In the Multiverse of Poverty: A Näive Bayes Approach to Multidimensional Poverty Classification

AI201 Mini Project

Jamlech Iram Gojo Cruz
National Graduate School of Engineering
University of the Philippines
Quezon City, Philippines
jngojocruz@up.edu.ph

Gyenn Neil Ibo
National Graduate School of Engineering
University of the Philippines
Quezon City, Philippines
jngojocruz@up.edu.ph

2 out of 5

individuals are classified as poor

INTRODUCTION

Poverty, typically defined by income, is a complex issue influenced by several factors

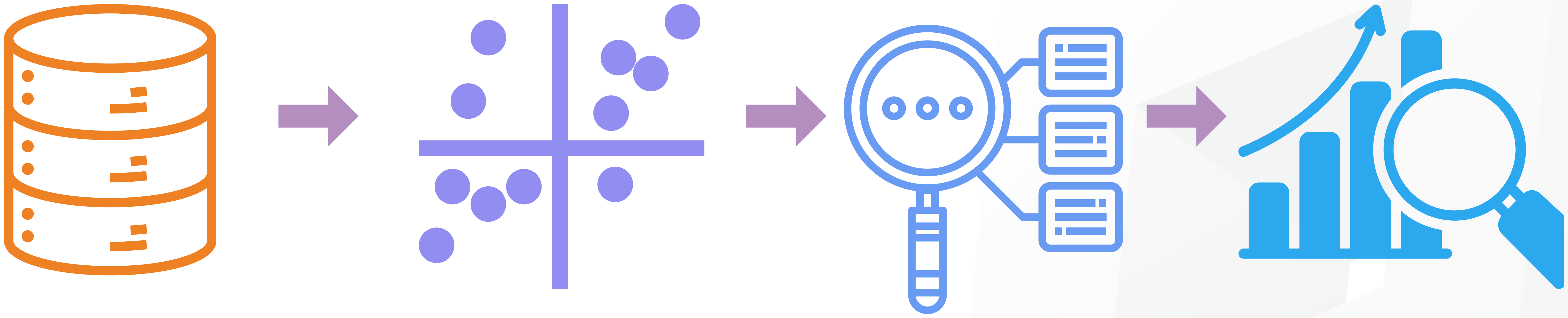
Mismatch between the outlined criteria in the national targeting system for the poorest families and the actual beneficiaries

Gap in the application of AI projects specifically targeting the UN SDG's No Poverty

OBJECTIVE

Develop a Naive Bayes model classifying individuals as either poor or non-poor while investigating the key factors contributing to poverty.

METHODS



Data Preprocessing

Feature Engineering

Training

Evaluation

METHODS

1. Data Processing

- **Description**
  - Dataset from a poverty prediction competition from datasciencecapstone.org
  - Comprising 58 features derived from the survey, the dataset includes 12,600 labeled samples and 8,400 unlabeled samples.
- **Cleaning**
  - Rows and columns containing mostly null values are removed.
- **Dealing with Imbalanced Data**
  - 7745 samples for poor, and 4323 for non-poor; undersampling is applied.
- **Data Partitioning**
  - 80% Training, 10% Validation, 10% Testing

