APPLYING STATISTICAL METHODS IN CLUSTERED EDUCATIONAL DATA

A Dissertation

by

WENTING WENG

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Wen Luo
Committee Members,	Oi-Man Kwok
	Noelle Wall Sweany
	Nicola L. Ritter
Head of Department,	Fuhui Tong

May 2022

Major Subject: Educational Psychology

ABSTRACT

As technologies have been used in education, data have been generated within technologies or collected outside. How educators can utilize data has become a challenge. Therefore, a systematic literature review was conducted in the first study. The review noted the impact from previous learning analytics and educational data mining studies uncovering sample and methodological characteristics of the studies. The findings showed every aspect of the studies, including research objectives, learning environments, education levels, data preprocessing tasks, data analysis methods, data tools, sample sizes, and feature information. Additionally, big data in education can support the application of learning theories into practices. The design and improvement of technologies can use these theories as underpinnings.

The second study applied mixed effects Random Forest (MERF), the random effects expectation-maximization recursive partitioning method (RE-EM Tree), hierarchical linear modeling (HLM), and regular Random Forest (RF). The comparison results of these methods have shown that MERF generated the most accurate models. RE-EM Tree and HLM achieved similar accuracy. The advantages and disadvantages of each method were explained. The results indicated that MERF was more appropriate than RF in clustered data and choosing which method depended on a research or project purpose. When the purpose is to predict students' learning performance, MERF can be the optimal method choice. When the purpose is to detect the relationship between predictor and response variables and examine each variable's impact, RE-EM Tree and HLM will better serve the purpose. Whether we should select RE-EM Tree or HLM can depend on the size of data dimension.

Considering the data dimension, HLM was applied in the third study to examine the relationship between student information and communications technology (ICT) related factors and learning performance in mathematics and science moderated by school-level factors. The results showed the importance of ICT related factors and indicated that schools with higher students' socio-economic status yielded better learning outcomes in mathematics and science as well as better supported ICT use. The shortage of school resources had an interaction effect with students' ICT use at school. School size was also important for students' mathematics achievements.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Wen Luo, and my committee members, Dr. Oi-man Kwok, Dr. Nicola L. Ritter, and Dr. Noelle Sweany, for their guidance and support throughout the course of this research. I sincerely want to express my gratitude and appreciation from the bottom of my heart to Dr. Luo for her mentorship and contributions of time, assistance, and support on my Ph.D. study. I also appreciate the support from Dr. Ritter.

Thanks to my parents for their support and encouragement during hard times in my Ph.D. pursuit.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Wen Luo, Oi-man Kwok, and Noelle Wall Sweany of the Department of Educational Psychology, and Professor Nicola L. Ritter of the Department of Veterinary Integrative Biosciences.

All work for the dissertation was completed independently by the student.

Funding Sources

This work was not funded.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

In the past decade, Learning Analytics (LA) and educational data mining (EDM) have been greatly impacting the education field. Many research institutions have collected various types of data and technologies used in classrooms or online courses have started tracking users' actions. How to handle the data to reveal the pattern beneath the data has raised great concerns. Therefore, the first goal of this dissertation is to provide a comprehensive review through the lens of how LA and EDM have been impacting education and discover the value and potential of the LA/EDM field. The first study includes the analyses of 113 articles selected from nine peer-review journals with high impact factors in the educational technology field. The study aims to reveal the impacts of these studies and detailed information of their sample and methodological characteristics.

Previous educational studies have adopted different data mining algorithms to solve research questions and practical issues. However, most data mining algorithms currently only consider single data levels (e.g., Martínez-Abad et al., 2018) and few studies using data mining algorithms in clustered educational data have considered different data levels and mixed effects. When the data has clustered structures, the methods without considering different data levels may yield misleading results. Therefore, the second goal of this dissertation is to overcome the drawbacks of traditional data mining methods such as Decision Tree (DT) and Random Forest (RF) and apply the mixed effects tree models to the clustered data. The study compares the results of RF, mixed effects Random Forest (MERF), the random effects expectation-maximization recursive partitioning method (RE-EM Tree), and hierarchical linear modeling (HLM). The study aims to reveal the optimal methods applied to certain circumstances in clustered data.

Considering each method’s advantages and disadvantages, the third goal of the dissertation is to apply HLM to examine the relationship between students’ information and communications technology (ICT) related factors and their learning achievements in mathematics and science moderated by the school-level factors. Previous studies focused more on student-level and country-level information. A few studies reported school-level analysis, but selected school-level factors based on research interests or preferences (e.g., Gómez-Fernández & Mediavilla, 2018). Therefore, the study applies DT method to select school-level factors in an unbiased and data-driven fashion. Two separate DT models are generated based on students’ mathematics and science achievements. Accordingly, two separate HLM models are generated to uncover the relationships among predictor and response variables as well as moderators.

CHAPTER II

LEARNING ANALYTICS AND EDUCATIONAL DATA MINING: A SYSTEMATIC LITERATURE REVIEW

Introduction

In the past decade, big data is becoming ubiquitous in many fields. Educational data mining (EDM) and Learning Analytics (LA) are growing interdisciplinary fields of studies. EDM was defined by the International Educational Data Mining Society as “an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in” (p. 601). EDM projects build models to improve teaching and learning experiences as well as institutional effectiveness (Dutt et al., 2017). EDM can also be adopted to evaluate educational systems so as to improve education (Macarini et al., 2020). EDM deploys statistical and data mining techniques to various educational data sets from different educational settings (Romero & Ventura, 2020). As technologies have been integrated into education, those educational settings and supportive tools, such as Massive Open Online Courses (MOOCs) and Learning Management Systems (LMSs), show great advantages in data analytics. For example, these tools can capture and store data information of students’ learning process and interaction (e.g., Castro et al., 2007). The overall EDM procedure to reveal information from educational data sets mainly includes data preprocessing, data analysis, and post processing (Romero et al., 2004).

Respectively, LA has historical roots that are related to academic analytics, action analytics, and predictive analytics (Weng et al., 2021). The first international Conference on Learning Analytics and Knowledge (LAK 2011) defined LA as “the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (https://tekri.athabascau.ca/analytics/). LA aims to gain insights, improve decision-making to resolve educational issues, and provide interventions in the learning process based on data information (Siemens, 2011). Like EDM, LA also follows a three-step procedure: data gathering, information processing, and knowledge application (Elias, 2011).

EDM and LA communities share common interests and similar goals, which are to help and promote data-driven decision-making in education (Papamitsiou & Economides, 2014; Siemens & Baker, 2012). The overall procedures of EDM and LA both focus on data collection, analysis, interpretation, and report of results in order to eventually improve individual or organizational performance (Papamitsiou & Economides, 2014). Baker and Yacef (2009) summarized five major analysis techniques in the fields of LA/EDM, which include 1) prediction, 2) clustering, 3) relationship mining, 4) distillation of data for human judgement, and 5) discovery with models. Although the communities share these commonalities, they still retain differences in several aspects, such as the fields of emphasis, the types of discovery, and the adaptation focus (Siemens & Baker, 2012). Research in LA considers leveraging human judgement, while EDM focuses more on automated discovery. LA emphasizes the understanding of the whole system, whilst EDM has a greater focus on the individual component. LA research aims to empower instructors and learners, while EDM research emphasizes automated adaptation of educational tools. Siemens and Baker (2012) also mentioned the differences in analysis methods for these two communities. For example, Siemens and Baker (2012) listed social network analysis as the main techniques and methods in the LA field, while classification was listed as the main techniques in the EDM field.

Several literature reviews of LA/EDM were conducted from a different perspective. For example, Dutt et al. (2017) conducted a three-decade long literature review focusing on clustering algorithm and its usability in the EDM studies. Other data mining algorithms are out-of-scope for this review work. Papamitsiou and Economides (2014) highlighted four major directions of the LA/EDM research through 40 selected articles, which were pedagogy-oriented issues, learning context, networked learning, and recommending educational resources. Although this review categorizes the types of data analysis methods for the studies, the review does not provide detailed information about those methods. Sin and Muthu (2015) reviewed 90 publications in the LA/EDM field and reported findings separately in the LA and EDM fields. However, the findings reported from this review are lacking clear organization, which mix data mining methods, research purposes, and learning contexts together. Romero and Ventura (2010) conducted a review about EDM and provided comprehensive information from the study data type to the research questions.

The fast development of the LA/EDM field in the last decade has motivated us to conduct this systematic literature review to reveal the latest trend in this field. This paper aims to carry out a comprehensive analysis of the empirical research studies in LA/EDM and answer the following two specific questions:

RQ1: *What is the impact of applied LA/EDM research on education?*

RQ2: *What are the sample and methodological characteristics of applied LA/EDM studies?*

The first research question intends to indicate how LA/EDM projects help improve data-driven decision making, teaching services, and development and implementation of technology tools in education. The synthesis of this information can provide educators or researchers with a perspective about the way of utilizing data mining algorithms to achieve their research or instructional purposes. The second research question aims to reveal characteristics of previous LA/EDM related studies based on the research objectives, studies' learning environments, data preprocessing and analysis, and other related elements. Interpreting these results can indicate the research trends in the EMD/LA field and give insight about the further development of the field.

Method

This review protocol is refined based on the guidelines by Hart (1998) and Tranfield et al. (2003), which consist of five stages: 1) definition of research questions; 2) key terms for search; 3) data selection based on the inclusion and exclusion criteria and databases; and 4) data procedure; and 5) data synthesis and report. Search database and search criteria for this review are noted below.

Data Collection

Search Database

A comprehensive search was conducted in two stages. In the first stage, we used the key search terms "Educational data mining" OR "Learning analytics" to search through various databases, such as IEEE Xplore, ScienceDirect, JSTOR, EBSCO, ERIC, SpringLink, Google Scholar, and Web of Science and identified journals that published articles in the LA/EDM field. Based on the findings in this stage, we targeted the most influential journals in the educational technology field. These journals are peer-reviewed and have high impact factors according to the citation indexes (i.e., social science citation index). In the second search stage, we narrowed down our focus on these selected journals (Table 2), applied our inclusion and exclusion criteria, and assessed full-text articles to further confirm the eligibility and determined the final selected articles.

Data Inclusion and Exclusion Criteria

The inclusion criteria included:

1) Publications from 2008 to September 2020 in the selected journals.

2) Publications must address K12 or higher education level.

3) Publications report empirical studies that involve technology use in teaching or learning.

4) Publications used data mining techniques to the gathered information.

The exclusion criteria included:

1) Articles were published earlier than 2008 or later than September 2020 in the selected journals.

2) Articles were published in the unselected journals.

3) Research studies were conducted beyond K12 or higher education scope such as for the professional development.

4) Publications focused on literature review or any other types of review.

5) Publications did not apply or report any use of data mining algorithms in any phase of research.

6) Publications included studies which were irrelevant to the topic of LA/EDM.

Initially, we found 4,171 publications from 2008 to September 2020 from those databases. After narrowing down our search to the selected journals and using the key search terms, we extracted 688 articles ready for applying the inclusion and exclusion criteria and full-text scanning. Eventually, we selected 113 articles for the coding procedure and excluded the rest of articles that did not meet the inclusion criteria. Table 1 shows the search process for this review, and Table 2 gives the information of the selected journals names, initial search results, and the number of final selected articles.

Table 1

PRISMA Chart Exhibiting the Search Process

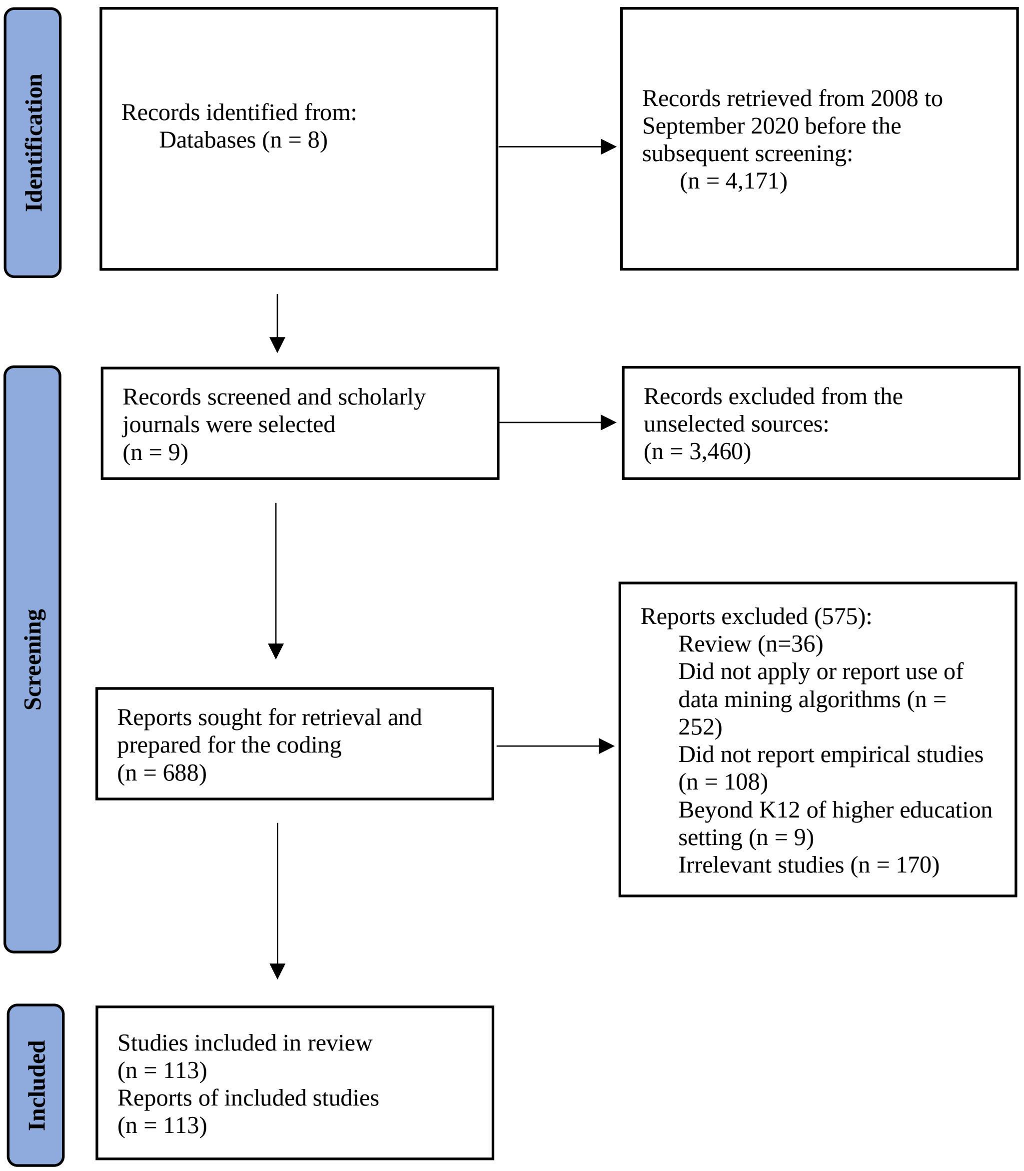

Table 2

Selected Journals and Search Results

Journal Name	Number of Articles from Initial Search	Number of Selected Articles for Coding
The Internet and Higher Education	32	14
Computers & Education	105	14
IEEE Transactions on Learning Technologies	44	15
British Journal of Educational Technology	138	14
Journal of Computer Assisted Learning	64	9
Journal of Computing in Higher Education	10	4
Educational Technology & Society	145	23
Educational Technology Research and Development	44	3
Interactive Learning Environments	106	17
Total	688	113

Data Analysis

Content analysis technique (Krippendorff, 2018) was applied to this study; moreover, this technique can use both quantitative and qualitive inquiries to reveal the content of communication from verbal discourse, written documents, and visual presentations (Krippendorff, 2018). This study applies both quantitative and qualitative content analysis, which presents the number of occurrences of the specific parameters within the selected articles to reveal the research topics, analysis methods, and other relevant information in the field of LA/EDM. The content analysis procedure for this study includes 1) developing a data encoding system, 2) implementing the encoding system to the selected articles, 3) organizing the coded information, and 4) interpreting the results. There are several main parameters included in the encoding system based on our research interests, which are 1) research objective, 2) learning environment, 3) education level, 4) data preprocessing methods, 5) data analysis methods, 6) data analysis tools, 7) sample size, and 8) variable information.

Research Objective

The research objectives are categorized and refined upon Papamitsiou and Economides (2014). These objectives include 1) discovering student engagement, 2) predicting student learning performance, 3) identifying student learning behavior, 4) enhancing assessment and feedback services, 5) increasing (self-) reflection and (self-) awareness, 6) building adaptive or recommender systems, and 7) improving the educational system.

Learning Environment

The learning environment refers to the context and environment where the learning experiences occur, which is mainly classified as below:

1) LMS. The LMS is an online portal that enables instructors to share classroom materials and support students to conduct online learning activities and virtually interact with each other (Adzharuddin & Ling, 2013).
2) MOOC. MOOCs provide learners with free and open online courses, involving open curricula and open-ended learning results (McAuley et al., 2010).
3) Intelligent systems. Intelligent systems refer to adaptive computer technology in education which scaffolds and supports student learning.
4) Games and simulations. Games and simulations refer to computer technology that provide students with learning experiences via gameplay or virtually simulated scenarios.

5) Other e-learning environments. Other e-learning environments include online assessment and scoring systems, e-book learning systems, recommender systems, standalone Wiki platforms, online learning communities outside of LMS, and other online e-learning platforms for different purposes.
6) Face-to-face classes. Attending face-to-face classes is the most common type for student learning.

Educational Level

The educational level for the selected studies focuses on either K12 or higher education.

Data Preprocessing

Data preprocessing is known as data preparation, which can be carried out through 1) attribution selection, 2) data cleaning, 3) transforming continuous attributes, and 4) data integration (Romero et al., 2004). Considering the complexity and different needs of data preprocessing, we further detail various data preprocessing activities as the following:

1) Balancing data. This activity refers to the handling of the imbalanced data where the number of observations is unequally distributed.
2) Normalization/standardization. It includes normalizing or standardizing the raw data points as well as transforming categorical or continuous attributes.
3) Dimension reduction. It refers to reducing the dimension via selecting features in a high dimensional data set.
4) Missing data. This activity refers to the approach of handling the missing data points.
5) Natural language preprocessing. This activity is an important step for natural language processing technique, which involves any tasks transforming text into a sort of features for the later analysis.

6) Generating new features. This includes aggregating or categorizing events, jointing all the data from multiple sources, and using other calculations to generate new features.
7) Outlier or noisy data removal. This activity removes outliers and other noisy data points before starting the analysis.
8) Absolute/relative measures. The original data feature was converted into the absolute/relative forms.

Studies which did not include any information about the data pre-processing tasks were marked as 'Not specified.'

Data Analysis Methods

Data analysis methods mainly include classification, clustering, regression, association rule mining, sequential pattern mining, text mining, and social network analysis. The basic definitions of these methods are listed below.

1) Classification – Classification is a supervised machine learning technique to categorize data to a given set of classes. Logistic regression is categorized into classification, considering its purpose of assigning observations to a discrete set (Wright, 1995).
2) Clustering – Clustering is an unsupervised machine learning technique that involves grouping of a set of objects based on their similarities through identifying the distance between them.
3) Regression – Regression is a supervised machine learning method that aims to predict values of a continuous response variable by examining the relationship with the input features.

4) Association rule miming – Association rule mining aims to discover correlations between variables from data sets.
5) Sequential pattern mining – Sequential pattern mining specializes in revealing sequential patterns in a sequence database.
6) Text mining – Text mining is a data mining technique to transform unstructured text data so as to identify the underlying patterns.
7) Social network analysis – Social network analysis is to identify behavior patterns and social structures of individuals based on the relation with each other on common interests.

Data Tools

Data tools used in data preprocessing and analyses are noted according to the selected articles. Several popular data mining tools are as the following:

1) R. R is a programming language and a free access software environment for statistical analyses and visualizations (Team, 2013).
2) Python. Python is a programming language and free open source that has been widely used for different purposes. In data science, Python greatly supports data analyses and applications (Van Rossum & Drake Jr, 1995).
3) SPSS. SPSS (Statistical Package for the Social Sciences) is a popular statistical program used in social sciences (Nie et al., 1975).
4) WEKA. WEKA (Waikato Environment for Knowledge Analysis) is a free access data mining software written in JavaScript developed at the University Waikato (New Zealand). This tool supports comprehensive data mining solutions and visualizations (Witten & Frank, 2002).

5) RapidMiner. RapidMiner is a data science platform written in JavaScript supporting multifaceted functions such as data preprocessing, data mining, prediction models development, and visualizations (Mierswa et al., 2006).

Sample Size

Sample size is the number of observations included in a study. We focus on the number of students participated in the research studies. When an article includes multiple case studies, the sample size is calculated using the average count of all the case studies.

Feature Information

Feature information relates to the type of input variables used for the analysis. User information includes user’s personal data, such as demographics, characteristics, and previous academic records. User learning activity tracks user’s learning actions occurring in the LMS or other learning environments, such as number of submitted assignments. User system action focuses on the interactions between the user and any tools used (e.g., times of login, total clicks). User dispositional data relates to emotions, motivation, self-regulation, perceptions, and attitudes during the learning process (Tempelaar et al., 2017).

Results

Quantitative Analysis

Distribution of the Publications

113 publications were selected and analyzed (Table A in Appendix). Figure 1 shows that the number of publications in the LA/EDM field has grown in the past few years. Specifically, most articles were published in the past five years. These research studies are conducted across 28 different locations. The top four locations for the studies are the United States (29%), Taiwan (10%), Australia (9%), and Spain (9%) (Table 3).

