ISSN: 2581-1274

Analysis of Unemployment Patterns in Indonesia Using K-Means Clustering and Identification of Dominant Factors Using Random Forest

Yuliana [a,1,*], Rima Fazri Ramadhani [a,2], Asep Abdul Latip [a,3] , Muhammad Farhan Harahap [a,4]

[a] *University Pamulang, Jl. Raya Puspiptek No. 46, Kel. Buaran, Kec. Serpong, South Tangerang City, Banten 15310, Indonesia*
[1] *dosen02557@gmail.com *;* [2] *rimafazriramadhani@gmail.com;* [3] *asepabdullatif741@gmail.com;* [4] *mfarhan0706@gmail.com*
** Corresponding author*

INFO ARTICLE

Riwayat article:
Published
April 12, 2026

Keywords:
Grouping of K-Means
Random Forest
Unemployment Patterns
Central Statistics Agency (BPS)
Economic Growth

ABSTRACT

The disparity in unemployment rates between provinces in Indonesia, exacerbated by the COVID-19 pandemic, is the main focus of this study. This study aims to (1) map the grouping of unemployment patterns in 34 provinces based on the Open Unemployment Rate (TPT) time series data for the 2020-2024 period, and (2) analyze the most significant socio-economic determinants of the formation of these patterns. Differentiating itself from previous research, this study uses a new two-stage hybrid framework that integrates the grouping of K-Means with the classification of Random Forest to address regional heterogeneity. Applying a two-stage methodology, cluster analysis using K-Means—validated through the Elbow Method and Silhouette Score of 0.456—succeeded in classifying the provinces into four different groups. The two prominent clusters identified were "Cluster 2: Pandemic Shock Pattern" (e.g., DKI Jakarta, West Java) which showed a surge in TPT above 10%, and "Cluster 3: Resilient Pattern" (e.g., Bali, DIY) which showed the lowest TPT rate and fastest recovery. Furthermore, the Random Forest Classifier analysis identified a hierarchy of determining factors, with the 2024 Average School Length (RLS) as the strongest predictor, followed by the 2024 Provincial Minimum Wage (UMP) and 2024 GDP. These findings underline that the quality of human resources (education) is a more crucial factor than economic output (GDP) in shaping the resilience of the labor market. The study concludes the need for different and cluster-specific unemployment policy interventions, rejecting a nationally uniform approach. Ultimately, this cluster-based analysis provides a strategic roadmap for policymakers to implement targeted interventions that prioritize human resource development over general economic policies.

I. Introduction

To date, unemployment is still one of the biggest challenges in Indonesia's economic progress that has not been fully resolved. The high unemployment rate not only affects the social conditions of the community, but also threatens the stability of the national economy and people's purchasing power. According to a report by the Central Statistics Agency, the Open Unemployment Rate (TPT) in Indonesia has experienced extreme fluctuations over the past five years. The unemployment rate has indeed begun to decline until 2024 after increasing sharply due to the shock of the COVID-19 pandemic in 2020–2021. However, at this stage of recovery, a new and urgent challenge arises: the

widening gap in post-pandemic recovery between provinces. Compared to agrarian areas such as West Nusa Tenggara or Central Sulawesi, the unemployment rate in provinces with high levels of industrialization such as West Java, DKI Jakarta, and Banten was recorded much higher. This shows that in the context of post-pandemic dynamics, economic structure and regional development inequality have a very significant influence on the distribution of job opportunities in Indonesia [1].

According to [2] Unemployment analysis needs to be carried out more systematically using a data-based scientific approach, not just through descriptive statistical comparisons. Analysis methods such as K-Means Clustering can be used to group provinces in Indonesia based on the similarity of socio-economic characteristics such as education level, economic growth rate, and labor force participation. With this approach, each region can be grouped into clusters that represent the same employment conditions and potential, so that employment policies can be directed more specifically and effectively. To group data based on its characteristics, the K-Means Clustering algorithm is used, since its implementation is easier and relatively faster [3].

Research [4] explained that the K-Means algorithm is included in the category of unsupervised learning which functions to separate data into groups (clusters) based on the similarity distance between variables. In the context of unemployment, K-Means can help the government identify provinces that have similar employment patterns and welfare levels [5]. Analysis results from [6] shows that the use of K-Means is able to reveal significant differences between regions based on human development indicators and labor productivity levels. In computer science and statistics, K-Means is a method for solving complex data segmentation problems. This technique works by organizing quantitative data into groups based on their common characteristics, so that members in one group are much more similar to each other than members in different groups[7].

