Applying Random Forest to Predict Unemployment in Medan City for Adaptive Labor Policy Formulation

Eva Mufida Padilla, Muhammad Syahputra Novelan

Abstract

Urban unemployment is a multidimensional policy problem because its dynamics are associated not only with economic growth but also with labor-force participation, human capital, poverty, inequality, inflation, entrepreneurship, and the absorptive capacity of the local economy. This study evaluates the preliminary feasibility of a Random Forest Regressor for forecasting unemployment in Medan City as a foundation for an adaptive labor-policy decision-support system. The experiment employs a structured simulated dataset informed by indicators contained in the Medan City Medium-Term Development Plan for 2025–2029. It comprises 20 observations, four subdistricts, the 2020–2024 period, 12 predictors, and one target variable. The dataset is neither official labor-force microdata nor an official subdistrict-level unemployment estimate; accordingly, the study is framed as an exploratory methodological proof of concept. Observations from 2020–2023 were used for training, while 2024 was retained as a temporal test set. The analytical pipeline incorporated data-quality auditing, one-hot encoding, a baseline Random Forest, year-grouped cross-validation, and randomized hyperparameter search over 30 candidate configurations. The optimized model used 100 trees, a maximum depth of 3, a minimum split size of 3, max_features of 0.7, and no bootstrap sampling. Optimization reduced MAE from 0.4587 to 0.3989, RMSE from 0.4821 to 0.4184, and MAPE from 6.9481% to 6.0382%. Nevertheless, the test R^2 remained negative at -1.4892, indicating inadequate generalization. Labor-force participation and economic growth accounted for a combined importance of 83.99%, but these results must be interpreted as predictive associations rather than causal effects. Policy deployment therefore requires a substantially larger, official, granular, and externally validated longitudinal dataset in github.com.

Keywords: *Random Forest; unemployment; labor policy; economic forecasting; panel data; Medan City*

Eva Mufida Padilla[1]
[1]Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: evamufidapadilla@gmail.com [1]

Muhammad Syahputra Novelan[2]
[2]Universitas Pembangunan Panca Budi, Indonesia
e-mail: putranovelan@dosen.pancabudi.ac.id [2]
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

Unemployment is a central indicator of an economy's ability to convert growth, investment, education, and business activity into productive employment. In an urban setting, unemployment is more complex than a simple shortage of vacancies. Large cities attract labor migration, produce substantial cohorts of new graduates, undergo sectoral restructuring, and experience mismatches between job-seeker competencies and industry requirements. Changes in unemployment may consequently emerge from nonlinear interactions among labor-force participation, economic growth, productivity, poverty, income inequality, inflation, human development, entrepreneurship, and territorial characteristics.

The Medan City Medium-Term Development Plan for 2025–2029 reports that the city's open unemployment rate was 10.74% in 2020, increased to 10.81% in 2021, and subsequently declined to 8.89% in 2022, 8.67% in 2023, and 8.13% in 2024. Despite this improvement, the 2024 rate remained above the North Sumatra provincial average of 5.60% and the national rate of 4.91%. The same planning document records a recovery in Medan's labor-force participation rate from 62.16% in 2021 to 66.44% in 2024, an entrepreneurship ratio reaching 6.84% in 2024, and continuing growth in regional gross domestic product per capita. These trends suggest that declining unemployment coincided with restored economic participation, expanding business activity, and broader economic improvement; they do not, however, establish a simple, linear, or causal relationship.

International research increasingly applies machine learning to labor-market forecasting. Gogas, Papadimitriou, and Sofianos show that Random Forest can outperform several competing methods when predicting the direction of euro-area unemployment[1]. Katris compares time-series and machine-learning models across country groups and emphasizes that performance depends on the forecast horizon and series characteristics[2]. Kreiner and Duca likewise find that machine learning based on economic indicators can provide competitive predictive power for unemployment [6]. Other studies use Internet search behavior as an early unemployment indicator [8], [9], while research conducted under unusual labor-market conditions indicates that nonlinear algorithms can adapt more rapidly to structural changes than simple autoregressive models [7].

Random Forest is well suited to tabular socioeconomic data because it can represent nonlinear relationships, interactions, and heterogeneity without imposing a predetermined functional form[3],[4]. It constructs multiple decision trees from variation in observations and features and then aggregates their predictions. Although relatively robust to noise and multicollinearity, its performance remains dependent on sample size, variable quality, validation design, hyperparameter configuration, and alignment between the predictive objective and the intended policy use[5], [15].

The central research problem is not merely whether a model can achieve a small numerical error, but whether the resulting evidence is sufficiently credible to support policy decisions. Within the framework of prediction policy problems, better forecasts create public value only when they improve the allocation, timing, or targeting of interventions [16]. Predictions derived from simulated data and a very small sample must therefore not be interpreted as an operational decision rule.

This study makes three methodological contributions. First, it develops a Random Forest pipeline that integrates demographic, labor-participation, human-development, macroeconomic, poverty, inequality, inflation, and entrepreneurship indicators within a single unemployment-prediction framework. Second, it uses a temporal holdout, ensuring that 2024 is excluded from the principal training stage. Third, it reports not only MAE, RMSE, MAPE, and feature importance, but also evidence of inadequate generalization through a negative R^2 and systematic residual patterns. The contribution is therefore not a claim of deployment readiness, but a transparent analytical framework that can be replicated and strengthened when adequate official data become available.

The objectives are to: (1) construct a Random Forest Regressor pipeline for unemployment prediction; (2) evaluate performance changes after hyperparameter optimization; (3) identify the features most frequently used by the fitted model; (4) analyze prediction errors on the 2024 temporal test set; and (5) specify the data and validation requirements needed to transform a proof of concept into an adaptive labor-policy instrument.

Literature Review

Random Forest was introduced as an ensemble of randomized trees whose regression outputs are averaged [1]. For B regression trees, the ensemble prediction is expressed as,

$$\hat{y}(x) = \left(\frac{1}{B}\right) \Sigma(b = 1 \; to \; B) T_{b(x)}$$

where T_b(x) denotes the prediction generated by tree b for observation x. Aggregation is intended to reduce the high variance commonly associated with a single decision tree. Biau and Scornet explain that Random Forest performance reflects the balance between individual-tree strength and diversity across trees [2]. The algorithm can accommodate complex data structures, although its outputs still require cautious interpretation.

