Article

Practical Foundations of Machine Learning for Addiction Research.
Part I. Methods and Techniques

Pablo Cresta Morgado [1,*], Martín Carusso [1], Laura Alonso Alemany [2] and Laura Acion [1,3]

[1] Instituto de Cálculo, FCEyN, Universidad de Buenos Aires - CONICET, Argentina
[2] FaMAF, Universidad Nacional de Córdoba, Argentina
[3] Department of Psychiatry, University of Iowa, Iowa, USA
* Correspondence: pablo_crestam@hotmail.com; Address: Intendente Güiraldes 2160, Ciudad Universitaria. (C1428EGA) Buenos Aires, Argentina.

Abstract: Machine learning assembles a broad set of methods and techniques to solve a wide range of problems, such as identifying individuals with substance use disorders (SUD), finding patterns in neuroimages, understanding SUD prognostic factors and their association, or determining addiction genetic underpinnings. However, machine learning use in the addiction research field continues to be insufficient. This two-part review focuses on machine learning tools and concepts and provides insights into their capabilities to facilitate their understanding and acquisition by addiction researchers. In this first part, we present supervised and unsupervised methods and techniques such as linear models, naive Bayes, support vector machines, artificial neural networks, k-means, or principal component analysis and examples of how these tools are already in use in addiction research. We also provide open-source programming tools to apply these techniques. Throughout this work, we link machine learning techniques to applied statistics. Machine learning tools and techniques can be applied to many addiction research problems and can improve addiction research.

Keywords: machine learning, data science, artificial intelligence, statistical methods, addiction

1. Introduction

Several biomedical research domains, including addiction, successfully applied machine learning methods and techniques in the last decade. Machine learning can be applied to solve a wide range of problems, such as identifying individuals with substance use disorders (SUD) (1), evaluating treatment success (2), finding patterns on brain images (3), understanding SUD prognostic factors and their association (4), or identifying addiction genetics underpinnings (5). In all these scenarios, the use of the rich analytical machine learning toolbox can improve results.

Despite its advantages, the application of machine learning in addiction research is still scarce (6). In a recent search in PubMed, we found less than 200 articles about machine learning and addiction (Figure 1). Other bibliographical surveys present similar figures (6). Although the number has been increasing in the last year, articles using machine learning make only 0.25There is no doubt about the improvements that machine learning can bring for many application domains, in particular, in the health sciences (7). Comprising a wide variety of methods, machine learning can help answer a broad set of research questions.

The use of machine learning in most of the subdomains of addiction research continues to be insufficient, probably, due to the lack of understanding about these methods, which are often seen as black boxes. Besides, machine learning jargon differs substantially from that of the familiar applied statistics field, even when several concepts are the same. The field of machine learning is sufficiently mature that many of the tools and techniques are accessible to researchers with a very reasonable amount of effort. For instance, free and open-source software implementations of all discussed methods are readily available.

This machine learning review aims to bring closer these concepts, methods, and implementations. We will open the black box and show what tools are available and how to use them. We will also relate machine learning concepts to more familiar statistical terms.

Citation: Cresta Morgado, P.; Carusso, M.; Alonso Alemany, L.; Acion, L. Machine Learning for Addiction Research. *Preprints* **2021**, *1*, 0. https://doi.org/

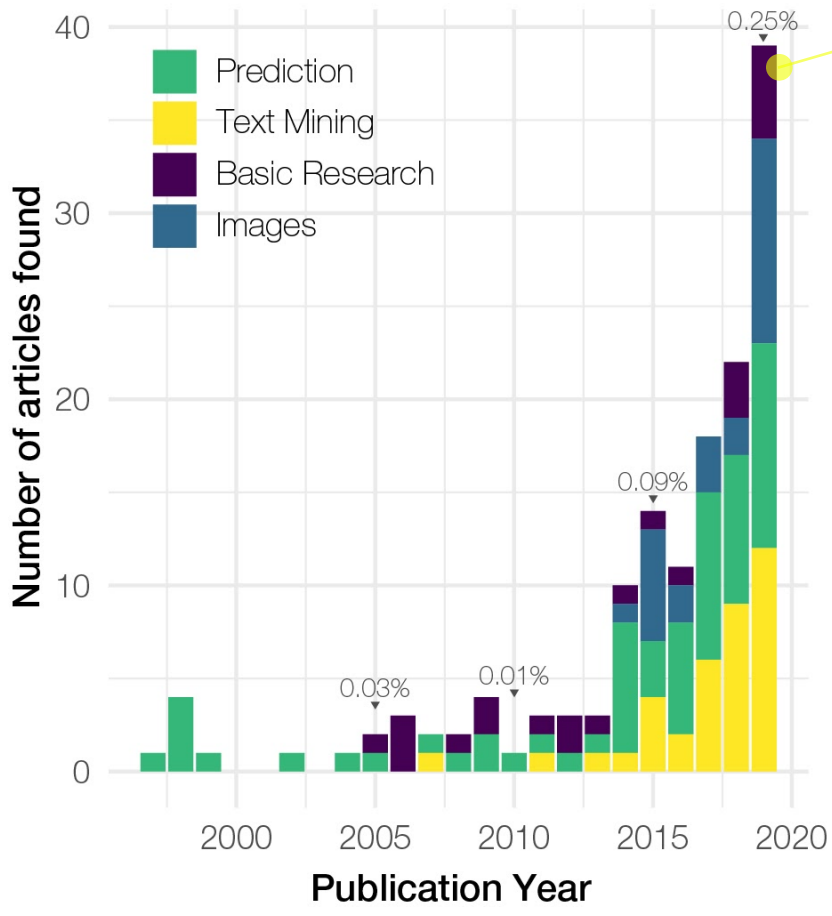
Maadhav Kothuri: Data mining application for ML

Maadhav Kothuri: Machine learning has more applications than just engineering and robotics. It can be useful really with any data-driven field

Maadhav Kothuri: Without knowledge of them, they can't actually be used

Maadhav Kothuri: Something mysterious and unknown

Maadhav Kothuri: I didn't know that applied statistics was so relevant to addiction research, but it does make sense, since there is likely a lot of data that needs to be analyzed through statistics in the field



Maadhav Kothuri: Even at the peak, articles found mentioning ML is only at 0.25%?? Seems that there is a wider application of machine learning that isn't being used just yet

Figure 1. Articles of addiction research applying machine learning according to a PubMed search. The basic research, images, prediction, and text mining groups identify articles according to their use of machine learning depending on the data or scientific context presented. The plot shows the evolution of the use of machine learning in addiction research since 1995.

