Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

# Statistical Data Analysis of Legatum Prosperity Index and Presentation of Prediction Models from Data Mining Process

**Ali Naghash Asadi[1,*], Amin Mastourinejad[1], Romina Aghili[1]**

[1] *Fouman Faculty of Engineering, College of Engineering, University of Tehran, Tehran, Iran*
**Corresponding author: naghashasadi@ut.ac.ir*

*Received: 2024-08-02 Revised: 2025-06-03 Accepted: 2025-08-25*

## Abstract

Data mining is a process to discover patterns from a dataset and it can play an important role in understanding the multifaceted aspects of development on a global scale. Policymakers and researchers can formulate more effective strategies to promote sustainable development by identifying key indexes of prosperity, their interrelationships, and the impact of each index on other indexes. In this paper, first, 12 indexes of prosperity provided by Legatum institute are introduced. Then the values of these indexes are collected for 167 countries and statistical analyzes are presented. In the following, a heat map of the mutual influence of indexes is presented, which shows the influence of each index on other indexes. Finally, prediction models are presented for each index using regression and neural network methods, and they will be evaluated based on different measures. The values of the evaluation measures prove that the results from the prediction models were up to 98% similar to the actual results in the best conditions. Therefore, prediction models can determine how much other indexes will change with the improvement of one index. Based on this, policymakers and researchers can plan and prioritize to improve the status of indexes.

## Keywords:

Statistical data analysis, Legatum prosperity index, Prediction models, Neural network, Regression.

## 1. Introduction

Organizations and governments generate and maintain large amounts of data related to economic development, social conditions, public health, education, environmental quality, governance, and other dimensions of national prosperity. In addition, international organizations and institutions publish extensive datasets that enable countries to be compared using standardized indexes. Despite the availability of these data, policymakers may face difficulties in converting large volumes of information into actionable knowledge [1]. Data mining can address this challenge by extracting meaningful patterns, relationships, and predictive models from existing datasets [2].

The multidimensional nature of prosperity makes this issue particularly important. Improvements in one aspect of development may be associated with changes in other aspects, and therefore evaluating individual indexes independently may not provide sufficient information for effective policy planning. Understanding the relationships among prosperity indexes can help policymakers identify areas that deserve greater attention and determine potential priorities for improving overall national prosperity.

The **Legatum Prosperity Index** provides a comprehensive framework for evaluating countries using 12 indexes [3]: **Safety Security**, **Personal Freedom**, **Governance**, **Social Capital**, **Investment Environment**, **Enterprise Conditions**, **Infrastructure & Market Access**, **Economic Quality**, **Living Conditions**, **Health**, **Education**, and **Natural Environment**. These indexes cover political, social, economic, infrastructural, health-related, educational, and environmental dimensions of prosperity. The availability of data for a large number of countries makes this index a suitable basis for statistical analysis and data mining.

Previous studies have applied data mining, statistical analysis, and machine learning techniques to financial, economic, and development-related problems. However, based on the review presented in the original study, relatively few studies have simultaneously analyzed the 12 global prosperity indexes and developed predictive models that estimate one index from the others [4] [5]. Some studies have examined the geographical or statistical differences among prosperity indexes without developing prediction models, while other data mining studies have focused on financial and economic variables with different objectives.

The main objective of this study is therefore to analyze the relationships among the 12 prosperity indexes for 167 countries and develop predictive models capable of estimating each index based on the remaining indexes. The main contribution of the study is the integration of statistical analysis, Pearson correlation analysis, linear regression, and artificial neural networks into a unified data mining framework for the analysis and prediction of global prosperity indexes. This framework can provide information about the indexes most strongly associated with a target index and can consequently assist policymakers and researchers in planning and prioritizing development strategies.

## 2. Methodology

The dataset used in this study consists of the values of the 12 Legatum prosperity indexes for 167 countries [3]. The Legatum Institute assigns a score between 0 and 100 to each index and calculates an overall prosperity score based on the indexes. The study analyzes the available data, with particular attention to the 2023 values and the changes in country rankings between 2018 and 2023.

Two complementary data mining approaches were employed. First, descriptive statistical analysis was used to investigate the distribution and relative status of the indexes. Second, predictive data mining was applied to estimate each target index from the other indexes. Linear regression and artificial neural networks (ANNs) were selected as the prediction methods. In addition, a Pearson correlation matrix was calculated for the 12 indexes and the overall prosperity score. The resulting matrix was visualized as a heat map to provide an overview of the relationships among indexes and to identify the strongest relationships associated with each target index. The Weka software [6] was used to develop the prediction models. The dataset was divided into training and testing subsets, with 80% of the observations used for training and the remaining 20% used for testing. Model performance was evaluated using three measures: correlation coefficient (CC), mean absolute error (MAE), and root mean square error (RMSE). A higher CC and lower MAE and RMSE indicate closer agreement between predicted and actual values.

