README.md

Implementation of privacy-oriented metrics and procedures to increase security after a data minimization process.

The added features aim to define useful metrics (Anonymity set size and l-diversity) and take action (injecting synthetic data) to increase those security related metrics after a data generalization/minimization process as the one suggested in Goldsteen's paper.

The extension is contained in the DPTAssignment folder.

The extension includes three main components:

- Anonymity set size calculator
- Extensive l-diversity calculator
- Synthetic data injector

The three components are available in the minimizerSupporter.py file

Anonymity set size calculator:

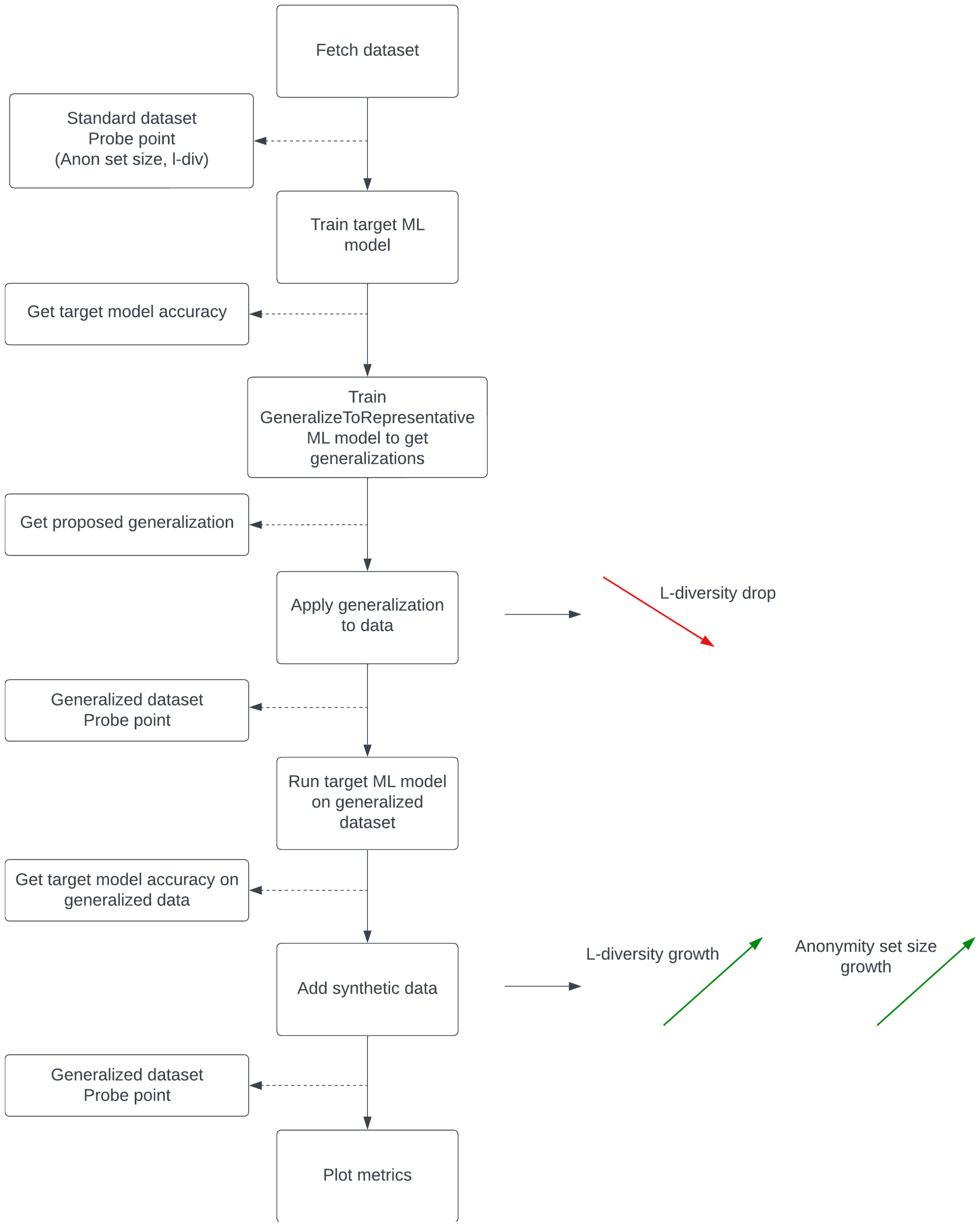
It is important to define not only the anonymity set size of a single set but also the whole dataset. To accomplish the task the get_anon_set_size function returns three values, the minimum anonymity set size, the average and the maximum. These values are calculated by calculating the anonymity set size of every single set in the dataset, a set is a group of entries that share the same quasi-identifiers.

Extensive l-diversity calculator:

L-diversity defines the variety of the sensitive data stored in a given set, a set is defined as a group of entries with the same quasi-identifier. A set is considered l-diverse when each entry has unique sensitive data among the set. The l-diversity metric depicts a rather black or white picture, therefore I tried to develop a more nuanced l-diversity metric, by calculating the l-diversity of each set in the dataset and then get average, minimum and maximum results. The l-diversity is expressed for each set in percentage as: *100*Number of distinct sensitive values/Number of entries in the set*

Synthetic data injector:

The synthetic data injector adds data to the available dataset mimicking the original dataset statistical properties. It is important to insert realistic data since in case of a leak it is important not to give any hint on what data is real and what is just dummy data. Being well aware of the anonymity-set-size and l-diversity of the dataset, the data injector has been designed to increase those values (more details in the documentation). The injector plays a key role in avoiding the drop of l-diversity after the data minimization. It's important to note that we don't care if the synthetic data doesn't belong to the range defined by the minimizer, since the ML target model will not be trained on those data and those data will not be used since they belong to no one.


Fetch dataset
Standard dataset
Probe point
(Anon set size, l-div)
Train target ML model
Get target model accuracy
Train GeneralizeToRepresentative ML model to get generalizations
Get proposed generalization
Apply generalization to data
L-diversity drop
Generalized dataset
Probe point
Run target ML model on generalized dataset
Get target model accuracy on generalized data
Add synthetic data
L-diversity growth
Anonymity set size growth
Generalized dataset
Probe point
Plot metrics

Demos:

The metrics are just used as a reference in these examples, in a real world scenario they could be used to define how many new entries to inject or to help to tune the generalized data prediction accuracy. Both demos can be run from CLI with:

```
cd DPTAssignment
pip3 install numpy==1.24.2 pandas==1.1.05 scipy==1.10.1
pip3 install scikit-learn>=0.22.2,<=1.1.3 torch>=1.8.0 tqdm>=4.64.1
pip3 install matplotlib>=3.7.0 adversarial-robustness-toolbox>=1.11.0
python3 diabetes.py
python3 sleepQuality.py
```

Do not use the IDE run since I've noticed some issues in rendering the matplotlib graphs. Sleep quality uses a non-sklearn dataset that has been modified before being run, the original dataset is sleep_health_and_lifestyle_dataset.csv and it's been transformed in the dataset used in the demo (processed_dataset.csv) by cleanDataSet.py (no need to clean it again).

Citation:

Abigail Goldsteen, Ola Saadi, Ron Shmelkin, Shlomit Shachor, Natalia Razinkov, "AI privacy toolkit", SoftwareX, Volume 22, 2023, 101352, ISSN 2352-7110, https://doi.org/10.1016/j.softx.2023.101352.