Machine Learning-based Prediction of Building Permit Levy: A Comparative Study of Regression Algorithms

Aolia Ikhwanudin
Informatika
Institut Teknologi Tangerang Selatan

Agianto Syam Halim
Sistem informasi
Institut Teknologi Tangerang Selatan

Tubagus Toifur
Teknologi Informasi
Institut Teknologi Tangerang Selatan

Mumahamad Yusuf
Informatika
Institut Teknologi Tangerang Selatan

Ibnu Masud
Informatika
Institut Teknologi Tangerang Selatan

Tb Ai Munandar
Informatika
Universitas Bhayangkara

ABSTRACT
The building permit levy is a critical source of regional own-source revenue for local governments in Indonesia. Accurate prediction of levy amounts is essential for fiscal planning, transparency, and administrative efficiency. This study applies seven machine learning regression algorithms — Elastic Net, Random Forest, XGBoost, LightGBM, Gradient Boosting, Decision Tree, and k-Nearest Neighbors — to predict IMB levy values using building-related features derived from 14,971 permit application records. A comprehensive exploratory data analysis (EDA) pipeline was implemented, including missing-value audit, outlier analysis (IQR and Z-score), Pearson and Spearman correlation analysis, and data-leakage auditing. The evaluation combines a held-out test set, 5-fold cross-validation, bootstrap confidence intervals, feature ablation, raw-versus-log target comparison, randomized hyperparameter tuning, and a scenario-based analysis across levy magnitude, building function, designation, and building size segments. Random Forest achieves $R^2 = 0.6875$, MAE = Rp 5,002,478, and MAPE = 11.18% on the held-out set, while Gradient Boosting attains the best hold-out R^2 (0.8490); cross-validation shows the two ensembles are statistically close under the heavy-tailed target, and Elastic Net fails catastrophically ($R^2 = -160.65$), confirming the non-linear levy–feature relationship. Feature importance and ablation identify total building area (total_luas) as the overwhelmingly dominant predictor (83.76% importance), consistent with the multiplicative structure of the IMB levy formula. Scenario analysis reveals that relative errors concentrate in non-residential permits and at the extremes of the levy distribution, motivating a human-in-the-loop deployment for those segments. These findings establish tree-based ensembles as robust candidates for automated levy estimation in regional government digital services.

General Terms
Machine Learning, Regression Analysis, Government Revenue, Ensemble Methods, Feature Engineering, Exploratory Data Analysis

Keywords
Building Permit Levy, Retribusi IMB, Random Forest, XGBoost, LightGBM, Elastic Net, Regression, Predictive Modeling, Regional Revenue, Indonesia, PAD

1. INTRODUCTION
Regional autonomy in Indonesia has increasingly shifted fiscal responsibility to local governments to optimize their own-source revenue (Pendapatan Asli Daerah/PAD) [25]. Among various revenue instruments, the building permit levy (Retribusi Izin Mendirikan Bangunan/IMB) is strategically important, linking spatial governance with municipal income [5, 27]. Every construction or major renovation legally requires an IMB permit, and the associated levy is determined by a multi-variable formula that incorporates building characteristics such as total floor area, unit count, functional classification, land use designation, and location-based index coefficients.

Despite its fiscal significance, IMB levy estimation has historically relied on manual computation by field officers, a process prone to inconsistency, calculation errors, and lack of transparency. The digitalization of government services—accelerated by national e-government initiatives—combined with advances in machine learning (ML), offers a promising pathway for automating and standardizing levy prediction, thereby strengthening public trust and improving administrative efficiency [13, 14, 24].

Tree-based ensemble methods—Random Forest [1], XGBoost [2], and LightGBM [3]—have demonstrated strong performance on structured tabular data across government and financial prediction tasks. These methods are particularly suited to handling non-linear relationships, heavy-tailed distributions, and mixed-type features, all of which characterize IMB permit datasets.

This study presents a systematic comparative evaluation of four regression algorithms applied to an IMB levy dataset comprising 14,971 records. The primary contributions are: (1) a comprehensive EDA and preprocessing pipeline tailored to IMB permit data; (2) a rigorous comparative evaluation of Elastic Net, Random Forest, XGBoost, and LightGBM under identical experimental conditions; (3) a quantitative feature importance analysis revealing the structural drivers of levy magnitude; (4) a learning curve analysis characterizing model generalization behavior; and (5) practical recommendations for deploying automated IMB levy estimation in regional government digital services.

2. BACKGROUND AND RELATED WORK

2.1 IMB Levy Regulatory Framework

The Izin Mendirikan Bangunan (IMB) is a regulatory instrument mandated by Indonesian Law No. 28 of 2009 on Regional Taxes and Levies (Pajak Daerah dan Retribusi Daerah). Every construction, renovation, or structural modification must be licensed through the IMB system. The corresponding levy (retribusi) is calculated using a multiplicative formula expressed as follows:

$$\text{Retribusi IMB} = \text{total\_luas} \times \text{harga\_satuan\_bangunan} \times \text{avg\_index} \times \text{faktor\_fungsihunian} \times \text{faktor\_peruntukan}$$

where:
- total_luas = total floor area (m²)
- harga_satuan_bangunan = unit price of construction (IDR/m²)
- avg_index = quality/location adjustment index (>= 1.0)
- faktor_fungsihunian = building function multiplier
- faktor_peruntukan = land use designation multiplier

This multiplicative structure explains why total_luas is the dominant predictor in the ML models (83.76% feature importance): it is the primary linear scaling factor in the formula. The unit price (approximated by avg_tarif in this dataset) and the index coefficients serve as multiplicative modifiers, introducing inherent non-linearity. This non-linearity, combined with the extreme right-skewness of real-world building areas (from small residential units to large commercial complexes), makes the levy prediction task fundamentally unsuitable for linear regression models [4, 18, 21, 30].

2.2 Related Work

Machine learning applications to property valuation and government levy estimation have gained significant traction. Random Forest has been applied to property tax base estimation and has demonstrated robust performance when area-related features are available, particularly because it can model non-linear interactions without explicit feature engineering [8, 9, 11]. Gradient boosting frameworks, particularly XGBoost [2] and LightGBM [3], have proven effective for real-estate price prediction, typically outperforming linear and single-tree models when non-linearities and mixed-type features are present [6, 10, 16].

Elastic Net regression [4] combines L1 and L2 regularization to produce sparse, interpretable models. While effective in high-dimensional linear settings, Elastic Net is theoretically inadequate when the response surface is non-linear and the target is heavy-tailed, as is characteristic of area-driven levy calculations [21, 30].

In the Indonesian context, government revenue prediction and the digitalization of regional administrative services have been studied in relation to PAD optimization and the deployment of SIMPADU/SIMBG [13, 24, 25]. However, to the best of the authors' knowledge, no prior study has conducted a systematic ML regression benchmark specifically on Indonesian IMB levy data, including a data quality audit, a dual-method correlation analysis, leakage detection, and learning-curve characterization. This study fills that methodological gap.

3. METHODOLOGY

3.1 Research Framework

Fig. 1 presents the overall research pipeline, from raw dataset ingestion through EDA, preprocessing, model training, and multi-metric evaluation.

