**Report on the Impact of Foreign Workforce and Students on Taiwan's Unemployment**

朱俊曉 **H24115328**

**Abstract**

This report investigates how foreign workers and foreign students relate to Taiwan's unemployment trends, addressing concerns about potential negative impacts on local job markets. Data spanning 2015–2023 from three sources (foreign workers, foreign students, and annual unemployment totals) were merged and analysed. Despite a small dataset, we believe that we can still find insightful information in determining the impacts of foreigners in Taiwan. Our initial hypotheses are that the increase in foreign workers and students will have a negative impact as more job opportunities will be taken from the local workforce. Our project results will highlight the importance of nuanced policy decisions: rather than strictly limiting foreign labour or student admissions, stakeholders might consider a balanced approach that fosters economic and educational benefits. Nonetheless, the limited annual data restricts definitive claims, underscoring the need for more granular, multi-variable information—such as monthly or quarterly data and additional macroeconomic indicators—to better capture the full complexity of Taiwan's labour landscape.

**Table of Contents**

## 1. Introduction

### 1.1 Background

In recent years, Taiwan has experienced a noticeable rise in the influx of foreign workers and international students. While these developments can contribute to economic growth, globalization, and cultural exchange, they also raise important questions about labour market dynamics—particularly concerning unemployment among local workers. This project aims to investigate the relationship between foreign workers, foreign students, and unemployment rates (or total unemployment numbers) in Taiwan using a combination of data analysis and predictive modelling. We seek to inform policymakers, universities, and the public on whether foreign labour or student enrolment significantly influences local unemployment figures. Our computational environment:

- **Operating system**: Vivobook_ASUSLaptop M3500QC
- **CPU**: MD Ryzen 9 5900HX with Radeon Graphics 3.30 GHz
- **GPU**: NVIDIA GeForce RTX 3050 (CUDA 12.6)
- **Language**: Python 3.12

### 1.2 Objectives

1. To analyse how changes in the number of foreign workers and international students correlate with unemployment in Taiwan.
2. To develop and evaluate predictive models that use foreign worker and foreign student data to forecast unemployment.
3. To offer scenario-based insights for policymakers (e.g., changes in foreign worker quotas or student admissions) and academic institutions regarding potential future labour market impacts.
4. To recommend steps for future research, including data improvements and additional modelling strategies.

## 2. Data Analysis

### 2.1 Dataset Overview

**1. Foreign Workers Dataset**

- **Original Contents:**
    - This dataset contained annual figures for foreign workers in Taiwan across various industries and sectors. It had multiple breakdowns, such as:
        - Workers by Industry Type.
        - Workers by Nationality.
        - Workers by Sex and Permit Authority.
    - The dataset was presented in tabular form with many categories irrelevant to our analysis.
    - The "Grand Total" column was embedded within rows rather than a dedicated column, making it necessary to extract this manually.
- **Preprocessing Steps:**
    - Extracted the "Year" and "Grand Total" columns.
    - Filtered the dataset to include only the years 2015–2023.
    - Removed irrelevant rows and unnecessary categorical data.

**2. Foreign Students Dataset**

- **Original Contents:**
    - This dataset detailed the number of foreign students in Taiwan, broken down by:
        - Student Type (e.g., degree-seeking foreign students, exchange students, etc.).
        - Academic Year (e.g., 99 學年度 for 2010).
    - The dataset was initially in wide format, with years as column headers.
    - Included many student categories, such as:
        - Short-Term Exchange Students.
        - Degree-Seeking Students.
        - Overseas Chinese Students.
- **Preprocessing Steps:**
    - Isolated specific student categories relevant to the study (e.g., degree-seeking foreign students, overseas Chinese students).

    - Converted the wide format to long format, making it easier to analyze trends by year.
    - Filtered the dataset to retain only totals for 2015–2023.

