Thesis

Education Content Provider Based on Particular Weaknesses and Strengths Using Machine Learning Algorithms

Supervisor : *Dr. Jannatun Noor*

Name : *Muhammad Nayeem Mubasshirul Haque*

ID : 19101115

Name : *Tasnim Akter Fariha*

ID : 20101405

Name : *Shabab Intishar Rahman*

ID : 18241010

Name : *Ammar Mohammad*

ID : 19301063

Name : *Shadman Ahmed*

ID : 20101031

Submission date : $17 - 09 - 2023$

**Department of
Computer Science and Engineering**

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.
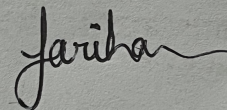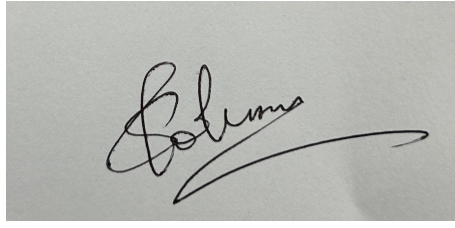
Student's Signature:

Approval

The thesis/project titled "Education Content Provider Based on Particular Weaknesses and Strengths Using Machine Learning Algorithms" submitted by

1. Muhammad Nayeem Mubasshirul Haque

2. Tasnim Akter Fariha

3. Shabab Intishar Rahman

4. Ammar Mohammad

5. Shadman Ahmed

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September 17, 2023.

Examining Committee:

Supervisor:
(Member)

———————————————

Dr. Jannatun Noor
Assistant Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

———————————————

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

To uncover underlying patterns in large datasets, a procedure called data mining is often utilized. By analyzing data gathered through Online Learning (OL) systems, data mining can be used to unearth hidden relationships between topics and trends in student performance. Here in this paper, we show how data mining techniques such as clustering and association rule algorithms can be used on historical data to develop a unique recommendation system module. In our implementation, we utilize historical data to generate association rules specifically for student test marks below a threshold of 60%. By focusing on marks below this threshold, we aim to identify and establish associations based on the patterns of weakness observed in the past data. Additionally, we leverage K-means clustering to provide instructors with visual representations of the generated associations. This strategy aids teachers in better comprehending the information and associations produced by the algorithms. K-means clustering helps visualize and organize the data in a way that makes it easier for instructors to analyze and gain insights, enabling them to support the verification of the relationship between topics. This can be a useful tool to deliver better feedback to students as well as provide better insights to instructors when developing their pedagogy. This paper further shows a prototype implementation of the above-mentioned concepts to gain opinions and insights about the usability and viability of the proposed system.

Keywords: Threshold, Weaknesses, Unsupervised algorithms, Associative pattern, E-learning sphere, Prototype implementation

Acknowledgement

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ϵ Epsilon

υ Upsilon

EDM Educational Data Mining

FC Forgetting Curve

GBL Game based Learning

OL Online Learning

$VOIP$ Voice Over Internet Protocol

$WCSS$ Within Cluster Sum Square

Chapter 1

Introduction

1.1 Educational Data

The spheres of education and technology have become intertwined in recent years. The development of various e-learning systems over the past few decades has greatly increased the usage of computers in learning and teaching. With covid-19 epidemic driving the global education sector into an emergency shift to OL, we all saw the scope and capabilities of this sector. Beyond this emergency shift, OL has been trending upward in terms of participants for the past few years as technology has become more accessible to the general populace. With over 150 million users partaking in OL, this field of education is also bound to produce massive amounts of data. As students engage in learning, explore various subjects, complete tests, and submit projects and assignments, they leave behind a sea of information. Hence this prompts the question: How can we leverage these extensive data archives? This brings us to the concept of Educational Data Mining (EDM). To explore data from educational contexts, EDM emerges as a paradigm focused on designing models, activities, methodologies, and algorithms. EDM seeks to identify trends and forecasts that represent learners' actions and accomplishments, as well as assessments, educational features, and applications [37]. In recent times we can identify several trends in EDM. Some of these trends include the inclusion of EDM modules as a standard component of computer-based educational systems. The use of EDM at various stages of the teaching and learning process is another trend. EDM assists in the initial stage's customization of the learning environment based on the profile of the student. EDM analyzes data as the student is interacting with the system and offers suggestions in real-time. The success of the education delivered, including services, results, user happiness, and resource usefulness, is assessed at the final step by EDM.

This paper's focus will be on developing an EDM module as a component of an existing OL system. Our module takes the historical student test marks database from an OL system. This database is filtered only to include marks below 60%. For this research, 60% is being held as the minimum threshold for a pass. This filtered data is then passed through an association algorithm. By analyzing historical data on students' performance in different topics, our system identifies associations between topics in which students typically struggle. These associations are then visualized using clustering graphs for the ease of understanding of the instructors. The output (the association of topics in which students are not performing up to

standard) can then be used to reinforce and integrate into the revision system of the OL platform upon which this module is to be integrated. This paper further shows a prototype demonstration of this module to instructors and teachers of various spheres of education to gather their valuable opinions regarding this proposed system. Additionally, A/B testing has also been performed on a volunteer group of students to obtain a quantitative evaluation of the effectiveness of our system.

1.2 Motivations

With the standards of education moving away from preparing students for the job market and more towards critical thinking and adaptive learning [55], it is imperative that we keep our tools for education up-to-date as well. To ensure that educators are receiving more insightful data from students, which can result in significantly better feedback [51] along with a better presence of the instructors [51], we must continue to improve our tools in all areas of education, but especially in online learning. Previous research has shown these factors to be of huge importance when it comes to user satisfaction with an OL learning environment and has consistently been the issue that most students raise when asked about the challenges of OL[50]. As most OL learning models are variations on the mastery model [1] of learning and teaching, it is high time we incorporate more modern algorithms to aid in further developing the models being applied to different OL platforms. It is indisputable that educational data mining is a significant emerging field of research. Efforts should be made in the creation of novel, intuitive, and simple-to-implement software solutions. Such software can be essential in promoting online learning and encouraging efficient teaching methods by utilizing the large data reservoirs and patterns generated through EDM. Adopting this strategy has the potential to bring about fundamental changes in the educational landscape by equipping both educators and students with insightful knowledge and effective tools for improved learning experiences. The possibility of having a positive impact on education's future is compelling.

1.3 Research Objective

OL has been a large part of the education sector since networking and computation have been readily available[6]. Besides its myriad of features[9], it also provides educators and learners with precisely calculated and curated data for further evaluation and subject recommendation. In recent years there has been a huge bloom in OL[15], especially during the covid-19 pandemic [58]. OL has several well-researched benefits and drawbacks. Much research has come forth on how to better the pedagogy of OL [46] and how implementing new technologies has led to better user satisfaction with the system through iterative design [57], hence it remains a priority to try and optimize and revamp designs and trends over time[15]. The paper will be focused on specifically Apriori Algorithm [23], and F-P Growth Algorithms as association algorithms for educational data mining alongside K-means clustering[2] as a clustering algorithm. Based on the data mining results, we suggest creating a recommendation system that makes use of the identified association rules to provide students with pertinent learning resources based on previously discovered trends of weakness. This kind of learning support aims to give students the tools they need

to address their areas of weakness and eventually enhance their academic performance. We want to create a thorough revision module that combines the identified themes using historical data, allowing new students to learn from the mistakes of their forebears. Secondly, these trends and patterns unearthed through EDM can be useful intel for instructors. We also highlight a visual representation of our analytical results to aid instructors with their pedagogical decision-making using the EDM results. In addition, the research paper also showcases a prototype implementation of the above-stated module to a volunteer group of students and instructors to gain insights into the viability, effectiveness, and usability of the proposed model. By contributing to the broader field of educational data mining, our research seeks to advance knowledge and understanding of the effective implementation of EDM in an educational content provider module which can be applied to other OL platforms.

1.4   Research Contributions

Our paper makes three significant contributions to the sphere of OL and EDM. They are as such :

- The methodological landscape of educational data mining and recommendation systems is improved by this study. Our novel use of the Apriori and F-P growth algorithms to analyze educational data offers a fresh method for identifying undiscovered correlations in student learning patterns. We found complex topic association patterns using the algorithms on historical student data. This not only improves our comprehension of how students' mastery of one topic affects their performance in adjacent topics, but it also offers insightful information for educators and course planners. For the purpose of creating more efficient and individualized learning experiences that would ultimately improve students' academic progress, it is essential to identify topic weaknesses and dependencies.

- The prototype and validation of a component or feature of a recommendation system is another contribution. We developed a component of a recommendation system that directs students toward specific remedial content and resources based on their shortcomings by utilizing the insights we learned from our association algorithm study. We outlined the viability of our prototype in enhancing students' learning outcomes through A/B testing. The success of our recommendation system's implementation highlights both its usefulness in real-world settings and its capacity to improve learning outcomes in online learning settings. Furthermore, instructors were interviewed to gain their expert opinion on the developed prototype.

- Our study advances the use of relatively more tailored instruction in online education. We create the foundation for a more individualized and adaptive educational experience by utilizing data-driven insights to detect and address topic shortcomings. This is especially important when it comes to online learning, where a variety of student demographics need individualized help to succeed. Thus, our research provides a framework for further study and the creation of recommendation systems that might assist both instructors and students in enhancing the online learning environment.

1.5 Thesis Organisation

This thesis is organized into the following chapters, each of which contributes to the overall understanding of the research topic and its implications.

- Introduction :

  - Educational Data: A brief idea on the sphere of OL and EDM.
  - Motivation: Discussion on why EDM is important and why a novel recommendation system is important.
  - Research Objective: What are the overarching goals we wish to accomplish with this paper.
  - Research Contribution: What are the different contributions our paper has actually managed to make.

- Related Work: An exploration of previous relevant literature associated with our thesis topic.

- Background: In-depth discussion on the architecture and technical aspects of all utilized algorithms and methodology alongside comparing and contrasting different association algorithms.

- Dataset: Details on how data was collected and description of the utilized dataset.

- Methodology:

  - Data Collection: How we have collected our data.
  - Data Pre-Processing: What pre-processing methods were applied to the dataset to be usable for our application.
  - Experimental Test-Bed: Information about all software and hardware utilized during our research.
  - Applied Methodology: describes what methods were used during our entire research paper, including details about our prototypes.

- Experimental Evaluation: Describes all results obtained from applied methodology and discusses the implications and findings from these results, including the results from our validity tests.

- Conclusion: Summarizes the entire thesis along with future works to continue and make the research more robust.

