Yuzhu Xia
CSASS Certificate Paper

# A brief introduction on machine learning as a potential solution in addressing methodological issues in educational research with large datasets

## *Introduction*

With the development and availability of large educational datasets in the United States and across the world, big data analysis (Daniel, 2015, 2019) has been gaining attention from various disciplines, including education. The term big data generally refers to datasets with massive size that is beyond the ability of common tools (Buyya, Vecchiola, & Selvi, 2013). Daniel (2019) proposed Data Science as the fourth research tradition/methodology that started to take shape in the 2000s. It stimulates new approaches to framing research questions, designing studies, and analyzing data in educational research. Data science is a field that is primarily concerned with extracting information from complex data and utilizing that information to create knowledge through the development and use of tools in computational platforms. The strategic combination of data science and educational research makes it possible for educational researchers to make further use of available large datasets to explore patterns and make predictions about many educational aspects such as administration (e.g. resource allocation, policy making, *etc.*), student learning (e.g. learning environment, learning outcomes, student retention, *etc.*), and teachers/delivery (e.g. quality of instruction, effect of instruction on student learning, *etc*.) (Daniel, 2015, 2019; McKenney & Mor, 2015).

Extensive literature has examined student learning/performance and its related factors in the past decades (Darling-Hammond, 2002; Greenberg, Rhodes, Ye & Stancavage, 2004; Greenwald, Hedges & Laine, 1996), and specifically, literature that investigates student learning/performance using large datasets has emerged and developed in the past thirty years worldwide, though the majority remain in ~~the~~ western countries (Daniel, 2019; Baker & Yacef, 2009; Romero & Ventura, 2010). As a relatively new field, the exploration and analysis of big data in scientific research has been growing fast with many potential issues, especially in educational research, partly due to improper handling of large datasets, the high statistical/computational threshold of big data analytics for education scholars, and conflicting worldviews (Baker & Yacef, 2009; Kline, 1994; Romero & Ventura, 2010; Woltman, Feldstain, MacKay, & Rocchi, 2012).

Machine learning, as one of the fastest growing bodies of knowledge in handling large datasets, is an option that has not been sufficiently studied and utilized by educational researchers. This paper aims to briefly analyze how machine learning could be used in educational research and how it can answer educational questions that cannot be properly addressed by traditional statistical modeling methods. In the sections below, I first present some of the common issues in analyzing large educational datasets, then explore each of the problems and see if machine learning is a good alternative. Specifically, I present how machine learning is a better tool in generating evolving models in predicting student performance within nested large datasets. I will be using the terms large datasets, large enough datasets, nationally representative datasets, and big data interchangeably to refer to datasets that are large enough for machine learning to perform well in this paper, though they have slightly different emphasis in their definitions.

## ***Issues in Using Large Educational Datasets***

### *1. Missing Data Handling*

One of the common issues in current educational research using large datasets is handling missing data. Missing data is common in educational research, especially in survey-based research, due to noncoverage, partial nonresponse, total nonresponse, and item nonresponse (Brick & Kalton, 1996; Cheema, 2014; Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2004). How missing data is handled can have different degrees of influence on data analysis and results depending on the amount of missing data and on the reason why data is missing (e.g., if the data missing is at random or are there other non-random issues). Not able to handle missing data appropriately has been an issue documented by many (Bodner, 2006; Enders, 2010; Peugh & Enders, 2004), and the reasons vary, including not having sufficient instruction and training, negative attitudes toward statistics, and the inability of linking theory to practice (Cheema, 2014; Murtonen & Lehtinen, 2003).

In general, there are two categories of methods for handling missing data. One focuses on the deletion of missing data/samples, and the other focuses on mathematical/statistical imputation methods to generate values to replace the missing data. A large number of educational research that used large educational datasets has been implementing deletion methods (listwise deletion or pairwise deletion, depending on the degree to which variables are deleted), but more scholars are moving away from these methods because of bias inherent in such replacement approaches. This is problematic for several reasons. First, the sample size will be decreased. Deleting cases when one or more missing values exist in any given variable simply decreases the sample size, and if there is a large portion of missing values in a given dataset, researchers might lose the precision and power of statistical testing of the remaining cases. Second, the generalizability will be changed (Little 1988; Little & Rubin, 1989). When cases are being deleted from a sample, researchers would not be able to generalize their results to the target population, because the current sample after deletion no longer represents the original sample population.

Some educational researchers also use the various imputation methods in handling missing data (e.g., mean imputation, maximum likelihood expectation-maximization imputation, multiple imputation, regression imputation, single random imputation, and zero imputation) (e.g., Bradley, 2012; Olinsky, Chen, & Harlow, 2003; Pigott, 2001). In general, imputation methods are better ways than the deletion methods to handle missing data, because by using imputation methods, we are not decreasing the sample size, and therefore we are not losing the ability to generalize the results to a larger population (Anderson, Basilevsky, & Hum, 1983; King, Honaker, Joseph, & Scheve, 2001; Little, 1988; Pampaka, Hutcheson & Williams, 2016).

However, there is a significant difference when we impute different values to replace the missing values. For example, in mean imputation, researchers use the mean of the nonmissing values for that variable to replace the missing value. There are other imputation methods in this type of situation, such as replacing the missing value with the median or the mode of the other nonmissing values for that variable (Cheema, 2014). This method has also been used in education research by many, however, if these measures are biased, the bias will be extended to missing cases after replacement. Additionally, it increases the risk of the possibility to wrongly reject the null hypothesis. This can have significant consequences if results are wrongly concluded due to mean imputation.

Multiple imputation is a better way than mean/median/mode imputation when the data are missing at random (MAR) or missing completely at random (MCAR) (King, Honaker, Joseph, & Scheve, 2001; Pampaka, Hutcheson & Williams, 2016). Multiple imputation is iterative in a way that the distribution of the existing data is used to estimate multiple values that reflect the unknown true value. Multiple imputation is a more ideal way to handle missing data, because the logic of multiple imputation is that first, it generates multiple hypothetical datasets with the estimated values and fitting models to reflect the different variations; then, multiple imputation averages across the many sets of values generated by the many rounds of estimating; and finally, it generates unbiased estimates of missing values simulating the natural variation in missing data (Cheema, 2014; Little, 1988).