Hyperparameters determine model capacity. The number of trees affects aggregation stability; max_depth controls tree complexity; min_samples_split and min_samples_leaf restrict data fragmentation; max_features governs feature diversity at each split; and bootstrap determines the sampling mechanism used to construct trees. Probst, Wright, and Boulesteix note that default settings are often competitive, but tuning may improve performance for a particular dataset [3]. With a small sample, limiting tree depth can act as regularization and reduce the tendency to memorize idiosyncratic patterns.

In unemployment forecasting, Gogas et al. compare decision trees, Random Forest, support vector machines, and elastic-net logistic regression for predicting the direction of euro-area unemployment [4]. Their out-of-sample evaluation favors Random Forest. Katris evaluates statistical and machine-learning models across several national unemployment series and shows that no single method dominates every horizon and data pattern [5]. Kreiner and Duca find that machine-learning models combining economic indicators can improve U.S. unemployment forecasts [6]. Orozco-Castañeda, Sierra-Suárez, and Vidal extend labor-market forecasting by integrating macroeconomic indicators, labor metrics, and online search data [7].

High-frequency information is also increasingly used for nowcasting. D'Amuri and Marcucci demonstrate that Google job-search intensity contains predictive information for U.S. unemployment [8]. Naccarato et al. combine official statistics and Google Trends to forecast Italian youth unemployment [9]. These studies show that unemployment prediction may benefit from signals that move faster than conventional statistical releases, while also demonstrating that forecast quality depends on stable digital behavior and a persistent relationship between online indicators and labor-market conditions.

Validation design is fundamental to forecasting. Bergmeir, Hyndman, and Koo discuss the conditions under which cross-validation can be valid for autoregressive prediction [11]. Hyndman and Koehler examine the limitations of forecast-accuracy measures and stress that metrics must be selected according to scale and decision purpose [12]. For limited samples, Vabalas et al. show that performance estimates may be highly biased and unstable [13]. Varma and Simon further warn that using the same folds for model selection and error estimation can produce optimistic results, supporting the use of nested validation or an independent test set [17].

Feature-importance interpretation is also constrained. Mean-decrease-in-impurity importance can favor variables because of scale, category count, correlation, or tree-splitting structure [14]. It must not therefore be treated as a measure of causal influence. In economic research, machine learning is primarily designed to minimize predictive error, whereas causal inference requires a distinct identification strategy [15]. This distinction becomes critical when model outputs are translated into policy [16].

The literature therefore leaves room for a city-level unemployment framework that combines cross-domain indicators, preserves a temporal test set, and explicitly assesses policy readiness. This study addresses that gap as an exploratory analysis. Unlike studies using hundreds of monthly observations or high-dimensional macroeconomic datasets, the present experiment relies on a very small simulated dataset. Its primary emphasis is consequently on process validity, transparent limitations, and the specification of future data requirements.

Research Methodology

The study uses an exploratory quantitative design based on a machine-learning experiment. The unit of analysis is a subdistrict-year observation. Implementation was conducted in Python on Google Colab using pandas, NumPy, Matplotlib, joblib, and scikit-learn. A scikit-learn Pipeline combines preprocessing and model estimation so that transformations learned from the training data are applied consistently to the test data.

The design is not intended to estimate causal effects. Its methodological objective is to test preliminary predictive feasibility and identify the limits of model generalization. Feature importance is therefore used to describe model behavior, not to claim that changing a predictor would cause a corresponding change in unemployment.

Results

4.1 ata Source, Status, and Coverage

The Medan City Medium-Term Development Plan for 2025–2029 was used as a conceptual basis for selecting relevant indicators, including population, working-age population, labor-force participation, human development, GRDP per capita, economic growth, inflation, poverty, the Gini index, entrepreneurship, and unemployment. The experimental dataset contains 20 observations covering four subdistricts—Medan Baru, Medan Kota, Medan Denai, and Medan Tuntungan—during 2020–2024.

Data-integrity statement: the experimental values constitute a structured simulated dataset previously created to demonstrate Random Forest implementation. They are not direct extractions from subdistrict tables in the development plan, are not derived from official labor-force survey microdata, and must not be presented as official government statistics. The official city-level unemployment figures are used only as policy context and are not the direct target for each simulated row. This separation prevents empirical claims from exceeding the evidentiary support of the data.

Table 1. Research variables and operational definitions

No.	Variable	Type	Role	Operational definition
1	Year	Numeric	Predictor	Observation period, 2020–2024.
2	Subdistrict	Categorical	Predictor	Simulated geographic identity: Medan Baru, Medan Kota, Medan Denai, and Medan Tuntungan.
3	Population	Numeric	Predictor	Population in the relevant subdistrict-year observation.
4	Working_Age_Population	Numeric	Predictor	Population represented as belonging to the working-age group.
5	Labor_Force_Participation	Numeric	Predictor	Percentage of the working-age population employed or actively seeking work.
6	HDI	Numeric	Predictor	Composite indicator of human development.
7	GRDP_Per_Capita_Million	Numeric	Predictor	Economic output per resident, represented in millions of rupiah.
8	Economic_Growth	Numeric	Predictor	Annual percentage change in economic activity.
9	Inflation	Numeric	Predictor	Annual percentage change in the general price level.
10	Poverty	Numeric	Predictor	Percentage of residents represented as living below the poverty line.
11	Gini_Index	Numeric	Predictor	Indicator of expenditure or income-distribution inequality.
12	Entrepreneurship_Ratio	Numeric	Predictor	Representation of the proportion of entrepreneurs or business owners.
13	Unemployment_Rate	Numeric	Target	Unemployment percentage predicted by the model.