Chapter 2

Related Work

The boom of technological advancement during the early 2000s [48], greatly fueled the usage of computers and the Internet in learning environments. E-learning or Online Learning is described as the use of information and communication technology in the education sector to enable the provision of services aimed at improving academic results [31]. When the issues and challenges of implementing OL [19] started to be overcome as time passed, we began to see more and more users subscribing to the idea of OL. As suggested by the Technology Acceptance Model (TAM) [33], the younger generation initially looked through unofficial lectures on various topics mainly on video-sharing platforms such as YouTube[44]. One of the biggest and most popular channels on YouTube (in the context of OL) was known as "Khan Academy" [17]. "Khan Academy" had recorded lectures explaining various topics on a broad spectrum of difficulty levels, thus pioneering the 'e-learning'(EL) model. Khan Academy eventually grew into its website, adding the next most important feature in the OL: testing. Testing allowed the advent of OL to venture into the mastery model of learning [1]. The mastery model can be described as a learning model which expects users to show a certain acceptable level of competence on a certain topic/subject before they are allowed to delve further into the learning materials. This was of course not just limited to Khan Academy but other learning models also followed this formula for OL. The type of OL described so far can be categorized as asynchronous learning [5]. Asynchronous learning has two main advantages [61] :

- Flexibility: Since this type of learning takes advantage of recorded lectures and videos, students and learners can choose to view these lessons at their convenience instead of in real time. This allows learners to absorb the materials at their own pace, whenever they want.

- Deeper Learning: since asynchronous learning does not follow real-time lectures or have hard and fast traditional deadlines, learners are encouraged to research further into their materials and content. This can allow learners to identify further context and literature about their concerned subjects.

The other type of learning that gained traction was synchronous learning. With Voice Over Internet Protocol(VOIP) [16] technology advancing to enable low-latency communication, voice, and video calls started to become prevalent [16]. This allowed OL classes to be taken virtually and remotely, where learners and instructors could communicate in real-time.
Hybrid classes [53], became a possibility and thus another sphere of OL was created.

Synchronous OL has 2 significant advantages [61] :

- Interaction: Since lessons are being carried out in real-time, the instructor and learner may talk to each other, ask questions and overall have a social presence in the classroom

- Accountability: Since synchronous learning requires the learner to be online and active during a certain time of day, it fosters the feeling of being part of a regulated program which helps students with time management and maintaining a routine.

Trends in OL [38] have since been constantly changing, growing, and being iterated. One of the latest trends in OL is introducing game-based learning (GBL) [24]. GBL uses many different gamification techniques to help facilitate and motivate learners using a game-like environment and reward systems. Giving learners virtual points or badges upon achieving certain goals, gives them immediate feedback and positive reinforcement for their efforts, and previous research has shown this technique to positively affect a student's experience with OL and their learning rate [34].

With new technology and research bringing forth new ideas such as GBl, we must be open-minded in assuming that constant iteration and testing are needed to always try to reinforce OL.

One factor that we believe to be a disadvantage of the widely adopted mastery learning model is that it does not promote revision of a topic [36], additionally, some learners may face more difficulty achieving the level of merit required by the model to move past a certain problem topic. The mastery model does not typically have any tools to address these issues.

Ol typically also does not consider the Forgetting Curve (FC) [56], which states we are prone to forgetting things when not looked at over time. This theory can be applied to education as well [43]. FC was initially researched by Hermann Ebbinghaus and does not seem to have seen much mention in OL pedagogy or applications. Although, work has been done researching how students forget and learn [13][27], some acknowledgments have been made on how we can mitigate the effects of the forgetting curve in education.

One of the suggestions for mitigating the FC and OL pedagogy is making the content more engaging. If we look at recent research into OL pedagogy [20], we can see that more appropriate feedback alongside engaged and active instructor presence is a top priority.

Previous work has been done to evaluate student satisfaction with online learning, A survey at Deakin University in Australia [18] revealed that most students were satisfied with the accessibility of information and the ease of submission of assignments, but were not satisfied with the level of feedback that they received. In the context of Bangladesh, [47] showed that the flexibility y and quality of information and assessment had the greatest effect on user satisfaction (for public university students). Another study [54], found that most students faced telecommunication issues during online learning which had a heavy negative impact on their satisfaction with online learning. The same study also concludes that students had issues such as feedback, lack of interaction, and failure to understand materials during online learning. Students also seem to prioritize the instructor's capacity to operate the tools for an online class and engage with students even in an online environment as a big contributing factor to how effective online learning can be.

The design of online learning has also evolved with the iteration and integration of new pedagogy and student needs in mind [35], such as gamification is one of the newer trends in online learning platforms

Bangladeshi students have now returned to in-person classes for the past year. With their unique retrospective insight, we can generate new ideas with their unique insights, as a good iterative design process [4].

One such idea is the use of machine learning with online learning [41] . Machine learning has the potential to unlock many new and important features that can improve online learning in ways we have already talked about and the solutions that will be proposed below. One such algorithm that can be applied is clustering algorithms[29]. Alongside that association algorithms can also be used to make design new tools and create new trends in the online learning space[42], leading to better pedagogy of online learning environments.

Most learning models employed in an online learning environment are variations on the Mastery model but forget to take into account other learning requirements and shortcomings, such as the forgetting curve [21].

Personalized recommendation systems are a prominent area of research in OL. Xiao et al.[26] had a significant impact in pushing forward this idea. Since then several innovative research and applications have been done for recommendation systems. Wei Xu and Yuhan Zhou, [49] proposed using deep learning for recommending the best courses based on a course's metadata and user content filtering. Furthermore, Hua Wang et al. [30] have given precedence to the use of association algorithms for recommendation systems, which can also be seen in a 2016 paper by Fang Liu et al. [39], which uses a variation of Apriori algorithm for university course recommendation.

Association Algorithms, specifically the Apriori [45] and F-P growth algorithm [10], have been previously used in research for data mining of education data. These algorithms find the frequent itemsets and generate association rules based on a minimum support threshold. One potential use of these algorithms can be as follows :

If we have a dataset that contains the marks of students divided into topics in each column, we can filter out data based on which subjects each student is failing These itemsets can be stored in a separate array structure. This array can be passed through the algorithms to generate association rules based on which topics students are struggling with. This would allow instructors to better understand the root cause of the problems that learners are facing. This idea will be focused on later in the paper. A similar ideology can be applied to generate personalized revision routines for each student according to the association between subjects that they are struggling with by the algorithms on a massive scale; such as an OL website based on asynchronous learning or in a hybrid real-time classroom.

Additionally, clustering algorithms have also been deployed on educational data [59], they can be used to find students who are struggling on similar subjects as well. However, we should also be aware that parameters and metric settings matter a lot for these algorithms to perform well.

Instead of guessing how many clusters we shall utilize, we can use the 'Elbow Method' This method takes the dataset and applies the k-means algorithm to them. For each value of K, we calculate the Within-Cluster Sum of Squares. This calculates the compactness of the clusters. Where the graph starts to flatten out, we call

this the elbow of the graph, and is usually the optimal choice to find out the number of clusters. In case this method is unclear we can also refer to the Silhouette score for determining optimal clusters [60]. A summary comparing and contrasting our proposed method with related research is shown below:

Research	Dataset Size	Method	Validation
Our Proposed System	399 students with 16 topics	Apriori algorithm on filtered data based on threshold	A/B Testing and Instructor's Interview
Xu, W., and Zhou, Y. (2020)	1853 learners of 5479 courses	Deep learning with content and collaborative filtering by extracting watch time and video meta-data	AUC score (a metric for binary classification models)
Jun Xiao et al.(2018)	-	Combination of sparsity,content and collaborative filtering	-
Hua Wang et al. (2014)	500 students	Improved apriori algorithm which process dataset vertically	Run-time comparison
Fang Liu et al. (2016)	-	Apriori Algorithm combined with collaborative filtering	-

Table 2.1: Comparison with related work

Chapter 3

Background

We are mainly focused on using an unsupervised association algorithm, specifically: Apriori Algorithm and F-P Growth algorithm. Alongside that, we will also be applying K-means Clustering. Finally, we shall also look at A/B testing as a validation technique.

3.1   Clustering

Clustering is a machine-learning technique that involves grouping similar data points based on their characteristics or attributes. The goal of clustering is to find patterns and relationships within the data, which can then be used to gain insights into the underlying structure and organization of the information.

The fundamental idea of clustering is rather straightforward. The program analyzes a set of data points and groups them based on similarities it finds between them. Different criteria, such as distance metrics (such as the Euclidean distance), correlation coefficients, or other measurements, can be used by the algorithm to identify similarity.

The k-means algorithm is one typical method of grouping. The user selects the number of clusters (k) they want to produce when using this approach. The method then determines the centroids (the center point) of each cluster after randomly assigning each data point to one of the k clusters. After that, until the clusters take on a stable configuration, iteratively reassigns data points to the nearest centroid and recalculates the centroids.

The Euclidean distance is the distance metric that K-means clustering uses the most frequently. In a d-dimensional space, it calculates the straight-line distance between two data points. The Euclidean distance between two points, x, and y, is determined as follows:

The Euclidean distance between two points x and y in a d-dimensional space is given by:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_d - y_d)^2}$$

Cluster centroid updates are necessary following data point assignment. The mean vector of all data points in a cluster makes up the centroid. The average of the

coordinates of all the data points in a cluster, C, is used to determine the updated centroid coordinates:

The formula for calculating the new centroid in k-means clustering is given by:

$$\text{new\_centroid} = \frac{1}{|C|} \sum_i x_i$$

Here, $|C|$ represents the number of data points in cluster C, and $\sum_i x_i$ denotes the sum of the coordinates of all data points in the cluster.

To determine the optimal number of clusters in K-means clustering we can use the "Elbow Method" and "Silhouette Score". The Elbow Method is a heuristic method for determining the K-means clustering's ideal number of clusters (K). This is how it goes:

You calculate the sum of squared distances (SSD) between each data point and its designated cluster center for various values of K (often ranging from 1 to some upper limit). The "within-cluster sum of squares" (WCSS) is another name for this. The WCSS formula is:

$$WCSS(K) = \sum_{i=1}^{N} \left(\text{distance}(data\_point_i, cluster\_center_i)^2 \right)$$

in this equation: N represents the total number of data points. data_point_i represents each data point. cluster_center_i represents the cluster center to which data_point_i is assigned. distance(data_point_i, cluster_center_i) represents the distance between data_point_i and cluster_center_i. The summation symbol, Σ, iterates over all data points, from 1 to N. The caret (ˆ) denotes exponentiation, so we're squaring the distance.

Plot the WCSS scores for various K values. The plot frequently resembles an "elbow," with the WCSS declining quickly at first (with a rising K) before beginning to level out. This is how the Elbow Method got its name.

In case it is still unclear on the optimal number of clusters, we can further refer to the silhouette score method. The silhouette score is a measure to find the optimal number of clusters in K-means clustering. It quantifies how well-defined and separate the clusters are. Here's how it works:

- For each data point, calculate:

    - "a": Average distance to other points in the same cluster.
    - "b": Average distance to points in the nearest different cluster.

- Compute the silhouette score for each point: $\frac{b-a}{\max(a,b)}$

- Calculate the average silhouette score for all points in the dataset.

- Repeat this process for different cluster numbers (k) and choose the k with the highest silhouette score as the optimal number of clusters.

Higher silhouette scores indicate better-defined clusters, helping us find the best cluster count for our data.