Extensive literature has documented the advantages and disadvantages of using each method compared with others, dating back to the 1960s (Afifi & Elashoff, 1966; Haitovsky, 1968). Young, Weckman, and Holland (2011) provided a few cut-off percentages of missing data for researchers to consider which missing data handling method to use. Missing data handling methods do not make a great impact on results if missing data is less than 1%; and for less than 5% of missing data, simple methods can be used, for example deletion methods and less advanced imputation methods; for less than 15% of missing data, advanced imputation methods should be used, such as multiple imputation (Young et al., 2011). Pampaka et al. (2016) provided a list of available procedures in *R* to perform multiple imputation, including but are not limited to: Amelia II (Honaker, King, & Blackwell, 2011), cat for categorical-variable data sets with missing values (Schafer, 1997), and impute (Hastie et al., 2014). Please refer to Rubin (1987), Little (1986), Peugh and Enders (2004), and Kim (2004) for more detailed statistical calculation, rationale, and examples for multiple imputation.

In selecting missing data handling methods, a large number of educational researchers use deletion methods due to the ease of application, and it is the default method in SPSS, a software that many, if not most, educational researchers use. Only a small proportion of studies used advanced missing data handling methods such as multiple imputation (Enders, 2010; Peugh & Enders, 2004).

### *2. Modeling and Assumptions*

Another issue in educational research using large datasets is the modeling methods, including addressing violations of assumptions. Education research, like any discipline of scientific inquiry, partly relies on the ability of interpretation and generalizability of sample findings (Cunnings, 2012). During the past century, quantitative methods has played an important role in examining student performance, however a large amount of literature limits its research methods to somewhat basic quantitative measures, despite the advancement in statistical techniques developed in statistics, such as mixed-effects modeling, or multilevel modeling, and other advanced data analysis tools (e.g. Baayen et al., 2008; Quene & van den Bergh, 2008).

Mixed models, including multilevel modeling, or hierarchical linear modeling (HLM), are statistical models of parameters that vary at more than one level (Raudenbush & Bryk, 2002). Mixed models are an extension of simple linear models, which is a statistical method that allows researchers to study the relationships between two quantitative variables, to allow for both fixed factors and random factors, especially when there is non independence in the data, for example, in hierarchical structures, such as classrooms and schools. It is also necessary to adjust standard errors resulting from data clustering. Mixed modeling has been regarded as an advanced and ideal approach to generate models in large and nested datasets to represent relationships among variables and make predictions. Education data are sometimes collected and presented in a hierarchical fashion of individual students within classrooms, within schools, and within school districts (Woltman, Feldstain, MacKay, & Rocchi, 2012). Mixed models have been demonstrated to work well for hierarchical data because they are better models than simply aggregating or analyzing each higher level of unit separately, which fail to take full advantage of available data.

Mixed-effects models have specific benefits especially when compared with simple linear modeling, one of which is that a mixed-effects model can involve multiple independent variables of interest, including categorical variables, continuous variables, or the mixture of the two (Cunnings, 2012). Another major benefit is that a mixed-effects model can capture the changes reflected by the data over time, not having to assume the change to be linear (Ortega & Byrnes, 2008; Ortega & Iberri-Shea, 2005).

However, even though mixed models are more advanced than simple linear models and are an effective approach for nested large educational datasets, there are several major limitations. First, it is complex to implement and time consuming to conduct (Woltman, Feldstainm MacKay, & Rocchi, 2012). More importantly, mixed models rely on quite a few assumptions, including normal distribution, linearity, and homoscedasticity, just to name a few important and common assumptions. Violations of such assumptions may result in inaccuracy of parameter estimates and standard errors, which will affect the interpretation of data findings. It should be noted that the true relationships in any large datasets may or may not be linear, and we cannot guarantee that the datasets we are using, completely satisfy those assumptions. Slight violations of those assumptions may not be a huge issue when the dataset is large enough, and performing mixed modeling may still be an acceptable approach, but quantitative researchers cannot say mixed modeling is the perfect solution.

### *3. After Modeling*

For the majority of empirical studies in educational research, after generating a model with adequate R-squared and statistical significance, the only job left is to discuss the results and connect with literature. The model remains unchanged. Though researchers may try multiple models with different predictors to obtain their “best fitted” model to explain some phenomenon, once they have that model, that model remains unchallenged. The model obtained may seem to be the best model among the many traditional statistical models researchers try out; however, we as researchers will never know if our “best” model truly represents the real trend, or if our model is the optimal model.

Influential advancements in quantitative research methodologies have been focusing on the methodological advantages of certain techniques, for example, in handling missing data or generating models (e.g. Grimm, Mazza, & Mazzocco, 2016; Rehmtulla & Hancock, 2016). However, what could or needs to be done to improve the generated models is usually beyond the scope of the studies.

Norris, Plonsky, Ross, and Schoonen (2015) specified the guidelines for reporting quantitative methods and results in language learning research, and elaborated in detail the sections and contents that a good quality quantitative study should include, such as study population, sampling, participants in the sample, measurement, design, procedures, and analysis. However, nothing was mentioned regarding perfecting the models or any potential concerns regarding the generated model. In the paper discussing the potential issues with mathematical modeling in educational research, Varaki and Earl (2006) pointed out three problems: (1) if the educational issue under study is in an open system or closed system (most education settings are open systems); (2) the relationship between association and causation; and (3) the neglect of the intentional dimension of educational life. The first issue is most prominent and has the greatest impacts, because in open systems like educational settings, agents operate not as consistently as closed systems do, and there are almost always agents with power and capabilities that not only reside in the individual level, but also institutional level that can influence the relationships among them. Existing modeling methods of large datasets handle the relationship between variables as static numbers. However, this relationship may be fluid rather than static. Machine learning, on the other hand, learns about the relationship between variables and allows the relationship to be fluid or static as a reflection of the data. Machine learning can take place with or without labels of input information, and therefore, simulate the human brain's neural network in digesting the information, and generates models that best capture characteristics of all layers to get the optimal model.

Book chapters introducing quantitative research methods in educational research also either focus on the distinction between qualitative and quantitative methodology, or the selection and use of statistical tools for different purposes (e.g. Creswell & Creswell, 2014; Sheard, 2018), but still leave what happens after generating a “best fit” model undiscovered.

