A New Data Fusion Framework of Business Intelligence for Mining Educational Data

Nissreen El Saber*, Aya Gamal Mohamed, Khalid A. Eldrandaly

[1]Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt.

Emails: naelsaber@fci.zu.edu.eg; aya.gamal@fci.zu.edu.eg; Khalid_Eldrandaly66@zu.edu.eg

Abstract

Student academic performance can be affected by social, economic, and educational factors. Many research works studied these factors applying to different levels in the educational organizations' models. The importance spans giving professional educational advice to vulnerable students, supporting the student's development of special education-related skills, and encouraging students to handle their education challenges. For educational organizations, dealing with pandemics and other obstacles has proven to be essential for education sustainability. One way is to be proactive and use the power of exploring and discovering educational data to predict students' performance and attitude. Mining educational data can benefit from Business Intelligence (BI) in visualizing, organizing, and extracting insights for student's performance. Educational Data Mining (EDM) is used in this research to predict students' performance. A novel data fusion framework is introduced for Business Intelligence using educational data mining. This study aims to show the techniques that predict students' performance and the most effective methods for each of them. The proposed framework used the advantage of business intelligence concepts and tools to highlight the metrics providing better statistical and analytical understanding.

Keywords: data mining; educational data mining; business intelligence; learning management systems; sustainability

1 Introduction

Students' success rate is one of the main measures for any education system. Monitoring students' performance in education systems is considered essential for student success. Data is stored for a variety of reasons, including learning, accessing, understanding, and so on. Large amounts of data must be stored using a high-quality, relatively expensive, system, in terms of storing, retrieving, and reserving privacy [1]. Data mining can be beneficial for solving these issues. It provides tools and methods to investigate huge amounts of data allowing for discovery and organization [2]. With data mining tools and methods classify, cluster, predict, relationship identification, and prepare for the decision-making process [3]. One of the promising applications of educational data mining is to predict student performance which can help in an efficient and sustainable educational system.

Surveying the literature (summarized in Sections 2 and 3), we found that there is a need for educational data analysis as the number of students continues to rise with their data filed in many Terabytes over years of education. The results can lead to teaching process improvements in different education facilities. Predicting students' performance is one of the most pressing challenges in educational data mining (EDM) [4].

By predicting student performance, the design of the curriculum can improve, and measures for academic support and advice on the education given to students can be planned [2],[4]. Learning analytics has been the focus of several

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

research studies as an entirely new technique for learning and a tool for improving learning performance[5]. The education management system must be able to predict performance and identify successful students early to avoid the risk of students' academic failure in the learning process as early as feasible. The process allows educators to intervene and guide students early enough by using classification or regression algorithms. In this work, we presented a set of different DM methods that were used to predict student performance by different algorithms and others using business intelligence to improve predicting education and give better results. Moreover, a framework is presented and explained for using BI in the EDM for predicting student performance.

Student failure is a possibility, that is, this risk should be assessed, and predicted and avoidance plans should be adopted. Students who are at risk of academic failure are identified by predicting their performance and advising instructors to take steps to support students as soon as feasible. This may include providing guidance or interventions or conducting ongoing evaluations of learner resource suggestions [6], [7]. Higher educational organizations can be supplied with effective ways to increase the efficiency of the organization and the learning process of students with the use of EDM [7]. This study will help education institutes/organizations identify and support students who need extra care toward better academic growth.

The development of techniques for making discoveries within the special categories of data gathered from educational settings and applying those techniques to learn more about students and the environments in which they learn are the goals of educational data mining (EDM)[2], [8]. EDM approaches are characterized by directly utilizing the numerous levels of relevant hierarchy in educational data to conduct an efficient analysis[8]. EDM is an interdisciplinary research field that uses statistics, information systems, educational psychology, data mining, machine learning, and other theories and methods to analyze educational data, assisting people in effectively resolving a variety of educational issues[2], [3], [8], [9]. Hence, EDM is an excellent option for data mining, making it an excellent opportunity when creating and utilizing new techniques to examine educational data and comprehend the learning environment of students.

Business Intelligence (BI) is an umbrella term for multiple processes and technologies [6] that are designed to enable an organization to gather, analyze, store, and access substantial amounts of data[6]. It is the process of converting business data into meaningful insights that can enable groups and organizations to make informed decisions and successful strategies. BI and DM are both subsets within the field of data analytics. BI focuses on the use of data and analytics to inform decisions, improve operations, optimize processes, and identify opportunities and trends within an organization. DM is the process of extracting useful patterns from large datasets. The general outcome of the DM methods are general patterns that can be used for future results prediction or organizational decision guidance.

In the following work, we introduce a new data fusion framework of BI for mining Educational Data. EDM and BI can provide a more comprehensive picture of the educational environment. We have used about 8,000 data samples, to train different algorithms to give more appropriate mining results. We used the WEKA application and applied different algorithms to the data then the result ran through a random tree algorithm. After that Power BI was used to generate and visualize reports to the administrator.

In Section 1, EDM concepts and methods, and EDM are introduced. Related previous works in EDM for student performance prediction are elaborated, and compared highlighting the limitations. Then Section 2 introduces the basic concepts of BI for education, briefing the previous studies combining BI and EDM for a better education decision experience. After that, the proposed BI framework for EDM is presented and explained in Section 3. In Section 4, the practical results of our experiment are presented and commented on. Finally, Section 5 concludes the study.

