Module 1: Training, Testing and validation dataset

The fundamental purpose for splitting the dataset is to assess how effective will the trained model be in generalizing to new data. This split can be achieved by using **train_test_split** function of scikit-learn.

Training Set

This is the actual dataset from which a model trains .i.e. the model sees and learns from this data to predict the outcome or to make the right decisions. Most of the training data is collected from several resources and then preprocessed and organized to provide proper performance of the model. Type of training data hugely determines the ability of the model to generalize .i.e. the better the quality and diversity of training data, the better will be the performance of the model. This data is more than 60% of the total data available for the project.

Testing Set

This dataset is independent of the training set but has a somewhat similar type of probability distribution of classes and is used as a benchmark to evaluate the model, used only after the training of the model is complete. Testing set is usually a properly organized dataset having all kinds of data for scenarios that the model would probably be facing when used in the real world. Often the validation and testing set combined is used as a testing set which is not considered a good practice. If the accuracy of the model on training data is greater than that on testing data then the model is said to have overfitting. This data is approximately 20-25% of the total data available for the project.

Validation Set

The validation set is used to fine-tune the hyperparameters of the model and is considered a part of the training of the model. The model only sees this data for evaluation but does not learn from this data, providing an objective unbiased evaluation of the model. Validation

dataset can be utilized for regression as well by interrupting training of model when loss of validation dataset becomes greater than loss of training dataset .i.e. reducing bias and variance. This data is approximately 10-15% of the total data available for the project but this can change depending upon the number of hyperparameters .i.e. if model has quite many hyperparameters then using large validation set will give better results. Now, whenever the accuracy of model on validation data is greater than that on training data then the model is said to have generalized well.