This article is the first part of this review of machine learning in addiction research. We start by introducing foundational concepts of machine learning, followed by a description of tools and available methods. For every machine learning technique, we offer a reference to open source software implementing it. The second part illustrates the machine learning process, intending to consolidate the concepts, methods, and techniques presented in the first part (8). We provide a glossary (in the supplemental material) with standard machine learning terms (featured in italics) and their equivalents in applied statistics, the most relevant ones are displayed in Appendix 5.

Maadhav Kothuri: !!! Open source software can make it much easier to get hands-on experience with these concepts, and could even be used in my original work and final product!!!

2. Potential of Machine Learning

From a black-box perspective, machine learning is a set of different computational and statistical tools commonly used to provide an automated solution for a repetitive task, based on examples that have been typically solved by human experts. Far from a magic device, machine learning tools are a natural extension of traditional statistical approaches towards automation. There is a continuum between traditional statistical models (with components specified by humans) and a fully machine-guided data analysis (9).

Maadhav Kothuri: Can train a model to do something over and over and over again relatively easily

Maadhav Kothuri: Explains why so much of ML models relies on formulas and math (much more than I expected)

Machine learning assembles techniques that learn from experience to improve their performance at a given task (e.g., clustering, classification, prediction) (10). Experience is leveraged from observed data, from which actionable information is extracted in the form of a model using an algorithm. There are three elements in common in every learning problem: the task's class to be resolved, the observed data, and the performance measure to be improved (10). Model performance is evaluated during the learning process in order to improve it by estimating algorithm parameters. There is a wide variety of use cases of machine learning in addiction research, including but not limited to SUD prediction, discerning a binging from a non-binging event, identifying genes linked to addiction, or detecting and connecting words from tweets related to opioid unhealthy use. For the problems above, observed data can be from different sources: electronic health records, questionnaires or interviews, genomic data, or thousands of tweets, respectively. The performance measure should also be in agreement with the task.

Maadhav Kothuri: Fundamental elements of an ML system

It is common to classify machine learning methods into two main categories: supervised and unsupervised (10, 11). Supervised methods require an outcome, response, or dependent variable, which is the task target. The task target guides or supervises (as is commonly said in the machine learning community) the learning process. A set of independent variables or features relevant to predict or explain the response represent each observation. Thus, a machine learning method finds and uses a relationship between the independent variables and the response, inferring a model that links them. The model is then used to accurately predict a previously unseen observation (a classification or prediction task) or to reach a better understanding between the response and the independent variables (an inference task) (11). Supervised methods can answer the following questions: How can we identify a risk population? What are the most relevant variables associated with SUD? How can we distinguish patients with and without SUD?

Maadhav Kothuri: The output result that helps the model train itself

There are many approaches to discover the relation between response and independent variables, and usually, several good relations to be discovered in a given dataset. However, it is usually unfeasible to find and evaluate all possible relations, so methods find only the most promising. When evaluating relations between independent variables and predictions, methods have to balance bias and variance. If variance is prioritized, the resulting model tends to suffer from overfitting, that is, a model that reproduces training examples without generalization. Conversely, if bias is prioritized, then the resulting model will be very similar to the starting hypothesis and valuable information from the data will probably be missed. It is crucial that machine learning methods find a good balance between these two objectives in order to obtain valuable models of data. For a deeper discussion of these and other foundational concepts, see Appendix 5.

Maadhav Kothuri: Really? I hadn't realized this before since I've always thought of it as being one variable to determine the outcome, but this makes much more sense

Maadhav Kothuri: Measures how far data is spread out

Maadhav Kothuri: So it wouldn't be able to recognize new examples

While supervised methods imply an underlying relationship between the outcome and the independent variables, unsupervised methods most often help explore and discover

patterns in the data. They work without a response variable. There is not a 'right answer' to supervise the analysis (11). Unsupervised machine learning can group observations with certain similarities into clusters and combine (partially) redundant features into new ones. These methods can handle questions such as: Is there a way to find, in a high number of participants, groups that share characteristics, even when we do not know what may relate them in the first place? How can Twitter users be grouped considering their tweet production regarding alcohol use? Can we reduce the dimension or compile items from extensive questionnaires in a more useful way? Some specific areas of machine learning, like image and natural language processing, time series analysis, or reinforcement learning, require an amount of specific knowledge that is beyond the scope of this review. However, they can still be used at a high level of abstraction as black boxes. An outline of such areas and approaches can be seen in Box 2.

Next, we dissect in more detail supervised and unsupervised machine learning methods and their application to addiction research. All discussed techniques are readily available through free and open-source software implementations in R or Python programming languages (13,14). The functions for each algorithm are usually in specialized libraries. Appendix 5 provides a primer toolbox for both languages.

3. Using supervised machine learning for prediction

One of the most frequent uses of machine learning is to make predictions (6,15), for example, to distinguish individuals (e.g., with and without a binge drinking behavior (16)) or predict an event (e.g., an opioid overdose (17)). To address this kind of problem, we can use machine learning supervised methods. In this type of approach, we can identify a set of predictor variables and a response or target variable. According to the type of response, these methods are regressions or classifications (10, 11). In the former, the response is a number such as a score (18) or the age of SUD onset (19). In classification problems, the response most often organizes into categories (e.g., the occurrence of an opioid overdose (17)).

The complexity of the prediction model is another aspect to consider. It depends on the kind of question we seek to answer, what data are available, and the kind of relationship between predictors and the outcome (7). Generalized linear models often work well to address questions for which we have relatively small datasets with a relatively reduced number of predictors (7, 8). In many cases, linear models are the best alternative and often recommended, at a minimum, as a starting model. However, prediction tasks with a high number of features and known dependencies between predictors and observations, such as neuroimaging, require techniques capable of modeling more complex phenomena such as artificial neural networks (7, 17).