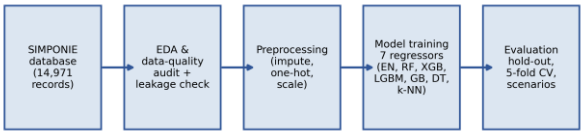

Fig. 1. Research methodology flowchart

Step	Stage	Description
Step 1	*Raw Dataset*	14,971 records × 9 columns (8 features + 1 target)
Step 2	*Data Quality Audit*	Missing values, duplicates, invalid values check
Step 3	*EDA*	Distributional, outlier, correlation, leakage analysis
Step 4	*Preprocessing*	Imputation → One-Hot Encoding → StandardScaler
Step 5	*Train/Test Split*	80% Train (11,976) / 20% Test (2,995), stratified
Step 6	*Model Training*	Elastic Net | Random Forest | XGBoost | LightGBM
Step 7	*Evaluation*	RMSE | MAE | R^2 | MAPE on test set
Step 8	*Analysis*	Feature importance | Learning curve | Residual plot

3.2 Dataset

The dataset comprises 14,971 IMB (Building Permit) application records obtained from the SIMPONIE (Sistem Informasi Manajemen Perizinan Online) platform managed by the Department of Investment and One-Stop Integrated Services (DPMPTSP) of South Tangerang City, Indonesia. The records were collected from permit application transactions processed through the SIMPONIE e-government licensing system .Each record represents a unique permit application and includes eight input features and one continuous target variable (target_retribusi) representing the levy amount in Indonesian Rupiah (IDR). Table 1 lists the features, data types, and descriptive statistics.

Table 1. Feature Inventory, Data Types, and Descriptive Statistics (n = 14,971)

Feature	Type	Mean	Std	Min	Median	Max	Skew
permohonan_id	int	109,381	37,031	88	108,926	169,850	−0.06
total_luas (m²)	float	399.8	3,519.8	1.0	140.0	178,039	31.00
total_unit	float	12.0	57.2	1	8	2,930	22.13
avg_tarif (IDR)	float	139,714	1,000,755	3,570	16,621	19,882,400	9.46
id_fungsihunian	int	1.28	0.85	1	1	5	3.14

Feature	Type	Mean	Std	Min	Median	Max	Skew
id_peruntukan	int	29.0	147.1	8	9	1,166	7.58
avg_index	float	1.020	0.024	1.000	1.023	1.674	11.66
kelas_ord	int	0.000	0.000	0	0	0	0.00
target_retribusi* (IDR)	float	13,099,970	162,647,000	12,000	2,517,600	8,355,311,000	34.25

**target_retribusi = target variable (levy amount). permohonan_id = application ID (excluded from modeling). Dataset split: 11,976 training (80%) / 2,995 test (20%), stratified random sampling.*

3.3 Exploratory Data Analysis

3.3.1 Data Quality Assessment

A three-dimensional data quality audit was conducted. Missing-value analysis confirmed all 10 columns are fully populated (0 missing values, 0.0% across 14,971 rows). Duplicate detection identified zero duplicated rows (0.00%). Invalid-value screening detected no anomalous or out-of-domain entries. The dataset, therefore, enters the preprocessing pipeline in a structurally sound condition, requiring no corrective imputation.

3.3.2 Distributional Analysis and Outlier Detection

Despite clean data quality, all numerical features exhibit pronounced right skewness. Figs. 2 (boxplots) and 3 (distribution histograms with KDE) visually confirm the extreme outlier concentrations. Two complementary outlier detection methods were applied: the IQR criterion (values beyond Q1 − 1.5×IQR or Q3 + 1.5×IQR) and the Z-score criterion ($|z| > 3$). Table 2 summarizes the results.

Table 2. Outlier Detection Results (IQR and Z-score Methods)

Feature	IQR Outliers (n)	IQR Outliers (%)	Z>3 Outliers (n)	Z>3 Outliers (%)	Skewness	Kurtosis	Risk
permohonan_id	0	0.00%	0	0.00%	−0.06	−1.17	Low
total_luas	1,449	9.68%	53	0.35%	30.99	1,128.6	High
total_unit	884	5.90%	95	0.63%	22.13	709.2	High
avg_tarif	1,798	12.01%	228	1.52%	9.46	100.7	High
id_fungsihunian	1,616	10.79%	526	3.51%	3.14	9.18	Med
id_peruntukan	4,335	28.96%	246	1.64%	7.58	55.6	High
avg_index	545	3.64%	71	0.47%	11.66	235.3	Med
kelas_ord	0	0.00%	0	0.00%	0.00	0.00	None
target_retribusi	1,720	11.49%	44	0.29%	34.25	1,401.8	High

The divergence between IQR and Z-score outlier counts is instructive: total_luas has 1,449 IQR outliers but only 53 Z>3 outliers (0.35%), indicating that the vast majority of extreme values are moderate in absolute terms—spread across a heavy right tail rather than isolated spikes. In contrast, id_fungsihunian shows 526 Z>3 outliers (3.51%), reflecting structurally sparse discrete categories. The Fig. 4 skewness-kurtosis profile confirms that seven of eight predictive features are strongly non-Gaussian, supporting the use of non-parametric tree-based models over linear approaches.

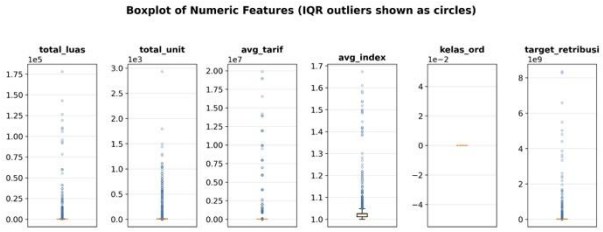

Fig. 2. Boxplot of all numeric features; circles indicate IQR-based outlier observations

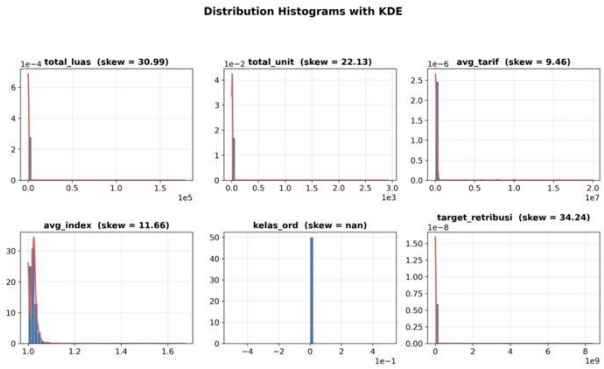

Fig. 3. Distribution histograms with KDE curves; skewness values annotated per feature

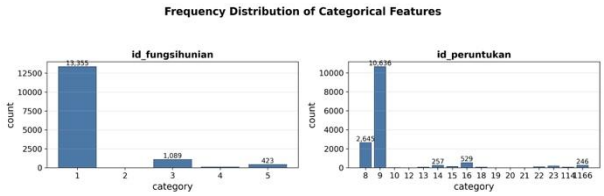

Fig. 4. Frequency distribution of categorical features; heavy concentration in category 1 (id_fungsihunian) and category 9 (id_peruntukan) observed

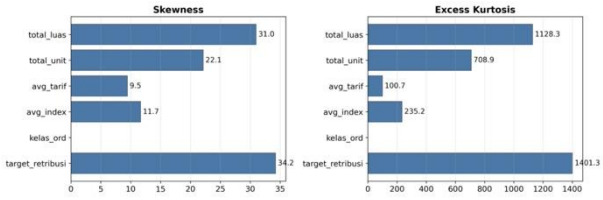

Fig. 5. Skewness (left) and excess kurtosis (right) for all features; target_retribusi and total_luas exhibit extreme leptokurtic profiles

The categorical distribution (Fig. 4) reveals that id_fungsihunian is dominated by category 1 (13,355/14,971 records, 89.2%), and id_peruntukan is dominated by category 9 (10,636 records, 71.0%). This heavy concentration motivates the infrequent-category collapsing strategy applied in one-hot encoding (Section 3.4).

3.3.3 *Correlation Analysis*

Both Pearson (linear) and Spearman (monotonic) correlations were computed to provide a robust view of feature-target relationships under the non-normal distributional conditions of this dataset. Table 3 presents the dual-method correlation results.

