Published in final edited form as:
Behav Ther. 2020 September ; 51(5): 675–687. doi:10.1016/j.beth.2020.05.002.

Supervised machine learning: A brief primer

Tammy Jiang, MPH[1], **Jaimie L. Gradus, DMSc, DSc, MPH**[1,2], **Anthony J. Rosellini, PhD**[3]

[1]Department of Epidemiology, Boston University School of Public Health, Boston, USA

[2]Department of Psychiatry, Boston University School of Medicine, Boston, USA

[3]Center for Anxiety and Related Disorders, Department of Psychological and Brain Sciences, Boston University

Abstract

Machine learning is increasingly used in mental health research and has the potential to advance our understanding of how to characterize, predict, and treat mental disorders and associated adverse health outcomes (e.g., suicidal behavior). Machine learning offers new tools to overcome challenges for which traditional statistical methods are not well-suited. This manuscript provides an overview of machine learning with a specific focus on supervised learning (i.e., methods that are designed to predict or classify an outcome of interest). Several common supervised learning methods are described, along with applied examples from the published literature. We also provide an overview of supervised learning model building, validation, and performance evaluation. Finally, challenges in creating robust and generalizable machine learning algorithms are discussed.

Keywords

machine learning; supervised learning; ensemble methods

Introduction

Machine learning is a branch of computer science that aims to learn patterns from data to improve performance at various tasks (e.g., prediction; Mitchell, 1997). In applied healthcare research, machine learning is typically used to describe automatized, highly flexible, and computationally intense approaches to identifying patterns in complex data structures (e.g., nonlinear associations, interactions, underlying dimensions or subgroups). This definition is often used in contrast to “traditional” parametric methods that involve numerous statistical assumptions, and require a priori specification of dimensions or

Correspondence. Correspondence concerning this article should be addressed to Anthony J. Rosellini, Center for Anxiety and Related Disorders, 900 Commonwealth Avenue (East), 2nd Floor, Boston, MA, 02215; phone: 617-353-9610, fax: 617-353-9609, ajrosell@bu.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Declaration of interests. None.

subgroups of interest, the functional form of the relationship between predictors and the outcome, and interactions among predictors.

In clinical psychology and psychiatry, researchers are conducting increasingly broad and multi-modal assessments of mental disorder phenotypes and associated risk and prognostic factors (e.g., self-report measures, physiological factors, imaging data). It is now common for datasets to contain thousands of measurements, often assessed repeatedly over time. Machine learning can be applied to such complex data structures to aid the conceptualization of mental disorders (Galatzer-Levy et al., 2018), detect and predict of the risk and trajectory of symptoms (Shatte et al., 2019), and study treatment outcome and differential treatment response (Kessler et al., 2019). The goal of this paper is to provide a primer in supervised machine learning (i.e., machine learning for prediction) including commonly used terminology, algorithms, and modeling building, validation, and evaluation procedures. Prior to discussing supervised learning, however, it is first necessary to understand its distinction from unsupervised learning, and the types of research tasks for which each may be used.

When Can You Use Machine Learning?

The first step to determine if (and which) machine learning method(s) to use is specifying one’s research question. Using Hernán and colleagues’ framework, there are three major data science research tasks: description, prediction, and causal inference (Hernán et al., 2019). Machine learning can be used for each of these three tasks, but traditional statistical methods may prove to be sufficient and more appropriate depending on the specific research question.

Description.

Description tasks involve using data to provide a quantitative summary of certain variables. An example of a descriptive question is: what is the proportion of individuals with posttraumatic stress disorder (PTSD) in a healthcare database? The answer to this question might be obtained straightforwardly using basic statistics (e.g., frequency of a diagnostic code), and machine learning might not be necessary. However, if there is a lot of misclassification of PTSD in the healthcare database, machine learning could be used to attempt to construct a PTSD profile that combines multiple sources of data such as other diagnostic codes (e.g., depression, generalized anxiety disorder), medications (e.g., prazosin), and free-text from clinician notes (e.g., description of avoidance behaviors and nightmares).

Unsupervised machine learning methods are particularly useful in description tasks because they aim to find relationships in a data structure without having a measured outcome. This category of machine learning is referred to as unsupervised because it lacks a response variable that can supervise the analysis (James et al., 2013). The goal of unsupervised learning is to identify underlying dimensions, components, clusters, or trajectories within a data structure. Several approaches commonly used in mental health classification and psychometric research fall under the umbrella of unsupervised learning including principal components analysis, factor analysis, and mixture modeling (one reason for our focus on supervised learning below). Accordingly, unsupervised learning can be used to identify underlying or unobserved mental health dimensions and trajectories, and to determine how to best categorize dimensions into subtypes (e.g., diagnostic groups). Identification of data-driven dimensions and subgroups may lead to the formulation of novel hypotheses surrounding a new symptom, phenotype, or diagnosis. When prospective data are available, unsupervised learning methods (e.g., growth mixture modeling; Jo et al., 2017) also can be used to identify heterogeneity in symptom or phenotype progression over time, which may improve our understanding of the pathogenesis, chronicity, and remission of mental disorders (i.e., using unsupervised learning to define data-driven mental health outcomes). Accordingly, unsupervised learning is relevant for studying mental disorders within conceptual frameworks such as the Research Domain Criteria (RDoC; Cuthbert & Insel, 2013) or Hierarchical Taxonomy of Psychopathology (Kotov et al., 2017). Along these lines, unsupervised learning has been used to empirically derive classes of psychiatric symptoms that are transdiagnostic (Kircanski et al., 2017; Rosellini & Brown, 2014), and incorporate multiple levels of data (e.g., genes, circuits, physiology, behaviors; Wu et al., 2017), to develop alternative approaches to classification.

Prediction.

The second major data science task is prediction, which can involve a range of specific goals. For example, one conceptually driven prediction goal is to test the hypothesis that personality traits and trauma exposure increase risk for the subsequent development of major depressive disorder (i.e., testing the extent to which vulnerability and stress predicts psychopathology). Traditional statistical methods (e.g., logistic regression) can be used to test this hypothesis by obtaining interpretable estimates of the nature and statistical significance of associations between predictors and the outcome (e.g., odds ratio).

Another prediction task could involve using hundreds of variables to develop a risk score that accurately identifies who is most likely to experience major depressive disorder within three months of trauma exposure. Machine learning may be preferable to traditional statistical methods if the goal is prediction optimization in large/complex data structures because such methods have fewer and less restrictive statistical assumptions compared to traditional parametric methods (e.g., linear relationships, absence of multicollinearity). A related prediction goal is identifying the variables that most strongly contribute to prediction accuracy (e.g., identifying biological markers of a behavioral phenotype). Accordingly, some machine learning methods provide variable importance metrics (e.g., recursive partitioning, described below) that may be useful in generating novel hypotheses surrounding risk/prognostic factors and their interactions (cf. testing a hypothesis focused on a specific piece of a conceptual model).

