Systematic Analysis of Datamining Methods in Education

Husam A.S Khalifa

Computer Science Department, Faculty of Science, Bani Waleed University, Libya

husam.khalifa@bwu.edu.ly

Submitted: 01/11/2024. Accepted: 23/11/2024. Published: 01/12/2024

Abstract

Educational Data Mining (EDM) is a rapidly expanding discipline focused on exploring data generated within educational settings. This paper conducts a systematic analysis of the EDM research literature to: (a) investigate the data mining methods used in academia and non-academic institutions; (b) review and compare these methods; and (c) classify them to understand their potential for solving educational issues. Based on a structured review of key EDM journals and conferences, the study categorizes common methods into major groups, including classification (e.g., C4.5, CHAID), clustering (e.g., K-means), association rule learning, and regression analysis. The paper also examines the primary data sources (e.g., Learning Management Systems, Student Information Systems) and essential data preprocessing techniques (e.g., data cleaning, transformation, reduction). Key applications are explored through case studies, such as predicting student performance and identifying at-risk students. Finally, the analysis addresses significant challenges (e.g., data privacy, data quality, algorithmic bias, scalability) and outlines future directions, including the integration of AI, real-time data analysis, and the development of personalized learning experiences.

Keywords: Educational Data Mining (EDM), Systematic Analysis, Machine Learning, Classification, Clustering, Regression, Student Performance Prediction, At-Risk Students, Data Privacy, Algorithmic Bias, Personalized Learning.

1. Introduction

Educational data mining (EDM) is a discipline that investigates methods for exploring and analyzing any type of data, in particular data generated in a learning context. It was born in 2009 under the initiative of a group of researchers from different universities and countries. Together with the Journal of Educational Data Mining and the International Educational Data Mining Society, it has grown into a substantial and active interdisciplinary research community with a highly regarded annual conference, a journal, and a professional society.

Besides the community, some other aspects of EDM must also be considered. Education and learning not only in schools need to be supported by the community. Data mining methods must be adapted to meet the educational system needs. There need to be computer applications using and applying those methods, or algorithms, to an educational context. So information from information appliances must be collected, selected, and stored concerns. In fact, processes to transform that information into knowledge must be developed. The resulting knowledge must be put into an appropriate form, so information appliances can use it. Finally, each appliance must be applied to and perform its tasks in a particular educational setting (Moscoso-Zea et al., 2019).

This paper conducts a systematic evaluation of the EDM research literature. It aims to (a) investigate what data-mining methods are being used in academia and non-academic institutions; (b) review and compare the identified data-mining methods and describe their characteristics; and (c) classify these methods. All of the above will be used to investigate the potentialities of them in solving educational issues. Although there are some general reviews of EDM, they ignore to explore the state and characteristics of methodologies, which forms the basis of any research area. Understanding the methodologies used in a research area would help newcomers when starting research in a small field. It would also help researchers establish collaboration in heterogeneous collaborations.

To achieve the aims stated above, the EDM research literature has been systematically searched, analyzed, and coded. The results are presented and discussed in the light of the development of EDM as an upcoming discipline. This paper is organized as follows: Section 2 examines the methodology underlining the research, Section 3 reports the results; Section 4 discusses the future of EDM, and finally, conclusions can be found in Section 5.

2. Overview of Datamining in Education

In the present world, Education plays a significant role and gives a pathway for individual growth in all aspects. In today's scenario, a student's performance cannot be assessed only by marks mathematically. There are several factors for a student's academic performance in their educational domain and predicting a student's performance in their studies is one of them. In this paper, the factors that affect the student's academic performance at Shree Rayeshwar Institute of Engineering and Information Technology (SRIEIT) are studied. Individual subjects in a semester syllabus are considered and depending on the student's subject marks the effective subjects, good, average, and poor academic performances of the student are predicted. For this purpose, classification and clustering techniques of data mining are used. Decision trees, and K-means clustering techniques are used on more than 4600 records of 4 years of data in Rapid Miner software. Academic and demographic attributes are used in the data analysis. The results show the best subject that shows the positive influence on others and they are good, average, and poor student categories depending on their marks in individual subjects. Educational Data Mining is emerging as an interdisciplinary research area with the goal of developing methods to analyze data from educational settings to better understand students and the settings in which they learn. Since educational data mining gathers a variety of techniques from machine learning, analytics, statistics, visualizations, and others, this panel intends to elicit discussions on the similarities and differences of the approaches, algorithm space, and the potential synergy between research areas (Satyanarayana & Nuckowski, 2016).

3. Types of Datamining Methods

Many educational institutions possess large databases filled with relevant information which may be analysed giving insights that identify trends and patterns (Moscoso-Zea et al., 2019). In this paper the term of Data Mining is used to refer to the group of software tools that have been created in order to mine data

Several methods were developed in the last decades to help in the process of mining data into knowledge, by means of extraction of useful information. The group of methods is large and usually it is appealing to work only on one methodology, discarding the other possibilities. This work aims to provide a systematic framework applicable on the context of educational data mining methods, it is expected that the systematization makes easy not only to review the literature but also to identify possible research opportunities in the field. Firstly, the methods available on the literature are screened in order to categorize them. Each method is analysed in terms of its functionality, main characteristics and possible software tools available.

A protocol to systematically identify the research papers on educational data mining was developed. The search domain has been focused on three venues: Journal on Educational Data Mining, proceedings of International Conference on Educational Data Mining and proceedings of International Conference on Artificial Intelligence in Education. The literature was analysed by conference year in order to tract its development. For each conference year, all the papers were read and the corresponding information was noted in the database. Educational Datamining methods were classified into four main categories: Preparation, Descriptive, Predictive and Visualisation. The principal characteristics of these methods are reported. A systematic and comprehensive overview of methods in educational datamining. Systematic and comprehensive account of the available methods in educational datamining. Educational data mining conferences and journals represented a clear clustering of publications. A large heterogeneity in substantive research problems. Topics of social and affective matters are relatively less addressed. Research opportunities were highlighted. Descriptive analytics, based on methods such as clustering, text mining and multi-dimensionality, were scarce in contrast with prediction methods.

3.1. Classification Techniques

Many classification techniques have been applied on educational data and six methods among these have been reviewed for understanding their objectives and techniques. These classification methods as applied on educational databases are classified into two categories: Classifiers based on Symbolic Decision Trees and Classifiers not based on Symbolic Decision Trees (Kumar Yadav & Pal, 2012).

Classifiers based on Symbolic Decision Trees Classifiers based on Symbolic Decision Trees detect the classification model in the form of decision trees that can be represented in a human-readable format. Six among these are reviewed.

