Learning Analytics: A Data Mining and Machine Learning Perspective

Salam Ullah Khan[1], Kifayat Ullah[2], Mahvash Arsalan Lodhi[3], Sadaqat Ali Khan Bangash[4]

(Received December 26, 2017; Revised November 20, 2015; Accepted December 10, 2018)

DOI: 10.33317/SSURJ.V8.I2.72

***Abstract*— Tremendous proliferation in data generation in the past few years has paved the way for new research and the development of new and improved techniques and algorithms in different fields of science and education. Initially terms like educational data mining emerged as a branch of data mining borrowing techniques from its ancestor. The challenges brought about by this large and heterogeneous data are diverse and needs a greater serious technical treatment. New and emerging fields like learning analytics have been introduced to manage the complexities of this data deluge. Learning analytics deals with data in the context of learner and the learning environment to improve the overall learning experience. The ultimate aim of the field is to make use of the data about learners and their environments to gain insights into the learning process using some of the well-known techniques and algorithms from the fields of data mining and machine learning. The process involves collecting, analysis of data and reporting the results to understand and optimize the learning experience. The fields of data mining and academic analytics closely related to learning analytics. Systematic Literature Review (SLR) is a robust, organized and rigorous literature review and reporting process aimed at identifying, collecting and synthesizing the relevant literature on a research question according to specified criteria. The process is more unbiased and balanced by systematic sequence of steps. This paper presents a systematic literature review by first developing the systematic literature review protocol and then discussing the main findings of the literature review by especially focusing on the applications and uses of machine learning and data mining techniques in the domain of learning analytics.**

***Index Terms*— Systematic Literature Review (SLR), Learning Analytics (LA), Big Data, Educational Data Mining (EDM), Machine Learning (ML).**

I. INTRODUCTION

Data mining is sometimes also referred to as Knowledge Discovery in Databases (KDD) and Data Extraction. The goal of data mining is to find inherent and any significantly meaningful patterns in the data collected in any domain of interest. Data mining is a mature field with many applications and useful contributions in different domains. The techniques of data mining have widely been used in education, science, health, business and other spheres of society. Among the many established data mining models, few models like Cross-Industry Standard Process for Data Mining (CRISP-DM) have become a de-facto standard for discovering useful knowledge from the large amount of data available[1].

Artificial Intelligence (AI) in general and its subfield of Machine Learning (ML) specifically relate to teaching computers, how to learn different patterns from data and to create models and use those models in a much broader and generalized context. Hence, these insights and patterns could be used for future predictions. The fields of data mining and machine learning are closely associated. In fact, both fields work in parallel to find meaningful patterns from data and then make predications on any further data. Most of the power of machine learning comes from techniques in supervised learning such as classification, regression as well as unsupervised, active and semi-supervised learning [2].

Owing to the tremendous growth of big data, fields like education and learning science also have substantially benefitted from this setup thereby giving rise to notion of fields like Educational Data Mining (EDM) and Learning Analytics (LA).Some authors elaborate on the concept of their relationship by stating the key contribution of the EDM in LA and borrowing of concepts primarily from EDM, [3]. They give an overview of the earlier pioneering work in the fields giving the narrative of the methods used and the evolution to the status quo. The methods ofEDM in particular have contributed to the development of LA. As the field of LA gets more and more mature, the gap is getting thinner. LA and EDM have the power to turn data from learning and educational domains into more meaningful form thereby making it more suitable for gaining insights and decision making which is equally important for all stakeholders including managers involved in strategic decision making, teachers and students itself. Take the example of teachers, they can utilize the insights obtained from the analytical treatment of data to fine-tune their teaching and adopt more sophisticated procedures to better suit their students' needs. The word, 'Learning Analytics' itself is sometimes mixed up with EDM and academic analytics. While they are closely associated and have some common attributes, they still demonstrate their own peculiarities. Academic analytics focus more on applying BI techniques at macro or organizational level. EDM has its uses primarily in employing techniques to solve research challenges at micro level or individual level. LA, on the other hand, focus on the process of learning and the learner itself. Applying these quantitative measures to the data about learners can extend our knowledge and give use insights about how to improve the overall experience for a learner. Thus, learning resources could be managed in an optimal way and help us in better decision making [4].

[1]Lecturer, Department of Computer Science, University of Science & Technology, Bannu, Pakistan. Salamullah2003@yahoo.com
[2]Lecturer, Department of Electrical Engineering, University of Science & Technology, Bannu, Pakistan. umerkaif@yahoo.com
[3]Lecturer, Department of Computer Engineering, Sir Syed University of Engineering & Technology, Karachi, Pakistan. m.iq2010@hotmail.com
[4]Assistant Professor, Department of Computer Science, University of Science & Technology, Bannu, Pakistan. sabangash.81@gmail.com

In order to learn the relationship of ML and data mining with LA, the authors study the applications of data mining and ML in the field of LA. We perform a systematic review of the relevant literature to discover the current state-of-the-art and potential challenges. The research study gives us insights and technical details of the applications of these fields in LA. We explore in detail the applications, research challenges and possible future directions for applying ML and data mining techniques to LA.

As per the specified objectives of this research, a protocol was developed with research questions to help and guide the research study in a particular direction:

- RQ1: What are the critical applications of data mining and ML in the field of LA?
- RQ2: What are the challenges, as found in literature, in applying data mining and ML techniques to the field of LA?
- RQ3: What are the approaches to the challenges of applying data mining and ML techniques to LA?

