Journal of Methods and Measurement in the Social Sciences,
Vol. 13, No.1, 1-14, 2022

Machine Learning Method for High-Dimensional Education Data

Haiyan Bai
University of Central Florida
Xing Liu
Eastern Connecticut State University
Fangxing Bai
University of Cincinnati
Yuting Chen
University of Central Florida
Randyll Pandohie
University of Central Florida

Machine learning has become one of the important methods to process big data. It has made a breakthrough in the limitations of traditional statistical models dealing with high-dimensional data. The current study is to introduce and discuss about how machine learning method can be implemented in high-dimensional education data and help with increasing the model efficacy in dealing with high-dimensional education data. A demonstration of the implementation with an empirical data set is also provided.

Key words: Machine learning, high-dimensional data, educational research

The increasing development of the technology and internet advanced the data collection and storage techniques (Kidzinski et al., 2016) that allow big data available in every field. Machine learning (ML) has become one of the important methods to process big data, having made a breakthrough in the limitations of traditional statistical models dealing with high-dimensional data. ML methods have been widely used in many fields, such as artificial intelligence development (e.g., Bengio, 2009; Kamar, 2016, Monostori, 2003), medical studies (e.g., Goecks et al., 2020; Kononenko, 2001; Wang & Summers, 2012), agriculture (e.g., Duro et al, 2012; Rehman et al., 2019), business (Bose & Mahapatra, 2001; Dean, 2014; Lieder, 2020), industry (e.g., Ge et al., 2017; Paek & Pieraccini, 2008; Tsakanikas et al., 2020), and many other areas. However, the journal publications on implementation of ML method in solving high-dimensional education data are sparse, except some conference papers starting to focus on the applications of ML in processing educational data (e.g., Ciolacu et al., 2017). The literature gap is noticeable that ML still has not been well used in educational research, while ML will allow educational researchers to address problems which were not accessible previously (Kidzinski et al., 2016).

What is machine learning?

Machine learning algorithms use statistics to find patterns in the data (Michie et al., 1994) which encompasses a lot of information with numbers, words, images, clicks, and many forms of data that we can collect. Once we can store them in computer, the data can be fed into a machine-learning algorithm.

ML is usually classified into *supervised learning* and *unsupervised learning* (Sathya & Abraham, 2013). In *supervised learning*, which is most prevalent, the data are labeled to tell the machine exactly what patterns it should look for. Supervised learning conducts the ML task of inferring a function from training and testing data. It basically uses the known data to do classification and regression. For classification, it usually predicts discrete valued output. It is most frequently encountered decision-making tasks (Jordan & Rumelhart, 1992). For regression, it commonly performs predictions using the recursive resampling procedures to predict continues output (Al Hasan, 2006). *Unsupervised learning* refers to the technique of finding hidden structure in unlabeled data. In unsupervised learning, it mainly clusters data based on the features (Figueiredo & Jain, 2002). The key point for unsupervised learning is that the information input to the machine does not have either supervised target outputs, or rewards from its environments; therefore, the ML model will explore the patterns of the data to build representation of the input for decision making (Ghahramani, 2003). Clustering and dimensionality reduction are usually the classic tasks for unsupervised learning.

There is another type of ML named reinforcement learning which is the latest frontier of ML. It is mainly using reinforcement algorithm learned by a trial and error to achieve a clear objective. This is mainly used in AI development like Google's AlphaGo, the program that famously beat the best human players in the complex game of Go (Hao, 2018; Inventado, 2012); but this is not the focus of the current study for educational high-dimensional data.

Machine learning method for education data

There are many factors related to education variables, such as predicting career paths, precise grading, and more personalization in the classroom. However, studying the actual behavior of teachers and students has always been a difficult and expensive proposition (Petrilli, 2018). Machine learning makes it possible to analyze the high-dimensional data containing rich information for better understanding student behaviors and performance or psychological properties with rich information collected from variety of instruments including tests, surveys, videos, interviews, images to capture class activities, and teaching activities with teachers' questions and students' responses (Inventado, 2012). Machine learning enables us to build a research enterprise that improves classroom instruction, regardless of how traditional or technology-infused the instruction might be (Petrilli, 2018).

