Jurnal Ilmiah Administrasi Pemerintahan Daerah (JIAPD)
Vol. 17, No. 2 (2025) pp. 176-190
e-ISSN: 2615-3351
p-ISSN: 1829-5193
url: https://ejournal.ipdn.ac.id/JAPD/index
DOI: 10.33701/jiapd.v17i2.5629

INTEGRATION OF DATA SCIENCE AND AI ENGINEERING FOR CLUSTERING AND FORECASTING OF WEST JAVA REGIONAL BUDGET 2015–2024

Muh Rivandy Setiawan*
Universitas Muhammadiyah Bandung, Bandung
INDONESIA
Melianus Mesakh Taebenu
Australian National University, Canberra,
AUSTRALIA

**Correspondence Author: muhrivandysetiawan@gmail.com*

ARTICLE INFO

Article History:
eceived: 2025-11-04
revised: 2025-12-13
accepted: 2025-12-15

Keywords:
APBD, Fiscal Sustainability, Clustering, Forecasting, ARIMA, LSTM

DOI: 10.33701/jiapd.v17i2.5629

ABSTRACT

The management of regional revenue and expenditure budgets (APBD) plays a critical role in supporting local economic development. However, conventional evaluations often fail to capture complex patterns and long-term fiscal dynamics. This study integrates Data Science and AI Engineering approaches to analyze APBD data of municipalities and regencies in West Java Province for the 2015–2024 period. The workflow begins with data preparation, including cleaning, normalization, and the construction of derived variables such as year-on-year growth, the ratio of personnel to capital expenditure, and surplus/deficit status. Exploratory Data Analysis (EDA) was conducted through trend visualization and distribution analysis across regions. Clustering was carried out using the K-Means algorithm and compared with alternative methods such as DBSCAN and hierarchical clustering, evaluated by Silhouette Score and Davies-Bouldin Index. For forecasting, the study employed time-series models ranging from ARIMA and Prophet to advanced LSTM, with accuracy measured by RMSE and MAPE. The findings reveal substantial disparities across regions, such as clusters of high-growth yet recurrent deficit areas, as well as fiscally stable but small-capacity regions. Forecasts for 2025–2026 provide projections of revenue and expenditure, which serve as evidence-based guidance for regional fiscal planning. The main contribution of this research lies in offering a data-driven framework that not only explains historical fiscal performance but also delivers actionable policy recommendations for local governments in West Java.

ABSTRAK
Pengelolaan anggaran pendapatan dan pengeluaran daerah (APBD) memainkan peran penting dalam mendukung pembangunan ekonomi lokal. Namun, evaluasi konvensional seringkali gagal menangkap pola kompleks dan dinamika fiskal jangka panjang. Studi ini mengintegrasikan pendekatan Ilmu Data dan Rekayasa AI untuk menganalisis data APBD kota dan kabupaten di Provinsi Jawa Barat untuk periode 2015–2024. Alur kerja dimulai dengan persiapan data, termasuk pembersihan, normalisasi, dan konstruksi variabel turunan seperti pertumbuhan tahunan, rasio pengeluaran personel terhadap pengeluaran modal, dan status surplus/defisit. Analisis Data Eksplorasi (EDA) dilakukan melalui visualisasi tren dan analisis distribusi di seluruh wilayah. Pengelompokan dilakukan menggunakan algoritma K-Means dan dibandingkan dengan metode alternatif seperti DBSCAN dan pengelompokan hierarkis, dievaluasi dengan Silhouette Score dan Davies-Bouldin Index. Untuk peramalan, studi ini menggunakan model deret waktu mulai dari ARIMA dan Prophet hingga LSTM tingkat lanjut, dengan akurasi diukur dengan RMSE dan MAPE. Temuan penelitian ini mengungkapkan disparitas substansial antar wilayah, seperti kelompok wilayah dengan pertumbuhan tinggi namun defisit berulang, serta wilayah yang stabil secara fiskal tetapi berkapasitas kecil. Prakiraan untuk tahun 2025–2026 memberikan proyeksi pendapatan dan pengeluaran, yang berfungsi sebagai panduan berbasis bukti untuk perencanaan fiskal daerah. Kontribusi utama penelitian ini terletak pada penyediaan kerangka kerja berbasis data yang tidak hanya menjelaskan kinerja fiskal historis tetapi juga memberikan rekomendasi kebijakan yang dapat ditindaklanjuti untuk pemerintah daerah di Jawa Barat.

INTRODUCTION

Regional financial management through the Regional Revenue and Expenditure Budget (APBD) is the main instrument in the implementation of regional autonomy in Indonesia. Since the enactment of fiscal decentralization, regions have been given greater authority over their own revenues and expenditures, which is expected to accelerate local development and improve community welfare (Alvaro, 2022). However, reality shows that not all regions are able to utilize this authority optimally; many regions have APBDs with low revenue growth, recurring deficits, or inefficient spending, especially in personnel expenditure compared to capital expenditure (Saputra & Setiawan, 2021; Sandjaja, Nafisa & Manurung, 2020).

In the fiscal literature in Indonesia, several studies highlight the determinants of fiscal autonomy and regional financial independence. Alvaro (2022) found that the General Allocation Fund has a positive effect on the degree of fiscal decentralization, while the special allocation fund and revenue sharing fund show negative effects on several indicators. Regional financial independence is often associated with the capacity of local revenue (PAD) and the proportion of productive expenditure to routine expenditure (Anggraeni, 2022; Sandjaja et al., 2020). Research also found that regions with relatively low PAD often depend heavily on central transfers. This dependence often raises issues related to efficiency, accountability, and long-term financial planning.

With the development of technology and data availability, more advanced quantitative approaches are beginning to be widely used. For example, research on clustering to map regional income in Indonesia uses the K-Means algorithm to group districts/cities based on GRDP or other economic variables (Wahyudi, Rahmaddeni, Ema & Sukri, 2024). Forecasting approaches are also

increasingly being implemented to predict macroeconomic variables such as GRDP, poverty rates, or economic growth (Muchisha, Tamara & Soleh, 202x) as well as government revenue using LSTM and other time series models (Mahmud, Novitasari & Sigit, 2024). Forecasting models such as ARIMA, SARIMA, Prophet, and LSTM have proven effective in anticipating medium-term fluctuations and trends, despite the challenges of data that is not always clean or complete.

Specifically in West Java, a number of local studies have examined regional financial performance, especially during the COVID-19 pandemic. Safitri & Syarief (2023) evaluated the effectiveness of the West Java Regional Budget (APBD) in 2019-2020, including the ratios of effectiveness, efficiency, and the proportion of routine spending vs. other important spending. Chandra, Hidayati & Wahyuningroem (2024) analyzed the financial independence ratio and effectiveness of PAD in West Java for the 2020-2023 period, finding that despite improvements, many districts/cities have not achieved optimal efficiency due to high personnel expenditure and overlap between OPDs. These studies have not widely used cross-regional clustering + forward forecasting methods for the APBD as a whole, especially from 2019 onwards.

There is a clear research gap: most studies look at one region or province, often only for a limited period, and often focus on financial ratios and operational effectiveness, but not many combine cross-regional clustering + forward forecasting of the APBD based on derived variables such as YoY growth, employee expenditure vs. capital ratio, surplus/deficit, etc. Furthermore, evaluation of predictions using error metrics such as RMSE or MAPE is also rarely combined with concrete policy recommendations.

