**👤 Member 2: ML Engineer – Myra.**

**Title:** Machine Learning Paradigm & Model Justification for Solar Energy Forecasting

**🧠 ML Paradigm & Model Justification**

This project focuses on predicting solar energy generation in rural areas of Ghana and Kenya using weather-based data. Because we have access to historical data that includes both the input features (e.g., temperature, humidity, irradiance) and the target output (energy produced), this is clearly a **Supervised Learning** problem. More specifically, we are dealing with a **regression task**, as the output variable — solar energy output — is a continuous numeric value rather than a category.

We initially considered **Linear Regression** as a baseline model. Linear Regression is known for its simplicity and interpretability, and it provides insight into how each individual feature (such as temperature or irradiance) influences the output. This makes it a strong choice for initial modeling and exploratory analysis.

However, given the complex, often non-linear nature of environmental and weather data, we selected the **Random Forest Regressor** as our final model. Random Forests are ensemble models that combine multiple decision trees, making them highly effective at handling noisy, real-world datasets. They are well-suited for modeling non-linear relationships and are robust to outliers and irrelevant features. Additionally, Random Forests allow us to calculate **feature importance**, which is critical for understanding which weather conditions most strongly influence solar generation.

In conclusion, while Linear Regression was considered for baseline comparison, the **Random Forest Regressor** was chosen for its ability to balance **accuracy**, **robustness**, and **interpretability** — making it the best fit for our solar energy prediction task in this project.

**🤖 Proposed Model: Gradient Boosting Regressor (XGBoost or LightGBM)**

We also propose **Gradient Boosting Regression** as an optional or complementary approach. It’s a robust ensemble technique that combines multiple weak learners (typically decision trees) into a strong predictive model. It’s particularly powerful for tabular data and excels in capturing non-linear relationships.

**Why XGBoost or LightGBM?**

* Handles missing values and noisy data effectively
* Offers high accuracy without requiring deep neural networks
* Supports model interpretability, which is critical for real-world energy planning
* Easily tunable for performance improvement

Alternative models like Linear Regression may underperform due to the likely non-linear interactions in environmental data. Deep learning models (e.g., CNNs, RNNs) may be unnecessarily complex unless large time-series datasets are available.

**🔁 Model Pipeline Overview**

**1. Data Collection**  
– Gather historical weather and solar generation data for Ghana and Kenya.

**2. Data Preprocessing**  
– Clean missing values  
– Normalize/standardize features  
– Engineer time-based features (e.g., hour, season)

**3. Feature Selection**  
– Select key features: irradiance, temperature, humidity, cloud cover, etc.

**4. Train/Test Split**  
– Divide dataset into training and testing sets (e.g., 80/20 split)

**5. Model Training**  
– Train Random Forest or XGBoost/LightGBM on training data

**6. Model Evaluation**  
– Use metrics like RMSE, MAE, and R² to evaluate performance

**7. Deployment/Prediction**  
– Use the trained model to predict future solar energy generation

#### NOTES

Contributor: MYRA.

Supported with machine learning design and evaluation for Week 2