OLAP Mining with Educational Data Mart to Predict Students' Performance

Ihab Ahmed Najm[1], Jasim Mohammed Dahr[2], Alaa Khalaf Hamoud[3], Ali Salah Alasady[3], Wid Akeel Awadh[3], Mohammed B. M. Kamel[4] and Aqeel Majeed Humadi[5]
Email: ihab@tu.edu.iq, jmd20586@gmail.com, alaa.hamoud@uobasrah.edu.iq, alishashim2009@googlemail.com, wid.jawad@uobasrah.edu.iq, mkamel@hs-furtwangen.de, aqeelm16@gmail.com

[1]College of Computer and Mathematic Sciences, Tikrit University, Iraq

[2]Directorate of Education in Basrah, Iraq

[3]College of Computer Science and Information Technology, University of Basrah, Iraq

[4]Eotvos Lorand University, Budapest, Hungary
Hochschule Furtwangen University, Furtwangen im Schwarzwald, Germany
University of Kufa, Kufa, Iraq

[5]College of software engineering, Islamic Azad University of Khorasgan.

Keywords: OLAP mining, data mining algorithms, data mart, educational cube, OLAP.

Received: December 7, 2021

Academic institutions always try to use a solid platform for supporting their short-to-long term decisions related to academic performance. These platforms utilize historical data and turn them into strategic decisions. The hidden patterns in the data need tools and approaches to be discovered. This paper aims to present a short roadmap for implementing educational data mart based on a data set from Alexandria Private Elementary School, located in the Basrah province of Iraq in the 2017-2018 academic year. The educational data mart is implemented, then the cube is constructed to perform OLAP operations and present OLAP reports. Next, OLAP mining is performed on the educational cube using nine algorithms, namely: decision tree with score method (entropy) and split method (complete)), decision tree with score method (entropy) and split method (complete)), decision tree with score method (entropy) and split method (both)), Logistic, Naïve Bayes, Neural Network, clustering with expectation maximization, clustering with K-means clustering, and association rules mining. According to a comparison of all algorithms, clustering with expectation-maximization proved the highest accuracy with 96.76% for predicting the students' performance and 96.12% for predicting students' grades amongst all other algorithms.

Povzetek:

1 Introduction

The significance of data repositories has become prominent with the increase in the number of large institutions. Departments organize and oversee their own databases (administrative, academic, marketing, financial, etc.) which control a large amount of common data [1]. The term Data warehouse (DW) can be described as time-variant, subject-oriented, non-volatile and integrated data that can be used in supporting strategic decision making [2, 3]. DW accommodates a huge set of permanent historical data that are useful in administrative decision-making when it comes to data access and retrieval for time analysis, decision making, and knowledge discovery [4]. It is precisely developed for the purpose of extracting, processing, and representing data in an appropriate format [5, 6]. The extracted data from different sources, systems, rules, and places are described as a database containing a huge collection of existing data that is stored to assist in decision-making within an organization [7]. An indicator is required for the interior intention of a user. Indicators are required systems for studying, analyzing, and presenting the enterprise data in a way that permits senior management to produce decisions [8]. The choice of a DW adopted by an organization is greatly influenced by the manner in which the organization operates as well as the type of decision support system (DSS) required by the organization [9]. Another example of a DW is the data mart. The data mart is a single-use or limited-use system used in the analyses of precise information in a precise area or for a particular sector of production [10, 11]. Datastores mostly hold only summary data but can be connected to operational systems for the purpose of partitioning into various transaction details when required. Datastores are controlled by information technology departments in organizations as well as users in a particular group or department [12, 13]. Many types of DWs and data marts are implemented in many fields such as educational health DW [9], clinical path [14-16], cancer

DW [17], HR DW [18], compliant DW [19], and retail DW [20] to support strategic decisions.

The educational staff is always seeking novel technologies and applications for decision support. They are in search of applications and tools for decision support built on varying techniques and algorithms such as data mining and machine learning algorithms, as well as web-based and mobile learning tools [14, 21-28]. Educational DW is one of the most promising solutions that produce online analytical results based on multidimensional OLAP (MOLAP) queries to benefit all educational stakeholders such as lecturers, professors, decision-makers, and managers. Educational DW grants a large perspective of students' performance and permits the detection of obstacles along the way. It is also of significance to future applications of data mining for the implementation of several algorithms and techniques designed for handling large amounts of educational historical records or data as against small datasets. The utilization of educational DW can significantly mitigate educational mistakes that have great negative impacts on academic decision-making. Datamart can be utilized as a platform for deploying and implementing varying algorithms and tools to acquire adequate analysis results. Key performance indicator (KPI), data mining, and MOLAP algorithms can be integrated to develop a DSS for coming up with absolute strategic decisions [29, 30].

