

Classification Index Village Building (IDM) Using Random Forest And Gradient Boosting

M. Rizal Wahyudi ¹, Yunifa Miftachul Arif ^{2*}, M. Imamudin ³

^{1,2,3} Maulana Malik Ibrahim State Islamic University, Malang City, Indonesia

Article Information

Article history :

Accepted September 24, 2026

Revised October 05, 2026

Accepted October 06, 2026

Keywords :

Village Funds
Index Village Building (IDM)
Random Forest
Gradient Boosting
SMOTE
Gridsearch

ABSTRACT

Village development requires mapping conditions based on the Village Development Index (IDM) into categories of Very Underdeveloped, Underdeveloped, Developing, Advanced, and Independent. This study aims to develop and evaluate a classification model for IDM status in Probolinggo Regency using the Random Forest and Gradient Boosting algorithms integrated with the Synthetic Minority Over-sampling Technique (SMOTE) and GridSearchCV. SMOTE was applied to address data imbalance, while Grid Search was used for hyperparameter optimization and feature importance analysis. The results showed that Random Forest produced better classification performance than Gradient Boosting before optimization, with accuracies of 93.10% and 89.66%, respectively. After optimization using GridSearchCV, the accuracy of Gradient Boosting increased to 93.10%, while Random Forest remained at 93.10%. These results indicate that hyperparameter optimization has different effects on each algorithm. The application of SMOTE and GridSearchCV can support the IDM status classification process on data with an unbalanced class distribution. The use of IDM indicator variables as well as the Village Fund ceiling and realization provides additional information in the classification process, while feature importance is used to identify the relative contribution of variables to the model.

This article is article access open below license [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/).



Corresponding Author :

Yunifa Miftachul Arif,

Maulana Malik Ibrahim State Islamic University, Malang City, Indonesia

Email: yunif4@ti.uin-malang.ac.id Orcid Id : [0000-0002-2183-0762](https://orcid.org/0000-0002-2183-0762)

1. INTRODUCTION

Village development is a crucial component of national development, aimed at improving community welfare and village independence. Its implementation is supported by Law Number 6 of 2014 concerning Villages and Village Fund management policies. In the context of Village Fund management, Minister of Finance Regulation Number 7 of 2026 emphasizes that Village Fund allocations must consider village needs, priority programs, and relevant indicators. Therefore, information that describes the condition and level of village development is needed to support development planning.

One of the instruments for measuring the level of village development and independence is the Village Development Index (IDM) as regulated in the Regulation of the Minister of Villages, Development of Disadvantaged Regions, and Transmigration Number 2 of 2016. IDM measures village development and independence based on three dimensions, namely the Social Resilience Index (IKS), the Economic Resilience Index (IKE), and the Environmental/Ecological Resilience Index (IKL) [1]. IKS describes the condition of social resilience, IKE reflects economic resilience, while IKL describes the village's environmental resilience. These three dimensions complement each other in forming the IDM value which then becomes the basis for grouping villages into five statuses, namely Very Disadvantaged, Disadvantaged, Developing, Advanced, and Independent [1]. Thus, differences in social, economic, and environmental conditions can result in differences in village development status. Differences in the characteristics of each village produce diverse data patterns so that an analytical approach is needed that is able to classify objectively and systematically.

Utilizing machine learning can be an approach in identifying patterns and performing classification based on historical data [20]. Previous studies have applied various algorithms for IDM status classification.

Research [24] used C4.5 and obtained an accuracy of 69.96%, while research [25] used a Multiclass Support Vector Machine (SVM) with an accuracy of 97.75%. Research [26] applied XGBoost to West Java IDM data and obtained an accuracy of 89%, while research [27] compared Backpropagation and Naïve Bayes on East Java IDM data in 2024 and obtained the best accuracy of 81%. These results indicate that algorithm performance can differ based on the characteristics of the data and the region used.

Random Forest and Gradient Boosting are ensemble learning algorithms that are capable of handling classification problems with complex data characteristics [21][22]. However, village status classification data can have an unbalanced class distribution, potentially causing the model to predominantly learn the majority class. To overcome this problem, this study uses the Synthetic Minority Over-sampling Technique (SMOTE) to improve the representation of the minority class [11]. In addition, GridSearchCV is used to obtain the optimal combination of hyperparameters so that model performance can be improved [3].

