Part 1: Short Answer Questions (30 points)

1. Problem Definition (6 points)

Hypothetical AI problem: Predicting property price trends in a local real estate market.

Objectives:

- Forecast future property prices based on historical data.
- Identify factors influencing price fluctuations.
- Provide pricing recommendations for buyers and sellers.

Stakeholders:

- Real estate agents.
- Property investors.

Key Performance Indicator (KPI):

Mean Absolute Error (MAE) between predicted and actual property prices to measure prediction accuracy.

2. Data Collection & Preprocessing (8 points)

Data sources:

- Zillow API for property listings and price history.
- Public property records from local government databases.

Potential bias:

Data may be biased toward properties listed online, excluding off-market or private sales, which can skew price trends.

Preprocessing steps:

- Handle missing values by imputing or removing incomplete records.
- Normalize numerical features like property size and price.
- Encode categorical variables such as property type or neighborhood.

3. Model Development (8 points)

Model choice:

Random Forest, because it handles nonlinear relationships well, is robust to outliers, and interpretable for real estate features.

Data splitting:

Split data into 70% training, 15% validation, and 15% test sets to train the model, tune hyperparameters, and evaluate performance fairly.

Hyperparameters to tune:

- Number of trees (to balance bias and variance).
- Maximum tree depth (to prevent overfitting).

4. Evaluation & Deployment (8 points)

Evaluation metrics:

- Mean Absolute Error (MAE) to measure average prediction error in price units.
- R-squared (R^2) to assess how well the model explains price variability.

Concept drift:

Changes in market conditions or economic factors over time that affect property prices differently from training data. Monitor it by regularly comparing prediction errors on new data and retraining the model as needed.

Technical challenge:

Scalability—handling large volumes of real-time property data and delivering fast predictions to users without delays.