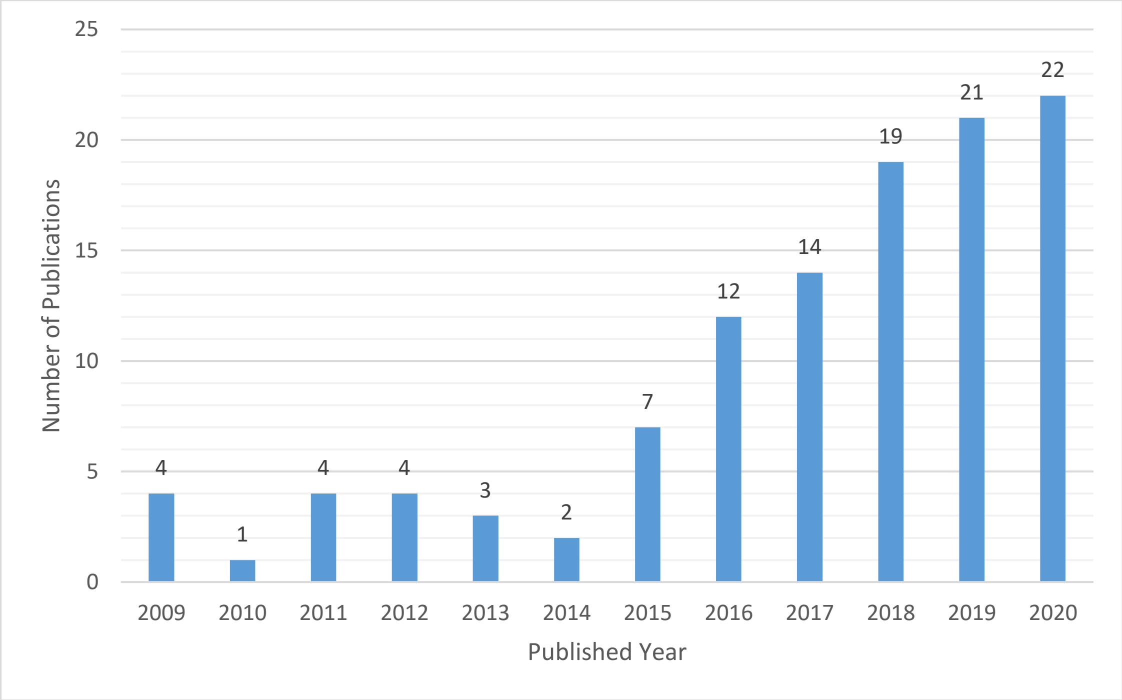

Figure 1. Number of Publications Included Each Year (n = 113).

Table 3

Top 4 Locations of the Studies

Study Location	Number of Studies	Ratio
United States	29	26%
Taiwan	11	10%
Australia	10	9%
Spain	10	9%

Educational Level and Sample Size

82% studies target higher education as the educational level (n = 93) and 15% studies focus on K-12 (Table 4). 3% studies are conducted both in higher education and K-12.

Table 4

Educational Levels of the Selected Studies

Educational Level	Number of Studies	Ratio
Higher Education	93	82%
K-12	17	15%
Both Higher Education & K-12	3	3%

Among the 113 publications, 107 studies reported sample sizes. Descriptive statistics of the studies' sample sizes is shown in Table 5. Although the range of sample sizes for the studies is large, 75% studies had 1,136 or fewer participants. The study with the largest number of observations is from Macarini et al. (2020), which used the data from the secondary school in Uruguay, reaching up to about 135,000 students. In contrast, the study of Shen et al. (2009) only had one single subject for the study but gathered data across many weeks of time.

Table 5

Descriptive Statistics of the Studies' Sample Sizes

	Average Value	Standard Deviation	25% Quantile	50% Quantile	75% Quantile
Number of Observations	4,545	16,171	97	238	1,136

Data Tool and Preprocessing

43% studies (n = 49) note the data tools used in the preprocessing or analysis stage (Figure 2). The popular data tools that researchers utilized include R (30%), SPSS (14%), and Python (13%). Other (27%) mainly included WEKA, RapidMiner, MATLAB, Java, and

JavaScript. Some studies used more than one tool to preprocess and analyze data sets (e.g., Lemay & Doleck, 2020).

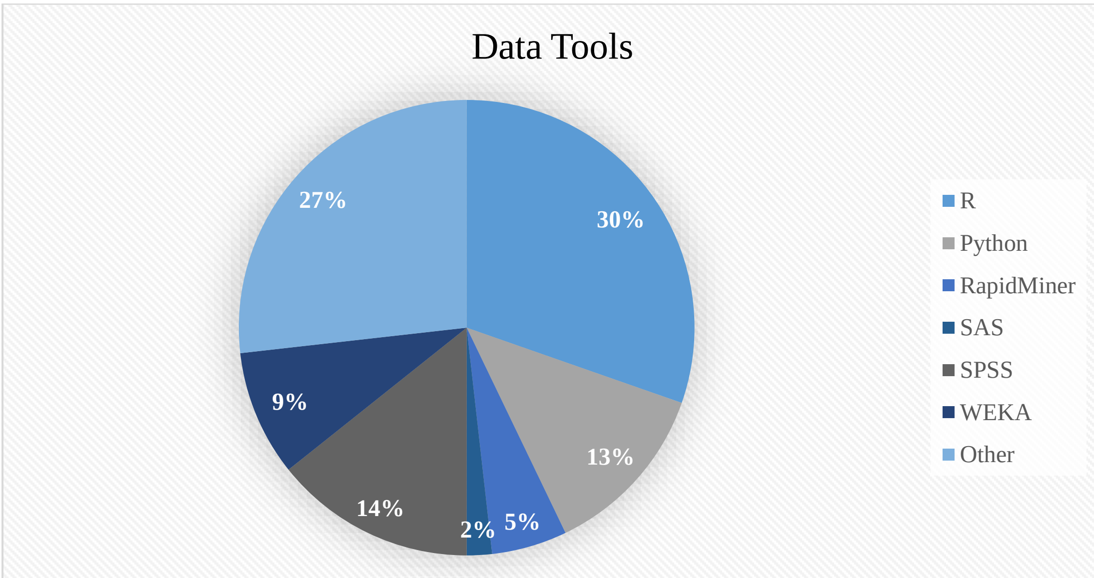

Figure 2. Data Tools.

76% studies (n = 86) reported their data preprocessing activities (Figure 3). In those studies, 35% studies (n = 39) used a combination of several preprocessing activities. The top 3 preprocessing activities are 1) generating new features (33%), 2) excluding outliers or noisy data (22%), and 3) normalizing or standardizing data before starting analysis (17%). Dimension reduction (10%) is also helpful, especially when the data size is large. Natural language preprocessing (10%) is an important step when the data set is in the text format.

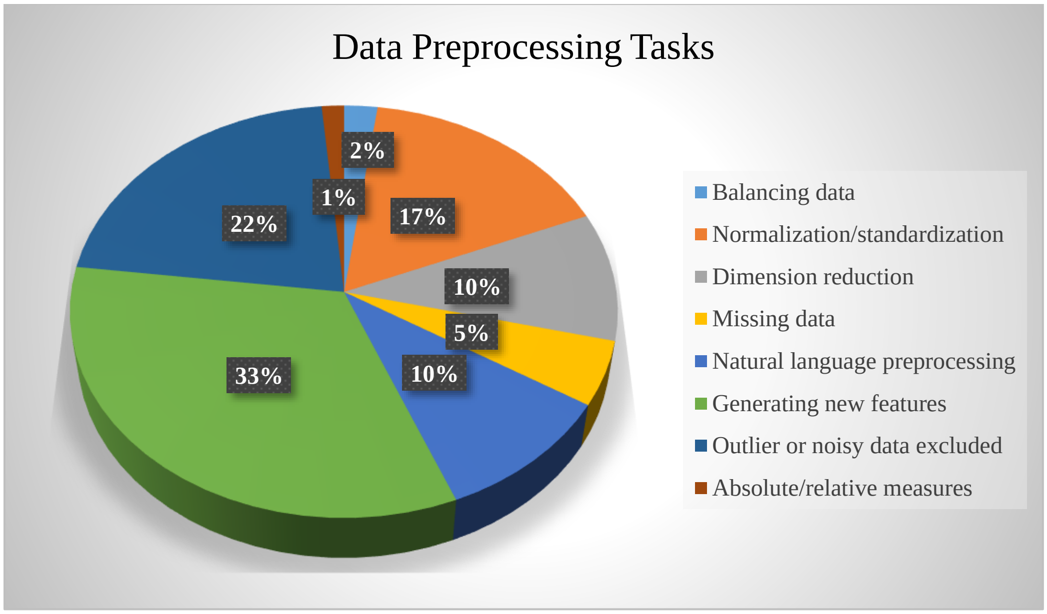

Figure 3. Data Preprocessing Tasks.

Qualitative Analysis

Research Objective

The majority of studies have focused on prediction of student learning performance, discovery of student learning engagement, and identification of student learning behavior (Table 6). Prediction of student learning performance includes 1) predicting dropout rates from a course (Olivé et al., 2020), 2) generating early alert from identifying at-risk students (Hung et al., 2019), and 3) forecasting achievements at different time points (Kostopoulos et al., 2019). Discovering student engagement includes 1) measuring student collaboration with peers to facilitate learning (Viswanathan & VanLehn, 2018), 2) revealing learner interests through the interactions with learning resources (Wu et al., 2018) or technology tools (Cerezo et al., 2016), and 3) examining student participation level in a course via assessing the communication structure with each other (Ergün & Usluel, 2016). The purposes of identifying student learning behavior are to 1) detect unusual behavior such as cheating (Ruiperez-Valiente et al., 2019), disengagement (Cocea &

Weibelzahl, 2011), procrastination (You, 2015); 2) unveil patterns indicative of learning strategies underlying the behavior (Fincham et al., 2019); and 3) develop adaptive computer-mediated systems to respond to learners based on their learning behavior (Chih-Ming & Ying-You, 2020).

The objective of increasing self-reflection and self-awareness is to increase teachers' or students' awareness on learning preferences (Clewley et al., 2011) and understand cognitive or affective states (Shen et al., 2009) in order to provide adaptive support and improve learning design. This objective can be pursued through the exploration of learning experiences in LMS, MOOCs, intelligent systems, or other technology tools. Enhancing assessment and feedback services aims to provide appropriate types of feedback to students or improve assessment for the purpose of increasing student learning success (Xing et al., 2015). This can be achieved through computer-mediated adaptive intelligent learning systems (Holmes et al., 2018) or in-game assessments (Cheng et al., 2017). Building adaptive or recommender systems is to adapt learning or provide recommendations tailored to learners' abilities with respect to their personalized activities (Khribi et al., 2009), recommended links to visit (Romero et al., 2009), content to review (Albatayneh et al., 2017), books to read (Chien et al., 2017), searching strategies (Liu, et al., 2013), and the selection of courses (Xu & Zhou, 2020).

Improving the educational system promotes data driven decision making and educational policies through the invaluable information retrieved from educational data sets. For example, a nationwide dataset was analyzed to detect the relationship between students' persistence in schools and learning performance on different subjects, such as History, Biology, Spanish, and English (Macarini et al., 2020). The visualizations of results in this study have shown the changes of students' performance for each year. These findings can be used to improve the educational authorities' decision making and implementation of policies. In addition, overlaps may occur between these research objectives. Therefore, some studies have two or more research objectives at the same time (e.g., Hu et al., 2018; Zhang et al., 2018).

Table 6

Research Objectives

Research Objective	Count	Author and Publication Year (Article Reference)
Discovering student engagement	21	Alario-Hoyos et al., 2016; Angeli & Valanides, 2013; Cerezo et al., 2016; Chen et al., 2018; Dawson, 2010; Ergün & Usluel, 2016; Hershkovitz & Nachmias, 2011; Howard et al., 2018; Hu et al., 2018; Khalil & Ebner, 2017; Mirriahi et al., 2016; Moon et al., 2020; Moore et al., 2019; Niemelä et al., 2020; Shibani et al., 2017; Sun et al., 2019; Viswanathan & VanLehn, 2018; Wu et al., 2018; Xie et al., 2018; Xie et al., 2014; Zhu et al., 2019
Predicting student learning performance	34	Abdous, Wu, & Yen, 2012; Alonso-Fernández et al., 2020; Asif et al., 2017; Bernacki et al., 2020; Cano & Leonard, 2019; Choi et al., 2018; Conijn et al., 2018; Conijn et al., 2017; Ellis et al., 2017; Er et al., 2019; Gašević et al., 2016; Gkontzis et al., 2019; Gray & Perkins, 2019; Huang et al., 2020; Hung et al., 2012; Hung et al., 2019; Junco & Clem,

		2015; Kostopoulos et al., 2019; Lee, 2015; Lemay & Doleck, 2020; Lu et al., 2018; Marbouti, Diefes-Dux, & Madhavan, 2016; Mubarak et al., 2020; Olive et al., 2019; Olivé et al., 2020; Ortigosa et al., 2019; Ruiz et al., 2018; Sandoval et al., 2018; Soffer & Cohen, 2018; Wakelam et al, 2020; Wu et al., 2020; Xing et al., 2019; Yu et al., 2017; Zacharis, 2015
Identifying student learning behavior	24	Abdous & He, 2011; Aguilar et al., 2019; Ahmad Uzir et al., 2020; Botelho et al., 2019; Brooks et al., 2014; Chang et al., 2015; Chih-Ming & Ying-You, 2020; Cocea & Weibelzahl, 2011; Codish et al., 2019; de Barba et al., 2020; Fincham et al., 2019; Geng et al., 2020; Howard et al., 2018; Jovanović et al., 2017; Kim et al., 2018; Li et al., 2020; Paquette & Baker, 2019; Park et al., 2016; Pereira et al., 2020; Riofrio-Luzcando et al., 2017; Ruiperez-Valiente et al., 2019; Valsamidis et al., 2012; Wang et al., 2017; Xia, 2020
Enhancing assessment and feedback services	7	Araya et al., 2012; Cerezo et al., 2020; Cheng et al., 2017; Holmes et al., 2018; Lin et al., 2013; Mensink & King, 2020; Xing et al., 2015
Increasing self-reflection and self-awareness	5	Cho & Yoo, 2017; Clewley et al., 2011; Li & Chen, 2009; Martín-García et al., 2019; Mouri et al., 2018

Building adaptive or recommender systems	9	Albatayneh et al., 2018; Chien et al., 2017; Hooshyar et al., 2018; Khribi et al, 2009; Liu et al., 2013; Mahnane, 2017; Romero et al., 2009; Shen et al., 2009; Xu & Zhou, 2020
Improve the educational system	1	Macarini et al., 2020
Multiple objectives	12	Chung & Paredes, 2015; Cela et al., 2016; Kim et al., 2016; Jo et al., 2016; Baker et al., 2016; Chen et al., 2019; You, 2015; Poitras et al., 2019; Schwarzenberg et al., 2020; Tempelaar et al., 2017; You, 2016; Zhang et al., 2018

Learning Environment

As shown in Table 7, most studies were conducted related to LMSs. Those studies collect log data from LMSs including students' learning activities and interactive actions within the platform systems. The popular research interests using LMS data include prediction of students' learning performance, identification of student learning behavior, and discover student engagement. For example, Hung et al. (2019) proposed a predictive modeling method to identify at-risk students who took the fully online courses hosted on the LMS. Fincham et al. (2019) collected learning trace data from a LMS to examine the relationship between students' behavioral patterns indicative of learning strategies and their academic performances. Kim et al. (2016) used log data to predict student performance and investigate student engagement in the online discussion.

Other e-learning environments are also noted by previous studies. For example, Abdous and He (2011) investigated problems that students encountered when using live video streaming.

Pereira et al. (2020) used an online evaluation platform CodeBench to score students' programme assignments. Some studies focus on recommender systems, which aim to provide learners with effective recommendations to improve their learning performance (e.g., Albatayneh et al., 2018; Romero et al., 2009). Online assessments and scoring systems can support researchers to understand students' learning behavior, interaction with the software, or predict student learning performance (e.g., Baker et al., 2016; Holmes et al., 2018). More examples of other e-learning environments include computer-mediated communication (e.g., Chih-Ming & Ying-You, 2020; Ergün & Usluel, 2016), hypermedia systems (e.g., Howard et al., 2018; Mirriahi et al, 2016), e-book learning systems (e.g., Mouri et al., 2018), and Wiki context (e.g., Hu et al., 2018). Another popular learning environment is MOOCs. The extensive amounts of data collected in MOOCs can be used to predict student learning performance (Conijn et al., 2018), identify students at risk of dropout (Olivé et al., 2020), discover student engagement using social tools in a MOOC (Alario-Hoyos et al., 2016), detect cheating behavior in MOOCs (Ruiperez-Valiente et al., 2019), and investigate learners' implicit learning attitudes (Geng et al., 2020).

Games and simulations, intelligent systems, and face-to-face classes can also be used in the LA/EDM studies. In the game and simulations context, previous studies deliver different research topics, such as tracking learners' engagement within the game (Moon et al., 2020), identifying student learning behavior (Li et al., 2020), predicting students' knowledge after gameplay (Alonso-Fernández et al., 2020), and building an adaptive system for the student learning (Hooshyar et al., 2017). Studies related to intelligent systems focus on modeling student behavior within the system (Paquette & Baker, 2019) and detecting any issues of student learning behavior when using the online tutoring system (Cocea & Weibelzahl, 2011). Some researchers have conducted LA/EDM studies via collecting data directly through face-to-face classes. For example, Choi et al. (2018) used the in-class students' responses data and student academic information to identify at-risk students. Gray and Perkins (2019) utilized students' weekly class attendance information to predict learning outcomes. Additionally, Codish et al. (2019) conducted two case studies: one used a gamified course on a LMS and the other used a MOOC. Martín-García et al. (2019) adopted an online survey to examine the stages of instructors' intentions of implementing the blended learning modality.

Table 7

Learning Environment of the Studies

Learning Environment	Count	Authors and Publication Year (Article Reference)
Face-to-face classes	10	Araya et al., 2012; Asif et al., 2017; Choi et al., 2018; Gray & Perkins, 2019; Macarini et al., 2020; Marbouti et al., 2016; Wakelam et al., 2020; Wang et al., 2017; Wu et al., 2020; Yu et al., 2017
Games and simulations	7	Alonso-Fernández et al., 2020; Cheng et al., 2017; Hooshyar et al., 2018; Li et al., 2020; Moon et al., 2020; Niemelä et al., 2020; Xing et al., 2019
Intelligent systems	5	Botelho et al., 2019; Cocea & Weibelzahl, 2011; Lee, 2015; Paquette & Baker, 2019; Riofrio-Luzcando et al., 2017
LMS	47	Aguilar et al., 2019; Ahmad Uzir et al., 2020; Bernacki et al., 2020; Brooks et al., 2014; Cano & Leonard, 2019; Cela et al., 2016; Cerezo et al., 2020; Cerezo et al., 2016; Chen et al., 2019; Chen et al., 2018; Cho & Yoo, 2017; Chung & Paredes, 2015; Clewley et al., 2011; Conijn et al., 2017; Dawson, 2010; Ellis et al., 2017; Fincham et al., 2019; Gašević et al., 2016; Gkontzis et al., 2019;

		Hershkovitz & Nachmias, 2011; Howard et al., 2018; Hung et al., 2012; Hung et al., 2019; Jo et al., 2016; Jovanović et al., 2017; Kim et al., 2016; Kim et al., 2018; Kostopoulos et al., 2019; Mensink & King, 2020; Mubarak et al., 2020; Olive et al., 2019; Olivé, Huynh, Reynolds, Dougiamas, & Wiese, 2020; Ortigosa et al., 2019; Park et al., 2016; Sandoval et al., 2018; Soffer & Cohen, 2018; Sun et al., 2019; Tempelaar et al., 2017; Valsamidis et al., 2012; Xia, 2020; Xie et al., 2018; Xie et al., 2014; Xu & Zhou, 2020; You, 2015; You, 2016; Zacharis, 2015; Zhang et al., 2018
MOOC	13	Alario-Hoyos et al., 2016; Chang et al., 2015; Conijn et al., 2018; de Barba et al., 2020; Er et al., 2019; Geng et al., 2020; Huang et al., 2020; Khalil & Ebner, 2017; Lemay & Doleck, 2020; Lu et al., 2018; Moore et al., 2019; Ruiperez-Valiente et al., 2019; Schwarzenberg et al., 2020
Other e-learning environments	29	Abdous & He, 2011; Abdous et al., 2012; Albatayneh et al., 2018; Angeli & Valanides, 2013; Baker et al., 2016; Chien et al., 2017; Chih-Ming & Ying-You, 2020; Ergün & Usluel, 2016; Holmes et al., 2018; Howard et al., 2018; Hu et al., 2018; Junco & Clem, 2015; Khribi et al., 2009; Li & Chen, 2009; Lin et al., 2013; Liu et al., 2013; Mahnane, 2017; Mirriahi et al., 2016; Mouri et al., 2018; Pereira et al., 2020; Poitras et al., 2019; Romero et al., 2009; Ruiz et al., 2018; Shen et al., 2009; Shibani et al., 2017; Viswanathan & VanLehn, 2018; Wu et al., 2018; Xing et al., 2015; Zhu et al., 2019
LMS & MOOC	1	Codish et al., 2019
None	1	Martín-García et al., 2019

Feature Information and Analysis Method

The studies gathered data from various sources, including log files extracted from the software, questionnaires, open data sets, and student records. Most studies use information retrieved from student learning activities (46%), student interaction with the software (30%), and personnel basic information (15%) (Figure 4). Some studies gathered student dispositional data (8%) that involve student cognitive and emotional states, sentiment information from self-evaluations, and learning motivation and perception (e.g., Moon et al., 2020; Yu et al., 2017). 75% studies (n = 85) used the combination of different types of information, such as collecting students' information, learning activities data and how they use the tools for learning (e.g., Howard et al., 2018). Other data information includes students' self-reported questionnaires about their online learning experiences (Ellis et a l., 2017).

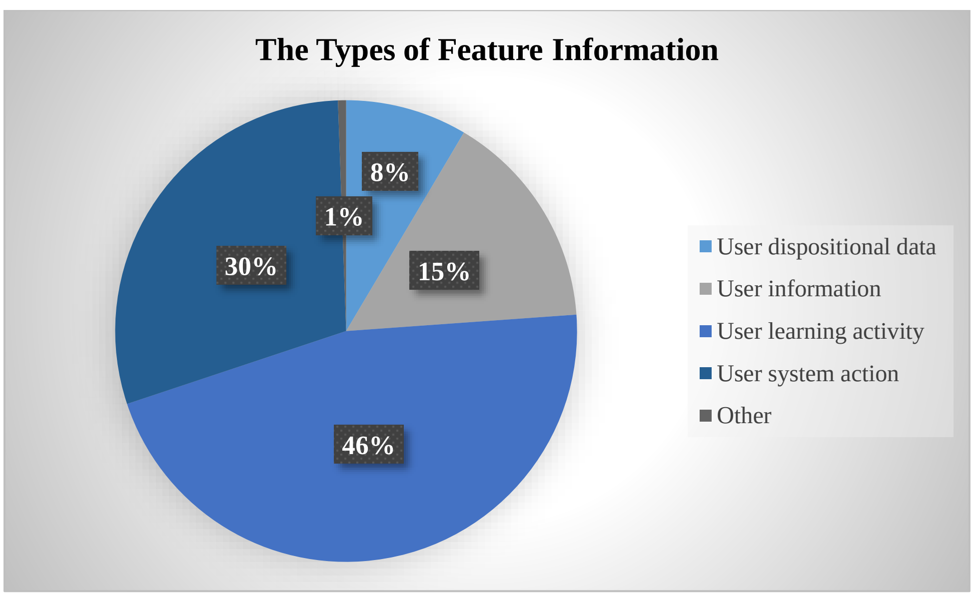

Figure 4. Feature Information Types.