Based on previous research, there is still a research gap in the application of the IDM classification specifically using village data from Probolinggo Regency in 2024 by integrating IDM variables and Village Fund management variables, particularly the Village Fund budget and realization. Previous research also has not specifically combined the comparison of Random Forest and Gradient Boosting with SMOTE, GridSearchCV optimization, and feature importance analysis in the context of the region. Therefore, the novelty of this research lies in the integration of IDM characteristics and Village Fund management in the classification of village status and the application of an ensemble learning approach combined with data imbalance handling, hyperparameter optimization, and interpretation of variable contributions.

This study aims to develop and compare Random Forest and Gradient Boosting models for classifying village IDM status in Probolinggo Regency in 2024. SMOTE is used to handle class imbalance, GridSearchCV for hyperparameter optimization, and feature importance to identify variables that have a relative contribution to the classification results. The results of this study are expected to provide an optimal classification model as well as information on the characteristics of variables that contribute to village development status.

2. METHOD

A series of research processes that will be done can seen in Figure 1. The research flow is arranged as following :

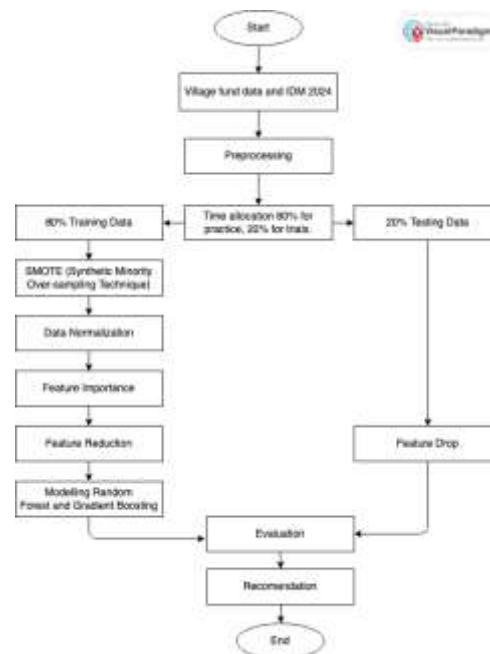


Figure 1. Research Flow

Research process This involving a number of step main, where the process starts with initial data processing For ensure the data is ready for use. Next, the village fund ceiling realization of village funds Fund absorption and IDM Status. The dataset is divided become training data by 80% and test data by 20%. Furthermore, the variable independent (x) and target variables (y) are separated. Next, data balancing is performed use SMOTE method for ensure distribution balanced class. Data normalization is done so that the data is more proportional and not creates bias in the learning model machine.

Researchers using two algorithms classification Random Forest and Gradient Boosting find best hyperparameters, we will use GridSearchCV with cross-validation. Metrics the evaluation that will be used is $f1_weighted$ Because own problem classification multiclass and previous experience imbalance class . $f1_weighted$ consider recall and precision for every class and calculate the weighted average based on support (number of true instances) of each label. For measure author model performance use Confusion Matrix Accuracy, Precision, Recall, F1-Score ROC-AUC Score.

2.1 DATA COLLECTION

Data collection is the initial step to collect all the variables needed in modeling the IDM Status classification in Probolinggo Regency. Data used in study This is a secondary dataset results integration between indicator IDM composite from the Ministry of Villages, Development of Disadvantaged Regions , and Transmigration (Kemendesa PDRT) 2024 as well as budget and realization data utilization of Village Funds from the Central Statistics Agency (BPS Kab. Probolinggo) in 2024. Research location in the region Regency Probolinggo is located in the Province East Java.

