Journal Management, Informatics, Software Engineering and Communication Technology
Vol. 5 No. 1, December (2026) pp. 110-115
Published by: Marq & Cha Institute
Journal Management, Informatics, Software Engineering and Communication Technology
Journal homepage: www.ejournal.marqchainstitute.or.id/index.php/jmirte

Predicting Social Assistance Recipient Eligibility Using the Explainable Boosting Machine Algorithm on Demographic Data

Silvy Handayani Maharaja[1], Hevlie Winda Nazry S [2]
[1,2] Program Studi Teknologi Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

Article Info

Article history

Received : Sep 2, 2026
Revised : Sep 28, 2026
Accepted : Oct 10, 2026

Keywords:

Explainable Boosting Machine;
Social Assistance;
Machine Learning;
Demographic Data;
Eligibility Prediction.

Abstract

This study aims to predict the eligibility of social assistance recipients using an Explainable Boosting Machine (EBM) algorithm based on demographic data. The main problem in social assistance distribution is inaccurate targeting, which leads to inclusion and exclusion errors due to simple and inaccurate selection methods. This study employed a quantitative approach with supervised learning on 500 demographic data sets covering variables such as income, employment, education, and number of dependents. The data underwent preprocessing, including labeling, transformation, and balancing using SMOTE, resulting in a total dataset of 924 data points. The model was then divided into 80% training data and 20% testing data. The results showed that the EBM model achieved very high performance, with accuracy, precision, recall, and F1-score values each reaching 100%. Furthermore, EBM was able to interpret the contribution of each variable in determining recipient eligibility. Thus, this model is not only accurate but also transparent, supporting more objective, effective, and accountable decision-making in social assistance distribution.

Corresponding Author:

Silvy Handayani Maharaja,
Program Studi Teknologi Informasi,
Universitas Muhammadiyah Sumatera Utara,
Jl. Kapten Mukhtar Basri No. 3, Medan, 20238, Indonesia.
Email : maharajahandayani11.11@gmail.com

This is an open access article under the CC BY-NC license.

1. Introduction

Poverty remains a major challenge for developing nations, including Indonesia, where segments of the population struggle to meet basic needs such as employment, adequate income, housing, education, and the ability to support their families (Pratiwi et al., 2022). Social assistance policies serve as a primary government instrument to reduce poverty rates through the proportional redistribution of resources. However, the effectiveness of social assistance distribution relies heavily on the accuracy of beneficiary selection criteria. Conventional selection mechanisms often manual or based on simple criteria are prone to subjective bias, administrative errors, and issues regarding inclusion errors (ineligible recipients) and exclusion errors (eligible individuals being overlooked) (Sari et al., 2023). Such targeting inaccuracies lead to budgetary waste and erode public trust in government programs.

To address this, a data-driven approach to determining eligibility leveraging machine learning technology is required. Although machine learning algorithms such as Random Forest or Gradient

Boosting offer high predictive accuracy, they often function as "black-box" models, making it difficult to explain their decision-making processes to policymakers. This lack of transparency can create obstacles during audits and hinder public accountability. As a solution, this study employs the Explainable Boosting Machine (EBM) algorithm. EBM is an interpretable model based on generalized additive models (GAMs) that combines the high predictive power characteristic of boosting techniques with full transparency regarding the contribution of each independent variable (Caruana et al., 2015).

The innovation and novelty of this research lie in the application of EBM to model demographic and socioeconomic population data utilizing a sample of 500 secondary data points from Simalungun Regency to predict the eligibility status ("eligible" or "ineligible") of social assistance recipients. Unlike conventional machine learning approaches, EBM specifically offers interpretable visualizations, such as feature importance plots and individual conditional expectation curves. This enables policymakers to identify in a measurable and transparent manner the most dominant demographic variables (such as income, number of dependents, or employment status) that influence eligibility decisions.

The primary objective of this study is to model and evaluate the application of the Explainable Boosting Machine (EBM) algorithm in predicting social assistance eligibility based on demographic data, while simultaneously quantifying the contribution of each variable to the decision outcome. It is anticipated that the proposed approach will not only reduce rates of inclusion and exclusion errors but also provide an objective foundation for government agencies to design accountable decision support systems (utilizing explainable AI).

2. Research Methodolgy

This study employs a quantitative approach based on supervised learning to analyze numerical and categorical demographic data for predicting the eligibility of social assistance recipients (Cronje, 2020). The Explainable Boosting Machine (EBM) approach is utilized to address the limitations of "black-box" models and traditional classification algorithms, which often fail to provide transparency regarding variables and their contributions. The research was conducted in a systematic, chronological manner, comprising data collection, preprocessing, EBM modeling, performance evaluation, and system testing. The chronological workflow of the study is illustrated in Figure 1.

