p. 348-363

THE JOURNAL OF INDONESIA SUSTAINABLE DEVELOPMENT PLANNING

VOL. 7 NO. 2 – AUGUST 2026

E-ISSN: 2722-0842 | P-ISSN: 2721-8309

Available online at
http://journal.pusbindiklatren.bappenas.go.id/

Research Paper

Determinants of the Village Development Index (IDM) on Madura Island: A Classical and Machine Learning Comparison

Dita Amelia[1*], Addina Nurkamila[2], Anggita Nariswari[3], and Fadli Akbar Pambudi[4]

[1,2,3]Department of Mathematics, Universitas Airlangga, Surabaya, Indonesia.
[4]Ministry of National Development Planning, Bappenas, Republic of Indonesia, Jakarta Pusat, Indonesia.

***Correspondence author**: dita.amelia@fst.unair.ac.id

Abstract

Development disparities on Madura Island are characterized by an uneven distribution of basic social infrastructure, reflected in persistently low human development indicators and high poverty rates relative to other regencies in East Java. This study aims to identify fiscal and social infrastructure determinants of IDM status and to compare the performance of ordinal logistic regression (OLR), K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM) in classifying IDM status across 62 sub-districts on Madura Island using 2024 data, evaluated under Stratified 10-Fold Cross Validation. K-NN achieved the highest overall accuracy (71.00%), followed by OLR (67.62%) and SVM (50.00%); however, the macro-averaged F1-Score of 0.42 for both K-NN and SVM indicates limited performance on minority classes. OLR reveals that healthcare facilities and equitable school access are significantly associated with IDM status. The models are best positioned as early-screening tools to support more responsive village development policies, with findings relevant to SDGs 1, 8, and 16.

Keywords: Village Development Index; Ordinal Logistic Regression; K-Nearest Neighbors; Support Vector Machine; Madura Island.

ARTICLE INFO
Received: January 10, 2026
Received in revised form: March 10, 2026
Accepted: August 30, 2026

doi: 10.46456/jisdep.v7i2.1036

THE JOURNAL OF INDONESIA SUSTAINABLE DEVELOPMENT PLANNING
Published by Centre for Planners' Development, Education, and Training (Pusbindiklatren), Ministry of National Development Planning/National Development Planning Agency (Bappenas), Republic of Indonesia

Address: Jalan Proklamasi 70, Central Jakarta, Indonesia 10320
Phone: +62 21 31928280/31928285
Fax: +62 21 31928281
E-mail: journal.pusbindiklatren@bappenas.go.id

Supported by Indonesian Development Planners Association (PPPI)

Please cite this article in APA Style as:
Amelia, D., Nurkamila, A., Nariswari, A., & Pambudi, F. A. (2026). Determinants of the Village Development Index (IDM) on Madura Island: A Classical and Machine Learning Comparison. *The Journal of Indonesia Sustainable Development Planning, Vol 7(2),* 348-363. https://doi.org/10.46456/jisdep.v7i2.1036

1. Introduction

Village development is one of the strategic priorities in Indonesia's national development agenda. The government, through the Ministry of Villages and Development of Disadvantaged Regions, has developed the Village Development Index (IDM) as an instrument to measure the progress and self-reliance of villages based on three main dimensions, namely social resilience, economic resilience, and environmental resilience (Kemendesa, 2025). The concept of the Village Development Index (IDM) is explained in the Village Ministry Regulation (Permendes) No. 2 of 2016 concerning the pattern and map of village development. The IDM concept integrates several categories, including the Economic Resilience Index (IKE), the Environmental Resilience Index (IKL), and the Social Resilience Index (IKS) (Sukarno, 2020). Village development strata are assessed against these three dimensions. Beyond categorising villages by development status, the IDM serves as an evidence base for formulating more targeted village development policies. In this way, the IDM indirectly captures improvements in the welfare of rural communities.

The national picture has improved steadily over the past decade. The number of villages holding self-reliance status reached 17,203 in 2024, up from 11,456 the year before, and the share of self-reliance villages met the 22.85% target set in the 2025–2029 National Medium-Term Development Plan (Kementerian Perencanaan Pembangunan Nasional, 2025). East Java Province recorded 4,019 villages with self-reliance status, 2,924 with Progressing status, and 778 with Growing status, with no villages remaining at Underdeveloped status (Satu Data Indonesia, 2024). Of this total, BAPPEDA Jawa Timur, (2025) reports that 2,874 administrative villages, equivalent to 37.22% of all villages in East Java, have achieved self-reliance status, with the difference attributed to the inclusion of urban village in the national figure. Aggregate provincial figures, however, conceal wide differences between the regencies and cities that make up East Java. Among the 38 regencies and cities in the province, the four located on Madura Island consistently occupy the lower ranks. In 2019, the economic growth rate of most Madura regencies fell below the East Java average of 5.52% (Manshur & Suprapto, 2020). The gap extends to human development: the Human Development Index of Sumenep Regency in 2024 stood at 69.78, against 75.35 for the province as a whole (Badan Pusat Statistik Kabupaten Sumenep, 2024). Poverty follows the same pattern, with Sampang Regency at 20.71% and Sumenep Regency at 19.40%, the two highest rates among East Java regencies (Manshur & Suprapto, 2020). Comparing regencies within a single province in this way holds provincial policy constant, since all four Madura regencies operate under the same East Java Provincial Spatial Plan 2011–2031, one mission of which is balanced development between regions (Majida & Handayeni, 2019). The remaining disparity is therefore internal to the province rather than a product of differing provincial regimes.

As an instrument for monitoring village development, the IDM also reveals where development remains uneven, particularly in areas that face persistent structural challenges. One relevant area for study is Madura, which comprises four districts, namely Bangkalan, Sampang, Pamekasan, and Sumenep, which have different socio-economic characteristics compared to other areas in East Java Province. Despite significant progress through the Village Fund, several regencies such as Sampang, Bangkalan, and some sub-districts in Pamekasan still have a proportion of villages with growing or underdeveloped status based on the latest IDM. This pattern points to structural disparities in development equity, particularly in access to education, healthcare, and fiscal support, and motivates an empirical analysis of the factors that drive IDM variation across sub-districts (Majida & Handayeni, 2019).

Village budget allocations through the Village Fund and the Village Fund Allocation (ADD) are the central government's main fiscal instrument for accelerating village economic development and self-reliance. Article 6 of Government Regulation No. 60 of 2014 governs the transfer of these funds from the regency budget (APBD) to the village budget (APB Desa) (Yani, 2023). The funds finance basic infrastructure such as roads, irrigation and sanitation, together with education and health services, all of which feed into the three dimensions of the IDM. Evidence on their effect is mixed. Iftitah & Wibowo (2022), studying Gowa Regency, found that village fund participation in BUMDes capital and village-generated revenue raise the IDM, whereas the village fund allocation itself has no measurable effect. Yusuf and Khoirunurrofik (2022), using village-level data across Indonesia, report that the contribution of village funds to economic development depends on the financial management capacity of the village government. Ahadis et al. (2025) reach a comparable conclusion in North Lampung Regency, where the village fund allocation shapes the IDM only in interaction with poverty and demographic conditions. Read together, these studies indicate that both the absorption and the effect of village budget allocations vary across regions, driven by differences in village apparatus capacity, supervision quality and local management priorities.

Education is an important aspect of the social resilience dimension of the IDM because it is a component of the Social Resilience Index, alongside health and social capital. The availability of adequate schools improves community access to education, thereby helping to build quality human resources and supporting social and economic participation through improved skills and work capacity (Alhumami, 2025). In addition, the number of villages with school facilities reflects the level of equitable access to education between regions, which plays a role in reducing educational development disparities by providing more equal opportunities to all levels of society in different areas (Sugiarto, 2025). The Human Development Index has been demonstrated to be positively and significantly impacted by government spending on the health sector, which can increase public access to and the quality of health care by improving facilities and the availability of medications (Asnidar et al., 2025). Thus, it is hypothesized that the inclusion of health and education infrastructure as factors in this study will play a strategic role in raising the IDM value by promoting social resilience and sustainable human development (Tarigan & Saharuddin, 2025).