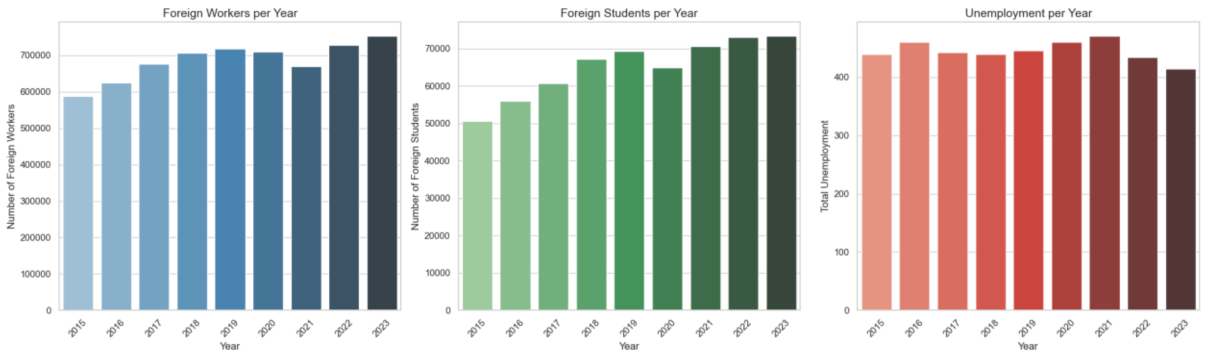
**3. Unemployment Dataset**

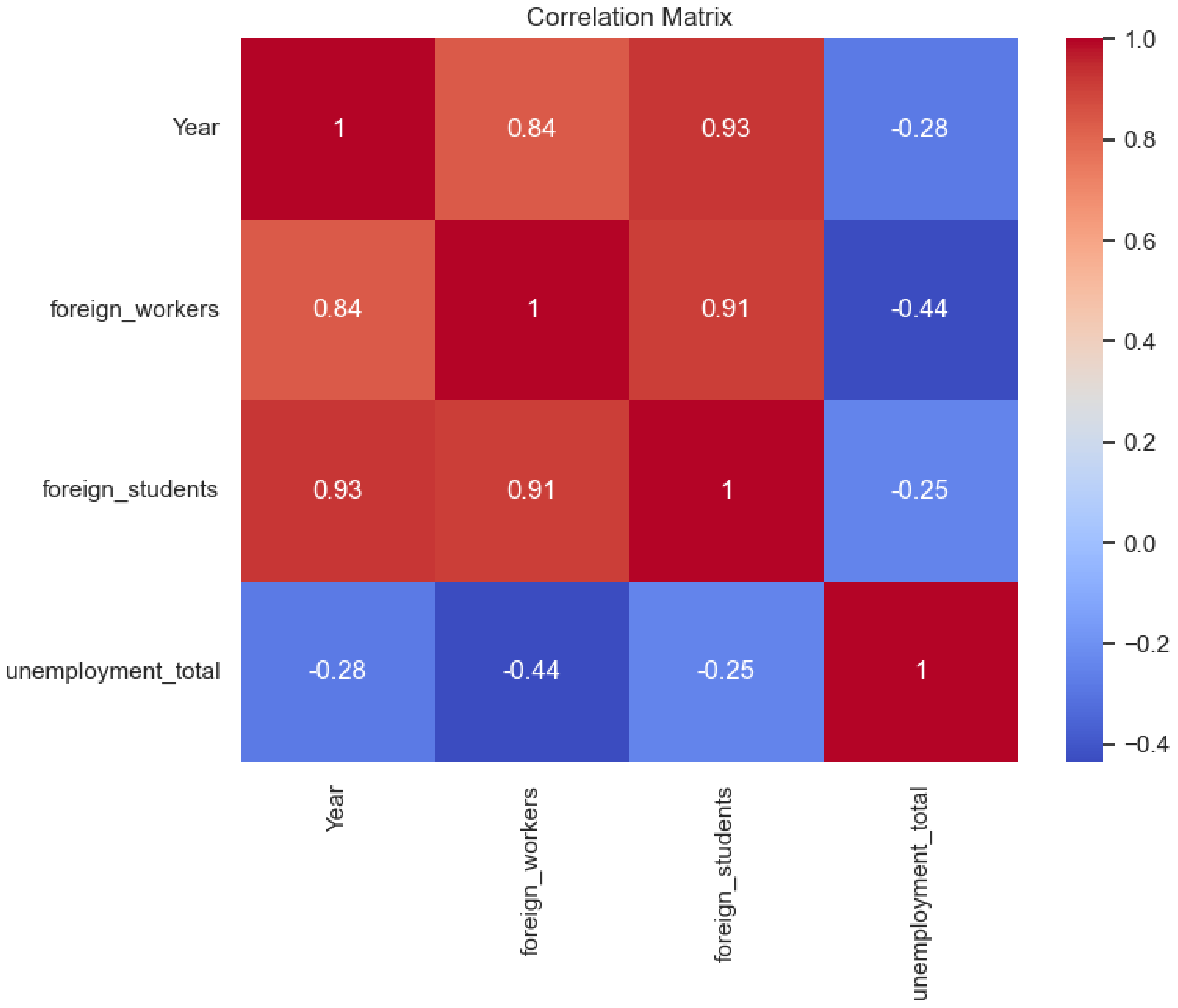
- **Original Contents:**
    - This dataset provided annual unemployment data, broken down by:
        - Reasons for Unemployment (e.g., new job seekers, laid-off workers).
        - Annual Growth Rate (percentage changes from previous years).
    - The dataset included metadata rows, footnotes, and blank rows that made direct analysis challenging.
    - Some years, like 2014 and 2015, were present but misaligned due to inconsistent formatting.
- **Preprocessing Steps:**
    - Extracted the columns for Year and Total Unemployed (in thousands).
    - Calculated or retained the Annual Growth Rate where applicable.
    - Filtered the dataset for the years 2015–2023.