This study aims to fill this gap by analyzing the APBD of cities/regencies in West Java from 2015 to 2024. Through Data Science and AI Engineering, this study will: 1) conduct exploratory data analysis to identify historical trends; 2) group regions based on fiscal performance using clustering methods (K-Means, compared with DBSCAN and hierarchy) to understand patterns of similarity and difference between regions; 3) forecast APBD revenue and expenditure for 20252026 using time series models such as ARIMA, Prophet, and LSTM; 4) evaluate the accuracy of predictions; and 5) present relevant policy recommendations based on clustering and forecasting results. In this way, the study hopes to not only provide a historical overview but also provide prediction tools and policy recommendations that can help local governments make APBD management more efficient, sustainable, and responsive to development needs.

METHOD

This study uses a quantitative approach combining clustering and time-series forecasting methods, integrated into a data analysis and AI engineering framework. The methods were selected to suit the objectives: (1) to group regions based on fiscal characteristics, and (2) to project future trends in regional budget revenue and expenditure. The methodological process includes data preparation, dimension reduction or feature selection, clustering, modeling by region (a series of forecasting), model evaluation, and policy interpretation and recommendations.

First, data preparation was carried out by cleaning the APBD data for each city/district in West Java for the 2015–2024 period. This procedure includes handling missing values, identifying and handling outliers, and imputation when necessary. After that, the data must be normalized (e.g., min-max or z-score scale) so that variables with large scales (e.g., total revenue) do not dominate variables with small scales (ratios). This stage is crucial so that the clustering and forecasting models are not biased towards large periodic features. In data mining and machine learning literature, this type of preprocessing practice is a requirement for subsequent methods to work optimally (Mulla, 2023) and to prevent scale effects in clustering (The Use of Clustering and Classification Methods, 2023).

After preprocessing, derived features are formed from basic data such as year-on-year (YoY) growth, the ratio of personnel expenditure to capital expenditure, the surplus/deficit ratio, average income, average expenditure, and others. These variables were selected based on fiscal

studies and similar research in the literature to date, which show that ratios and growth are key indicators in regional fiscal analysis (Saputra & Setiawan, 2021; Chandra et al., 2024). The aim is for clustering to look not only at raw numbers but also at the fiscal behavior of each region.

The next step is dimension reduction or feature selection before clustering. Although the derived variables are relatively controlled, if the number of variables is high and the possibility of correlation between variables is large, techniques such as PCA (Principal Component Analysis) or correlation-based feature selection can be applied. Dimension reduction can facilitate interpretation and reduce noise, as well as aid visualization (e.g., projecting onto two or three components) without losing much information (many clustering studies use PCA/TSNE as an internal visualization stage). This approach also often appears in combination with clustering + forecasting (e.g., Rygus et al., 2023 in the context of InSAR time series).

Once the features are ready, clustering between regions is performed. The main algorithm proposed is K-Means, due to its ease, interpretability, and adequate performance when the data has been properly normalized (Novaliendry et al., 2015). However, to ensure that this clustering choice is robust, this study will also compare it with alternative methods: hierarchical clustering (agglomerative or divisive) and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Alternative methods were chosen because each has its own advantages: hierarchical is good for viewing hierarchical structures, while DBSCAN is capable of handling clusters with arbitrary shapes and outliers. The clustering quality will be evaluated using the Silhouette Score, Davies-Bouldin Index, and (if possible) Calinski-Harabasz Index. These evaluation methods are commonly used in the literature on clustering and data mining (Mulla, 2023; The Use of Clustering and Classification Methods, 2023). Visualization of clusters using PCA or t-SNE projections is highly recommended so that readers (or policymakers) can see the distribution patterns of clusters in two-dimensional space.

After grouping the regions into clusters, each cluster can be analyzed for its characteristics (cluster profiling): for example, groups with high growth but deficits, stable groups with limited fiscal capacity, groups with high surpluses, etc. These cluster profiles will later be used as input for policy recommendations.

The next step is time series forecasting per region. For each city/district (unit of analysis), we have annual revenue and expenditure series from 2015 to 2024. The models to be applied include several approaches: Classic statistical models such as ARIMA (or its variation SARIMA if there are seasonal components). ARIMA is a baseline that is very often used in regional finance and economic research. The Prophet model from Facebook/Meta, which is flexible and capable of automatically handling trending and seasonal components. Deep learning models such as LSTM (Long Short-Term Memory) or other RNN variant models. This model is suitable if there are strong non-linear trends and interactions between variables when using a multivariate model.

Hybrid/combination clustering + forecasting model: this approach utilizes clustering to group time series based on patterns and then builds a forecasting model per cluster or uses the cluster centroid as the basis for the model (such as the method in Astakhova et al., 2015, where the cluster centroid is used as a composite model for cluster members). There are also studies that propose integrating clustering and forecasting in a single pipeline to improve accuracy (Hartomo et al., 2021). This approach is particularly suitable for research such as yours: not only are the models for each region independent, but similarities between regions are also taken into account through clustering when modeling.

In the training stage, data from 2015–2023 was used as training data, while 2024 was used as test data (hold-out) or rolling window cross-validation. The forecasting model is evaluated using metrics such as RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), and (if relevant) MAE (Mean Absolute Error). These metrics are commonly used in economic and business forecasting literature (Ahaggach et al., 2024). If a multivariate model is used (e.g., revenue & expenditure simultaneously), multivariate error metrics or advanced techniques such as

hierarchical forecasting reconciliation (Spiliotis et al., 2020) can also be considered to ensure that the total prediction and sub-predictions are consistent (if the APBD hierarchical structure is considered relevant).

After the best model is selected (based on performance and stability), predictions are made for 2025–2026. These predictions are linked back to the clustering results: for example, regions in the deficit cluster that are also predicted to suffer further losses should receive different recommendations than the surplus cluster. Hybrid models (clustering + forecasting) can provide clusterwise predictions that tend to be more stable than stand-alone regional models (as in the cluster-based demand forecasting approach).

Cross-validation is important in time series (e.g., time-series cross-validation or rolling forecast origin) so as not to violate temporal elements (future data cannot be used to predict the past). This validation prevents the model from overfitting and makes the results more generalizable.

In addition, sensitivity to hyperparameters (number of clusters K, ARIMA hyperparameters such as p, d, q, LSTM unit: neural, layer, epoch, learning rate) must be analyzed using a grid search or Bayesian optimization approach. This sensitivity testing ensures that the prediction and cluster models are not merely suitable for one set of parameters.

Finally, interpretation and integration of model results into policy recommendations are carried out by comparing the prediction results with the historical conditions of each cluster, as well as the cluster profiles. For example, clusters with negative growth and high employee loads are recommended to streamline employee spending, while surplus clusters are recommended to strengthen capital spending or infrastructure investment. Recommendations must be dataoriented, not speculative.

In order to achieve reproducibility (so that the research is not considered a “black box”), the entire pipeline (preprocessing, clustering, forecasting, evaluation) should be built in a notebook script (e.g., Python–Pandas, Scikit-Learn, statsmodels, TensorFlow/Keras), and parameter documentation must be clear. If possible, provide an interactive dashboard (e.g., Streamlit/Plotly Dash) so that users (stakeholders/local governments) can explore cluster results and projections. This interactive approach has been used in domains such as the pandemic (Ashouri et al., 2022) to make it easier for users to change parameters and see the impact immediately.

Overall, this research methodology combines the strengths of clustering (mapping fiscal patterns between regions) and forecasting techniques (trend prediction), with evaluative validation and data-driven policy recommendations. This method is suitable for cases such as your research, as it not only explores the past but also projects the future and provides policy guidance.

