CHAPTER - 2

LITERATURE REVIEW

2.1 Introduction

Education is described as the act, process, or application of mental discipline or a process of developing moral character. Education is believed to influence or shape a person's social behavior. Education is a continuous activity usually used to suggest a happy attitude. Intellectual growth, skill acquisition, and attitude formation are all cumulative processes that shape our diverse worldviews and propensities for behavior in general.

Educational data mining is one of the most crucial methods for successfully analyzing large educational data. As an interdisciplinary study area, EDM analyses educational data using machine learning (ML), statistics, data mining, educational psychology, cognitive psychology, and other theories and approaches, assisting people in efficiently resolving various educational issues.

Educational information is gathered from various sources, including interactions between teachers and students, usage of educational technology, administrative data, demographic information, and student affectivity (such as motivation and emotional states). An enormous amount of potential data from numerous sources, in various formats, with varying degrees of granularity, or in various levels of useful hierarchy, can be stored in educational contexts. Preprocessing is required since simply collecting and integrating all of this raw data for mining is difficult.

Most conventional data mining methods, such as visualization, Classification, clustering, and association analysis, have been used effectively in education. This chapter explains the importance of educational data mining, machine learning techniques, sentiment analysis, cluster techniques, association rule mining and recommender systems in education.

2.2 Survey on Educational Data Mining

Educational data mining is an emerging research area that applies data mining techniques to educational data to reveal new information on student behaviors and learning strategies and enhance teaching and learning practices. The major objectives of EDM are to comprehend the student's learning process, anticipate the student's learning outcomes, offer a better understanding of education-related phenomena [49], and assist educational institutions in comprehending and improving their educational processes. Academic institutions are increasingly focused on enhancing their instruction methods, students' academic performance, and learning environments. EDM discusses methods for comprehending learning processes and finding patterns in data to assist academic institutions in making decisions. This section explains some different techniques used in educational data mining.

Predicting student performance determines the grade a student will earn before registering for a course or completing an exam. Zhang et al., [50] thoroughly analyses how machine learning and data mining can be used to predict student success. The five stages of this review are data collection, problem formalization, model development, prediction, and application.

A machine learning-based framework was proposed by Czibula et al. [51], and it helps improve the performance of data mining activities. It comprises components for feature analysis, as well as unsupervised and supervised learning-based mining. An EDM case study using real-time data assesses and analyses the framework's efficacy.

A web-based software solution for student profiling is presented by Prada et al. [52], aiding instructing professionals without a background in data science. The analysis and forecasting of students' performance, as measured by their observable test results and degree completion, are the main objectives of the tool that is being provided.

Rahman et al. [53] suggest an educational data mining system to support learning to program using unsupervised methods. The following activities are part of the framework: data collection, preprocessing, clustering, feature extraction, data pattern and association rule mining with a recommendation.

Students' academic performance is influenced by various factors, including their behavioral, academic, and personal characteristics. Predicting students' academic success at an Indian technical institution is the topic of the study [54]. A dataset was gathered by employing a survey with a questionnaire and the academic division of the selected university. The given dataset has undergone data preprocessing and factor analysis to remove anomalies, reduce the dimensionality, and find the most correlated feature.

Romero et al. [55] held that EDM, which may be referred to as the use of data mining technology, is a method designed to study special data kinds from the educational environment with the aim of better understanding how students learn. Using a clustering method, Moises et al. [56] identified six distinct student groups, examined the communication patterns of each type and discovered that these interactions would recur early in the course.

Karthikeyan et al. [57] created a hybrid educational data mining technique to analyze student performance and improve students' educational experience. Finally, to analyze student academic performance data, Crivei et al. [58] investigated applying unsupervised machine learning techniques, specifically principal component analysis and relational association rule mining.

Delgado et al. [59] suggest a novel self-organizing mapping-based unsupervised clustering method results in accurate and diverse user clustering following student behavior data. Okoye et al. [60] presented an educational process and data mining combined with a machine learning model to

contextually examine the teachers' performance and provide recommendations based on data collected from students' evaluations of instruction. Kumar [61] created the Multi-Tier Student Performance Assessment Model using single and ensemble classifiers. This methodology collects and analyses student data from higher education institutions based on important variables that substantially impact students' performance and outcomes.

Duhayyim [62] developed an enhanced evolutionary algorithm-based feature subsets election with neuro-fuzzy Classification for data mining in education. Notably, this is the first comprehensive analysis of EDM studies that solely considers classroom instruction and emphasizes the time component.

Yağcı [63] suggests a new model based on machine learning techniques to forecast undergraduate students' final test marks using the results of their midterm exam as the primary data. Various machine learning algorithms were determined and compared to forecast the students' final test marks. To examine the learning behaviors in the data produced by students in a blended learning course, EDM approaches are used by Sokkhey et al. [64] to provide an early-stage prediction for intervention and enhance student performance based on a created academic performance prediction system. The technology offers quicker and more comfortable ways for users to obtain real-time information on students to improve academic outcomes by anticipating student performance levels and learning patterns early.

