

POVERTY RISK PREDICTION BASED ON SOCIOECONOMIC FACTORS USING MACHINE LEARNING APPROACH

Nur Farhana Adibah, Balkiah Moktar

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perlis Branch, Malaysia

e-mail: farhanazawari03@gmail.com, balkiah@uitm.edu.my

Abstract

Poverty remains a persistent socio-economic challenge in Malaysia, affecting access to basic needs such as healthcare, education, and employment. Traditional measurement methods like the Poverty Line Income (PLI) often oversimplify poverty classification by relying solely on income, ignoring multidimensional factors that influence well-being. This study aims to develop a predictive model for assessing poverty risk using supervised machine learning techniques based on a wider range of socioeconomic factors. A secondary dataset involving 635 households from eight districts in Terengganu was analyzed. Attributes such as age, income, education, occupation, health, gender, marital status, and savings were considered. Data preprocessing, including cleaning and transformation, was conducted, followed by feature selection using the Information Gain method. Four machine learning classifiers, such as Logistic Regression, Random Forest, J48, and Logit Boost, were evaluated using 10-fold cross-validation and a 70:30 training-testing split. Logistic Regression achieved the best overall performance with 99.06% accuracy in cross-validation and 98.42% in data splitting, along with superior precision, recall, and F1-score values. Age was identified as the most significant factor of poverty risk, followed by income and occupation. These results demonstrate that Logistic Regression is a reliable, interpretable, and stable model for classifying poverty risk using structured socioeconomic data. Although the study is geographically limited, its findings support the potential of machine learning as a data-driven tool in social policy development. This approach may enhance poverty targeting strategies and contribute to achieving Malaysia's Sustainable Development Goals (SDG) 1: No Poverty and SDG 2: Zero Hunger.

Keywords: *Poverty Risk, Machine Learning, Socioeconomic Factors, Logistic Regression, Information Gain, Predictive Modelling*

1. Introduction

Poverty remains a serious concern in Malaysia, influenced by factors such as income, education, occupation, and health. Traditional tools like the Poverty Line Income (PLI) often fail to capture the multidimensional nature of poverty. This study introduces a machine learning approach to classify poverty risk more effectively by analyzing household-level attributes. The main objective is to identify a predictive model for assessing the poverty risk based on socioeconomic factors by using supervised machine learning techniques. Specifically, the study aims to identify poverty risk factors based on socioeconomic factors such as gender, age, marital status, occupation, education, health, income, and savings, and to compare the performance of machine learning models. The secondary dataset of 635 Muslim households from eight districts in Terengganu by Rashid et al. (2021) was used to get the insights and contribute to a deeper understanding of poverty risk factors in the Malaysian context and support more data-informed intervention strategies.

2. Methodology

This study utilized a supervised machine learning approach, beginning with feature selection, with Information Gain was used to rank the features based on their predictive value using WEKA software. Four classification models, namely Logistic Regression, Random Forest, J48, and Logit Boost, were trained and evaluated using two validation methods, 10-fold cross-validation and a 70:30 train-test

splitting. Performance of the model was measured by using Accuracy, Precision, Recall, and F1-Score (Moscato et al., 2021).

3. Results / Findings

Feature importance analysis using Information Gain revealed that age was the most influential factor, followed by income, occupation, and health. These results not only align with findings from previous research but also highlight age as a newly emphasized predictor in this context. Results also confirmed that Logistic Regression was the most outperformed classifier for poverty risk prediction, followed by Random Forest, J48, and Logit Boost across all performance metrics. Using 10-fold cross-validation, it achieved an accuracy of 99.06% and 98.42% with a 70:30 train-test split. It also recorded the highest precision, recall, and F1-score values (up to 0.991), indicating strong consistency in detecting poverty cases.. The model remained stable even when features were reduced, further proving its robustness. While Random Forest and Logit Boost also showed good performance, J48 consistently had lower accuracy and generalization, especially under data splitting. Overall, the findings validate Logistic Regression as a promising model for classifying poverty risk using structured socioeconomic data

4. Novelty of Research / Product / Project

This study introduces the novel application of supervised machine learning to poverty prediction within the Malaysian context. Unlike traditional income-based classifications, the model incorporates multidimensional socioeconomic features to improve prediction accuracy. The integration of Information Gain for feature selection enhances interpretability, which is crucial for policymaking. While previous studies focused on ensemble models like Random Forest, this research demonstrates that a simpler model like Logistic Regression can outperform others in both accuracy and explainability. This makes the model more suitable for real-world implementation in government and social welfare systems. The approach aligns with national priorities such as RMK13 and SDG targets, supporting the digital transformation of social data analytics (Economic Planning Unit, 2019).

5. Conclusion or Future Work

Logistic Regression was found to be the most accurate and interpretable model for predicting poverty risk. Future work may include expanding the study to other states, incorporating geographic or real-time data, and exploring deep learning models for improved scalability and accuracy.

REFERENCE LIST

- Economic Planning Unit. (2019). Shared Prosperity Vision 2030. <https://ekonomi.gov.my/sites/default/files/202002/Shared%20Prosperity%20Vision%202030.pdf>
- Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165(113986). <https://doi.org/10.1016/j.eswa.2020.113986>
- Rashid, N. K. A., Sulaiman, N. F. C., Anang, Z., Possumah, B. T., Muhamad, S., Sa'at, N. H., Hassan, F. A., & Wahid, H. (2021). Data format Raw Analysed Descriptive Statistical Parameters for data collection Sustainable consumption, income and consumption expenditure. <https://doi.org/10.17632/snw2pbpbn2.1>