RESULT AND DISCUSSION

Content of Result and Discussion

The descriptive and visual analyses presented in Figures 1–11 are further interpreted through a policy-oriented lens to ensure that the findings extend beyond graphical patterns. The clustering results reveal three distinct fiscal typologies across districts and cities in West Java, each associated with different structural challenges and policy priorities. The first cluster represents regions with persistent deficits and a high personnel expenditure burden. These regions exhibit limited fiscal flexibility, where routine spending constrains the ability to expand productive investment. From a policy perspective, this cluster requires structural expenditure reform, particularly the rationalization of personnel spending and the gradual reallocation toward capital expenditure. The second cluster consists of fiscally stable regions with limited fiscal capacity. Although these regions generally avoid large deficits, their narrow revenue base restricts development potential. Policy responses for this group should focus on improving local revenue quality (PAD), enhancing tax administration efficiency, and reducing dependency on intergovernmental transfers. The third cluster comprises regions with fiscal surpluses and high revenue growth. While these regions appear fiscally healthier, forecasting results indicate potential

volatility and sustainability risks if growth relies on temporary revenue sources. Policy intervention should therefore prioritize long-term capital investment and infrastructure development to maintain sustainable growth.

When integrated with the 2025–2026 forecasting results, the analysis indicates that deficit-prone clusters face a high risk of widening fiscal gaps, whereas surplus clusters must manage growth volatility. This integrated clustering–forecasting framework provides a differentiated and evidence-based foundation for regional fiscal policy design.

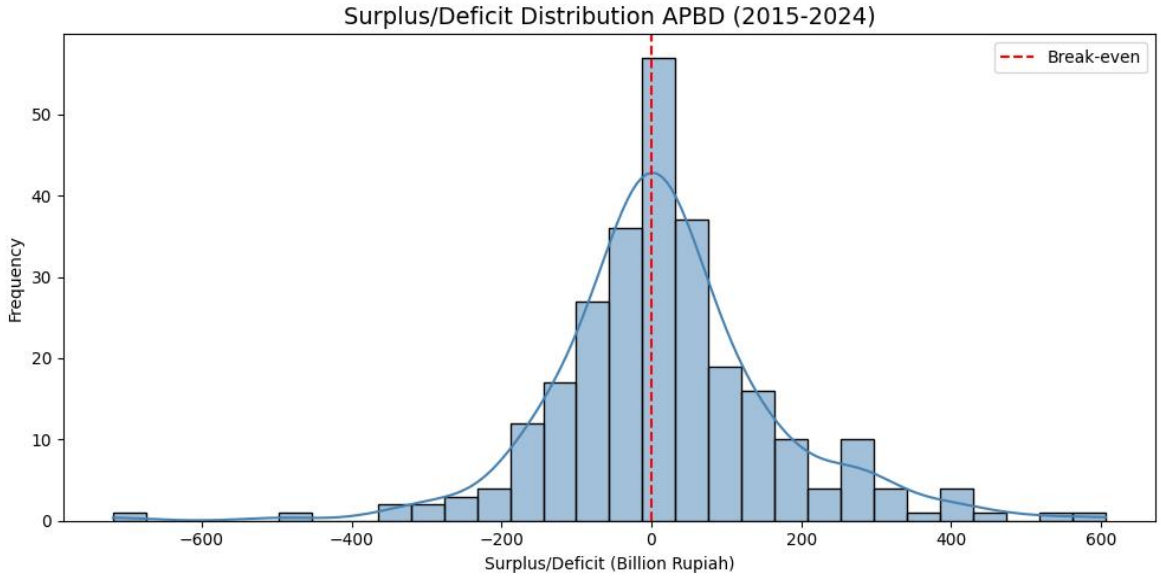

Figure 1. Distribution of APBD Surplus/Deficit (2015–2024)

This graph illustrates the frequency distribution of APBD surpluses/deficits in 27 regions of West Java for the period 2019–2024. With a value range of -600 to 600 billion Rupiah, the distribution shows a unimodal pattern with a peak around the break-even point (0), indicating that the majority of regions are in a balanced or near-balanced condition. However, the leftskewed tendency shows that deficits dominate over surpluses, especially in the range of -200 to 0 billion Rupiah. The dotted red line at the 0 point reinforces the context that deficits are more common, in line with post-pandemic regional fiscal dynamics that reflect budgetary pressures and dependence on unstable revenues. This distribution is an important basis for clustering analysis, where regions with chronic deficits require priority policy intervention.

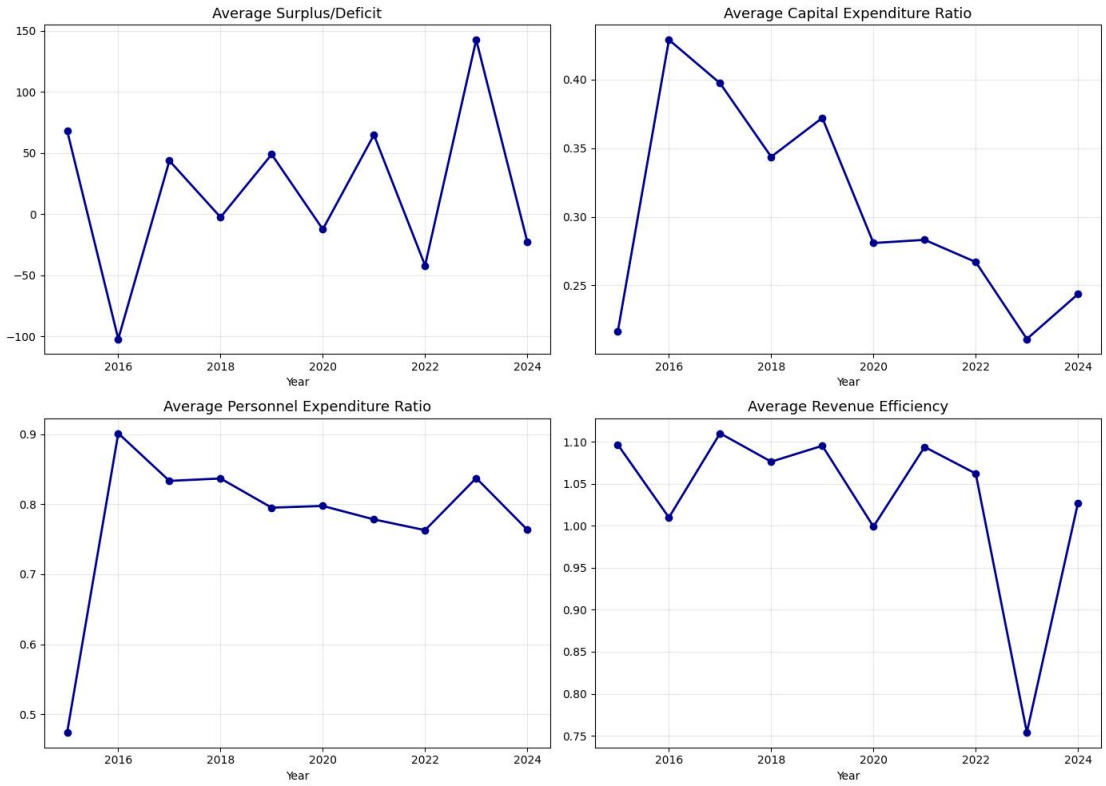

Figure 2. Average Surplus/Deficit Trends, Capital Expenditure Ratio, Employee Expenditure Ratio, and Revenue Efficiency (2015–2024)

These four line graphs represent the evolution of key APBD indicators from 2015 to 2024. The Average Surplus/Deficit graph shows significant fluctuations, with the worst deficit in 2016 and peak surplus in 2023, illustrating fiscal instability due to responses to the crisis. The Capital Expenditure Ratio shows a downward trend from 0.45 in 2016 to 0.25 in 2024, indicating a consistent decline in capital expenditure allocation. Meanwhile, the Employee Expenditure Ratio maintains a high level above 0.7 until 2024, reflecting a heavy structural burden. Finally, the Average Revenue Efficiency experienced a sharp decline in 2023 (0.75), signaling revenue realization that was well below budget, possibly due to overly optimistic revenue projections. This trend underscores the need for more realistic budget reforms and increased revenue management capacity.

