Spam vs Ham Detection Case Study Rubric

DS 4002 - Fall 2025 - Raj Bhowmic
Due: Dec 9
Submission format: Upload pdf and link to GitHub repository on UVA Canvas

Individual Assignment

Why am I doing this?

This case study provides you an opportunity to apply your skills in data science to a common problem in society with the rise of spam text messages. With how common SMS spam messages are and how hard it is to detect them, data-driven approaches are needed to protect users from phishing, fraud, and harmful content. This case study is similar to what you could face in a researching or industry role where you transform raw datasets into models providing insightful analysis on everyday common problems.

What am I going to do?

The GitHub repository for this case study can be found at https://github.com/Raj-B-1/DS4002_CS3. You will analyze a real dataset of SMS messages from this respiratory (located under “DATA” folder) and design a machine learning model capable of distinguishing spam from “ham” (valid) messages. Drawing on your understanding of preprocessing, modeling, and evaluation, you will create a system that can reliably detect harmful spam messages. Deliverables include:

- A written PDF report
- A GitHub repository containing code and necessary data used

How will I know I have succeeded?

You will meet expectations on this case study when you successfully follow and complete the criteria in the rubric below:

Spec Category	Spec Details
Formatting	• One GitHub Repository (Submitted via link on Canvas) ○ Create new repository for this assignment titled ‘CS2_SMSSpamClassification’ that

	contains: ■ README.md ■ LICENSE.md ■ Source Code Files ■ Your data (i.e., the images you choose) ■ Written PDF
README.md	● Brief summary of what you produced for the case study, provide enough information to navigate people through repository
Written Portion	Discuss analysis of study and thought process by writing ● One paragraph on summarizing the problem ● One paragraph on the plan to achieve deliverables ○ Include graphic demonstrating the analysis plan ● Discuss results and significance of the study ● One reflection paragraph describing challenges met while replicating the case study and how you solved them, discuss what you could do to improve in the future
Source Code	Code should include ● An exploratory data analysis (EDA) to explore the dataset through class distribution plots, message length histograms, and other visualizations ● Trained models that implement Logistic Regression, Linear SVM, and Random Forest on both the original dataset and the SMOTE-balanced dataset ● Tuning to optimize performance for each model Screenshots of: ● Comparative metrics such as accuracy, precision, recall, and F1-scores, as well as visualizations to support the findings
References	● At the end of the written portion of the assignment, include a list of references in IEEE citation style not in the provided reference materials