www.ijecs.in

International Journal Of Engineering And Computer Science
Volume 15 Issue 01 January 2026, Page No. 28138-28155
ISSN: 2319-7242 DOI: 10.18535/ijecs/v15i01.5387

Discovery of Learners' Academic Patterns from a Multi-dimensional Educational Dataset: A Case Study of the Punjab School Education Board, India

Dr Karan Sukhija, Dr Shabina Sukhija

Assistant Professor SSET, JGND PSOU, Patiala Punjab India

Abstract:

Educational Data Mining is an emerging trend in learning management, aimed at identifying changes within the educational system and uncovering hidden factors that influence learners' academic performance. This study utilizes six consecutive years of student examination data, comprising approximately 2 million records from the Punjab School Education Board, India, organized in a database format to identify dependencies among learner attributes. To effectively analyze students' academic performance, this paper proposes a framework combining frequent pattern mining and expert rule mining. It refines the frequent pattern mining algorithm and generates new rules to uncover significant patterns among various attributes in the dataset. Additionally, validation checks are conducted on the resulting pattern set to identify the optimal hypotheses for improving learner academic performance.

Keywords: Education Data Mining (EDM); Multidimensional Data Extraction; Customized association and classification; Frequent Pattern and Expert Rule mining framework; Visual analytics.

1. Introduction

The education landscape in developing countries has undergone rapid changes in size and diversity over the past few decades. Education professionals are now working with vast amounts of data in a complex environment, where they are tasked with identifying associations and correlations among various factors influencing academic performance (Sukhija et al., 2016). While the growth of the education system has expanded access and availability, the quality of education, reflected in success and retention rates, remains a concern. According to the Ministry of Human Resource Development's education census report, India, the dropout and failure rates in school education are alarmingly high. After the primary level, the dropout rate is 28.9%, rising to 42.4% after middle school and further escalating to 52.8% after high school. Figure 1 illustrates the enrolment and pass percentages for matriculation and senior secondary courses from 1999 to 2015, according to the Punjab School Education Board, India.

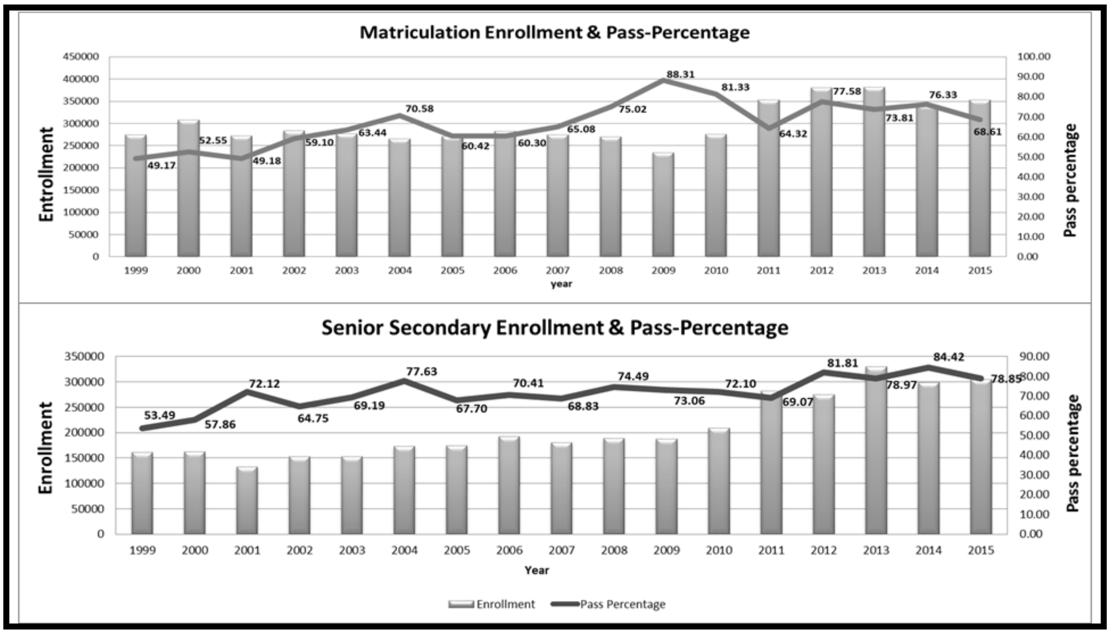


Fig. 1 Statistics on student enrolment and pass percentages for the Punjab School Education Board, India, from 1999 to 2015.

This study examines the variability in enrolment and pass percentages over different years, prompting an exploration of additional factors influencing the performance of education systems in developing countries (Fernandes et al., 2019).

Although significant work has been reported in the literature on education data mining, most existing studies have focused primarily on simplistic metric analysis (Chandra and Nandhini, 2007), often limited to individual institutions or single courses. Numerous issues still need to be identified (Wang et al., 2017) and addressed to enhance the success rates of school education systems. Furthermore, other studies highlight the potential for further research, uncovering additional contributing factors (Bennett et al., 2017) and examining the underlying causes of specific outcomes (Calvet et al., 2015).

In this study, we address these issues by proposing a novel methodology in the form of an EDM (Education Data Mining) analytical tool designed to analyze learner behavior and uncover relationships among various attributes. We utilized a dataset spanning six years (2011–2016), sourced from the Punjab School Education Board, India, comprising approximately 2 million records of Class X and XII students, including their personal, academic, and school-related details.

The proposed EDM analytical tool aims to enhance the quality of education by analyzing student performance (Angel et al., 2015) and identifying the factors influencing academic outcomes (Rosalind James, 2016). The steps involved in developing the EDM analytical tool are as follows:

1.1. Data set generation

The multidimensional dataset employed as input for the EDM analytical tool was meticulously gathered from diverse federated sources, ensuring comprehensive coverage of relevant information. To prepare the data for analysis, an extensive preprocessing phase was undertaken. This phase involved addressing and resolving anomalies to ensure data integrity (Dara et al., 2016), imputing missing values to maintain consistency, and transforming the dataset into the requisite format, optimizing it for seamless algorithm implementation and robust analytical outcomes.

1.2. Data set extraction

The input data for the EDM analytical tool is sourced from either multidimensional datasets or external repositories, ensuring a diverse and robust foundation for analysis. To elevate the effectiveness and user experience of the graphical user interface (GUI), a visual analytics feature has been seamlessly integrated. This feature enables clear and interactive graphical representations of the loaded attribute sets, allowing users to easily explore, interpret, and gain actionable insights from the data.

1.3. Customisation of algorithms

Custom association rule mining and classification algorithms were meticulously designed and implemented using the Java programming language, tailored to meet the specific needs of educational data analysis. These advanced algorithms were then applied to educational datasets to identify meaningful relationships among various attributes and to categorize them systematically. This approach enhances the predictive accuracy of course outcomes, providing valuable insights that contribute to improved decision-making and educational planning.

1.4. Expert rule incorporation

This feature enables domain experts to manually input rule sets that were not generated by the mining algorithms. The incorporated rules are then subjected to validation through various parameter checks, ensuring their relevance, accuracy, and alignment with the system's analytical framework.

The steps outlined above were followed to construct a comprehensive framework for the EDM analytical tool tailored to the school education system. This paper is organized as follows:

Section 2 presents a review of background research related to the education domain, providing context for the study.

Section 3 introduces the proposed system framework, along with the experimental setup based on the aforementioned educational dataset.

Section 4 delivers an in-depth analysis of the experimental results obtained using the proposed research methodology.

Section 5 provides an empirical analysis, emphasizing the results through visual analytics to derive optimal hypotheses for improving the education system.

Section 6 concludes the paper, summarizing key findings and their implications.

2. Background work

The exploration of the educational domain can be effectively carried out using a combination of psychometrics, statistics, and cognitive psychology, as highlighted by Winter et al. (2006). While statistics encompass all facets of data, psychometrics specifically focuses on measuring and quantifying educational outcomes. When combined with Education Data Mining (EDM), these disciplines offer novel insights into complex educational data. Jindal et al. (2013) emphasized the availability of a wide range of tools and techniques in EDM that can be applied to educational datasets. These tools address diverse objectives and empower various stakeholders, including educators and policymakers, to make informed and effective decisions based on data-driven insights.

Igor et al. (2009) contributed to this field by outlining the construction of data warehouses and demonstrating the efficient use of Extract, Transform, Load (ETL) tools for data abstraction, facilitating a deeper understanding of educational information. Similarly, Kovacic et al. (2010) explored the use of socio-demographic variables as a foundation for educational data and applied classification methods to predict student performance during initial phases of learning. Their findings affirmed the suitability of data mining techniques for accurately assessing and predicting learner outcomes. Quinlan et al. (1994) further advanced the field by introducing classification decision trees, structured hierarchically, as a powerful method for data organization and analysis.