According to [8] Random Forest is a combined technique that works by building a set of Decision trees during the training process. The final prediction is then determined by combining the results from all the trees to improve accuracy. Meanwhile, the Random Forest approach provides additional capabilities in predictively analyzing the factors that cause unemployment. It can process various variables simultaneously and determine the factors that have the most influence on the unemployment rate. Random Forest is included in supervised learning that uses a set of decision trees to improve the accuracy of prediction results. With this method, factors such as education, economic growth, and investment can be sorted according to their importance in influencing unemployment in each province[9].

In addition, the development of the Geographically Weighted Random Forest (GWRF) method that allows unemployment analysis to be carried out spatially, taking into account the differences in conditions between regions. This model has been shown to be more accurate in mapping the relationship between socioeconomic factors and geographical location, so that the results can provide a more realistic picture of unemployment patterns in Indonesia [10]. However, existing studies rarely incorporate these methods into an integrated framework to address the complexities of post pandemic recovery.

High economic growth does not necessarily automatically reduce the unemployment rate, because not all sectors are able to absorb large amounts of labor. This reinforces the view that development policies are needed that are not only oriented towards increasing GDP, but also on equal distribution of employment opportunities and improving the quality of human resources.

Overall, the application of the K-Means and Random Forest methods is an important step to understand the pattern of unemployment in Indonesia in depth. Through data-driven analysis, governments and research institutions can leverage the results of these modeling to strengthen national economic and employment development strategies, while ensuring that the policies implemented are truly in line with the socio-economic characteristics of each region.

Thus, the background shows that this study explicitly aims to fill the research gap in previous research on employment. Previous research has been limited to mapping regional clusters based on unemployment characteristics, without conducting further analysis of the causes of the formation of

such patterns. There has not been a comprehensive objective mapping that shows the unemployment cluster and the most dominant socio-economic factors within it.

This study uses a new two-stage hybrid methodology to address these differences. This study uses the K-Means algorithm to address regional heterogeneity first by forming different groups. Then, they used Random Forest to extract a hierarchy of specific determinants for each cluster, which differed from a single-stage approach. As a result, the objectives of this study are:

1. Conducting mapping and study of unemployment clusters in 34 provinces of Indonesia during the 2020-2024 period, by applying *the K-Means Clustering* algorithm.

2. Analyse the level of significance of socio-economic factors (including GDP, Average School Age, and UMP) as the main determinants influencing the formation of each unemployment cluster, using the *Random Forest*[11].

II. Method

This research was designed quantitatively by integrating two computational methods into one hybrid framework. The first phase relies on *an unsupervised learning* algorithm (*K-Means Clustering*) to segment the heterogeneity of the regional unemployment rate. Meanwhile, the second phase applies *a supervised learning* algorithm (*Random Forest Classifier*) purely for the purpose of analytical exploration—i.e. identifying the most dominant socio-economic factors—and is not intended as a predictive model. The research workflow continues from data preprocessing to cluster formation, and ends with feature significance analysis.

A. Research Workflow

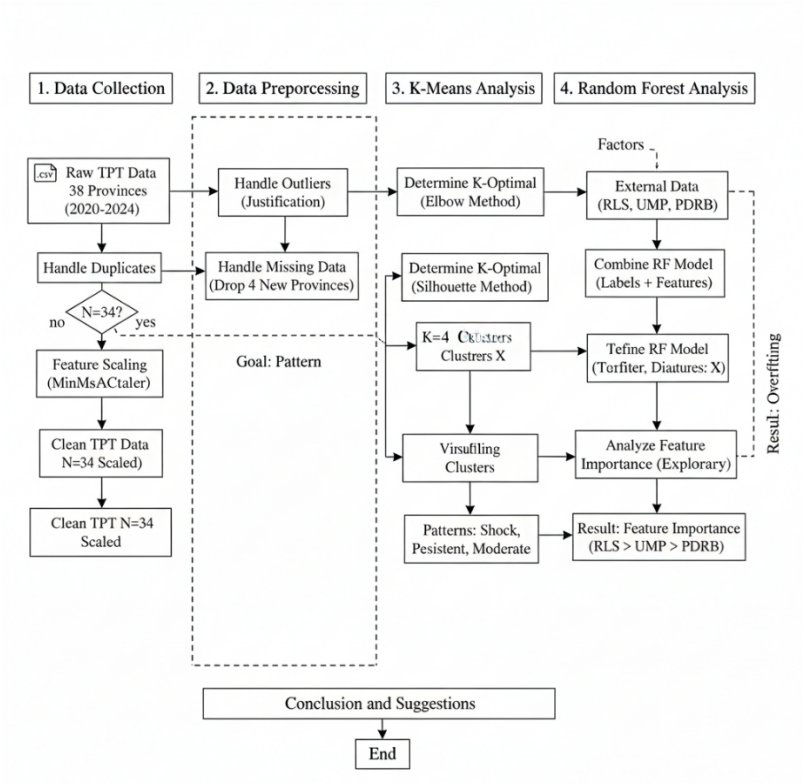

Fig 1. Research workflow

The first step in the flow is to collect the data that is already available for analysis, the next step is to clean the data of missing values, duplicate the data, check the outliers and change the data so that it can be read by the algorithm, the third step is to implement the K-Means algorithm to look for

cluster patterns from the analyzed data, the last part is to use the Random Forest algorithm to find the factors that affect the unemployment rate.

