Country Development Level Analysis Based on Multidimensional Feature Clustering and LightGBM Classification

Lingyu Zhao, Xiaofeng Li*
College of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China
**Corresponding Author*

Abstract: As globalization accelerates, significant differences emerge across countries in multi-dimensional indicators such as economy, education, health, and social welfare, making traditional single indices inadequate for comprehensively capturing their development patterns. This paper proposes a national development classification framework that combines unsupervised clustering with supervised classification enhanced by monotonicity. First, using seven macroeconomic indicators include per capita GDP, urbanization rate, higher education enrollment rate, life expectancy, minimum wage, fertility rate, and CPI change rate—the K-means algorithm is applied to classify countries into four development tiers (A/B/C/D), enabling an interpretable classification of the clustering results; Using the clustering levels as supervised labels, a LightGBM multi-classification model with monotonicity constraints is constructed for prediction. Positive correlation constraints are applied to positive indicators, while negative correlation constraints are applied to negative indicators, such as fertility rates, ensuring compliance with the corresponding monotonicity logic. Subsequently, consistency tests were conducted against the United Nations Human Development Index (HDI) categories, with correlation coefficients reaching high levels. This validated that the clustering classification generally maintains monotonic consistency with the HDI but is not entirely identical, demonstrating the superiority of this clustering method. The final model demonstrated excellent classification performance on the validation set. This method provides a comprehensive and logically sound assessment classification scheme for evaluating national development levels while balancing interpretability and predictive accuracy.

Keywords: K-means; LightGBM; Country classification; United Nations Human Development Index

1. Overall Research Approach

1.1 Introduction

With the development of globalization, differences in economic, educational, health, and infrastructure development among countries have gradually emerged. The traditional Human Development Index (HDI)[1] proposed by the United Nations Development Programme (UNDP) is a core indicator for measuring a country's social development. However, HDI only reflects a country's development from three aspects—life expectancy per capita, education level, and gross national income per capita—and cannot capture factors such as environment, urbanization, or labor force participation, nor the interactions between different indicators. Thus, HDI can easily overlook regional and group-level disparities.

To address these issues, this paper proposes a clustering method for national development levels based on multidimensional features. We cluster countries using a series of socio-economic indicators—such as GDP per capita, urbanization rate, education enrollment rate, minimum wage level, and CPI change rate—to identify similarities and differences in countries' development patterns. The clustering and grading method not only compensates for HDI's limited dimensionality but also uncovers latent group structures within macro indicators, which can help policymakers obtain a more segmented view of development.

This study first applies the K-means[2] clustering algorithm to the multidimensional data of countries and assigns four levels (A, B, C, D) from high to low. K-means is an unsupervised

learning algorithm whose goal is to find inherent structure in unordered data, i.e., to determine a 'National Integrated Development Level' (NIDL) for each country based on its multidimensional indicators. Then, using the labels generated by K-means as targets, we train a LightGBM[3] model using the original multidimensional national data as features. LightGBM is an efficient and powerful gradient boosting decision tree algorithm that will learn the country-level labels. After training, the model can accurately predict categorization labels from input data. Therefore, as new country data becomes available, the trained model can predict the most likely level for the new country. This provides a dynamic tool for understanding the development status of each country.

1.2 Overview of National Development Assessment Methods

Current measurements of national development primarily rely on authoritative composite indices, such as the Human Development Index (HDI) and the Social Progress Index (SPI)[4]. This study expands on traditional indicators by incorporating more core features to reveal national performance across economic, social, and humanistic dimensions.

The HDI proposed by the UNDP is one widely used composite indicator. It calculates a value between 0 and 1 by weighting three core dimensions: life expectancy per capita, the education index (mean years of schooling and expected years of schooling), and gross national income per capita, and divides countries into low, medium, high, and very high categories. HDI's advantage lies in having few indicators and strong comparability, but because it only includes three dimensions, it cannot fully reflect employment status, environmental quality, social inequality, and other important areas.

To compensate for HDI's limitations, the Social Progress Index introduces richer social and environmental dimensions. SPI starts from three major dimensions—Basic Human Needs, Foundations of Wellbeing, and Opportunity—further broken down into 12 components. It covers basic guarantees such as food and water security and basic medical care, as well as modern social progress factors like personal rights and information and communications. Compared with HDI, SPI does not take economic growth as a premise but more directly measures social welfare in areas such as education, health, environment, and civil rights.