Bidgoli et al. (2003) conducted a comprehensive evaluation of various classification techniques to predict students' examination performance, demonstrating that decision trees can be effectively transformed into sets of "if-then" rules. This transformation simplifies the interpretability of results and supports practical applications. Garcia et al. (2011) designed a collaborative EDM tool based on association rule mining, which was aimed at improving e-learning courses. This tool enabled instructors teaching similar courses to share, annotate, and utilize the discovered patterns, fostering collaboration and enhancing instructional quality. Building on these advancements, Anupama Kumar (2016) incorporated visual analytics feature into an EDM framework, significantly improving the representation of academic performance data. This enhancement supports better decision-making by offering clear and actionable insights.

Expanding on this body of work, our research introduces a constraint-based multidimensional association rule classification system, specifically applied to a non-volatile dataset. This system is designed to uncover hidden patterns in students' learning behaviours, offering deeper insights into the factors influencing academic success. The proposed framework provides a robust foundation for developing decision support systems (Donald et al., 2018) tailored to the school education sector, aiding in the formulation of strategies to enhance educational outcomes.

3. System Overview

The proposed EDM analytical framework encompasses several well-defined phases, including dataset generation, multidimensional data extraction, customization of frequent pattern mining algorithms, and the integration of expert rule sets, as depicted in Figure 2. Each phase plays a crucial role in the systematic analysis and interpretation of educational data, contributing to the overall effectiveness of the framework. The framework is structured into the following key phases:

Dataset Generation: This phase involves acquiring relevant data from federated educational sources. The process includes addressing anomalies, handling missing values, defining key working attributes, and transforming the raw data into a structured format suitable for analysis. This foundational step ensures data integrity and readiness for subsequent phases.

Multidimensional Data Extraction: In this phase, the prepared dataset is loaded into the EDM analytical tool. The dataset serves as the input for executing mining algorithms, enabling the discovery of complex patterns and relationships across multiple dimensions of educational data.

Customization of Frequent Pattern Mining: The framework customizes frequent pattern mining algorithms to uncover hidden dependencies among attributes within the dataset. By tailoring these algorithms to educational data, the framework can extract meaningful insights, providing a deeper understanding of learner behavior and performance.

Integration of Expert Rule Sets: Domain experts are empowered to incorporate their own rule sets, which may not be derived through automated mining algorithms. These expert rules are validated against the educational dataset to ensure relevance and accuracy. This phase bridges the gap between automated analysis and expert intuition, enriching the overall analytical process.

Visualization and Analytics: To enhance interpretability, the framework incorporates a visual analytics feature. This component generates intuitive graphical representations of course outcomes, enabling stakeholders to easily comprehend and analyze trends, patterns, and key findings in the educational data.

Key Benefits of the Framework:

- Enables comprehensive analysis of large-scale educational data.
- Supports customization and expert input for tailored insights.
- Provides intuitive visualizations for better decision-making.

The proposed EDM analytical tool framework is designed to empower educators, administrators, and policymakers to optimize strategies, improve learning outcomes, and drive advancements in the educational landscape.

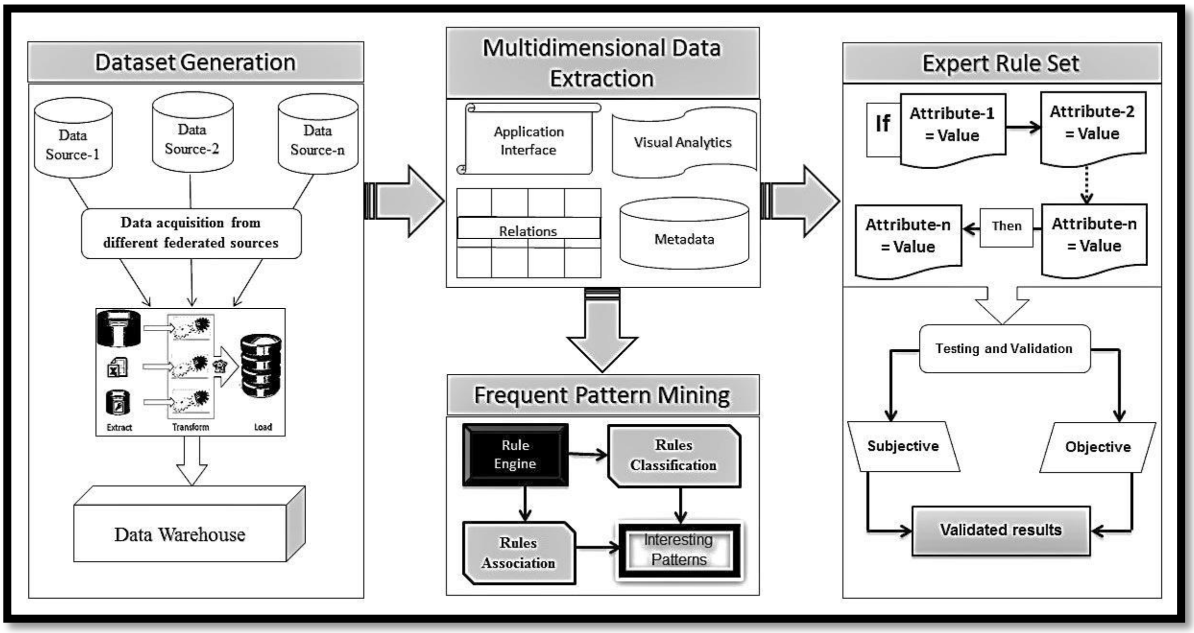


Fig. 2. Framework of proposed EDM analytical tool

The framework of the proposed EDM analytical tool, as depicted in Figure 2, outlines several distinct phases, including dataset generation, multidimensional data extraction, customization of frequent pattern mining, and incorporation of expert rule sets. Each phase is designed to execute specific tasks aimed at uncovering hidden dependencies and relationships among attribute sets. These phases work cohesively to provide a comprehensive and insightful analysis of educational data, ultimately supporting improved decision-making and strategic planning.

3.1. Data set generation

The system's structure begins with the crucial process of dataset generation. During this phase, the data acquisition methodology must be highly reliable to ensure that the authenticity and accuracy of the collected data remain unquestionable (Ihantola et al., 2015). To maintain these standards, a comprehensive student dataset, encompassing various examination and personal fields, was collected through legitimate and authorized channels with due permission from the concerned authorities.

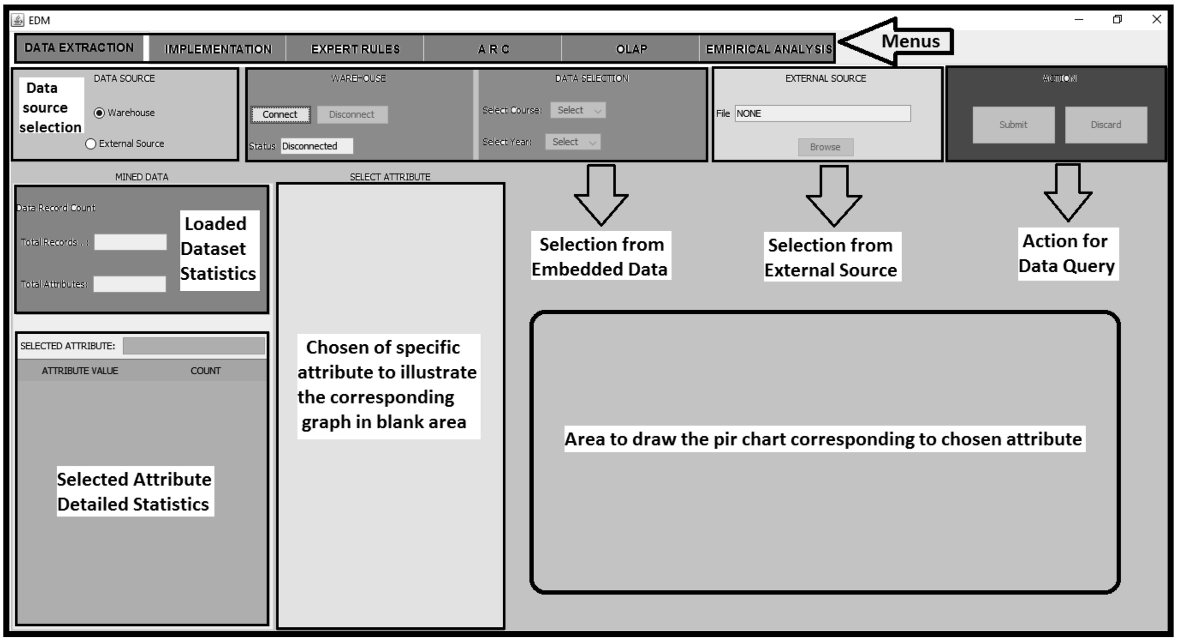
As part of this process, we successfully compiled a dataset of approximately 2 million records, covering a six-year academic timeline (2011–2016), sourced from the Punjab School Education Board, India. This dataset serves as the foundation for analysing trends and patterns within the academic records, enabling a detailed evaluation of educational progress over the specified period (Sukhija et al., 2016).

