DATA 620 Final Project: Spam Detection

\*\*Author:\*\* Ariba Mandavia

\*\*Course:\*\* DATA 620

\*\*Project Title:\*\* Spam Classification Using Logistic Regression and Naive Bayes

# 📘 Project Overview

This project focuses on classifying emails as spam or ham using the Kaggle spam-ham dataset. It implements a full NLP pipeline, including text preprocessing, TF-IDF vectorization, model training, evaluation, and network analysis.

# 🔍 Methods Used

- Text preprocessing (lowercasing, punctuation removal, stopword removal, stemming)  
- TF-IDF vectorization with unigram and bigram features  
- Classification using Multinomial Naive Bayes and Logistic Regression  
- Evaluation using accuracy, precision, recall, F1-score, and ROC-AUC  
- Error analysis and cross-validation  
- Word co-occurrence network analysis on spam emails

# 📈 Results Summary

Logistic Regression achieved the best performance with 98.1% accuracy and 0.99 recall for spam. The model effectively identified spam-indicative features such as 'click', 'money', and 'free'. Network analysis revealed clusters of commonly co-occurring spam words, adding interpretability to the model's behavior.

# 🛠️ Tools & Libraries

- Python (Pandas, NumPy, Scikit-learn, NLTK, NetworkX, Matplotlib, Seaborn)  
- Jupyter Notebook for reproducible analysis  
- Dataset: spam\_ham\_dataset.csv from Kaggle

# 🧠 Key Takeaway

With proper preprocessing and feature engineering, even simple, interpretable models can achieve high performance in spam detection. Combining machine learning with NLP and network analysis enhances both model accuracy and explanatory power.