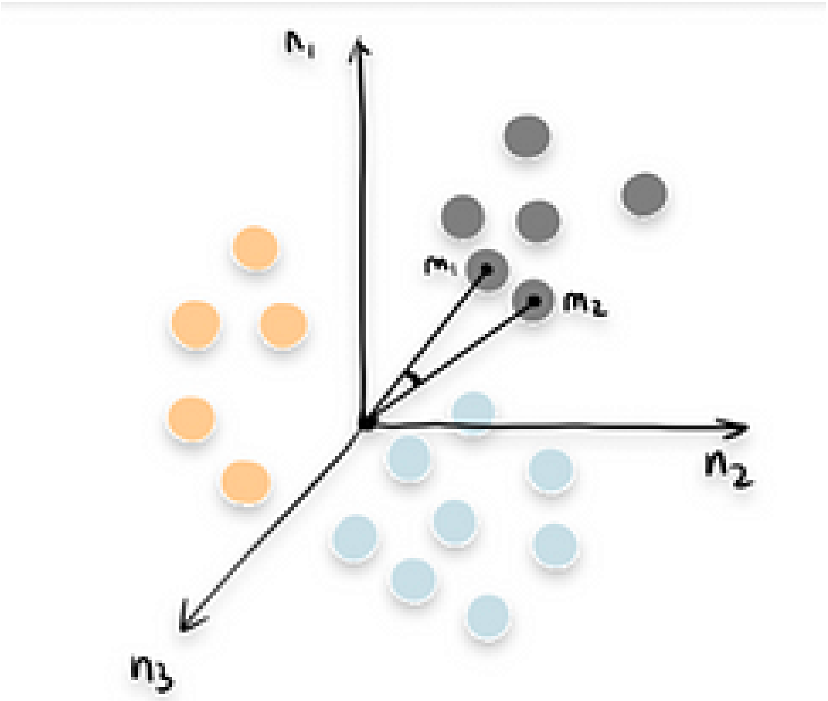
METHODS

2. Feature Engineering

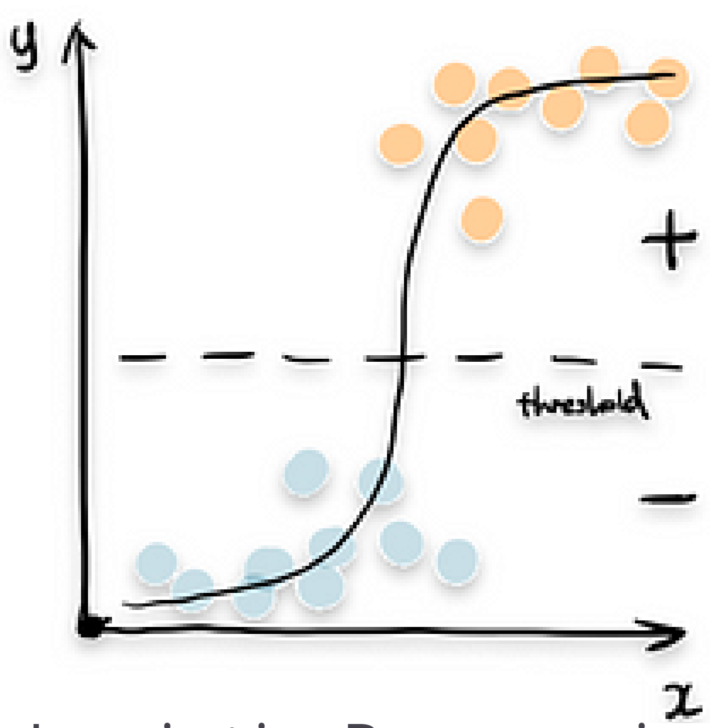
- **Numerical to Categorical**
  - Numerical columns were transformed into categories by grouping them into bins (e.g. on age)
- **Aggregation**
  - Reduced the number of unique values in some columns; some values are grouped together to balance representation.