Supervised learning is used to describe prediction tasks because the goal is to forecast/classify a specific outcome of interest (e.g., presence or absence of a mental disorder). Supervised learning has been applied to large data structures including demographic, clinical, and social predictors in order to develop risk scores predicting the onset and trajectory of a range of mental disorders (e.g., anxiety, depression, and trauma-related disorders) and suicidal behavior (Galatzer-Levy, 2015; Gradus et al., 2020; Kessler et al., 2014; Rosellini et al., 2018; Rosellini et al., 2020). Supervised learning also has been used to develop prognostic scores predicting the likelihood of responding to a particular mental health intervention (e.g., cognitive behavioral therapy for depression; obsessive-compulsive disorder program, Hasanpour et al., 2018; Webb et al., 2020). Risk and prognostic scores may be particularly useful for identifying individuals in need of preventive interventions or tailored/intensive treatment (based on high or low predicted risk/prognosis). Although most studies develop risk and prognostic scores using predictors assessed at a single time point (e.g., pre-treatment), there are a variety of ways to integrate repeated measures (time-varying predictors) into supervised learning algorithms (e.g., operationalizing predictors based on timing prior to outcome; utilizing a person-time data structure).

Causal Inference.

Causal inference involves estimating effects by comparing the outcomes among the exposed with the counterfactual outcomes if they had instead not been exposed. An example of this is estimating the suicide rate that would have been observed if all individuals in a study population had received a psychosocial suicide prevention intervention compared with the suicide rate if they had not received the intervention (i.e., determining the effect of treatment exposure on a specific outcome). To examine this type of causal question, clinical psychologists typically use a randomized controlled trial (RCT) design and traditional statistical methods.

However, causal inference can be applied to observational (i.e., non-randomized) data in a way that mimics a hypothetical randomized trial (Hernán et al., 2016). Causal inference methods have existed for decades (Rubin, 1974), but the integration of supervised learning methods is more recent. For example, classification and regression trees and random forests have been used to generate propensity score models used to adjust treatment effect estimates based on individual differences in the (non-random) likelihood of receiving a specific intervention (Lee et al., 2010). One increasingly popular approach to causal inference is targeted maximum likelihood estimation (TMLE; van der Laan & Rubin, 2006). TMLE is an attractive method because it has a double robustness property, meaning that the average treatment effect will remain unbiased as long as either the exposure regression *or* outcome regression is correctly estimated. Simulation studies have found that implementation of TMLE using supervised learning (i.e., ensemble methods, described below) helps protect against bias in treatment effect estimates (Schuler & Rose, 2017). It also is noteworthy that methods exist to use TMLE and supervised learning to develop optimized treatment rules (from RCT or observational data) for matching patients to the intervention most likely to provide benefit (Kessler et al., 2019).

In summary, machine learning may or may not be necessary depending on the goal of the study (e.g., interpretability versus prediction accuracy). The first question to consider is: what is the nature of the research question (e.g., descriptive, predictive, causal?). The follow up question is: what is the most appropriate statistical tool to answer this question with (e.g., traditional descriptive statistics or parametric regression versus more flexible unsupervised or supervised machine learning)?[1]

Overview of Terminology and Supervised Machine Learning for Prediction Tasks

As described above, there are similarities in the broad tasks/goals of traditional statistical approaches and supervised machine learning. At the same time, this overlap is often missed because the machine learning literature uses different terminology (see Table 1). For example, rather than discussing *predictors* or *covariates* for an *outcome* or *dependent variable*, the machine learning literature refers to *features* that may be used to classify *outputs* or *targets*. Further, supervised learning to predict a categorical outcome is referred to as *classification* in the machine learning literature (cf. logistic regression), while prediction of a continuous outcome is referred to as *regression* (cf. linear regression). Below we summarize some of the most commonly used supervised learning methods and associated terminology. See also Table 2 for a brief description of each method, their strengths and weakness, and corresponding R packages[2].

Recursive partitioning.

One of the most popular approaches to supervised learning are decision tree methods. Decision tree methods are non-parametric and thus highly flexible. For example, classification and regression trees (CART) involve dividing all possible values for all predictors into distinct and non-overlapping regions (i.e., the predictor space). This process, which automatizes detection of main effects (including non-linear associations) and interactions, is referred to as recursive partitioning. A tree is grown based on these non-overlapping regions, with those at the bottom of the tree referred to as *terminal nodes* (distinguishable subgroups). For categorical outcomes, observations are predicted to belong to the most commonly occurring class/category in a node (James et al., 2013). For continuous outcomes, every observation that belongs to a particular node is predicted to have the mean of the response values within that node. To grow a tree, splits can be made based on a range of criteria. Using CART, for instance, splits are based on minimizing classification error (for categorical outcomes) or residual sum of squares (for continuous outcomes). In comparison, conditional inference trees are grown based on the strength of univariate associations (see Strobl et al., 2007). Individual decision trees have high visual interpretability and are useful for identifying nonlinear associations and interactions between variables (see Strobl, Malley, & Tutz, 2009 for a review of recursive partitioning).

Random forests are another recursive partitioning method that involves prediction based on a collection of individual decision trees. Random forests are referred to as an *ensemble method* -- multiple models (e.g., trees) are combined into a single random forest. To do this, the method creates bootstrapped copies of the original data and a single tree is estimated in each bootstrap. Random forests have a lower risk of overfitting than a single tree because

[1] Although we broadly distinguish between supervised and unsupervised machine learning methods, *semi-supervised machine learning* also exists (i.e., learning based on a combination of labeled data/known outcomes and unlabeled/unknown underlying dimensions or subgroups). Semi-supervised methods are not reviewed here as there are fewer applied studies using mental health data.

[2] Most supervised learning approaches have a number of sub-configurations that can influence model performance and risk of overfit, referred to as *hyper parameters* or *tuning parameters* (e.g., classification tree depth; number of trees from which to build a random forest; lambda - the elastic net penalty term). Tuning parameters must be selected and adjusted (and ideally cross-validated) with careful consideration of risk of model overfit.

multiple trees are averaged together to provide more accurate and stable predictions (James et al., 2013). Overfitting arises when a model corresponds too closely to a particular dataset and fails to accurately predict events in new samples (Hawkins, 2004). A disadvantage of random forests is that a collection of hundreds or thousands of trees is not as visually interpretable as a single tree. However, random forests compute measures of variable importance, which provide information on the extent to which each predictor improves or worsens model accuracy (e.g., mean decrease in accuracy) or discrimination ability (e.g., mean decrease in Gini). Of note, alternate parameterizations of CART and random forests algorithms (and variable importance metrics) should be used for data structures involving a combination of binary, categorical, and continuous variables (e.g., conditional inference trees; Hothorn et al., 2006; Strobl et al., 2007).