Educational Data Mining aims to develop a detailed understanding of each student to provide personalized learning tailored to their specific needs. This approach considers a wide range of factors affecting learning, including cognitive abilities, emotional states, psychological traits, and social contexts, to overcome barriers to success and boost motivation and engagement. Despite progress in predicting academic performance and identifying at-risk students, there's a notable gap in fully integrating diverse factors into

comprehensive student models. This gap highlights the need for more inclusive and integrative strategies in education that address the complexity of learning beyond traditional metrics.

2.3 Machine learning techniques for the education

The education sector is greatly impacted by machine learning, which uses new intelligent technology to forecast the future makeup of the educational environment. Improving teaching abilities with tools created with machine learning and artificial intelligence may be beneficial. It has the potential to operate without teachers' involvement and assist them. These applications are mostly used in the following:

- Intelligent teaching tools are flexible learning tools that can interact with students, respond to their questions, and provide feedback.
- In terms of learning materials, learning sequence, and student understanding of various topics, adaptive tutoring systems can be tailored to meet the needs of each student. Also, it helps students with special needs recognize facial emotions.
- Automatic methods are very effective in determining the student's level of comprehension because they can change the difficulty of subsequent questions based on answers provided in the past.
- Machine learning algorithms could carry out the common tasks of recording attendance, assessing assignments, and creating questions.

Some machine learning advantages in the field of education:

- *Predict Student Performance:* A fundamental feature of machine learning is its capacity to estimate student achievement. The system "learns" about each student, allowing it to pinpoint shortcomings and recommend useful study aids like more practice exams for each learner.

- *Properly Grading Students:* Machine learning can also fairly grade students by eliminating human bias.
- *Effective Content Organization:* Machine learning can better content organization by recognizing its flaws. For instance, students move on to the next after learning one skill, continuously expanding their knowledge.
- *Recommended learning path:* After the software evaluates the student's performance, it may recommend a more effective route to learn new content. It begins with an examination of the curriculum's current body of knowledge. Students also receive recommendations for materials and additional learning techniques after identifying their weak points.
- Student abilities, interests, and preferences can all be tracked by machine learning algorithms. It examines the behaviors and emotions of students. The analysis helps forecast the student's potential strengths in terms of interests.
- *Students and teachers should be grouped* following their needs and availability as another way that machine learning will enhance education.

Machine learning is now frequently employed in fields including detecting at-risk students, forecasting students' final exam results, and early failure detection. A prerequisite for instructors to conduct extra research to support their performance is identifying at-risk students.

The enhanced conditional generative adversarial network-based deep support vector machine approach was suggested by Chui et al. [65]. Researchers attempted to predict student performance under-supported learning using their devised method. With enhanced CGAN, they produced new training data and stressed the significance of producing new data for the model.

Adnan et al. [66] provide a predictive model that evaluates the issues experienced by at-risk students, making it easier for teachers to prompt intervention to persuade students to raise their study engagements and

enhance their study performance. Several machine learning and deep learning (DL) techniques are used to train and test the predictive model, which uses student study factors to characterize students' learning behavior.

Behr et al. [67] use random forests based on conditional inference trees and a large German data set containing various aspects of student life and study courses to predict university dropout. By sequentially modelling students' journey from high school (pre-study) through the study-choice phase (decision phase) and into their first semesters of university, the author focuses on the very early identification of learner dropout.

Pek et al. [68] present the performance outcomes discovered utilizing machine learning techniques to recognize at-risk students and reduce student dropout. The major goal of this research is to use the ensemble stacking method to build a hybrid model and use that model to identify at-risk students. Bujang et al. [69] thoroughly review machine learning strategies to forecast the final student grades in the first semester courses by enhancing the performance of prediction accuracy. A multiclass prediction model is also suggested to minimize overfit and misclassification.

Ko et al. [70] employ supervised and unsupervised machine learning approaches to identify the key components of an effective learner in a computer course. Seven supervised ML algorithms and ensembles are used to examine the accuracy, precision, and sensitivity of classifier performance. The association rule and clustering are also used to identify the crucial characteristics of effective students. Rafique et al. [71] create a system that can forecast student performance and assist teachers in introducing remedial interventions on time to improve the performance of underperforming learners. Additionally, it investigates the potential of collaborative learning as an intervention that may be used in conjunction with the prediction system to raise student performance. A visualization system is also created to accompany these improvements to track and monitor student performance

across groups, the entire class, and individual students. This system aids teachers in regrouping students based on performance. Finally, students' performance is predicted using several well-known machine learning models.