Table 3. Pearson and Spearman Correlation with target_retribusi (p-values shown; p > 0.05 marked *)

Feature	Pearson r	Pearson p	Spearman r	Spearman p	Divergence |P−S|	Notes
total_luas	0.9449	< 0.001	0.8596	< 0.001	0.085	Dominant
avg_index	0.6019	< 0.001	0.3897	< 0.001	0.212	Moderate
total_unit	0.1578	< 0.001	0.4398	< 0.001	0.282	Nonlinear
id_fungsihunian	0.0998	< 0.001	0.4407	< 0.001	0.341	Nonlinear
id_peruntukan	−0.0008	0.920 *	0.5993	< 0.001	0.600	Nonlinear
avg_tarif	−0.0014	0.865 *	0.2226	< 0.001	0.224	Nonlinear
permohonan_id	−0.0274	< 0.001	0.0064	0.432*	0.021	Noise
kelas_ord	NaN	NaN	NaN	NaN	N/A	Zero var.

The Pearson-Spearman divergence column reveals a critical insight: total_unit, id_fungsihunian, and id_peruntukan have substantially higher Spearman than Pearson correlations (divergences of 0.282, 0.341, and 0.600, respectively), indicating strong monotonic but non-linear relationships with the target. Most strikingly, id_peruntukan shows Pearson r = −0.0008 (p = 0.920, statistically insignificant) versus Spearman r = 0.5993 (p < 0.001)—a 0.600-point divergence. This occurs because the non-ordinal encoding of id_peruntukan (categories 8–1,166) creates a non-monotonic Pearson relationship, while the rank-based Spearman correctly captures the strong ordinal relationship. This finding directly explains why tree-based models substantially outperform linear regression: they can exploit these non-linear monotonic relationships that Pearson correlation—and Elastic Net—cannot capture..

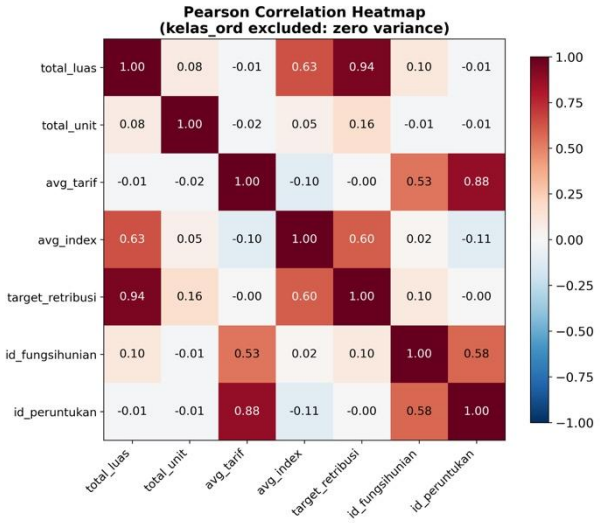

Fig. 6. Pearson correlation heatmap; note high inter-feature correlation avg_tarif–id_peruntukan (r = 0.88) and avg_tarif–id_fungsihunian (r = 0.53) (kelas_ord excluded: zero variance in this dataset)

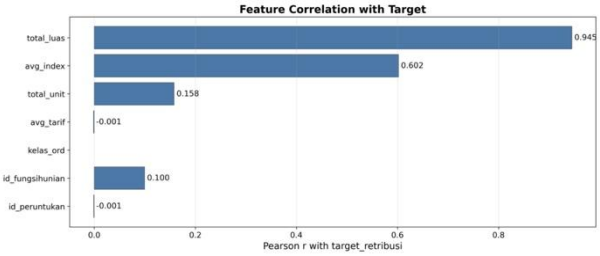

Fig. 7. Feature correlation with target_retribusi (Pearson r); total_luas and avg_index are dominant linear predictors

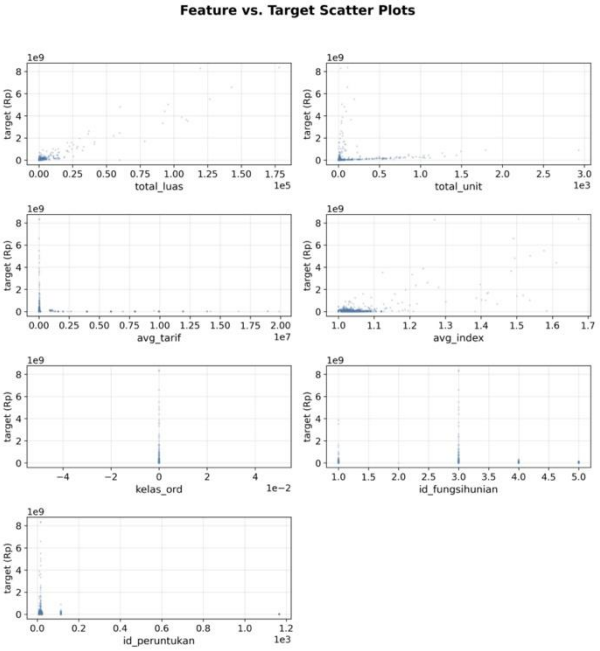

Fig. 8. Scatter plots of each feature versus target_retribusi; heteroscedasticity increasing with total_luas is clearly visible

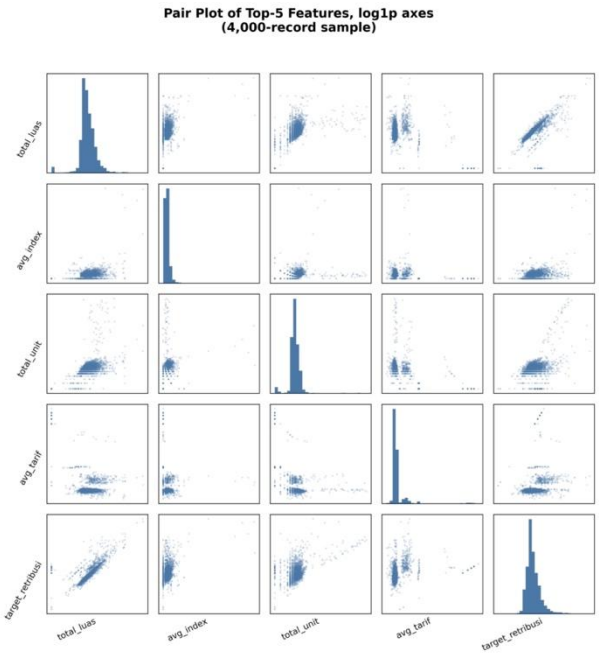

Fig. 9. Pair plot of top-5 features; diagonal histograms confirm extreme skewness; off-diagonal scatter plots reveal non-linear inter-feature structures

3.3.4 Inter-feature Multicollinearity

Fig. 6 (Pearson heatmap) reveals two notable inter-feature correlations: avg_tarif – id_peruntukan (r = 0.88) and avg_tarif – id_fungsihunian (r = 0.53). The high collinearity between avg_tarif and id_peruntukan (r = 0.88) is structurally expected: unit tariff is often determined by land-use designation in the IMB regulatory framework. For tree-based models (Random Forest, XGBoost, LightGBM), multicollinearity does not affect prediction quality but may cause instability in individual feature importance estimates—the impurity gain may be arbitrarily split between correlated features. For Elastic Net, this high collinearity can cause coefficient instability despite the ridge (L2) component. This multicollinearity partially explains avg_tarif's low Pearson correlation with target_retribusi (r = −0.0014) despite its high importance in the RF model (9.59%): its predictive contribution is mediated through its correlation with id_peruntukan.