Table 1. Stages and Detailed Activities of the Research Workflow

Research Stage	Detailed Activities & Sub-processes	Output / Result Generated
1. Data Collection	Observation and recording of population demographics.	Secondary dataset (500 samples of demographic & socio-economic data).
	Coordination with local government agencies in Simalungun Regency.	
2. Data Preprocessing	Data cleaning (handling missing values & outliers).	Cleaned dataset ready for training and evaluation (*train/test split*).
	Categorical encoding (one-hot / target encoding).	
	Dataset splitting (80% Training : 20% Testing).	
3. EBM Modeling & Training	Initialization of additive functions and residual calculation.	Trained *Explainable Boosting Machine* (EBM) model.
	Iterative boosting process (feature-by-feature learning).	

	Feature interaction extraction & hyperparameter tuning.	
4. Model Evaluation & Interpretation	Test set evaluation using a Confusion Matrix. Calculation of performance metrics: Accuracy, Precision, Recall, and F1-Score. Analysis of Feature Importance graphs and individual variable contributions.	Precision performance metrics and feature weight/influence contributions.
5.System Implementation & Testing	Integration of the model into a Decision Support System (DSS). Web-based / Python API dashboard interface development. • User access role testing (*Admin* & *Officer*).	

Data Collection and Research Variables

The data used consist of secondary data comprising a sample of 500 demographic records obtained from the local government of Simalungun Regency. Data collection was conducted through observation and coordination with sub-district and village authorities to ensure the validity of the demographic variables (Fryer & Dinsmore, 2020). The research variables are categorized into independent and dependent variables, as outlined in the Table.

Table 2. Research Variables for Social Assistance Eligibility Prediction

Variable Type	Variable Name	Data Type	Description / Value Range
Independent (Features)	Occupation Type	Categorical	Primary work sector of the resident
	Monthly Income	Numeric	Monthly income amount (IDR)
	Number of Dependents	Numeric	Number of dependent family members
	Education Level	Ordinal	Highest level of completed education
	Housing Condition	Categorical	Livability and homeownership status
Dependent (Target)	Eligibility Status	Binary	1 = Eligible, 0 = Ineligible

Data Preprocessing

Before being fed into the EBM model, the raw dataset undergoes preprocessing stages to ensure data quality. This process includes data cleaning to handle missing values, categorical encoding (using one-hot or target encoding), and dataset splitting. The 500-sample dataset is randomly divided into a training set (80% or 400 samples) and a testing set (20% or 100 samples).

Performance Evaluation and System Testing Procedures

Model performance is evaluated by testing the model against the testing set (20% of the data) using a confusion matrix. Performance metrics specifically Accuracy, Precision, Recall, and F1-Score are employed to ensure the model maintains precision in minimizing inclusion and exclusion errors.

The developed decision support system is implemented using Python and a web-based framework. The system architecture supports distinct roles for Officers (responsible for inputting demographic data and processing predictions) and Administrators (responsible for database management and model updates). The system's output provides not only a classification of "Eligible" or "Ineligible" status but also displays contribution weights (feature importance) and eligibility probabilities to ensure decision transparency.

3. Results And Discussions

This section presents the experimental results of the Explainable Boosting Machine (EBM) modeling, performance evaluation, and a comprehensive discussion of the findings (Grieshaber, 2020).

Dataset and Data Preprocessing

The dataset was obtained directly from the Serbelawan Village Office, Dolok Batu Nanggar District, Simalungun Regency. The raw dataset consists of 500 population data samples. Prior to modeling, preprocessing was conducted, which included target labeling based on income and dependency criteria, categorical variable encoding, and class imbalance handling.

The initial data showed a significant class imbalance, where the "Ineligible" class dominated with 462 samples (92.4%), while the "Eligible" class comprised only 38 samples (7.6%). To prevent prediction bias toward the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. This balancing process produced a total of 924 data samples, evenly distributed with 462 samples per class. The balanced dataset was then split using an 80:20 ratio, resulting in 739 training samples and 185 testing samples, as detailed in Table 3.

Table 3. Dataset Distribution Before and After SMOTE Processing

Dataset Condition	Ineligible Class	Eligible Class	Total Data	Training Set (80%)	Testing Set (20%)
Initial Raw Data	462	38	500	-	-
After SMOTE	462	462	924	739	185

Performance Evaluation of the Explainable Boosting Machine

Model training was conducted using the balanced dataset. The classification performance was evaluated using a Confusion Matrix on the test data. The confusion matrix yielded flawless classification results with 93 True Negative (TN) samples, 8 True Positive (TP) samples, and 0 False Positive (FP) and False Negative (FN) samples.