II. BACKGROUND

A. The scope of Learning Analytics

The increase usage of ICT tools for learning and teaching has brought up many research challenges. A new set of tools, techniques and improved services are required to cope with these challenges. The ultimate benefit of this endeavor is to improve the learning process and offer an overall customized experience for the learners. Students and teachers alike would benefit from this improvement. It emphasize on collaboration, among experts of different fields, like statistics, education, technology and related domains. This collaboration would center around the data and manipulation of this data. The cyclic and iterative nature of activities in LA focuses on collection of data of student-teacher and student-subject interactions. Next as per the requirements of data mining and knowledge discovery process, the data must be prepared according to the nature of the problem. Thereafter, a set of analytical techniques and algorithms give useful insights about the learners and the learning process as a whole. Visualization techniques can help us better understand the structure and inherent patterns in the data. Experts from relevant fields can help guide the process by sharing their valuable experience and knowledge in the process. Besides getting useful trends and insights, learning resources could be adapted and optimized for the learners. Additionally, usability and accessibility improvement, and increased collaboration could prove very beneficial in achieving short term and long term learning goals [5].

New and emerging forms of technologies involving pervasive computing, custom classroom settings, and high tech visual displays are actively being deployed to revolutionize education. This places emphasis more on the 'physical' or 'hard' aspect of the technologies involved, while the 'logical' or 'soft' aspect relies on the algorithms and techniques for the data analysis e.g. big data analytics. The field of LA itself involves rigorous testing and experimentation in terms of application of new techniques and algorithms. Researchers are still posed with many challenges. The integration of big data analytics has to play an important role in the field of education, and this trend is constantly rising. They could play an important role in shaping educational reforms and improving the overall learning and teaching process. LA provides the necessary platform for instructors and the learners and it makes available the intuitions and insights into the whole process of learning and how to improve it. For managers and others involved in decision making life could get a lot easier and a clear picture can be acquired. This could remove the uncertainty and improve budget planning in the face of ambiguity. Overall improvement in resource allocation, managing healthy competition, quality improvement and accomplishing the goals of learning could be easily achieved [6].

Increasing interest in collecting and analyzing data from the fields of learning and education have led to incorporating traditional methods and the formulation of new and better analysis methods, techniques and tools with better predictive and decision making capabilities. Research communities for Learning Analytics and Knowledge (LAK) and EDM have made significant contribution towards this goal. Few of the authors emphasize on increased collaboration and communication among the two research communities to share the research output of EDM and LAK [7].

LA emerged in the last decade and it involves extensive use of technology based learning. Fig. 1 shows the Google Trends graph for the search term "Learning Analytics" for the last three years, i.e., July 2015 to August, 2018. The graph shows a constant growth in search trend for the increased interest in the field in general.

A comprehensive coverage of the different factors which have led to the development of newer analytical trends in the field of education and learning has been researched. The authors then discuss the association between EDM , LA and academic analytics. At the end, they pose some of the emerging research challenges and future directions in the field [8].

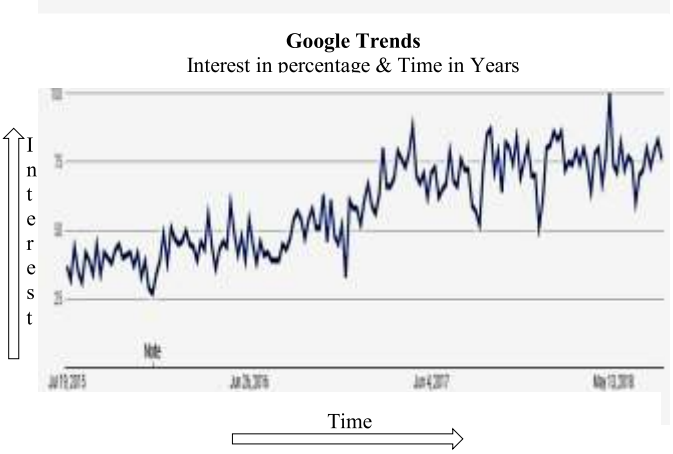

Fig. 1: Google Trends Showing Growing Interest in the Field of Learning Analytics over Time

LA is already on its way to becoming a prevalent instrument for obtaining insights about the learners and the learning process and using those insights for making better and more accurate decisions to improve the whole process of learning. It can help create a personalized learning experience by highlighting individual needs and an improved performance.

Still, there exists among a few, confusion about the scope of LA and its applications.

LA and heavy use of data from the education comes with some issues. A group of authors performed a review by outlining the dimensional scope of LA, the problem areas of more critical nature and potential pitfalls in using data from the educational context. A generic framework is charted to work as a map with guidelines for the researchers and practitioners in areas like guidance and counselling, development of curriculum and improving the overall quality of teaching. Some of the obstacles and restrictions are delineated. Key skills and experiences required for all the stakeholders for working successfully within the domain of LA are discussed. Ways of handling issues like security and privacy, ethical considerations and guidelines for adopting policies and procedures are outlined in the context of LA [9].

Research interest is increasing in LA employing educational datasets to improve the learning process. LA has evolved into a multidisciplinary field involving information retrieval, AI, statistics, ML. Another review onLA and associated fields maps them together into a reference model in four dimensions. A review of recent literature further identifies different challenges. The authors map them four different dimensions in the Reference Model. A narration of the opportunities and challenges is given [10].