Simpler models usually have stronger assumptions. A linear relationship, the base of linear models, can be a much too restrictive assumption about a given problem, unable to capture complex data relationships. Several methods relax the linearity constraint resulting in more expressive models that can capture more complex data relationships. Nevertheless, there is a tradeoff between expressivity and generalization. Methods that model data too closely may incorporate anecdotal phenomena, errors, or noise. They cannot generalize correctly and, at the same time, are more sensitive to small data changes. When a model fits the data following anecdotal information too closely, it suffers from overfitting (11). Overfitting is undesirable because a model excessively fitted to one set of data will not make accurate predictions for unseen data. A useful analogy to understand model overfitting is to equate the model to a student who fails a test because they prepared only by memorizing answers from previous tests, instead of learning general patterns that apply to new test questions. In terms of interpretability, more straightforward methods are usually easier to understand than more complex and expressive ones. For instance, with some methods, such as neural networks, it can be completely opaque how predictors relate to the response. Thus, interpretability can be another factor to consider when choosing a method.

Maadhav Kothuri: ***This is huge! Unsupervised ML models are just for patterns (kind of like graph analysis) so it would be better for just data and not for recognizing something

Maadhav Kothuri: Based on the trend

Maadhav Kothuri: Used for modeling complex phenomena

Maadhav Kothuri: Interferes with generalization

In what follows, we describe prediction models starting with the simplest linear models and finishing with algorithms capable of representing more complex phenomena.

3.1. Linear models for continuous outcomes

Regarding regression methods, the simplest and most restrictive model is multiple linear regression. It is one of the most widely used and considered a good starting point to solve problems with a numeric response variable. For example, Locke et al (21) used linear regression to examine the relationship of an interpersonal guilt score, a numerical response variable, with SUD while adjusting by sociodemographic covariates. The simplicity of linear regression makes it unbeatable at interpretation and computational requirements.

A more sophisticated variant of linear regression is penalized or regularized regression. Penalized regression is slightly more computationally expensive but powerful due to using a different method of estimation. Penalized regression incorporates a penalty (or regularization) term in the linear model, which allows the model to obtain a smooth model that leaves out anecdotal or noisy information from the data to obtain better generalizations (2,11). The least absolute shrinkage and selection operator (lasso), ridge regression, and elastic net are different penalized regression parameterizations. While lasso forces some coefficients to be zero, ridge regression keeps them in the final model, and the elastic net is an intermediate between the strong lasso and the less restrictive ridge regularizations. Lasso can eliminate less useful predictors, which is most suitable for problems with many predictors that exceed the number of observations (also known as the curse of dimensionality or k > n problem). Ridge regression is better than lasso when variables are highly correlated, while lasso keeps the interpretability of multiple linear regression, making it possible to deal with an exceedingly high number of predictors (2). In addiction research, Morozova et al. (22) compared the performance of lasso and stepwise regression for selecting relevant variables for the association between SUD-related variables and a quality-of-life score, showing disparities between the variables that each method had selected.

When linearity is too simple for the problem at hand, polynomial regression allows accommodating non-linear relationships between predictors and a target by raising each of the original predictors to a power (11). Alternatively, step functions cut the range of a variable into several distinct regions and generate a categorical variable, fitting a piecewise constant function to each region (11). Regression splines are an extension of polynomial regression and step functions; they divide the range of the predictors into several regions and fit in each region a polynomial function (11). Interestingly, Linden-Carmichael et al. (23) applied regression splines to predict alcohol use disorder (AUD) prevalence according to the number of drinks in the past year. The authors used the regression spline curves to find a threshold for the number of drinks for which the AUD prevalence rate stabilizes. Smoothing splines modify regression splines by adding a penalty term that constraints their expressivity, and thus their capacity to overfit the data (11).

Generalized additive models (often referred to as GAMs) maintain the principle of additivity of all models mentioned before (i.e., predictors are summed to each other rather than multiplied by each other to predict an outcome) using a different function for accounting for each predictor. Generalized additive models use building blocks where each predictor can relate to the target through its functional form (11). Chilcoat and Schütz (24) used generalized additive models with smoothing splines to analyze the association between the use of hallucinogens (a binary outcome) with age, adjusting by sociodemographic covariates. This approach allowed them to detect a nonlinear relationship between age and hallucinogen use.

3.2. Linear models for binary outcomes

The method most similar to linear regression for predicting binary and other categorical outcomes is logistic regression (25). This is the baseline technique for categorical prediction, usually applied to compare with more complex approaches. However simple, if the representation of the problem is adequate, logistic regression provides excellent results,

Maadhav Kothuri: Conversely, this could make it less sensitive to subtle data changes

Maadhav Kothuri: More features = more error and more runtime

Maadhav Kothuri: Allows for more accurate models that find a greater medium between generalization and sensitivity

as seen in the study of Sears and Anthony (26), where logistic regression obtained the same performance as the far more complex artificial neural networks (presented at the end of this section) to assess adolescent marijuana use based on survey questions about the history of alcohol and tobacco consumptions. Frequently, when logistic regression is compared with other machine learning methods, it falls within those with the best performance (2,4).

Support vector machines (often referred to as SVMs) are also a handy and cost-efficient linear classifier, and they include a generic method to improve the representation of the problem for classification purposes. They avoid the computational cost of high dimensionality by exploiting mathematical properties and avoid overfitting. In the addiction domain, Kornfield et al. (27) used support vector machines to triage messages of persons seeking online support from an addiction recovery forum. Mete et al. (28) applied support vector machines to brain images and identified patients with cocaine use disorder from those without it.

3.3. Bayesian approaches

Another straightforward, widely used machine learning method for classification is the Naive Bayes algorithm. This method naively assumes independence between predictive variables. Naive Bayes selects the class with the highest probability, according to the conditional probability of the features, the probability of a class given the predictor variables probability (11,12). Mumtaz et al. (29) applied Naive Bayes, logistic regression, and support vector machine to identify individuals with AUD from healthy controls, using resting-state EEG-derived features, obtaining similar performance for the three methods with slightly better accuracy, but less sensitivity, for SVM than Naive Bayes. However, while it may perform well in some scenarios, in many others, the underlying assumption that each variable is independent of the rest does not hold, and the oversimplification of the Naive Bayes assumptions may provide inadequate results (30).