4.2 Data Audit and Cleaning

The initial dataset contains 20 rows and 13 columns. The audit examined data types, missing values, unique values, and duplicates. All numeric columns were stored as integer or floating-point variables, whereas Subdistrict was categorical. No missing values or duplicate

rows were found. The code nevertheless includes median imputation for numeric variables should missing values appear in an expanded dataset. Subdistrict labels were stripped of leading and trailing spaces.

No normalization or standardization was applied because decision trees split observations according to order and thresholds and are not scale-sensitive in the same way as distance-based models. The Subdistrict variable was transformed using OneHotEncoder with handle_unknown="ignore". The four subdistrict categories generated four dummy features, while eleven numeric predictors were passed through without transformation, producing a final feature space of 15 variables.

4.3 Descriptive Analysis

Across all observations, the target has a mean of 7.6770%, a standard deviation of 0.8138%, a minimum of 6.4100%, and a maximum of 9.2100%. Mean labor-force participation is 68.8100%; mean HDI is 82.4800; mean GRDP per capita is 81.2650 million; mean economic growth is 4.5375%; mean inflation is 2.8980%; mean poverty is 7.1400%; mean Gini index is 0.33275; and mean entrepreneurship ratio is 2.5360.

Table 2. Descriptive statistics of the principal variables

Variable	Mean	Standard deviation	Minimum	Maximum
Unemployment rate (%)	7.6770	0.8138	6.4100	9.2100
Labor-force participation (%)	68.8100	1.4201	66.1000	71.2000
HDI	82.4800	0.8301	80.9000	84.1000
GRDP per capita (million)	81.2650	5.3354	71.2000	91.3000
Economic growth (%)	4.5375	0.9317	2.9500	5.5100
Inflation (%)	2.8980	1.4389	1.8500	5.6400
Poverty (%)	7.1400	0.4385	6.4300	8.1000
Gini index	0.3328	0.0072	0.3210	0.3470
Entrepreneurship ratio	2.5360	0.2442	2.0500	2.9500

4.4 Data Partitioning

The partition preserves temporal order. Sixteen observations from 2020–2023 were used for training and optimization, while all four observations from 2024 were retained as the test set. This strategy prevents the model from observing 2024 target values while learning predictor-target relationships.

Table 3. **Experimental data partition**

Component	Years	Number of observations	Proportion
Training data	2020–2023	16	80%
Test data	2024	4	20%
Total	2020–2024	20	100%

4.5 Baseline Model

The baseline Random Forest was configured with n_estimators=300, max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features="sqrt", random_state=42, and n_jobs=-1. Preprocessing and estimation were combined in a single Pipeline. The random state was fixed to support reproducibility.

4.6 Cross-Validation and Hyperparameter Optimization

Training observations were grouped by Year. GroupKFold with four groups was used so that all observations from the same year remained within the same fold. Baseline cross-validation was evaluated using MAE, RMSE, and R^2. RandomizedSearchCV then evaluated 30 parameter configurations over four folds, resulting in 120 model fits. The tuning objective was negative mean absolute error.

The search space was defined as follows:

a. n_estimators: 100, 200, 300, 500, and 700;
b. max_depth: None, 3, 5, 7, and 10;

c. min_samples_split: 2, 3, 4, and 5;
d. min_samples_leaf: 1, 2, and 3;
e. max_features: sqrt, log2, 0.7, and 1.0;
f. bootstrap: True and False.

GroupKFold prevents observations from the same year from appearing in both training and validation portions of a fold, but it does not always preserve chronological direction. Final evaluation therefore remains based on the 2024 temporal holdout. Rolling-origin validation or TimeSeriesSplit should be preferred when a longer series becomes available.

4.7 Evaluation Metrics

Four regression metrics were used:

$$MAE = (1/n)\ \Sigma(i=1 \text{ to } n)\ |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{[(1/n)\ \Sigma(i=1 \text{ to } n)\ (y_i - \hat{y}_i)^2]}$$

$$R^2 = 1 - [\Sigma(i=1 \text{ to } n)(y_i - \hat{y}_i)^2 / \Sigma(i=1 \text{ to } n)(y_i - \bar{y})^2]$$

$$MAPE = (100\%/n)\ \Sigma(i=1 \text{ to } n)\ |(y_i - \hat{y}_i)/y_i|$$

MAE expresses the average absolute error in unemployment percentage points. RMSE imposes a larger penalty on extreme errors. R^2 compares the model with a mean-based predictor, while MAPE expresses average relative error as a percentage and must be interpreted together with scale-dependent metrics [12].

4.8 Feature-Importance Analysis

Importance was extracted from the fitted scikit-learn Random Forest through feature_importances_, which is based on impurity reduction. The feature values sum to one. They describe which variables most frequently contributed to tree splits; they do not estimate causal effects. Potential bias arising from predictor correlation and data structure was considered in accordance with Strobl et al. [14].

4.9 Scenario Prediction

The code constructs a 2025 scenario for Medan Baru with labor-force participation of 72.0%, HDI of 84.5, GRDP per capita of 95.0 million, economic growth of 5.60%, inflation of 2.30%, poverty of 6.30%, a Gini index of 0.318, and an entrepreneurship ratio of 3.10. This scenario is used solely to demonstrate the inference pipeline and is not an official projection.

4.10 Reproducibility

The optimized model was saved in JOBLIB format. The cleaned dataset, evaluation metrics, 2024 predictions, feature-importance values, and scenario output were exported as CSV files and compressed into a ZIP archive, enabling the inputs, analytical process, and outputs to be audited.

Results

5.1 Descriptive Dataset Pattern

The simulated dataset shows a decline in mean unemployment from 8.7675% in 2020 to 8.2175% in 2021, 7.5900% in 2022, 7.1175% in 2023, and 6.6925% in 2024. The average absolute decline over the period is 2.0750 percentage points, equivalent to approximately 23.67% of the 2020 mean.

Medan Denai has the highest five-year mean unemployment rate at 8.1120%, whereas Medan Tuntungan has the lowest at 7.3920%. In 2024, the minimum target is 6.4100% in Medan Tuntungan and the maximum is 7.1100% in Medan Denai. Across the four subdistricts, the 2020–2024 decline ranges from 2.01 to 2.11 percentage points.

