1. Problem Definition (6 points)

Hypothetical AI Problem: Predicting crop yield based on farm inputs and environmental factors.

Objectives:

- Estimate future yield for optimized resource allocation.
- Identify high-yield crop-soil-irrigation combinations.
- Reduce water/fertilizer usage without compromising yield.

Stakeholders:

- Agricultural policy makers.
- Farm managers and agronomists.

KPI:

- **Root Mean Squared Error (RMSE)** on crop yield predictions.

2. Data Collection & Preprocessing (8 points)

Data Sources:

- Agriculture_dataset.csv (keggle database).
- Remote sensing imagery (e.g., satellite data for vegetation indices and soil moisture).

Potential Bias:

- **Geographic bias** — dataset may be collected from specific regions or seasons, reducing generalizability to other climates or soil types.

Preprocessing Steps:

1. **Normalization:** Scale features like "Farm_Area" and "Water_Usage" for model stability.
2. **One-hot Encoding:** Convert categorical features like "Crop_Type", "Soil_Type", and "Season" into numeric format.
3. **Outlier Detection:** Remove farms with extreme yield or input usage that might skew training.

3. Model Development (8 points)

Chosen Model:

- **Random Forest Regressor** — it's robust to non-linear relationships, handles mixed feature types, and is interpretable via feature importance.

Data Split:

- 70% training
- 15% validation
- 15% test (Stratified if classifying crop types later.)

Hyperparameters to Tune:

- max_depth: Controls model complexity to reduce overfitting.
- n_estimators: Number of trees in the forest; more can improve accuracy but increase computation.

4. Evaluation & Deployment (8 points)

Evaluation Metrics:

- **RMSE**: Sensitive to large errors in crop yield prediction.
- **R^2 Score**: Measures how well the model captures variance in yield.

Concept Drift:

- Occurs when input-output relationships change over time (e.g., climate effects shifting crop response).
- Monitor by periodically re-evaluating model on new farm data and tracking performance metrics like RMSE.

Technical Challenge:

- **Scalability** — deploying the model across thousands of farms with real-time input updates may require distributed systems and optimized inference pipelines.