Jurnal Indonesia : Manajemen Informatika dan Komunikasi
www.journal.amikindonesia.ac.id/jimik/ Vol 4 No 2, May (2023)
E-ISSN: 2723-7079, P-ISSN: 2776-8074

CLASSIFYING VILLAGE FUND IN WEST JAVA, INDONESIA USING CATBOOST ALGORITHM

Muhammad Alfathan Harriz [1*], Nurhaliza Vania Akbariani [2], Harlis Setiyowati [3], Handri Santoso [4]

[1*,3,4] Universitas Pradita, Tangerang Regency, Banten Province, Indonesia.
[2] Sekolah Tinggi Teknologi Terpadu Nurul Fikri, City of South Jakarta, Special Capital Region of Jakarta, Indonesia.

Corresponding Email: harrizsb@gmail.com [1*]

Article History:
Submited March 23, 2023; Accepted in revised April 24, 2023; Accepted April 30, 2023; Published May 20, 2023.

Abstrak

Dengan lebih dari 261 juta penduduk, Indonesia memiliki sekitar 15.000 desa, menurut Kementerian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi. Di antaranya, 1.406 berada di Jawa Barat. Dari jumlah tersebut, 504 di antaranya maju, 464 berkembang, 390 tertinggal, dan 48 sangat tertinggal. Model pembelajaran mesin CatBoost digunakan untuk mengklasifikasikan dana desa di Jawa Barat dari tahun 2018 hingga 2021 dan memiliki peringkat akurasi 75%, peringkat presisi 79%, penarikan kembali 79%, dan skor f1 79%, menunjukkan kinerjanya yang sangat baik. Namun, poin data yang hilang harus dihilangkan dari analisis dan disarankan agar metode yang lebih canggih untuk menangani nilai yang hilang harus digunakan dalam penelitian selanjutnya. Selain itu, penyetelan hyperparameter dapat digunakan untuk meningkatkan performa model, dan berbagai metrik dapat digunakan untuk menilai hasil secara akurat. Secara keseluruhan, CatBoost dapat bermanfaat bagi Pemerintah Indonesia untuk mengklasifikasikan dana desa berdasarkan statusnya, menyalurkan dana secara lebih akurat dan efisien, serta mengamati situasi desa dari tahun ke tahun.

Kata Kunci: Algoritma CatBoost; Klasifikasi; Indonesia; Dana Desa.

Abstract

With over 261 million inhabitants, Indonesia is home to approximately 15,000 villages, according to the Ministry of Villages, Disadvantaged Regions, and Transmigration. Among these, 1,406 are in West Java. Of these, 504 of them are advanced, 464 are developing, 390 are disadvantaged, and 48 are very disadvantaged. The CatBoost machine learning model was used to classify village funds in West Java from 2018 to 2021 and had an accuracy rating of 75%, precision rating of 79%, recall of 79%, and f1 score of 79%, demonstrating its excellent performance. However, missing data points had to be removed from the analysis and it is suggested that a more sophisticated method for handling missing values should be used in future studies. In addition, hyperparameter tuning could be employed to increase the model's performance, and a variety of metrics could be used to accurately assess the results. Overall, CatBoost may be of benefit to the Indonesian Government in order to classify village funds according to their status, channel funds more accurately and efficiently, and observe the situation of a village year-over-year.

Keyword: CatBoost Algorithm; Classification; Indonesia; Village Fund.

1. Introduction

A country with more than 261 million inhabitants [1], Indonesia is one of the world's ten largest economies and has a per capita income of US$9,270 (PPP) in 2013 [2]. Its unique geography, comprising an archipelago of more than 17,000 islands distributed over five large islands and 33 provinces, adds to its cultural diversity, as well as presenting significant challenges in terms of infrastructure and connectivity between different regions. This diversity and complexity of Indonesia's geography contribute to its distinctiveness, while also highlighting the need for tailored approaches to development and governance. It has been revealed that Indonesia has approximately 15,000 villages and 1,406 of them are in West Java, according to data provided by the Ministry of Villages, Disadvantaged Regions, and Transmigration at the Dissemination Meeting in Jakarta on April 1, 2016. Of these, 504 of them are advanced, 464 are developing, 390 are disadvantaged, and 48 are very disadvantaged. According to Law No. 6 of 2014 [1][3][4], a village, recognized as part of the Indonesian unitary state system, is authorized to regulate and manage local governance affairs and community interests based on traditional and customary rights, as well as community initiatives, with territorial boundaries and authority designated for the community unit [3]–[5]. Through the Village Law, villages will be able to become economically independent, democratic, and progressive, laying the groundwork for effective governance [5]. It has been achieved through the implementation of various empowerment and development initiatives in villages.

