Part 2: Case Study Application (40 points)

Scenario: A hospital wants an AI system to predict patient readmission risk within 30 days of discharge.

Tasks:

1. **Problem Scope (5 points):** Define the problem, objectives, and stakeholders.
    - The task is to build a binary classification system that **predicts the probability of a patient being readmitted to the hospital within 30 days after discharge.** This is an unplanned readmission **risk model** aimed at **supporting early post-discharge** interventions such as follow-up calls, medication reconciliation, or home-care coordination.
    - The **objectives** are:
      → *(primary)* To identify high-risk patients with at least 85% recall so that limited care-management resources can be directed where they are most needed
      → *(secondary)* To reduce the hospital's overall 30-day unplanned readmission rate by at least 18% within the first two years of deployment.
      → To deliver an interpretable risk score and the top contributing factors directly into the clinician's workflow inside the existing Electronic Health Record (EHR) system.
    - **Key stakeholders** include hospital care managers and discharge planners who will act on the predictions, attending physicians and nurses who need transparent reasoning for trust, the finance and quality department that monitors CMS penalties and value-based care metrics, and, most importantly, the patients and their families whose care experience and outcomes will be directly affected by the interventions triggered by the model.

2. **Data Strategy (10 points):**

    - Propose data sources (e.g., EHRs, demographics).

      The core data will come from the hospital's Electronic Health Records (EHR) system, specifically admission-discharge-transfer (ADT) records, coded diagnoses (ICD-10), procedures, medication administration records, laboratory results, vital signs during the stay, and discharge summaries. Additional valuable sources include historical claims data (to capture prior admissions and emergency-department visits in the past 12 months) and socio-demographic variables such as age, gender, primary language, insurance type, and ZIP code, which serve as proxies for social determinants of health. Where available, structured social-risk screening data (e.g., PRAPARE or AHC scores) will also be incorporated.

- Identify **2 ethical concerns** (e.g., patient privacy).
    1. Patient privacy and the risk of re-identification are significant because combinations of rare diagnoses, specific procedures, age, gender, and ZIP code can sometimes uniquely identify individuals, violating HIPAA's safe-harbor de-identification standard.
    2. The data reflect historical care patterns that may embed systemic biases; for example, patients with Medicaid or from underserved neighborhoods historically experience higher readmission rates partly because of access barriers rather than clinical need, which can perpetuate inequitable resource allocation if not corrected.

- Design a preprocessing pipeline (include feature engineering steps).

  The preprocessing and feature-engineering pipeline is designed as follows:

  - ★ Records for patients with length of stay less than 24 hours or planned readmissions (e.g. chemotherapy cycles) are excluded.
  - ★ Diagnoses and procedures are aggregated using Clinical Classifications Software (CCS) groups, and binary flags are created for the 50 conditions most associated with readmission risk.
  - ★ The LACE index components (Length of stay, Acuity, Comorbidities, Emergency visits) are calculated directly, supplemented by the Elixhauser comorbidity score and count of admissions in the prior year.
  - ★ Missing laboratory values are imputed with the median value for the specific admission type and ward, while extreme outliers are clipped at the 1st and 99th percentiles.
  - ★ Categorical variables such as discharge disposition, admission source, and payer are one-hot encoded, and high-cardinality ICD-10 chapters are target-encoded using the training-set readmission rate.
  - ★ All continuous features are scaled to [0,1] using Min-Max scaling derived only from the training split to prevent data leakage.

3. **Model Development (10 points):**

- Select a model and justify it.

  For this readmission-risk prediction task, **LightGBM** (Light Gradient Boosting Machine) is my selected primary model.

  Justification: It **consistently outperforms other algorithms on medium-sized tabular EHR datasets** while offering **decisive practical advantages** such as native handling of missing values and categorical features, very fast training and

inference even on standard hospital servers, low memory footprint, and excellent calibration of predicted probabilities.
Equally important in the clinical setting, LightGBM provides **reliable global and local feature importance through SHAP values**, which can be displayed to clinicians as “top reasons for this patient’s high risk,” significantly increasing trust and adoption compared to black-box deep learning models.

- Create a confusion matrix and calculate precision/recall (hypothetical data).

  The dataset is split using a **time-based stratified approach** to respect the **temporal nature of patient admissions** and **prevent leakage**: all discharges from the earliest available year(s) up to a chosen cut-off constitute the training set (approximately 70%), the following 6 - 9 months form the validation set for early stopping and hyperparameter tuning (15%), and the most recent 12 - 18 months are held out as the final test set (15%). **Stratification** ensures that the proportion of readmissions remains identical across splits.
  A hypothetical but realistic **confusion matrix** on the held-out test set of 2,000 patients is presented below:

	Predicted No Readmission	**Predicted High Risk**
Actual No (1,760)	1,650	110
Actual Yes (240)	72	168

From this matrix:

Precision = 168 / (168 + 110) = 0.604 (60.4%)

Recall = 168 / (168 + 72) = 0.700 (70.0%)

F1-score ≈ 0.649

These figures reflect a deliberate operating point that prioritizes recall (catching 70% of true readmissions) while keeping false positives manageable for care-management teams.

4. **Deployment (10 points):**

   - Outline steps to integrate the model into the hospital's system.

     The trained LightGBM model is **exported as a single joblib/pickle file** and **wrapped in a lightweight FastAPI** service, **containerized with Docker** for consistent deployment across development, staging, and production environments.
     Integration into the hospital's clinical workflow is achieved through two pathways:

     - Real-time scoring is triggered automatically upon finalization of the discharge order: an HL7 ADT message fires a secure HTTPS call to the prediction service, which returns the risk probability, risk tier (low/medium/high), and the top five contributing factors via SHAP values; these are written back into the EHR (Epic or Cerner) using FHIR API resources (Observation and RiskAssessment) so clinicians see the score directly on the discharge navigator screen.
     - A parallel batch process runs nightly to score all patients expected to be discharged the following day, populating a care-management dashboard for proactive outreach planning.

   - How would you ensure compliance with healthcare regulations (e.g., HIPAA)?

     Ensuring HIPAA compliance is implemented at multiple layers.

     - All data in transit uses TLS 1.3 encryption, and data at rest on the inference server is encrypted with AES-256.
     - Access to the prediction endpoint is restricted by mutual TLS certificates and role-based authorization tied to Active Directory groups; every request and response is logged with patient MRN, timestamp, and requesting user for audit purposes.
     - No raw PHI is stored in the model service – only the minimum necessary structured features are sent, and the service is covered under a signed Business Associate Agreement with the cloud or on-premise hosting provider.
     - Regular penetration testing and quarterly HIPAA risk assessments are scheduled as part of the deployment governance process.

5. **Optimization (5 points):** Propose **1 method** to address overfitting.

   To prevent overfitting – especially critical in healthcare where over-optimism on historical data can lead to missed high-risk patients in real-world use – the primary technique employed is **robust early stopping** combined with **monotonic constraints**.
   During training, LightGBM **monitors recall on the validation set and stops if no improvement occurs for 100 consecutive rounds**, ensuring the **model does not memorize noise in the training data.**
   Additionally, clinical domain knowledge is enforced through **monotonic constraints on key features known to have directional relationships with readmission risk**: age, length of stay, number of prior admissions, Elixhauser comorbidity score, and lab values such as serum albumin and creatinine are constrained to be monotonically increasing (higher values must not decrease predicted risk), while hemoglobin is set as monotonically decreasing.
   This approach not only reduces overfitting by **limiting model flexibility in clinically implausible regions** but also **improves calibration** and clinician trust in the predictions.

MELISSA