Working Paper 2025.1.2.2
- Vol 1, No 2

ỨNG DỤNG MÔ HÌNH TRÍ TUỆ NHÂN TẠO TRONG VIỆC LỰA CHỌN ĐẶC TRƯNG: CHỈ SỐ HIỆU SUẤT HỒI QUY VÀ CÁC YẾU TỐ KINH TẾ

Phạm Thị Anh Thư[1], Hồ Minh Trung, Trần Ngọc Bảo Trâm

Sinh viên K61 – Kinh tế đối ngoại

Trường Đại học Ngoại thương Cơ sở II, TP Hồ Chí Minh

Đặng Lê Quan

Giảng viên Cơ sở II

Trường Đại học Ngoại thương Cơ sở II, TP Hồ Chí Minh

Tóm tắt

Nghiên cứu này tập trung vào cách khai thác dữ liệu kinh tế thực tế để xác định các yếu tố kinh kế ảnh hưởng đến hiệu suất Logistics, được đo lường bằng chỉ số Hiệu suất Logistics (LPI). Phương pháp phân tích sử dụng các kỹ thuật học máy (ML) tiên tiến kết hợp với các phương pháp lựa chọn đặc trưng cho mô hình dự báo và hồi quy dựa trên các yếu tố kinh tế có liên quan. Mục tiêu chính là xác định bộ các yếu tố kinh tế tối ưu nhất để dự đoán hiệu suất logistics của một quốc gia. Ngoài ra, các thuât toán hồi quy khác cũng được nhóm tác giả thử nghiệm nhằm nâng cao độ chính xác của dự báo. Các kỹ thuật được lựa chọn bao gồm các phương pháp lọc dựa trên tương quan và phân tích thành phần chính (PCA), cùng với các phương pháp như hồi quy LASSO và Elastic-net. Các phương pháp khác như kiểm định ANOVA F-test, loại bỏ đặc trưng đệ quy (RFE) và phương pháp cây quyết định cũng được thử nghiệm nhưng cho ra kết quả ít ý nghĩa hơn so với PCA. Dựa trên các đặc trưng được chọn từ PCA, biến phụ thuộc (LPI) được dự đoán bằng các phương pháp hồi quy như mô hình Cây Quyết định, XGB Regressor, K-Nearest Neighbors, Hồi quy Rừng ngẫu nhiên, MLP (Multi-Layer Perceptron) và Hồi quy Máy Vectơ Hỗ trợ (SVM). Hiệu suất mô hình được đánh giá qua các chỉ số MAE, MAPE, RMSE, R² và R² điều chỉnh. Kết quả chỉ ra rằng bộ đặc trưng PCA và Elastic-net cung cấp hiệu suất đáng tin cậy nhất dựa trên các tiêu chí đo lỗi. Một chiến lược phù hợp nhất được áp dụng

[1] Tác giả liên hệ, Email: k61.2211115114@ftu.edu.vn

để tinh chỉnh lựa chọn kết hợp với các bộ đặc trưng mang lại kết quả tốt nhất. Các phát hiện cho thấy rằng các thuật toán học máy hỗ trợ hiệu quả trong việc lựa chọn các yếu tố kinh tế có liên quan nhất để đánh giá hiệu suất logistics của một quốc gia. Hơn nữa, nghiên cứu cũng chỉ ra rằng Rừng ngẫu nhiên là mô hình dự đoán hiệu quả nhất.

Keywords: Lựa chọn đặc trưng, hồi quy học máy, chỉ số Hiệu suất Logistics, thuộc tính kinh tế.

AN APPLICATION OF ARTIFICIAL INTELLIGENCE MODELING TO FEATURE SELECTION: THE LOGISTIC PERFORMANCE INDEX AND ECONOMIC FACTORS

Abstract

This study highlights how to leverage real-time dynamic economic big data to identify key economic factors influencing logistics performance, as measured by the Logistics Performance Index (LPI). The analytical approach utilizes advanced machine learning (ML) techniques and feature selection methods for predictive modeling and regression using relevant economic attributes. The primary objective is to determine the optimal set of economic indicators that best predict a country's logistics performance. Additionally, various ML regression algorithms are explored to enhance prediction accuracy. Feature selection techniques include correlation-based filter methods and principal component analysis (PCA), alongside embedded methods such as LASSO and Elastic-net regression. Other methods like the ANOVA F-test, Recursive Feature Elimination, and tree-based approaches are also tested but yield less significant results compared to PCA. Based on the selected PCA features, the dependent variable (LPI) is predicted using Decision Tree Regression, XGB Regressor, K-Nearest Neighbors, Random Forest Regressor, Multi-Layer Perceptron, and Support Vector Machine regressions. Model performance is assessed using MAE, MAPE, RMSE, R², and adjusted R² metrics. The results indicate that PCA and Elastic-net feature sets provide the most reliable performance based on error measurement criteria. A feature union and intersection strategy are applied to refine the selection, with the union of feature sets yielding the best outcomes. The findings suggest that ML algorithms effectively aid in selecting the most relevant economic factors for assessing a country's logistics performance. Moreover, the study identifies Random Forests as the most effective prediction model.

Keywords: Feature selection, machine learning regression, Logistics Performance Index, economic attributes.

1. Introduction

The LPI is an often-cited instrument that provides critical information for policymakers to evaluate a country's logistics performance (World Bank, 2018). It is now an integral instrument in trade facilitation alongside other tools that target the development of the global economy. The World Bank data reveals an extensive analysis that assists policymakers in identifying opportunities to enhance the global supply chain business including the efficiency of customs, trade, and transportation infrastructure (Gerschberger et al., 2017). While it is helpful to condense and analyze such information, there are critical benefits to having uninterrupted, real, and current information on a country’s logistics efficiency. This data would enable the monitoring of changes to various factors, analysis of certain trends, and supporting logistics performance evaluation predictions. This system would help policymakers access advanced estimates on performance levels in a timely manner which would help bolster the country’s logistics and supply chain potentials. Other studies indicate that new institutional or resource-based policies are highly likely to enhance logistics performance. As Wong et al. (2018) claim, the presence of low corruption and a stable political environment within a country tend to be associated with a higher level of logistics performance. Furthermore, changes in resource availability such as infrastructure, technology, labor, and education greatly improve logistics performance and contribute to increasing a country’s competitiveness.

World Bank (2018) underscores the point that a weak government or social discontent may negatively impact performance. However, the factors analyzed in earlier research are primary in nature and some aspects were captured through a survey. After those studies, an important correlation has developed between a country's logistics performance and other economic indicators, which is captured by the LPI scores. D’Aleo et al (2017) showed the correlation between logistics performance and various economic phenomena, including, GDP per capita, the volume of exports and imports, and economic growth. These LPI components are found to have a substantial positive influence on the expansion of international trade, both for imports and exports (Takele 2019). Still, important economic variables that affect logistics performance are known to exist but have not been widely researched. This research intends to use economic big data with the purpose of determining economic variables, which exhibit logistics performance according to their LPI. The proposed analytical approach incorporates machine learning (ML) techniques with an emphasis on prediction and regression modeling in relation to some selected economic features. The accuracy of ML predictions is based on the model structure, training algorithm, and to a certain degree the feature space that is built out of the initial feature set and the feature analysis algorithm (Chandrashekar et al., 2014). Preprocessing in ML applications often includes feature selection, which is the extraction of a subset of features that provide the highest predictive value and thus remove information-poor variables (Vieira et al., 2010). This paper sets out to accomplish two primary aims: (1) the identification of selected economic features that most represent the predicted variable in forecasting a country’s logistics performance, and (2) increase prediction accuracy by using different ML regression algorithms. This study attempts to address two critical issues: first, can ML algorithms be utilized to identify selected appropriate subsets of economic features that represent a nation’s logistics performance? Second, what type of ML regression is optimal to use on logistics performance with specific economic indicators?

The structure of this paper is as follows: Section 2 reviews relevant literature on feature selection and ML regression techniques. Section 3 presents the methodology, including the feature selection process, data sources, data preparation, analysis, and parameter configuration. Sections 4 and 5 provide the results and discussion, followed by conclusions and directions for future research.

2. Literature review

2.1. Machine Learning and Decision-Making Policy

The quick spread of machine learning (ML) in the era of big data greatly enhances decision-making ability in many different fields. Jordan & Mitchell (2015) underlined in their careful analysis that the integration of machine learning into data-intensive environments helps to extract complex patterns and predictions from large datasets, therefore affecting fields including urban planning and healthcare. Machine learning models, a subset of artificial intelligence, integrate many concepts by using the exponential growth of data (Zhu et al., 2021). Machine learning methods focus on how computers replicate human learning behaviors to acquire new information and improve predictive accuracy over time. Predictive or classification analytics is an essential function of machine learning. The core concept of machine learning is to employ computational algorithms to comprehend and derive insights from data. Fresh data facilitates machine learning algorithms to extend previously learned knowledge to provide predictions, hence enabling decision-making in new settings (Ray & Chaudhiri, 2021). While simultaneously posing issues relating to data privacy, security, and ethical algorithmic governance, this ability improves outcomes by process efficiency. Consequently, there is a growing necessity for comprehensive regulatory frameworks to tackle these concerns, guaranteeing that machine learning applications in decision-making are both efficient and ethically principled.

The role of ML learning in decision-making policies is critical due to its ability to uncover patterns and insights from large datasets that human analysts might miss. This is an important foundation for creating superior commercial value and comprehensive economic development, especially in the era of big data is the key since it provides the correlations between data inputs and decision outputs (Coyle & Weller, 2020). Furthermore, good application of ML in developing policies not only enhances the capacity for national level operation but also for the corporate level (Souza et al., 2019). Application of machine learning (ML) to public policy-making has demonstrated interesting outcomes in the ever-changing field. In this context, Kreif et al., 2022 shown how ML may assist government officials can review of evaluation of prior health insurance expansions are able to maybe guide the re-design of the eligibility criteria for subsidized health insurance in Indonesia. Furthermore, in the field of energy management, Kumar et al. (2023) investigated the benefits and drawbacks of using machine learning for energy optimization in smart homes. This study suggests using the Stochastic Gradient Descent (SGD) algorithm to maximize energy use in smart homes, however several obstacles remain, such as data privacy, data gathering accuracy, and cost, which can impede broad adoption of the technique. According to Ranjan et al. (2022), several years of trading and the growing popularity of Bitcoin have attracted significant attention from society, especially economic policymakers, in the efficacy of ML algorithms for predicting Bitcoin prices. The design of the ML system produces policies continuously as they adapt and grow over time (Mulligan & Bamberger, 2019). Besides, to fully leverage the potential benefits of AI for SCM, Min, H. (2009) revealed numerous sub-fields of AI that are most suitable for tackling real challenges important to SCM. In doing so, this article examines the history of successful applications of artificial intelligence to supply chain management and identifies the areas of supply chain management that are most suitable for the application of artificial intelligence. The use of dynamic economic big data as inputs to predict decision outputs in order to assist with policy making in all economic and commercial sectors will become more attractive. Therefore, further investigation in these fields is necessary.