METHODS

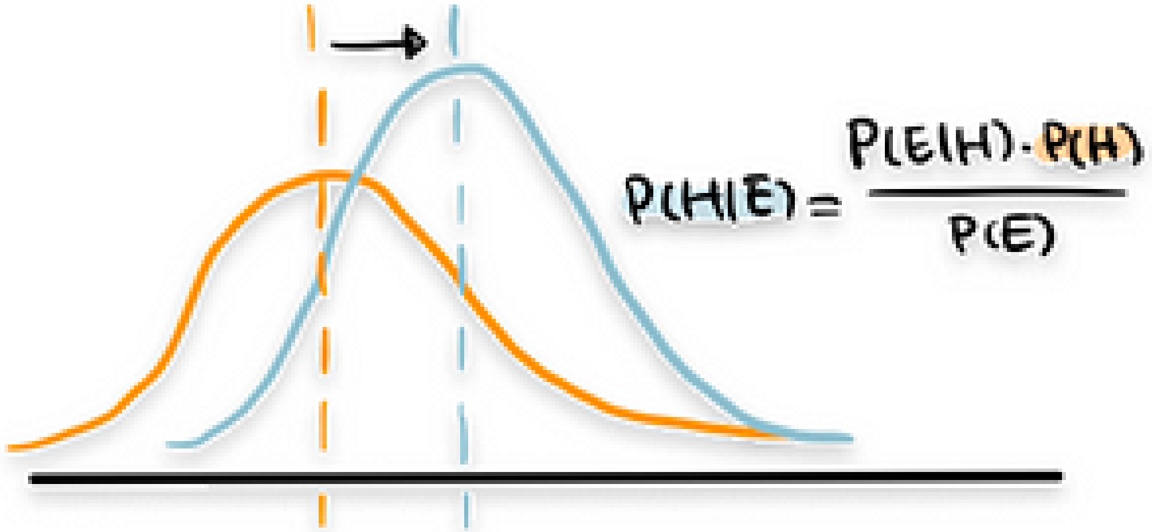
3. Machine Learning Models



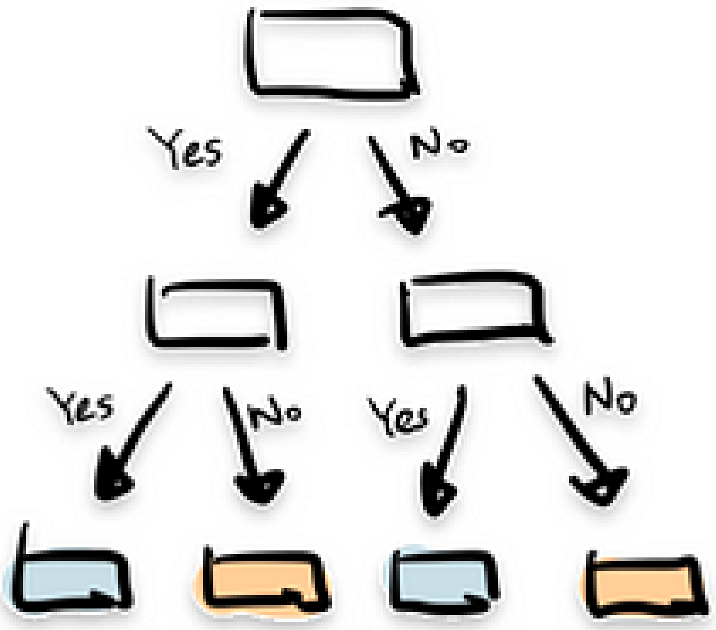
k-Nearest Neighbors



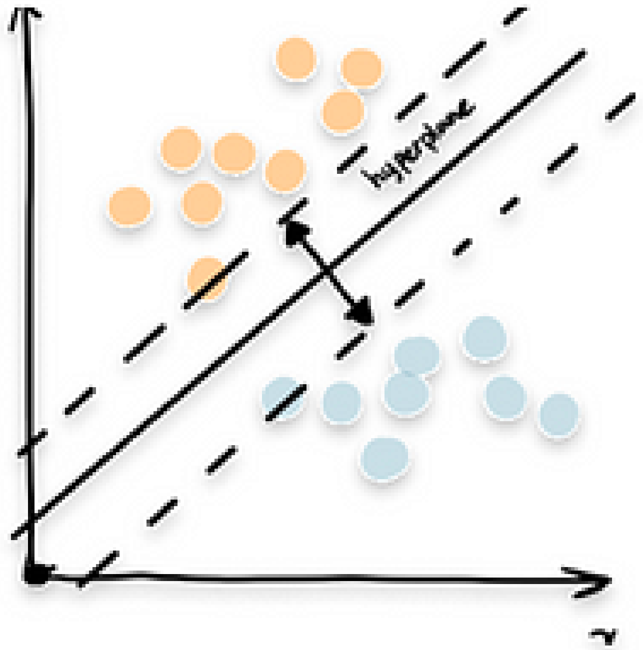
Logistic Regression



Naive Bayes



Decision Tree



Support Vector Machines

IMAGES: Destin Gong (2022), Top 6 Machine Learning Algorithms for Classification
https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

METHODS

4. Training

- **Classifier Evaluation**
  - The k-fold cross-validation method is used against training data to evaluate the classifier
  - k = {5, 10, 15, 20, 25, 30}
  - Tested across different experimental setup
- **Feature Selection**
  - Leave-one-out method was used
  - The set of features that boosted the evaluation metrics are retained, the rest are removed