Data analysis methods is another important parameter (Table 8). In the LA/EDM field, the most popular analysis method is classification, followed by clustering and regression. Other main methods include sequential pattern mining, association rule mining, social network analysis, and text mining. Some studies adopt only one analysis method (e.g., Lin et al., 2013; Ruiz et al., 2018), while some adopt two or more methods to achieve their research purposes (e.g., Martín-García et al., 2019; Pereira et al., 2020). Aside from these methods, Cerezo et al. (2020) applied Inductive Miner algorithm in process mining to discover student self-regulation learning process in an online course.

Classification. Classification algorithms have been used in previous studies to achieve different research objectives. For example, Neural Networks technique has been used to train models to predict at-risk students in the courses (Olive et al., 2019), identify learners' cognitive states (Holmes et al., 2018), and classify students posted messages (Wu et al., 2020). Logistic Regression (LR) can be used to predict student academic success (Lee, 2015) and completion of learning assignments (Lemay & Doleck, 2020). K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) methods can also be used to predict student emotions during learning (Shen et al., 2009), forecast student achievement (Chen et al., 2019), and identify at-risk students (Marbouti et al., 2016). Decision Tree (DT) has been applied to predict student online persistence in the web-supported courses (Hershkovitz & Nachmias, 2011), student academic success (Asif et al., 2017), and dropout risks (Ortigosa et al., 2019). Besides learning performance prediction (Hung et al., 2019), Random Forest (RF) can be used to detect student online cheating behavior (Ruiperez-Valiente et al., 2019).

A comparison of different classification algorithms has been carried out to identify at-risk students in higher education and K-12 online courses (Hung et al., 2019). Hung et al. (2019)

reported that Neural Networks and RF predict more accurately than SVM. Another comparison from Marbouti et al. (2016) noted that Naïve Bayes Classifier outperforms LR, SVM, DT, KNN, and Neutral Networks in the early prediction of at-risk students using standards-based grading information. Other classification algorithms comparisons are conducted to predict the possibility of student dropouts (Ortigosa et al., 2019); to predict student wheel spinning in the intelligent tutoring system (Botelho et al., 2019); to detect student disengagement in online learning (Cocea & Weibelzahl, 2011); and to classify collaboration ways with peers (Viswanathan & VanLehn, 2018).

Regression. Regression techniques have been used to solve different research questions. For example, Conijn et al. (2018) used multiple linear regression to predict students' final exam grades in a graduate-level MOOC. Zacharis (2015) employed a stepwise multivariate regression to predict student outcomes in the blended learning courses. Chih-Ming and Ying-You (2020) compared linear regression with other algorithms to model student learning behavior in order to predict computer-mediated communication competences. Tempelaar et al. (2017) predicted student use of digital tools via student dispositional information using hierarchical linear regression. Kostopoulos et al. (2019) predicted undergraduate students' final exam grades of a distance course using linear regression and classification algorithms. A comparison among Random Forest, linear regression and robust linear regression was conducted to predict student performance using LMS usage data, student information and academic records (Sandoval et al., 2018). Choi et al. (2018) compared the results of identifying at-risk students by using hierarchical linear regression versus hierarchical logistic regression. The findings have suggested that using hierarchical linear regression can provide more information about students' exam scores and yield better prediction models to detect at-risk students.

Clustering and sequential pattern mining. Studies have applied clustering solutions to group the subjects based on their learning behavior (e.g., Cerezo et al., 2016; de Barba et al., 2020). Clustering methods have been adopted to build recommendation systems (Khribi et al., 2009). Expectation Maximization (EM) algorithm and K-means are the two representative clustering algorithms used in previous studies (e.g., Brooks et al., 2014; Chang et al., 2015). Cerezo et al. (2016) claimed that EM and K-means could yield similar clustering results. Riofrio-Luzcando et al. (2017) compared three clustering algorithms (i.e. Xmeans, Expectation Maximization, Microsoft Sequence Clustering) versus without using clustering. The findings have shown that the differences between these methods are small but using sequence clustering method can yield the best outcome to divide students into groups. Different tutoring feedback can be accordingly provided to the clustered groups (Riofrio-Luzcando et al., 2017).

Other clustering algorithms include hierarchical clustering, Markov Clustering, and K-medoids. Ellis et al. (2017) used hierarchical clustering based on Ward's method (HCW) to find out two cluster solutions as the optimal to detect the association between student learning experience and academic performance. Fincham et al. (2019) also used the HCW approach, which generated four clusters to identify student learning strategies and the associations with academic success. Valsamidis et al. (2012) used Markov Clustering algorithm to separate 39 courses into two clusters, which were high activity courses and low activity courses. Kim et al. (2018) performed K-medoids clustering to identify three types of students' self-regulated learning profiles.

Sequential pattern mining can be used to understand student learning traces and examine the patterns of learning behavior. It can be applied with clustering analysis in the studies (Fincham et al., 2019; Jovanović et al., 2017). For example, Jovanović et al. (2017) used student learning sequences and their underlying learning strategies to cluster students via the HCW approach.

Association Rule Mining. This data mining approach is to reveal relationships between variables in a data set and present the patterns in the form following the "if-then" rules (Zhang & Zhang, 2002). Association rule miming has been applied to identify students' comment categories in Wiki in order to suggest appropriate early interventions and improve students' higher order thinking skills (Hu et al., 2018). Pereira et al. (2020) employed association rules analysis to identify effective learning behavior. Howard et al. (2018) applied this approach to identify students' use patterns of different mobile applications as to understand how this influenced their learning.

Social Network Analysis and Text Mining. Social network analysis (SNA) and Text mining are analytical techniques applied in social learning analytics (Buckingham et al., 2012). SNA aims to probe the relationship among individuals who build a social network with others. SNA employs visualizations to show the magnitude and density of interactions among nodes, which indicates the interactions among users (Zhang et al., 2018). The higher density of the network indicates its cohesiveness, which means individuals are more closely connected with others (Xie et al., 2014). SNA has been successfully deployed to detect interactive behavior when using the online communication tools. For example, SNA has been adopted to establish students' interactive modes using online forums (Zhang et al., 2018). Ergün and Usluel (2016) used SNA to detect student interactivity over a 14-week period discussion data and noted that instructors' participation greatly increased student interactivity.

A study has shown the development of a toolkit which used SNA to reveal student social interactions in the LMS discussion forums and provided visualizations of these interactions

(Chen et al., 2018). A learning analytics tool (VASCORLL 2.0) was built using SNA to bridge the gap between the eBook learning and the real-life scenarios as to help students better apply their knowledge into the practical situations (Mouri et al., 2018).

Another method is text mining which is the process of extracting the invaluable information from written resources. Text mining can help identify learning activities and tools preferred by the students based on students' posts on the social media (Aguilar et al., 2019). SNA and text mining have been adopted together to analyze the text data in the studies in order to measure and detect team leadership (Xie et al., 2018), probe the influence of students' interactions via the online posts (Aguilar et al., 2019), and examine students' cognitive and emotional engagements in learning (Geng et al., 2020).

Table 8

Classification of Data Analysis Methods

Data Analysis Method	Count	Authors and Publication Year, Article Reference;
Regression	10	Alario-Hoyos et al., 2016; Chih-Ming & Ying-You, 2020; Conijn et al., 2018; Conijn et al., 2017; Junco & Clem, 2015; Lu et al., 2018; Moore et al., 2019; Sandoval et al., 2018; You, 2015; You, 2016
Classification	39	Asif et al., 2017; Baker et al., 2016; Bernacki e t al., 2020; Botelho et al., 2019; Cano & Leonard, 2019; Chen et al., 2019; Cheng et al., 2017; Clewley et al., 2011; Cocea & Weibelzahl, 2011; Er et al., 2019; Gašević et al., 2016; Gkontzis et al., 2019; Gray & Perkins, 2019; Hershkovitz & Nachmias, 2011; Holmes et al., 2018; Hooshyar et al., 2018; Huang et al., 2020; Hung et al., 2019; Kim et al., 2016; Lee, 2015; Lemay & Doleck, 2020; Li & Chen, 2009; Lin et al., 2013; Marbouti et al., 2016; Mensink & King, 2020;

		Mubarak et al., 2020; Olive et al., 2019; Olivé et al., 2020; Ortigosa et al., 2019; Paquette & Baker, 2019; Poitras et al., 2019; Ruiperez-Valiente et al., 2019; Ruiz et al., 2018; Shen et al., 2009; Viswanathan & VanLehn, 2018; Wakelam et al., 2020; Xing et al., 2019; Xu & Zhou, 2020; Yu et al., 2017
Clustering	15	Brooks et al., 2014; Cerezo et al., 2016; Chang et al., 2015; de Barba et al., 2020; Ellis et al., 2017; Howard et al., 2018; Khalil & Ebner, 2017; Li et al., 2020; Liu et al., 2013; Mirriahi et al., 2016; Niemelä et al., 2020; Park et al., 2016; Riofrio-Luzcando et al., 2017; Valsamidis et al., 2012; Xing et al., 2015
Association rule mining	2	Hu et al., 2018; Xia, 2020
Sequential pattern mining	6	Chien et al., 2017; Fincham et al., 2019; Jovanović et al., 2017; Sun et al., 2019; Wang et al., 2017; Zhu et al., 2019
Social network analysis	8	Cela et al., 2016; Chen et al., 2018; Chung & Paredes, 2015; Dawson, 2010; Ergün & Usluel, 2016; Mouri et al., 2018; Xie et al., 2014; Zhang et al., 2018
Text mining	4	Abdous & He, 2011; Albatayneh et al., 2018; Geng et al., 2020; Wu et al., 2018
Other, i.e., Process mining;	1	Cerezo et al., 2020
Multiple methods	30	Abdous et al., 2012; Aguilar et al., 2019; Ahmad Uzir et al., 2020; Alonso-Fernández et al., 2020; Angeli & Valanides, 2013; Araya et al., 2012; Cho & Yoo, 2017; Choi et al., 2018; Codish et al., 2019; Fincham et al., 2019; Howard et al., 2018; Hung et al., 2012; Jo et al., 2016; Jovanović et al., 2017; Khribi et al., 2009; Kim et al., 2018; Kostopoulos et al., 2019; Macarini et al., 2020; Mahnane, 2017; Martín-García et al., 2019; Moon et al., 2020; Pereira et al., 2020;

Romero et al., 2009; Schwarzenberg et al., 2020; Shibani et al., 2017; Soffer & Cohen, 2018; Tempelaar et al., 2017; Xie et al., 2018; Wu et al., 2020; Zacharis, 2015

Discussion

RQ1: *What is the impact of applied LA/EDM research on education?*

The knowledge discovered from the studies in this review shows three main benefits of implementing LA/EDM: 1) improving institutions' decision making, 2) enhancing teaching services, and 3) improving the development of educational technology tools. First, these studies can help institutions consider how to utilize data and employ data mining methods to achieve data-driven decision making. Data-driven decision making is the systematic procedure of collecting, analyzing, and applying various data sets in order to improve student academic success and institutional effectiveness (Marsh et al., 2006). It has been recognized by the regulation such as the American Recovery and Reinvestment Act of 2009 (Act, 2009), which illustrates its expectation of using data to inform policy and practice.

Previous studies have shown the factors that are most significant for students' academic success (e.g., Baker et al. 2015). Although there is no one-size-fits-all prediction model, the studies have shown the impact of the early prediction on the instruction and administration. Specifically, it can trigger instructors to adopt early interventions to avoid student learning failure, which eventually can help meet institutional retention goals (Gray & Perkins, 2019). The results of the early prediction can also reveal the most relevant factors that institutional stakeholders should consider when making any curriculum change or improving course design as well as pedagogical practices (van Leeuwen, 2018; Weng et al., 2020). Besides predicting student performance, other research purposes in the LA/EDM studies, such as identifying learning behavior or discovering engagement, can also support instructors to better understand students' learning states and detect more reasons regarding student success (e.g., Schwarzenberg et al., 2020).

On the other hand, the implementation of LA/EDM can help improve the development of educational technology tools internally or externally. Educational technology tools include but are not limited to data warehouses, student records systems, LMSs, assessment systems, educational games, intelligent systems, and other systems that help teachers and students. From an internal approach, for example, an adaptive personalization system can be developed and integrated with a recommender engine using students' learning sequences data (Romero et al., 2009). In this situation, student usage data was systematically applied to develop the system, refine the system's adaptivity performance, and improve its recommendation accuracy. From an external approach, the implementation of LA/EDM evaluates the effectiveness of using a tool in learning. For instance, previous studies have predicted student learning performance after using tutoring systems (e.g., Lee, 2015). The prediction results can evaluate how the tutor has supported learning as well as provide an insight on what improvement the tutor still needs in order to provide a more effective automatic tutoring service.

Additionally, LA/EDM information can provide data-driven perspective of integrating the existing learning theories and technologies in practice. The findings from previous LA/EDM studies can be integrated into course design practices and help develop or improve educational technologies in the field. For example, previous studies have analyzed student learning interaction with each other in discussion forums via social network analysis to detect their course participation levels (e.g., Ergün & Usluel, 2016). Students who were in the low participation level can be identified through this analysis, so that instructors would be able to provide timely intervention. Increasing student learning interactions with each other is supported by a new learning theory, Connectivism, proposed by Siemens (2017). Connectivism notes the significant trends in learning in the last two decades since technologies have changed every aspect of living, communicating, and learning. Connectivism addresses learners' abilities of critical thinking, information search and discernment, decision-making, building connections between ideas and concepts, and corporations in activities (Siemens, 2017). Instructors and instructional designers can think of different methods to increase student learning interactions in the course design process. Previous findings from LA/EDM studies also support the learning framework called Community of Inquiry (Garrison et al., 2010). This framework introduces three elements of generating educational experiences, which include social, cognitive, and teaching presence (Garrison et al., 2010). The social element shows the importance of learning interactions between students and with instructors. The cognitive element shows student cognitive process during the learning experience, from trigging a problem to applying the knowledge to real scenarios. The teaching presence consists of providing different methods of learning content and conducting learning activities. These three elements can be used as underpinnings not only for pedagogy, but also for developing an educational game, intelligent tutoring system or other adaptive learning system. Previous LA/EDM findings have noted the adaptive systems used to support student learning (e.g., Albatayneh et al., 2018). These systems were data-driven and developed by considering students' learning behavior and cognitive process in learning. The data collected within the systems can be evidence of supporting the existing theoretical frameworks as well as updating the theories with new findings.

RQ2: *What are the sample and methodological characteristics of applied LA/EDM studies?*

The above results have shown that LA/EDM studies cover varieties of research objectives. Among those research purposes, most studies have focused on predicting student learning performance. This prediction is beneficial for stakeholders, including institutions, instructors, and students (Weng et al., 2020). Other research objectives can also help understand student learning from a different perspective. For instance, some LA/EDM studies have explored based on their log trace data how students interact with learning tools and how they react to learning when using the tools (Chih-Ming & Ying-You, 2020). These studies have collected data from LMSs, MOOCs, games, intelligent tutoring systems, face-to-face classes, and other e-learning environments. Regarding data collection, the range of sample sizes based on the previous studies is quite large, which indicates that there is no restrict requirement on sample sizes when conducting LA/EDM studies.

Unlike other fields emphasizing the big data's volume, big data for education mainly refers to school administrative data and student learning process data (National Academy of Education, 2017). It aims to mine student learning information and provide insights of learning performance and approaches (West, 2012). After data collection, data preprocessing techniques are applied before analyses. The most popular data preprocessing tasks include generating new features, excluding outliers, and normalizing or standardizing data. Data preprocessing is an important step, which has shown its benefits on improving the analysis accuracy (Chandrasekar et al., 2017).

From the results above, data mining algorithms have been used to support various research needs in education. The most popular data mining methods are classification, clustering, and regression. These methods can be applied to different student data, ranging from student academic data including learning activity and interaction using tools to personal records such as previous academic performance before enrollment. Student dispositional information (e.g., student cognitive and emotional states, learning attitude) is also considered in the previous studies (e.g., Clewley et al., 2011). Different data mining algorithms have their own advantages and disadvantages that should be considered when being applied to research. For instance, RF is a well-known “black-box” data mining method in which the predictors in a model are not transparent in their impact direction but usually outperforms than the “white-box” methods that make the models more visible (Villagrá-Arnedo et al., 2017). This limitation indicates that RF models cannot provide a full insight on the relationships among predictors and response variables. Therefore, it poses a challenge of identifying the exact reasons behind learning issues and developing appropriate interventions to improve student success even after identifying at-risk students or the possible course dropout. Another algorithm, Neural Networks, is a deep learning solution that has demonstrated predictive power in the LA/EDM studies. This technique normally requires big data sets and may lead to mislabeling errors in adversarial data (Szegady et al., 2013).

Overall these LA/EDM studies have shown great potentials of this field and revealed the directions of educational research. More institutions may focus on data-driven decision making to improve academic programs in terms of increasing student retention and enrollment. On the other hand, educational technologies will be more adaptive to support student learning. To achieve these goals, future studies in the field may address several concerns regarding data collection, data analysis, and data management. First, as more data are generated and collected, data quality based on project goals needs assurance. Currently, few studies in LA/EDM field have addressed and generated a comprehensive standard metric measuring data quality. The unknown status of data quality may yield inaccurate results and undermine data value. Second, current algorithms may need further development to fit education contexts. Big data in education has its uniqueness and complexity compared with other fields. For example, educational data can be multilevel, such as individual level, school level, and country level. Student learning performance can be influenced by not only student-level factors, but also factors in other levels. However, most data mining algorithms currently only consider one level of data, which could overestimate or underestimate the results. Therefore, studies developing advanced data mining algorithms fitting education contexts are expected. Additionally, a great concern has been raised in terms of data management and governance. Data ownership, privacy and security need clearly addressing. Will students be able to access their data? How long will students' data remain? Those related questions should be addressed when planning and implementing any LA/EDM projects. Although some studies have mentioned these concerns (e.g. Rubel & Jones, 2016), few empirical studies have noted the practical and consistent solutions for these questions. Future research is expected to provide a comprehensive standard practical guideline or template regarding data management and governance.

CHAPTER III

DATA MINING TECHNIQUES AND MIXED EFFECTS METHODS

Introduction

Hierarchical or clustered data has multilevel sampling with observations, including the lower-level units (individuals) nested within the higher-level units (clusters). This type of data involves individual-level attributes and cluster-level attributes to probe the variations among individuals within and between different clusters. Observations within the same cluster tend to have more similarities than from different clusters. Understanding both similarities and differences across clusters may lead to more accurate results. Clustered data sets are common in educational research. For example, Programme for International Student Assessment (PISA) data, measuring fifteen-year-old students' reading, mathematics, and science achievements, has a clustered structure and has been studied by scholars (e.g., Hu et al., 2018; Park & Weng, 2020).

Tree-based methods were proposed known as classification and regression trees (i.e., CART) by Breiman et al. (1984). CART is non-parametric and allows to handle big data with infinite attributes without being selected in advance. CART can robustly handle outliers, compared to some traditional statistical methods such as linear regression. However, CART may yield unstable results in some circumstances (e.g., the modification of observations), which can lead to high variability and poor predictive performance (Hastie et al., 2009). To improve this situation, a tree-based ensemble method was proposed – Random Forest (RF) by Breiman (2001). RF ensembles a large number of regression trees to improve predictions as its goal. RF has been applied to educational research to predict students' learning performance (e.g., Sandoval et al., 2018). However, RF only considers fixed effects of attributes even when the data structure is clustered. Considering this situation, a new method established upon the CART was proposed, which was called the random effects expectation minimization recursive partitioning method (RE-EM tree) (Sela and Simonoff, 2012). Later, another mixed effects Random Forest approach (MERF) was proposed, which adds random effects to RF (Hajjem et al., 2014).

This paper focuses on a survey of various tree-based data mining algorithms and hierarchical linear modeling (HLM), one of the most popular approaches applied to clustered educational data sets. The comparative study explores non-mixed effects tree models (i.e., RF) versus mixed-effects tree models (i.e., RE-EM tree, MERF) as well as HLM approach. This comparison will reveal advantages and disadvantages of each method, which can provide insights into the selection and adoption of these methods in clustered educational data sets.

In the following sections of the paper, I briefly cover non-mixed-effects tree-based method (i.e., RF), mixed-effects tree-based methods (i.e., RE-EM tree, MERF), and HLM approach. I then present a comparative study to discover the optimal method using PISA 2018 clustered data set. Finally, I report the results, discuss the findings, implications, and limitations.

Theoretical Framework

Tree-based Method: Random Forest

Random Forest (RF), introduced by Breiman (2001), has been widely used for prediction and classification (e.g., Fernández-Delgado et al., 2014), even in high-dimensional settings (Chen & Ishwaran, 2012). RF is a forest of regression trees that integrate the bagging procedure with randomization in splitting variables. Bagging (Breiman, 1996) generates random bootstrap samples from the original data. The bootstrap samples are repeatedly drawn from the original data pool and have the same length as the original data. Each tree is established upon selecting random features from bootstrap samples. The RF predictions are decided by averaging the output of the trees. The biggest challenge of RF is its difficult interpretation due to the combination of regression trees. However, RF can still indicate the relevance of input attributes. The out-of-bag observations are not part of the bootstrap sample when training a model. These observations yield the out-of-bag error to evaluate the accuracy of a RF and to select optimal values for tuning parameters, such as the number of candidate attributes that are randomly drawn for a split (Breiman, 1996).

Mixed Effects Methods

Hierarchical Linear Modeling

Hierarchical linear modeling (HLM), also referred to multilevel modeling, is widely employed in clustered data which has individuals (lower-level units) nested within clusters (higher-level units). This approach is frequently used in educational research, where sampling individuals are nested within classes and schools (e.g., Winitzky-Stephens & Pickavance, 2017). In a two-level model, one level examines the relationship among the lower-level units, and the other detects how this relationship is varying across higher-level units (Woltman et al., 2012). Take a random intercept model as an example. The model can be written as:

$$y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \varepsilon_{ij} \qquad (1)$$

where:

y_{ij} = response variable value for the individual i nested within the jth cluster unit;

β_{0j} = intercept for the jth cluster unit;

β_1 = regression slope associated with the attribute X_{ij} for the jth cluster unit;

X_{ij} = attribute value of X for the individual i in the jth cluster unit;

ε_{ij} = random error for the individual i in the jth cluster unit.

In the model formula (1), β_{0j} can be written as:

$$\beta_{0j} = \gamma_{00} + U_{0j} \quad (2)$$

where:

γ_{00} = mean intercept across all clustered units, which is a fixed effect;

U_{0j} = a random effect of the jth cluster unit on the intercept.

A combined model can be created using Equation (1) and Equation (2):

$$y_{ij} = \gamma_{00} + U_{0j} + \beta_1 X_{ij} + \varepsilon_{ij} \quad (3)$$

$$\varepsilon_{ij} \sim N\ (0, \sigma_\varepsilon^2)$$

$$U_{0j} \sim N\ (0, \sigma_U^2)$$

In this random intercept only model, the parameters are estimated via the variance components σ_ε^2 and σ_U^2. σ_ε^2 represents the unexplained variation at the lower level when controlling the attribute X_{ij}, while σ_U^2 is the unexplained variation at the higher level.