3.2. Multidimensional data extraction

Several analytical tools support the analytical process by providing static visualizations and incorporating animation features to enhance user engagement and understanding (Geryk et al., 2013). The interfaces of various standalone applications enable analysts to integrate their functionality seamlessly with traditional data mining approaches.

In the proposed EDM analytical tool, the integration of visual analytics (Wei Koutrika et al., 2014) and advanced visualization techniques (Geryk et al., 2013) has been implemented to improve its overall efficiency and usability. The tool leverages Java's Swing API to design a user-friendly windows-based interface, while the JFreeChart library is utilized to create dynamic and interactive charts for visualizing educational data. These visual components facilitate deeper exploration and analysis of the educational system, providing stakeholders with intuitive insights into the data (Geryk et al., 2014).

Figure 3, as shown below, illustrates the home window of the multidimensional data extraction module, demonstrating the tool’s capability to manage and visualize complex datasets effectively.

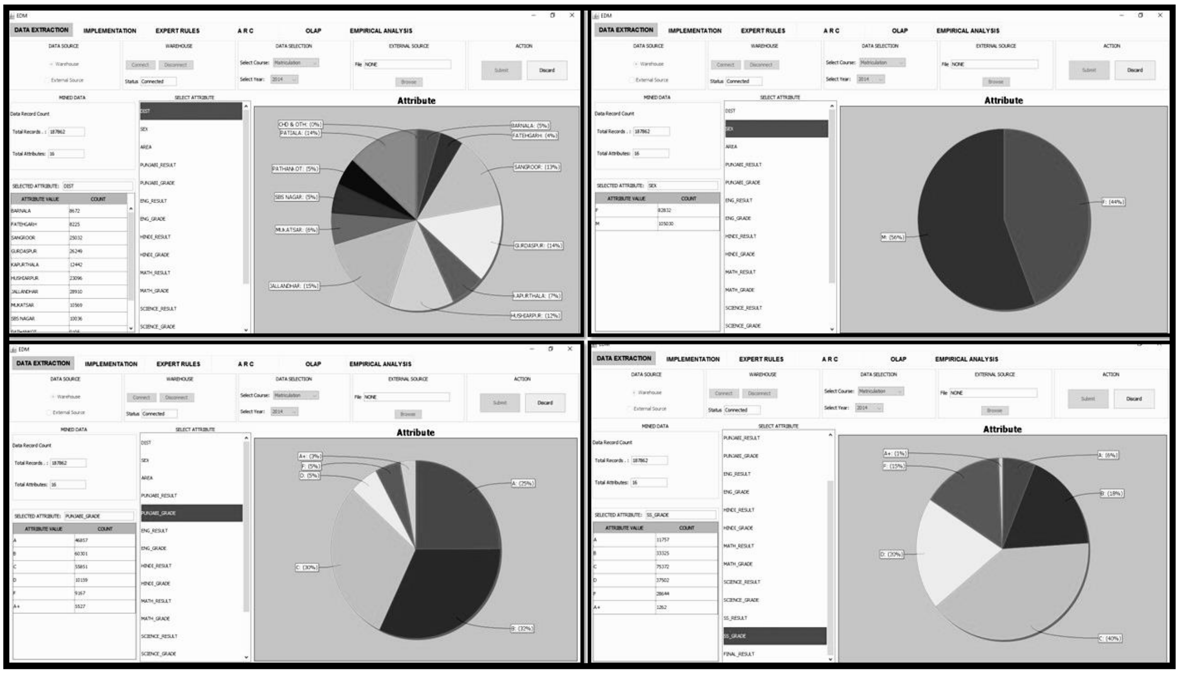


Status: Disconnected

File: NONE

Fig. 3. Multidimensional Data Extraction module home window

The Multidimensional Data Extraction module is responsible for retrieving multidimensional datasets from the embedded data warehouse based on specified constraints. This ensures that the extracted data aligns with the analytical objectives and requirements. Once extracted, the data is transformed into a transactional-based format, facilitating its compatibility with subsequent analytical processes. A filtration step is then applied to the transactional data, isolating the requisite or limited attribute sets essential for targeted analysis. The filtered dataset is subsequently loaded into the analytical tool, where all attribute sets, along with their associated metadata, are systematically described and displayed. This comprehensive view of the dataset ensures clarity and allows for detailed exploration. Figure 4 illustrates the layout and functionality of the tool, showcasing how it manages and presents the loaded datasets, enabling analysts to gain deeper insights into the data's structure and content.



Status: Connected

Select Course: Matriculation

Select Year: 2014

File: NONE

Data Record Count

Total Records . : 187862

Total Attributes: 16

SELECTED ATTRIBUTE: DIST

ATTRIBUTE VALUE	COUNT
BARNALA	8672
FATEHGARH	8225
SANGROOR	25032
GURDASPUR	26249
KAPURTHALA	12442
HUSHEARPUR	23096
JALLANDHAR	28910
MUKATSAR	10569
SBS NAGAR	10036

SELECTED ATTRIBUTE: SEX

ATTRIBUTE VALUE	COUNT
F	82832
M	105030

SELECTED ATTRIBUTE: PUNJABI_GRADE

ATTRIBUTE VALUE	COUNT
A	46857
B	60301
C	55851
D	10159
F	9167
A+	5527

SELECTED ATTRIBUTE: SS_GRADE

ATTRIBUTE VALUE	COUNT
A	11757
B	33325
C	75372
D	37502
F	28644
A+	1262

Fig. 4. Data Extraction explore the attributes set corresponding to loaded data

The figure mentioned above presents a summary of the dataset, including the total number of records and the total attributes that define each record within the dataset. This information provides an overview of the dataset's scope and complexity, serving as a foundational reference for subsequent analyses.

Furthermore, data extraction from the loaded dataset is performed based on specific attribute sets such as caste, gender (Alom et al., 2018), geographic area, result, or grade. These attributes serve as key filters, enabling the isolation of relevant data subsets for detailed analysis. This targeted extraction process facilitates a more focused exploration of patterns and relationships, ultimately contributing to a better understanding of the dataset's underlying trends and insights.

3.3. Customisation of frequent pattern mining

Association rules This section outlines the implementation of enhanced association and classification mining algorithms designed to uncover hidden dependencies among attribute sets (Garcia et al., 2011), which play a critical role in improving the education system (Anupama Kumar, 2016). While several contemporary data mining frameworks and tools can execute association and classification algorithms effectively (Wang et al., 2017), they often lack the efficiency to handle transactional-based educational data specifically.

In the proposed analytical tool, customized versions of association and classification algorithms have been developed and embedded to address this limitation. These algorithms are tailored to process transactional-based educational data with greater accuracy and efficiency. The tool's implementation is carried out using the Java programming language, ensuring robustness and scalability. The proposed enhanced frequent pattern-mining algorithm integrates association rule mining and decision tree classification, which are divided into the following subsections:

Association Rule Mining: Designed to identify and extract meaningful patterns and relationships among attributes in the dataset. These rules reveal dependencies that can inform decision-making and strategic improvements.

Decision Tree Classification: Implements a hierarchical structure to classify data attributes, enabling the prediction of outcomes and performance metrics based on the input dataset.

This dual approach not only enhances the discovery of hidden insights but also provides actionable outputs for stakeholders to optimize the educational system effectively.

3.3.1. Association rule mining

Association rules are a popular method for representing discovered knowledge and illustrating strong correlations (Garcia et al., 2011) between frequently occurring items in a database. An association of the form X implies Y signifies a close relationship between the attribute value sets, where the presence of X is strongly linked to the presence of Y. Most association rule mining algorithms are designed to identify all relationships that meet user-specified constraints and parameters (Sukhija et al., 2016). Typically, users define thresholds such as minimum and maximum support, confidence, and other criteria to guide the mining process (Garcia et al., 2011).

Support: This metric represents the probability that both X and Y occur together in a dataset. It measures the rule's prevalence within the database.

Confidence: This metric indicates the probability that Y is satisfied given that X is satisfied. It reflects the rule's predictive strength.

