

**Capstone Project Concept Note and Implementation Plan**

**Project Title:** **Classifying Socioeconomic Vulnerability Using the OSDG Community Dataset**

**Team Members:**

1. **Dugo Gadisa**
2. **Galata Waqwaya**
3. **Wabi Jifara**

**April 2025**

Contents

[1. Project Overview 1](#_Toc196147495)

[**2. Objectives** 1](#_Toc196147496)

[3. Background 1](#_Toc196147497)

[4. Methodology 2](#_Toc196147498)

[4.1. Architecture Design Diagram 2](#_Toc196147499)

[4.2. Data Sources 4](#_Toc196147510)

[5. Literature Review 4](#_Toc196147511)

[6. Implementation Plan 4](#_Toc196147512)

[6.1. Technology Stack 4](#_Toc196147513)

[6.2. Timeline 4](#_Toc196147514)

[*6.*3. Milestones 5](#_Toc196147515)

[6.4. Challenges and Mitigations 5](#_Toc196147516)

[6.5. Ethical Considerations 5](#_Toc196147517)

[6.6. Task Distribution Matrix 5](#_Toc196147518)

[6. References 6](#_Toc196147519)

**List of Tables**

[Table 1: Technology Stack 4](#_Toc196147211)

[Table 2: Timeline 4](#_Toc196147212)

[Table 3: Challenges and Mitigations 5](#_Toc196147213)

[Table 4:Task Distribution Matrix 5](#_Toc196147214)

# 1. Project Overview

This project aims to develop a machine learning model to classify socioeconomic vulnerability using the OSDG Community Dataset (OSDG-CD), aligning with SDG 1(No Poverty) and SDG 10 (Reduced Inequalities). Traditional methods have varied from rule-based systems, highlighting the need for manual keyword refinement to advanced transfer based machine learning approaches(Fankhauser *et al.*, 2023). By using multidimensional indicators (e.g., income, education, healthcare), the model will identify at-risk populations to enable targeted policy interventions. Considering this we proposed the solution that addresses gaps in traditional vulnerability assessments by providing a scalable, data-driven tool for policymakers. The Potential Impact of the project will beimproved resource allocation for vulnerable communities and real-time monitoring via an interactive dashboard (Streamlet).To this end this project will come up with appropriate dashboard that help end users that suggest policy recommendation based on the model obtained from OSDG-CD using machine learning classifiers.

**2. Objectives**

* **To prepare and preprocess** the OSDG Community Dataset (OSDG-CD) for machine learning models
* **To develop and train** an interpretable ML model classifier capable of accurately predicting socioeconomic vulnerability levels.
* **To design and deploy** a user-friendly interactive dashboard (using Streamlit) that enables policymakers to visualize model insights.
* **To align** model outputs with Sustainable Development Goals (SDGs) 1 and 10, ensuring actionable recommendations for targeted interventions.

# 3. Background

The SDGs strive to encompass and cultivate global aspirations in various domains linked to sustainable development (Niels, 2023). Existing solutions like the Multidimensional Poverty Index (Alkire & Santos, 2014) rely on composite indicators but lack real-time adaptability. Machine learning approaches (e.g., Chen et al., 2020) demonstrate high accuracy but often neglect SDG integration.

Recent advancements in Natural Language Processing (NLP), particularly in Large Language Models (LLMs) and generative AI, offer promising avenues for automating the classification of scientific abstracts into relevant SDGs(Fankhauser *et al.*, 2023). Our project bridges this gap by combining: SDG-aligned data (OSDG-CD), Interpretable ML classifiers with feature importance) and Policy-ready outputs (Streamlit dashboard).

# 4. Methodology

In this section, different methods, techniques, and approaches used to identify optimal algorithms for OSDG policy recommendation will be discussed. The architecture, which shows the overall description of this project, was designed, and its components are sequentially described. Machine Learning Algorithms like Random Forest (chosen for interpretability and handling tabular data). Key Steps should be followed includes:

**Step-1-Data Preprocessing**-Data preprocessing tasks like handling missing values, redundancy and normalize features (Scikit-learn) will be applied.

**Step-2-Feature Selection**: Following the preprocessing tasks feature importance selection will be applied using correlation analysis and domain knowledge.