RE-EM Tree

The random effects expectation-maximization recursive partitioning method (RE-EM tree) was proposed by Sela and Simonoff (2012) specializing for clustered and longitudinal data using CART (Breiman et al., 1984) as the underlying regression tree. In Sela and Simonoff (2012), we have sampling individuals or objects *i* = 1, ..., *I* at times *t* = 1, ..., T_i. An observation of an individual for a single time is referred as (*i*, *t*). An individual can have multiple observations across different times. For each observation, we have a vector of *j* attributes, $x_{it} = (x_{it1}, \dots, x_{itj})'$. The attributes may be constant among individuals over time or differ across time and individuals. To detect differences for individuals over time, we have a known design matrix Q_{it} and a vector of unknown individual-specific random effects intercept w_i being uncorrelated with the attributes. A general effects model can be written as:

$$\mathrm{y}_{it} = Q_{it}\mathrm{w}_i + \mathrm{f}(x_{it1}, \ldots, x_{itj}) + e_{it} \tag{1}$$

$$\begin{pmatrix} e_{i1} \\ \vdots \\ e_{iT_i} \end{pmatrix} \sim \mathit{Normal}\ (0, R_i) \tag{2}$$

and

$$\mathrm{w}_i \sim \mathit{Normal}\ (0, D) \tag{3}$$

The e_{it} are random errors that are independent and not associated with the random effects, w_i. R_i is a non-diagonal matrix that allows an autocorrelation structure within the errors for an individual. The RE-EM tree uses a tree structure to estimate f as well as the individual-specific random intercept w_i. Compared with a linear mixed effects model (where $f = x\beta$), the RE-EM tree has more flexible assumptions, which admit that the functional form of f is normally unknown. The RE-EM tree can also better handle with missing values and overfitting issues. The estimation process of a RE-EM tree is shown as below (Sela and Simonoff, 2012):

1. Initially set the estimated random effects, $\widehat{\mathrm{w}}_i$ to zero.
2. Run iterations through the steps a–c until the estimated random effects, $\widehat{\mathrm{w}}_i$, converge by considering change in the likelihood or restricted likelihood function being less than the tolerance value.
    a. Fit a regression tree to the data to predict the response variable using the attributes, $(x_{it1}, \ldots, x_{itj})$, for objects i = 1, ..., I at times t = 1, ..., T_i. The tree includes a set of indicator features, $I\ (x_{it} \in g_p)$, where g_p ranges over all of the terminal nodes in the tree.

b. Estimate the linear mixed effects model, $\mathrm{y}_{it} = Q_{it}\mathrm{w}_i + I\,(x_{it} \in g_p)\mu_p + e_{it}$ using the response variable and the attributes.

c. Extract the estimated random effects $\hat{\mathrm{w}}_i$ from the estimated linear mixed effects model.

3. Replace the predicted values of the response variable at each terminal node of the tree in the step 2a with the population-level predicted mean response $\hat{y}_i$ from the linear mixed effects model in step 2b.

Any tree algorithm can be applied to step 2a. Sela and Simonoff (2012) implemented the CART tree algorithm based on the R package – rpart in the step 2a and developed the R package, REEMtree. The RE-EM tree algorithm maximizes the reduction in sum of squares when splitting a node. Maximum likelihood or restricted maximum likelihood (REML) can be used in step 2b. The splitting process continues as long as the improvement in proportion of variability accounted for by the tree (termed complexity parameter), which determines the optimal size of the tree. In the example of Sela and Simonoff (2012), the value of complexity parameter (cp) was set at least 0.001, and the number of observations in the node was set at least 20. A 10-fold cross validation was applied to prune the tree once the initial tree was settled. The final split of the tree had the largest cp value and obtained the minimized validation error that was less than one standard error above the minimized value. The RE-EM tree allows for autocorrelation within individuals, which may yield more effective models comparing with no autocorrelation structure (Sela and Simonoff, 2012).

Mixed Effects Random Forest

Hajjem et al. (2011) extended the CART algorithm (Breiman et al., 1984) and proposed a mixed effects regression tree (MERT) approach for a continuous outcome to handle clustered data. MERT is to estimate the random components using the expectation-maximization (EM) algorithm and then apply a standard tree to estimate the fixed effects after removing the random component. This approach allows to examine the non-linearity between the fixed components and response values. To improve the prediction accuracy, Hajjem et al. (2014) developed a mixed effects Random Forest (MERF) where a regression tree is replaced by a Random Forest. Later, Hajjem et al. (2017) extended the MERT approach to non-Gaussian response variables and proposed a generalized mixed effects regression tree (GMERT) to solve classification problems.

The MERF algorithm can be defined as follows:

$$\mathrm{y}_i \ = \ \mathrm{f}(A_i) + Z_i w_i + e_i \tag{4}$$

$$w_i \sim N(0, D), \ \ e_i \sim N(0, R_i) \tag{5}$$

$$i \ = \ 1, \ldots, n_i \ , \tag{6}$$

where y_i = $[\mathrm{y}_{i1}, \ldots, \mathrm{y}_{in_i}]^T$ is the $n_i \times 1$ vector of responses for the n_i observations in the cluster i, A_i = $[\mathrm{A}_{i1}, \ldots, \mathrm{A}_{in_i}]^T$ is the matrix of fixed effects attributes, and $\mathrm{f}(A_i)$ is estimated using Breiman's Random Forest (2001). $Z_i \ = [\mathrm{Z}_{i1}, \ldots, \mathrm{Z}_{in_i}]^T$ represents the $n_i \times q$ matrix of random effects attributes for the cluster i, $w_i \ = \ (\mathrm{w}_{i1}, \ldots, \mathrm{w}_{in_i})^T$ is the $q \times 1$ matrix of random effects coefficients for the cluster i, and $e_i = \ (\mathrm{e}_{i1}, \ldots, \mathrm{e}_{in_i})^T$is the $n_i \times 1$ vector of errors. D is the covariance matrix of w_i, while R_i is the covariance matrix of e_i. In the MERF algorithm, $Z_i w_i$ is assumed linear with the response variable, the random component $Z_i w_i \ and \ e_i$ is assumed to be independent and normally distributed. The covariance matrix of the response is assumed to be V_i = Cov(y_i) = $Z_i D Z_i{}^T$+ R_i, and V = Cov(y) = diag(V_1,…, V_n), where y = $[y_1^T, \ldots, y_n^T]^{\ T}$. Another assumption is the between-clusters are independent. Fitting the MERF allows us to predict new observations in the clusters considering the cluster-level random effects. The correlation is assumed to occur only via the between-cluster variations, where R_i is diagonal ($R_i = \sigma^2 I_n$, i = 1,…, n).

The overall steps of the MERF algorithm, as described in Hajjem et al. (2014), can be outlined as follows:

1. Set r = 0 and the initial values for the parameters, which are $\widehat{w}_{i(0)} = 0$, $\widehat{\sigma}^2_{(0)} = 1$, $\widehat{D}_{(0)} = I_q$.

2. Set r = r + 1. Update the response corrected for the random effects $y^*_{i(r)}$, random forest of the fixed effects $\hat{f}(A_i)_{(r)}$, the random component $\widehat{w}_{i(r)}$:

(i) Set $y^*_{i(r)} = Z_i \widehat{w}_{i(r-1)}$, i = 1,…, n.

(ii) Build a RF with $y^*_{ij(r)}$ as the response and a_{ij} as the corresponding training set of attributes, i = 1,…, n, j = 1,…, n_j. The bootstrap training samples are repeatedly drawn from the training set ($y^*_{ij(r)}$, a_{ij}).

(iii) Estimate $\hat{f}(A_i)_{(r)}$ using the out-of-bag prediction of the RF, that is, estimate each $\hat{f}(a_{ij})$ using the bootstrap samples to build the trees not containing observation a_{ij}.

(iv) Set $\widehat{w}_{i(r)} = \widehat{D}_{(r-1)} Z_i{}^T \hat{V}^{-1}_{i(r-1)}(y_i - \hat{f}(A_i)_{(r)})$, i = 1,…, n, where $\hat{V}_{i(r-1)} = Z_i \widehat{D}_{(r-1)} Z_i{}^T + \widehat{\sigma}^2_{(r-1)} I_q$, for i = 1,…,n.

3. Update $\widehat{\sigma}^2_{(r)}$ and $\widehat{D}_{(r)}$ following

$$\widehat{\sigma}^2_{(r)} = \frac{1}{N} \sum_{i=1}^{n} \{\hat{e}^T_{i(r)}\, \hat{e}_{i(r)} + \widehat{\sigma}^2_{(r-1)} [n_i - \widehat{\sigma}^2_{(r-1)}\, tr(\hat{V}_{i(r-1)})]\}$$

$$\widehat{D}_{(r)} = \frac{1}{N} \sum_{i=1}^{n} \{\widehat{w}^T_{i(r)}\, \widehat{w}_{i(r)} + [\widehat{D}_{(r-1)} - \widehat{D}_{(r-1)}\, Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i \widehat{D}_{(r-1)}]\},$$

where $\hat{e}_{i(r)} = y_i - \hat{f}(A_i)_{(r)} - Z_i \widehat{w}_{i(r)}$.

4. Iterate the previous steps until convergence. Apply the generalized log-likelihood (GLL) criterion to confirm the convergence:

$$\mathrm{GLL}(f, w_i|y) = \sum_{i=1}^{n}\{[\,\mathrm{y}_i - \mathrm{f}(A_i) - Z_i w_i]^T R_i^{-1}[\mathrm{y}_i - \mathrm{f}(A_i) - Z_i w_i] + b_i^T D^{-1} w_i + \log|D| + \log|R_i|\}.$$

When predicting a new observation *j* from known cluster *i*, we can use the population-averaged RF prediction $\hat{\mathrm{f}}(A_{ij})$ and the random component $Z_i \widehat{w}_i$. If a new observation is from an unknown cluster not included in the sample, we use only the population-averaged RF prediction.

Methods

Data

This study employed the PISA 2018 data set from Organization for Economic Co-operation and Development (OECD). The PISA 2018 survey focused on 15-year-old students' knowledge and skills in the domains of mathematics, reading, and science across 79 participating countries and regions. Additionally, 52 countries distributed a questionnaire about student familiarity with the use of information and communications technologies (ICT). In this study, we merely focused on student reading competencies (PV1READ) as the response variable. After cleaning the missing values, two countries with different number of observations were chosen for this study. These countries were Kazakhstan (n_1 = 10,040) and the United States (n_2 = 2,592).

The study used 31 attributes including ICT related attributes, reading attributes, and other student relevant information to probe their relationships with student reading competencies. Table 9 lists these attributes and their brief descriptions.

Table 9

Attributes Information

Attribute Name	Description
PV1READ	Student reading performance score (WLE)
ICTHOME	ICT available at home
ICTSCH	ICT available at school
ICTRES	ICT resources (WLE)
INTICT	Student interest in ICT (WLE)
COMPICT	Perceived ICT competence (WLE)
AUTICT	Perceived autonomy related to ICT use (WLE)
SOCIAICT	ICT as a topic in social interaction (WLE)
ICTCLASS	Subject-related ICT use during lessons (WLE)
ICTOUTSIDE	Subject-related ICT use outside of lessons (WLE)
ENTUSE	ICT use for leisure outside of school (WLE)
HOMESCH	Use of ICT for schoolwork activities outside of school (WLE)
USESCH	Use of ICT at school in general (WLE)
PERFEED	Perceived Feedback from teachers (WLE)
EMOSUPS	Parental emotional support perceived by student (WLE)
LMINS	Learning time (minutes per week)
ESCS	Index of economic, social and cultural status (WLE)
UNDREM	Meta-cognition: understanding and remembering
METASUM	Meta-cognition: summarizing
METASPAM	Meta-cognition: assess credibility
HEDRES	Home educational resources (WLE)
STIMREAD	Teachers' stimulation of reading engagement perceived by student (WLE)
ADAPTIVITY	Adaptation of instruction (WLE)
TEACHINT	Perceived teacher's interest in teaching (WLE)
JOYREAD	Joy/Like reading (WLE)
SCREADCOMP	Self-concept of reading: Perception of competence (WLE)
SCREADDIFF	Self-concept of reading: Perception of difficulty (WLE)
PISADIFF	Perception of difficulty of the PISA test (WLE)
PERCOMP	Perception of competitiveness at school (WLE)
PERCOOP	Perception of cooperation at school (WLE)

ATTLNACT	Attitude towards school: learning activities (WLE)
BELONG	Subjective well-being: Sense of belonging to school (WLE)

Some attributes in PISA 2018 were generated by adopting the transformed weighted likelihood estimates (WLE) (Warm, 1989). The formula of transformation is as below:

$$W_t' = \frac{W_o - \bar{W}_{OECD}}{\sigma_{W_{OECD}}}$$

where W_t' is the final metric of the WLE scores after transformation, $W_o$is the original WLEs in logits, $\bar{W}_{OECD}$ is the mean score based on the equally weighted OECD country samples, and $\sigma_{W_{OECD}}$ is the standard deviation of the initial WLEs for the OECD samples.

The PISA 2018 applied plausible values for each student reading competency. Plausible values refer to a possible range of student competencies. Wu (2005) noted that " instead of obtaining a point estimate for θ, a range of possible values for a student's θ, with an associated probability for each of these values, is estimated. Plausible values are random draws from this (estimated) distribution for a student's θ. This distribution is referred to as the posterior distribution for a student's θ." (p. 116).

Some attributes adopted in this study were related to student engagement with teachers. For example, these attributes included teachers' stimulation of reading engagement (STIMREAD), the perceived teacher feedback (PERFREED), and the teacher's interest in teaching as perceived by the students (TEACHINT). Other attributes were related to student meta-cognition of reading, such as understanding and remembering (UNDREM), summarizing

(METASUM), assessing credibility (METASPAM), and student enjoyment of reading (JOYREAD). The learning related attributes also included learning time spent in test language (LMINS), student adaptivity of instruction in test language lessons (ADAPTIVITY), and student self-concept of reading (i.e. perception of competence in reading - SCREADCOMP, perception of difficulty in reading - SCREADDIFF). The student perception of difficulty of the PISA 2018 test (PISADIFF) was considered as well.

Regarding students' background information, several attributes were included in the analysis. For example, the index of student economic, social, and cultural status (ESCS) in PISA 2018 data set was computed by considering three aspects about family background. These aspects involve (1) parents' highest level of education, (2) highest occupational status (HISEI), and home possessions (e.g., number of books in the home). Other attributes included the household possessions such as home educational resources (HEDRES) and parental emotional support (EMOSUPS).

Considering the school effect on student learning, the attributes representing student perception of school environment were considered. These attributes were student perception of school competitiveness (PERCOMP), school cooperation (PERCOOP), attitude towards school (ATTLNACT), and school climate assessed with the scale on student sense of belonging to school (BELONG).

Data Analysis

Two countries' information were retrieved from the raw data set and separated as individual data sets. Two data sets were cleaned before analyses, including removing missing and noisy data points. Each sampling data set was then partitioned into 70% training and 30% testing datasets using random resampling without replacement within clusters. The training data sets were used to build RF regression, RE-EM tree, MERF, and HLM. The testing data sets were not involved in the model development phase but used to evaluate the models built in the training step. To apply RF regression, RE-EM tree, MERF, and HLM, each clustered sampling data set considered the fixed effects of all the selected attributes as well as the variation based on the schools.

Building a RF model

The *randomForest* package (Liaw & Wiener, 2002) in R (version 3.5.2) was applied to implement the RF algorithm. The following hyperparameters of RF were applied in the tuning process:

1) Number of trees (*ntreeTry*). The default setting of number of trees (*ntreeTry* = 500) was adopted. In this study, 500 trees were sufficient to produce solid results.

2) The *stepFactor* is the value by which the number of features sampled when constructing each tree (mtry) is inflated or deflated. This value was set as 1.5.

3) The improvement value in the minimum out-of-bag (OOB) error (*improve*) to continue the search was set as 0.01.

4) Number of features sampled when constructing each tree (*mtry*). The default value of *mtry* was calculated using the formula, *mtry* = number of attributes / 3. The starting value of *mtry* follows *mtry* = default value / *stepFactor*. The ending value of *mtry* follows *mtry* = default value * *stepFactor*. Therefore, we used tuneRF function to confirm the best value of *mtry* based on the OOB error. In both the Kazakhstan and USA data sets, the tuning process showed that *mtry* = 7 was the optimal value.

Building a RE-EM model

The *REEMtree* package (Sela & Simonoff, 2021) in R (version 3.5.2) was applied in the analyses. In the RE-EM tree analyses, 10-fold cross validation was applied when building the models, and complexity parameter (cp) was set as 0.01 for pruning the trees in order to select the optimal tree size based on the lowest cross validation error.

Building a MERF model

The *merf* package in Python (version 3.8) was used to run the MERF regression. In this study, we set 300 trees generated in the random forest and 50 as the maximum number of iterations until convergence for both sampling data sets.

Applying HLM

The HLM method was conducted in R (version 3.5.2) using the package *lme4* (Bates et al., 2015). The adjusted and conditional Intraclass Correlation Coefficient (ICC) was first run for each data set to estimate the variance explained by the school clustered structure. A random intercept model was employed for this study.

Evaluation Criteria

After generating the estimated models by running RF regression, RE-EM tree, MERF, and HLM, the testing data sets were used for evaluating the performance of the models. The measures, including the mean square error (MSE), mean absolute error (MAE), the mean absolute percent error (MAPE), and Accuracy (i.e. 100%*(1−MAPE)), were used to report the differences between the actual values and the predicted values and compare different model performance. These measures were successfully adopted in previous research studies (e.g. De Myttenaere, Golden, Le Grand, & Rossi, 2016). Below are the formulas of MSE, MAE, and MAPE:

$$MSE = \frac{1}{n}\sum_{q=1}^{n}(y_q - \hat{y_q})^2$$

$$MAE = \frac{1}{n}\sum_{q=1}^{n}|y_q - \hat{y_q}|$$

$$MAPE = \frac{1}{n}\sum_{q=1}^{n}|\frac{y_q - \hat{y_q}}{y_q}|$$

where n is the sample size, y_q is the actual value, $\hat{y_q}$ is the predicted value. The smaller values of MSE, MAE, and MAPE refer to the smaller differences between the estimated model and the actual situation, which indicates a better model performance.

Results

According to the results, the intraclass correlations in the baseline models for Kazakhstan and the United States are 0.387 and 0.15 accordingly. 38.7% of the variation in student reading achievement is attributable to school effects using the Kazakhstan dataset, and 15% of the variation in student reading scores is attributable to school effects using the United States dataset. According to the result of the random intercept model for the United States, seven ICT-related attributes showed significant effects on student reading achievement. These attributes were HOMESCH, INTICT, AUTICT, SOIAICT, ICTCLASS, ICTHOME, and ICTSCH. Three teacher related attributes were significant, which included PERFEED, STIMREAD, and TEACHINT. The significant impacts on student reading were caused by student reading-related attributes including UNDREM, METASUM, METASPAM, SCREADCOMP, and JOYREAD. Other significant attributes were EMOSUPS, HEDRES, ESCS, PISADIFF, PERCOOP, and BELONG. The overall HLM model reached 88.22% accuracy. Compared with the United States, the HLM model based on the Kazakhstan showed different significant attributes. For example, ENTUSE, USESCH, COMPICT, and ICTRES significantly influenced student reading scores, while HOMESCH, AUTICT, and ICTSCH were unsignificant. In addition, other attributes such as LMINS, ADAPTIVITY, and SCREADDIFF were significant when predicting Kazakhstan students' reading performance, while these attributes were unsignificant in the United States dataset. ESCS and BELONG were insignificant for Kazakhstan students' reading performance. Overall, the HLM model for Kazakhstan reached 89.8% accuracy.

According to the results of the RF models, 49.43% variance were explained by the model using the United States dataset, and 53.17% variance were explained in the Kazakhstan dataset. The top five most important attributes in the RF model using the United States dataset were METASPAM, PISADIFF, ESCS, JOYREAD, and METASUM. The RF model for Kazakhstan showed that METASUM, UDREM, PISADIFF, METASPAM, and SCREADDIFF were most important. The accuracy of the RF models for the United States and Kazakhstan were 92.61% and 93.72% respectively. The RF models performed better than the RE-EM Tree models, which were only 86.72% and 89.03% in accuracy for the United States and Kazakhstan datasets. Figure 5 and Figure 6 showed the RE-EM Tree models. The RE-EM Tree structure for the United States dataset were simpler than the Kazakhstan dataset. METASPAM, PISADIFF, and METASUM were the significant attributes contributing to the modeling structures for both datasets.