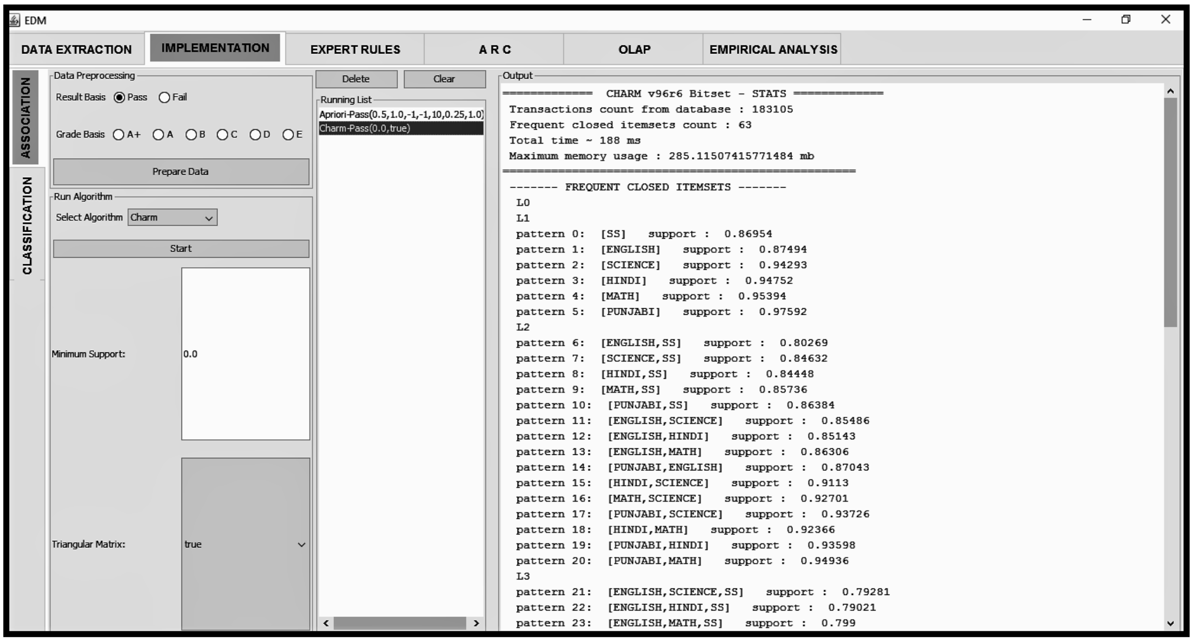
In this study, the customization and implementation of various association rule mining algorithms have been performed specifically for transactional-based educational datasets. These algorithms are tailored to work effectively with the unique attributes and structure of educational data. The proposed methods focus on uncovering meaningful patterns and correlations that can provide actionable insights to improve the education system. The following sections detail the customization and application of these association rule mining algorithms to analyze the educational dataset comprehensively.

```
$\boldsymbol{Algorithm\ ASSOCIATION}(\boldsymbol{DATA})$
$F[1] = get - frequent - 1 - itemset(DATA)$
$set\ k = 2$
$until\ F[k-1] = \emptyset\ do$
    $C[k] = generate - frequent - itemset(F[k-1])$
    $for\ each\ transaction\ t \in DATA\ do$
        $C[t] = c$
        $for\ each\ CANDIDATE\ c \in C[t]\ do$
            $count[c] = count[c] + 1$
        $done$
    $done$
    $F[k] = \{c \in C[k]\ AND\ count[c] \geq \varepsilon$
    $k = k + 1$
$done$
$return\ F[1] \cup F[2] \cup F[3] \dots\dots\dots\dots \cup F[k]$
```

```
$\boldsymbol{Algorithm\ GENERATE - FEREQUENT - ITEMSET}(\boldsymbol{F}[\boldsymbol{k-1}])$
$for\ each\ item\ I \in F[k-1]\ do$
    $for\ each\ item\ J \in F[k-1]\ do$
        $check\ if\ I[1] = J[1] AND\ I[2] = J[2] \dots\dots..AND\ I[k-1] = J[k-1]\ then$
            $c = I \times J$
        $check\ if\ HAS - INFREQUENT - SUBSET - FOR(c, F[k-1])\ then$
            $DISCARD\ c$
        $otherwise$
            $insert\ c\ into\ C[k]$
    $done$
$done$
$return\ C[k]$
```

```
$\boldsymbol{Algorithm\ HAS - INFREQUENT - SUBSET - FOR}(\boldsymbol{c}, \boldsymbol{F}[\boldsymbol{K-1}])$
$for\ each\ k-1\ subset\ s\ of\ c$
$check\ if\ s \notin F[k-1]\ then$
        $return\ \boldsymbol{true}$
$otherwise$
        $return\ \boldsymbol{false}$
```

The above-mentioned algorithm separates the frequent rule set from the infrequent rule set by eliminating the infrequent subsets. Further, the frequent item sets inspected for their reoccurrence counts in the data set. This reoccurrence count considered as threshold to get the relevant association rules. Figure 5 demonstrates the results in form of association rules returned by customised association algorithms.



Result Basis ● Pass ○ Fail

Grade Basis ○ A+ ○ A ○ B ○ C ○ D ○ E

Select Algorithm: Charm

Minimum Support: 0.0

Triangular Matrix: true

Running List:
Apriori-Pass(0.5,1.0,-1,-1,10,0.25,1.0)
Charm-Pass(0.0,true)

```
============= CHARM v96r6 Bitset - STATS =============
 Transactions count from database : 183105
 Frequent closed itemsets count : 63
 Total time ~ 188 ms
 Maximum memory usage : 285.11507415771484 mb
===================================================
 ------- FREQUENT CLOSED ITEMSETS -------
  L0
  L1
  pattern 0:  [SS]   support :  0.86954
  pattern 1:  [ENGLISH]   support :  0.87494
  pattern 2:  [SCIENCE]   support :  0.94293
  pattern 3:  [HINDI]   support :  0.94752
  pattern 4:  [MATH]   support :  0.95394
  pattern 5:  [PUNJABI]   support :  0.97592
  L2
  pattern 6:  [ENGLISH,SS]   support :  0.80269
  pattern 7:  [SCIENCE,SS]   support :  0.84632
  pattern 8:  [HINDI,SS]   support :  0.84448
  pattern 9:  [MATH,SS]   support :  0.85736
  pattern 10:  [PUNJABI,SS]   support :  0.86384
  pattern 11:  [ENGLISH,SCIENCE]   support :  0.85486
  pattern 12:  [ENGLISH,HINDI]   support :  0.85143
  pattern 13:  [ENGLISH,MATH]   support :  0.86306
  pattern 14:  [PUNJABI,ENGLISH]   support :  0.87043
  pattern 15:  [HINDI,SCIENCE]   support :  0.9113
  pattern 16:  [MATH,SCIENCE]   support :  0.92701
  pattern 17:  [PUNJABI,SCIENCE]   support :  0.93726
  pattern 18:  [HINDI,MATH]   support :  0.92366
  pattern 19:  [PUNJABI,HINDI]   support :  0.93598
  pattern 20:  [PUNJABI,MATH]   support :  0.94936
  L3
  pattern 21:  [ENGLISH,SCIENCE,SS]   support :  0.79281
  pattern 22:  [ENGLISH,HINDI,SS]   support :  0.79021
  pattern 23:  [ENGLISH,MATH,SS]   support :  0.799
```

Fig. 5. Running list of Association rule mining module

In the implementation phase of the association mining module, data preprocessing is conducted based on two primary options: **Result** and **Grade** of the learners. The Result option has two possible outcomes, namely pass or fail, while the Grade option includes possible values such as A+, A, B, C, D, and E.

Following the preprocessing phase (Sukhija et al., 2016), the dataset is filtered to isolate relevant information and subsequently transformed into a transactional format. This transformation ensures that the data is appropriately structured for the execution of association mining algorithms, enabling the discovery of meaningful patterns and relationships within the educational dataset.

The prepared data from the data preprocessing section serves as the input for the "Run Algorithm" section. This section forms the core of the association mining module and integrates five advanced association algorithms (Sukhija et al., 2015): Apriori, Apriori Close, FP-Growth, FP-Growth Close, and Charm. These algorithms have been specifically enhanced to handle transactional-based educational datasets effectively. Each algorithm is designed to uncover hidden relationships and dependencies within the data, making it possible to derive actionable insights. Additionally, the module includes configurable constraints, allowing users to specify parameter values, such as support and confidence thresholds, to guide the execution of the algorithms. The implementation of these algorithms has been carried out in the Java programming language, ensuring efficient processing and adaptability for educational data (Sukhija et al., 2016).

The execution of the selected algorithm is detailed in two key sections. The "Running List" section displays the entry of the algorithm chosen for execution, while the "Output" section presents the results of the executed algorithm (Garcia et al., 2011). Each association algorithm follows a specific execution methodology, driven by the defined constraints:

- Apriori: This algorithm relies on parameters such as minimum support, maximum support, minimum items, maximum items, number of rules, association type, minimum confidence, and maximum confidence.
- Apriori Close: The execution of this algorithm is based on minimum support, minimum items, number of rules, association type, and minimum confidence.
- FP-Growth and FP-Growth Close: Both algorithms require constraints like minimum support, minimum items, number of rules, association type, and minimum confidence.
- Charm Algorithm: This algorithm operates using only two constraints: minimum support and triangular matrix constraints.

These well-defined parameters allow each algorithm to efficiently mine patterns and relationships within the dataset, catering to the specific analytical requirements of the educational domain.

