1 NO POVERTY

THE EFFECTIVENESS OF MACHINE LEARNING CLASSIFICATION MODEL IN ANALYZING POVERTY FACTORS IN INDONESIA

School of Computer Science - Data Science
BINA NUSANTARA UNIVERSITY

DTSC6006001 - LB09
1. Crysantha Monica Lim - 2602090076
2. Maxcell Rimba - 2602110046
3. Andi Izzat Zaky Ashari - 2501977172
4. Teofilus San Prasetya - 2602081922

INTRODUCTION

BACKGROUND

Poverty poses a serious challenge worldwide, including in Indonesia, which is the focal point of Sustainable Development Goal number 1, "No Poverty." Indonesia is among the 100 poorest countries globally. According to Global Finance, Indonesia ranks 91st among the world's poorest countries in 2023. Based on data from the Central Statistics Agency (BPS), the percentage of the population living in poverty as of March 2023 reached 9.36%. The number of people living in poverty in March 2023 amounted to 25.90 million.

BENEFITS

- This research aims to implement the most effective Machine Learning classification model in analyzing economic and social factors influencing the poverty rate in Indonesia.
- The implementation of the classification model can serve as a powerful predictive tool to aid in identifying cities with high poverty rates and designing appropriate and effective measures or actions to help alleviate poverty levels.

PROBLEM LIMITATIONS

This research will be focused on analyzing poverty rates in various cities in Indonesia, with limitations on specific economic and social factors identified as predictive variables. The classification model will be restricted to identifying patterns and relationships, not to provide a final solution for poverty-related issues.

METHODOLOGY

DATA

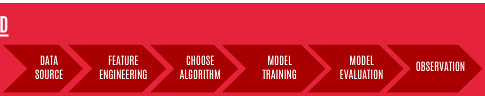
DESCRIPTION:

This dataset encompasses a broader understanding of poverty beyond mere financial inadequacy, involving considerations of health, malnutrition, lack of clean water, electricity, low-quality employment, and insufficient education. It employs the Alkire Foster (AF) method, generating the Global Multidimensional Poverty Index (MPI).

The dataset provides information on multidimensional poverty levels in various countries for the year 2017, with a focus on regional levels within each country. Key variables in this dataset include:

1. ISO (Unique ID for country): Unique identification for each country.
2. Country (country name): Name of the respective country.
3. Sub-national region: Information about regions within the country (sub-national) under investigation.
4. World Region: Classification of the global region to which the country belongs.
5. MPI National:Value of the national-level multidimensional poverty index.
6. MPI Regional:Value of the regional-level multidimensional poverty index within the country.
7. Headcount Ratio Regional: Percentage of the population considered poor at the regional level.
8. Intensity of Deprivation Regional:Average distance below the poverty line for those considered poor at the regional level.

METHOD


DATA SOURCE → FEATURE ENGINEERING → CHOOSE ALGORITHM → MODEL TRAINING → MODEL EVALUATION → OBSERVATION

- **Data Source**

The dataset used was downloaded from Kaggle at https://www.kaggle.com/datasets/ophi/mpi.

- **Feature Engineering**

The feature engineering stage involves assigning values to the "Label" column based on the following conditions: if the value in the "MPI" column is less than 0.10, "No Poverty" is assigned; if the value is greater than 0.10, "Poverty" is assigned; and if neither condition is met, an empty value ('') is assigned to the "Label" column. This code aids in classifying data based on poverty levels measured by the MPI.

- **Machine Learning Algorithm Selection**

1. Logistic Regression $Y = \beta_0 + \beta_1 X + \varepsilon$
2. KNN
3. Naive Bayes $P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)}$
4. Decision Tree
5. Ensamble Model : Gradient Boosting Classifier $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$

- **Model Training**

In the first stage, the dataset is used to train a machine-learning classification model. This model learns intricate associations between features. The Alkire Foster (AF) method is applied, with the Global Multidimensional Poverty Index (MPI) as the target variable. The training process fine-tunes the model parameters iteratively to minimize the difference between predicted and actual poverty levels.

- **Model Evaluation**

After training the model, we thoroughly evaluate how well it works. We use metrics like accuracy, precision, recall, and F1 score to measure its effectiveness. If the model makes any mistakes, we look into them to understand where it can be better.

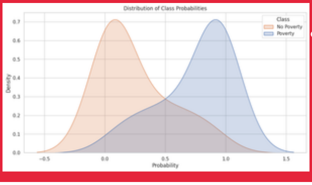
- **Observation**

The last step is to look at the results from the model. We may need to keep an eye on the model and make improvements over time to make sure it stays accurate and useful as conditions change.

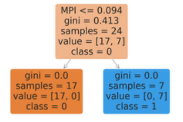
RESULT

MODEL PREVIEW

- Naive Bayes



- Decision Tree



MODEL COMPARISON (notes : n.p. = no poverty)

Model	accuracy	precision		recall		F1-score	
		n.p.	poverty	n.p.	poverty	n.p.	poverty
Logistic Regression	1.00	1.00	1.00	1.00	1.00	1.00	1.00
KNN	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Naive Bayes	0.89	0.86	1.00	1.00	0.67	0.92	0.80
Decision Tree	0.89	0.86	1.00	1.00	0.67	0.92	0.80
Gradient Boosting	0.90	0.88	1.00	1.00	0.67	0.93	0.80

SUMMARY

The table presents the performance metrics of different machine learning models in classifying instances into "poverty" and "no poverty" categories.

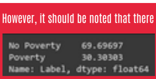
- **Logistic Regression and KNN:**

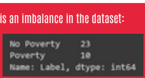
Both Logistic Regression and KNN models achieved perfect accuracy (1.00), precision, recall, and F1-score for both "no poverty" and "poverty" classes. These models demonstrate robust performance in distinguishing between the two categories.

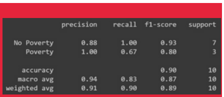
- **Naive Bayes, Decision Tree, and Gradient Boosting:**

  - Naive Bayes, Decision Tree, and Gradient Boosting models also performed well but showed slightly lower accuracy (0.89-0.90) compared to Logistic Regression and KNN.
  - While these models had high precision, recall, and F1-score for the "no poverty" class, there was a slight decrease in recall for the "poverty" class (0.67-1.00).

However, it should be noted that there is an imbalance in the dataset:


```
No Poverty    69.69697
Poverty       30.30303
Name: Label, dtype: float64
```


```
No Poverty    23
Poverty       10
Name: Label, dtype: int64
```


```
              precision    recall  f1-score   support

  No Poverty       0.88      1.00      0.93         7
     Poverty       1.00      0.67      0.80         3

    accuracy                           0.90        10
   macro avg       0.94      0.83      0.87        10
weighted avg       0.91      0.90      0.89        10
```

Despite attempting oversampling, the model results remain less reliable due to an accuracy of 1.00. This occurrence may be attributed to the limited dataset, consisting of only 33 samples representing Indonesian data.

Therefore, the most trustworthy model in terms of accuracy and precision is the Gradient Boosting Classifier. Even after addressing the sample imbalance, this model has proven to be more robust in handling the dataset characteristics and demonstrates a more realistic evaluation of poverty classification.