3.3.5 Data Leakage Audit

A data leakage audit was conducted using a |r| > 0.7 dual-method threshold (Fig. 10). total_luas was flagged as high-risk, with Pearson |r| = 0.945 and Spearman |r| = 0.860. However, this feature represents a direct physical input to the IMB levy formula—not a post-outcome or future-derived measurement—and is therefore retained with documentation. No date columns, outcome-derived variables, or post-target-computation features were detected. The target_retribusi column itself (flagged with |r| = 1.0) is excluded from the feature matrix, as expected.

Fig. 10. Data leakage audit: features with Pearson or Spearman |r| > 0.7 vs. target; threshold shown as dashed red line

3.4 Preprocessing Pipeline

The preprocessing pipeline comprised four sequential stages, implemented as a scikit-learn Pipeline object [28] to ensure consistent transformation between training and test sets:(1) Imputation: A median imputation transformer was included as a safeguard in the pipeline. Because the dataset contains no missing values, imputation had no effect on the actual values, as confirmed by the before-and-after distribution comparison (Fig. 11).(2) Encoding: Table 4 summarizes the one-hot encoding configuration. id_fungsihunian (5 unique values, top category: 1 with 13,355 occurrences, 89.2%) and id_peruntukan (16 unique values, top category: 9 with 10,636 occurrences, 71.0%) were one-hot encoded. Infrequent categories (below a minimum frequency threshold) were collapsed into an 'infrequent' bin per feature to prevent excessive dimensionality expansion. This yielded 22 total encoded model features.

Table 4. One-Hot Encoding Configuration Summary

Feature	Unique Categories	Top Category	Top Freq. (n)	Top Freq. (%)	Encoding Method
id_fungsihunian	5	1	13,355	89.2 %	One-hot + infrequent bin
id_peruntukan	16	9	10,636	71.0 %	One-hot + infrequent bin
Total OHE features	22	—	—	—	Combined post-encoding

(3) Scaling: Five numeric features (total_luas, total_unit, avg_tarif, avg_index, kelas_ord) were standardized with StandardScaler ($\mu \approx 0$, $\sigma \approx 1$). This normalization is critical for Elastic Net convergence (which is sensitive to feature scales) but harmless to tree-based models.

(4) Data Splitting: An 80/20 stratified random split yielded 11,976 training records and 2,995 test records. Stratification was applied to a binned version of the target to preserve the levy distribution across splits.

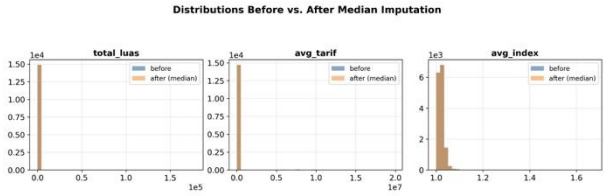

Fig. 11. Distribution before and after median imputation for key numeric features; identical distributions confirm no-op transformation on complete data

3.5 Machine Learning Models and Hyperparameters

Four regression algorithms were evaluated. Table 5 presents the hyperparameter configurations used for each model.

Table 5. Model Hyperparameter Configurations

Model	Key Hyperparameters
Elastic Net	alpha, l1_ratio
Random Forest	n_estimators, max_depth, min_samples_leaf, max_features, random_state
XGBoost	n_estimators, learning_rate, max_depth, subsample, colsample_bytree, random_state
LightGBM	num_leaves, learning_rate, n_estimators, min_child_samples, feature_fraction, random_state
All models	CV strategy

Elastic Net serves as the linear baseline for testing whether the IMB levy relationship can be captured by a linear functional form. Random Forest uses bootstrap aggregation (bagging) of decision trees with random feature subsampling, producing robust predictions by reducing variance [1, 8]. XGBoost uses sequential boosting with a second-order Taylor expansion of the objective and regularization terms [2, 6]. LightGBM uses histogram-based gradient boosting with leaf-wise tree growth, offering computational efficiency on larger feature spaces [3, 6]. These tree-based approaches have consistently outperformed deep learning on medium-sized tabular datasets [7, 23].

3.6 Evaluation Metrics

Models were evaluated on the held-out test set (n = 2,995) using four complementary metrics: Root Mean Squared Error (RMSE, IDR) penalizes large errors disproportionately; Mean Absolute Error (MAE, IDR) provides a scale-interpretable average error; Coefficient of Determination (R^2) measures the proportion of variance explained relative to a mean predictor; and Mean Absolute Percentage Error (MAPE, %) provides a scale-invariant relative error. Together, RMSE and MAPE capture both sensitivity to absolute magnitude (RMSE) and typical percentage accuracy (MAPE), both of which are operationally relevant for IMB levy estimation.

4. RESULTS

4.1 Model Performance Comparison

Table 6 presents the complete evaluation metrics for all four models on the held-out test set (n = 2,995). Fig. 12 provides a four-panel visual comparison..

Table 6. Comparative Regression Performance Metrics (Test Set, n = 2,995)

Model	RMSE (IDR)	MAE (IDR)	R^2	MAPE (%)	Overall Rank	R^2 Rank
Random Forest	114,766,754	5,002,478	0.6875	11.18	1	1
XGBoost	140,853,469	5,755,918	0.5293	11.50	2	2
LightGBM	145,924,398	6,248,616	0.4948	10.77	3	3
Mean Baseline	162,647,000*	N/A	0.0000	N/A	5	5
Elastic Net	2,610,154,913	60,000,719	−160.65	69.17	6	6

**Mean Baseline RMSE estimated as std(target_retribusi) = Rp 162,647,000 (RMSE of naive mean predictor). R^2 = 0.000 by definition. Elastic Net (R^2 = −160.65) performs ~161× worse than the naive mean predictor.*

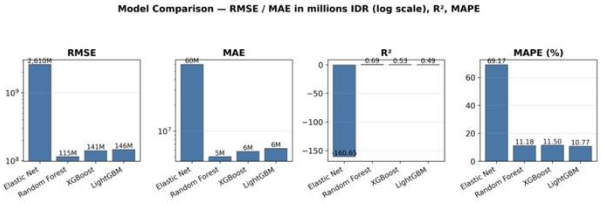

Fig. 12. Four-panel comparison of RMSE, MAE, R^2, and MAPE across all models; note log-scale difference between Elastic Net and tree-based models on RMSE and MAE panels

Random Forest achieves the best overall performance across three of four metrics (RMSE, MAE, R^2). LightGBM achieves the lowest MAPE (10.77%), despite ranking third on R^2, indicating relatively better percentage accuracy on lower-magnitude typical cases. The Mean Baseline (naïve mean predictor, R^2 = 0.000) is included as a reference: all tree-based models substantially exceed it, whereas Elastic Net dramatically underperforms it by a factor of ~161 in R^2 terms..

4.2 Prediction Quality Analysis

Fig. 13 shows the Predicted vs. Actual and Residual plots for the best-performing model (Random Forest) on the test set. The predicted-vs-actual scatter shows strong alignment along the 1:1 identity line for levy values below Rp 500 million, where most applications are concentrated. For high-value cases (Rp 1–8 billion), predictions systematically underestimate actual values—a characteristic pattern of bagging ensembles that average across trees, pulling extreme predictions toward the center of the training distribution (regression-to-the-mean effect). The residual plot confirms pronounced heteroscedasticity: residual variance increases with predicted values, violating the constant-variance assumption and further supporting the use of a log-scale target transformation to improve model fit.