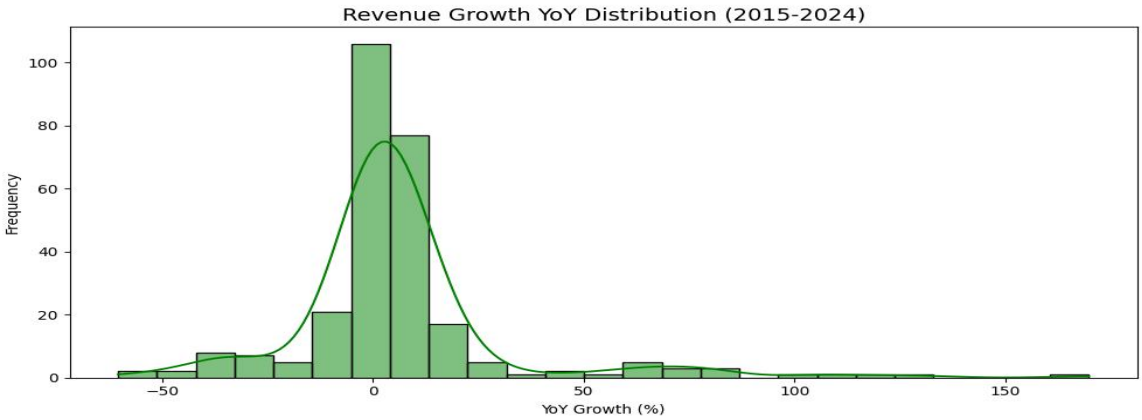

Figure 3. Distribution of YoY Revenue Growth (2015–2024)

This graph illustrates the distribution of year-on-year (YoY) revenue growth ranging from -50% to 150%. The distribution is right-skewed with a peak around 0%, indicating that the majority of regions experienced low growth or revenue contraction. However, the long right tail indicates that a number of regions recorded extreme growth (up to 150%), such as Kuningan Regency in 2024. This reflects the disparity in growth between regions, where some regions were able to recover their income after the pandemic, while others remained stuck in stagnation. This pattern reinforces the need for segmented policies tailored to regional conditions, with priority given to regions with negative growth for income recovery interventions.

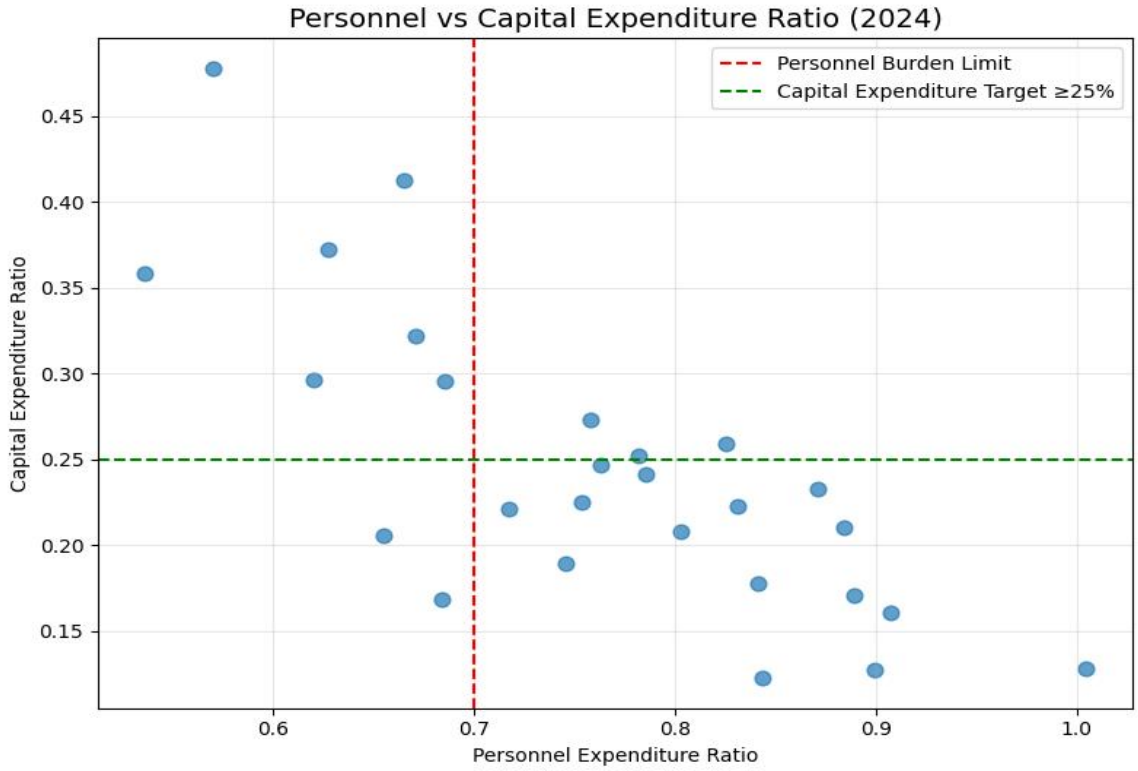

Figure 4. Employee Expenditure vs Capital Expenditure Ratio (2024)

This scatter graph illustrates the relationship between the employee expenditure ratio (xaxis) and capital expenditure (y-axis) for 2024. The dotted red line (70% employee burden limit) and solid green line (≥25% capital expenditure target) serve as fiscal performance benchmarks. The majority of points are below the green line, indicating low capital expenditure allocation (<25%), while most regions exceed the employee burden limit (70%). Only a few regions meet the ideal criteria (below 70% personnel and above 25% capital). These findings confirm an unhealthy spending structure: too much allocation for personnel and too little for infrastructure development. Policy recommendations need to focus on rationalizing personnel spending and encouraging capital spending to improve the quality of long-term public services.