Building on existing assessment indicators, this study proposes an unsupervised clustering and grading method based on multidimensional features. By introducing seven indicators—GDP per capita, urbanization rate, school enrollment rate, minimum wage, CPI change rate, fertility rate, and life expectancy—and using K-means clustering, countries are automatically assigned to four levels (A/B/C/D) from high to low. This data-driven clustering can reveal similarities in countries' development patterns, supplement the multidimensional blind spots of existing indices, and validate its reasonableness by comparison with HDI. Furthermore, a monotonicity constraint is introduced in the supervised classification stage to ensure that improvements in certain indicators will not lead to a lower predicted level, yielding a framework that combines completeness, objectivity, and logical consistency.

1.3 Experimental Approach and Preprocessing

The dataset used in this experiment is a public dataset on Kaggle. This comprehensive dataset contains indicators for most countries in the world, including demographic, economic, health, and education statistics. The dataset offers a global perspective for comprehensive analysis and comparison. HDI data are taken from UNDP public releases.

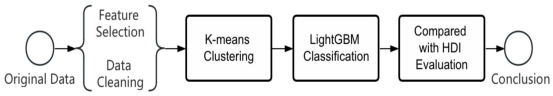

Figure 1. Flowchart of the Experiment

The overall workflow of the experiment is shown in Figure 1. First, raw country data are subjected to feature selection[5] and cleaning. Using unsupervised K-means clustering in the multidimensional feature space, countries are categorized and mapped to A–D levels based on cluster center means; these labels serve as targets for LightGBM training. Data preprocessing is critical to ensure the accuracy and stability of subsequent clustering and monotonic classification models. We defined unified cleaning functions for signed fields in macroeconomic and social indicators to remove non-numeric characters and convert values to floats, standardizing all indicators into computable numeric formats. Cleaned macro indicators were then merged with HDI group tables by country name to form a complete

dataset. The HDI textual labels (Low/Medium/High/Very High) were mapped to numeric labels 0–3. We retained only the features of interest and removed rows with any missing values to avoid bias from imputation.

For modeling needs, numeric features underwent two transformations: standardization[6] to mean 0 and variance 1, followed by normalization[7] to the [0,1] range to ensure features participate in clustering and classification on the same scale. The cluster labels (A/B/C/D) were converted to integers 0–3, and all preprocessors (scalers, encoders) and intermediate results were persisted for later reuse in model training and prediction modules. This standardized preprocessing pipeline provides a solid foundation for discovering development patterns via clustering and for monotonic LightGBM classification.

2. Multidimensional Feature Clustering and Grading

2.1 Feature Selection and Design

In this project, seven key indicators were selected from economic, social, educational, and health dimensions to capture differences in countries' development. GDP per capita (gdp_per_capita) and urbanization rate (urban_pct) reflect macroeconomic strength and infrastructure maturity[8]. GDP per capita removes the effect of population size and more fairly measures average national wealth; the urbanization rate reflects resource aggregation and public service efficiency that result from population concentration in cities. Both are closely related to living standards and industrial upgrading, so they are treated as positively correlated features—higher values should raise the predicted level.

In contrast, life expectancy and gross tertiary enrollment reflect social welfare and future development potential from health and education perspectives[9], respectively. Life expectancy is an intuitive indicator of public health, medical systems, and environmental management; tertiary enrollment is closely tied to a country's innovation capacity and industrial competitiveness. Including these two indicators helps capture long-term trends and latent drivers. Minimum wage represents social equity and income level, directly affecting purchasing power and domestic demand[10]. Fertility rate and CPI change rate characterize public service pressures and macroeconomic stability from the perspectives of population structure and price stability[11]. Because higher fertility or inflation generally correlates with lower development levels, these features are constrained to be negatively correlated. These design choices and directional constraints enhance the interpretability of the grading results.

2.2 K-means Clustering Algorithm

K-means is a common unsupervised learning method that aims to partition a dataset into K clusters so that points within the same cluster are as similar as possible, while points in different clusters are as dissimilar as possible. Similarity is usually measured by the distance between data points, such as Euclidean distance.

The core objective of K-means is to minimize the within-cluster sum of squared distances between samples and their cluster centers. The algorithm iteratively assigns each point to the nearest centroid and then recalculates centroids as the mean of points in each cluster. This process continues until convergence—centroids no longer move or a maximum number of iterations is reached. K-means is one of the simplest and most effective methods for data analysis and pattern recognition.