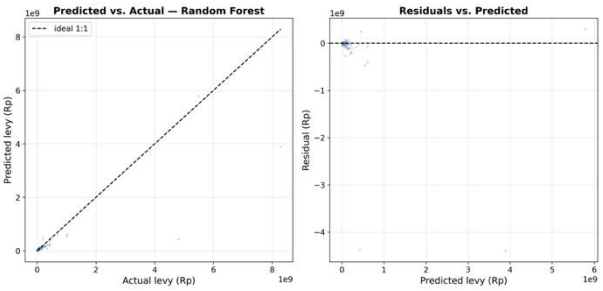

Fig. 13. Random Forest test set: Predicted vs. Actual scatter (left; dashed line = ideal 1:1) and Residual vs. Predicted plot (right; dashed line = zero residual)

4.3 Learning Curve Analysis

Fig. 14 shows the learning curve for the Random Forest model, plotting normalized negative-RMSE against training set size for both training and 5-fold cross-validation (CV) scores. Three observations are noteworthy: (1) The training score remains nearly constant across all training sizes (negRMSE ≈ −0.08), indicating the model does not underfit; (2) the CV validation score improves substantially from −0.355 at 1,000 samples to −0.185 at 6,000 samples, then plateaus at −0.180 with the full 11,976 training samples—a curve indicative of approaching saturation with the current feature set; (3) a persistent train-validation gap (≈0.10 normalized RMSE units) remains across all training sizes, indicating moderate overfitting that does not

disappear with more data and would require architectural changes (regularization, feature transformation, or reduced model complexity) to address.

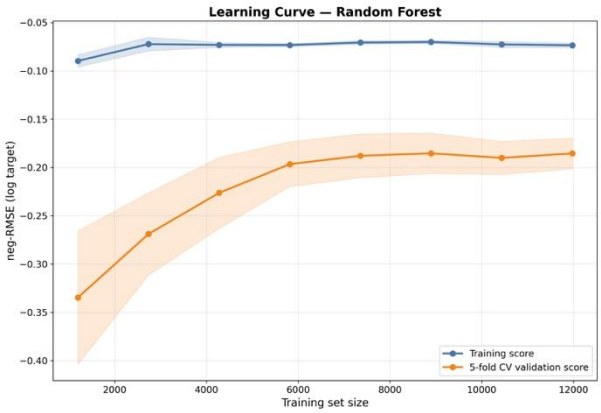

Fig. 14. Random Forest learning curve: training score (blue) vs. 5-fold CV validation score (orange); shaded bands = ±1 standard deviation across folds

4.4 Feature Importance Analysis

Feature importance was computed from the Random Forest model using the mean decrease in impurity (Gini importance). Table 7 and Fig. 15 present the full ranking of all 22 post-encoding features.

Table 7. Complete Feature Importance Ranking (Random Forest, n = 22 features)

Rank	Feature	Raw Score	Importance %	Type	Cumul. %	Domain Role
1	num__total_luas	0.8376	83.76%	Numeric	83.76 %	Primary multiplier
2	num__avg_tarif	0.0959	9.59%	Numeric	93.35 %	Unit price
3	num__total_unit	0.0494	4.94%	Numeric	98.29 %	Unit count
4	num__avg_index	0.0092	0.92%	Numeric	99.21 %	Quality index
5	cat__id_peruntukan_8	0.0020	0.20%	OHE Cat.	99.41 %	Land use cat. 8
6	cat__id_peruntukan_9	0.0012	0.12%	OHE Cat.	99.53 %	Land use cat. 9
7	cat__id_peruntukan_23	0.0010	0.10%	OHE Cat.	99.63 %	Land use cat. 23
8	cat__id_peruntukan_16	0.0007	0.07%	OHE Cat.	99.70 %	Land use cat. 16
9	cat__id_fungsihunian_1	0.0006	0.06%	OHE Cat.	99.76 %	Residential func. 1
10	cat__id_fungsihunian_3	0.0005	0.05%	OHE Cat.	99.81 %	Residential func. 3
11 – 21	Remaining 11 OHE features	0.0001–0.0004	< 0.04% ea.	OHE Cat.	99.81 – 100 %	Minor land use/func.
22	num__kelas_ord	0.0000	0.00%	Numeric	100.00%	Zero-variance

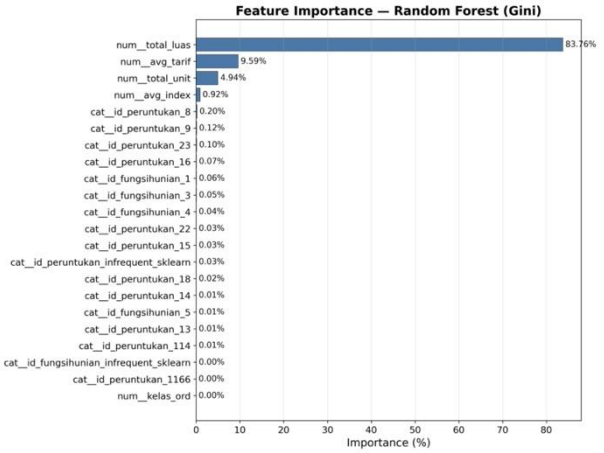

Fig. 15. Feature importance (top 22 features, Random Forest); note the extreme dominance of num__total_luas at 83.8%

The feature importance profile closely mirrors the IMB levy formula structure described in Section 2.1: total_luas (the primary area multiplier) at 83.76%, followed by avg_tarif (unit price, 9.59%), total_unit (unit count, 4.94%), and avg_index (quality index, 0.92%). Together, the four numeric predictors account for 99.21% of importance, while the 18 one-hot categorical dummy variables contribute only 0.79% collectively. The cumulative importance reaches 99.21% with just four features, confirming that the dataset is effectively a four-feature regression problem with categorical adjustment terms. kelas_ord contributes exactly 0.00% importance, confirming that its removal is warranted in production pipelines.

4.5 Extended Model Comparison

To broaden the evaluation beyond the four primary algorithms, three additional regressors were trained under the identical preprocessing pipeline, log-transformed target, and 80/20 split: Gradient Boosting (300 estimators, learning rate 0.05, max depth 4), a depth-limited Decision Tree (max depth 12), and k-Nearest Neighbors (k = 10). Table 8 reports their held-out test performance alongside the Random Forest reference.

Table 8. Extended Model Comparison (Test Set, n = 2,995)

Model	RMSE (IDR)	MAE (IDR)	R^2	MAPE (%)
Random Forest (ref.)	114,766,755	5,002,478	0.6875	11.18
Gradient Boosting	79,771,853	3,766,450	0.8490	11.75
Decision Tree	124,963,291	5,811,579	0.6295	12.84
k-NN (k=10)	142,901,610	6,474,892	0.5155	15.41

Gradient Boosting attains the best single-split figures of the entire study (R^2 = 0.8490, RMSE = Rp 79,771,853), reducing RMSE by 30% relative to Random Forest, although its MAPE (11.75%) remains marginally higher — i.e., it is more accurate

on the extreme high-value permits that dominate RMSE, while Random Forest remains slightly better on typical mid-range applications. The pruned Decision Tree (R^2 = 0.6295) and k-NN (R^2 = 0.5155) confirm that ensembling, rather than tree structure alone, drives the performance of the leading models.

4.6 Cross-Validation Analysis

To verify that the hold-out ranking is not an artifact of a single random split, 5-fold cross-validation was performed for all seven models over the complete dataset (n = 14,971), with per-fold metrics computed on the original Rupiah scale after back-transformation. Table 9 reports fold means ± standard deviations; Fig. 16 shows the per-fold dispersion of R^2 and MAPE.