A framework for analyzing students' learning abilities was established by Lin et al. [72] by integrating machine learning-based techniques into the examination of students' academic achievement and utilizing rank model learning-based approaches. Identifying student weaknesses can help teachers raise test scores and improve learning. Alsariera et al. [73] employed the most recent ML algorithms and variables to forecast student academic achievement. The study's findings revealed a recent significant increase in studies in this area. However, predicting students' performance is substantially influenced by academic variables, internal assessments, demographics, and family/personal traits.

Machine Learning (ML) in education is pushing the boundaries of personalized learning but faces scalability challenges and a need for models that are both interpretable and fair. Key research areas include integrating ML with pedagogical theories, providing real-time feedback, and conducting longitudinal studies to understand long-term effects. Ethical use of data and equipping teachers with ML tools are vital for better educational outcomes. Moreover, applying ML in varied settings like special education offers opportunities for broadening its impact. Addressing these challenges requires interdisciplinary collaboration to make ML in education more effective and inclusive.

2.4 Survey on Sentiment Analysis in Education

Sentiment analysis combines computational linguistics, natural language processing, and content interpretation that examine the viewpoint and identifies and extracts emotional polarity from the material. In general, the polarity of opinion is either positive (confident and exuberant), negative (confused, raucous, and enraged), or neutral. Sentiment analysis is used in

domains like e-commerce, financial market, customer relationship, social media and education. This section examines sentiment analysis in the context of education.

The teaching-learning process can be improved with the help of students' feedback, which is a powerful technique. A methodology for automatically analyzing the opinions of students given in reviews is proposed by Kastrati et al. [74]. The technique, which is based on aspect-level sentiment analysis, tries to automatically detect sentiment or opinion polarity expressed towards a specific MOOC-related element.

Tao et al. [75] created an ensemble technique for modelling student behavior utilizing features of engagement, semantics, and sentiment/stress collected from a MOOC discussion forum dataset. The goal was to examine the impact of stress on academic achievement with a second study comparing cohorts before and after COVID. The multi-attention fusion model is proposed by Zhai et al. in [76] for aspect-based sentiment analysis tasks. Using global attention, this approach considers how local attention may affect contextual representation. To get better classification results, it manages the weight of the fusion of local and global attention through the gating layer. This model offers a new approach to teaching assessment in the educational field.

An overview of sentiment analysis methods for education was presented by Han et al. in [77]. For multimodal fusions, the authors of this study presented a framework for sentiment discovery and analysis. It presents all elements of sentiment analysis of educational data methodically, concentrating just on textual data. It is crucial to consider student feedback and affective states to adjust and enhance educational content, especially when learning occurs in Intelligent Learning Environments. The educational content of the learning environment can be improved with the help of this feedback, which can also assist teachers in understanding student behavior.

The technique of building a corpus for the educational domain in the area of programming languages is described by Estrada et al. in [78]. With two new datasets of educational opinions tagged with learning-centered emotions, this work's key innovation is a comparison of various classifiers, including one that uses an evolutionary strategy, in an educational setting.

The Greek language sentiment analysis system described by Spatiotis et al. in [79] analyses texts written in Greek and produces feature vectors that, combined with classification algorithms, allow us to categorize Greek texts based on the opinions and levels of satisfaction expressed. The Greek school network's hybrid educational systems, which provide courses for lifelong learning, now include the sentiment analysis module. The module provides a wide range of opportunities for lecturers, decision-makers, and educational institutions that take part in the training process and provide life-long learning courses to comprehend how their learners view learning activities and define which components of the learning activities they enjoy and which they do not.

Asghar et al. [80] develop a fuzzy-based sentiment analysis system to analyze student feedback and satisfaction. This method assigns appropriate sentiment ratings to opinion terms and polarity shifters in the input reviews. This technique first calculates the sentiment score of student feedback reviews to examine and quantify student satisfaction at the fine-grained level.

He et al. [81] suggest a meta-based self-training approach with a meta-weightier. Properly choosing symbolic representations and efficient learning control in a neural system can lead to a generalizable model. As a result, it develops a teacher model that produces domain-specific knowledge and a student model that uses the generated pseudo-labels for supervised learning. Then, the meta-weightier of this is jointly trained with the student model to give each instance sub-task-specific weights, coordinating their convergence rates, balancing class labels, and mitigating the noise effects introduced by self-training.

Nikoli et al. [82] used sentence-level analysis to extract one or more sentence-level components and categorize them as having a positive or negative attitude. The authors combined SVM, cascade classifier, and rule-based approaches for aspect extraction. The sentiment was also identified by combining an SVM component with a dictionary-based method.

