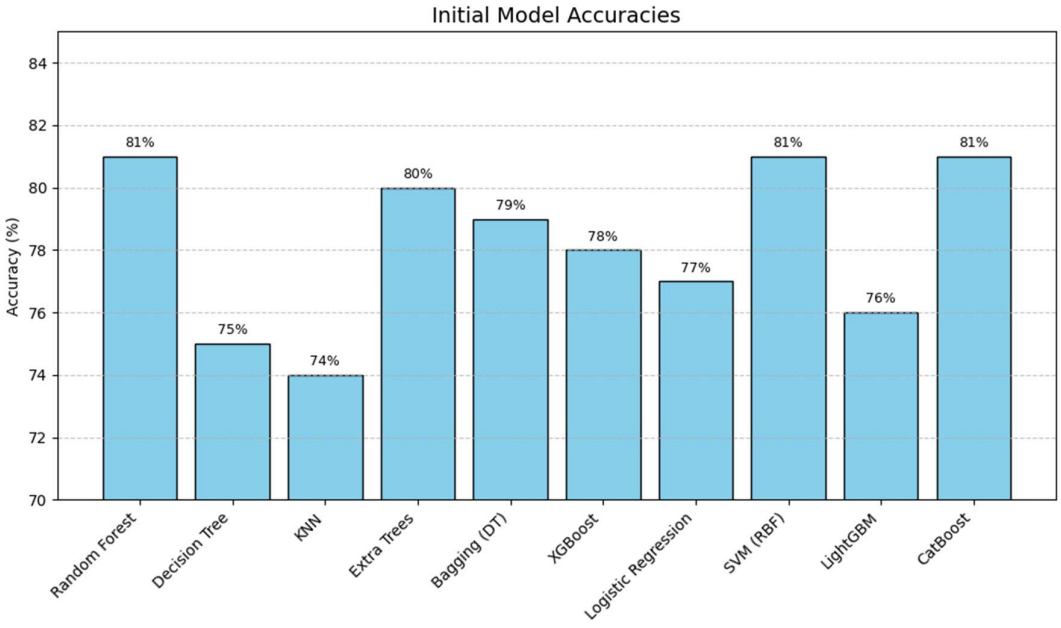

Figure 4.6: Accuracy analysis of different models.

The picture gives us a good side by side overview of the initial accuracy performance for a few popular machine learning models. The Random Forest and CatBoost models remain in the lead with both having an accuracy of 81%. This is an indication that these models, which are generally able to cope with intricate data, are particularly well-suited for the task at hand. The extra trees ensemble also gets an accuracy of 80% indicating the effectiveness of ensembles in dealing with data variance and preventing overfitting. Also Bagging(DT) which is also a ensemble of decision tree achieves 79%, which shows the ensemble models can help increase the performance of models. XGBoost, an extremely powerful and popular gradient boosting technique, has an accuracy of 78% while SVM (RBF), which is a famous model of handling non-linear relationship between features via kernel tricks, is also with 78%. Summary These experiments demonstrate that while these models do quite well, they are still not quite at the level of the top performers in this population. Logistic Regression has a score of 77%, reflecting the fact that it is very simple and attempts to capture linear relationships — which isn't ideal as the relationship quickly breaks down. Another gradient boosting model LightGBM gets 76%, indicating it is fast to compute and less

accuracy as other boosting models like CatBoost. Ultimately, the easier models such as KNN and Decision Tree come at the bottom with accuracies of 74 and 75%. These findings suggest that more basic models that do not capture full complexity of features maybe inappropriate, particularly if the feature interactions are strong. In general, we can conclude that ensemble methods are best for this task: Random Forest, Extra Trees and CatBoost perform best as they handle the complexity of data because it is a case of multi-class classification, not just a binary one. Also, I will say gradient boosting models such as XGBoost and LightGBM do well, but they still fall behind stiff competiton with the top models. Simpler models like KNN and Decision Tree may require additional tuning and feature engineering to improve accuracy. That's why, for the highest accuracy models, the decision should count on more advanced models like Random Forest or CatBoost, and for simpler tasks, models that are simpler can be useful, such as those with the purpose of interpretability.

4.4 Summary

The implementation portion of a project describes the formalized steps necessary to make the strategies off the planning document a reality. It contains the procedure followed in building, testing and deploying the system or the solution, such as the tools, technologies and methodologies implemented. This part also must discuss perceived difficulties in the implementation phase and how they were addressed.

The Results section describes the project s outputs, such as findings, data, or performance PROPRIETARY AND CONTENTS 3 measures. It reports the real results of the process compared to the desired results, seeking the success factors for the process transfer. This part of the report might contain visual aids such as graphs, tables, and or charts that illustrate key results and provide some information about the overall effectiveness of the project. It also addresses any lessons learned and opportunities for improvement.