2.2. The Application of ML in Logistics and Supply Chain Management

2.2.1. Logistics Performance Index

Efficient transportation and logistics activities benefit not only the foreign but also the home economies. Superior logistical and operational systems can assist international trade because they connect the domestic and foreign economies. Evaluating a country's trade facilitation and logistics is crucial for its competitiveness, particularly in emerging markets. Since 2010, the World Bank has issued the LPI every two years, ranking 160 countries. The first publication was in 2007. LPI provides an overview of country-specific customs procedures, logistical costs, land, and marine transport infrastructure, and more. Countries base their strategic development strategies and targets on their LPI score.

The Logistics Performance Index, or LPI, is a powerful and all-encompassing indicator that has been used in numerous studies to examine the general logistics operations of groups of nations in the context of the robust growth of multimodal transport services. Shepherd et al. (2023) laid the foundation for performance-tracking policymakers and researchers, particularly in smaller and lower-income countries, by using and choosing the best parameters for machine learning models to account for LPI scores from 30 more countries and 13 more years. In contrast to earlier research, the authors examined a wider range of explanatory factors for LPI scores.

According to World Bank, The LPI (Logistics Performance Index) is a global benchmarking instrument designed to assess a country's efficiency in trade and transport facilitation. It specifically evaluates aspects of trade and logistics processes, enabling nations to pinpoint critical obstacles and uncover potential areas for enhancement. The LPI summarizes the performance of countries through six dimensions that capture the most important aspects of the logistics environment:

1. Customs; efficiency of the customs clearance process.

2. Infrastructure; quality of trade and transport-related infrastructure.

3. International Shipments; ease of arranging competitively priced shipments.

4. Logistics Quality; competence and quality of logistics services.

5. Tracking and Tracing; ability to track and trace consignments.

6. Timeliness; frequency with which shipments reach the consignee within the scheduled or expected time (Arvis et al.2014).

The LPI offers a thorough evaluation of global logistics performance, along with an analysis of performance trends, enabling an understanding of how logistics efficiency evolves over time. The performance is measured using a 5-point scale, with the overall LPI calculated as a weighted average across six key areas of logistics performance. Additionally, the LPI incorporates domestic performance indicators, which are not reflected in the country's overall score. It is further supplemented by quantitative data on specific elements of international supply chains in the respondents' countries, including import/export activities, lead times, supply chain costs, customs procedures, and the proportion of shipments subject to physical inspections (Arvis et al., 2012).

In term of international economics, there is ample evidence that the effectiveness of logistics networks is a crucial factor of bilateral trade, although the amount of the influence varies depending on economic and geographical characteristics (Çelebi, D., 2019). According to Ojala (2015) the efficiency of the transport system and the profitability of the industry are closely interconnected. Key factors such as reducing inventory through rapid turnover, the ability to adapt to fluctuating demand, minimizing lead times, and achieving the lowest transportation costs are vital for a company's competitive edge. Consequently, transportation systems are viewed as a crucial production element and a significant factor in decisions regarding facility locations. Allowing for comparisons across 160 countries, the LPI helps businesses identify challenges and opportunities related to the transport infrastructure, logistics expertise, and efficient supply chains in the receiving country. Arvis et al. (2007) also concluded that countries with the most predictable, efficient, and well-managed transport routes and trade procedures are also those most likely to benefit from technological advancements, economic liberalization, and greater access to international markets. As a result, the index ranking tends to place developed countries at the top, while emerging nations are positioned more variably across the spectrum. In this context, the LPI serves as a key indicator of the host country’s trade logistics performance and a benchmark when selecting sites for various operations. This is why countries often prioritize their ranking over improvements in the actual values of the LPI indicators. However, by improving the LPI index, a country can indirectly enhance its ranking on the logistics positioning map, which in turn helps boost its international trade and exchange.

2.2.2. Feature Selection

Feature selection involves picking a limited number of variables that are most critical for building the model. Proper selection of features enhances the model performance through minimization of information loss and the elimination of redundant, irrelevant, or highly correlated features contained in the data. It is often used to make the model more comprehensible and to improve generality by decreasing variance (R, Muthukrishnan & Rohini, R., 2016). Feature selection incorporates three general strategies: filter, wrapper and embedded. Filter methods divide features into groups on the basis of their statistical measures and subsequently rank them, effectively turning feature selection into a ranking task. These methods are not dependent on the machine learning technique associated with the chosen features. These include mutual information, correlation-based methods, Chi-square tests, and analyze of principal components. Due to the speed with which they process data, filter methods are generally preferred when dealing with high dimensional datasets (R, Muthukrishnan & Rohini, R., 2016).

On the one hand, embedded methods offer a balanced approach to feature selection by integrating it directly into the model training process, providing a middle ground between filter and wrapper methods. These methods simultaneously return the learned model and the selected features (Lu, M., 2019). The key characteristic of embedded techniques is that the feature selection and model learning components are inseparable (Lal, T. N. et al., 2006). Regularization is often used within embedded methods during training, such as in regularized models like linear discriminant analysis, support vector machines (SVM), and LASSO, which are common embedded approaches (Lu, M., 2019). For example, LASSO applies an L1-norm penalty to normalize the parameters of a linear model, effectively reducing less relevant coefficients to zero. Additionally, sparse learning approaches for multi-class classification, such as L2, 1-norm regularized regression, have been proposed (Lu, M., 2019).

Additionally, it is widespread that Principal component analysis (PCA) has been utilized in data mining to examine data structure. By maximizing the variance of the data, PCA generates new orthogonal variables, often known as latent variables or main components. In order to visualize the data in a low-dimensional PC space, the number of latent factors is significantly smaller than the number of original variables (Guo et al, 2002). Since PCA employs all of the original variables to create the new latent variables (principal components), it significantly decreases the dimensionality of the space but does not decrease the number of original variables. In data mining, selecting a small subset of variables that can effectively represent the structure of the complete dataset is a critical task. The goal is to retain as much information as possible while reducing dimensionality, which enhances model performance and interpretability. Krzanowski (1987) proposed a method based on Procrustes analysis for feature selection, which aims to identify a subset of variables that closely mirrors the structure of the full dataset. The method employs a stepwise procedure, specifically backward elimination, where variables are removed one at a time to improve the subset's representation of the data. However, while this approach offers a structured way to select variables, it does not guarantee finding the best possible global subset. Since the method uses a stepwise procedure, it is prone to local optima, meaning that the selection process may not yield the most optimal subset of features. This limitation becomes particularly evident when working with datasets containing hundreds or thousands of variables, a common scenario in data mining. Additionally, Procrustes analysis requires performing Principal Component Analysis (PCA) at each step of the elimination process, which can be computationally intensive. As the number of variables increases, the computational cost grows exponentially, making it less efficient for large datasets (Guo et al, 2002).

For feature selection in this study, embedding techniques such as Lasso, Ridge, and Elastic Net are used. These are advantageous because they enable feature selection and prediction to happen at the same time. The correlations between economic variables are also assessed using correlation-based testing techniques. These techniques, which together aid in determining the most pertinent features for the model, include Principal Component Analysis (PCA), Analysis of Variance (ANOVA), F-test, and Recursive Feature Elimination (RFE). By choosing the most informative features and removing unnecessary or redundant ones, these methods seek to increase the accuracy and efficiency of the model.

2.2.3. Machine Learning of Regression

When it comes to predicting LPI and/or similarity indices, some of the most common ML methods is supervised learning, which trains the system using a collection of known or unknown patterns. There are several approaches employed, including regression and classification. A foundational concept in ML is the ability of algorithms to “learn” from historical data. During the training phase, these algorithms analyze large datasets to identify underlying trends, correlations, and structures within the data. For instance, supervised learning algorithms such as linear regression, decision trees, and support vector machines are widely used to forecast future outcomes based on labeled historical data Burr, T. (2008). A few studies have concentrated on the focus of learning in ML systems in relation to predictive capabilities, stressing their iterative nature. Algorithms increase their performance accuracy and model building as more data is processed. This is crucial in predictive maintenance of equipment failures in machine learning, as more operational data sets are analyzed. As told by Jordan & Mitchell (2015), these patterns are crucial to consider especially in dynamic environments where the conditions change over time. Patterns and algorithms that utilize machine learning, K-Nearest neighbors, support vector classifier (linear and nonlinear kernels), decision trees, random forests, AdaBoost, and Artificial Neural Networks, are commonly adopted by Zhang et al. (2023) to anticipate in the continuous-casting process by measuring time series data of machines. The **figure 1** below illustrate the general predictive ML framework for logistics performance prediction based on supervised learning algorithm.

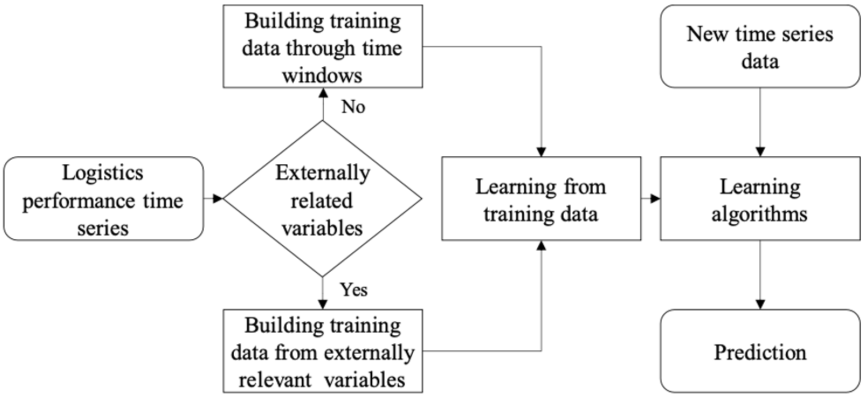
Building training data through time windows
No
Logistics performance time series
Externally related variables
Yes
Building training data from externally relevant variables
Learning from training data
New time series data
Learning algorithms
Prediction

Figure 1. General predictive ML framework for logistics performance prediction

Source: D'Aleo and Sergi (2017)

Regarding the predicting of the LPI, it is necessary to include some external factors while estimating the logistics performance predictor variables. If there are no such external variables, the data for the predictor is usually built with training sets based on historical periods. On the one hand, training data can also be enhanced by adding relevant external factors. Hwang et al. (2023) identified the most important macro-level logistics performance determinants of industrial policy priorities, strategic infrastructure development, public-private logistics market growth, and communications network configurations. These issues are vital for considering the macro logistics performance of China, Japan, and Korea. In the same vein, Zeng et al. (2021) state that using external precise multivariate metrics will improve the quality of logistics performance predictions. The text demonstrates the relevance of including variables that have strong correlation with the logistics system performance in formulating multivariate models. D’Aleo and Sergi (2017) further examine the impact of systematic factors that might influence performance and efficiency of logistics, such as transport policy and other external factors. This study highlights the importance of this. On the one hand, these techniques may not be as successful without the proper choice of external variables. In particular, researchers need to vary these external factors for every step of the model training and prediction process to quantify their contribution. In general, the prediction accuracy is high when the corresponding variable is easy to understand (Zeng et al., 2021). As reported in the literature, one of the dominant ANN approaches is widely used (Yuhong Li, & Weihua Ma, 2010). ANN models are constructed as a nonlinear combination of a set of elements (Tealab et al., 2017). More complex algorithms are necessary to resolve particularly nonlinear relationships due to the complex macroeconomic and microeconomic factors.