Other Bayesian methods can capture complex relations between phenomena, but the mathematical machinery to do that, for example, Bayesian nets, is very complex and beyond this work's scope.

3.4. K-nearest neighbors

The k-nearest neighbors' algorithm (often referred to as k-NN) is a non-parametric method used for classification and regression (31). In both cases, the input consists of the training examples and their location in the space defined by features or independent variables. For a new instance that needs to be classified, the k-nearest neighbors' algorithm finds the training instances closest to it in the feature space and assigns the new instance with the value provided by training those instances, either discrete (by voting) or continuous (by averaging, possibly weighted by distance). In this approach, the feature space fully determines the output.

3.5. Ensemble models

There are several other powerful methods in machine learning classified under the umbrella of ensemble methods. This term refers to machine learning algorithms that combine multiple models. The underlying idea is that an ensemble could be better than any of its constituents, capturing more complex characteristics, decreasing overfitting or prediction variability. These models apply the same algorithm to different versions of a data set (e.g., random forests) or applying a combination of algorithms (e.g., super learning) (12).

There are several widely used ensemble methods in which constituents are decision trees. Tree-based methods create a series of decision rules splitting the predictor space to predict a target. A decision tree summarizes these rules (11,12). Each branch of the tree represents a split of the predictor space. Without strong assumptions about the data, decision trees are very prone to overfitting. Several methods are available to avoid overfitting and improving decision trees for analytical purposes, with the concurrent cost

Maadhav Kothuri: The amount of data that is given can determine which models should be used

Maadhav Kothuri: These help avoid the limitations of complex ML models

Maadhav Kothuri: Decreases complexity but is risky depending on the situation in terms of accuracy

Maadhav Kothuri: Variables linked with a specific outcome

Maadhav Kothuri: !!!! Don't use this for highly interconnected situations

Maadhav Kothuri: Seems faster but less sensitive to details

of losing interpretability. Random forests are the most outstanding and useful prediction tree-based method. The random forests algorithm creates a set of datasets by sampling randomly with replacement from the original data (also known as bootstrapping) and then fits a tree for every new sample aggregating all predictions (i.e., bootstrap aggregation or bagging). This aggregation makes random forests an ensemble method. The creation of each branch in random forests includes only a random sample of predictors, which prevents overfitting to features that do not generalize to other samples, thus overcoming individual decision trees' tendency to overfitting (32). Like lasso, random forests can also deal with the dimensionality curse and serve variable selection well. Unlike lasso, random forests' interpretability is not as stellar because it accommodates highly non-linear relationships between the predictors and the outcome. Squeglia et al. (33) used random forests to select features from questionnaires and semi-structured interviews predicting the risk of developing a SUD. In this work, random forests had the best prediction accuracy compared to six other machine learning methods. Squeglia et al. is an example of a complex study being successfully addressed by random forests. Another family of ensemble methods uses boosting, a technique that trains models sequentially so that each model learns from the errors made by its predecessor (12,34). Boosting primarily reduces bias and also variance. In particular, gradient boosting works sequentially feeding each model, for example, a regression tree, with the residual errors made by the previous one (34). This technique deals with overfitting through the learning rate in each sequential step. For gradient boosting with regression trees, the learning rate corresponds to what each tree learns. If the rate is low, the possibility of overfitting decreases. A large number of sequential models could also end in overfitting. Therefore, the number of models to adjust sequentially is a parameter to consider when using gradient boosting. Lo-Ciganic et al. (35) developed a model to predict opioid overdose among Medicare beneficiaries. These authors used five machine learning methods and found gradient boosting and deep neural networks (see next section) had a similar global performance, but gradient boosting presented higher specificity and lower sensitivity than deep neural networks.

Maadhav Kothuri: Could use linear regression models to increase interpretability?

3.6. Artificial Neural Networks

Among the most expressive models, artificial neural networks (also known as neural networks) process the information from predictor variables through successive layers stacked on top of each other (36). Each layer transforms the data and the last layer produces the prediction. During neural network model training, the model predicted values are compared with actual observations, obtaining a measurement of how near the prediction was from the observation. An optimization algorithm then carries out the learning process, adjusting how the data are transformed within each layer to reduce the error between prediction and observation (36). This whole process runs iteratively until obtaining the best performance. Neural networks can capture dependencies and complicated relationships and incorporate many methods to avoid overfitting, but it is usually hard to understand each predictor variable's contribution to the outcome. Neural networks are very flexible. There are several architectures for different types of problems, with differing: numbers of layers, numbers of layer components (or neurons), ways to connect neurons, and ways to change the connections' weight (36). Some of the most successful architectures are well-established convolutional neural networks (for example, the VGG family of pre-trained networks for image analysis (37)), generative adversarial networks, or neural language models (for example, Bert (38) or the GPT family (39)). The latter three are part of the umbrella of methods often called deep learning; nevertheless, technical terms for the newest methodologies are continually evolving. Neural networks have been used in addiction research to identify cocaine-dependent participants using functional magnetic resonance imaging data (3). In another example, neural networks helped predict whether patients will become long- or short-term opioid users based on information available from electronic medical records (1).

Maadhav Kothuri: Unique element to neural networks?

4. Using unsupervised machine learning for clustering and pattern recognition

The task of discovering emerging relationships and groupings within the data without any predefined target is the domain of unsupervised learning methods. These tools are usually applied to the preliminary project phases, such as during data exploration, to understand a problem better.

4.1. Clustering

Clustering techniques aim to group elements in a dataset, so objects within a group are more similar than those in other groups. For this purpose, it is critical to define a similarity metric that mirrors the intuitive notion of human-expert-based similarity. Classical similarities used in clustering are different distances (e.g., Euclidean, cosine) between vectors representing objects. In this representation, each dimension in the vectors corresponds to one of the variables or features describing objects. Groups that result from applying clustering techniques can be non-overlapping (hard clustering) or overlapping (soft clustering), where each object has a probability of belonging to each cluster. Clustering can be flat or hierarchical if the algorithm produces a tree-like structure, recursively finding smaller clusters contained in bigger ones. A domain expert must analyze the groups obtained to interpret the latent causes governing the grouping (or clustering solution). This analysis may start at a quantitative characterization of the values of variables in each group, but it usually requires more insightful interpretation to add value to the technique.