An aspect-based sentiment analysis approach using a machine learning algorithm was proposed by Melba Rosalind and Suguna [83] to determine student satisfaction with Coursera's online courses. First, unsupervised and semi-supervised LDA approaches obtained certain elements from the student reviews. Next, sentences from the real reviews were separated, and the aspects and their sentiment polarity were evaluated. Next, the sentiment polarity of the sentences was determined using a customized vocabulary; it produced positive polarity for learning elements. Finally, classifying several attributes and attitudes was taught and evaluated using the maximum entropy classifier.

To evaluate the course materials, OSMANOLU et al. [84] studied student feedback from a university. The authors categorized the materials using machine learning techniques into positive, negative, or neutral attitudes. They subsequently used the negative input to improve their semester's course materials. After preprocessing the student comments, six machine-learning classifiers were employed.

A hybrid framework based on the random forest, logistic regression, and SVM models was proposed by Kaur et al. in [85] Using the lexicon based Senti wordnet approach, the authors classified student remarks as either favourable or negative. The hybrid classifier was trained for the labelled dataset, and performance against the SVM model at various test-training ratios was assessed. The hybrid classifier performed better in every Classification metric than the SVM model.

Sangeetha and Prabha [86] presented a multi-head attention fusion model to analyze student comments' sentiment. Using word and context embeddings like Glove and Cove, the feedback's input sequences of sentences are analyzed concurrently across the multi-head attention layer. The LSTM deep learning model receives the outputs of the two multi-head attention layers and the embeddings. To increase the model's accuracy, the dropouts of these layers are controlled. Comparing the suggested fusion model to individual multi-head attention and LSTM, it was more accurate at classifying the positive, negative, and neutral sentiment orientations.

Sentiment analysis for closed-ended questionnaires, which primarily generate quantitative data like Likert scale responses, presents unique challenges not found in textual analysis due to the absence of direct sentiment expression. Key research areas include merging quantitative data with qualitative insights to reveal sentiments, setting sentiment benchmarks for numerical responses, and creating methods for detecting emotions from such data. Moreover, the need to account for sentiment interpretation's cultural and linguistic variability in closed-ended formats underscores the importance of developing accurate models and tools. Advancing in these areas would improve sentiment analysis's effectiveness in surveys, providing deeper understanding of respondent attitudes and emotions.

2.5 Data preprocessing techniques

Data preprocessing is essential in machine learning since the learning capacity of the models is directly impacted by the quality of the data and the underlying knowledge collected during the preprocessing stage. The preprocessing stage in contemporary automated procedures typically consists of a straightforward modification that serves primarily to put the dataset into an acceptable format for the algorithm rather than to increase the performance of the models [87]. Data preprocessing encompasses a range of operations like cleaning, encoding, scaling, and dimensionality reduction. Data cleaning is the

process of correcting data errors. Missing values, incorrect datatypes, and repetitive rows are common data problems. Data cleaning aims to remove noise, inconsistencies, and missing values from the data. This section reviews some data preprocessing techniques.

Due to noise and missing values, real-world data is inconsistent. The dataset has missing information for various reasons, including storage space limitations, disputes over data uploading, compromised input devices, and occasionally security concerns. The missing variables negatively impact the performance and dependability of the machine learning or deep learning models [88]. Consequently, before using learning models, missing data must be resolved. Many strategies can be used to fill in these missing values, such as computing attribute means and likelihood or occasionally ignoring data rows. The missing values should be filled in to ensure that the information is consistent and noise-free.

Raja et al. [89] concentrate on applying unsupervised machine learning methods to handle missing values. A novel way to handle the missing data is created by combining soft computation approaches with clustering techniques. This method aids in the resolution of consistency issues. Finally, a deep learning approach for missing value reconstruction in multidimensional time-series data is presented by Bansal et al. in [90]. It combines coarse- and fine-grained patterns along a time series and patterns from related series across categorical dimensions using a neural network. The parameters and their training are carefully crafted to generalize across various missing block positions and data properties.

When using categorical data for ML models, encoding is a necessary preprocessing step that can be done in various ways [91]. The choice of approach can significantly impact the effectiveness of a model. For example, the technique of assigning a numerical number to each possible value of a category attribute is known as label encoding. While comparing the abilities

of various classifiers, Duan [92] finds that one-hot encoding produces a good result in developing a neural network and outperforms other ML techniques for encoding qualitative information.

Row-wise for reducing data samples and column-wise for reducing data variables are the two usual directions for data reduction. Row-wise data reduction is possible using various data sampling methods, including random and stratified sampling. While choosing a data sample, random sampling is frequently used to replicate a random process. Contrarily, stratified sampling is used to preserve the ratios of data samples that correspond to various categories [93]. The three basic approaches to column-wise data variable reduction are as follows. The first is to choose variables of interest directly using domain knowledge. The second is to choose significant variables for additional study using statistical feature selection techniques. Finally, the third step entails implementing feature extraction techniques to create valuable features for data analysis.