Chapter 5

Engineering Standards and Design Challenges

This section covers conformance to different engineering specifications, such as software, hardware, and communication specifications. It also examines the societal context of engineering, ethics, sustainability, and problems encountered in project management and financial analysis. The chapter finishes with a consideration of complex engineering problems and problem solving.

5.1 Compilance with Standards

5.1.1 Software Standards

Software Requirements:

Python programming language: Python is primarily used in this options programming for data analysis and machine learning.

Pandas: To load, manipulate, and analyze data; to handle missing values and transform data.

NumPy: To work with arrays and for numerical calculations.

Scikit-learn: A full suite of different machine learning algorithms, model selection tools, and evaluation metrics. It also supports various utilities for preprocessing.

Matplotlib and Seaborn: For data visualization in exploring the datasets and showing results.

Imbalanced-learn: Tools for estimating imbalanced data, which is SMOTE in this project.

LightGBM and XGBoost: Libraries of gradient boosting for efficient building of classification models.

5.1.2 Hardware Standards

Processor Needs:

The processes loading, cleaning data, engineering features, and training different classification models require a powerful CPU runtime. The more powerful the

CPU, the faster these processes proceed.

Sufficient Memory (RAM):

RAM is required to load the dataset into memory so that the models can access and process it during training. The necessary RAM size depends on the dataset size and model complexity.

Storage:

A large volume of storage is required to store the dataset and trained models.

Graphics Processing Unit (GPU) (optional but highly recommended for some models and tuning)

For some models at least (eg Logistic Regression or simple Decision Trees), this is nice-to-have, but not a must-have; however, if you want speed up the training for harder models (eg to train a MLPClassifier, or some ensembles with XGBoost/LightGBM for some configs) - particularly when conducting hyperparameter tuning via cross-validation, then you really want a GPU - also when working with bigger data or large tuning attempts.

5.1.3 Communications Standards

The Communication Standard sub-section, provided in the Compliance and Standards section, documents the critical communication protocol and behavior to support an open effective communications process throughout the project. It defines the mechanisms of communication -emails, meetings or project management tools- and makes sure the information is produced at each of these touchpoints within the organization. The protocol also outlines how often updates or meetings, such as weekly reports or daily check-ins, will be conducted, ensuring that all parties remain informed. It also underscores the need to record all communications and decisions on a file for future reference and for compliance. Explicit instructions to involve stakeholders and to remain professional in terms of tone and language are also specified, thus enable effective joint work and minimizing misinterpretation. This is to keep the communication within the project clear, organised, and aligned to your high-level objectives.So these standards should be maintained.

5.2 Impact on Society

The results in this work and the forecast framework developed in this project could have societal implications across many sectors, particularly for public policy making, child safety and immigration patrol in Bangladesh. By shifting the study of migration from reflection with retrograde optics to a predictive science based on metrics, the work introduces an instrument for potentially more proactive, efficient and humane informant-led interventions. From a government agency and policy NGO perspective, this system provides a simple way to identify youth before they consider migrating who may be at increased risk for psychological distress. Instead of relying on expensive, long-established traditional surveys that provide nothing more than a snapshot in time, this machine learning system might allow for a continually updated understanding of migration patterns. "Let's do a better job of finding ways to divert resources, whether it's mental health or job counseling or financial planning for those who are vulnerable, their families and the community. By understanding more about the determinants of decisionmaking on migration and the underlying sources of distress prompting migrants to move, policymakers may be able to design intervention mechanisms that address root causes such as unemployment better and manage -'migration types more strategically.

Central to the framework, though, is the shift in standpoint that it affords people as they are able to make choices (in this case, the young people at the heart of this study) and potentially feel confident in so doing. The personalized recommendation module is designed to predict not only the psychological stressors but also an individual barrier that people are likely to face in their subsequent visit. That could help potential migrants better prepare themselves, and think more skeptically about the wisdom of their migration plans. As a mechanism for providing end-users with something sensible they can understand and act on, this tool represents an important type of decision-support that may serve to mitigate some of the mental health problems far too often witnessed in poorly planned or difficult migration journeys.

At a theoretical level, this research also contributes to a deeper understanding of the complex interplay of economic aspirations, peer pressures and mental well-

being that shapes youth mobility decisions in Bangladesh. Results can contribute to discourse and debate about migration in the public domain or help create a climate that is more supportive towards young people whether they move or stay at home. Finally, by deploying technology to provide foresight and personalized advice, this project could help make migration less dangerous, confusing and traumatic for millions of people's and families' lives – which ultimately can create a more stable and successful society.