Table 1. Data Variables

NO	VARIABLE NAME	INFORMATION
1	KODE DESA	: Official identity code for village administrative areas (Ministry of Home Affairs/BPS)
2	KECAMATAN	: Name of the sub-district where the village is located
3	NAMA DESA	: Official name of the village
4	JUMLAH PENDUDUK	: Total number of people residing in the village
5	KEPADATAN PENDUDUK	: Ratio of the number of people per square kilometer (people/km ²)
6	JUMLAH KK	: Total number of registered heads of families
7	LUAS WILAYAH	: Land area of the village area (km ² or ha)
8	PAGU DANA DESA	: Total budget allocation ceiling for Village Funds (Rupiah)
9	REALISASI DANA DESA	: Total village funds successfully absorbed and spent (Rupiah)
10	SERAPAN DANA DESA	: Percentage of effectiveness of absorption realization against the ceiling (%)
11	IKS	: Social Resilience Index achievement value
12	IKE	: Economic Resilience Index achievement value
13	IKL	: Environmental/Ecological Resilience Index achievement value
14	NILAI IDM 2024	: IDM average composite score in 2024
15	STATUS IDM 2024	: Village status classification class (Very Underdeveloped, Underdeveloped, Developing, Advanced, Independent)

Table 2. Dataset

No	Village Name	Village Fund Ceiling	Village Fund Realization	Village Fund Absorption		IDM Value	IDM Status
1	Dringu	843,899,000	864,352,203	102%		7,617	Proceed
2	Kalirejo	1,292,407,000	1,237,416,700	96%		7,724	Proceed
3	Kalisalam	810,288,000	771,793,050	95%		7,562	Independent
4	Kedungdalem	1,134,701,000	1,260,187,000	111%		7,963	Proceed
5	Mranggon Lawang	846,293,000	868,458,573	103%		7,075	Proceed
6	Ngepoh	776,109,000	774,456,650	100%		7,144	Proceed
7	Pabean	885,247,000	886,187,600	100%		8,206	Independent
8	Randuputih	1,035,136,000	1,025,978,781	99%		8,322	Proceed

9	Sekarkare	671,694,000	682,813,750	102%		7,235	Proceed
10	Sumberagung	787,470,000	770,031,437	98%		7,698	Proceed
...	...	...	...	...	...	...	...
325	Sumberkledung	962,730,000	927,257,785	96.32%		7,243	Proceed

2.2 PREPROCESSING

The process begins with load index data Village Building (IDM) from Google Sheets. The data is then through stage comprehensive pre-processing, including :

1. Feature Engineering

First step in engineering feature is inspect information base DataFrame For understand column data type , number non-null values, and the use of memory. This will help in identify columns that require handling mark is lost or conversion data type .

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 143 entries, 0 to 142
Data columns (total 16 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   NO                                     143 non-null    int64
1   KODE DESA                             143 non-null    int64
2   KECAMATAN                             143 non-null    object
3   NAMA DESA                             143 non-null    object
4   JUMAH PENDUDUK                        142 non-null    float64
5   KEPADATAN PENDUDUK                    142 non-null    float64
6   JUMLAH KK                             142 non-null    float64
7   LUAS WILAYAH                          142 non-null    object
8   PAGU DANA DESA                        142 non-null    object
9   REALISASI DANA DESA                   142 non-null    object
10  SERAPAN DANA DESA                     142 non-null    object
11  IKS                                    143 non-null    float64
12  IKE                                    143 non-null    float64
13  IKL                                    143 non-null    float64
14  NILAI IDM 2024                         143 non-null    float64
15  STATUS IDM 2024                        143 non-null    object
dtypes: float64(7), int64(2), object(7)
memory usage: 18.0+ KB
```

Figure 2. Data Frame

2. Handling Missing Values

Based on Figure 2.1 several column own missing values (Non-Null Count < 143) . For handle This is for To fill in the empty data, median imputation is used. Median imputation is technique Fill in empty data (*missing values*) using mark middle from the data available on a variable .

3. Cleanup and Conversion of 'Object Column to Numeric

A number of the column that should be numeric moment This own data type 'object' because existence non-numeric characters (e.g. , comma, period) as separator thousands, or sign percent). Before median imputation can done, columns This need cleaned and converted to numeric data type (float or integer). We will see mark unique in some column the For understand pattern necessary cleaning.

4. Target Variable Encoding

The target variable encoding stage is performed to convert categorical data in the IDM status variable into numeric form so it can be processed by the machine learning algorithm. The IDM status, consisting of several categories, is converted into consistent numeric labels without changing the information or meaning of each class. The encoding results are then used as target variables in the training and evaluation of the classification model.