**Key Challenges**

1. **Foreign Workers**: The dataset's complexity arose from having multiple breakdowns (e.g., by nationality, sex, and permit authority) that were irrelevant to this study.
2. **Foreign Students**: The wide format required transformation, and irrelevant categories (e.g., short-term students) had to be removed.
3. **Unemployment**: Misaligned rows and inconsistent formatting required careful inspection to ensure all years were correctly included.

**Visualization：**





## 2.2 Exploratory Data Analysis (EDA)

- **Correlation Check**: A quick correlation matrix suggested that foreign workers and foreign students do not have a simple, direct positive correlation with unemployment; in some slices of the data, foreign workers appeared **negatively** correlated, indicating higher foreign labor might coincide with stable or lower unemployment.

- **Trend Plots**: Basic line charts over years revealed a steady increase in both foreign workers and students, while unemployment did not spike correspondingly.

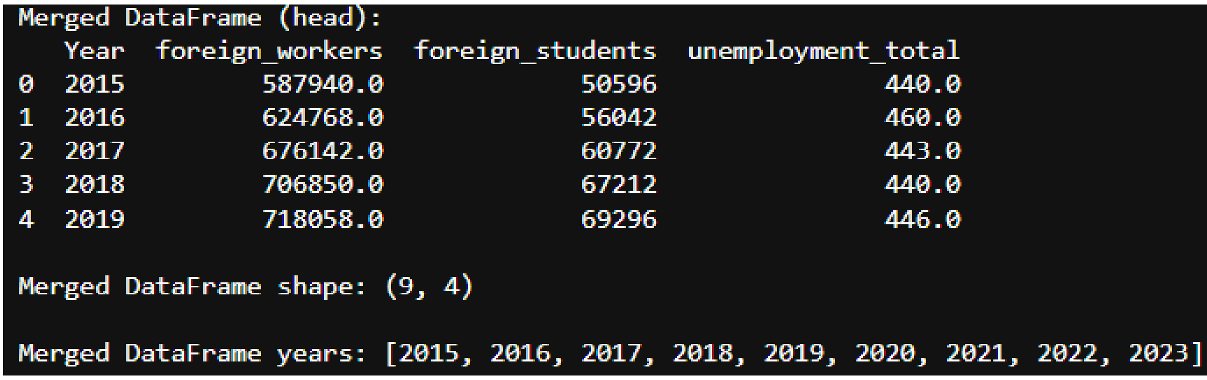
**2.3 Data Preprocessing**

- **Missing Values**:
  - **Foreign Workers Dataset**:
    - Entries without a valid **"Grand Total"** (e.g., rows for specific industries or categories) were excluded.
  - **Foreign Students Dataset**:
    - Missing or irrelevant student categories (e.g., short-term exchange students) were removed.
  - **Unemployment Dataset**:
    - Rows containing non-numeric years or footnotes were dropped to retain only valid unemployment data.
- **Filtering**:
  - Data for years outside the **2015–2023** range was removed to ensure alignment across all datasets.
- **Renaming Columns**:
  - For clarity and consistency:
    - In the **Foreign Workers Dataset**, **"Grand Total"** was renamed to **"foreign_workers"**.
    - In the **Foreign Students Dataset**, the total foreign student row was renamed to **"foreign_students"** after isolating and melting the dataset from wide to long format.
    - In the **Unemployment Dataset**:
      - **"Period"** was renamed to **"Year"**.
      - **"Total"** (total unemployed) was renamed to **"unemployment_total"**.
- **Transformation and Format Changes**:
  - **Foreign Workers Dataset**:
    - Originally contained multiple rows for different worker categories (e.g., by industry, nationality).
    - These were collapsed into a single column summarizing the total number of foreign workers annually.

- **Foreign Students Dataset**:
    - The dataset was initially in a **wide format** with academic years as column headers.
    - It was transformed into a **long format**, with the academic year (converted to a calendar year) in a single **"Year"** column and corresponding totals in a **"foreign_students"** column.