## 3. Discussion and Results

The statistical analysis shows considerable differences in the distribution of prosperity indexes among countries. The analysis of Iran provides a specific example of these differences. According to the 2023 data, Iran ranked 126th overall. Among the 12 indexes, Iran had relatively better positions in Health, Education, and Living Conditions, while its weaker positions were associated with Personal Freedom, Business Conditions, and Environment. The ranking analysis for 2018–2023 also indicates changes in the relative position of Iran across different indexes.

According to Figure 1, the correlation-based heat map provides important information about the relationships among the indexes. Investment Environment had the strongest overall relationship with the other indexes and showed a Pearson correlation coefficient of 0.95 with the overall prosperity score. In contrast, Environment showed the weakest overall relationship with the other indexes and the overall score. These results suggest that economic and institutional conditions represented by the Investment Environment are closely associated with several dimensions of prosperity.

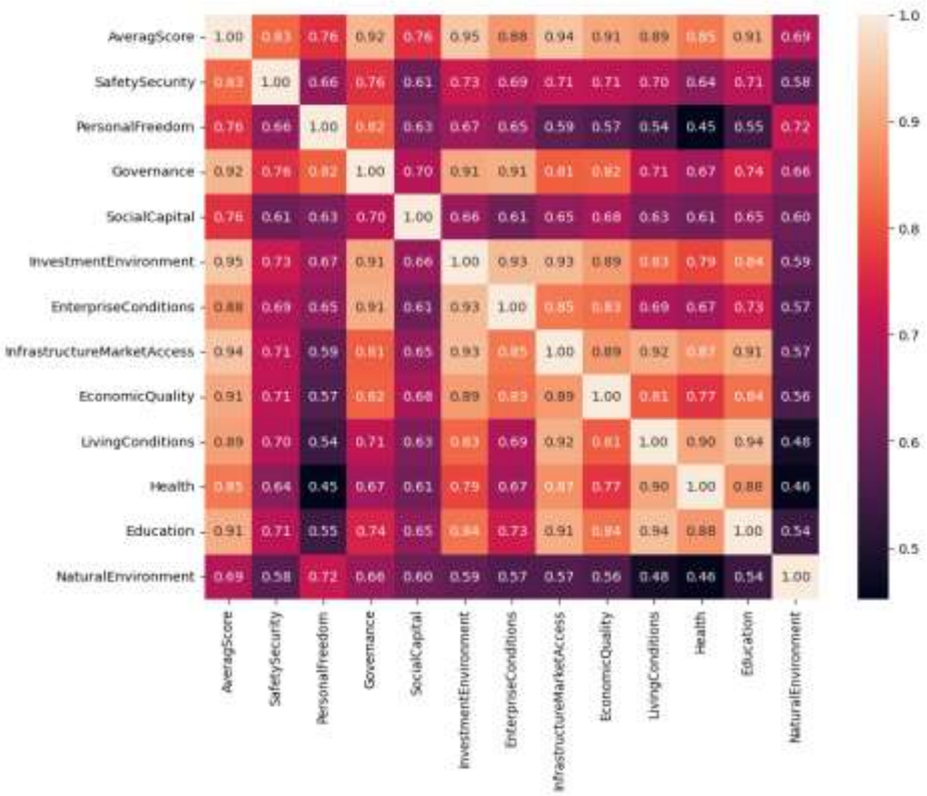
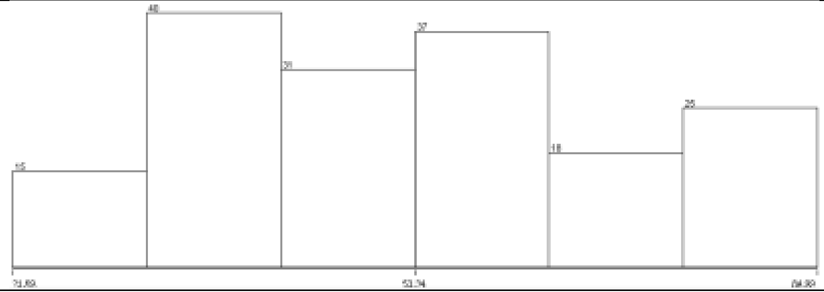



**Figure 1. Heat map diagram of 12 indexes and average score**

The prediction models further demonstrate the feasibility of estimating each prosperity index from the remaining indexes. For all target indexes, both linear regression and artificial neural network models produced acceptable results according to the evaluation measures. For example, Table 1 shows statistical information and prediction models for the Investment Environment. Based on CC, the performance of the developed models ranged from approximately 0.75 to 0.98. Thus, in the best-performing cases, the predicted values showed up to 98% correlation with the actual values.