```

*** Custom Mapping STATUS IDM (Encoded):
Sangat Tertinggal: 1
Tertinggal: 2
BERKEMBANG: 3
MAJU: 4
MANDIRI: 5

DataFrame setelah encoding variabel target:
  STATUS IDM 2024  STATUS IDM (Encoded)
0             MAJU                    4
1             MAJU                    4
2          MANDIRI                    5
3             MAJU                    4
4             MAJU                    4

Distribusi nilai STATUS IDM (Encoded):
STATUS IDM (Encoded)
4      77
5      48
3      18
Name: count, dtype: int64

```

Figure 3. Custom Mapping Status IDM

The IDM STATUS 2024 column is encoded to mark numeric (1-5) to be used as target variable.

2.3 DATA SHARING

After preprocessing, author separate data into feature (x) and target (y). Our target is IDM STATUS column (Encoded). Next, the data will be shared into a training set of 80% and a testing set of 20%. Because the target is variables possible categorical No balanced distribution, then writer using SMOTE (Synthetic Minority Over-sampling Technique) on training data For balance amount class.

2.3.1 SMOTE (Synthetic Minority Over-sampling Technique)

For understand SMOTE impact, author do visualize distribution a number of feature important on `x_train_resampled` and compare it based on target class `y_train_resampled`.

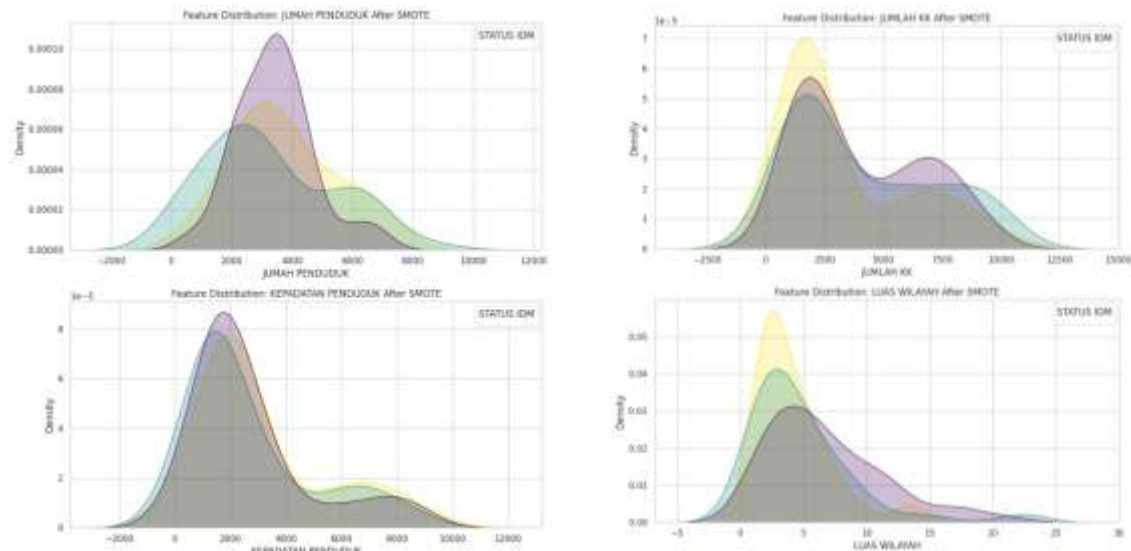


Figure 4. Graph Visualization Distribution SMOTE feature

Visualization distribution feature after SMOTE has been succeed created. Graphics This show distribution density For four feature first POPULATION NUMBER, POPULATION DENSITY, NUMBER OF FAMILY HOMES, and AREA) from the resampled training data (`x_train_resampled`), separated based on IDM STATUS class (Encoded).

From figure 4 we Can see that SMOTE has help balance amount samples in each target class is shown with color different . Although distribution original Possible No balanced , after SMOTE, density For every class on features This show more patterns evenly or balanced, which is objective from SMOTE. This means your model will trained with more representation fair For every IDM STATUS category.

2.3.2 DATA NORMALIZATION

After SMOTE, the author do normalize features numeric use StandardScaler . StandardScaler is tool deep data preprocessing machine learning that works For standardize feature numerical to have the mean value is the same with 0 and standard deviation The same with 1. Things are done so that all feature own the same scale , so that No There is features that dominate the learning model machine Because range its greater value big . We will do fit_transform on resampled training data (x_train_resampled) and only transform on test data (x_test) for avoid data leak.