K-means clustering is one of the most common algorithms, an excellent tradeoff between computational complexity and the resulting solution's optimality. It has been used, for instance, to group subjects with SUD with similar psychosocial or clinical features (40,41). In this method, users manually provide the k parameter for the optimal number of clusters. Since clustering usually pertains to the projects' exploratory stage, k tends to be a guess. Experimenting with different ks is standard practice and recommended to find satisfactory results. Given k, the k-means algorithm follows two steps iteratively until partitioning remains unchanged (12). In the first step, each observation is assigned to its nearest cluster geometric center. In the next step, the geometric center location moves to the mean of all data points assigned to that cluster. These two steps alternate until finding an assignment of objects to clusters where objects within a cluster are closest to each other and most distant to objects outside the cluster. The first location of each cluster center is random; then, the algorithm follows the learning process until it reaches a state where no reassignment could result in better optimization. Thus, the result of k-means clustering has the advantage of being easy to interpret, but a slight difference in k or the location of the initial centers can produce a different result (12). Another widely used clustering algorithm is Latent Dirichlet Allocation. It is often used for analyzing natural language datasets, for example, it has been successfully used to identify Twitter discussions associated with overdose death rate (42). However, a detailed description of its functioning is beyond the scope of this review (43).

Clustering techniques also allow detecting anomalies or outliers. If we assume that cluster centers characterize paradigmatic cases, those in each cluster's periphery could be anomalies, and we may subject them to specific inspection.

4.2. Representation Learning

Another family of tools within unsupervised methods generates changes in the feature space, like collapsing features into more meaningful entities for dimensionality reduction or embeddings. Within a supervised setting, where the target outcome is well-defined, variable selection helps reduce dimensionality. In unsupervised settings, the paradigmatic method for unsupervised dimensionality reduction is principal component analysis (also called PCA). Principal component analysis assists data visualization or is a preprocessing step to eliminate noise or reduce a dataset's size before applying supervised techniques. Each principal component is a linear combination of every feature in a way that retains the highest variance (11). Thus, the first component is a new dimension representing the

Maadhav Kothuri: Classification

Maadhav Kothuri: Could be subjective?

Maadhav Kothuri: Optimized but takes more time and processing power because number of comparisons that have to be made

Maadhav Kothuri: Helps streamline dataset to be more efficient and reduce outliers

most data variability in a single dimension. The second component then represents the most dataset variability that a dimension can capture, given the first component, and so on. Principal component analysis results are a set of few uncorrelated principal components available as an improved feature space for automatic classifiers/regressors or human analysis (11). Factor analysis is often associated with principal component analysis, but its characteristics for application to questionnaire design are more desirable as explained by Lloret-Segura et al. (44) and by Jolliffe (45).

Like principal component analysis, embeddings find a feature space where objects are represented differently, so that it is more feasible for a learning algorithm to find a model for a given task. Embeddings are the process of projecting the original space into a different space, of higher or lower dimension, where separations between classes may be easier to find, even linear separations in problems that cannot be linearly separated in the original space. In the context of neural networks, intermediate representations of a network may be a good representation of objects for related problems, improving the performance of tasks related to images and text (46–48).

5. Discussion

We have described a broad palette of machine learning tools and how they apply in the addiction domain. Many more addiction research projects can benefit from enhancing their current analytical toolbox by resorting to the robust, high-level implementation of these tools. For over two decades now, open-source software communities have boosted reliable, robust, and auditable software where the latest developments of machine learning and applied statistics are readily available for application to addiction research problems and replicating reference research. These communities regularly welcome new adopters and have reproducibility, replicability, ethical issues, and fairness as central objectives (see, for example, The Turing Way (49)) We believe many in addiction research stay away from machine learning because, often due to its jargon, it appears unrelated to well-established analytical practices in our field. This review shows how machine learning can be viewed as a natural instrumentalization of familiar and well-established statistical concepts. Differences between machine learning and statistics are likely motivated by each research field's particular evolution and in which subbranch of applied mathematics (i.e., statistics or computer sciences) methods were developed.

Notwithstanding, machine learning is not a field without shortcomings. Most of the mentioned methods (e.g., neural networks) have many parameters and require large amounts of data and considerable computational power. Lack of generalization (or overfitting) is a recurrent pitfall, and most complex models do not address interpretability. Nevertheless, interpretability is a growing study area. Some solutions integrate complementary strategies to complex models that provide explainability for machine learning decisions.

Another limitation in the current machine learning literature is the lack of consideration to inference and informing variability of results, an area where classical applied statistics shines. However, addressing these shortcomings is also part of the mainstream research agenda in machine learning.

Unstructured data sources, such as electronic health records, medical images, or social media data, make available vast quantities of secondary data for advancing research on addiction. However, processing these sources requires particular techniques, which lay beyond the scope of this review. Luckily, high-level software tools can be readily used off-the-shelf as black-box solutions to obtain representations of images or texts to input them to standard machine learning processes. Box 2 includes an outline of such areas and approaches.

This two-part review is a brief introduction to the capabilities and applicability of machine learning to addiction research. All the presented concepts constitute a research area full of nuances, potential, and conceptual depth. Readers are encouraged to take advantage of the many resources cited here and in the second part of this review (8), to

Maadhav Kothuri: Used to reduce curse of dimensionality and preserve processing power

Maadhav Kothuri: There are resources available to apply ML to these fields in a cost-effective manner!

Maadhav Kothuri: Pretty big limitations since not everybody has access to these

Maadhav Kothuri: Don't necessarily need to know HOW they work, just need to use them

deepen their understanding of the presented concepts. In this first part, we highlighted methods and techniques. In its second part (8) a machine learning analysis workflow and use cases in addiction research are a good starting point for a hands-on approach to machine learning in addiction research. In conclusion, machine learning encompasses several useful tools that need to be in the addiction researchers' toolbox. Some of them are well-known already, such as logistic regression, yet others will expand the capabilities in addiction research. We hope this text encourages addiction researchers to consider using machine learning tools in current and upcoming studies.