Machine learning could aid in this matter. I will elaborate on this with more details in the following section. I will briefly introduce machine learning, including types of machine learning, applications and uses in the sciences and social sciences, and its advantages and disadvantages, and then discuss how machine learning might be a good alternative to traditional statistical modeling methods in addressing the above issues in current educational research.

## ***Machine Learning***

### *Context and Definitions*

Machine learning has been applied in certain fields for almost 30 years since its start in the 1950s, though it remains understudied in many disciplines, including education. Samuel (1959), the first scholar who defined the term, described machine learning as the field of study where computers can learn without being explicitly programmed. About two decades later, Mitchell (1967) described machine learning from an engineering-oriented perspective as, “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E” (p. 2).

One of the most well known applications of machine learning, as described by Géron (2017), is spam filtering in our mailboxes. Géron (2017) pointed out that with traditional detection algorithms, programmers would need to write separate algorithms for each of the spam-like words that people can think of, such as “4U”, “free”, and “credit card”. Programmers would need to repeat this process until they believed they had come up with a somewhat exhaustive list of spam-like words and then run the long lists of complex algorithms, which can be really hard to achieve and maintain. By contrast, machine learning may be able to detect which words are good predictors of spam emails by detecting existing spam emails day by day, and the program is shorter and easier to maintain. Géron (2017) presented the following two figures as a comparison of how the processes are different:

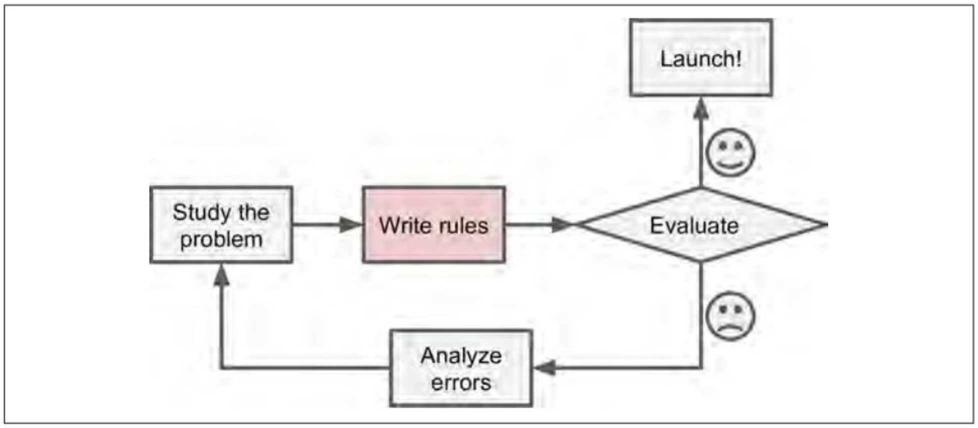


*Figure 1. The traditional approach (p. 5)*

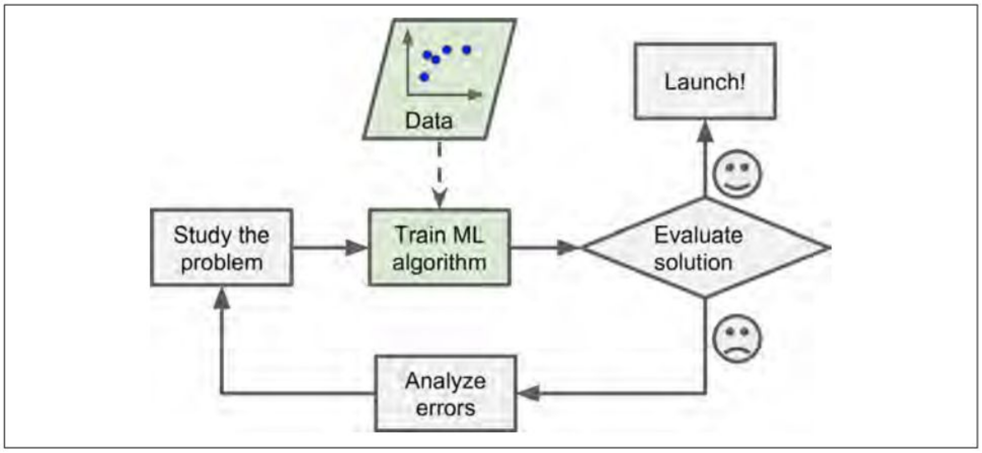


*Figure 2. Machine Learning approach (p. 5)*

In the traditional approach, programmers would have to repeat the cycle of the lower part of the figure until they have removed all bugs and errors before launching; in the machine learning approach, programmers would feed training data into the system to train the algorithm to learn from the data, and this step would save time and efforts of human power to complete the cycle before launching.

Machine learning has also been applied in more advanced and complex fields, such as voice recognition and humans learning in information technology, cancer research in medical and pharmaceutical research, and the social sciences as well, including but are not limited to studying networks, structures, and relationships in sociology and psychology, and learning and teaching in education, just to name a few.

## *Types of Machine Learning*

Some key words in differentiating types of machine learning include supervision, online/offline learning, and detecting patterns (Bishop, 2006; Goodfellow, Bengio, & Courville, 2016; Géron, 2017; Hastie, Tibshirani, & Friedman, 2009; Russell & Norvig, 2002). According to Géron (2017), machine learning systems can be categorized into three broad types, supervised/unsupervised learning; batch/online learning; and instance-based/model-based learning.

### *Supervised/Unsupervised Learning*

The amount and type of supervision systems receive determine if the machine learning process is supervised or not. In a supervised learning environment, researchers feed labels to the system, such as the spam-like words in the email filtering technique, or predictors, such as features of a car to predict the price of a car. In unsupervised learning, the system learns without any given labels or predictors through various algorithms (e.g. clustering, association rule learning). Semisupervised learning lies in between the above two, usually with a small portion of labeled data and the rest are unlabeled data.

There is another subtype in this category called Reinforcement Learning, such as robot implementation. In a reinforcement learning environment, the system, or an agent (Géron, 2017), learns by observation and trial and error. The agent makes progress by getting rewards when making good decisions and penalties when choosing bad strategies, or policies, which “defines what action the agent should choose when it is in a given situation” (p. 13).