3. RESULTS AND DISCUSSION

Modeling process involving implementation as well as training of two algorithms ensemble leading, namely Random Forest Classifier and Gradient Boosting Classifier, using training data that has been through stage pre-processing. Evaluation performance furthermore done against test data with utilise a number of metric main, which includes level accuracy, reports classification covering mark precision, recall, and F1-score, as well as matrix confusion .

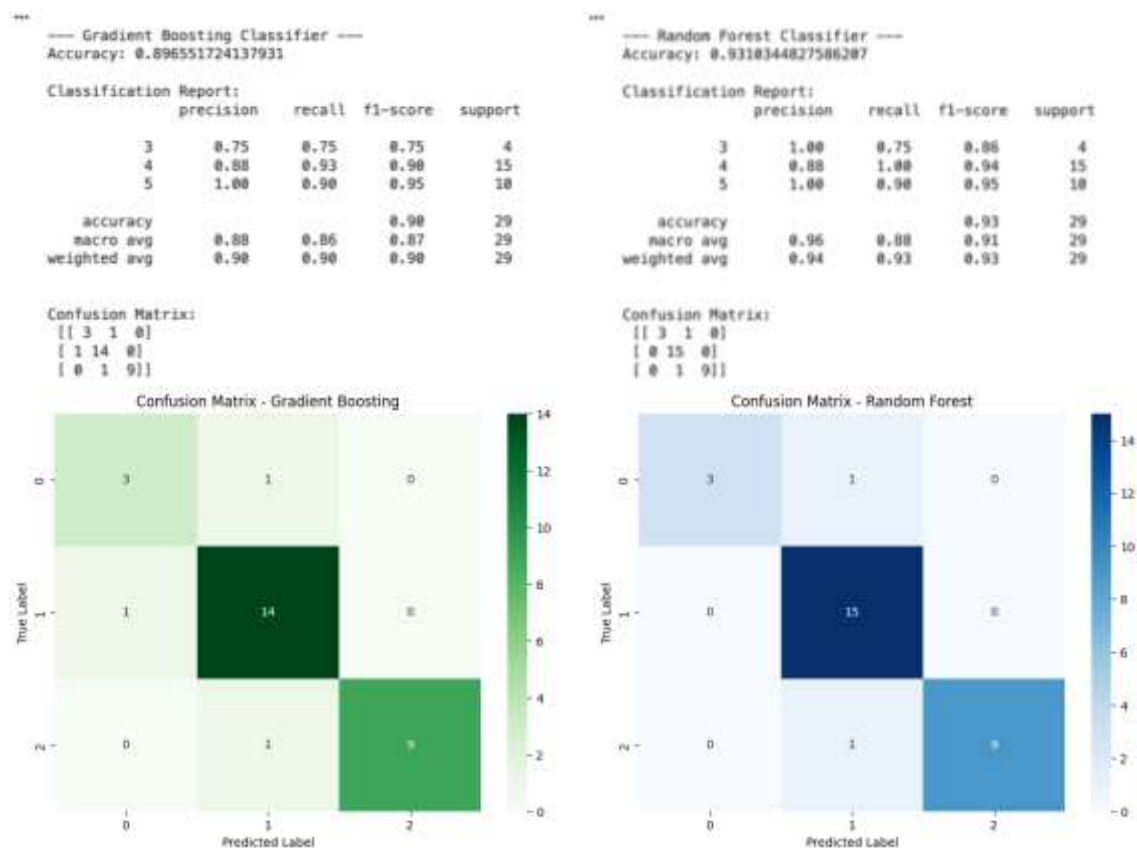


Figure 5. Test Results

Based on testing the Random Forest model succeed record level accuracy around 93.10%, which reflects capability very optimal predictive in classification of IDM status and consistency reliable performance throughout category class. Meanwhile that is, the Gradient Boosting model get accuracy in the range of 89.66%, although this global achievement A little be under Random Forest , algorithm This still categorized as a strong and competent role model in map as well as classifying class data in a way accurate.

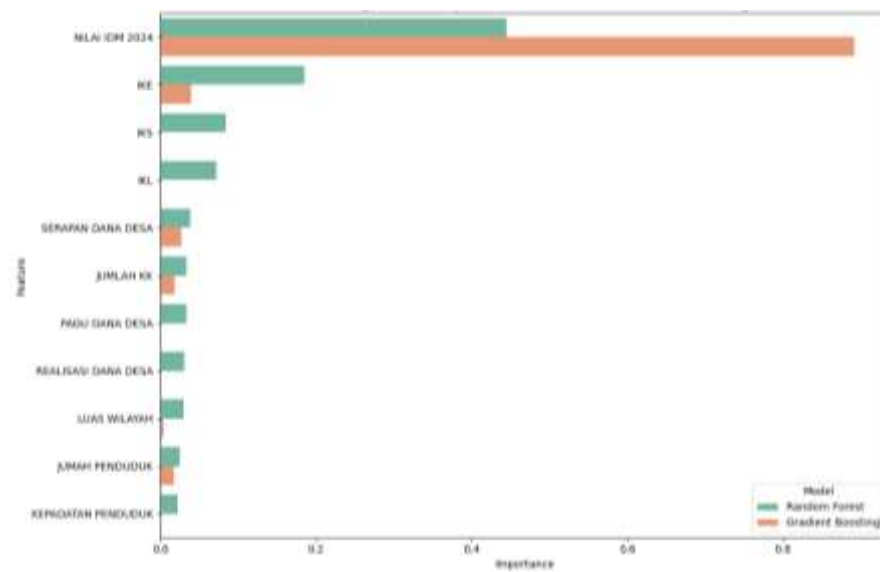


Figure 6. Feature Importance Result

Analysis level feature importance executed use identify to what extent each variable contributes to results model prediction. Based on Figure 6, the test results show that IDM 2024 VALUE variable in general consistent dominate as the most influential features on both model architecture. Dominance This appears very prominent in the Gradient Boosting model , which indicates level high dependency from the model to mark the aggregate in question.

Beside that, component index The IDM compilers, namely IKE, IKS, and IKL, also show level substantial interest, particularly in the Random Forest model, which confirms that index component the act as predictor key. While that, the variables supporters others, such as aspect demographics as well as condition finance village, giving relative contribution more small However still hold relevant roles in the prediction process.

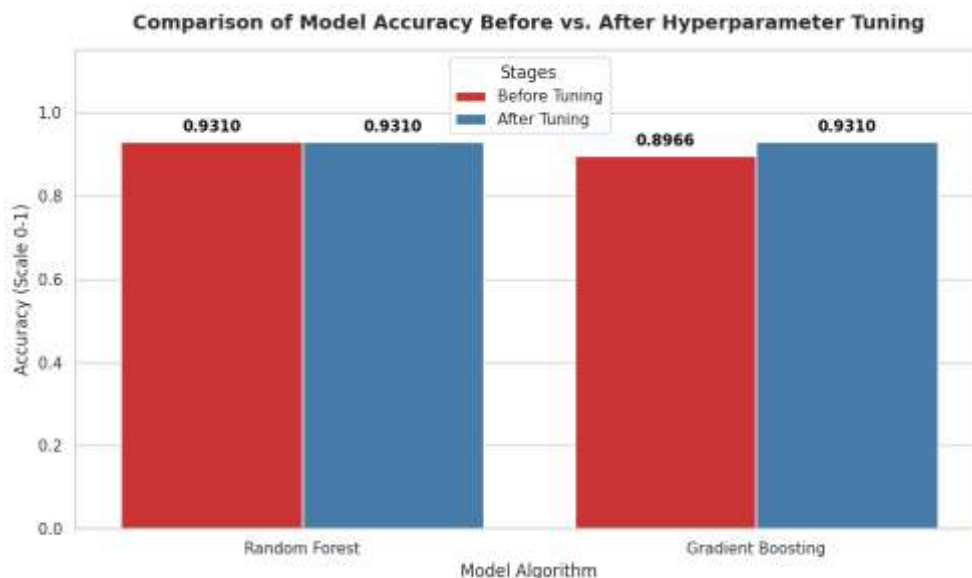


Figure 7. Comparison Model Accuracy Before and After Hyperparameter Tuning

