Part 3: Critical Thinking (20 points)

Ethics & Bias (10 points)

- **How might biased training data affect patient outcomes in the case study?**
  Biased training data, often reflecting historical disparities in healthcare access and quality, can lead to the AI model making inaccurate or unfair predictions for certain patient groups. For instance:

  - If the training data disproportionately represents healthier or more privileged populations, the model might underpredict readmission risk for underserved communities (e.g., based on race, socioeconomic status, or geographic location). This could lead to these high-risk patients not receiving necessary post-discharge interventions, exacerbating health inequities.

  - Conversely, it might overpredict risk for certain groups, leading to unnecessary interventions or stigmatization.

  - Errors in data collection or missing data for specific demographics could also introduce bias, making the model less effective or even harmful for those groups.

- **Suggest 1 strategy to mitigate this bias. Fairness-Aware Data Collection and Preprocessing:**

  - **Strategy:**
    Actively seek to collect more representative data from diverse patient populations, ensuring that all demographic groups are adequately represented in the training dataset. During preprocessing, techniques like **resampling (oversampling minority classes or undersampling majority classes)**, **reweighing (assigning different weights to data points based on their group membership)**, or **adversarial debiasing** can be used to reduce the impact of existing biases in the data before model training. Additionally, **feature selection should carefully consider proxy variables** that might inadvertently carry sensitive attribute information (e.g., zip code acting as a proxy for race or income).

Trade-offs (10 points)

- **Discuss the trade-off between model interpretability and accuracy in healthcare.**
    - **Interpretability:** Refers to the extent to which humans can understand the reasoning behind a model's predictions. Simple models like linear regression or decision trees are highly interpretable, as their decision rules are transparent.
    - **Accuracy:** Refers to how well the model's predictions match the actual outcomes. Complex models, such as deep neural networks or ensemble methods like Gradient Boosting Machines (GBMs), often achieve higher accuracy by capturing intricate non-linear relationships in the data.
    - **Trade-off in Healthcare:** In healthcare, there's a significant tension. While highly accurate models are desirable for critical predictions like readmission risk, clinicians often need to understand *why* a prediction was made to trust the system, justify interventions, and explain decisions to patients. A "black box" model, even if highly accurate, might be resisted by medical professionals due to a lack of transparency and accountability. For instance, if a model predicts high readmission risk, a clinician needs to know *which factors* (e.g., specific comorbidities, recent lab values, medication non-adherence) contributed most to that prediction to formulate an effective care plan. This need for interpretability can sometimes mean sacrificing a small degree of predictive accuracy by choosing a slightly less complex but more transparent model, or by employing explainable AI (XAI) techniques on complex models.

- **If the hospital has limited computational resources, how might this impact model choice?**
  Limited computational resources (e.g., less powerful servers, limited GPU access, restricted cloud computing budget) would significantly impact model choice in several ways:
    - **Preference for Simpler Models:** The hospital would likely need to opt for simpler, less computationally intensive models. Instead of deep learning models or large ensemble methods like complex GBMs, they might choose logistic regression, simpler decision trees, or smaller random forests. These models require less processing power and memory for training and inference.
    - **Reduced Feature Engineering Complexity:** Extensive feature engineering, especially involving complex transformations or interactions, can be

computationally demanding. The hospital might need to limit the number and complexity of features used.

- **Smaller Data Subsets:** Training on the full dataset might be infeasible. The hospital might need to sample smaller subsets of data for training or use techniques that allow for incremental learning.
- **Batch vs. Real-time Inference:** Real-time predictions require more immediate computational power. With limited resources, the hospital might be restricted to batch predictions, where data is processed periodically (e.g., overnight) rather than on demand.
- **Impact on Model Performance:** There's a risk that simpler models, chosen due to resource constraints, might not capture all the nuances in the data, potentially leading to lower predictive accuracy compared to what could be achieved with more powerful resources.