1. Problem Definition

Hypothetical AI Problem:

Predicting Student Dropout Rates in an online university.

Objectives:

Identify students at risk of dropping out early.

Recommend interventions based on risk level (e.g., mentorship, resources).

Improve student retention by 20% over one academic year.

Stakeholders:

University Administration – Interested in reducing dropout rates and improving institutional performance.

Students – Directly benefit from early interventions and support systems.

Key Performance Indicator (KPI):

Dropout Prediction Accuracy at Mid-Term (i.e., accuracy of correctly identifying students who eventually drop out by week 6).

2. Data Collection & Preprocessing

Data Sources:

Student Academic Records – GPA, course completions, grades, logins.

Learning Management System (LMS) Logs – Activity frequency, quiz scores, discussion forum posts.

Potential Bias:

Socioeconomic bias: Students from low-income backgrounds may be underrepresented or misrepresented in data, leading to biased predictions.

Preprocessing Steps:

Handle Missing Data – Fill or drop rows/columns with null values (e.g., missing quiz scores).

Normalize Features – Standardize numerical features (e.g., login count, GPA) to improve model performance.

Encode Categorical Data – Convert gender, major, or status into numerical format (e.g., one-hot encoding).

3. Model Development

Model Choice:

Random Forest Classifier

Justification: Handles both numerical and categorical data well, resistant to overfitting, and provides feature importance.

Data Splitting:

Training set: 70%

Validation set: 15%

Test set: 15%

This split ensures sufficient data for both training and generalization evaluation.

Hyperparameters to Tune:
n_estimators – Number of trees in the forest; controls model robustness and performance.

max_depth – Prevents overfitting by limiting how deep trees can grow.

4. Evaluation & Deployment
Evaluation Metrics:
F1-Score – Useful in handling imbalanced classes (e.g., more students stay than drop out).

AUC-ROC – Measures model's ability to distinguish between dropout and non-dropout classes.

Concept Drift:
Definition: When the underlying data distribution changes over time, causing the model to become less accurate.

Monitoring: Use periodic re-evaluation on new student cohorts and track performance drops.

Technical Deployment Challenge:
Scalability: The model needs to handle real-time predictions for thousands of students simultaneously. Solutions include model optimization or using cloud-based scalable services (e.g., AWS SageMaker, Google Vertex AI).