The Datamart is utilized in this study as the base of the data-driven educational DSS. It is implemented on specific school data and is also used in the analysis of executive information systems. This study presents the results of implementing data mining algorithms with MOLAP to produce OLAP mining based on educational data mart to predict both students' performance and final grades. As a part of the study, a comparative study between data mining algorithms is performed to acquire the most appropriate algorithm for the prediction of students' performance. The use of OLAP mining in the design of this model adds more analytical tools to uncover hidden patterns in students' behavior. OLAP mining encourages additional interactive exploration in a previously constructed educational cube. The implementation of the model begins with converting the paper-based educational data of student records into electronic educational records followed by performing all extraction, transformation, and loading (ETL) processes on the data mart schema (star schema). The educational cube is built to implement MOLAP queries, KPIs, and OLAP mining algorithms. SQL Server Management Studio (SSMS) 2014 is used to manage educational data mart objects, i.e. data mart schema and staging area. The ETL is designed and implemented by SQL Service Integration Service (SSIS) 2014, whereas SQL Server Analytical Service (SSAS) 2014 is implemented in the design and construction of the educational cube, KPIs, and OLAP mining. All of the projects are deployed by SQL Server Data Tools (SSDT) tool, relying on its ability to create and deploy (SSIS, SSAS, and SQL Server Reporting Service (SSRS)) projects. The rest of this paper is organized as follows: Section 2 analyses and discusses the literature review and presents the strength of the work. Section 3 explains the model implementation roadmap as well as the details of each step of the model implementation. Section 4 concludes and discusses the points derived from the model result and presents future works.

2 Literature review

Wang, Hairong, Pan Huang, and Xu Chen in [31] proposed a model to analyze and find the patterns in more than 21000 records of university data based on OLAP cube and multidimensional association rule mining of Apriori algorithm. These techniques are used for fast calculations to find the patterns that affect the performance of students. Khadija Letrache, Omar El Beggar, and Mohammed Ramdani [32] proposed a dynamic partitioning for OLAP cubes using association rules mining. An analyzing process is performed with queries of users in a specific period for both partitioning and non-partitioning data cube for the evaluation process.

Anjana Yadava and Anand Kumar Tripath [33] introduced a model that improved the OLAP performance by using a clustering approach. The multidimensional clinical data is divided into several clusters where each cluster achieves the query processing. Javier García-Tobar [34] proposed a new methodology based on OLAP and clustering algorithm in the field of analyzing radon measurements. OLAP cube has been implemented based on SSAS to analyze thousands of data measurements.

Xiao-Han Zhou and Xiao-Mei Zhang [35] proposed a model for book lending based on OLAP cube and a generalized rule induction algorithm to analyze library lending data. While Asma Lamani, Brahim Erraha, Malika Elkyal, and Abdallah Sair [36] proposed an approach to combine the data warehouse as a platform with OLAP cube and data mining algorithm for decision making. Their model aimed to integrate the data mining prediction with OLAP cube based on the clustering algorithm. Clustering is used for partitioning the initial data cube into sub-cubes in order to enhance OLAP work and the prediction results. The regression tree is applied to the subcube to find the prediction results of the cell. Finally, D Putri and I S Sitanggang [37] proposed a model to analyze the data of horticultural crops based on spatial OLAP cube and K-means clustering algorithm. This model can explore the distribution of the horticultural crops over Indonesia country. SpagoBI used to build this model to analyze data between the years 2000 to 2013. This tool can also visualize the results in many forms such as tabular tables and maps. However, all the discussed models either proposed a model to analyze data and OLAP cube without data mining algorithms or introduced a data mining algorithm for discovering and exploring hidden patterns in OLAP data cube. In many papers, the data mining algorithms are used with OLAP cube results and in many of them, the data mining algorithms are used to divide and partition data cube to enhance the results and improve query performance. In most papers, clustering or association rules mining is used and no comparative study among all algorithms is presented.

