# Capstone Project

Canadian Immigration Prediction

Submitted By:

Tirth Patel

Nachiketbhai Prajapati

Submitted to: - Rahim Samei

### Introduction

This document outlines the challenges faced, solutions implemented, and methodologies applied during the development of a predictive model for Canadian immigration trends from 2024 to 2028. The project utilized datasets sourced from Kaggle.com and CanadaImmigration.com, focusing on immigration statistics across multiple countries.

### Project Challenges and Solutions

#### Data Acquisition and Understanding

* **Challenges:**
  + The dataset contained inconsistent formats and missing values.
  + Certain columns were irrelevant or redundant.
* **Solutions:**
  + Performed initial data exploration to understand the structure and content.
  + Dropped unnecessary columns and handled missing values through imputation techniques.

#### Data Preprocessing

* **Challenges:**
  + Null values in crucial columns affected data reliability.
  + High-dimensional data required dimensionality reduction.
* **Solutions:**
  + Applied null value imputation techniques using statistical methods (mean/mode imputation).
  + Utilized Linear Discriminant Analysis (LDA) for dimensionality reduction.

#### Modeling

* **Challenges:**
  + Choosing appropriate models to predict immigration trends accurately.
  + Balancing model interpretability and performance.
* **Solutions:**
  + Implemented multiple models: Linear Regression for trend prediction, Time Series Analysis for patterns, and Random Forest for feature importance.
  + Smoothed time-series data to enhance prediction accuracy.

### Methodology

#### Dataset Preparation

* Collected datasets from Kaggle.com and CanadaImmigration.com.
* Preprocessed data through:
  + Handling null values.
  + Removing redundant columns.
  + Dimensionality reduction with LDA.

#### Exploratory Data Analysis (EDA)

* Identified top 5 and bottom 5 countries by immigration statistics.
* Visualized relationships and trends using Seaborn for better insights into data distribution.

#### Modeling Techniques

* **Linear Regression:**
  + Predicted immigration trends for 2024-2028 with a focus on India and random countries.
* **Time Series Analysis:**
  + Smoothed data for accurate forecasting.
* **Random Forest Classifier:**
  + Determined feature importance for immigration trends.
* **LDA:**
  + Reduced data dimensionality for better model performance.

### Key Findings

* The project successfully predicted immigration trends for Canada with notable insights:
  + Linear Regression effectively modeled trends for India and other selected countries.
  + Random Forest identified significant features influencing immigration.
  + Time Series Analysis highlighted long-term patterns and trends.

### Challenges Faced

1. **Understanding and Cleaning Data:**
   * Resolved issues of missing and inconsistent data using systematic cleaning techniques.
2. **Model Selection and Implementation:**
   * Selected models based on their suitability for trend prediction and feature importance.
3. **Data Visualization:**
   * Leveraged Seaborn for creating intuitive and presentation-ready visuals.

### How Challenges Were Resolved

* Conducted thorough EDA to gain insights into the data.
* Applied appropriate preprocessing steps to address missing and inconsistent data.
* Experimented with different models to determine the best fit for our dataset and objectives.

### Contributions

* **Tirth:**
  + Conducted Linear Regression and Time Series Analysis.
  + Handled dataset preparation and EDA.
* **Nachiket:**
  + Implemented Random Forest Classifier and LDA.
  + Focused on data smoothing and feature selection.

### Summary of Findings

* The project showcased the potential of predictive models in analyzing immigration trends, offering:
  + Insights into country-wise immigration patterns.
  + Key features influencing Canadian immigration.
  + Effective trend predictions using a combination of statistical and machine learning models.

### Future Enhancements

* Expand the dataset to include additional variables like economic indicators.
* Utilize advanced models, such as neural networks, for enhanced predictions.
* Apply findings to real-world scenarios, such as policy-making and resource allocation.