2 Educational Data Mining

The goal of educational data mining (EDM) is to provide techniques for analyzing the different types of data that are generated in educational settings. EDM employs a multitude of aspects, including social, behavioral, cognitive, and motivational factors[10]. The related methods are customized for the use of a better understanding of students' performance progress concerning the learning settings [3]. In concept, EDM originated from computer science, statistics, and education integration as shown in Figure 1. Machine Learning (ML) resulted from the intersection between computer science and statistics and computer-based education happens due to the integration of computer science and education disciplines. And finally, Learning Analytics (LA) integrates the education systems and statistics. All are depicted in Figure 1. EDM focuses on analyzing educational data to better understand student learning and improve the educational system and helped also to improve E-learning for obtaining information about teachers' and students' behavior and level of use on these platforms. Data mining offers a variety of methods for obtaining

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

knowledge from data [11]: it can identify the business problem, mine the data to produce actionable information, act on the information, and measure its results.

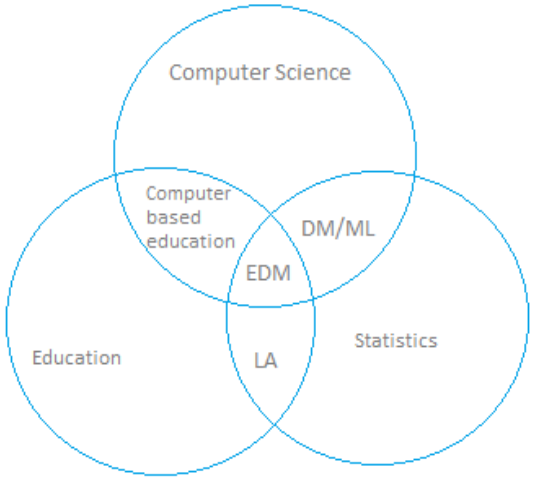

Figure 1: EDM is an interdisciplinary system (Adapted from [8])

Generally, an EDM case can be one of the following five problem types. These problem types are originally discussed in [3] and used in this work for organizing literature in EDM. The problem types are affected by the nature of EDM as an interdisciplinary system allowing for contributions from data mining, machine learning, psychometrics, statistics, and information systems among others. Here is a brief list of the considered EDM problem types:

1) **Prediction:** to predict unknown variables based on historical data for the same variable. Prediction can be based on **a) Classification** uses past knowledge to build a learning model using a binary or categorical variable for the new data, and **b) Regression** used to predict variables. It predicts continuous variables, such as linear regression and neural networks, and **c) Density estimation** using kernel functions, e.g., Gaussian functions.
2) **Clustering:** to separate data into separate groups based on certain common features. Here, the data labels are unknown giving a broader view of the dataset.
3) **Relationship mining:** to find relationships between different variables in multivariable datasets. That is, to determine relations among variables measuring its strength.
4) **Discovery with models** to make data more understandable. Ideally, presenting the allown different distinct ways allows more details to arise.
5) **Distillation of data for human judgment:** to visually present education-related data for classification and identification.

For the following EDM literature review, we searched from 2015 to 2022 on Google Scholar and used key terms: data mining, educational data mining, and prediction. In this section assets of literate investigated in EDM using AI approaches and analytics. The survey focused on the prediction problem type for EDM. A summary is provided in Table 1.

A prediction problem was introduced and handled using students' pass-or-fail assignment results to predict their performance [12]. Tools for classification and prediction are in Table 1. Artificial Neural Networks (ANN) ran over the students' cumulative GPA for performance prediction in [13]. Having many mining methods, authors in [14] used classification and regression for results prediction using R language with automatic application of algorithms (listed in Table 1). For high school students' performance prediction, a clustering and prediction model is built in [15] with K-Means and SVM. Classification, rule-based learning, ensemble methods, and neural network-based algorithms were applied in [14] for student performance prediction. Traditionally, ensemble methods are better at dealing with structured data than neural networks [14].

Another application of EDM was introduced in [16] to predict the department a student studies in with respect to student's performance historical data using rule-grounded literacy, ensemble styles, and neural network. In [12], researchers used students' demographic features and performance measures, based on certain assignments, to explain and predict their performance later. The tools used are highlighted in Table 1.

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

For vulnerable students, a sample of 2000 students' data is used for performance prediction with the assistance of different mining techniques [17]. Vulnerable students were identified through a Decision tree and Logistic Regression predicted the performance effectively. For online education, the research in [18] examined the ability to predict student performance using DEEDS dataset logs. The model used Random Forest (RF), SVM, Naïve Bayes, Logistic Regression, and Multilayer Perceptron. In [19], authors studied the performance of 2000 students to extract students with difficulties before dropping out in credit-hour systems. The model used WEKA to apply five main ML algorithms proving that the Kernel Logistic Regression (KLR) algorithm yields better results.

