2.1 Machine learning and datasets

In general, ML techniques can be classified according to the amount and shape of the provided data (i.e., the dataset).

In *supervised* learning, the dataset includes the labels that the ML algorithm needs to learn. A particular type of learning task is *regression*, in which the system learns to predict a numerical variable (e.g., a person height). Another type of task is *classification*, in which the system learns to predict a categorical variable (e.g., hair color). The existence of a dataset, such as the one proposed in this work, will enable classification applications associated with the management of large model repositories [19], like attaching tags automatically to models to help the user's navigation and the automatic detection of anomalous models to discard them, among others. Moreover, the labels of a dataset are not only useful for supervised tasks, but they also play a role in stratified sampling to make sure that, when the data is split, the models within each split preserve the percentage of sample for each class.

In contrast, in *unsupervised* learning, the dataset does not need to be labelled since the system tries to identify patterns by itself. A typical task is clustering, in which the system identifies groups of similar examples according to some criteria [11]. Other unsupervised tasks include learning modelling patterns which arises in a dataset, for instance, with the aim of creating smart modelling environments, including recommenders for model editors [31,33] and supporting the interaction with bots [42,43]. It is important to note that a labelled dataset is useful for testing clustering techniques since there are quality metrics that require the ground truth to be computed (e.g., Rand Index, NMI, AMI, etc). On the other hand, reinforcement learning approaches have been used in the modelling domain to apply automatic repairs [26].

Regarding clustering techniques, many algorithms have been proposed. Among the most popular ones, there are K-Means, hierarchical clustering and DBSCAN. All of them requires a distance measure and the first two admit the number of clusters as hyperparameter. DBSCAN does not require setting the number of clusters upfront, but it requires other parameters. We use these three algorithms as a baseline to compare our labelling method.

Altogether, the existence of high-quality datasets is a pre-requisite for applying some of the ML techniques mentioned above to modelling. Moreover, the availability of curated datasets will enable the development of model