In addition to providing a venue for religious activity, the presence of suitable facilities such as houses of worship fosters social connection and unity in rural communities (Prasetyo et al., 2024). Through various collective activities, religious facilities contribute to the formation of trust and social cohesion as part of the community's social capital (Rinanda et al., 2025). Strong social cohesion promotes social stability and increases the capacity of villages to face development challenges, including poverty and regional inequality (Rojak & Fajar, 2025). Therefore, the number of places of worship can serve as a quantitative indicator of the strength of social capital in supporting the progress and self-reliance of villages on Madura Island.

Beyond fiscal variables, empirical work has examined the wider determinants of the IDM. Arina et al. (2021) and Akbar et al. (2024) show that village funds, village fund allocations and village revenue are positively associated with the IDM, confirming that fiscal capacity is one determinant of village development status (Yusuf & Khoirunurrofik, 2022). Naufal et al. (2023) Analyze the role of social infrastructure in rural poverty using the IDM as a moderating variable, while Muhtarom et al. (2018) compare the IDM with the Rural Development Index (IPD) as alternative measures of village progress.

Three gaps remain in this body of work. First, many studies still rely on the Rural Development Index (IPD), which is narrower in scope than the IDM, a composite designed to capture interrelated social, economic, and ecological resilience (Muhtarom et al., 2018). Second, fiscal and social variables are rarely tested simultaneously within one model; Naufal et al. (2023), for instance, examine social infrastructure and the IDM without entering fiscal capacity into the same specification, which leaves their relative weight unknown. Third, the unit of analysis is almost always the village or the regency, so the sub-district level at which regency governments actually distribute technical assistance and supervision, and which sits between the granularity of village-level IDM research and the aggregation of regency-level fiscal analysis, remains thinly covered. A methodological gap runs alongside these. IDM studies commonly apply linear regression, panel regression, or structural equation modeling (Iftitah & Wibowo, 2022; Akbar et al., 2024), yet IDM status is an ordinal variable, for which ordinal logistic regression is the appropriate specification. Machine learning classifiers such as SVM and K-NN have seen only limited use in classifying village development status, with Azies, (2024) among the few applications in the Indonesian context.

Against this literature, the contribution of this study can be stated in three points, ordered by weight. The primary contribution is methodological alignment: IDM status is treated as what it is, an ordinal variable, and is modeled with ordinal logistic regression rather than with the linear or panel regression commonly applied to the IDM composite score, so that the estimated effects speak directly to the probability of a sub-district occupying a higher development category. Supporting this, the study works at the sub-district level for all 62 sub-districts on Madura Island, a level between the village-level studies dominating IDM research and the regency-level studies used for fiscal analysis, and one that corresponds to the level at which regency governments allocate technical assistance and supervision. Third, fiscal capacity and social infrastructure enter the same specification simultaneously, which allows their relative weight to be compared rather than inferred from separate models estimated on different samples. A substantive finding emerging from this design is that the equitable distribution of basic social infrastructure, particularly healthcare facility access and village-level school coverage, is more strongly associated with IDM status than the volume of village budget allocation, a result that challenges fiscal-centered policy narratives and cannot be obtained from studies that enter these variables in separate models.

The role of machine learning in this design needs to be stated precisely, because the two approaches answer different questions and are used sequentially rather than competitively. Ordinal logistic regression is an explanatory instrument: it estimates the direction and magnitude of each determinant through odds ratios and tests their significance, which is what a policy maker requires when deciding where to allocate resources. K-NN and SVM are predictive instruments: they assume no distribution and impose no functional form, so they can capture non-linear boundaries between IDM categories, but they return no interpretable coefficients. The classical model therefore identifies which determinants matter and by how much, while the machine learning classifiers test whether that same set of predictors is sufficient to reproduce IDM status out of sample under stratified cross-validation. Agreement between the two indicates that the determinants identified by the conventional assessment are robust and not an artifact of its parametric assumptions; disagreement signals that the ordinal structure or the class distribution imposes a limit the conventional assessment alone would not reveal. In operational terms, ordinal logistic regression explains the IDM status a sub-district currently holds, whereas the trained classifier can be re-run on updated administrative data to flag sub-districts at risk of stagnating before the next annual IDM evaluation is published.

Madura provides a suitable setting for this design. The island is separated from the East Java mainland by the Madura Strait, a physical condition that has shaped its development trajectory (Heriqbaldi et al., 2025). The Suramadu Bridge improved connectivity, yet limited internal accessibility and the uneven distribution of infrastructure mean that per capita GRDP and HDI in the four Madura regencies remain below the East Java average (Fauzi et al., 2019). The findings of this study are also relevant to national efforts toward the Sustainable Development Goals, in particular Goal 1 on poverty reduction, Goal 8 on decent work and economic growth, and Goal 16 on strong and accountable institutions. This study therefore pursues two objectives: to identify the fiscal and social determinants of IDM status across all 62 sub-districts on Madura Island, and to compare the classification performance of ordinal logistic regression with K-NN and SVM under stratified 10-fold cross validation, a resampling scheme chosen to give a robust performance estimate given the small sample ($n = 62$).

2. Methods

This study uses a quantitative approach by comparing the classical method, namely ordinal logistic regression, which aims to interpret the influence of each variable, and methods with a machine learning approach, namely K-Nearest Neighbours (K-NN) and Support Vector Machine (SVM).

2.1. Analysis of Determinants and Classification of IDM using a Classical Approach

In modelling with ordinal logistic regression, a classic assumption is required, namely a multicollinearity test, which is used to detect strong linear relationships between predictor variables in a regression model. In general, the criteria commonly used in research indicate that if $VIF > 10$ or tolerance $< 0{,}10$ indicates serious multicollinearity, $5 < VIF \leq 10$ can be interpreted as moderate multicollinearity, $VIF \leq 5$ is generally considered to be within acceptable limits. If all variables have tolerance values $\geq 0{,}10$ and $VIF \leq 10$, it can be concluded that there are no significant multicollinearity issues in the model, so parameter estimates remain efficient and the analysis results can be interpreted validly.

Following the multicollinearity check, ordinal logistic regression was estimated to analyze the relationship between the ordinal response variable and the predictor variables (Hosmer et al., 2013). The logit model, a cumulative logit model, is the model utilized in ordinal logistic regression. Suppose there is a random sample from a joint distribution $(Y, \boldsymbol{X})$ where Y is an ordinal response and $\boldsymbol{X} = (X_1, X_2, \dots, X_P)$ is a vector of predictor variables. If $\pi_k(\boldsymbol{X}) - P(Y = k|\boldsymbol{X})$ is the probability that $Y = k$ if X is known, for $k = 1,2, \dots, q$ and $\gamma_k(\boldsymbol{X}) = P(Y \leq k|\boldsymbol{X})$ is the cumulative probability of the response variable category k

for $k = 1,2, \dots, q-1$. Relationship between probability vectors $\pi = (\pi_1(\boldsymbol{X}), \pi_2(\boldsymbol{X}), \dots, \pi_q(\boldsymbol{X}))$ with ordinal response variables expressed in the logit model as follows.

$$g(\gamma_k(\boldsymbol{X})) = \theta_k + \sum_{j=1}^{p} \beta_j X_j \,; k = 1,2, \dots, q-1$$