B. Data Collection and Preparation

The data collection that will be analyzed in this study uses secondary methods, namely quantitative data obtained from official publications of government agencies, including:

1. Open Unemployment Rate (TPT), Average School Length (RLS), and Gross Regional Domestic Product (GDP): which are a reference in the socio-economic pattern of society [12]. Sourced from the Central Statistics Agency (BPS) through https://www.bps.go.id period 2020-2024.
2. Provincial Minimum Wage (UMP): Sourced from the Ministry of Manpower through https://satudata.kemnaker.go.id for the same period.

Furthermore, the data preparation stage is the stage where data is processed which aims to provide relevant information and process the data to be clean, consistent and provide the best performance model. In the preparation of data, there are important stages, namely:

1. Handling duplicate data is part of the data cleansing process, which is to ensure that there is no duplication of data. The findings from the data used that there are no duplicates at all
2. Four new autonomous provinces in the Papua region were excluded due to the absence of equivalent historical time series data, so the number of observations was fixed to N = 34 provinces.
3. Outliers detected in major industrial provinces during the period August 2020 – August 2021 were deliberately maintained. This data is not deleted because it is a factual representation of structural anomalies due to the "Pandemic Shock".
4. To prevent weight inequality in the calculation of distances on the K-Means algorithm, TPT (10-dimensional) time series data is normalized using *the Min-Max Scaler*. This method converts the entire range of values into a scale [0,1] through the following mathematical equation:

$$X_scaled = \frac{(X - X_min)}{(X_max - X_min)} \quad (1)$$

C. Analysis and Evaluation Methods

The variables p and q in the formula represent the representation of the provincial data point and *its centroid* in the n-dimensional space. In order to determine the most optimal value of K, the *Elbow Method* approach is applied by tracking the extreme drop point in the *Within-Cluster Sum of Squares* (WCSS) value. The results of clustering were then tested for quality—both in terms of cohesion and separation between clusters—using *Silhouette Score*. The closer it is to +1, the more perfect the separation of the cluster, according to the following formula:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (2)$$

Where and is the province and centroid while n it is the 10 dimension or feature of time pq

1. Iteration, the model will repeat until the centroid converges.

K-Means evaluation, this evaluation is carried out with the aim of finding out the results of the accuracy of the model to be used. Here is the evaluation at K-Means.

1. The Elbow method is used to determine the optimal number of clusters by observing a drastic decrease in the Within-Cluster Sum of Squares (WCSS). Meanwhile, the Silhouette Score is a model evaluation stage that determines the quality of cluster separation to be further validated. From this score determination, it will measure how well a province with compatibility characteristics with its cluster $(a(i))$ compared to other clusters. In the results of the evaluation, the average score was 0.456 which was applied to confirm K= 4. Mathematically, silhouette scores are written as follows$(b(i))$.

$$s(i) = \frac{(b(i) - a(i))}{max(a(i), b(i))} \quad (3)$$

2. PCA visualization aims to reduce the 10 dimensions of the feature to 2 dimensions to visualize the cluster's results for further analysis.

Entering the second stage, the *Random Forest Classifier* algorithm [11] is run by making the results of labeling K-Means as a dependent variable (Y), while the socio-economic indicators (GDP, RLS, UMP) act as the predictor variable (X). Because the dataset used covers the entire population but with a very limited sample size (N=34), this model has a high risk of *overfitting*. As an anticipatory step, a rigorous evaluation was carried out by reviewing the comparison between *training accuracy*, *testing*, and *Out-of-Bag* (OOB) scores. This evaluation is essential to reinforce the limitations of the study: that the use *of Random Forest* here is focused purely on exploratory analysis to extract the significance of features (*Feature Importance*) based on *Gini Impurity* reductions, rather than designed for future forecasting.

III. Results and Discussion

The first part to analyze unemployment patterns using K-Means is to define the optimal (K) which aims to group data from 34 provinces into clusters. (K) optimal is obtained from the results of the Elbow Method which determines the number or number of data groups in which the Elbow Method is evaluated with a Distortion Score or also known as WCSS. This distortion score indicates that each centroid or data point to center is measured based on the calculation of the total squared distance. If there is a solid cluster, it indicates that the distortion value is low and vice versa.

