Rubric: Spam or Ham Detection

DS4002 Case Study – Spring 2024 – Natalie Assaad

Submission Format: upload link to GitHub repository in Canvas.

Individual Assignment

Why am I doing this? This case study gives you the opportunity to practice your machine learning skills in studying three different models: K-Nearest Neighbors, Logistic Regression, and Decision Trees. Through the reference materials and your work in completing the code, you will be exposed to a real-world application of machine learning in identifying spam emails.

What am I going to do? In assignment, you will use the GitHub repository provided in the Hook Document to access the necessary reference materials, datasets, and scripts to replicate the study. First you will read through the reference documents to familiarize yourself with the topic. Then, you will download the preprocessing and analysis scripts which will each be run in Jupyter Notebook. Each script is partially complete, but you will have to use your data science knowledge as well as the reference materials to write the code where instructed. You will be writing the code for preprocessing the textual data in the preprocessing script as well as the code for the KNN and Decision Tree models in the analysis script. Deliverables include:

- GitHub repository including:
    - Datasets
    - Completed preprocessing and analysis scripts, including descriptive comments
    - Output files
    - Write-up analyzing the results and drawing a conclusion on the most effective model – one page maximum, PDF format
    - README.md
    - REFERENCES.md

Tips for Success:

- Include copious amounts of comments in your code to help keep track of your steps
- Create informative names for your variables – with three different models, this is vital to ensure your code is readable and runs successfully
- Read the reference materials if you have difficulty completing/understanding any part of the code

How will I know I have Succeeded? You will meet expectations on this assignment when you follow the criteria in the rubric below.

Category	Details
Formatting	● GitHub Repository ○ Submit a link to your repository on Canvas

	○ All materials used in this project should be in the repository ○ Contents ■ Data ■ Code ■ Outputs ■ Write-up in PDF format ■ README.md ■ REFERENCES.md
GitHub Repository	● Goal: This repository will contain all materials that you used in the case study. ● DATA folder ○ Include each dataset used in CSV format (emails.csv, preprocessed_data.csv) ● CODE folder ○ Include the completed preprocessing script and completed analysis script (i.e., completed the code for textual preprocessing and for KNN and Decision Tree models) ○ Ensure that each file has descriptive comments and an effective header comment to orient others to the script ● OUTPUT folder ○ Include relevant results, graphs, tables, etc. with informative file names ○ You do not need to include visuals from the EDA ● Write-up ○ ~1 page ○ PDF file ○ Describe the result of your analysis and draw conclusions on the most effective model based on precision and accuracy values ○ Questions you may consider answering in your write-up: Which model was most effective? Why do you think this was the case? How did the precision and accuracy values compare across models? How can this study be improved? ● README.md ○ Describe the project goal and each part of the repository to orient new viewers to the study ● REFERENCES.md ○ Include all references used in IEEE format

Acknowledgements: Thank you to Professor Alonzi for providing the structure of this rubric.