Dimensionality reduction reduces the number of dimensions in a high-dimensional data representation [94]. Various dimensionality reduction techniques have become widespread in various applications due to the enormous increase in high-dimensional data. Also, contemporary methods are always evolving. A high-dimensional dataset is transformed into a low-dimensional dataset using dimensionality reduction techniques, preserving as much of the data's original meaning as is feasible. The dimensionality-curse issue is partly solved by representing the original data in a low-dimensional manner. Low-dimensional data is very simple to process, visualize, and analyze [95]

The two main categories of dimensionality-reduction strategies, feature selection and feature extraction, can be used to categorize and compare dimensionality-reduction methodologies [96]. Finding the subset of characteristics that best defines the data through feature selection is a

technique for reducing the dimensionality of the data. It chooses the characteristics from the original data that are essential and pertinent to the ML goal and eliminates the pointless and redundant features. Furthermore, it helps locate a useful subset of traits pertinent to the current task [97].

The main objective of feature selection is to provide a subset of features that accurately captures the essential characteristics of the entire input while remaining as small as is practical [98]. Feature selection lessens the amount of data needed, requires less storage, increases the accuracy of predictions, prevents overfitting, and speeds up training and execution. Subset Creation and Subset Evaluation are the two stages of the feature selection method. Subset Evaluations determine whether the final subset is optimal after Subset Creation asks us to create a subset from the input dataset [99].

Data preprocessing, essential for enhancing data quality and analysis efficiency, faces several research challenges. Key areas needing advancement include developing sophisticated automated cleaning techniques for large or complex datasets, improving methods for handling high-dimensional data and missing values to preserve data integrity, and devising real-time preprocessing for streaming data. Moreover, there's a need for more adaptive feature selection and extraction methods, scalability and efficiency improvements, and privacy-preserving techniques. Integrating preprocessing seamlessly with machine learning pipelines, ensuring cross-domain adaptability, and establishing standardized evaluation metrics for preprocessing effectiveness are also critical. Addressing these gaps promises to significantly advance preprocessing methods, thereby boosting the quality and reliability of data analysis and machine learning across various domains.

2.6 Clustering technique for the educational research

A data mining approach called clustering analysis divides data objects into smaller groups. These subsets or clusters are used to group objects to be comparable inside a cluster while being different from objects in other

clusters. This section examines how educational data, including social-economic, demographic, higher education access average, and academic scores, are used in clustering algorithms to identify obstacles to academic success and forecast students' academic performance.

Marbouti et al. [100] used cluster analysis to examine trends in student success based on demographic data and academic achievement. First, students were split into two groups based on how they were admitted to getting better results (i.e., freshman or transfer). Then, according to the ideal number of clusters determined by the Elbow approach, bisecting k-means was used to group the students into several groups.

Delgado et al. [59] introduce a unique unsupervised clustering method based on the Self-Organizing Map (SOM) artificial neural network model to assess student behavior. SOM performs user clustering that is accurate and comprehensive based on student data. Findings show a direct correlation between user participation and better performance, with some clusters being linked to the university's average intake profile.

To study the behaviors of various university students, Chang et al. [101] offered K-Means and clustering by quick search and discovery of density peaks. The behavioral traits of a particular group of students in academic performance and personal conduct may be reflected in clustering centers. Therefore, this algorithm may specify the ideal clustering center and the proper k value directly. In other words, the algorithm could calculate the number of different student conduct categories and behavior scores for each university.

Zhang et al. [102] suggested an enhanced version of the k-means clustering algorithm for student big data portraiture. The preliminary clustering process is carried out using the canopy method, and the outcomes are given to k-means as a reference for the K value. By eliminating the issue

of arbitrarily choosing the initial centers to fall into the local optimum, the max-min distance algorithm chooses K samples with reasonably large distances as the starting centers of k-means. With the help of this algorithm, college administrators may identify students who exhibit various behavioral traits and support their decision-making accordingly.

McBroom et al. [103] describe a unique divisive hierarchical clustering method that prioritizes the discovery of behavioral patterns by using time information in its objective function. The resulting clusters have a hierarchy of clusters determined by decision rules on features, resembling a decision tree in structure. Ahmed et al. [104] describe a case study in which student groups were created using evaluation information from two online platforms. The personalization mechanisms take clustering information into account. The k-means clustering algorithm is used to group skills.

Almasri et al. [105] suggest developing a brand-new supervised cluster-based classifier model using a unified framework. Using the clustering approach, the unified framework groups historical student records into a collection of homogeneous groupings. The final unified classifiers and the centroids at each cluster are utilized as the basis for the cluster-based classifier model, which is then constructed.

