## Part 1: Short Answer Questions (30 points)

**1. Problem Definition (6 points)**

**An example of a hypothetical AI problem is predicting hospital readmission risk for diabetic patients. The goal is to use patient health data to identify individuals likely to be readmitted within 30 days after discharge. This helps hospitals take early action to prevent avoidable readmissions, improve patient outcomes, and reduce healthcare costs. The main stakeholders are healthcare providers, who use the predictions to guide care, and patients, who benefit from targeted support and better health management.**

- **List 3 objectives and 2 stakeholders.**

**Objectives:**

1. **To predict which diabetic patients are at risk of being readmitted within 30 days.**
2. **To analyze patient data (age, medications, test results) for patterns.**
3. **To support healthcare staff in making better care decisions.**

---

**Stakeholders:**

- **Healthcare providers (doctors, nurses, hospital management)**
- **Diabetic patients**
  -

**Propose 1 Key Performance Indicator (KPI) to measure success.**

**Key Performance Indicator (KPI):**
**Percentage reduction in 30-day readmission rates among diabetic patients after deploying the AI model.**

**2. Data Collection & Preprocessing (8 points)**

- **Identify 2 data sources for your problem.**

**Data Sources:**

- ☐ **Electronic Health Records (EHRs): Includes patient demographics, medical history, diagnoses, medications, lab results, and previous hospital admissions.**
- ☐ **Hospital Discharge Summaries: Contains details about the patient's condition at discharge, treatment given, and follow-up instructions, which are useful for predicting readmission risk.**

## Explain 1 potential bias in the data.

**Potential Bias:**
**If the data mostly comes from urban hospitals, the model may not perform well for patients in rural areas, leading to biased predictions and unequal healthcare outcomes.**

- **Outline 3 preprocessing steps (e.g., handling missing data, normalization).**

**Preprocessing Steps:**

1. **Handling Missing Data: Fill in missing values using techniques like mean imputation or remove records with excessive missing fields.**
2. **Normalization: Scale numerical features (e.g., age, blood glucose levels) to a standard range to improve model performance.**
3. **Encoding Categorical Variables: Convert non-numerical data such as gender or diagnosis codes into numerical format using one-hot encoding or label encoding.**

## 3. Model Development (8 points)

- Choose a model (e.g., Random Forest, Neural Network) and justify your choice.

## 3. Model Development (8 points)

Question: Choose a model (e.g., Random Forest, Neural Network) and justify your choice.
Answer: I would choose the Random Forest model. It is suitable for classification problems such as predicting hospital readmission risk. Random Forest handles both numerical and categorical data, is robust to overfitting, and provides feature importance, which helps in understanding the most influential factors. It also performs well with missing or noisy data.

Question: Describe how you would split data into training/validation/test sets.
Answer: I would split the data into 70% training, 15% validation, and 15% testing. The training set is used to train the model, the validation set is used for hyperparameter tuning, and the test set is used to evaluate final model performance on unseen data.

Question: Name 2 hyperparameters you would tune and why.
Answer:

1. Number of trees in the Random Forest: Controls the number of decision trees; tuning it helps balance accuracy and computation time.
2. Maximum depth of the trees: Limits how deep the tree can grow; tuning prevents overfitting by avoiding overly complex trees.

## 4. Evaluation & Deployment (8 points)

Question: Select 2 evaluation metrics and explain their relevance.
Answer:

1. Precision: Measures how many of the patients predicted as high-risk were actually readmitted. Important to reduce false alarms.
2. Recall: Measures how many of the actual readmissions were correctly predicted. Important for capturing all at-risk patients.