*Batch/Online Learning*

Whether the learning happens offline with all available data or online as new data come determines if the process is batch learning or online learning (Géron, 2017). In batch learning, or offline learning, all available data will be fed to the system, and once the system is trained, it concludes its learning process. In contrast, in an online learning environment, the system learns as new data comes. Understandably, online learning depends on the amount and quality of new data. A potential solution could be human supervision, where researchers can switch off the learning process when a noticeable decline in data quality or learning performance appears.

### *Instance-Based/Model-Based Learning*

As suggested by the two names, the distinction between the two subtypes is whether the system learns by case or learns by models generated from multiple cases. In instance-based learning, the system learns by each case (Géron, 2017). Take the same spam filter example, when the system learns by instance-based algorithms, it measures the similarity between a stored email and the new incoming email, and decides if the new email is spam or not; when the system learns by model-based algorithms, it generalizes from existing emails to build a model first, then uses the model to predict if new emails are spam or not.

## *Applications and Uses of Machine Learning*

In this section I will present some representative applications and empirical studies in the hard sciences and social sciences below as an overview of the development of the field, and discuss what is needed in the field of education, especially when scholars are handling large datasets.

### *Application of Machine Learning in the Sciences*

Machine learning, as one of the most popular applications of artificial intelligence (AI), has been applied in the sciences for the past few decades, especially in the recent thirty years. We are surrounded by different kinds of machine learning applications such as personal assistants (e.g. Siri, Alexa, and Google Now) and recommendation systems (e.g. Amazon and social media suggestions). Researchers have been utilizing machine learning to advance their knowledge across disciplines (Langley & Simon, 1995), especially in medical and pharmaceutical fields in cancer prediction, prognosis and diagnosis, and drug development (e.g. Cruz & Wishart, 2006; Guan, Zhang, Quang, Wang, Parker, Pappas, Kremer, & Zhu, 2019; Negi & Mathew, 2018; Puri, 2019; Rahman, Ali, Altwijri, Alqahtani, Ahmed, & Ahamed, 2019); in information technology and engineering in optimization, virtualization, and back-end and front-end user interface development (e.g. Hong, Razaviyayn, Luo, & Pang, 2015; Ismael, Song, Ha, Gilbert, & Xue, 2017; Kundu, Rangaswami, Gulati, Zhao, & Dutta, 2012; Shabtai, Fledel, & Elovici, 2010); and disciplines such as mathematics, statistics, and environmental science, *etc*. (e.g. Kim & Boyd, 2008; Recknagel, 2001). Various disciplines in the sciences have benefited from the applications of machine learning, if not all. However, machine learning has not been widely accepted or used in the social sciences, especially in educational research.

### *Application of Machine Learning in Social Sciences*

Machine learning has been a useful tool for the social sciences, though not as significantly or broadly utilized as for the hard sciences. The few fields utilizing machine learning in the social sciences include economics, in the building of models in general and specifically in econometrics  (e.g. Athey, 2018; Chalfin, Danieli, Hillis, Jelveh, Luca, Ludwig, & Mullainathan, 2016; Mullainathan & Spiess, 2017); social media and communication in facial recognition and user interface development (e.g. Galán-García, Puerta, Gómez, Santos, & Bringas, 2016; Ge & Qiu, 2011; Hamilton, Ying, & Leskovec, 2018; van Zoonen, & Toni, 2016), public policy and sociology in generating models for policy prediction and theory development, (e.g. Burnap & Williams, 2015; Samii, Paler, & Daly, 2016; Toch, Lerner, Ben-Zion, & Ben-Gal, 2019); psychology and anthropology in building models to analyze and interpret structures, networks, and relationships (e.g. Cunningham, 1996; Valletta, Torney, Kings, Thornton, & Madden, 2017; Yarkoni & Westfall, 2017); and education (Acharya & Sinha, 2014; Berland, Baker, & Blikstein, 2014; Kotsiantis, 2012). Kučak, Juričić, and Đambić (2018) examined 67 existing studies that are representative in the field of education, which used machine learning as their methods, and summarized four categories of most popular uses: grading students, improving student retention, predicting student performance, and testing students. They further concluded that the application of machine learning in educational research could be beneficial in many different ways.

The application of machine learning in education has been limited in the fields related to sub-areas of student performance, such as testing and grading, and this application includes a specific approach, educational data mining (EDM), which I will go into more details below. No research has been identified to examine how traditional statistical modeling methods and machine learning can answer educational questions such as student performance differently. With different roots and processing algorithms, traditional statistical modeling methods and machine learning have different strengths in the different stages of data processing and data analysis, and research documenting these differences and contributions to educational data analysis will yield valuable information regarding the various educational research questions and the methodologies approaching the research questions.

*Educational Data Mining (EDM)*

Comparing the applications of machine learning in the social sciences with the sciences, I see a much smaller number of published articles within a somewhat limited scope of fields. If we look at education, even a smaller number of research utilize machine learning in analyzing large datasets, and only a subset of these research use machine learning to predict student academic achievement. For example, an ERIC search with terms “Machine Learning” and “Education” only yields 2,255 results in total, when a search with the term “Student Achievement” yields 132,278 results (less than 2%). Most existing educational research in this realm has been using data mining as a tool to explore patterns in educational data.

Simply put, data mining can function as an information source for machine learning to pull from. Educational data mining (EDM), as defined by Romero and Ventura (2010), is “an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context” that uses “computational approaches to analyze educational data in order to study educational questions” (p. 1). Unlike the limited use of machine learning, education scholars have been using data mining to analyze educational data for the past couple decades, especially after the 2000s (Romero & Ventura, 2010). Some popular trends and topics for EDM researchers include clustering, relationship mining, prediction, distillation of data for human judgement, and discovery with models (Baker & Yacef, 2009; Baker & Inventado, 2016; Romero & Ventura, 2010). Most cited papers using EDM investigate issues such as online courses (Zaïane, 2001), e-learning systems (Tang & McCalla, 2005; Zaïane, 2002), and student model development in their behavior, emotional, and engagement, *etc*. (Beck & Woolf, 2000; Baker & Yacef, 2009). Among the above topics, discovery of models is still an emerging subarea of EDM (Baker & Yacef, 2009), especially when it is strategically combined with prediction. Admittedly, model discovery and generating predictive models require complex skills in writing algorithms, which is the base of machine learning. ~~that machine learning relies on to .~~

So far, only a limited amount of research utilizes machine learning to further explore advanced models and improve the evolving models to predict student performance or knowledge using EDM results, partly due to the fact that large-scale, nationally representative educational datasets only started to appear recently (Baker & Yacef, 2009), and the high threshold of machine learning for education major scholars (Romero & Ventura, 2010).