The topic of village funds in Indonesia has been discussed in several studies. A study conducted of priority program selection of village fund using the k-means method [6], as a result of investigating the prioritization of village fund programs using the k-means method, this study aims to assist village officials in determining which programs should receive priority funding for their village funds, providing help to those determining how they can improve their village funds. In addition, grouping of village status in west java province using the Manhattan, Euclidean and Chebyshev methods on the k-mean algorithm [7], according to this study, the k-means algorithm is useful for clustering village data and prioritizing fund distribution, with Chebyshev being the most efficient distance calculation method, and Euclidean being the most efficient method for determining cluster status using the Davies Index. Other study of decision support system for determination of village fund allocation using AHP method [8], the study analyzed village fund management in Siborna village using the AHP method to identify factors that hinder or support it, with the village of Vegetable Farm being recommended for a government fund allocation of 3,0000. Moreover another study of implementation of linear regression algorithm and support vector regression in building prediction models fish catches of fishermen in ciparagejaya village [9] concluded linear regression algorithm and support vector regression algorithm resulting in a smallest RMSE value of 0. Furthermore, a study of technique for order preference method by similarity to ideal solution (TOPSIS) for decision support system in determining the priority for receiving village fund assistance [10] has been conducted and shows the use of a Decision Support System (DSS) with TOPSIS method can resolve problems faced by the STM Hulu TigaJuhar sub-district in accurately determining priority for receiving Village Fund assistance.

In accordance with the study mentioned above, village funds are distributed in a manner inconsistent with the existing priority scale [6], [10] as well ineffective and inefficient [7], [8]. Aside from that, most of the studies conducted by the researchers used an unsupervised learning approach. We aim to evaluate the categorical boosting (CatBoost) algorithm as a supervised learning method for classifying data of village fund in west java [11] in this study. Due to the importance of rural development, this study on the distribution of village funds is extremely necessary, because the distribution of funds accurately and fairly can have a significant impact on the lives of people living in rural areas. Using supervised learning techniques, like the CatBoost algorithm, a more accurate classification process can be achieved in order to distribute funds more equitably and to increase process efficiency. Likewise, the results of this study can assist policymakers and government officials in determining how to allocate resources to rural development in the future based on the findings.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi, Volume 4 No 2, May (2023), pp. 691-697
DOI : https://doi.org/10.35870/jimik.v4i2.269

2. Methods

2.1 Preprocessing

Missing values, noise, inconsistencies, and voluminous data generated from various sources are possible characteristics of data obtained from various sources [12]. As a result of these imperfect data, a data preparation stage is required for cleaning and preparing the data for analysis. Among the many meaningful steps that machine learning provides, data preprocessing is one of the most important [13]. Because of these factors, this step typically takes a substantial amount of time and must be completed carefully to ensure that the modeling process is as efficient as possible. In order to reduce complex, noisy, and irrelevant elements, data pre-processing involves several steps, including data preparation, integration, cleaning, normalization, scaling, and data reduction techniques [14], including feature selection and discretization [15]–[18]. Here, the goal is to develop a final dataset suitable for machine learning (ML) analysis.

2.2 Categorical Boosting (CatBoost)

A machine learning framework known as CatBoost was developed in 2018 by Dorogush [19] as an improved variation of the Gradient Boosting Decision Tree (GBDT) toolkit [16], similar to XGBoost [20]. By using this algorithm, gradient bias and prediction shift problems can be addressed. With this model, categorical features are automatically regarded as numerical characteristics by an algorithm, and a combination of category features can be utilized to benefit from connections between features [16], [21]. Furthermore, the tree model is perfectly symmetrical and reduces overfitting while improving accuracy and generalizability. In addition, the formula for categorical boosting algorithm is explained in formula 1.

$$y = \Sigma_i = 1^T f_i(x_i) \quad (1)$$

Where y is the predicted target value, T is the number of trees in the ensemble, f_i is the i-th tree and x_i is the feature vector for the i-th tree. In words, this equation states that the estimated value y has the same value to the sum of the predicted values of each individual decision tree in the ensemble, with each decision tree being a function of the input feature vector x_i. The model combines the predictions of all the trees to generate a final prediction.