Table 1: EDM-Related Prediction Literature

Reference (chronically)	EDM Problem Type		Techniques Used
2015[20]	Prediction	Classification Regression	Decision trees, SVM, Random Forest, AdaBoost
2016[13]		Regression	Artificial Neural Network technique
2017[14]		Classification Regression	Linear regression, Decision trees, and Naïve Bayes classifier
2019[15]		Density estimates Clustering	SVM, Linear Regression, k-means
2019[12]		Classification Regression	Rule-based learning, Ensemble, and Artificial Neural Network
2020[16]		Classification	Machine Learning, Collaborative Filtering, Recommender Systems, and Artificial Neural Networks
2021[18]		Classification	K-nearest Neighbors and Multilayer Perceptron algorithms
2021[21]		Classification	Multilayer Perceptron, Random Forest, and Support Vector Machine
2021[22]		Classification Regression	The Explainable models
2021[23]		Clustering	Cluster sampling and Knowledge discovery is databases (KDD)
2022[24]		Classification Regression	Kernel Logistic Regression (KLR). LIBSVM. Artificial Neural Network. Naïve Bayes classifier. Decision Tree-J48
2022[25]		Classification Regression	Random Forest (RF), SVM, Naïve Bayes, Logistic Regression, Multilayer Perceptron.
2022[26]		Classification Regression	Decision Tree (DT), Support Vector Machines (SVM), Multi-Layer perceptron (MLP), Naïve Bayes (NB), Logistic regression
2023[27]		Classification	random Forest, Logistic Regression, Decision Tree, Naïve Bayes, Support Vector Machine and K nearest Neighbor

As shown in Table 1, most studies share the same problem type (i.e., prediction) for predicting student performance using EDM. However, they vary in the type of prediction methods and techniques used. Some used classification techniques based on students' data (e.g., grades, marks, levels) and regression analysis to predict how well the student's performance was supposed to be noticed. For the classification techniques, SVM, and RF are most used as classifiers while ANN, Linear, and logistic regression are most used for regression.

In addition to that, the surveyed studies have some drawbacks highlighted collectively. First, there is no significant difference in the prediction algorithms, although proved to have differences [3]. Second, the small size of datasets affects the results as data mining is about large amounts of data to allow the algorithms to effectively be applied [16]. In addition, the datasets are not publicly available. The algorithms used need to be customized for the education field or new algorithms can be developed for the special nature of the education-related problems. Most used algorithms come from machine learning or data mining fields [15]. That is, prediction models need more enhancements in performance and accuracy.

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

3 Bussiness Intelligence

Business Intelligence (BI) is considered the procedure of turning data into information [28]. This information will be used by the leadership to guide organizational leaders to make better business decisions [29]. The use of mathematical models, the enterprises' operating rules, and providing scientific decision-making material are benefits of using BI in organizations [11]. Educational institutions are good candidate organizations for the application of the BI system, due to the highly competitive nature in recruiting students both locally and internationally. The role of the BI system appears in collecting correct and precise information allowing a smooth and beneficial decision-making process. From this perspective, the applications of BI in organizations provide leaders with different perspectives to view, analyze, and generate data.

We surveyed research articles from 2017 to 2023 that are indexed by the Elsevier Scopus library through the Egyptian Knowledge Bank (EKB). The following key search terms are used: EDM, BI, prediction, WEKA, and data mining. The following are sample-related articles using BI techniques and/or tools for educational data mining management [17], [18]. Table 2 summarizes the findings of the survey. Using educational data stored in central cloud storage, authors in [17]provided analysis and insights over all sorts of stored data using the ERMS model. Meanwhile, engineering students were provided with personalized education plans based on their capabilities [28]. The student's capabilities are extracted from the data and associations through mining techniques. In [11], Pentaho software, a data integration and analytics application, was introduced in the context of education-related organizations. The effect of using data analytics techniques on educators and students has been discussed in [29]. They highlighted the need for some level of technical knowledge for educators to be able to benefit from the use of BI applications in the education management field. Villegas-Ch, et al.[30] designed a BI-based support framework for Learning Management Systems (LMS) to analyze the students' leave rates for 3,207 distance-learning students. For the aim of decision support in educational organizations, [17] used ML algorithms on students' performance data in a semester in higher education organizations. The introduced framework is promising with a detailed case study.

Table 2: BI in EDM Literature Summary

Ref.	Tools	
	BI	**Data Mining**
[31]	Enterprise Resource Management System (ERMS)	Advanced Java and Bootstrap
[30]	ETL process, Data Warehouse	Microsoft Cluster Algorithm Learning Management System (LMS)
[32]	Pentaho Business Analytics	Decision Support Software WEKA
[19]	Hefesto Methodology	Moodle LMS Algorithm J48 (an implementation of the C4.5 algorithm)
[33]	ETL process, Data warehouse and Kimball methodology	Algorithm J48
[34]	business decision support system (Data Layer-business layer-application layer)	Bayesian Network

4 THE PROPOSED FRAMEWORK

The proposed data fusion framework combines the benefits of data mining (DM) & business intelligence (BI) highlighted in the previous literature. In this section, the framework is built and discussed. The framework for business intelligence is a strategic method for collecting data to use in decision-making. It aids in ensuring that all valuable information has been used, that the data is accurate and validated, and that the data is used effectively to improve their results, it is a method that enables educational organizations to find the best answers to information-related challenges and provide the educators and decision-makers with the required evidence for proceeding in effective decisions.

