Part 1: Short Answer Questions

---

1. Problem Definition

AI Problem:

Predicting Student Dropout Rates

Many educational institutions struggle to identify students who are at risk of dropping out. An AI system can analyze academic, behavioral, and socio-economic data to predict which students are likely to leave school early and allow timely interventions.

Objectives:

1. To accurately predict students who are at high risk of dropping out.
2. To help educators design personalized intervention plans for at-risk students.
3. To improve student retention rates and overall institutional performance.

Stakeholders:

1. **Students** – benefit from early support and improved academic outcomes.
2. **School Administrators and Educators** – use insights to allocate resources effectively.

Key Performance Indicator (KPI):

- **Student Retention Rate (%)** – the percentage of students who remain enrolled after targeted interventions.

---

2. Data Collection & Preprocessing

Data Sources:

1. **School Management System (SMS):** Attendance records, grades, and disciplinary data.
2. **Student Surveys or Socio-Economic Databases:** Information about family background, financial status, and motivation levels.

Potential Bias:

- **Socio-economic bias:** Students from low-income backgrounds may be over-represented as “at risk,” leading to unfair predictions that reflect social inequality rather than true academic potential.

Preprocessing Steps:

1. **Handling Missing Data:** Impute missing attendance or grade records using mean/median or interpolation techniques.
2. **Normalization:** Scale numeric features like test scores to a uniform range to ensure fair comparison across variables.
3. **Encoding Categorical Variables:** Convert non-numeric data such as gender or region into numerical form (e.g., one-hot encoding).

---

3. Model Development

Model Choice:

Random Forest Classifier

Justification:

- Handles both numeric and categorical data efficiently.
- Robust to noise and overfitting due to ensemble averaging.
- Provides feature importance scores, which are useful for interpreting key dropout factors.

Data Splitting Strategy:

- **Training Set:** 70% of data (for model learning)
- **Validation Set:** 15% (for tuning hyperparameters)
- **Test Set:** 15% (for final model evaluation)

Hyperparameters to Tune:

1. **Number of Trees (n_estimators):** Affects model performance and computation time.
2. **Maximum Depth (max_depth):** Controls model complexity and prevents overfitting.

---

4. Evaluation & Deployment

Evaluation Metrics:

1. **Accuracy:** Measures the overall percentage of correct predictions.

2. **F1-Score:** Balances precision and recall, especially important when dropout cases are relatively few compared to non-dropouts (class imbalance).

Concept Drift:

- **Definition:** When the statistical properties of input data change over time, making the model's predictions less accurate.
- **Monitoring Approach:** Periodically retrain the model using recent data and compare real-world outcomes with predicted results.

Technical Deployment Challenge:

- **Scalability:** Deploying the model across multiple schools with varying data systems may require a cloud-based infrastructure and standardized data pipelines.

References:

1. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier.
2. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
3. Brownlee, J. (2020). *Machine Learning Mastery*. Machine Learning Mastery Press.