ELSI Machine Learning Project Overview

Molly Millar

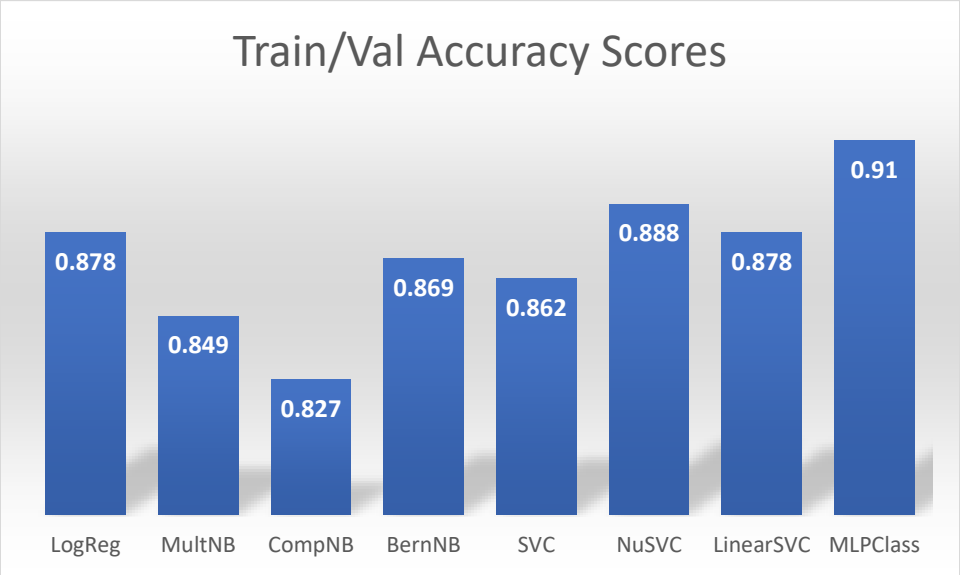
Summary: To determine whether machine learning is a better text mining method to find ELSI papers than a library search string, a dataset of ELSI papers and non-ELSI papers were compiled to train and test multiple supervised machine learning methods from Scikit Learn for efficiency and accuracy of predicting ELSI papers. After testing each of the different methods, each of the accuracy scores were high but the Linear SVC method proved to be the most effective with a score of 0.873.

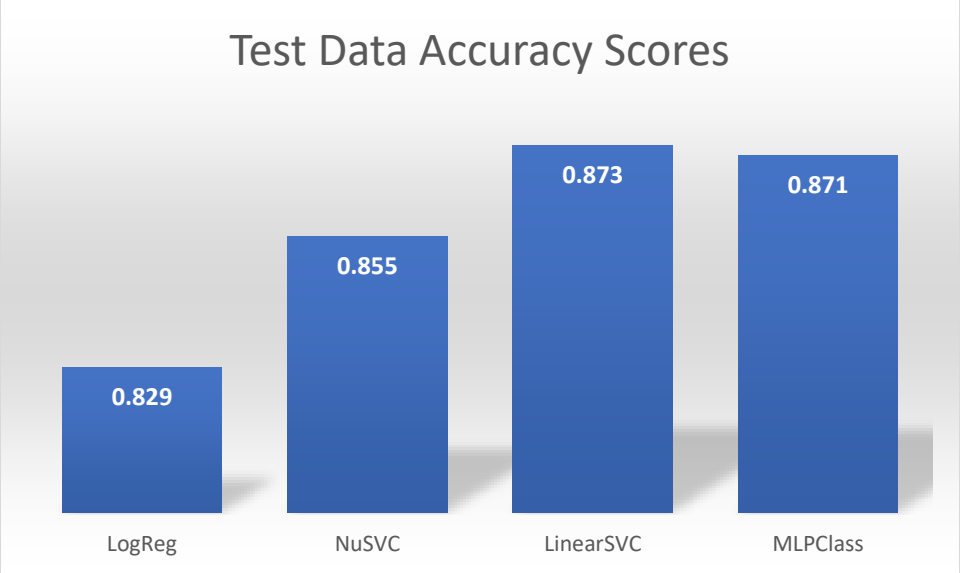
Background: The ELSI research hub compiles papers on the ethics of genetics and genomics research across multiple disciplines. The team compiling this list of papers would like a more efficient and effective way of finding ELSI research papers rather than a manual search with a library search string. The goal is to build and test if a machine learning approach for text mining is a more efficient and effective method for discovering new ELSI research papers.

Data: For this project, the data was composed of 779 "positive" papers and 779 "negative" papers. The positive papers were ELSI research papers gathered from the list already compiled for the ELSI research hub and marked as "True" under the feature "is_elsi_paper". The negative papers were research papers pulled from PubMed that had similar MeSH terms but concluded to be not completely ethics and genomics related and marked as "False" under the feature "is_elsi_paper". Once gathered, the titles and abstracts from the 1558 papers were randomly split into training, validation and testing datasets (20%, 20% and 60%, respectively).

Methods: The data was run through a word vectorizer to build a vocabulary of words used in the titles and abstracts, then split into the training, validation and testing data. From there, eight different supervised machine learning methods from Scikit Learn (Logistic Regression, Multinomial Naive Bayes, Complement Naive Bayes, Bernoulli Naive Bayes, Support Vector Classifier, Nu-Support Vector Classifier, Linear Support Vector Classifier, and Multilayer Perception Classifier) were trained using the training data and tested using the validation data to predict "True" or "False" for "is_elsi_paper". Accuracy scores and confusion matrices were gathered from the results of the tests on the validation data. After comparing the scores from the validation data, the top four methods were chosen and tested using the testing data.

Results: While all the machine learning methods produced accuracy scores with the training and validation data, the highest scores came from Logistic Regression, NuSVC, Linear SVC, and MLP Classifier. These four methods were tested again using the testing data, where Linear SVC proved to have the highest score, meaning Linear SVC predicted ELSI papers with the most accuracy.





Conclusion: The Linear SVC method proves to be the most accurate machine learning method for predicting ELSI papers, with the highest accuracy score. The next steps would be to use this model on an untested corpus of documents to see if it can find new ELSI papers which I had started to do with a corpus of 10000 papers. The accuracy scores from the machine learning models can also be tested against the accuracy of a final library search string provided by those at the ELSI hub to check if machine learning is truly more efficient and effective. Another option is to find which words are the biggest contributors to predicting ELSI papers and in return which words could be producing false positives or false negatives.