asm_manuscript

Joanna Colova

- Abstract
- Importance
  - publishing data helps get more use out of research
  - helps eliminate file drawer effect as it shows more negative data
  - want to incentivize authors to publish/make available their original data
- Keywords
  - data accessibility
  - data reproducibility
- Introduction
  - NIH funded research must make data available as of January 2023 (Policy for Data Management and Sharing (NOT-OD-21-013))
  - investigate metrics of making data publicly available in 12 ASM journals
  - DNA sequencing efforts are commonly uploaded to databases
    - want to evaluate how well this community is using reporducible data analysis
- Results
- Discussion
- Materials and Methods
  - oringinal dataset from adena (which i think is from crossref)
  - hand identifying 500 papers for da and nsd status
  - training model using mikropml methodology
  - picking a model (glmnet, rf, xgbtree, picked rf)
  - training of the models
  - hypertuning parameters (rf = mtry)
  - Snakemake/python
  - crossref gathering of DOIs
  - webscrape using httr2/rcrossref/wget
  - cleaning of html of each indiviudal file
  - tokenizing, stemming/lemmitization
  - formatting/applying zscore (and replicating that for the rest of the datasets)

  - using the model to predict da/nsd

- Supplemental Material file list (where applicable)
- Acknowledgments
- References