Where θ is the intercept parameter vector and $\boldsymbol{\beta}' = (\beta_1, \beta_2, \dots, \beta_t)$. Then the probability value for each category when there are three response variables $k = 1,2,3$ is as follows:

$$\pi_1(\boldsymbol{X}) = P(Y \le 1|\boldsymbol{X}) = \gamma_1(\boldsymbol{X}) = \frac{\exp\,(\theta_1 + \sum_{k=1}^{p} \beta_k X_k}{1 + (\exp(\theta_1 + \sum_{k=1}^{p} \beta_k X_k))}$$

$$\pi_2(\boldsymbol{X}) = P(Y \le 2|\boldsymbol{X}) - P(Y \le 1|\boldsymbol{X}) = \frac{\exp\,(\theta_2 + \sum_{k=1}^{p} \beta_k X_k}{1 + (\exp(\theta_2 + \sum_{k=1}^{p} \beta_k X_k))} - \frac{\exp\,(\theta_1 + \sum_{k=1}^{p} \beta_k X_k}{1 + (\exp(\theta_1 + \sum_{k=1}^{p} \beta_k X_k))}$$

$$\pi_3(\boldsymbol{X}) = 1 - \pi_1(\boldsymbol{X}) - \pi_2(X) = 1 - P(Y \le 2|\boldsymbol{X}) = \frac{1}{1 + (\exp(\theta_2 + \sum_{k=1}^{p} \beta_k X_k))}$$

In ordinal logistic regression, a parameter significance test is also conducted using the likelihood ratio test. The parameter significance test aims to determine the effect of each variable. There are two parameter significance tests, namely the simultaneous test using G statistics and the individual test using the Wald test. The odds ratio quantifies the change in the probability of an observation falling in a given category for a one-unit increase in a predictor variable. The calculation of odds values in the ordinal logit model for category $Y \le j$ is as follows.

$$Odd = \frac{P(Y \le j|\boldsymbol{X})}{P(Y > j|\boldsymbol{X})} = \exp(\theta_j + \boldsymbol{X\beta})\,;\ j = 1,2, \dots q-1$$

Based on the Odds value in the equation, the Odds Ratio (OR) value of the predictor variable X_k in category $Y \le j; j = 1,2, \dots, q-1$ is as follows:

$$OR_k = \frac{\dfrac{P(Y \le j|X_k = x_k + 1)}{P(Y > j|X_k = x_k + 1)}}{\dfrac{P(Y \le j|X_k = x_k)}{P(Y > j|X_k = x_k)}} = \exp(\beta_k)$$

The interpretation of the OR value is as follows (Sharma et al., 2020).

1. $OR < 1$ indicates that exposure to odds potentially reduces the occurrence of the group being studied.
2. $OR = 1$ indicates that exposure to odds does not cause the outcome in the group being studied.
3. $OR > 1$ indicates that exposure to odds potentially increases the occurrence of the group being studied.

The sign convention above follows the standard cumulative logit formulation. However, the ordinal logistic regression model in this study was estimated using the *OrderedModel function,* which is based on a latent variable formulation, $P(Y \le k) = F(\tau_k - X\beta)$ where an increase in $X\beta$ decreases $P(Y \le k)$. In the logit model, the odds ratio describes the change, either an increase or decrease, in the tendency for each increase of one unit of the predictor variable X_k, if the predictor variable is continuous, or the difference in tendency between categories if the predictor variable is categorical (Budyanra & Azzahra, 2017).

2.2. IDM Classification Using a Machine Learning Approach

In this study, the classification of the Village Development Index (IDM) status was carried out using a machine learning approach with two algorithms: K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM). Both methods were used to identify classification patterns based on predictor variables and to compare model performance to determine the optimal method for classifying IDM status.

2.2.1 K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) is a supervised learning classification technique classified as instance-based learning or lazy learning because it saves all training data for use during prediction rather than explicitly constructing a parametric model (Bhatia & Vandana, 2010). K-NN selects a number of k nearest neighbors based on a certain distance metric after computing the distance between a test observation and all training data. Assuming that objects near each other in feature space typically have similar attributes or labels, the majority label of the k nearest neighbors is used to identify the class of the test observation. The non-parametric K-NN technique does not rely on any particular distributional assumptions, but its performance is influenced by the number of samples and feature dimensions. In high-dimensional data, the curse of dimensionality phenomenon can reduce the accuracy of inter-observation distances (Rimanic et al., 2020). The proximity measure in K-NN is determined using a distance function. Three distance measures are commonly used: Euclidean distance, Manhattan distance, and Minkowski distance (Mahapatra & Chakraborty, 2015).

Table 1. Mathematical Equations of Distance Types in the K-NN Method

Distance Type	Formula
Euclidean Distance	$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
Manhattan Distance	$d(x,y) = \sum_{i=1}^{p}\lvert x_i - y_i\rvert$
Minkowski Distance	$d(x,y) = \left(\sum_{i=1}^{p}(\lvert x_i - y_i\rvert)^q\right)^{1/q}$

Minkowski distance is a generalization of Manhattan distance (for $q = 1$) and Euclidean distance (for $q = 2$). The selection of the k value is a crucial aspect of the K-NN method. A value of k that is too small can cause the model to be sensitive to noise and local variations, thereby increasing the model's variance. Conversely, a value of k that is too large can result in overly general classification boundaries and increase bias. K-NN has the advantages of simple implementation, the ability to handle non-linear decision boundaries, and no data distribution assumptions. However, this method has high computational complexity because distance calculations are performed on all training data of order $O(n)$ per prediction, and it is sensitive to variable scale, thus requiring data normalization or standardization.

The classification of IDM status using the K-NN algorithm was used to classify IDM status in several processing stages. Before modeling, all predictor variables were standardized using StandardScaler to avoid bias from scale differences between variables, given that K-NN is highly sensitive to data scale. To prevent data leakage, the scaler was fitted exclusively on the training fold within each cross-validation iteration using a Pipeline. To ensure a directly comparable evaluation against SVM, K-NN hyperparameter tuning was conducted using the same nested cross validation protocol: an outer Stratified 10-fold loop for performance estimation and an inner Stratified 3-fold loop for GridSearchCV, searching over $k \in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$ combined with Euclidean, Manhattan, and Minkowski (p = 2, 3) distance metrics. The most frequently selected configuration across the 10 outer folds was k = 9 with the Euclidean distance metric, yielding a mean outer-fold accuracy of 63.10% with a macro-averaged F1-Score of 0.454. All analyses were implemented in Python using scikit-learn with random_state = 42.

2.2.2 Support Vector Machine (SVM)

Cortes & Vapnik (1995) The Support Vector Machine (SVM) classification technique is introduced, which is based on the structural risk minimization (SRM) principle of statistical learning theory. The primary objective of SVM is to find the optimal separating hyperplane, that is, the plane that maximizes the margin between data classes. Support vectors, or data points near the margin's edge, are crucial in establishing the hyperplane's position. Mathematically, the linear separating function is defined as $f(\boldsymbol{x}) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + b$,, where $\boldsymbol{w}$ is the weight vector, $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ is the input feature vector, and b is the bias. The hyperplane is chosen such that it maximizes the margin between the positive and negative classes, which can be geometrically measured as $\frac{2}{\|\mathbf{w}\|}$. In the case of perfect separation, the positive class data satisfies $\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + b = 1$ and the negative class data satisfies $\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + b = -1$.

Data are frequently not linearly separable in practice. SVM employs the kernel method to address this by mapping data to a higher-dimensional space without explicitly computing the transformation. In this study, the linear kernel was selected because it provides a more stable and interpretable decision boundary for small-sample, imbalanced datasets (n = 62), and consistently produced the best inner cross-validation accuracy across all outer folds.