In this study, A range of regression techniques are used to understand the relationship between goal variables and economic factors. First, Multiple Linear Regression (MLR) is used to understand the linear relationship between predictors and outcomes. MLR provides valuable insight into how significant and impactful each feature is. Also, to increase predicted accuracy to account for complexities, non-linear correlations are captured through Support Vector Machines (SVM), Random Forest Regression, and XGBoost. To avoid overfitting and to impose a penalty on excessive model complexity, more complex regression techniques such as Lasso, Ridge, and Elastic-Net are used. These regression models are suitable for the analysis of economic data in this study, since these models can accommodate structures of data that are both linear and non-linear.

3. Data and Methodology

3.1. Data sources and data preparation

3.1.1. Data sources

The authors use the secondary data from OurWorldInData and WorldBank statistics over 14 years, from 2010 to 2023 (7 periods), including the data of 91 countries with available LPI information.

Typically, the variables the authors use in this research include economic component and population component, which show a substantial correlation between the logistics performance of a nation of LPI score and a factor, for example GDP per Capita (GDP_C), Population Growth (Pop_Gr), Export (Exp) and Import (Imp) that LPI has a significant positive effect on increasing international trade for both import and export.

Table 1. Summary of variables

Summary of variables	Definition	Sources
LPI	Logistics Performance Index	WorldBank
CPI	Consumer Price Index	WorldBank
Exp	Export	WorldBank
Imp	Imports	WorldBank

Summary of variables	Definition	Sources
GDP_Gr	GDP Growth	WorldBank
GDP_C	GDP per capita	WorldBank
GD/GDP	General government gross debt	WorldBank
GLB/GDP	General government net lending or borrowing	WorldBank
GS/GDP	Gross savings GDP	WorldBank
GE/GDP	Government expenditure of GDP	WorldBank
IRA	Inflation rate average consumer prices	WorldBank
IRE	Inflation rate end of period consumer prices	WorldBank
LB	Labor force	WorldBank
Net_In	Net inflows	WorldBank
Net_Out	Net outflows	WorldBank
N_GDP	Nominal GDP	WorldBank
GEE/GDP	Government expenditure on education, total (% of GDP)	OurWorldInData
GEE/GE	Government expenditure on education, total (% of government expenditure)	OurWorldInData
Pop_Gr	Population growth	WorldBank
Pop	Population	WorldBank
SE	Secondary education enrollment	OurWorldInData
CAB/GDP	Current account balance GDP	WorldBank
CAB	Current account balance	WorldBank

Source: Summarize by the authors, 2025

3.1.2. Data Pre-Processing

In this research, data were aggregated from seven cycles conducted between 2010 and 2023, culminating in a dataset encompassing 23 features across 91 countries, derived from the mapping of the Logistics Performance Index (LPI) data. Ensuring the integrity of the compiled dataset was a critical priority, as any data loss during the mapping process could potentially undermine the validity and precision of subsequent analytical procedures.

Recognizing the detrimental effects of missing data on model performance and the inherent risks of introducing bias especially when utilizing imputation techniques such as mean or median substitution, rigorous measures were implemented to minimize data incompleteness. Specifically, all missing values related to the LPI variables were excluded to maintain the robustness and reliability of these primary indicators. For the remaining variables, where the incidence of missing data was minimal, imputation was conducted using the mean values of the respective variables, thereby optimizing data retention without compromising the dataset's overall integrity.

The finalized dataset consists of over 630 observations, corresponding to the seven cycles across the 91 countries. For the development and evaluation of the machine learning model, the dataset was partitioned, with 80% designated for training purposes and the remaining 20% reserved for testing. This stratification facilitates a robust validation process, ensuring the model's performance is accurately assessed.

3.2. Empirical framework

The economic component plays a critical role in the logistics performance of a nation, as evidenced by a substantial correlation between a country's Logistics Performance Index (LPI) score and key economic factors, such as GDP per capita (Word Bank, 2024), export and import volumes (OurWorldInData, 2024). According to Takele (2019), the components of the LPI significantly influence the growth of international trade, affecting both imports and exports. These factors are integral to understanding the relationship between logistics performance and economic development, providing valuable insights for generating policies aimed at improving logistics performance based on predictive correlation parameters.

This study focuses on supply chain efficiency, particularly in terms of the financial efficiency of supply networks, with a specific emphasis on the 91 countries. The performance of logistics is closely tied to how effectively supply chains connect firms to both domestic and international opportunities (World Bank, 2018). The analysis presented in this research utilizes training data derived from both macroeconomic and microeconomic variables. **Figure 2** illustrates the construction of this training data, incorporating relevant economic features that influence logistics performance.

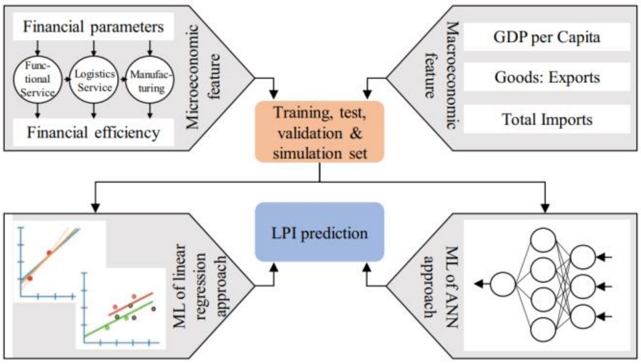

Figure 2. A methodological framework of LPI prediction procedure

Source: Suriyan Jomthanachai et al., 2023

Seven ML models and one MLP-ANN were prepared. The methodologies employed encompass a spectrum ranging from traditional statistical techniques to cutting-edge machine learning algorithms. Specifically, the models implemented include Multiple Linear Regression (MLR), XGBoost Regression (XGB), Random Forest Regression (RF), Support Vector Machines (SVM) for regression, KNearest Neighbors Regression (KNN), Decision Tree

Regression, an embedded approach based on penalized linear regression, and Multilayer Perceptron Artificial Neural Network (MLP-ANN) Regression. This comprehensive modeling framework facilitated a rigorous evaluation of both predictive performance and model generalizability, thereby elucidating the relationships between process parameters and the LPI.

3.3. Data analysis and parameter setting

3.3.1. Feature selection

3.3.1.1. Correlation method

In this study, the Pearson correlation coefficient is employed to quantify the linear association between the variables x and y (Pearson, 1895). This framework facilitates a systematic evaluation of the linear relationships among the studied variables, thereby supporting the robustness of the model's feature selection process.

3.3.1.2. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a well-established multivariate technique used to reduce data dimensionality and is commonly applied to consolidate multiple well-being indicators into a single composite index. In the present study, PCA was implemented in Python to extract principal components from both input and output variables, following the centering and scaling of the datasets to standardize the data.

A loading threshold of 0.3 (in absolute value) was established, ensuring that only variables with loadings meeting or exceeding this criterion were considered for subsequent analysis. To further enhance the interpretability of the extracted factors, a varimax rotation was applied, which minimizes cross-loadings by diminishing the influence of variables with loadings below the designated significance threshold (Lawrence S et al., 2013).

For both feature selection and visualization, PCA biplots were generated (Abimbola O-PP et al., 2020). In these biplots, the first principal component (PC0) delineates the primary dimension, while the second principal component (PC1) represents the secondary dimension. This biplot-based approach offers a graphical depiction of the interrelationships among the variables, with vectors representing the contributions of the original parameters (Zhang H, Srinivasan R., 2021). It is noteworthy that the centering and scaling of the attributes during preprocessing were critical in ensuring the robustness and reliability of the PCA results.

3.3.1.3. ANOVA F-test

In this study, we employed a filte-based feature selection approach, specifically, the ANOVA-F test, to identify the most pertinent features from both datasets. Filter-based methods utilize various statistical measures, such as similarity, dependence, information, and distance metrics, to elucidate significant dependencies or correlations between the input features and the target variable. Analysis of Variance (ANOVA) comprises a family of parametric statistical models and estimation procedures designed to evaluate whether the means of two or more samples originate from the same distribution. The F-test, also known as the F-statistic, involves calculating the ratio of variances to determine statistical significance.

In this context, the ANOVA-F test functions as a univariate statistical method whereby each feature is individually compared to the target variable to assess the existence of statistically significant relationships. This technique is particularly advantageous in classification scenarios where the input features are numerical and the target variable is numerical. The ANOVA-F test was implemented in Python using the f_classif() function provided by the scikit-learn library. This function serves as the scoring mechanism within the SelectKBest class, which ranks features based on their computed scores and selects those with the highest values. In our analysis, the f_classif() function was employed as the scoring function - representing the ANOVA-F test to discern and retain the most critical features from the datasets (Han Zhuang et al., 2021).

3.3.1.4. Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a backward feature selection technique that begins by constructing a model using the complete set of available features and computing an importance score for each. Subsequently, features with the lowest importance scores are iteratively removed, with the model being retrained at each step to recalculate the scores. This recursive process continues until a specified number of features remains. Notably, users can define both the number of features to evaluate and the size of each subset, making the subset size a critical tuning parameter. The subset that optimizes the performance criteria is ultimately selected for training the final model (Baffa et al., 2022).

The reduced dataset, comprising the selected features and therefore exhibiting lower dimensionality compared to the original dataset, was subsequently partitioned into training and testing sets. A 10-fold cross-validation procedure was then applied, wherein the dataset was divided into 10 distinct folds. In each iteration, one-fold was designated as the testing set while the remaining folds served as the training set. The training data in each iteration was fed into an ensemble classifier to facilitate model training.

3.3.2. Methodology of the ML and ANN Models

3.3.2.1. Multi Linear Regression (MLR) Algorithm

Multiple Linear Regression (MLR) is one of the most widely employed linear regression models. As a multivariate statistical technique, MLR is used to elucidate the relationship between a set of independent variables (X_1, X_2, ..., X_n) and a dependent variable (Y) with an explanation and prediction as objectives: explanation and prediction. From an explanatory standpoint, the focus is on the regression coefficients evaluating their magnitude, sign, and statistical significance to understand the influence of each predictor. In terms of prediction, the model assesses the extent to which the independent variables can accurately estimate the dependent variable (Hair, J.F., 2010).

3.3.2.2. XGB Regression Algorithm

XGBoost (eXtreme Gradient Boosting) is an advanced implementation of gradient boosting algorithms that emphasizes computational speed and enhanced performance. It is widely recognized as an efficient and scalable end-to-end tree boosting framework. The objective function in XGBoost is composed of two key elements: a loss function and a regularization term. The loss function measures the discrepancy between the model’s predictions and the observed data typically represented as L(θ) for a set of n predictions, where θ denotes the model parameters. The inclusion of the regularization term serves to penalize excessive model complexity, thereby reducing the risk of overfitting and enhancing the model's generalization capabilities.