B. Reference Model for Learning Analytics

LA has attracted a lot more researchers form different domains. The motivation for this is to better understand the learning process by employing the data collected and the urge to add "intelligence" and "customization" into educational systems and learning content. Research communities like SoLAR (Society for Learning Analytics and Research) and International Educational Data Mining Society (IEDMS) are formed to provide a platform and to help guide researchers and advance research in this field. The theme of all such societies is to use educational data for gaining knowledge and insights about the learners and the learning process. A detailed narrative of the fields which have significantly contributed towards the development of LA is outlined. Contemporary research issues, challenges in LA and their existing solutions and the techniques used are also described in detail. The authors emphasize taking into account increasing reliance on data, thereby enabling us to mirror the complexity of learning process more precisely. Issues like privacy, ethical challenges and ownership of data must be considered seriously to make these efforts successful [11].

Learning and education through online resources have been popular since the inception of digital resources and massively open online platforms. Some well-known examples of such platforms are Coursera, FutureLearn, Udacity, MIT OCW, EdX etc. The characteristics of these platforms include keeping track of learner's progress and provide instant feedback and visual illustration through real-time dashboard interfaces. Learners participate in these courses coming from diverse backgrounds, demographic characteristics and different time zones. The issue with analyzing data from such systems is their non-availability for public use in research. Replicated versions of these could mock the behavior of such systems in a laboratory setting. Some authors use visual techniques by taking into account the parameters like relative importance of learning content, context and their response time during the interaction. The proposed system is then tested by using software engineering principles and practices with academic data of five years. They discover that visual learning analytics can prove to be very beneficial in making better decisions and the improvement of the learning process as a whole [12].

Others have studied the applications of LA within specific domains and in specific contextual settings like few of the authors apply these techniques to know the ins and outs of the driving process. In their main findings they state that by applying the techniques from LA, the process can become more efficient and costs and human efforts could be saved in this way [13].

C. Machine Learning and Data Mining techniques used in Learning Analytics

Data mining and ML techniques play a central role in activities like predictive modeling and data driven decision making involving past data. Supervised learning techniques like regression, classification, KNN and unsupervised learning techniques like clustering, including K-means are widely used in making predictions and intelligent decision making. In addition to using recommendation systems and visual displays especially in learning dashboard designs are highly useful. The authors gave a detailed account of usage patterns and scenarios of these techniques [14]:

- Recording and analyzing classroom interaction data could give us a rough estimate of the academic performance of the learners.
- Using ML and data mining techniques can be helpful in predicting risk of student attrition and rate of retention.
- By employing academic data one can get accurate measure of the learner's interest and inclination towards certain courses. This can also be helpful in recommending courses in line with their interests. Potential field selection can become easier with this analysis.
- Techniques from data science like data analysis, data visualization and reporting are useful in understanding inherent patterns and trends in the data.
- Learning related issues in the learning environment like communication and collaboration could be easily handled.
- Techniques fromLA could be used to handle learners' feedback about the environment, the tutors and the learning tools.
- Detecting learner's behavior, estimating their skills and its modeling.

D. The era of Big Data

The increase in the generation of heterogeneous data comprising of structured data from operational systems and unstructured data in the form of text, audio and video has given rise to the era of 'Big Data' [15]. The number of applications using big data has been increased manifold. The

problem with using typical database systems for such large data lays in the characteristics of big (volume, velocity, variety) data itself. The need for, scalability with fault-tolerant systems to handle such data is fulfilled by the development of a new set of big data tools. These tools support the characteristics of big data like volume, velocity, variety and veracity [16]. Sources from which the data comes include widely deployed sensors and devices, social media and social interactions, web usage, human-machine interactions [17]. This data is heterogeneous and does not follow any schema or data model. The sheer size and velocity of the data generation can roughly be appreciated from social media. Facebook, on the average, receives nearly two hundred and fifty million posts per hour, while Twitter, another social platform receives above twenty million tweets per hour. Similar statistics for other online platforms and systems give us an idea of the characteristics of this data.

Data in any form can prove to be very valuable for organizations. Big data can prove to be very useful for organizations if managed and analyzed properly. It gives organizations a competitive advantage in better decision making. The center of gravity for the field of big data analysis is 'analysis' in a novel way where traditional database tools are not suitable to be used [18].

As discussed earlier, big data has the greatest potential both as an asset and as a powerhouse to produce better decisions for businesses and other organizations. Industries like finance, manufacturing, health, science, defense, government organizations, education and human development are all benefitted by the big data paradigm.

Despite all its advantages, big data has its own issues and limitations. Issues like security, privacy, management of the data as well as optimizing the network data with efficient use of energy are some of the common research challenges for big data researchers. There are many issue and challenges inherent within the big data domain. A comprehensive review of the available literature on these issues, their potential solutions has been studied in detail. The authors pointing out that still some of these issues pose a prompt challenge for the researcher due to heterogeneous nature of big data [19].

E. Big Data in Education

The big data paradigm is implemented by an updated set of tools and techniques, keeping the development of a scalable and fault-tolerant system in mind. The whole standard is covered by the umbrella of cloud computing platforms and infrastructures. Some authors propose a platform comprising, Massive Open Online Courses (MOOCs), based course management and recommendation system with LA support with a specific focus on scalability and fault tolerance [20]. Some other researchers provide a detailed coverage of the recent trends and developments in predicting and quantifying academic performance of learners within the context of MOOCs [21]. More recently, deep learning techniques and tensors have been used for recommending courses and specific academic content in the context of LA [22].