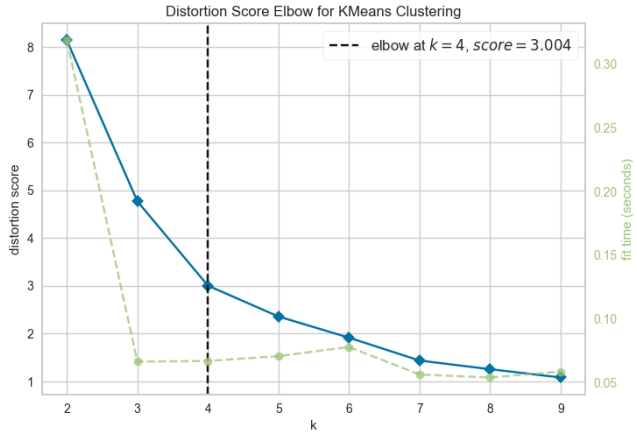

Fig 2. Elbows to Find the Optimal K

The figure shows that the test range is in the number of clusters in the range K = 2 to K = 9. Gbar shows blue graphs ranging from K = 2 scores ≈ 8.1 to K = 4 scores ≈ 3,004 graphs showing a significant decrease in distortion. However, at K=4 there is a vertical line separating the steep graph and the slope graph, which means that the addition of other clusters such as K=5, K=6 and so on is not capable of providing a commensurate decrease in model complexity. Then the optimal K = 4.

The evaluation of the Elbow results is to evaluate the optimal K justification results of 4 clusters using Silhouette Score to provide the quality parameters of separation between clusters. Its function is to calculate each sample with its similarity characteristics or called cohesion compared to other clusters.

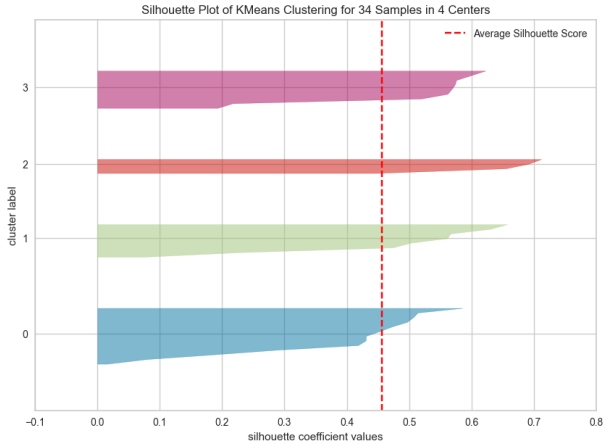

Fig 3. Visualization of Elbow Evaluation with Silhouette Score

The results of Elbow's evaluation were carried out using the Silhouette Score which showed that the red vertical dotted line was the result of the average Silhouette Score score, which was 0.456. This means that this score is considered the best because it shows a reasonable cluster structure. The reason for this is the indication of the Silhouette value which has a range of -1 to +1. If the result is -1 it is interpreted as a bad result and causes an error in the grouping, if the result is 0 it means that the sample is right at the border between the two clusters, and if +1 indicates perfect separation where the sample is more similar to its own sample characteristics than the rest of the cluster.

In addition, from the four clusters, no indication of negative coefficient values was found, meaning that no province was indicated to be misclassified. The image above means that clusters 2 and 3 show the best scores, both clusters are characterized by very dense characteristics and cluster members are above average. Below that cluster is cluster 1 which is also a good cluster with average members. However, a cluster of 0 indicates that the cluster is weak because it scores below average or is to the left of the red line, meaning that there is not a single sample or province that is wrong in the grouping but still has a weak score and can still be said to be relevant within the cluster itself.

The next stage is to create a visualization of the Key Component Analysis (PCA) to demonstrate the deployment and validate the cluster separation. In the loaded data, the data has 10 dimensions which are a crossover of features and time series. The function of PCA in analyzing data in the K-Means algorithm is to reduce the data from 10 dimensions to two dimensions of the main components, where the two components represent the largest data variance.

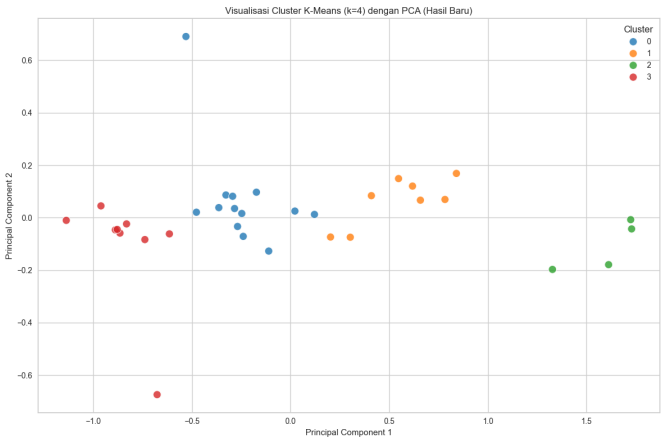

Fig 4. Visualization of Key Component Analysis (PCA)

From the images generated from the data of 34 provinces that have been processed and digitized by the PCA, the PCA presents four clusters with varying distributions. Clusters can be described by interpreting cluster separation from 0 to 3 with a clear indication of the separation between clusters. Cluster patterns 2 (green) and 3 (red) show a very significant and isolated separation from the right and left sides of the plot, cluster separation means there is a unique data pattern at the open unemployment rate. On the other hand, cluster 1 is orange and 0 is blue, indicating that the cluster is in the center and visually separated with a less significant cluster spread.