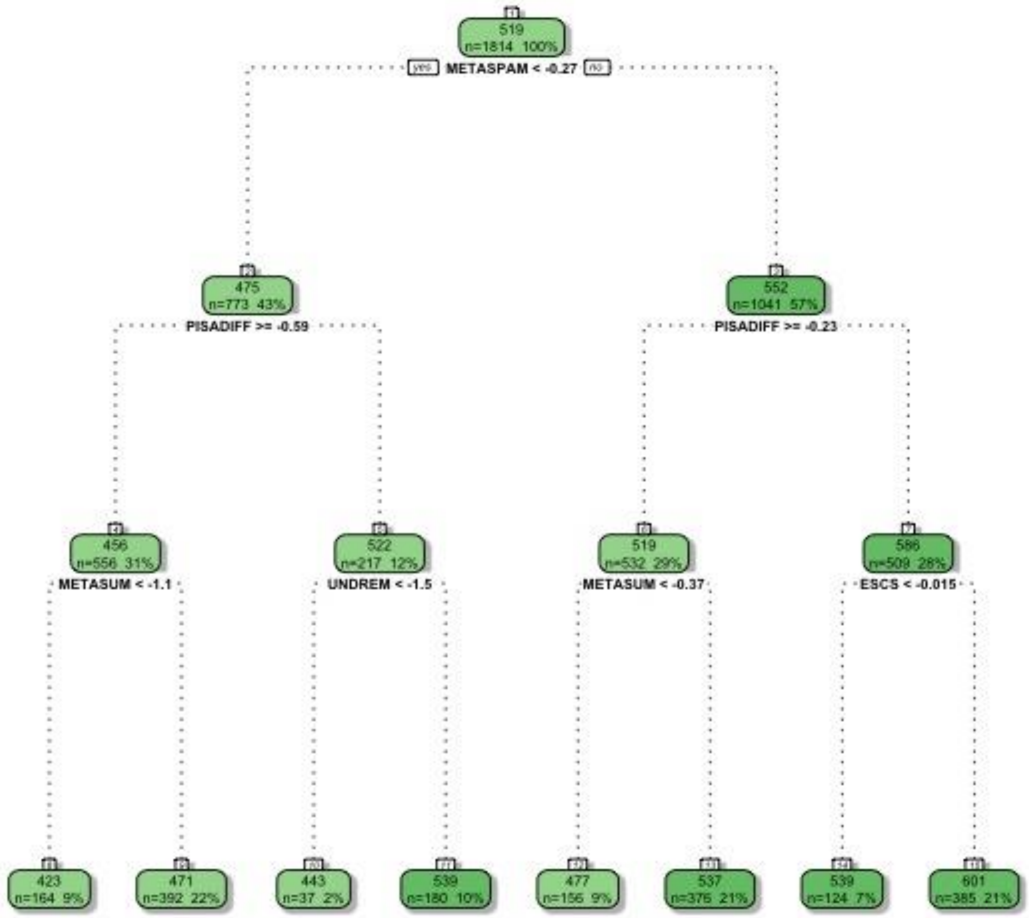

Figure 5. RE-EM Tree Model Result for the United States Data

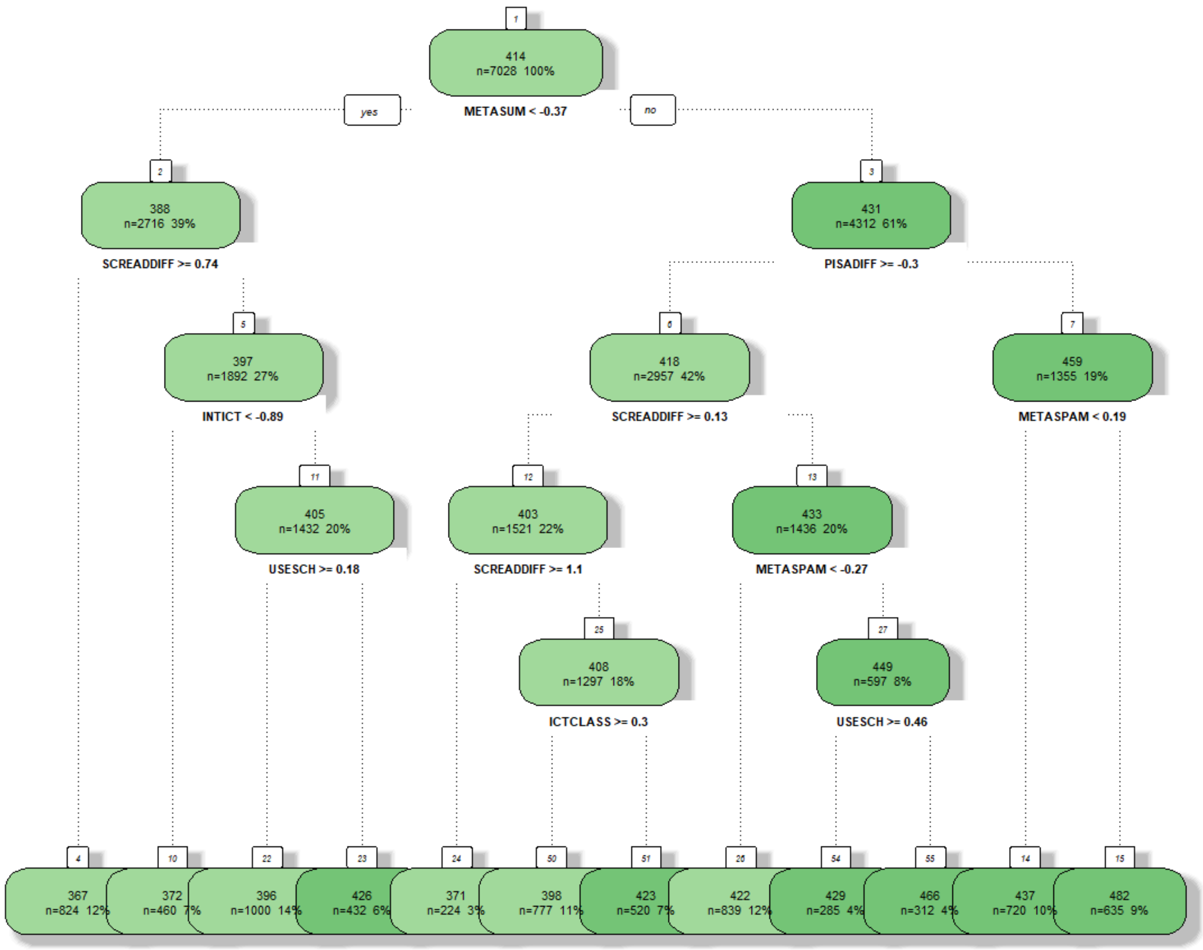

Figure 6. RE-EM Tree Model Result for the Kazakhstan Data

Regarding the MERF models, the models performed the best compared with other methods in both datasets (see Table 10 and Table 11). The MERF model using the United States dataset predicted students' reading performance with the accuracy of 93.16%, and the MERF model for Kazakhstan was 94.38% in accuracy. Other evaluation metrics for the MEFF models were the lowest values compared with other methods, which were consistent with the accuracy comparison results.

Table 10

The Evaluation Metrics Result of Each Model for the United States Data

	MSE	MAE	MAPE	ACCURACY
RF	2371.006	34.6963	0.0739	92.61%
RE-EM Tree	6238.66	62.8526	0.1328	86.72%
MERF	2207.5367	20.2245	0.0684	93.16%
HLM	4956.902	56.0686	0.1178	88.22%

Table 11

The Evaluation Metrics Result of Each Model for the Kazakhstan Data

	MSE	MAE	MAPE	ACCURACY
RF	1295.416	25.6777	0.0628	93.72%
RE-EM Tree	3227.529	45.0954	0.1097	89.03%
MERF	1143.1682	14.6682	0.0562	94.38%
HLM	2837.556	42.138	0.102	89.8%

As the Figure 7 and Figure 8 indicated, METASPAM, PISADIFF, and METASUM influenced students' reading performance among top five important attributes for both datasets. These results were consistent with the results from the RF models, though the MERF models slightly improved the accuracy in both datasets compared with the RF models.

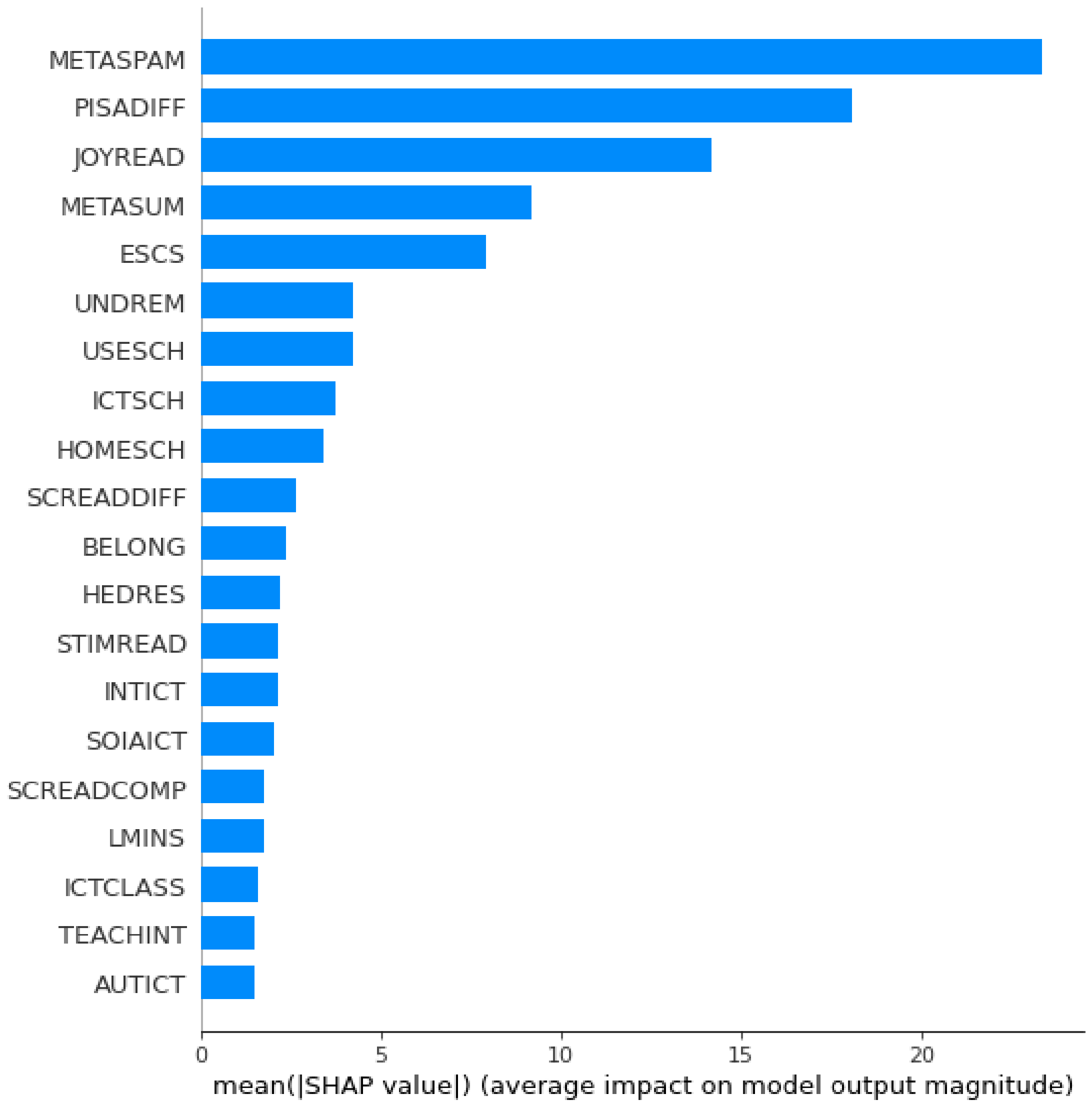

Figure 7. The Importance of Predictors in MERF Model for the United States Data.

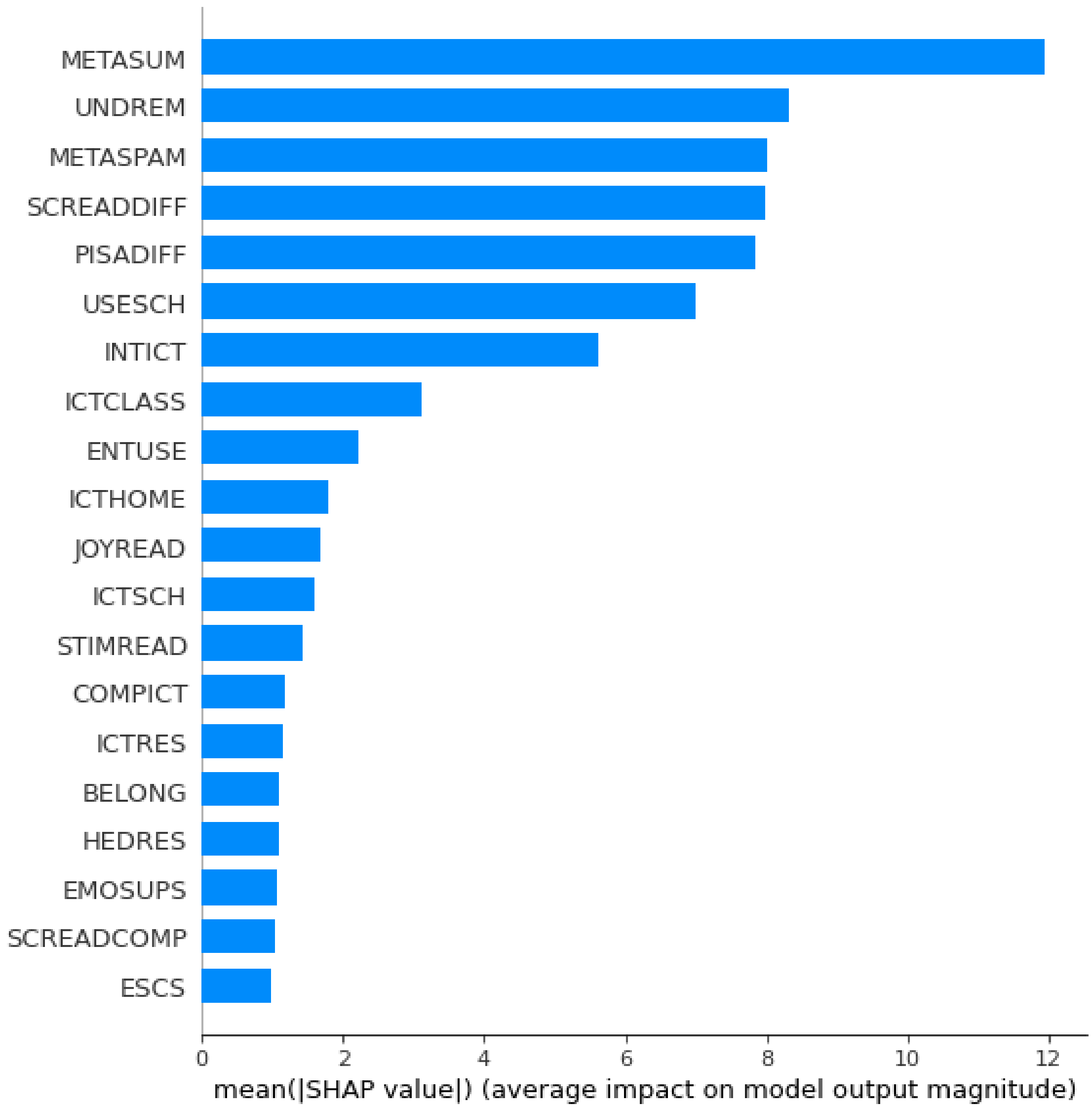

Figure 8. The Importance of Predictors in MERF Model for the Kazakhstan Data.

Discussion

MERF generated the most accurate models among all the methods applied to both the United States dataset and Kazakhstan dataset. MEFR inherits the advantages of RF method, which include 1) reducing the overfitting issue, 2) being less sensitive to data outliers, 3) easily setting parameters, and 4) automatically generating variable importance (Horning, 2013). MERF is more suitable and accurate than RF in clustering data because MERF considers both fixed effects and random effects of variables. Overall MERF using bagging scheme improves the accurate predictions, which can be helpful in predicting students' learning outcomes. Previous study by Pellagatti et al. (2021) developed a method of generalized mixed-effects Random Forest (GMERF) for classification and demonstrated the successful adoption of this method on predicting university student dropout.

However, MERF has a major disadvantage as RF, which is its 'black box' nature causing difficult interpretations of the relationships examined between predictor and response variables. The assembling tree structures confound the interpretation of each tree that does not allow to differentiate the exact directions and magnitudes of variables' impacts, though the model result can show the information of variable importance. Considering this major drawback, CART-based RE-EM Tree method is more friendly on interpreting the results of relationships among variables. RE-EM Tree inherits the advantages of both a regression tree algorithm and a linear mixed effects regression algorithm (Sela & Simonoff, 2012). RE-EM Tree method is robust to outliers because its tree splitting process can isolate outliers in individual tree nodes (Timofeev, 2004). In a high dimensional dataset, RE-EM Tree does not require preselected variables, which allows flexibility of data capturing. However, RE-EM Tree method may generate unstable decision trees due to different splitting ways that the tree structure can adopt.

Comparing data mining methods with HLM in the clustering educational data setting, data mining methods such as MERF and RE-EM Tree perform better for high dimensional data because they do not require specification of any functional form and can better handle missing data values. Depending on the purposes of research studies or applications, MERF and RE-EM Tree can be used in different settings. For example, when developing an early alert system of identifying student dropout or course grades, MERF or GMERF can be used in the system to yield accurate prediction results. MERF or GMERF may also have great potentials of being used in other related systems in the future. Previous studies have noted several technologies to predict students' learning performance, such as intelligent tutoring systems (e.g., Baker et al., 2011), educational games (e.g., Tadayon et al., 2021), and recommender systems (e.g., Thai-Nghe et al., 2010). On the other hand, when the main purpose is to examine the relationships among variables from a big data for education collected in technology systems or from multiple resources, RE-EM Tree can be more appropriate for being applied.

In addition, HLM is still a useful method for educational clustering data, especially when the data is not high dimensional and does not have serious issues of outliers or missing values. For example, Xu et al. (2018) applied HLM to detect the relationship between students' ICT and learning performances in mathematics, science and reading. Hew et al. (2020) adopted HLM to predict student satisfaction with massive open online courses. In a non-high dimensional dataset, this study has proved the advantage of applying HLM, which even showed a slightly higher accuracy than the RE-EM Tree model.

CHAPTER IV

EXPLORING THE INFLUENCE OF STUDENTS' ICT USE ON MATHEMATICS AND SCIENCE MODERATED BY SCHOOL RELATED FACTORS

Introduction

In the last decades, Information and Communication Technology (ICT) has dramatically influenced the way of sharing information and communicating with each other (OECD, 2005). The development of ICT has extended its impact to the realm of education by equipping classrooms and individuals with ICT tools to promote student achievement as well as enhance equal access to educational resources (UNESCO, 2015). The prevalence of ICT in education has increased the overall education quality (Murthy et al., 2015), stimulates initiative and creativity (Wheeler et al., 2002), and enables student learning personalization (Abell, 2006). These changes have motivated educators and researchers to explore the relationship between the use of technology and student learning. Previous studies have addressed the impacts of ICT related factors, such as students' interests in technology, frequent use of technology, ICT competence (e.g., Park & Weng, 2020; Skrybin et al., 2015). However, no consensus has been reached regarding the impacts of ICT related factors on student learning performance. For instance, some studies reported the positive correlation between using ICT for entertainment and learning achievements (e.g., Gumus & Atalmis, 2011), whilst other studies showed an insignificant impact (Bulut & Cutumisu, 2017).

Moreover, most of those studies emphasized on identifying student-level or country-level ICT factors (Park & Weng, 2020). Although school-level factors were considered in some studies (e.g., Gómez-Fernández & Mediavilla, 2018), those studies selected school-level factors based on researchers' subjectivity and interests, which can lead to the lack of full consideration in terms of school's background. Such selection bias can cause misleading causal inferences (Berk, 1983). Therefore, it is crucial to use a less subjective method in feature selection when considering school's background to examine the impacts of ICT factors on student academic performance. In this study, a data mining approach was used to select school-level factors based on their importance on student academic performance. The goal of this feature selection is to yield higher accuracy in the subsequent data analysis, which can accurately reveal hidden relationships.

Thus, this study aims to estimate the relationship of ICT related factors and student academic performance in mathematics and science and the moderating effects of school-level factors on the relationship based on the results of the Program for International Student Assessment (PISA) in 2018. Particularly, (1) each ICT-related factor effect on student mathematics and science achievement, (2) impacts of school-level significant factors on student learning achievement, and (3) the cross-level interaction effects will be probed.

Theoretical Framework

The Relationship Between Students' ICT Use and Academic Performance

ICT use at school or classroom addresses students' use of computers or other technologies for the educational purposes or to communicate with peers. Some studies have found the positive influence regarding the relationship between students' usage of ICT for education and their learning achievements (e.g., Luu & Freeman, 2011), while others reported negative or no significant relationships (e.g., Chiao & Chiu, 2018).

ICT interest refers to the attitude, emotion, and motivation of using ICT tools (Zylka et al., 2015). Although previous research showed that ICT interest is positively correlated with student engagement of using ICT (Hu et al., 2018), researchers have reported mixed findings regarding the impact of ICT interest on learning achievement. For example, some research reported that student ICT interest had positive influence on students' mathematics and science achievements (e.g., Hu et al., 2018; Meng et al., 2018). Nevertheless, some research reported that ICT interest was not a significant attribute on student learning scores (e.g., Juhaňák et al., 2018).

ICT competence is related to ICT knowledge and skills to perform the ICT tasks (Meng et al., 2018). Previous research showed conflicting findings of ICT competence and student learning performance. For example, positive relations between ICT competence and student mathematics scores were found in some studies (Hu et al., 2018; Martínez-Abad et al, 2018), whereas others revealed a null relationship (Meng et al., 2018; Juhaňák et al., 2018) or negative impacts between ICT competence and student academic achievement (Xiao et al., 2019).

ICT autonomy is the student control in the use of ICT (Fu, 2013). The relationship between student ICT autonomy and academic achievement was found to be consistent based on previous research findings. Research revealed positive associations between student ICT autonomy and learning performance (e.g., Hu et al., 2018; Juhaňák et al., 2018).

ICT as a topic in social interaction is one aspect of ICT engagement addressing "the extent to which students make ICT a subject of interpersonal communication and interaction" (Zylka et al., 2015, p. 151). This factor refers to the development of ICT skills in the informal learning contexts (Kunina-Habenicht & Goldhammer, 2020). Previous studies have reported contradictory findings between this ICT inclusion in social interaction factor and student learning performance. For example, a positive association was found between this ICT inclusion in social interaction and Spanish student mathematics scores in PISA 2015 studies (Martínez-Abad et al., 2018), while several other studies reported negative impacts of social interactions involving ICT on the mathematics scores (e.g., Hu et al., 2018; Juhaňák et al., 2018).

School's ICT Readiness, Background, and Students' Academic Performance

The effects of school's ICT readiness on student academic performance have been analyzed in previous literature. However, these studies have not reach agreement on whether school's ICT readiness affects student academic performance. School's ICT readiness is any investment in ICT, especially the availability of computers in schools. Some studies have claimed a positive correlation between the investment in ICT and student academic performance (e.g., Machin et al., 2010), while other studies have shown no significant effects observed in the student academic performance when increasing the investment in ICT (e.g., Cristia et al., 2014). School's internet access is part of school ICT infrastructures, which has been reported having the conflicted findings regarding its impact on student academic performance (Zhang & Liu, 2016).

Research concerning the relationship between school's background and students' academic performance focuses on school average socio-economic status and schools' educational resources. School average socio-economic status has been reported to have significantly impacts on student math and science achievement (Zhang & Liu, 2016), as well as students' ICT use (Aypay, 2010). The high socio-economic status groups performed better in learning than the low socio-economic status groups (Ahmar & Anwar, 2013). Educational resources measures school principals' perceptions of the potential factors regarding human resources or educational material as obstacles for school instructions. The findings have reported that the impacts of educational resources on students' academic performance can be mixed (Zhang & Liu, 2016).

Methods

Data

This study employed PISA 2018 dataset from Organization for Economic Co-operation and Development (OECD) including 79 participating countries and regions. PISA is a large-scale assessment globally delivered every three years to measure fifteen-year-old students' abilities in reading, mathematics, and science by the Organization for Cooperation and Development (OECD). The PISA 2018 study also collected participants' contextual data, such as their demographic information, students' learning attitude, and information related to their parents and schools. A questionnaire about student familiarity with the use of information and communications technologies (ICT) was administered as an additional survey in PISA 2018. In this study, we focused on students' learning performance in mathematics and science in the United States. After cleaning the missing values, the data had 2,592 observations.