Similar to traditional data analytics approaches to quantification and measurement of academic data, big data approach when applied to LA, can help us find attrition risk, forecasting future academic performance, interactive visualization, analysis, reporting, reward, feedback, course recommendations, group collaboration and behavior assessment [23]. Smart learning systems involving active learning are used in a variety of ways of to use the full potential of LA [24].

F. Previous Work on Systematic Reviews in Learning Analytics

Previous literature reviews exist on the general applications of LA and EDM, while systematic literature reviews of data mining and ML techniques in general are predominantly available, no significant literature work exist on applications of ML and data mining in the context of LA. Few of the authors suggest that ML, data mining and statistical analysis techniques could be very efficiently used to make predictions about students' academic achievements and improve their learning experience. They provide extensive coverage of uses of EDM in student academic performance analysis [25]. A group of authors provide a general concept of the fields of LA and EDM. They support their work through evidence based literature and case studies and emphasize critical potential applications of both in the field of LA [26]. Also, another group of authors review existing literature the context, potential sources of data, factors, challenges and opportunities within visual LA and the LA framework [27].

III. METHODOLOGY

In this particular research work, the aim is to conduct a systematic literature review of the applications and research challenges of data mining and ML techniques within the domain of LA. The earliest known systematic literature review was conducted in software engineering and medical sciences. Procedure and guidelines to conduct systematic reviews in these fields are discussed in details by some researchers [28-30]. Other researchers have developed the extended versions of these guidelines for fields like computer science [31]. Being a more organized approach to literature review [32-34], systematic literature review is reliable and minimizes any research bias [35]. This includes a three-step process involving planning, conducting and documenting the review [36]. To get started on the process, a systematic literature review protocol is developed. The rest of the steps are all followed in the 'systematic' pattern as discussed above.

A. Development of Search String (s)

Taking into account the research questions and extracting search terms and keywords, a search string is formulated. Given below is a detailed narrative of research questions with specification of population, intervention, context, outcome and design of research instrument, see Table I.

Table I: Formulation of Search String

Specification	Research Questions
Population	Learning Analytics
Intervention	Applications, Techniques, Challenges, Limitations, Approaches
Context	Machine Learning, Data Mining
Outcomes of relevance	To appropriately analyze data related to the learning process and the learners using machine learning and data mining
Methods and techniques	Surveys and experimental studies, theoretical models

The aforementioned concepts can be demonstrated using the research questions discussed next.

B. Search Plan

As part of the search strategy, an initial trial search on mainstream digital resources including IEEE Xplore, Google Scholar, ACM Digital Library is conducted to broadly find scope and boundaries of our search. The following query strings were used:

- For Data Mining: ("Learning Analytics") AND (Applications OR Uses) AND ("Data Mining")
- For Machine Learning: ("Learning Analytics") AND (Applications OR Uses) AND ("Machine Learning")

Table II: Total Papers from Different Databases for RQ1

Publication Search Results			
Library/Database	**Total Papers**	**Conference**	**Journal**
IEEE Xplore	349	289	50
ACM DL	5624	4370	404
Google Scholar	2910	NA	NA
ScienceDirect	93	NA	NA

Table II, gives a general overview of number of papers retrieved from different digital databases. The papers retrieved from different digital databases are not mutually exclusive. These papers are used as guiding factor in the research direction and final validation of the protocol.

C. Isolation of Search Terms

Taking into account the preliminary search results, actual research questions and other guiding factors, the following procedure is adopted to isolate search terms and search strings:

- First, the research questions are used to extract major search terms e.g., PICO (Population, Intervention, Context and Outcome).
- Next, a list of synonyms andalternative meanings is formulated.
- As preliminary verification, the relevant keywords and synonyms are checked in the corresponding research papers.
- Boolean operators 'OR' and 'AND' are used to join key search terms and synonyms.
- Any modification to the search string if needed.

These steps are used for all the research questions. A sample for Research Question 1 is given as under:

Research Question 1: What are the critical applications of data mining and ML in the field of LA?

- For Applications: "Applications" OR "Practices" OR "Procedures" OR "Uses".
- For Data Mining: "Data mining" OR DM OR "Data manipulation" OR "Data Extraction".
- For Machine Learning: "Machine Learning" OR ML OR "Statistical Learning" OR "Machine Intelligence".
- For Learning Analytics: "Learning Analytics" OR LA OR "Academic analytics".

D. Resources used

We conduct our search for the search strings by using the following research databases and online libraries:

- IEEE Xplore digital Library.
- ACM Digital Library.
- Google Scholar.
- SpringerLink.
- ScienceDirect.
- CiteSeerX.

E. Restrictions on Search Criteria and Validation

All the above mentioned research databases are searched with the initial query string to get a preliminary idea of the potential applications and research challenges. The search criteria is not delimited by any restriction in publication date. By using trial search, alist of relevant publication is retrieved as already discussed previously. No restrictions and limitations are applied at this stage to avoid the possibility of missing any relevant publication. Almost all of the digital resourcesM, including IEEE Xplore (ieeexplore.ieee.org), Google Scholar (scholar.google.com), ACM DL (dl.acm.org) and Science Direct were used for search using the search terms:

- For Data Mining: ("Learning Analytics") AND (Applications OR Practices OR Procedures OR Uses) AND ("Data Mining").
- For Machine Learning: ("Learning Analytics") AND (Applications OR Practices OR Procedures OR Uses) AND ("Machine Learning").