2.3 Clustering Results Analysis

In this experiment, clustering was used to classify countries according to their multidimensional development indicators: GDP per capita, urbanization rate, tertiary enrollment, average life expectancy, minimum wage, fertility rate, and CPI change rate. Using K-means, countries were successfully divided into four clusters, each representing a different development tier. Based on the cluster means for these indicators, clusters were labeled A/B/C/D, with A being the highest level and D the lowest.

Table 1. Four-level Country Preview

Level	Country	Level	Country
Level A	Australia	Level C	Thailand
	Canada		Iraq
	Greece		India
	UK		Botswana
	Spain		Jamaica
Level B	Mexico	Level D	Afghanistan
	Serbia		Kenya
	Malaysia		Pakistan
	Türkiye		Haiti
	Brazil		Uganda

Table 1 previews five example countries from

each level. Level A countries are mainly in Europe, North America, and Australia, and score high on all positive indicators: GDP per capita, urbanization, tertiary enrollment, and life expectancy[12]. They also tend to have low fertility and stable inflation, so they are assigned to higher tiers.

Conversely, D-level countries are often located in Sub-Saharan Africa, South Asia, and other developing regions. Their common features are low GDP per capita, short life expectancy, low education enrollment, and large CPI volatility. These low scores reflect challenges in economic development, social welfare, and infrastructure.

The clustering clarifies global differences in development and provides a reference framework for policymakers and international organizations. Further inspection of cluster center statistics shows that A-cluster means for GDP per capita are roughly twice the global average, while tertiary enrollment and life expectancy lead other clusters by 20–30%. D-cluster GDP per capita is around 30% of the global average, and life expectancy and enrollment are below the median. B and C clusters fall in between: B countries invest more in education and health than C, while C countries are often in Eastern Europe and Latin America and face uneven education, medical resources, and inflation pressures[13].