Table 4. Mean target values by year

Year	Number of observations	Mean (%)	Minimum (%)	Maximum (%)	Standard deviation
2020	4	8.7675	8.4200	9.2100	0.3393
2021	4	8.2175	7.9400	8.7100	0.3428

2022	4	7.5900	7.3600	8.0100	0.2920
2023	4	7.1175	6.8300	7.5200	0.2967
2024	4	6.6925	6.4100	7.1100	0.3062

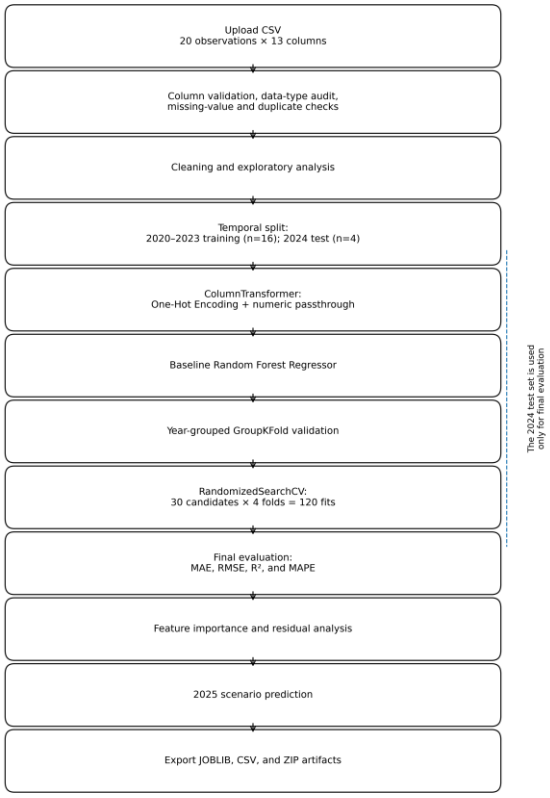


Figure 1. **Random Forest research workflow.**

The diagram traces the complete analytical sequence: CSV upload; validation of the 13 required columns; data-type, missing-value, and duplicate audits; cleaning and exploratory analysis; temporal partitioning into 2020–2023 training data (n=16) and 2024 test data (n=4); one-hot encoding of the subdistrict variable; baseline Random Forest estimation; year-grouped GroupKFold validation; RandomizedSearchCV with 30 candidates and 120 fits; final evaluation using MAE, RMSE, R^2, and MAPE; feature-importance and residual analysis; 2025 scenario prediction; and export of JOBLIB, CSV, and ZIP artifacts. The 2024 holdout enters only the final evaluation stage.

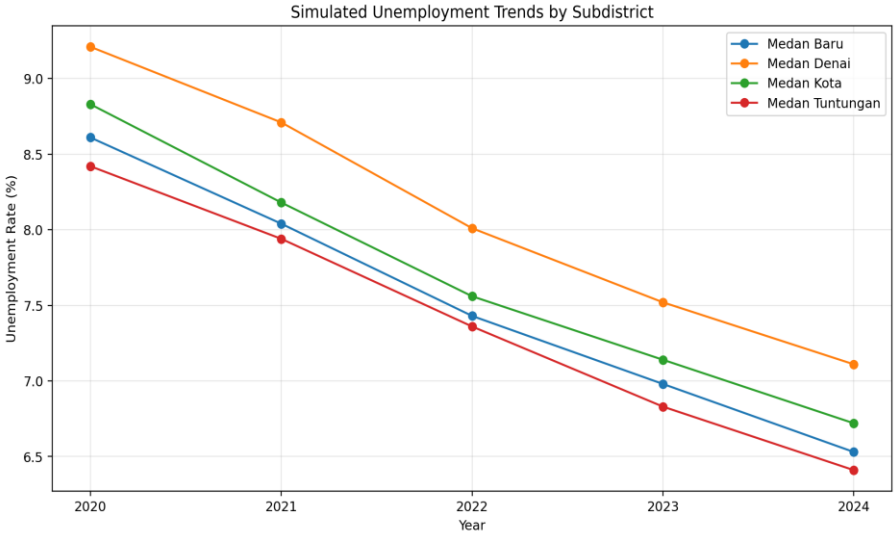


Figure 2. Simulated unemployment trends in four subdistricts during 2020–2024.

The horizontal axis represents year and the vertical axis represents unemployment percentage. Every subdistrict follows a relatively consistent downward trajectory. Medan Denai remains highest, declining from 9.21% in 2020

to 7.11% in 2024, while Medan Tuntungan is lowest for most of the period, declining from 8.42% to 6.41%. The four-subdistrict mean falls from 8.7675% to 6.6925%. These are structured simulated values, not official subdistrict statistics.

5.2 Correlation among Variables

Pearson correlations indicate very strong relationships between the target and several predictors. Labor-force participation correlates negatively with unemployment at -0.9848; entrepreneurship ratio at -0.9518; year at -0.9360; economic growth at -0.9331; GRDP per capita at -0.9214; and HDI at -0.8646. Poverty correlates positively at 0.9681, while the Gini index correlates positively at 0.9765.

These extreme correlations require substantial caution. The dataset contains only 20 observations and was constructed with relatively regular trends. Many predictors move jointly over time. High correlations may therefore reflect shared trends, simulation design, and multicollinearity rather than independent structural relationships. The same feature also helps explain why a small set of variables can dominate tree splits even though out-of-sample generalization remains poor.

