**Project Title**

Robert Chavez, Kiera Conway

**CSC 723 – Machine Learning for Cyber Security – Project Topic Proposal – Spring 2023**

**Proposal Content**

Our group’s final project for the Spring 2023 semester of CSC 723 will utilize an SMS spam dataset to train, test, and analyze various classification algorithms to find the most precise option. Precision and accuracy will be determined using a combination of model scores, confusion matrices, and other model metrics, when applicable. While the math deriving most of these metrics will be computed using the Python library , any formulas used will be thoroughly explained. Furthermore, the training and testing algorithms will be created using Python and submitted alongside the final project report.

Any scripts written for this project will be created using the Jupyter Notebook environment Kaggle and will be submitted in the .ipynb Kaggle file format. We chose Kaggle as our main environment due to the abundance of datasets, the ability to directly import said datasets, and the facilitation of cooperation between group members. While each algorithm will be trained and tested in a separate notebook, the final report will consolidate the data in an organized manner. Any additional materials which require group collaboration, including outlines, notes, and resources, will be shared via a Github repository.

As we will be classifying discrete, fixed, binary values (SPAM v HAM), our group will use various Supervised Classification algorithms. The three algorithms we plan to test, train, analyze, and compare are K-Nearest Neighbors (KNN), Decision Tree, and Naive Bayes (including Bayes’ Theorem to calculate the joint density of dependent events). While each algorithm script may contain additional, unique libraries, the primary libraries utilized in will be , , and . Each algorithm will be trained and tested using the SMSCollection.csv from the spam-or-ham dataset from Kaggle. Prior to any testing, the dataset will be thoroughly reviewed in Python; This review will contain general information, specific observations, and statistical information. Then, each step during the training and testing process with be explained using code examples for support. Lastly, all three algorithms will be tested and their model metrics compared to find the best option for the SPAM classification algorithm.

All team member contributions will be equal, as we share responsibility for research, scripting, consolidation, writing, and reviewing. As previously mentioned, all tools have been selected with cooperation as the driving factor. All research materials are submitted to a shared ‘Data Dump’ folder, scripting is done via a shared notebook, and written documents are regularly updated on our shared repository. We have also set up a tentative schedule to track our progress throughout this project, and meet weekly, via a Discord call, to discuss plans and strategies. Furthermore, the algorithms and report sections will be split evenly to maintain workload symmetry.

**Resources**

https://www.kaggle.com/datasets/arunasivapragasam/spam-or-ham