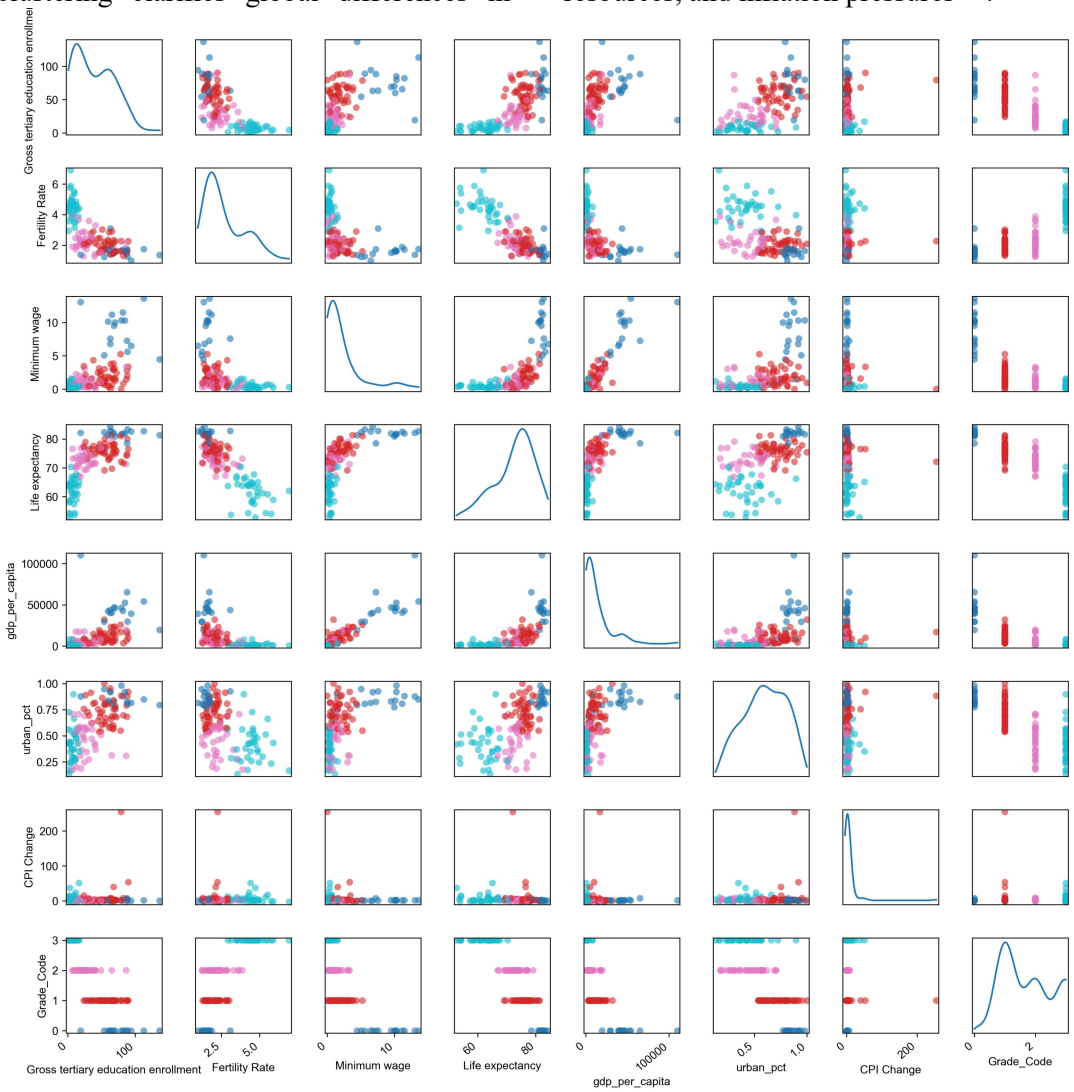

Figure 2. Scatterplot of Two-by-two Features

To analyze relationships between dimensions, pairwise feature plots were created (Figure 2), with diagonal panels showing each feature's distribution density. Each point represents a country and is colored by cluster label, allowing a visual inspection of relationships and level distributions.

The plots show clear positive correlations among GDP per capita, life expectancy, and tertiary enrollment: points tend to extend from the lower-left to the upper-right. Fertility rate is negatively correlated with those features, with

points roughly arranged from the upper-left to the lower-right in corresponding subplots, indicating that high fertility countries tend to have lower GDP and life expectancy.

Overall, K-means effectively partitions countries into four clear development tiers, consistent with HDI trends. The data-driven grades not only supplement the dimensional limits of traditional composite indices but also provide interpretable labels for subsequent monotonic classification modeling.

3. Monotonic LightGBM Classification Model

3.1 Monotonic Constraints in LightGBM

Monotonic constraints in LightGBM incorporate prior knowledge into the gradient boosting tree training process by enforcing that model predictions satisfy certain monotonic relationships with given features. Specifically, if a feature is specified as positively (or negatively) correlated, splits are restricted so that increasing the feature cannot decrease (or increase) the model output. This ensures the model's predictions are monotonic with respect to those features.

Introducing monotonic constraints narrows the search space of tree structures, which improves interpretability and trustworthiness—especially in policy evaluation or risk rating scenarios—by preventing local 'reversals' caused by noise or weak correlations. However, monotonic constraints require additional validation at split time, which can slow training, and overly strict constraints may hurt model fit. Therefore, a balance must be struck between monotonicity requirements and predictive performance.

3.2 Model Principles and Training

LightGBM (Light Gradient Boosting Machine) is a machine learning algorithm based on gradient boosted decision trees (GBDT). Like traditional gradient boosting, LightGBM iteratively trains weak learners and accumulates their predictions to form a strong model. Each iteration fits a tree to the residuals of the previous ensemble, optimizing a loss function and adding a regularization term to prevent overfitting.

The objective function and leaf weight formulas govern the training process: the total loss measures discrepancy between predictions and true values and includes regularization to control complexity. Leaf weights are updated based on first- and second-order gradients and are scaled by regularization terms. The model updates predictions by adding the new tree's output (scaled by the learning rate) to previous predictions.

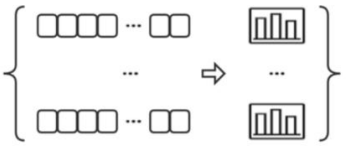

Figure 3. Histogram Algorithm Construction Process

LightGBM achieves efficiency in building individual trees by using a histogram-based algorithm, as shown in Figure 3, to discretize continuous values into bins and then searching for splits on these histograms. This reduces computation and memory consumption, making LightGBM highly efficient for large-scale, high-dimensional datasets.

In this study, LightGBM was used for country-level classification. The K-means clusters (A/B/C/D) serve as labels. Multidimensional country indicators were used as input features, and LightGBM iteratively learned to map these features to levels. During training, new trees correct cumulative prediction errors, and hyperparameters (learning rate, number of trees, max leaves, and regularization) were tuned to avoid overfitting and improve generalization.

The trained model can predict which level (A, B, C, or D) a country belongs to, capturing nonlinear relationships between levels. LightGBM performs well on large-sample, high-dimensional classification tasks, offering both speed and accuracy. The trained model is saved for later loading to predict new countries, completing the prediction system.

