Summary Report

1. Objective and Approach

The primary goal of this project was to develop and deploy a scalable, interpretable machine learning model that predicts the likelihood of household struggle. The task involved handling a dataset with socio-economic variables and building a machine learning model to predict the "ProgressStatus" of households, which is categorized as follows:

- On Track (>= 2.15)
- At Risk (>= 1.77)
- Struggling (>= 1.25)
- Severely Struggling (< 1.25)

The steps taken in this project include data cleaning, exploratory data analysis (EDA), handling class imbalances, model development, evaluation, and finally, deploying the model using a Django-based API.

2. Data Preprocessing and Handling Imbalance

Data Cleaning:

- Outlier Detection and Removal: Outliers in the "AgricultureLand" variable were detected and removed using the interquartile range (IQR) method to avoid skewing the model.
- Categorical Encoding: Categorical variables were label-encoded for machine learning compatibility. This converted qualitative data such as district and household head sex into numerical representations.

Imbalance Handling:

- The target variable "ProgressStatus" exhibited class imbalance. To address this:
- Random Under-Sampling was applied using the *RandomUnderSampler* from *imblearn*. This helped to balance the dataset by under-sampling the majority classes, ensuring that all classes had equal representation, which is crucial for unbiased model predictions.

3. Exploratory Data Analysis (EDA)

Distribution and Relationships:

- Histograms and density plots were used to visualize the distribution of the variables.

4. Model Development

Model:

- A Random Forest Classifier was selected for more advanced modeling. It was chosen for its robustness to overfitting, interpretability (via feature importance), and ability to handle non-linear relationships between features.

Cross-Validation:

- Cross-validation was used to evaluate model performance, ensuring that the model did not overfit the training data. This was crucial given the complexity of the socio-economic data, which had many interdependent variables.

5. Model Evaluation

Evaluation Metrics:

- Accuracy, F1-score, precision, recall, and confusion matrices were used to evaluate model performance.

6. Model Deployment

Inference Endpoint:

- A Django REST Framework (DRF)-based API was developed to serve the trained model. This allows real-time prediction of household progress statuses based on input features.

Logging:

- Logging was set up to capture incoming request details, errors, and predictions. This helps in debugging, auditing model performance, and tracking prediction usage.