Glossary+

+These features are part of StataNow.

bagging. A model agnostic procedure that generates perturbation of the dataset by random and independent drawings (Breiman 1996).

base learner. A learner whose error rate is only slightly better than random guessing.

beeswarm plot. A type of data visualization used to display the individual data points as dots such that the points do not overlap, resulting in a "swarm" of points. This type of plot is used by `h2omlgraph shapsummary`.

bias-variance tradeoff. This controls the tension between learning and generalization. The tradeoff concerns how to lower generalization error by reducing the bias and variance of the machine learning methods. For details, see *Fundamentals of machine learning* in [H2OML] **Intro**.

black box method. A machine learning method that is difficult to interpret by design. For example, linear models and decision trees belong to the class of interpretable models, but ensemble methods, and neural networks are considered black box methods.

boosting. A model agnostic deterministic procedure that generates perturbation of the dataset by sequentially reweighting it (Freund and Schapire 1997).

categorical encoding. A process of transforming categorical predictors into numerical representations so that they can be used in machine learning models. For details, see [H2OML] ***encode_option***.

classification. A type of supervised machine learning task where the goal is to predict the category or class of a response based on predictors.

classifier. A machine learning method that is designed for classification. When the response variable in the supervised learning method is categorical, then the method implements classification.

DOT language. A plain-text graph description language used in the Graphviz software.

ensemble method. A mechanism that forms a smart committee of incompetent but carefully selected members to solve a machine learning problem. For details, see *Ensemble methods* in [H2OML] **Intro**.

explainable method. A technique used in machine learning that enables explaining the predictions of a model.

feature. Same as *predictor*.

fitting. A process of training a model on data by adjusting its hyperparameters to improve performance.

generalization. A process where the model not only performs well on the training data but also generalizes to new (testing) data.

generalization error. A quantitative measure of how well a machine learning model can predict outcomes for new (testing) data. Generalization error is the expected error on new data (the testing set).

grid search. A process of evaluating different hyperparameter configurations in the hyperparameter space to find the best configuration that improves performance of a model.

hyperparameter. A parameter whose value is adjusted to control and improve the training process.

hyperparameter space. Possible values and ranges of the hyperparameters.

hyperparameter tuning. A process where the hyperparameters of a model are optimized to improve performance.

impurity measure. A measure to quantify the goodness of fit of a split in the regression or classification trees.

k-fold cross-validation. A process of splitting a dataset into k parts. For each of k iterations, it uses one part for validation and the remaining $k-1$ parts as a training subset for model fitting.

learn. In the machine learning context, learning refers to the process when a model uses data to adjust its parameters to increase prediction accuracy.

learner. A machine learning method such as random forest and gradient boosting machine used for learning.

majority-vote rule. A classification rule that returns a class that is the most commonly occurring one among the predictors. Majority-vote rule is used in bagging and random forest to predict the class.

manifold hypothesis. The manifold hypothesis states that the observed high-dimensional data lie on a low-dimensional manifold.

metric scoring. A process of evaluating the performance of a machine learning algorithm by using a specified metric.

model agnostic. A methodology whose implementation does not directly require a particular model.

model selection. The process of building an optimal model by exploring a range of possible hyperparameters and selecting the ones that result in the best-performing model.

one-hot encoding. A process that decomposes categories of a categorical predictor into binary variables.

optimism bias. Bias that occurs when a sufficiently complex machine learning model memorizes the patterns in the training data.

out-of-bag observations. Observations that are not used to grow the tree after bootstrap.

overfitting. A process of fitting a machine learning method too well on the training data so the method fails to generalize to testing data. For details, see *Fundamentals of machine learning* in [H2OML] **Intro**.

performance metric. A quantitative measure used to evaluate the performance of a model.

pessimistic bias. Bias that occurs when the validation set is small and the machine learning model fails to reach its full capacity.

predictive modeling. A process of developing a model that generates accurate predictions.

predictor importance. The degree to which a predictor influences the model's predictions.

predictors. The inputs for a machine learning model. In classical statistics, these may be referred to as independent variables, covariates, x variables, or predictors. In machine learning literature, they are also referred to as features.

proportion predictor importance. A type of predictor importance calculated by dividing the importance of each predictor by the total sum of the importance of all predictors.

pruning. A process to optimize hyperparameters for regression and classification trees (Breiman et al. 1984).

response. The outputs for a machine learning model. In classical statistics, these may be referred to as dependent variables, y variables, or outcomes. In machine learning literature, they are also referred to as targets.

root node. A node in the graph or tree that does not have parents. For details, see *Decision trees* in [H2OML] **Intro**.

scaled predictor importance. A type of predictor importance calculated by dividing the importance of each predictor by the largest importance score of the predictors.

stopping criteria. In growing decision trees, the stopping criteria determine what will be used to halt the additional splitting of the node. Examples of stopping criteria are the depth of the tree, minimum number of observations in each tree, etc.

stump. A decision tree with depth equal to one. Stumps are weak learners.

supervised learning. A type of machine learning in which a method is trained on data where there is an associated response for each observation. Linear regression, random forest, and gradient boosting machine are examples of supervised learning.

surrogate model. An explainable model that approximates the prediction of the machine learning model.

target. See *response*.

terminal node. A node in the graph that does not have children. For details, see *Decision trees* in [H2OML] **Intro**.

testing set. New data used to estimate the generalization error of the machine learning method.

three-way holdout. A process of splitting the dataset into three parts: training, validation, and testing datasets. This method is used to evaluate model performance.

training set. Data used to train a machine learning method.

tuning budget. Time or computational resources allocated for hyperparameter tuning.

two-way holdout. A process of splitting the dataset into two parts: training and testing datasets. This method is used to evaluate model performance.

underfitting. Underfitting occurs when a machine learning model is not complex enough to capture the hidden patterns of the data, resulting in poor performance on the training and testing data.

unsupervised learning. A type of machine learning where there is no response variable.

validation dataset. A subset of data separated during the training process of a machine learning model and used to evaluate the model's performance during hyperparameter tuning.

variable importance. See *predictor importance*.

weak learner. See *base learner*.

References

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24: 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Boca Raton, FL: Chapman and Hall/CRC.

Freund, Y., and R. E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55: 119–139. https://doi.org/10.1006/jcss.1997.1504.