XGBoost extends the gradient boosting framework by iteratively adding new predictors (trees), each designed to rectify the residual errors of the preceding ensemble. At each iteration, the model employs a gradient descent approach to minimize the overall objective function, updating the predictions by following the negative gradient of the loss function as evaluated on the current predictions (Sun, Z. et al., 2024).

XGBoost is designed to autonomously determine the optimal strategy for handling missing values during training. When a split point contains missing data, the model learns whether assigning these observations to the left or right branch maximizes the gain, thereby enhancing its ability to effectively manage incomplete datasets. Unlike conventional gradient boosting methods that cease splitting a node once no further improvement is detected, XGBoost initially expands the tree to its full depth. It then prunes branches that offer minimal contribution to the overall prediction performance, using the gain from the objective function as the pruning criterion (Wan, A. et al., 2024).

3.3.2.3. Random Forest Regression Algorithm

The random forest (RF) model is an ensemble learning method that synthesizes predictions from multiple decision trees, thereby yielding results that are typically more accurate and stable than those derived from any single decision tree. During the training process, RF constructs a collection of decision trees, each generated from a bootstrap sample of the training data. Additionally, at each node split, only a randomly selected subset of features is evaluated. This inherent randomness contributes to the model's robustness and significantly reduces its susceptibility to overfitting.

This method provides several advantages, including enhanced predictive accuracy through the aggregation of outputs from multiple trees and increased robustness against overfitting, particularly as the number of trees grows. Additionally, these models demonstrate resilience to missing data, maintaining performance even when a significant portion of the data is absent. Additionally, the primary limitations include increased computational complexity and cost associated with large ensembles, as well as diminished interpretability compared to a single decision tree (Sun, Z. et al., 2024).

3.3.2.4. Support Vector Machines Algorithm

Support Vector Machines (SVM), initially proposed by Vapnik, have been extensively applied to address non-linear regression challenges (Vapnik, V., 1995). A considerable body of literature exists that elucidates the theoretical foundations of SVM (Desai, S.S. et al., 2019). In this study, an ε-SVM regression model was employed, which necessitates the use of a training dataset for effective model development.

For the ε-SVM regression model employing a diametral basis function (RBF) kernel, the generalization capability is optimized by tuning three hyperparameters: the regularization parameter C, the ε-insensitive loss parameter, and the kernel parameter γ. In this study, optimal values for these parameters were determined via a trial-and-error approach. Specifically, 80% of the dataset was randomly allocated for training and the remaining 20% for validation, with performance further evaluated through ten-fold cross validation. All support vector machine models were implemented using Python.

3.3.2.5. KNeighbors Regression Algorithm

KNearest Neighbor (KNN) algorithm classifies an object by calculating the Euclidean distance between data points and assigning the object to the class most common among its KNearest Neighbors (J. Huang, 2018). The value of K is positive integer and usually small. The classifier's performance is highly dependent on the choice of value K. Usually the value of K is chosen as an odd number for binary classification to avoid ties in the voting process. If the value of K is chosen 1 then the object is simply assigned the class of its single nearest neighbor. Value of K chosen should be optimal, if the value of K is small, then it could be underfitting, while a very large value can cause overfitting of the model.

3.3.2.6. Decision Tree Regression Algorithm

The Decision Tree (DT) algorithm is widely employed for classification tasks in machine learning due to its versatility in handling both categorical and continuous data (Charbuty & Abdulazeez, 2021). DTs represent data through a hierarchical tree structure that extends from a single root node to multiple leaf nodes. The tree construction process initiates at the root, where the selection of the splitting feature is guided by impurity metrics such as the Gini index or entropy (Kingsford et al., 2008).

3.3.2.7. Embedded technique (ML of Penalized Linear Regression Technique)

In the presence of noisy data, conventional linear regression methods such as ordinary least squares (OLS) regression, are prone to overfitting, resulting in models that perform well on training data yet fail to generalize to new or unseen samples. In contrast, regularization techniques employed in ridge regression, LASSO regression, and Elastic-net regression mitigate overfitting by constraining model complexity, thereby enhancing the generalizability of predictions on unseen data (Cui & Gong, 2018).

a) Ridge regression:

Ridge regression approach effectively shrinks the magnitude of regression coefficients, thereby enhancing the model's generalizability when predicting new, unseen data. A regularization parameter, often referred to as the penalty factor, is employed to balance the trade-off between minimizing the training data's prediction error and imposing L2-norm regularization, thereby controlling the bias-variance trade-off (Zou & Hastie, 2005).

The primary advantages of this method are its capacity to handle strongly correlated environmental variables and its efficacy in scenarios with relatively modest data volumes. However, a notable disadvantage is that the resulting parameter estimates may be biased (Ahmadi-Nedushan et al., 2006).

b) LASSO regression:

The L1-norm regularization is applied to the OLS loss function in LASSO regression, with the goal of minimizing the sum of the absolute values of the regression coefficients (Tibshirani, 1996).

L1-norm regularization typically forces most coefficients to zero, retaining only one feature among groups of correlated predictors (Zou & Hastie, 2005). Consequently, LASSO regression generates a highly sparse predictive model that facilitates variable selection and reduces model complexity. However, this sparsity may be problematic when the number of features is high relative to the number of samples (Efron et al., 2004).

The primary advantage of LASSO lies in its ability to yield interpretable models by selecting a subset of predictors that most strongly influence the response variable, a feature particularly beneficial when data are scarce. Conversely, a key limitation is that, among sets of highly collinear variables, the model tends to arbitrarily select a single covariate while excluding the others (Boucher et al., 2015).

c) Elastic-net regression:

Elastic-net regression seeks to overcome the limitations of the LASSO technique (Zou & Hastie, 2005). Elastic-net regression represents a hybrid of LASSO and ridge regression techniques, enabling the selection of a number of features that may exceed the sample size while still promoting model sparsity (Zou & Hastie, 2005). A mixing parameter α is utilized to balance the contributions of the L1-norm (associated with LASSO) and L2-norm (associated with ridge regression). The values for α of Elastic-net lie between Ridge (α = 0) and LASSO (α = 1).

One notable advantage of the Elastic-net approach is its robust performance in high-dimensional settings, particularly when the number of predictors surpasses the number of observations. Moreover, it often yields a model that is more stable and interpretable compared to LASSO alone. Conversely, a potential drawback is that when data availability is limited, the model may incorporate an excessive number of variables, thereby overwhelming the dataset (Boucher et al., 2015). In summary, Elastic-net is a regularized regression method that integrates both L1 and L2 penalties, making it particularly effective in scenarios involving multiple correlated features.

3.3.2.8. MLP-ANN Regression Algorithm

The multilayer perceptron (MLP) is a fundamental architecture within artificial neural networks and serves as a cornerstone of deep learning methodologies. It is characterized by the presence of at least three distinct layers: an input layer, one or more hidden layers, and an output layer. Training of the MLP is typically conducted using the supervised learning algorithm known as backpropagation (Gürkan Işık et al., 2023).

In this architecture, the input layer is responsible for receiving the raw data, which is subsequently propagated through the network via one or more hidden layers. Within these hidden layers, individual neurons perform computations by processing the incoming signals and transmitting the resulting outputs to subsequent layers. The overall complexity and representational capacity of the MLP are contingent upon the number of hidden layers and the number of neurons contained within each layer (Chai Meijuan, 2021).

Ultimately, the processed information reaches the output layer, where the network generates its final prediction or classification. At the level of each neuron, a weighted sum of its inputs is computed, to which a bias is added; this aggregate is then transformed through an activation function to yield the neuron's output.

Activation functions endow the network with non-linear properties, which are essential for learning and representing complex data patterns. Among the widely used activation functions are Sigmoid, Tanh, and ReLU (Rectified Linear Unit), with the Softmax function typically employed in the output layer for classification tasks (Shomope, I., et al, 2025).

The network is trained using the backpropagation algorithm, which entails computing the gradient of the loss function with respect to each weight via the chain rule. This error gradient is then propagated backward from the output layer to the input layer.

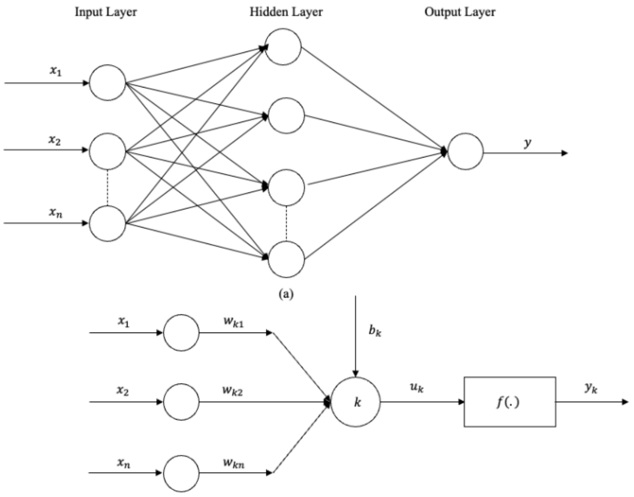

Figure 3. Architecture of neural network

Source: Osisanwo et al., 2017

Where x_1, x_2, …, x_n are the inputs, w_{k1}, w_{k2}, …, w_{kn} are the neuron weights, uk is the computation outcome of weighted inputs, bk is the bias term, f(.) is the activation function, and y_k is the output. There are several algorithms with which a network can be trained (Osisanwo et al., 2017).

The loss function quantitatively evaluates the discrepancy between the network's predicted outputs and the actual target values. For regression tasks, mean squared error (MSE) is typically employed, whereas cross-entropy loss is generally preferred for classification problems. To minimize this loss, optimization algorithms are utilized. The foundational approach is gradient descent, with its more advanced variants, such as stochastic gradient descent (SGD), Adam, and RMSprop frequently implemented to enhance convergence and overall training performance (Ghadery-Fahliyany, H. et al., 2024).

Artificial neural networks (ANNs) offer the distinct advantage of adeptly discerning complex patterns and generating highly accurate predictions. However, a significant limitation is that the network architectures used for function approximation typically necessitate a large volume of training data (Syam & Sharma, 2018).

3.3.3. Hyperparameter tuning

In the context of hyperparameter tuning for neural networks, the "black box" problem pertains to the inherent difficulty in discerning the impact of individual hyperparameter modifications on overall model performance and outcomes. This challenge is particularly salient in deep learning, where a multitude of hyperparameters - such as the number of neurons per layer, learning rate, regularization techniques, batch size, activation functions, and even the number of estimators in ensemble methods - can profoundly influence model effectiveness, generalizability, and the computational resources required for training (Ogunsanya, M. et al., 2023).

To systematically navigate the hyperparameter space and identify optimal configurations, several automated techniques have been developed, including grid search, random search, Bayesian optimization, and evolutionary algorithms. Complementary visualization tools (e.g., Matplotlib, Seaborn, Plotly, Weights & Biases, TensorBoard, Mlflow, Scikit-learn) facilitate a deeper understanding by graphically depicting changes in performance during the learning process as hyperparameters are adjusted (Malakouti, S.M. et al., 2023).

