Capstone Project Concept Note and Implementation Plan

Project Title: Predicting Poverty Levels Using World Bank Indicators

Team Members

1. Sultan Mansour Raofi
2. Shabir Khairzad
3. Arezo Mohammadi

Concept Note

1. Project Overview

The capstone project entitled “Predicting Poverty Levels Using World Bank Indicators” is designed to tackle the significant challenge posed by insufficient or outdated poverty data, which hampers the effective allocation of resources by decision-makers. By leveraging socioeconomic and developmental indicators sourced from the World Bank, this initiative aims to create a machine learning model that can reliably forecast poverty levels in diverse regions. This endeavor is closely aligned with Sustainable Development Goal 1 (SDG 1): No Poverty, as it offers critical insights for identifying and focusing efforts on areas with elevated poverty rates. By generating precise predictions, the project can assist governments and organizations in prioritizing the necessary interventions to reduce poverty and foster economic development.**2.**

Objectives

The specific objectives of this project include:

- Developing a machine learning model that can predict poverty levels based on various World Bank development indicators.
- Analyzing socioeconomic factors to understand their correlation with poverty levels.
- Providing actionable insights to policymakers for efficient resource allocation to high-poverty regions.

Evaluating model performance through various metrics to ensure its reliability and effectiveness in real-world scenarios.

By achieving these objectives, the project will contribute significantly to addressing the identified problem of insufficient poverty data, enabling informed decision-making and targeted interventions to reduce poverty.

3. Background

Poverty is a significant challenge around the world, and many areas struggle with the absence of precise and up-to-date information that is crucial for creating effective policies. Organizations like the World Bank often depend on old data, which can result in resources being distributed ineffectively. While there are some efforts to update poverty indicators every few months, these updates are not comprehensive enough to capture the full picture. Utilizing machine learning can be a game-changer in this context, as it can process large amounts of data quickly, analyze it efficiently, and make predictions based on various factors. This approach can help fill the gaps in current methods used to evaluate poverty, leading to better-informed decisions and more effective solutions.

4. Methodology

To implement this project, we plan to utilize several machine learning techniques:

- Decision Trees: For initial analysis and feature importance evaluation.
- Random Forest: To enhance predictive accuracy and reduce overfitting.
- Gradient Boosting: For fine-tuning predictions by combining weak learners.

These algorithms are well-suited due to the structured and tabular nature of the dataset. They allow for interpretability, which is critical for policymakers to understand the influence of various indicators on poverty levels.

5. Architecture Design Diagram


World Development Indicators (WDI) Dataset
Timeline and Milestones
Programming Language: Python
Challenges and Ethical Considerations
Data Preprocessing
Other Tools: GitHub
Other Tools: Jupyter
Team Member 1: Data Cleaning
Data Cleaning Module
Normalization Module
Libraries: Pandas
Imputation Techniques
Feature Selection Block
Team Member 2: Feature Selection
Machine Learning Models Block
Libraries: Scikit-learn
Model Training & Evaluation Module
Team Member 3: Model Training
Results Analysis
Data Visualization Block
Model Performance Metrics Block
Libraries: Matplotlib/Seaborn

6. Data Sources

The foundation of this project is the World Development Indicators (WDI) dataset, which contains more than 1,600 different indicators collected over various years and across numerous countries. This dataset is vital because it provides important socioeconomic information, including gross domestic product (GDP), literacy rates, employment statistics, and access to healthcare services. The preprocessing phase will focus on selecting the necessary features, filling in any gaps in the data through imputation techniques, and normalizing the dataset to make it suitable for use with machine learning algorithms.

7. Literature Review

Previous research highlights the effectiveness of machine learning in poverty prediction. For instance, H. Zixi (2021) emphasizes the effectiveness of machine learning techniques in identifying poverty levels using socioeconomic indicators. The review by Aziza Usmanova underscores the variety of artificial intelligence implementation strategies for poverty prediction, demonstrating the utility of structured data analysis. Our project builds upon these foundational works by employing advanced machine learning techniques to provide a nuanced understanding of poverty dynamics.

Implementation Plan

1. Technology Stack

Programming Languages: Python

Libraries:

- Pandas (data manipulation)
- NumPy (numerical computations)
- Scikit-learn (machine learning)
- Matplotlib & Seaborn (data visualization)

Other Tools:

- Jupyter for interactive data analysis
- GitHub for version control

2. Timeline

Phase	Tasks	Start Date	End Date
Data Collection	Gather WDI dataset	Second week	Third week
Data Preprocessing	Clean data, handle missing values, normalization	Third week	Fourth week
Model Development	Build and train models	Fourth week	Fifth week
Model Training	Evaluate model and adjust parameters	Fifth week	Sixth week
Deployment	Implement and test web application	Sixth week	End of Sixth week

Task Distribution:

Arezo Mohammadi (Data Collection and Preprocessing)

Responsible for gathering data, cleaning, and preprocessing.

Shabir Khairzad (Model Development and Evaluation)

Focus on building and training machine learning models, evaluating performance.

Sultan Mansour Raofi (Deployment and Documentation)

Handle deployment of the final model and preparation of project documentation.

3. Milestones

- Completion of data collection
- Successful data preprocessing and feature extraction
- Development and training of machine learning models
- Achieving target model performance metrics
- Deployment of the application for user access

4. Challenges and Mitigations

- Data Quality: Ensure the integrity of data by implementing robust preprocessing methods.
- Model Performance: Regularly evaluate and retrain the model to incorporate new data.

- Technical Constraints: Ensure computational resources are sufficient by utilizing cloud-based services if necessary.

5. Ethical Considerations

Key ethical considerations involve safeguarding data privacy, mitigating bias in predictions, and assessing the potential effects on marginalized groups. Responsible data management and the incorporation of bias assessments within the model are crucial to guarantee equitable treatment among various demographics.

6. References

1. H. Zixi, "Poverty Prediction Through Machine Learning," 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), Hangzhou, China, 2021, pp. 314-324, doi: 10.1109/ECIT52743.2021.00073.
2. Usmanova, A. (2021). "Utilities of Artificial Intelligence in Poverty Prediction: A Review." Journal of Artificial Intelligence Research, 72, 301-317. doi:10.1613/jair.1.12853.
3. World Bank. World Development Indicators. [https://databank.worldbank.org/source/world-development-indicators]