Abstract— Subsequent to the data deluge of the internet era and the recent advancement in big data technologies, it is easy to affirm the continuous application of such technological innovation to tackling a wide array of students' educational needs. The field of artificial intelligence and machine learning have improved education learning outcomes. However, the problem of generalized traditional supportive collaboration scripts for all students irrespective of the student's learning traits and position on the learning spectrum leads to less than optimum result in their educational pursuits. This paper presents a novel approach that uses data mining algorithm to optimize the selection of educational resources for students based on their learning traits and the six factors that confound instructional content and delivery with a focus on students with learning disabilities for STEM subjects. Keywords— STEM education, data mining, knowledge discovery, teaching and learning analytics, intelligent tutoring system

I. INTRODUCTION

Turning data into knowledge historically often relies on a traditional method that utilizes manual analysis and interpretation. The internet has powered the most significant data dump in the history of the world-leading to the big data concept that alludes to a large amount of generated data from the technological evolution and the interaction of people in digital spaces [12]; [24]. Consequently, the probing of the data set using the manual process was slow, expensive, and highly subjective [7][11] further noted that the ability of commonly used software to manage, query, process, and analyze data within a tolerable elapsed time becomes impracticable as data volume increases in size in the era of big data. Specifically, big data refers to the large generated data volume from technology development and the continuous interaction and actions of users in the digital space. Big data and its associated concepts like data learning mining or learning analytics help with understanding and making sense of the data [18]. The handling of the large amount of data that includes analyzing patterns and serving as an object of research after technological advances and platforms development for interaction between users and the content reached its greatest apogee in the decade 2010-2020 [12] [18]. In the education field, the analysis of big data coming from students constitutes a new effort towards improving the teaching-learning process and serves as the bedrock of this study.

Education is a data-driven enterprise where teachers provide or facilitate the instruction, and the students internalize the material [15]. The improvement in technology has encouraged the collection, aggregation, and analysis of student data on both micro and macro scale. [15] has pointed out that on the macro level, data sets from different countries allow policymakers to adjust the goals of their national education system-based students achievement data. The micro-scale solution facilitates improvements in instructional technology and helps teachers assess students' progress in real-time and make spontaneous changes to their lessons. Other students' data include attendance, discipline, and demographic data in addition to their grades and test results that allowed schools to form tolerably complete pictures of their students.

Intelligent Tutoring System (ITS) is an adaptive learning environment that can monitor and interpret user's activities, explain user's needs and preferences based on activities, and makes changes to the learning process [6]. [13] said ITS is a good fit for delivering customized instruction due to their ability to measure and diagnose the knowledge, use the information to determine variances in actual and anticipated level of knowledge, and provide learning tasks that are an appropriate fit to the observed difference. ITS can serve as elements of modern education. It can also serve as constituents of a comprehensive, widely adaptive system comprising of learning styles/sensory preferences, student's knowledge identification, assigning suitable learning objects, and creating a personalized study plan that can be tested and run in practice [6]. Customized user instruction is a critical advantage of ITS, but research results pointed to the fact that using computers to enhance learning has fallen short of the initial expectations [13].

This paper is structured as follows. The second section explores contemporary related research. The third section provides the problem statement for the study and presents the hypothesis. The fourth section describes the methodology for the work. The fifth section shows a few application examples of the
extension. The final section offers the conclusion and future work.

II. RELATED WORK

The two popular growing areas in the inclusion and exploration of significant data capabilities in education are EDM and LA and other related communities [21]. There is a growing utilization of EDM and LA in dealing with associated issues in providing instructional strategies that support and enhances the collaboration process among students. Also, big data offers the opportunity for better implementation of a non-intrusive and in-process evaluation strategy for online courses that complements the traditional and time-consuming ways to collect feedback [10].

A. Education Data Mining

Romero and Ventura (as cited in [21] p.49) noted that the EDM field deals with developing, researching, and applying algorithms to detect patterns in big data that would be near impossible to analyze with other methods. Furthermore, [3] described EDM as techniques, tools, and research designs that facilitate the process of extracting meaningful information from captured data from the learning activities of learners in educational settings. Additionally, EDM utilizes techniques from data mining, machine learning, and statistics to data from educational environments like universities, colleges, and tutorial systems [1] [3] [20]. [3] further described the ultimate goal of EDM as the prediction of student's learning behavior, analysis of the result of the provided educational support, new teaching model discovery including the improvement of the existing models, and the advancement of scientific knowledge among the students.

Education data may include student demographic data or captured data reflecting navigation behavior within a learning environment. The data may also be learning activities data like quizzes, interactive class exercises, and activities; it may also be data from students working together or data from a text chat forum [19]; [26]). [4] stated that EDM collects direct measures from the learning management system (LMS) logs, database queries, and analytics. [28] described the commonly utilized EDM methods under five categories: prediction, clustering, relationship mining, a discovery with models, and human judgment distillation of data. Xue further stated that prediction, clustering, and relationship mining are standard in traditional mining while EDM utilizes discovery models and distillation of data and EDM emphasizes the importance of these techniques.