3 Methodology

The flowchart in figure (1) represents the implementation steps of the proposed model. The model consists mainly of three areas: data preparation, data mart, and presentation areas. The paper-based records are turned into electronic records and then extracted into the staging area of the educational data mart. All transformation processes are performed on the data in order to get consolidated data. The result data after transformation are loaded into fact and dimension tables and fact. These tables are used to construct the educational cube to be used later to build MOLAP reports, KPIs, and data mining algorithms. The presentation area serves the academic stakeholders such as senior managers, decision-makers, analysts, and teachers to get the insight analysis results.

The educational data set consists of 4170 records taken from Alexandria Private Elementary School in Basrah province, Iraq during the 2017-208 academic year. All these data are in the form of paper-based records where they were transferred into electronic tables using SQL Server Management Studio (SSMS) 2014. SSMS is used to manage the staging area of the data mart, building the schema and holding all the dimension and fact tables. Next, the data profiling step is performed on the data, and the results are conducted in the table (1). The table shows the distinct values and null values in each column of the data set. The birth column holds 104 records of null values. This table helps in performing the data preprocessing to handle these null values. The ratio of null values compared with the overall size of the data set is not critical, so these values can be ignored or replaced with a value to clean these missing values.

Table (2) of data profiling represents the class distribution according to subjects. Six classes (1-6) with four groups (A, B, C, and D) of students in the elementary school with different subjects are taken in each class. The marks are divided from mid-exam to four exams before mid-exam and two exams after the mid-exam. The final exam grade is used as a class label to predict. The grade is either pass (P) or fail (F). The subjects taught in the school are listed in the table (2). The subjects differ in each of the classes. For example, Islamic studies and Arabic studies are taught in the first class to sixth class, while the English language is taught in fifth and sixth classes.

For the patterns distributions in the data set, table (3) shows a sample of patterns in the data set. The final grade holds two distinct values (P and F) where the percentage of the (F) is greater than the percentage of (P). For the groups of the students (A, B, C, and D), group B toke the highest percentage with 36%, followed by groups A, C, and D with 27%, 23%, and 10%, respectively. For Mid grade, again (F) represented the highest percentage with 67% followed by E, V, M, and G with 13, 6%, 5%, and 5%, respectively. The other patterns in the data set are very useful in the preprocessing step of building the data mart and using ETL tools.

There are two main areas in the data mart: staging area and schema tables of the data mart. The staging area represents the intermediate area between data sources and data mart where all the processes of ETL starting from data sources and ending with data mart tables are involved in the staging area [7, 38, 39]. The five dimension tables of educational data mart are connected to the fact table through surrogate keys. Datamart star schema was implemented by SSMS where it consisted of five dimension tables (degree, subject, enrollment, information, and address) which connect to the

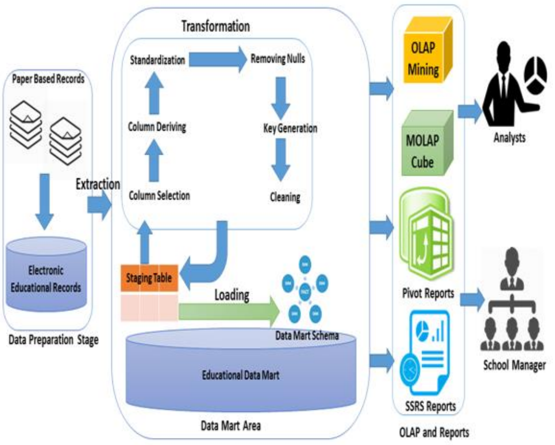


Figure 1: Model Implementation Roadmap.

Field Name	***Number of Distinct Values***	***Null Ratio***
Birth	528	104
Address	11	0
Groups	5	0
Month	613	0
Number_of_Students	599	0
Class	6	0
School_Name	1	0
Student_Number	599	0
Mark	613	0
School_Address	1	0
Subject	12	0
Student_Name	599	0
Year	1	0

Table 1: Columns statistics.

Subject	***Class***
Islamic Studies	1 to 6
English Language	5 to 6
Arabic Language	1 to 6
Physical Education	1 to 6
Arts	1 to 6
Mathematics	1 to 6
Social Studies	4 to 6
Science	1 to 6

Table 2: Subject per classes.

Column	*Pattern*	*Percentage*
Final Grade	F	65%
	P	35%
Groups	A	27%
	B	36%
	C	23%
	D	10%
Mid Grad	F	67%
	M	5%
	E	13%
	V	6%
	G	5%

Table 3: Patterns distributions.