Researchers also conducted trials For increase accuracy with hyperparameter tuning with GridSearchCV, the Gradient Boosting model experienced improvement accuracy from 89.66% to 93.10%, equaling The accuracy of the Random Forest model is stable at 93.10%. The use of The SMOTE technique has proven to be crucial. in equalize representation target class (Very Disadvantaged, Disadvantaged, Developing,

Advanced, Independent), so that the model is not biased and produces even and high f1-score For all over IDM status category.

4. CONCLUSION

The results showed that Random Forest produced better classification performance than Gradient Boosting before optimization, with accuracies of 93.10% and 89.66%, respectively. After optimization using GridSearchCV, the accuracy of Gradient Boosting increased to 93.10%, while Random Forest remained at 93.10%. These results indicate that hyperparameter optimization has different effects on each algorithm. The application of SMOTE and GridSearchCV can support the IDM status classification process on data with an unbalanced class distribution. The use of IDM indicator variables and the Village Fund budget and realization provides additional information in the classification process, while feature importance is used to identify the relative contribution of variables to the model.

ACKNOWLEDGEMENTS

The author would like to thank the supervisor, institution, data providers, and all parties who provided support, guidance, and contributions during the research and the preparation of this article. This support was instrumental in completing the research and publishing this article.

REFERENCE

- [1] Amanda, W., & Voutama, A (2025). Classification Income Use Random Forest Algorithm : A Study The Case of Adult Income Dataset. *Journal Scientific Global Informatics* Volume 16 No. 2 August 2025.
- [2] Chuenmee, N., Phothi, N., Chamniprasart, K., Khaengkarn, S., & Srisertpol, J. (2025). Machine learning for predicting resistance spot weld quality in automotive manufacturing. *Results in Engineering*, 25. <https://doi.org/10.1016/j.rineng.2024.103570>
- [3] HC Morama, DE Ratnawati, and I. Arwani, (2022) "Analysis Sentiment based Aspect to Hotel Tentrem Yogyakarta reviews using Random Forest Classifier Algorithm," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 4, pp. 1702–1708. Available at: <https://jptiik.ub.ac.id/index.php/jptiik/article/view/10908>
- [4] Harriz, MA, Akbariani, NV, Setiyowati, H., & Santoso, H. (2023). Classifying Village Fund In West Java, Indonesia Using Catboost Algorithm. *Journal Indonesia : Management Informatics and Communication*, 4 (2), 691-697. <https://doi.org/10.35870/jimik.v4i2.269>