B. Learning Analytics

As defined in [3] learning analytics is data mining plus interpretation and action. Learning analytics deals with the utilization of intelligent data, learner produced data, and analysis models to identify information and social connections for predicting and advising the learning process [16] [17]. Furthermore, [22] defined learning analytics as a system that enhances stakeholders' decision making at multiple levels through a feedback mechanism. A holistic view of LA techniques revealed its objectives as 1.) measuring educational data and improving educational effectiveness, and 2.) focusing mainly on the learner and how to analyze their activities to enhance learning outcomes [9] [19] [27]). [23] discussed further that learning analytics goals are to collect, analyze, and visualize students' data to gain insights that are meaningful and capable of offering learning environments improvements. [3] said LA lays more emphasis on concurrent investigation of collected data alongside human observations in teaching context and learning, which stands in contrast to EDM that focuses on the development of new methods of computation for data analysis. [3] said LA processes focus on applying known techniques to answer the vital questions that have an impact on students' and organizational learning systems.

LA can be useful in many ways, like placing models on both learners and questions and using the models to estimate the competence of learners and predict the success level of future learners [25]. [9] pointed out that data for predictive modeling emanate from many sources, like student characteristics that encompass attributes like disposition and demographics. [8] discussed one of the ways organizations are utilizing LA by citing an example of a software application - Instructure's Canvas Dropout Detective that empowers faculties to identify student's performance issues that may put them at-risk for unsuccessful outcomes. [5] said their software suite consisting of analytics applications could help schools with such tasks as "predictive analytics" or help with "assessment and accreditation.".

III. PROBLEM STATEMENT, HYPOTHESIS STATEMENT, AND RESEARCH QUESTION

A. Problem Statement

Currently, ITS does not mitigate inaccurate resource selection by students with learning disabilities.

B. Hypothesis Statement

The study introduces a new composite variable to achieve the primary goal of the research. The variable is student learning index (SLI) that helps to document observation of difficult attributes like intelligence or health. Many research described creating a composite variable as a process that begins with critically analyzing the theoretically intended meaning of a concept to capture the composite and logical meaning. This study adopted meaningful grouping to develop SLI from the observation of students' and learning resource attributes. The student learning index consists of student's prior achievement, students' perception of self-efficacy, the content of instruction, management of instruction, educators' effort to evaluate and improve instruction, and educators' beliefs about the nature of effective instruction. Subsequently, we can generate a function that expresses the effectiveness of selecting an educational resource using SLI as the hypothesis statement as shown below.

Hypothesis: Learning resource selection can be expressed as a function of students’ learning index.

C. Research Question

How will the utilization of machine learning algorithm in an ITS affect the accuracy of educational resources selection for students with learning disabilities in STEM subjects?
IV. METHODOLOGY

The methodology of the framework provides a prototypical structure that predicts the efficiency of a learning resource based on the effectiveness of instruction and students' learning attributes. The framework utilizes a decision tree algorithm to build a decision tree that can accurately predict the degree of success of the education materials based on the class attributes. In this section, we will discuss the data mining method and its implementation using WEKA software. Subsequently, we discuss the result of the machine learning methods and evaluation.

A. Decision Tree Algorithms

The decision tree is a comfortable and widely accepted algorithm to understand and interpret. The algorithm is a member of the supervised learning algorithms group with the ability to solve both regression and classification problems. The classification process in machine learning is a two-step process, the learning step, and the prediction step. The learning phase deals with building the model based on the training data. The prediction step predicts the response to the given data. Decision trees have two types of target variables: categorical and continuous variables. Decision trees classify problems by sorting down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. An individual node in the tree serves as attribute test cases, and each edge from the node corresponds to the possible answers to the test case.

Decision trees utilize multiple algorithms to split a node into sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes leading to an increased in the purity of the node concerning the target variable. This study will utilize the iterative dichotomiser (ID3), C4.5, C5.0 and CART algorithm selection method that uses a top-down greedy search approach through the space of possible branches with no backtracking. A greedy algorithm always makes the best choice for the moment.

(i) Steps in ID3, C4.5, C5.0 and CART algorithm

The simplicity of the tree algorithms plays to the strength of its choice for the research, and the following steps will be the path of execution 1) the root node will be the original set consisting of the possible class attribute values; 2) each iteration of the algorithm will iterate through the unused attribute of the set S and the corresponding calculation of Entropy (H) and Information gain (IG); 3) the selection of the next attribute will depend on the smallest Entropy or Largest Information gain; 4) splitting of the set S will occur based on the selected attribute to generate a subset of the data; 5) the algorithm continues to recur on each subset considering unutilized attributes.