Results obtained using this method is helpful in guiding the research work and final validation of search strings.

The final search strings developed are given as under:

a) Search String for Research Question 1

("Learning Analytics" OR LA OR "academic analytics") (Applications OR Practices OR Procedures OR Uses) AND ("Data mining" OR "Data manipulation" OR DM OR "Data Extraction" OR "Data analysis" OR "Data analytics") AND ("Machine Learning" OR ML OR "Machine Intelligence" OR "Statistical Learning") AND (Applications OR Procedures OR Uses OR Practices).

b) Search String for Research Question 2

("Learning Analytics" OR LA OR "academic analytics") AND ("Data mining" OR "Data manipulation" OR DM OR "Data Extraction" OR "Data analysis" OR "Data analytics") AND ("Machine Learning" OR ML OR "Machine Intelligence" OR "Statistical Learning") AND (Challenges OR Issues OR Problems OR Barriers OR Hurdles OR Difficulties OR Obstacles).

c) Search String for Research Question 3

("Learning Analytics" OR LA OR "academic analytics") AND ("Data mining" OR "Data manipulation" OR DM OR "Data Extraction" OR "Data analysis" OR "Data analytics") AND ("Machine Learning" OR ML OR "Machine Intelligence" OR "Statistical Learning") AND (Approaches

OR Methods OR Narratives OR Tasks OR Experiments OR Solutions OR Explanations).

F. Search Documentation

The results of online databases and library search are documented in the following predefined format in a well-known spreadsheet application like SPSS or MS Excel:

- Search Date.
- Name of Online Database/Library.
- Numbers of years included in the search.
- The formatted search string.
- Numbers of publications retrieved.
- Initially selected publications.
- Final selection of publications.

G. Organizing the Search Results

Electronic record of papers retrieved from the different research databases is maintained in a separate directory with specific subdirectories created for each database. For additional formatted resources like HTML pages and image resources, separate subdirectories are maintained. The format given in Table III shows a sample of the directory structure. "Tracking Number" is a composite two-part field with the first part being resource identifier and the second part as the serial number of the resource retrieved from the corresponding database. As an example 'IEEE-16' means the entry is at serial number 16 and that it has been extracted from IEEE Xplore database. Any redundant entries or similar resources found in different databases are excluded. However, appropriate record keeping procedures for those entries are used subsequently.

Table III: Organizing the Search Results

SLR Search Results			
S No.	**Tracking No.**	**Digital Resource Used**	**Paper Title**
1.	GS-16	Google Scholar	Title of the paper
2.	CS-55	CiteSeer	Actual title
3.	IEEE-16	IEEE Xplore	Actual title
4.	----------	INMIC'16	Actual title

H. Publication Selection

A thorough resource selection procedure is adopted for inclusion and exclusion of publications and related resources. Only resource fulfilling the search criteria is retained. Other resources including books, original research work and technical reports are also carefully examined in order to include only most relevant literature.

a) Inclusion Criteria

A rigorous inclusion criteria comprising literature related to LA is formulated. The resource to be selected conformed to the following conditions:

- Literature focusing on applications of data mining and ML to LA.
- Literature discussing challenges in applications of data mining and ML to LA.
- Literature resources that describe approaches to the issues and challenges mentioned above.
- Literature discussing limitations of the above mentioned solutions.

b) Exclusion Criteria

The exclusion criteria given below ignore any resource not satisfying inclusion criteria:

- Literature without any significance for the research questions.
- Literature disregarding discussion of LA resources with no primary focus on LA.
- Literature with no narration of resources and also with no particular discussion on applications and challenges of ML and data mining to the domain of LA.

I. Selection of Primary Resources and Quality Assessment

In a systematic review of literature, while making the decision about including or excluding a particular resource we look at paper title, abstract and keywords. The material only fulfilling the selection criteria and corresponding research questions is selected. A full text review further clarifies things with respect to inclusion and exclusion of the resource. Ambiguity or items irreverent to the research question and string criteria, is removed and only a subset of resources is retained. Each and every step is properly documented with justification for a particular resource selection.

On completing the selection of literature resources, the quality assessment process follows. This is done side by side with resource extraction. The following checklist is considered for each publication with keywords 'Yes', 'No', 'NA' and 'Partial':

- Are applications of ML and data mining in LA clearly discussed?
- Does the resource discuss issues and challenges of ML and data mining in LA?
- Whether the techniques and procedures for tacking those challenges stated exclusively?

The process is constantly monitored for validation and verification.

J. Data Extraction Scheme

a) Identification and Selection of Primary Data

As already discussed the objective of performing systematic literature review is to search and document resources guided through an organized approach defined by research questions. Detail of data and meta-data of the resources are recorded including title, authors and coauthors, conference or journal information.

In order to better cope with research questions the following portion of data is identified:

Research Question 1: Context and applications of ML and data mining in LA.

Research Question 2: Context and challenges of data mining and ML applications in LA.

Research Question 3: Context including approaches to tackle the challenges of applications of ML and data mining to LA.

