Data Management Plan: Improving the Generalisation Performance of the Weisfeiler-Lehman Subtree Kernel

Antoine Origer

April 28, 2024

General Information

- **Project Title:** Improving the generalisation performance of the Weisfeiler-Lehman subtree kernel
- **Researcher:** Antoine Origer - ORCID ID 0009-0006-7102-8571
- **DMP Version:** 1.0
- **Funding:** None

1 Data Description

1.1 Data Re-use

- Data sets used include AIDS, PROTEINS, MUTAG, Mutagenicity, IMDB-BINARY, SYNTHETICnew, COIL-DEL, NCI1.
- All data sets are well-known and publicly available, primarily used in chemoinformatics and machine learning research.
- Data provenance is documented in README file with references to sources as provided by the tutor.

1.2 Data Types and Formats

- Used data consists of graph representations in text files (TXT format).
- Output data includes accuracy measures stored in Excel (XLSX format).
- Data volumes range from 2MB to 600MB per dataset with an output file under 1MB.
- All the programming is done in using Python (.py files). These are combine just under 1 MB in size.

2 Documentation and Data Quality

2.1 Metadata and Documentation

- Metadata includes hyperparameters used in SVM computations (iterations, thresholds).
- Documentation will be provided in README file detailing the methodology and software requirements.

2.2 Data Quality Control Measures

- Quality control is maintained by consistent use of data and code validation through manual checks and computational reproducibility.
- Used Datasets are well known and were downloaded and added into to project to keep them consistent. They are not being changed during all the research, only accessed.
- To ensure output data quality each hyperparameter combination runs 10 times and the mean value of these results is noted down in the excel file.

3 Storage and Backup During the Research Process

- Data stored on a local device and backed up weekly to Google Drive and GitHub.
- No sensitive data involved; standard security measures in place with data stored on secure platforms.

4 Legal and Ethical Requirements, Codes of Conduct

- No personal data processed.
- Intellectual property: Data created and used are open and free for reuse with proper citation following CC-BY
- Code written and shared will also be open and free for reuse following MIT

5 Data Sharing and Long-Term Preservation

- Data and results in their entirety will be shared via GitHub, no part of this reaserch project will be kept secret or destroyed.
- As for long term preservation the data remains available on GitHub indefinitely. In case Github no longer exists, a locally stored copy as well as a backup on a Drive exist. A reupload would be done in that case.

6 Access and Reuse Tools

- Python scripts and Excel are required to access and use the data and results.
- A DOI will be assigned by the university (TU Vienna) for the final project dataset and code repository.

7 Data Management Responsibilities and Resources

- Antoine Origer is responsible for all aspects of data management.
- Resources required include personal computing resources, GitHub, and Google Drive accounts. No additional financial resources are necessary or available.