4. Experimental Results and Analysis

4.1 Consistency Check between Clustering and HDI

In data analysis and machine learning—especially when grading or classifying entities—a consistency check[14] is necessary to examine how closely a model's categorization agrees with a recognized standard. Consistency checks validate the reasonableness and credibility of clustering or grading results and strengthen their persuasive power[15].

In this experiment, countries were clustered by K-means and then classified by LightGBM. K-means discovers natural structures in an

unsupervised manner; to verify whether these structures align with human development levels (as measured by HDI), we performed consistency testing against UNDP HDI categories[16]. HDI combines variables such as life expectancy, education, and standard of living, and is an internationally recognized measure of development, so comparing cluster grades with official HDI grades is an important test of objectivity and effectiveness. High consistency indicates that meaningful development patterns similar to HDI exist in our clustering. However, excessively high consistency could mean the clustering adds little new information, so an appropriate consistency measure is required to support later analysis[17].

To quantify consistency between clustering and official HDI categories, two rank correlation coefficients were used: Spearman's[18] rho and Kendall's[19] tau. Both statistics measure ordinal association between variables and do not assume normality or linearity, making them suitable for comparing graded data.

The comparison with HDI produced a Spearman correlation coefficient of 0.839 and a Kendall coefficient of 0.771, indicating a positive association between cluster grades and HDI categories—higher cluster grades (from D to A) correspond to higher official HDI levels. The Spearman coefficient is well-suited for monotonic relationships, and a value of 0.839 supports the claim that clustering grades relates meaningfully to HDI. Kendall's tau also shows high agreement and is robust to outliers; a value of 0.771 suggests that in most pairwise comparisons, a country with a higher cluster grade also has a higher HDI grade. Importantly, the coefficient is not overly high, indicating that clustering retains useful differences from HDI.

Both coefficients conclude that the clustering grades are highly statistically significantly consistent with official HDI grades. This enhances confidence in the clustering results and implies that the four levels derived from K-means successfully reflect differences in human development. This supports the LightGBM training and ensures the prediction framework is practical and interpretable.

4.2 Comparative Analysis of Clustering and HDI

A violin plot (Figure 4) compares the HDI distributions across the four cluster levels, with sample counts annotated. Overall, countries in level A fall in the high HDI range, while most level D countries are in the low HDI range. Levels B and C show wider or multimodal distributions, indicating within-group heterogeneity and suggesting the presence of mid-to-high and mid-to-low subtypes within those groups.

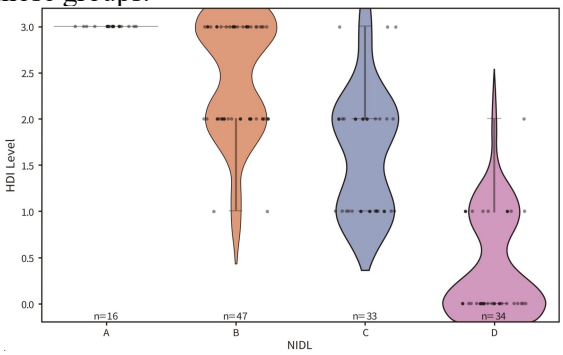

Figure 4. Comparison of cluster grading and HDI distribution

Countries with high HDI but low cluster grade typically perform well on HDI's three components but lag on other macro dimensions such as minimum wage, urbanization, or CPI stability. Conversely, countries with low HDI but high cluster grade may have strong GDP per capita, income, or urbanization while lagging on HDI components like education or health, or differences in data definitions may cause lower HDI.

Compared with a single HDI indicator, clustering based on multidimensional features offers clear advantages. Clustering uses more macroeconomic and social indicators to capture urbanization, GDP per capita, enrollment, minimum wage, and CPI changes, providing a more comprehensive profile. The multimodality and widespread use within B/C groups reveal within-group differences not captured by HDI, which is valuable for layered policy design and identifying subgroups that require targeted intervention. Clustering can also highlight outliers or marginal countries that differ from HDI, supporting case studies and data quality checks[20].

5. Conclusion and Future Work

5.1 Main Conclusions

This paper proposed an efficient method for automatic classification and prediction of national development levels by combining K-means clustering with a monotonic LightGBM classifier. First, countries were clustered based on multiple socio-economic indicators and

assigned A–D levels according to cluster center means, constructing an initial grading system. Then a LightGBM multiclass model was trained to learn the mapping between indicators and grades, enabling prediction for new country data while enforcing monotonicity for features with clear directional effects.