3.2   Apriori Association

The association is a type of unsupervised machine learning that takes transaction data and sorts them according to the frequency of seeing items together. In other words, it tries to find patterns of items occurring together. It is an iterative algorithm, it parses through the data and looks for frequent individual items. then, it parses the dataset again and looks for how often a pair of items from the frequent individual items appear together. It accepts or declines these itemsets based on a preset parameter such as "support". The iterations continue until no more frequent itemsets can be generated or all the generated itemsets no longer meet the minimum support threshold. The resulting frequent itemsets represent the associations that occur frequently in the dataset.

Association rules are also generated, which indicate exactly the conditional probability that has to occur to gain a certain frequent item set.

The 'support' is calculated using the following equations :

The equation for calculating the support of an itemset in association algorithms is given by:

$$\text{Support}(A) = \frac{\text{Number of transactions containing itemset A}}{\text{Total number of transactions}}$$

In this equation, A represents the itemset for which we want to calculate the support.

Besides support, the "strength" of the association is measured using conditional probability using a parameter known as 'confidence'. Confidence in association rule is a measure of the accuracy of a rule. In simple terms, if items A and B are in the association rule, their confidence will show how often B is seen together with item A.

The formula for confidence is given by:

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

Broadly, The Apriori algorithm functions like this :

- When a candidate itemset is created, it combines the item (k-1)-itemset that was obtained in the previous iteration. The availability of candidate k-itemset trimming with a subset comprising k-1 items but not in the high-frequency pattern with a length of k-1 is one feature of the Apriori method.

- The evaluation of each candidate's k-itemset's support. The number of transactions that contain all the items in each of our candidate candidates is determined by scanning the database for each candidate candidate. This is a characteristic of the Apriori algorithm as well, which necessitates calculation by thoroughly searching the database, starting with the longest items.

- Set a pattern with a high frequency. The candidate kitet with support larger than the minimal support is used to determine a high-frequency pattern with k items or itemset.

- The process is terminated if no new high-frequency pattern is discovered. If not, perform k plus 1 and go back to phase 1

3.3 Frequent Pattern Growth Algorithm (F-P Growth)

In the field of creating associations through item set mining, the Frequent Pattern Growth (FP-growth) algorithm is one of the fastest and most popular algorithms. [8] FP-growth algorithm is primarily based on FP-tree which is a prefix tree representation of the given database of transactions. [11] The FP-tree represents the dataset in the form of a tree where there is no need for candidate generation
In the preprocessing of the FP-growth algorithm, all individually infrequent items are deleted from the transaction database before turning into an FP tree. This selection can be done with the help of a threshold value. In the FP-tree, each path represents a set of transactions that share the same prefix, each node corresponding to one item where all nodes referring to the same item are linked together.

- Firstly, an FP-tree is generated where the FP-tree represents the frequency and relationships of frequent itemsets from the input database

- The inputs are scanned from the database to find the frequency of each individual item. This way headers are formed which contain tables that show the links to all occurrences of all items

- Sub databases for each specific item are created this way from the database which forms the conditional pattern base for each frequent item. Each sub-database contains only the transactions of a specific frequent item.

- Using the sub-databases, the FP-Growth algorithm recursively constructs a conditional FP-tree for each frequent item. Then, the conditional FP-trees are linked to the main FP-tree based on the shared frequent item in their paths

- The FP-growth algorithm avoids generating candidate itemsets with the help of the main FP tree and conditional FP trees unlike the Apriori algorithm

- After fully completing the FP tree, the FP-growth algorithm recursively mines frequent itemsets by exploring the tree structure. Then frequent itemsets are generated by combining the suffix patterns (conditional FP trees) with the prefix paths in the main FP tree.

- When there are no more frequent itemsets to be found, the algorithm finishes the process.

3.4 A/B testing

A/B testing is a very highly used approach that is used by many modern-day web-tech companies for development and data-driven approach.[32] The A/B testing framework consists of two steps: sampling and estimation. In the sampling steps who gets which treatment is determined and in estimation, a framework is provided to compute the average treatment effect.

- The first objective of A/B testing is to determine the goal and specific hypotheses upon which we will evaluate.

- The next step is to gather samples, by dividing the users into control and experimental groups and selecting elements to test.

- After gathering the data fulfilling these criteria, appropriate statistical methods are applied to compare the control and experimental groups.

- By analyzing the test results and findings of the statistical tests, a decision is made whether to accept or reject the initial hypotheses.

- Finally, proper changes are made based on the decision and again the same approach is used for similar or different hypotheses as A/B testing is an iterative process.

3.5   Comparison of Different Association Algorithms

The Apriori algorithm is one of the simplest and most well-known association algorithms which helps to determine co-relation and find frequent itemsets in scenarios where k itemsets generate k+1 itemsets. [28] Finding common itemsets is one of the most important factors in the field of data mining and apriori algorithm is very well established in fulfilling this purpose. [14] When it comes to creating association relationships between items in the field of data mining there are several association algorithms that we can mention for comparison. For example, the F-P growth algorithm, Eclat algorithm, Fuzzy association rule mining, RuleGrowth algorithm, multi-level association rule mining, etc. The Apriori algorithm has its own strengths and weaknesses compared to these other association algorithms.

- Apriori and F-P Growth: F-P growth algorithm is primarily based on a prefix tree representation of the database it is given which forms an FP-tree of the transactions. [11] FP-growth has a similar approach towards item set mining to the Apriori algorithm. In the FP-growth algorithm, the frequencies of the items (support of single element item) are determined first and then all item sets that appear less times than the user set threshold are discarded.

  The FP growth algorithm is generally faster than the Apriori algorithm in most cases. [52] FP growth algorithm solves the problem of Apriori where each iteration generates a large number of candidate frequent item sets and requires a scan of the dataset which may become problematic if the data size increases[52].

  However, in our observation, the Apriori algorithm and FP growth algorithm both produced similar associations for our dataset. With the help of a threshold system and itemsets of different marks gained by the students in different subjects, generating association by Apriori and FP growth algorithm produced ideal outputs.

- Equivalence Class Transformation (Eclat) : For creating the association, the Eclat algorithm represents a set of transactions as a bit matrix and intersects rows to give the support of item sets following a depth-first traversal of a prefix tree. [25] Eclat's out might not be as interpretable as Apriori since it directly mines frequent itemsets without generating candidate itemsets explicitly. For closed itemsets with efficient filtering (in our case threshold filtering), the

Apriori algorithm gains a clear edge over the Eclat algorithm[8]. In datasets with a small number of transactions, Eclat will not offer a significant advantage over the Apriori algorithm

- Fuzzy Association : The fuzzy association rule is also one of the known algorithms in the field of data mining but is much used as Apriori or FP-growth algorithm. In fuzzy association, determining the degree of membership that each leaf attribute belongs to of its previous parent. [7] Then using these membership degrees, we try to find frequent itemsets called frequent itemsets using the summation of count values that are greater than min-support x |T|. [7] Then the frequent itemsets are used to generate associations whose degrees of confidence are greater than user-specified min-confidence.

  Therefore, the fuzzy association rule deals with uncertain or fuzzy data where items have degrees of membership to itemsets rather than being strictly present or absent. However, in our case, we give every subject the same degree and itemsets the same confidence initially as every subject can have the same chance of being a weakness for weakness in another subject. Thus, association algorithms such as Apriori and FP-growth will be more helpful compared to fuzzy association in this regard

- RuleGrowth : The RuleGrowth algorithm depends on a pattern-growth approach instead of a candidate-and-test approach for sequential pattern mining in datasets. [22] Firstly, it finds rules between two items and then grows them recursively by scanning the database for single items that could expand left (for processes) or right (for expansions).[22]

  However, the RuleGrowth algorithm will not be suitable for determining weaknesses for a specific subject. Sequential patterns often have temporal dependencies, where the order of items matters and sometimes makes it difficult to distinguish meaningful patterns from random occurrences. RuleGrowth takes as parameters a sequence database and the min-sup and minconf thresholds. [22] Therefore, Apriori and FP-growth algorithms will not face this problem and would be more applicable in this scenario.

- Maximal sequential algorithms : Maximal sequential algorithms are primarily used for solving problems where patterns are too long which may generate an exponential number of results. [12] Maximal sequential algorithms (MSA) focus on finding the longest sequential patterns that appear in a set of sequences.

  However, for our scenario where we are required to find the co-relations between subjects and find their weaknesses, finding the longest sequential patterns will not contribute much compared to other algorithms. On the other hand, FP-Growth is an algorithm used for mining frequent itemsets from databases which is ideal for determining the weakness and co-relations between subjects that students encounter.

Chapter 4

Dataset

We have created a totally new data set. Since we wish to test the algorithms and their effectiveness with real-world data, it was imperative to generate a dataset from online and hybrid learning classes.

4.1 Ethical Considerations

Since this paper is dealing with OL, we set out to find student data from OL spheres. During the Covid-19 pandemic, almost all schools and educational institutions shifted to OL. Hence, there should be an abundance of data from the years 2019 to 2021.

It is also important to understand that information about an institution's students is considered sensitive and confidential data is not usually publicly available. Instructors and learners may also feel uncomfortable sharing personal data such as their names and emails with others.

Keeping these obstacles in mind, we approached instructors with letters from our supervisor to ensure that both learners and instructors were presented with the assurance of the fact this information would not be revealed to the public and would strictly be used for research purposes. The letter also clearly stated that no personal information of the learners or instructors would be recorded or used during the research.

This letter was then submitted to various educational institutes. Two after-school coaching centers responded positively and agreed to provide data from their time teaching online and during hybrid classes.

4.2 Description

The data was presented in an Excel format, with each column representing a specific topic and each row representing a specific student.

In both cases, the data was obtained from physics instructors and learners from grade 10.

16 topics of physics are present alongside 4 mock exams. Each student was assigned a unique ID to be able to identify them without the need for their names.

In total, there are 399 unique students whose data we have collected. All of the marks in the dataset have been converted to be out of 100 for the sake of consistency by

the centers and missing data (as in students who had not appeared for the majority of the test) was not collected and hence, not sent from the centers themselves. The dataset showed students with mean marks of 69.547 and a standard deviation of 20.5. This indicates a limitation of the dataset, where most students did not receive very high marks, and the data itself is quite spread out.

	Id	Motion - H	Units - Re	Units and	Measuren	Displacem	Measuren	Motion Ec	Hookes La	Motion Ec	Moments	Forces and
2	123abc	12	52	88	39	41	43	58	93	4	48	28
3	456def	78	83	24	18	90	18	15	13	58	8	47
4	789ghi	63	98	94	45	57	9	89	67	65	31	8
5	101jkl	67	92	58	15	16	5	26	98	55	32	81
6	112mno	93	97	15	22	50	10	40	10	38	8	96
7	123pqr	80	52	84	90	69	98	82	31	12	65	12
8	134stu	15	100	86	7	55	92	13	14	30	39	66
9	145vwx	8	2	47	39	2	55	33	27	31	53	52
10	156yz	15	97	85	37	55	38	9	97	22	34	51
11	167abc	45	100	31	66	76	52	51	80	36	78	98
12	178def	36	25	18	57	66	41	50	5	82	59	10
13	189ghi	58	91	25	34	29	31	15	11	36	65	77
14	190jkl	96	97	81	53	10	97	39	78	18	72	80
15	201mno	26	31	68	77	87	51	69	10	55	44	47
16	212pqr	97	59	35	31	42	82	83	29	10	33	20

Figure 4.1: Dataset snippet

Chapter 5

Methodology

5.1 Data Collection

As mentioned previously we are to be working with OL data and hence could not collect data from traditional in-person classes.

