AI Development Workflow Assignment

Part 1: Short Answer Questions (30 points)

1. Problem Definition (6 points)

1) Define a hypothetical AI problem
   "Detecting fake news online"

2) List **3 objectives** and **2 stakeholders**.

Objectives

a) Accurately classify articles as fake or real.
b) Reduce misinformation spread on digital platforms.
c) Provide interpretability to users on why an article was flagged

Stakeholders

a) News consumers (general public).
b) Media regulators/platforms

3) Propose **1 Key Performance Indicator (KPI)** to measure success.

F1-Score – (balancing precision and recall).

2. Data Collection & Preprocessing (8 points)

1) Identify **2 data sources** for your problem.

   a) Kaggle Fake News Dataset
   b) Credible news sources' RSS feeds

2) Explain **1 potential bias** in the data.

   a) Political bias in labeled training data could skew results toward certain narratives.

3) Outline **3 preprocessing steps** (e.g., handling missing data, normalization).

   a) Handle missing headlines/text.

Problem:

- Some news articles may have missing headlines or body text.

Solution:

- Remove records with completely empty text fields (as they provide no information).
- For partially missing fields, you may use placeholders (e.g., "unknown") or impute with average text length or titles.

2) Text normalization (lowercasing, punctuation removal).

Goal:

- Standardize the text so the model can learn patterns better.

Steps:

- Convert all text to lowercase (`"Breaking News"` → `"breaking news"`).

- Remove punctuation and special characters.
- Strip extra white spaces and tabs.

3) Tokenization and TF-IDF vectorization.

- **Tokenization**: Break sentences into words/tokens.
  Example: `"COVID cases rise"` → `["covid", "cases", "rise"]`
- **Stopword Removal**: Remove common words like "the", "is", "in" that don't add much meaning.
- **Stemmin**: Reduce words to their base forms.
  `"running", "ran"` → `"run"`
- **Vectorization**: Convert text to numbers using:

  a. **TF-IDF (Term Frequency-Inverse Document Frequency)** or
  b. **CountVectorizer** for bag-of-words model.

3. Model Development (8 points)

a. Choose a modeland justify your choice.

- Logistic Regression or Random Forest (interpretable, effective for text features)

b. Describe how you would split data into training/validation/test sets.

- 70% Train / 15% Validation / 15% Test

c. Name **2 hyperparameters** you would tune and why.

- `max_depth` (to avoid overfitting in Random Forest)
- `n_estimators` (to improve model robustness)

4. Evaluation & Deployment (8 points)

a) Select **2 evaluation metrics** and explain their relevance.

- Precision (important to avoid false positives)
- Recall (important to catch all fake news)

b) What is **concept drift**? How would you monitor it post-deployment?

- **Concept Drift**:
  Occurs when patterns in news change over time.
- **Monitoring**: Track accuracy and re-train with recent data.

c) Describe **1 technical challenge** during deployment (e.g., scalability).

- **Scalability**—real-time classification at high traffic loads.