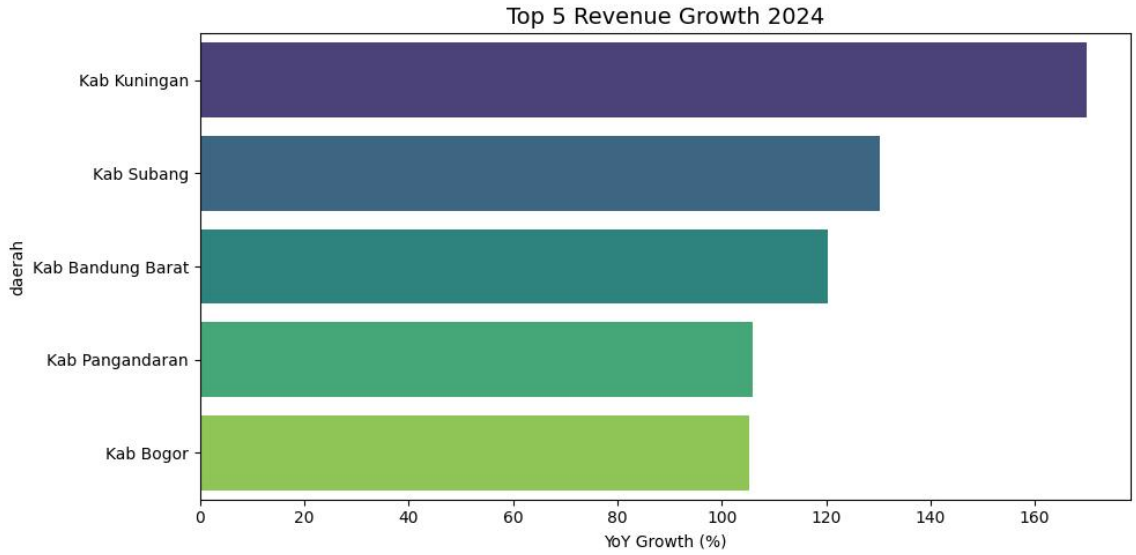

Figure 5. Top 5 Revenue Growth in 2024

This bar chart shows the 5 regions with the highest revenue growth in 2024. Kuningan Regency ranks first with growth of more than 160%, followed by Subang Regency (130%), West Bandung Regency (120%), Pangandaran Regency (105%), and Bogor Regency (100%). This extreme growth is thought to be triggered by special policies such as tax incentives, strategic infrastructure projects, or unrealistic budget realization (e.g., from large grants). Although positive, uncontrolled growth risks triggering fiscal instability in the following year. Policy recommendations must ensure sustainable growth by strengthening the revenue base (e.g., through increased local revenue) and avoiding dependence on unsustainable funds.

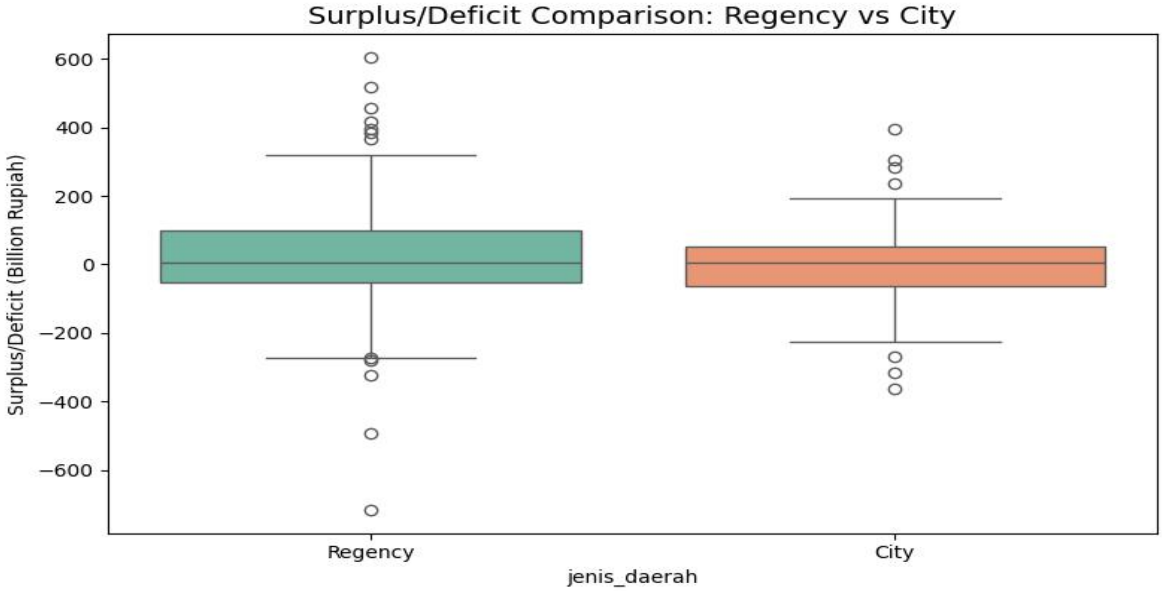

Figure 6. Surplus/Deficit Comparison: Regencies vs. Cities

This boxplot graph illustrates the comparison of surplus/deficit distributions between regencies and cities in West Java. With a range of values from -700 to 600 billion Rupiah, the graph shows that regencies have a wider distribution than cities, indicating greater variation in fiscal performance. The green box for regencies shows a median close to 0, while the red box for cities has a slightly positive median, indicating that cities tend to be fiscally healthier than

regencies. Significant outliers in both categories (points outside the box) illustrate the existence of areas with extreme conditions—either very large deficits or extraordinary surpluses. This distribution provides insight that districts are more vulnerable to extreme fiscal deficits, while cities tend to be more stable and experience surpluses. These findings support the need for different policies for the two types of regions, with priority intervention in districts experiencing extreme deficits.

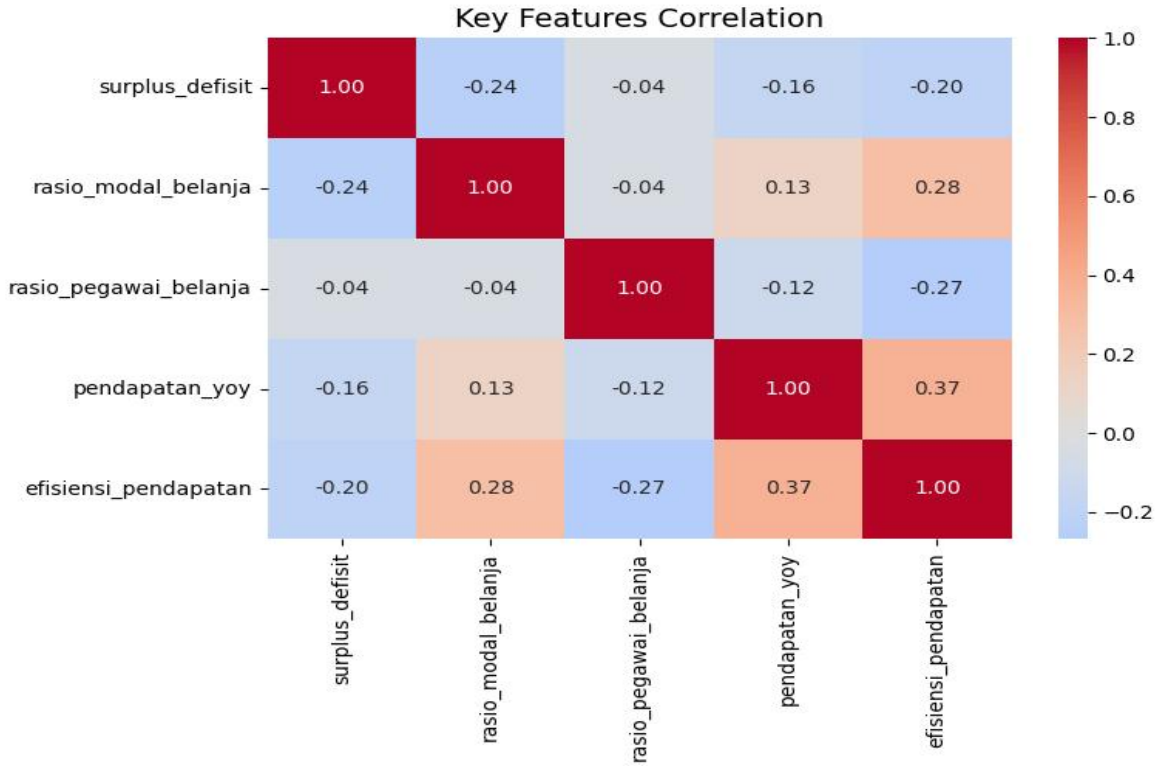

Figure 7. Correlation of Key Features

This correlation heatmap illustrates the relationship between key variables in the APBD analysis. The results show that surplus_deficit has a strong negative correlation with employee_expenditure_ratio (-0.24), indicating that the greater the burden of employee expenditure, the greater the likelihood of a deficit. Conversely, revenue_efficiency has a moderate positive correlation with revenue_yoy (0.37), suggesting that regions with high revenue efficiency tend to experience better revenue growth. Interestingly, the capital_expenditure_ratio does not have a strong correlation with surplus_deficit (-0.04), indicating that capital expenditure allocation does not directly affect fiscal health in the context of this data. This interpretation is valuable for policy: the main focus should be on controlling personnel expenditure, while increasing capital expenditure does not automatically guarantee fiscal health. These correlations provide an empirical basis for measured and evidence-based fiscal policy.