Supervised Learning for Education Data. Classification and regression are two types of supervised learning that are useful for handling high-dimensional data in education. In educational studies, there are many situations that we may need to identify students or teachers into predefined groups based on many known factors, such as substance use, students with learning disabilities, student drop-out, and teachers' attrition. For regression, it commonly predicts continuous output (Al Hasan, 2006). For example, in education, student academic performance is always influenced by many factors, and once we have the data with many related variables, we can use ML to predict the student academic performance more accurately by considering the high-dimensional data with rich education related information. There are many other education outcomes with many influential factors, but we cannot use all of them as predictors in classic statistical models. In this case, ML can handle the high-dimensional data well with more accurate predictions because ML predicts performance using various performance metrics, such as accuracy, precision-recall, F-values, squared error etc. with multiple-fold cross validation procedures utilizing resampling techniques (Al Hasan, 2006; Shipp et al., 2002).

Unsupervised learning for education data. Unsupervised learning is usually focusing on clustering and dimensionality reduction. It is also useful in educational research. It is very common for educational studies to collect many data related to students' behaviors and psychological data. In this case, we may need to cluster the behaviors into different categories or cluster the students by the cognitive and affective elements into groups to better understand students' needs for educational services. Another example to implement unsupervised learning in education is that student learning behavior may be related to learning environment with many influential features interplaying between emotion, learning and non-learning related activities; in this case, we can cluster the student groups by the influential features to explore the important characters for building a heathy learning environment for students (Inventado, 2012).

Although classical statistical techniques may still apply, large datasets allow us to discover deeper patterns and to provide more accurate predictions of student's behaviors and learning outcomes (Kidzinski et al., 2015).

Available software to conduct the ML project

There are many different types of machine learning software. These software packages include caret in R, TensorFlow, Shogun in C++ (open-source), Apache Spark MLlib, Oryx 2, H20.ai, Pytorch, RapidMiner, Weka, KNIME, and Keras. They support various languages like Python, R, Scala, C#, Ruby etc., to meet the user's needs.

In the current study, we use the caret in R to demonstrate an example to implement ML in high-dimensional education data. caret **(Classification And Regression Training)** is a comprehensive package in R for solving supervised machine learning problems (Kuhn & Johnson, 2013). It contains several machine learning algorithms and standardizes various other tasks such as data splitting, pre-processing, feature selection, variable importance estimation with graphics.

The purpose of the current study is to discuss about and demonstrate how machine learning method can be implemented in high-dimensional education data to help with increasing the prediction efficacy in dealing with high-dimensional education data.

Method

Data sources

The data used for this demonstration are from an online public resource (Cortez & Silva, 2008). The demonstration does not intend to derive any inferences from the data, but only for the purpose of demonstrating the implementation of ML in high dimensional education data. The data were collected from 788 high school students in Portugal during 2005-2006 school year from two public schools. During the school year, students were evaluated in three periods and the last evaluation (G3 of Table 1) corresponded to the final grade. The data contained 33 variables including student demographic data, family backgrounds, school related information, and students' behavior, and academic performance data. The 33 variables are described in Table 1. The short variable labels from Table 1 are used in Graph 1.

Table 1

The 33 variables related to students in the dataset

#	Label	Description	Coding
1	school	student's school	(binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2	sex	student's sex	(binary: 'F' - female or 'M' - male)
3	age	student's age	(numeric: from 15 to 22)
4	address	student's home address type	(binary: 'U' - urban or 'R' - rural)
5	famsize	family size	(binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6	Pstatus	parent's cohabitation status	(binary: 'T' - living together or 'A' - apart)
7	Medu	mother's education	(numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8	Fedu	father's education	(numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9	Mjob	mother's job	(nominal: 'teacher', 'health' care related, civil 'services' e.g. administrative or police), 'at_home' or 'other')
10	Fjob	father's job	(nominal: 'teacher', 'health' care related, civil 'services' e.g. administrative or police), 'at_home' or 'other')
11	reason	reason to choose this school	(nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	guardian	student's guardian	(nominal: 'mother', 'father' or 'other')
13	traveltime	home to school travel time	(numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	studytime	weekly study time	(numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	number of past class failures	(numeric: n if 1<=n<3, else 4)

Table 1 (continued)