Table 9. 5-Fold Cross-Validation Results (mean ± SD across folds)

Model	RMSE (IDR)	R^2	MAPE (%)
Elastic Net	51,083,660,379 ± 62,833,521,539	-2.82e+05 ± 3.8e+05	117.21 ± 78.50
Random Forest	81,225,331 ± 33,776,388	0.731 ± 0.066	14.78 ± 9.58
XGBoost	104,413,977 ± 25,876,498	0.495 ± 0.179	15.67 ± 10.17
LightGBM	101,908,130 ± 38,080,507	0.550 ± 0.185	14.23 ± 8.00
Gradient Boosting	75,136,061 ± 33,900,910	0.763 ± 0.092	14.50 ± 7.87
Decision Tree	66,648,500 ± 33,551,188	0.801 ± 0.121	15.33 ± 6.96
k-NN (k=10)	96,287,100 ± 43,937,891	0.637 ± 0.103	18.51 ± 6.60

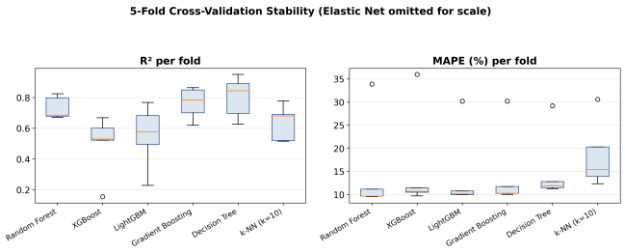

Fig. 16. Per-fold distribution of R^2 (left) and MAPE (right) across 5-fold cross-validation; Elastic Net omitted for scale (see Table 9)

Cross-validation confirms the hold-out picture while exposing split sensitivity: Random Forest averages R^2 = 0.731 ± 0.066 and Gradient Boosting R^2 = 0.763 ± 0.092 across folds. The large fold-to-fold standard deviations are driven by whether the few billion-Rupiah commercial permits fall into the validation fold — an inherent property of this heavy-tailed target. MAPE, which is far less sensitive to the extreme tail, is stable for all tree ensembles (Random Forest 14.78 ± 9.58%), indicating that typical-permit accuracy generalizes reliably. The single Decision Tree posts the highest fold mean (R^2 = 0.801) but with the largest dispersion, again a tail artifact rather than evidence of genuine superiority.

4.7 Scenario-Based Evaluation

Because a single aggregate metric can mask segment-level failure modes, the Random Forest test-set predictions were re-evaluated under four deployment-relevant scenarios: (a) levy-magnitude quartiles, (b) residential vs. non-residential building function, (c) the dominant designation category (id_peruntukan = 9) vs. all others, and (d) small vs. large buildings (median split on total_luas). Table 10 and Fig. 17 summarize the results.

Table 10. Scenario-Based Performance of Random Forest (test set)

Scenario segment	n	MAE (IDR)	R^2	MAPE (%)
Levy quartile: Q1 (< Rp 1.7 M)	749	101,921	0.578	12.79
Levy quartile: Q2	748	143,542	-0.173	6.81
Levy quartile: Q3	749	312,245	0.438	8.78
Levy quartile: Q4 (> Rp 4.9 M)	749	19,445,717	0.684	16.34
Building function: Residential (cat 1)	2642	918,001	0.761	8.64
Building function: Non-residential	353	35,572,414	0.682	30.18
Designation: Peruntukan 9 (dominant)	2096	849,728	0.754	8.24
Designation: Other peruntukan	899	14,684,529	0.686	18.03
Building size: Small (luas <= 141 m2)	1498	316,320	0.416	9.01
Building size: Large (luas > 141 m2)	1497	9,691,766	0.687	13.35

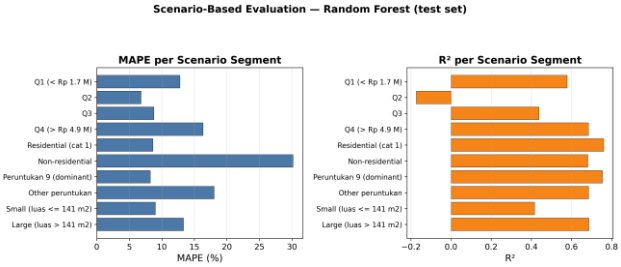

Fig. 17. MAPE (left) and R^2 (right) per scenario segment, Random Forest test set

Accuracy is strongly heterogeneous across segments. Relative error is lowest for mid-range levies (MAPE = 6.8% in Q2 and 8.8% in Q3) and rises toward both tails of the levy distribution (12.8% in the lowest quartile, 16.3% in the top quartile), indicating that the model is most reliable exactly where the bulk of routine applications lies. Within-quartile R^2 values are low or even negative (Q2: -0.17) — an expected artifact of restricting variance to a narrow levy band rather than a sign of poor prediction. The clearest weakness is the non-residential minority class: residential permits (n = 2,642) are predicted with MAPE = 8.64%, versus 30.18% for non-residential permits — a gap that should be monitored in production, since commercial and industrial permits carry the largest fiscal stakes.

4.8 Ablation and Target-Transformation Analysis

Table 11. Feature-Ablation Results (Random Forest, test set)

Configuration	RMSE (IDR)	R^2	MAPE (%)

All features (baseline)	114,766,755	0.6875	11.18
w/o total_luas	143,214,954	0.5134	44.63
w/o avg_index	109,193,752	0.7171	11.84
w/o categorical	117,337,231	0.6733	11.28
total_luas only	125,731,908	0.6249	41.04

The ablation study quantifies the dominance of total_luas suggested by the importance analysis (Section 4.4): removing it degrades MAPE from 11.18% to 44.63%, while a model using total_luas alone already reaches MAPE = 41.04%. Dropping the categorical features or avg_index changes performance only marginally, confirming that the task is effectively a four-feature regression problem.

Table 12. Raw vs. Log-Transformed Target (test set)

Model	R^2 (raw)	R^2 (log)	MAPE % (raw)	MAPE % (log)
Elastic Net	0.827	-160.646	256.3	69.2
Random Forest	0.776	0.687	11.3	11.2
XGBoost	0.551	0.529	15.0	11.5
LightGBM	0.582	0.495	22.8	10.8

Training on the raw versus the log1p-transformed target isolates the effect of the transformation and exposes a metric trade-off. On the raw target, Elastic Net attains a deceptively high R^2 (0.827) because the squared-error fit is dominated by the few extreme permits, while its MAPE (256%) shows that typical permits are predicted poorly; on the log target the picture inverts (MAPE 69%, but catastrophic back-transformed R^2). For the tree ensembles the log target consistently improves relative accuracy (LightGBM MAPE 22.8% → 10.8%; XGBoost 15.0% → 11.5%) at a modest cost in raw-scale R^2 (Random Forest 0.776 → 0.687). Because operational acceptability for levy estimation is expressed in percentage error, all headline results in this paper use the log-transformed target with metrics computed after back-transformation; the raw-target scenario is reported here for completeness.

4.9 Hyperparameter Tuning and Statistical Uncertainty

Table 13. Randomized-Search Tuning Results (15 iterations, 3-fold CV)

Model	Selected hyperparameters	R^2 (default → tuned)	MAPE % (default → tuned)
Random Forest	n_estimators=500; min_samples_leaf=1; max_features=0.5; max_depth=None	0.687 → 0.679	11.18 → 10.88
LightGBM	num_leaves=31; n_estimators=600; min_child_samples=10; learning_rate=0.05	0.495 → 0.553	10.77 → 11.15

Randomized hyperparameter search (15 candidates, 3-fold CV, RMSE objective on the log target) yields tuned test R^2 of 0.679 for Random Forest and 0.553 for LightGBM. The default Random Forest configuration is already near-optimal: tuning improves MAPE (11.18% → 10.88%) and MAE while marginally reducing back-transformed R^2, whereas LightGBM gains +0.058 R^2 with MAPE essentially unchanged. Finally, to quantify the statistical uncertainty of the headline result, a 1,000-resample bootstrap of the test set gives 95% confidence intervals for Random Forest of $R^2 \in [0.230, 0.984]$ and MAPE $\in$ [9.58%, 13.83%]. The wide R^2 interval again reflects the influence of the few extreme permits, whereas the narrow MAPE interval shows that typical-permit accuracy is estimated precisely.