(ii) Population and Sample

The classification is a two-step process, the learning step, and the prediction step in machine learning. In the learning step, the training data serves the purpose of building the training model, and in the prediction step, the trained model will predict based on the given data. The research will utilize computer-based simulation, also referred to as analytical simulation. While the research chose decision-tree algorithm from plethora of machine learning algorithm, the study introduces a new dataset with a foundation in previous research that identified the factors that confound instructional content and delivery for students with learning disabilities. To facilitate understanding the research dataset and the machine learning algorithm, it is vital to understand the following concepts: instances, dataset, attribute, and instance space. Instances refer to individual observations, and dataset refers to a set of observations. Attributes are a set of values that measure specific instance properties. Instance space is the space of all possible combinations of attribute values. Inductive learning algorithms induce a domain model from a given set of observations by utilizing domain data as input and producing a model of the domain's structure as output.

(iii) Machine Learning Dataset

[14] noted that the body of research on mathematics instruction for students with learning disabilities (LD) is still developing and does not describe a specific and comprehensive set of well-researched practices. However, it provides sufficient procedures and issues as clearly associated with practical instruction and increased student achievement to help set the training data's attributes that constitute student learning index components: prior achievement, perception of self-efficacy, the nature of example, explicitness, parsimony, time on task, level of success, and content coverage. Each attribute can take on one of the values: low, medium, and high; the possible value of the student learning index, which is the outcome attribute, can be: failure, success, and highly successful. An STI failure value indicates the educational resource selection will yield less desirable results, while success or high success value indicates a high probability of success.

The solution space encapsulates all the possible combinations of the attribute values leading to all possible outcomes. The study aims to model the solution space by representing all the possible scenarios and feeding it to the machine learning algorithm to deploy a classifier that can accurately predict a subset match for queries. Since the solution space generalization includes every possible occurrence, it is possible to quickly adapt the solution to other problem spaces with minimum changes. For example, our study has a scope limitation of addressing students with a disability. However, with a slight modification, the classifier can grow to address other problems by modifying the attributes and the underlying algorithm.

An excel macro generated all the possible combinations of attribute values for the study's class attributes. We can calculate all the possible scenarios using different mathematical formulas since we know all the attribute values. However, for this study, we adopted an excel macro that automatically calculates a possible combination of output with a given input. The excel macro generated a total of 19690 possible instances representing all the scenarios. The generated scenarios included lots of duplicates and improbable scenarios. A preprocessing phase cleaned up unneeded data to prevent ambiguous and unreliable results.
(iv) Verification Dataset

The plan for verifying the machine learning model includes designing an independent test set that closely mimics students’ attributes with learning disabilities and mapping it to different educational resources. An independent test set consisting of traits that closely resembles students with learning disabilities as highlighted by [14] matched with video learning resources from study.com with varying attribute values to verify which combination will yield an outcome that corresponds to a successful student learning index.

B. Evaluating the Trained Model

The classifier output contains the result of the experiment with six distinct sections: run information, classifier model (full training set), prediction on test data, stratified cross-validation, precise Accuracy by class, and confusion matrix. This section discusses the summary, detail accuracy by class, and the confusion matrix.

TABLE I. EVALUATION OF TRAINING SET

<table>
<thead>
<tr>
<th>Summary</th>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
<th>Kappa statistic</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Relative absolute error</th>
<th>Root relative squared error</th>
<th>Total Number of Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12415</td>
<td>708</td>
<td>0.8529</td>
<td>0.0487</td>
<td>0.1548</td>
<td>68.3549 %</td>
<td>77.1291 %</td>
<td>13123</td>
</tr>
</tbody>
</table>

(ii) Detailed Accuracy by Class

This section consists of the 'TP Rate' is the true positive rate, and 'FP Rate' is the false positive rate. The true positives are the number of items that are correctly identified, divided by the total number of all the items represented as a percentage. Conversely, the number of incorrectly identified items divided by the total number of items is the false positive, also represented as a percentage. The precision and recall are vital factors that help to examine the strength of the model. Precision is a reflection of how accurately the classification method performed by correctly identifying the right category for the instances; it is the fraction of relevant instances among the retrieved instances. Recall determines the number of retrieved items that are relevant. Precision in machine learning represents a measure of quality. In contrast, recall is a measure of quantity with the implication that higher precision implies the ability of an algorithm to return relevant results than irrelevant ones, and recall suggests that an algorithm returns most of the appropriate result.

The F-Measure is the combination of precision and recall, which results in the harmonic mean of both precision and recall. The 'ROC Area' receiver operator characteristics are a measure of Accuracy that determines the percentage of accurate prediction; it uses true positives and false-positive rates as the axis to draw a curve. ROC area is an area under the curve, and values above 0.5 reflect better ROC. The 'PRC Area' under the curve deals with precision in recall and uses the precision in recall to draw a curve and does not account for true positives in any of the statistics. The PRC Area may be useful for unbalanced data.