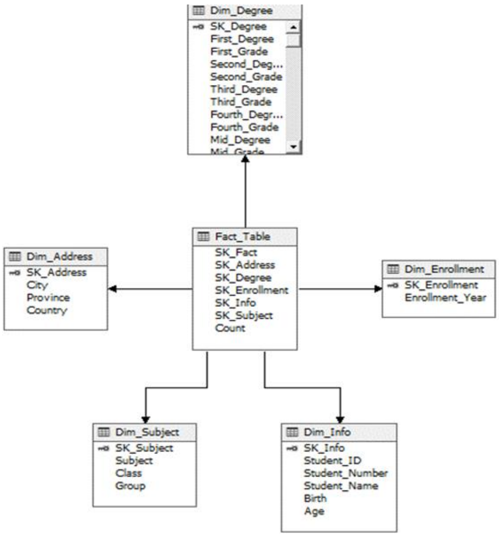


Figure 2: Data mart schema.

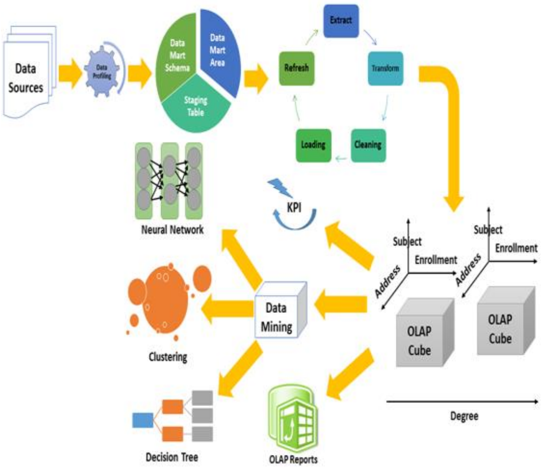


Figure 3: Proposed Model Structure.

educational fact table. The measurement in the educational fact table is (count) which represents the count function to find the number of result students with MOLAP queries. The data mart schema is shown in figure (2).

All ETL processes are performed in a staging table to produce integrated and cleaned data to be loaded into dimension tables for fast OLAP queries [31, 40, 41, 42, 43]. The approach of designing an independent educational data mart is a bottom-up approach due to the short duration of implementation and the size of data sources. Figure (3) presents the structure of the proposed model. The first step involves examination and fast analysis of the data source to check the quality of data. This process is called data profiling where the data is assessed and checked to determine the required processes for cleaning data. A full summary is conducted from this process involving null ratio, domain range, data type, and length. Data profiling is one of preprocessing steps that affect the success of data mart implementation. As mentioned earlier, the data mart area involves the data mart tables and staging area. The ETL processes are performed on the data in the staging table in order to consolidate, integrate, and standardize the data to be loaded in dimension tables. SSIS package used to implement all these processes where the data in staging table first loaded into dimension tables and then all the concatenated surrogate keys with count are loaded into the fact table. Next, the cube construction based on the dimension tables is performed using SSAS.

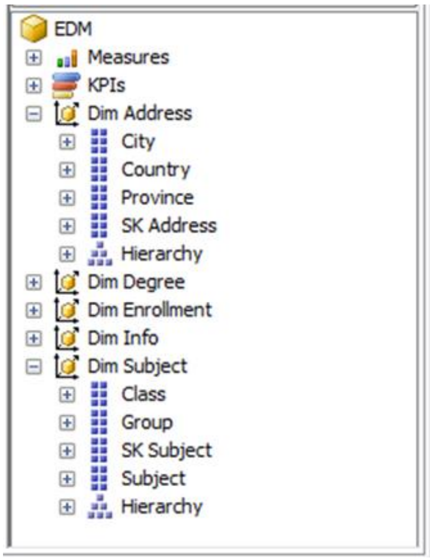
The educational cube in figure (4) holds all the dimension tables with their hierarchies. The cube is the analytical platform for the MOLAP queries. It can be used for constructing reports based on MOLAP queries and for analyzing data based on an excel pivot table. This cube is also used as a platform for data mining algorithms due to its fast and excellent response. This fast response resulted from the way of storing and indexing information in the cube.

After the educational cube is constructed, the data mining structures are implemented on the MOLAP cube using nine data mining algorithms (DT-1, DT-2, DT-3, Logistic, NB, NN, clustering with Expectation-Maximization (EM), clustering (K-means), and association rules mining (Apriori)).