The proposed framework consists of three phases in two primary areas. The first two phases are basically performed as a data mining solution (i.e., activities like determining data sources, preprocessing process, and mined data results). The third phase is the business intelligence basic activities to visualize and provide statistical insights to the decision-makers. Figure 2 illustrates the proposed data fusion framework showing its three phases as A, B, and C

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

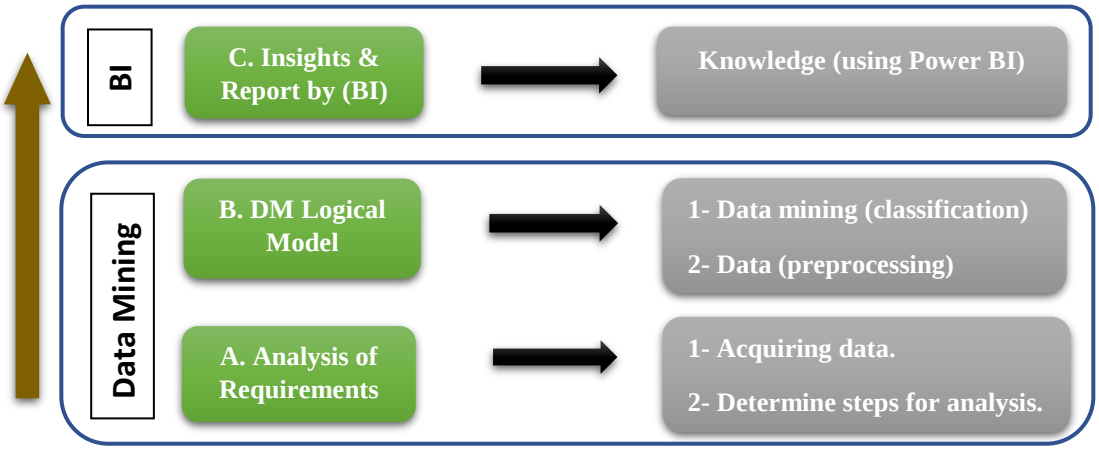

Figure 2: The Proposed Data Fusion BI Framework for EDM

A. Analysis of Requirements:

First, we use the dataset from (College Attending Plan Classification | Kaggle). It is mainly designed with students' data to allow predictions of educational decisions. One of the decisions is whether high school students are willing to go to college. This information helps the college prepare for the upcoming number of students at a suitable time and helps the government put plans for others to join the market. Data were prepared by using student IDs instead of names for privacy. Other variables considered are gender, the student's IQ, the parent's income, and if parents support students at college.

Second, the best supports the user's requirements and needs selected to develop the logical model of the data mining, and as a result, tables of dimensions, tables of conformed dimensions, and tables of data are defined for each perspective.

B. Data Mining Logical Model

In this phase of the methodology, the process of Extraction, Transformation, and Loading of data (ETL) from the sources detected for the analysis, for the software used for data analysis, mining, and reporting: The ETL process carried out, which allowed connecting the various data sources and transforming them to load them within the DM structure. We use WEKA programming and explore data by graphical user interfaces. Weka provides a variety of visualization tools and algorithms for data analysis and predictive modeling, a comprehensive set of data preparation, and modeling methods. Weka provides many types of standard data mining operations, particularly data preparation, clustering, classification, regression, visualization, cleaning data, and transforming it into appropriate structures for an analytical model and feature selection. Data Mining uses different algorithms to know the best by using the WEKA program. Knowledge is available to the end user using testing modelings such as random trees, Naive Bayes, function logistics, and decision tables. Result comparisons are illustrated in Figures 4,5,6,7 below. Figure 3 shows the order of steps to be performed when using WEKA. Results related to the case study are highlighted in Section 5.

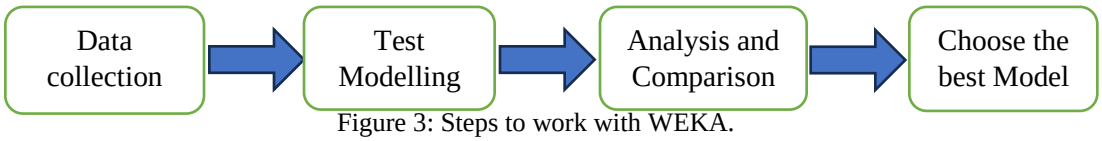

Figure 3: Steps to work with WEKA.

From Sections 2 and 3, after surveying the literature, we apply certain algorithms and techniques highlighted below for different stages of the proposed framework. Working with the data mining algorithms, the following algorithms for supervised machine learning are used:

- **Random Forest (RF):** The RF method similarly generates a tree, but this algorithm generates multiple trees using the values of random samples in the dataset, and the result is based on the results of many of the constructed trees. RF significantly improves a model's classification accuracy by constructing a set of trees

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

that generate results separately, collecting those results, and determining which class got the most votes [35]. The expression h(x, k, k=1, ...) represents a tree classifier, where x is the input and k is a set of equally split random vectors. The group of classifiers can be expressed as h1(x), h2(x)..., and h K(x), each of which provides a result for the class with the highest probability[35].

- **Naive Bayes:** A native Bayesian classifier is a statistical supervised machine learning method that predicts the probability of class membership. When used to a large dataset, NB achieves excellent accuracy and speed, but it also performs very well on tiny datasets [35], which can be characterized as follows.

$$P(A \mid B) = P(B \mid A)\, P(A) \quad P(B)$$

.

P(A) expresses the chance of A occurring, while P(B) denotes the likelihood of B occurring. The probabilities of A and B were distinct values that had no bearing on one another. The probability of A occurring given that B has occurred is known as the posterior probability. Finally, P(B|A) denotes the degree of probability, which is determined by the likelihood of B occurring if A is true.