The following data with metadata is extracted from the collection of papers selected from the literature:

- Date of Review.
- Title of the Paper.
- Author/List of authors.
- Reference/Citations (available).
- Library/Database used.
- Research Methodology used (case study, standard research paper, technical report, survey).
- Attributes about quality of the paper.
- Year published.
- Research area/subarea within the LA domain.
- Detail of the specific data mining/ ML techniques used.
- Input data (description of the data used).
- Model/architecture or reference framework if used.
- Description of the proposed technique.

b) Data Extraction Process.

The data extraction process consists of the first author as primary review with the other authors assisting the first author in guiding and documenting the necessary observations. Following this review steps are used to validate the data extraction phase. The other reviewers choose some publications at random from the list of selected publications. From the chosen publications, the supporting reviewers extract data at random. The comparison of the two reviewers would better validate the data. Others reviewers, if available, would help supervise and guide the whole process.

c) Storage of Data

The publications with their summaries and any metadata are kept in electronic form in a predefined documented structure. The data will be available for review and a safe and a secure backup mechanism is adopted to avoid any loss of this data. Any updates to the data are instantly propagated to all the backup sources to reflect the updates.

K. Data Synthesis

The data synthesis phase comprises a three-step process i.e., one step for each of the research questions. Data for each of the applications, challenges and approaches is kept in a proper tabular form. For RQ1, we keep track of the applications of ML and data mining to the domain of LA. For Research Question 2 and 3, we document the detail on challenges and currently used approaches to those challenges with any issues and limitations if available.

L. Protocol Validation

As per standard criteria and literature, the validation process takes places by involving experts and the necessary suggestions and amendments were incorporated accordingly. Their expert opinion helped improve the protocol and its research questions. The first author developed the initial draft of the protocol and the coauthors reviewed it. The points highlighted by the coauthors were helpful in first refinement of the protocol. Some other senior faculty members, including those mentioned in the acknowledgement section were constantly involved in the whole process. The final draft was a result of these reviews and suggestions by the coauthors and others experts in the field.

IV. DISCUSSION AND MAIN FINDINGS

A. Applications of Machine Learning and Data Mining in the domain of Learning Analytics

Different ML and data mining techniques in the context of big data can be used in LA. These techniques could be used in predicting performance of individual students, measuring student retention rate and attrition risk, data visualization and reporting, recommending courses for students, intelligent feedback, student behavior detection and estimation of student skill level.

Big data techniques and frameworks have played vital role in different fields. Most of these techniques and frameworks are most suitable for learning analytics. Integration of these big data techniques plays a significant role in the field of learning analytics due to large volume and heterogeneous nature of data. Some research work done earlier does not explicitly mention LA. The authors propose a generic big data based framework to incorporate a unified system for the collection, processing, storage and analysis and visualization of data and to subsequently create alerts system for the end users. These including components like data gathering components, data storage and management facilities, data analysis, data visualization and action subsystems.

B. Machine Learning and Data Mining Techniques used in Learning Analytics

There are various machine learning and data mining techniques available. These techniques and their uses can be seen in Table IV.

V. ACKNOWLEDGMENT

We acknowledge the guidance and constant help of the members of Department of Computer Science, University of Science and Technology, Bannu especially, Dr. Ihsan Rabbi and Dr. Abdul Wahid Khan at every stage of the study. Further, encouragement and support from Dr. Umar Farooq among others helped us accomplish our goal of carrying out this research work well within the stipulated boundaries.

VI. CONCLUSION AND FUTURE WORK

The results obtained from the systematic literature review of the list of selected publications from the digital resources related to our research questions show a growing research interest in the domain of learning analytics. More and more research work is appearing with underlying applications of ML and data mining techniques in LA.

ML and data mining techniques like classification, clustering, artificial neural networks, collaborative filtering and recommendation systems, social network analysis and visualization are some of the most frequently used techniques in recent research work. Keeping in view these growing trends, it is estimated that the applications of ML and data mining will increase in the field of LA. As part of future work, newest, learning techniques and models like deep learning and semantic interpretation could be reviewed with reference to their role and applications in LA using systematic literature review procedures.

Table IV: Machine Learning and Data Mining Techniques and their Uses in Learning Analytics