### *Advantages and Limitations of Machine Learning*

There are many known advantages to machine learning, one of which is the ability to identify trends/patterns and to predict values more easily. Compared with traditional statistical modeling methods, machine learning usually takes less time to identify the underlying structures and patterns that might not be apparent to human eyes. Second, models generated by machine learning can improve over time depending on the quantity and quality of data. Unlike traditional statistical methods where researchers establish models by specifying the equations to answer their research questions, machine learning keeps learning from new data as they come, and therefore generates evolving models as the best fit. Third, machine learning is more effective in handling large and multi-dimensional datasets, which might be difficult for other statistical software to handle. Additionally, machine learning does not assume all relationships among variables are linear, whereas most statistical modeling in educational research assumes linearity, which limits the accuracy and scope of the model and further interpretations.

Although it has been gaining popularity across various disciplines at different levels of application, machine learning also has its challenges and limitations. First, it suffers when there is insufficient quantity of training data (Géron, 2017; Romero & Ventura, 2010). Regardless of types of learning, the system depends on some amount and type of training data to learn before applying what it has learned to new data or situations. It limits the usage of machine learning when there is not enough training data for the machine to learn. In educational research, it is not always easy to obtain large amounts of data for machine learning to perform well. Therefore, even though machine learning shows great advantage over other modeling methods in analyzing large datasets, it may not always be possible to be applied in educational research when there is not sufficient data. A second challenge of machine learning is the quality of data (Géron, 2017; Gudivada, Apon, & Ding, 2017; Sheng, Provost, & Ipeirotis, 2008). When the training data fed into the system is not of good quality, for example, too many errors and noise due to various reasons, it makes it difficult for the system to detect the underlying patterns and subsequently harder to provide accurate predictions. Another issue with machine learning is that researchers need to select representative enough data for the systems to learn. Understandably, when we feed nonrepresentative training data into the system, the model it generates and predictions it makes may fail to represent the true situation that researchers anticipate. In educational research, it is not always possible or appropriate to collect representative enough data depending on the research questions. In cases where researchers are investigating specific groups of individuals that are not representative of their population, machine learning might not be their first choice of data analysis tool.

There are more limitations to machine learning, such as overfitting (when the models learn the training data too well including the errors and noise) or underfitting (when the models do not fit the data well enough and cannot capture the underlying trend of the data). The above three are more general and crucial at the initial stage of research because they involve what and how much data is needed to be fed into the system for the systems to learn. More technical considerations come later.

### *Machine Learning to Address Methodological Issues in Current Research Using Large Datasets*

Machine learning is an application of artificial intelligence that enables systems to learn by itself and improve from the learning experience without explicit commands. It is not a software per se, but a tool that can be applied on software platforms such as *R* or

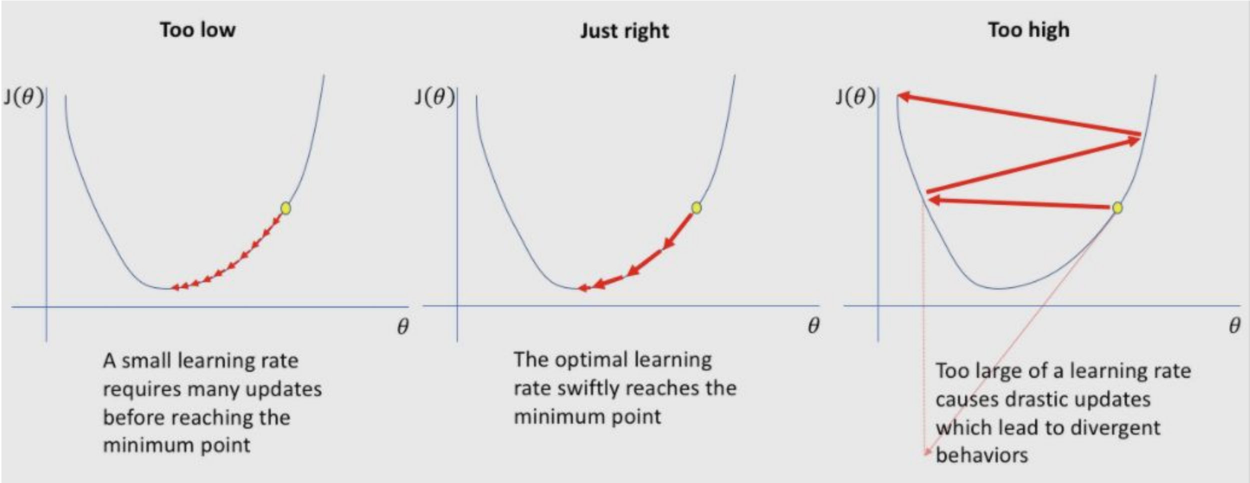
Python. It is designed and best used for big data analysis, especially for generating models and perfecting the models through training/learning. Machine learning may not show clear advantages over traditional statistical softwares in processes such as data cleaning, because packages and functions in *R* can properly perform multiple imputation just as Python, but supervised machine learning performs particularly well in imputing missing values (Molina & Garip, 2019; Sovil, Eirola, Miche, Björk, Nian, Akusok, & Lendasse, 2016). Machine learning is also a good tool to illustrate how statistical imputation methods, multiple imputation in particular, are overall better approaches in handling missing data (Richman, Trafalis, & Adrianto, 2007). Comparing the different approaches in handling missing data with statistical methods and machine learning methods, Jerez, Molina, García-Laencina, Alba, Ribelles, Martín, and Franco (2010) listed the steps and results of various missing data handling methods with mean, standard deviation, and mean squared error (MSE), followed by predictive models generated with the different missing data handling methods applied in the dataset. Given the logic of multiple imputation (using an appropriate model that incorporates random variation in generating the multiple sets of values to replace missing values), machine learning can be a good alternative in the sense that it can generate more accurate models by taking the weights of different layers/dimensions when learning the data, and the neighboring relations between nodes (neurons) into considerations. Jerez et al. (2010) found that overall, machine learning methods (multi-layer perceptron MLP, self-organization maps SOM, and k-nearest neighbors KNN) showed clear advantages in decreasing standard deviations and MSEs when imputing values to handle missing data, and subsequently led to better predictive models.