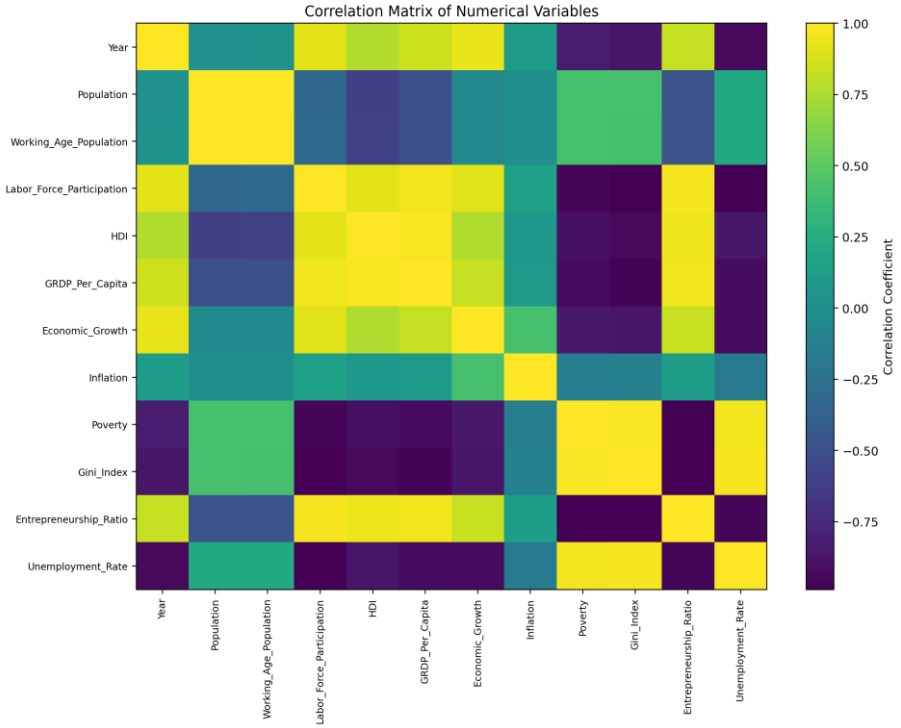


Figure 3. Pearson correlation matrix for all numerical variables.

Color intensity represents correlation from -1 to +1. The unemployment target is strongly negatively associated with labor-force participation (-0.9848), entrepreneurship ratio (-0.9518), economic growth (-0.9331), GRDP per capita (-0.9214), and HDI (-0.8646), and strongly positively associated with the Gini index (0.9765) and poverty (0.9681). Because the matrix is calculated from 20 simulated observations with shared time trends, it does not establish causality and may contain spurious correlation.

5.3 Baseline Cross-Validation

The four-group GroupKFold procedure produces highly variable performance. Fold-specific MAE values are 0.4845, 0.1231, 0.1369, and 0.4551. Mean MAE is 0.2999 with a standard deviation of 0.1966. Mean RMSE is 0.3220 with a standard deviation of 0.2098. Fold-specific R^2 values are -2.8935, 0.7562, 0.7225, and -1.8929, yielding a mean of -0.8269 and a standard deviation of 1.8542.

Table 5. GroupKFold results for the baseline model

Fold	MAE	RMSE	R^2
1	0.4845	0.5070	-2.8935
2	0.1231	0.1249	0.7562
3	0.1369	0.1564	0.7225
4	0.4551	0.4997	-1.8929

Mean	0.2999	0.3220	-0.8269
Standard deviation	0.1966	0.2098	1.8542

The wide variation across folds indicates that performance is highly sensitive to the year assigned to validation. Two folds produce positive R^2 values, while two produce strongly negative values. This instability is consistent with evidence that performance estimates from very small machine-learning samples can have high variance [13]. GroupKFold also does not guarantee that every model is trained on the past and validated on the future. Cross-validation is therefore used for internal tuning, whereas the principal performance claim is based on the 2024 temporal holdout.

5.4 Hyperparameter Optimization

RandomizedSearchCV selected n_estimators=100, min_samples_split=3, min_samples_leaf=1, max_features=0.7, max_depth=3, and bootstrap=False. The best validation MAE was 0.2448.

Table 6. Optimized model hyperparameters

Hyperparameter	Optimized value	Interpretation
n_estimators	100	The ensemble aggregates 100 regression trees.
max_depth	3	Each tree is restricted to a maximum depth of three.
min_samples_split	3	A node can split only when it contains at least three observations.
min_samples_leaf	1	A terminal leaf may contain a minimum of one observation.
max_features	0.7	Each split considers 70% of the available features.
bootstrap	False	Trees are constructed without bootstrap sampling.
random_state	42	The randomization is reproducible.

The selected max_depth of 3 suggests that shallow trees are more appropriate for this very small dataset than unrestricted trees. Greater depth would increase the risk of learning unique combinations within the 16 training observations. The reduction from 300 trees in the baseline model to 100 trees does not imply that fewer trees are universally superior; it only identifies the configuration preferred within the specified search space and validation folds [3].

5.5 Comparison of the Baseline and Optimized Models

On the 2024 test set, the baseline model produces MAE of 0.4587, RMSE of 0.4821, R^2 of -2.3051, and MAPE of 6.9481%. The optimized model produces MAE of 0.3989, RMSE of 0.4184, R^2 of -1.4892, and MAPE of 6.0382%.

Table 7. Model performance on the 2024 test set

Model	MAE	RMSE	R^2	MAPE (%)
Baseline Random Forest	0.4587	0.4821	-2.3051	6.9481
Optimized Random Forest	0.3989	0.4184	-1.4892	6.0382
Relative change	-13.05%	-13.22%	Improved, but still negative	-13.10%

Optimization reduces MAE by approximately 13.05%, RMSE by 13.22%, and MAPE by 13.10%. Tuning therefore improves absolute error. R^2 nevertheless remains negative, meaning that the model's sum of squared errors exceeds the total sum of squares around the test-set mean. This outcome is influenced by a test set containing only four observations within a narrow target range of 6.41–7.11%, while absolute errors range from 0.2453 to 0.5747 percentage points.

The results require a distinction between apparently small absolute errors and adequate generalization. A MAPE near 6% is not sufficient evidence of a strong model when it still performs worse than a mean-based benchmark in explaining target variation. The optimized model is therefore not suitable for setting program targets, allocating budgets, or prioritizing labor-market interventions.