Publication	Technique Used	Detail/Application	Context and Limitations
[37]	Naïve Bayes and Decision Tree	Prediction of degradation of academic performance	Academic attrition prediction at Universidad Nacional de Colombia
[38]	Various types of Clustering Techniques	Review of Clustering Techniques in educational environment	Educational Data Clustering (EDC)
[39]	Induction Rules and Decision Trees	Student school dropout or failure prediction	Over 600 school student's data from Mexico
[40]	Artificial Neural Networks	Propose, design, develop and test a near real-time framework COMPASS to classify comprehension level of a learner during e-learning activity.	44 Undergraduate students on screen activity data used for classifying comprehension level of the students
[41]	Decision Tree and Artificial Neural Network Classification	Predict Student academic retention and risk assessment	Over 900 Engineering students' data in Colombia
[42]	Single-label text classification	Classification of Teacher Online professional development level of reflection and evolution	Sampling of data from 17,624 online posts 6,650 in-service K12 teachers
[43]	Classification	Propose and develop a classification system to measure learners' collaboration in real-time	Audio and video data collected from students' verbal interactions and action logs while solving complex math problems
[27]	Visualization	The proposed system includes integration of learning tools to discover learning issues and content relevance during learning and interaction sessions.	Data involving academic Interaction logs of five years for the subject of software engineering
[22]	Social Network Analysis	Use of Social Network analysis techniques with learning analytics framework	Develop student's self-regulated learning (SRL) skills during course
[46]	Data preprocessing	Identification of data preprocessing phases necessary for educational and learning analytics data	Quantitative and Qualitative data from VLEs (Virtual Learning Environments)
[22]	Collaborative filtering/Recommender Systems	To monitor and take full advantage of temporal and other contextual setting collaborative filtering techniques are developed using two types of matrix factorization	Predict grades and context from the data gathered a University of Minnesota
[48]	Classification	Predict academic success factors and chances of student success	CS Student data used to assess performance and chances of academic success
[49]	Reduced training vector-based support vector machine (RTV-SVM)	Predict marginal, at-risk in a virtual university learning environment	Prediction accuracy between 91.2 and 93.5% for marginal and 92.2 to 93.8% for at-risk students
[48]	Classification/Deep Learning	Using agile development with Deep learning algorithms to Model learning analytics for Internet of Everything (IoE)	Discuss uses, issues and challenges of deep learning in the context of education using IoE data
[25][49][50]	Systematic Literature Review	Systematic Literature review of educational data mining and learning analytics and visual LA	Restricted to the aforementioned domains only

REFERENCES

[1] Cios K. J. et al., (2007). *Data mining: a knowledge discovery approach*, USA: Springe r Science & Business Media. Retrieved from https://www.springer.com/gp/book/9780387333335.

[2] Han, J., Pei, J.& Kamber, M. (2011). *Data mining: concepts and techniques*, USAMorgan Kaufmann. Retrieved from https://www.elsevier.com/books/data-mining-concepts-and-techniques/han/978-0-12-381479-1.

[3] Baker, R. S. & Inventado, P. S. (2014). *Educational Data Mining and Learning Analytics (61-75)*, New York, NY: Springer New York.

[4] Doug, C. (2013). *An overview of learning analytics* : *Teaching in Higher Education (683-695),* Milton Keynes, UK :Routledge vol. Retrieved from https://www.ingentaconnect.com/content/routledg/cthe/2013/00000018/00000006/art00009.

[5] Guenaga M., & Garaizar, P. (2016)From Analysis to Improvement: Challenges and Opportunities for Learning Analytics. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (The IEEE Journal of Latin-American Learning Technologies)*, *11*(1), 146-147.

[6] Siemens, G., &Long, P.(2011). *Penetrating the fog: Analytics in Learning and . Education* Retrieved from https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education.

[7] Siemens, G., & Baker, R. S. J. d. (2012). *Learning analytics and educational data mining: towards communication and collaboration.* Presented at the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, British Columbia: Canada.

[8] Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(1), 304-317.

[9] Greller, W., &Drachsler, H. (2012). *Translating learning into numbers: A generic framework for learning analytics.* Retrieved from https://www.researchgate.net/publication/234057371_Translating_Learning_into_Numbers_A_Generic_Framework_for_Learning_Analytics.

[10] Chatti, M. A. et al., (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(1), 318-331.

[11] Siemens, G. (2013).Learning analytics: The emergence of a discipline. *American Behavioral Scientist, 57*(1), 1380-1400.

[12] Conde, M. Á.e t al., (2015). Exploring Software Engineering Subjects by Using Visual Learning Analytics Techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (The IEEE Journal of Latin-American Learning Technologies),, 10*(1), 242-252.

[13] Pañeda, A. G. et al., (2016). An Architecture for a Learning Analytics System Applied to Efficient Driving. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (The IEEE Journal of Latin-American Learning Technologies), 11*(1) , 137-145.

[14] Roy, S. & Singh, S. N. (2017). *Emerging trends in applications of big data in educational data miningand learning analytics.* 7th International Conference on Cloud Computing, Data Science & Engineering Confluence,(193-198), Noida : India.

[15] Chen, M., Mao, S., & Liu, Y. (2014). Big Data : A Survey, *Mobile Networks and Applications, 19*(1), 171-209.

[16] Yadav, R., & Sharma, A. (2016). *A research review on approaches/techniques used in big data environment.*P resented at Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), (242-252), Waknaghat, India : IEEE.

[17] Zafar, R., et al., (2016). *Big Data: The NoSQL and RDBMS review.* Presented at 2016 International Conference on Information and Communication Technology (ICICTM), (120-126), Kuala Lumpur, Malaysia : IEEE.

[18] Arora, Y., & Goyal, D. (2016).*Big data: A review of analytics methods & techniques.* Presented at 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), (225-230), Greater Noida, India : IEEE.

[19] Nelson, B., &Olovsson, T. (2016). *Security and privacy for big data: A systematic literature review.* IEEE International Conference on Big Data (Big Data), (3693-3702), Washington DC, USA : IEEE.

[20] Ruipérez-Valiente, J. A., et al., (2017). Scaling to Massiveness With ANALYSE: A Learning Analytics Tool for Open edX. *IEEE Transactions on Human-Machine Systems, 1*(1), 1-6.

[21] Duru, I., Dogan, G. & Diri, B. (2016). *An overview of studies about students' performance analysis and learning analytics in MOOCs.* 2016 IEEE International Conference on Big Data (Big Data), (1719-1723), Washington DC, USA : IEEE.