Machine learning also shows a great advantage in the data analysis processes, including generating models through training and testing, and perfecting the models through tuning parameters, such as the learning rate, how many times the systems learn, and error rate. I will go into more details below.

In traditional statistical modeling, we are performing regression analysis assuming that certain assumptions were true, including but are not limited to normality and linearity. Generally speaking, when the dataset is large enough, observed data tend to be normally distributed, and the relationships among the variables tend to be linear. However, we cannot be entirely sure that it represents the true situation. If the true relationships among the variables from different layers are slightly non-linear, it might be acceptable to use mixed modeling, but it may not be the most ideal approach. If we are using machine learning, we are not assuming any of the standard statistical assumptions such as normality or linearity. Machine learning is designed to simulate how the human brains work, with neurons in neural networks, and it relies on other types of assumptions, not the statistical ones. For example, machine learning relies on manifold learning in semi-supervised learning, but not the statistical assumptions such as random sampling or linearity. Therefore, when we feed data into the system, we are not assuming all relationships are linear, and the generated models may or may not be different. The systems learn about the true relationships among variables through detecting various features simulating the neural networks, and improve the knowledge about the relationships as the system learns time after time.

Specifically, machine learning allows us to improve the models by tuning the parameters we set for the systems to learn: the learning rate, the error rate, and the number of times we want the machine to learn (as shown in below figure). As the figure below suggests, the x axis represents the entirety of our independent variables (the Xs); the y axis represents the difference between our predictions and the true observed outcomes; the line is one example of how the predictions can look like. The lowest point represents the optimal point when the difference between the predictions and the true observed scores is the smallest.

Learning rate refers to the degree of change in the neurons in each learning epoch (Géron, 2017). For example, we feed in the training data, and have a model generated. Imagine we ask the machine to learn the training data at a high learning rate, which means each time the systems learn, the generated model modifies a lot to fit the training data (the difference between each red dot represents the learning rate of that epoch). When the learning rate is high, we may miss the optimal model by jumping too far; but when the learning rate is low, it may take a long time to reach the optimal model. We need to adjust the learning rate to reach a reasonable speed where the systems can learn fast enough and at the same time not having to miss the optimal point, which is the optimal model the systems can generate (the picture in the middle) (Jordan, J. 2018).


Too low
Just right
Too high
J(θ)
J(θ)
J(θ)
θ
θ
θ
A small learning rate requires many updates before reaching the minimum point
The optimal learning rate swiftly reaches the minimum point
Too large of a learning rate causes drastic updates which lead to divergent behaviors

A second important parameter to tune our generated model is the error rate. The error rate is defined as the number of cases the systems predict wrong out of all cases. It is used to measure if the generated model is accurate enough to represent the reality. With this value available for each generated model, we will be able to know which of the generated models is the closest to the optimal model.

A third parameter to improve the generated models is how many times we ask the systems to learn. With this parameter we need to try and see what is a good number that strikes a balance between the speed that the systems learn and represent the reality as accurately as possible, and how much time it takes. We decide which model is the best fit at different learning times with different learning rates by looking at the error rate each model generates. This is a judgement call that researchers have to make through reviewing literature and practical limitations.

***Summary***

The collection and availability of large educational datasets makes it possible for educational researchers to study patterns and make predictions about larger populations more effectively. However, there are methodological issues in existing educational research using large datasets. This paper presents three of the common methodological issues concerning data processing and data analysis, including missing data handling, modeling and assumptions, and what comes after generating the models. Machine learning, as an application of artificial intelligence that enables systems to learn by themselves and improve from the learning experience without explicit commands, generates more accurate estimated values and provides the technical proof of the advantage of multiple imputation in handling missing data; machine learning could also be a good alternative in generating models and improve the models by tuning parameters to let the systems learn better.

As illustrated above, traditional statistical modeling methods rely on several common and important assumptions, such as normality and linearity. Slight violations of those assumptions may not be a serious problem that would significantly affect analyzing and interpreting the results, but will make the traditional methods weak if better solutions exist. Machine learning is designed best for creating predictive models because on the one hand, it does not rely on the statistical assumptions traditional statistical modeling methods rely on, and on the other, the systems learn about the true relationships among variables through detecting various features simulating neural networks, and improve the knowledge about the relationships as the systems learn time after time.

Additionally, machine learning can improve models through tuning parameters such as the learning rate, error rate, and how many times the systems learn the training data. This brings the existing modeling process to a further step, because creating a model to interpret data should not be the end, we should critically examine our model to determine if the generated model is the optimal one that best represents the reality. This not only helps by making sure the generated model is the optimal fit among the various models, but also establishes a procedure to improve modeling through adjusting the learning process, which can also be applied in future analysis as new data comes in.

In conclusion, this paper presents how machine learning could be an effective tool to address methodological issues existing in current educational research using large datasets, and answer educational questions that researchers and educators may not have been able to answer before. This is not saying that we should abandon traditional statistical analysis methods altogether, but rather suggesting that we should consider using either one or both when they can make the greatest contribution in processing, analyzing, and interpreting data to explain and address educational issues. One thing that we as educators and researchers always have to remember is that regardless of the analysis approach, we are trying to understand, present, and explain some phenomena that are driven by theoretical and/or practical needs, and not for the methods themselves.

**References:**

Acharya, A., & Sinha, D. (2014). Application of feature selection methods in educational data mining. *International Journal of Computer Applications, 103*(2). Doi: 10.5120/18048-8951.

Afifi, A., & Elashoff, R. (1966). Missing observations in multivariate statistics: I. Review of the literature. *Journal of the American Statistical Association, 61*(315), 595–604. Doi:10.2307/2282773.

Anderson, A. B., Basilevsky, A. T., & Hum, D. (1983). Missing Data: A Review of the Literature.  InP. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of Survey Research* (pp. 415-494). New York: Academic Press.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507-547). University of Chicago Press.

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005.

Baker, S., & Inventado, P. S. (2016). Educational data mining and learning analytics: Potentials and possibilities for online education. In G. Veletsianos (Ed.), *Emergence and Innovation in Digital Learning* (83–98). doi:10.15215/aupress/9781771991490.01.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM| Journal of Educational Data Mining, 1*(1), 3-17. https://doi.org/10.5281/zenodo.3554657.