The PCA graph has a distribution that is far apart, namely in cluster 0 and cluster 3. The blue cluster 0 is far above Main Component 2 with a score of 0.7, while the red cluster 3 coordinates far below Main Component 2 with a score of -0.7. These points are outliers of the analyzed data, the reason for maintaining these outliers is due to the impact of the COVID-19 pandemic and causes a high number or percentage of layoffs recorded by the Central Statistics Agency as open unemployment with the peak of the pandemic in 2020-2021 precisely from August 2020 to August 2021. With cases like this, it proves that PCA not only records general patterns but also records unique patterns of extreme data variance.

After the analysis is carried out, the next stage is to conduct cluster profiling. Cluster profiles have the function of searching or understanding characteristics and identifying unique characters in the composition and analyzing data patterns to provide insights into each group. The first part in the findings of the provincial composition in each cluster is presented in the table below.

Table 1. Clusters and Number of Provinces in Each Cluster

Cluster	Number of Provinces	Name of Province
0	13	Riau, Jambi, South Sumatra, Lampung, Bangka Belitung Province, East Java, West Kalimantan, Central Kalimantan, South Kalimantan, North Kalimantan, South Sulawesi, North Maluku, Papua
1	8	Aceh, North Sumatra, West Sumatra, Central Java, East Kalimantan, North Sulawesi, Maluku, West Papua
2	4	Riau District, DKI Jakarta, West Java, Banten
3	9	Bengkulu, DI Yogyakarta, Bali, West Nusa Tenggara, East Nusa Tenggara, Central Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi

The next stage is to display an analysis table of average unemployment patterns in the range of 2020 – 2024 in each cluster, this table is used to find the cause of the grouping of 34 provinces carried out by the K-Means algorithm. Below is a table of average per cluster features.

Table 2. Average Features per Cluster

Cluster	February 2020	August 2020	February 2021	August 2021
0	4.206154	5.273846	4.864615	4.891538
1	5.638750	6.933750	6.492500	6.470000
2	6.707500	10.597500	9.140000	9.302500
3	2.788889	4.302222	3.934444	3.796667

Cluster	February 2022	August 2022	February 2023	August 2023
0	4.565385	4.471538	4.264615	4.255385
1	6.107500	6.093750	5.787500	5.761250
2	8.225000	7.952500	7.760000	7.072500
3	3.674444	3.351111	3.401111	3.018889

Cluster	February 2024	August 2024
0	4.201538	4.363077
1	5.355000	5.388750
2	6.725000	6.507500
3	3.021111	2.885556

Furthermore, table visualization is done to make it easier to find contrast differences in data patterns. Here are the plot results from the table to the line graph.

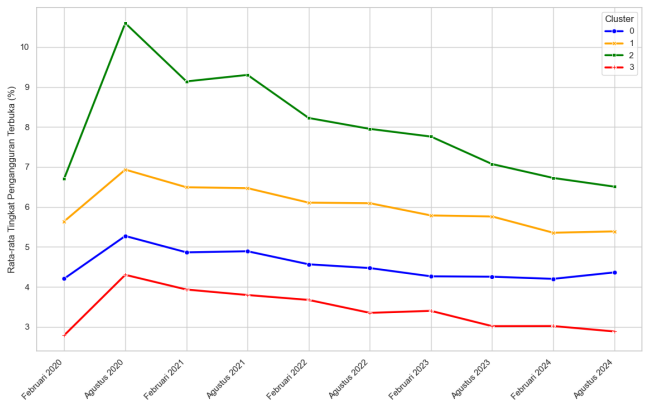

Fig 5. Graph of TPT average pattern per cluster

From the tables and figures that contain the average pattern of the open unemployment rate that is comprehensively analyzed, it is possible to identify and name each unique cluster that contains the pattern of unemployment in Indonesia. Among them are the following:

1. Cluster 2 is a green line, the pattern shows the shock of the pandemic and is heavily affected. The cluster covers four provinces: DKI Jakarta, West Java, Banten, and Riau Islands, all of which are major centres of industrial, service, and large-scale population movement activities in Indonesia. As shown in Figure 4.3, the provinces in this group already had a relatively high open unemployment rate before the pandemic, which was around 6.7%. During the COVID-19 pandemic, this group experienced the most significant increase in open unemployment compared to other clusters, with an average of 10.6% in August 2020. Despite the downward trend in the following period, the unemployment rate remained high, at around 6.5%, making this group the group with the highest open unemployment overall until the end of the observation period.