[22] Almutairi F. M.,Sidiropoulos, N. D& Karypis, G. (2017) Context-Aware Recommendation-Based Learning Analytics Using Tensor and Coupled Matrix Factorization. *IEEE Journal of Selected Topics in Signal Processing, 11*(5), 729-741.

[23] Merceron, A., Blikstein, P. & Siemens, G. (2016). Learning analytics: from big data to meaningful data. *Journal of Learning Analytics, 2(*1), 4-8.

[24] Hammad, R., & Ludlow, D. (2016). *Towards a Smart Learning Environment for Smart City Governance.* 2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC), (185-190), Shanghai, China : IEEE/ACM.

[25] Vieira, C., Parsons, P. & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education, 122(*1), 119-135.

[26] Papamitsiou, Z. & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society, 17(*4), 49-64.

[27] Conde, M. Á.et al., (2015). Exploring Software Engineering Subjects by Using Visual Learning Analytics Techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (The IEEE Journal of Latin-American Learning Technologies), 10*(1), 242-252.

[28] Kitchenham, B. (2004). Procedures forperforming systematic reviews (Joint Technical Report, Keele University, Keele, UK). Retrieved from http://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf.

[29] Kitchenham, B. et al., (2002). Preliminary guidelines for empirical research in software engineering.*IEEE Transactio ns on software engineering, 28(*1),721-734.

[30] Kitchenham, B., & Stuart C. (2007). Guidelines for performing systematic literature reviews in software engineering. (Joint Technical Report, Keele University, Keele, UK). Retrieved from https://userpages.uni-koblenz.de/~laemmel/esecourse/slides/slr.pdf.

[31] Weidt, F. & Silva, R. (2016). Systematic Literature Review in Computer Science : A Practical Guide (Technical Report, Federal University of Judge De Fora (UFJF), Juiz de Fora, Brazil). Retrieved from https://nrc.ice.ufjf.br/seer/index.php/relate/

[32] Kitchenham, B. (2004).Procedures forPerforming Systematic Reviews (Joint Technical Report, Keele University, Keele, UK). Retrieved from http://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf.

[33] Kitchenham, B. & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software (Joint Technical Report, Keele University,University of Durham). Retrieved from https://userpages.uni-koblenz.de/~laemmel/esecourse/slides/slr.pdf

[34] Staples, M. & Niazi, M. (2008). Systematic Review of Organizational Motivations for Adopting CMM-based SPI. *Information and Software Technology Journa*l, *50*(1), 605-620.

[35] Mulrow, C. D. (1994). Systematic Reviews: Rationale for Systematic Reviews. *British Medical Journal, 309*(1), 597-599.

[36] Brereton, O. P. et al., (2005). Service-Based Systems: A Systematic Literature Review of Issues (Computer Science Technical Repor, Keele University , Keele, U.K).

[37] Guarín, C. E. L., Guzmán, E. L. & González, F. A. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (The IEEE Journal of Latin-American Learning Technologies)*, *10*(1), 119-125.

[38] Dutt, A. , Ismail, M. A.& Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access, 5*(1), 15991-16005.

[39] Marquez-Vera, C., Morales, C. R., & Soto, S. V. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, (The IEEE Journal of Latin-American Learning Technologies), 8(*1), 7-14.

[40] Holmes, M. et al., (2018). Near Real -Time Comprehension Classification with Artificial Neural Networks: Decoding e-Learner Non-Verbal Behavior. *IEEE Transactions on Learning Technologies, 11(*1), 5-12.

[41] Rubiano, S. M. M. & Garcia, J. A. D. (2016). Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance. *IEEE Latin America Transactions* , *14*(1), 2783-2788.

[42] Liu, Q. (2018). Mining Online Discussion Data for Understanding Teachers’ Reflective Thinking. *IEEE Transactions on Learning Technologies, 11(*1), 243-254.

[43] Viswanathan, S. A. & VanLehn, K. (2018). Using the Tablet Gestures and Speech of Pairs of Students to Classify Their Collaboration. *IEEE Transactions on Learning Technologies, 11(*1), 230-242.

[44] Gewerc, A. Rodríguez-Groba, A. & Martínez-Piñeiro, E. (2016). Academic Social Networks and Learning Analytics to Explore Self-Regulated Learning: a Case Study. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (The IEEE Journal of Latin-American Learning Technologies),, 11*(1), 159-166.

[45] Munk, M. et al., (2017).Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques. *IEEE Access,5* (1), 8989-9004.

[46] Salmeron-Majadas, S. (2018). A Machine Learning Approach to Leverage Individual Keyboard and Mouse Interaction Behavior From Multiple Users in Real-World Learning Scenarios.*IEEE Access, 6*(1), 39154-39179.

[47] Chui, K. T. et al., (2018). Predicting at -risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior.* Retrieved from https://doi.org/10.1016/j.chb.2018.06.032.

[48] Ahad, M. A., Tripathi, G. & Agarwal, P. (2018). Learning analytics for IoE based educational model using deep learning techniques: architecture, challenges and applications. *Smart Learning Environments, 5(*1), 7-12.

[49] Roy, S. & Garg, A. (2017). *Analyzing performance of students by using data mining techniques a literature survey.* Papers presented at 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), (130-133), Uttar Pradesh : India.

[50] Papamitsiou, Z. & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society, 17*(1), 49-64.