- **Logistics Function Algorithm:** The logistic function is also known as the sigmoid function or the cost function. A logistic function is a Shaped curve that takes the input (numeric value) and changes it to a value between 0 and 1.

$$y = \frac{e^{a0+a1*x+a2x2}}{1+e^{a0+a1*x+a2x2}}$$

where y is the predicted value, a0 is the y-intercept, a1 is the principal component's coefficient of the independent variable x1, a2 is the value of the independent variable x2, and e is the natural algorithm's base. In this investigation, the independent variables x1 and x2 are replaced by the principal components (pc1 and pc2). The maximum likelihood method is used to compute the y-intercept and regression coefficients rather than the least squares method [36].

- **Decision Table:** Decision Table generates a majority classifier for a decision table. It can do cross-validation and assess feature subsets using best-first search. Based on the same set of features, an alternative strategy, rather than the table's global majority, is used to select the class for each instance that is not covered by a decision table entry. [37]

C. Insights and Reports by BI (Knowledge):

It can exploit the BI information in diverse ways: With pre-defined reports by using Power BI, to monitor the performance of a specific area or the entire company, with an operational, tactical, and strategic vision. Power BI is a group of software services, applications, and connectors that combine to transform disparate data sources into coherent, attractive graphics, and interactive insights. Your data may be stored in a hybrid data warehouse that is both cloud-based and on-premises, or it may be an Excel spreadsheet. Power BI makes it simple to connect to your data sources, visualize the data, identify the key information, and share it with anyone you want. To validate our proposed framework that was solved in section 3, a case study was used. We will case study to view the result of WEKA and Power BI.

5 Case Study

We entered data on the WEKA program and applied different algorithms classification: random tree, naïve Bayes, functional logistics, and decision table. We will display their results and know what is appropriate for our data. The performance of each model was tested using 10-fold cross-validation and compared based on major accuracy measures, properly or erroneously classified instances of kappa, mean absolute error, and time required to develop the model.

5.1 Framework Application

First, we applied a random tree algorithm that is 10-fold cross-validation, classifier model (full training set) on plan column that has taken time to build the model: .38 seconds. The size of the tree was 2937. The result of correctly classified instances is 78.1% and incorrectly classified instances is 21 .8%. Naïve Bayes algorithm that is 10-fold

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

cross-validation, classifier model (full training set. on plan column that takes time to build the model: .01 seconds. The result of correctly classified instances is 82.6% and incorrectly classified instances is 17.4%. function logistics algorithm that is 10-fold cross-validation, classifier model (full training set. on plan column that takes time to build the model: .11 seconds. The result of correctly classified instances is 83.1% and incorrectly classified instances is 16.9%. Decision table algorithm that is 10-fold cross-validation, classifier model (full training set. on plan column that has taken time to build a model: .44 seconds. The result of correctly classified instances is 83.8% and incorrectly classified instances is 16.2%.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6253               78.1625 %
Incorrectly Classified Instances      1747               21.8375 %
Kappa statistic                          0.5028
Mean absolute error                      0.2184
Root mean squared error                  0.4673
Relative absolute error                 49.8099 %
Root relative squared error             99.8117 %
Total Number of Instances             8000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.163    0.662      0.667    0.665      0.503    0.752     0.550     plan
                 0.837    0.333    0.840      0.837    0.838      0.503    0.752     0.813     not plan
Weighted Avg.    0.782    0.278    0.782      0.782    0.782      0.503    0.752     0.727

=== Confusion Matrix ===

    a    b   <-- classified as
 1732  864 |    a = plan
  883 4521 |    b = not plan
```

Figure 3: Random tree results

The results of the random tree on WEKA, it was taken time to build the model 0.07 second. The correctly classified instance was 78.16% and the incorrectly classified instance was 21.83%. The size of the tree was 2973.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6608               82.6    %
Incorrectly Classified Instances      1392               17.4    %
Kappa statistic                          0.6068
Mean absolute error                      0.2208
Root mean squared error                  0.3516
Relative absolute error                 50.3735 %
Root relative squared error             75.0943 %
Total Number of Instances             8000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.137    0.724      0.750    0.737      0.607    0.891     0.822     plan
                 0.863    0.250    0.878      0.863    0.870      0.607    0.891     0.937     not plan
Weighted Avg.    0.826    0.213    0.828      0.826    0.827      0.607    0.891     0.900

=== Confusion Matrix ===

    a    b   <-- classified as
 1947  649 |    a = plan
  743 4661 |    b = not plan
```

. Figure 4: Naïve Bayes results.

