Dataset and Goal

For our project, we decided to work with the the Spam data that will allow us to classify emails as spam or not spam based on the frequency of some words.
The dataset has 4601 examples and 57 attributes which includes the label $(0 : notspam, 1 : spam)$.
Link to the dataset: $https://archive.ics.uci.edu/ml/datasets/Spambase$.

Software/Methods

We plan to use Decision Trees, Naive Bayes and Neural Networks on our dataset. We will be using sklearn to run the Decision Tree and Naive Bayes algorithms and tensorflow to run Neural Networks.
Decision Tree Classifier:
$https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html$
Naive Bayes (we will use the Multinomial Naive Bayes option):
$https://scikit-learn.org/stable/modules/naive\_bayes.html$

Scientific Question

Our hypothesis is that Naive Bayes will achieve the highest accuracy when guessing if something is spam or not because it assumes independence among its predictors.

We are interested in investigating and looking into this dataset because we talked about spam classification in class briefly, but it is something that we take for granted every day when we use our email and other applications. We are interested in investigating deeper to see how it works and to see how machine learning is applied in people's everyday life.

Evaluate and Interpret Results

To visualize the results of our data we will use the mathplotlib package to create graphs to see which algorithms have the best accuracy on the data and which hyperparameter value worked best. For example, if Naive Bayes had the best accuracy out of the three, which hyperparameter value could improve this accuracy?

References

Xavier Carreras and Lluis Marquez, "Boosting Trees for Anti-Spam Email Filtering" Proceedings of RANLP-2001 (2001).
S. K. Tuteja and N. Bogiri, "Email Spam filtering using BPNN classification algorithm" 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (2016).