Results show the method effectively distinguishes countries at different development levels, and the trained model maintained good monotonicity on key features. The final model generalizes well to new data and can rapidly grade new countries. NIDL, compared with HDI, is a more comprehensive assessment that better supports cross-country comparison, policy making, and resource allocation. Overall, the framework integrates data-driven structure discovery with domain-informed constraints and offers a generalizable solution for measuring development levels.

5.2 Future Research Directions

Future work can expand feature selection and weighting beyond the fixed socio-economic indicators used here. New dimensions such as environmental sustainability, digital infrastructure, and governance quality could enrich the depiction of development. Automatic feature selection mechanisms or expert-informed weighting schemes may further improve accuracy and interpretability.

Regarding monotonic constraints, future research could explore methods to learn or adapt monotonic directions during training rather than relying solely on manual specification. Combining monotonic constraints with model interpretability methods (e.g., SHAP, LIME) may reveal hidden structural biases or inconsistencies.

Finally, from an application perspective, this method could be deployed as a policy-oriented assessment tool and extended with temporal modeling to track development trends over time, constructing a dynamic multidimensional development index system to better serve global monitoring and international cooperation goals.

References

[1] Programme U N D ,UNDP.Human Development Report 1990: Concept and Measurement of Human Development. Undp, 1990.

[2] Selim, Shokri Z., and M. A. Ismail. "K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality." IEEE Transactions on Pattern Analysis and Machine Intelligence 6.1(1984):81-87.

[3] Meng, Qi. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." Neural Information Processing Systems Curran Associates Inc. 2017.

[4] Stern, S., Wares, A., Orzell, S., & O'Sullivan, P. (2014). Social progress index 2014. Methodol Approach Wash Soc Prog Imp, 6.

[5] Kira, K., & Rendell, L. A. (1992). A Practical Approach to Feature Selection. Proceedings of the Ninth International Workshop on Machine Learning (ML 1992), Aberdeen, Scotland, UK, July 1-3, 1992. Morgan Kaufmann Publishers Inc.

[6] Hastie T, Tibshirani R, Friedman J .The elements of statistical learning. 2001. Journal of the Royal Statistical Society, 2004, 167(1):192-192.

[7] Freedman D, Pisani R, Purves R. Statistics: Fourth international student edition. WW Nort Co Httpswww Amaz ComStatistics-Fourth-Int-Stud-Free Accessed, 2020, 22.

[8] Chen M, Zhang H , Liu W ,et al. The Global Pattern of Urbanization and Economic Growth: Evidence from the Last Three Decades. Plos One, 2014, 9.

[9] Miladinov G .Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries. Genus, 2020, 76(1):1-20.

[10] Villarreal K .Urban Institute. John Wiley & Sons, Inc. 2017.

[11] Barro R J .Determinants of Economic Growth: A Cross-Country Empirical Study. American Political Science Review, 2003, 92(2):145-477.

[12] Ott J. World Bank World Development Indicators//Encyclopedia of Quality of Life and Well-Being Research. Cham: Springer International Publishing, 2024: 7858-7858.

[13] Cornia G A. Economic integration, inequality and growth: Latin America versus the European economies in transition. Review of Economics and Institutions, 2011, 2(2).

[14] Batini C, Scannapieca M .Data Quality: Concepts, Methodologies and Techniques. 2006.

[15] Shu, Xiaoling, and Yiwan Ye. "Knowledge Discovery: Methods from data mining and

machine learning." Social Science Research 110 (2023): 102817.

[16]Fahmiyah I, Ningrum R A. Human development clustering in Indonesia: Using K-Means method and based on Human Development Index categories. Journal of Advanced Technology and Multidiscipline, 2023, 2(1): 27-33.

[17]do Nascimento E R, de Albuquerque M A, de Oliveira Barros K N N, et al. Cluster analysis applied to the human development index (HDI) of Brazilian states. Research, Society and Development, 2022, 11(2): e18011225747-e18011225747.

[18]Hamilton, Martin A., R. C. Russo, and R. V. Thurston. "Trimmed Spearman-Karber method for estimating median lethal concentrations in toxicity bioassays." Environmental Science & Technology

[19]Lindskog, Filip, A. Mcneil, and U. Schmock. "Kendall's Tau for Elliptical Distributions." Contributions to Economics (2003).

[20]Mylevaganam S. The analysis of Human Development Index (HDI) for categorizing the member states of the United Nations (UN). Open Journal of Applied Sciences, 2017, 7(12): 661-690.