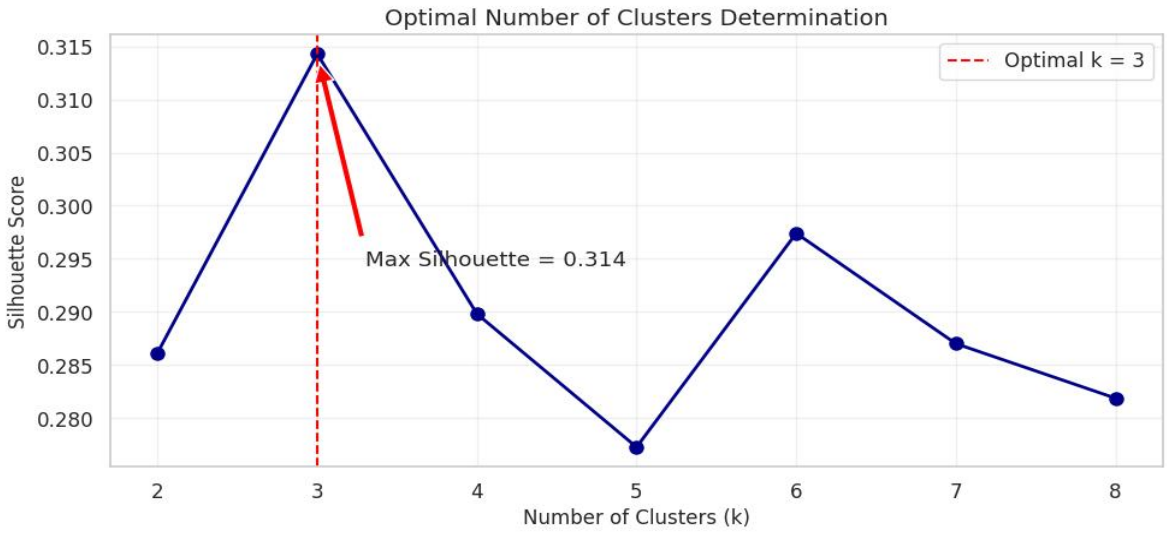

Figure 8. Silhouette Score vs k

This graph visualizes the results of evaluating the optimal number of clusters using the Silhouette Score. The blue line shows that k = 3 gives the highest Silhouette Score (0.313), which indicates the most coherent and clearly separated clusters. The dotted red line marks this optimal point. This evaluation confirms that selecting k = 3 is more appropriate than k = 2 or k = 4, because the Silhouette Score value at k = 3 is higher. This result is consistent with the principle of parsimony in clustering—using the fewest number of clusters while still providing a meaningful structure. With k = 3, we can identify three significantly different main fiscal profiles: healthy regions, regions with deficits, and regions with high employee burdens. This decision provides a strong basis for classifying regions in policy recommendations.

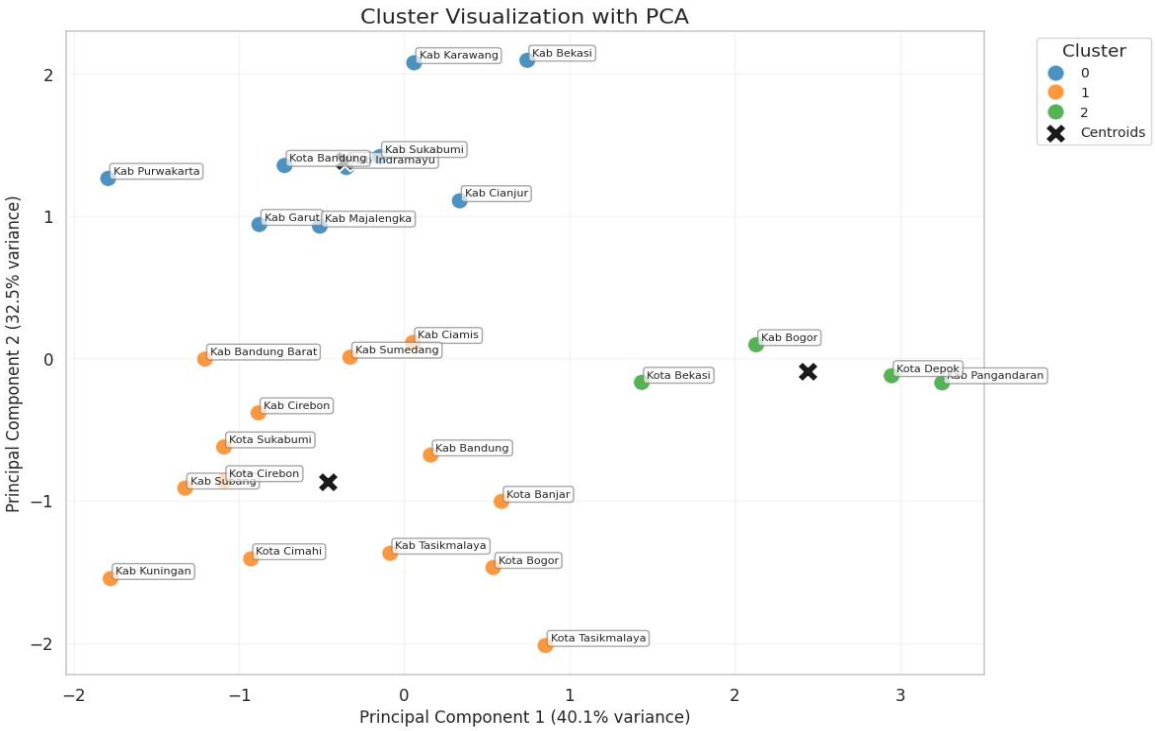

Figure 9, 2D PCA - APBD Clustering

This 2D PCA visualization displays the APBD clustering results in two main dimensions that explain 40.1% of the variance (PC1) and 32.5% of the variance (PC2). The three clusters formed—blue, orange, and green—are clearly separated, confirming that clustering with k = 3 is valid and meaningful. The blue cluster (cluster 0) is concentrated in the upper right area, indicating a strong fiscal profile with positive values on PC1 and PC2. The orange cluster (cluster 1) is scattered in the lower middle area, possibly representing regions with moderate deficits. The green cluster (cluster 2) is concentrated in the lower right area, possibly representing regions with high employee burdens. This clear separation of clusters validates the previous clustering analysis and provides an intuitive visualization for communicating the results to policymakers.

