# AI Project: Spam Email Classifier

## Summary

This project aims to develop an AI model to classify emails as "Spam" or "Not Spam" using machine learning. It uses a dataset of 1000 email messages, with a focus on optimizing classification accuracy for imbalanced data sets where the number of legitimate emails vastly exceeds the number of spam emails.

This is a Building AI course project.

## Background

The problem of email spam is common, with an increasing number of spam messages cluttering people's inboxes. Filtering these messages is essential for productivity and data security. The goal is to build a classifier that can accurately distinguish between legitimate emails and spam.

- \*\*Problem 1:\*\* The high volume of spam emails.

- \*\*Problem 2:\*\* The imbalance between spam and non-spam emails in typical datasets.

- \*\*Personal Motivation:\*\* As spam emails can be a huge distraction and security threat, an effective classification model can save time and resources for individuals and organizations.

## How is it used?

The solution will be used in email filtering systems to classify incoming messages. Users will benefit from a cleaner inbox, while organizations can protect their data from potential threats. The system will need to handle both training and real-time classification, adapting to new types of spam that evolve over time.

## Data sources and AI methods

Data for this project will come from a publicly available dataset of email messages. The classifier will be trained using the following methods:

- \*\*Naive Bayes Classifier\*\*

- \*\*Logistic Regression\*\*

- \*\*Support Vector Machine (SVM)\*\*

These methods are effective for text classification tasks, particularly with imbalanced datasets.

## Challenges

This project doesn't solve the problem of completely eliminating spam, as there will always be new methods for spammers to evade filters. Additionally, ethical concerns include privacy, as users' email contents must be handled securely, without violating privacy laws.

## What next?

The project could grow by incorporating more advanced techniques, such as deep learning, for better accuracy in detecting spam emails. A key next step is to expand the dataset and fine-tune the models. Collaboration with cybersecurity experts could improve the solution further.

## Acknowledgments

- Dataset from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/spambase)

- [OpenAI](https://openai.com) for inspiration in using advanced NLP techniques.