5.2.1 Impact on Life

However, But the most basic effects are those upon the human lives it's supposed to save: the young waifs who ply there. It is not a mere prediction engine; rather it was intended as an individualized decision support machine having the ethos of facilitating the young obstinate individual (making its life altering move i.e. migrating) to make his/her decision differently.

This is important for a young person thinking about making that journey, as the community structure is an essential period of self questioning and reflection. Rather than giving a black box estimate, we provide the user with transparent view of what (business pressure, social motivation or educational aspiration) is driving their intention. This can inspire them to consider their motivation more rationally and make an informed decision in the long run.

Certainly life style effects such as decreased psychological distress fall into this domain. Migration in itself is stressful and this device helps to identify potential stressors before, well, entering the country. By identifying issues such as concerns over family expectations, or even lack of financial security, the system allows people to build coping strategies and find support before it all gets on top of them – with positive effects on mental health throughout their entrepreneurial journey.And not just that, but the personal advice can be lived as a life plan. For instance, when the model highlights financial barrier as a major cause, the recommendation module can make suggestions of such resources as those involving financial planning or skills enhancement. That makes the abstract into a set of things that one can do and now be better positioned to encounter the experiences of migration.

And ultimately, it's life-changing for the individual. By unravelling the complex interplay of the various determinants of migration, and by allowing evidence-informed tailored advise, this a tool to empower young people in preparation for one of the major decisions in their life: it strengthens them and promotes well-being.

5.2.2 Impact on Society & Environment

The impacts of migration on society and the environment are multi-dimensional and affect various dimensions of receiving and sending societies. Societally, migration can lead to substantial changes in the demographics of a society by affecting not only the size and age distribution of the resident population but also its ethnic make-up. What the answer lies in, more often than not, is urbanisation and depopulation from rural regions to the cities — a well known demographic fact among developing countries. This transfer can cause overcrowding in populations, overloading public services and increasing the housing, educational and health needs. While migration may offer new economic opportunities to migrants, it also presents challenges to host societies such as demand for social integration and employment of newcomers and cultural conflict that can arise due to the potentially large scale redistribution of per capita income amongst workers and their punitive-benefit families in home nations.

In ecological language, migration has its good and bad sides. Migration-driven by urbanisation is often accompanied with increased environmental degradation, pollution, depletion of natural resources (water, land and food security) and energy consumption. Moreover, facilities constructed in the name of providing infrastructure for floating population can cut forests, lead species to extinction and have unfavourable effects on local ecology. But migration might also have thanksable environmental dividends, especially as it involves the transfer of resources and knowledge of eco-friendly lifestyle habits. Similarly, rural uprooted people might bring with them new agricultural techniques or environmental awareness — such as a greater appreciation for organic produce or an understanding of renewable resources which may lead in the direction of sustainability.

Overall migration has an ambivalent role in shaping societies and economies, and natural environments, both in origin and destination areas. It can be a driver of economic development and cultural exchange, but it also entails problems of resource use, social integration and environmental protection. Dealing with them in a coherent manner requires an integrated approach that balances the opportunities offered by migration with those for good urban development and management that is sustainable.

5.2.3 Ethical Aspects

This experiment challenged the ethics of using sensitive abroad migration data and AI in medicine. Additionally, patient privacy and data security need to be considered - such as through data anonymization – as well as any other regulation that is applicable. Consent The consent obtained from the participants and also transparency in the research process is vital. They're also going to put AI models through the wringer for bias, prioritizing fair balance so no single demographic has a runaway advantage. In addition, the AI models are explainable and interpretable; thus, the workers can have confidence to rely on/validate their predictions. Finally, broader societal implications for the introduction of AI in health care are discussed such as efforts to promote fair deployment and access to health care. And finally, lead in the AI tech that make you trust AI as a human with a doctor on top. These high-level ethics themes are woven through the programmed of research to support the development of technology ethically, for people in society.

5.2.4 Sustainability Plan

Sustaining the research and findings is important in generating enduring impact, lasting benefit to health care and technology users. It revolves around environmental and social sustainability in the emerging AI in health solutions. From the environment standpoint Environmentally, software is a more sustainable resource for two big reasons; software requires much less going abroad and migration infrastructure and therefore also has a lower carbon footprint and waste-intense model when compared to traditional model like that.

It should refine these AI models further; we can design ones that actually do run efficiently: not as energy intensive as the most accurate. This is in alignment with the white sustainability global priorities of managing the environmental harm done from delivering health care.