- **Unemployment Dataset**:
    - Rows with reasons for unemployment (e.g., "Laid Off," "Job Seekers") were removed to focus on the overall **total unemployed**.
    - Any rows containing annual growth rates without corresponding totals were excluded.
    - Numerical inconsistencies, such as commas in values (e.g., "4,460"), were cleaned to allow for numeric processing.

- **Final Structure**:

- The datasets were aligned into a consistent format with the following columns:
    - **Year**: The calendar year (2015–2023).
    - **foreign_workers**: Total number of foreign workers per year.
    - **foreign_students**: Total number of foreign students enrolled in Taiwanese universities per year.
    - **unemployment_total**: Total number of unemployed individuals per year (in thousands).



```
Merged DataFrame (head):
   Year  foreign_workers  foreign_students  unemployment_total
0  2015         587940.0             50596               440.0
1  2016         624768.0             56042               460.0
2  2017         676142.0             60772               443.0
3  2018         706850.0             67212               440.0
4  2019         718058.0             69296               446.0

Merged DataFrame shape: (9, 4)

Merged DataFrame years: [2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023]
```

### 2.4 Insights from Data Analysis

- The dataset size remained relatively small (less than 10 yearly data points once merged). This introduced challenges for complex predictive modelling and time-series splits.

- Despite the limited data, preliminary visualization hinted that neither foreign workers nor foreign students strongly "pushes up" unemployment. Instead, some evidence suggests foreign labour might be associated with stable or slightly lower unemployment, though we cannot confirm cause and effect with such limited data.

## 3. Methods and Algorithms

### 3.1 Algorithm Selection

**- Timeseries Split:** Since our data is more stable and is set chronologically, along with our approach in predicting future years, we have to first approach the data as time-series data instead of normal machine learning data.

- **Baseline**: We started with simple regression models (Linear Regression) to quickly assess relationships.

- **Tree-Based Models** (Random Forest, Gradient Boosting, XGBoost, CatBoost, LightGBM): Tree-based methods can capture more complex patterns in a small dataset compared to purely linear models, especially if any nonlinearities exist.

- **Scenario-Based Analysis**: Instead of heavily relying on time-series forecasting with a tiny dataset, we adopted a "what-if" approach—varying foreign worker and student counts to see how the model predicts unemployment might change.

