### Determining train, validation and test set size

In machine learning, classification is based on the statistical analysis of features occurring in the training set, and on the probability inference that samples belong to a class depending on their features. The model will predict incorrectly if the distribution of the features varies substantially between train set and test set. Choosing the correct ratio for training, validation and test set is therefore fundamental to make the latter sets statistically representative of the whole set [1].

According to the central limit theorem, for samples with independent features – as it is the case for this study - the larger the sample size is, the more similar its mean and standard deviation will be to the whole population’s mean and standard deviation [2]. It can be therefore inferred that, in supervised learning, the training set should be large enough for the model to learn effectively from its features, while the validation and test sets should be large enough to be statistically representative of the whole set. Also, according to the law of large numbers, the frequency of occurrence of a feature will tend to be equal to its probability as the number of trials increases [3]. It can be therefore inferred that smaller samples will tend to present uneven distribution of their features, compared to larger ones.

After an analysis of these factors it was decided that, for a small sample such as the one proposed in this study, a higher ratio of validation set and test set to train set, in the range traditionally proposed for supervised learning [4], was necessary for the model to predict accurately. A ratio of 64% training set to 16% validation set and 20% test set was therefore chosen.

[1] Bartlett, J. E., II; Kotrlik, J. W.; Higgins, C. (2001). [Organizational research: Determining appropriate sample size for survey research](http://www.osra.org/itlpj/bartlettkotrlikhiggins.pdf). Information Technology, Learning, and Performance Journal. **19** (1): 43–50. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.486.8295&rep=rep1&type=pdf>

[2] Rouaud, Mathieu (2013). [Probability, Statistics and Estimation](http://www.incertitudes.fr/book.pdf) (PDF). p.10.

[3] Law of large numbers. *Encyclopedia of Mathematics.* URL: <http://www.encyclopediaofmath.org/index.php?title=Law_of_large_numbers&oldid=26552>

[4] Moindrot, O.; Genthial, G. Splitting into train, dev and test sets. Jan 24th, 2018. URL: [https://cs230-stanford.github.io/train-dev-test-split.html#theory-how-to-choose-the-train-train-dev-dev-and-test-sets](https://cs230-stanford.github.io/train-dev-test-split.html" \l "theory-how-to-choose-the-train-train-dev-dev-and-test-sets)