C4.5 C4.5 building classifier generates a decision tree from a set of training data using a divide-and-conquer algorithm. The tree. Internally, nodes are of three types: Decision nodes, leaf nodes, and root nodes. Each decision node tests an attribute and has one branch for each possible outcome of the test. Each leaf node is associated with a class label. C4.5 supports several functions for pruning trees when they grow too complex. An initial tree is constructed from the training set. Then this tree is pruned by removing branches that do not contribute significantly to the accuracy of the tree on a separate pruning set. After pruning has been completed, the classification accuracy of the final tree is estimated from this other pruning set. This information is used by an adjustment algorithm to take an existing set of class probability estimates for the leaves of the pruned tree and revise the probabilities to make them more accurate.

CHAID CHi-squared Automatic Interaction Detection builds trees in a way similar to C4.5 but tests only nominal attributes. It tests for association between an attribute and the target variable using the Pearson Chi-squared statistic for categorical variables and the analysis of variance for continuous variables. It splits each of the selected nodes into child nodes and continues recursively, unless there are too few cases in the nodes, splits with no statistical significance, or one of the child nodes has zero observations. C45S Classifier for building classification trees, like C4.5, but relying on symbolic attributes. 70% of the classes were used for building the model and 30% as test data.

3.2. Clustering Techniques

Data mining, as the extraction of implied, previously unknown and potentially useful information and knowledge from a large amount of data, is a comprehensive and typical frontier field of computer science. Data mining is to automatically discover and extract information rules from large amount of data, which is an information processing technology to make comprehensive, deep and efficient use of the existing data. Many researchers have investigated the algorithms, mainly focused on the specification and exemplarification of the domains, attributes, databases, actions and

constraints of data mining, as well as on the graphical expressions of the rules and knowledge discovered. Recently, recent topics include active data mining, distributed data mining, big data mining, geographic data mining, web content mining and web usage mining.

Clustering is a type of unsupervised learning. A clustering method is to classify the examples into groups according to some degree of similarity. By giving a description of the learned clusters, data abstraction can be provided . A robust clustering method will return clusters of similar sizes and shapes (e.g., spherical and isotropic clusters). Statistical approaches based clustering rely on the assumption of underlying distribution of the data. Most of clustering methods make a hard assignment of examples to clusters. Nevertheless, cases often cannot belong purely to one cluster. A cluster of overlapping boundaries can make it unsatisfactory to know which cluster a point belongs to. Density-based clustering works well for arbitrary shapes. Nevertheless, they often fail for high-dimensional data. A -nearest neighbour graph captures high-dimensional structure but is difficult to construct for large data sets due to its O(n2) complexity. Model-based clustering formulates clustering as an optimization problem, but the optimization problems are often NP-hard.

Clustering to several classes or clusters enables high similarity among objects in identical clusters, with significant differences among various clusters. Through aggregation, one can identify dense and sparse regions and discover global distribution patterns among data attributes. Oftentimes, an inferior hybrid clustering result is obtained. Since every method has its preference, the disadvantage of one is hidden and new information is mined with the combination of several methods. Various clustering approaches, partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods, have evolved with recurrently different techniques and different mathematical constructs. In each approach, many collaboration processes have been developed to reallocate cluster assignments. Fifty-five clustering methods are regarded, which greatly broaden the choice of clustering methods in remedial comparison.

3.3. Association Rule Learning

In education, students' results are vital, providing significant information for decision-making. Academic planners cannot rely only on the processes of generating students' results from the transactional system database, as nothing is mined from such results generated. There are important decisions to be made based on such repository. Therefore, knowledge discovery techniques need to be applied to this vital repository. Knowledge representing, storing, and manipulating it for decision-making process is highly required. Discovering relationships in students' results repository is imperative to enhancing academic planners' decision-making and improving students' performance. Education system in Nigerian institutions does not differ from the grading system in other institutions globally, and this system considers students' performance in courses with letter grade designation ranging from A to F.

Analysis of students' academic failure is a cold process in many higher institutions worldwide. Techniques are employed to analyse performance as a metric, with minimum insight into academic failures. Different analysis has been done on students' result repository but failed courses in isolation have never been analysed in an academic environment for hidden and important patterns, which could be of great importance to academic planners in making better decisions. The analysis of failure is necessary, and its associated rules on failed courses using association rule learning will be beneficial to educational planners.

Association rule mining associates one or more attributes of a dataset with another attribute, discovering hidden relationships between the attributes and producing an if-then statement concerning attribute values in the form of rules. Knowledge of academic failure of students supports better curriculum delivery by strengthening weak topics or areas.

The association rules in the successes' repository would enhance the establishment of partner related courses that promote better performance due to the known strengths of the affected students. Students who fail a course are much likely to fail partner courses which prerequisite such a course. The rules could also expose courses that are too difficult for the batch or related items that constitute rigid hurdles for the students. The discovered rules will be helpful for educational planners in quickly identifying lots of relationships between students' failings. Community of different course failures can be identified, and recommendations made on course delivery to enhance teaching and learning.

3.4. Regression Analysis

Regression analysis is one of the traditional methods of data mining that has been often used in educational data mining. It examines the relationship between dependent and independent variables and is based on the statistical procedure of regression modeling. There are some different types of regression analysis such as Linear Regression, Multiple Regression, Logistic Regression, and others.

Using Multiple Regression, the relationship of three independent variables of height, gender, and age with the dependent variable of weight has been on students in schools in Tabriz, Iran. A theoretical model was created with which the required weight of a student can be specified based on height, gender, and age. This model could also be used for prediction. The results show that the created regression model could significantly estimate students' weights based only on the height, gender, and age variables (Hajizadeh & Ahmadzadeh, 2014).

A method was needed for estimating university students' overall academic performance based on Cumulative Grade Point Averages (CGPA) and the result of an English Language Proficiency Test. A statistical technique known as Multiple Linear Regression was applied for formulating a model for estimating academic results. Using the independent variables, this model accurately predicted students' academic performance. To assess the performance of the formulated models based on regression, Mean Absolute Percentage Error (MAPE), and R2 values were investigated. The models formulated from obtaining normalized data with the application of transformation were obtained to possess better values of performance matrices than the models of raw data (Johnson et al., 2012).

4. Data Sources in Educational Datamining

Educational data mining (EDM) is defined as the area of scientific inquiry centered around the development of methods for making discoveries within data from educational settings and using those methods to better understand students (Rudresh Shirwaikar & Rajadhyax, 2012). EDM encompasses a wide spectrum of data from disciplinary perceptions and algorithms from data mining and machine learning. With the advancement of technology, data in education is growing tremendously from principles such as learning analytics, big data, learning management systems, online learning systems, and automated assessment systems. EDM differs fundamentally from traditional educational research in that many empirical hypotheses are based on data mining directly.

This paper presents a systematic review of the existing literature of data mining in education published between 2004 and 2016 and proposes a systematic research agenda with respect to future developing and exploiting purified, scaled data mining tools for educational settings. The systematic review was rigorously designed with respect to sourcing data, choosing analysis tools, and reporting results. The results suggest that the publications covering processes, validating, or applying data mining in education are growing exponentially. Data mining in education is mainly studied in North America, West Europe, and Northeast Asia.