2. The red line of cluster 3, with a pattern that shows the lowest resilience and unemployment, the number of provinces in cluster 3 consists of nine provinces, including Bali, DI Yogyakarta, NTB, and NTT. Cluster 3 is the lowest unemployment graph compared to other clusters In cluster 3 in the period from August 2020 to August 2021 there was a significant spike, the surge was caused by the COVID-19 pandemic which caused an increase from 2.8% to 4.3%. Despite the impact of the pandemic, this cluster is the lowest of the others and shows a very rapid recovery with the open unemployment rate in the last period 2.9% lower than before the pandemic.

3. Cluster 1 is the yellow line, with a pattern indicating a pattern of open unemployment rates with high percentages and slow recovery. This cluster consists of eight provinces, including Aceh, North Sumatra, Central Java, and East Kalimantan. . The graph showing cluster 1 tends to persist at high levels (below after cluster graph 2 is green). In the observation of the cluster 1 graph, the open unemployment rate started at the 5.6% figure recorded before the COVID-19 pandemic. However, in the August 2020-2021 period, there was a surge from 5.6% to around 6.9% during the peak of the pandemic. However, the decline occurred after the peak of COVID with the endemic period or new normal enforced, causing a decrease in the percentage of open unemployment rate. Post-pandemic recovery in cluster 1 tends to be slow. This condition shows that having problems with the unemployment rate is chronic and structural compared to other clusters.

4. Cluster 0 is a blue line, with moderate and stable indications, meaning that the pattern in this social dynamic shows moderate and geographically diverse social dynamics, the cluster with the most provinces is 13 provinces consisting of the provinces of East Java, Riau, Papua, and others that are included in cluster 0 in the table. The characteristics of the cluster 0 pattern are the characteristics of the national average which began with the open unemployment rate in the range of 4.2% and at the peak of the pandemic there was an increase from 4.2% to 5.3% and recovered stably at the endemic or end of the pandemic to 4.3%.

After analyzing the data using the K-Means algorithm, the next stage is to look for the factors that affect the high and low unemployment percentages in 34 provinces in Indonesia. The algorithm used in this factor analysis uses the Random Forest Classifier algorithm. This algorithm provides insight from the results of the analysis of factors that affect the GDP variable as X1, the average length of school (RLS) as X2, and the provincial minimum wage as the X3 variable against the Y variable, which is a cluster containing all the provinces analyzed in the K-Means study. The results of using the Random Forest Classifier have two parts, namely the evaluation of the model and the results of the influence of variables on the analysis of the Random Forest Classifier.

In the analysis of the open unemployment rate for the evaluation of the Random Forest Classifier model, it was found that the results of the evaluation were on the training data (Training Accuracy) of 1.0000 (100%), the OOB (Out-of-Bag) score of 0.5556 (55.6%), and the accuracy of the test data (Test Accuracy) of 0.4286 (42.9%). There is a very striking difference between the model's performance on the training data and the test data, as shown by the evaluation matrix analysis. With a hundred percent accuracy rate on the training data, the model was able to classify all 27 training samples without errors. However, when tested on seven new samples that had never been tested before, the accuracy dropped dramatically, reaching only 42.9 percent to 55.6 percent. This condition is a strong overfitting signal, which occurs when the model is overturned to training data so as to capture noise rather than generalizable patterns. This kind of situation is understandable given the very small data set (N = 34). Complex models like Random Forest are more susceptible to excessive training data learning. Therefore, this model is unreliable for prediction purposes, such as projecting the formation of new provincial clusters in the future.

But this study uses Random Forest for exploratory analysis rather than predictive capabilities. The main goal is to examine the internal structure of the model and find out which variables are considered most influential by the algorithm during the learning process, even if the learning process is too tight. For the purposes of this exploration, an analysis of important characteristics is still relevant and worth presenting.

After the results of the model evaluation are found, the next step is to present the results of the analysis of open unemployment data using the Random Forest Classifier.