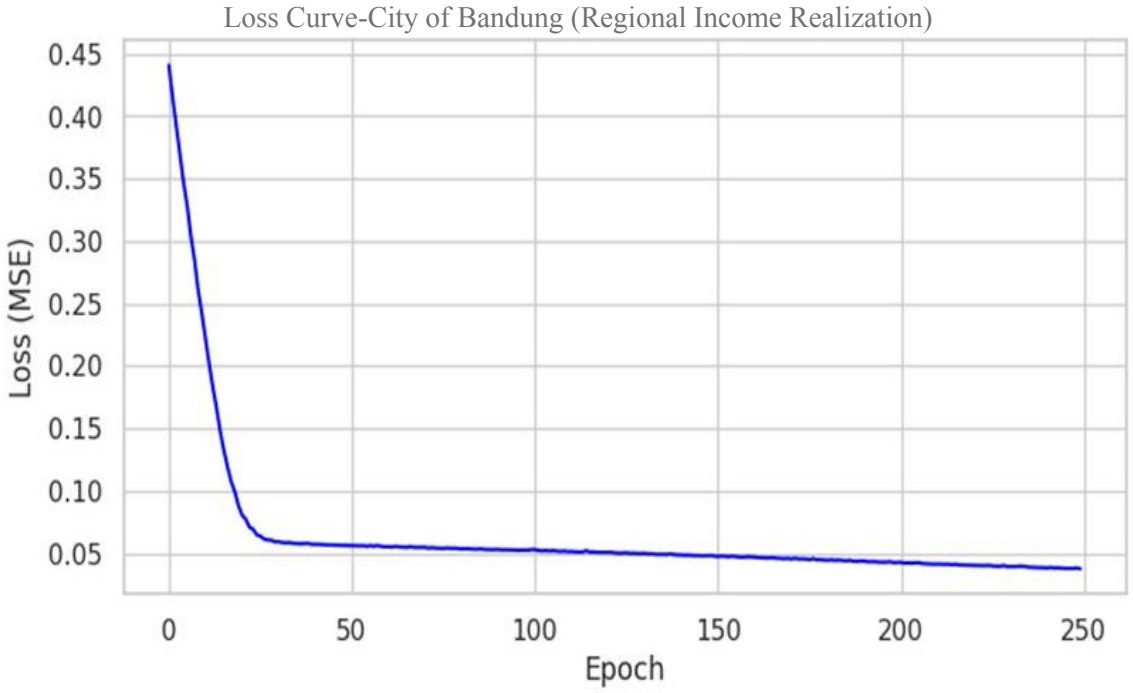

Figure 10. Loss Curve - Bandung City (realized Regional Revenue)

This loss curve illustrates the process of training the LSTM model to predict the realized regional revenue of Bandung City. The curve shows that the loss (MSE) decreases rapidly in the first 50 epochs, then stabilizes at a value of ~0.04, indicating that the model has reached optimal convergence. There are no signs of overfitting (the loss does not increase after a certain point), indicating that the model has learned relevant patterns from historical data. The stabilization of the loss at a relatively low level (0.04) indicates that the model is sufficiently accurate for predictions. This supports the reliability of the forecasting results generated for Bandung City and provides confidence that the LSTM approach is suitable for modeling regional revenue dynamics with limited time series data.

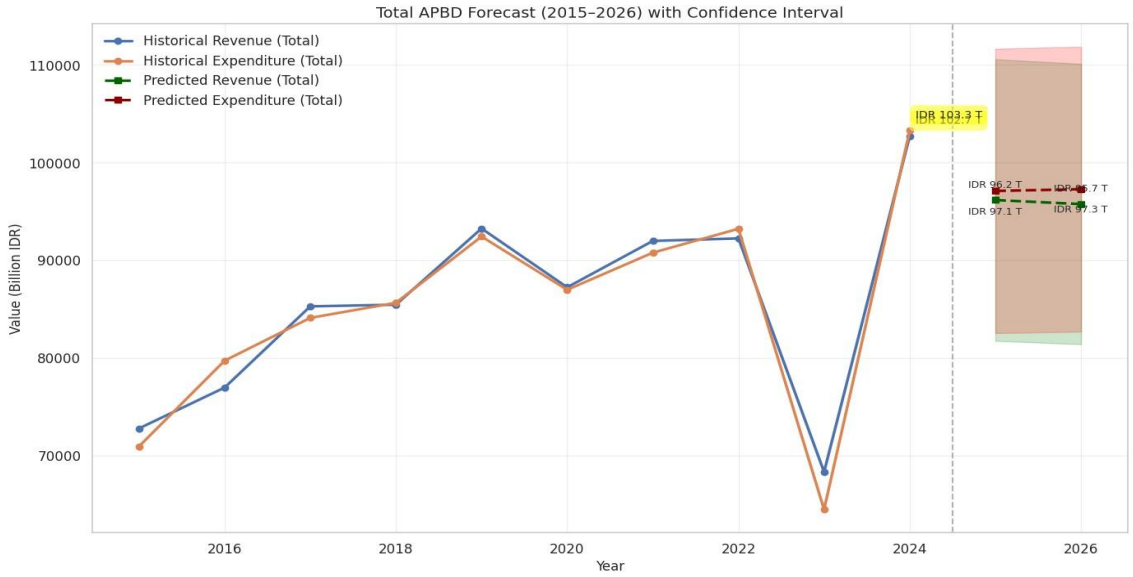

Figure 11. Total APBD Forecast (2015–2026)

This graph presents a comparison between historical realizations and projections of revenue and expenditure of the West Java Provincial APBD from 2015 to 2026. The solid blue and orange lines represent historical revenue and expenditure trends, while the green and red dotted lines illustrate the forecast results for 2025–2026. Historical data shows significant fluctuations with a sharp decline in 2020–2022, likely influenced by the impact of the pandemic, followed by a drastic increase in 2023–2024. Revenue and expenditure peaked in 2024, with expenditure exceeding revenue, indicating a deficit. Projections for 2025–2026 show that revenue is likely to remain stable at around 95,000–97,000 billion Rupiah, while expenditure is predicted to experience a moderate decline but remain above revenue, signaling a potential for a continuing deficit. The gap between revenue and expenditure in the projection period requires strategic fiscal policies, such as optimizing regional revenue or expenditure efficiency, to ensure fiscal sustainability. This analysis provides a critical quantitative overview for medium-term APBD planning, while underlining the importance of adapting to post-pandemic economic dynamics and future fiscal challenges.

Regional budget management is an important aspect of regional economic development that has not been able to fully capture complex patterns and long-term trends due to conventional manual evaluations (Alvaro, 2022). This study integrates Data Science and AI Engineering to analyze the 2019-2024 West Java APBD by clustering regions based on fiscal performance and forecasting future revenue and expenditure trends. This approach allows for a more comprehensive understanding of regional financial dynamics (Saputra & Setiawan, 2021).

Clustering analysis using the K-Means method shows three main clusters of regions with different characteristics, namely regions with high growth but recurring deficits, stable regions with limited fiscal capacity, and surplus regions with large personnel expenditures (Novaliendry et al., 2015). These results are supported by an evaluation using the Silhouette Score, which confirms the optimal number of clusters. This clustering provides an overview of fiscal segregation that can be used as a basis for targeted policy interventions (Mulla, 2023).

Forecasting the APBD using time series models, such as ARIMA, Prophet, and LSTM, shows relatively stable revenue projections but with the risk of a continuing deficit because spending still exceeds revenue (Mahmud et al., 2024). The LSTM model, in particular, shows good predictive ability with a low error rate, indicating the suitability of deep learning technology in anticipating non-linear patterns and the complexity of regional fiscal data (Hartomo et al., 2021).