(iii) Confusion Matrix

The confusion matrix shows how well the classifier is performing and represent a good measure of the effectiveness of the machine learning program. The confusion matrix shows the number of classes by rows. The sum of the row values indicates the number of instances in the model. There are a total of 12277 instances of the class (a) - Failure, 828 instances of the class (b) - Success, and 18 instances of the class (c) – high Success. A closer examination reveals that 618 instances of a are incorrectly identified as instances of b while 90 instances of class b are incorrectly identified as instances of a. All the 18 instances of class c had correct identification.

(iv) Evaluation Result

To verify our classifier (trained model), we ran an independent set of test data to get a fair and unbiased evaluation result in line with best practices in the industry. The goal here is to have different independent samples from an infinite population. We created three sets of data tagged allfailuretest, allsuccessest, allhighsuccessest. The data represents the traits that are dominant in students with learning disabilities; the students' characteristics range from low to medium and mapped to different configurations of educational resources. The resulting data set is now fed to the classifier for evaluation.

(i) Stratified Cross-validation Summary

Table 1 shows the summary section and discusses the correctly and incorrectly classified instances from the implementation of the algorithm. The decision tree data mining method in this experiment correctly classified 12415 instances and incorrectly classified 708 instances, translating to 94.60% and 5.39%.
The Research Question and the Classifier Evaluation

Table 1 shows the classifier result of an approximately 95% prediction rate. The Precision values for the classes are 0.992 for Failure, 0.544 for Success, and 1.0 for High_Success. The Recall values are 0.950 for Failure, 0.891 for Success, and 1.0 for High_Success. The weighted avg for F-Measure, MCC, ROC Area, and PRC Area is 0.952, 0.677, 0.966, and 0.974, respectively. Conclusively, we can conclude that we have a robust model that can predict the choice of educational resources for students based on students and instruction content attributes with a very high accuracy. If students can utilize resources that have the best chances of improving their learning, we can reasonably conclude that the classifier optimization of educational resources for students can only improve their learning outcome.

TABLE II. INDEPENDENT TEST VALIDATION RESULT FOR ALL SUCCESS AND FAILURE SCENARIOS

<table>
<thead>
<tr>
<th>inst#</th>
<th>actual</th>
<th>predicted</th>
<th>error</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1:Success</td>
<td>1:Success</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2:Failure</td>
<td>2:Failure</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2:Failure</td>
<td>2:Failure</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2:Failure</td>
<td>2:Failure</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2:Failure</td>
<td>2:Failure</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>2:Failure</td>
<td>2:Failure</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2:Failure</td>
<td>2:Failure</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2:Failure</td>
<td>2:Failure</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2:Failure</td>
<td>2:Failure</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Correctly Classified Instances 17 100%
Incorrectly Classified Instances 0 0%
Kappa statistic 1

Mean absolute error 0
Root Mean absolute error 0
Relative absolute error 0%
Root relative squared error 0%
Total Number of Instances 17

Confusion Matrix

\[ \begin{array}{ccc}
9 & 0 & 0 \\
0 & 8 & 0 \\
0 & 0 & 0 \\
\end{array} \]

V. APPLICATION EXAMPLE

There are several ways to deploy the classifier and utilize it in an education environment. The classifier can be helpful when
deployed in the implementation of ITS as a simplified method of selecting the best resources for a learner that presents the best opportunity for success. The student's attribute information may be obtained by a form of a questionnaire that determines the prior achievement of the learner concerning the topic at hand. The questionnaire can also help to determine the level of self-efficacy attribute for Success or Failure. The other attributes that relate to the instruction are primarily in the domain of the educators. A teacher may set all the attributes for a particular lesson or may choose to do so or other existing resources. Our classifier will utilize the attribute to predict the chances of success based on each student. This is vital to help create a differentiated script of education resource utilization for student solving the problem of generalized delivery that affect students.

This solution has paid attention to students with learning disabilities and modeled the attributes in a way that mimics their learning style. Still, the applicability of the solution extends beyond the scope of what we utilized for the study demonstration and has the possibility of a broader solution in the education community.

VI. CONCLUSION

In response to the influx of data in education, many data analytics fields have responded to the opportunity with learning analytics emerging as a burgeoning field for improving student education outcomes.

The study utilized data learning mining techniques and procedures for extracting useful and relevant information to improve the learning process. The derived data from the pedagogical agents (such as teachers and learners) serves as input in the analytics process to improve the quality and the learning process experience of the digital environments [17]. Several possible future works are validating the learning material selected through this approach with various groups of students and faculty, integrating the model with any existing Learning Management System, and creating the model with any E-Tutoring software or service system.

REFERENCES