However, the concentration differences of these publications suggest that there is considerable room for further dissemination and adoption in other countries and regions. Data sources used in research of data mining in education,

which were proposed as those from which large volumes of data can contribute to understanding educational processes, may be classified as 1. Educational environments and platforms that do not emphasize assessment, including learning management systems, virtual learning environments, online courses, and communities and social networks; 2. Learning and assessment agents, systems, and recommender systems; Personalized instructional agents, intelligent tutoring systems, and intelligent agents; and 3. Administrative systems of educational organizations, including institutional records and applications for admission, enrollment, course registration, and grades.

The considerations in choosing backgrounds and design challenge the preferred education data mining research strategy to be one of investigating furniture styles in apartments over which theories and principles can effectively prosperently (Adam Whitley, 2018). A broader and richer research strategy resulted from the supporting researched background of the focus on education data mining can yield deeper insights and more insights of educational systems or enhance the exploration of subjects and the performance of models.

4.1. Learning Management Systems

Web-based educational systems, also known as Learning Management Systems (LMSs), Course Management Systems, or Learning Content Management Systems, are systems offering several facilities for information sharing and communication between the participants. LMSs are e-learning platforms allowing the distribution of the information to the students and facilitate content production, assignment preparation, distance class management, discussions, and collaborative learning. Because of their features and affordances, these systems are now ubiquitous in schools, colleges, universities, open and distance learning institutions, enterprises, service organizations, and even at home. Further, there are both commercial and free-of-cost open-source LMSs such as Moodle, Claroline, Sakai, DOCEBO, and others. Mostly Moodle is the widely used LMS. Open source Moodle system that features all capabilities for asynchronous online learning. It is also well scalable with capacity for several millions of students in various courses (J. Carmona & Elizondo, 2012). Most LMSs have a relational database format in which student activity and usage information is registered (e.g., number of logins, number of posts, times reading, quiz grades). As a rule, these variables are captured automatically and periodically. Conditioned to an educational model, this trace of the students' actions in the platform generates a huge amount of data. However, in cases with a considerably high number of students, the information contained in the relational database makes it complicated for the tutor to manage it. Several wasted database records are returned by standard queries. More importantly, a huge amount of information is logged by the LMS on student activity, but it is not possible this way to retrieve useful information to improve the learning process. Data Mining techniques allow obtaining a pattern related to the use of the platform (what happens) to interpret the activity of the students of the course (why does it happen) and provide feedback for instruction (how to change it) and knowledge about Learning on the LMS (what could have happened). These have been applied to provide feedback to courseware authors about the design of the courses and/or about learner patterns (e.g., predictive models to discard students at risk of failing). Some of these techniques borrowed from the fields of statistics, social network analysis, and artificial intelligence, have been used in the e-learning problem (Zhang et al., 2024).

4.2. Student Information Systems

Particularly in the computer-assisted education area, where students experience ubiquitous interaction with technology, Student Information Systems (SIS) have become a central focus of data mining research. Such systems collect every type of data regarding the students, including the educational resources they accessed and the specific actions they performed. In particular, the wide adoption of Learning Management Systems (LMS) and its widespread

use have produced a wealth of educational data. These systems can track both the actions of students regarding learning resources and their learning process, where each step of the learning path is recorded (Willyanto Santoso, 2018). Message boards and other features used for collaborative learning have also produced much unstructured but educationally valuable data.

Analyzing and interpreting this mass of data regarding student performance and behavior is essential to understanding, improving, and personalizing student learning and development processes. While there is no unique way of analyzing educational data, data mining is a field known for its automatic and semi-automatic algorithms that analyze enormous databases and discover hidden and potentially useful patterns. For various educational problems, data mining techniques have been successfully applied to extract knowledge from educational databases. As Educational Data Mining covers the application of data mining to educational contexts, it can complement existing research in Educational Data Mining by producing tools and techniques that, in conjunction with the development of Information and Communication Technology, can provide quantitative and qualitative answers to these important questions on an institutional level. Moreover, a better understanding of the strategies and interactions of students and teachers and the organizations where they are immersed can improve learning and teaching processes and general institution tracking and monitoring.

4.3. Surveys and Assessments

This section reviews and summarizes current efforts in applying data mining methods to educational domain knowledge, focusing on ways of forming or developing students' knowledge according to specific models. With similar infrastructures, most current implementations belong to similar development frameworks. Basic data descriptions are shared, and the following steps of pre-processing and modifying data on the practical educational implementations, as well as methods used, are introduced. In particular, different scoring details regarding item answering behaviour in the log data are carefully classified. Afterwards, some selected educational implementations on domains and methodologies are more closely described to illustrate the data mining techniques. At last, there is a summary of this chapter with a short discussion on the strengths, weaknesses, and future research directions of the current implementation.

A survey is first conducted on various types of educational knowledge and log data based on the knowledge type they belong to. For each kind of knowledge, there may be various types of log data gathering it. Even with the same type of data, they can be implemented in different ways to fit different types of educational knowledge. By classifying concrete cases that deal with knowledge and log data of different types, development frameworks of current implementations are summarized based on common compatibility in infrastructure and tasks to be done (Qiao & Jiao, 2018). Further, basic data descriptions and pre-processing refinements that are similar to these frameworks must be processed with these types of log data, regardless of the methods used thereafter to gain more concrete educational knowledge.

5. Data Preprocessing Techniques

Pre-processing or data cleaning refers to the removal of errors, inconsistencies and outliers in data. Cleaning data is very important because mining data with imperfections can have dramatic consequences on the results of the analysis. The global process of transforming a raw dataset to a correct one ready for analysis is called data pre-processing (Gibert et al., 2016). Pre-processing consists of several steps such as general procedure, descriptive statistics, detect outliers, deal with outliers, graphical methods (one variable and many variables), transformations, clustering, smoothing (rough data and noise), discretization and unsupervised classification.

Error detection, which aims to identify errors in a dataset, is usually the first step in cleaning data. Accounting applications are recognized as critical in developing data cleaning capabilities. Tools like "The Progress Report" are also

used extensively in manufacturing and service sectors (Hajizadeh & Ahmadzadeh, 2014). Error detection generally describes how gridding anomaly data is overlain with deterministic models. This approach is not typically targeted at detecting data failures; rather it is used primarily for ensuring that the anomaly under analysis has an appropriate hypothesis persistence. A likely unrecognized model failure invokes co-located anomalies, so model fit is failtocast for future run states. The roles of Models, Statistical Process Control, and Domain Knowledge are studied in error detection in the oil industry. Assessment of current processes for identification and correction of errors in estimating wind turbine annual energy.