METHODS

5. Evaluation Metrics

- **Accuracy**
  - For balanced dataset
- **Precision**
- **Recall**
- **Specificity**
- **F2-Score**
  - wrongly identifying a poor person as not-poor (False Negative) is a worse error than wrongly predicting a not-poor person as poor (False Positive)
- **Balanced Accuracy**
  - For imbalanced dataset

RESULTS

- **Experimental setup:**
  a. **Experiment 1:** Imbalanced, with numerical & categorical features
  b. **Experiment 2:** Balanced (undersampling), numerical & categorical with feature selection (top 11)
  c. **Experiment 3:** Balanced (undersampling), all categorical with feature selection (top 10 features)
  d. **Experiment 4:** Balanced (undersampling), all categorical with feature engineering (top 14)
  e. **Experiment 5:** Imbalanced, all categorical with feature engineering (top 14)
  f. **Experiment 6:** Imbalanced, all categorical with feature engineering and feature selection (top 14)
  g. **Experiment 7:** Imbalanced, all categorical with feature engineering and feature selection based on balanced accuracy (top 18)

RESULTS

Table 1: Results of the K-Fold Cross Validation Experiments

		1	2	3	4	5	6	7
K								
5	**Accuracy**		68.98%	69.30%	69.47%			
	F2 Score	75.56%				79.92%	82.61%	79.42%
	Balanced Accuracy	67.83%						70.23%
10	**Accuracy**		70.25%	70.27%	70.88%			
	F2 Score	75.51%				79.21%	81.48%	79.15%
	Balanced Accuracy	67.10%						69.97%
20	**Accuracy**		68.49%	67.20%	67.48%			
	F2 Score	75.04%				79.04%	82.16%	79.03%
	Balanced Accuracy	67.38%						70.01%

RESULTS

Table 3: Top Features Indicating Poverty (using the leave-one-out method) based on Balanced Accuracy

Top	Feature	Decrease (LOOCV)
1	country	69.47%
2	is_urban	69.91%
3	education_level	71.06%
4	can_make_transaction	72.03%
5	num_shocks_last_year	72.06%
6	nonreg_active_mm_user	72.43%
7	reg_bank_acct	72.44%
8	active_mm_user	72.52%
9	borrowing_recency	72.56%
10	married	72.57%
11	borrowed_for_emergency_last_year	72.57%
12	income_friends_family_last_year	72.64%
13	literacy	72.70%
14	employment_type_last_year	72.71%
15	income_public_sector_last_year	72.71%
16	active_informal_nbfi_user	72.90%
17	phone_ownership	72.90%
18	can_call	73.09%