There are many recent applications of recursive partitioning in psychology and psychiatry (Gradus et al., 2020; Walsh et al., 2017). For example, Gradus and colleagues (2020) used classification trees and random forests to predict death by suicide from Danish population-based registry data. The classification trees revealed specific combinations of risk factors associated with high risk of suicide (e.g., not being prescribed antidepressants, antipsychotics, or anxiolytics, but having a prior suicide attempt and having lower income). The random forests identified the most important variables for accurately predicting suicide deaths, which included antipsychotic prescriptions, alcohol-related disorders, and schizophrenia (Gradus et al., 2020). In another application of random forests, Walsh, Ribeiro and Franklin (2017) used electronic medical records to predict suicide attempts among 5,167 adult patients with a claim code for self-injury. They found that prediction accuracy improved as the suicide attempt became more imminent (from 720 days to 7 days before the suicide attempt) and that predictor importance shifted over time. For example, some predictors were consistently important across time such as psychotic disorders, recurrent depression, and poisoning, whereas other predictors were important only several months or years before the suicide attempt (e.g., prescriptions for selective serotonin reuptake inhibitors, benzodiazepines, acetaminophen, and recent inpatient, outpatient, and emergency department visits, Walsh et al., 2017).

Support vector machines.

Support vector machines are intended for classification in which there are two classes (e.g., cases or controls) in a multidimensional space (i.e., across many variables). To differentiate between classes of individuals, support vector machines identify a *hyperplane*, which is a boundary that maximally separates classes (i.e., outcome categories). Support vector machines apply a data transformation that project the data into a higher dimensional space to find a separating decision surface. This process is referred to as a *kernel function*. Although support vector machines can result in highly accurate prediction using flexible non-linear kernels, they are also limited by being a *black box* approach in the sense that metrics are not provided for how predictors are combined to optimize the hyperplane.

In one of the first applied psychopathology forecasting studies, Galatzer-Levy and colleagues (2014) used support vector machines and longitudinal data to predict a chronic trajectory of PTSD symptoms among individuals admitted to an emergency department following the experience a traumatic event (Galatzer-Levy et al., 2014). This study compared the prediction accuracy of three models which used: 1) all available predictors (n = 68), 2) a subset of predictors selected using a predictor selection algorithm (n = 16), and 3) only acute stress disorder symptoms. The model with only 16 predictors (e.g., event/injury characteristics, use of tranquilizers, psychological symptoms, etc.) performed just as well as the model using all available predictors, which suggests that it is possible to accurately forecast PTSD using a small number of clinical assessment tools (Galatzer-Levy et al., 2014). In another application of support vector machines, Månsson and colleagues (2015) assessed neural predictors of long-term treatment outcome among patients with social anxiety disorder one year after completion of an internet-delivered cognitive behavioral therapy intervention. Support vector machine classification revealed that the initial blood oxygen level-dependent responses to self-referential criticism in the anterior cingulate cortex predicted long-term response with high accuracy (Månsson et al., 2015).

Regularization.

Regularization (or penalization) methods are an extension of conventional regression that involve shrinking coefficients to zero (least absolute shrinkage and selection operator (LASSO) regularization), toward zero (ridge regularization), or some combination of the two (elastic net), to optimize prediction accuracy while preventing model overfit (Hastie, et al., 2009). LASSO, ridge, and elastic net use regularization terms that penalize a model for increasing complexity (e.g., having more predictors than observations; large number of collinear predictors). This is beneficial because increasing model complexity may lead to increased risk of overfitting -- models may find trivial patterns (*noise*) that are unique to a specific dataset but are not generalizable to external datasets. LASSO is often used for variable selection and may be particularly applicable in settings with a large number of predictors (and small number of observations), and when the goal is to achieve a sparse model with only a subset of the variables predicting an outcome. At the same time, LASSO is limited in that it can arbitrarily select predictor coefficients to shrink to zero among highly correlated sets, which may not always be desirable (e.g., for well-established or conceptually important predictors). In comparison, ridge regression can result in non-parsimonious models that retain all predictors. Elastic net was designed to combine ridge and LASSO penalties, permitting the development of parsimonious models with greater stability and accuracy than ridge or LASSO (Zou & Hastie, 2005). Elastic net identifies the optimal penalty term using an internal cross validation procedure (see Model Building and Validation for a definition of cross validation).

One applied example is from Kessler and colleagues (2015), who used elastic net to identify a subset of predictors that yielded high model accuracy for predicting suicides in the 12 months after psychiatric hospitalization in U.S. Army soldiers, with the goal of targeting expanded post-hospital care. Elastic net was specifically used to address the problem of multicollinearity among the original set of hundreds of predictors. The best performing model included sociodemographic characteristics (e.g., male, late age of enlistment), criminal offenses (e.g., weapons possession, verbal violence), prior suicidality, aspects of prior psychiatric inpatient and outpatient treatment, and specific disorders diagnosed during the hospitalizations (Kessler et al., 2015). Regularization methods also have been used to predict trajectory of depressive symptoms (Chekroud et al., 2017). For example, Chekroud and colleagues (2017) used elastic net to identify 25 out of 164 variables that optimally predicted change in three empirically defined depressive symptom clusters over treatment: core emotional symptoms, sleep problems, and atypical symptoms (e.g., psychomotor agitation).

Comparing algorithms and stacking methods.

Although we review some of the most popular supervised machine learning algorithms used in the clinical psychology and psychiatry literature above, a growing number of other approaches exist (e.g., neural networks, gradient boosting, Friedman, 2002; Ripley & Hjort, 1995). Different algorithms may result in better or worse performance, and it is crucial to test different algorithms to compare performance. Most supervised learning methods can be applied to complex data structures including a combination of categorical and continuous predictors. However, some methods have been adapted in order to handle predictors that vary in their scale of measurement or number of categories (e.g., conditional inference trees is an adaption of CART). Further, some methods tend to perform better or worse with binary versus continuous predictors (see Kotsiantis et al., 2007). For these reasons, researchers interested in comparing different algorithms may need to operationalize predictors using different approaches (e.g., using continuous predictors as well as creating dummy variables based on which quintile a score falls in).

