Part 3: Critical Thinking (20 points)

1. **Ethics & Bias (10 points):**

    - How might biased training data affect patient outcomes in the case study?
      Biased training data in the readmission model can have serious real-world consequences. If historical EHR data **over-represents readmissions from low-income ZIP codes or Medicaid patients because of poorer access to primary care** rather than purely clinical factors, the model will learn to assign higher risk scores to these groups even when their current clinical profile is similar to privately insured patients.
      This creates a feedback loop: already disadvantaged patients receive more interventions (which is good) but also risk stigmatisation, over-treatment, or resource misallocation away from other high-risk groups. In extreme cases, insurers could misuse such scores for reimbursement decisions, further entrenching health inequity.

    - Suggest **1 strategy** to mitigate this bias.

      A practical mitigation strategy is to **implement fairness-aware preprocessing combined with post-hoc calibration**. During training, we apply **re-weighting of samples** (or adversarial debiasing) so that the **model's error rates are balanced across protected attributes** such as insurance type and race/ethnicity proxies. After deployment, we enforce demographic parity on the final risk tiers by **adjusting thresholds per subgroup** (e.g. slightly higher threshold for Medicaid patients while maintaining overall recall ≥ 70%). All fairness metrics are monitored monthly alongside traditional performance metrics.

2. **Trade-offs (10 points):**

    - Discuss the trade-off between model interpretability and accuracy in healthcare.

      In healthcare, the classic trade-off exists between **model accuracy** and **interpretability**. **Complex ensembles** or **deep neural networks** typically achieve 1 - 3 points higher AUROC on readmission tasks, but clinicians and regulators demand transparent reasoning. **LightGBM with SHAP values** strikes an excellent balance: it matches or exceeds most deep learning baselines on tabular EHR data while providing both global rankings and patient-level explanations that can be directly shown in Epic/Cerner. When interpretability is non-negotiable (as in this case), sacrificing a small amount of raw accuracy is justified and often required for clinical adoption and regulatory approval.

- If the hospital has limited computational resources, how might this impact model choice?

  If the hospital has **severely limited computational resources** (e.g. only on-premise servers with no GPU), we would **favour even lighter models such as a calibrated logistic regression on top of carefully engineered features** (LACE+ index, Elixhauser score, prior admissions) or a **compact XGBoost/LightGBM with ≤ 50 trees**. These **trade some predictive performance** (AUROC ~0.76–0.78 vs 0.81) for dramatically **lower inference latency** (< 10 ms per patient) and easier maintenance, which is often more important than marginal accuracy gains in resource-constrained settings.

Part 4: Reflection & Workflow Diagram (10 points)

1. **Reflection (5 points):**

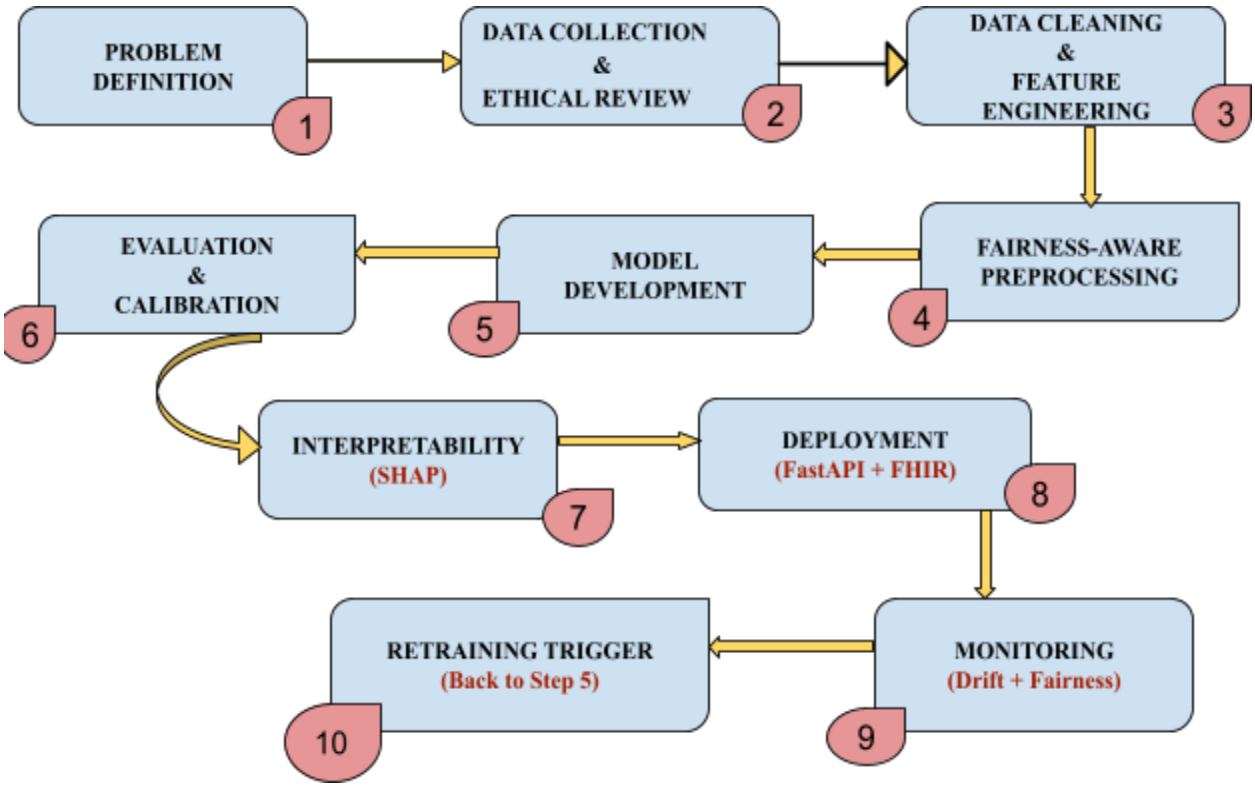
   - What was the most challenging part of the workflow? Why?

     The most challenging part of the entire workflow was **managing concept drift** and **ensuring long-term fairness** in a live clinical environment. While the technical aspects of training and evaluation are straightforward, **guaranteeing that the model remains safe and equitable** once patient demographics, coding practices, or care protocols change over months/years requires robust MLOps infrastructure (automated retraining triggers, drift dashboards, fairness audits) that many hospitals still lack.

   - How would you improve your approach with more time/resources?

     With more time and resources, I would **implement a full continuous training pipeline** using Feature Store + model registry (e.g., Feast + MLflow), **add automated fairness testing** in CI/CD, and **conduct a prospective clinical pilot with real-time clinician feedback before full rollout**.
     I would also collect structured “outcome of intervention” labels to enable reinforcement learning from human decisions in future iterations.

2. **Diagram (5 points):**
    - Sketch a flowchart of the AI Development Workflow, labeling all stages.



MELISSA