5.6 Subdistrict-Level Predictions on the Test Set

Table 8. Actual and predicted values for 2024

Subdistrict	Actual (%)	Predicted (%)	Residual: actual-predicted	Absolute error	APE (%)
Medan Baru	6.5300	6.9848	-0.4548	0.4548	6.9640
Medan Kota	6.7200	7.0407	-0.3207	0.3207	4.7723
Medan Denai	7.1100	7.3553	-0.2453	0.2453	3.4501
Medan Tuntungan	6.4100	6.9848	-0.5748	0.5748	8.9665

All residuals are negative because every prediction exceeds its actual value, indicating systematic overprediction on the 2024 holdout. The smallest absolute error occurs in Medan Denai at 0.2453 percentage points, whereas the largest occurs in Medan Tuntungan at 0.5748 percentage points. The identical prediction for Medan Baru and Medan Tuntungan, 6.9848%, suggests that shallow trees placed the two observations in equivalent terminal nodes or produced the same ensemble average.

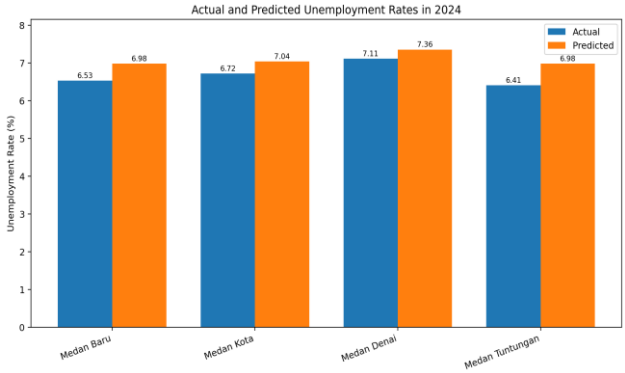


Figure 4. Actual and predicted unemployment rates in the 2024 test set.

Paired bars show Medan Baru at 6.53% actual and 6.9848% predicted; Medan Kota at 6.72% and 7.0407%; Medan Denai at 7.11% and 7.3553%; and Medan Tuntungan at 6.41% and 6.9848%. Every predicted value exceeds the corresponding actual value, revealing systematic overprediction. The test set contains only four observations and does not establish population-level generalization.

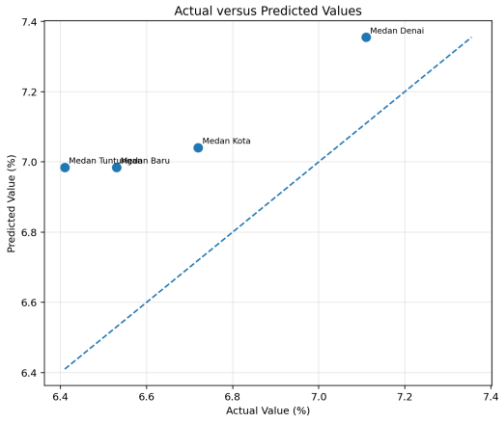


Figure 5. Scatter plot of actual and predicted values for the 2024 test set.

The horizontal axis represents actual unemployment and the vertical axis represents the optimized Random Forest prediction. The dashed y=x line denotes perfect prediction. All four points lie above the diagonal because the model consistently overpredicts. The prediction range, approximately 6.9848–7.3553%, is narrower than the actual range, indicating regression toward the mean and limited ability to distinguish subdistrict variation.

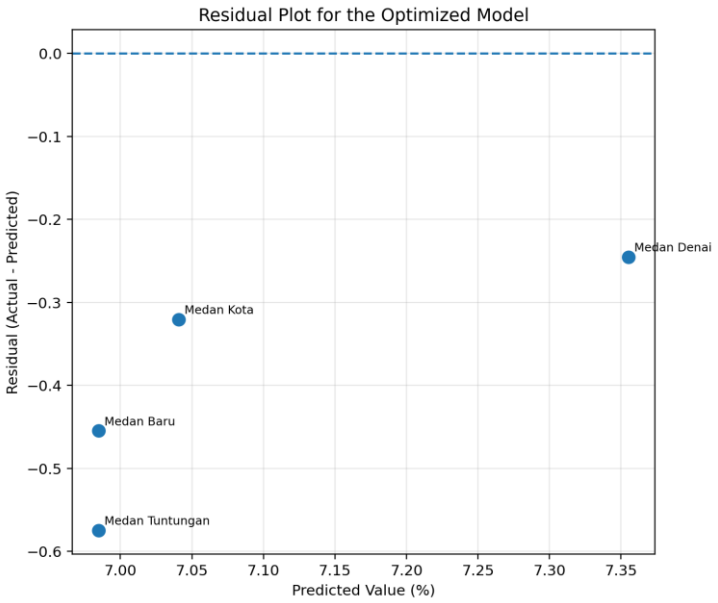


Figure 6. Residual plot for the optimized Random Forest on the 2024 test set.

The horizontal axis shows predicted values and the vertical axis shows residuals calculated as actual minus predicted. All residuals are below zero: -0.4548 for Medan Baru, -0.3207 for Medan Kota, -0.2453 for Medan Denai, and -0.5748 for Medan Tuntungan. The absence of positive residuals confirms directional bias. Four observations are insufficient to diagnose heteroskedasticity, but they clearly show that the model is not calibrated to the 2024 target level.

5.7 Feature Importance

Table 9. Feature importance of the optimized Random Forest

Rank	Feature	Importance	Percentage
1	Labor_Force_Participation	0.466904	46.6904%
2	Economic_Growth	0.372982	37.2982%
3	Inflation	0.052848	5.2848%
4	Gini_Index	0.033201	3.3201%
5	Poverty	0.028200	2.8200%
6	Entrepreneurship_Ratio	0.017496	1.7496%
7	Year	0.015645	1.5645%
8	HDI	0.006770	0.6770%
9	GRDP_Per_Capita_Million	0.004295	0.4295%
10	Medan_Kota_Dummy	0.001659	0.1659%
11	Working_Age_Population	0.000000	0.0000%
12	Medan_Baru_Dummy	0.000000	0.0000%
13	Population	0.000000	0.0000%
14	Medan_Tuntungan_Dummy	0.000000	0.0000%
15	Medan_Denai_Dummy	0.000000	0.0000%

Labor-force participation contributes 46.69% and economic growth contributes 37.30%, together accounting for 83.99% of total importance. The model therefore forms most of its predictions through thresholds involving these two variables. Inflation ranks third, but its contribution is substantially smaller at 5.28%. Demographic variables and subdistrict dummy variables receive little or no importance.