According to their success in academic activities, Chaves et al. [106] analyze students' profiles while considering two separate evaluation systems: knowledge-based assessment and work-based assessment. A clustering strategy and supervised feature selection were used to analyze the student profiles. According to the findings, using both evaluation methods, two student profiles may be recognized based on how well they performed in the course. These two profiles match students who pass and fail the course, respectively. The analysis's results also show that some features are unnecessary or useless.

The two-step clustering technique is introduced, and the clustering model's algorithm flow is discussed by Chu et al [107]. Then, it examines the mental health evaluation tool test's clustering algorithm, builds a data mining model, mines the database, examines the state features of various college students' mental health, and offers pertinent remedies. Finally, the clustering analysis algorithm is used to cluster the data to fulfil the requirements of the emotional management system based on the analytical clustering technique.

Guo et al. [108] explore emotional education in the framework of college physical education. This work suggests enhanced approaches for text features and classification algorithms utilizing the original structure model based on emotional feature clustering. The emotion recognition method calculates the two points with the least similarity as the beginning cluster center before each clustering. The two classes with the shortest distance between them are joined after clustering, in which case the distance between each class is determined. The emotion recognition method in emotion feature clustering puts the objective of information transfer at the core of emotional education.

K-means were used by Hassan et al. [109] to categorize students. A hybrid model is presented that integrates regression-based prediction with the k-means clustering phase based on student similarity. Both static and spatiotemporal features of the students were clustered. The use of student behavior evaluation and study models based on clustering technology was introduced by Li et al. [110] It employs the application study of student behavior analysis and research model based on clustering technology, analyses student behavior using clustering technology, and fairly assesses the viability of campus data mining and the k-means method. Students are divided into groups using the cluster analysis algorithm following their features, and each group is subsequently subjected to data analysis and mining for data association rules.

Li et al. [111] offer an unsupervised ensemble clustering approach to find patterns in student behavior using behavioral data. To find behavioral patterns, this framework first extracts behavior features from the two perspectives of statistics and entropy before combining density-based spatial application clustering with noise and k-means algorithms. Priyambada et al. [112] developed a profile-based cluster evolution analysis for dynamic data that occasionally arise. This strategy was tested using educational data, where the profiles corresponded to the alignment of the course sequence and the semester grade point average of the students. As a result of the effort, stakeholders can identify students who will graduate on time or drop out by looking at the learning behavior patterns of the students. Furthermore, by observing how students moved between clusters each semester, students altered their learning habits. Finally,Chi et al. [113] use the K-Means approach in data mining to perform cluster analysis on the sample data set based on the final grade data of students majoring in software and information services at a certain university.

Clustering algorithms in education grapple with challenges unique to educational data, which is diverse in nature, including both numerical and textual information. These algorithms need optimization to handle this heterogeneity and must be dynamic enough to adapt to the evolving nature of learner data. A critical challenge is scaling personalized recommendations for large student bodies without losing accuracy or computational efficiency. Additionally, there's a need to make these algorithms more interpretable and to align them with educational theories for more effective teaching and learning strategies. The lack of standardized metrics for evaluating their impact in educational contexts, combined with the need to address data anomalies, bias, privacy, and resource constraints, underscores the importance of targeted research. Addressing these research gaps could significantly improve hybrid clustering algorithms' application in education, leading to enhanced personalized learning experiences and outcomes.

2.7 Survey on Association Rule mining

Association rule mining is a data mining technique which discovers the association between things. This section explains how association rule mining and classification-based association rule mining are used in the education research field.

From the data on college students' lives and learning, association rule mining algorithms can identify possible connections between various characteristics of college students. Teachers can then use this information to identify the problems and personal strengths of various students and tailor their instruction. A modified script-based associative rule extraction approach is recommended by Lei [114] to determine the relationship between college students' aptitude and quality. Analysis and research are done on the college students' quality assessment data.

Jia et al. [115] suggest a multi-resource mining method based on association rules to gather data on learners' activity in MOOCs and create a repository of instructional resources for them. The association rules of educational resources are created with a view towards the attribute set of information association positioning. To locate and mine a variety of MOOC teaching-associated resources, the association rules are also combined with the shortest path scheduling scheme of teaching resources.

For college students, Liu et al. [116] developed a mental health assessment system and used data mining technologies in the psychological system. The association rule decision tree algorithm is used in this work to create and optimize the management system for consultations on mental health education.

Student behavioral data from association rule mining is crucial to smart campus data analysis. It can improve student management research and increase data analysis techniques for smart campus construction. Wang et al.

[42] concentrate on using association rule mining to study student behavioral data. Campus managers will have a solid foundation on which to base their decisions due to association rule mining, which can intuitively reflect the relationship between students' behavioral characteristics and further evaluate the student management knowledge in the data.

Applying association rule mining algorithms can reveal valuable hidden information for raising academic performance. To investigate necessary subjects' influence on dependent subjects' academic outcomes, Das et al. [117] introduce association rule mining tools. The study clearly illustrates the major impact of preparatory topics on the academic result of dependent subjects students.

