AI Development Workflow Assignment Report

1. Problem Definition

Short Answer (Part 1 Q1)

Problem Definition

Hypothetical Problem: Predicting student dropout rates using academic and demographic data.

Objectives:

1. Identify at-risk students with 80% accuracy.

2. Reduce dropout rates by 20% within one academic year.

3. Allocate counseling resources efficiently.

Stakeholders: School administrators, teachers.

KPI: Precision@80% recall (prioritize minimizing false negatives).

Case Study: Hospital Readmission Prediction

Problem: Predict 30-day readmission risk using EHR data.

Objectives:

1. Reduce avoidable readmissions by 15%.

2. Flag high-risk patients for post-discharge follow-ups.

Stakeholders: Clinicians, hospital administrators, insurers.

---

2. Data & Preprocessing

Short Answer (Part 1 Q2)

Data Collection & Preprocessing

Data Sources:

1. Student: Grades, attendance, socioeconomic status.

2. Institutional: Course difficulty, teacher ratios.

Potential Bias: Dataset lacks part-time student representation.

Preprocessing Steps:

1. Impute missing grades with subject-wise medians.

2. Normalize test scores (Z-score standardization).

3. One-hot encode categorical variables (e.g., school branch).

Case Study Data Strategy

Data Sources:

1. Structured EHRs (labs, diagnoses, medications).

2. Unstructured discharge summaries (NLP extraction).

Ethical Concerns:

1. Privacy: Anonymize PHI (Protected Health Information) per HIPAA.

2. Bias: Audit model for disparities by race/insurance status.

---

3. Model Development

Short Answer (Part 1 Q3)

Model Development

Chosen Model: Random Forest (handles mixed data types, robust to outliers).

Data Splits: 60% train, 20% validation, 20% test (stratified by dropout status).

Hyperparameters:

1. `max_depth=5` (avoid overfitting).

2. `class_weight='balanced'` (address class imbalance).

Case Study Model

Model: Logistic Regression (prioritize interpretability for clinicians).

Confusion Matrix:

	Predicted: No	Predicted: Yes
Actual: No	150	20
Actual: Yes	30	100

Metrics:

- Precision = 83% (TP/(TP+FP) = 100/120).
- Recall = 77% (TP/(TP+FN) = 100/130).

---

4. Deployment & Ethics

Short Answer (Part 1 Q4)

Evaluation & Deployment

Metrics:

1. AUC-ROC (handles class imbalance well).
2. F1-score (balances precision/recall).

Concept Drift: Monitor via monthly KS-tests on feature distributions.

Deployment Challenge: Latency → optimize with feature selection.

Case Study Deployment

Steps:

1. Dockerize model as REST API.

2. Integrate with hospital EHR using HL7/FHIR standards.

HIPAA Compliance:

- Data encryption in transit/at rest.

- Role-based access control (RBAC).

---

5. Critical Thinking

Ethics & Bias

Impact of Bias: Underrepresentation of uninsured patients could worsen care disparities.

Mitigation: Adversarial debasing during training.

Trade-offs

Interpretability vs. Accuracy: Use LIME/SHAP explanations with Logistic Regression (5% accuracy trade-off justified for trust).

Resource Limits: Prioritize lightweight models (e.g., Logistic Regression over deep learning).

---

6. Workflow Diagram & GitHub Setup

AI Development Workflow

```mermaid
```

```
flowchart TD
    A[Problem Scope] --> B[Data Collection]
    B --> C[Preprocessing]
    C --> D[Model Training]
    D --> E[Evaluation]
    E --> F[Deployment]
    F --> G[Monitoring]
```

GitHub Repository

Files:

1. preprocess.py:

python

```
Handle missing data
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='median')
X_processed = imputer.fit_transform(X_raw)
```

2. train_model.py:

python

```
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(max_depth=5)
model.fit(X_train, y_train)
```

3. README.md:

markdown

```
# AI Assignment
**Objective**: Predict student dropout/hospital readmission.
**KPI**: Precision@80% recall.
```