# MLCQ\_metrics\_LM\_multicat.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 |
| **constructor** | method is a constructor (1) or not (0) |
| **hasJavaDoc** | method is accompanied by JavaDoc (1) or not (0) |
| **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, values for columns **cbo,**and **rfc** may be approximated rather than the exact value. This column is potentially helpful for outlier analysis. |

# MLCQ\_metrics\_LM\_binary.xlsx

Same as *MLCQ\_metrics\_LM\_multicat.xlsx*, but the label is binary: smell (TRUE) or non-smell (FALSE)

# MLCQ\_long\_method\_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\_LM\_binary.xlsx**, the following colums are discarded:

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

All colums are normalized except: **constructor** and **hasJavaDoc**.

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