Methods also are available to *stack (ensemble*) several different supervised learning approaches into a single composite algorithm with optimized prediction. Super learning, for example, involves generating a consolidated algorithm from multiple supervised learning methods (van der Laan et al., 2007). Briefly, super learning is implemented four steps. First, a user-specified library of algorithms is specified (e.g., classification trees, random forests, support vector machines, elastic net, LASSO, and ridge regression). Second, each algorithm in the library is implemented using k-fold cross validation (see Model Building and Validation) in order to compute individual-level predicted values. Third, the outcome is regressed onto the predicted values from each algorithm in the library. This determines the best weighted combination of the individual algorithms. Fourth, each algorithm is fit on the full dataset and combined with the weights, creating a super learner prediction function.

In one recent application, Rosellini and colleagues (2020) used super learning to develop algorithms predicting the onset of internalizing disorders between Waves 1 and 2 of the National Epidemiological Survey on Alcohol and Related Conditions (n = 34,653). Risk scores were developed for five disorders (generalized anxiety, panic, social phobia, depression, mania) using over 200 predictors and a library of nine different supervised learning algorithms (Rosellini et al., 2020). As expected, the composite super learner ensemble resulted in better prediction than individual algorithms from which it was developed. Nevertheless, the difference in performance across algorithms was sometimes small, including when compared to logistic regression. At the same time, it is important to note that other studies have found substantially better performance of super learning relative to individual algorithms including logistic regression (Bergquist et al., 2017; Kessler et al., 2014).

Sample size considerations.

It is not possible to provide precise rules about sample size requirements for supervised learning. In general, prediction performance improves as sample size increases. The methods described in the section above have been implemented in samples as small as a few hundred observations (Askland et al., 2015; Poulin et al., 2014), as well as in samples with over 100,000 observations (Gradus et al., 2020; Ilgen et al., 2009; Kessler et al., 2017). Nevertheless, other algorithms (e.g., neural networks) may require especially large samples to develop accurate predictions from certain types of data (e.g., text or image data). In smaller samples, performance may be poor or very similar to traditional regression methods. Simulation can be used to estimate how prediction error (e.g., mean squared error) may vary at different sample sizes. Along these lines, the machine learning literature sometimes generates *learning curves* to plot how model performance would improve as the number of observations increases (Figueroa et al., 2012).

Model Building and Validation

Predictor selection.

Mental health is shaped by the complex interplay between genetics, physiology, health behaviors, and social and environmental factors. In rich datasets that capture these multiple levels of information, the number of measured variables per subject can substantially exceed the number of subjects. Researchers should be cognizant that preparing such complex data structures for supervised learning (e.g., creating thousands of predictors) can be very time consuming. Moreover, one might assume that supervised learning accuracy will improve as the number of variables used to fit a model increases. In reality, test error (i.e., poor performance in independent samples) tends to increase as the number of input variables increases, unless the variables are truly associated with the response (James et al., 2013). Overfitting is more likely to occur when the *dimensionality* of the dataset (i.e., number of predictors, intercorrelations) is excessively high relative to the number of observations. Indeed, a central issue surrounding supervised learning is the *bias-variance tradeoff*; fitting a model that is complex and flexible enough to accurately predict an outcome (i.e., minimizing bias), but not so complex and flexible that *noise* is used to predict the outcome (i.e., minimizing variance/model overfit).

To mitigate these biases, predictor selection or data reduction typically occurs prior to supervised learning. Reducing the number of variables reduces the dimensionality of the data, and several approaches exist (e.g., for a review see Heinze et al., 2018). For example, variables that are highly correlated may provide little independent information. As mentioned above, one approach to selecting individual predictors among highly correlated sets is to use to use LASSO regression and retain predictors with non-zero coefficients. Alternatively, if correlations among predictors are due to substantive, conceptual overlap, unsupervised machine learning methods first may be used to operationalize a smaller set of underlying dimensions, components, or classes (Huys et al., 2016).

Cross validation.

After data processing and reduction, supervised machine learning methods may be implemented[2]. Ideally, prediction algorithms are developed (i.e., fitted) in one sample and then evaluated (i.e., applied) in an independent sample. In the machine learning literature, this is referred to as having a *training set* versus *test set* of observations. However, researchers may not have immediate access to an external dataset for validation. When external observations are not available, internal validation methods typically are used to estimate test error. The test error is the average error of a method used to predict a response on a new observation (i.e., a data point not used in training/fitting).

One approach to internal validation is the simple method of randomly dividing a dataset into a training set and a test set (i.e., hold-out set), fitting the model on the training set, and then applying the model to the validation set. Although this *validation set* approach is conceptually simple and easy to implement, it has disadvantages including fluctuating test set error (depending on which observations are included in the training and validation sets), limited overall performance because of reduced sample size (i.e., some observations cannot be used for training, (James et al., 2013), and increased risk of false positive findings (Källberg et al., 2010). K-folds cross-validation is a refinement of the validation set approach that addresses its disadvantages. K-fold cross validation involves randomly dividing the set of observations into *k* groups (i.e., folds) of approximately equal size. The supervised learning method is fit on *k-1* folds, with the left-out folds treated as validation sets. This procedure is repeated *k* times, thus a mutually exclusive group of observations is treated as the validation set each time. This process results in *k* estimates of the test error, and the k-fold cross validation estimates are computed by averaging these estimates. Although k-fold cross validation is a common and preferred approach (James et al., 2013), leave-one-out cross-validation may be more appropriate for smaller datasets with no more than a few hundred observations (i.e., k-folds cross-validation where k is equal to the number of observations minus 1).

Performance evaluation.

For continuous outcomes, supervised learning performance is typically assessed based on mean squared error or R-squared (referred to as the *coefficient of determination* in the machine learning literature). However, when the outcome is binary, performance is more often evaluated using sensitivity, specificity, positive and negative predictive value, area under the receiver operating characteristic curve (AUC), and calibration. Sensitivity (*recall* in the machine learning literature) is the proportion of subjects predicted to be positive (have the outcome occur) among all those who are truly positive, while specificity is the proportion of subjects predicted to be negative among all those who are truly negative. Positive predictive value *precision* in the machine learning literature) is the proportion of subjects who are truly positive among all those who are predicted to be positive, and negative predictive value is the proportion of subjects who are truly negative among all those who are predicted to be negative. A simple 2×2 contingency table that displays the actual outcome x predicted outcome (*confusion matrix* in the machine learning literature) may be used to calculate these performance metrics. Researchers may wish to calculate these metrics using several different risk thresholds, particularly when it is more important to identify true cases than true non-cases, or vice versa (i.e., prioritizing sensitivity and positive predictive value versus specificity and negative predictive value; Kessler et al., 2015; Rosellini et al., 2020).