Descriptive statistics have been developed to identify quality control issues which could impact assessing the annual energy production from a wind farm. These events tend to invoke multiple turbine anomalies which are indicative of a model misfit or a sensor failure. Effective quality control measures mitigate the impact of failures and reduce the likelihood of a significant number of co-located anomalies. The limitation of the method is the cost of data collection, especially when testing many variables and many levels of a factor. Additional tests in combination with more efficient designs would reduce the need for costly data collection. Mathematical programming approaches have superior solution quality in terms of identified levels of the factors, but they do not guarantee equidistance from background. However, more test points are often suggested, which either cannot be accommodated or imply additional data collection cost.

5.1. Data Cleaning

The first stage of the survey methodology is to perform data cleaning to ensure the dataset is free from outliers and discrepancies before applying various data mining techniques. Although various learning frameworks are presently available to data scientists, data cleaning is always the first stage of any chosen solution. According to a survey by (Young Lee et al., 2021), data cleaning is the most crucial step in data preprocessing and plays an important role in determining successful predictive analytical tasks. This process entails ensuring that the dataset is devoid of incorrect or erroneous data. Tools like R, Rattle, MYSQL, data wrangling with Python, and AVAZA offer a variety of methods to clean the dataset manually. Data can also be cleaned automatically using a computer program. Data cleaning entails a slew of procedures that make the data ready for analysis, including handling duplicate records, dealing with values, transforming data types, correcting inconsistencies, clearing unintended values, standardizing values, and many more. Ensuring that the data is appropriate is the first step in any endeavor. Errors or inconsistencies can directly influence the accuracy of the remaining data analysis or mining steps. Thus, there is a growing interest in the development of an efficient and effective data-cleaning framework that can easily decompose a variety of patterns in data. The various types of errors and data-cleaning processes are widely discussed in the following component. Further, some of the most recent advancements of data cleaning approaches to assist the data processing in building a model are explored for their effectiveness, and some future research directions are suggested.

5.2. Data Transformation

Data transformations are a common procedure to accommodate violations of these assumptions and enhance data interpretability. Unfortunately, poor understanding of the purpose and potential effects of transformations can undermine the success of the exercise. Below is a brief description of the transformations most commonly discussed in statistics tests which include square root and log transformations (Osborne, 2019).

Square root transformations are used to address cardinal variables measured as counts. For example, if the number of hospital admissions for a population of individuals in a year is counted it is reasonable to expect that some individuals have had no admissions whilst others have had one or two. Some individuals, however, may have had a

higher number of admissions. In this case, the Poisson distribution specifies that the expected number of events is needed. In these cases, a square root transformation may be in order (Pek et al., 2019). The square root transformation would have the effect of compressing the scale of items with large counts thereby minimizing asymmetry. It should be noted that this transformation naturally makes it impossible to state that there were on average two admissions since the average of square root transformed values will not mathematically understand average levels of admissions unless post transformation values are square by invoking further complex mathematical structures.

Log transformations are also considered robust to violations of symmetry about means where the range spans many orders of magnitude. For example, presumably the rates of ethnic intermarriage in the United States may be expected to be more than just orientated about a united distribution. It is expected that some states would have low rates whilst states with more intermingled portions of each ancestry group would display a larger dispersion. It is often the case that the measurement instrument is ill-specified for these data given the bounded nature of the scales. It can be expected that there are bounds in the smaller ranges where very few transformations are possible. In these cases, an initial log transformation can help flatten the observed variance in the data. However, it should also be noted that this also means that it is impossible to state that there were on average 2.34 states with such rates, mathematically speaking.

5.3. Data Reduction

Data reduction is an essential technique to minimize and optimize data size. This process can improve the performance efficiency of a data mining algorithm. Data reduction is an important pre-processing procedure in DM (Data Mining) due to the enormous volume of data that is being handled today. Data reduction has to be done before proceeding to mining. It is aimed at reducing the size of the data set and the data representation, while retaining the integrity of the dataset. It can be accomplished by reducing the amount of data, attribute size, dimensionality, or granularity of the data. By minimizing the amount of data, relevant information and information patterns of the data can be more readily revealed (Hajizadeh & Ahmadzadeh, 2014).

There are several methods for data reduction, including dimensionality reduction, numerosity reduction, aggregating, and data cubes. Data cubing allows one to process data by reduction, projection, and aggregation. Only the most essential information is saved, allowing for storage savings of up to three orders of magnitude in the density of data. Aggregating is the longest stage in data summarization; while the routine means are useful in data mining, they are not like data cubes. Numerosity reduction consists of algorithms that create a model representation of the data, allowing for a reduction of the amount of input to the algorithm (Satyanarayana & Nuckowski, 2016).

Dimensionality reduction approaches can be grouped into feature selection, feature extraction, and feature transformation. Feature extraction is the process of extracting a much smaller set than input features, yielding a set of new features that captures most of the important information present in the input data. Feature selection consists of selecting a small subset of existing features while maintaining or improving the original representation. Feature transformation consists of changing the representation of the data without changing its intrinsic dimension. The goal of feature transformation is to limit the curse of dimensionality while making generalization easy in the transformed space.

6. Evaluation Metrics for Datamining Models

6.1. Accuracy

6.2. Precision and Recall

6.3. F1 Score

6.4. AUC-ROC

7. Case Studies of Datamining in Education

Since the research into data mining and education is still in its infancy, both in-depth and broad studies have been few and far between. Some oft-cited examples, along with their contributions and applications, are as follows. On the variety of data mining methods used to analyze student behavior and performance, some methods are found to be employed across a wide variety of contexts. Time constraints appear to be a major obstacle for more in-depth analyses of experimental and tutorial data, which rely on significant prior knowledge of the educational technology being used and labor-intensive code-reviewing. The 'behavioural' data mining literature is represented by two wide-ranging reviews, while relatively few reviews cover prediction models built on demographic data and learning analytics on free-text or video-based data (Adam Whitley, 2018). Some countries, particularly in North America and Northern Europe, are engaged in comprehensive government-driven initiatives to further their learning analytics agendas. Two broad systems employing predictive algorithms adopted in different contexts are covered in-depth, addressing issues of transparency, pareto optimal solutions, ethics, and bias. Court cases brought against one of these systems are noted as the first for an automated data-mining system aimed at educational accountability. For many fledgling educational data mining groups, a single case and the consideration of what stripped-down analytics are both reasonably doable and timely.

7.1. Predicting Student Performance

In recent years, educational data mining for student performance prediction has gained widespread popularity. Based on historical data, EDM techniques can be used to aid educational decision-making for an educational institution. Predicating student performance refers to the prediction of future performance of students based on various attributes. The moral of the work is to address the following major research problem: "how to analyze an educational dataset effectively to predict student performance?" Classification, regression, association rules are the most commonly used methods in EDM. The rapid growth of educational data has increased the realization of educational data mining. EDM uses exploratory data mining techniques to find visible and hidden patterns in student performance data. Many works have been done in the educational field using EDM and they consist of pattern discovery, which is the most commonly used method in EDM. Several works have also been done for student performance prediction using EDM. However, the educational dataset consists of hundreds of attributes which results in overfitting, regularization will deal with preventing overfitting. Existing works consist of a classification method, however, most of the works considered using a single method. To the best of the authors' knowledge, this is the first work in the educational field dealing with ensemble classifiers with different properties on the student detailed information's dataset.