-**Training and Validation Splits:**

**Train: 2015-2022**

**Test: 2023**

### 3.2 Challenges Encountered and Solutions

**- Insufficient Granularity**: We only had annual data for foreign workers and students. Ideally, monthly or quarterly data would help, but such data is not freely available.

- **Solution**: Used scenario-based or "what-if" methods to simulate how changes in annual totals may affect unemployment in future years.

**- Data Scarcity:** Fewer than 10 data points for some years forced us to be cautious about overfitting. We wanted more data points, however due to some of the datasets not providing monthly or quarterly data, this is not feasible unless we were willing to have many assumptions. Cross-validation or time-series splits often led to very small training folds, producing unstable metrics. Moreover, some datasets aren't readily available for free (such as unemployment rates per year), requiring signing up to expensive memberships to access them.

- **Solution:** We recognized the need for more macroeconomic variables or scenario simulations to glean insights rather than purely accurate forecasts.

**-Correlation and Causation:** Our project as of now is only a surface level analysis on what it could be. As such it might not have the most realistic conditions and variables, instead filling them with unrealistic assumptions. For example, since we're only comparing numbers of foreign students and workers with unemployment, we cannot know for sure that unemployment numbers change solely because of foreigners or other factors. We did not take into account other factors, such as natural causes, population growth, etc. We also couldn't use unemployment rates to portray actual unemployment to workforce ratio. We cannot directly prove that either foreign workforce or students have causational effects towards unemployment, even if we found that they indeed have some level of correlation.

## 4. Hyperparameter Tuning

### 4.1 Initial Tuning and Best Hyperparameters

- We used **RandomizedSearchCV** or **GridSearchCV** with a small set of hyperparameter ranges for Random Forest, XGBoost, and other boosting algorithms.

- With so few annual data points, the model's cross-validation scores varied widely; best hyperparameters sometimes did not generalize well due to limited samples.

### 4.2 Handling Hard to Fit Datasets

**Approach**:

1. When cross-validation yielded negative $R^2$ for certain splits, it became clear that the dataset lacked enough variance or size for robust out-of-sample performance.

