**Email Spam Detection**

**1. Collect Data**

Use public datasets:

* *SMS Spam Collection Dataset (UCI ML Repository / Kaggle)*
* *Enron Email Dataset*
* *Or scrape emails/messages (if allowed)*

Data format: Each row should have text (message/email) and label (spam/ham).

**2. Preprocess Data**

* *Clean the text so the model can learn patterns.*
* *Lowercasing*
* *Removing punctuation, numbers, stopwords*
* *Tokenization (splitting into words)*
* *Lemmatization or stemming (normalize words)*

**3. Feature Extraction**

Convert text into numbers that ML models can use.

1. *Bag of Words (BoW)*
2. *TF-IDF (Term Frequency–Inverse Document Frequency)*
3. *Word embeddings (Word2Vec, GloVe, BERT) – advanced*

**4. Train ML Models**

Popular algorithms for spam classification:

* *Naive Bayes (MultinomialNB) → simple, works very well*
* *Logistic Regression*
* *Random Forest*
* *SVM*
* *Deep learning (LSTMs, Transformers) – for advanced stage*

**5. Evaluate**

* *Use accuracy, precision, recall, F1-score*
* *Precision & recall are important because:*
* *High false positives → valid emails marked spam*
* *High false negatives → spam emails missed*

**6. Save Model & Deploy**

* *Save model with joblib or pickle*
* *Create a Flask / FastAPI app or deploy with Streamlit*
* *User inputs a message → model predicts spam/ham*
* *Example save & load:*