Project Idea

The main idea of the project is spam vs ham classification. Various traditional machine learning and an LSTM based deep learning model with word embedding are employed for the task. The machine learning algorithms include the following.

- Logistic Regression.
- Support Vector Machine.
- Multinomial Naïve Bayes.
- Decision Tree Classifier.
- K Nearest Neighbor.
- Random Forest Classifier.

The dataset used for the task is from the UCI datasets known as spambase. The dataset is preprocessed and the features are already extracted. The features include word frequency count of various words. The dataset is split into training and testing set. The training set is used to train the model while the testing set is used to evaluate the performance of the models.

For the second approach, LSTM based deep learning model is used with a word embedding layer. The embedding layer weights are loaded from the GLOVE model which is a pre-trained word embedding model available as open source. The model is trained and evaluated on the SMSspam dataset obtained from the same UCI datasets.