REPORT ON POVERTY PREDICTION USING PSLM-2020 DATASET

Jawad Ahmad

What is the PSLM Survey?

The Pakistan Social and Living Standards Measurement (PSLM) Survey is a national survey that collects information about the lives of people in Pakistan. The PSLM-2020 dataset focuses on important topics like education, health, jobs, income, and how people live. It includes data from both cities and villages, covering thousands of families to represent the whole country. This dataset helps understand how much people earn, what jobs they do, and how many are living in poverty. It's very useful for creating plans and policies to improve the lives of people in Pakistan

Why Understanding Poverty Matters?

Poverty remains a critical global challenge, and understanding its determinants is essential for effective policy-making. The World Bank defines poverty as a state of deprivation where individuals lack the resources to meet their basic needs. Using this definition as a guiding principle, the PSLM-2020 dataset was explored to predict and estimate poverty levels. This report outlines the steps taken to process the dataset, prepare it for analysis, and implement predictive modeling.

DATA EXPLORATION AND CLEANING

Dataset Sections Used:

- **Employment Data (Section E):** Captures job status and income levels.
- **Housing Data (Section F1):** Includes information on living conditions.
- **WASH Data (Section F2):** Details access to water, sanitation, and hygiene.
- **Solid Waste Data (Section F3):** Provides waste management details.
- **Assets Data (Section G):** Lists household assets.
- **Durable Items Data (Section H):** Reports ownership of durable goods.
- **Education Data (Section C1):** Captures education levels and literacy rates.

Handling Missing Values

Missing values in income-related columns were replaced with 0 to simplify calculations and ensure that all rows had complete income data for analysis.

Additional Preprocessing Steps:

- Renamed column names from code-based labels to meaningful names for better interpretability.
- Selected relevant columns from various data files to retain only useful features.
- Standardized the 'Household_Code' column across datasets to ensure proper alignment.
- Merged relevant data files into a single consolidated dataframe for streamlined analysis.
- Normalized categorical variables to maintain data uniformity.

FEATURE ENGINEERING

Key Features Extracted:
Total Household Income: Summed income from all employment sources.
Housing Quality Index: Derived from housing conditions.
Asset Score: Based on the number of valuable household assets.
Sanitation Score: Evaluated based on access to clean water and waste management.
Ownership of Agricultural Land: Determines land ownership status and its impact on poverty.
Irrigation Status of Land: Assesses the availability of irrigation for agricultural land.
Water Treatment: Evaluates whether households have access to treated water for consumption.
Payment for Waste Services: Identifies households paying for waste disposal services, indicating financial capability.

TARGET VARIABLE

Poverty status is classified using the World Bank threshold **($2.15/day or $784.75/year).** Households below this threshold are labeled as 'Poor,' others as 'Non-Poor.'

MODEL TRAINING AND EVALUATION

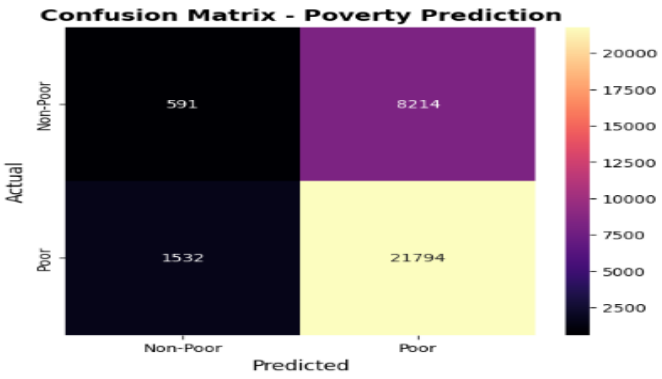
Random Forest Classifier: A robust ensemble learning method that captures nonlinear relationships in the data. The model was trained using an 80-20 train-test split with stratification to maintain class balance.
Feature Standardization: Applied using StandardScaler to normalize the feature values.
Hyperparameters: Default settings with random_state=42 for reproducibility.