In the present study, hyperparameter tuning for both machine learning (ML) and artificial neural network (ANN) models was conducted using the grid search method. This technique involves an exhaustive exploration of a predefined hyperparameter space, wherein each configuration is systematically evaluated - often via cross-validation—to ascertain its performance based on specific criteria, such as accuracy. Although grid search is appreciated for its simplicity and ease of implementation, its primary drawback lies in its computational expense, particularly when applied to extensive hyperparameter spaces (Qu, Z. et al., 2021).

3.3.4. Model Evaluation Metrics

All models were implemented in Python, utilizing robust machine learning libraries such as Scikit-learn and XGBoost. To rigorously evaluate the performance of the optimized models, six key performance metrics were employed: the coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), root mean squared error (RMSE), and accuracy (Wu, Y. et al., 2022).

4. Result and discussion

4.1. The result of correlation method

Figure 4 depicts the result of the correlation study performed using Python tool. When a regression type prediction is used, the input of a correlation model spanning both the dependent and predictor variables is used.

We begin by constructing a feature set of predictor factors that have a direct good or outstanding correlation to the dependent variable of LPI (r >= 0.4, as shown in the red border in **Figure 4**) (namely set A – Cor_direct). A set A's predictor variables are GDP_C, r = 0.77, GE/GDP, r = 0.62), CAB/GDP with r = 0.41, Exp, r = 0.6, and Imp, r = 0.54), for a total of five features.

Furthermore, the predictor variables that have a strong or outstanding correlation with a member of set A (r > 0.45, as shown in the yellow border in **Figure 4**) is taken into account and subsequently extended to a member of set B. The additional predictor variables of a set A into set B (Cor_related) include GLB/GDP (with CAB/GDP), r= 0.52, GS/GDP (with CAB/GDP) and r = 0.68, N_GDP (with Exp and Imp) and r = 0.84, r = 0.9. The total number of features in set B is 5 + 3 = 8.

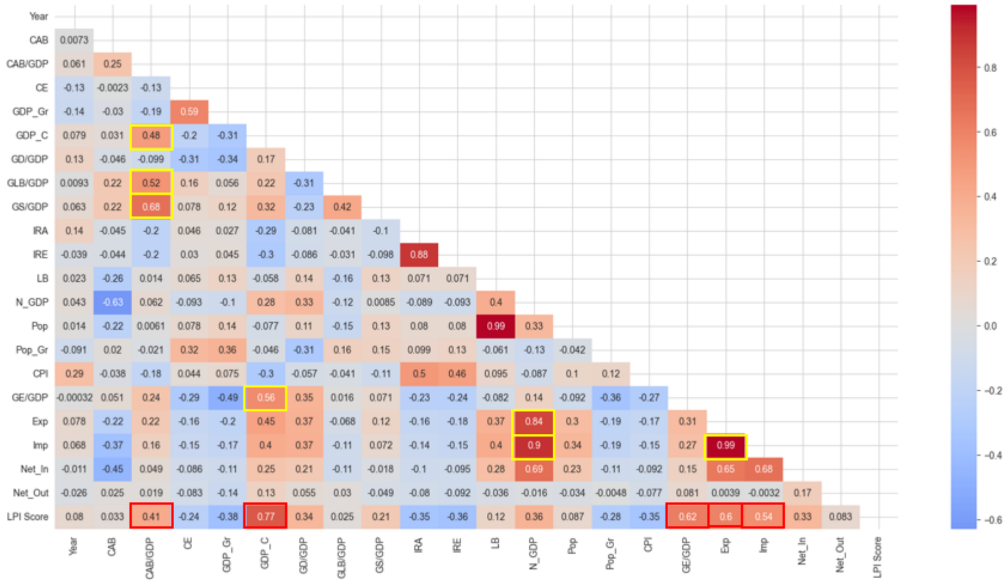

Figure 4. Correlation Matrix

Source: The authors, 2025

4.2. The result of PCA method

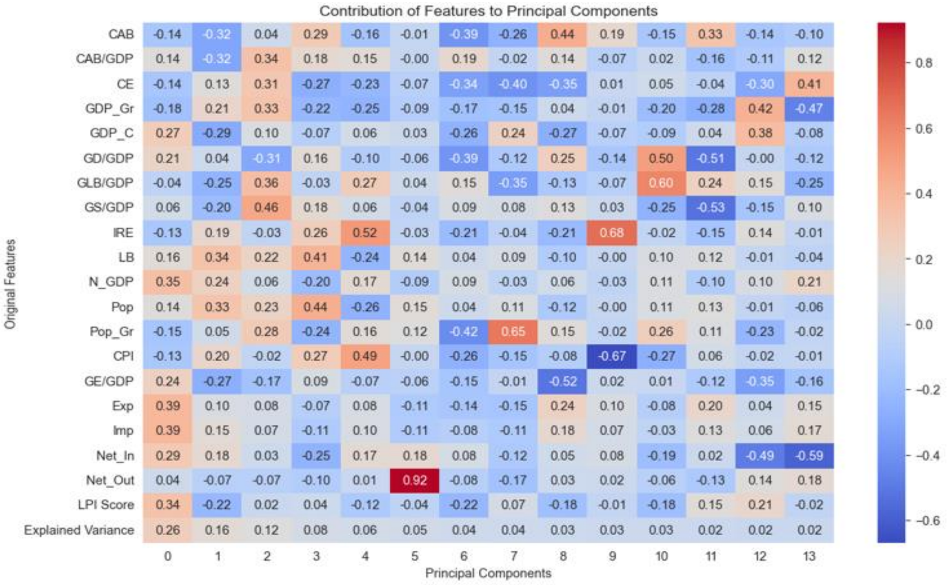

Figure 5. Contribution of Features to Principal Component

Source: The authors, 2025

The results of contributions of Features to Principal components are displayed in **Figure 5**. The variables, with more than 95% confidence interval, are predicted to affect this study's prediction model. The principal components are divided into 13 components, each component represents its impact on the prediction mode, written in the Explained Variance row and in the order from largest to smallest, from left to right. In the Figure, PC0 explains 26% of the variance, which is the highest among principal components, while PC13 only explains 2% of the variance.

The PCA result is generated by Python in which the dependent variable and predictor variables are used as PCA model input to select the feature for the regression purpose. **Figure 5** depicts the proportion of variance of each principal component based on the overall result (only PC0 to PC9 out of a total of 14 PCs).

From the results of contributions of Features to Principal, the authors proceed to select features based on the level of variance, which are from 1-3, 1-5, 1-10. The variance of each feature must be above 30% to be taken into consideration. PC0 through PC9 may encompass roughly 90 percent of the variation (90.22 percent). Furthermore, when PC0 to PC2 were evaluated, the variation was 46.2 percent, which is more than half of the range of PC0 to PC9. When PC0 to PC4 is considered half of the 10 PCs from PC0 to PC9, the variance is 53.22 percent.

Table 2. Contribution of Features to Principal Component

Feature	Contribution (PC0-PC9)	Contribution (PC0-PC4)	Contribution (PC0-PC2)
Imp	0.389	0.389	0.389
Exp	0.388	0.388	0.388
N_GDP	0.345	0.345	0.345
LPI Score	0.338	0.338	0.338
LB	0.336	0.336	0.336
Pop	0.327	0.327	0.327
CAB/GDP	0.32	0.32	0.317
CAB	0.317	0.317	0.317
GS/GDP	0.426	0.426	**0.463**
GLB/GDP	0.357	0.357	0.357
GDP_Gr	0.326	0.326	0.326
CE	0.312	0.312	0.312
GD/GDP	0.306	0.306	0.306
IRE	0.489	**0.522**	-

Feature	Contribution (PC0-PC9)	Contribution (PC0-PC4)	Contribution (PC0-PC2)
CPI	0.49	0.489	-
Net_Out	**0.923**	-	-
Pop_Gr	0.421	-	-

Source: The authors, 2025

To construct a collection of selected features, we examined the attribute that provides a high loading on a factor (equal to or greater than 0.3). The detected attributes in PC0 to PC2 (46.2 percent variance), PC0 to PC4 (53.22 percent variation), and PC0 to PC9 (92.27 percent variation) are allocated to feature sets C, D, and E, respectively.

Set C such as Imp, Exp, N_GDP, LB, Pop, CAB/GDP, CAB, GS/GDP, GLB/GDP, GDP_Gr, CE, GD/GDP, 12 features in total. Set D of 14 features is set C plus 2 features which are IRE and CPI. And set E has a total of 17 from the overall 23 features that include 3 more variables Net_Out, Pop_Gr, GE/GDP.

Furthermore, the feature selection while constructing a PCA-biplot is illustrated in Figure 6, with the selected features are those who have strong interrelations with LPI score. The selection is motivated by the interrelationships of each feature to LPI. The direction of the feature vector reflects the positive or negative correlations. When a feature has a comparable direction that is the smallest in the angle of the vector relative to the LPI vector, it indicates the strongest positive correlations, while the opposite direction indicates negative correlations.

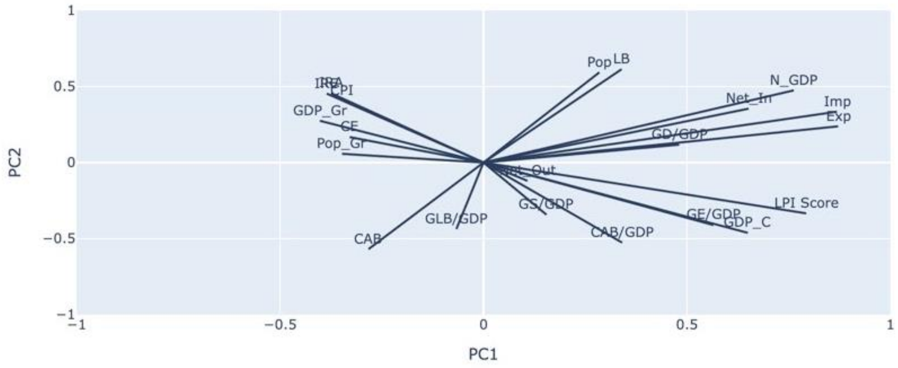

Figure 6. PCA Biplot

Source: The authors, 2025

Vectors close to perpendicular to the LPI vector, on the other hand, are weakly correlated (GLB/GDP and GS/GDP in **Figure 6**) Based on the PCA-biplot, the selected features of set F out of 23 includes GD/GDP, Net_In, N_GDP, Imp, Exp, GE/GDP, GDP_C, which are 7 features in total. This is due to the relatively high Squared R when compared to other groups of features, and the authors recognized set F is the most suitable set to run regression and machine learning to have a good outcome. The summary of the subset of features selected using the PCA method is shown in **Table 3.**

Table 3. Summary of the subset of features selected using the PCA method

Economic Feature	Correlation		Principle Component Analysis				Penalized Linear Regression			
	Cor_direct (set A)	Correlated (set B)	PC1-3 (set C)	PC1-5 (set D)	PC1-10 (set E)	Biplot (set F)	LASSO (set G)	E-net_0.9 (set H)	E-net_0.5 (set I)	E-net_0.1 (set J)
CAB			o	o	o					o
CAB/GDP	o	o	o	o	o					o
CE			o	o	o					
GDP_Gr			o	o	o					o
GDP_C	o	o				o	o	o	o	o
GD/GDP			o	o	o	o			o	
GLB/GDP		o	o	o	o					o
GS/GDP		o	o	o	o					
IRA									o	o
IRE				o	o				o	o
LB			o	o	o					o
N_GDP		o	o	o	o	o				o
Pop			o	o	o					o
Pop_Gr					o			o	o	o
CPI				o	o				o	o
GE/GDP	o	o			o	o	o	o	o	o
Exp	o	o	o	o	o	o	o	o	o	o

Economic Feature	**Correlation**			**Principle Component Analysis**				**Penalized Linear Regression**		
	Cor_direct	Correlated	PC1-3	PC1-5	PC1-10	Biplot	LASSO	E-net_0.9	E-net_0.5	E-net_0.1
	(set A)	(set B)	(set C)	(set D)	(set E)	(set F)	(set G)	(set H)	(set I)	(set J)
Imp	o	o	o	o	o	o				o
Net_In						o				o
Net_out					o					
LPI Score										

Source: The authors, 2025

4.3. The result of penalized linear regression method

Table 4 also displays the results of the **Embedded technique - LASSO and Elastic Net regression**. Using Python, the model for LASSO regression has been reducing the predictor parameters from 23 to 3, which offer various interception values and parameter significance. The 3 features selected by LASSO (set G) include GDP_C, GE/GDP, and Exp.