- [5] Caring, Endiana & Sunarsih (2025). Good Governance and Trust in Village Fund: The Role of Accountability and Community Participation. *HKJU-CCPA*, 25(4), 651–675, <https://doi.org/10.31297/hkju.25.4.4>
- [6] Mulyana Mulyana & Muhammad Al Habib (2024). Effectiveness of Village Funds In Community Development and Empowerment In Village Tambusai Batang Dui District Inner Solapan Regency Bengkalis Riau Province. *JEKP (Journal of Economics and Finance) Public* Vol. 11 No. 1. <https://doi.org/10.33701/jekp.v11i1.4087>
- [7] PPE Indarbensyah and N. Rochmawati, (2021) "Implementation of N-Gram using Random Forest Algorithm and Naïve Bayes Classifier in Analysis Sentiment PPKM Policy 2021," *J. Informatics Comput. Sci.*, vol. 2, no. 4, pp. 235–244. Available at: <https://ejournal.unesa.ac.id/index.php/jinacs/article/view/39845>
- [8] Permatasari, J., Sulistianingsih, N., & Dinata, MI (2026). Comparison of Machine Learning Models for Analysis Factor Dominant Index Village Building in NTB. *Journal Informatics Technology and Science (Jinteks)*.
- [9] Qaddara, I., Alraba'nah, Y. Enhancing Requirements Classification Using Machine Learning Techniques. *SN COMPUT. SCI.* 6, 649 (2025). <https://doi.org/10.1007/s42979-025-04158-z>
- [10] Samudra, JT, Hayadi, BH, & Ramadhan, PS (2022). Number 2. Volume, 5, 127–133. <https://ojs.trigunadharma.ac.id/index.php/jsk/index>
- [11] Sukarman Kamuli, Yowan Guest, Mustaqim Pabbajah, Irwan Abdullah, Sastro M. Wantu & Yanti Aneta (2025). Analyzing the ineffectiveness of village fund utilization in improving food security. *African Journal* Vol. 64, no. 3-4 https://hdl.handle.net/10520/ejc-agrekon_v64_n3-4_a313
- [12] Sutikno, Zulkefly Abdul Karim & Alifah Rokhmah Idialis (2025). Analysis of village fund efficiency and the variables affecting it in Indonesia. *Public and Municipal Finance*, 14 (2), 1-14. doi: [10.21511/pmf.14\(2\).2025.01](https://doi.org/10.21511/pmf.14(2).2025.01)
- [13] M, R., Hasiah, H., & Azmi, A. (2023). Village Fund Management In Upgrading Index Status Village Building. *Supremacy Law : Journal Legal Research*, 31 (2), 99–109. <https://doi.org/10.33369/jsh.31.2.99-109>
- [14] Ramadhani, Arsy (2022) "Review of the Effectiveness Policy Priority Use of Village Funds Year 2021 Budget against Index Village Village Development (IDM) Wonoketro". KTTA thesis, Polytechnic State Finance STAN.