Data Preprocessing

In this study, the data preprocessing was to select the most influential school-level variables. A Decision Tree (DT) approach was applied to investigate the relationships of school related variables and student learning performance in mathematics and science. DT is a multivariate and non-parametric supervised learning approach to examine the associations between attributes and response variables. This technique can handle continuous or categorical response variables and accordingly build regression or classification trees. It is a top-down recursive partition starting from a root node (also known as top decision node) that can be understood by following the IF-THEN rules (Romero et al., 2008). A root node is split to form internal nodes (also known as decision nodes) and further split the instance space into sub-spaces with leaf nodes (also known as terminal nodes). These nodes are the attributes selected using attribute selection measures (ASM) such as Information Gain (Quinlan, 1986) or Gini Index (Steinberg & Colla, 2009), which are the popular splitting criterion to partition data in the best way. The DT tree model generated process includes two stages: 1) tree building and 2) tree pruning. The R package rpart (Therneau et al., 2013) was applied to generate the DT models.

Although a single DT could yield an unstable model, there are multiple advantages of applying the DT approach. The main advantages include 1) its easy interpretability through its tree structure, 2) its comprehensibility to reveal data structure for both large and small data (Shahiri & Husain, 2015), and 3) its efficient computation (Singh & Gupta, 2014).

Variables

Considering the applied dataset having a clustered structure, variables include two levels, student level and school level, to examine the relationship between students' ICT use and their learning achievement in mathematics and science. The response variables are students' mathematics and science scores. Therefore, separate models and analyses were conducted for each response variable. The student-level and school-level variables are described in detail below.

Student-level Variables

Six ICT use related variables were included at the student level (see Table 12). On the PISA ICT questionnaire, students were asked their general use of ICT at school (USESCH), ICT usage for their daily social life (SOIAICT), their perceived autonomy related to ICT usage (AUTICT), their perceived competence in ICT usage (COMPICT), and the subject-related use of digital devices during their classroom lessons (ICTCLASS), and ICT resources regarding the household possessions (ICTRES).

Table 12

ICT Use Related Variable Information

Variable Name	Description
COMPICT	Students' Perceived ICT competence. This index includes five questions on a four-point Likert scale: strongly agree, agree, disagree, and strongly disagree. IC014: Thinking about your experience with digital media and digital devices: to what extent do you disagree or agree with the following statements? IC014Q03. I feel comfortable using digital devices that I am less familiar with. IC014Q04. If my friends and relatives want to buy new digital devices or applications, I can give them advice. IC014Q06. I feel comfortable using my digital devices at home. IC014Q08. When I come across problems with digital devices, I think I can solve them. IC014Q09. If my friends and relatives have a problem with digital devices, I can help them.
AUTICT	Students' perceived autonomy related to ICT use. This index is scaled based on the five questions on a four-point Likert scale: strongly agree, agree, disagree, and strongly disagree. IC015. Thinking about your experience with digital media and digital devices: to what extent do you disagree or agree with the following statements? IC015Q02. If I need new software, I install it by myself. IC015Q03. I read information about digital devices to be independent. IC015Q05. I use digital devices as I want to use them. IC015Q07. If I have a problem with digital devices I start to solve it on my own. IC015Q09. If I need a new application, I choose it by myself.
SOIAICT	Students' ICT use for social networking was measured specifically via five statements on a four-point Likert scale: strongly agree, agree, disagree, and strongly disagree. IC016. Thinking about your experience with digital media and digital devices: to what extent do you disagree or agree with the following statements? IC016Q01. To learn something new about digital devices, I like to talk about them with my friends. IC016Q02. I like to exchange solutions to problems with digital devices with others on the Internet. IC016Q04. I like to meet friends and play computer and video games with them. IC016Q05. I like to share information about digital devices with my friends.

	IC016Q07. I learn a lot about digital media by discussing with my friends and relatives.
ICTCLASS	Subject-related use of digital devices during classroom lessons. It included nine items. The response format is a five-point Likert scale: “No time”, “1-30 minutes a week”, “31-60 minutes a week”, “More than 60 minutes a week”, and “I do not study this subject”. IC150. In a typical school week, how much time do you spend using digital devices during classroom lessons? (Test language lessons/ Mathematics/ Science/ Foreign language/ Social sciences/ Music/ Sports/ Performing arts/ Visual arts)
USESCH	Use of ICT at school in general. This index is calculated based on 10 questions on a five-point Likert scale: “Never or hardly ever”, “Once or twice a month”, “Once or twice a week”, “Almost every day”, and “Every day”. IC011. How often do you use digital devices for the following activities at school? IC011Q01. Chatting online at school IC011Q02. Using email at school. IC011Q03. Browsing the Internet for schoolwork. IC011Q04. Downloading, uploading or browsing material from the school’s website IC011Q05. Posting my work on the school’s website. IC011Q06. Playing simulations at school. IC011Q07. Practicing and drilling, such as for foreign language learning or mathematics IC011Q08. Doing homework on a school computer. IC011Q09. Using school computers for group work and communication with other students. IC011Q010. Using learning apps or learning websites.
ICTRES	ICT resources. This index measured the availability of six household items at home. (Educational software/ A link to the Internet/ Cell phone with Internet access/ Computers/ Tablet computers/ ebook readers)

School-level Variables

We initially considered seven school-level variables to build the DT models. These variables contain five primary aspects, which are school-level socio-economic status, school size, ICT development, school resources, and teacher human resources. School-level socio-economic status was calculated by averaging student-level socio-economic status based on schools (AverageofESCS). School-level ICT development is shown via the availability of computers at school (AverageofRATCMP1) and the ratio of computers connected to the internet (AverageofRATCMP2). School resources is represented by the shortage of educational materials (AverageofEDUSHORT), which considers both the amount and the quality of educational resources and physical infrastructure at school. A positive value of this index means that the school's amount and/or quality of human or educational resource could hinder instruction at school. Teacher human resources refer to the total number of teachers at school (AverageofTOTAT) and the student-teacher ratio (AverageofSTRATIO).

When building the DT models, the dataset was randomly split into 70% train dataset and 30% test dataset. The models were built through the train dataset and evaluated using the test dataset. The results have shown that the DT model predicting students' mathematics performance reached the accuracy at 95.53% (RMSE = 21.1436), and the DT model predicting students' science performance reached 95.82% accuracy (RMSE = 20.9972). According to the model results displayed in Figure 9 and Figure 10, school-level socio-economic status is important when predict students' mathematics and science performance. The shortage of educational material can significantly influence students' science performance. Additionally, the DT results have also shown that the availability of computers at school, total number of teachers, and school size are influential variables regarding student learning performance in mathematics. Therefore, these important variables were included and centered at the school level in the subsequent data analysis.

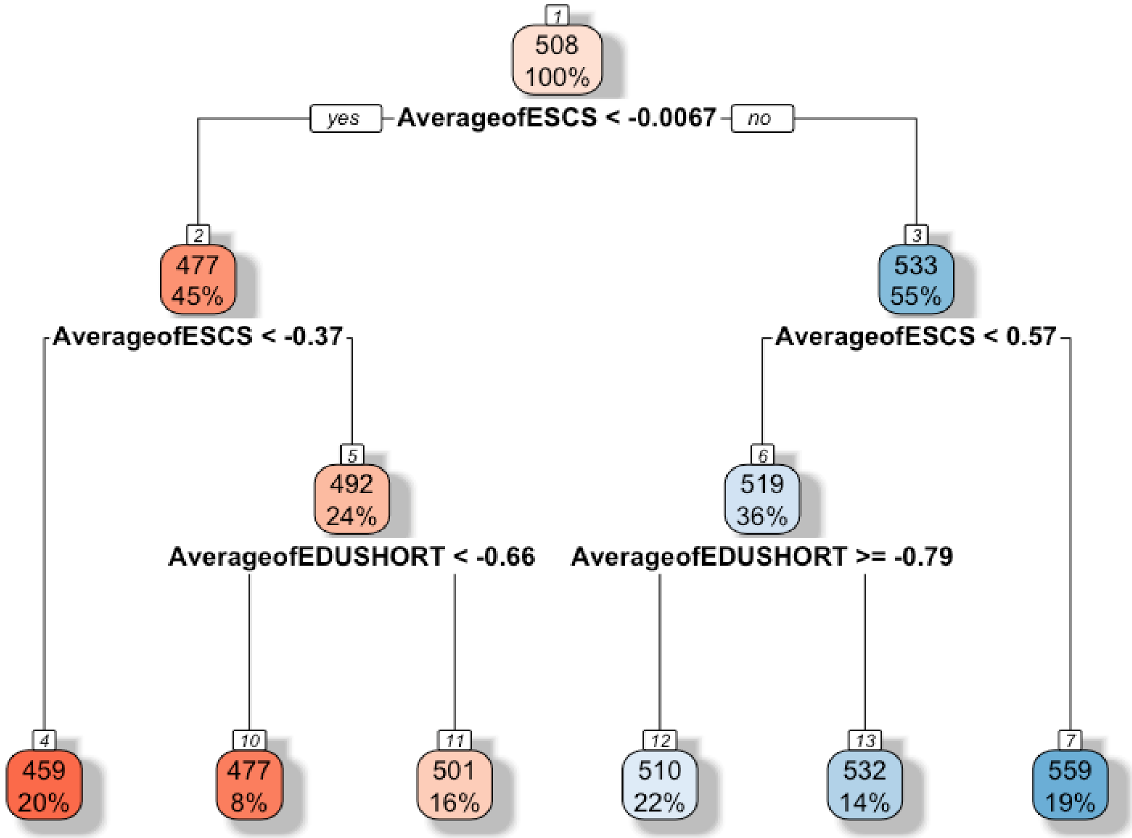

Figure 9. The DT Model to Predict Student Learning Performance in Science.

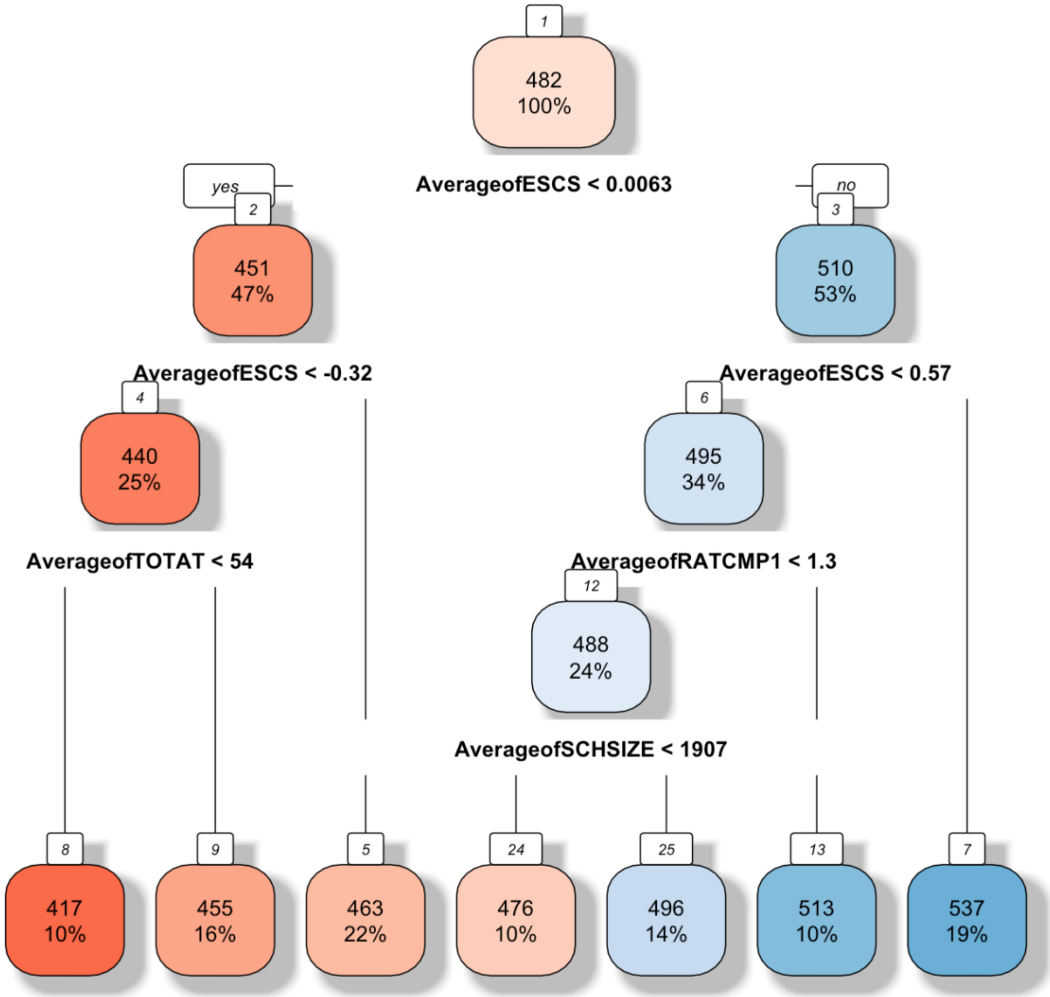

Figure 10. The DT Model to Predict Student Learning Performance in Mathematics.

Data Analysis

Hierarchical linear modeling (HLM) was used to build models with school- and student-level variables. Since students within the PISA 2018 were nested within schools, the intercepts- and slopes-as-outcomes models were conducted to analyze both all level predictors and interactions among two-level variables while considering variances of each level predictors. All student-level variables were group-centered, and the school-level variables were grand-centered for the analysis. The HLM analyses were carried out using the lme4 package (Bates et al., 2007) in R (version 3.6.1).

First, the null models with random effects were created to partition the variance of the outcome variable in within-and between-group components (Raudenbush & Bryk, 2002). Then, include all the predictors at the student-level in the models. Lastly, we added all school-level predictors in the models. According to the results of the DT models, we selected the influential school-level variables and applied them into the HLM. We ran two separate HLM focusing on students' mathematics and science accordingly. The first level model for both HLM can be written as:

$Y_{mn} = \beta_{0n} + \beta_{1n}*(level1\_USESCH_{mn}) + \beta_{2n}*(level1\_SOIAICT_{mn}) +$
$\beta_{3n}*(level1\_COMPICT_{mn}) + \beta_{4n}*(level1\_ICTCLASS_{mn}) + \beta_{5n}*(level1\_AUTICT_{mn}) +$
$\beta_{6n}*(Slevel1\_ICTRES_{mn}) + e_{mn}$

Where

Y_{mn} are the mathematics or science scores for the m^{th} unit of the n^{th} school,

β_{0n} is the constant term,

β_{1n}, β_{2n},…, β_{6n} represents the slope parameters for the first level and is estimated by the second level sub-models. $level1\_USESCH_{mn}$, $level1\_SOIAICT_{mn}$, $level1\_COMPICT_{mn}$, $level1\_ICTCLASS_{mn}$, $level1\_AUTICT_{mn}$, and $level1\_ICTRES_{mn}$ are the predictors for the m^{th} unit of the n^{th} group.

e_{mn} is the error term.

The second level sub-models which focus on student mathematics can be written as:

β_{0n} = γ_{00}+γ_{01}*c_ESCS_n+γ_{02}*$c\_RATCMP1_n$+$\gamma_{03} * c\_TOTAT_n$+$\gamma_{04} * c\_SCHSIZE_n$+$u_{0n}$

β_{1n} = γ_{10}+γ_{11}*c_ESCS_n +γ_{12}*$c\_RATCMP1_n$+$\gamma_{13} * c\_TOTAT_n$ +$\gamma_{14} * c\_SCHSIZE_n$

β_{2n} = γ_{20}+γ_{21}*c_ESCS_n+γ_{22}*$c\_RATCMP1_n$+$\gamma_{23} * c\_TOTAT_n$ +$\gamma_{24} * c\_SCHSIZE_n$

β_{3n} = γ_{30}+γ_{31}*c_ESCS_n +γ_{32}*$c\_RATCMP1_n$+$\gamma_{33} * c\_TOTAT_n$ +$\gamma_{34} * c\_SCHSIZE_n$

β_{4n} = γ_{40}+γ_{41}*c_ESCS_n +γ_{42}*$c\_RATCMP1_n$+$\gamma_{43} * c\_TOTAT_n$ +$\gamma_{44} * c\_SCHSIZE_n$

β_{5n} = γ_{50}+γ_{51}*c_ESCS_n+γ_{52}*$c\_RATCMP1_n$+$\gamma_{53} * c\_TOTAT_n$+$\gamma_{54} * c\_SCHSIZE_n$

β_{6n} = γ_{60}+γ_{61}*c_ESCS_n+γ_{62}*$c\_RATCMP1_n$ +$\gamma_{63} * c\_TOTAT_n$+$\gamma_{64} * c\_SCHSIZE_n$

Where

γ_{00}, γ_{10},..., $\gamma_{60}$are the constant terms,

γ_{01}, γ_{02},…, γ_{62} are the slope parameters of the sub-models.

c_ESCS_n, $c\_RATCMP1_n$, c_TOTAT_n, $c\_SCHSIZE_n$ represent the value of second-level predictors.

$u_{0n}$is the error term.

Overall, the combined model for student mathematics can be written as:

Y_{mn} =γ_{00} + γ_{01}*c_ESCS_n + $\gamma_{02} * c\_TOTAT_n$+ γ_{03}*$c\_RATCMP1_n$+$\gamma_{04} *$ $c\_SCHSIZE_n$+γ_{10}*$level1\_USESCH_{mn}$+γ_{11}*$level1\_USESCH_{mn}$*c_ESCS_n +γ_{12}* $level1\_USESCH_{mn}$*$c\_RATCMP1_n$+γ_{13}*$level1\_USESCH_{mn}$*c_TOTAT_n+$\gamma_{14} *$ $level1\_USESCH_{mn}$*$c\_SCHSIZE_n$+γ_{20}*$level1\_SOIAICT_{mn}$+γ_{21}*$level1\_SOIAICT_{mn}$* c_ESCS_n +γ_{22}*$level1\_SOIAICT_{mn}$*$c\_RATCMP1_n$+γ_{23}*$level1\_SOIAICT_{mn}$*c_TOTAT_n+γ_{24}* $level1\_SOIAICT_{mn}$*$c\_SCHSIZE_n$+γ_{30}*$level1\_COMPICT_{mn}$+γ_{31}*$level1\_COMPICT_{mn}$* c_ESCS_n +γ_{32}*$level1\_COMPICT_{mn}$*$c\_RATCMP1_n$+γ_{33}*$level1\_COMPICT_{mn}$*c_TOTAT_n+γ_{34}

$*level1\_COMPICT_{mn}*c\_SCHSIZE_n+\gamma_{40}*level1\_ICTCLASS_{mn}+\gamma_{41}*level1\_ICTCLASS_{mn}*$

$c\_ESCS_n +\gamma_{42}*level1\_ICTCLASS_{mn}*c\_RATCMP1_n+\gamma_{43}*level1\_ICTCLASS_{mn}*c\_TOTAT_n+\gamma_{44}$

$*level1\_ICTCLASS_{mn}*c\_SCHSIZE_n+\gamma_{50}*level1\_AUTICT_{mn}+\gamma_{51}*level1\_AUTICT_{mn}*$

$c\_ESCS_n +\gamma_{52}*level1\_AUTICT_{mn}*c\_RATCMP1_n+\gamma_{53}*level1\_AUTICT_{mn}*c\_TOTAT_n+\gamma_{54}*$

$level1\_AUTICT_{mn}*c\_SCHSIZE_n+\gamma_{60}*level1\_ICTRES_{mn}+\gamma_{61}*level1\_ICTRES_{mn}*c\_ESCS_n +$

$\gamma_{62}*level1\_ICTRES_{mn}*c\_RATCMP1_n+\gamma_{63}*level1\_ICTRES_{mn}$

$*c\_TOTAT_n+\gamma_{64}*level1\_ICTRES_{mn} *c\_SCHSIZE_n+u_{0n}+e_{mn}$

The second level sub-models which focus on student science can be written as:

$\beta_{0n} = \gamma_{00}+\gamma_{01}*c\_ESCS_n +\gamma_{02}*c\_EDUSHORT_n+u_{0n}$

$\beta_{1n} = \gamma_{10}+\gamma_{11}*c\_ESCS_n+\gamma_{12}*c\_EDUSHORT_n$

$\beta_{2n} = \gamma_{20}+\gamma_{21}*c\_ESCS_n+\gamma_{22}*c\_EDUSHORT_n$

$\beta_{3n} = \gamma_{30}+\gamma_{31}*c\_ESCS_n+\gamma_{32}*c\_EDUSHORT_n$

$\beta_{4n} = \gamma_{40}+\gamma_{41}*c\_ESCS_n+\gamma_{42}*c\_EDUSHORT_n$

$\beta_{5n} = \gamma_{50}+\gamma_{51}*c\_ESCS_n+\gamma_{52}*c\_EDUSHORT_n$

$\beta_{6n} = \gamma_{60}+\gamma_{61}*c\_ESCS_n +\gamma_{62}*c\_EDUSHORT_n$

Where

γ_{00}, γ_{10},..., $\gamma_{60}$are the constant terms,

γ_{01}, γ_{02},…, γ_{62} are the slope parameters of the sub-models.

c_ESCS_n, $c\_EDUSHORT_n$ represent the value of second-level predictors.

$u_{0n}$is the error term.

Overall, the combined model for student science can be written as:

$$Y_{mn} = \gamma_{00} + \gamma_{01} * c\_ESCS_n + \gamma_{02} * c\_EDUSHORT_n + \gamma_{10} * level1\_USESCH_{mn} + \gamma_{11} * level1\_USESCH_{mn} * c\_ESCS_n + \gamma_{12} * level1\_USESCH_{mn} * c\_EDUSHORT_n + \gamma_{20} * level1\_SOIAICT_{mn} + \gamma_{21} * level1\_SOIAICT_{mn} * c\_ESCS_n + \gamma_{22} * level1\_SOIAICT_{mn} * c\_EDUSHORT_n + \gamma_{30} * level1\_COMPICT_{mn} + \gamma_{31} * level1\_COMPICT_{mn} * c\_ESCS_n + \gamma_{32} * level1\_COMPICT_{mn} * c\_EDUSHORT_n + \gamma_{40} * level1\_ICTCLASS_{mn} + \gamma_{41} * level1\_ICTCLASS_{mn} * c\_ESCS_n + \gamma_{42} * level1\_ICTCLASS_{mn} * c\_EDUSHORT_n + \gamma_{50} * level1\_AUTICT_{mn} + \gamma_{51} * level1\_AUTICT_{mn} * c\_ESCS_n + \gamma_{52} * level1\_AUTICT_{mn} * c\_EDUSHORT_n + \gamma_{60} * level1\_ICTRES_{mn} + \gamma_{61} * level1\_ICTRES_{mn} * c\_ESCS_n + \gamma_{62} * level1\_ICTRES_{mn} * c\_EDUSHORT_n + u_{0n} + e_{mn}$$

Results

Student-Level ICT Variables

Table 13 and Table 14 report the estimation results of the HLM regarding students' science and mathematics achievements. At the within-school level, holding school-level variables (i.e. ESCS, EDUSHORT) constant, all ICT-related variables showed significant effects on students' science achievement. Both USESCH (-12.3401) and SOIAICT (-14.9715) negatively associated with students' science achievement at the within-school level. On the other hand, students' ICTCLASS (9.268), COMPICT (6.4155), AUTICT (13.0166), and ICTRES (4.692) showed significantly positive relationships with students' science achievement. Since there were significant interaction effects of ESCS and EDUSHORT, the effects of ICTCLASS and USESCH on science achievement were conditional. Regarding students' mathematics achievement, students' ICT competence showed an insignificant effect, but all other ICT-related variables had significant effects. Among these significant ICT-related variables, USESCH and SOIAICT negatively influenced students' mathematics achievement, while others had positive associations with mathematics achievement. The coefficients for these ICT-related variables were -10.04 (USESCH), -9.461 (SOIAICT), 8.286 (ICTCLASS), 14.22 (AUTICT), and 8.601 (ICTRES). The impact of ICTCLASS on mathematics achievement was conditional because there was an interaction effect with ESCS.