#	Label	Description	Coding
16	schoolsup	extra educational support	(binary: yes or no)
17	famsup	family educational support	(binary: yes or no)
18	paid	extra paid classes within the course subject (Math or Portuguese)	(binary: yes or no)
19	activities	extra-curricular activities	(binary: yes or no)
20	nursery	attended nursery school	(binary: yes or no)
21	higher	wants to take higher education	(binary: yes or no)
22	internet	Internet access at home	(binary: yes or no)
23	romantic	with a romantic relationship	(binary: yes or no)
24	famrel	quality of family relationships	(numeric: from 1 - very bad to 5 - excellent)
25	freetime	free time after school	(numeric: from 1 - very low to 5 - very high)
26	goout	going out with friends	(numeric: from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption	(numeric: from 1 - very low to 5 - very high)
28	Walc	weekend alcohol consumption	(numeric: from 1 - very low to 5 - very high)
29	health	current health status	(numeric: from 1 - very bad to 5 - very good)
30	absences	number of school absences	(numeric: from 0 to 93)
31	G1	first period grade	(numeric: from 0 to 20)
31	G2	second period grade	(numeric: from 0 to 20)
32	G3	final grade	(numeric: from 0 to 20, output target)

Suppose that we want to predict student mathematic learning outcome G3 (final grade of numeric data: from 0 to 20), it is clear that we are not able to include all the 32 variables in our multiple regression (MR) model as the predictors due to the concerns for the reduction of statistical power; thus, variable selection for MR model or most classic statistical predictive models could be challenging because the model accuracy will be an issue if we exclude any essential variables in the model. However, to use the ML method, we are able to put all the available variables into the model to run the importance of the predictors for identifying major predictors using ML techniques. ML uses multiple procedures with bootstrap resampling technique in random forest functions to assess the importance of the predictors; therefore, the selection of the significant predictors is considered more accurate, and the procedures are efficient.

To implement the ML procedures for selecting importance of the variables in prediction of the student math final grade score of G3, in this demonstration, we used the caret package in R 4.2.1.

Data Analysis Procedures and Results

To demonstrate that ML can be implemented in high-dimension education data, we focus one issue of how ML selects importance of the variables for prediction on. Therefore, in this proposal we only demonstrate the procedures of how caret package runs importance analysis.

Specifically, the major procedures are as follows: To run the importance analysis for which variables are important predictors for student mathematic learning outcome G3 in the Portugal High School data, we first prepare the data to run the ML feature selection model. The data preparation including missing data imputation, scaling data, checking and reducing skewness, and removing the multicollinearity of the data. Those are the normal data cleaning procedures which are omitted from this paper. The following steps are the major steps after the data cleaning to run the importance analysis using caret.

#Step 1. To install and load caret and other related packages for the current study

```
install.packages(c('caret', 'skimr', 'RANN', 'randomForest',
'fastAdaboost', 'gbm', 'xgboost', 'caretEnsemble', 'C50', 'earth',
'mlbench'))
library(caret)
```

#Step 2. To import the dataset (for our current example is student-mat.csv dataset)

```
d1_math=read.table(“student-mat.csv”, sep=”;”,header=TRUE
```

#Step 3. After the dataset imported, we created the training and test datasets by splitting the dataset into training (80%) and test (20%) or any other ratio (e.g., training 70% vs. test 30%) is also fine with the training sample of a larger potion.)

```
sample <- sample(c(TRUE, FALSE), nrow(d1_math),
replace=TRUE, prob=c(0.8,0.2))
train_math  <- d1_math[sample, ]
test_math   <- d1_math[!sample, ]
```

#Step 4. Prepare data and run the importance analysis: using random forest functions and resampling procedures at 10 folds to conduct importance analysis

```
control <- trainControl(method=”repeatedcv”, number=10,
repeats=3
model <- caret::train(G3~., data=train_math, method=”rf”,
preProcess=”scale”, trControl=control, na.action = na.omit)
importance <- varImp(model, scale=FALSE)
```

#Step 5. Summary of the statistics of the ranking of important variables

```
print(importance)
```

#Step 6. Plot the ranking of important variables

```
plot(importance)
```

The above code performs the recursive feature selection using the outer resampling method of 10-fold to cross-validate the selection results with each group of the selected variables appear in the same group together. The code produces statistics related to each model for the subset of the data (See Table 2).

In Table 2, based on the statistics, we can see that caret identifies 5 top variables as important predictors in the order as *G2* (second period grade), *absences* (the absences times), *famrel* (quality of family relationships), *failures* (number of past class failures), and *studytime* (weekly study time). The star at the row 5 indicates the best model size out of the provided models sizes (in subsets) is 5. The five variables were selected through recursive feature selection with 10-fold Cross-Validated procedures by resampling performance over subset size, that is the five variables appear in each subset repeatedly. Comparing to the multiple regression model, caret can put much more variables in the model and using resampling procedures to cross validate the selection results, which make the predictions more accurate, and the reliability of the predictions are improved.

caret can also produce the graphs to present the import of all the variables in the data in the prediction using the plot function as presented in Step 6.

