Spam Email Detection using Logistic Regression

---

1. Objective

The objective of this project is to develop a machine learning model that can automatically detect whether an email is spam or not. This helps improve email security and user experience by filtering unwanted messages.

---

2. Dataset

The dataset used is `spam.csv` (SMS Spam Collection Dataset from Kaggle/UCI).
It contains two columns:

- `label` → 0 = Not Spam (Ham), 1 = Spam
- `message` → Text content of the email or SMS

---

3. Preprocessing

Before training the model, the text data was preprocessed as follows:

1. Converted all text to lowercase
2. Removed punctuation and numbers
3. Removed stopwords (common words like “the”, “is”, “in”)
4. Tokenized the text
5. Converted text into numerical features using **TF-IDF vectorization**

---

4. Model Development

- **Algorithm Used:** Logistic Regression

- **Train-Test Split:** 70% training, 30% testing
- Optional comparison with other models such as Naive Bayes or SVM can also be done.

---

5. Model Performance

The Logistic Regression model achieved an accuracy of approximately **0.96** *(replace with your actual result)*.
The model is able to correctly classify most spam and non-spam emails, making it effective for practical spam detection.

---

6. Important Words for Spam Detection

The model identified the following words as most indicative of spam:

- free, win, prize, click, urgent, cash, claim, now, call, txt

---

7. Conclusion

- Logistic Regression effectively classifies emails as spam or not spam.
- Performance can be further improved with more data or by trying other classifiers.
- The Streamlit app allows users to test new email content interactively.
- This project demonstrates a simple yet practical approach to spam detection using text processing and machine learning.
