AI Assignment Report

Part 1: Short Answer Questions (30 points)

1. Problem Definition (6 points)

Problem:
Predicting Student Dropout Rates in Universities.

Objectives:

Identify at-risk students early using historical academic and behavioral data.

Enable timely intervention through academic support systems.

Reduce institutional dropout rates by 20% over two academic years.

Stakeholders:

University Administration

Academic Counselors

Key Performance Indicator (KPI):

Dropout Prediction Accuracy (Target ≥ 85%)

2. **Data Collection & Preprocessing (8 points)**

Data Sources:

1. Student Information System (SIS):

Grades

Attendance

Disciplinary records

2. Learning Management System (LMS):

Logins

Assignment submissions

Discussion participation

Potential Bias:

Socioeconomic bias: Students from underprivileged backgrounds may have less access to LMS, leading to skewed data.

Preprocessing Steps:

Handle missing values (e.g., imputation for attendance).

Normalize numerical features (e.g., grades, time online).

One-hot encodes categorical variables (e.g., program type, department).

3. **Model Development (8 points)**

Model Choice:

Random Forest Classifier

Justification:

Robust to overfitting

Handles missing data

Provides feature importance

Data Split:

70% Training

15% Validation

15% Test

Key Hyperparameters:

1. n_estimators:
Controls the number of trees.

Too few = underfitting

Too many = slow training

2. max_depth:
Prevents overly complex trees and overfitting.

4. **Evaluation & Deployment (8 points)**

Evaluation Metrics:

F1-Score: Balances precision and recall; ideal for imbalanced datasets.

AUC-ROC: Measures the model's ability to distinguish between classes.

Concept Drift:

Definition: When statistical properties of input data change over time.

Monitoring Techniques:

Rolling window accuracy

KL-divergence-based drift detection

Deployment Challenge:

Scalability: The model must handle thousands of predictions during peak admission cycles.

Part 2: Case Study Application – Predicting 30-Day Hospital Readmission Risk (40 points)

1. Problem Scope (5 points)

Problem:
Predict patients at risk of being readmitted within 30 days post-discharge.

Objectives:

Reduce unnecessary hospital readmissions.

Optimize post-discharge patient care.

Stakeholders:

Hospital Administration

Healthcare Providers (Doctors, Nurses)

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

Data Sources:

1. Electronic Health Records (EHR):

Diagnoses

Medications

Previous admissions

2. Demographic Data:

Age

Gender

Insurance type

<u>Ethical Concerns:</u>

1. Patient Privacy: Risk of exposing sensitive health data.

2. Algorithmic Discrimination: Potential bias against elderly or minority patients.

Preprocessing Pipeline:

Handle missing vitals via mean/mode imputation.

Encode categorical variables (e.g., diagnosis codes via label encoding).

Feature Engineering:

Average length of stay

Number of past visits in the last 6 months

Discharge-to-follow-up duration

3. **<u>Model Development (10 points)</u>**

Model:

XGBoost

Justification:

Performs well on structured/tabular data.

Effectively handles class imbalance.

Hypothetical Confusion Matrix:

	Predicted: Yes	Predicted: No
Actual: Yes	45	10
Actual: No	15	130

Metrics:

Precision = 45 / (45 + 15) = 0.75

Recall = 45 / (45 + 10) = 0.82

4. **Deployment (10 points)**

Integration Steps:

1. Package model as a REST API using Flask or FastAPI.

2. Securely connect API to hospital EHR system.

3. Build a prediction dashboard showing patient risk scores to doctors.

Compliance Measures:

Use HTTPS & AES for data encryption.

Follow HIPAA protocols:

Role-based access control

Data logging

Retention policies

5. **<u>Optimization (5 points</u>**)

Method:

Apply dropout regularization or early stopping during training to prevent overfitting.

<u>Part 3: Critical Thinking (20 points)</u>

1. <u>Ethics & Bias (10 points)</u>

Impact of Bias:

Biased models may underpredict risk for marginalized patients.

Leads to inadequate post-care and higher complication rates.

Mitigation Strategy:

Perform fairness audits (e.g., demographic parity testing).

Retrain using a stratified and demographically balanced dataset.

2. <u>Trade-offs (10 points)</u>

Interpretability vs. Accuracy:

High-performing models (e.g., XGBoost) lack transparency.

Clinician trust and legal compliance demand explainability.

Solution:

Use SHAP values to explain individual predictions.

Computational Trade-off:

With limited resources, choose simpler models like Logistic Regression or LightGBM over deep learning.

Part 4: Reflection & Workflow Diagram (10 points)

Reflection (5 points)

Most Challenging Part:

Balancing model complexity with interpretability in healthcare scenarios.

Improvements if Given More Time:

Collect more longitudinal data.

Use AutoML for hyperparameter optimization.

Workflow Diagram (5 points)

```
+----------------------------+
| 1. Problem Definition      |
+----------------------------+
            ↓
+----------------------------+
| 2. Data Collection & Cleaning |
+----------------------------+
            ↓
+----------------------------+
| 3. Feature Engineering     |
+----------------------------+
            ↓
+----------------------------+
| 4. Model Training          |
+----------------------------+
            ↓
+----------------------------+
```

```
| 5. Model Evaluation        |
+----------------------------+
            ↓
+----------------------------+
| 6. Deployment              |
+----------------------------+
            ↓
+----------------------------+
| 7. Monitoring & Updates    |
+----------------------------+
```

References

scikit-learn documentation: https://scikit-learn.org

Azure Machine Learning: https://azure.microsoft.com/en-us/services/machine-learning/

SHAP library: https://github.com/slundberg/shap

HIPAA Guidelines: https://www.hhs.gov/hipaa