The concentration of importance in two variables must not be translated into a claim that higher labor-force participation or economic growth automatically causes lower unemployment. First, labor-force participation and the target move systematically over time in the simulated dataset. Second, the economic predictors are mutually correlated. Third, impurity-based importance may distribute credit unstably among correlated variables [14]. Fourth, Random Forest optimizes prediction rather than causal identification [15].

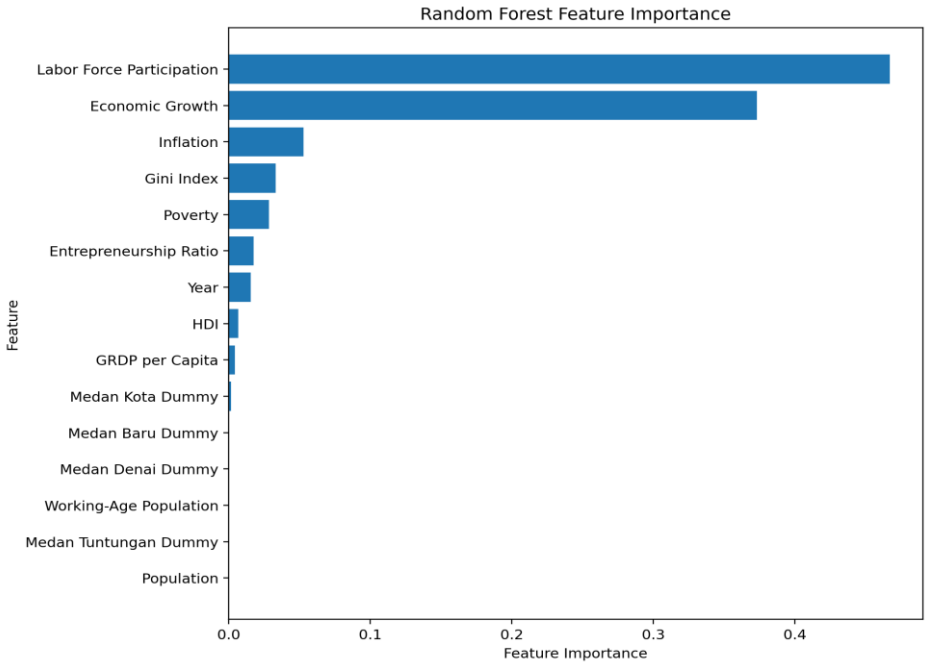


Figure 7. Impurity-based feature importance for the optimized Random Forest.

Labor-force participation is the dominant feature at 0.4669, followed by economic growth at 0.3730. Together they account for 0.8399, or 83.99%, of total importance. Inflation contributes 0.0528; Gini index 0.0332; poverty 0.0282; entrepreneurship ratio 0.0175; year 0.0156; HDI 0.0068; GRDP per capita 0.0043; and the Medan Kota dummy 0.0017. Five remaining features have zero importance. Impurity-based importance may be biased when predictors are correlated and must not be interpreted as a causal effect size.

5.8 2025 Scenario Prediction

The model predicts a 2025 unemployment rate of 6.98475% for the Medan Baru scenario. This is identical to its 2024 prediction for Medan Baru, illustrating the limited ability of Random Forest to extrapolate numerically beyond combinations observed during training. Decision trees predict terminal-node averages and do not naturally extend a linear trend outside the training range.

Table 10. Simulated Medan Baru scenario for 2025

Variable	Scenario value
Year	2025
Subdistrict	Medan Baru
Population	44,600
Working-age population	30,200
Labor-force participation	72.00%
HDI	84.50
GRDP per capita	95.00 million
Economic growth	5.60%
Inflation	2.30%
Poverty	6.30%
Gini index	0.318
Entrepreneurship ratio	3.10
Predicted unemployment rate	6.98475%

The scenario output must not be treated as an official forecast. It is generated from simulated data, the fitted model has a negative R^2, and no prediction interval is provided. The scenario only demonstrates that the pipeline can accept a new row, apply the fitted transformations, load the stored model, and produce an output.

5.9 Implications for Adaptive Labor Policy

Conceptually, a forecasting system could support government in four functions. First, early warning could identify areas or groups expected to experience rising unemployment. Second, targeting could prioritize training, certification, placement services, business incubation, and job matching for high-risk groups. Third, scenario analysis could evaluate changes in predicted unemployment under alternative assumptions for labor-force participation, growth, poverty, or entrepreneurship. Fourth, monitoring could compare forecasts with subsequent realizations to detect model instability and changes in labor-market structure.

These functions require much stronger data. A policy-grade dataset should cover all 21 subdistricts over a longer monthly or quarterly period and include sex, age group, educational attainment, recent-graduate status, business sector, occupation, wage, job-search duration, informal status, disability, migration, vacancies, realized investment, active micro and small enterprises, training participation, placement outcomes, and digital job-search indicators. It also requires consistent definitions, metadata, update schedules, and personal-data protection mechanisms.

Policy decisions should not rely on a single algorithm. Random Forest should be benchmarked against naive forecasts, historical means, linear regression, regularized regression, ARIMA/ETS, support vector regression, gradient boosting, XGBoost, CatBoost, and ensemble methods. Evaluation should employ rolling-origin temporal validation, nested tuning, and an external test set. Prediction uncertainty may be quantified using quantile regression forests, conformal prediction, or temporally appropriate bootstrap procedures.

5.10 Research Limitations

The first limitation is the structured simulated status of the dataset. Subdistrict-level values are not official statistics and cannot establish factual local conditions. The second limitation is the sample of only 20 observations, comprising 16 training and four test cases. This size is insufficient for stable machine-learning performance estimation [13].