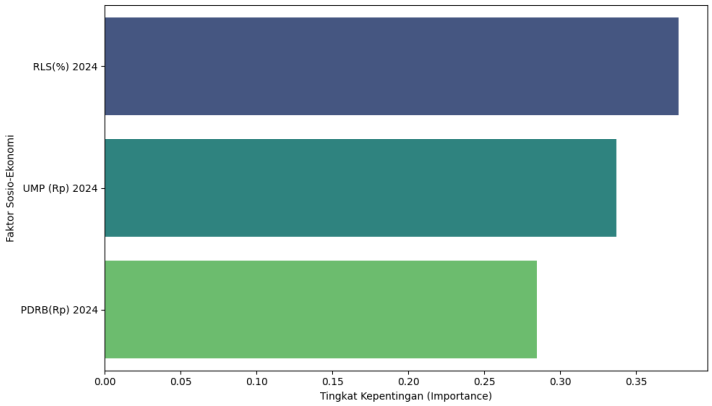

Fig 6. Ranking of factors influencing unemployment

From the results of the Random Forest Classifier model on the importance of features, three factors that affect inter-provincial unemployment in Indonesia were obtained. The factors shown in the figure above are Gross Regional Domestic Product (GDP), Provincial Minimum Wage (UMP) and Average School Length (RLS). The explanation is as follows.

1. RLS (%) 2024 with a score of 0.378. The variable that contributes the most is the average school time. This shows that the main factor that distinguishes the characteristics of unemployment in each province is the quality of human resources. Notably, this variable shows important differences between provinces in the High Persistent Pattern (Cluster 1) and provinces in the Resilient Pattern (Cluster 3).
2. UMP (Rp) 2024 with a score of 0.337. The second most influential factor is the Provincial Minimum Wage. These results are consistent with the findings of K-Means, which found that provinces with high industrial and service wage structures—such as DKI Jakarta and West Java in Cluster 2 of the Pandemic Shock Pattern—display different patterns of unemployment dynamics compared to agrarian provinces in other clusters.

3. GDP (Rp) 2024 with a score of 0.285. The third most significant factor is the Gross Regional Domestic Product. Although it has a significant impact, this factor is not the main factor, showing that the scale of the regional economy is the only factor that causes differences in unemployment patterns.

Overall, Random Forest's findings show that the dynamics of unemployment in Indonesia are more influenced by the labor cost structure (UMP) and the quality of human resources (RLS), rather than just the measure of the regional economy (GDP). These results provide a broader view of the components that make up the vulnerability and resilience of interprovincial unemployment.

The study moves beyond descriptive grouping to analyze the structural determinants of unemployment. The findings from the Random Forest analysis, which highlighted Average School Length (RLS) and Minimum Wage (UMP) as the most dominant factors, offer important insights into Indonesia's labor market.

1. Empirical facts that show that the significance of RLS (0.378) far exceeds GDP (0.285) confirm that the unemployment problem in Indonesia is rooted in structural issues, not just cyclical fluctuations. From a theoretical point of view, these results are in line with the framework of Human Development and Human Capital Theory. The theory asserts that the success of the absorption of productive labour is determined more by the suitability of education and skill levels, rather than relying solely on aggregate *economic output*. Furthermore, the high unemployment rate that remains in Cluster 1 and Cluster 2 emphasizes the occurrence of *skill mismatch* (competency mismatch). In this situation, the increase in GDP from the industrial sector fails to absorb the local workforce automatically due to the imbalance between the qualifications of graduates of educational institutions and the real needs of the industrial world.
2. Urban Wage and Fine Rigidity (Cluster 2) Identification of Cluster 2 (DKI Jakarta, West Java, Banten) as a "Pandemic Shock" group highlights the vulnerability of industrial estates. UMP's high interest score (0.337) supports the theory of wage rigidity. In this high-wage province, industries facing the shock of the pandemic preferred layoffs to wage reductions to maintain efficiency, leading to a drastic spike in unemployment (up to 10.6%). This phenomenon underscores the "Urban Penalty" during the crisis, where regions that are highly integrated into formal global markets are less resilient than agrarian regions.
3. Informal Economic Resilience (Cluster 3) On the other hand, Cluster 3 (Bali, DIY, NTB) showed the fastest recovery ("Resilient Pattern"). This resilience can be attributed to the elasticity of the informal and agricultural sectors, which act as a buffer when the formal tourism sector collapses. Lower minimum wage pressures in the region allow for more flexible labor absorption during the recovery phase.
4. Cluster-Specific Policy Recommendations Based on these findings, the national "one size fits all" policy is ineffective. We propose the following targeted interventions:

    a. For Cluster 2 (Pandemic Shock Pattern): Policies should focus on the Active Labor Market Policy (ALMP), especially re-skilling and up-skilling programs for urban workers displaced due to industrial shifts. Strengthening urban social safety nets is essential to mitigate future shocks.
    b. For Cluster 3 (Resilient Pattern): Interventions must focus on modernizing the informal sector and MSMEs (MSMEs). Digitalization and access to credit for micro enterprises will increase their productivity without disrupting their labor absorption capacity.
    c. For Cluster 1 (High Persistent Pattern): The government should address the root causes of "skills mismatch" by revitalizing the vocational education curriculum to align with the demands of local industries, rather than focusing solely on increasing GDP.