RESULTS

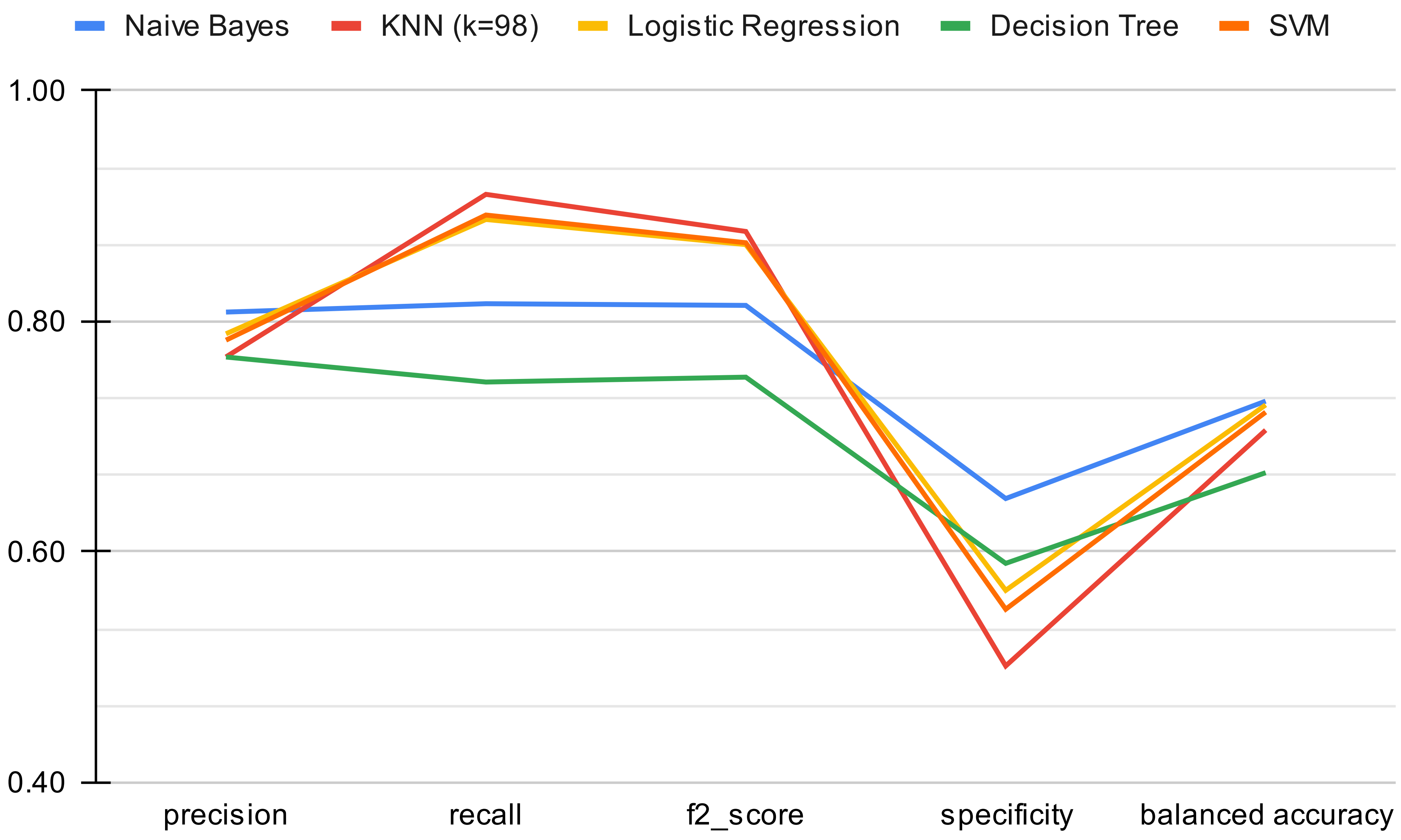


Table 2: Comparison of Different Classifiers

	precision	recall	f2_score	specificity	balanced accuracy
Naive Bayes	81%	82%	81%	65%	73%
KNN (k=98)	77%	91%	88%	50%	71%
Logistic Regression	79%	89%	87%	57%	73%
Decision Tree	77%	75%	75%	59%	67%
SVM	78%	89%	87%	55%	72%

RESULTS

In the final assessment, we evaluated the Naive Bayes classifier using unseen data. The results yielded a **balanced accuracy of 69%**, accompanied by **precision, recall, and F2-score metrics at 78%, 81%, and 81%**, respectively. These findings underscore the classifier's ability to generalize well to new, unseen data.

CONCLUSION

- **Data Preprocessing Techniques**
  - Involves cleaning, balancing, and partitioning a dataset from a poverty prediction competition.
  - Applies feature engineering, including numerical to categorical transformations and aggregation, to enhance model performance.
  - Data preprocessing has major impact in classifier evaluation
- **Key Poverty Indicators Revealed**
  - Identifies key indicators influencing poverty, including country, living area, marital status, and income.
  - Enhances policymakers' understanding of factors influencing poverty.
- **Contributions to Poverty Alleviation Discourse**
  - Highlights the potential of machine learning in addressing complex socio-economic challenges.

In the Multiverse of Poverty: A Näive Bayes Approach to Multidimensional Poverty Classification

AI201 Mini Project

Jamlech Iram Gojo Cruz
National Graduate School of Engineering
University of the Philippines
Quezon City, Philippines
jngojocruz@up.edu.ph

Gyenn Neil Ibo
National Graduate School of Engineering
University of the Philippines
Quezon City, Philippines
jngojocruz@up.edu.ph