Based on the confusion matrix, performance metrics were calculated using Equations (1) through (4):

Evaluation Metric	Mathematical Formula	Substitution & Calculation	Result / Value
Accuracy	(TP + TN) / (TP + TN + FP + FN)	(8 + 93) / (8 + 93 + 0 + 0) = 101 / 101 = 1.00	1.00 (100%)
Precision	TP / (TP + FP)	8 / (8 + 0) = 8 / 8 = 1.00	1.00 (100%)
Recall	TP / (TP + FN)	8 / (8 + 0) = 8 / 8 = 1.00	1.00 (100%)
F1-Score	2 × (Precision × Recall) / (Precision + Recall)	2 × (1.00 × 1.00) / (1.00 + 1.00) = 2 × 1.00 / 2.00 = 1.00	1.00 (100%)

The performance evaluation metrics for each class and the overall averages are summarized in Table 4.

Table 4. EBM Model Evaluation Performance Metrics

Classification / Class	Precision	Recall	F1-Score	Support
Ineligible	1.00	1.00	1.00	93
Eligible	1.00	1.00	1.00	8

Accuracy	-	-	1.00	101
Macro Average	1.00	1.00	1.00	101
Weighted Average	1.00	1.00	1.00	101

Discussion and Novelty Analysis

The experimental results demonstrate that the EBM algorithm achieved optimal performance with 100% accuracy, precision, recall, and F1-score across both non-SMOTE and SMOTE scenarios. This indicates that the EBM model effectively learned the underlying decision boundaries.

The perfect classification performance is attributed to the distinct decision boundaries inherent in the demographic data of the study location, where income thresholds and dependency ratios are linearly separable. Nevertheless, applying SMOTE remains crucial to ensure the model avoids majority class bias when generalized to newer, more complex demographic datasets.

The main advantage of EBM over traditional black-box models (such as Random Forest or XGBoost) lies in the transparency of its decision-making process. Through EBM's additive functions, individual feature contributions were extracted explicitly:

1. Monthly Income and Number of Dependents emerged as the primary feature importance contributors determining eligibility status.
2. Secondary variables such as Occupation Type and Housing Condition provided fine-tuning effects on the final decision probabilities.

Implementation of the Decision Support System

To apply the empirical findings in practice, the trained EBM model (serialized as a .pkl file) was integrated into a web-based Decision Support System (DSS). The system allows village administrators and authorized officers to input demographic data and receive predictions ("Eligible" or "Ineligible"), accompanied by probability scores and feature contribution graphs, thereby promoting accountability and transparency in public assistance distribution.

4. Conclusion

The Explainable Boosting Machine (EBM) algorithm based on demographic data was successfully developed and implemented to predict social assistance eligibility status ("Eligible" and "Ineligible"). The experimental pipeline—comprising target labeling, categorical encoding, class balancing via SMOTE, and an 80:20 train-test dataset split—produced optimal classification performance. The EBM model achieved 100% accuracy, precision, recall, and F1-score with zero False Positives or False Negatives on the testing set. The consistent performance observed across both SMOTE and non-SMOTE scenarios indicates that the population dataset from the study site features clear and distinct decision boundaries, although SMOTE remains essential for preventing majority class bias. In practical terms, the decision support system built upon the trained EBM model effectively streamlines and structures the decision-making process for distributing social assistance at the village level, ensuring both objectivity and transparency. However, the flawless performance metrics highlight a research limitation concerning dataset complexity. Future studies should evaluate the model's generalization capabilities across larger, more dynamic, and highly heterogeneous demographic datasets, as well as incorporate automated real-time model updating within the system's architecture.

References

Alahmadi, R., Almujibah, H., Alotaibi, S., Elshekh, A. E. A., Alsharif, M., & Bakri, M. (2023). Explainable Boosting Machine: A Contemporary Glass-Box Model to Analyze Work Zone-Related Road Traffic Crashes. *Safety*, *9*(4), 1–15. https://doi.org/10.3390/safety9040083

Arslan, A. K., Yagin, F. H., Algarni, A., Al-hashem, F., & Paolo, L. (2024).

Combining the Strengths of the Explainable Boosting Machine and Metabolomics Approaches for Biomarker Discovery in Acute Myocardial Infarction.

Artha, D. P., & Misdawita. (2023). Ekonomi dan Kebijakan Publik Indonesia Keywords. *Jurnal Ekonomi Dan Kebijakan Publik Indonesia*, *10*(1), 65–80.