Table 2. Mathematical Equations of the Linear Kernel Function Used in SVM

Kernel Type	Formula
Linear	$K(x_i, x) = x_i'x$
Polynomial	$K(x_i, x) = (\gamma \cdot x_i'x + r)^p$
RBF	$K(x_i, x) = \exp\left(-\gamma \lvert\lvert x - x_i'x \rvert\rvert^2\right)$
Sigmoid	$K(x_i, x) = \tanh(\gamma \cdot x_i'x + r)$

For the linear kernel, the key parameter is C, the regularisation term that controls the trade-off between maximizing the margin and minimizing classification errors. A smaller value of C enforces stronger regularisation, which is preferred when the sample size is limited. In this study, C was optimized via GridSearchCV with an inner 3-fold cross-validation over the candidate set {0.001, 0.005, 0.01, 0.05}. The most frequently selected value of C across the 10 outer folds was $C = 0.05$, reflecting a preference for moderate regularisation whilst maintaining sufficient margin flexibility given the limited sample size ($n = 62$).

The classification decision on the test data was made using a combination of support vectors and Lagrange weights, so that the decision function was written as:

$$f(x_t) = \sum_{s=1}^{ns} a_s y_s x_s \cdot x_t + b$$

with x_t as the feature vector in the test data, x_s as the selected support vector, α_s as the Lagrange multipliers from quadratic programming optimization, and y_s as the class label of the support vector. In the context of multi-class classification, SVM was extended using the one-vs-one strategy, the default multi-class approach in scikit-learn's SVC implementation, which selects the class with the highest number of votes across all pairwise classifiers (Hsu et al., 2003).

Additionally, class_weight='balanced' was applied in the SVM model to address the unequal class distribution in the IDM dataset (Progressing: 62.90%; Growing: 33.90%; Self-Reliance: 3.20%). This setting automatically adjusts the penalty parameter C proportionally to the inverse of class frequencies, thereby preventing the model from being biased towards the majority class. Unlike SVM, ordinal logistic regression does not require an explicit class-balancing mechanism, as it estimates cumulative probabilities across ordinal categories and is inherently less sensitive to class frequency imbalance than margin-based classifiers; the class distribution is therefore reflected naturally in the estimated intercept thresholds rather than requiring reweighting of the likelihood function. To avoid optimistic bias in performance estimation, hyperparameter tuning for SVM was conducted using nested cross-validation, with an outer 10-fold loop for performance evaluation and an inner 3-fold loop for GridSearchCV, ensuring that model selection and performance assessment were performed on independent data partitions.

2.3. Data and Data Sources

The secondary data utilized in this investigation, obtained at the sub-district level, came from the official websites of the Ministry of Villages and Development of Disadvantaged Regions and the Central Statistics Agency. The data analyzed were the Village Development Index in four regencies on Madura Island, namely Bangkalan Regency, Sampang Regency, Pamekasan Regency, and Sumenep Regency in 2024, with a total of 62 observations. Districts that were not included were excluded from the analysis due to missing data for one of the research variables, rendering them unsuitable for the model estimation process. This exclusion is acknowledged as a research limitation, as it could introduce selection bias if the excluded districts possess systematic characteristics that differ from those retained in the analysis. Based on the 2024 IDM classification, the status of villages on Madura Island is divided into three categories: Growing, Progressing, and Self-Reliance. Thus, the response variable in this study is ordinal, whilst the predictor variables consist of five variables believed to influence the IDM score, as presented in Table 3.

Table 3. Research Variables

Response Variable	Scale	Description	Data Sources
Village Development Index (Y)	Ordinal	1. Growing: 0.600 – 0.707 2. Progressing: 0.708 – 0.815 3. Self-reliance: > 0.815	(Kementerian Desa Pembangunan Daerah Tertinggal, 2024)

Predictor Variable	Scale	Unit	Data Sources
Village Budget Allocation (X_1)	Ratio	Millions of Rupiah	(Badan Pusat Statistik Kabupaten Pamekasan, 2025; Badan Pusat Statistik Kabupaten Sampang, 2025; Badan Pusat Statistik Kabupaten Bangkalan, 2025; Badan Pusat Statistik Kabupaten Sumenep, 2025)
Number of Schools (X_2)	Ratio	Unit	
Number of Healthcare Facilities (X_3)	Ratio	Unit	
Number of Places of Worship (X_4)	Ratio	Unit	
Number of Villages with School Facilities (X_5)	Ratio	Village	

The collection of variables was the first step in this study, followed by data preprocessing to detect missing values, normalization, and data partitioning using Stratified 10-fold cross-validation, in which the data were divided into ten subsets whilst preserving the class proportion in each fold. Each fold served alternately as the test set whilst the remaining nine folds were used for training, ensuring that every observation was tested exactly once and that the minority class (self-reliance, n = 2) was represented across all folds. The data were then explored through descriptive analysis and tested for multicollinearity assumptions using VIF to ensure no high linear relationships between independent variables. The IDM category modeling on Madura Island in 2024 was carried out using ordinal logistic regression, including the calculation and interpretation of the Odds Ratio, which aimed to determine the magnitude of the predictor variables' influence on the response variables and to calculate the percentage of classification accuracy using the APPER value as a measure of model performance. In addition, classification was also modeled using the K-Nearest Neighbor (K-NN) algorithm with the determination of the k value, as well as Support Vector Machine (SVM) with kernel selection and training the model to form the optimal separating hyperplane. Finally, the classification results from ordinal logistic regression, K-NN, and SVM were compared to determine the method with the best performance in classifying the IDM category.

3. Results and Discussions

3.1. Overview of Data Characteristics

As a first step in the analysis, the distribution of IDM status across all sub-districts on Madura Island was examined to determine the proportion of each category, as shown in Table 4.

Table 4. IDM Status Distribution

Category	Number	Percentage
Growing	21	33.87%
Progressing	39	62.90%
Self-reliance	2	3.23%
Total	62	100%

According to Table 4, all sub-districts with 'Self-reliance' status are located in Sumenep Regency, namely Sumenep City and Gapura sub-districts, indicating that urban centers and areas with better public service facilities tend to achieve a higher IDM status. Conversely, sub-districts with a 'Growing' status are concentrated in Bangkalan and Sampang Regencies, where almost all sub-districts in Bangkalan Regency remain in the 'Growing' category, such as Kokop, Galis, and Klampis sub-districts. This situation indicates that although the Suramadu Bridge has improved connectivity, its benefits have not been evenly distributed across the entire Madura region, as evidenced by the low IDM status in Bangkalan Regency, which is actually the closest to the bridge. This underscores that physical accessibility alone is insufficient without the strengthening of social infrastructure and village institutions. This pronounced class imbalance, particularly the extremely small proportion of self-reliance sub-districts ($\boldsymbol{n}$ = 2, 3.23%), has direct implications for model training and evaluation, and is addressed in the classification modeling stage through the application of class_weight='balanced' in the SVM model and Stratified 10-fold cross-validation to ensure proportional class representation across all folds.

Table 5. Characteristics of IDM on Madura Island and the Predictive Variables that Influence It

Variable	N	Mean	Variance	Minimum		Maximum	
				Value	Subdistrict	Value	Subdistrict
X_1	62	514.70	1531215	3.83	Karang Penang	9003.82	Banyuates
X_2	62	141.31	3639.30	38	Nonggunong	285	Palengaan
X_3	62	6.81	26.32	1	Kadur	31	Bangkalan
X_4	62	139.27	18273.10	8	Masalembu	630	Tlanakan
X_5	62	30.82	211.76	10	Kamal	63	Bluto

Based on Table 5, variations in characteristics between sub-districts on Madura Island reflect a fairly striking disparity in development compared to the East Java average. Fiscal disparities in village budget allocations ($\boldsymbol{X}_1$) indicate an uneven distribution across sub-districts, which has implications for villages' ability to provide basic services to their communities. This situation is exacerbated by the uneven distribution of educational infrastructure, as reflected in the variation in the number of schools ($\boldsymbol{X}_2$) and the number of villages with schools ($\boldsymbol{X}_5$), which contributes to Madura's low HDI, with Sumenep's HDI standing at just 69.78, well below the East Java average of 75.35 (Badan Pusat Statistik Kabupaten Sumenep, 2024). From a healthcare perspective, a number of remote sub-districts have only one healthcare facility ($\boldsymbol{X}_3$). This situation reflects the limited coverage of basic healthcare services in the Madura region, which has a direct impact on the quality of life and productivity of the rural population. On the other hand, the wide variation in the number of places of worship ($\boldsymbol{X}_4$) in fact reflects the strength of religious social capital, which is a distinctive feature of Maduran culture and has the potential to serve as an endogenous force in fostering social cohesion and community participation in village development.