Table 2

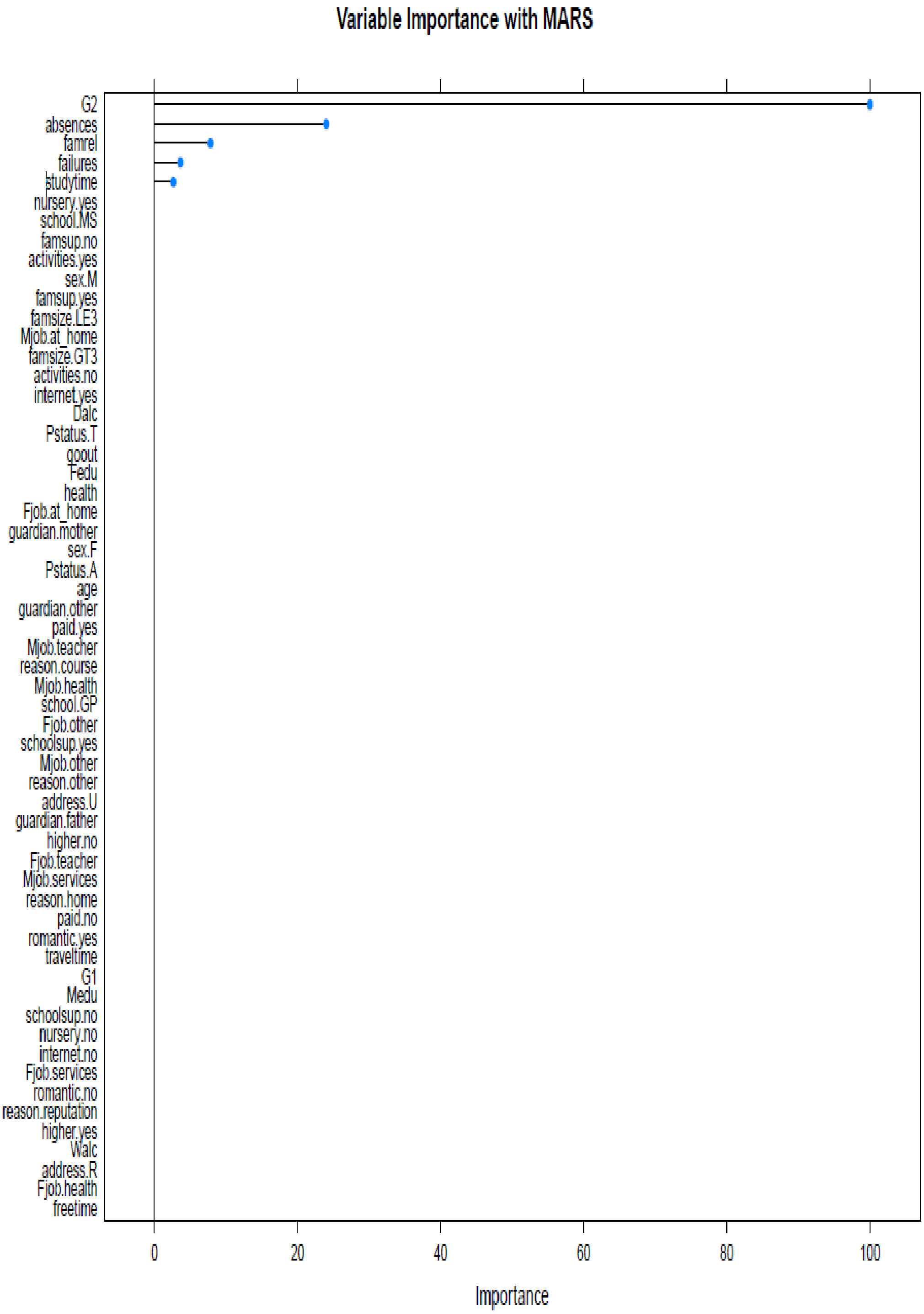
Recursive feature selection by resampling performance over subset size