It has been shown that classification using voting method significantly improves prediction performance. Voting helps to maximize model diversity by aggregating multi-classifier system predictions. Despite achieving high accuracies with multi-classifier systems, how to optimize this aggregation is an important challenge. As previously mentioned, there are different aggregation methods and further research may explore this issue. The majority voting aggregation method

gives equal weights to every base classifier while the weighted majority voting method set weights to classifiers based on their prediction performance. Predicting student performance in course enrollment with collaborative seasons has gained considerable attention since it can serve as an effective decision support system for advisors and institutions. In this study, the potential usage of ensemble classifiers for improved prediction of student academic performance on a course enrollment dataset is examined (Satyanarayana & Nuckowski, 2016).

7.2. Identifying At-Risk Students

In recent years, predictive systems for learning environments have been increasingly referred to as students' risk monitors, early warning systems, or drop-out prediction systems (A. Palacios et al., 2021). These systems have grown due to the increasing relevance and popularity of on-line education platforms. These platforms collect and store vast amounts of information regarding student actions. Educational Data Mining (EDM) intends to take advantage of this information to gain insight into the behavior of students and, most importantly, to devise and implement solutions for the problems identified thereafter. Solutions to address these problems include the recognition of at-risk students earlier and more reliably, so that educational institutions can intervene and provide support to change their behavior and increase the chances of achieving course completion.

Dropping out of a learning program can have many ramifications that are detrimental to both the educational institutions that provide the programs and to the students that drop out. High levels of left courses have cost implications for institutions. Consequently, such institutions have a mutual interest in devising and implementing solutions to improve retention rates. Conversely, students dropping out from a learning program can be looked at as a failure. Students experience financial loss, wasted effort, and time. Failing to complete a chosen learning program may have broader implications in a student's lifelong learning development. Such ramifications entail the need for the provision of new and better student retention solutions for educational institutions.

In general, EDM aims to exploit the experience derived from the past to improve the future. All actions taken in these learning programs leave a trace within the technological interfaces of the platforms. This is data that can be mined by appropriate data mining techniques to increase the understanding of the relevant behaviors of the learning processes and, afterwards, can be used to develop predictive models that can help to anticipate future actions and provide educational institutions with an opportunity for intervention before dropping out occurs. Nonparametric methods, such as Genetic Programming, are ideal for data mining since they can analyze data independent of previous assumptions regarding the underlying model of the data or problem.

7.3. Enhancing Curriculum Design

The unstructured design of the curriculum only provides a general idea of the relationship between content items, which may be sufficient for specialists or teachers. In order to provide a more meaningful basis to stakeholders at all levels, the exploration of curriculum content in a more systematic and detailed manner is extremely important (Pahl, 2004). By using other fields' theories of knowledge structure, ontology and taxonomy, a curriculum knowledge description model can be provided in finer granularity. The former focuses on the meaning of knowledge, i.e. the subject's conceptual level or deeper semantic structure; while the latter focuses more on the surface structure of knowledge, i.e. knowledge realization level for a certain subject at a relatively independent level of a discipline (Rehman, 2015). Mixed structure in examining knowledge can provide a more systematic understanding of contents with own characteristics. The intention of this content enhancement is to assist in discovering two educational perspectives upon teaching and learning, known as the pedagogic and conceptual lens. In order to assist in the design of assessment strategy

based on content understanding, instructional designers are attracted to educational affordance while investigating the theory of evaluation expects comprehensively.

Curriculum is regarded as a collection of content, not knowledge. It is difficult to establish the interrelations between knowledge items on the knowledge granularity of subject content. The taxonomy-based only content expansion prevents deeper exploration of the interrelations, thus missing teaching perspective knowledge items regarding group and sequence. Researchers have begun to utilise Network Theory to examine a specific relationship, but only on the relationship level rather than the effects or reasons. A structured curriculum can then be used under unclear knowledge flow and without considering assessment strategy. The objective of this direction is to conduct an exploration of knowledge interrelations on a wider range of queries and its educational understanding. In order to proceed to a more systematic enhancement of the unstructured curriculum with respect to its pedagogic understanding.

8. Challenges in Educational Datamining

8.1. Lost in Data

A poorly organized educational system can lead to errors, confusion, and even disenchantment for students. The educational system in Afghanistan has been incapable of organizing this vast territory and providing guidance and insights to students wishing to select the best route for university entrance exam preparation. Public universities offer several majors, and the initial placement of students in those majors is determined by a national entrance exam. There is a huge amount of data about the exam. More than fifteen years of experience conducting the exam has produced huge amounts of data that have yet to be mined. Today, the examination is simply undergoing; what are the facts and figures for the current year? It is hard to believe that such a huge amount of data could be produced without being organized and utilized for prediction. The simple question is; what majors should receive which students in the next entrance exam? Clustering algorithms can easily and efficiently achieve such facts and others using an arbitrary, random numbering of students. More sophisticated methods are required to achieve predictions based on previous entrants or any records. Educational data mining is a developing discipline, which means that work on developing methodologies is still active and evolving.

This study examines the prospects and challenges of applying educational data mining (EDM) to educational systems in Afghanistan. As in many other countries with developing or poor education systems, the meeting of the complexities and challenges of educational bigger data in such countries can lead to the development of new theories about and methods for organizing, managing, and properly utilizing big data to prevent students from falling behind, leaving, or failing at the previous estimation stages and direction (Moscoso-Zea et al., 2019). There are lots of records, raw data, and big non-time-series data that can be learned from and about whether recommendations can be made to educational institutions to improve the situation. Well-designed questions implying more complex targeting than just the number of subjects may lead to the development of new methods and theories for educational big data. A challenge is to identify whether this kind of applicability is achievable with the available number of records.

8.2. Data Privacy Concerns

The increasing role of data mining applications in education has prompted concerns about data privacy. These concerns stem from two aspects of privacy involving data mining models: the methods used by data mining researchers to handle academic data and the means adopted by educational institutions to collect and disseminate sensitive academic data. A few works on privacy-preserving educational data mining systematically analyze privacy concerns. However, most of these works do not address rigorous privacy definitions. Starting from a rigorous formulation of privacy, a general

privacy-preserving data mining method is proposed to protect differential privacy. Then, the open challenges of incorporating privacy-promising algorithms into the data mining pipeline are explored.