School-Level Factors

Socio-economic status showed significantly positive impacts on students' science and mathematics achievements with the coefficients 58.9521 and 57.25 accordingly. School size also had a significant relationship with students' mathematics achievement. Its coefficient was 0.012. However, other school-level variables showed insignificant impacts on students' science or mathematics achievements. For example, the shortage of school materials had no significant impact on students' science achievement. The total number of teachers at school and the availability of computers at school showed no significant impacts on students' mathematics achievement.

In terms of cross-level interactions, socio-economic status exhibited significantly positive interaction effects on the relationships between ICT use during the class (ICTCLASS) and students' mathematics (10.39) and science achievements (11.0974). School resource shortage had a significantly negative interaction effect on the relationship between ICT use at school and students' science achievements (-6.0222). However, other interaction effects between the school-level variables and ICT-related variables on students' mathematics and science achievements were insignificant.

Table 13

The HLM Results Regarding Student Science Achievement

	Within-School Model		
	Estimate	SD Error	p value
USESCH	-12.3401	2.6611	**<.0001**
SOIAICT	-14.9715	2.3316	**<.0001**
COMPICT	6.4155	2.8814	**.0277**
ICTCLASS	9.268	2.4062	**.0001**
AUTICT	13.0166	2.5711	**<.0001**
ICTRES	4.692	1.886	**.0138**
	Between-School Model		
	Estimate	SD Error	p value
ESCS	58.9521	5.0328	**<.0001**
EDUSHORT	-0.3014	2.8468	0.9159
ESCS*USESCH	-0.511	5.0746	0.9199
EDUSHORT *USESCH	-6.0222	2.8993	**.0397**
ESCS*SOIAICT	-0.5202	4.4629	0.9073
EDUSHORT*SOIAICT	1.1691	2.3274	0.6162
ESCS*COMPICT	-4.1173	5.4885	0.4544
EDUSHORT*COMPICT	-3.068	2.8904	0.291
ESCS*ICTCLASS	11.0974	4.7962	**.0213**
EDUSHORT*ICTCLASS	-0.8834	2.5858	0.7329
ESCS*AUTICT	2.5036	4.9813	0.6154

EDUSHORT*AUTICT	-0.5015	2.5129	0.8419
ESCS*ICTRES	-2.345	3.5768	0.5128
EDUSHORT*ICTRES	0.025	1.9539	0.9898

Table 14

The HLM Results Regarding Student Mathematics Achievement

	Within-School Model		
	Estimate	SD Error	p value
USESCH	-10.04	2.65	**.0003**
SOIAICT	-9.461	2.149	**<.0001**
COMPICT	2.753	2.653	0.3017
ICTCLASS	8.286	2.183	**.0002**
AUTICT	14.22	2.287	**<.0001**
ICTRES	8.061	1.644	**<.0001**
	Between-School Model		
	Estimate	SD Error	p value
ESCS	57.25	4.709	**<.0001**
RATCMP1	-1.016	1.841	0.5822
SCHSIZE	0.012	0.0042	**.0052**
TOTAT	0.1377	0.0806	0.0906
ESCS*USESCH	-1.468	5.104	0.7743
RATCMP1 *USESCH	-0.7482	1.785	0.6766
SCHSIZE*USESCH	-0.004	0.0048	0.3982

TOTAT*USESCH	0.0701	0.0959	0.4658
ESCS*SOIAICT	-0.5955	4.214	0.8878
RATCMP1*SOIAICT	-0.8421	1.591	0.5972
SCHSIZE*SOIAICT	-0.0023	0.0034	0.5009
TOTAT*SOIAICT	0.0062	0.0653	0.9241
ESCS*COMPICT	-3.627	5.181	0.4851
RATCMP1*COMPICT	-1.481	1.884	0.4336
SCHSIZE*COMPICT	0.0034	0.0043	0.4336
TOTAT*COMPICT	-0.0557	-0.084	0.5009
ESCS*ICTCLASS	10.39	4.485	**.021**
RATCMP1*ICTCLASS	1.324	1.533	0.3883
SCHSIZE*ICTCLASS	0.0037	0.0038	0.3302
TOTAT*ICTCLASS	-0.0947	0.0762	0.2146
ESCS*AUTICT	5.92	4.575	0.1959
RATCMP1*AUTICT	1.145	1.709	0.5031
SCHSIZE*AUTICT	-0.0026	0.0036	0.4726
TOTAT*AUTICT	0.0858	0.0698	0.2193
ESCS*ICTRES	-4.793	3.202	0.1352
RATCMP1*ICTRES	0.2636	1.154	0.8195
SCHSIZE*ICTRES	-0.0045	-0.0026	0.0859
TOTAT*ICTRES	0.1045	0.0541	0.0543

Discussion

Student-Level ICT-Related Variables and Learning Performance

At the within-school level, students' use of ICT at school negatively influenced their mathematics and science achievements. This finding supports previous studies that reported negative impacts on students' academic performance due to using ICT at school (Chiao & Chiu, 2018; Park & Weng, 2020). Although some studies reported positive relationships between students' ICT use at school and their academic performances (e.g., Luu & Freeman, 2011), these studies analyzed other countries' data of which could have different situations from the United States. Therefore, it is highly likely that students' use of ICT at school can negatively influence their academic achievement across different schools at the United States. The possible reason of the negative impact could be the higher frequency of using ICT for the non-learning activities at school rather than using ICT for the learning purposes. Similarly, students' use of ICT for social interactions had negative impacts on their mathematics and science achievements. This finding aligns with previous studies (Hu et al., 2018; Meng et al. 2018), whilst Martínez-Abad et al. (2018) claimed a positive relation between students' ICT use for social interactions and their mathematics performance. The possible reason of the negative impacts can be that students perhaps used digital devices for entertainment or leisure to support their social interactions instead of conducting learning related activities, such as learning discussion with their peers.

The results showed that students' perceived ICT autonomy had significantly positive effects on students' mathematics and science achievements. This finding agrees with previous literature (e.g., Meng et al., 2019; Xiao et al., 2019). Students' perceived ICT autonomy indicates their self-regulation in learning and usage of ICT in their learning process (Fu, 2013). ICT autonomy can have more powerful impacts on student learning performance than other ICT-related factors (Park & Weng, 2020). Therefore, it is important for students to increase their autonomy in ICT use in order to improve their learning performance by using technologies. Meanwhile, students' use of ICT in classes also showed significantly positive impacts on their mathematics and science performance. This finding agrees with previous literature, which indicates that the more use of ICT in classes can yield better learning results (KAYA & Sibel, 2021). In contrast with this finding, Erdoğdu and Erdoğdu (2015) reported negative relationships between this ICT factor and students' performance and claimed students' learning distraction probably due to using ICT at school. Overall using ICT in classes can be beneficial if ICT is used effectively in learning.

The availability of students' ICT resources at home can significantly improve their mathematics and science achievements. However, previous literature findings have shown a negative relationship (Hu et al., 2018) or mixed impacts (Gubbels et al., 2020; Lee & Wu, 2012). The possible reason is that the availability of students' ICT resources does not equal to how students utilize these resources for their learning. The students in previous literature may utilize these ICT home resources less than the students in this study. More specific data will be needed to investigate the details and other possible reasons behind.

Students' ICT competence showed a significantly positive impact on their science achievement but had no relationship with their mathematics achievement. Previous literature also reported mixed findings (Hu et al., 2018; Xiao et al., 2019). The possible reasons causing the mixed findings depend on students' other covariates, such as demographics or prior academic performance. As students enter to higher grade levels, they tend to be more competent overall and the impact of their ICT competence on their learning performance may reduce (Selwyn & Husen, 2010).

Main and Moderating Effects of School-Level Factors

According to the results above, school size has a significantly positive main effect on students' mathematics achievement. School size has been an important factor to be considered in educational research when investigating its relationship with student learning performance. This study finding is in agreement with some existing studies (e.g., Luyten,m 2014; Scheerens et al., 2014). However, the bulk of studies have noted that school size could yield negative impacts on student outcomes in mathematics and reading (e.g., Egalite & Kisida, 2016). The reasons causing the inconsistency of the findings regarding the impact of school sizes can be complicated. One possible reason could be due to different samplings. For example, we included 118 schools attending PISA 2018 across the United States in our study, while Egalite and Kisida (2016) covered 2,679 unique schools from 2007 through 2011. Apparently, more data samples can yield more robust findings, but the differences among school districts also remained and might influence the results. Other reasons may include school types (i.e. private schools versus public schools), different grade levels, impacts on certain subjects, and schools' resources.

The shortage of school resources had only the moderating effect on student science achievement. Considering the interaction effect of students' ICT use at school with the shortage of school resources, lacking school resources could worsen students' ICT use at schools. Hence, increasing school resources including ICT resources can provide students more support of using ICT at schools. Although previous literature has claimed the benefits of adding school resources (e.g., Smanova, 2021), how students utilize the resources, especially the ICT resources, is the key for their learning improvement. Setting and implementing clear policies and strategies of using school resources can be helpful in terms of regulating students' learning behavior at school.

Socio-economic status (ESCS) had significant both main effects and moderating effects on student mathematics and science achievements. PISA measures ESCS through several aspects including parental educational attainment (in years), parental highest occupational status on the "International Socio-Economic Index" (ISEI) scale (OECD, 2019), and household possessions. The findings of ESCS at the school level indicates that schools having students from families with higher economic, social, and cultural backgrounds in average performed better in mathematics and science than schools having students from families with lower economic, social, and cultural backgrounds in average. Previous studies also have claimed that students' ESCS can positively impact their learning outcomes (e.g., Luu & Freeman, 2011). First, students from high-ESCS families tend to attend better schools which have more advantages in different aspects (Schulz, 2005). Second, schools with higher ESCS are highly likely to have more school-level ICT resources such as ICT infrastructure, ICT support, and ICT use with subscriptions of educational technologies. This impact is also noted from the interaction effect between students' ESCS and ICT class use in our study. In addition, students with higher levels of ESCS may start using technologies in their earlier ages and be more competent in using ICT than students with lower levels of ESCS (Luu & Freeman, 2011). To better support students' ICT use in classrooms, improving teachers' ICT skills can be helpful. Teachers' professional development from district and school support can provide useful resources to build their ICT skills (Wayne, 2002).

CHAPTER V

CONCLUSIONS

This dissertation provided a comprehensive insight of LA/EDM and statistical methods in clustered educational data. Previous LA/EDM studies were analyzed from several aspects from impacts to methodological characteristics. The information from this dissertation can provide educators a better understanding of potential topics that can be supported from LA/EDM, the benefits of implementing LA/EDM into practices, and a statistical lens of method selection.

The dissertation noted the potentials of applying different statistical methods, including RF, RE-EM Tree, MERF, and HLM in clustered educational data. The comparison results of these four different statistical methods indicated the advantages and disadvantages of each statistical method and the optimal selection of these methods in certain circumstances. Particularly, RE-EM Tree and MERF are rarely used in clustered educational data. The dissertation showed the possibility of applying these mixed effects methods in educational research. Among all four methods, MERF yielded the most accurate result. RE-EM Tree and HLM reached similar accuracy. Although RF yielded more accurate results than RE-EM Tree and HLM, it is recommended to adopt MERF instead of RF because RF only considers single data level.

Although mixed effects tree methods can perform well in clustered educational data, HLM is still a very useful statistical method in a non-high dimensional data. Therefore, the dissertation adopted HLM to examine the relationship between students' ICT and their learning achievements in mathematics and science moderated by school-level factors. The study selected school-level factors using DT to ensure an unbiased and data-driven process. The findings indicated that students' learning performance in mathematics and science were influenced by certain ICT related factors. ESCS was found to be a significant moderator that influenced some of ICT related factors and students' learning achievements in mathematics and science. The interaction effect between the shortage of school resources and ICT use at school also showed a significantly negative impact on student learning performance in science. Schools' administration may consider formulating policies and processes of how students use ICT tools to best support their learning at school. School size was found to be important regarding students' mathematics achievement. This school-level factor has been greatly noted due to its significance in previous literature. Those studies reported both positive and negative impacts of school size on student learning achievement. The mixed findings indicated the importance of increasing data samples or using school size with other school related covariates (e.g., school types) when examining the effects of school factors.

REFERENCES

Abdous, M. H., & He, W. (2011). Using text mining to uncover students' technology-related problems in live video streaming. *British Journal of Educational Technology*, *42*(1), 40-49.

Abdous, M. H., He, W., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Educational Technology & Society*, *15*(3), 77-88.

Abell, M. (2006). Individualizing learning using intelligent technology and universally designed curriculum. The journal of Technology, Learning and Assessment, 5(3):5- 10.

Act, R. (2009). The American Recovery and Reinvestment Act of 2009. *Public Law*, *111*(5), 5-30.

Adzharuddin, N. A., & Ling, L. H. (2013). Learning management system (LMS) among university students: Does it work. *International Journal of e-Education, e-Business, e-Management and e-Learning*, *3*(3), 248-252.

Aguilar, J., Buendia, O., Pinto, A., & Gutiérrez, J. (2019). Social learning analytics for determining learning styles in a smart classroom. *Interactive Learning Environments*, 1-17.

Ahmar, F., & Anwar, E. (2013). Socio economic status and its relation to academic achievement of higher secondary school students. *IOSR Journal of Humanities and Social Science*, *13*(6), 13-20.

Ahmad Uzir, N. A., Gašević, D., Matcha, W., Jovanović, J., & Pardo, A. (2020). Analytics of time management strategies in a flipped classroom. *Journal of Computer Assisted Learning*, *36*(1), 70-88.

Alario-Hoyos, C., Muñoz-Merino, P. J., Pérez-Sanagustín, M., Delgado Kloos, C., & Parada G, H. A. (2016). Who are the top contributors in a MOOC? Relating participants' performance and contributions. *Journal of Computer Assisted Learning*, *32*(3), 232-243.

Albatayneh, N. A., Ghauth, K. I., & Chua, F. F. (2018). Utilizing learners’ negative ratings in semantic content-based recommender system for e-learning forum. *Journal of Educational Technology & Society*, *21*(1), 112-125.

Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., & Fernández-Manjón, B. (2020). Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *Journal of Computer Assisted Learning*, *36*(3), 350-358.

Angeli, C., & Valanides, N. (2013). Using educational data mining methods to assess field-dependent and field-independent learners’ complex problem solving. *Educational Technology Research and Development*, *61*(3), 521-548.

Araya, R., Plana, F., Dartnell, P., Soto-Andrade, J., Luci, G., Salinas, E., & Araya, M. (2012). Estimation of teacher practices based on text transcripts of teacher speech using a support vector machine algorithm. *British Journal of Educational Technology*, *43*(6), 837-846.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177-194.

Aypay, A. (2010). Information and communication technology (ICT) usage and achievement of Turkish students in PISA 2006. *The Turkish Online Journal of Educational Technology*, *9*, 116-124.

Baker, R. S., Clarke-Midura, J., & Ocumpaugh, J. (2016). Towards general models of effective science inquiry in virtual performance assessments. *Journal of Computer Assisted Learning*, *32*(3), 267-280.

Baker, R. S., Lindrum, D., Lindrum, M. J., & Perkowski, D. (2015). Analyzing Early At-Risk Factors in Higher Education E-Learning Courses. *International Educational Data Mining Society*.

Baker, R. S., Pardos, Z. A., Gowda, S. M., Nooraei, B. B., & Heffernan, N. T. (2011, July). Ensembling predictions of student knowledge within intelligent tutoring systems. In *International conference on user modeling, adaptation, and personalization* (pp. 13-24). Springer, Berlin, Heidelberg.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, *1*(1), 3-17.

Bates, D., Maechler, M., & Bolker, B. (2015). Walker., S. Fitting linear mixed-effects models using lme4. *J Stat Softw*, *67*(1), 1-48.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R package version*, *2*(1), 74.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American sociological review*, 386-398.

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education*, *158*, 103999.

Botelho, A. F., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S. A., & Beck, J. E. (2019). Developing early detectors of student attrition and wheel spinning using deep learning. *IEEE Transactions on Learning Technologies*, *12*(2), 158-170.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Cart. *Classification and Regression Trees; Wadsworth and Brooks/Cole: Monterey, CA, USA*.

Brooks, C., Erickson, G., Greer, J., & Gutwin, C. (2014). Modelling and quantifying the behaviours of students in lecture capture environments. *Computers & Education*, *75*, 282-292.

Cano, A., & Leonard, J. D. (2019). Interpretable multiview early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies*, *12*(2), 198-211.

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In *Evolution of teaching and learning paradigms in intelligent environment* (pp. 183-221). Springer, Berlin, Heidelberg.

Cela, K., Sicilia, M. Á., & Sánchez-Alonso, S. (2016). Influence of learning styles on social structures in online learning environments. *British Journal of Educational Technology*, *47*(6), 1065-1082.

Cerezo, R., Bogarín, A., Esteban, M., & Romero, C. (2020). Process mining for self-regulated learning assessment in e-learning. *Journal of Computing in Higher Education*, *32*(1), 74-88.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, *96*, 42-54.

Chandrasekar, P., Qian, K., Shahriar, H., & Bhattacharya, P. (2017, July). Improving the prediction accuracy of decision tree mining with data preprocessing. In *2017 IEEE 41st*

Annual Computer Software and Applications Conference (COMPSAC) (Vol. 2, pp. 481-484). IEEE.

Chang, R. I., Hung, Y. H., & Lin, C. F. (2015). Survey of learning experiences and influence of learning style preferences on user intentions regarding MOOC s. *British Journal of Educational Technology*, *46*(3), 528-541.

Chen, B., Chang, Y. H., Ouyang, F., & Zhou, W. (2018). Fostering student engagement in online discussion through social learning analytics. *The Internet and Higher Education*, *37*, 21-30.

Chen, W., Brinton, C. G., Cao, D., Mason-Singh, A., Lu, C., & Chiang, M. (2018). Early detection prediction of learning outcomes in online short-courses via learning behaviors. *IEEE Transactions on Learning Technologies*, *12*(1), 44-58.

Cheng, M. T., Rosenheck, L., Lin, C. Y., & Klopfer, E. (2017). Analyzing gameplay data to inform feedback loops in The Radix Endeavor. *Computers & Education*, *111*, 60-73.

Chien, T. C., Chen, Z. H., & Chan, T. W. (2017). Exploring long-term behavior patterns in a book recommendation system for reading. *Journal of Educational Technology & Society*, *20*(2), 27-36.

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99*(6), 323-329.

Chiao, C., & Chiu, C. H. (2018). The mediating effect of ICT usage on the relationship between students' socioeconomic status and achievement. *The Asia-Pacific Education Researcher*, *27*(2), 109-121.

Chih-Ming, C., & Ying-You, L. (2020). Developing a computer-mediated communication competence forecasting model based on learning behavior features. *Computers and Education: Artificial Intelligence*, *1*, 100004.

Cho, M. H., & Yoo, J. S. (2017). Exploring online students' self-regulated learning with self-reported surveys and log files: a data mining approach. *Interactive Learning Environments*, *25*(8), 970-982.

Choi, S. P., Lam, S. S., Li, K. C., & Wong, B. T. (2018). Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. *Journal of Educational Technology & Society*, *21*(2), 273-290.

Chung, K. S. K., & Paredes, W. C. (2015). Towards a social networks model for online learning & performance. *Journal of Educational Technology & Society*, *18*(3), 240-253.

Clewley, N., Chen, S. Y., & Liu, X. (2011). Mining learning preferences in web-based instruction: Holists vs. serialists. *Journal of Educational Technology & Society*, *14*(4), 266-277.

Cocea, M., & Weibelzahl, S. (2010). Disengagement detection in online learning: Validation studies and perspectives. *IEEE transactions on learning technologies*, *4*(2), 114-124.

Codish, D., Rabin, E., & Ravid, G. (2019). User behavior pattern detection in unstructured processes–a learning management system case study. *Interactive Learning Environments*, *27*(5-6), 699-725.

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, *10*(1), 17-29.

Conijn, R., Van den Beemt, A., & Cuijpers, P. (2018). Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*, *34*(5), 615-628.

Cristia, J., Czerwonko, A. & Garofalo, P. (2014). Does technology in schools affect repetition, dropout and enrollment? Evidence from Peru. *Journal of Applied Economics*, *17*(1), 89-111.

Dawson, S. (2010). 'Seeing'the learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology*, *41*(5), 736-752.

de Barba, P. G., Malekian, D., Oliveira, E. A., Bailey, J., Ryan, T., & Kennedy, G. (2020). The importance and meaning of session behaviour in a MOOC. *Computers & Education*, *146*, 103772.

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, *192*, 38-48.

Dutta, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access*, *5*, 15991-16005.

Egalite, A. J., & Kisida, B. (2016). School size and student achievement: A longitudinal analysis. *School Effectiveness and School Improvement*, *27*(3), 406-417.

Elias, T. (2011). Learning analytics. *Learning*, 1-22.

Ellis, R. A., Han, F., & Pardo, A. (2017). Improving learning analytics–combining observational and self-report data on student learning. *Journal of Educational Technology & Society*, *20*(3), 158-169.

Er, E., Gómez-Sánchez, E., Dimitriadis, Y., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., & Álvarez-Álvarez, S. (2019). Aligning learning design and learning analytics through instructor involvement: A MOOC case study. *Interactive Learning Environments*, *27*(5-6), 685-698.

Erdogdu, F., & Erdogdu, E. (2015). The impact of access to ICT, student background and school/home environment on academic success of students in Turkey: An international comparative analysis. *Computers & Education*, *82*, 26-49.