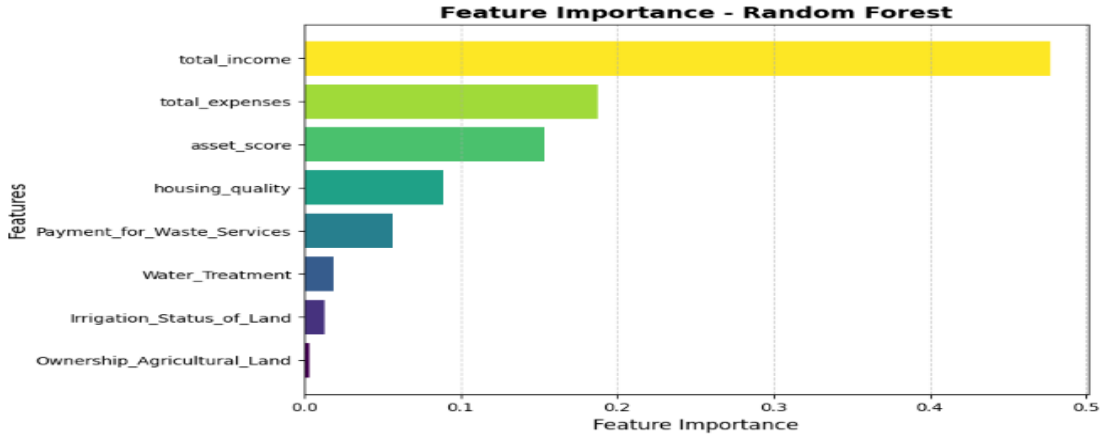
PERFORMANCE METRICS:

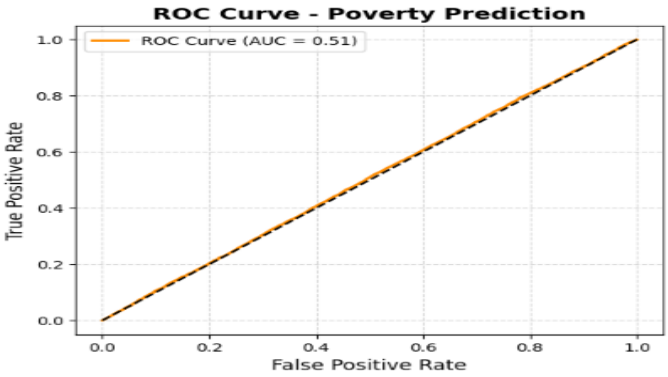
Accuracy: 69.67%
Classification Report: Precision, Recall, and F1-score indicate reasonable model performance in distinguishing poverty status.

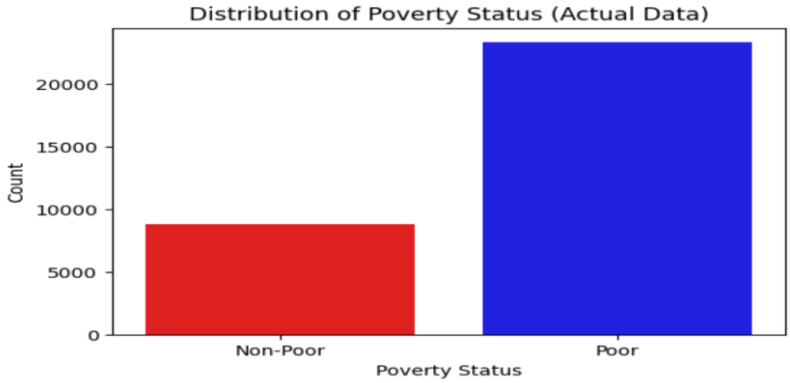
Confusion Matrix: Analyzed to understand misclassification patterns.

DATA VISUALIZATION


Confusion Matrix - Poverty Prediction
591
8214
1532
21794
Non-Poor
Poor
Actual
Non-Poor
Poor
Predicted
20000
17500
15000
12500
10000
7500
5000
2500


Feature Importance - Random Forest
total_income
total_expenses
asset_score
housing_quality
Payment_for_Waste_Services
Water_Treatment
Irrigation_Status_of_Land
Ownership_Agricultural_Land
Features
0.0
0.1
0.2
0.3
0.4
0.5
Feature Importance


ROC Curve - Poverty Prediction
ROC Curve (AUC = 0.51)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Distribution of Poverty Status (Actual Data)
Count
20000
15000
10000
5000
0
Non-Poor
Poor
Poverty Status