IV. Conclusion

This study analyzed the open unemployment rate using two different algorithmic approaches, namely K-Means clustering and Random Forest Classifier, which were applied to interprovincial data with a total of 34 observations. The results showed that the K-Means algorithm managed to identify the optimal cluster division of the four groups, supported by a silhouette score of 0.456. These clusters revealed different patterns of vulnerability and resilience during the pandemic period, where one cluster experienced severe unemployment shocks exceeding 10 percent, while another cluster showed greater resilience with consistently lower unemployment rates and faster recovery.

Methodologically, this study contributes by combining unsupervised and supervised machine learning techniques to explore structural patterns in regional unemployment. While K-Means captures heterogeneity between provinces, the Random Forest model provides insight into the relative importance of socio-economic determinants. A feature significance analysis shows that education-related factors play a more dominant role than purely economic indicators, with RLS emerging as the most influential variable, followed by UMP and GDP. These findings underscore the importance of the quality of human resources in shaping regional labor market outcomes.

From a policy perspective, the results suggest that efforts to reduce open unemployment should not rely solely on economic growth measures, but also emphasize investment in education and workforce development to strengthen regional resilience.

However, this study has some limitations. The small sample size leads to overfitting in the Random Forest model and limits the generalization of the findings. In addition, the analysis is limited by the use of only three socio-economic variables. Therefore, the results should be interpreted primarily for exploration purposes. Future research is encouraged to incorporate more detailed variables, larger datasets, and comparative modeling approaches to improve resilience and predictive performance.

References

[1] K. Y. Mahendra, M. Susilawati, N. Luh, and P. Suciptawati, "Modeling the Open Unemployment Rate," E-Journal of Mathematics, vol. 10, no. 1, pp. 20–25, 2021.

[2] R. Maliki, K. Falgenti, S. Priani, F. Fithri, M. Suherman, and D. S. Nugraha, "Comparison of Provincial Open Unemployment Rates in Indonesia Based on the K-Means Clustering Method," Journal of Economic Science, vol. 2, no. 2, pp. 109–116, 2022.

[3] H. D. Dermawan, R. Kurniawan, and Y. A. Wijaya, "Sales Data Analysis Using K-Means Clustering Algorithm," CESS (Journal of Computer Engineering System and Science), vol. 9, no. 1, pp. 175–191, 2024.

[4] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive Review of K-Means Clustering Algorithms," International Journal of Advances in Scientific Research and Engineering (IJASRE), vol. 7, no. 8, pp. 22–23, 2021, doi: 10.31695/IJASRE.2021.34050.

[5] G. H. Fatimah and G. E. Setyowisnu, "District/City Grouping in West Java Province Based on Open Unemployment Rate in 2023 Using K-Means Clustering," Journal of Statistics, vol. 11, no. 2, pp. 279–291, 2025.

[6] J. E. Simarmata, D. Chrisinta, and M. Purnomo, "Implementation of K-Means Clustering to Human Development Indicators in East Nusa Tenggara," Journal of Research in Mathematics Trends and Technology, vol. 6, no. 2, pp. 46–56, 2024, doi: 10.32734/jormtt.v6i2.17066.

[7] A. I. Ramadhan, P. D. Atika, and K. F. Ramdhania, "K-Means Clustering Analysis for Mapping Open Unemployment Rates in Indonesian Provinces in 2013-2023," Journal of Computer Science, vol. 5, no. 2, pp. 109–122, 2025.

[8] S. Sinaga and E. S. Negara, "K-Means Clustering Algorithm for Mapping the Unemployment Rate in Indonesia," 2020 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 2020, pp. 1-5, doi: 10.1109/ICoDSA50139.2020.9212854.

[9] K. Angkatan et al., "Workforce classification in West Java 2018 with Random Forest," Journal of Mathematics, Statistics and Computing (JMSK), vol. 17, no. 2, pp. 240–251, 2021, doi: 10.20956/jmsk.v17i2.11680.

[10] M. R. G. Ramos, A. M. D. C. Ramos, and M. L. M. Santos, "Predicting Unemployment Rates using

Machine Learning Algorithms," 2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Manila, Philippines, 2021, pp. 1-5.

[11] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[12] D. Junaedi, T. I. Wahyuni, N. Masruroh, R. Rachmawati, L. Q. K. Aini, S. A. Rustina, and S. Nurhalisa, "Analysis of Economic Development Stages on Human Development, Population Dynamics, and Unemployment in Probolinggo Regency," Journal of Artificial Intelligence and Digital Business (RIGGS), vol. 4, no. 4, pp. 13298-13305, 2026.