3. Result and Discussion

Data of village funds in West Java from 2018 to 2021, obtained from Kaggle, consisted of 22187 values. The summary of data is described in table 1.

Table 1. Raw data of village fund

Year	Kabupaten	Desa	Anggaran	Penyaluran
2021	Bogor	Cisarua	3233745000	3233745000
2020	Bogor	Cisarua	2960868000	2960863248
2020	Bekasi	Babelan Kota	2820159000	2820159000
...	...	...	...	...
2018	Sukabumi	Warungreja	953036000	953036000

Upon examination we've found some null values in the data. These null values will impact the performance of the machine learning model. Therefore, as part of preprocessing process, some null values were removed by filtering out the null values of the Anggaran and Penyaluran columns. This resulted in a total of 21173 data. Since there was no categorical status provided in the village data, a status category was created based on the median of the data. This classification was based on the

Jurnal Indonesia : Manajemen Informatika dan Komunikasi, Volume 4 No 2, May (2023), pp. 691-697
DOI : https://doi.org/10.35870/jimik.v4i2.269

categories provided by the Ministry of Villages, Disadvantaged Regions, and Transmigration at the Dissemination Meeting, which consisted of "very disadvantaged," "disadvantaged," "developing," and "advanced." The division of village status was done by comparing the village's "Anggaran" column to the median budget of all villages: if a village's budget was less than half of the median, it was classified as "very disadvantaged"; if it was between half and the median, it was classified as "disadvantaged"; if it was between the median and 1.5 times the median, it was classified as "developing"; and in case it was more than 1.5 times the median, it was classified as "advanced." Furthermore, we put the value of division village into "anggaran_classification" column. The data with "anggaran_classification" is describes as in Table 2.

Table 2. Data of village fund with status of village

Tahun	Kabupaten	Desa	Anggaran	Penyaluran	anggaran_classification
2021	Bogor	Cisarua	3233745000	3233745000	advanced
2020	Bogor	Cisarua	2960868000	2960863248	advanced
2020	Bekasi	Babelan Kota	2820159000	2820159000	advanced
...	...	...	...	...	...
2018	Sukabumi	Warungreja	953036000	953036000	disadvantaged

In the event that the data has been normalized to a sufficiently high degree, then the CatBoost algorithm may be applied. CatBoost's effectiveness was confirmed by a variety of evaluation metrics, including accuracy, precision, recall, and F1 score, which are used to measure CatBoost's effectiveness. A measure of accuracy is calculated by dividing the number of true positives by the total number of predictions that have been made [22]–[25]. Likewise, precision is expressed in terms of the ratio of the number of correctly predicted positive instances to the total number of correctly predicted positive instances [26]. When it comes to recall, it is calculated by comparing the number of instances that were correctly predicted as positive to the number of instances that were actually positive, and it is rated accordingly [26]. Further to this, the F1 score is calculated by calculating the harmonic mean of the precision and recall measurements [22]. This score is used in order to evaluate the overall performance of the model as a whole. It was determined from the results of the evaluation that the accuracy of the model was 75%, the precision of the model was 79%, the recall was 79%, and the F1 score was 79%.