Hence, we determined the most effective method for obtaining data was through purposive sampling.

2 after school coaching institutions agreed to provide their data from when they had shifted to OL during the pandemic period.

The data was presented in 45 and 75 student batches. The batches were combined into one single dataset containing a total of 399 entries. The data was collected in two phases. Once during January 17th, 2023, and again on 12th February 2023.

5.2 Data Pre-Processing

Within the data set 20 columns were present. The first column represented a unique student ID as a "key" identifier, to maintain the learner's privacy.

The rest of the columns' headers were the specific topics on which they were tested. Hence, each row represented a unique student.

Any row with missing data was dropped and no imputation was done. We believed that there was no way to accurately predict a student's performance on a specific topic without introducing some form of bias.

We have also opted to drop the columns containing marks for their mock exam of paper 6. According to Cambridge IGCSE, paper 6 is testing the candidates on a different set of skills, quite different from the theoretical concepts on papers 4 and 2. Beyond these, no other pre-processing was deemed necessary.

5.3 Exipermental Test-Bed

Initial testing of the association algorithm, specifically the Apriori algorithm, was done using the cloud-based development environment "Google Colaboratory". The environment was given 0.8 gigabytes of RAM and 26.3 gigabytes of solid-state drive space.

Since this is based on cloud technology and is based around the Python language, it was chosen. Python has a vast amount of libraries and resources, dedicated to

unsupervised algorithms and working with datasets and data frames.

To ensure the algorithm was producing the expected output a snippet of the original collected data was used for testing. We only utilized 7 students and 7 subjects to apply the algorithm. The code was iterated until the expected output of frequent itemsets was produced.

Afterward, Jupyter Notebook running the Python kernel on our personal computers was utilized to run the algorithm that we have just tested online. This time, the entire dataset was utilized. The local machine had the following hardware specifications :

- 16 Gigabytes of Random Access Memory.

- AMD Ryzen5 5600G Graphical Processing Unit.

- 500 Gigabytes Solid-State Drive Memory.

The clustering algorithm was applied using the same methods.

5.4   Applied Methodology

We have utilized the pandas and sk. learn libraries in Python 3.10.8. The dataset was stored locally as a .CSV file and turned into a Python data frame using the pd.read_csv function.

The next step was to take the data frame and convert it into transaction data, so that we may apply the algorithm to them.

All the column headings were stored in a list, alongside an empty list of transaction data. We took the unique identifier, (in this case the student ids) and iterated over every column for each unique student.

While iterating, we only chose students who had marks below 60 (This was taken as the threshold for passing a topic) These chosen marks were then appended to the initially empty transaction data list. Then the transactions are fitted onto the transaction encoder essentially takes the transactions and assigns a unique binary value to each unique transaction. It is placed into a binary matrix, where each row is a transaction and each column is an item. This binary matrix is then transformed and stored as a data frame. Visual Studio Code was used to compile and run the algorithms, taking up 984 megabytes of memory with 5% Central Processing unit usage. This new data frame containing the transactions is then passed as the parameter onto the Apriori Algorithm, which produces frequent item sets. These item sets are generated based on conditional probability, with a minimum support threshold of 50%. This method was repeated to apply the F-P Growth algorithm as well.

A similar approach was utilized when applying the K-means clustering algorithm, by loading the. CSV dataset into a Python data frame. Then, we took two subjects that our association algorithm had provided to be frequent itemsets and selected those columns only. So that we can gather some useful information from the clustering, we selected only the passing values from one subject and failing marks from another. This new filtered data was fitted onto the K-Means method and plotted to give us a clustering graph. Each point on the clustering graph represents a unique student. The number of clusters was determined by the "Elbow Method" where the within-cluster sum of squares (WCSS) was plotted against the number of clusters. Since
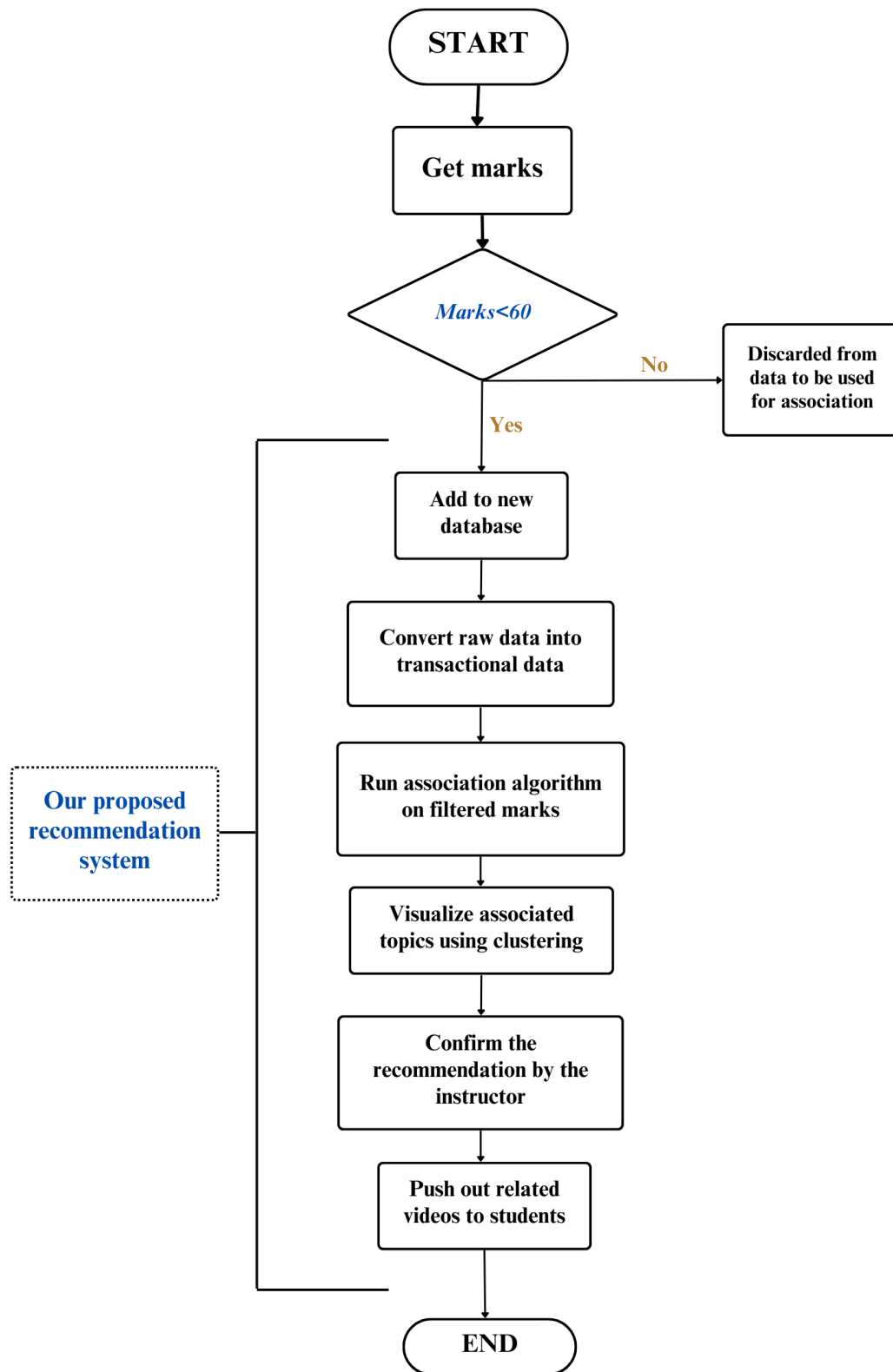
Figure 5.1: Content provider flow diagram