Question: What is concept drift? How would you monitor it post-deployment?
Answer: Concept drift refers to changes in the data patterns over time that reduce model

**accuracy. It can be monitored by regularly checking performance metrics and retraining the model if accuracy drops.**

**Question: Describe 1 technical challenge during deployment (e.g., scalability).**
**Answer: A key challenge is scalability—handling predictions for many patients in real-time. This requires optimizing the model and using cloud services or APIs to ensure the system responds quickly.**

**Part 2: Case Study Application (40 points)**

**Problem Scope (5 points)**

**Question: Define the problem, objectives, and stakeholders.**
**Answer: The problem is predicting patient readmission risk within 30 days of discharge. Objectives include identifying high-risk patients, supporting early interventions, and improving healthcare outcomes. Stakeholders are healthcare providers and patients.**

**Data Strategy (10 points)**

**Question: Propose data sources (e.g., EHRs, demographics).**
**Answer: Data sources include Electronic Health Records (EHRs), discharge summaries, patient demographics, lab test results, and medication history.**

**Question: Identify 2 ethical concerns (e.g., patient privacy).**
**Answer:**

1. **Patient privacy: Ensuring personal health information is protected and anonymized.**

2. **Data bias: Unequal representation in the data can lead to biased predictions against certain groups.**

**Question: Design a preprocessing pipeline (include feature engineering steps).**
**Answer:**

1. **Handle missing values using imputation.**

2. **Normalize numerical features (e.g., age, glucose levels).**

3. **Encode categorical variables (e.g., gender, diagnosis).**

4. **Engineer features such as number of previous admissions, time since last visit, and medication count.**

**Model Development (10 points)**

**Question: Select a model and justify it.**
**Answer: Random Forest is chosen for its robustness, ability to handle mixed data types, and feature importance insights.**

**Question: Create a confusion matrix and calculate precision/recall (hypothetical data).**
**Answer:**
**Hypothetical Confusion Matrix:**

- **True Positives (TP): 40**
- **False Positives (FP): 10**
- **False Negatives (FN): 20**
- **True Negatives (TN): 130**

**Precision = TP / (TP + FP) = 40 / (40 + 10) = 0.80**
**Recall = TP / (TP + FN) = 40 / (40 + 20) = 0.67**

**Deployment (10 points)**

**Question: Outline steps to integrate the model into the hospital’s system.**
**Answer:**

1. **Develop API to serve model predictions.**
2. **Integrate with hospital's EHR system.**
3. **Train staff on system use.**
4. **Monitor system performance and retrain as needed.**

**Question: How would you ensure compliance with healthcare regulations (e.g., HIPAA)?**
**Answer: Ensure patient data is encrypted, access is restricted, and processing complies with data protection standards. Regular audits and training on data privacy are also essential.**

**Optimization (5 points)**

**Question: Propose 1 method to address overfitting.**
**Answer: Use cross-validation and prune the trees in the Random Forest to limit complexity and improve generalization.**

**Part 3: Critical Thinking (20 points)**

**Ethics & Bias (10 points)**

**Question: How might biased training data affect patient outcomes in the case study?**
**Answer: Biased data can lead to unfair predictions, such as underestimating risk for certain demographic groups, resulting in unequal care and worse outcomes.**

**Question: Suggest 1 strategy to mitigate this bias.**
**Answer: Use balanced datasets and perform fairness audits to detect and correct bias during model training.**

**Trade-offs (10 points)**

**Question: Discuss the trade-off between model interpretability and accuracy in healthcare.**
**Answer: More accurate models like deep neural networks are often less interpretable, making it hard for doctors to trust them. Simpler models like Random Forests or decision trees may be slightly less accurate but offer clearer insights into predictions.**

**Question: If the hospital has limited computational resources, how might this impact model choice?**
**Answer: The hospital may need to choose lightweight models like logistic regression or decision trees that require less processing power and memory.**

**Part 4: Reflection & Workflow Diagram (10 points)**

**Reflection (5 points)**

**Question: What was the most challenging part of the workflow? Why?**
**Answer: The most challenging part was data preprocessing because real-world data is often incomplete, messy, and requires careful handling to ensure model accuracy.**

**Question: How would you improve your approach with more time/resources?**
**Answer: With more time and resources, I would collect more diverse data, test multiple models, and use advanced techniques like ensemble learning for better performance**