# MLCQ\_metrics\_blob.xlsx

All potentially useful columns:

|  |  |
| --- | --- |
| **parts** | whether the example belongs to the training or test set |
| **label** | sample label (critical, major, minor, or none) |
| **sample\_id** | unique sample id |
| **type** | class (value 1), innerclass (value 2), or interface (value 3) |
| **from\_project** | whether the CK tool metrics were extracted from the project (value TRUE) or a single file (value FALSE). If extracted from a single file (80 examples), values for columns **type**, **CBO**, **DIT**, **RFC**, **TCC,** **LCC**, **NOSI,**and **uniqueWordsQty** may be approximated rather than the exact value. This column is potentially helpful for outlier analysis. |
| **TCC** and **LCC** | The value of this metric should be between 0 and 1. Value -1 denotes that the metric cannot be calculated (classes with 0 or 1 method). Semantically, value -1 is closer to 1 (perfect cohesion). |
| **RM\_** prefix | Denotes that the metric was extracted using the Repository Miner tool. All these metrics are calculated using the whole project rather than a single file (and thus should be exact). |

# MLCQ\_god\_class\_data\_splitted.csv

|  |  |
| --- | --- |
| **files** | label (critical, major, minor, or none)/sample\_id |
| **parts** | whether the example belongs to the training or test set |

# train\_X, test\_X, train\_y, test\_y

**train\_X** and **test\_X** are the attributes (class metrics). Compared to **MLCQ\_metrics\_blob.xlsx**, the following colums are discarded:

* parts
* label
* sample\_id
* from\_project

All colums are normalized except:

* tcc and lcc (typical value is between 0 and 1; -1 denotes a special case where these metrics cannot be calculated )
* type (a categorical variable denoting whether the examined class is a class (1), an innerclass (2), or an interface(3)).

**train\_y** and **test\_y** are the corresponding labels (row *k* in train\_y corresponds to row *k* in train\_X)

# Otvorena pitanja

* Da li iz skupa podataka eliminisati sve klase koje imaju TCC i LCC -1 (trivijalni slučajevi koji nisu GC)?
* [@Nikola Luburić](https://cleancadet.slack.com/team/URU5LMCE4) zabrinjava me što sam analizom iz pojedinačnih fajlova uglavnom dobijala TCC i LCC negde između 0 i 1 (što mi je očekivano), dok su sada te vrednosti uglavnom ekstremi 0 ili 1. Bilo bi dobro da možda baciš pogled na primer-dva samnom da li su korektni.