Project Proposal

COMP 7118

Data Mining

Spring 2022

Krishnaveni Mandadi  
U000829899

Contact email: [kmandadi@memphis.edu](mailto:kmandadi@memphis.edu)

Should be submitted by February 9, 2019

# What is the problem/application?

Determine whether a given email is spam or not

# What methods will be tested or implemented?

* Binary classification (with Support Vector Machine (SVM) Algorithm)

# What data sets will be used?

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that the machine learning community uses to test machine learning algorithms empirically. Dataset is available at [UCI Machine Learning Repository: Spambase Data Set](http://archive.ics.uci.edu/ml/datasets/Spambase)

The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets.

Dataset is all about spam emails, Spam can take many forms: product/website ads, get-rich-quick scams, chain letters, pornography, and so on. Spam e-mails were sent to us by our postmaster and spammers.

# What are the potential challenges for implementation?

I am anticipating some challenges concerning selection of right algorithm and time management in completing within due dates.

# What are the expected deliverables?

Initial observations

* Nature of source datasets
* Previous/similar works

How the solution is implemented:

* Discussion of algorithms used
* Prepare trained data and test data

Results and analysis:

* Visualizing the distribution of data & their relationships
* boosting algorithms perform the best on the current dataset
* Calculate accuracy and F1 score.

Conclusions

* Thoughts on potential future works