Variables Selected	RMSE[1]	R^2	MAE[2]	RMSE *SD*	R^2 *SD*	MAE *SD*
1	4.575	.0610	3.523	0.4630	0.0631	0.3321
2	4.617	.0459	3.534	0.4646	0.0575	0.3399
3	4.606	.0425	3.512	0.4479	0.0540	0.3149
4	4.586	.0474	3.492	0.4489	0.0527	0.3125
5	4.572	.0552	3.485	0.4532	0.0631	0.3125*
10	4.599	.0753	3.563	0.5082	0.0854	0.3938
15	4.688	.0651	3.645	0.5229	0.0719	0.3985
18	4.678	.0651	3.635	0.5313	0.0683	0.4043
31	4.689	.0576	3.623	0.5067	0.0666	0.3834

Note. [1]RMSE = Root Mean Square Error which is the standard deviation of the residuals (prediction errors); [2]MAE=mean absolute error which is a measure of errors between paired observations expressing the same phenomenon. * Indicates the MAE for 5 variables selected is the lowest and this coupled with the variable importance means that these 5 variables may be significant in predicting the final grade for the students. Outer resampling method: Cross-Validated (10-fold, repeated 10 times) results are presented.

In Graph 1, we can see that the ML randomForest procedure selects five important predictors for predicting G3 (final grade). These five predictors are *G2* (second period grade), *absences* (the number of absences), *famrel* (quality of family relationships (numeric: from 1 – very bad to 5 – excellent)), *failures* (number of past class failures (numeric: n if 1 ≤ n < 3, else 4)), and *studytime* (weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)).

When we examine the selected important predictors ranked by the ML procedures, we confirmed that all the predictors are theoretically supportive as predictors for students' learning outcome, such as students' previous learning outcomes (Hailikari, Nevgi, & Komulainen, 2008), absence of classes (Gottfried, 2009), family relationships (Roksa & Kinsley, 2019; Tomul, Önder, & Taslidere, 2021), class failures (Salal, Abdullaev, & Kumar, 2019), and study time (Xu, 2022).



Graph 1. *Variable Importance*

Discussions and Conclusions

The demonstration results produced only one type of methodological procedures for how to use ML to help with the improvement of classical statistical modeling with high dimensional education data. In educational research, regression models are commonly used for predictions; however, when we have many variables, we cannot include all of them as predictors in the regular regression models, such as multiple or logistic regression for predictions. In the cases of high dimensional data, such as dataset with more than 20 variable, factor analysis has been used to reduce the data dimensions. However, it is very common that factor analysis may group the unrelated constructs or concepts into one factor, which is problematic. In that, it will significantly challenge the model prediction accuracy when using a variable created through grouping some unimportant variables together in a cluster as a variable from unrelated constructs or concepts in the regression model. As for the existing advanced methods, latent class analysis (LCA) can produce better results for multivariate variables than factor analysis in terms of clustering the variables in meaningful groups; however, the LCA models need to be manually built based on the hypotheses or theories, while ML can use the training sample to train the machine to learn the patterns of the data. For applications, ML procedures as demonstrated in this study are much easier than LCA due to the complicated procedures of LCA modeling for empirical researchers to implement the methods in their real study. Therefore, the ML importance analysis is a better alternative strategy to solve this high-dimensional data issue in classic statistical modeling.

The limitations of the current study are that we only demonstrate one type of ML model to analyze education data. We also only presented the use of random forest method for recursive important predictor selections. There are many other ways to use ML in education data to answer many research questions. For further study, we will (1) compare the other ML methods for variable selections, and (2) demonstrate other implementations of the ML in high dimensional education data, such as dimension reductions.

It is evident that it is beneficial to apply ML in high dimensional education data. This study briefly introduces the fundamental concept of ML and some specific educational areas that have potential to implement ML to solve related questions with high-dimensional data. The demonstration presents how ML can work in identifying important predictors and improving the accuracy for variable selections for predictions.

This study is expected to promote the use of ML method in educational studies to fill in the literature gap. It is significant to educational studies by helping educational researchers use the available high-dimensional data to solve some data issues that classic statistical models cannot deal with and to help increase the prediction accuracy with high dimensional data.

Author Notes. Corresponding Author: Haiyan Bai, Ph.D. Professor, Quantitative Methodology Dept. of Learning Sciences & Educational Research, College of Community Innovation and Education, University of Central Florida, P.O. Box 161250, Orlando, FL 32816-1250, email: Haiyan.Bai@ucf.edu

References

Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. (2006, April). Link prediction using supervised learning. In *SDM06: Workshop on link analysis, counterterrorism and security* (Vol. 30, pp. 798-805).

Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers Inc.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. *Information & Management*, *39*, 211-225.

Ciolacu, M., Tehrani, A. F., Beer, R., & Popp, H. (2017, October). Education 4.0—Fostering student's performance with machine learning methods. In *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)* (pp. 438-443).

Cortez, P., & Silva, A. M. G. (2008, April). Using data mining to predict secondary school student performance. In A. Brito and J. Teixeira Eds., *Proceedings of 5th FUture BUsiness TEChnology Conference* (FUBUTEC 2008) pp. 5-12, Porto, Portugal, EUROSIS, ISBN 978-9077381-39-7.

Dean, J. (2014). *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. New York: John Wiley & Sons.

Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, *118*, 259-272.

Figueiredo, M. A. T., & Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, *24*, 381-396.

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data mining and analytics in the process industry: The role of machine learning. *IEEE Access*, *5*, 20590-20616.

Ghahramani, Z. (2003, February). Unsupervised learning. In *Summer School on Machine Learning* (pp. 72-112). Berlin, Heidelberg:Springer,.

Goecks, J., Jalili, V., Heiser, L. M., & Gray, J. W. (2020). How machine learning will transform biomedicine. *Cell*, *181*, 92-101.

Gottfried, M. A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis, 31*, 392-415.

Hailikari, T., Nevgi, A., & Komulainen, E. (2008). Academic self-beliefs and prior knowledge as predictors of student achievement in Mathematics: A structural model. *Educational Psychology, 28*, 59-71.

Hao, K. (2018). What is machine learning? *MIT Technology Review. URL*:https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/

Inventado, P. S., Legaspi, R., Cabredo, R., & Numao, M. (2012, December). Student learning behavior in an unsupervised learning environment. In *Proceedings of the 20th international conference on computers in education* (pp. 730-737).

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*, 307-354.

Kamar, E. (2016, July). Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *International Joint Conferences on Artificial Intelligence Organization* (pp. 4070-4073).

Kidziński, Ł., Giannakos, M., Sampson, D. G., & Dillenbourg, P. (2016). A tutorial on machine learning in educational science. In *State-of-the-Art and Future Directions of Smart Learning* (pp. 453-459). Singapore: Springer.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23, 89-109.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Lieder, M., Asif, F. M., & Rashid, A. (2020). A choice behavior experiment with circular business models using machine learning and simulation modeling. *Journal of Cleaner Production*, 120894.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning. *Neural and Statistical Classification*, *13*, 1-298.

Monostori, L. (2003). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *Engineering Applications of Artificial Intelligence*, *16*, 277-291.

Paek, T., & Pieraccini, R. (2008). Automating spoken dialogue management design using machine learning: An industry perspective. *Speech Communication*, *50*, 716-729.

Petrilli, M. J. (2018). Big data transforms education research. *Education Next, Winter*, 86-87.

Rehman, T. U., Mahmud, M. S., Chang, Y. K., Jin, J., & Shin, J. (2019). Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Computers and Electronics in Agriculture*, *156*, 585-605.

Roksa, J., & Kinsley, P. (2019). The role of family support in facilitating academic success of low-income students. *Research in Higher Education*, *60*, 415-436.

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*, 34-38.

Salal, Y. K., Abdullaev, S. M., & Kumar, M. (2019). Educational data mining: Student performance prediction in academic. *International Journal of Engineering and Advanced Tech*, *8*, 54-59.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., ... & Ray, T. S. (2002). Diffuse large B-cell lymphoma outcome prediction by geneexpression profiling and supervised machine learning. *Nature Medicine*, *8*, 68-74.

Tomul, E., Önder, E., & Taslidere, E. (2021). The relative effect of student, family and school-related factors on math achievement by location of the school. *Large-scale Assessments in Education*, 9, 1-29.

Tsakanikas, P., Karnavas, A., Panagou, E. Z., & Nychas, G. J. (2020). A machine learning workflow for raw food spectroscopic classification in a future industry. *Scientific Reports*, *10*, 1-11.

Wang, S., & Summers, R. M. (2012). Machine learning and radiology. *Medical Image Analysis*, *16*, 933-951.

Xu, J. (2022). A latent profile analysis of homework time, frequency, quality, interest, and favorability: implications for homework effort, completion, and math achievement. *European Journal of Psychology of Education*, 1-25.