the graph did not have a clear "elbow", we also applied the "Silhouette" score graph to uncover the optimal number of clusters. A higher score is optimal, hence the number of clusters that produce the global maxima (in our it is 6) is to be chosen. We selected the value where the Silhouette score value seems to be at a global maximum, in our case this being 6.
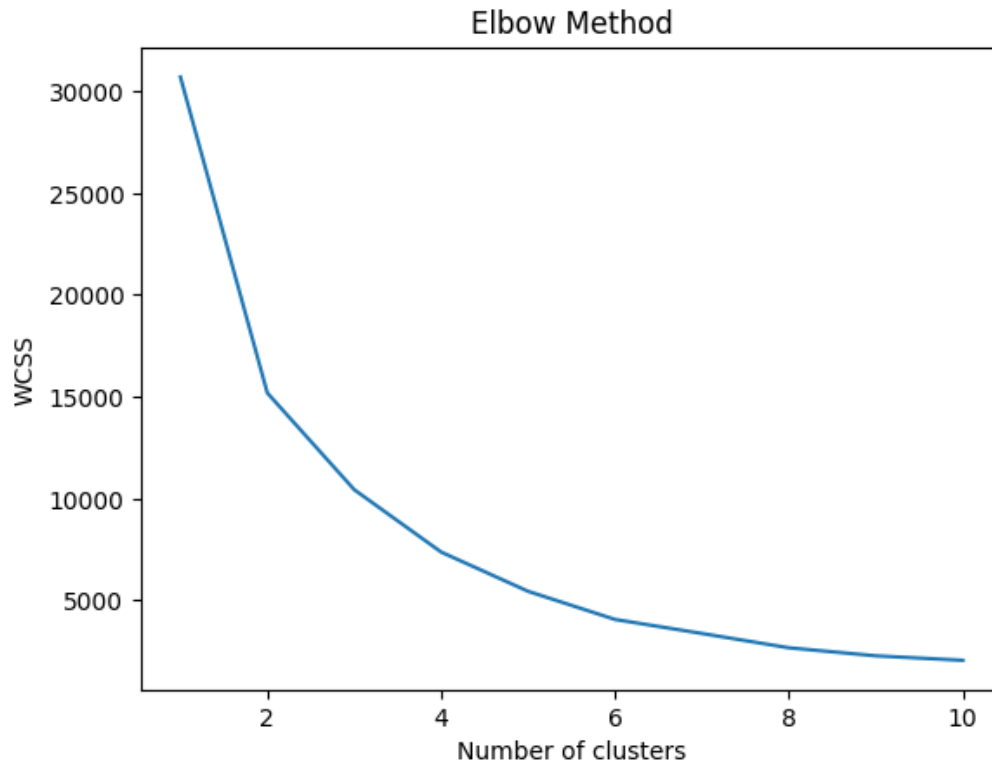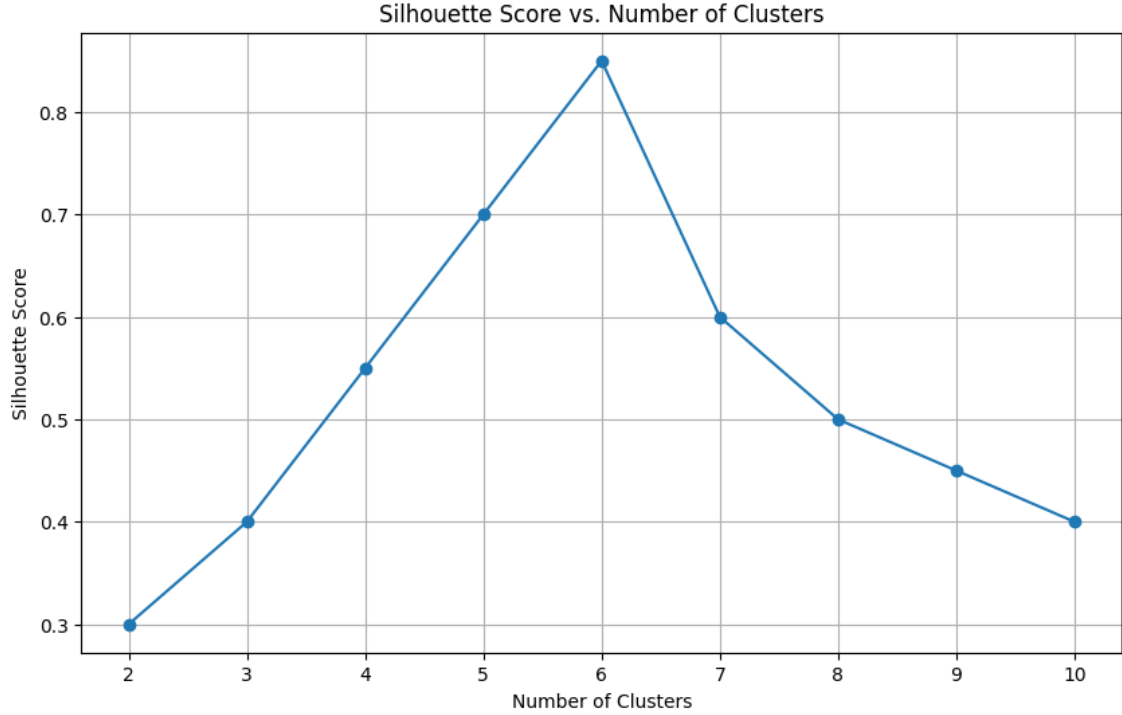
Figure 5.2: Elbow method graph

Figure 5.3: Silhouette score vs. optimal cluster graph

Additionally, Using relevant generated output two high-fidelity, horizontal web application prototypes were generated [3]. These are categorized as such :

- Instructor Side: The instructor's portal allowed for the upload of a . CSV file database as showcased in this paper beforehand. The web application would take the dataset and generate a frequent itemset using the properties and constraints already described in this paper. This item is presented visually as a table, along with the support of each association. Furthermore, the clustering graphs from the associations are also shown. Two K-means clustering graphs are used to graphically depict the association analysis. We can better grasp the correlation between students' performance across several topics thanks to these graphs. In the first graph, we compare students who have scored over the cutoff in one subject to those who have fallen short in a different topic(as shown in figure 5.5 captioned "Visualisation for instructors 1"). This enables us to determine whether the performance in these two subjects is related. We can more easily spot trends and patterns thanks to the clustering algorithm, which pairs up students with similar performance patterns. The students whose scores fell short of the cutoff in both subjects are the topic of the second graph (As shown in the figure captioned 5.6 "Visualisation for instructors 2"). We can identify the pupils who struggle in both disciplines by grouping this subset of individuals. We can also visually see the number of cluster points on the second graph to be far larger, supporting our generated association. Instructors need this information to understand whether the number of students who are having difficulties in both topics is greater or smaller. These visualizations give teachers insightful information about how their pupils are performing on certain topics. They can modify their teaching methods based on this knowledge and offer more assistance to individuals who
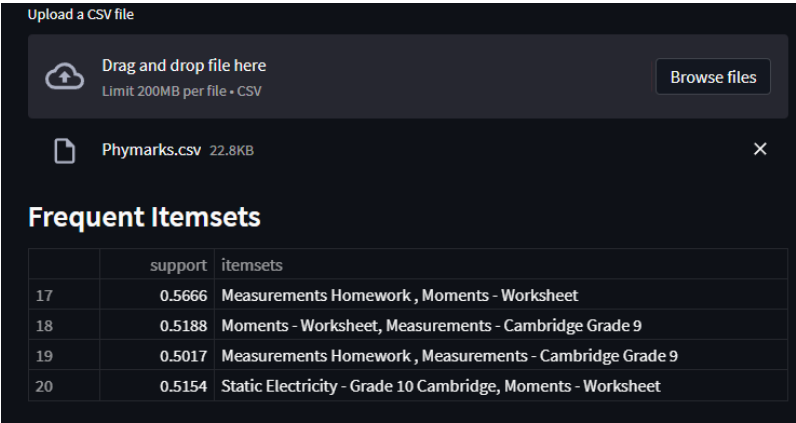
need it the most.

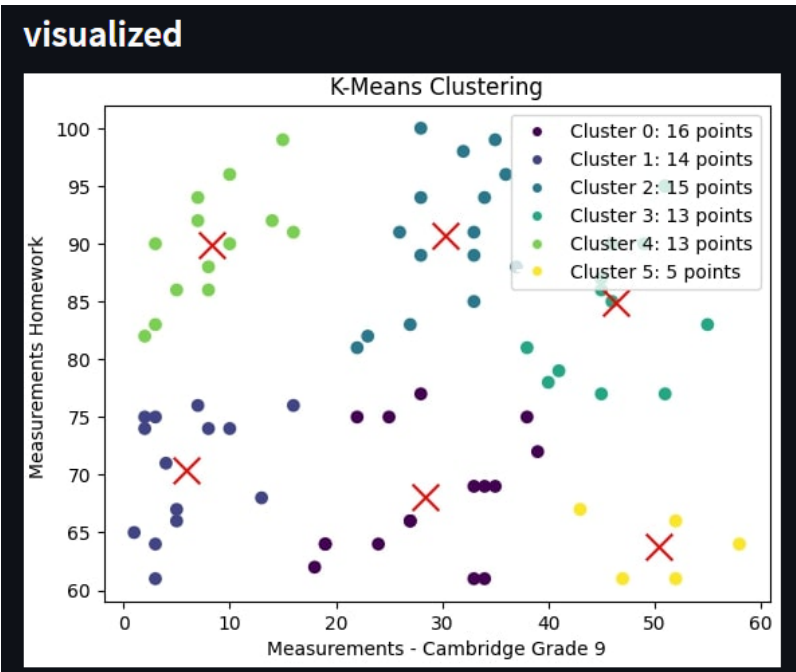Figure 5.4: Instructor portal 1

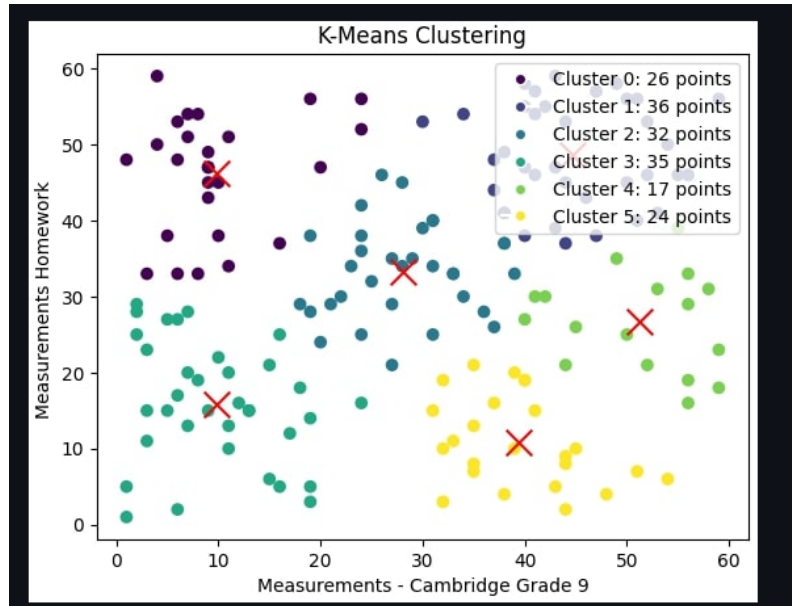Figure 5.5: Visualization for instructors 1

Figure 5.6: Visualisation for instructors 2

- Student Side : Based on the output of our algorithm, it can be seen that there's an association between Measurements Homework and Measurements - Cambridge Grade 9 exam, Moments – Worksheet and Measurements - Cambridge Grade 9 exam, Moments – Worksheet and Measurements Homework, and lastly between Moments – Worksheet and Static Electricity - Grade 10 Cambridge exam, all of these will be discussed further into this paper. We are utilizing our database of videos and are not presenting any new algorithms. So, if a student does well (marks greater than 60) in all subjects, there is no suggestion for him as he did well in all subjects. Our algorithm only targets the students who got lower than 60 in any subject Secondly, if a student's marks are below 60% in a topic that has a known association with another, the system suggests a video for each of the associated topics. These recommended videos can be taken in from 2 sources.

    – Platform's own database : in this method, if the recommendation system has already been applied to a OL platform , they can suggest the relevant videos from their own video bank

    – Key-word search : Alternatively, we can search the keyword of a specific topic on publically available video sharing sites such as www.youtube.com.

Figure 5.7: Student portal example 1

Figure 5.8: Student portal example 2

Finally, to evaluate the effectiveness of our proposed content provider module 2 different methods were used

- Qualitative Analysis: 10 Instructors were shown the instructor side prototype and asked a series of questions. Their answers were further analyzed to gain insight into their opinions and concerns regarding the proposed module. The Questions asked during their interview are summarised below. Please be noted that the Linkert scale used was a 5-point scale, namely :

  Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree

Open-ended Questions:
* How could the e-learning module improve your ability to deliver course content effectively? * In what aspects of your pedagogy would you find the e-learning module most beneficial?

Linkert Scale Questions:
* The E-Learning Module has Provided Opportunities for New Instruction. * The e-learning module can influence my assessment and feedback strategies. * The e-learning module can improve my ability to provide timely and constructive feedback to students. (Follow up: If you have answered above a 3 in the previous question, can you explain a bit how?)

Table 5.1: Instructor Feedback Questions

- Quantitative Analysis: 70 volunteer students from high school were recruited and A/B testing was performed. In this case, Version A was defined as our novel module, whereas Version B is the more traditional approach. The test was separated into two phases:

  - In the first phase of the test, all 70 students were given a test. The syllabus of the test was one of the subjects our module had found associations with, namely: "Moment of a Force". The students who had gotten below the 60% threshold were then further requested to participate in Phase 2.