3.2. Analysis of Determinants and Classification of IDM using the Classical Approach

Ordinal logistic regression was chosen because the dependent variable (IDM status) consists of three ordered categories, allowing the identification of factors that influence IDM on Madura Island. Prior to parameter estimation, a multicollinearity test was conducted to ensure the absence of high correlation among the independent variables. The results indicate that the Variance Inflation Factor (VIF) values for all predictor variables are below the threshold of 10, suggesting no serious multicollinearity among the independent variables in the model. Given the relatively small sample size (n = 62), a significance level of α = 10% was adopted to mitigate the reduced statistical power associated with small-sample inference. This is further supported by the view of Doan (2004), who states that social science research can employ the α level more flexibly, depending on several factors, such as the type and objectives of the study. Nevertheless, this constitutes one of the limitations of the study because it increases the risk of a Type I error compared to the conventional 5% threshold. The parameter estimates are presented in Table 6 as follows.

Table 6. Parameter Estimation Results Using Ordinal Logistic Regression

Predictors	Estimate	SE Estimate	Wald	Sig.	Decision
θ_1	0.62	0.88	0.50	0.48	
θ_2	1.66	0.20	68.95	0.00	
β_1	-0.00	0.00	0.62	0.43	Not significant
β_2	-0.01	0.01	2.30	0.13	Not Significant
β_3	0.14	0.07	3.60	0.06	Significant
β_4	-0.01	0.00	3.10	0.08	Significant
β_5	0.11	0.04	9.40	0.00	Significant

Based on Table 6, the following logit model was obtained.

$$g(\gamma_1(\boldsymbol{X})) = 0.62 - 0.00X_1 - 0.01X_2 + 0.14X_3 - 0.00X_4 + 0.11X_5$$
$$g(\gamma_2(\boldsymbol{X})) = 1.66 - 0.00X_1 - 0.01X_2 + 0.14X_3 - 0.01X_4 + 0.11X_5$$

From the resulting logit model, the following logistic regression equation is obtained.

$$\hat{\pi}_1 = P(\boldsymbol{Y} \leq 1|\boldsymbol{X})$$
$$= \frac{\exp\,(0.62 - 0.00X_1 - 0.01X_2 + 0.14X_3 - 0.01X_4 + 0.11X_5\,)}{1 + \exp\,(0.62 - 0.00X_1 - 0.01X_2 + 0.14X_3 - 0.01X_4 + 0.11X_5\,)}$$
$$\hat{\pi}_2 = P(\boldsymbol{Y} \leq 2|\boldsymbol{X}) - P(\boldsymbol{Y} \leq 1|\boldsymbol{X})$$
$$= \frac{\exp(1.66 - 0.00X_1 - 0.01X_2 + 0.14X_3 - 0.01X_4 + 0.11X_5)}{1 + \exp(1.66 - 0.00X_1 - 0.01X_2 + 0.14X_3 - 0.01X_4 + 0.11X_5)}$$

$$-\frac{\exp(0.62-0.00X_1-0.01X_2+0.14X_3-0.01X_4+0.11X_5)}{1+\exp(0.62-0.00X_1-0.01X_2+0.14X_3-0.01X_4+0.11X_5)}$$
$$\hat{\pi}_3 = 1-P(\boldsymbol{Y}\le 2|\boldsymbol{X})$$
$$=\frac{1}{1+\exp(1.66-0.00X_1-0.01X_2+0.14X_3-0.01X_4+0.11X_5)}$$

After formulating the ordinal logistic regression equation, we proceeded to test the significance of the model, both simultaneously and individually. The results of the simultaneous test are presented in Table 7 below.

Table 7. Results of the Simultaneous Test of the Ordinal Logit Model

df	Chi-Square	Sig.
5	70.08	0.00

Based on Table 7, the decision is to reject H_0 because $G = 70.08 > \chi^2_{(0.1;5)} = 9.24$ and the $P-value = 0.00 < 0.10$. From this, it can be concluded that there is a combined effect of all variables used on IDM status on Madura Island. Additionally, significance tests were also conducted individually using the Wald test. Based on Table 6, the variables number of health facilities (X_3), number of places of worship (X_4), and number of villages with school facilities (X_5) have a major impact on Madura Island's IDM status, whereas the budget allocation variable (X_1) and number of schools (X_2) does not have a significant individual effect. The finding that the budget allocation variable (X_1) does not have a significant individual effect contradicts the results of previous studies by Arina et al. (2021), Akbar et al. (2024), and Iftitah & Wibowo (2022), which indicated that this variable significantly influences the IDM. This discrepancy may stem from several factors, including the use of the sub-district level as the unit of analysis—which can obscure fiscal variations between villages that would otherwise be observable at the village level—and the possibility that IDM status is driven more by the quality and effectiveness of budget utilization than by the sheer magnitude of the budget. It should be noted, however, that X_3 and X_4 are significant only at α = 10%; at the conventional α = 5% threshold, only X_5 remains significant. As X_3 is central to this study's findings, this result should be interpreted with caution, as it may partly reflect limited statistical power from the small sample size (n = 62) rather than a robust effect.

Table 8. Brant Test for the Proportional Odds Assumption

Variables	Chi-Square	df	Sig.
X_1	0.00	1	1.00
X_2	0.00	1	1.00
X_3	0.00	1	1.00
X_4	0.00	1	1.00
X_5	0.00	1	1.00
Omnibus	0.00	5	1.00

Prior to evaluating the overall goodness of fit, the proportional odds assumption underlying the ordinal logistic regression model was examined using the Brant test. As shown in Table 8, the omnibus test yielded a Chi-Square value of 0.00 with 5 degrees of freedom and a p-value of 1.00, which is greater than the α = 0.10 threshold. This result indicates that the null hypothesis of proportional odds is not rejected, and the same conclusion holds for each individual predictor (X_1, X_2, X_3, X_4, and X_5), all of which returned p-values of 1.00. These findings confirm that the proportional odds assumption is satisfied across all variables, supporting the appropriateness of the ordinal logistic regression model used in this study. In addition to the proportional odds assumption, the overall goodness of fit of the model was also assessed, with the results presented in Table 9 below.

Table 9. Goodness of Fit

	Chi-Square	Df	Sig.
Lipsizt	7.45	18	0.99

Based on Table 9, the Chi-Square value obtained using the Lipsizt method is 7.45 with 18 degrees of freedom, indicating that $7.45 < \chi^2{}_{(0.1;18)} = 25.99$. Furthermore, the significance value of 0.99 is greater than 0.10, leading to the conclusion that the obtained logit model is suitable for use. The results of the IDM status analysis on Madura Island can be interpreted by examining the Odds Ratio (OR) values of the significant independent variables in the logit model. Based on the results of the ordinal logistic regression modeling performed, the odds ratio values for each variable are presented in Table 10 as follows.