AUC, in comparison, reflects the overall ability of a test to distinguish between subjects with and without the outcome. AUC is also known as the concordance statistic or c-statistic, and is one of the most commonly reported measures of supervised learning performance because of its ease of interpretation. Values range from 0 to 1, with <0.50 indicating prediction no better than chance; 0.50–0.70 indicating poor prediction; 0.70–0.79 indicating acceptable prediction; 0.80–0.89 indicating excellent prediction; and >0.90 indicating outstanding prediction (Hosmer & Lemeshow, 2013). However, because it is a measure of overall performance, the utility of AUC is limited if sensitivity or positive predictive value (i.e., true positives) are more important than specificity or negative predictive value (e.g., when determining suicide risk; see also Wald & Bestwick, 2014). Finally, one underreported metric of model performance is calibration, or goodness of fit (Bouwmeester et al., 2012). Calibration refers to the extent to which a model correctly estimates the absolute risk. In other words, calibration is the level of agreement between the values predicted by the model and the observed values. Poor calibration may lead to underestimation or overestimation of the outcome of interest (Alba et al., 2017).

Challenges and Future Directions

A major challenge in developing useful machine learning algorithms is measurement error. Measurement error arises when information is not correctly captured in the study database due to faulty instruments, respondents provide inaccurate information due to poor recall or sensitive issues, or mistakes are made in coding data (Lash et al., 2009). Measurement error can be large for complex phenomena such as neural and psychological processes. In the machine learning literature, measurement error is referred to as *label noise* (Frénay & Verleysen, 2014). Label noise has been shown to lead to inaccurate predictor selection and rankings (Frénay et al., 2014; Gerlach & Stamey, 2007; Shanab et al., 2012) and to decrease prediction performance (Lachenbruch, 1966; Nettleton et al., 2010; Wilson & Martinez, 2000). Thus, it is important to use well-validated (e.g., gold-standard) measures whenever possible. A prediction can only be useful if it is based on valid measurements, and serious consideration should be given to the quality of the data used to build a model. Nevertheless, in the absence of (or in addition to) well-validated measures, many statistical methods exist for mitigating the impact of measurement error, including using latent variable measurement models (i.e., operationalizing a construct using several measures, Brown, 2015), quantitative bias analysis (Lash et al., 2009), multiple imputation-based corrections that treat measurement error as a missing data problem (Cole et al., 2006; Edwards et al., 2013), Bayesian methods (Hubbard et al., 2019), and more. Currently, measurement error adjustment methods are seldom applied to machine learning. A potential reason for this is the lack of validation studies that provide estimates of a misclassified variable's classification probabilities (e.g., sensitivity and specificity), which are often needed for measurement error correction methods. Measurement error could lead to machine learning algorithms not being transportable to different settings and populations.

A second and related major challenge is external validation. This challenge has considerable overlap with current issues surrounding replication and reproducibility in psychological science (Tackett et al., 2017). Most studies use a single dataset for model development and validation. Although internal cross-validation may reduce the risk of overfitting a model, it is better to examine (i.e., replicate) performance of a model in an external sample. At the same time, a challenge of validating an algorithm in an external dataset is heterogeneity in the way that variables are measured across studies/samples (both predictors and outcomes). Variables in one dataset may be measured differently in the external validation dataset. This may occur when variables are assessed using different instruments (e.g., Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, 5th edition vs. Composite International Diagnostic Interview), or categorized using different cutoff values. Measurement heterogeneity can lead to miscalibration and affects discrimination and prediction accuracy (Luijken et al., 2019). To improve our ability to reproduce and replicate results, it is important for researchers to transparently report all measurement instruments administered and how variables were operationalized, machine learning algorithms and tuning parameters tested, software used, and also provide open sharing of data and code. Such efforts are possible using study preregistration, open access/science portals, and detailed supplemental materials. Virtually all of the studies reviewed above included supplemental materials providing additional details surrounding the creation of study variables, how supervised learning methods were implemented (e.g., tuning parameters), and results (e.g., relative performance of different classifiers). Other transparency methods are increasingly being used (see current projects on https://osf.io/, e.g., Hsu et al., 2020). Although more common in the physical health prediction literature, researchers also should follow standard reporting guidelines such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement (Collins et al., 2015), which are currently being adapted for machine learning applications (Collins & Moons, 2019).

A third challenge is the computational intensity and efficiency of many machine learning methods, especially with large datasets (e.g., memory limitations; runtimes ranging from several hours to several weeks). Advances in computing may alleviate some concerns regarding computational intensity, but there is a large carbon footprint associated with using exceptionally large cloud computing resources (Strubell et al., 2019). Further, there is the potential for negligible incremental performance compared to traditional approaches and thus machine learning may not be preferable or efficient for certain data structures (Christodoulou et al., 2019). Moreover, even if incremental performance over traditional methods is large, it may not be clinically meaningful. For example, rare outcomes such as suicide mortality are often predicted with low positive predictive values, even when using machine learning (Belsher et al., 2019). Unlike sensitivity, the positive predictive value of an algorithm is dependent on the prevalence of the outcome/disease in the study population. The relative importance of sensitivity versus positive predictive value depends on intervention cost-effectiveness and availability of resources. A test with good sensitivity but poor positive predictive value will capture most cases, but there will be many false positives. Accordingly, algorithms with low positive predictive value could be used in conjunction with affordable/low-resource interventions (e.g., suicide assessment at annual physical examination), but expensive/high-resource interventions such as a 12-week cognitive behavioral therapy program would not be practical because most “high risk” individuals will not actually experience the outcome.

Conclusions

Researchers must be thoughtful in determining if machine learning is the method best suited to their research question of interest, as well as cautious in the application and interpretation of such highly flexible methods. Whenever possible, studies using machine learning methods should compare performance against traditional statistical approaches that may perform similarly and be much more interpretable. Additional research is also needed to determine how measurement error affects various machine learning algorithms, and develop ways to harmonize measures to make external validation (i.e., replication) is more feasible. If methods are developed to address these challenges, machine learning has the potential to vastly improve our ability to understand and predict mental disorders and associated adverse outcomes.

Funding.