Table 4. Feature selection techniques results

Economic Feature	**Univariate Selection**	**Recursive Feature Elimination**	**Tree-based Methods**
	ANOVA F-test	RFE	Random Forest
CAB	o	o	o
CAB/GDP	o	o	o
CE	o	o	o
GDP_Gr	o	o	o
GDP_C	o	o	o
GD/GDP	o	o	o
GLB/GDP	o	o	o
GS/GDP	o	o	
IRA	o	o	
IRE	o	o	
LB	o	o	
N_GDP	o	o	
Pop	o	o	
Pop_Gr	o	o	
CPI	o		
GE/GDP			
Exp			
Imp			
Net_In			
Net_out			
LPI Score			

Source: The authors, 2025

For Elastic-net related, in this study, we vary the a as 0.1, 0.5, and 0.9. The results of feature selection from Python which provides the preferred parameters that the model does not shrink are displayed in **Table 4**.

It was found that when a = 0.9 the set of selected features is nearly similar to the results of LASSO, but added Pop_Gr. When a is assigned with the value of 0.5, they provide the likely set of 8 selected features (set H). Finally, for a = 0.1, we found that 17 parameters were nonshrink (set I). Set H contains all attributes of set G which GD/GDP, CPI, IRE and IRA are added. And set I comprised all elements of set H with CAB, CAB/GDP, GDP_Gr, GLB/GDP, LB, N_GDP, Pop, Imp, Net_In combined. The summary of the subset of features selected using penalized linear regression method is shown in **Table 4**.

We also use other features selection methods to validate the results of LASSO and E-net, including ANOVA F-test, Recursive Feature Elimination, and Tree-based Methods.

The ANOVA F-test for univariate selection is utilized. After running ANOVA F-test and got the scores of those features, the authors recognized that 20 is a suitable score to select out features (feature must has score more than 20 to be selected). There are 15 out of 23 features selected, and this group of features is nearly alike set I. The Recursive Feature Elimination is also run, with a total of 14 features selected.

Finally, Tree-based Methods (Random Forest) is used, with a minimum score of 0.01 for feature to be selected. There are 7 features selected, which are CAB, CAB/GDP, CE, GDP_Gr, GDP_C, GD/GDP. GLB/GDP. With the rfr scores of features, the authors consider the score 0.01 is good enough to be used for the selection.

In summary, after running multiple features selection methods, we conclude that each method has its own strength and weakness, with various groups of selected features. These groups will be used for the later validation of selected features.

4.4. Regression and validation result

Table 5. Average Results for Each model across all datasets

Model	MAE	RMSE	MAPE	NSE	R^2
LR	0.279	0.346	9.056	0.631	0.631
XGB	0.214	0.264	7.056	0.79	0.79
RFR	0.189	0.235	6.232	0.832	0.832
SVR	0.218	0.279	7.249	0.76	0.76
KNN	0.233	0.266	6.838	0.785	0.785
DTR	0.24	0.31	7.953	0.788	0.788
MLP-ANN	0.217	0.274	7.195	0.769	0.769
Ridge	0.278	0.346	9.057	0.63	0.63
LASSO	0.325	0.39	10.519	0.538	0.538

Model	MAE	RMSE	MAPE	NSE	R²
E-net 0.5	0.34	0.367	9.866	0.539	0.539
E-net 0.1	0.288	0.353	9.351	0.616	0.616

Source: The authors, 2025

For all datasets, 80% of datasets are trained utilizing identified ML methods such as LR, XGB, RFR, SVR, KNN, DTR, MLP-ANN. Furthermore, the LASSO and E-net models constantly train their datasets using only the selected feature set that they have been trained on. Furthermore, the entire collection of all features is compared. The test sets are then utilized to validate the model. The validation findings are represented by a performance evaluator or criterion such as MAE, MAPE, RMSE, NSE, and R2.

Table 6. Average Results for Each model across datasets applied PCA method

Model	MAE	RMSE	MAPE	NSE	R²
Full_dataset	0.225	0.277	7.429	0.768	0.768
Corr_a (set A)	0.234	0.287	7.735	0.752	0.752
Corr_b (set B)	0.387	0.488	12.29	0.28	0.28
PC1_3 (set C)	0.276	0.358	9.683	0.609	0.609
PC1_5 (set D)	0.263	0.339	8.619	0.656	0.656
PC1_10 (set E)	0.251	0.317	8.272	0.693	0.693
Biplot (set F)	0.252	0.286	7.46	0.768	0.768
Lasso (set G)	0.25	0.287	7.643	0.764	0.764
Enet_10 (set J)	0.225	0.275	7.452	0.77	0.77
Enet_50 (set I)	0.23	0.284	7.589	0.757	0.757
Enet_90 (set H)	0.231	0.279	7.635	0.764	0.764
ANOVA - Ftest	0.245	0.301	7.833	0.727	0.727
RFE	0.236	0.291	7.728	0.741	0.741
RFR	0.251	0.316	8.256	0.695	0.695

Source: The authors, 2025

The average score of all ML methods of the subset of selected features (set A to set J, and F-test, rfe and rfr) is then displayed. Compared to other sets, Enet_10 has the greatest performance of all criterion. It has the greatest MAE performance (*0.224) (minimum value), while this model with a comparable set has the best MAPE performance as the lowest value (*4.41), RMSE as the lowest of 0.275. NSE values range between 0.28 to 0.768 indicates the prediction performance, whereas NSE values close to 1 indicate best prediction performance that the Enet_10 achieves the most excellent performance (*0.775). Finally, we proceed to predict result base on the dataset selected by PCA Biplot to see if the result is as significant as that of Enet_10. The dataset of 7 features selected by

PCA Biplot is also divided into 2, with 80% is for trained and 20% is for predict. The authors utilized regressions such as Decision Tree Regression, XGBoost, K-Nearest Neighbors, Random Forest Regression, and Support Vector Machine. The table below show the results for PCA Biplot dataset with different methods of regression.

Table 7. The results of different methods of regression**.**

Model	Train Score	Validation Score	Test Score
Decision Tree Regression	0.908	0.736	0.821
XGBoost	0.849	0.768	0.85
K-Nearest Neighbors	1	0.819	0.887
Random Forest Regression	0.959	0.822	0.894
Support Vector Machine	0.844	0.789	0.873

Source: The authors, 2025

Overall the score looks good, except the overfitting problem of K-Nearest Neighbors. The test score shows that Random Forest Regression has the highest score, which indicates that this method can give the most appropriate predicted score for LPI Index in the near time.

The findings demonstrate that the PCA Biplot and Elastic-net 10 feature sets give the closest to adequate performance based on the error measurement criteria. The findings also suggest that ML algorithms are capable of assisting in the selection of a proper set of economic factors that indicate a country’s logistics performance. Furthermore, Random Forest Regression was shown to be the best effective prediction model in this investigation.

4.5. Discussion

4.5.1. Technical discussion

Finally, as shown in Table 10, we discussed the finding outcomes based on both feature selection techniques of filter and embedding method which is focused on the suggested statistical property and ML algorithm. The discussion describes the advantages and disadvantages of models that influence the findings of this study. To get good results, effective wrapper strategies, such as sequential search, or evolutionary algorithms, such as Particle Swarm Optimization (PSO) or Genetic Algorithm (GA), provide local optimum solutions and are computationally viable, are utilized. Because of the potential of overfitting and computationally costly (Takele TB, 2019), wrappers have a significant disadvantage, particularly in terms of computational inefficiency, which becomes more obvious as the feature space develops. The wrapper technique is thus eliminated from this analysis, although it will be significant in future studies.

Table 8. Finding discussion based on the study model

Feature selection strategy	**Approach**		**Advantage**	**Disadvantage**	**Finding discussion**
	Statistical property	ML algorithm			
Filter	Correlation		For many features, efficient, and fast (Guo, 2019)	The weak correlation subgroup may contain certain potentially advantageous traits that cannot be properly utilized (Guo, 2019).	The dependent and predictor variables exhibit a substantial association with the less advantageous feature set of A and B. This model might not be as effective for this research data set due to its drawbacks with the feature selection approach.
	PCA		Lack of features and low complexity (Fauvel, 2009) Reduced time and computational expense (Bolo, 2013) Strong capacity for generalization (Kwak, 2002)	The results could be significantly altered by just scaling some of the criteria (Kwak, 2002).	The potentially significant feature set of Sets C, D, E, and particularly Set F is provided by the PCA technique. One of its benefits is that it lacks few important features. However, a significant alteration that will alter the feature set's outcome based on its disadvantage is the difference in criteria, namely the factor loading and percentage of variation.