Table 10. Odds Ratio Value

Variable	X_3	X_4	X_5
Odds Ratio	$e^{0.139} = 1.15$	$e^{-0.005} = 0.99$	$e^{0.107} = 1.11$

Based on Table 10, if the number of health facilities (￼) increases by one unit, the odds that a region's IDM status falls into the same or a higher group increase by a factor of 1.15 compared with falling into a lower category. This is supported by previous research showing that the number of health facilities plays a key role in improving public well-being (Tambaip et al., 2023). If the variable for the number of places of worship (X_4) increases by one unit, the probability that a region falls into the same or a higher group decreases by a factor of 0.99; that figure is nearly equal to 1, indicating that its substantive impact on the IDM can be disregarded. A one-unit increase in X_5 is associated with 1.11 times higher odds of the sub-district falling into the same or a higher IDM category, relative to a lower category, holding other variables constant. This provides empirical evidence that an increase in the number of health facilities in a specific region improves the quality of life for the local community. The number of school facilities in each subdistrict does not have the same significance as the number of villages with schools. This indicates that an increase in the number of villages with school facilities reflects the level of availability and equitable access to education in Madura; that is, the more villages that have school facilities, the easier it will be for the community to access education, which ultimately improves the quality of human resources.

To evaluate the accuray of the model's classification, a confusion matrix was used, which compares actual categories with predicted categories. The classification table obtained from the analysis results is presented in Table 11 below.

Table 11. Classification Performance Table for the Ordinal Logit Model

Fold	1	2	3	4	5	6	7	8	9	10
Accuracy	0.71	0.71	0.67	0.67	0.83	0.33	1.00	0.67	0.50	0.67
Average	0.68									

Table 11 shows that the average classification performance is 0.68, with subset 7 having the best classification performance, achieving an accuracy of 1.00. Additionally, Table 12 presents a confusion matrix indicating that subdistricts with a "Growing" status are less sensitive to the minority class due to the dominance of the "Progressing" category. Furthermore, the "Self-reliance" category has not been able to learn the overall data pattern due to the small data distribution.$(n = 2)$. This results in the model's attempt to correctly classify that class actually causing misclassification in the more frequently occurring classes, thereby lowering the overall accuracy. These results highlight a fundamental limitation: with only two observations in the self-reliance class, neither model can adequately learn the decision boundary, regardless of the balancing strategy used.

Table 12. Classification Report for the Ordinal Logit Model

Fold	**Precision**	**Recall**	**F1-Score**	**Support**
Growing	0.58	0.52	0.55	21
Progressing	0.72	0.79	0.76	39
Self-reliance	0.00	0.00	0.00	2
Accuracy		0.68		

3.3. IDM Classification Using a Machine Learning Approach

Following the nested cross-validation protocol described in Section 2.2, both K-NN and SVM were trained and evaluated, and the complete classification results are presented in Table 13.

Table 13. Classification results for IDM using K-NN and SVM

	Category IDM	**Precision**	**Recall**	**F1 Score**	**Support**
K-NN	Growing	0.78	0.33	0.47	21
	Progressing	0.70	0.95	0.80	39
	Self-reliance	0.00	0.00	0.00	2
	Accuracy		0.71		

	Category IDM	Precision	Recall	F1 Score	Support
SVM	Growing	0.45	0.95	0.62	21
	Progressing	0.83	0.26	0.39	39
	Self-reliance	0.17	0.50	0.25	2
	Accuracy			0.50	

Based on the results of the IDM classification above, the K-NN model demonstrated limitations in identifying sub-districts with a 'Growing' status, indicating that this algorithm is less sensitive to the minority class due to the dominance of the 'Progressing' class in the data. This situation is a direct consequence of the unbalanced class distribution (Progressing: 62.90%; Growing: 33.90%; Self-Reliance: 3.23%), whereby K-NN tends to be biased towards the majority class when determining nearest neighbours. Overall, the K-NN model achieved an accuracy of 71.00% averaged across 10-fold. The SVM model achieved an overall accuracy of 50.00% averaged across 10-fold, which is lower than both K-NN (71%) and ordinal logistic regression (67.62%). To mitigate class imbalance, the SVM model was trained with class_weight='balanced', which automatically adjusts the penalty parameter C proportionally to the inverse of class frequencies. Despite the linear kernel combined with strong regularisation (small C) being theoretically well-suited for small-sample datasets, this balancing strategy caused the model to allocate disproportionately greater weight to the self-reliance class (n = 2, 3.23%). As this class is too rare to be learned reliably, the model's attempts to correctly classify it resulted in misclassifications in the more frequent classes, thereby reducing overall accuracy. This result highlights a fundamental limitation: with only two observations in the self-reliance class, neither model can learn its decision boundary adequately, regardless of the balancing strategy applied.

The finding that K-NN outperforms SVM in overall accuracy is consistent with Azies (2024), who demonstrates that machine learning approaches, particularly instance-based methods, can effectively classify village development status in the Indonesian context. However, the macro F1-Score results (K-NN: 0.42; SVM: 0.42) underscore the challenge of classifying imbalanced ordinal data, a limitation acknowledged in the broader machine learning literature as the "class imbalance problem" (He & Garcia, 2009). Future studies may address this limitation through expanding the dataset to a multi-year panel or to all sub-districts in East Java to obtain a sufficient number of self-reliance observations, or by collapsing the response variable into a binary outcome to reduce the class imbalance problem.

It should be noted that the self-reliance class (n = 2) was too rare to be reliably learned or evaluated by either model. Precision, Recall, and F1-Score for this class are reported as 0.00 for K-NN and near-zero for SVM across all folds, and these results must be interpreted with caution. The performance of both models on the self-Reliance category cannot be generalised and represents a key limitation of this study attributable to the extreme class imbalance in the dataset.

3.4. A Comparison of Model Performance in Classifying the Status of the Village Development Index

A comparison of model performance was conducted based on the accuracy scores obtained in order to determine the approach with the greatest classification performance for identifying the state of the Village Development Index (IDM). All classification models were evaluated using Stratified 10-fold cross validation to ensure robust and unbiased performance estimation across the entire dataset of 62 observations. Table 14 displays the outcomes of the performance evaluation comparison for every classification technique.

Table 14. Comparison of IDM Classification Results

Model	Accuracy
Classification of IDM using the Classical Approach	67.62%
Classification of IDM using K-NN	71.00%
Classification of IDM using SVM	50.00%

As shown in Table 14, the K-NN model achieved an overall accuracy of 71.00%, exceeding ordinal logistic regression (67.62%) and SVM (50.00%). However, given the strongly imbalanced class distribution, overall accuracy alone may be misleading — K-NN's macro-averaged F1-Score of 0.42 reflects limited performance on the minority classes, as evidenced by its low recall for 'Growing' (0.33) and zero performance for 'Self-Reliance'. Whilst K-NN recorded the highest accuracy among the three methods, this margin should be interpreted with caution given the extreme class imbalance in the dataset.

The higher accuracy of OLR relative to SVM (67.62% vs 50.00%) warrants further discussion. Unlike SVM, OLR does not require an explicit class-balancing mechanism. The model estimates cumulative probabilities across ordinal categories and is therefore less sensitive to class frequency imbalance than distance-based or margin-based classifiers. In contrast, the application of class_weight='balanced' in the SVM model, whilst theoretically designed to address class imbalance, caused the model to over-compensate for the extremely rare self-reliance class ($\boldsymbol{n}$ = 2), leading to increased misclassifications in the dominant Progressing class and ultimately lowering overall accuracy. K-NN, despite not employing an explicit balancing mechanism, recorded the highest overall accuracy (71.00%) by leveraging local neighborhood structures, which proved more robust under the given data distribution. Each model nonetheless offers a different kind of value: OLR provides interpretability through statistically significant coefficients and Odds Ratios, enabling policymakers to understand which variables are associated with IDM status; K-NN offers stronger classification accuracy for screening applications; and SVM, despite its lower accuracy in this study, provides a theoretically sound framework that may perform better with a larger and more balanced dataset.

