Deliverable 1 : Spam Email Classifier

1. Choice of dataset
   The dataset I chose is the Phishing Email Dataset from Kaggle. It was created by combining multiple datasets such as Enron, Ling, Nigerian Fraud and more. The dataset includes real-world spam and phishing emails which makes it ideal for building a spam classifier. It is relatively large, containing 82 000 emails (50% labeled as spam and the other 50% as non spam) giving me a lot of data to train my model.

2. Methodology
   a. Data Preprocessing
      The dataset contains features such as the body of the email, and also the which are all useful to build the classifier. However, some entries only contain the body. To process my data, I will convert all text to lowercase, remove punctuation and stop words. Then, I'll use TF-IDF (Term Frequency-Inverse Document Frequency) to vectorize text.

   b. Machine Learning Model
      I propose to use a random forest classifier because it provides feature importance scores. This will help me identify which features (specific keywords for example) are most indicative of spam.

   c. Evaluation matrix
      Accuracy, Precision, Recall, and F1-score. I want to achieve an accuracy of at least 90% and an F1-score above 0.85.

3. Application
   I will build a web app using Flask where users can input emails. The model will classify the input as either Spam or Not Spam with a confidence score. I would also like, if possible, to display the factors influencing the classification, such as presence of certain keywords or suspicious patterns in the email body.