Embedded	LASSO		Provides a clear and interpretable model by selecting key predictors; useful when limited data is available (Boucher, 2015)	Randomly selects one correlated variable while ignoring others, potentially losing valuable information (Boucher, 2015)	LASSO is valuable for feature selection but limited in comparison to other ML algorithms when used for regression tasks
	E-net		Performs well when there are more parameters than samples; offers greater stability compared to LASSO (Boucher, 2015)	Ineffective for feature selection with limited data, as it struggles with high variable counts (Boucher, 2015)	This approach provides significant feature sets but is limited compared to filter-based ML algorithms
Other	ANOVA F-test		A flexible statistical method applicable to diverse conditions (Optimus, 2023)	Struggles with varying conditions and non-uniform effects (Luepsen, 2021)	Subset generation results are slower for ANOVA F-test compared to other methods
	RFE		Effectively removes redundant and irrelevant features for high-dimensional data (Yau, 2014)	Doesn't perform well on diverse datasets; lacks automatic stopping criteria for optimal feature subsets (Chen, 2007)	Less effective in handling large datasets
	Random Forest		Has advantages in identifying key predictor variables (Lu, 2022)	Selection bias and challenges in identifying informative variables (Speiser, 2019)	Random Forest yields weaker index results compared to other methods
		LR	Provides flexibility for label adjustments and better class separation (Fang, 2018)	Relies heavily on assumptions and struggles with high-dimensional and non-linear data (Yu, 2024)	SVR achieves good performance but faces challenges with complex data issues
		XGBoost	High predictive accuracy in various applications, superior to other models (Moore, 2022)	Requires complex hyperparameter tuning,	XGBoost achieves acceptable results but faces challenges in hyperparameter optimization

			demanding time and expertise (Pesantez-Narvaez, 2019)	
	K-Nearest Neighbors	Can handle various data types and forecasting problems (Burba, 2009)	Sensitive to noisy data and outliers, impacting performance (Song, 2017)	KNN demonstrates acceptable performance but performs poorly with noisy datasets
	Decision Tree Regression	Less sensitive to outliers compared to other models (Jena, 2020)	Can overfit if the tree is too deep, capturing noise instead of patterns (Huang, 2024)	Decision Tree Regression achieves acceptable results despite challenges with noisy data
	MLP-ANN	Useful for high-complexity data with non-linear correlations (Hundi, 2020)	Slow convergence and learning difficulties due to large parameter numbers (Huang, 2021)	MLP-ANN faces limitations in reaching optimal effectiveness due to network design restrictions
	SVR	Works well with a linear or non-linear kernel (Zhu, 2021)	Challenging to solve non-linear problems and kernel selection (Ahmadi, 2006)	SVR achieves acceptable results but struggles with kernel-related challenges
	RFR	Can handle high-dimensional data efficiently (Zhu, 2021)	Risk of overfitting and unstable results in noisy classification (Zhu, 2021)	RFR achieves acceptable results but suffers from overfitting in noisy data scenarios
	Ridge	Handles multicollinearity effectively and performs well with smaller datasets (Ahmadi, 2006)	Estimations may be biased (Ahmadi, 2006)	Ridge regression faces restrictions due to poorly correlated variables

Source: The authors, 2025

4.5.2. Economic discussion

The findings demonstrate that the PCA Biplot and E-net_10 feature sets give the closest to adequate performance based on the error measurement criteria. Based on the PCA-biplot, the selected features of set F out of 23 includes GD/GDP, Net_In, N_GDP, Imp, Exp, GE/GDP, GDP_C. Meanwhile, E-net_10 set only excludes Net_Out, GS/GDP, GD/GDP, CE throughout the full dataset.

Based on these findings, the authors recognize that normal economic growth indexes related to gross domestic product growth such as GDP and GE/GDP can be effective economic attributes to predict LPI scores. Other factors regarding trade openness and trading activity such as Imp, Exp also prove to be useful when being included in the predictive dataset. This result is consistent with Alnipak (2021), who stated that GDP per capita, the percentage of commercial service imports, and the liner shipping connectivity index significantly affect the logistics performance index at country level. Similarly, Bhatt (2021) also found that the relationship between LPI and trade is significantly affected by trade flow. This means that the developed framework can be used by countries to benchmark and implement relevant logistics policies, ultimately improving their LPI scores and global trade performances.

In addition, the study suggests possible datasets for future research related to logistics indices. The study identifies several datasets that could be valuable for future research related to logistics indices. These datasets include global supply chain performance metrics, transportation network efficiency data, warehouse optimization statistics, and trade flow analytics. By analyzing these datasets, researchers can develop more accurate models to assess logistics performance, predict disruptions, and optimize supply chain operations. Additionally, incorporating real-time tracking data and economic indicators can enhance the precision of logistics indices, leading to better decision-making for businesses and policymakers. Therefore, the study not only highlights existing datasets but also provides a foundation for further exploration and improvement in the field of logistics research.

5. Conclusion

In summary, this study demonstrates the application of machine learning regression for feature selection. It examines the impact of logistics performance using the Logistics Performance Index (LPI) alongside macroeconomic data from the World Bank. The dataset spans from 2010 to 2023 and initially includes 23 economic features. A trade-off is made between maximizing the number of instances and minimizing missing values in national economic data during the first stage of feature selection. Correlation-based filtering and Principal Component Analysis (PCA) are employed in the proposed feature selection process. Various machine learning regression models, including Linear Regression (LR), XGBoost, K-Nearest Neighbors (KNN), Multi-Layer Perceptron Artificial Neural Networks (MLP-ANN), Support Vector Regression (SVR), Random Forest Regression (RFR), and Ridge Regression, are used to train and validate the dataset based on selected features. Additionally, embedded methods such as penalized linear regression techniques, including LASSO and Elastic Net (E-net), are applied to refine feature selection, followed by continuous training and validation. Based on performance metrics such as MAE, MAPE, RMSE, NSE, and R^2, the PCA Biplot feature set (Set F) and the E-net feature set (Set J) yield the most reliable results. These feature sets can serve as viable alternatives, providing performance close to the best while ensuring maximum instances and minimal missing data in the dataset.

6. Limitations and future work

Future research may explore the integration of more diverse feature dimensions alongside economic attributes. Key factors linked to megatrends, such as carbon emission rates, fuel and renewable energy costs and consumption, and the expansion of the e-commerce market, could provide deeper insights into logistics performance in the evolving global supply landscape.

REFERENCES

Abimbola, O. P., Mittelstet, A. R., Messer, T. L., Berry, E. D., Bartelt-Hunt, S. L., & Hansen, S. P. (2020) “Predicting Escherichia coli loads in cascading dams with machine learning: An integration of hydrometeorology, animal density and grazing pattern”, *The Science of the Total Environment,* Vol. 722, pp.137894.

Ahmadi-Nedushan, B. et al. (2006), “A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment”, *River Res Appl,* Vol.22 No.5, pp. 503–523.

Alnıpak, S., Isikli, E., & Apak, S. (2021), “The propellants of the Logistics Performance Index: an empirical panel investigation of the European region”, *International Journal of Logistics Research and Applications*, Vol.26, pp. 894 - 916.

Arvis, J.F., Mustra, M., Panzer, J., Ojala, L. and Naula, T. (2007), “Connecting to Compete: Tradelogistics in the global economy World Bank: Washington”.

Arvis, J-F., Mustra, M. A., Ojala, L., Shepherd, B. and D. Saslavsky, (2012), “Connecting to compete 2012: Trade logistics in the global economy: the logistics performance index and its indicators”, *The World Bank*, 2012.

Arvis, J-F., Ojala, L., Shepherd, B. and D. Saslavsky, Busch, C. and A. Raj. (2014), “Connecting to compete 2014: Trade logistics in the global economy: the logistics performance index and its indicators”, *The World Bank*, 2014.

Arvis, J-F., Ojala, L., Shepherd, B. and Raj, A., Dairabayeva, K., Kiiski, T. (2018), “Connecting to compete 2014: Trade logistics in the global economy: the logistics performance index and its indicators”, *The World Bank*, 2018.

Babayigit, B., Gürbüz, F., & Denizhan, B. (2022), “Logistics performance index estimating with artificial intelligence”, *International Journal of Shipping and Transport Logistics,* Vol.1.

Baffa Sani Mahmoud et al. (2022), “A Machine Learning Model for Malware Detection Using Recursive Feature Elimination (RFE) For Feature Selection and Ensemble Technique”.

Bahzad Taha Jijo, Adnan Mohsin Abdulazeez (2021), “Classification Based on Decision Tree Algorithm for Machine Learning*”, Journal of Applied Science and Technology Trends*, Vol.2 No.01, pp. 20-28.

Bhatt, P., & Professor, A. (2023), "IMPACT OF LOGISTICS PERFORMANCE INDEX ON INTERNATIONAL TRADE", *International Research Journal of Modernization in Engineering Technology and Science,* Vol.5, pp.115-120

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013), "A review of feature selection methods on synthetic data", *Knowledge and Information Systems,* Vol.34 No.3, pp. 483–519.

Boucher, T. F., Ozanne, M. V., Carmosino, M. L., Dyar, M. D., Mahadevan, S., Breves, E. A., Lepore, K. H., & Clegg, S. M. (2015), "A study of machine learning regression methods for major elemental analysis of rocks using laser-induced breakdown spectroscopy", *Spectrochimica Acta. Part B: Atomic Spectroscopy,* Vol.107, pp. 1–10.

Burba, F., Ferraty, F., & Vieu, P. (2009), "k-Nearest Neighbour method in functional nonparametric regression", *Journal of Nonparametric Statistics,* Vol.21 No.4, pp. 453–469.

Burr, T. (2008), "Pattern Recognition and Machine Learning: Review of Pattern Recognition and Machine Learning*", Journal of the American Statistical Association*, Vol.103 No.482, pp. 886–887.

Çelebi, D. (2019), "The role of logistics performance in promoting trade", *Maritime Economics & Logistics,* Vol.21 No.3, pp. 307–323.

Chai, Meijuan (2021), "Application of ANN technique to predict the thermal conductivity of nanofluids: a review", *Journal of Thermal Analysis and Calorimetry*.

Chandrashekar, G. & Sahin, F. (2014), "A survey on feature selection methods*", Computers & Electrical Engineering,* Vol.40 No.1, pp. 16–28.

Chen, X., & Jeong, J. (2007), "Enhanced recursive feature elimination", *Sixth International Conference on Machine Learning and Applications (ICMLA 2007),* pp. 429-435.

Chen, Y. (2023), "Research on the Prediction of Boston House Price Based on Linear Regression, Random Forest, Xgboost and SVM Models", *Highlights in Business, Economics and Management*.

Coyle, D., & Weller, A. (2020), "Explaining' machine learning reveals policy challenges", *Science (American Association for the Advancement of Science),* Vol.368 No.6498, pp. 1433–1434.

D'Aleo, V., & Sergi, B. S. (2017), "Does logistics influence economic growth? The European experience", *Management Decision,* Vol.55 No.8, pp. 1613–1628.

Fang, X., Li, X., Lai, Z., Wong, W., & Fang, B. (2018), "Regularized Label Relaxation Linear Regression", *IEEE Transactions on Neural Networks and Learning Systems,* Vol.29, pp. 1006-1018.

Fauvel, M., Chanussot, J., & Benediktsson, J. A. (2009), "Kernel Principal Component Analysis for the Classification of Hyperspectral Remote Sensing Data over Urban Areas", *EURASIP Journal on Advances in Signal Processing,* Vol. 2009 No. 1.

Feizabadi, J. (2022), "Machine learning demand forecasting and supply chain performance", *International Journal of Logistics,* Vol.25 No.2, pp. 119–142.

Gerschberger, M., Manuj, I., & Freinberger, P.P. (2017), "Investigating supplier-induced complexity in supply chains", *International Journal of Physical Distribution & Logistics Management,* Vol.47 No.8, pp. 688–711.