Socially, it is our ambition to democratize going abroad dicision maker — elevate the frequency of equitable access to abroad prediction system by using AI technologies in response with specifically issued accessibility-aimed challenges from underserved areas of the world. With wearables and smartphone apps, physiological pressure for abroad dicision making could come to people's communities without the need for specialized care. It could also reach difficult-to-reat populations, who have geography or economics or some other impediment to traditional models. The study is also the AI system construction promising taking into account various potential ethical problems such as proposal, fairness and inclusivity for all users. To keep the research result we would have to: (1) get the future model update, (2) maintain the system, and (3) scale in size in the future. The model will also go through periodic reviews and retraining using different up-to-date data to keep the model fresh and accurate over time. There will also be work with clinicians, governments and tech providers to ensure that these AI-focused solutions are adopted in existing healthcare provision systems and their further development encouraged. By connecting the findings of the present study with sustainability visions for the future, it will help to further capitalize on gains made in access to healthcare and environmental social benefit.

Its investable plans are therefore to reduce environmental damage, increase social Justice, develop an AI health science and technology developable plan and target the sustainable character of AI health technologies. This is an effort to wring as much from research as it can at a time when the returns that such research has given us in recent decades have, arguably, supported us as a society.

5.3 Project Management and Financial Analysis

Table 5.1: Project Management and Financial Analysis Table

Activities	Duration (Week)	Estimated Cost (BDT)
Research and Planning	3	500
Visiting Data Collection	9	6000
Online Tools (1)	3	1000
Preprocessing and Feature Engineering	6	No (Colab Use)
Model Training and Evaluation	12	No (Colab Use)
Online Tools (2)	2	1500
Recommendation System and Integration	5	No (Colab Use)
Final Model Evaluation	3	No (Colab Use)
Documentation and Report Writings	5	2600
Total	**48**	**11600**

.

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

In this section, provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Use Table 5.2). For P1, you need to put another mapping with Knowledge profile and rational thereof.

Table 5.2: Mapping with Complex Engineering Problem.

EP Definition	Justification (with Knowledge Profile)
EP1 Dept of Knowledge	This work demonstrates the basic principles of K3 through applying machine learning techniques including Random Forest, Logistic Regression and LightGBM models to predict migration decision and physiological conditions based on migration data. It demonstrates expert (K4) use of advanced techniques such as nonlinear data learning algorithm for improved accuracy in predicting of youth migration problem. The work also uses the engineering practice & design (K5) as the dataset is gathered, cleaned, features engineered and model evaluated in a systematic way. From an Engineering Practice & Technology point of view (K6), it applies ML models of clinical diagnostics. It relates to K8 (Research literature) by referencing and synthesis of literature of state of the art of similar studies in predictive modeling of diseases via machine learning.
EP2 Range Of Conflicting Requirements	The project focuses on EP-2 by detecting inadequate requirement in the data analysis and model deployment phases at the imbalanced class's concern of the migration dicision-related dataset. It incorporates approaches such as SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset and for getting proper and better model training to enhance the precision and recall for the small data class.
EP3 Depth of Analysis	It fulfills EP-3 by extracting and evaluating numerous machine learning algorithms and models, as well as comparing their performance, with respect to performance metrics accuracy, precision, recall, and f1 score. Special focus is given to Random Forest, and XgBoost, which are particularly useful in predicting youth migration decision related in the ability of specifics driven more through deeper analysis such as hyperparamter tuning and cross validation.
EP4 Familiarity of Issues	This interdisciplinary work goes beyond computer science and engineering, adding to the

	Migration dicision, and improving the youth migration related issues which is EP-4.
EP5 Extent ofApplicable Codes	N/A
EP6 Extent Of Stake-holder Involvement	N/A
EP7 Interdependence	There's preprocessing, there's featture selection, there's model building and evaluation which all share information in a lot of ways. All these stages are associated with each other — hence when one stage (as an example, feature engineering) is optimized it can lead to better models performance, which in turn make possible to obtain better predictions of migration disicion related section.

Mapping with Knowledge Profile

This section is designed to map the overall problem and EP1 ***(multiple between K3, K4, K5, K6, K8 for attaining EP1)*** to the Knowledge Profile.

Table 5.3: Mapping with Knowledge Profile.

K1 Natural Science	Experience with advanced systems, physics, or environmental data, applied to training machine learning models in natural science.
K2 Mathematics	Higher level math such as linear algebra, calculus, probability, statistics which are necessary to write machine learning algorithms.
K3 Engineering Fundamentals	Fundamental understanding of algorithms, data structures and computation principles underlying machine learning systems.
K4 Specialist Knowledge	Strong understanding of machine learning algorithms, neural networks, natural language processing, deep learning, etc.

K5 Engineering Design	Building machine learning systems - from data collection to model deployment - which can operate at scale with high utilization.
K6 Engineering Practice	Training from scratch and deploying machine learning models to address real-world problems (including model evaluation, optimization, and integration.
K7 Comprehension	Ability to interpret and analyse patterns in data, to select and apply appropriate models and treatments, and to appreciate the limitations and likely levels of error in their results.
K8 Research Literature	Scanning and summarizing recent ML research papers, journals & work, to direct and seed current work and search novel trends.

