Data Accessibility Paper

Joanna Colovas

- Abstract
- Importance
  - Incentivize authors to publish/make available their original data
  - Publishing data helps get more use out of research
  - Helps eliminate file drawer effect as it shows negative 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 reproducible data analysis as a metric
    - International Nucleotide Sequence Database Collaboration (INDSC) databases
  - COVID sequencing and data availability was essential to vaccine development
- Results
  - Models
  - Model prediction results
  - Confusion Matrices for each model
    - Spot checking and error methodology
  - Regression modeling
  - Regression modeling/confusion matrices for papers that contain new sequence data
  - Citations corrected for time
- Discussion
  - Making data available provides more citations per paper than not doing so.

  - Allows for replication of studies
  -

- Materials and Methods
  - Original data set from Adena (which I think is downloaded from crossref)
  - Hand identifying 500 papers for da and nsd status
  - Training model using mikropml methodology
  - Training of the models
  - Picking the best model (glmnet, rf, xgbtree, picked rf)
  - Hypertuning parameters (rf = mtry)
  - Snakemake/python
  - Crossref gathering of DOIs (146K)
  - 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 for new data
- Supplemental Material file list (where applicable)
- Acknowledgments
- References