Beck, J. E., & Woolf, B. P. (2000, June). High-level student modeling with machine learning. In *International Conference on Intelligent Tutoring Systems* (pp. 584-593). Springer, Berlin, Heidelberg.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning, 19*(1-2), 205-220. Doi: 10.1007/s10758-014-9223-7.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bodner, T. (2006). Missing data: Prevalence and reporting practices. *Psychological Reports, 99*, 675–680. Doi:10.2466/PR0.99.3.675–680.

Bradley, E. (1994). Missing Data, Imputation, and the Bootstrap. *Journal of the American Statistical Association, 89*(426), 463-475. https://doi.org/10.1080/01621459.1994.10476768.

Brick, J., & Kalton, J. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5*, 215–238. Doi:10.1177/096228029600500302.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet, 7*(2), 223-242. https://doi.org/10.1002/poi3.85.

Buyya, R., Selvi, S. T., & Vecchiola, C. (2013). *Mastering Cloud Computing*. Elsevier, Inc.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review, 106*(5), 124-27.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 84*(4), 487-508. https://doi-org.oca.ucsc.edu/10.3102/0034654314532697

Creswell, J. W., & Creswell, J. D. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics, 2*, 117693510600200030. https://doi.org/10.1177/117693510600200030.

Cunningham, S. J. (1996). Machine learning applications in anthropology: automated discovery over kinship structures. *Computers and the Humanities, 30*(6), 401-406. https://doi.org/10.1177/0267658312443651.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369–382. https://doi.org/10.1177/0267658312443651.

Daniel, B. K. (2015). Big Data and analytics in higher education: opportunities and challenges. *British Journal of Educational Technology, 46*, 904–920. doi:10.1111/bjet.12230.

Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology, 50*(1), 101-113. Doi: 10.1111/bjet.12595.

Darling-Hammond, L. (2002). The research and rhetoric on teacher certification: A response to “Teacher certification reconsidered.” *Educational Policy Analysis Archives, 10*(36), 1–55. https://doi.org/10.14507/epaa.v10n36.2002.

Enders, C. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Galán-García, P., Puerta, J. G. D. L., Gómez, C. L., Santos, I., & Bringas, P. G. (2016). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL, 24*(1), 42-53. https://doi.org/10.1093/jigpal/jzv048.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

Ge, Y., & Qiu, Q. (2011, June). Dynamic thermal management for multimedia applications using machine learning. In *2011 48th ACM/EDAC/IEEE Design Automation Conference (DAC)* (pp. 95-100). IEEE.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Greenberg, E., Rhodes, D., Ye, X. & Stancavage, F. (2004). *Prepared to teach: Teacher preparation and student achievement in eighth-grade mathematics.* Washington, DC: American Institute for Research.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of educational research, 66*(3), 361-396. https://doi.org/10.3102/00346543066003361.

Grimm, K. J., Mazza, G. L., & Mazzocco, M. M. M. (2016). Advances in methods for assessing longitudinal change. *Educational Psychologist, 51*, 342–353. https://doi-org.oca.ucsc.edu/10.1080/00461520.2016.1208569.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley.

Guan, Y., Zhang, H., Quang, D., Wang, Z., Parker, S. C., Pappas, D. A., ... & Zhu, F. (2019). Machine Learning to Predict Anti–Tumor Necrosis Factor Drug Responses of Rheumatoid Arthritis Patients by Integrating Clinical and Genetic

Markers. *Arthritis & Rheumatology, 71*(12), 1987-1996. https://doi.org/10.1002/art.41056.

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software, 10*(1), 1-20.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society: Series B, Methodological, 30*, 67–82.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T., R. Tibshirani, B. Narasimhan, & G. Chu. (2014). Impute: Imputation for Microarray Data. R Package Version 1.32.0. Accessed January 15, 2014. http://www.bioconductor.org/packages/release/bioc/manuals/impute/man/impute.pdf.

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*.

Honaker, J., G. King, & M. Blackwell. (2011). Amelia II: A Programme for Missing Data. *Journal of Statistical Software 45*(7): 1– 47. Doi: 10.18637/jss.v045.i07.

Hong, M., Razaviyayn, M., Luo, Z. Q., & Pang, J. S. (2015). A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine, 33*(1), 57-77. DOI: 10.1109/MSP.2015.2481563.

Ismael, O. A., Song, D., Ha, P. T., Gilbert, P. J., & Xue, H. (2017). *U.S. Patent No. 9,594,905*. Washington, DC: U.S. Patent and Trademark Office.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine, 50*(2), 105-115. https://doi-org.oca.ucsc.edu/10.1016/j.artmed.2010.05.002.

Jordan, J. (2018). Setting the learning rate of your neural network. Retrieved from: https://www.jeremyjordan.me/nn-learning-rate/.

Kim, J. (2004). Finite sample properties of multiple imputation estimators. *Annals of Statistics, 32*, 766–783. Doi:10.1214/009053604000000175.

Kim, S. J., & Boyd, S. (2008). A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal on Optimization, 19*(3), 1344-1367. https://doi.org/10.1137/060677586.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American political science review, 95*(1), 49-69. https://doi-org.oca.ucsc.edu/10.1017/S0003055401000235

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review, 37*(4), 331-344. https://doi-org.oca.ucsc.edu/10.1007/s10462-011-9234-x.

Kučak, D., Juričić, V., & Đambić, G. (2018). Machine learning in education: A survey of current research trends. *Proceedings of the 29th DAAAM International Symposium,* 406-410. Doi: 10.2507/29th.daaam.proceedings.059

Kundu, S., Rangaswami, R., Gulati, A., Zhao, M., & Dutta, K. (2012, March). Modeling virtualized applications using machine learning techniques. In *Proceedings of the 8th ACM SIGPLAN/SIGOPS conference on Virtual Execution Environments* (pp. 3-14). https://doi.org/10.1145/2151024.2151028.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM, 38*(11), 54-64. https://doi.org/10.1145/219717.219768.

Little, R. J. A. 1988. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3): 287–296.