5.4.2 Engineering Activities

In this section, provide a mapping with engineering activities. For each mapping, add subsections to put rationale (Use Table 5.3).

Mapping with Complex Engineering Activities

This section is designed to map the overall problem and EA's ***(multiple).***

Table 5.4: Mapping with Complex Engineering Activities.

EA1 Range of resources	This study used different tools which including High-Performance Compute (CPUs-GPUs), ML libraries (e.g., Numpy, pandas, Scikit-learn, Matplotlib and Seaborn) and Annotated Youth Migration related Datasets. These resources enable systematic investigation and might contribute to the advancement of predictive tool in terms of physiological condition (e.g., stress, tension). The study further points at ethics aspects such as patient data privacy, which is fundamental for the ethical application of abroad migration data in AI-based diagnostics.
EA2 Level of Interaction	
EA3 Innovation	

EA4 Consequences for society andenvironment	This paper is of great social benefit, to improve the health system as we can provide more accurate predictions for anyone going abroad (Y/N). It enhances the capacity to diagnose and comprehend such conditions through AI, and can help produce better prediction outcomes. Also, the project promotes sustainable friendly environmental sustainability by reducing energy consumption in computational works and utilizes cloud technologies. Ethical application of technology in medicines upholds the privacy of the individual through the moral treatment of abroad going data.
EA5 Familiarity	This builds on previous work in youth migration and machine learning and draws attention to aspects of predicting migration dicision respect of others factors. A detailed comparison of various machine learning models is carried out and novel applications of nonlinear data and the state-of-the-art, as far as we are aware, of diagnostic methods are developed.

5.5 Summary

The essence of this system lies in viewing macroscopic machine learning studies from a systemic point of view. The corpus tackles the availability of crucial resources such as toolkits, datasets, and computation for developing and evaluating machine learning models. The level of interactivity demonstrates the way in which the machine learned systems interact with users and other systems that will implement them in actual applications. Innovation highlights the continued growth and new challenges in machine learning, a wide-ranging field with applications throughout science and engineering. Societal and environmental consequences Finally, social and environmental consequences cover society-level impacts of machine learning (ethical issues, privacy concerns) and potential social changes resulting from this technology. Finally, knowledge refers to the degree of familiarity an idea has within research and practices communities, and influences the planning and use of new methods. They, in combination provide a comprehensive view of machine learning research as it stands today and what is possible to come.

Chapter 6

Summary, Conclusion, Recommendation and Implication for Future Research

Key findings of the study and their implications Summary and conclusions This chapter presents summaries conclusion from the study. It describes the limitations of the research, makes suggestions for future studies, discusses the implications for subsequent research and possible improvements.

6.1 Summary of the Study

In terms of modeling the influence on migration decision-making, our research aims to use machine-learning techniques for predicting migration-related outcomes. Mixing and meeting all the different types of ML algorithms (Random Forest, K-Nearest Neighbors, XGBoost and Support Vector Machines) with our data about personal details and behavior patterns, the project aims to develop a trustable system that is also scalable there for migration prediction. To date the models have been trained on diverse data, such as age group, occupation, levels of psychological stress and the impact of social media. This enabled us to gain further insights into the factors that drive people's movement (migration) decision.

Pre-processing: a few pre-processing steps such as missing value imputation, feature scaling and one-hot encoding have been applied to pre-pare the input data for the model. Moreover, methods for feature selection such as SelectKBest and Mutual Information techniques were used to pick the most predictive features. The model was greatly improved when tuning the hyperparameters (using grid search and randomized search), but we focused on random forest, since it always achieved a good accuracy in all phase). Interpretation methods based on feature interpretability, for example SHAP and LIME, were applied in study. These techniques succeeded in describing what is important and how the models make

their decisions. For the study, a recommendation module was created that provided actionable feedback based on model predictions rather than regular model predictions while also including tailored insights and short-snapped advice for curious movers.

The study illustrates how machine-learning models could contribute to migration-sensitive decision-making by enabling better pre-screening of candidates for localised access to prevention (such as early ART, targeted testing in seasonal workers), and thus optimising resource allocation. This paper also discusses broader social, environmental and ethical impacts of migration prediction systems, specifically fairness, privacy and transparency in model deployment. Ultimately, the book aims to provide a substantial contribution to predictive analytics as well as digital health and migration governance worldwide by providing a set of experiences that are at once diverse and actionable.