This work was supported by grants awarded to Dr. Gradus (R01MH109507, R01MH110453) and Dr. Rosellini (K01MH106710, R21MH119492) from the National Institute of Mental Health. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

References

Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, McGinn T, & Guyatt G (2017). Discrimination and calibration of clinical prediction models: Users’ guides to the medical literature. JAMA, 318(14), 1377–1384. 10.1001/jama.2017.12126 [PubMed: 29049590]

Askland KD, Garnaat S, Sibrava NJ, Boisseau CL, Strong D, Mancebo M, Greenberg B, Rasmussen S, & Eisen J (2015). Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy. International Journal of Methods in Psychiatric Research, 24(2), 156–169. 10.1002/mpr.1463 [PubMed: 25994109]

Belsher BE, Smolenski DJ, Pruitt LD, Bush NE, Beech EH, Workman DE, Morgan RL, Evatt DP, Tucker J, & Skopp NA (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. JAMA Psychiatry, 76(6), 642–651. 10.1001/jamapsychiatry.2019.0174 [PubMed: 30865249]

Bergquist SL, Brooks GA, Keating NL, Landrum MB, & Rose S (2017). Classifying lung cancer severity with ensemble machine learning in health care claims data. Proceedings of Machine Learning Research, 68, 25–38. [PubMed: 30542673]

Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, & Moons KGM (2012). Reporting and methods in clinical prediction research: A systematic review. PLOS Medicine, 9(5), 1–12. 10.1371/journal.pmed.1001221

Breiman L (2001). Random forests. Machine Learning, 45(1), 5–32. 10.1023/A:1010933404324

Breiman L, Friedman J, Stone CJ, & Olshen RA (1984). Classification and Regression Trees. Taylor & Francis.

Brown TA (2015). Confirmatory Factor Analysis for Applied Research, Second Edition Guilford Publications.

Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, & McCarthy G (2017). Reevaluating the efficacy and predictability of antidepressant treatments. JAMA Psychiatry, 74(4), 370–378. 10.1001/jamapsychiatry.2017.0025 [PubMed: 28241180]

Chipman HA, George EI, & McCulloch RE (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1), 266–298. 10.1214/09-AOAS285

Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, & Van Calster B (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. 10.1016/j.jclinepi.2019.02.004 [PubMed: 30763612]

Cole SR, Chu H, & Greenland S (2006). Multiple-imputation for measurement-error correction. International Journal of Epidemiology, 35(4), 1074–1081. 10.1093/ije/dyl097 [PubMed: 16709616]

Collins GS, & Moons KGM (2019). Reporting of artificial intelligence prediction models. The Lancet, 393(10181), 1577–1579. 10.1016/S0140-6736(19)30037-6

Collins GS, Reitsma JB, Altman DG, & Moons KGM (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. Annals of Internal Medicine, 162(1), 55 10.7326/M14-0697 [PubMed: 25560714]

Cuthbert BN, & Insel TR (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. BMC Medicine, 11, 126 10.1186/1741-7015-11-126 [PubMed: 23672542]

Edwards JK, Cole SR, Troester MA, & Richardson DB (2013). Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. American Journal of Epidemiology, 177(9), 904–912. 10.1093/aje/kws340 [PubMed: 24627573]

Figueroa R, Zeng-Treitler Q, Kandula S, & Ngo L (2012). Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making, 12(8), 1–10. 10.1186/1472-6947-12-8 [PubMed: 22217121]

Frénay B, Doquire G, & Verleysen M (2014). Estimating mutual information for feature selection in the presence of label noise. Computational Statistics & Data Analysis, 71, 832–848. 10.1016/j.csda.2013.05.001

Frénay B, & Verleysen M (2014). Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems, 25(5), 845–869. 10.1109/TNNLS.2013.2292894 [PubMed: 24808033]

Friedman J (2002). Stochastic gradient boosting. *Computational Statistics &* Data Analysis, 38(4), 367–378.

Friedman J, Hastie T, & Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22. [PubMed: 20808728]

Galatzer-Levy IR (2015). Applications of latent growth mixture modeling and allied methods to posttraumatic stress response data. European Journal of Psychotraumatology, 6, 27515 10.3402/ejpt.v6.27515 [PubMed: 25735414]

Galatzer-Levy IR, Karstoft K-I, Statnikov A, & Shalev AY (2014). Quantitative forecasting of PTSD from early trauma responses: A machine learning application. Journal of Psychiatric Research, 59, 68–76. 10.1016/j.jpsychires.2014.08.017 [PubMed: 25260752]

Galatzer-Levy IR, Ruggles K, & Chen Z (2018). Data science in the research domain criteria era: Relevance of machine learning to the study of stress pathology, recovery, and resilience. Chronic Stress, 2, Epub. 10.1177/2470547017747553

Gerlach R, & Stamey J (2007). Bayesian model selection for logistic regression with misclassified outcomes: Statistical Modelling, 7(3), 255–273. 10.1177/1471082X0700700303

Gradus JL, Rosellini AJ, Horváth-Puhó E, Street AE, Galatzer-Levy IR, Jiang T, Lash TL, & Sørensen HT (2020). Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. JAMA Psychiatry, 77(1), 25–34. 10.1001/jamapsychiatry.2019.2905

Hasanpour H, Ghavamizadeh Meibodi R, Navi K, & Asadi S (2018). Novel ensemble method for the prediction of response to fluvoxamine treatment of obsessive-compulsive disorder. Neuropsychiatric Disease and Treatment, 14, 2027–2038. 10.2147/NDT.S173388 [PubMed: 30127613]

Hawkins DM (2004). The problem of overfitting. Journal of Chemical Information and Computer Sciences, 44(1), 1–12. 10.1021/ci0342472 [PubMed: 14741005]

Heinze G, Wallisch C, & Dunkler D (2018). Variable selection-A review and recommendations for the practicing statistician. Biometrical Journal. Biometrische Zeitschrift, 60(3), 431–449. 10.1002/bimj.201700067 [PubMed: 29292533]

Hernán MA, Hsu J, & Healy B (2019). A second chance to get causal inference right: A classification of data science tasks. Chance, 32(1), 42–49. 10.1080/09332480.2019.1579578

Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, & Shrier I (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of Clinical Epidemiology, 79, 70–75. 10.1016/j.jclinepi.2016.04.014 [PubMed: 27237061]

Hosmer D, & Lemeshow S (2013). Applied Logistic Regression (3 edition). John Wiley & Sons.

Hothorn T, Hornik K, & Zeileis A (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3), 651–674. 10.1198/106186006X133933

Hsu KJ, McNamara M, Shumake J, Alario A, Gonzalez GDS, Schnyer DM, … Beevers CG (2020, 5 7). Neurocognitive predictors of self-reported reward responsivity and approach motivation in depression: a machine learning approach. Retrieved from osf.io/wcg69

Hubbard RA, Huang J, Harton J, Oganisian A, Choi G, Utidjian L, Eneli I, Bailey LC, & Chen Y (2019). A Bayesian latent class approach for EHR-based phenotyping. Statistics in Medicine, 38(1), 74–87. 10.1002/sim.7953 [PubMed: 30252148]

Huys QJM, Maia TV, & Frank MJ (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience, 19(3), 404–413. 10.1038/nn.4238 [PubMed: 26906507]

Ilgen MA, Downing K, Zivin K, Hoggatt KJ, Kim HM, Ganoczy D, Austin KL, McCarthy JF, Patel JM, & Valenstein M (2009). Exploratory data mining analysis identifying subgroups of patients with depression who are at high risk for suicide. The Journal of Clinical Psychiatry, 70(11), 1495–1500. 10.4088/JCP.08m04795 [PubMed: 20031094]

