🧑‍💻 AI Interview Task: Email Spam Classifier

Objective

Build a machine learning model that can classify whether an email is spam or not spam based on its text content.

Step 1. Data

Use the Spam SMS/Email dataset (e.g., SMS Spam Collection from Kaggle).

Columns: label (spam/ham), message (text).
Target: label.

Step 2. Preprocessing

1. Clean text (lowercase, remove punctuation, etc.)
2. Convert text into numeric features using:

- Bag of Words (CountVectorizer)
- TF-IDF (TfidfVectorizer)

Step 3. Build a Model

Train at least one ML classifier:

- Naive Bayes (recommended for text)
- Logistic Regression
- Random Forest

Step 4. Train & Evaluate

Split dataset: 70% train, 30% test.

Report:

- Accuracy
- Confusion Matrix

Step 5. Deliverables

The candidate must provide the following:

1. Dataset used (or source link)
2. Codebase (Jupyter Notebook or Python scripts)
3. Video demonstration of running the project
4. Comparative analysis of algorithms (e.g., Naive Bayes vs Logistic Regression vs Random Forest)
5. Full report documenting the entire process:
   - Collecting dataset
   - Cleaning and preprocessing
   - Building and training models

  - Evaluation metrics
  - Comparison of results

6. A short README (few lines) explaining how to run the code.

Hints for Beginners

- Use pandas for data handling
- Use scikit-learn for vectorization & ML models
- Naive Bayes is a great baseline for text classification