6.2 Limitations & Conclusions

These models could also be used to predict migration decisions ahead of time, and the products applied to provide individual or community personalized advice for effective management of the migrations. Using different machine-learning classification algorithms (Random Forest, K-Nearest Neighbors, XGBoost and Support Vector Machines), the paper evaluates how well these computational models are capable to predict migration intentions and outcomes with respect to different assortments of socioeconomic and/or behavioral predictors.

The study brings to fore interesting findings, it says of elements that heavily influence decisions on migration. These are psychological stress, employment, impact of social media and couriosity in the locality. The best performance among the models built using these features was achieved by the Random Forest model after hyperparameter tuning. Added as well that the model, through model explainability tools like SHAP and Lime was interpretable. It made those predictions they made, transparent as well and it instilled that trust, transparency in terms of the decision making. Results of Such Study The result of this study was an opinionated feedback recommendation module based on predictions by aforementioned machine learning models. More than just a model

to forecast migration decisions, this module aims to provide personalized advice for people, so that they can better assess their migration plans. Through linking potential migration diagnoses to lifestyle interventions and side-effect pattern, the recommendation system extends prediction models beyond simple predictions unto a tool useful for decision makers.

From a social point of view, the results in this study may help local governments to establish scalable, fair and sustainable migration forecasts. It also refers to the ethical implications of sharing personal sensitive information. Access to privacy, equity and transparency throughout the entire process were paramount throughout the research. Yet that evidence has the power to help shape policy and prepare systems, allocate resources more smartly to keep migration manageable and less stressful for individuals and communities.

Finally, the study demonstrates the potential utility of machine learning for migration decision making and public health. And, cruelly, it can not only forecast that with its models but also offer valuable insights that might be used to halt migrations and provide societal benefits. Our work is an extension to previous works that discovered similar mechanism to support the fact of insect migrations by applying machine learning and can be used with some applied tasks. It would represent a valuable resource for policy makers, healthcare providers or members of the public.

6.3 Implication for Future Study

This article offers an insight in relation to machine learning in predicting migration trends and delivering personalised guidance in personal migration, although many areas of future research would be details of such developments, improved practical representation of these models. Future work should continue to build on this work and overcome its limitations by testing potential new ways to improve the accuracy, fairness, and utility of the different models.

One consideration for future work is to explore using more diverse and comprehensive data sources. Having employed the socio-political-economic factors as resources for its analysis up to this point, this study found that what might be called 'migration' is in fact seen to take place under a profoundly wide

range of external influences, namely, political, environmental and the cultural. The follow-up might be better models based on more-detailed global migration statistics, climate change information or policies from local governments and other groups. By learning from data collected from different locations and circumstances, we might be able to modify our model to incorporate the various ways in which local conditions influence an individual's decision to migrate, and become a more locally, adaptable global model. Another topic on which future research should focus is how to further promote fairness and how to defend against bias. Attempts were included to weight the dataset for fair predictions to be applied across all demographics, however biases may remain. Many, so state-of-the-art techniques to detect bias as well as prevent unfair predictions, will be necessary in the future studies of this topic: to guarantee that the migration forecasts are both fair and just for every people, in particular for those populations most disadvantaged. These could include applying adversarial training — or embedding constraints of equity in machine learning models to govern predictions, so they aren't unfairly affecting a certain demographic group the most when weighing results they generate.

Put another way, the recommendation system developed in this work is still far from mature. Right now all the good advice it offers can be taken or left depending on the player-emigration and living conditions the players choose; with further tuning it can perhaps tell the player what to do, and for each different person's condition. Absorbing more targeted data on the individual patient's mental health, family dynamics and specific wants (not to say needs) as opposed to general needs could enable an even more fitting recommendation. Future research studies should also consider how to best integrate real-time data, such as financial indicators or work in the market trends, to offer the man contemplating leaving the home timely and contextually relevant advice. Stipulate that the ethical and privacy considerations reflected in this research remain as areas for future research. The more machine learning models are deployed for making predictions on critical life issues such as migration, the more important they become, as model development must be transparent and ethical. Future work should further explore how data privacy can be enhanced, how to

present the model, and look out for ethical issues such as the use of personal information in predictive systems. Worth thinking about privacy-preservation ML techniques too - federated learning is again something to think about here for opening up protocol uses of sensitive data while preserving individual privacy. Last but not least is the intersection of machine learning with policy making and social planning is an interesting area for further research. It may even be feasible, with insights of migration forecasting models, to orient the support in policy for sustainable migration management when cooperating with stakeholders in politics, urban planning and society at large. In the future, it may be interesting to explore ways these models can be used to encourage flexibility in human decision making, such as towards resource allocation, planning cities, and taking in refugees, so that ensure flows that achieve both the needs of the individuals and if possible the benefits of the communities.