2. We emphasized a scenario-based approach to glean directional insights rather than strict numeric forecasts.

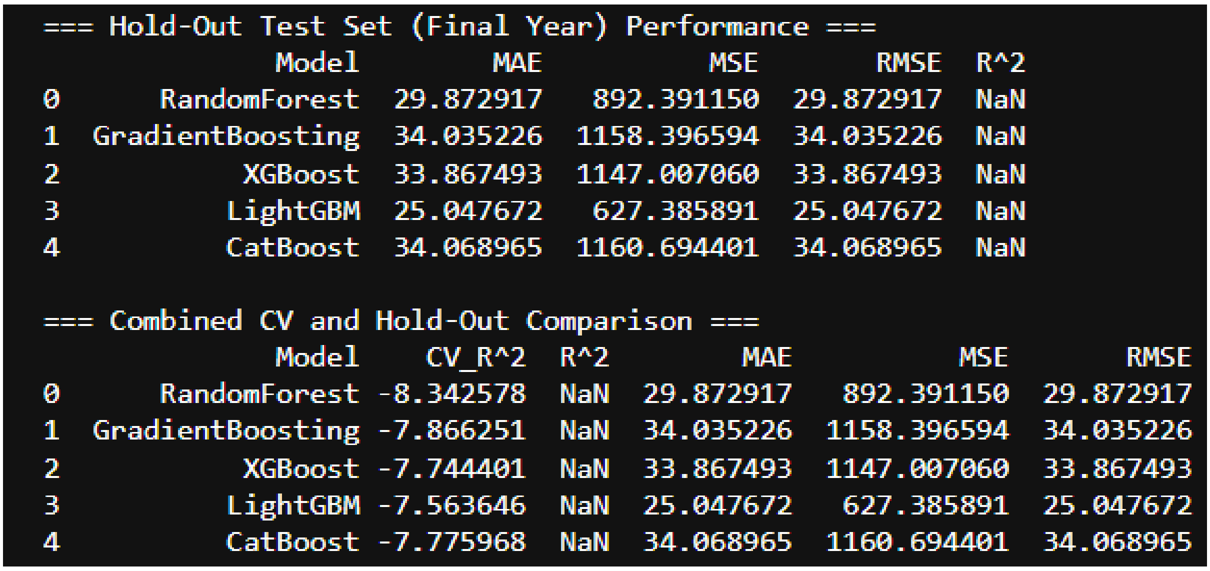
## 5. Performance Evaluation

### 5.1 Model Performance Metrics

- R-squared ($R^2$): Indicated how much variance was captured by the model. Some folds produced negative values due to sample insufficiency. On the entire dataset (train set), we saw an $R^2$ up to ~0.84 for a Random Forest model—indicating decent fit on historical data but likely overfitting.

- MAE / MSE: The best model often had an MAE of roughly 5–6 unemployment units (depending on whether we treat them as thousands of persons or raw counts).

### 5.2 Comparative Analysis

- Random Forest vs. Gradient Boosting vs. XGBoost: Each had similar performance on the small dataset, with minor differences in error metrics. Although we found that Random Forest had the best performance, the main conclusion was that model differences were overshadowed by the dataset's small size and limited features.
- Scenario Analysis: Proved more informative, revealing that decreasing foreign workers or students might raise predicted unemployment, and that foreign workers had a slightly bigger impact in the model's predictions.
- 

```
=== Hold-Out Test Set (Final Year) Performance ===
              Model        MAE          MSE       RMSE  R^2
0      RandomForest  29.872917   892.391150  29.872917  NaN
1  GradientBoosting  34.035226  1158.396594  34.035226  NaN
2           XGBoost  33.867493  1147.007060  33.867493  NaN
3          LightGBM  25.047672   627.385891  25.047672  NaN
4          CatBoost  34.068965  1160.694401  34.068965  NaN

=== Combined CV and Hold-Out Comparison ===
              Model    CV_R^2  R^2        MAE          MSE       RMSE
0      RandomForest -8.342578  NaN  29.872917   892.391150  29.872917
1  GradientBoosting -7.866251  NaN  34.035226  1158.396594  34.035226
2           XGBoost -7.744401  NaN  33.867493  1147.007060  33.867493
3          LightGBM -7.563646  NaN  25.047672   627.385891  25.047672
4          CatBoost -7.775968  NaN  34.068965  1160.694401  34.068965
```

**6. Discussion and Insights**

**6.1 Key Findings**

- **Model Suggests Negative Correlation**: Higher foreign labour enrolment does not appear to spike unemployment. In some scenarios, a drop in foreign workers correlated with a rise in unemployment predictions.

- **Foreign Students**: Increasing or decreasing foreign students had less of an effect compared to foreign workers—though the dataset is too small to claim absolute certainty.

- **No Clear Evidence** that foreigners "take" local jobs in a direct sense based on this annual dataset alone. This was outside our initial hypothesis, as logically speaking, an introduction, especially an increase in outsourced workers should lead to more job opportunities taken.

**6.2 Interpretation of Results**

- **Potential Economic Explanation**: Foreign labour may fill positions that locals do not want or cannot fill, hence not raising overall unemployment. Foreign students may also contribute tuition and consumption to local economies, having no major negative effect on local job markets. Aside from that, foreign students might also join the foreign labour force, however since not all foreign students are guaranteed to join the workforce, their impact on unemployment shifts aren't as big.

- **Taiwan's Job Opportunities:** Based on the results we have gotten, we can reasonably assume that the job opportunities provided in Taiwan can keep up with the increasing amount of foreigners coming in, at least for now.

- **Caution**: The results are heavily data-dependent and do not account for macroeconomic shifts (e.g., global recessions, industrial restructuring).