Educational data mining findings, including patterns, outliers, and trends, may identify students and teachers if they are not protected by privacy. Therefore, analyzing data that may contain personal information raises privacy concerns. Data-mining applications in education have prompted concerns about privacy. These concerns stem from two aspects of privacy involving data mining models: the methods used by data mining researchers to handle academic data and the means adopted by educational institutions to collect and disseminate sensitive academic data. Two types of disclosures of sensitive academic data or mining models are distinguished. The first is the reciprocal data mining. Educational institutions disclose their data in exchange for data analysis or transparent mining models. In this case, a "middleman" may disclose sensitive student data to attack institutions. For instance, intelligent tutoring systems are able to analyze course scores, wrong question explanations, exhibition time, and tutoring scores to generate feedback on personalized learning, which prompts institutions to disclose educational data. Therefore, trends of data mining publications prompt organizations to consider revealing. However, if sensitive data are leaked, it is impossible to trace the source educational institution or operator. The second is the one-way data mining. In this case, education institutions may disseminate information about students—both by releasing original academic data and by abstracting sensitive information. For instance, if the whereabouts of a student are disclosed, a teacher may infer that she/he has skipped class. Such data breaches have been common in some countries, especially where the education system is not strictly supervised, and prompt educational institutions to pay close attention to the privacy concerns involved in data mining applications in education.

8.3. Data Quality Issues

Every data mining method is practically an approximation of the desired output, firstly because local phenomena are monitored using measurements that can only be approximated themselves, no a better grid and/or coverage of the monitored data limits the search to some spatial area and other sampling problems. These data quality issues are either in the data fitness domain (completeness, correctness, and consistency) or affected by the data structure (dimensionality and time issues) (Rahman Sherzad, 2017). Almost all the methods examined in this study might be affected by these issues, though not all of the issues might impact all the methods the same way.

Many of the data quality issues are partially resolvable either through educated guess work or by statistical techniques that minimize risks and chances of errors occurring. Hence visualizations where possible should be put to contextual data, results, and boundaries to better explain either presumed observations or predicted consequences. When these visualizations are given to experts familiar with the domain at question, many obvious problems become visible. Educational databases that are used by high schools or universities should be commonly integrated. Software might be tried to automatically gather the data from the original compiled databases. A higher uniformity with respect to the data quality is more preferable since it is much harder to compare dimensionalities and time issues when a diversified educational database is used (Jaber Almalki, 2021). On the contrary, no educational institution is likely to use the same assessment methods, so as the nature of the data, their representation varies, resulting in different number of candidates that are tested (spatial dimension), properties of those testing times (temporal dimension), and distributions within the decomposed test categories.

8.3. Scalability of Methods

Data mining is still a nascent discipline, there are still a great deal of key arguments remain unaddressed. As it can be seen from the literature review, the treating of scalability is rare. Very few address the arguments directly. A comprehensive understanding of scalability in educational datamining is limited. The deluge of educational data arises from the rapid adoption of Learning Management Systems, educational games, and Intelligent Tutoring Systems. These data may be used to improve educational practice, enhance learning experience. Educational data mining is concerned with developing and using methods for exploring the unique types of data from educational settings, with the aim of better understanding students, and the educational systems which are designed to serve them. Scalability can be divided into two distinct related issues. Target data size will increase as educational datamining applications are deployed. As a result, people remain uncertain of the feasibility of the existing models when applied against very large datasets. Scalability can be considered an accepted feature of a modeling algorithm or approach if there exists a high level of confidence that it can work in practice, given a certain implementation. Scalability is a comprehensive term, and may refer in a range of subjects, from algorithmic approaches down to the actual system implementation (Zhang, 2018).

Most papers neglected to consider the potential scalability challenges, if really large amounts of data are anticipated, when designing their methodologies or experiments. System implementation detail of generalisation algorithm is explicitly compared across six systems. Several methods are compared at multiple data sizes, while additional methods are attempted but provided no results. It is not only necessary to assess scalability, but also to understand the cost, benefit, and assumptions, in order to facilitate further development. It is possible that two very different methods can yield similar growth rates, but understanding their performance principles is key to developing scalable implementations. Peer-reviewers are much more appreciative of works which are also comprehensive, and quality papers which is presented not only methods but also issues in order to stimulate meaningful interactions from the community are often in demand. Another consideration is the proper presentation of benchmarking results; deceptively simple diagrams are frequently used, without demonstration of their construction, with a danger of being misleading.

9. Future Directions in Educational Datamining

The rise of big data in education has raised important issues in research, technology, and policy. This article discusses use scenarios of big data, analyzing risks and concerns on three levels: technical and ethical concerns raised by big data technologies, policy frameworks national and international for big data in education, and what the future holds for big data in education research (Zhang, 2018). Current technologies for collecting educational big data include applications, social software, and sensor data sources. New big data analytic methods, such as machine learning and decision trees, provide wealth analytical tools for educational big data.

New policies and frameworks for the collection and use of educational big data by policy makers are surveyed. At both technical and political levels, there are exciting possibilities, but at the same time, dangerous risks exist. The importance of improving educational data literacy for teachers, researchers, and policy makers is addressed. To gain an edge in the emerging big data revolution, educational researchers need to learn powerful new analytic methods, think critically about issues raised by technology and policy, and pursue new opportunities in education and neighboring disciplines. Educational data mining is a field concerned with developing methods for exploring the trace data from computers and the internet. This popularization has primarily taken the form of surveys of specific EDM approaches and reviews of substantive analyses (Moscoso-Zea et al., 2019).

To complement these descriptive studies, the present circumstances within and around EDM are reviewed systematically. Specifically: (1) what have been and are the main sources of input data for EDM studies? (2) what

measures have been used to characterize this input data? (3) what analyses and data mining methods have educational data miners applied to data? And (4) what outputs have these analyses produced? Finally, it is discussed what data sources, measures, and analyses have been scarcely considered in the past, and some ideas for novel data mining approaches to enrich the EDM field in the near future are discussed.

9.1. Integration of AI and Machine Learning

To explore the social media, education, and large-scale Data Mining communities, and to reveal the impact/interest of an author/topic on the social media, education, or general audience, a systematic mapping study in the field of Data Mining methods/systems in education is reported. It is totally different from a survey. The task of a survey is to summarize work that has already been done. The task of a systematic mapping study is to summarize the relevant research effort in terms of the quantity, status, type, or other categorization of the primary studies (Zhang, 2018). A comprehensive list of 662 articles included in the study from a perfect search strategy is reported. A complete document retrieval will allow researchers to better explore specific topics of interest over time. There is an urgent need of systematic analysis of Data Mining methods in Education. Data Mining methods, Educational data Mining, General Data Mining methods, Educational settings, Data Mining systems, and Data Mining in MOOCs are coding schemes in the classification scheme. There are two checklists: one is used in the coding phase, and the other is used to ensure satisfactory quality level of the process. The classification scheme, checklists, and relevant coding results are available for researchers. Mining in LU or higher junior college with less than 500 frequent courses, basic system, and comprehensive Data Mining functions is preferred.