- [15] Netra, Ekawati (2022) Use of Village Funds and Its Implications To Index Village Build (IDM) In the Regency West Pasaman. Master's thesis, Andalas University.
- [16] D Nayanti, et al (2025). Effectiveness Village Fund Management In Physical Development in the Village Balida Subdistrict Paringin Regency Balangan, *JKDPK*, vol. 2, no. 1, pp. 56–65, March. 2025, Accessed: Sep. 18, 2026. [Online]. Available: <https://ejurnal.stiaamuntai.ac.id/index.php/JKDPK/article/view/1190>
- [17] Anam, C., Plaček, M., Valentinov, V. et al. Village funds and poverty reduction in Indonesia: new policy insight. *Discov glob soc* 1, 14 (2023). <https://doi.org/10.1007/s44282-023-00016-6>

M. Rizal Wahyudi : Classification Index Village Building ...

- [18] Rijani *et al.* , (2026). Effectiveness of Village Funds In Improving Development in Muara Payang Village Muara Komam District Regency Paser , *JKDPK* , vol. 1, no. 2, pp. 214–223, Oct. 2024, Accessed: Sep. 18, 2026. [Online]. Available: <https://ejurnal.stiaamuntai.ac.id/index.php/JKDPK/article/view/734>
- [19] F. Akmaliah , Y. Umaidah , and I. Maulana, “ Comparative Decision Tree and Gradient Boosting Algorithms for Prediction Recipient Help Social in the Village West Kalijati ”, *ikraith-informatika* , vol. 10, no. 2, pp. 163-171, May 2026.
- [20] Aravind, T., Reddy, B.S., Avinash, S., & Jeyakumar , G. (2019). Information of Undergraduate Students. Proceedings of the Third International Conference on I-SMAC, 542–546
- [21] Sihombing , PR, & Yuliati , IF (2021). Implementation Machine Learning Methods in Classification Risk Incident Low Birth Weight in Indonesia. *MATRIX: Journal Management , Informatics and Engineering Computers* , 20(2), 417–426. <https://doi.org/10.30812/matrik.v20i2.1174>
- [22] Putra, PP (2022). Molecular Docking Theory and Tutorial Using AutoDock Vina. *Insight Knowledge*
- [23] AF Anjani, D. Anggraeni , and IM Tirta, “ Random Forest Implementation Using SMOTE for Analysis Sentiment Review Sister Application for Students UNEJ”, *TEKNOSI* , vol. 9, no. 2, pp . 163–172, Sep 2023. <https://orcid.org/0009-0001-2525-8708>
- [24] MB Prayoga , N. Cahyono , Subektiningsih , and Kamarudin , " Application of Grid Search for Machine Learning Model Optimization in Classification Sentiment YouTube Comments ," *JATI (Journal Informatics Engineering Students)* , vol. 9, no. 3, pp. 3817–3824, Jun. 2025. DOI: <https://doi.org/10.36040/jati.v9i3.13375>