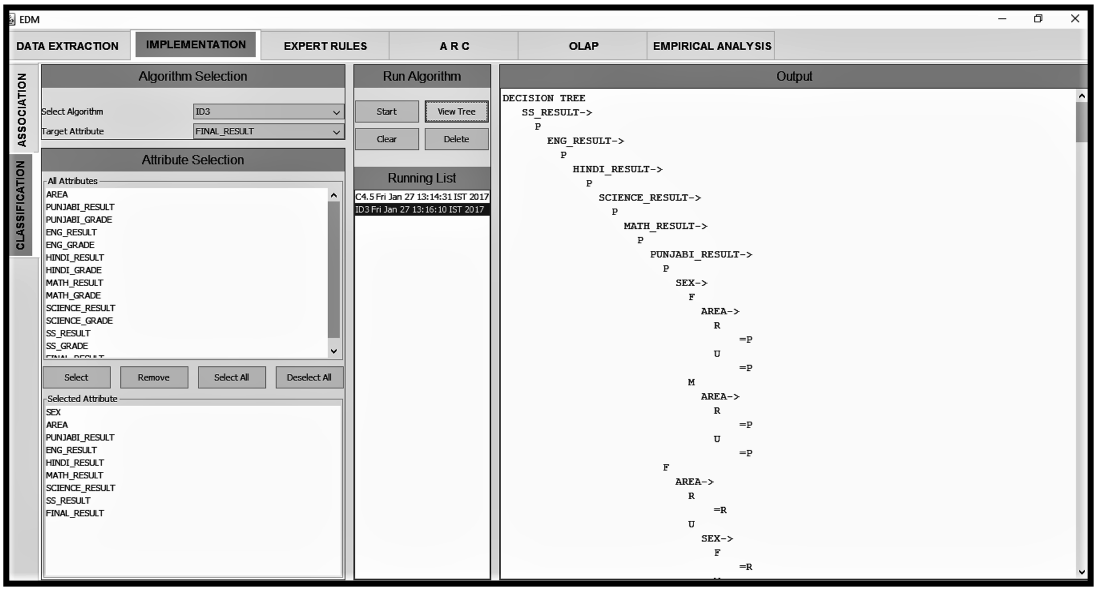
3.3.2. Decision Tree Classification

Classification analysis is a powerful technique for predicting labels and generating a set of labelled patterns through supervised classification (Sorabi et al., 2016). It is widely applied in educational data mining to identify students with similar characteristics and responses to specific features. This approach is instrumental in predicting student performance, evaluating the significance of various attributes, and uncovering common misconceptions that may impact course success (Sorabi et al., 2016). The classification algorithm described below takes association rules as input and classifies them into item sets corresponding to a specified target attribute. These target attributes are defined by the end user and are used to generate classes that align with the decision-making requirements of stakeholders. By doing so, the algorithm enables meaningful outcomes that support data-driven decisions for improving educational outcomes.

```
$\boldsymbol{Algorithm\ CLASSIFIACTION}(\boldsymbol{DATA})$
    $Let\ T\ be\ an\ empty\ Decision\ Tree \mid T.left = T.right = \emptyset$
    $check\ if\ DATA\ is\ \boldsymbol{pure}\ or\ other\ stopping\ criteria\ meet\ then$
        $return\ DATA$
    $for\ each\ attribute\ a\ \in DATA\ do$
        $calculate\ entropy\ and\ information\ gain$
    $done$
    $a_{best} = \{a \mid \max(information\ gain(a_i)\ \forall i\ , i = 1,2,3 \ldots .\}$
    $T.insert(a_{best})$
    $DATA_R = induced\ DATA\ based\ on\ a_{best}$
    $for\ each\ D\ \in DATA_R\ do$
        $T_R = \boldsymbol{CLASSIFICATION}(\boldsymbol{DATA_R})$
        $T = T\ \cup T_R$
    $done$
    $return\ T$
```

The proposed EDM analytical tool incorporates customized versions of two widely used classification algorithms, ID3 and C4.5 (Sukhija et al., 2016), specifically tailored to work with educational data (Sukhija et al., 2015). These algorithms have been implemented in the Java programming language, ensuring platform independence and making the analytical tool highly versatile and adaptable to various systems. This customization enhances the tool's ability to process and analyze educational datasets effectively, delivering valuable insights for stakeholders.

The classification methodology is structured into several distinct phases: algorithm selection, attribute selection, run algorithm, running list, and result set repository, as illustrated in Figure 6. Each phase plays a crucial role in the systematic execution of classification tasks, ensuring a streamlined process for analyzing and interpreting educational data effectively.



Select Algorithm: ID3
Target Attribute: FINAL_RESULT

All Attributes
AREA
PUNJABI_RESULT
PUNJABI_GRADE
ENG_RESULT
ENG_GRADE
HINDI_RESULT
HINDI_GRADE
MATH_RESULT
MATH_GRADE
SCIENCE_RESULT
SCIENCE_GRADE
SS_RESULT
SS_GRADE

Selected Attribute
SEX
AREA
PUNJABI_RESULT
ENG_RESULT
HINDI_RESULT
MATH_RESULT
SCIENCE_RESULT
SS_RESULT
FINAL_RESULT

C4.5 Fri Jan 27 13:14:31 IST 2017
ID3 Fri Jan 27 13:16:10 IST 2017

```
DECISION TREE
   SS_RESULT->
     P
       ENG_RESULT->
         P
           HINDI_RESULT->
             P
               SCIENCE_RESULT->
                 P
                   MATH_RESULT->
                     P
                       PUNJABI_RESULT->
                         P
                           SEX->
                             F
                               AREA->
                                 R
                                     =P
                                 U
                                     =P
                             M
                               AREA->
                                 R
                                     =P
                                 U
                                     =P
                         F
                           AREA->
                             R
                                 =R
                             U
                               SEX->
                                 F
                                     =R
```

Fig. 6. Running list of decision tree algorithms corresponding to mentioned data set

Figure 6 illustrates the execution of the ID3 and C4.5 algorithms within the classification module. The classification process is organized into several key windows, each serving a distinct purpose:

- **Algorithm Selection Window**: In this window, users select the desired algorithm and the target attribute. The chosen algorithm determines the execution methodology, while the target attribute specifies the parameter for classification.
- **Attribute Selection Window**: This window facilitates the selection of attributes for classification. The first pane displays all the attributes available in the loaded dataset, while the second pane lists the attributes chosen for execution. Attribute filtration is performed here to finalize the attributes for the classification process (Sukhija et al., 2016).
- **Run Algorithm Window**: The primary component of the classification process, this window initiates the execution of the selected algorithm. A "view tree" option is integrated, allowing users to visualize the classification results in a decision tree format for better interpretability (Sukhija et al., 2015).
- **Running List Window**: This window maintains a log history of all implemented algorithms. Users can select specific entries from the list to view their corresponding outputs.
- **Output Window**: The classification results are displayed here. The output includes the results of the selected algorithm, the chosen attributes, and their classification in relation to the target attribute.

This modular design ensures a streamlined and user-friendly process for executing and analyzing classification tasks, enhancing the usability and functionality of the proposed EDM analytical tool.

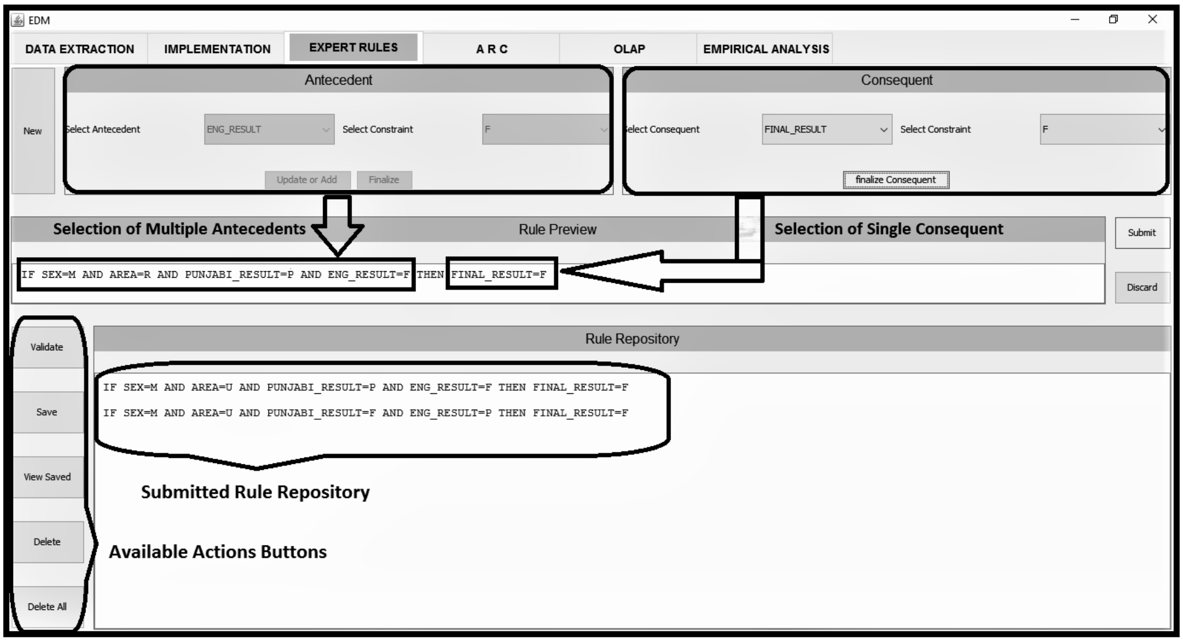
3.4. Expert Rule Set Incorporation

Although the improvised frequent pattern mining algorithms for educational data yield optimal results, they primarily generate outputs in the form of association rules and classification hierarchies derived from frequent pattern analysis. To further enhance the frequent pattern mining methodology, we introduce an expert rule set repository (Garcia et al., 2011). This repository facilitates both subjective and objective analyses of newly incorporated rules by domain experts. The graphical interface of the expert rule set window is depicted in Figure 7.