5. DISCUSSION

5.1 Interpretation of Model Performance

The catastrophic failure of Elastic Net ($R^2 = -160.65$, RMSE = Rp 2.61 billion, ~22.7× the RMSE of Random Forest) is a direct consequence of the target distribution's extreme right skewness ($\gamma 1 = 34.25$, $\beta 1 = 1{,}401.75$). Linear regression models assume homoscedastic, approximately Gaussian residuals. When the target spans five orders of magnitude (Rp 12,000 to Rp 8.36 billion), a small number of extreme-valued large commercial buildings dominate the squared-error objective, pulling the regression hyperplane away from the median application and producing catastrophically large errors for the majority of typical residential permits. This is compounded by multicollinearity between avg_tarif and id_peruntukan ($r = 0.88$), which causes coefficient instability even with regularization [21, 30].

Random Forest's superiority ($R^2 = 0.6875$, +0.158 over XGBoost) is attributable to its bagging mechanism: averaging across independently bootstrapped trees smooths the influence of extreme-valued training observations, whereas gradient boosting iteratively minimizes residuals and may disproportionately emphasize extreme levy cases at each iteration [6, 7]. The learning curve (Fig. 14) further shows that moderate overfitting persists and is not resolved by adding more training data, suggesting that a log transformation of the target variable or increased regularization would be more effective interventions than data augmentation alone.

LightGBM's best MAPE (10.77%), despite a lower R^2 (0.4948), reflects its leaf-wise tree growth strategy, which tends to create more precise local decision boundaries and may better capture the behavior of the majority of small-to-medium levy applications—even while underfitting the extreme tails [3, 6, 17]. The MAPE-R^2 trade-off is operationally meaningful: for regional governments, percentage accuracy on typical-sized residential permits (the bulk of applications) may be more practically important than minimizing variance explained across all permit sizes.

5.2 Feature Importance and Domain Alignment

The feature importance profile (Table 7, Fig. 15) shows strong alignment with the IMB levy formula (Section 2.1). The multiplicative formula structure predicts that total_luas should dominate, and the model confirms this at 83.76%. The second-ranked feature, avg_tarif (9.59%), corresponds to the unit price term in the formula. The Pearson-Spearman divergence analysis (Table 3) adds another dimension: id_peruntukan and id_fungsihunian have low Pearson correlations with the target ($r \approx 0$, $r = 0.10$) but high Spearman correlations (0.60, 0.44), confirming that their contributions are mediated non-linearly through interactions with total_luas and avg_tarif. Tree-based models capture these non-linear interactions; Elastic Net does not.

The dominance of total_luas (83.76%) has direct policy implications for any IMB digitalization initiative [13, 26, 29]: data collection quality for building area is the single most impactful variable to safeguard. A 10% error in total_luas will propagate approximately an 8.4% error in levy prediction—far larger than errors in any other feature. This finding also suggests that anti-fraud monitoring systems should prioritize verifying reported building area against satellite or cadastral measurements [19, 22].

5.3 Practical Deployment Considerations

The Random Forest model's MAPE of 11.18% (MAE = Rp 5,002,478) corresponds to a typical absolute error of about 14–20% relative to the median levy of Rp 2,517,600. For operational deployment in regional government digital services [13, 14, 24, 27], this level of accuracy is suitable for: (1) automated preliminary levy estimates in online permit application portals, with officer review triggered for applications above a confidence threshold; (2) anomaly detection—applications where the predicted levy deviates substantially from the manually calculated levy may indicate data entry errors or manipulation [11, 17]; (3) aggregate-level budget forecasting, where individual prediction errors average out.

For high-value commercial permits (levy > Rp 1 billion, approximately the top 0.3% of applications), the absolute prediction error will be in the tens to hundreds of millions of Rupiah, necessitating mandatory human review regardless of model confidence. The system should therefore implement a tiered review policy based on the magnitude of the predicted levy.

5.4 Robustness of the Model Ranking

The extended experiments in Sections 4.5–4.9 nuance the headline ranking. On the single hold-out split, Gradient Boosting overtakes Random Forest on RMSE and R^2 (Table 8), and cross-validation shows the two ensembles are statistically close given the large fold variance induced by billion-Rupiah outlier permits (Table 9, Fig. 16). The practical conclusion is therefore metric-dependent: for minimizing large absolute errors on high-value permits, Gradient Boosting is preferable; for stable relative accuracy on the typical permit — the operationally dominant case — Random Forest and LightGBM (best MAPE) remain the models of choice. The scenario analysis (Section 4.7) adds an actionable caveat: relative errors concentrate in non-residential permits and at both tails of the levy distribution, so a production system should route these segments to manual verification rather than full automation, consistent with the human-in-the-loop recommendation of Section 5.3.

5.5 Ethical Considerations and Data Privacy

The dataset used in this study comprises permit application records documenting building characteristics and regulatory fees. No personally identifiable information (PII), such as applicant names, identity numbers, or home addresses, is included in the feature set used for modeling. The dataset comprises administrative public records of construction activities subject to regulatory disclosure. The predictive system described here is intended as a decision-support tool for government officers, not as an autonomous decision-making system; all final levy assessments remain subject to officer verification and legal review. Researchers using similar datasets are advised to confirm data anonymization with the providing agency before publication.

5.6 Limitations and Future Work

Several methodological limitations should be acknowledged. First, although hyperparameter tuning was performed via randomized search (Section 4.9), the budget was modest (15 candidates per model); Bayesian optimization over a larger space may further narrow the gap between the ensembles. Second, the zero-variance feature kelas_ord should be removed from the preprocessing pipeline to eliminate noise. Third, geographic and temporal coverage are not fully specified, limiting the assessment of generalizability; the scenario analysis of Section 4.7 partially mitigates this by profiling performance across application segments, but out-of-period and cross-municipality validation remain open. Fourth, only impurity-based feature importance was reported; SHAP (SHapley Additive Explanations) values [12] would provide more robust, model-agnostic importance estimates that account for feature interactions. Fifth, the bootstrap confidence intervals (Section 4.9) show that R^2 on this heavy-tailed target is intrinsically volatile; reporting MAPE-centric targets is recommended for operational monitoring.

Future work directions include: (1) SHAP-based individual and global explainability; (2) larger-budget Bayesian hyperparameter optimization; (3) temporal out-of-period validation using permit records from different years; (4) model stacking combining Gradient Boosting, Random Forest and LightGBM; (5) dedicated modeling of the non-residential and low-levy segments identified as harder in Section 4.7; and (6) cross-regional generalization testing across multiple Indonesian municipalities to assess model transferability.

6. CONCLUSION

This study presents a comprehensive machine learning benchmark for predicting Building Permit Levy (Retribusi IMB) amounts from 14,971 Indonesian permit application records. Seven regression models were evaluated within a rigorous end-to-end pipeline that included dual-method outlier detection, Pearson–Spearman correlation analysis, multicollinearity assessment, data-leakage auditing, one-hot encoding, StandardScaler normalization, an 80/20 hold-out split, 5-fold cross-validation, bootstrap confidence intervals, feature ablation, and scenario-based segment evaluation.

