Daniel Plascencia

CS4680 – Prompt Engineering Assignment 1

1. Problem Identification
   1. There are many scam messages out there in the world attempting to lull people into a scam where their information can be stolen for nefarious purposes.
   2. I am to predict if a given SMS message is a legitimate message or a spam message meant to deceive receivers
   3. Given that this is a classification problem the only two real features necessary for classification would be the message itself as well as the identification detailing if the message is real or spam
2. Data Collection.
   1. Data will be provided by Kaggle utilizing a classification model consisting of a label identifying whether the given message is a spam message or not as well as the message itself as the only necessary features
   2. The dataset at Kaggle "uciml/sms-spam-collection-dataset" contains about 5,500 datapoint entries where 80% will be utilized for training and 20% utilized for testing
3. Model Development
   1. The model used within scikit-learn will be LogisticRegression where the model will predict the probability of the given data belonging to a certain class (in this case determining the probability of belonging to the ham or spam classes)
4. Model Evaluation
   1. The model contains a 95% accuracy in identifying spam and ham messages  
      A screenshot of a computer screen

      AI-generated content may be incorrect.
   2. The model is inherently made for classification due to the nature of needing to classify the message between the two available labels but can be used within a regression model where instead each message is given a likelihood of being spam value between 0 and 1. Though the essence stays the same of classifying messages between spam or ham.
5. Documentation and Code Submission
   1. The code is available on GitHub with comments detailing the various steps of the process.