

Short Answer Questions

1. Problem Definition (6 points)

Hypothetical AI Problem: *Predicting crop disease outbreaks in smallholder farms using satellite and sensor data.*

Objectives:

- Detect early signs of crop disease from environmental patterns.
- Alert farmers via mobile platforms for timely intervention.
- Reduce crop loss and improve food security.

Stakeholders:

- Smallholder farmers in East Africa.
- Agricultural extension officers and policymakers.

Key Performance Indicator (KPI):

- **Disease prediction accuracy (%):** Percentage of correctly predicted disease outbreaks compared to actual occurrences.

2. Data Collection & Pre-processing (8 points)

Data Sources:

- Satellite imagery datasets (e.g., Sentinel-2, Landsat).
- IoT sensor data from soil moisture, temperature, and humidity sensors.

Potential Bias:

- **Geographic bias:** Data may overrepresent well-resourced regions with better sensor coverage, underrepresenting remote or underserved areas.

Pre-processing Steps:

1. **Handling missing data:** Impute missing sensor readings using interpolation or mean values.
2. **Normalization:** Scale environmental features (e.g., temperature, moisture) to a common range for model stability.
3. **Feature engineering:** Extract vegetation indices (e.g., NDVI) from satellite images to enhance disease detection.

3. Model Development (8 points)

Model Choice:

- **Random Forest Classifier** — robust to noisy data, interpretable, and effective for tabular and mixed-type features.

Data Splitting Strategy:

- **70% training, 15% validation, 15% test** — stratified sampling to preserve class distribution (e.g., diseased vs. healthy crops).

Hyperparameters to Tune:

1. **Number of trees** (`n_estimators`) — affects model complexity and performance.
2. **Maximum tree depth** (`max_depth`) — controls overfitting and generalization.

4. Evaluation & Deployment (8 points)

Evaluation Metrics:

- **F1-score:** Balances precision and recall, crucial for imbalanced datasets (e.g., rare disease outbreaks).
- **ROC-AUC:** Measures model's ability to distinguish between classes across thresholds.

Concept Drift:

- **Definition:** Change in data distribution over time that affects model performance (e.g., climate shifts altering disease patterns).
- **Monitoring Strategy:** Use rolling window evaluation and retrain periodically with new labelled data.

Technical Challenge:

- **Scalability:** Deploying the model to serve thousands of farmers across regions with limited internet access requires lightweight, offline-compatible inference solutions (e.g., edge computing or SMS-based alerts).