Part 1: Short Answer Questions (30 points)

Problem definition (6 points)

Hypothetical AI Problem:

Predicting daily traffic congestion in Nairobi using AI.

Objectives:

i. Use real-time and historical data to forecast traffic in major Nairobi roads.
ii. Assist commuters and logistics companies in optimizing routes and travel times.
iii. Help city planners in finding the congestion spots that are at high risk to improve infrastructure.

Stakeholders:

i. Nairobi County Transport Department
ii. Daily commuters and delivery companies

Key Performance Indicator (KPI):

Accuracy of traffic congestion predictability (e.g., Mean Absolute Error below a set criterion limit of traffic congestion)

Preprocessing and Data Collection (8 points)

Data Sources:

i. The analytics of past traffic data through the Google Maps API or the Nairobi County sensors
ii. The real-time GPS records of the apps used to hail a ride or a taxi

Potential Bias:

Device and network bias: Smartphone users or more recent car-owning customers are the only ones to feed the GPS data, potentially at the expense of informal transport (matatus, boda bodas) in which most people travel.

Preprocessing Steps:

i. Handle missing data: Interpolate or eliminate records with missing timestamps or locations.

ii. Feature engineering: Transform timestamps into day; hour; weekday/weekend; and rush hour.
iii. Normalization: To train a model, scale the GPS coordinates, speed, and distance features.

Model Development (8 points)

Model Chosen:

Random Forest Regressor

Justification:

It is robust to outliers, can also deal with non-linear relationships, and does not need feature scaling as rigorously as other models.

Strategy of Splitting Data:

70% training, 15% validation, 15% testing to train and evaluate generalization and tune the hyperparameters without overfitting.

Hyperparameters to Tune:

- n_estimators – Number of trees in the forest; more trees can improve accuracy but increase computational cost.
- max_depth – Limits tree depth to prevent overfitting on noisy traffic data

Evaluation & Deployment (8 points)

Evaluation Metrics:

i. Mean Absolute Error (MAE): The average error in estimating congestion level (readily comprehensible).
ii. R2 Score: Measures the model's quality of fit on the congestion pattern's variance.

Concept Drift:

A situation in which the underlying distribution of the data varies with time, e.g., because of construction, weather, or seasonal impacts on traffic.

Monitoring Strategy post-deployment:

Record model inputs and model predictions continuously. Compare the prediction and actual traffic levels by using the dashboard. Monthly, update the model using new data and retrain it.

Deployment Challenge:

Scalability: Quick inference over various routes in real-time traffic prediction is needed.

Solution: API with lightweight models that offer full batch prediction and static route predictions of cache during peak hours.