Naïve Bayes took 0.04 seconds to build. The correctly classified instance was 82.6% and incorrectly classified instances were 17.4%

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

```
Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6652               83.15    %
Incorrectly Classified Instances      1348               16.85    %
Kappa statistic                          0.6112
Mean absolute error                      0.2425
Root mean squared error                  0.347
Relative absolute error                 55.3234 %
Root relative squared error             74.1196 %
Total Number of Instances             8000

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.718     0.114     0.751       0.718   0.735      0.612    0.891     0.821     plan
                 0.886     0.282     0.868       0.886   0.877      0.612    0.891     0.936     not plan
Weighted Avg.    0.832     0.227     0.830       0.832   0.830      0.612    0.891     0.899

=== Confusion Matrix ===

    a    b   <-- classified as
 1865  731 |    a = plan
  617 4787 |    b = not plan
```

Figure 5: Logistics Function results

The logistics function took time to build model 0.11 second. The correctly classified instance was 83.15% and the incorrectly classified instance was 16.85%

```
Time taken to build model: 0.44 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6706               83.825   %
Incorrectly Classified Instances      1294               16.175   %
Kappa statistic                          0.6233
Mean absolute error                      0.2316
Root mean squared error                  0.3388
Relative absolute error                 52.8189 %
Root relative squared error             72.3604 %
Total Number of Instances             8000

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.711     0.101     0.772       0.711   0.740      0.624    0.897     0.834     plan
                 0.899     0.289     0.866       0.899   0.883      0.624    0.897     0.939     not plan
Weighted Avg.    0.838     0.228     0.836       0.838   0.836      0.624    0.897     0.905

=== Confusion Matrix ===

    a    b   <-- classified as
 1846  750 |    a = plan
  544 4860 |    b = not plan
```

Figure 6: Decision table

The decision table was taken time to build model 0.44 seconds. The correctly classified instance was 83.825%and the incorrect instance was 16.175%.

5.2 Performance of Classification Algorithms

The reliability of the predictions is an important parameter to consider when comparing the effectiveness of various categorization algorithms. Which algorithm is performing better on the given task can be determined by the proportion of examples that were correctly classified. Seven different classification methods are compared in Figure 7 based on the proportion of occurrences that are correctly identified. The Decision table method recorded the largest percentage of cases that were successfully identified, 83.8%, while the random tree algorithm recorded the lowest percentage of instances that were correctly classified, 78.1%. Table 3 shows the list of rates computed for the above confusion matrix.

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

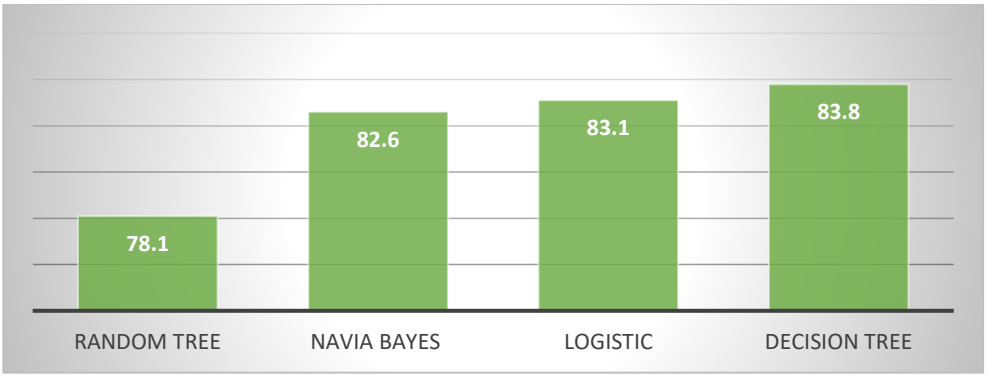

Figure 7: Percentage of Correctly Classified Instances

Table 3: Confusion Matrix of All Features by Classification Algorithm:

	A	B	←-- Classifier as
Random Tree	1732 883	864 4521	a= plan b=not plan
Naïve Bayes	1947 743	649 4661	a= plan b=not plan
Logistics	1865 617	731 4787	a= plan b=not plan
Decision tree	1846 544	750 4860	a= plan b=not plan

A. Dashboard using Power BI

Power BI first collects data, transforms it, models it, visualizes it, and finally can share reports. Power BI Desktop combines Microsoft Query Engine (M) with visualization and data modeling for creating interactive reports. For more advanced visuals, users familiar with the R or Python scripting languages can also add charts created from these graphics' libraries. In addition to these built-in data visualization options, Microsoft has also enabled open-source visual effects in the Power BI platform[38].

Data Preparation

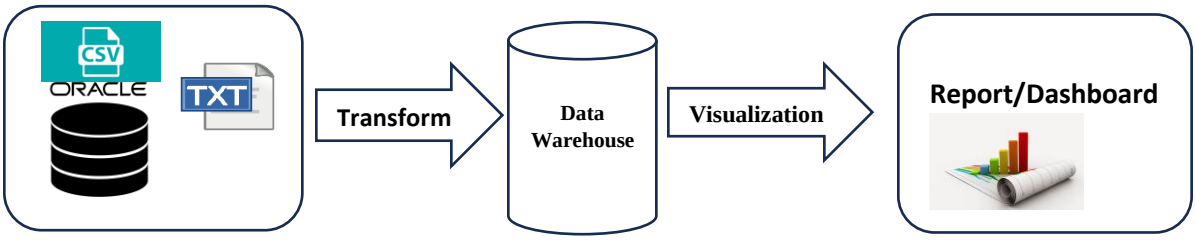

Figure 5: Power BI Architecture

Using student database of 8000 records Dashboard created in Power BI. The main attributes for the dashboard are Student ID, plan, Parent income, IQ, Gender, and Encouragement. We used pie charts, donut charts, Gauges, line charts, clustered bar charts, and stacked column charts. As shown from the results on the WEKA & Power BI, the data was classified using different algorithms. Power BI can show different level views on how an organization is running, what gain, and pain points they have, and live stats to update an organization's internal or external stakeholders. All of this is presented in a visually appealing way compared to rows and rows of numbers in an Excel file.

