Part 2: Tasks (40 points)

1. Problem Scope (5 points)

- **Problem:**
  Predict the likelihood of a patient being readmitted within 30 days after hospital discharge.
- **Objectives:**
  - Reduce unnecessary readmissions.
  - Improve patient outcomes and care continuity.
  - Optimize resource allocation and lower healthcare costs.
- **Stakeholders:**
  - Hospital administrators
  - Clinicians and care coordinators
  - Data scientists/IT staff
  - Patients

2. Data Strategy (10 points)

a. Data Sources:

- **Electronic Health Records (EHR):** Diagnoses, medications, procedures, discharge summaries.
- **Demographics:** Age, gender, ethnicity, insurance status.
- **Utilization History:** Past admissions, ER visits.
- **Social Determinants of Health (SDoH):** Zip code, income bracket (optional).
- **Lab Results and Vital Signs**

b. Ethical Concerns:

1. **Patient Privacy:** Sensitive health data must be protected from breaches and misuse.

2. **Informed Consent & Data Usage:** Patients may be unaware their data is used for modeling; transparency is essential.

c. Preprocessing Pipeline:

1. **Data Cleaning:** Handle missing values, duplicates, and erroneous entries.
2. **Feature Engineering:**
   - Time since last admission
   - Comorbidity count (Charlson Index)
   - Length of stay
   - Discharge disposition (e.g., home, rehab)
3. **Normalization:** For lab results and vitals
4. **Encoding:** One-hot encoding for categorical features (e.g., diagnosis codes)
5. **Splitting:** Train-test split (e.g., 80-20), with cross-validation

3. Model Development (10 points)

a. Model Choice:
Gradient Boosting (e.g., XGBoost or LightGBM)

- Justification: Handles tabular data well, captures non-linearities, provides feature importance, and performs robustly with imbalanced data.

b. Confusion Matrix (Hypothetical):

	Predicted Readmit	**Predicted No Readmit**
Actual Readmit	80	20
Actual No Readmit	40	160

Precision:
= TP / (TP + FP) = 80 / (80 + 40) = **0.67**

Recall (Sensitivity):
= TP / (TP + FN) = 80 / (80 + 20) = **0.80**

4. Deployment (10 points)

a. Integration Steps:

1. Model API hosted securely (e.g., RESTful API on cloud/on-prem server)
2. EHR system sends discharge data to model
3. Model returns risk score to clinician dashboard
4. Care team receives alert for high-risk patients
5. Regular model retraining (monthly/quarterly)

b. Regulatory Compliance (e.g., HIPAA):

- De-identify patient data where possible
- Encrypt data in transit and at rest
- Access control for model and logs
- Audit trails for data access and model decisions
- Partner with hospital compliance officers

5. Optimization (5 points)

Overfitting Solution:
Use Regularization (e.g., L2 or Tree Pruning in XGBoost)

- Prevents the model from capturing noise in training data
- Cross-validation ensures generalizability