Conclusion

This research shows that machine learning techniques, specifically K-NN, are capable of classifying IDM status on Madura Island with the highest accuracy of 71.00%, outperforming ordinal logistic regression at 67.62% and SVM at 50.00% based on 10-fold cross-validation. Nevertheless, these three methods are complementary rather than mutually exclusive. Whilst machine learning methods offer predictive accuracy, ordinal logistic regression provides interpretative depth through odds ratios, revealing that the number of healthcare facilities ($\boldsymbol{X}_3$) and the number of villages with school facilities ($\boldsymbol{X}_5$) are the most significant determinants of an increase in IDM status. Odds ratios of 1.15 and 1.11, respectively, indicate that higher values of these variables are associated with a greater likelihood of a sub-district being classified in a higher IDM category. Based on the 2024 sub-district-level data from Madura Island, the availability and equitable distribution of basic social infrastructure, particularly healthcare facilities and village-level school access, are more strongly associated with IDM status than the volume of village budget allocation. This finding is exploratory and cannot establish causal direction given the study's cross-sectional design of the study. This study also provides new empirical contributions to village development research of applying a comparative approach between classical statistical and machine learning methods at the sub-district level, an analytical scale that has hitherto been limited in IDM studies in Indonesia.

The models developed in this study are best positioned as early-screening tools for identifying sub-districts at risk of stagnating IDM status, rather than as definitive classification systems for routine policy targeting. Given the current performance limitations, particularly the failure to reliably identify the self-reliance class (F1 = 0.00 for K-NN) and the low recall for the Growing class (0.33), operationalization in a governance context would require substantial improvements in model performance before deployment. Specifically, the model should achieve a minimum balanced accuracy of 70.00% and a macro F1-Score above 0.60 across all IDM categories before it is considered suitable for operational use. Field verification would remain essential before using any model-based classifications to inform resource allocation decisions. Substantively, development priorities should be directed toward strengthening healthcare services in remote sub-districts and ensuring equitable distribution of school facilities across all villages, which directly contribute to efforts relevant to the Sustainable Development Goals (SDGs), particularly Goal 1 (No Poverty), Goal 8 (Decent Work and Economic Growth), and Goal 16 (Peace, Justice, and Strong Institutions) on Madura Island.

Future research should address the class-imbalance limitation with more appropriate strategies than synthetic oversampling. With only two self-reliance observations, SMOTE-based interpolation lacks statistical validity. More defensible alternatives include expanding the dataset to a multi-year panel or broader geographic scope, collapsing the response to a binary outcome, or modeling the continuous IDM score directly. Future studies should also consider including additional variables that may substantively influence IDM status, such as village infrastructure quality, local government capacity, and geographic accessibility indicators.

Several limitations of this study must be explicitly acknowledged. First, the small sample size ($\boldsymbol{n}$ = 62) limits statistical power and the generalisability of findings beyond Madura Island. Second, the cross-sectional design based on a single year (2024) prevents causal inference; the observed associations between predictor variables and IDM status may reflect reverse causality or omitted variable bias. Third, measurement limitations including the use of absolute rather than per-capita infrastructure counts and the aggregation of village-level IDM scores to the sub-district level may introduce imprecision in the estimated associations. Fourth, the severe class imbalance particularly the self-reliance class (n = 2) fundamentally limits the reliability of classification models for the minority class, which is precisely the category most relevant for targeted policy intervention.

Declaration of Generative AI Use

During the preparation of this manuscript, the authors used a generative AI tool to assist with language editing and improving clarity of expression. All AI-assisted content was reviewed, verified, and edited by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

Author Contributions

The first author contributed to the main research idea, conceptualization, methodology design, and overall manuscript structure. The second author contributed to the literature review, data collection, and refinement of the discussion section. The third author contributed to the statistical validation of the ordinal logistic regression approach and machine learning implementation, including K-NN and SVM, result visualization, and drafting of the results section. The fourth author contributed to policy interpretation, manuscript revision, and validation of the final academic contribution. All authors reviewed, revised, and approved the final version of the article.

Data Availability Statement

The datasets analyzed in this study are publicly available. Village Development Index (IDM) data were obtained from the official website of the Ministry of Villages and Development of Disadvantaged Regions (Kemendes PDT) and Transmigration of the Republic of Indonesia, while supporting predictor variable data were obtained from the official website of the Central Statistics Agency (Badan Pusat Statistik) of the respective regencies. Both sources are accessible to the public upon request or through their respective official portals.

Funding Statement

No funding was received for this manuscript. This research was conducted as independent research, registered under Surat Keterangan Penelitian Mandiri No. 1667/B/DST/UN3.DRI/PT.01.05/2026, Universitas Airlangga.

Acknowledgments

The authors express sincere gratitude to the National Development Planning Agency of Indonesia (Bappenas), the Ministry of Villages and Development of Disadvantaged Regions (Kemendes PDT), and Statistics Indonesia (BPS) for providing valuable data and information that greatly supported this research. Their contributions have been essential in enabling the analysis of village development indicators and related socio-economic conditions. The authors also extend their appreciation to Universitas Airlangga for the academic support, research environment, and facilities that facilitated the completion of this study. The institutional support has played an important role in ensuring the successful conduct of this research.

Conflict of Interest

The authors declare no conflicts of interest.

References

Ahadis, M. I. H., Endaryanto, T., & Murniati, K. (2025). Factors Affecting the Developing Village Index (IDM) in the North Lampung Regency Area. *Journal of Social Research*, *4*(6), 1154–1171. https://doi.org/10.55324/josr.v4i6.2546

Akbar, F. S. A., Dirgantara, N., Wahyuni, S., & Wibowo, H. (2024). Faktor Determinan Indeks Desa Membangun. *JIMEA: Jurnal Ilmiah MEA (Manajemen, Ekonomi, Dan Akuntansi)*, *8*(3), 2793–2816.

Alhumami, A. (2025). Meningkatkan Akses dan Kualitas Pendidikan untuk Mewujudkan SDM Unggul yang Produktif. *Bappenas Working Papers*, *8*(3), 537–550. https://doi.org/10.47266/bwp.v8i3.473

Arina, A. I. S., Masinambow, V., & Walewangko, E. N. (2021). Pengaruh Dana Desa dan Alokasi Dana Desa Terhadap Indeks Desa Membangun di Kabupaten Minahasa Tenggara. *Jurnal Pembangunan Ekonomi Dan Keuangan Daerah*, *22*(3), 22–41. https://doi.org/10.35794/jpekd.35490.22.3.2021

Asnidar, Putri, A., & Hanum, N. (2025). Pengaruh Pengeluaran Pemerintah Sektor Pendidikan, Kesehatan dan Infrastruktur terhadap Pertumbuhan Ekonomi dan Indeks Pembangunan Manusia Universitas Samudra Langsa, Indonesia. *Jurnal Ekonomi Dan Pembangunan Indonesia*, *3*(2), 166–181. https://doi.org/10.61132/jepi.v3i2.1374

Azies, H. Al. (2024). AI-Based Models for Identifying Underdeveloped Villages in Indonesia’s Rural Development. *The Journal of Indonesia Sustainable Development Planning*, *5*(3), 192–202. https://doi.org/10.46456/jisdep.v5i3.611

Badan Pusat Statistik Kabupaten Bangkalan. (2025). *Kabupaten Bangkalan Dalam Angka*.

Badan Pusat Statistik Kabupaten Pamekasan. (2025). *Kabupaten Pamekasan Dalam Angka*.

Badan Pusat Statistik Kabupaten Sumenep. (2025). *Kabupaten Sumenep Dalam Angka*

Badan Pusat Statistik Kabupaten Sampang. (2025). *Kabupaten Sampang dalam Angka* (Vol. 15).

Badan Pusat Statistik Kabupaten Sumenep. (2024). *Indeks Pembangunan Manusia (IPM) Kabupaten Sumenep Tahun 2024*. Badan Pusat Statistik Kabupaten Sumenep. https://sumenepkab.bps.go.id/id/pressrelease/2024/12/02/165/indeks-pembangunan-manusia--ipm--kabupaten-sumenep-tahun-2024.html

BAPPEDA Jawa Timur. (2025). *Ringkasan Rencana Pembangunan Jangka Menengah Daerah (RPJMD) Provinsi Jawa Timur Tahun 2025-2029*.