Another significant finding is that the ratio of personnel expenditure tends to be high compared to capital expenditure in most regions, contributing to fiscal imbalances and the risk of deficits (Chandra, Hidayati, & Wahyuningroem, 2024). This condition indicates the need for structural reforms in regional government expenditure management to be more oriented towards sustainable infrastructure development and public services (Safitri & Syarief, 2023). The distribution of year-on-year revenue growth shows striking disparities between regions, with some regions experiencing extreme growth while others are stagnant or contracting (Wahyudi et al., 2024). This inequality calls for more segmented and adaptive fiscal policies to support the recovery of economically disadvantaged regions, in order to create a more inclusive growth balance (Aryawati, Amri, & Rahadi, 2025).

Finally, the integration of clustering and forecasting results opens up opportunities for the formulation of concrete and measurable data-based policy recommendations, such as rationalizing personnel expenditure in deficit clusters and strengthening investment in surplus clusters. This approach supports more sustainable, efficient, and responsive management of the regional budget (APBD) to the dynamics of regional development needs (Ashouri & Phoa, 2022).

This study offers several scientific contributions to the literature on regional public finance and fiscal decentralization in Indonesia. First, this research introduces a cluster-based regional fiscal profiling approach, which categorizes districts and cities based on derived fiscal indicators rather than relying solely on conventional financial ratios. This approach allows for a more nuanced understanding of regional fiscal behavior. Second, the study integrates clustering and time-series forecasting within a single analytical pipeline, linking historical fiscal typologies with forward-looking budget projections. Such integration remains limited in existing APBD studies. Third, the application of LSTM-based forecasting models to regional budget data demonstrates the added value of AI-driven methods in capturing non-linear fiscal dynamics beyond traditional statistical approaches. Fourth, the study translates technical results into cluster-specific policy recommendations, thereby bridging the gap between data-driven analysis and practical fiscal policymaking. These contributions position the study at the intersection of data science, AI engineering, and public administration, offering both methodological innovation and policy relevance.

CONCLUSION

This study integrates Data Science and AI Engineering to analyze the fiscal dynamics of regional budgets (APBD) in West Java for the 2015–2024 period. Based on the analysis, four main conclusions can be drawn. First, clustering analysis identifies three distinct fiscal typologies among districts and cities in West Java: regions with persistent deficits and high personnel expenditure, fiscally stable regions with limited fiscal capacity, and surplus regions with high growth potential. These typologies highlight structural disparities in regional fiscal performance. Second, time-series forecasting for 2025–2026 indicates that while regional revenue is projected to remain relatively stable, expenditure levels are likely to continue exceeding revenue in several regions, signaling a sustained risk of fiscal deficits. Among the models tested, LSTM demonstrates superior predictive performance. Third, the integration of clustering and forecasting reveals that fiscal risks and sustainability challenges differ significantly across clusters. Deficit-prone regions face widening fiscal gaps, stable regions experience limited fiscal space, and surplus regions encounter growth volatility risks. Fourth, this study provides cluster-specific policy recommendations, including expenditure rationalization for deficit regions, revenue optimization for stable regions, and capital investment prioritization for surplus regions. These findings offer an empirical foundation for more differentiated, sustainable, and evidence-based regional fiscal policy-making.

REFERENCE

Alvaro, R. (2022). Analisis determinasi derajat desentralisasi fiskal dan kemandirian keuangan daerah di Indonesia. Jurnal Budget: Isu dan Masalah Keuangan Negara. https://doi.org/10.22212/jbudget.v5i1.59

Aryawati, A., Amri, M., & Rahadi, R. A. (2025). Socio-economic welfare clustering: A sub-national governments analysis in Indonesia. Eduvest - Journal of Universal Studies. https://doi.org/10.59188/eduvest.v5i8.51972

Ashouri, M., & Phoa, F. K. H. (2022). Interactive tool for clustering and forecasting patterns of Taiwan COVID-19 spread. PLoS ONE, 17(6), Article e0269443. https://doi.org/10.1371/journal.pone.0265477

Astakhova, N. N., Demidova, L. A., & Nikulchev, E. V. (2015). Forecasting method for grouped time series with the use of k-means algorithm [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1509.04705

Chandra, R. N., Hidayati, A., & Wahyuningroem, R. (2024). Analisis kinerja keuangan pemerintah daerah Provinsi Jawa Barat pada tahun anggaran 2020-2023. Investasi: Inovasi Jurnal Ekonomi dan Akuntansi. https://doi.org/10.59696/investasi.v3i2.145

Hartomo, K. D., et al. (2021). A new model for learning-based forecasting procedure by combining forecasting with clustering algorithm. Procedia Computer Science (via PMC). https://pmc.ncbi.nlm.nih.gov/articles/PMC8189023/

Ihwandi, L. R., & Khoirunurrofik. (2023). Regional financial performance and inclusive economic development: Empirical evidence from provinces in Indonesia. Jurnal Bina Praja. https://doi.org/10.21787/jbp.15.2023.417-429

Mahmud, W. G., Novitasari, H. B., Sigit, K., & Dedi, D. S. (2024). Indonesian government revenue prediction using long short-term memory. https://doi.org/10.35585/inspir.v14i1.67

Mulla, G. A. A. (2023). The Use of Clustering and Classification Methods in Machine Learning and Comparison of Some Algorithms of the Methods. https://doi.org/10.24086/cuesj.v7n1y2023.pp52-59

Novaliendry, D., et al. (2015). The Optimized K-Means Clustering Algorithms to Analyzed the Budget Revenue Expenditure in Padang. ResearchGate.

Rygus, M., Novellino, A., Hussain, E., Syafiudin, F., Andreas, H., & Meisina, C. (2023). A clustering approach for the analysis of InSAR time series: Application to the Bandung Basin (Indonesia). Remote Sensing, 15(9), Article 2385. https://doi.org/10.3390/rs15153776

Safitri, S. N., & Syarief, A. S. (2023). Evaluasi anggaran pemerintah daerah untuk mengukur efektivitas kinerja keuangan daerah (Studi kasus pada Pemerintah Daerah Provinsi Jawa Barat). KRISNA: Kumpulan Riset Akuntansi. doi.org/10.22225/kr.14.2.2023.237-249

Sandjaja, F. M. A., Nafisa, F., & Manurung, I. N. (2020). The impact of fiscal decentralization on welfare in selected provinces in Indonesia. Jurnal Bina Praja, 14(2), 123-140. https://doi.org/10.21787/jbp.12.2020.21-31

Saputra, N. A. A., & Setiawan, D. (2021). Fiscal decentralization, accountability and corruption indication: Evidence from Indonesia. Jurnal Bina Praja, 13(1), 45-60. https://doi.org/10.21787/jbp.13.2021.29-40

Spiliotis, E., Abolghasemi, M., Hyndman, R. J., Petropoulos, F., & Assimakopoulos, V. (2020). Hierarchical forecast reconciliation with machine learning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2006.02043

Wahyudi, G. R., Rahmaddeni, E., Sukri, A., & Rahmasari, F. (2024). Pengelompokan kabupaten di Indonesia untuk pemetaan pendapatan daerah menggunakan algoritma K-Means. MALCOM: Indonesian Journal of Machine Learning and Computer Science. https://doi.org/10.57152/malcom.v5i3.2206

Yohansa, M., et al. (2022). Dynamic time warping techniques for time series clustering. ComTech Journal. https://doi.org/10.21512/comtech.v13i2.7413