Maadhav Kothuri: Helpful to know use cases and different options before diving into a hands-on use of ML

References

1. Che Z, St Sauver J, Liu H, Liu Y. Deep Learning Solutions for Classifying Patients on Opioid Use. AMIA Annu Symp Proc AMIA Symp. 2018;2017:525–34.
2. Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. Niaura R, editor. PLOS ONE. 2017 Apr 10;12(4):e0175383.
3. Sakoglu U, Mete M, Esquivel J, Rubia K, Briggs R, Adinoff B. Classification of cocaine-dependent participants with dynamic functional connectivity from functional magnetic resonance imaging data. J Neurosci Res. 2019 Jul;97(7):790–803.
4. Jing Y, Hu Z, Fan P, Xue Y, Wang L, Tarter RE, et al. Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder. Drug Alcohol Depend. 2020 Jan;206:1–6.
5. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet. 2019 Feb;51(2):237–44.
6. Mak KK, Lee K, Park C. Applications of machine learning in addiction studies: A systematic review. Psychiatry Res. 2019 May;275:53–60.
7. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019 Apr 4;380(14):1347–58.
8. Cresta Morgado P, Carusso M, Alonso Alemany L, Acion L. Practical foundations of machine learning for addiction research. Part II. Workflow and use cases. Am J Drug Alcohol Abuse. Forthcoming.
9. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018 Apr 3;319(13):1317.
10. Mitchell TM. Machine Learning. New York: McGraw-Hill; 1997. 414 p. (McGraw-Hill series in computer science).
11. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning [Internet]. New York, NY: Springer New York; 2013 [cited 2020 Jun 10]. 426 p. (Springer Texts in Statistics; vol. 103). Available from: http://link.springer.com/10.1007/978-1-4614-7138-7
12. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. Am J Epidemiol. 2019 Oct 21;188(12):2222–39.
13. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria.: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/
14. Van Rossum G, Drake Jr FL. Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam; 1995.
15. Barenholtz E, Fitzgerald ND, Hahn WE. Machine-learning approaches to substance-abuse research: emerging trends and their implications. Curr Opin Psychiatry. 2020 Jul;33(4):334–42.
16. Gowin JL, Manza P, Ramchandani VA, Volkow ND. Neuropsychosocial markers of binge drinking in young adults [published online ahead of print, 2020 May 12]. Mol Psychiatry [Internet]. 2020 May 12 [cited 2020 Jun 10];10.1038/s41380-020-0771-z. Available from: http://www.nature.com/articles/s41380-020-0771-z

17. Dong X, Rashidian S, Wang Y, Hajagos J, Kong J, Saltz M, et al. Machine Learning Based Opioid Overdose Prediction Using Electronic Health Records. AMIA Annu Symp Proc. 2020 Mar 4;2019:389–98.
18. Locke GW, Shilkret R, Everett JE, Petry NM. Interpersonal Guilt and Substance Use in College Students. Subst Abuse. 2015 Jan 2;36(1):113–8.
19. Gattamorta KA, Mena MP, Ainsley JB, Santisteban DA. The Comorbidity of Psychiatric and Substance Use Disorders Among Hispanic Adolescents. J Dual Diagn. 2017 Oct 2;13(4):254–63.
20. Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N. Ten Simple Rules for Effective Statistical Practice. Lewitter F, editor. PLOS Comput Biol. 2016 Jun 9;12(6):e1004961.
21. Locke GW, Shilkret R, Everett JE, Petry NM. Interpersonal Guilt and Substance Use in College Students. Subst Abuse. 2015;36:113–8.
22. Morozova O, Levina O, Uusküla A, Heimer R. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC Med Res Methodol. 2015 Dec;15(1):71.
23. Linden-Carmichael AN, Russell MA, Lanza ST. Flexibly modeling alcohol use disorder risk: How many drinks should we count? Psychol Addict Behav. 2019 Feb;33(1):50–7.
24. Chilcoat HD, Schütz CG. Age-specific patterns of hallucinogen use in the US population: an analysis using generalized additive models. Drug Alcohol Depend. 1996 Dec;43(3):143–53.
25. Hosmer D, Lemeshow S. Applied Logistic Regression. Second. John Wiley & Sons, Inc; 2000. (Wiley Series In Probability and Statistics).
26. Sears ES, Anthony JC. Artificial Neural Networks for Adolescent Marijuana Use and Clinical Features of Marijuana Dependence. Subst Use Misuse. 2004 Jan 1;39(1):107–34.
27. Kornfield R, Sarma PK, Shah DV, McTavish F, Landucci G, Pe-Romashko K, et al. Detecting Recovery Problems Just in Time: Application of Automated Linguistic Analysis and Supervised Machine Learning to an Online Substance Abuse Forum. J Med Internet Res. 2018 Jun 12;20(6):e10136.
28. Mete M, Sakoglu U, Spence JS, Devous MD, Harris TS, Adinoff B. Successful classification of cocaine dependence using brain imaging: a generalizable machine learning approach. BMC Bioinformatics. 2016 Oct;17(S13):357.
29. Mumtaz W, Saad MN b M, Kamel N, Ali SSA, Malik AS. An EEG-based functional connectivity measure for automatic detection of alcohol use disorder. Artif Intell Med. 2018 Jan;84:79–89.
30. Russek E, Kronmal RA, Fisher LD. The effect of assuming independence in applying Bayes' Theorem to risk estimation and classification in diagnosis. Comput Biomed Res. 1983 Dec;16(6):537–52.
31. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967 Jan;13(1):21–7.
32. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
33. Squeglia LM, Ball TM, Jacobus J, Brumback T, McKenna BS, Nguyen-Louie TT, et al. Neural Predictors of Initiating Alcohol Use During Adolescence. Am J Psychiatry. 2017 Feb;174(2):172–85.
34. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. California, USA: O'Reilly Media, Inc; 2019. 851 p.
35. Lo-Ciganic W-H, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. JAMA Netw Open. 2019 Mar 22;2(3):e190968.
36. Chollet F, Allaire JJ. Deep learning with R. Shelter Island, NY: Manning Publications Co; 2018. 335 p.
37. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [Internet]. [cited 2020 Nov 25]. Available from: rXiv:1409.1556v6 [cs.CV]

38. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. [cited 2020 May 11]. Available from: arXiv:1810.04805 [cs.CL]
39. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners [Internet]. [cited 2020 Jan 12]. Available from: https://www.bibsonomy.org/bibtex/ce8168300081d74707849ed488e2a458#export
40. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Vigués M, Pujol-Ribera E, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. BMC Fam Pract. 2018 Dec;19(1):108.
41. Sun J, Bi J, Chan G, Oslin D, Farrer L, Gelernter J, et al. Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors. Addict Behav. 2012 Oct;37(10):1138–44.
42. Graves RL, Tufts C, Meisel ZF, Polsky D, Ungar L, Merchant RM. Opioid Discussion in the Twittersphere. Subst Use Misuse. 2018 Nov 10;53(13):2132–9.
43. Blei DM, Ng A, Jordan M. Latent Dirichlet Allocation. J Mach Learn Res. 2003;3:993–1022.
44. Lloret-Segura S, Ferreres-Traver A, Hernández-Baeza A, Tomás-Marco I. El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. An Psicol. 2014 Oct 1;30(3):1151–69.
45. Jolliffe IT. Principal Component Analysis. Second. UK: Springer; 2002. (Springer Series in Statistics).
46. Yoshua B, Réjean D, Pascal V, Christian J. A Neural Probabilistic Language Model. J Mach Learn Res. 2003;3:1137–55.
47. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: NIPS 2013 [Internet]. 2013. p. 3111–9. Available from: https://dblp.org/rec/conf/nips/MikolovSCCD13
48. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, et al. DeViSE: A Deep Visual-Semantic Embedding Model. In: NIPS [Internet]. 2013. p. 2121–9. Available from: http://dblp.uni-trier.de/db/conf/nips/nips2013.html#FromeCSBDRM13
49. The Turing Way Community, Becky A, Bowler L, Gibson S, Herterich P, Higman R, et al. The Turing Way: A Handbook for Reproducible Data Science (Version v0.0.4) [Internet]. Zenodo. 2019 [cited 2020 Dec 1].
Available from: http://doi.org/10.5281/zenodo.3233986
50. Hastings JS, Howison M, Inman SE. Predicting high-risk opioid prescriptions before they are given. Proc Natl Acad Sci. 2020 Jan 28;117(4):1917–23.
51. Katardjiev N, McKeever S, Hamfelt A. A machine learning-based approach toforecasting alcoholic relapses. In: ITISE 2019 (6th International conference on Time Series and Forecasting). Granada, Spain; 2019.
52. Wang JM, Zhu L, Brown VM, De La Garza R, Newton T, King-Casas B, et al. In Cocaine Dependence, Neural Prediction Errors During Loss Avoidance Are Increased With Cocaine Deprivation and Predict Drug Use. Biol Psychiatry Cogn Neurosci Neuroimaging. 2019 Mar;4(3):291–9.
53. Baker TE, Zeighami Y, Dagher A, Holroyd CB. Smoking Decisions: Altered Reinforcement Learning Signals Induced by Nicotine State. Nicotine Tob Res. 2018 Jun 30;22(2):164–71.

Machine Learning Methods implemented in R and Python

This section is a primer showing a few of many existing open-source software readily available to apply machine learning methods. It is not intended as a thorough guide. For further information about R packages and how to use them, visit the rdocumentation.org site. Helpful tips: start with Google, adding "R" or "Python" to your question; if you still find nothing you also can upload a question with an example. Try with stackexchange.com and stackoverflow.com to search existing questions or to ask new ones tagging your questions with "R" or "Python". When no "library()" statement is listed for R, the functions listed are included with R's initial installation. When functions are listed after an R library, it refers to some of the main (many times out of many) functions within the library that are relevant to each analytic technique. In Python, most implementations are found within the scikit-learn library, with the exception of deep learning models, which may be found in the Keras or Pytorch packages.

Supervised Methods

Linear Regression

Python: sklearn.linear_model: LinearRegression
R: lm(y x, ...)

Ridge Regression

Python: sklearn.linear_model: Ridge
R: $library(glmnet), glmnet(x, y, alpha = 0, ...), cv.glmnet(x[train,], y[train,], alpha = 0)$

Lasso Regression

Python: sklearn.linear_model: Lasso
R: $library(glmnet), glmnet(..., alpha = 1, ...), cv.glmnet()$

Polynomial Regression

Python: sklearn.preprocessing: PolynomialFeatures + linear model
R: $lm(y\ poly(...), ...)$

Splines

Python: scipy.interpolate: UnivariateSpline
R: $library(splines), lm(y\ bs(...), ...) or lm(y\ ns(...), ...)$

Smoothing Splines

Python: statsmodels.gam.api: GLMGam
R: $library(gam), gam(y\ ...) or library(mgcv)$

Generative Additive Models

Python: statsmodels.gam.api: GLMGam
R: $library(gam), gam(y\ \dots) or library(mgcv)$

Decision Trees/Random Forests

Python: sklearn.tree: DecisionTreeRegressor sklearn.ensemble: RandomForestRegressor sklearn.tree: DecisionTreeClassifier sklearn.ensemble: RandomForestClassifier
R: $library(randomForest), randomForest(y\ x, ...)$

Gradient Boosting

Python: XGBoost
R: $library(xgboost), xgboost(...)$

K-nn

Python: sklearn.neighbors: KNeighborsRegressor
R: $library(class), knn(train.x, test.x, train.y, k = ...)$

Logistic Regression

Python: sklearn.linear_model: LogisticRegression
R: $glm(y\ x, ..., family = binomial)$

Naive Bayes Algorithms

Python: MultinomialNB, GaussianNB
R: $library(naivebayes), naive_bayes(y\ x, ...)$