Ergün, E., & Usluel, Y. K. (2016). An analysis of density and degree-centrality according to the social networking structure formed in an online learning environment. *Journal of Educational Technology & Society*, *19*(4), 34-46.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The journal of machine learning research*, *15*(1), 3133-3181.

Fincham, E., Gašević, D., Jovanović, J., & Pardo, A. (2018). From study tactics to learning strategies: An analytical method for extracting interpretable representations. *IEEE Transactions on Learning Technologies*, *12*(1), 59-72.

Fu, J. (2013). Complexity of ICT in education: A critical literature review and its implications. *International Journal of education and Development using ICT*, *9*(1), 112-125.

Garrison, D. R., Anderson, T., & Archer, W. (2010). The first decade of the community of inquiry framework: A retrospective. *The internet and higher education*, *13*(1-2), 5-9.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68-84.

Geng, S., Niu, B., Feng, Y., & Huang, M. (2020). Understanding the focal points and sentiment of learners in MOOC reviews: A machine learning and SC-LIWC-based approach. *British Journal of Educational Technology*, *51*(5), 1785-1803.

Ghavifekr, S., Razak, A. Z. A., Ghani, M. F. A., Ran, N. Y., Meixi, Y., & Tengyue, Z. (2014). ICT integration in education: Incorporation for teaching & learning improvement. *Malaysian Online Journal of Educational Technology*, *2*(2), 24-45.

Gkontzis, A. F., Kotsiantis, S., Panagiotakopoulos, C. T., & Verykios, V. S. (2019). A predictive analytics framework as a countermeasure for attrition of students. *Interactive Learning Environments*, 1-16.

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, *131*, 22-32.

Gubbels, J., Swart, N. M., & Groen, M. A. (2020). Everything in moderation: ICT and reading performance of Dutch 15-year-olds. *Large-scale Assessments in Education*, *8*(1), 1-17.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning* (pp. 9-41). Springer, New York, NY.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & probability letters*, *81*(4), 451-459.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313-1328.

Hajjem, A., Larocque, D., & Bellavance, F. (2017). Generalized mixed effects regression trees. *Statistics & Probability Letters*, *126*, 114-118.

Hershkovitz, A., & Nachmias, R. (2011). Online persistence in higher education web-supported courses. *The Internet and Higher Education*, *14*(2), 98-106.

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, *145*, 103724.

Holmes, M., Latham, A., Crockett, K., & O'Shea, J. D. (2017). Near real-time comprehension classification with artificial neural networks: Decoding e-learner non-verbal behavior. *IEEE Transactions on Learning Technologies*, *11*(1), 5-12.

Hooshyar, D., Yousefi, M., Wang, M., & Lim, H. (2018). A data-driven procedural-content-generation approach for educational games. *Journal of Computer Assisted Learning*, *34*(6), 731-739.

Horning, N. (2013). Introduction to decision trees and random forests. *Am. Mus. Nat. Hist*, *2*, 1-27.

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education*, *37*, 66-75.

Howard, S. K., Yang, J., Ma, J., Maton, K., & Rennie, E. (2018). App clusters: Exploring patterns of multiple app use in primary learning contexts. *Computers & Education*, *127*, 154-164.

Hu, X., Cheong, C. W. L., & Chu, S. K. W. (2018). Developing a multidimensional framework for analyzing student comments in wikis. *Journal of Educational Technology & Society*, *21*(4), 26-38.

Hu, X., Gong, Y., Lai, C., & Leung, F. K. (2018). The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis. *Computers & Education*, *125*, 1-13.

Huang, A. Y., Lu, O. H., Huang, J. C., Yin, C. J., & Yang, S. J. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interactive Learning Environments*, *28*(2), 206-230.

Hung, J. L., Hsu, Y. C., & Rice, K. (2012). Integrating data mining in program evaluation of K-12 online education. *Journal of Educational Technology & Society*, *15*(3), 27-41.

Hung, J. L., Shelton, B. E., Yang, J., & Du, X. (2019). Improving predictive modeling for at-risk student identification: A multistage approach. *IEEE Transactions on Learning Technologies*, *12*(2), 148-157.

International Educational Data Mining Society. (2011). [Online]. Available: http://www.educationaldatamining.org/

Jiang, Y., Bosch, N., Baker, R. S., Paquette, L., Ocumpaugh, J., Andres, J. M. A. L., ... & Biswas, G. (2018, June). Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection? In *International conference on artificial intelligence in education* (pp. 198-211). Springer, Cham.

Jo, I., Park, Y., & Lee, H. (2017). Three interaction patterns on asynchronous online discussion behaviours: A methodological comparison. *Journal of Computer Assisted Learning*, *33*(2), 106-122.

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, *33*(4), 74-85.

Juhaňák, L., Zounek, J., Záleská, K., Bárta, O., & Vlčková, K. (2018). The Relationship between Students' ICT Use and Their School Performance: Evidence from PISA 2015 in the Czech Republic. *Orbis scholae*, *12*(2).

Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education*, *27*, 54-63.

KAYA, V. H., & Sibel, İ. N. C. İ. (2021). How does Information and Communications Technology Influence Turkish Students' Science Achievement?. *Journal of Computer and Education Research*, *9*(18), 754-770.

Khalil, M., & Ebner, M. (2017). Clustering patterns of engagement in Massive Open Online Courses, MOOCs: the use of learning analytics to reveal student categories. *Journal of computing in higher education*, *29*(1), 114-132.

Khribi, M. K., Jemni, M., & Nasraoui, O. (2009). Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. *Educational Technology & Society*, *12* (4), 30–42.

Kim, D., Park, Y., Yoon, M., & Jo, I. H. (2016). Toward evidence-based learning analytics: Using proxy variables to improve asynchronous online discussion environments. *The Internet and Higher Education*, *30*, 30-43.

Kim, D., Yoon, M., Jo, I. H., & Branch, R. M. (2018). Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea. *Computers & Education*, *127*, 233-251.

Kostopoulos, G., Karlos, S., & Kotsiantis, S. (2019). Multiview learning for early prognosis of academic performance: a case study. *IEEE Transactions on Learning Technologies*, *12*(2), 212-224.

Kunina-Habenicht, O., & Goldhammer, F. (2020). ICT Engagement: a new construct and its assessment in PISA 2015. *Large-scale Assessments in Education*, *8*(1), 1-21.

Lee, Y.-J. (2015). Analyzing log files to predict students' problem solving performance in a computer-based physics tutor. *Educational Technology & Society, 18* (2), 225–236.

Lee, Y. H., & Wu, J. Y. (2012). The effect of individual differences in the inner and outer states of ICT on engagement in online reading activities and PISA 2009 reading literacy: Exploring the relationship between the old and new reading literacy. *Learning and Individual Differences*, *22*(3), 336-342.

Lemay, D. J., & Doleck, T. (2020). Predicting completion of massive open online course (MOOC) assignments from video viewing behavior. *Interactive Learning Environments*, 1-12.

Li, L. Y., & Chen, G. D. (2009). A coursework support system for offering challenges and assistance by analyzing students' web portfolios. *Journal of Educational Technology & Society*, *12*(2), 205-221.

Li, S., Chen, G., Xing, W., Zheng, J., & Xie, C. (2020). Longitudinal clustering of students' self-regulated learning behaviors in engineering design. *Computers & Education*, *153*, 103899.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Lin, C. F., Yeh, Y. C., Hung, Y. H., & Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: An application of decision trees. *Computers & Education*, *68*, 199-210.

Liu, C. C., Chang, C. J., & Tseng, J. M. (2013). The effect of recommendation systems on I nternet-based learning for different learners: A data mining analysis. *British Journal of Educational Technology*, *44*(5), 758-773.

Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying learning analytics for the early prediction of Students' academic performance in blended learning. *Journal of Educational Technology & Society*, *21*(2), 220-232.

Luu, K., & Freeman, J. G. (2011). An analysis of the relationship between information and communication technology (ICT) and scientific literacy in Canada and Australia. *Computers & Education*, *56*(4), 1072-1082.

Luyten, H. (2014). Quantitative summary of research findings. In H. Luyten, M. Hendriks, & J. Scheerens (Eds.), *School size effects revisited: A qualitative and quantitative review of the research evidence in primary and secondary education* (pp. 177–218). Dordrecht, The Netherlands: Springer.

Macarini, L. A., Lemos dos Santos, H., Cechinel, C., Ochoa, X., Rodés, V., Pérez Casas, A., ... & Díaz, P. (2020). Towards the implementation of a countrywide K-12 learning analytics initiative in Uruguay. *Interactive Learning Environments*, *28*(2), 166-190.

Machin, S., McNally, S. & Meghir, C. (2010). Resources and standards in urban schools. *Journal of Human Capital*, *4*(4), 365-393.

Mahnane, L. (2017). Recommending learning activities in social network using data mining algorithms. *Journal of Educational Technology & Society*, *20*(4), 11-23.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1-15.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). Making Sense of Data-Driven Decision Making in Education: Evidence from Recent RAND Research. Occasional Paper. *Rand Corporation*.

Martínez-Abad, F., Gamazo, A., & Rodríguez-Conde, M. J. (2018, October). Big data in education: detection of ICT factors associated with school effectiveness with data mining techniques. In *Proceedings of the sixth international conference on technological ecosystems for enhancing multiculturality* (pp. 145-150).

Martín-García, A. V., Martínez-Abad, F., & Reyes-González, D. (2019). TAM and stages of adoption of blended learning in higher education by application of data mining techniques. *British Journal of Educational Technology*, *50*(5), 2484-2500.

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice.

Meng, L., Qiu, C., & Boyd-Wilson, B. (2019). Measurement invariance of the ICT engagement construct and its association with students' performance in China and Germany: Evidence from PISA 2015 data. *British Journal of Educational Technology*, *50*(6), 3233-3251.

Mensink, P. J., & King, K., 2020. Student access of online feedback is modified by the availability of assessment marks, gender and academic performance. *British Journal of Educational Technology*, *51*(1), 10-22.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006, August). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 935-940.

Mirriahi, N., Liaqat, D., Dawson, S., & Gašević, D. (2016). Uncovering student learning profiles with a video annotation tool: reflective learning with and without instructional norms. *Educational Technology Research and Development*, *64*(6), 1083-1106.

Moon, J., Ke, F., & Sokolikj, Z. (2020). Automatic assessment of cognitive and emotional states in virtual reality-based flexibility training for four adolescents with autism. *British Journal of Educational Technology*, *51*(5), 1766-1784.

Moore, R. L., Oliver, K. M., & Wang, C. (2019). Setting the pace: examining cognitive processing in MOOC discussion forums with automatic text analysis. *Interactive Learning Environments*, *27*(5-6), 655-669.

Mouri, K., Uosaki, N., & Ogata, H. (2018). Learning analytics for supporting seamless language learning using e-book with ubiquitous learning system. *Journal of Educational Technology & Society*, *21*(2), 150-163.

Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, 1-20.

National Academy of Education. (2017). *Big data in education: Balancing the benefits of educational research and student privacy: Workshop summary.*

Nie, N. H., Bent, D. H., & Hull, C. H., 1975. *SPSS: Statistical package for the social sciences*, Vol. 227. New York: McGraw-Hill.

Niemelä, M., Kärkkäinen, T., Äyrämö, S., Ronimus, M., Richardson, U., & Lyytinen, H. (2020). Game learning analytics for understanding reading skills in transparent writing system. *British Journal of Educational Technology*, *51*(6), 2376-2390.

OECD. (2019). *PISA 2018 technical report*. Paris: PISA, OECD Publishing.

Olive, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Transactions on Learning Technologies*, *12*(2), 171-183.

Olivé, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2020). A supervised learning framework: Using assessment to identify students at risk of dropping out of a MOOC. *Journal of Computing in Higher Education*, *32*(1), 9-26.

Ortigosa, A., Carro, R. M., Bravo-Agapito, J., Lizcano, D., Alcolea, J. J., & Blanco, O. (2019). From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system. *IEEE transactions on learning technologies*, *12*(2), 264-277.

Papamitsiou, Z. K., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *J. Educ. Technol. Soc.*, *17*(4), 49-64.

Paquette, L., & Baker, R. S. (2019). Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. *Interactive Learning Environments*, *27*(5-6), 585-597.

Park, S., & Weng, W. (2020). The relationship between ICT-related factors and student academic achievement and the moderating effect of country economic index across 39 countries. *Educational Technology & Society*, *23*(3), 1-15.

Park, Y., Yu, J. H., & Jo, I. H. (2016). Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. *The Internet and Higher Education*, *29*, 1-11.

Pellagatti, M., Masci, C., Ieva, F., & Paganoni, A. M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *14*(3), 241-257.

Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Cristea, A. I., Carvalho, L. S., Fonseca, S. C., ... & Isotani, S. (2020). Using learning analytics in the Amazonas: understanding students' behaviour in introductory programming. *British journal of educational technology*, *51*(4), 955-972.

Poitras, E., Butcher, K. R., Orr, M., Hudson, M. A., & Larson, M. (2019). Predicting student understanding by modeling interactive exploration of evidence during an online science investigation. *Interactive Learning Environments*, 1-13.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Richards, G. (2011). Measuring engagement: Learning analytics in online learning. *electronic Kazan*, *2011*.

Riofrıo-Luzcando, D., Ramırez, J., & Berrocal-Lobo, M. (2017). Predicting student actions in a procedural training environment. *IEEE Transactions on Learning Technologies*, *10*(4), 463-474.

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.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384.

Romero, C., Ventura, S., Zafra, A., & De Bra, P. (2009). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computers & Education*, *53*(3), 828-840.

Rubel, A., & Jones, K. M. (2016). Student privacy in learning analytics: An information ethics perspective. *The information society*, *32*(2), 143-159.

Ruiperez-Valiente, J. A., Muñoz-Merino, P. J., Alexandron, G., & Pritchard, D. E. (2017). Using machine learning to detect 'multiple-account'cheating and analyze the influence of student and problem features. *IEEE transactions on learning technologies*, *12*(1), 112-122.

Ruiz, S., Urretavizcaya, M., Rodríguez, C., & Fernández-Castro, I. (2020). Predicting students' outcomes from emotional response in the classroom and attendance. *Interactive Learning Environments*, *28*(1), 107-129.

Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an institutional context. *The Internet and Higher Education*, *37*, 76-89.

Scheerens, J., Hendriks, M., & Luyten, H. (2014). School size effects: Review and conceptual analysis. In H. Luyten, M. Hendriks, & J. Scheerens (Eds.), *School size effects revisited: A qualitative and quantitative review of the research evidence in primary and secondary education* (pp. 7–40). Dordrecht, The Netherlands: Springer.

Schulz, W. (2005). Measuring the Socio-Economic Background of Students and Its Effect on Achievement on PISA 2000 and PISA 2003. *American Educational Research Association 2005*.

Schwarzenberg, P., Navon, J., & Pérez-Sanagustín, M. (2020). Models to provide guidance in flipped classes using online activity. *Journal of Computing in Higher Education*, *32*(2), 282-306.

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine learning*, *86*(2), 169-207.

Selwyn, N., & Husen, O. (2010). The educational benefits of technological competence: an investigation of students' perceptions. *Evaluation & Research in Education*, *23*(2), 137-141.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422.

Shen, L., Wang, M., & Shen, R. (2009). Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Journal of Educational Technology & Society*, *12*(2), 176-189.

Shibani, A., Koh, E., Lai, V., & Shim, K. J. (2017). Assessing the language of chat for teamwork dialogue. *Journal of Educational Technology & Society*, *20*(2), 224-237.

Shum, S. B., & Ferguson, R. (2012). Social learning analytics. *Journal of educational technology & society*, *15*(3), 3-26.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380-1400.

Siemens, G. (2017). Connectivism. *Foundations of learning and instructional design technology.*

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254).

Sin, K., & Muthu, L. (2015). APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS--A LITERATURE REVIEW. *ICTACT journal on soft computing*, *5*(4).

Singh, S., & Gupta, P. (2014). Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, *27*(27), 97-103.

Smanova, N. (2021, July). Can We Overcome the Achievement Gap between Urban and Rural Students in Kazakhstan through School Resources: Evidence from PISA 2018. In *2021 5th International Conference on Education and Multimedia Technology* (pp. 321-326).

Soffer, T., & Cohen, A. (2019). Students' engagement characteristics predict success and completion of online courses. *Journal of Computer Assisted Learning*, *35*(3), 378-389.

Steinberg, D., & Colla, P. (2009). CART: classification and regression trees. *The top ten algorithms in data mining*, *9*, 179.

Sun, D., Cheng, G., Xu, P., Zheng, Q., & Chen, L. (2019). Using HMM to compare interaction activity patterns of student groups with different achievements in MPOCs. *Interactive Learning Environments*, *27*(5-6), 766-781.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tadayon, M., & Pottie, G. J. (2020). Predicting student performance in an educational game using a hidden Markov model. *IEEE Transactions on Education*, *63*(4), 299-304.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tempelaar, D. T., Rienties, B., & Nguyen, Q. (2017). Towards actionable learning analytics using dispositions. *IEEE Transactions on Learning Technologies*, *10*(1), 6-16.

Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Procedia Computer Science*, *1*(2), 2811-2819.

Therneau, T., Atkinson, B., & Ripley, B. (2013). Rpart: Recursive Partitioning. R Package Version 4.1-3. http://CRAN.R-project.org/package=rpart

Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. *Humboldt University*, *Berlin*, *54*.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, *14*(3), 207-222.

Xiao, Y., & Hu, J. (2019). Regression analysis of ICT impact factors on early adolescents' reading proficiency in five high-performing countries. *Frontiers in psychology*, *10*, 1646.

van Leeuwen, A. (2018). Teachers' perceptions of the usability of learning analytics reports in a fipped university course: When and how does information become actionable knowledge? *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-018-09639-y

Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., & Karakos, A. (2012). A clustering methodology of web log data for learning management systems. *Journal of Educational Technology & Society*, *15*(2), 154-167.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Amsterdam: Centrum voor Wiskunde en Informatica.

Villagrá-Arnedo, C. J., Gallego-Durán, F. J., Llorens-Largo, F., Compañ-Rosique, P., Satorre-Cuerda, R., & Molina-Carmona, R. (2017). Improving the expressiveness of black-box models for predicting student performance. *Computers in Human Behavior*, *72*, 621-631.

Viswanathan, S. A., & VanLehn, K. (2017). Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Transactions on Learning Technologies*, *11*(2), 230-242.

Wakelam, E., Jefferies, A., Davey, N., & Sun, Y. (2020). The potential for student performance prediction in small cohorts with minimal available attributes. *British Journal of Educational Technology*, *51*(2), 347-370.

Wang, S. M., Hou, H. T., & Wu, S. Y. (2017). Analyzing the knowledge construction and cognitive patterns of blog-based instructional activities using four frequent interactive strategies, problem solving, peer assessment, role playing and peer tutoring: A preliminary study. *Educational Technology Research and Development*, *65*(2), 301-323.

Warm. T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.

Wayne, A. J. (2002). Teacher Inequality. *education policy analysis archives*, *10*, 30.

Weng, W., Ritter, N. L., Cornell, K., & Gonzales, M. (2021). Adopting Learning Analytics in a First-Year Veterinarian Professional Program: What We Could Know in Advance about Student Learning Progress. *Journal of Veterinary Medical Education*, e20200045.

West, D. M. (2012). Big data for education: Data mining, data analytics, and web dashboards. *Governance studies at Brookings*, *4*(1), 1-10.

Wheeler, S., Waite, S. & Bromfield, C. (2002). Promoting creative thinking through the use of ICT. Journal of Computer Assisted Learning, 18(3), 367-378.

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and soil*, *340*(1), 7-24.

Winitzky-Stephens, J. R., & Pickavance, J. (2017). Open educational resources and student course outcomes: A multilevel analysis. *International Review of Research in Open and Distributed Learning*, *18*(4), 35-49.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), 76-77.

Wright, R. E. (1995). Logistic regression.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2-3), 114-128.

Wu, J. Y., Hsiao, Y. C., & Nian, M. W. (2020). Using supervised machine learning on large-scale online forums to classify course-related Facebook messages in predicting learning achievement within the personal learning environment. *Interactive Learning Environments*, *28*(1), 65-80.

Wu, P., Yu, S., & Wang, D. (2018). Using a learner-topic model for mining learner interests in open learning environments. *Journal of Educational Technology & Society*, *21*(2), 192-204.

Xia, X. (2020). Learning behavior mining and decision recommendation based on association rules in interactive learning environment. *Interactive Learning Environments*, 1-16.

Xie, K., Di Tosto, G., Lu, L., & Cho, Y. S. (2018). Detecting leadership in peer-moderated online collaborative learning through text mining and social network analysis. *The Internet and Higher Education*, *38*, 9-17.

Xie, K., Yu, C., & Bradshaw, A. C. (2014). Impacts of role assignment and participation in asynchronous discussions in college-level online classes. *The Internet and Higher Education*, *20*, 10-19.

Xing, W., Pei, B., Li, S., Chen, G., & Xie, C. (2019). Using learning analytics to support students' engineering design: the angle of prediction. *Interactive Learning Environments*, 1-18.

Xing, W., Wadholm, R., Petakovic, E., & Goggins, S. (2015). Group learning assessment: Developing a theory-informed analytics. *Journal of Educational Technology & Society*, *18*(2), 110-128.

Xu, W., & Zhou, Y. (2020). Course video recommendation with multimodal information in online learning platforms: A deep learning framework. *British Journal of Educational Technology*, *51*(5), 1734-1747.

You, J. W. (2015). Examining the effect of academic procrastination on achievement using LMS data in e-learning. *Journal of educational technology & society*, *18*(3), 64-74.

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education*, *29*, 23-30.

Yu, L. C., Lee, C. W., Pan, H. I., Chou, C. Y., Chao, P. Y., Chen, Z. H., ... & Lai, K. R. (2018). Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning*, *34*(4), 358-365.

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, *27*, 44-53.

Zhang, D., & Liu, L. (2016). How does ICT use influence students' achievements in math and science over time? Evidence from PISA 2000 to 2012. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*(9), 2431-2449.

Zhang, C., & Zhang, S. (2003). *Association rule mining: models and algorithms* (Vol. 2307). Springer.

Zhang, J. H., Zhang, Y. X., Zou, Q., & Huang, S. (2018). What learning analytics tells us: Group behavior analysis and individual learning diagnosis based on long-term and large-scale data. *Journal of Educational Technology & Society*, *21*(2), 245-258.

Zhu, G., Xing, W., & Popov, V. (2019). Uncovering the sequential patterns in transformative and non-transformative discourse during collaborative inquiry learning. *The Internet and Higher Education*, *41*, 51-61.

Zylka, J., Christoph, G., Kroehne, U., Hartig, J., & Goldhammer, F. (2015). Moving beyond cognitive elements of ICT literacy. First evidence on the structure of ICT engagement. *Computers in Human Behavior,53,* 149–160. https://doi.org/10.1016/j.chb.2015.07.008.