□ Internship Report – Resume Classification (Option A)

□ 1. Problem Statement

The goal of this project is to build a machine learning model that can classify resumes into different job categories, similar to an Applicant Tracking System (ATS). This helps recruiters automatically filter and categorize resumes.

□ 2. Dataset

- Source: Kaggle Resume Dataset
- Shape: ~965 resumes across multiple categories (e.g., Data Science, Java Developer, HR, etc.)
- Columns:
  - `Resume_str` → Raw resume text
  - `Category` → Target label (job domain)

□ 3. Approach

Step 1: Preprocessing

- Converted resumes to lowercase
- Removed numbers, punctuation, and special characters
- Removed stopwords (NLTK)
- Stored cleaned text

Step 2: Feature Extraction

- Used **TF-IDF Vectorizer** with `max_features=5000`
- Converted text into numerical vectors

Step 3: Model Training

- Split data into **80% train / 20% test**
- Trained a **Logistic Regression classifier** (baseline model)

Step 4: Evaluation

- Evaluated using **Accuracy, Precision, Recall, F1-score**
- Generated **Confusion Matrix**

□ 4. Results

Classification Report (sample output)

```
              precision    recall  f1-score   support

   Category1       0.92      0.91      0.91        XX
   Category2       0.89      0.87      0.88        XX
   ...
   Accuracy                             0.90       XXX
   Macro avg      0.89      0.89      0.89       XXX
   Weighted avg   0.90      0.90      0.90       XXX
```

Confusion Matrix (screenshot)

(Insert the heatmap from your Colab here)

□ 5. Tools & Libraries

- Python (Google Colab)
- Libraries: Pandas, NumPy, scikit-learn, NLTK, Matplotlib, Seaborn
- Dataset: Kaggle Resume Dataset

□ 6. Conclusion & Future Work

- Logistic Regression with TF-IDF achieved good baseline accuracy (~85–90%).
- The model can classify resumes into categories reliably.
- **Future improvements:**
    - Use **BERT / Transformer embeddings** for deeper semantic understanding
    - Build a **web app (Streamlit/Flask)** for real-time classification (Option B)