Project Proposal

2466025 Park Hyerin, 2135017 Lee Eunju, 2329006 Kim Seungwon,
2391028 Jo Seungyeon, 2380022 Park Jieun

1. Business Value

1.1 Background

Most countries experience changes in employment rates due to a range of factors, including economic growth, infectious diseases, and demographic shifts. Among these, aging populations, education levels, and healthcare infrastructure play particularly significant roles in shaping labor market outcomes. By examining the employment strategies and policy responses of various nations, we can identify valuable insights for improving South Korea's employment environment. This analysis aims to compare and evaluate the key factors affecting employment across these countries and based on the findings, propose strategic directions for fostering a sustainable and resilient employment environment in South Korea. Additionally, understanding how different countries have adapted to recent global challenges—such as pandemics and automation—can further inform South Korea's approach to long-term workforce development.

1.2 Problem Definition & Value Proposition

South Korea is facing a critical challenge with persistently low employment rates, which poses a significant risk to the country's economic and social stability. Factors such as demographic shifts, including an aging population and declining birth rates, further exacerbate the problem. Additionally, global economic uncertainties, such as the impact of the pandemic and financial market volatility, have intensified labor market vulnerabilities.

To address this issue, it is crucial to gain insights from other advanced economies that have demonstrated resilience and adaptability in their labor markets. By analyzing the socioeconomic indicators from countries with higher employment rates, we can identify best practices and strategies that may be applicable to South Korea.

Our data, sourced from the World Bank's official World Development Indicators (WDI) database, includes comprehensive information on employment, education, health, and population dynamics from 20 advanced economies, covering the period from 2011 to 2021. This dataset provides valuable insights into labor market structures, educational investments, healthcare infrastructure, and youth engagement, among other critical factors.

By leveraging comparative analysis, we aim to extract actionable insights that can inform policy recommendations and strategic interventions to improve South Korea's employment rates. This cross-national analysis will help identify effective approaches to fostering workforce participation, reducing youth unemployment, and enhancing labor market resilience in the face of socio-economic challenges.

2. Available Data

2.1 Data Acquisition

This dataset was collected by Kaggle and includes information on population, education, healthcare, and employment in the developed countries in the top 19 countries of the WDI. Based on the current situation about the low employment rate in South Korea, this dataset was selected to analyze the relationship between various variables and the employment rate and observe which factors affect how.

2.2 Data Review

The dataset is structured with countries as rows and socioeconomic factors as columns. The columns include diverse variables such as health (medical), education level, education expenditure, and population structure, all of which can influence employment rates.

In the case of the United Arab Emirates and Japan, the number of NaN values was too high that we decided to delete these two rows.

After applying linear interpolation, any columns with remaining NaN values were removed.

We categorized the factors and created heatmaps for education and employment rates, health and employment rates, and demographic structure and employment rates. Based on these heatmaps, we then generated a correlation bar chart.

3. Formulation

3.1 Choice of Algorithms and Rationale

We are considering several algorithm options to extract actionable insights from the data. However, the final choice of model will be determined after further analysis of the dataset and team discussions.

1. Regression (Supervised Learning): For analyzing the relationship between socioeconomic factors and employment rates, we suggest panel data regression and consider Ridge/Lasso regression as potential tools.

- Panel Data Regression: Useful for handling the cross-sectional and temporal nature of the data, accounting for country-specific and yearly effects. We will explore both Fixed Effects and Random Effects models.
- Ridge/Lasso Regression: These methods may help in case of multicollinearity or handling high-dimensional

features. However, these methods will be considered based on further exploration of the dataset.

2. Clustering (Unsupervised Learning): To group countries with similar labor market characteristics as South Korea, we employ clustering algorithms, such as K-Means or Hierarchical Clustering. This step helps to:

- Identify groups of countries with similar employment patterns and socioeconomic profiles.
- Pinpoint benchmark countries that share characteristics with South Korea, helping to guide policy comparison and adaptation.
- Analyze the temporal stability of these clusters to understand how countries' labor market dynamics evolve over time.
- Incorporate time-series clustering techniques (e.g., Dynamic Time Warping) to better capture temporal trends in employment rates.

For the rationale, while we have several potential algorithms in mind, the final choice will depend on the specific characteristics of the data and the insights we aim to derive. We will choose the most suitable approach based on:

- The goals of the analysis (e.g., understanding patterns, predicting employment rates).
- The need to handle challenges such as multicollinearity, high-dimensionality, or feature selection.

3.2 Input & Output Variables

Input variables for the prediction model include:

- T_public_edu_exp
- F_bachelor_25p
- M_bachelor_25p
- T_bachelor_25p
- F_master_25p
- M_master_25p
- T_master_25p
- F_doctoral_25p
- M_doctoral_25p
- T_doctoral_25p
- RD_expenditure
- sci_tech_journals
- health_expgov_health_exp
- hospital_beds
- physicians_per_1K
- nurse_midwife_per_1K
- annual_population_growth
- F_population
- M_population
- T_population
- F_survival_to_65
- M_survival_to_65

Output variable for the prediction model is Employers, total

3.3 Expected Challenges and Alternatives

Handling missing data posed a significant challenge due to the extensive NaN values across our dataset. Given the temporal and geographical correlations, simple imputation methods like mean or mode filling were judged to be statistically inappropriate. Initial attempts to source missing data from national statistics, OECD, World Bank, and other organizations yielded limited results and modeling the missing values using other indicators was considered but also found to be inefficient. Consequently, we used linear interpolation where there were over half indicators in a specific column of a country by year, and used polynomial interpolation where there were fewer. In common with many of the 19 countries, columns without indicators for all 10 years were inevitably deleted, and rows corresponding to the United Arab Emirates and Japan were also deleted where almost all columns were empty overall.

Another expected challenge is the scope of modeling. The limited data—only 10 annual observations per country from 2011 to 2021—raises concerns about the robustness of the analysis. Additionally, attempting to model all countries simultaneously may obscure year-to-year trends, complicating the identification of meaningful patterns. Therefore, we are planning to adopt regularized regression techniques, specifically Ridge regression (L2 regularization) and Lasso regression (L1 regularization). These methods are particularly suited to our dataset, which includes a high number of features and potential multicollinearity among them. Regularization can help prevent overfitting and enhance model stability. However, a key concern with applying Ridge or Lasso regression is that these techniques may overlook temporal and country-specific characteristics. Since they focus on identifying global patterns across the dataset, they may not adequately reflect time-sensitive trends or the unique economic or social contexts of individual countries such as Korea. This could limit the effectiveness in achieving our primary goal. To address this limitation, we are currently discussing alternative or complementary modeling approaches within our team. These may include incorporating time-series models, panel data regression, or country-level clustering to better capture year-to-year trends and country-specific behaviors. The approaches will be further explored and evaluated as the project progresses.