Support Vector Machines

Python: SVC
R: $library(e1071), svm(y\ x, ...)$

Artificial Neural Networks

Python: pytorch, keras
R: $library(keras)$, various functions within the library

Unsupervised Methods

k-means

Python: sklearn.cluster: KMeans
R: $kmeans(x.matrix, centers, nstart = ..., ...)$

Principal Component Analysis

Python: sklearn.decomposition: PCA
R: $library(mixOmics), pca(x.matrix, ...)$

Latent Dirichlet Allocation

Python: sklearn.decomposition: LatentDirichletAllocation
R: $R = library(topicmodels), LDA(dataset, k = ..., ...)$

Box 1: Machine learning explained through its keywords

Learning - Broadly, learning involves acquiring general concepts from particular instances. An algorithm learns when it improves its performance at solving a given task. To learn, an algorithm extracts information from the experience coming through examples (e.g., outcomes and with their characteristics as organized in a dataset). An algorithm solves a given task using the information extracted and obtains a measure of its performance at that moment of the learning process. An algorithm captures this learning in parameters or weights of a model of the data from which the information is extracted. Frequently, this procedure occurs iteratively, in alternative steps of measuring the performance at each step and adjusting the parameters.

Algorithm - The function or computational process which allows to learn from the data. It is often a synonym for technique or method. For instance: "We applied three machine learning algorithms: random forest, deep neural networks, and a support vector machine".

Training - This is the process through which the algorithm learns from data. As a result, the parameters or weights are calculated and a model is fitted.

Parameters, coefficients, weights - They capture what the algorithm learned from the data. Finding the weights that reduce error the most is the result of the learning process.

Model - A model is a function that provides a variable as a combination of the describing variables of an entity (e.g., object, event). It is the result of a machine learning process, applying a particular algorithm to a specific dataset.

Dependent variable, response variable, or target - It is the variable of interest, the outcome, corresponds to what one wants to predict or classify. It is typically denoted by the letter y. For example, y could be that a participant has a SUD.

Independent variable, predictor variable, feature, attribute - Variable(s) used by the algorithm to predict or classify. Each instance to be processed is characterized or represented by these variables. These attributes are typically denoted using the letter x, with numbered subscripts when there is more than one attribute. (i.e., an instance with two attributes would have x1 and x2 as attributes). For instance: when we seek to predict the presence or absence of a SUD in a participant (y), the attributes gender (x1) and age (x2) are relevant.

Model performance - It is a measurement of how well a model solves the task of interest, for example, how well the model can classify participants with and without SUD. Performance can be measured with several metrics that may also be useful at comparing models.

Bias-variance trade-off - Variance refers to the sensitivity of a model to small data changes. In a high variance scenario, model performance changes a lot with different data, even if the data come from the same population. Bias is a systematic error, which is the result of solving a complex real-life problem with a simpler model. There is a trade-off between variance and bias. More flexible models have higher variance and lower bias than simpler ones, and vice versa. This trade-off is the core of most of the limitations of machine learning techniques.

Overfitting - This is an undesired phenomenon produced by a model, when it follows the errors, or noise in the data (rather than the signal or information in the data), too closely. Overfitting occurs when a fitted model follows every detail of a dataset, the algorithm learns anecdotal information building a model with an excellent performance on that dataset but with poor performance on unseen data. Overfitting is associated with high model variance.

Inference - Models can be used for predicting and understanding dependent variables. When seeking to understand how predictor variables are associated with the outcome,

the magnitude of some model parameters can be assessed through statistical inference (also known as statistical hypothesis testing).

Clustering, cluster analysis - It is a set of learning methods aimed to make groups with observations that share similar characteristics. These techniques find feature patterns, which are given by a set of related values of the variables analyzed. These patterns generate clusters that group similar observations.

Ensemble methods - These are tools that combine several techniques to improve the final model performance. Ensembles often outperform each method separately, can capture more complex characteristics and may decrease overfitting. Examples include random forests, bagging, and gradient boosting.

Regularization - It is a technique used to prevent overfitting by imposing additional constraints to a model, usually in the presence of a relatively high number of predictors, that leads to some type of parameter shrinkage approach to fit a model. Lasso and ridge regression are two methods that apply regularization.

Cross-validation - It is a group of techniques that split the training dataset, leaving out a subset to evaluate the model fitted using the data left out during training. Cross-validation implements these steps (splitting, leaving out, and evaluating in the left-out data) repeatedly, allowing us to fit the model and validate it using different subsets. Considering the need of evaluating the model with unseen observations, unlike simpler techniques to split datasets, cross-validation evaluates models by maximizing the use of the whole dataset. There are several techniques, such as k-fold and or leave-one-out cross-validation.

Box 2: Specific domains within machine learning

Image analysis involves specific challenges besides proficient domain knowledge, such as mastery of highly specialized techniques, the use of specific software for pre-processing, post-processing and the modeling itself, the high dimensionality problem (hundreds of thousands of features per image), and having to consider time in the case of functional imaging. Mete et al (28) analyzed 162 neuroimages from two groups (participants with cocaine use disorder and participants without it). Considering that each image had 517,845 voxels (a volumetric pixel, the minimal unit in a 3D image), the authors used several strategies towards dimensionality reduction including both neuroanatomic considerations and machine learning methods. After obtaining high classification performance, relevant brain regions that differed between both groups were explored to interpret their medical significance.

Natural language processing frameworks can process individuals' health records, including free text, and detect relevant categories from thousands of diagnoses and procedures (50). Natural language processing is also used for data collected from social media such as the analysis of Twitter messages and their association with overdose death rate studied by Graves et al (42). Natural language processing also involves very specific techniques and complex pre-processing.

Time series is another area with extensive specific techniques often used in high-level solutions to discover relevant, paradigmatic, or anomalous patterns in data series over time. For example, time series analysis helped forecast alcohol consumption relapses (51).

In some cases, a problem that evolves can be more adequately modeled as a **reinforcement learning** problem if there are many observable states but only some of them can be interpreted, such as cigarette smoking or cocaine use disorder (52,53). The main difficulty of using reinforcement learning is the complexity of the targeted model. Defining a model to be learned requires a deep analysis of the problem and experience in formalizing problems into the reinforcement learning analytical framework.