**Step3-Model Training**: After OSDG dataset is cleaned and important feature is selected the data will split in to training and testing. The training data set will used to construct the model and the test dataset is used to validate the constructed model. This will be done by optimize hyperparameters via grid search.

**Step 4- Evaluation**: In this steps, evaluation metrics like overall accuracy, precision-recall, and SHAP values will be used for explain ability.

## 4.1. Architecture Design Diagram

The proposed architecture focus on the general workflow of the project starting from the dataset preparation to model development and utilization. The proposed architecture has the overall layers and key activities that should be carried out in each layers. These common layers are presented as follows:

* **Data Layer**: OSDG-CD (CSV) → Pandas/NumPy preprocessing.
* **Model Layer**: Scikit-learn pipeline ML clasifiers
* **Application Layer**: Streamlit dashboard for visualization.

The following designed architecture shows the proposed architecture of the proposed system.

OSDG-Community Dataset

Feature Selection

(Correlation, Domain Knowledge)

Data Preprocessing (Cleaning, Scaling and Imbalance fix)

Model Training (Using ML algorithms)

Model Evaluation (Accuracy,SHAP)

Streamlet-Dashboard

Policy Recommendations

Figure 1: Proposed Architecture diagram

## 4.2. Data Sources

**Primary Dataset**: OSDG Community Dataset (publicly available CSV).

* **Features**: Income, education access, healthcare, geographic region.
* **Preprocessing**: Handle missing values (imputation), scale numerical features, address class imbalance (SMOTE).

# 5. Literature Review

* **Alkire & Santos (2014)**: Validates multidimensional indicators for poverty assessment.
* **Chen et al. (2020)**: Demonstrates ML efficacy (XGBoost) for socioeconomic classification.
* **Plataniotis et al. (2023)**: Links SDGs to policy tools, justifying our SDG-aligned approach.

# 6. Implementation Plan

## 6.1. Technology Stack

**Table 1: Technology Stack**

|  |  |  |
| --- | --- | --- |
| **Category** | **Tools** | **Purpose** |
| Programming Language | Python | Core development language |
| ML Libraries | Scikit-learn, Pandas, NumPy | Data handling, model training. |
| Visualization | Streamlit, Matplotlib | Interactive dashboard and plots. |
| Version Control | GitHub | Collaboration and Code Management |

## 6.2. Timeline

**Table 2: Timeline**

|  |  |  |
| --- | --- | --- |
| **Phase** | **Tasks** | **Duration** |
| Week1 | Data Collection and Preprocessing | 1 week |
| Week 2 | EDA & Feature Engineering | 1 week |
| Week 3 | Model Training and tuning | 1 week |
| Week4 | Dashboard development and deployment | 1 week |

## *6.*3. Milestones

* **Data Ready**: Cleaned dataset by Week 1.
* **Model Validated**: >85% accuracy by Week 2.
* **Dashboard Live**: Functional prototype by Week 3.

## 6.4. Challenges and Mitigations

Table 3: Challenges and Mitigations

|  |  |
| --- | --- |
| **Challenges** | **Mitigation Strategy** |
| Class Imbalance | Use SMOTE or class-weighted sampling. |
| Model interpretability | Integrate SHAP values for explanations. |
| Dashboard Scalability | Migrate to Dash if user load increases. |

## 6.5. Ethical Considerations

* **Bias**: Audit model for demographic biases (e.g., regional representation).
* **Privacy**: Use aggregated data; avoid individual identifiers.
* **Transparency**: Publish model metrics and limitations for policymakers.

## 6.6. Task Distribution Matrix

Table 4:Task Distribution Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Dugo Gadisa** | **Galata Waqwaya** | **Wabi Jifara** |
| Data Preprocessing | Lead | Support | Support |
| Model Development | Support | Support | Lead |
| Dashboard Deployment | Support | Lead | Support |

# 6. References

Alkire, S., & Santos, M. E. (2014). *World Development*.

Chen, Y., et al. (2020). *Applied Artificial Intelligence*.

OSDG Community Dataset. (n.d.). [https://www.osdg.ai](https://www.osdg.ai/)

Fankhauser, T., Clematide, S., et al., 2023. SDG Classification Using Instruction-Tuned LLMs.

Niels, B., 2023. MSc Business Administration and Data Science Sustainable Development Goals in Corporate Reporting : Investigating the Incorporation of SDGs with Multi-label Text Classification Supervised by.