



Final Capstone Project SPAM Dataset

The SMS Spam Collection Dataset is a set of SMS tagged messages that have been collected for SMS Spam research. The SMS messages are in English composed of 5,572 messages, tagged according being legit or spam. The SMS Spam CSV file has a total of 4,825 SMS legitimate messages (86.6%) and a total of 747 (13.41%) spam messages.

Create a predictive model to predict Spam and Legit. You can use any methodology (classical or modern). The main KPI is the **average of class accuracy** so if you just predict 100% legit (100% accuracy class legit) , and 0% spam (0% accuracy class spam), then your grade will just be 50%. You can use R or Python.

Submit a word file detailing the preprocessing steps you have done, the modelling (parameters used) and the final generalization error values you have obtained. Also include the *.r or *.py code in your submission. Submit it via a zip folder named “Surname-198 Finals.zip” in UVLE.

Deadline is on Tuesday, January 18, 2022, 12 noon.