Part 1: Short Answer Questions

1. Problem Definition

AI Problem: Predicting student dropout risk in rural high schools.
Objectives:

- Identify at-risk students early.
- Reduce dropout rates through timely interventions.
- Improve overall school performance and retention.

Stakeholders:

- School administrators
- Ministry of Education

Key Performance Indicator (KPI):

- Dropout prediction accuracy (e.g., percentage of correct predictions)

---

2. Data Collection & Preprocessing

Data Sources:

- School attendance and academic records
- Household socioeconomic data from national statistics

Potential Bias:

- Underrepresentation of students from remote regions, leading to lower prediction accuracy for them.

Preprocessing Steps:

1. Handle missing values (e.g., impute missing grades or income).
2. Normalize features (e.g., scale scores and income levels).
3. Encode categorical data (e.g., parental education levels).

---

3. Model Development

Model: Random Forest
Justification:

- Handles non-linear relationships and categorical features well.
- Robust against overfitting and easy to interpret.

Data Split:

- 70% training, 15% validation, 15% testing

Hyperparameters to Tune:

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

---

4. Evaluation & Deployment

Evaluation Metrics:

- **F1 Score**: Balances precision and recall (useful with imbalanced data)
- **ROC-AUC**: Measures ability to distinguish dropout vs. non-dropout

Concept Drift:

- When student behavior patterns change over time (e.g., due to new policies).
  Monitoring:

- Track accuracy/F1 over time; retrain if performance drops.

Deployment Challenge:

- **Scalability**: Serving predictions for thousands of schools with limited internet infrastructure.

Part 2: Case Study Application

Problem Scope

Problem:
Predict if a patient will be readmitted within 30 days post-discharge.
Objectives:

- Reduce avoidable readmissions

- Improve patient care quality
  Stakeholders:

- Hospital management

- Clinicians and care teams

Data Strategy

Data Sources:

- Electronic Health Records (EHRs)

- Patient demographics and previous admission history

Ethical Concerns:

1. Patient privacy and data security
2. Bias against certain patient groups (e.g., low-income or elderly)

Preprocessing Pipeline:

- Remove duplicates and irrelevant entries
- Impute missing values (e.g., lab results)
- Feature engineering:
    - Length of stay
    - Number of previous visits
    - Chronic condition flags (e.g., diabetes, hypertension)

Model Development

Model: Logistic Regression
Justification:

- Easy to interpret and explain to medical professionals
- Effective for binary classification

Hypothetical Confusion Matrix (100 patients):

	Predicted Readmit	**Predicted Not Readmit**
Actual Readmit	30 (TP)	10 (FN)

Actual Not	15 (FP)	45 (TN)

Precision: 30 / (30 + 15) = 0.67
Recall: 30 / (30 + 10) = 0.75

Deployment

Integration Steps:

- Embed model into EHR system via API
- Trigger prediction at discharge
- Display risk score in patient's digital file

Compliance:

- Use encrypted servers and access control
- Ensure data collection and usage align with **POPIA** (South Africa) or **HIPAA** (US)

Optimization

Overfitting Solution:

- Apply **regularization** (e.g., L2) to penalize model complexity

---

Part 3: Critical Thinking

Ethics & Bias

Impact of Bias:

- Biased training data may lead to underestimating risk for underrepresented groups (e.g., rural patients), causing worse outcomes.

Mitigation Strategy:

- Use stratified sampling and fairness-aware algorithms to ensure balanced representation during training.

Trade-offs

Interpretability vs. Accuracy:

- Complex models (e.g., Neural Networks) may perform better but are hard to explain to doctors.
- Interpretable models (e.g., Decision Trees) offer trust but may be less accurate.

Limited Resources:

- Favor lightweight models (e.g., Logistic Regression) over GPU-dependent deep learning.
- Reduce model complexity to allow on-premise deployment without cloud dependency.

Part 4: Reflection & Workflow Diagram

Reflection

Most Challenging Part:

- Handling bias and ensuring fairness due to limited, imbalanced data.

Improvement with More Time:

- Collect larger, more diverse datasets
- Involve healthcare staff in model co-design for better alignment

---

Diagram (5 points)

lua
CopyEdit

```
 +-------------------------+
  |  Problem Definition     |
  +-------------------------+
               ↓
  +-------------------------+
  |  Data Collection        |
  +-------------------------+
               ↓
  +-------------------------+
  |  Data Preprocessing     |
  +-------------------------+
               ↓
  +-------------------------+
  |  Model Development      |
  +-------------------------+
               ↓
  +-------------------------+
  |  Evaluation             |
  +-------------------------+
               ↓
  +-------------------------+
  |  Deployment             |
  +-------------------------+
               ↓
```