The incorporation of this feature into the proposed EDM analytical tool enables meaningful interaction between domain experts and the frequent pattern mining process. While frequent pattern mining algorithms (Garcia et al., 2011) rely solely on input datasets for generating results, there are instances where the output fails to meet user expectations. To address this gap, the expert rule set framework provides a mechanism for users to define and integrate new rules that may not be captured by traditional frequent pattern mining algorithms.

The expert rule set feature is built on IF-THEN-ELSE logic (Romero et al., 2008; Garcia et al., 2011), allowing domain experts to contribute rules that reveal significant insights about the classification of educational data fields. When constructing new rule sets, users can specify multiple antecedents linked to a single consequent (Garcia et al., 2011). Additionally, the framework ensures that attributes or fields in the rule-building process are connected exclusively using the AND (&) clause.

This innovative approach bridges the gap between automated analysis and expert knowledge, enabling the proposed EDM analytical tool to uncover deeper insights and offer more targeted solutions to challenges in the educational domain.



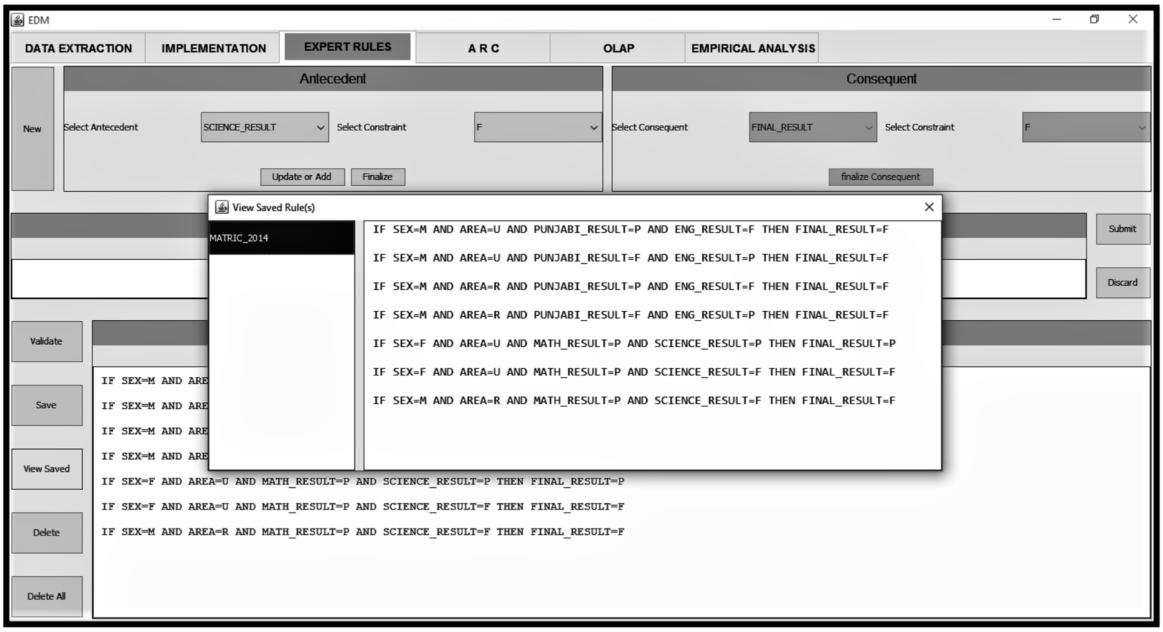
Select Antecedent: ENG_RESULT
Select Constraint: F
Select Consequent: FINAL_RESULT
Select Constraint: F

```
IF SEX=M AND AREA=R AND PUNJABI_RESULT=P AND ENG_RESULT=F THEN FINAL_RESULT=F
```

```
IF SEX=M AND AREA=U AND PUNJABI_RESULT=P AND ENG_RESULT=F THEN FINAL_RESULT=F
IF SEX=M AND AREA=U AND PUNJABI_RESULT=F AND ENG_RESULT=P THEN FINAL_RESULT=F
```

Fig. 7. Home window view of expert rule set incorporation process against the loaded educational data set.

The figure 8 depicts the phases of construction of new rules with multiple antecedents corresponding to single consequent. Now question arises that, these rules are independent from the frequent pattern mining algorithm, so how to validate these rules? Without validation of these rules set, it is uncertain to use it for further execution.



Select Antecedent: SCIENCE_RESULT
Select Constraint: F
Select Consequent: FINAL_RESULT
Select Constraint: F

MATRIC_2014

```
IF SEX=M AND AREA=U AND PUNJABI_RESULT=P AND ENG_RESULT=F THEN FINAL_RESULT=F
IF SEX=M AND AREA=U AND PUNJABI_RESULT=F AND ENG_RESULT=P THEN FINAL_RESULT=F
IF SEX=M AND AREA=R AND PUNJABI_RESULT=P AND ENG_RESULT=F THEN FINAL_RESULT=F
IF SEX=M AND AREA=R AND PUNJABI_RESULT=F AND ENG_RESULT=P THEN FINAL_RESULT=F
IF SEX=F AND AREA=U AND MATH_RESULT=P AND SCIENCE_RESULT=P THEN FINAL_RESULT=P
IF SEX=F AND AREA=U AND MATH_RESULT=P AND SCIENCE_RESULT=F THEN FINAL_RESULT=F
IF SEX=M AND AREA=R AND MATH_RESULT=P AND SCIENCE_RESULT=F THEN FINAL_RESULT=F
```

```
IF SEX=M AND ARE
IF SEX=M AND ARE
IF SEX=M AND ARE
IF SEX=M AND ARE
IF SEX=F AND AREA=U AND MATH_RESULT=P AND SCIENCE_RESULT=P THEN FINAL_RESULT=P
IF SEX=F AND AREA=U AND MATH_RESULT=P AND SCIENCE_RESULT=F THEN FINAL_RESULT=F
IF SEX=M AND AREA=R AND MATH_RESULT=P AND SCIENCE_RESULT=F THEN FINAL_RESULT=F
```

Fig. 8. Incorporated expert rules set corresponding to educational data set.

The validation measures for various frequent pattern mining algorithms are integrated within this window. A "Validate" button is provided to perform a comprehensive validation check on the newly constructed rule sets defined by the domain expert. These expert rules are evaluated against the loaded dataset to ensure their accuracy and applicability. The process involves identifying multiple antecedents corresponding to a single consequent within the dataset. Based on these findings, a detailed substantiation report is generated for the rule set, outlining its consistency, reliability, and alignment with the dataset. This validation mechanism ensures the expert rules are logically sound and suitable for integration into the analytical process.

4. Deliverable Analysis

The proposed integrated framework, combining educational data mining and visual analytics, aims to predict students' academic performance and derive effective solutions for improving the education system. By customizing mining algorithms specifically for educational data, the framework optimizes existing scenarios, offering efficient approaches for educational decision support systems. The innovative concept of an expert rule set enables the incorporation of diverse rules that can account for a wide range of values within the available educational dataset. These rules are validated against multiple thresholds aligned with the loaded dataset, uncovering optimal hidden patterns within the educational context (Rosalind James, 2016). Additionally, the integration of a visual analytics feature extends the tool’s capabilities to include exploratory analysis of the educational dataset.

The experiments were conducted on time-varying student data collected from the aforementioned educational sources to analyze and compare academic performance over specific years. During the implementation phase, various patterns were discovered using customized association and classification mining techniques. Furthermore, association rules deemed valuable by subject matter experts, but not returned by standard mining algorithms, can be integrated via the expert rule set module.

Table 1 presents the validation results of the submitted rule sets across three key parameters: support, confidence, and constraint satisfaction. These validation results quantify the reliability of the rules by assessing their support and confidence levels, while the "constraint satisfied" attribute indicates the rules’ evaluation outcomes in Boolean form (i.e., true or false). This systematic validation process provides users with a clear rationale for selecting rule sets that are both feasible and valuable for the educational decision support system (Sukhija et al., 2016). Moreover, as shown in Figure 9, these rule sets are further validated against the loaded dataset to identify and select the most optimal rules.