Ghadery-Fahliyany, H., Ansari, S., Mohammadi, M.-R., Jafari, S., Schaffie, M., Ghaedi, M., & Hemmati-Sarapardeh, A. (2024), "Toward predicting thermal conductivity of hybrid nanofluids: Application of a committee of robust neural networks, theoretical, and empirical models", *Powder Technology,* Vol.437, pp. 119506.

Guo, J., Yang, L., Bie, R., Yu, J., Gao, Y., Shen, Y., & Kos, A. (2019), "An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring", *Computer Networks (Amsterdam, Netherlands: 1999),* Vol.151, pp. 166–180.

Guo, Q., Wu, W., Massart, D. L., Boucon, C., & de Jong, S. (2002), "Feature selection in principal component analysis of analytical data", *Chemometrics and Intelligent Laboratory Systems,* Vol.61 No.1, pp. 123–132.

Işık, G., Öğüt, H., & Mutlu, M. (2023), "Deep learning based electricity demand forecasting to minimize the cost of energy imbalance: A real case application with some fortune 500 companies in Türkiye", *Engineering Applications of Artificial Intelligence,* Vol.118, pp.105664-.

Huang, R., Ma, C., Ma, J., Huangfu, X., & He, Q. (2021), "Machine learning in natural and engineered water systems", *Water Research (Oxford),* Vol.205, pp. 117666.

Huang, X. (2024), "Predictive Models: Regression, Decision Trees, and Clustering", *Applied and Computational Engineering,* Vol.79, pp. 124-133.

Hundi, P., & Shahsavari, R. (2020), "Comparative studies among machine learning models for performance estimation and health monitoring of thermal power plants", *Applied Energy,* Vol.265, pp. 114775.

Islam, Yaseen & Siddiqui, Danish. (2019), "The Effect of External Factors towards the Logistics Performance in Efficiency of Material Supply for the Pipeline Construction Based Projects: Evidence from the Oil and Gas Industry in Pakistan", *SSRN Electronic Journal.*

Jena, M., & Dehuri, S. (2020), "DecisionTree for Classification and Regression: A State-of-the Art Review", *Informatica (Slovenia),* Vol.44.

Jordan, M. I., & Mitchell, T. M. (2015), "Machine learning: Trends, perspectives, and prospects", *Science (American Association for the Advancement of Science),* Vol.349 No.6245, pp. 255–260.

Kingsford, C. & Salzberg, S. L. (2008), "What are decision trees?", Nature Biotechnology, Vol.26 No.9, pp. 1011.

Kreif, N., DiazOrdaz, K., Moreno-Serra, R., Mirelman, A., Hidayat, T., & Suhrcke, M. (2022), "Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia", *Health Services and Outcomes Research Methodology,* Vol.22 No.2, pp. 192–227.

Krzanowski, W. J. (1987), "Selection of Variables to Preserve Multivariate Data Structure, Using Principal Components", *Applied Statistics,* Vol.36 No.1, pp. 22–33.

Kumar, N., Sundaram, K., R., R., & S., M. (2023), "Optimizing Energy Consumption in Smart Homes Using Machine Learning Techniques", *E3S Web of Conferences,* Vol.387, pp. 2002.

Kwak, N. & Choi, C.H. (2002), “Input feature selection for classification problems”, *IEEE Transactions on Neural Networks,* Vol.13 No.1, pp. 143–159.

Kıyak, B., Öztop, H. F., Ertam, F., & Aksoy, İ. G. (2024), “An intelligent approach to investigate the effects of container orientation for PCM melting based on an XGBoost regression model”, *Engineering Analysis with Boundary Elements,* Vol.161, pp. 202–213.

L'opez-De-Castro, M., Garc'ia-Galindo, A., & Armañanzas, R. (2024), “Conformal Recursive Feature Elimination”, ArXiv, abs/2405.19429.

Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006), “Studies in Fuzziness and Soft Computing”, pp. 137–165.

Lawrence, S. et al. (2013), “Source apportionment of traffic emissions of particulate matter using tunnel measurements”, pp. 548–557.

Li, Y., & Ma, W. (2010), “Applications of Artificial Neural Networks in Financial Economics: A Survey*”, International Symposium on Computational Intelligence and Design,* Vol.1, pp. 211–214.

Loh, W. (2011), “Classification and regression trees*”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* Vol.1.

Lu, M. (2019), “Embedded feature selection accounting for unknown data heterogeneity”, *Expert Systems with Applications,* Vol.119, pp. 350–361.

Lu, T. (2022), *Research on the Variable Selection Methods Based on Random Forests,* 2022 7th International Conference on Computational Intelligence and Applications (ICCIA), p. 59.

Luepsen, H. (2021), “ANOVA with binary variables: the F-test and some alternatives”, *Communications in Statistics - Simulation and Computation,* Vol.52, pp. 745-769.

Malakouti, S. M., Menhaj, M. B., & Suratgar, A. A. (2023), “The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction”, *Cleaner Engineering and Technology,* Vol.15, pp. 100664.

Min, H. (2010), “Artificial intelligence in supply chain management: theory and applications”, *International Journal of Logistics,* Vol.13 No.1, pp. 13-39.

Moore, A., & Bell, M. (2022), “XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study”, *Clinical Medicine Insights: Cardiology,* Vol.16.

Mulligan, D. K., & Bamberger, K. A. (2019), “Procurement as Policy: Administrative Process for Machine Learning”, *Berkeley Technology Law Journal*, Vol.34 No.3, pp. 773-852.

Muthukrishnan, R., & Rohini, R. (2016), “LASSO: A feature selection technique in predictive modeling for machine learning”, *in: Proceeding of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA),* pp. 18–20.

Ogunsanya, M., Isichei, J., & Desai, S. (2023), “Grid search hyperparameter tuning in additive manufacturing processes”. *Manufacturing Letters,* Vol.35, pp.1031–1042.

Ojala, L., & Çelebi, D. (2015), "The World Bank's Logistics Performance Index (LPI) and drivers of logistics performance".

Pearson, K. (1895), "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material", *Philosophical Transactions of the Royal Society of London. A,* Vol.186, pp. 343–414.

Pesantez-Narvaez, J., Guillén, M., & Alcañiz, M. (2019), "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression", Risks.

Qu, Z., Xu, J., Wang, Z., Chi, R., & Liu, H. (2021), "Prediction of electricity generation from a combined cycle power plant based on a stacking ensemble and its hyperparameter optimization with a grid-search metho",. *Energy (Oxford),* Vol.227, pp.120309-.

Ranjan, S., Kayal, P., & Saraf, M. (2023), "Bitcoin Price Prediction: A Machine Learning Sample Dimension Approach", *Computational Economics,* Vol.61 No.4, pp. 1617–1636.

Ray, A., & Chaudhuri, A. K. (2021), "Smart healthcare disease diagnosis and patient management: innovation, improvement and skill development", *Machine Learning with Applications,* Vol.3, pp. 100011.

Shepherd, B., & Sriklay, T. (2023), "Extending and understanding: an application of machine learning to the World Bank's logistics performance index", *International Journal of Physical Distribution & Logistics Management,* Vol.53 No.9, pp. 985–1014.

Shomope, I., Tawalbeh, M., Al-Othman, A., & Almomani, F. (2025), "Predicting biohydrogen production from dark fermentation of organic waste biomass using multilayer perceptron artificial neural network (MLP–ANN)", *Computers & Chemical Engineering,* Vol.192, pp.108900.

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017), "An efficient instance selection algorithm for k nearest neighbor regression", *Neurocomputing,* Vol.251, pp. 26-34.

Souza, J. T. d., Francisco, A. C. d., Piekarski, C. M., & Prado, G. F. d. (2019), "Data Mining and Machine Learning to Promote Smart Cities: A Systematic Review from 2000 to 2018", *Sustainability,* Vol.11 No.4, pp. 1077.

Speiser, J., Miller, M., Tooze, J., & Ip, E. (2019), "A comparison of random forest variable selection methods for classification prediction modeling", *Expert Systems with Applications,* Vol.134, pp. 93-101.

Sun, Z., Li, Y., Yang, Y., Su, L., & Xie, S. (2024), "Splitting tensile strength of basalt fiber reinforced coral aggregate concrete: Optimized XGBoost models and experimental validation", *Construction & Building Materials,* Vol.416, pp. 135133.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024), "An improved random forest based on the classification accuracy and correlation measurement of decision trees", *Expert Systems with Applications,* Vol.237, pp. 121549.

Syam, N., & Sharma, A. (2018), "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice", *Industrial Marketing Management,* Vol.69, pp. 135–146.

Takele, T. B. (2019), “The relevance of coordinated regional trade logistics for the implementation of regional free trade area of Africa*”, Journal of Transport and Supply Chain Management,* Vol.13 No.1, pp. 1–11.

Tealab, A., Hefny, H., & Badr, A. (2017), “Forecasting of nonlinear time series using ANN”, *Future Computing and Informatics Journal,* Vol.2 No.1, pp. 39–47.

Vieira, S. M., Sousa, J. M., & Runkler, T. A. (2010), “Two cooperative ant colonies for feature selection using fuzzy models”, *Expert Systems with Applications,* Vol.37 No.4, pp. 2714–2723.

Wong, W. P., & Tang, C. F. (2018), “The major determinants of logistic performance in a global perspective: evidence from panel data analysis”, *International Journal of Logistics,* Vol.21 No.4, pp. 431–443.

Wu, Y., & Zhou, Y. (2022). “Hybrid machine learning model and Shapley additive explanations for compressive strength of sustainable concrete”, *Construction & Building Materials,* Vol.330, pp.127298.

You, W., Yang, Z., & Ji, G. (2014), “Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination”, *Expert Systems with Applications,* Vol.41, pp. 1463-1475.

Yu, H., Fernando, R., & Dekkers, J. (2024), “Use of the linear regression method to evaluate population accuracy of predictions from non-linear models”, *Frontiers in Genetics,* Vol.15.

Zeng, D. D., Cao, Z., & Neill, D. (2021), “Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control”, *Technical Basis and Clinical Applications,* Vol.22, pp. 437-453.

Zhang, H., & Srinivasan, R. (2021), “A Biplot-Based PCA Approach to Study the Relations between Indoor and Outdoor Air Pollutants Using Case Study Buildings”, *Buildings (Basel),* Vol.11 No.5, pp. 218.

Zhang, Y., Gao, Z., Sun, J., & Liu, L. (2023), “Machine-Learning Algorithms for Process Condition Data-Based Inclusion Prediction in Continuous-Casting Process: A Case Study”, *Sensors,* Vol.23 No.15, pp. 6719.

Zhu, R., Hu, X., Hou, J., & Li, X. (2021), “Application of machine learning techniques for predicting the consequences of construction accidents in China”, *Process Safety and Environmental Protection*, Vol.145, pp. 293–302.

Zhuang, H., Liu, X., Wang, H., Qin, C., Li, Y., Li, W., & Shi, Y. (2021), “Diagnosis of early stage Parkinson's disease on quantitative susceptibility mapping using complex network with one-way ANOVA F-test feature selection”, *Journal of Mechanics in Medicine and Biology,* Vol.21 No.05, pp. 2140026.