  - In Phase 2, 40 students had been filtered as per the threshold. 10 students were asked to revise using Version A, the rest using Version B. After revisions, another test on the same topic was given to all 40 students. The results along with conversion ratios are discussed further in this paper.

  - conversion rate:

  $$\text{Conversion Rate} = \frac{\text{Number of Conversions}}{\text{Number of Visitors}} \times 100\%$$

  was also calculated for each version, where the Number of Conversions was set to be the number of students who scored above the threshold, and the Number of Visitors was set to be the total number of students per version, to better understand the comparison.

Chapter 6

Experimental Evaluation

6.1 Experimental Findings

The Apriori algorithm gave an output of frequent itemsets, based on the following metrics (as discussed before) :

- Transaction data only contained marks below 60

- Minimum support was set at 50%

- Minimum confidence was set at 65%

The F-P Growth algorithm was run with the same parameters and produced the same results. Confidence measures how often a specific association rule is true. It quantifies the likelihood that when one set of items is purchased, another set of items is also purchased. Whereas Support in association rule mining measures how often an itemset (a combination of items) appears in the dataset. It quantifies the frequency of occurrence of that itemset among all transactions. High support indicates that the itemset is common in the dataset, while low support suggests it is infrequent

There was a significant run-time difference between the two association algoriths with the F-P growth algorithm having a 3.7 times faster runtime, as portrayed below.

Algorithm	Runtime (seconds)
Apriori	0.06801557
F-P Growth	0.018004179

Table 6.1: Comparison of Algorithm Runtimes

if we analyze the results, we can decipher some interesting results. To be noted , items are separated using a comma (,) and not a hyphen (-). Firstly, single transaction itemsets are the majority of the output, meaning the poor results in these topics were usually not associated with or dependent on other topics. If we look at this result in terms of pedagogy, we can see [40] that reasons for poor performance in standardized testing can be multi-faceted and multi-factored and not always directly linked to other topics of the same subject. Of the association with multiple items, there does seem to be a relation between the topics, which can be seen as valuable

```
     support                                                               itemsets
0   0.645051  (Displacement - Construction )
1   0.655290  (Electrical Quantities Exam)
2   0.679181  (Forces and Moments - Cambridge)
3   0.587031  (Hookes Law - Graph)
4   0.682594  (Measurements - Cambridge Grade 9)
5   0.709898  (Measurements Homework )
6   0.593857  (Mock 1 - P4)
7   0.627986  (Mock 2 - P4)
8   0.723549  (Moments - Worksheet)
9   0.655290  (Motion - Homework)
10  0.627986  (Motion Equations and Graphs)
11  0.648464  (Motion Equations and Graphs - Exam)
12  0.638225  (Sound and Electromagnetism)
13  0.675768  (Static Electricity - Grade 10 Cambridge)
14  0.621160  (Thermal Physics Assignment)
15  0.627986  (Units - Recap Quiz)
16  0.645051  (Units and Prefixes - Grade 9 - Quiz)
17  0.501706  (Measurements Homework , Measurements - Cambridge Grade 9)
18  0.518771  (Moments - Worksheet, Measurements - Cambridge Grade 9)
19  0.566553  (Moments - Worksheet, Measurements Homework )
20  0.515358  (Moments - Worksheet, Static Electricity - Grade 10 Cambridge)
```

Figure 6.1: All generated associations from our dataset

information extracted from the data set, so that instructors are better aware of this link and can focus on ensuring learners also see this connection. The association rules with multiple items are shown in the table below :

Association Rules	Support	Confidence
Measurements Homework and Measurements Grade 9 Test	0.50	0.79
Moments Worksheet and Measurements Grade 9 Test	0.51	0.76
Moments Worksheet and Measurements Homework	0.56	0.74
Moments Worksheet and Static Electricity	0.52	0.76

Table 6.2: Association rules with multiple items, support values, and confidence values

We can see that 'Measurements Homework' and 'Measurements - Cambridge Grade 9' are frequent items in terms of below-par marks. If a learner is not performing well on their homework, they are unlikely to perform well on their class test.
'Moments - Worksheet' seems to be associated with 'Measurements Homework' and 'Measurements - Cambridge Grade 9'. Instructors can look deeper into the syllabus of both these topics to find out further correlations and how to instruct learners better. As for the learners, they can know to focus on these topics to do well in all three or vice-versa.
'Static Electricity - Grade 10 Cambridge' seems to have low marks associated with 'Moments - Worksheet', which we have not been able to find out as to how they are related, in terms of actual syllabus content, beyond the fact that both topics require some mathematical calculation and application of formulae.

When applying clustering, a unique problem emerged. As mentioned earlier we chose subjects according to the frequent itemsets that the Apriori Algorithm had provided. If we were to just apply clustering on all 399 rows of marks from the two columns, there are too many data points to provide a meaningful graph to extract any meaningful information
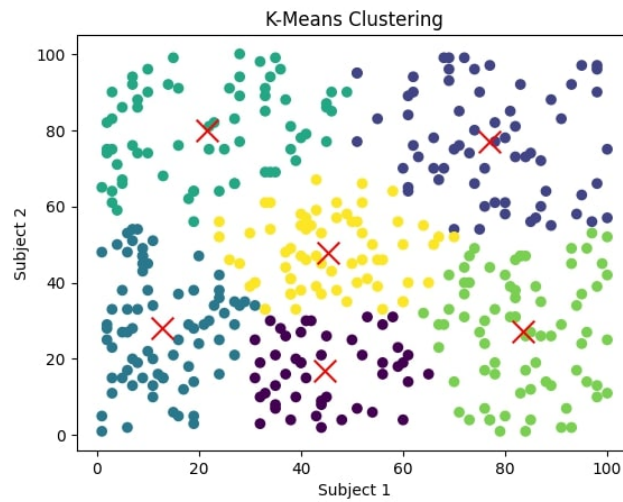

Figure 6.2: All 399 points for 2 subjects

Hence, to combat this issue we plotted clustering according to only the failing marks from the two particular columns. This again did not reveal too much useful information, as it just showed students who had failed in both topics and thus we could not draw any meaningful conclusions or information. However, if we plotted the clustering where we take the marks of one subject where students had failed against the other subject where students had passed, we could draw some useful information. Here we can see clusters 0 and 2 have the most number of students, meaning
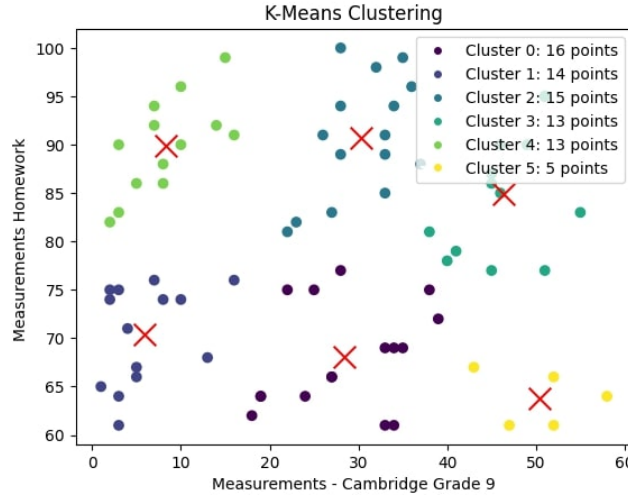

Figure 6.3: Filtered data clustering 1

that students who are barely above the passing grades for 'Measurements Homework' are doing very poorly in the 'Measurements - Cambridge Grade 9' test. This information was not picked up by the association algorithm and may be very useful information during the pedagogy design or design of content providers. We can also compare this graph with the K means Clustering graph where we take filtered marks where students had failed both subjects. This gives us more visual confirmation of the association.
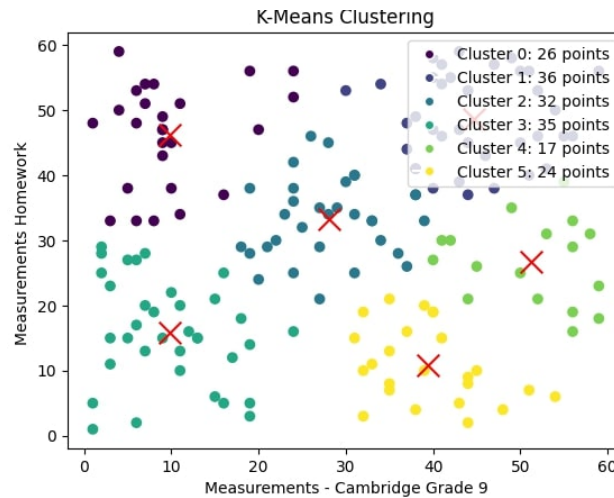

Figure 6.4: Filtered data clustering 2

6.2 Results from A/B testing

Our proposed module needed to be reviewed by actual students so that we can analyze the fruitfulness of our research and website. Hence, to quantitatively analyze the proposed module we opted for A/B testing. As mentioned before, the test was divided into 2 phases. During phase 1, 70 volunteer high-school students (From grades 10-11) were asked to sit for a Physics test on the topic "Moment of a Force". This specific topic was chosen as our historical data suggested an association between the "Moment of a Force" and "Measurements". 55% of the students failed to meet the threshold in the quiz considering a pass was above 60% marks.

In phase 2, 20 students were shown 1 video (This can be considered Version A), whereas 20 students were shown 2 videos (This can be considered Version B). Version A showed a conversion rate of 20% whereas Version B showed a conversion rate of 70%. This shows a big difference in student performance using version B and points to it being the superior method of revision for students. These results are summarised in the figure below:

Total Students = 40	Number of Students	
	Passed	Failed
Traditional Method (One video)	4	16
Our Proposed Method (Two videos)	14	6

Table 6.3: Comparison of pass and fail statistics

6.3 Analysis of Interview

15 instructors were interviewed to ascertain their views and opinions on the module and specifically the instructor portal. Most instructors requested to keep personal information anonymous, hence their names are not presented in this paper. They revealed having the correlations of topics that pertain to student performance can shift how they approach those topics from the beginning. Shifting their pedagogy to address any overlap of syllabus content to aid students. 8 of the 15 participants also revealed that they would give extra emphasis on related topics so that students have a stronger foundation moving forward. Dr. Navid Rahman(MBBS and PGT), an instructor of high-school chemistry for 14 years said,

> "I can get the data from beforehand about link of topics and be ready to give feedback when they do bad."

Additionally, Dr.Dewan Chowdhury a high-school physics instructor for over 10 years commented,

> "I would make plans in my lesson plans knowing these links exist, so I can show them what the overlaps are in topic content."

Almost all instructors agreed that knowing links between topics can help them prepare their pedagogy to know which topics lay a good foundation of understanding for students.