The suitable option and assessment of learning outcomes significantly impact students' academic progress. Therefore, the results of the students' tests serve as key benchmarks for their academic progress. The correlation between student exam performance and learning outcomes was examined by Salahli et al. in [118]. A strategy utilizing data mining techniques and fuzzy logic is taken into consideration. Following an analysis of the weights assigned to the learning outcomes in the questions, the Apriori method was used to determine the correlations between the outcomes and scores. The most frequent things identified as the outcomes of the Apriori method were then used to generate fuzzy inference rules.

Shatnawi et al. [119] suggest using association rule mining to assist advisers and students in deciding and ranking courses. By identifying relationships across courses, association rules enable students to choose courses based on their performance in earlier courses. The association rules mining process is applied to thousands of student records to discover relationships between courses that students have registered for in numerous prior semesters. The system has successfully generated a list of association rules that direct a specific student in choosing their courses.

Rajab [120] suggests Active Pruning Rules (APR) as a new approach for improving classifier performance, particularly prediction accuracy and decreasing rule redundancy. It preserves the rules in the classifier with actual data coverage and discards any rule whose data frequency has dropped below the permissible minimum frequency.Segatori et al. [121] propose an efficient distributed fuzzy associative categorization technique based on the MapReduce paradigm. The method uses a unique distributed discretizer based on fuzzy entropy to efficiently construct fuzzy partitions of the characteristics. Then, a list of potential fuzzy association rules is created using a distributed fuzzy variant of the well-known FP-Growth method.

Kumi et al. [122] use URL and webpage content attributes to detect malicious URLs using CBA. The CBA algorithm employs a training dataset of URLs as historical data to develop an accurate classifier to uncover association rules. Using an evolving memetic algorithm based on the random walk algorithm, Siddique Ibrahim et al. [123] propose an efficient weighted decision support system for diagnosing cardiac disease. Pal et al. [124] present heuristic approaches for efficient rule development and selection in Associative Classification. It first uses the database to reduce the search space and then explicitly investigates the most powerful class association rules from the optimized database. This combines rule generation and classifier construction to shorten the total classifier development cycle.

Mattiev et al. [125] describe a new strategy for reducing the number of class association rules produced by classical class association rule classifiers while retaining an accurate classification model comparable to that produced by state-of-the-art classification algorithms. To be more specific, CMAC is a new associative classifier that employs agglomerative hierarchical clustering as a post-processing step to minimize the number of rules.Finally, song and Lee [126] explore the predictive ability of candidate rules and develop a novel AC mining approach dubbed predictability-based collective class association

rule (PCAR) to build a high-predictability classifier by appropriately applying cross-validation and aggregating the final rules. According to theoretical analysis and testing, their PCAR technique keeps candidate rules with high predictability while simultaneously removing superfluous, redundant, and low predictable rules.

Alwidian et al. [127] introduce the concept of weight and propose WCBA, a new weighted classification algorithm based on association rules. In their WCBA approach, domain field experts define attribute weight based on its importance and allocate it to the associated attributes. They also classify breast cancer data using the WCBA method.

Associative Classification (AC) Rule Mining, blending association rule mining with classification, faces challenges despite its potential across sectors like retail and healthcare. Key research gaps include developing scalable algorithms for large, high-dimensional datasets, improving rule quality to prevent overfitting, and enhancing rule interpretability for users. Additionally, adapting AC rule mining to dynamic or streaming data, integrating it with other machine learning techniques, and expanding its application to multi-label and multi-class classification are areas needing exploration. Addressing bias and ensuring fairness in AC models, along with making these techniques adaptable across various domains without domain-specific adjustments, are critical for broadening their applicability. Tackling these challenges could make AC rule mining more efficient, accurate, and user-friendly, significantly extending its impact and utility in diverse applications.

2.8 Survey on Recommender System in Education

The recommender system uses information filtering, data mining, and prediction algorithms to assist users in making decisions. Regarding recommendation systems in education, the usefulness of suggested educational resources will enhance students' learning. A sizable number of recommendation systems have been developed in teaching, academic advisory

services, and education [128]. The recommended components in the educational domain include educational resources, teaching objects, articles, institutions, course information, student performance, and the subject of study. The target users within the educational domain are learners, trainers, and academic advisers.

A unique neural network approach for session-based thread recommendation is proposed by Zhang et al. in [129]. This approach uses the self-attention mechanism to understand the relationships between threads and makes recommendations to students based on the threads they have already viewed in the current session. Based on a learner's prior performance and learning history, Mondal et al. [130] suggest using machine learning to suggest appropriate courses to the learner. The framework uses the k-means clustering algorithm to first categorize students based on their prior performance. To recommend a few acceptable courses, collaborative filtering will be used in the cluster.