Damuri, A., Riyanto, U., Rusdianto, H., & Aminudin, M. (2021). Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako. *JURIKOM (Jurnal Riset Komputer)*, *8*(6), 219. https://doi.org/10.30865/jurikom.v8i6.3655

Jatipaningrum, M. T., Azhari, S. E., & Suryowati, K. (2022). Pengelompokan Kabupaten Dan Kota Di Provinsi Jawa Timur Berdasarkan Tingkat Kesejahteraan Dengan Metode K-Means Dan Density-Based Spatial Clustering Of Applications With Noise. *Jurnal Derivat: Jurnal Matematika Dan Pendidikan Matematika*, *9*(1), 70–81. https://doi.org/10.31316/j.derivat.v9i1.2832

Jurnal, H., Fathurohman FKIP, A., & Fisika, P. (2021). Machine Learning Untuk Pendidikan: Mengapa Dan Bagaimana. *Jurnal Informatika Dan Teknologi Komputer*, *1*(3), 57–62.

Khattak, A., Zhang, J., Chan, P. W., Chen, F., & Almujibah, H. (2024).

Explainable Boosting Machine: A Contemporary Glass-Box Strategy for the Assessment of Wind Shear Severity in the Runway Vicinity Based on the Doppler Light Detection and Ranging Data. *Atmosphere*, *15*(1), 1–20. https://doi.org/10.3390/atmos15010020

Kiss, O., Baker, F. C., Palovics, R., Dooley, E. E., Pettee Gabriel, K., & Nagata, J.M. (2024). Using explainable machine learning and fitbit data to investigate predictors of adolescent obesity. *Scientific Reports*, *14*(1), 1–13. https://doi.org/10.1038/s41598-024-60811-2

Maxwell, A. E., Sharma, M., & Donaldson, K. A. (2021). Explainable boosting machines for slope failure spatial predictive modeling. *Remote Sensing*,

13(24). https://doi.org/10.3390/rs13244991

Normah, Rifai, B., Vambudi, S., & Maulana, R. (2022). Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE. *Jurnal Teknik Komputer AMIK BSI*, *8*(2), 174–180. https://doi.org/10.31294/jtk.v4i2

Of, E., Assistance, S., Eligibility, R., On, B., Of, N., Using, D., Fuzzy, T. H. E., & Approach, M. (2025). *Evaluasi Kelayakan Penerima Bantuan Sosial BerdasarkanPenghasilan Dan Jumlah Tanggungan Dengan Pendekatan FuzzyMamdani*. *1*(2), 67–78.

Perhan, A. I., Yustiana, I., & Sanjaya, I. (2025). Implementation of Machine Learning Using Decision Tree Method for Social Assistance Recipient Classification. *Bit-Tech*, *8*(1), 901–909. https://doi.org/10.32877/bt.v8i1.2755

Puspitasari, N., & Sitohang, F. K. S. (2023). Pengolahan Data Demografi Penduduk Menggunakan Microsoft Excel di Kelurahan Sempaja Timur, Kota Samarinda. *JAPI (Jurnal Akses Pengabdian Indonesia)*, *7*(3), 181–190. https://doi.org/10.33366/japi.v7i3.3774

Setya, F., Sri, I., Zahra, M., Kristian, M., Jasmina, Z., Studi Administrasi Pembangunan Negara, P., & STIA LAN Bandung, P. (2022). "My Nutera" Di Kelurahan Tamansari Bandung Digitalization of Equivalent Social Assistance With Innovation "My Nutera" in Tamansari Village, Bandung. *SeTIA Mengabdi-Jurnal Pengabdian Kepada Masyarakat*, *3*(3), 55–62.

Sihombing, P. R., Muslianti, D., & Yunita. (2022). RETRACTION NOTICE TO "Apakah Dana Desa dan Fungsi Belanja APBD Mampu Mengatasi Kemiskinan di Indonesia?" *Jurnal Ekonomi Dan Statistik Indonesia*, *2*(2), 236–243. https://doi.org/10.11594/jesi.02.02.12

Sri Wahyuni, W., & Prima Putra, F. (2025). Analisis Karakteristik Rumah Tangga Miskin Penerima Program Keluarga Harapan dan Penerima Keluarga Harapan Sekaligus Bantuan Pangan Non-Tunai di Indonesia. *Jurnal Informatika Ekonomi Bisnis*, *7*(2), 141–147. https://doi.org/10.37034/infeb.v7i2.1112

Suyasa, I. M. G., Retnosari, L., & Taufiq, N. (2022). Tepatkah Program Bantuan Sosial dalam Pengentasan Kemiskinan di Indonesia? *Salus Cultura: Jurnal Pembangunan Manusia Dan Kebudayaan*, *4*(1), 42–52. https://doi.org/10.55480/saluscultura.v4i1.156

Wicaksono, H., Riyanto, A., Darmawan, R., & Hidayat, M. F. (2025). *Explainable*