4. Conclusion

It was found that CatBoost machine learning model worked well on the village funds in West Java from 2018 to 2021 and had an accuracy rating of 75%, precision rating is 79%, recall is 79%, and f1 score is 79%, demonstrating its excellent performance. It was essential to introduce a distinct column for the status of each village based on the median value of the data, since the original data itself did not contain much information of this type. Even though some data points were missing due to null values, metrics indicated that CatBoost could be used to classify village funds in a manner that is accurate. A more sophisticated method of handling missing values such as imputation techniques is suggested for future studies than simply removing them from the study and just removing them from the analysis. Moreover, it is necessary to determine the status of a village within the framework of applicable laws and regulations. As a further enhancement, hyperparameter tuning could be employed to increase the model's performance, and a variety of metrics such as ROC curve, AUC or logarithmic loss could be employed to accurately assess the results. CatBoost may be of benefit to the Indonesian government to classify village funds according to their status in order to channel funds more accurately and efficiently. Furthermore, the government might be able to observe the situation

Jurnal Indonesia : Manajemen Informatika dan Komunikasi, Volume 4 No 2, May (2023), pp. 691-697
DOI : https://doi.org/10.35870/jimik.v4i2.269

of a village year-over-year, and if it remains in the same condition year-over-year, it would be able to determine what improvements are needed.

5. Reference

[1] Permatasari, P., Ilman, A.S., Tilt, C.A., Lestari, D., Islam, S., Tenrini, R.H., Rahman, A.B., Samosir, A.P. and Wardhana, I.W., 2021. The village fund program in Indonesia: Measuring the effectiveness and alignment to sustainable development goals. *Sustainability*, *13*(21), p.12294. DOI: https://doi.org/10.3390/su132112294.

[2] Kurniawan, H., de Groot, H.L. and Mulder, P., 2019. Are poor provinces catching-up the rich provinces in Indonesia?. *Regional Science Policy & Practice*, *11*(1), pp.89-108. DOI: https://doi.org/10.1111/rsp3.12160.

[3] Bawono, I.R., 2019. *Optimalisasi potensi desa di Indonesia.* Gramedia Widiasarana Indonesia.

[4] Bawono, I.R., 2019. *Panduan penggunaan dan pengelolaan dana desa.* Gramedia Widiasarana Indonesia.

[5] Husmayanti, R., 2021. Tata Kelola Dana Desa Berbasis Perencanaan Partisipatif di Desa Pantai Cermin Kiri Kabupaten Serdang Bedagai. *Jurnal Ilmiah Mahasiswa Ilmu Sosial dan Politik [JIMSIPOL]*, *1*(3).

[6] Sulastri, A., Khalifah, S., Lestari, D., Gustian, D., Muslih, M. and Rafaelevna, K.I., 2020. PRIORITY PROGRAM SELECTION OF VILLAGE FUND USING THE K-MEANS METHOD. *INTERNATIONAL JOURNAL ENGINEERING AND APPLIED TECHNOLOGY (IJEAT)*, *3*(2), pp.75-85. DOI: https://doi.org/10.52005/ijeat.v3i2.61.

[7] Pranoto, G.T., Hadikristanto, W. and Religia, Y., 2022. Grouping of Village Status in West Java Province Using the Manhattan, Euclidean and Chebyshev Methods on the K-Mean Algorithm. *JISA (Jurnal Informatika dan Sains)*, *5*(1), pp.28-34. DOI: https://doi.org/10.31326/jisa.v5i1.1097.

[8] Sianipar, V.V., Wanto, A. and Safii, M., 2020. Decision Support System for Determination of Village Fund Allocation Using AHP Method. *The IJICS (International Journal of Informatics and Computer Science)*, *4*(1), pp.20-28. DOI: http://dx.doi.org/10.30865/ijics.v4i1.2101.

[9] F. Mahendra, A. M. Siregar, K. A. Baihaqi, B. Priyatna, and L. Setyani, 2023. Implementation Of Linear Regression Algorithm And Support Vector Regression In Building Prediction Models Fish Catches Of Fishermen In Ciparagejaya Village. *Edutran Computer Science and Information Technology*, *1*(1).

[10] Andayani, M., Yanti, N. and Lusiyanti, L., 2022. TECHNIQUE FOR ORDER PREFERENCE METHOD BY SIMILARITY TO IDEAL SOLUTION (TOPSIS) FOR DECISION SUPPORT SYSTEM IN DETERMINING THE PRIORITY FOR RECEIVING VILLAGE FUND ASSISTANCE. *Jurnal Mantik*, *6*(1), pp.560-567.

[11] Village Fund Data in Jawa Barat Indonesia. 2023. Available at: https://www.kaggle.com/datasets/eki1381/village-fund-data-in-jawa-barat (accessed Apr. 23, 2023).

