**About ready\_for\_training.csv**

This file is the clean, feature-rich, and final dataset for our models. It combines both the phishing and normal emails and has been shuffled randomly.

**Key Characteristics:**

* **Composition:** It contains approximately 20% phishing emails and 80% normal emails.
* **Target Variable:** The crucial column for training is label.
  + label = 1 means the email is **Phishing/Scam**.
  + label = 0 means the email is **Normal/Legitimate**.

**Feature Columns:**

The dataset contains two types of features that we engineered:

**1. Numerical Features:** These columns are ready to be used directly by the models. They capture important characteristics of the emails.

|  |  |
| --- | --- |
| Feature Name | Description |
| num\_hyperlinks | The total count of URLs in the email. |
| num\_suspicious\_links | Count of links from known risky domains (e.g., forms.gle, .ipfs). |
| num\_ip\_urls | Count of links that use a raw IP address instead of a domain name. |
| body\_char\_count | The total number of characters in the email body. |
| subject\_char\_count | The total number of characters in the subject line. |
| urgency\_keyword\_count | Count of words like "urgent", "verify", "payment", "suspension". |
| num\_capital\_words | The number of words written in ALL CAPS. |
| special\_char\_ratio | The ratio of special characters (!, $, &) to total characters. |
| has\_javascript | 1 if the email contains a <script> tag, 0 otherwise. |
| has\_form\_tag | 1 if the email contains a <form> tag, 0 otherwise. |

**2. Text Feature:** This column requires vectorization before it can be used.

|  |  |
| --- | --- |
| Feature Name | Description |
| processed\_text | The raw email text after being fully cleaned (HTML removed, tokenized, lemmatized, and stop words removed). **This column is the intended input for a TF-IDF vectorizer.** |

**Next Steps**

You can now load ready\_for\_training.csv and begin training the models as planned (Naive Bayes, Logistic Regression, Random Forest).