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

Figure 6: Statistical results (power BI) on the case study

6. Conclusion

Educational data mining is a promising field of application for business intelligence. To name a few, exploring educational management data, finding the vulnerable students' performance patterns, and predicting many effective situations and plans for problems and issues. This research introduces a state-of-the-art survey for the related EDM and BI studies, trying to understand the application of BI in EDM. A data fusion framework of BI for EDM is developed and applied to a case study. The results are highlighted to show how BI is powerful in presenting and providing insights for decision-makers in educational organizations. The provided case study is an educational data mining mined using WEKA then visualized for insights using MS Power BI allowing for better decision making. Naïve Bayes provides the least time to run and the Decision table provides the best result of correctly and incorrectly classified instances.

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

Reference

[1] S. S., V. P. K. M. S., M. M. A., S. R. G. S., K. S., T. S. Mangalapalli, "Data Mining in Education: A Review of Current Practices," *https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152430632&doi=10.1109%2fICAIS56108.2023.10073932&partnerID=40&md5=8aa10dc18e9eb2473ded4f06eaea5d6*, 2023.

[2] A. A. Alqarni, "A New Data Fusion Framework of Business Intelligence and Analytics in Economy, Finance and Management," 2021 2nd International Conference on Computing and Data Science (CDS), 2021, pp. 1-6, doi: 10.1109/CDS52072.2021.00009.

[3] A. 2016 Algarni and abdulmohsen algarni, "Data Mining in Education," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 7, pp. 456–461, Jun. 2016.

[4] M. M. Abuteir, A. Mustafa El-Halees, M. M. Abu Tair, and A. M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study Arabic Sarcasm Sentiment Analysis View project Distributed Arabic Text Classification View project Mining Educational Data to Improve Students' Performance: A Case Study," vol. 2, no. 2, 2012,

[5] B. Wibawa, J. S. Siregar, D. A. Asrorie, and H. Syakdiyah, "Learning analytic and educational data mining for learning science and technology," 2021, p. 060001. doi: 10.1063/5.0041844.

[6] A. M. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student Engagement Level in an e-Learning Environment: Clustering Using K-means," American Journal of Distance Education, vol. 34, no. 2, pp. 137-156, 2020, doi: 10.1080/08923647.2020.1696140.

[7] R. A. Huebner, "A survey of educational data A survey of educational data-mining research mining research."

[8] W. Xiao, P. Ji, and J. Hu, "A survey on educational data mining methods used for predicting students' performance," *Engineering Reports*, vol. 4, no. 5. John Wiley and Sons Inc, May 01, 2022. doi: 10.1002/eng2.12482.

[9] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," in *Procedia Computer Science*, Elsevier, 2015, pp. 500–508. doi: 10.1016/j.procs.2015.07.372.

[10] H. Pallathadka, S. Jain, S. Kamble, and K. Tongkachok, "Educational Data Mining: A Comprehensive Review and Future Challenges," *ECS Trans*, vol. 107, no. 1, pp. 16129–16136, Apr. 2022, doi: 10.1149/10701.16129ecst.

[11] Á. Rocha, J. L. Reis, M. K. Peter, and Z. Bogdanović, "Marketing and Smart Technologies Proceedings of ICMarkTech 2019." [Online]. Available: http://www.springer.com/series/8767

[12] I. D. Shetty, D. Shetty, and S. Roundhal, "Student Performance Prediction," *International Journal of Computer Applications Technology and Research*, vol. 8, no. 5, pp. 157–160, Apr. 2019, doi: 10.7753/IJCATR0805.1003.

[13] M. F. Sikder, M. J. Uddin, and S. Halder, "Predicting students yearly performance using neural network: A case study of BSMRSTU," in *2016 5th International Conference on Informatics, Electronics and Vision, ICIEV 2016*, Institute of Electrical and Electronics Engineers Inc., Nov. 2016, pp. 524–529. doi: 10.1109/ICIEV.2016.7760058.

[14] M. Pojon, "Using Machine Learning to Predict Student Performance," 2017.

[15] A. U. Khasanah and H. Harwati, "Educational data mining techniques approach to predict student's performance," *International Journal of Information and Education Technology*, vol. 9, no. 2, pp. 115–118, Feb. 2019, doi: 10.18178/ijiet.2019.9.2.1184.

[16] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Applied Sciences (Switzerland)*, vol. 10, no. 3. MDPI AG, Feb. 01, 2020. doi: 10.3390/app10031042.

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

[17] M. D. S. Mussa, S. C. de Souza, E. F. da S. Freire, R. G. Cordeiro, and H. R. M. da Hora, "BUSINESS INTELLIGENCE IN EDUCATION: AN APPLICATION OF PENTAHO SOFTWARE," *Revista Produção e Desenvolvimento*, vol. 4, no. 2, pp. 29–41, Jul. 2018, doi: 10.32358/rpd. 2018.v4.274.

[18] H. El Aouifi, M. El Hajji, Y. Es-Saady, and H. Douzi, "Predicting learner's performance through video sequences viewing behavior analysis using educational data mining," *Educ Inf Technol (Dordr)*, vol. 26, no. 5, pp. 5799–5814, Sep. 2021, doi: 10.1007/s10639-021-10512-4.