This study further improves the primary study selection technique in multi-database environment identified previously, and proposes a comprehensive set of research suggestions on Data Mining methods in Education. It recognizes ten important & valuable educational settings for future research and professional development. Similarly, high priority Data Mining systems based on a major popular Education-related forum are identified. Data Mining methods in MOOCs are initially reviewed and analyzed. This study will stimulate research and activities in these valuable directions. A large volume of work has been done for exploring educational data and improving learning and teaching behaviors using various Data Mining tools Meth/Tech. In most reviews, what have been done is summed up. Before answering ‘what has been done’, some background knowledge should be provided to indicate the readers ‘why has been done?’.

9.2. Real-Time Data Analysis

Researchers investigated the efficacy of real-time data collection in a system that contains information on both students and the wide variety of resources, with the goal of offering a recommendation system that groups collaborative resources and students to enhance learning. It was critical to determine whether the resources were being utilized and whether the groups were functioning, which would inform the proposed updating system. Simulation scenarios were designed for 20 different values of parameters such as BC, ES, and AS, and different rates of application. Changes in certain metrics would indicate that resources were not used or groups were malfunctioning. Since alternative models had to be evaluated, some simplifications were made. A simulation environment and models were created to implement these parameters and make predictions about the effects of application (Adam Whitley, 2018).

During the modelling process, it was decided to present a simplified version of the collaborative system with a smaller numbers of potential resources and an OSN to observe the connections. Nevertheless, some indicators of overall system performance such as final achievement levels of students, which is related to resource utilization, learning

outcomes, and abodes based on resources affecting students were calculated. The proposed models were implemented in the simulation environment, which is capable of simulating all collaborative learning and community principles, and the desired metrics were derived for different values of application rates.

It was found that the collaborative learning process would fall into different modes based on the competition and redundancy coefficients. With regard to students, some groups would last in states with low equity, high deviation, and high activity, while the completion. When checking the final connection strengths to see whether all students had collaborated with the resources, it was found that some resources would not be used at all. There were two main outcomes based on the simulation experiments using the same set of parameters. Some scenarios showed a "gradual growth and extinction" learning mode. Data were dynamically created, updated, and improved by posting the connection strength and metric values to a Broker Service.

9.3. Personalized Learning Experiences

In recent years, advances in information technology and artificial intelligence have paved the way for novel web-based education systems. The popularity of existing systems stems from the flexibility they bring to education, making it available to a much larger (and previously unreachable) demographic. These systems offer an alternative to conventional classroom teaching, where students attend classes taught by a single teacher and rely on a static material database. Students can learn at their own pace in an adaptable environment where learning comes from one's initiative. Students can take their learning into their own hands, moving past the fixed one-size-fits-all teaching method. However, although this methodology is much more beneficial, it is still largely ineffective with existing systems.

Existing web-based education systems take the course material online without making proper use of personalization. They just put the learning material into the internet for students to learn through the web, precisely like books. Students have no interaction other than looking at the slides and infographics and reading notes. And there is mainly a single mechanism for all students to communicate and access teaching material. This system is passive, whereas a proper teaching system must be active. Teaching should adapt to the learner's face unknown. Teaching and training depend on several parameters, including the learner, teaching strategies, and infrastructure. These dynamic variables are uncertain and can be partially unknown (Tekin & van der Schaar, 2014).

To properly comprehend the course and get the required training, each student must be educated through personalized teaching and training depending on their unique context, which include different assumptions about students' background knowledge, learning rate, expectation from the course, preference over teaching mechanisms, educational objectives, measuring strategies, and infrastructure to learn, such as the means of access and time of availability. However, whether a student should take extra teaching and additional examinations or skip to the training period depends on many parameters that must be learned for each student context. For each particular student context, the best teaching and training strategy must be searched. Teaching and training can be interactive: solving problems, quizzing students in different ways, and observing their performance. Students' answers/scores are feeding back to identify and understand the context better; teaching/training are adapted online based on students' scores/feedback (Y. Ayyoub & S. Al-Kadi, 2024).

10. Ethical Considerations in Datamining

From the creation of their first artificial intelligence in 1956, humans have wondered whether or not they should share their agency with these outside entities. Therefore, it is unsurprising that with the vast amounts of student data scholars now have at their disposal, similar discussions regarding the ethics of these data mining methods have arisen.

With the accessibility of new data mining methods, it is important to conduct research on how the methods work, the assumptions made and which types of educational data mining questions can be answered. However, it is just as important to consider if there are ethical concerns surrounding a specific type of data or scrutiny of students. This is especially true when considering new disciplines and data not typically used in educational research such as social media data, text mining, predictive modeling, or if and how students should be informed that their data is being scrutinized. Essentially there exists three layers of ethical scrutiny. The first encompasses ethical concerns surrounding the use of data, privacy, limits of data, etc. This is the backdrop against which discussion and scrutiny take place. Most of these ethical considerations are derived from pre-existing theories originating from the field of philosophy addressing the morality of human action. These theories can be used as lenses to observe educational data mining. As educational data mining has become a multidisciplinary field, it would be wise to keep both educational practices and available data in mind. The second layer encompasses more practical aspects of educational data mining within the ethical backdrop mentioned in the first layer. For instance, computer science may hold the more sophisticated computational skills, but prudence requires participation of the education science disciplines. Thus educational data mining becomes an interdisciplinary field striving to combine the technological and statistical knowledge of computer science with educational insights and pedagogy from education science. Be aware that interdisciplinary collaboration may create hidden agendas as theorists and practitioners view the same matters from different angles. This has implications for gaining the trust of educationalists and e-learning platform providers, and the responsible use of student data.

10.1. Informed Consent

Informed consent has been a hot topic in recent years in data mining and data privacy communities. Informed consent-related papers have focused on how to improve consent processes, in terms of both signing disclosures, ensuring that consent is thoughtful, and ongoing consent monitoring. This range of consent issues has culminated in two relevant papers. One examines consent as a two-sided negotiation process and describes a system that uses the challenges of ensuring understanding in one-sided disclosures as inspiration for improving disclosures and consent. This work is very relevant for designing alerts that communicate risks to privacy and data uses. The second paper is the first comprehensive framework to describe consent process activities, with examples from many domains (Hutton & Henderson, 2017). With an eye toward how those activities can be adapted or reused, a wide-ranging catalogue of techniques to better acquire informed consent, whose effectiveness has been evaluated using user assessment, is reviewed. Informed consent processes can broadly be classified as secured or sustained. Secured consent consists of consent gated by a single form at the beginning of the data collection process that is not revisited. In contrast, sustained consent consists of a continual reacquisition of consent over the period that the data are collected or used that can prompt the disclosure to be viewed as a new interaction. The scholarly debate on the merits of the two approaches has ranged widely in the literature. Sustained consent obviates the need to structure consent as a snapshot of expectations across time or ask users to think about what those expectations may be over a longer timescale. It has been argued that sustaining consent may ultimately provide better-informed consent, since gaining consent after the individual has had experience with a particular service or research study may be a better way for study subjects to make informed decisions about participation.

