AI System for Predicting Patient Readmission Risk

Problem Scope (5 points)

Problem Definition:
Develop an AI system to predict the probability of patient readmission within 30 days of discharge from the hospital.

Objectives:

1. Reduce Readmission Rates: Identify high-risk patients for targeted interventions
2. Improve Patient Outcomes: Enable proactive care management and follow-up
3. Optimize Resource Allocation: Efficiently deploy limited healthcare resources to highest-risk patients
4. Reduce Healthcare Costs: Avoid financial penalties associated with high readmission rates

Key Stakeholders:

- Patients: Directly affected by readmission risk and subsequent care
- Healthcare Providers: Doctors, nurses, and care managers who implement interventions
- Hospital Administration: Responsible for quality metrics and financial performance
- Insurance Payers: Have financial interest in reducing unnecessary readmissions

---

Data Strategy (10 points)

Proposed Data Sources:

1. Electronic Health Records (EHR): Medical history, diagnoses, medications, lab results
2. Demographic Data: Age, gender, race, socioeconomic status, insurance type
3. Clinical Data: Vital signs, treatment procedures, length of stay, discharge disposition
4. Social Determinants: Living situation, social support, transportation access

2 Ethical Concerns:

1. Patient Privacy & Data Security: Handling sensitive health information requires strict confidentiality measures and robust security protocols to prevent data breaches.
2. Algorithmic Bias: Risk of perpetuating healthcare disparities if model underperforms for minority groups or penalizes patients from disadvantaged backgrounds.

Preprocessing Pipeline:

1. Data Cleaning:
    - Handle missing values: Impute numerical data (median), categorical data (mode/missing category)
    - Remove duplicates and correct data entry errors
2. Feature Engineering:
    - Create composite features: Number of medications, comorbidity scores, previous admission frequency
    - Temporal features: Time since last admission, seasonal patterns
    - Clinical risk scores: Calculate established medical risk indices
3. Feature Transformation:
    - One-hot encoding for categorical variables (diagnosis codes, discharge locations)
    - Standardization of numerical features (age, lab values, vital signs)
    - Handling class imbalance using SMOTE or weighted loss functions

---

Model Development (10 points)

Selected Model: Gradient Boosting Classifier (XGBoost)

Justification:

- Handles mixed data types (numerical and categorical) effectively
- Provides feature importance for model interpretability
- Robust to outliers and missing data
- High performance on tabular healthcare data
- Can capture complex non-linear relationships in medical data

Confusion Matrix (Hypothetical Data):

	Predicted: No Readmit	Predicted: Readmit
Actual: No Readmit	850 (TN)	50 (FP)
Actual: Readmit	40 (FN)	60 (TP)

Calculations:

- Precision = TP / (TP + FP) = 60 / (60 + 50) = 60/110 = 0.545
- Recall = TP / (TP + FN) = 60 / (60 + 40) = 60/100 = 0.60

Interpretation: The model correctly identifies 60% of actual readmissions (recall) and 54.5% of predicted readmissions actually occur (precision).

Deployment (10 points)

Integration Steps:

1. API Development: Create RESTful API wrapper around the trained model
2. EHR Integration: Connect with hospital's electronic health record system
3. Real-time Scoring: Trigger predictions during discharge planning process
4. Dashboard Interface: Develop clinician-facing interface showing risk scores and contributing factors
5. Alert System: Implement automated alerts for high-risk patients to care coordinators

HIPAA Compliance Measures:

1. Data Anonymization: Remove personally identifiable information before model training
2. Encryption: Implement end-to-end encryption for data in transit and at rest
3. Access Controls: Role-based access with strict authentication protocols
4. Audit Logging: Comprehensive logging of all data accesses and model usage

5. Business Associate Agreements: Ensure all third-party vendors comply with HIPAA requirements
6. Data Minimization: Collect and use only necessary data elements for prediction

---

Optimization (5 points)

Method to Address Overfitting: Regularization via Cross-Validation

Implementation:

- Use Stratified K-Fold Cross-Validation (k=5 or k=10) during training to ensure robust performance estimation across different data splits
- Apply L2 Regularization in the Gradient Boosting model to penalize complex trees
- Implement Early Stopping based on validation performance to prevent over-training
- Use Feature Selection to eliminate redundant or noisy features that don't contribute meaningfully to predictions

Benefits: This approach ensures the model generalizes well to unseen patient data while maintaining predictive performance, crucial for reliable clinical decision support.