James G, Witten D, Hastie T, & Tibshirani R (2013). An Introduction to Statistical Learning: With Applications in R. Springer-Verlag //www.springer.com/us/book/9781461471370

Jo B, Findling RL, Wang C-P, Hastie TJ, Youngstrom EA, Arnold LE, Fristad MA, & Horwitz SM (2017). Targeted use of growth mixture modeling: A learning perspective. Statistics in Medicine, 36(4), 671–686. 10.1002/sim.7152 [PubMed: 27804177]

Källberg H, Alfredsson L, Feychting M, & Ahlbom A (2010). Don’t split your data. European Journal of Epidemiology, 25(5), 283–284. 10.1007/s10654-010-9447-3 [PubMed: 20339902]

Kessler RC, Stein MB, Petukhova MV, Bliese P, Bossarte RM, Bromet EJ, Fullerton CS, Gilman SE, Ivany C, Lewandowski-Romps L, Millikan Bell A, Naifeh JA, Nock MK, Reis BY, Rosellini AJ, Sampson NA, Zaslavsky AM, Ursano RJ, & Army STARRS Collaborators. (2017). Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Molecular Psychiatry, 22(4), 544–551. 10.1038/mp.2016.110 [PubMed: 27431294]

Kessler RC, Bossarte RM, Luedtke A, Zaslavsky AM, & Zubizarreta JR (2019). Machine learning methods for developing precision treatment rules with observational data. Behaviour Research and Therapy, 120, 103412 10.1016/j.brat.2019.103412 [PubMed: 31233922]

Kessler Ronald C., Rose S, Koenen KC, Karam EG, Stang PE, Stein DJ, Heeringa SG, Hill ED, Liberzon I, McLaughlin KA, McLean SA, Pennell BE, Petukhova M, Rosellini AJ, Ruscio AM, Shahly V, Shalev AY, Silove D, Zaslavsky AM, … Viana MC (2014). How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? An exploratory study in the WHO World Mental Health Surveys. World Psychiatry, 13(3), 265–274. 10.1002/wps.20150 [PubMed: 25273300]

Kessler Ronald C., Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, Brown M, Cai T, Colpe LJ, Cox KL, Fullerton CS, Gilman SE, Gruber MJ, Heeringa SG, Lewandowski-Romps L, Li J, Millikan-Bell AM, Naifeh JA, Nock MK, … Army STARRS Collaborators. (2015). Predicting suicides after psychiatric hospitalization in US Army soldiers: The Army Study To Assess Risk and Resilience in Servicemembers (Army STARRS). JAMA Psychiatry, 72(1), 49–57. 10.1001/jamapsychiatry.2014.1754 [PubMed: 25390793]

Kircanski K, Zhang S, Stringaris A, Wiggins JL, Towbin KE, Pine DS, Leibenluft E, & Brotman MA (2017). Empirically derived patterns of psychiatric symptoms in youth: A latent profile analysis. Journal of Affective Disorders, 216, 109–116. 10.1016/j.jad.2016.09.016 [PubMed: 27692699]

Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, Brown TA, Carpenter WT, Caspi A, Clark LA, Eaton NR, Forbes MK, Forbush KT, Goldberg D, Hasin D, Hyman SE, Ivanova MY, Lynam DR, Markon K, … Zimmerman M (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. Journal of Abnormal Psychology, 126–134(4), 454–477. 10.1037/abn0000258

Lachenbruch PA (1966). Discriminant analysis when the initial samples are misclassified. Technometrics, 8(4), 657–662. JSTOR. 10.2307/1266637

Lash TL, Fox MP, & Fink AK (2009). Applying Quantitative Bias Analysis to Epidemiologic Data. Springer-Verlag //www.springer.com/us/book/9780387879604

Lee BK, Lessler J, & Stuart EA (2010). Improving propensity score weighting using machine learning. Statistics in Medicine, 29(3), 337–346. 10.1002/sim.3782 [PubMed: 19960510]

Luijken K, Groenwold RHH, Calster BV, Steyerberg EW, & van Smeden M. (2019). Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. Statistics in Medicine, 38(18), 3444–3459. 10.1002/sim.8183 [PubMed: 31148207]

Månsson KNT, Frick A, Boraxbekk C-J, Marquand AF, Williams SCR, Carlbring P, Andersson G, & Furmark T (2015). Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Translational Psychiatry, 5(3), e530–e530. 10.1038/tp.2015.22 [PubMed: 25781229]

Mitchell TM (1997). Machine Learning. McGraw-Hill.

Nettleton DF, Orriols-Puig A, & Fornells A (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial Intelligence Review, 33(4), 275–306. 10.1007/s10462-010-9156-z

Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, Watts B, Flashman L, & McAllister T (2014). Predicting the risk of suicide by analyzing the text of clinical notes. PLOS ONE, 9(1), e85733 10.1371/journal.pone.0085733 [PubMed: 24489669]

Ripley BD, & Hjort NL (1995). Pattern Recognition and Neural Networks (1st ed.). Cambridge University Press.

Rosellini A, Dussaillant F, Zubizarreta JR, Kessler RC, & Rose S (2018). Predicting posttraumatic stress disorder following a natural disaster. Journal of Psychiatric Research, 96, 15–22. 10.1016/j.jpsychires.2017.09.010 [PubMed: 28950110]

Rosellini AJ, & Brown TA (2014). Initial interpretation and evaluation of a profile-based classification system for the anxiety and mood disorders: Incremental validity compared to DSM-IV categories. Psychological Assessment, 26(4), 1212–1224. 10.1037/pas0000023 [PubMed: 25265416]

Rosellini AJ, Liu S, Anderson GN, Sbi S, Tung ES, & Knyazhanskaya E (2020). Developing algorithms to predict adult onset internalizing disorders: An ensemble learning approach. Journal of Psychiatric Research, 121(1), 189–196. 10.1016/j.jpsychires.2019.12.006 [PubMed: 31864158]

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–701. 10.1037/h0037350

Schuler MS, & Rose S (2017). Targeted maximum likelihood estimation for causal inference in observational studies. American Journal of Epidemiology, 185(1), 65–73. 10.1093/aje/kww165 [PubMed: 27941068]

Shanab AA, Khoshgoftaar TM, & Wald R (2012). Robustness of threshold-based feature rankers with data sampling on noisy and imbalanced data. Twenty-Fifth International FLAIRS Conference. Twenty-Fifth International FLAIRS Conference https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS12/paper/view/4451

Shatte ABR, Hutchinson DM, & Teague SJ (2019). Machine learning in mental health: A scoping review of methods and applications. Psychological Medicine, 49(9), 1426–1448. 10.1017/S0033291719000151 [PubMed: 30744717]

