Question 1 week 5 Ai engineering

Part 1: Short Answer Questions

1. Problem Definition (6 points)

AI Problem: Predicting Infectious Disease Prevention Coverage (UHC Service Coverage Index)

Objectives:

1. Predict future infectious disease prevention coverage rates for countries/regions
2. Identify factors that most influence coverage levels
3. Enable proactive healthcare resource planning and policy decisions

Stakeholders:

1. World Health Organization (WHO) - For global health monitoring and policy recommendations
2. National Health Ministries - For country-level healthcare planning and resource allocation

KPI: Mean Absolute Error (MAE) of coverage predictions - measures how close predictions are to actual coverage percentages

2. Data Collection & Preprocessing (8 points)

Data Sources:

1. UHC_SCI_INFECT dataset (WHO infectious disease prevention coverage data)
2. Country demographic and economic indicators (GDP, population density, healthcare spending)

Potential Bias: Data is biased toward countries with better healthcare reporting systems, potentially underrepresenting coverage in developing nations

Preprocessing Steps:

1. Handle missing values in ParentLocationCode by filling with 'Unknown'
2. Convert categorical variables (country codes, region codes) to numerical using category codes
3. Split data into training (80%) and testing (20%) sets with random_state=42 for reproducibility

Model Development

Chosen Model: Random Forest Regressor

Justification:

- Handles mixed data types well (as shown in the notebook with categorical country codes)
- Robust to outliers and missing values
- Provides feature importance rankings
- Achieved good performance ($R^2 = 0.670$) in the implementation

Data Split:

- Training set: 80% of data (X_train, y_train)
- Test set: 20% of data (X_test, y_test)
- Using random_state=42 ensures reproducible results

Hyperparameters to Tune:

1. n_estimators - Number of trees in the forest (currently 100, could optimize for performance vs computation time)
2. max_depth - Maximum depth of trees (prevents overfitting, controls model complexity)

Evaluation & Deployment (8 points)

Evaluation Metrics:

1. **R^2 Score (0.670)** - Measures proportion of variance explained by the model, shows overall predictive power
2. **Mean Absolute Error** - Provides interpretable error in percentage points for coverage predictions

Concept Drift: When the statistical properties of the target variable change over time, making the model less accurate. In this case, changes in global healthcare policies, pandemics, or economic conditions could alter coverage patterns.

Monitoring: Track prediction accuracy over time by comparing predictions with newly collected actual coverage data, and retrain model when performance degrades beyond a threshold.

Technical Challenge: Feature Encoding Consistency - Ensuring that new country/region codes in future data match the encoding used during training. The model relies on consistent categorical encoding, so new countries would need to be properly mapped, or the model retrained.

Ethics & Bias

- **Impact of Biased Data:**
  Biased training data can lead to systematically worse predictions for specific patient groups (e.g., based on race, socioeconomic status, or insurance type). This could:
  1. **Under-predict risk** for disadvantaged groups, denying them necessary follow-up care and increasing their likelihood of complications.
  2. **Over-predict risk** for others, leading to unnecessary and costly interventions, straining hospital resources and patient trust.
- **Bias Mitigation Strategy:**
  Pre-process the training data using techniques like reweighting or resampling to ensure the dataset is representative across sensitive attributes (e.g., race, age, gender). This helps the model learn patterns that are fair across all subgroups.

Trade-offs

- **Interpretability vs. Accuracy:**
  In healthcare, interpretability is often prioritized. A slightly less accurate but interpretable model (like Logistic Regression or a shallow Decision Tree) allows clinicians to understand *why* a prediction was made, fostering trust and enabling clinical validation. A "black box" model like a complex ensemble may be more accurate but is dangerous if its reasoning cannot be scrutinized, potentially leading to uncaught errors or ethical issues.
- **Impact of Limited Computational Resources:**
  Limited resources would shift the model choice away from computationally intensive algorithms like deep neural networks or large ensembles. The focus would be on **simpler, more efficient models** like Logistic Regression, shallow Decision Trees, or Naive Bayes, which require less power for both training and inference and can run on standard hardware.