[19] M. Sanchez Peralta, "Business Intelligence in E-Learning, a Case Study of an Ecuadorian University," in *2018 XIII Latin American Conference on Learning Technologies (LACLO)*, IEEE, Oct. 2018, pp. 29–32. doi: 10.1109/LACLO.2018.00016.

[20] P. Strecht, L. Cruz, J. Mendes Moreira, C. Soares, J. Mendes-Moreira, and R. Abreu, "A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance," 2015.

[21] A. Triayudi and W. O. Widyarto, "Educational Data Mining Analysis Using Classification Techniques," *J Phys Conf Ser*, vol. 1933, no. 1, p. 012061, Jun. 2021, doi: 10.1088/1742-6596/1933/1/012061.

[22] R. Alamri and B. Alharbi, "Explainable Student Performance Prediction Models: A Systematic Review," *IEEE Access*, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 33132–33143, 2021. doi: 10.1109/ACCESS.2021.3061368.

[23] A. Abdulahi Hasan and H. Fang, "Data Mining in Education: Discussing Knowledge Discovery in Database (KDD) with Cluster Associative Study," in *2021 2nd International Conference on Artificial Intelligence and Information Systems*, New York, NY, USA: ACM, May 2021, pp. 1–6. doi: 10.1145/3469213.3471319.

[24] H. A. Abdelhafez and H. Elmannai, "Developing and Comparing Data Mining Algorithms That Work Best for Predicting Student Performance," *International Journal of Information and Communication Technology Education*, vol. 18, no. 1, pp. 1–14, Feb. 2022, doi: 10.4018/ijicte.293235.

[25] G. Ben Brahim, "Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features," *Arab J Sci Eng*, vol. 47, no. 8, pp. 10225–10243, Aug. 2022, doi: 10.1007/s13369-021-06548-w.

[26] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," *IEEE Access*, vol. 10. Institute of Electrical and Electronics Engineers Inc., pp. 72480–72503, 2022. doi: 10.1109/ACCESS.2022.3188767.

[27] S. M. Dol and Dr. P. M. Jawandhiya, "A Review of Data Mining in Education Sector," *Journal of Engineering Education Transformations*, vol. 36, no. S2, pp. 13–22, Jan. 2023, doi: 10.16920/jeet/2023/v36is2/23003.

[28] M. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The Current Landscape of Learning Analytics in Higher Education," Computers in Human Behavior, vol. 89, pp. 98-110, 2018, doi: 10.1016/j.chb.2018.07.027.

[29] S. Anardani, L. Sofyana STT, and A. Maghfur, "Analysis of business intelligence system design for student performance monitoring," *J Phys Conf Ser*, vol. 1381, no. 1, p. 012015, Nov. 2019, doi: 10.1088/1742-6596/1381/1/012015.

[30] C. da R. Brito *et al.*, *EDUNINE2018 - II IEEE World Engineering Education Conference: the role of professional associations in contemporaneous engineer careers: proceedings: March 11 to 14, 2018, Buenos Aires, Argentina*.

[31] A. Bansal, A. Singhal, Amity University. School of Engineering and Technology. Department of Computer Science and Engineering, Amity University, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, and Institute of Electrical and Electronics Engineers, *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering: 12th-13th January 2017, Amity University, Noida, Uttar Pradesh, India*.

[32] M. D. S. Mussa, S. C. de Souza, E. F. da S. Freire, R. G. Cordeiro, and H. R. M. da Hora, "BUSINESS INTELLIGENCE IN EDUCATION: AN APPLICATION OF PENTAHO SOFTWARE," *Revista Produção e Desenvolvimento*, vol. 4, no. 2, pp. 29–41, Jul. 2018, doi: 10.32358/rpd. 2018.v4.274.

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023

[33] W. Villegas-Ch, X. Palacios-Pacheco, and S. Luján-Mora, "A business intelligence framework for analyzing educational data," *Sustainability (Switzerland)*, vol. 12, no. 14, pp. 1–21, Jul. 2020, doi: 10.3390/su12145745.

[34] A. Ghyasi and H. Rashidi, "A New Approach Based on Business Intelligence and Bayesian Network for Analysis of Corporate Accounting Systems," 2023.

[35] C. N. Villavicencio, J. J. E. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Covid-19 prediction applying supervised machine learning algorithms with comparative analysis using Weka," *Algorithms*, vol. 14, no. 7, Jul. 2021, doi: 10.3390/a14070201.

[36] S. Shaik, C. Shravya, K. Pravalika, and S. Subhani, "CITATIONS 78 READS 1,763," 2019. [Online]. Available: https://www.researchgate.net/publication/363296423

[37] A. Çakır and B. Demirel, "A software tool for determination of breast cancer treatment methods using data mining approach," *J Med Syst*, vol. 35, no. 6, pp. 1503–1511, 2011, doi: 10.1007/s10916-009-9427-x.

[38] J. D. Wark, "Power Up: Combining Behavior Monitoring Software with Business Intelligence Tools to Enhance Proactive Animal Welfare Reporting," *Animals*, vol. 12, no. 13, p. 1606, Jun. 2022, doi: 10.3390/ani12131606.

Doi: https://doi.org/10.54216/FPA.130108
Received: March 22, 2023 Revised: June 26, 2023 Accepted: August 29, 2023