Steinwart I, & Christmann A (2008). Support Vector Machines. Springer-Verlag //www.springer.com/us/book/9780387772417

Strobl C, Boulesteix A-L, Zeileis A, & Hothorn T (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8, 25 10.1186/1471-2105-8-25 [PubMed: 17254353]

Strubell E, Ganesh A, & McCallum A (2019). Energy and policy considerations for deep learning in NLP. ArXiv:1906.02243 [Cs] http://arxiv.org/abs/1906.02243

Tackett JL, Lilienfeld SO, Patrick CJ, Johnson SL, Krueger RF, Miller JD, Oltmanns TF, & Shrout PE (2017). It’s time to broaden the replicability conversation: Thoughts for and from clinical psychological science. Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12(5), 742–756. 10.1177/1745691617690042 [PubMed: 28972844]

van der Laan MJ, Polley EC, & Hubbard AE (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6, Article 25. 10.2202/1544-6115.1309

van der Laan MJ, & Rubin D (2006). Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1), Article 11. 10.2202/1557-4679.1043

Wald N, & Bestwick J (2014). Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test? Journal of Medical Screening, 21(1), 51–56. 10.1177/0969141313517497 [PubMed: 24407586]

Walsh CG, Ribeiro JD, & Franklin JC (2017). Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science, 5(3), 457–469. 10.1177/2167702617691560

Webb CA, Cohen ZD, Beard C, Forgeard M, Peckham AD, & Björgvinsson T (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. Journal of Consulting and Clinical Psychology, 88(1), 25–38. 10.1037/ccp0000451 [PubMed: 31841022]

Wilson DR, & Martinez TR (2000). Reduction techniques for instance-based Learning algorithms. Machine Learning, 38(3), 257–286. 10.1023/A:1007626913721

Wu M-J, Mwangi B, Bauer IE, Passos IC, Sanches M, Zunta-Soares GB, Meyer TD, Hasan KM, & Soares JC (2017). Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. NeuroImage, 145(Pt B), 254–264. 10.1016/j.neuroimage.2016.02.016 [PubMed: 26883067]

Zou H, & Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. 10.1111/j.1467-9868.2005.00503.x

Highlights

- Machine learning may help characterize and predict psychiatric outcomes
- Description of commonly used supervised learning methods
- Introduction to model building, validation, and evaluation of algorithms
- Discussion of challenges and opportunities to move field forward

Table 1.

Comparison of terminology in traditional statistics vs. supervised machine learning

Traditional statistics	**Machine learning**
Prediction	Supervised learning
Predictors/covariates/independent variables	Features
Outcome/dependent variable	Output/target
Prediction of categorical outcomes	Classification
Prediction of continuous outcomes	Regression
Number/overlap of predictors	Dimensionality
R-squared	Coefficient of determination
Sensitivity	Recall
Positive predictive value	Precision
Contingency table	Confusion matrix

Table 2.

Brief descriptions of common approaches to supervised learning

Algorithm	R package	Description	Strengths/Limitations
Conventional Regression	*stats*	• Traditional parametric linear or logistic regression	Strengths: • Easily interpretable Limitations: • Prone to overfit if independent variables are highly collinear • Optimal functional form and interactions must be specified a priori
Decision tree Classification and regression trees (Breiman et al., 1984) Random forest (Breiman, 2001) Bayesian regression trees (Chipman et al., 2010)	*rpart* *randomFor est* *party* *BART*	• Decision tree methods automatize detection of interactions and non-linear main effects • Predictors are partitioned (based on values) and stacked to build decision trees and ensemble an aggregate "forest" • Random forests builds numerous trees in bootstrapped samples and generates an aggregate tree by averaging across trees (reducing overfit) • Bayesian trees are based on an underlying probability model (priors) for the structure and likelihood for data in terminal nodes; aggregate tree is generated by averaging across tree posteriors (reducing overfit)	Strengths: • Decision trees are visually interpretable • Random forests have lower risk of overfitting than a single tree • Variable importance metrics provided Limitations: • Individual classification trees may overfit the data • Forests are not visually interpretable (e.g., variable importance metrics do not indicate direction of effects)
Support vector machines (Steinwart & Christmann, 2008) Linear kernel Polynomial kernel Radial kernel	*e1401*	• Support vector machines treat each predictor as dimensions in high dimensional space and attempts to identify the best hyperplane to separate the sample into classes (e.g., cases and non-cases) • Goal is to find the hyperplane with the maximum margin between the two closest points in space • Captures linear associations, but alternate kernels can be used to capture nonlinearities	Strengths: • Performs well when with highly complex data (e.g., text, images) including when number of predictors is larger than the number of observations ("high dimensional" data) Limitations: • More prone to overfit than other algorithms (e.g., regularization) • "Black box" algorithm; metrics are not provided for how predictors are combined to optimize hyperplane (may be unclear *why* predictions are accurate)
Regularization (Friedman et al., 2010) Ridge Elastic net Lasso	*glmnet*	• Penalized regression reduces overfit due to collinear independent variables • Ridge regression shrinks coefficients for collinear independent variables *toward* zero, but does not fully-eliminate any independent variable • Elastic net regression allows various penalties where coefficients for collinear independent variables are shrunk *toward* zero (but not to	Strengths: • Penalization decreases risk of overfitting when variables are highly correlated • Can be used for variable selection Limitations: • LASSO may arbitrarily select which coefficients to shrink

Algorithm	R package	Description	Strengths/Limitations
		eliminating contributions to the predicted probability) and/or *to* zero (eliminating their contributions to the predicted probability) • Mixing parameter penalty (alpha) is set somewhere between .01 and .99. • Lasso regression shrinks coefficients for collinear covariate coefficients to zero, eliminating their contributions to the predicted probability	(e.g., conceptually important predictors) • Regression coefficients may not be interpretable after shrinkage • Cannot calculate standard errors for coefficients
Super learning	*SuperLearner*	• Ensembles a composite algorithm from any number user-specified approaches to supervised learning (e.g., a single algorithm based on all of the above approaches) • Outcome is regressed onto predicted values estimated by individual algorithms • Implemented using k-folds cross-validation	Strengths: • Combines predictions from multiple machine learning methods to optimize prediction performance Limitations: • Computational intensity and efficiency; individual algorithms (including logistic regression) may perform comparably
Other useful R packages	*swirl* *caret h2o mlr* *shiny*	• Text-based interface that teaches programming, data manipulation, and analysis in R • Broad machine learning packages able to implement many forms of supervised learning, feature selection, and cross-validation • Used to develop interactive web-apps - e.g., risk calculators based on prediction functions developed in R	