10.2. Bias in Algorithms

Algorithmic bias is an under-explored topic in educational data mining, despite being a major issue. Research has shown that education is one of the fields in which machine learning models are likely to exhibit algorithmic bias, as access to education and quality of education differ among demographic groups. Broadly speaking, algorithmic bias can

be defined as situations when the performance metrics for a machine learning model substantially differ across demographic groups of a population ((Švábenský et al., 2024)). As evident in the example of predicted grades of students in the Philippines, despite overall good performance of the model, specific demographic subgroups can be treated unfairly. However, this unfairness can be mitigated, so that every demographic subgroup can be treated equally fair.

Algorithms deployed in a higher education context can have a major impact on the future of a student. Therefore, it is particularly important to understand whether and how the algorithms we deploy might perpetuate or amplify existing inequalities. In most educational data mining (EDM) studies that predict student success, behavioral data from a learning management system (LMS) is used as input to a machine learning model to predict a numerical target variable of student success. This target variable is often discretized, resulting in models that classify students into success class labels. While this approach has shown a lot of promise, it is important to study how fairly students from different demographic groups are treated by these models, especially in contexts where these predictions are used to make educational decisions ((Maï Cock et al., 2022)).

11. Conclusion

When it comes about knowledge, the World is pretty different today as compared to the past. Three effects of this, operating together, are changing completely the nature of Knowledge, the Knowledge Process, and the Knowledge Institutions (Rahman Sherzad, 2017). The first is the introduction of Computer Based Technologies. As these technologies create the ability to capture, store and manipulate an exponentially increasing amount of information, the rise of what for the last couple of decades has been called the “Information Society” has been produced. More generally, it could be said that the World has entered the “Digital Age”. The second is the introduction of (theoretical and empirical) modeling approaches of complex social phenomena and societal organizations. The unfoldment of these approaches has been made possible thanks to the speed-power-up of computers and the rapid growth of databases concerning social realities – Networks, Dynamic Systems, Complex Systems, Learning Organizations, Sociocybernetics, and, more recently, the Web of People and Things in a Post-Internet Scenario, are some of the constructs becoming popular in this context (Moscoso-Zea et al., 2019). Finally, the third is the emergence of a radically new organization of social existences. It is related to the rapid development of Collective Smart Units.

An intermingled understanding of these three complementary and interrelated phenomena could be fruitful in making sense of the new (unforeseen) things that are happening in the Knowledge Domain, and a direction of wants and actions in confronting and coping with the challenges they entail. In this respect, an attempt has been made here by modeling the Knowledge Domain as a complex co-evolving system characterizing Contemporary Knowledge Societies. Thus, a bottom-up and emergent dynamics could be understood shaping the (sometimes, unexpected) behaviors and properties of socio-technical systems operating in the Knowledge domain (K-domain) of Society. Moreover, factors contributing to shape the emergent dynamics of the K-domain complex system as a whole have also been analyzed, as well as their impact on all aspects of societal Knowledge. Finally, it has been discussed what could be changed and how, through well-directed wants and actions, to move towards a more desirable K-domain world.

Organization could be considered a Ke-organization for its understanding, design and rich fully directed actions – MC K-Organization, including action-competence, wisdom knowledge, mindfulness cultural, knowledge development, and social actions – could be considered an activity-directed action-organization, on purpose to differently act on the K-domain activities. Based on the above analysis and modeling self-concept, the modeling, knowledge and other directed mature actions could also be addressed more simply. It has also been argued how collective outcomes and impacts of directed K-education programs, actions and projects for developing directed individual actions (to learn how to K-learn,

K-organize, K-know, K-innovate etc.) could be more Smart, participative, jointly funded, creativity and so K-benefit oriented more than possible (at least as produce) by the parallel individual actions (of smart individuals).

References:

Moscoso-Zea, O., Saa, P., & Luján-Mora, S. (2019). Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining.

Satyanarayana, A. & Nuckowski, M. (2016). Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance.

Kumar Yadav, S. & Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification.

Hajizadeh, N. & Ahmadzadeh, M. (2014). Analysis of factors that affect the students academic performance - Data Mining Approach.

Johnson, J., Johnson, G., & Cavanagh, R. (2012). Rough sets for mining educational data.

Rudresh Shirwaikar, P. & Rajadhyax, N. (2012). Data Mining on Educational Domain.

Adam Whitley, L. (2018). EDUCATIONAL DATA MINING AND ITS USES TO PREDICT THE MOST PROSPEROUS LEARNING ENVIRONMENT.

J. Carmona, C. & Elizondo, D. (2012). Analysing the Moodle e-learning platform through subgroup discovery algorithms based on evolutionary fuzzy systems.

Zhang, X., CS Lee, V., Xu, D., Chen, J., & S. Obaidat, M. (2024). An Effective Learning Management System for Revealing Student Performance Attributes.

Willyanto Santoso, L. (2018). Early Warning System for Academic using Data Mining.

Qiao, X. & Jiao, H. (2018). Data Mining Techniques in Analyzing Process Data: A Didactic. ncbi.nlm.nih.gov

Gibert, K., Sanchez-Marre, M., & Izquierdo Sebastián, J. (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining.

Young Lee, G., Alzamil, L., Doskenov, B., & Termehchy, A. (2021). A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance.

Osborne, J. (2019). Notes on the use of data transformations.

Pek, J., Wong, O., & C. Wong, A. (2019). Data Transformations for Inference with Linear Regression: Clarifications and Recommendations.

A. Palacios, C., A. Reyes-Suárez, J., A. Bearzotti, L., Leiva, V., & Marchant, C. (2021). Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile. ncbi.nlm.nih.gov

Pahl, C. (2004). Data mining technology for the evaluation of learning content interaction.

Rehman, M. (2015). Role of Educational Data Mining Model in university course selection by the students..

Rahman Sherzad, A. (2017). Applicability of Educational Data Mining in Afghanistan: Opportunities and Challenges.

Jaber Almalki, A. (2021). Accuracy analysis of Educational Data Mining using Feature Selection Algorithm.

Zhang, J. (2018). Exploring and Evaluating the Scalability and Efficiency of Apache Spark using Educational Datasets.

Tekin, C. & van der Schaar, M. (2014). eTutor: Online Learning for Personalized Education.

Y. Ayyoub, H. & S. Al-Kadi, O. (2024). Learning Style Identification Using Semi-Supervised Self-Taught Labeling.

Hutton, L. & Henderson, T. (2017). Beyond the EULA: Improving consent for data mining.

Švábenský, V., Verger, M., Mercedes T. Rodrigo, M., James G. Monterozo, C., S. Baker, R., Zenon Nicanor Lerias Saavedra, M., Lallé, S., & Shimada, A. (2024). Evaluating Algorithmic Bias in Models for Predicting Academic Performance of Filipino Students.

Maï Cock, J., Bilal, M., Davis, R., Marras, M., & Käser, T. (2022). Protected Attributes Tell Us Who, Behavior Tells Us How: A Comparison of Demographic and Behavioral Oversampling for Fair Student Success Modeling.