The third limitation is the highly regular trend structure. Labor-force participation, HDI, GRDP per capita, growth, poverty, the Gini index, and the target move jointly over time, allowing the model to learn the simulation trend rather than an economic mechanism. Fourth, GroupKFold does not fully preserve chronology during tuning. Fifth, no simple benchmark was included, so the incremental value of Random Forest over a mean forecast, linear trend, or conventional time-series model cannot be established.

The sixth limitation is the sensitivity of impurity-based importance to correlated predictors [14]. Seventh, the study provides neither prediction intervals nor an analysis of feature-importance stability. Eighth, it evaluates only one scenario without sensitivity analysis. Ninth, the negative R^2 directly indicates inadequate generalization. The study must therefore be classified as a proof of concept rather than a final policy evaluation.

Conclusion

This study constructs a structured pipeline for unemployment prediction using Random Forest Regressor, covering data auditing, categorical preprocessing, temporal partitioning, year-grouped validation, hyperparameter optimization, evaluation, feature interpretation, scenario prediction, and model persistence. RandomizedSearchCV selected 100 trees, max_depth=3, min_samples_split=3, min_samples_leaf=1, max_features=0.7, and bootstrap=False. Relative to the baseline model, the optimized model reduced MAE from 0.4587 to 0.3989, RMSE from 0.4821 to 0.4184, and MAPE from 6.9481% to 6.0382%, corresponding to an improvement of approximately 13%.

Despite lower absolute error, the test-set R^2 remains negative at -1.4892. Every 2024 prediction is also higher than the corresponding actual value, producing systematic overprediction. The model cannot therefore be regarded as accurate, stable, or suitable for

policy decisions. Labor-force participation and economic growth are the two dominant features with a combined importance of 83.99%, but this concentration primarily reflects predictive patterns in the simulated dataset and does not demonstrate causality. Random Forest can be implemented as a computational framework for unemployment prediction, but technical execution is not equivalent to policy readiness. The principal implication is the need for a broader, granular, longitudinal, and documented official panel dataset. Future work should include all subdistricts, higher-frequency observations, individual and sectoral labor variables, rolling-origin validation, multi-algorithm benchmarks, uncertainty intervals, and external validation. Once these requirements are satisfied, the model may be developed into an early-warning and targeting system for adaptive training, placement, and entrepreneurship programs.

References

[1] P. Gogas, T. Papadimitriou, and E. Sofianos, "Forecasting unemployment in the euro area with machine learning," *J. Forecast.*, vol. 41, no. 3, pp. 551–566, Apr. 2022, doi: 10.1002/FOR.2824;WGROUP:STRING:PUBLICATION.

[2] C. Katris, "Prediction of Unemployment Rates with Time Series and Machine Learning Techniques," *Comput. Econ.*, vol. 55, no. 2, pp. 673–706, Feb. 2020, doi: 10.1007/S10614-019-09908-9/TABLES/19.

[3] Y. Mansour and M. Schain, "Learning with Maximum-Entropy Distributions," *Mach. Learn.*, vol. 45, no. 2, pp. 123–145, Nov. 2001, doi: 10.1023/A:1010950718922/METRICS.

[4] G. Biau and E. Scornet, "A Random Forest Guided Tour," Nov. 2015, Accessed: Jul. 25, 2026. [Online]. Available: http://arxiv.org/abs/1511.05741

[5] A.-L. Boulesteix, M. N. Wright, P. Probst, and M. Wright, "Hyperparameters and Tuning Strategies for Random Forest Vegetation modeling in a spatiotemporal framework View project mlr: R-package for machine learning View project Philipp Probst Hyperparameters and Tuning Strategies for Random Forest," 2018, Accessed: Jul. 25, 2026. [Online]. Available: https://www.researchgate.net/publication/324438530

[6] A. Kreiner and J. V. Duca, "Can machine learning on economic data better forecast the unemployment rate?," Applied Economics Letters, vol. 27, no. 17, pp. 1434–1437, 2020, doi: 10.1080/13504851.2019.1688237.

[7] J. M. Orozco-Castañeda, L. P. Sierra-Suárez, and P. Vidal, "Labor market forecasting in unprecedented times: A machine learning approach," Bulletin of Economic Research, vol. 76, no. 4, pp. 893–915, 2024, doi: 10.1111/boer.12451.

[8] F. D'Amuri and J. Marcucci, "The predictive power of Google searches in forecasting US unemployment," International Journal of Forecasting, vol. 33, no. 4, pp. 801–816, 2017, doi: 10.1016/j.ijforecast.2017.03.004.

[9] A. Naccarato, S. Falorsi, S. Loriga, and A. Pierini, "Combining official and Google Trends data to forecast the Italian youth unemployment rate," Technological Forecasting and Social Change, vol. 130, pp. 114–122, 2018, doi: 10.1016/j.techfore.2017.11.022.

[10] A. R. S. Parmezan, V. M. A. Souza, and G. E. A. P. A. Batista, "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model," Information Sciences, vol. 484, pp. 302–337, 2019, doi: 10.1016/j.ins.2019.01.076.

[11] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," Computational Statistics & Data Analysis, vol. 120, pp. 70–83, 2018, doi: 10.1016/j.csda.2017.11.003.
[12] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," International Journal of Forecasting, vol. 22, no. 4, pp. 679–688, 2006, doi: 10.1016/j.ijforecast.2006.03.001.
[13] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," PLOS ONE, vol. 14, no. 11, Art. no. e0224365, 2019, doi: 10.1371/journal.pone.0224365.
[14] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics, vol. 8, Art. no. 25, 2007, doi: 10.1186/1471-2105-8-25.
[15] S. Mullainathan and J. Spiess, "Machine learning: An applied econometric approach," Journal of Economic Perspectives, vol. 31, no. 2, pp. 87–106, 2017, doi: 10.1257/jep.31.2.87.
[16] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, "Prediction policy problems," American Economic Review, vol. 105, no. 5, pp. 491–495, 2015, doi: 10.1257/aer.p20151023.
[17] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, Art. no. 91, 2006, doi: 10.1186/1471-2105-7-91.