Overall's there's not a whole lot more to an employee motivation than one would presume; there is therefore room for some fresh ideas. If the following issues can be resolved, the next prospective studies may significantly enhance the precision, equality and universality of migration forecasting systems. This makes them worth more to people whose sore eaten rearing.

References

[1] Md S. Islam, S. Afrin, and A. S. M. Taj Uddin, "Factors associated with perceived higher quality health among people that have migrated within Bangladesh," preprint, 2021. Available: https://doi.org/10.21203/rs.3.rs-6440902/v1.

[2] A. V. Bakina, S. V. Yaremtchuk, O. A. Orlova, and Y. V. Krasnoperova, "Life satisfaction and migration intention of youth," Int. Conf. on Soc. Futures, 2019. Available: https://doi.org/10.2991/iscfec-18.2019.158.

[3] Y. M. Sabti and S. S. Ramalu, "Home country economic, political, social push factors and intention to migrate in Iraq: psychological distress as mediator," J. Migr. Stud., vol. 1, no. 1, pp. 2299507, Dec. 2021. Available: https://doi.org/10.1080/23311975.2023.2299507.

[4] A. Hossain et al., "Predisplacement abuse and postdisplacement factors associated with mental health symptoms after forced migration among Rohingya refugees in Bangladesh," JAMA Network Open, vol. 4, no. 3, e211801, Mar. 16, 2021. doi: 10.1001/jamanetworkopen.2021.1801.

[5] Shuva, "Information, employment, and settlement of immigrants: Exploring the role of information behaviour in the settlement of Bangladesh immigrants in Canada," Ph.D. thesis, Western Univ., 2020. Available: https://ir.lib.uwo.ca/etd/6877.

[6] M. K. N. Sohad, G. Celi, and E. Sica, "Factors determining migration intentions in Bangladesh: from land to factory," J. Econ. Stud., vol. 51, no. 5, pp. 1058–1076, 2024. Available: https://doi.org/10.1108/JES-06-2023-0293.

[7] A. Tirado-Espín et al., "Analyzing communication and migration perceptions using machine learning: A feature-based approach," J. Media, vol. 6, no. 3, pp. 112, 2025. Available: https://doi.org/10.3390/journalmedia6030112.

[8] M. A. Siddik et al., "Impact of climate-induced migration on depression: A study between disaster-affected migrant and non-migrant adolescents," Climatic Change, vol. 178, art. 76, 2025. Available: https://doi.org/10.1007/s10584-025-03916-5.

[9] R. Ahsan, S. Karuppannan, and J. Kellett, "Climate migration and urban planning system: A study of Bangladesh," Environ. Justice, vol. 4, no. 3, pp. 163-170, 2011. Available: https://doi.org/10.1089/env.2011.0005.

[10] K. N. Koly et al., "Health-related quality of life among rural-urban migrants living in Dhaka slums: A cross-sectional survey in Bangladesh," Int. J. Environ. Res. Public Health, vol. 18, no. 19, art. 10507, 2021. Available: https://doi.org/10.3390/ijerph181910507.

[11] A. M. Williams et al., "The migration intentions of young adults in Europe: A comparative, multilevel analysis," J. Popul. Econ., vol. 31, no. 1, pp. 1–28, 2017. Available: https://doi.org/10.1002/psp.2123.

[12] H. S. Baggen et al., "The application of machine learning to rural population migration research," Population, Space and Place, vol. 29, no. 1, pp. 1–15, 2023. Available: https://doi.org/10.1002/psp.2664.

[13] K. Best et al., "Applying machine learning to social datasets: A study of migration in southwestern Bangladesh using random forests," Reg. Environ. Change, vol. 22, art. 52, 2022. Available: https://doi.org/10.1007/s10113-022-01915-1.

[14] K. Khatri Park et al., "Machine learning applications in studying mental health among immigrants and racial and ethnic minorities: An exploratory scoping review," BMC Med. Inform. Decis. Making, vol. 24, art. 298, 2024. Available: https://doi.org/10.1186/s12911-024-02663-4..

[15] A. Sîrbu et al., "Human migration: The big data perspective," Regular Paper, Population, Space and Place, vol. 11, pp. 341–360, Mar. 2021. Available: https://doi.org/10.1007/s41060-020-00213-5.

[16] M. M. Biswas et al., "Psychological implications of unemployment among higher educated migrant youth in Kolkata City, India," Sci. Rep., vol. 14, art. 10171, 2024. Available: https://doi.org/10.1038/s41598-024-60958-y.

[17] S. Sultana et al., "Mental health and associated factors among Bangladeshi migrants in Thailand: A cross-sectional study," Sci. Rep., vol. 15, art. 966, 2025. Available: https://doi.org/10.1038/s41598-024-84650-3.