Bhatia, N., & Vandana. (2010). Survey of Nearest Neighbor Techniques. *(IJCSIS) International Journal of Computer Science and Information Security*, *8*(2), 302–305. https://doi.org/10.48550/arXiv.1007.0085

Budyanra, B., & Azzahra, G. N. (2017). Penerapan Regresi Logistik Ordinal Proportional Odds Model pada Analisis Faktor-Faktor yang Mempengaruhi Kelengkapan Imunisasi Dasar Anak Balita di Provinsi Aceh Tahun 2015. *MEDIA STATISTIKA*, *10*(1), 37–47. https://doi.org/10.14710/medstat.10.1.37-47

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Doan, A. E. (2004). Type I and Type II Error. *Encyclopedia of Social Measurement, Three-Volume Set*, *3*(1). https://doi.org/10.1016/B0-12-369398-5/00110-9

Fauzi, M. R., Rustiadi, E., & Mulatsih, S. (2019). Ketimpangan, Pola Spasial, dan Kinerja Pembangunan Wilayah di Provinsi Jawa Timur. *Journal of Regional and Rural Development Planning (Jurnal Perencanaan Pembangunan Wilayah Dan Perdesaan)*, *3*(3), 157–171. https://doi.org/10.29244/jp2wd.2019.3.3.157-171

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heriqbaldi, U., Wisnu, W., Ramadhan, I., Ann, J., Basconcillo, Q., & Taniu, S. (2025). Social Sciences & Humanities Open Bridging the gap: The impact of suramadu bridge provision on poverty reduction in Madura Island, Indonesia. *Social Sciences & Humanities Open*, *12*. https://doi.org/10.1016/j.ssaho.2025.101740

Hosmer, D. W., Lemeshow, S., & Sturvidant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons, Inc.

Hsu, C., Chang, C., & Lin, C. (2003). *A Practical Guide to Support Vector Classification* (Vol. 1, Issue 1).

Iftitah, A. E., & Wibowo, P. (2022). Pengaruh Dana Desa, Alokasi Dana Desa, dan Pendapatan Asli Desa Terhadap Indeks Desa Membangun Di Kabupaten Gowa. *Jurnal Ilmu Pemerintahan Widya Praja*, *48*(1), 17–36. https://doi.org/10.33701/jipwp.v48i1.2331

Kemendesa. (2025). *Tentang Indeks Desa Membangun*. https://idm.kemendesa.go.id/view/detil/1/tentang-idm

Kementerian Desa dan Pembangunan Daerah Tertinggal. (2024). *Peringkat Indeks Desa Membangun (IDM) Tahun 2024*.

Kementerian Perencanaan Pembangunan Nasional. (2025). *Ringkasan Rencana Pembangunan Jangka Menengah Nasional Tahun 2025-2029*.

Mahapatra, P. R., & Chakraborty, P. S. (2015). Comparative Analysis of Nearest Neighbor Query Processing Techniques. *Procedia Computer Science*, *57*, 1289–1298. https://doi.org/10.1016/j.procs.2015.07.438

Majida, F., & Handayeni, D. M. E. (2019). Tipologi Ketertinggalan Wilayah pada Kabupaten Sampang. *JURNAL PENATAAN RUANG*, *14*(1), 24–31.

Manshur, A., & Suprapto, F. A. (2020). A Review of Suramadu Regional Development Acceleration towards Sustainable Development Concept. *The Journal of Indonesia Sustainable Development Planning*, *1*(2), 193–208. https://doi.org/10.46456/jisdep.v1i2.72

Muhtarom, Kusuma, N., & Purwanti, E. (2018). Village Index Analysis Building to Know The Village Development in Gadingrejo District of Pringsewu. *INOVASI PEMBANGUNAN: JURNAL KELITBANGAN*, *6*(2), 179–190. https://doi.org/10.35450/jip.v6i02.96

Naufal, H. A., Mulyanto, & Suryanto. (2023). The Effect of Social Infrastructure on Rural Poverty in Indonesia with The Village Development Index As Moderation Variable. *EKO-REGIONAL: JURNAL PEMBANGUNAN EKONOMI WILAYAH*, *18*(1), 87–98. https://doi.org/10.32424/1.erjpe.2022.18.1.3340

Prasetyo, D., Ikhwanuddin, El-Yunusi, M. Y. M., Priambodo, S., Darmawan, D., Djaelani, M., & Triono, B. (2024). Pengenalan Karakter Walisongo dengan Metode Mind Mapping pada Anak Remaja. *ARDHI: Jurnal Pengabdian Dalam Negri*, *2*(2), 46–56. https://doi.org/10.61132/ardhi.v2i2.522

Rimanic, L., Li, B., Renggli, C., & Zhang, C. (2020). On Convergence of Nearest Neighbor Classifiers over Feature Transformations. *Advances in Neural Information Processing Systems (NeurIPS)*, 1–22.

Rinanda, H., AB, Z., & Fajarni, S. (2025). Peran Masjid Agung Baitul Ghafur Sebagai Pusat Kegiatan Sosial dan Keagamaan di Kabupaten Aceh Daya Barat. *Arus Jurnal Sosial Dan Humaniora (AJSH)*, *5*(2), 1711–1721.

Rojak, J. A., & Fajar, A. S. M. (2025). Social Capital Dynamics in Supporting Inclusive Community Resilience and Welfare. *International Journal of Service Science, Management, Engineering, and Technology*, *7*(1), 27–34.

Satu Data Indonesia. (2024). *Indeks Desa Membangun 2024*. Satu Data Indonesia.

Sharma, S. K., Mudgal, S. K., & Chaturvedi, J. (2020). Expression and interpretation of relative risk and odds ratio in biomedical research studies. *Indian Journal of Community Health*, *32*(2), 344–347.

Sugiarto. (2025). Dampak Sistem Zonasi Terhadap Prestasi Belajar pada SMA N 1 Pangkah. *Jurnal Economic Edu*, *5*(2), 49–54.

Sukarno, M. (2020). Analisis Pengembangan Potensi Desa Berbasis Indeks Membangun Desa (IDM) (Studi Kasus: Desa Ponggok, Kecamatan Palohharjo , Kabupaten Klaten). *Prosiding Seminar Edusainstech FMIPA UNIMUS 2020*, *6*, 533–541.

Tambaip, B., Tjilen, A. P., & Ohoiwutun, Y. (2023). Peran Fasilitas Kesehatan untuk Kesejahteraan Masyarakat. *Jurnal Kebijakan Publik*, *14*(2), 189–196. https://doi.org/10.31258/jkp.v14i2.8245

Tarigan, R. A. B., & Saharuddin. (2025). The Role of Government Spending on Education, Health, and Infrastructure in Enhancing the Human Development Index in Aceh Province. *Journal Of Maliksussaleh Public Economics*, *8*(1), 34–47. https://doi.org/10.29103/jompe.v8i1.22437

Wahyuni, E., Qadri, U. I., & Elida, S. S. (2020). Implementasi Program Desa Mandiri Melalui Peningkatan Indeks Desa Membangun (IDM). *JURNAL PERSPEKTIF ADMINISTRASI DAN BISNIS*, *1*(2), 76–83. https://doi.org/10.31573/jpab.v1i2.9

Yani, S. L. (2023). Alokasi Dana Desa dalam Mendukung Kebijakan Desa di Kabupaten Sidoarjo. *AKUNTANSI HUMANIORA JURNAL PENGABDIAN MASYARAKAT*, *1*(3), 79–88. https://doi.org/10.38142/ahjpm.v2i3.670

Yusuf, M., & Khoirunurrofik, K. (2022). The Relationship of Village Funds With Village Economic Development: A Village Level Study in Indonesia. *Jurnal Bina Praja*, *14*(3), 493–504. https://doi.org/10.21787/jbp.14.2022.493-504