3 open-ended questions and 3 Linkert scale questions were presented to them. As mentioned before, the Linkert scale used was a 5-point scale, namely :
Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree Here we can see a general trend of positive responses deeming the module to be helpful in pedagogical decision-making. The responses to the Linkert scale questions are given below :
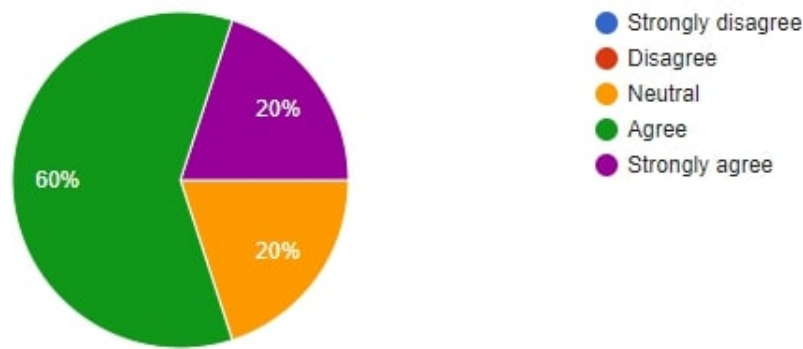

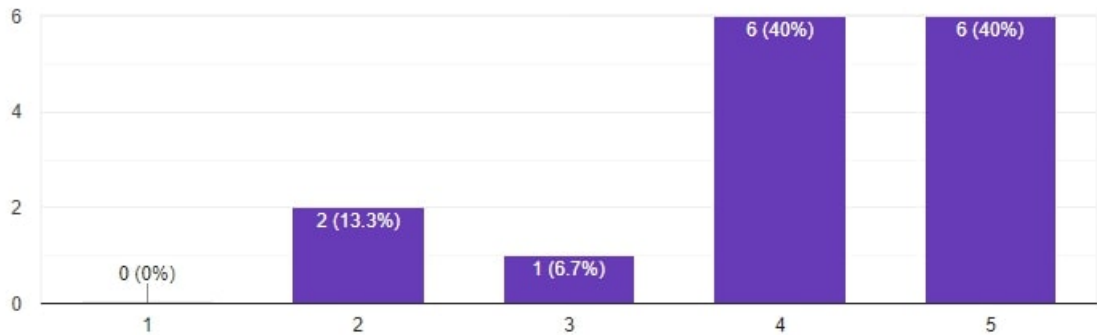Figure 6.5: Responses to the e-learning module has provided opportunities new instructions

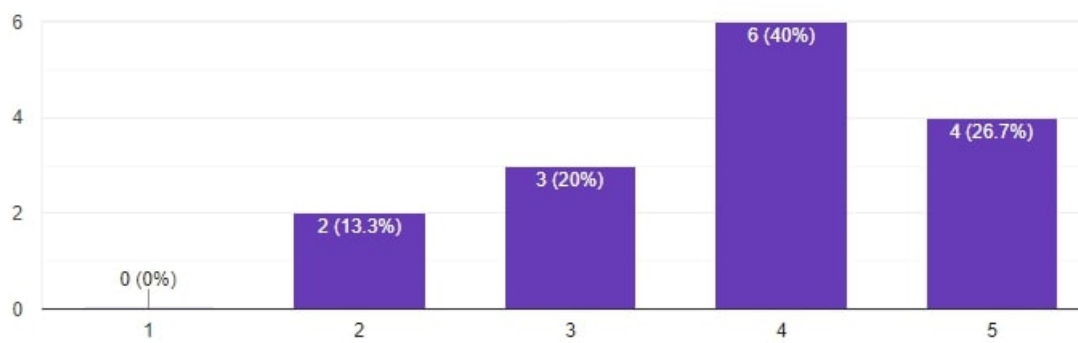Figure 6.6: Responses to the e-Learning module can influence my assessment and feedback strategies.

Figure 6.7: Responses to The e-Learning module can improve my ability to provide timely and constructive feedback to students.

Chapter 7

Conclusion

Through the provision of individualized feedback and improved learning resources, the use of machine learning algorithms has the potential to completely transform online learning. During the Covid-19 epidemic, real-world data from physics professors and students in grade 10 were collected for this project to establish a new dataset. The dataset included data in Excel format for 16 physics topics and 4 practice examinations. To ensure student anonymity, each student was given their own ID. By obtaining permission from educational institutions and ensuring that both students and instructors were given the reassurance that their personal information would not be presented to the public and would strictly be used for research purposes, the study was able to successfully address ethical considerations related to sensitive and confidential data. Data pre-processing included removing any rows with incomplete information and eliminating columns with mock test paper 6 marks. The applied methodology includes transforming the data frame into transaction data and employing the Apriori and F-P Growth algorithm with a minimum threshold of 60% to produce frequent item sets. The experimental results revealed that although single-item transactions made up the majority of the output, there were some intriguing correlations between themes that could be helpful to teachers and students. Examples of items that frequently received subpar grades are "Measurements Homework" and "Measurements - Cambridge Grade 9," showing that if a student is struggling with their homework, they are unlikely to succeed on their test in class. To discover more connections and understand how to train students more effectively, instructors might delve further into the syllabuses for both of these subjects. Additionally, 'Moments - Worksheet' appears to be connected to 'Measurements Homework' and 'Measurements - Cambridge Grade 9', suggesting that students should concentrate on these subjects to succeed in all three. The fact that 'Static Electricity - Grade 10 Cambridge' appears to have poor marks related to 'Moments - Worksheet' may also shed light on the syllabus's content and suggest ways to teach these subjects more effectively. The study provided insights into correlations between themes and potential causes for subpar performance on standardized tests, proving the efficacy of applying machine learning algorithms like the Apriori algorithm to enhance online learning. According to the results, learners could receive tailored feedback and improved learning materials based on their particular needs and learning preferences, enhancing the general standard of online education.

The study's findings emphasize the value of gathering real-world data when assessing association and clustering algorithms for online learning. Additionally, it shows that

machine learning algorithms may be able to shed light on the relationships between various subject areas and possible causes of subpar performance on standardized tests. To further enhance the pedagogy of online learning models, future studies might investigate the usage of additional unsupervised algorithms such as hierarchical clustering and CNN architecture. Overall, the study offers a solid framework for future research on the application of machine learning algorithms to transform online education and raise educational standards generally.

These generated data were then utilized to create a prototype implementation for both instructors and students. These prototypes were evaluated using qualitative and quantitative (A/B testing) respectively.

For our approach to be useful, this must be applied to a curriculum in which subjects have at least some correlations with each other. Otherwise, without interdependence, a student will not be able to improve by a significant margin as finding the root cause and improving on the detected subject will not be able to improve their performance in other subjects. Furthermore, it will be helpful to be able to group individuals by clustering which will also help us detect similar behaviors among them. [2] Moreover, by using clustering methods we will be able to achieve more useful visualization of our results. [3]

7.1 Future Works

While the methodology presented in this study for using data mining techniques to create a recommendation system module for Online Learning (OL) platforms is promising, there are still a number of areas that may use additional research and development to make the system more workable and scalable.

- Integration with Online Learning Platforms: The smooth integration of the suggested recommendation system module into fully functional online learning platforms is the following research agenda phase. For a variety of pupils, this connection would make it easier to analyze data in real time and provide individualized recommendations. To achieve a seamless and successful integration process, partnerships with educational technology companies and institutions are required.

- Extensive Testing and Evaluation: Thorough testing and evaluation on larger datasets are crucial to proving the usefulness and dependability of the recommendation system. It will be possible to learn more about the system's flexibility and generalizability by conducting experiments with a wide student population and a range of courses. To evaluate the effect of tailored recommendations on student engagement and performance, long-term research should be carried out.

- Pedagogical Impact Studies : Future work will focus heavily on integrating the recommendation system into institutions' and instructors' educational procedures. Studies should look into how teachers might use the system's findings to improve their instruction and modify their course materials to better suit the needs of certain students.

In conclusion, the upcoming projects listed above serve as a thorough roadmap for integrating and improving the suggested recommendation system for online learning

platforms. By focusing on these crucial areas, we can open the door for more efficient and individualized online learning experiences, which will ultimately be advantageous for both students and teachers.

Bibliography

[1] J. H. Block and R. B. Burns, "Mastery learning," *Review of Research in Education*, vol. 4, pp. 3–49, 1976. DOI: 10.2307/1167112.

[2] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979. DOI: 10.2307/2346830.

[3] R. Budde, K. Kautz, K. Kuhlenkamp, and H. Züllighoven, "What is prototyping?" *Information Technology People*, vol. 6, no. 2/3, pp. 89–95, 1990. DOI: 10.1108/eum0000000003546.

[4] H. Ishii, M. Kobayashi, and K. Arita, "Iterative design of seamless collaboration media," *Communications of the ACM*, vol. 37, no. 8, pp. 83–97, 1994.

[5] D. Jaffee, "Asynchronous learning: Technology and pedagogical strategy in a distance learning course," *Teaching Sociology*, vol. 25, no. 4, pp. 262–277, 1997. DOI: 10.2307/1319295.

[6] A. Molnar, "Computers in education: A brief history," *The Journal*, Jun. 1997. [Online]. Available: https://thejournal.com/articles/1997/06/01/computers-in-education-a-brief-history.aspx.

[7] G. Chen, Q. Wei, and E. E. Kerre, "Fuzzy data mining: Discovery of fuzzy generalized association rules+," *Recent Issues on Fuzzy Databases*, pp. 45–66, 2000.

[8] C. Borgelt, "Efficient implementations of apriori and eclat," in *FIMI'03: Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, vol. 90, Nov. 2003.

[9] S. Carliner, "An overview of online learning (2nd ed.)," *European Business Review*, vol. 16, Aug. 2004. DOI: 10.1108/09555340410561723.

[10] C. Borgelt, "An implementation of the fp-growth algorithm," in *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, ser. OSDM '05, Chicago, Illinois: Association for Computing Machinery, 2005, pp. 1–5, ISBN: 1595932100. DOI: 10.1145/1133905.1133907. [Online]. Available: https://doi.org/10.1145/1133905.1133907.

[11] ——, "An implementation of the FP-growth algorithm," in *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, Aug. 2005, pp. 1–5.

[12] E. Z. Guan, X. Y. Chang, Z. Wang, and C. G. Zhou, "Mining maximal sequential patterns," in *2005 International Conference on Neural Networks and Brain*, vol. 1, IEEE, Oct. 2005, pp. 525–528.

[13]   T. Sitzmann, K. Kraiger, D. Stewart, and R. Wisher, "The comparative effectiveness of web-based and classroom instruction: A meta-analysis," *Journal of Educational Psychology*, vol. 98, no. 4, pp. 832–845, 2006. DOI: 10.1037/0022-0663.98.4.832.

[14]   Y. Ye and C.-C. Chiang, "A parallel apriori algorithm for frequent itemsets mining," in *Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06)*, IEEE, 2006.

[15]   I. E. Allen and J. Seaman, *Online nation: Five years of growth in online learning.* 2007.