### 6.3 Limitations and Future Work

**- Data Coverage:** Limited to annual points for roughly 8–9 years, insufficient for robust time-series forecasting.

**- Lack of Key Macroeconomic Features:** GDP growth, wages, inflation, or sectoral breakdown could reveal more complex relationships.

**- Generalization:** Negative or small $R^2$ in cross-validation indicates the model might not generalize well to truly new conditions.

**- Future Work:**

- Collect or request more granular (monthly/quarterly) data from government agencies or educational institutions.
- Include additional macro variables (GDP, labour force participation, immigration policy changes).
- Explore formal time-series forecasting or advanced scenario planning with a bigger dataset.

## 7. Conclusion and Recommendations

### 7.1 Conclusion

This project analysed the link between foreign workers, foreign students, and unemployment in Taiwan. While the best model (Random Forest) performed adequately on the small historical dataset ($R^2$ up to ~0.84 in-sample), the limited data poses challenges for definitive conclusions. Nonetheless, scenario analyses consistently showed that **reducing** foreign workers or students (particularly workers) leads to **higher** unemployment predictions, suggesting these international populations may be **neutral or slightly beneficial** to local unemployment rates rather than detrimental.

### 7.2 Recommendations

- **Scenario-Based Analyses**: Given the small dataset, scenario modelling ("what if +10% foreign workers?") proved more insightful than raw forecasting. We recommend continuing or expanding this approach for policy debates or university planning.

- **Expand the Dataset**: Seek additional monthly/quarterly labour data or macroeconomic features to strengthen modelling.
- **Policy Caution**: With limited data, we cannot claim absolute proof that foreign labour lowers unemployment, but results do challenge the argument that foreign workers or students "steal jobs."
- **Proxy Variables**: If direct data on monthly foreign workers is unavailable, look for **proxy** or **correlated** data series that **track** or **approximate** foreign labour trends.
- **Broaden Feature Set:** If you cannot get finer-grained foreign worker data, you can **counterbalance** the limited detail by adding other **macroeconomic** or **demographic** variables known to impact unemployment, such as GDP, CPI, etc.
- **Synthetic Data Generation:** Creating a **synthetic** dataset that simulates monthly/quarterly trends for foreign workers based on known patterns or assumptions (growth curves, seasonal factors, etc.)

## 8. References

1. Taiwan Ministry of Labor (statistics website), for foreign worker counts and unemployment data.
   https://statdb.mol.gov.tw/html/mon/i0120020620e.htm
   https://nstatdb.dgbas.gov.tw/dgbasall/webMain.aspx?k=engmain
2. Ministry of Education Taiwan for foreign student enrolment data.
   https://data.gov.tw/en/datasets/42158?utm_source=chatgpt.com
3. International student mobility and labour market outcomes: an analysis
   https://link.springer.com/article/10.1007/s10734-020-00532-3
4. Studying Abroad and the Effect on International Labor Market Mobility
   https://docs.iza.org/dp3430.pdf
5. Erasmus Student Mobility as a Gateway to the International Labour Market?
   https://link.springer.com/chapter/10.1007/978-3-658-02439-0_13
6. Studying abroad and the effect on international labour market mobility
   https://ifs.org.uk/publications/studying-abroad-and-effect-international-labor-market-mobility-evidence-introduction
7. Does the effect of studying abroad on labour income vary by graduates' social origin?
   https://link.springer.com/article/10.1007/s10734-020-00579-2
8. Artificial intelligence and unemployment dynamics: an econometric analysis in high-income economies
   https://www.emerald.com/insight/content/doi/10.1108/techs-04-2024-0033/full/html
9. Predicting the contribution of artificial intelligence to unemployment
   https://link.springer.com/article/10.1007/s12197-023-09616-z
10. Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes
    https://www.nber.org/papers/w28399

**11.** Unemployment Rate Forecasting Using Supervised Machine Learning Model
https://rjpn.org/ijcspub/papers/IJCSP22D1055.pdf

**12.** Can Machine Learning on Economic Data Better Forecast the Unemployment Rate?
https://digitalcommons.oberlin.edu/honors/126/