Little, R. J. A., & D. B. Rubin. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

McKenney, S., & Mor, Y. (2015). Supporting teachers in data-informed educational design. *British Journal of Educational Technology, 46*, 265–279. https://doi.org/10.1111/bjet.12262.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill. ISBN 978-0-07-042807-2.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology, 45*, 27-45. https://doi-org.oca.ucsc.edu/10.1146/annurev-soc-073117-041106.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106. DOI: 10.1257/jep.31.2.87.

Murtonen, M., & Lehtinen, E. (2003). Difficulties experienced by education and sociology students in quantitative methods courses. *Studies in Higher Education, 28*, 171–185. Doi:10.1080/0307507032000058064.

Negi, R., & Mathew, R. (2018, December). Machine Learning Algorithms for Diagnosis of Breast Cancer. In *International conference on Computer Networks, Big data and IoT* (pp. 928-932). Springer, Cham.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning, 65*(2), 470-476. https://doi.org/10.1111/lang.12104.

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research, 151*(1), 53-79. https://doi.org/10.1016/S0377-2217(02)00578-7.

Ortega, L., & Byrnes, H. (2009). *The longitudinal study of advanced L2 capacities*. Routledge.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual review of applied linguistics, 25*, 26-45. https://doi.org/10.1017/S0267190505000024.

Pampaka, M., Hutcheson, G., & Williams, J. (2016). Handling missing data: analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education, 39*(1), 19-37. https://doi.org/10.1080/1743727X.2014.979146.

Peugh, J., & Enders, C. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525–556. Doi:10.3102/00346543074004525.

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation, 7*(4), 353-383. https://doi.org/10.1076/edre.7.4.353.8937.

Póczos, B., Xiong, L., & Schneider, J. (2012). Nonparametric divergence estimation with applications to machine learning on distributions. *arXiv preprint arXiv:1202.3758*.

Puri, M. (2019). Automated machine learning diagnostic support system as a computational biomarker for detecting drug-induced liver injury patterns in whole slide liver pathology images. Assay and drug development technologies. https://doi.org/10.1089/adt.2019.919.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002.

Rahman, S. M., Ali, M. A., Altwijri, O., Alqahtani, M., Ahmed, N., & Ahamed, N. U. (2019, July). Ensemble-Based Machine Learning Algorithms for Classifying Breast Tissue Based on Electrical Impedance Spectroscopy. In *International Conference on Applied Human Factors and Ergonomics* (pp. 260-266). Springer, Cham.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Recknagel, F. (2001). Applications of machine learning to ecological modelling. *Ecological modelling, 146*(1-3), 303-310. https://doi.org/10.1016/S0304-3800(01)00316-7.

Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist, 51*, 305–316. https://doi-org.oca.ucsc.edu/10.1080/00461520.2016.1208094.

Richman, M. B., Trafalis, T. B., & Adrianto, I. (2007, January). Multiple imputation through machine learning algorithms. In *Fifth conference on artificial intelligence applications to environmental science* (Vol. 3).

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601-618. 10.1109/TSMCC.2010.2053532.

Rubin, D. B. (1987). *Multiple imputation in sample surveys and censuses*. New York: John Wiley.

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, N.J: Prentice Hall.

Samii, C., Paler, L., & Daly, S. Z. (2016). Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia. *Political Analysis, 24*(4), 434-456. https://doi.org/10.1093/pan/mpw019.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development,* 44, 206-226. Doi: 10.1147/rd.441.0206.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L., and J. W. Graham. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2): 147 –177. https://doi.org/10.1037/1082-989X.7.2.147.

Shabtai, A., Fledel, Y., & Elovici, Y. (2010, December). Automated static code analysis for classifying android applications using machine learning. In *2010 International Conference on Computational Intelligence and Security* (pp. 329-333). IEEE. DOI: 10.1109/CIS.2010.77.

Sheard, J. (2018). Quantitative data analysis. In *Research Methods: Information, Systems, and Contexts*, (2nd ed., pp. 429-452). Elsevier.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008, August). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 614-622). https://doi-org.oca.ucsc.edu/10.1145/1401890.1401965.

Sovilj, D., Eirola, E., Miche, Y., Björk, K. M., Nian, R., Akusok, A., & Lendasse, A. (2016). Extreme learning machine for missing data using multiple imputations. *Neurocomputing, 174*, 220-231. https://doi-org.oca.ucsc.edu/10.1016/j.neucom.2015.03.108.

Tang, T. Y., & McCalla, G. (2003, July). Smart recommendation for an evolving e-learning system. In *Workshop on Technologies for Electronic Documents for Supporting Learning, International Conference on Artificial Intelligence in Education* (pp. 699-710).

Toch, E., Lerner, B., Ben-Zion, E., & Ben-Gal, I. (2019). Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems, 58*(3), 501-523. https://doi-org.oca.ucsc.edu/10.1007/s10115-018-1186-x.

Valletta, J. J., Torney, C., Kings, M., Thornton, A., & Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour, 124*, 203-220. https://doi.org/10.1016/j.anbehav.2016.12.005.

van Zoonen, W., & Toni, G. L. A. (2016). Social media research: The application of supervised machine learning in organizational communication research. *Computers in Human Behavior, 63*, 132-141. https://doi.org/10.1016/j.chb.2016.05.028.

Varaki, B. S., & Earl, L. (2006). Math Modeling in Educational Research: An Approach to Methodological Fallacies. *Australian Journal of Teacher Education, 31*(2). http://dx.doi.org/10.14221/ajte.2006v31n2.3

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in quantitative methods for psychology, 8*(1), 52-69. Doi: 10.20982/tqmp.08.1.p052.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100-1122. https://doi-org.oca.ucsc.edu/10.1177/1745691617693393.

Yi, M. (2019). Application of Principal Component Analysis in Teaching Evaluation. *Frontiers in Sport Research, 1*(1).

Young, W., Weckman, G., & Holland, W. (2011). A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits. *Theoretical Issues in Ergonomics Science, 12*, 15–43. Doi:10.1080/14639220903470205.

Zaïane, o., Luo, J. (2001). Web usage mining for a better web-based learning environment. In *Proceedings of Conference on Advanced Technology for Education* (pp. 60–64). Banff, Alberta.

Zaïane, O. (2002). Building A Recommender Agent for e-Learning Systems. In *Proceedings of the International Conference in Education, Auckland* (pp. 55-59), New Zealand.