[18] M. A. Islam, M. H. Rahman, and M. S. Tabassum, "A machine learning approach to predict the chances of dropping out students due to COVID-19 in university perspective Bangladesh," Daffodil International University, 2023. Available: https://doi.org/10.1007/s12864-019-6412-8.

[19] A. Tirado-Espín, A. Marcillo-Vera, K. Cáceres-Benítez, D. Almeida-Galárraga, N. Orozco Garzón, J. A. Moreno Guaicha, and H. Carvajal Mora, "Analyzing communication and migration perceptions using machine learning: A feature-based approach," Journalism and Media, vol. 6, no. 3, pp. 112, 2025. Available: https://doi.org/10.3390/journalmedia6030112.

[20] T. K. Reddy, Y. Govardhan, U. Vamsi, V. Vidhya, and M. Selvameena, "Predict migration using machine learning," International Journal on Science and Technology (IJSAT), vol. 16, no. 1, pp. 1–10, 2025. Available: https://www.ijsat.org.

[21] M. S. Islam, M. A. Rahim, N. K. Podder, M. N. Hossain, and M. I. Hossain, "Prediction of irregular Bangladesh-EU migration trends using machine learning techniques," 2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), pp. 1–10, 2024. Available: https://doi.org/10.1109/ICAEEE2024.

[22] K. B. Best, J. M. Gilligan, H. Baroud, A. R. Carrico, K. M. Donato, B. A. Ackerly, and B. Mallick, "Random forest analysis of two household surveys can identify important predictors of migration in Bangladesh," Journal of Computational Social Science, vol. 3, no. 2, pp. 1–35, 2020. Available: https://doi.org/10.1007/s42001-020-00066-9.

[23] H. Weber, "How well can the migration component of regional population change be predicted? A machine learning approach applied to German municipalities," Comparative Population Studies, vol. 45, pp. 143–178, 2020. Available: https://doi.org/10.12765/CPoS-2020-08en.

[24] GeeksforGeeks, "Random forest algorithm in machine learning," GeeksforGeeks, Sept. 1, 2025. Available: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/.

[25] Y. Li, D. M. Umbach, A. Bingham, et al., "Putative biomarkers for predicting tumor sample purity based on gene expression data," BMC Genomics, vol. 20, no. 1, pp. 1–14, 2019. Available: https://doi.org/10.1186/s12864-019-6412-8.

[26] M. M. Khan, M. Masud, S. Aljahdali, M. Kaur, and P. Singh, "Comparative analysis of machine learning algorithms to predict Alzheimer's disease," Journal of Healthcare Engineering, vol. 2021, pp. 1–15, 2021. Available: https://doi.org/10.1155/2021/9917919.

[27] Codecademy, "Python: Sklearn Kernel Ridge Regression," Oct. 1, 2024. Available: https://www.codecademy.com/resources/docs/sklearn/kernel-ridge-regression.

[28] Y. Lou, Y. Ye, Y. Yang, and W. Zuo, "Individualized empirical baselines for evaluating the energy performance of existing buildings," Science and Technology for the Built Environment, vol. 29, no. 1, pp. 1–15, Oct. 2022. Available: https://doi.org/10.1080/23744731.2022.2134680.

[29] Simplifying Marketing, "Decision tree or flowchart," Simplifying Marketing. Available: https://simplifyingmarketing.com/decision-tree-or-flowchart/

[30] Arize, "F1 Score In Machine Learning: What It Is and How To Use It," Arize, Aug. 28, 2024. Available: https://arize.com/blog-course/f1-score/.

[31] Z. Sadeghtabaghi, M. Talebkeikhah, A. R. Rabbani, "Prediction of vitrinite reflectance values using machine learning techniques: A new approach," Journal of Petroleum Exploration and Production Technology, vol. 11, no. 3–4, 2020. Available: https://doi.org/10.1007/s13202-020-01043-8.

213-15-4347

ORIGINALITY REPORT

16%	11%	9%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to United International University Student Paper	1%
4	Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Progressive Computational Intelligence, Information Technology and Networking", CRC Press, 2025 Publication	<1%
5	link.springer.com Internet Source	<1%
6	Bui Thanh Hung, M. Sekar, Ayhan Esi, R. Senthil Kumar. "Applications of Mathematics in Science and Technology - International Conference on Mathematical Appplications in Science and Technology", CRC Press, 2025 Publication	<1%
7	Md Sirajul Islam, Sabiha Afrin, A.S.M. Taj Uddin. "Factors are associated with perceived higher quality health among people that have migrated within Bangladesh", Springer Science and Business Media LLC, 2025 Publication	<1%
8	research-repository.st-andrews.ac.uk Internet Source	<1%

25% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

57 AI-generated only 25%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.