[16]   J. Newman, "Using voip technology for online courses in higher education," in *Proceedings of SITE 2007–Society for Information Technology Teacher Education International Conference*, Association for the Advancement of Computing in Education (AACE), San Antonio, Texas, USA, 2007, pp. 444–447. DOI: 10.24059/olj.v12i3.311. [Online]. Available: https://www.learntechlib.org/primary/p/24578/.

[17]   M. Prensky, "The role of technology," *Educational Technology*, vol. 48, no. 6, pp. 1–3, 2008. [Online]. Available: http://www.marcprensky.com/writing/Prensky-Khan%20Academy-EdTech-Jul-Aug2011.pdf.

[18]   G. J. Hwang, P. H. Wu, C. C. Chen, and S. H. Huang, "Effects of a peer-assessment strategy on online collaborative learning," *Journal of Computer Assisted Learning*, vol. 25, no. 5, pp. 438–448, 2009. DOI: 10.1111/j.1365-2729.2008.00294.x.

[19]   K. Mahmud and K. Gope, "Challenges of implementing e-learning for higher education in least developed countries: A case study on bangladesh," in *2009 International Conference on Information and Multimedia Technology*, Nov. 2009, pp. 155–159. DOI: 10.1109/ICIMT.2009.27.

[20]   J. Mansbach and Y. G. Bachner, "Collaborative learning via asynchronous discussion forums: A comparison of academic writing in l2 english and l1 hebrew," *CALICO Journal*, vol. 27, no. 2, pp. 237–256, 2010. [Online]. Available: https://www.calico.org/html/article728.pdf.

[21]   L. Averell and A. Heathcote, "The form of the forgetting curve and the fate of memories," *Journal of mathematical psychology*, vol. 55, no. 1, pp. 25–35, 2011.

[22]   P. Fournier-Viger, R. Nkambou, and V. S. M. Tseng, "Rulegrowth: Mining sequential rules common to several sequences by pattern-growth," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, Mar. 2011, pp. 956–961.

[23]   D. Hunyadi, "Performance comparison of apriori and fp-growth algorithms in generating association rules," in *Proceedings of the 5th European Conference on European Computing Conference*, Apr. 2011, pp. 376–381.

[24]   C. Rensing, R. Steinmetz, B. Frey, and N. Sattes, "Adaptive e-learning offers tailored support for learning factually dense content," in *Seamless Learning in the Age of Mobile Connectivity*, Springer US, 2012, pp. 425–427. DOI: 10.1007/978-1-4614-3185-5_38.

[25] M. Girotra, K. Nagpal, S. Minocha, and N. Sharma, "Comparative survey on association rule mining algorithms," *International Journal of Computer Applications*, vol. 84, no. 10, 2013.

[26] J. Xiao, M. Wang, L. Wang, and X. Zhu, "Design and implementation of c-ilearning: A cloud-based intelligent learning system," vol. 11, no. 3, Jul. 2013. DOI: 10.4018/JDET.2013070106.

[27] P. C. Brown, H. L. R. III, and M. A. McDaniel, *Make It Stick*. Harvard University Press, 2014.

[28] M. Al-Maolegi and B. Arkok, "An improved apriori algorithm for association rules," Unpublished, 2014.

[29] F. Schwenker and E. Trentin, "Pattern classification and clustering: A review of partially supervised learning approaches," *Pattern Recognition Letters*, vol. 37, pp. 4–14, 2014. DOI: 10.1016/j.patrec.2013.10.017.

[30] H. Wang, P. Liu, and H. Li, "Application of improved association rule algorithm in the courses management," in *2014 IEEE 5th International Conference on Software Engineering and Service Science*, 2014, pp. 804–807. DOI: 10.1109/ICSESS.2014.6933688.

[31] M. F. Baris, "Future of e-learning: Perspective of european teachers," *EURASIA Journal of Mathematics, Science and Technology Education*, vol. 11, no. 2, pp. 421–429, 2015. DOI: 10.12973/EURASIA.2015.1361A.

[32] H. Gui, Y. Xu, A. Bhasin, and J. Han, "Network A/B testing: From sampling to estimation," in *Proceedings of the 24th International Conference on World Wide Web*, May 2015, pp. 399–409.

[33] N. Marangunic and A. Granic, "Technology acceptance model: A literature review from 1986 to 2013," *Universal Access in the Information Society*, vol. 14, pp. 81–95, 2015. DOI: 10.1007/s10209-014-0348-1.

[34] K. Scott, "The impact of collaborative writing technologies on student learning," *Communications in Information Literacy*, vol. 9, no. 1, pp. 43–55, 2015. [Online]. Available: https://files.eric.ed.gov/fulltext/EJ1062107.pdf.

[35] C. A. Twigg, "Improving learning and reducing costs: Fifteen years of course redesign," *Change: The Magazine of Higher Learning*, vol. 47, no. 6, pp. 6–13, 2015. DOI: 10.1080/00091383.2015.1089753.

[36] G. Cheng and J. Chau, "Exploring the relationship between learning approaches, self-regulation, and academic achievement of medical students: A structural equation modeling analysis," *Advances in Medical Education and Practice*, vol. 7, pp. 389–396, 2016. DOI: 10.2147/AMEP.S131638.

[37] L. Khanna, S. N. Singh, and M. Alam, "Educational data mining and its role in determining factors affecting students academic performance: A systematic review," in *2016 1st India International Conference on Information Processing (IICIP)*, 2016, pp. 1–7. DOI: 10.1109/IICIP.2016.7975354.

[38] M. Lister and R. E. West, "Design of e-learning and online courses: A literature analysis," in *Advanced Methodologies and Technologies in Modern Education*, Springer International Publishing, 2016, pp. 216–235. DOI: 10.1007/978-3-319-39483-1_9.

[39] F. Liu, S. Zhang, J. Ge, F. Lu, and J. Zou, "Agricultural major courses recommendation using apriori algorithm applied in china open university system," in *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1, 2016, pp. 442–446. DOI: 10.1109/ISCID.2016.1109.

[40] *Beyond Test Scores*, pp. 14–51, 2017. DOI: 10.4159/9780674981157-002.

[41] G. N. Rayasad, "Association rule mining in educational recommender systems," Unpublished, 2017.

[42] D. Kučak, V. Juričić, and G. Dambić, "Machine learning in education - a survey of current research trends," in *Annals of DAAAM Proceedings*, vol. 29, 2018.

[43] Y. Liang, Z. Liu, and X. Li, "A conceptual framework for understanding e-learning in a performance simulation context," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, 2018, p. 443. DOI: 10.1145/3178158.3178206.

[44] S. Moghavvemi, A. Sulaiman, N. I. Jaafar, and N. Kasem, "Social media as a complementary learning tool for teaching and learning: The case of youtube," *The International Journal of Management Education*, vol. 16, no. 1, pp. 37–42, 2018.

[45] S. Panjaitan, Sulindawaty, M. Amin, S. Lindawati, R. Watrianthos, H. T. Sihotang, and B. Sinaga, "Implementation of apriori algorithm for analysis of consumer purchase patterns," *Journal of Physics: Conference Series*, vol. 1255, no. 1, p. 012 057, Aug. 2019. DOI: 10.1088/1742-6596/1255/1/012057. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1255/1/012057.

[46] J. Steele, R. Holbeck, and J. Mandernach, "Defining effective online pedagogy," *Journal of Instructional Research*, vol. 8, no. 2, pp. 5–8, 2019.

[47] M. G. Uddin, A. Isaac, and Osama, "Impact of the system, information, and service quality of online learning on user satisfaction among public universities students in bangladesh," in *Proceedings of the 3rd International Conference on Advanced Information and Communication Technology (ICAICT)*. 2019, vol. 3, pp. 1–10.

[48] M. Hillyer, "How has technology changed-and changed us-in the past 20 years," World Economic Forum, Tech. Rep., 2020.

[49] W. Xu and Y. Zhou, "Course video recommendation with multimodal information in online learning platforms: A deep learning framework," vol. 51, no. 5, Sep. 2020. DOI: 10.1111/BJET.12951.

[50] P. Chakraborty, P. Mittal, M. S. Gupta, S. Yadav, and A. Arora, "Opinion of students on online education during the covid-19 pandemic," *Human Behavior and Emerging Technologies*, vol. 3, pp. 357–365, 2021. DOI: 10.1002/hbe2.240.

[51] J. A. Cohen, "A fit for purpose pedagogy: Online learning designing and teaching," *Development and Learning in Organizations*, vol. 35, no. 4, pp. 15–17, 2021. DOI: 10.1108/DLO-08-2020-0174.

[52] A. Essam, M. A. Abdel-Fattah, and L. Abdelhamid, "Towards enhancing the performance of parallel fp-growth on spark," *IEEE Access*, vol. 10, pp. 286–296, 2021.

[53] J. Lepičnik-Vodopivec and A. Šorgo, "The impact of online learning on students' motivation and self-regulated learning," in *Advances in Intelligent Systems and Computing*, vol. 1233, Springer International Publishing, 2021, pp. 102–111. DOI: 10.1007/978-3-030-88520-5_9.

[54] M. S. Rahaman, I. H. Moral, M. M. Rahman, M. Sahabuddin, and A. B. Samuel, "Online learning in bangladesh during covid-19: Perceived effectiveness, challenges, and suggestions," *Journal of Education, Management and Development Studies*, vol. 1, no. 3, pp. 35–47, 2021. DOI: 10.52631/jemds.v1i3.51.

[55] F. A. Abdurazakov and F. B. Odinaboboev, "Pedagogical importance of using module educational technologies in the system of continuous education on the basis of modern approaches," *Web of Scientist: International Scientific Research Journal*, vol. 3, no. 1, pp. 173–180, 2022. DOI: 10.17605/OSF.IO/N9KSD.

[56] Author Name. (Feb. 2022). "Forgetting curve – importance of distribution as well as quantity in general practice teaching." British Journal of General Practice, [Online]. Available: https://bjgp.org/content/forgetting-curve-importance-distribution-well-quantity-general-practice-teaching.

[57] C. Hoadley and F. C. Campos, "Design-based research: What it is and why it matters to studying online learning," *Educational Psychologist*, vol. 57, no. 3, pp. 207–220, 2022.

[58] J. H. L. Koh and B. K. Daniel, "Shifting online during covid-19: A systematic review of teaching and learning strategies and their outcomes," *International Journal of Educational Technology in Higher Education*, vol. 19, 56 2022. DOI: 10.1186/s41239-022-00361-7.

[59] M. M. Rahman, Y. Watanobe, T. Matsumoto, R. U. Kiran, and K. Nakamura, "Educational data mining to support programming learning using problem-solving data," *IEEE Access*, vol. 10, pp. 26 186–26 202, 2022. DOI: 10.1109/ACCESS.2022.3157288.

[60] *How to determine the optimal k for k-means?* Sep. 2023. [Online]. Available: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb.

[61] M. Higley. (). "Benefits of synchronous and asynchronous e-learning." Retrieved April 8, 2020, [Online]. Available: https://www.example.com.