**Table 1. Statistical Information and Prediction Models for the Investment Environment Index**

| **Minimum Value** | 21.69 | **Maximum Value** | 84.99 |
|---|---|---|---|
| **Frequency Distribution** |  | | |
| **Regression Equation** | InvestmentEnvironment = 0.2096 * Governance + 0.3938 * EnterpriseConditions + 0.4387 * InfrastructureMarketAccess + 0.1079 * EconomicQuality - 8.8112 | | |
| **Most Influential Indicator According to the Regression Model** | | | Infrastructure & Market Access |
| **Evaluation Measures** | **CC** | **MAE** | **RMSE** |
| **Regression** | 0.9798 | 2.9221 | 3.5819 |
| **Neural Network** | 0.9768 | 3.2889 | 4.0239 |

The results also reveal some differences between the correlation analysis and the regression models. For example, Governance was identified through the heat map as the strongest associated index with Safety Security, and the regression model produced the same result. A similar consistency was observed for Personal Freedom, for which Governance was identified as the strongest predictor. Investment Environment and Enterprise Conditions were the strongest associated indexes for Governance, and the regression analysis confirmed this result.

For Social Capital, however, the heat map identified Governance as the strongest associated index, whereas Economic Quality was identified as the strongest predictor in the linear regression model. Similarly, Education was the strongest associated index for Living Conditions in the correlation analysis, while Infrastructure & Market Access was identified as the strongest predictor in the regression model. For Natural Environment, Personal Freedom showed the strongest Pearson correlation, whereas Infrastructure & Market Access was identified by the regression model as the strongest predictor.

These differences demonstrate the value of combining statistical correlation analysis with predictive modeling. Correlation analysis provides information about pairwise relationships, whereas regression and neural network models consider multiple input indexes simultaneously. Consequently, using both approaches provides a broader understanding of the structure of the prosperity indexes. Among the individual prediction results, the Social Capital models achieved CC values of 0.75 for linear regression and 0.83 for the neural network. For Enterprise Conditions, the corresponding CC values were 0.95 and 0.96, respectively. These results indicate that model performance varies according to the target index and that artificial neural networks can outperform linear regression for some indexes.

## 4. Conclusions

This study presented a data mining framework for statistical analysis and prediction of global prosperity indexes. Data corresponding to the 12 Legatum prosperity indexes for 167 countries were analyzed, and relationships among the indexes were investigated using Pearson correlation coefficients and a heat map. Subsequently, linear regression and artificial neural network models were developed to predict each index using the other 11 indexes as inputs.

The findings indicate that Investment Environment has the strongest overall relationship with the other prosperity indexes and with the overall prosperity score, whereas Natural Environment has the weakest overall relationship. The analysis also identified the strongest associated indexes for individual dimensions of prosperity, including Governance for Safety Security and Personal Freedom, Investment Environment for Enterprise Conditions and Infrastructure & Market Access, and Living Conditions for Health and Education.

The predictive models produced CC values between 0.75 and 0.98, demonstrating a substantial agreement between predicted and actual values. The models can therefore be used as analytical tools for estimating potential changes in one prosperity index based on changes in other indexes. Such information may assist policymakers and researchers in identifying priority areas and evaluating possible development strategies.

However, the findings should be interpreted as statistical relationships rather than direct causal effects. A major limitation of the present study is that the prediction models are primarily based on data for a specific period. As future work, data from multiple consecutive years can be incorporated to develop time-dependent prediction models. Such models may reduce the dependence of the results on a particular year and improve the robustness and reliability of the findings. In addition, other data mining and machine learning methods can be investigated and compared to determine whether higher predictive accuracy can be achieved.

## 5. References

[1] Schuh, Günther, Christina Reuter, Jan-Philipp Prote, Felix Brambring, and Julian Ays. "Increasing data integrity for improving decision making in production planning and control." CIRP Annals 66, no. 1 (2017): 425-428.

[2] Schuh, Günther, Gunther Reinhart, Jan-Philipp Prote, Frederick Sauermann, Julia Horsthofer, Florian Oppolzer, and Dino Knoll. "Data mining definitions and applications for the management of production complexity." Procedia Cirp 81 (2019): 874-879.

[3] "Rankings : Legatum Prosperity Index 2023," Legatum Institute, [Online]. Available: https://www.prosperity.com/rankings. [Accessed 28 07 2024].

[4] Budsaratragoon, Pornanong, and Boonlert Jitmaneeroj. "Reform priorities for prosperity of nations: The Legatum Index." Journal of Policy Modeling 43, no. 3 (2021): 657-672.

[5] Puente-López, José Luis, Jenny Paola Lis-Gutiérrez, and Jhonathan Steven Pulido-Flórez. "The Legatum Prosperity Index and non-cooperative tax jurisdictions (2021)." Procedia Computer Science 203 (2022): 514-519.

[6] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.