A personalized recommendation system architecture that can handle educational resources was created by Cheng et al. in [131]. A hybrid recommendation system based on content and collaborative filtering was created to satisfy the requirements for suggestion originality and address the cold start issue of the recommendation system. Using machine learning techniques, Dhar et al. [132] predict each student's best academic route based on their prior academic achievements and suggest the most appropriate academic programme for their higher studies. The optimal model is chosen for each educational programme based on each student's performance using multiple ML algorithms to forecast their academic achievement. As a result, the recommendation process considers the best-selected model and related aspects to offer the most appropriate academic path for fulfilling each student's professional goals.

An improved recommendation technique is presented by Chen et al. in [133] that employs learning resource adaptation by mining learners' behavioral data. Based on their preferred online learning methods, it groups learners into clusters. It uses collaborative filtering and association rule mining to extract each cluster's preferences and behavioral patterns and creates a unique, variable-size set of recommendations. A personalized online learning platform based on a collaborative filtering algorithm for proposing online education courses was created by Li et al. [134]. The outcomes of the course recommendation are more following the users' interests, which significantly increases the recommendation's efficacy and accuracy.

Using collaborative filtering and ontology, Agbonifo et al. [135] created an ontology-based personalized recommender system that is required to suggest appropriate learning content to learners. First, users take a pre-test to help classify them into learning categories appropriate for their ability level. Then, the learning materials are organized according to an ontology, and collaborative filtering is utilized to compile user preferences and then suggest the most popular information to users. Using various machine learning methods, Yanes et al. [136] offer a recommender system for anticipating appropriate actions based on course requirements, academic records, and course training outcomes assessments. The outcomes demonstrated that the suggested recommender system gives more suggestions for actions to enhance students' learning opportunities.

Educational recommender systems aim to tailor learning by recommending resources and courses but face significant research gaps. These include insufficient adaptation to diverse learning styles, lack of dynamic content updates based on student feedback, and inadequate alignment with long-term educational goals. Challenges also arise from the need to integrate these systems with educational theories, address emotional and motivational factors, and ensure bias and fairness in recommendations. The absence of

standardized evaluation methods and the struggle to balance scalability with personalized learning further complicate their effectiveness. Moreover, enhancing collaborative and social learning aspects remains essential. Addressing these issues could greatly improve recommender systems, making them more adaptable, equitable, and impactful for varied student needs.

2.9 Summary

In the realm of educational technology, significant strides are being made through Educational Data Mining (EDM), Machine Learning (ML), sentiment analysis, data preprocessing, clustering algorithms, Associative Classification (AC) Rule Mining, and educational recommender systems to tailor personalized learning experiences based on a comprehensive understanding of each student's unique needs. These approaches consider a multitude of factors from cognitive abilities to social contexts, aiming to enhance academic performance and engagement. However, challenges such as integrating diverse learning factors, ensuring scalability and fairness of ML models, and adapting technologies to dynamic educational settings underscore the need for more inclusive strategies and interdisciplinary collaboration. Additionally, the effective analysis of sentiment in closed-ended questionnaires and the advancement in data preprocessing techniques highlight the necessity for developing sophisticated methods that can handle the complexity and heterogeneity of educational data. Overcoming these challenges is crucial for improving the effectiveness, inclusivity, and impact of educational technologies, thereby fostering enhanced learning experiences and outcomes across diverse educational landscapes.

Table 2.1 Research gap identified.

SNo.	Research Domain	Research Gap
1.	Educational Datamining	Lack of full integration of diverse factors into comprehensive student models.
2.	Machine Learning	Integration with pedagogical theories
3.	Sentiment Analysis	Quantitative data like Likert scale responses, that present unique challenges not found in textual analysis due to the absence of direct sentiment expression.
4.	Data Preprocessing	Data quality, complex/ large dataset , data integrity and efficient data analysis.
5.	Clustering	Data Heterogeneity - Handling mixed datasets numerical and categorical data Dynamic adaption - Algorithms must adapt to the evolving nature of learner data to stay relevant and effective.
6.	Association Rule Mining	Rule quality - Improving rule generation to prevent overfitting and ensure the relevance and generalizability of the rules.
7.	Recommender System	Integration with educational theories to enhance teaching and learning strategies. Emotional and motivational factors, crucial for enhancing engagement and motivation in learning, are often disregarded in the design of recommender systems.

Based on the research gaps identified, with Bloom's educational theory as base, this research aims to develop a recommender system to predict the academic performance according to the influence of six different affective attributes. The research also focuses on developing a sentiment analysis evaluation method for closed ended questionnaire, a data preprocessing method that handle missing values, a clustering algorithm that handles heterogenous dataset and an association rule mining algorithm that results high quality and relevant dataset. The result of the proposed algorithms is integrated to develop a recommender system to predict the students' academic success.