[12] Eilertz, D., Mitterer, M. and Buescher, J.M., 2022. automRm: an r package for fully automatic LC-QQQ-MS data preprocessing powered by machine learning. *Analytical Chemistry*, *94*(16), pp.6163-6171. DOI: https://doi.org/10.1021/acs.analchem.1c05224.

[13] Ahmad, T. and Aziz, M.N., 2019. Data preprocessing and feature selection for machine learning intrusion detection systems. *ICIC Express Lett*, *13*(2), pp.93-101. DOI: https://doi.org/10.24507/icicel.13.02.93.

[14] Tong, Y., Lu, K., Yang, Y., Li, J., Lin, Y., Wu, D., Yang, A., Li, Y., Yu, S. and Qian, J., 2020. Can natural language processing help differentiate inflammatory intestinal diseases in China? Models applying random forest and convolutional neural network approaches. *BMC Medical Informatics and Decision Making*, *20*, pp.1-9. DOI: https://doi.org/10.1186/s12911-020-01277-w.

[15] Cho, E., Chang, T.W. and Hwang, G., 2022. Data preprocessing combination to improve the performance of quality classification in the manufacturing process. *Electronics*, *11*(3), p.477. DOI: https://doi.org/10.3390/electronics11030477.

[16] Hussain, S., Mustafa, M.W., Jumani, T.A., Baloch, S.K., Alotaibi, H., Khan, I. and Khan, A., 2021. A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. *Energy Reports*, *7*, pp.4425-4436. DOI: https://doi.org/10.1016/j.egyr.2021.07.008.

[17] Khanam, J.J. and Foo, S.Y., 2021. A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, *7*(4), pp.432-439. DOI: https://doi.org/10.1016/j.icte.2021.02.004.

[18] Misra, P. and Yadav, A.S., 2019, March. Impact of preprocessing methods on healthcare predictions. In *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*. DOI: https://doi.org/10.2139/ssrn.3349586.

[19] Kamran, M., 2021. A state of the art catboost-based T-distributed stochastic neighbor embedding technique to predict back-break at dewan cement limestone quarry. *Journal of Mining and Environment*, *12*(3), pp.679-691. DOI: https://doi.org/10.22044/JME.2021.11222.2104.

[20] Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S. and Sun, Y., 2021. Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. *Forests*, *12*(2), p.216. DOI: https://doi.org/10.3390/f12020216.

[21] Al Daoud, E., 2019. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, *13*(1), pp.6-10. DOI: doi.org/10.5281/zenodo.3607805.

[22] Agustyaningrum, C.I., Gata, W., Nurfalah, R., Radiyah, U. and Maulidah, M., 2020. Komparasi Algoritma Naive Bayes, Random Forest Dan Svm Untuk Memprediksi Niat Pembelanja Online. *Jurnal Informatika*, *20*(2). DOI: https://doi.org/10.30873/ji.v20i2.2402.

[23] Religia, Y., Pranoto, G.T. and Santosa, E.D., 2020. South German Credit Data Classification Using Random Forest Algorithm to Predict Bank Credit Receipts. *JISA (Jurnal Informatika dan Sains)*, *3*(2), pp.62-66. DOI: https://doi.org/10.31326/jisa.v3i2.837.

[24] Chairunisa, R. and Astuti, W., 2020. Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *4*(5), pp.805-812. DOI: https://doi.org/10.29207/resti.v4i5.2083.

[25] Sandy, W.K., Widodo, A.W. and Sari, Y.A., 2018. Penentuan Keaslian Tanda Tangan Menggunakan Shape Feature Extraction Techniques Dengan Metode Klasifikasi K Nearest Neighbor dan Mean Average Precision. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, *2548*, p.964X.

[26] Novianti, K.D.P., Setiawan, N.A. and Kusumawardani, S.S., 2015. Peningkatan Nilai Recall dan Precision pada Penelusuran Informasi Pustaka Berbasis Semantik (Studi Kasus: Sistem Informasi Ruang Referensi Jurusan Teknik Elektro dan Teknologi Informasi UGM). *Proceedings Konferensi Nasional Sistem dan Informatika (KNS&I)*.