<table>
<tr><th colspan="5">TABLE 1: Validated rule set corresponding to aforementioned educational data set</th></tr>
<tr><td colspan="5">Total Records: 187862</td></tr>
<tr><td>SR. NO.</td><td>RULE</td><td>SUPPORT</td><td>CONFIDENCE</td><td>CONSTRAINTS SATISFIED</td></tr>
<tr><td>1</td><td>IF SEX=F AND AREA=R AND PUNJABI_RESULT=P AND ENG_RESULT=F AND MATH_RESULT=F AND SCIENCE_RESULT=F THEN FINAL_RESULT=F</td><td>0.34</td><td>4.32</td><td>NO</td></tr>
<tr><td>2</td><td>IF SEX=F AND AREA=R AND PUNJABI_RESULT=F AND ENG_RESULT=P AND MATH_RESULT=P AND</td><td>0.00</td><td>0.00</td><td>NO</td></tr>
</table>

	SCIENCE_RESULT=P THEN FINAL_RESULT=F			
3	IF SEX=F AND AREA=U AND PUNJABI_RESULT=P AND ENG_RESULT=F AND MATH_RESULT=F AND SCIENCE_RESULT=F THEN FINAL_RESULT=F	0.15	1.93	NO
4	IF SEX=F AND AREA=UAND PUNJABI_RESULT=F AND ENG_RESULT=P AND MATH_RESULT=P AND SCIENCE_RESULT=P THEN FINAL_RESULT=F	0.00	0.01	NO
5	IF SEX=M AND AREA=R AND PUNJABI_RESULT=P AND ENG_RESULT=F AND MATH_RESULT=F AND SCIENCE_RESULT=F THEN FINAL_RESULT=F	0.58	7.52	NO
6	IF SEX=M AND AREA=R AND PUNJABI_RESULT=F AND ENG_RESULT=P AND MATH_RESULT=P AND SCIENCE_RESULT=P THEN FINAL_RESULT=F	0.00	0.04	NO
7	IF SEX=M AND AREA=U AND PUNJABI_RESULT=P AND ENG_RESULT=F AND MATH_RESULT=F AND SCIENCE_RESULT=F THEN FINAL_RESULT=F	0.21	2.68	NO
8	IF SEX=M AND AREA=U AND PUNJABI_RESULT=F AND ENG_RESULT=P AND MATH_RESULT=P AND SCIENCE_RESULT=P THEN FINAL_RESULT=F	0.00	0.02	NO

The combination of educational patterns identified by mining algorithms and those contributed by domain experts is seamlessly integrated into the visual analytics module. This module offers dynamic, graphical representations of the discovered results, enabling insightful analysis. Figure 9 highlights visual analytics generated from both algorithmic and expert-defined rules, based on selected attributes from the input dataset. The identified rules uncover dependencies among key learner attributes, such as examination subjects, grades, institutional areas, and residential locations. The visual elements in Figure 9 include:

- **Column Chart**: Displays performance across subjects including English, Hindi, Mathematics, Punjabi, Science, and Social Studies (Hema et al., 2018).
- **Line Chart**: Illustrates grade trends across these subjects, with grades categorized as A+, A, B, C, D, and E.

- **Pie Chart**: Represents the distribution of learners' residential areas, segmented into urban and rural.
- **Bar Chart**: Visualizes district-level statistics derived from the result set.

The visualizations are linked to a designated target attribute—in this case, the "Final Result" attribute, with possible values of Pass, Fail, and Reappear. The graphs dynamically update based on the selected target attribute value, and Figure 9 specifically illustrates analytics for the "Reappear" value. These graphical insights provide a clear, data-driven view of student performance and related factors. By combining algorithmic findings with expert input, the proposed system supports the development of effective strategies to enhance educational outcomes and improve system-wide performance.

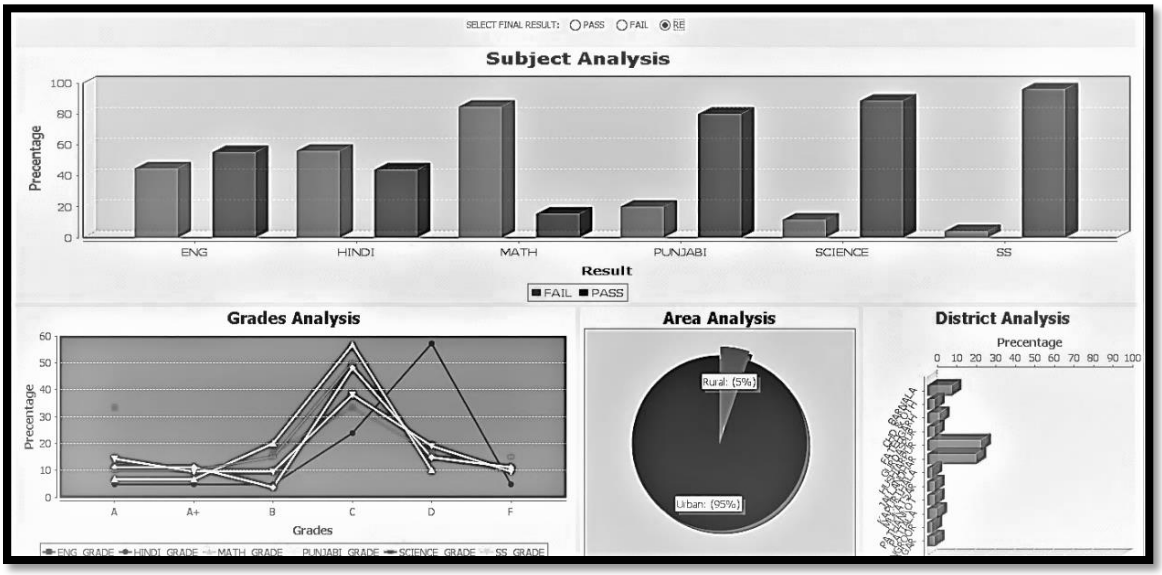


Fig. 9. Resultant graphs of online analytical module based on the discerned rule set.

5. Empirical analysis

In the experimental phase, an empirical analysis was conducted to validate time-sensitive educational data. The matriculation course data for the years 2014 and 2015, obtained from the Punjab School Education Board, India, was used to examine key dependent attributes within the education system. The analysis focused on common attributes derived from the resultant rule set. Initially, a counter-check was performed to verify the presence of specific attributes within the overall dataset. Subsequently, these selected attributes were validated against various performance parameters. This process was applied consistently across all relevant attributes, resulting in percentage-based statistics for the identified rule set. Figure 10 presents the resultant statistics along with a graphical representation of the dependent attributes analysed during this empirical study.

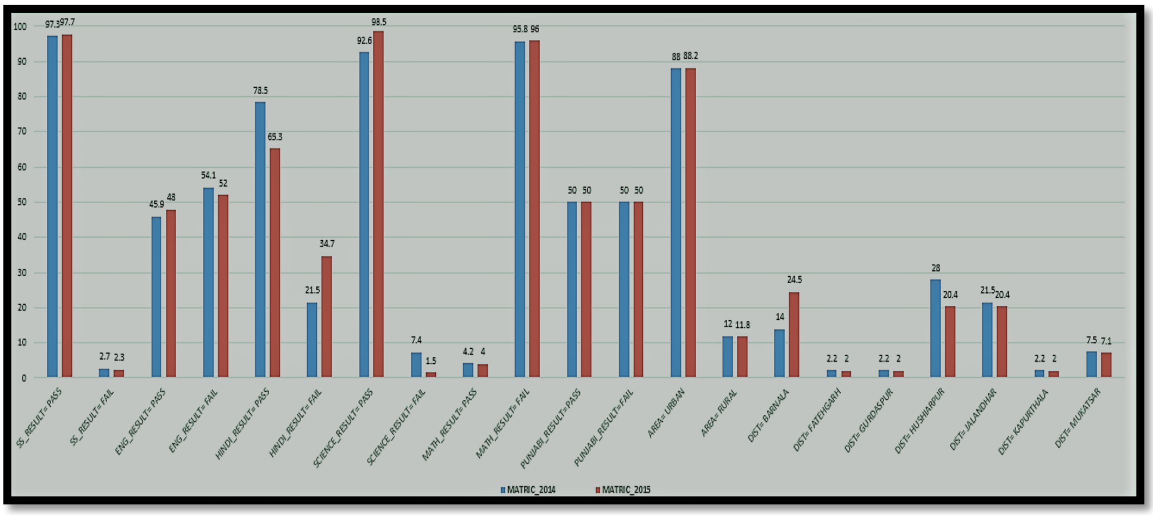


Fig. 10. Empirical analysis results against the two years data set (2014 & 2015) of matriculation course of Punjab School Education Board, India.