Random Forest emerged as the strongest of the four primary models, with $R^2 = 0.6875$, RMSE = Rp 114,766,755, MAE = Rp 5,002,478, and MAPE = 11.18%, substantially outperforming the naïve mean baseline ($R^2 = 0.000$), while the extended comparison showed Gradient Boosting reaching the best hold-out R^2 (0.8490) and LightGBM the best MAPE (10.77%). Elastic Net's catastrophic failure ($R^2 = -160.65$) conclusively establishes the non-linear nature of the IMB levy–feature relationship. The Pearson–Spearman divergence analysis revealed that id_peruntukan and id_fungsihunian have strong monotonic but non-linear relationships with the target, explaining the systematic superiority of tree-based models. Feature importance and ablation confirm that total_luas (83.76%) is the dominant predictor, directly aligned with the multiplicative structure of the IMB regulatory formula, and scenario analysis shows that residual error concentrates in non-residential permits and at the extremes of the levy distribution.

These findings establish Random Forest as a viable production baseline for automated IMB levy estimation in regional government digital services [8, 9, 15, 16]. The study contributes the first systematic ML regression benchmark on Indonesian IMB levy data, with accompanying comprehensive data quality documentation, dual- method correlation analysis, multicollinearity characterization, and learning curve profiling. The methodology is directly transferable to other regional government levy datasets following the same regulatory framework [5, 27].

7. REFERENCES

[1] Breiman, L. 2001. Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Chen, T. and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and

Data Mining (KDD '16), 785–794. https://doi.org/10.1145/2939672.2939785

[3] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems (NeurIPS), 30, 3146–3154.

[4] Zou, H. and Hastie, T. 2005. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society: Series B, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[5] Republic of Indonesia. 2009. Undang-Undang Nomor 28 Tahun 2009 tentang Pajak Daerah dan Retribusi Daerah. Lembaran Negara RI Tahun 2009 Nomor 130. Jakarta: Sekretariat Negara.

[6] Bentejac, C., Csorgo, A., and Martinez-Munoz, G. 2021. A Comparative Analysis of Gradient Boosting Algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[7] Grinsztajn, L., Oyallon, E., and Varoquaux, G. 2022. Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data? Advances in Neural Information Processing Systems (NeurIPS 2022), 35. https://doi.org/10.48550/arXiv.2207.08815

[8] Mayer, M., Bourassa, S. C., Hoesli, M., and Scognamiglio, D. 2022. Machine Learning Applications to Land and Structure Valuation. Journal of Risk and Financial Management, 15(5), 193. https://doi.org/10.3390/jrfm15050193

[9] Zilli, C. A. and Bastos, L. C. 2024. Machine Learning Models in Mass Appraisal for Property Tax Purposes: A Systematic Mapping Study. Aestimum, 84, 31–52. https://doi.org/10.36253/aestim-15792

[10] Sharma, H., Harsora, H., and Ogunleye, B. 2024. An Optimal House Price Prediction Algorithm: XGBoost. Analytics, 3(1), 30–45. https://doi.org/10.3390/analytics3010003

[11] Xu, C. and Kong, Y. 2024. Random Forest Model in Tax Risk Identification of Real Estate Enterprise Income Tax. PLOS ONE, 19(3), e0300928. https://doi.org/10.1371/journal.pone.0300928

[12] Lundberg, S. M. and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.

[13] Apriliyanti, I. D., Kusumasari, B., Pramusinto, A., and Setianto, W. A. 2021. Digital Divide in ASEAN Member States: Analyzing the Critical Factors for Successful E-Government Programs. Online Information Review, 45(2), 440–460. https://doi.org/10.1108/OIR-03-2020-0093

[14] Dwivedi, Y. K., Rana, N. P., Janssen, M., Lal, B., Williams, M. D., and Clement, M. 2023. An Empirical Validation of a Unified Model of Electronic Government Adoption (UMEGA). Government Information Quarterly, 40(2), 101819. https://doi.org/10.1016/j.giq.2022.101819

[15] Zilli, C. A., Bastos, L. C., and Da Silva, L. R. 2024. Mass Appraisal of Apartments Using Random Forest and Gradient Boosting Algorithms: Case Study of Florianopolis, Brazil. Revista do Departamento de Geografia, 44, e212297. https://doi.org/10.11606/eISSN.2236-2878.rdg.2024.212297

[16] Antipov, E. A. and Pokryshevskaya, E. B. 2023. Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach. The Journal of Real Estate Finance and Economics. https://doi.org/10.1007/s11146-023-09944-1

[17] Yang, L. 2025. Predictive Modeling of Tax Compliance Risks: A Comparative Study of Machine Learning Approaches. PLOS ONE, 20(9), e0331715. https://doi.org/10.1371/journal.pone.0331715

[18] Moreno-Lopez, M., Zofio, J. L., and Gonzalez-Arias, J. 2025. Comparative Analysis of Advanced Models for Predicting Housing Prices: A Review. Buildings, 9(2), 32. https://doi.org/10.3390/urbansci9020032

[19] Indah, Y. M., Sumarni, S., and Widiyastuti, T. 2024. Comparison of Random Forest, XGBoost and LightGBM Methods on the Human Development Index in Indonesia. Jurnal Sains dan Informatika, 10(2), 155–164. https://doi.org/10.34128/jsi.v10i2.651

[20] Yilmaz, A. K. and Kalayci, A. 2024. Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data. In Proceedings of the 2024 IEEE International Conference on Big Data,4521–4527. https://doi.org/10.1109/ICAD65464.2025.11114069

[21] Kepplinger, D. 2023. Robust Variable Selection and Estimation via Adaptive Elastic Net S-Estimators for Linear Regression. Computational Statistics & Data Analysis, 183, 107730. https://doi.org/10.1016/j.csda.2023.107730

[22] Tsagkis, P., Bakogiannis, E., and Nikitas, A. 2023. Analysing Urban Growth Using Machine Learning and Open Data: An Artificial Neural Network Modelled Case Study of Five Greek Cities. Sustainable Cities and Society, 89, 104337. https://doi.org/10.1016/j.scs.2022.104337

[23] Shwartz-Ziv, R. and Armon, A. 2022. Tabular Data: Deep Learning Is Not All You Need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

[24] Firmansyah, R. and Wicaksono, A. 2021. E-Government Implementation and Corruption Reduction: Evidence from Indonesian Local Governments. Asian Journal of Political Science, 29(4), 445–463. https://doi.org/10.1080/02185377.2021.1978143

[25] Shaputra, E., Ginting, B. S., and Nurhayati, N. 2021. Prediksi Pendapatan Asli Daerah (PAD) Kabupaten Langkat Menggunakan Metode Backpropagation Neural Network. JTIK (Jurnal Teknik Informatika Kaputama), 5(1), 69–75. https://doi.org/10.59697/jtik.v5i1.587

[26] Nagappan, M. and Daud, S. M. 2021. Machine Learning Predictors for Sustainable Urban Planning. International Journal of Advanced Computer Science and Applications (IJACSA), 12(7), 457–464. https://doi.org/10.14569/IJACSA.2021.0120754

[27] Republic of Indonesia. 2021. Peraturan Pemerintah Nomor 16 Tahun 2021 tentang Peraturan Pelaksanaan Undang-Undang Nomor 28 Tahun 2002 tentang Bangunan

Gedung. Lembaran Negara RI Tahun 2021 Nomor 26. Jakarta: Sekretariat Negara.

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. (Scopus-indexed; widely cited in ML pipeline implementations)

[29] Robi, R. K. and George, J. K. 2025. Application of Machine Learning Algorithms to Predict Urban Expansion. Journal of Urban Planning and Development, 151(2). https://doi.org/10.1061/JUPDDM.UPENG-5466

[30] Hachcha, F. and Kessentini, H. 2022. Elastic Net Penalized Quantile Regression Model. Journal of Computational and Applied Mathematics, 409, 113800. https://doi.org/10.1016/j.cam.2022.113800