This empirical analysis revealed both improvements and declines in students' academic performance over the specified educational data range. As shown in Table 1, the pass rate for the Hindi subject experienced a 13.2% performance decline, whereas the Science subject showed a 5.9% improvement. In conclusion, this methodology leverages educational datasets to uncover dependencies among various attributes, providing actionable insights that can contribute to the enhancement of the educational system.

6. Conclusion

This research paper introduces a novel methodology in the form of an Educational Data Mining (EDM) analytical tool, designed to leverage students' examination data to uncover dependencies among various attributes that significantly influence academic outcomes. Existing state-of-the-art methods often overlook the complex interdependencies among educational attributes, limiting their ability to achieve remarkable insights.

The core focus of this study is to analyze students' academic performance based on various attributes, including subjects, grades, gender, and residential area. The experimental process utilized six consecutive years (2011–2016) of examination data from matriculation and senior secondary courses of the Punjab School Education Board, India. Given the increasing scale of educational systems, the proposed framework combines frequent pattern mining with expert rule integration to extract meaningful insights from multi-dimensional datasets.

This paper presents an EDM analytical framework that enhances frequent pattern mining algorithms and introduces an expert rule set module. This module allows domain experts to incorporate new rules beyond the scope of mining algorithms. These derived rules are validated against thresholds such as support and confidence to identify optimal patterns. Additionally, the visual analytics feature within the EDM tool provides a graphical representation of the discovered patterns, making insights more accessible and actionable.

The empirical analysis conducted during validation extends the tool's utility to exploratory educational data analysis. The results demonstrate the framework's capability to verify hypotheses and support the enhancement of school education systems. By integrating expert-driven rules with data-driven mining, the EDM analytical framework delivers optimal results, identifying significant patterns that contribute to improving educational strategies and outcomes.

7. Future scope

While this research has successfully enhanced and implemented various frequent pattern mining algorithms for educational datasets, a key limitation remains: the lack of integration between association and classification techniques. This gap hinders the discovery of complex, multidimensional dependencies within attribute sets. To overcome this, there is a clear need for a hybrid approach that combines the strengths of these techniques, enabling constraint-based multidimensional frequent pattern mining within an association rule-based classification framework specifically designed for educational datasets. This hybrid methodology promises a deeper analysis of educational systems, offering a more nuanced understanding of students' academic performance and learning environments. Furthermore, a novel experiment will be conducted to validate and generalize the results, ensuring their applicability across diverse educational domains. By addressing this limitation, the proposed hybrid approach can significantly enhance the ability to derive meaningful insights, ultimately supporting more informed and effective educational decision-making.

8. References

1. Alom, B. M., and Matthew Courtney. (2018) "Educational Data Mining: A Case Study Perspective from Primary to University Education in Australia". International Journal of Information Technology and Computer Science Vol 2 Issue 2, pp. 1-9.
2. Angel, Gomez, Luna, Romero, Ventura. (2015). Discovering clues to avoid middle school failure at early stages. ACM Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY, USA, March 16 - 20, pp. 300-304.
3. Bakhshinategh, Behdad, Osmar R. Zaiane, Samira ElAtia, and Donald Ipperciel. (2018) "Educational data mining applications and tasks: A survey of the last 10 years." Education and Information Technologies Vol 23 Issue 1, pp. 537-553.
4. BH, Hema Malini, and L. Suresh. (2018) "Data Mining in Higher Education System and the Quality of Faculty Affecting Students Academic Performance: A Systematic Review". International Journal of Innovations & Advancement in Computer Science, Vol 7 Issue 3, pp. 66-70
5. Bennett, Sue, Shirley Agostinho, and Lori Lockyer. (2017) "The process of designing for learning: understanding university teachers' design work." Educational Technology Research and Development Vol 65 Issue 1, pp. 125-145.
6. Bidgoli, B., Kashy, Kortemeyer, Punch. (2003). Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA. Proceedings of ASEE/IEEE Frontiers in Education Conference, Boulder, Colorado, November 5 - 8, pp. 13-18.
7. Chandra and Nandhini. (2007). Predicting Student Performance using Classification Techniques. Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India, Dec 15, pp. 83-87.
8. Calvet Linan, L., & Juan Perez, A. A. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. International Journal of Educational Technology in Higher Education, Vol 12 Issue 3, pp. 98-112.
9. Dara, Raju, Satyanarayana, and Govardhan. (2016). A Novel Approach for Data Cleaning by Selecting the Optimal Data to Fill the Missing Values for Maintaining Reliable Data Warehouse. International Journal of Modern Education and Computer Science, Vol 8 Issue 5, pp. 64-70.
10. Fernandes, Eduardo, Maristela Holanda, Marcio Victorino, Vinicius Borges, Rommel Carvalho, and Gustavo Van Erven. (2019) "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil." Journal of Business Research, Vol 94, pp. 335-343.
11. Garcia, Enrique, Romero, Ventura, and Castro. (2011). A collaborative educational association rule mining tool. The Internet and Higher Education, Vol 14 Issue 2, pp. 77-88.
12. Geryk, Jan. "Visual analytics by animations in higher education. (2013) Proceedings of the 12th European Conference on e-Learning ECEL. Sophia Antipolis, France, Academic Conferences and Publishing International. pp. 565-572.
13. Geryk, Jan, and LubosPopelinsky. (2014). Analysis of Student Retention and Drop-out using Visual Analytics. The 7th International Conference on Educational Data Mining EDM. Institute of Education (IOE), London, UK. pp. 331-332.
14. Igor, LjiljanaBrkic, Mirta. (2009). Improving the ETL process and maintenance of Higher Education Information System Data Warehouse. Journal WSEAS Transactions on Computers archive. Vol 8 Issue 10, pp. 1681-1690.
15. Ihantola, Petri, ArtoVihavainen, Ahadi, Butler, Borstler, Edwards, EssiIsohanni et al. (2015). Educational data mining and learning analytics in programming: Literature review and case studies. In Proceedings of the ITiCSE on Working Group Reports, July 4–8, Vilnius, Lithuania, pp. 41-63.
16. Jindal. (2013). A Survey of Educational Data mining and Research Trends. International Journal of Database Management Systems. Vol 5 Issue 3, pp. 53-73.

17. Kovacic. (2010). Early Prediction of Student Success: Mining Students Enrollment Data. Proceedings of Informing Science & IT Education Conference Open Polytechnic, Wellington, New Zealand, pp. 647-665.
18. Kumar, Anupama. (2016). Edifice an Educational Framework using Educational Data Mining and Visual Analytics. International Journal of Education and Management Engineering. Vol 2, pp. 24-30.
19. Manhaes, Barbosa, Cruz, and Zimbrao. (2014). WAVE: an architecture for predicting dropout in undergraduate courses using EDM. In Proceedings of the 29th Annual ACM Symposium on Applied Computing. Gyeongju, Korea, March 24 - 28, pp. 243-247.
20. Merceron, Yacef. (2003). A Web-based Tutoring Tool with Mining Facilities to Improve Learning and Teaching. Proceedings of 11th International Conference on Artificial Intelligence in Education. Sydney, Australia, July 20-24, pp. 201-208.
21. Quinlan. (1994). C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, Vol 16 Issue 3, pp. 235–240.
22. Romero, Cristobal, Ventura, and Garcia. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education. Vol 51 Issue 1, pp. 368-384.
23. Rosalind James (2016), Tertiary student attitudes to invigilated, online summative examinations, International Journal of Educational Technology in Higher Education, pp. 13-19.
24. Simendinger, Earl, Abdul-Nasser El-Kassar, Maria Alejandra Gonzalez-Perez, John Crawford, Stephanie Thomason, Philippe Reynet, Björn Kjellander, and Judson Edwards. (2017) "Teaching effectiveness attributes in business schools." International Journal of Educational Management Vol 31, Issue 6, pp. 780-800.
25. Sohrabi, Karim, and Akbari. (2016). A comprehensive study on the effects of using data mining techniques to predict tie strength. Computers in Human Behaviour. Vol 60, pp. 534-541.
26. Sukhija, Jindal, and Aggarwal. (2015). The recent state of educational data mining: A survey and future visions. IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education, October 1-2, pp. 354-359.
27. Sukhija, Jindal, and Aggarwal. (2016). Educational data mining towards knowledge engineering: a review state. International Journal of Management in Education. Vol 10 Issue 1, pp. 65-76.
28. Wang, Shu-Ming, Huei-Tse Hou, and Sheng-Yi Wu. (2017) "Analyzing the knowledge construction and cognitive patterns of blog-based instructional activities using four frequent interactive strategies (problem solving, peer assessment, role playing and peer tutoring): a preliminary study." Educational Technology Research and Development Vol 65 Issue 2, pp. 301-323.
29. Wei, Koutrika, and Wu. (2014). Learn2Learn: A Visual Analysis Educational System for Study Planning. 17th International Conference on Extending Database Technology. Athens, Greece, March 24-28, pp. 656-659.
30. Winters. (2006). Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes-Based Assessment. A dissertation work of University of California, Riverside, pp. 105-151.