Microsoft DT algorithm is a classification and prediction algorithm used to predict both continuous and discrete attributes. Feature selection guides an algorithm to select the better features. Feature selection is used to improve the algorithm performance, enhance the resulting quality, and prevent unnecessary attributes from exhausting the processor. The entropy score method determines the interestingness score. An attribute is considered as low interesting if it has high entropy (i.e. random distribution) and low information gain [44, 45]. The feature entropy is calculated and compared with all other features such as:

Interestingness (feature) = - (m - Entropy(feature)) * (m - Entropy(feature))

Where m represents the central entropy. The central entropy is the entropy of all feature sets. After that, the entropy of the feature is subtracted from the central entropy to find the information of the given feature. The split methods are binary, complete, and both. The binary method divides the paths after nodes into two branches, where the complete method divides the tree into attributes values. Both methods allow the analysis service to



EDM
- Measures
- KPIs
- Dim Address
  - City
  - Country
  - Province
  - SK Address
  - Hierarchy
- Dim Degree
- Dim Enrollment
- Dim Info
- Dim Subject
  - Class
  - Group
  - SK Subject
  - Subject
  - Hierarchy

Figure 4: Educational cube.

determine the best method (complete or binary) and produce better results. Microsoft logistic regression is one of the most known statistical methods to determine the contribution of multiple values to a pair of values in the class label. A modified neural network is used in this approach to visualize the effect between the input and output features. Various values in input features are weighted in the final model. The logistic regression name comes from using the logistic information to compress the data curve in order to reduce the effect of the extreme value.

Microsoft naïve Bayes is one of the classification algorithms based on the Bayes theorem which can be used for both predictive and exploratory models. The Naïve name represents that the dependencies among features are not taken into account when implementing the Bayesian technique. The Bayes theorem is represented in the following equation [46]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where P(A|B) is the conditional probability of A giving that B is occurred, P(B|A) is the conditional probability of B giving that A is occurred, and P(A) is the probability of A, while P(B) is the probability of B.

This algorithm gives fast computational results compared to other algorithms, so it can be used to discover the relationship between the input features and output predicted features.

Microsoft neural network is the adaptable implementation of the neural network for machine learning. This algorithm tests the values of all input features against the values of predicted output features and calculates the probabilities of each combination according to the training set. This approach can utilize multiple outputs with multiple networks. The number of networks in the model is determined by the number of input features values as well as the number of output features values. The network consists of multiple layers of neurons to receive the input and produce the output. In the input layer, each neuron is connected to one or more neurons in the hidden layer, and the hidden layer neurons are also connected to multiple neurons in the output layer. It is noteworthy to mention that the neurons are not connected to each other within the same layer.

Microsoft Clustering algorithm combines the Markov chain analysis with the clustering algorithms to determine the clusters and their sequences. The used sequence data in this approach represents the sequence of events within the dataset such as the series of web clicks of a particular customer. The probabilities of all transactions are measured to find the differences among all the sequences to find the best sequence to use as clustering input. EM and K-means are utilized in the clustering algorithm. The algorithm in EM clustering refines the initial cluster in order to fit the data and define the probabilities of data within the cluster. This process will end when the probabilistic model and the data fit. A threshold is used to determine which clusters should be reseeded at new cluster points. The result of using EM is a model where each data point is assigned to a cluster based on probabilities. In the next clustering algorithm, K-means assigns cluster membership by reducing the differences among items and increasing the distances of the clusters. The centroid of the cluster (i.e. means) is an arbitrary point within the cluster. This value will be refined until it becomes the mean value of all cluster data points. K represents the number of arbitrary points within the cluster to determine the best number of clusters. Euclidian distances are calculated to measure the mean value. In k-means, each data point is assigned to one cluster, while the membership represents the distance between the data point and centroid [26].

Microsoft Association Rule Mining (ARM) is used to build a model with the dataset that holds different cases for different items within the dataset. Item set is the group of items within the dataset where the ARM is used to find the correlation among items in different cases. The result ruler from ARM can be used to predict the future cases within the dataset. ARM uses the Apriori algorithm to count the item sets. The scores are generated by Apriori to represent the support and confidence to evaluate the result rules [27].

Throughout the implementation of data mining structure, the final grade (P for Pass, and F for failed) is selected as a final class to get the confusion matrix and to measure students' performance. All of the attributes (group, class, subject, grade1 (October), grade2 (November), grade3 (December), grade4 (January), mid-grade, grade5 (March), grade6 (April), and city) are selected as input columns and the final grade is the final predicted class. The attributes from different dimensions are connected throughout keys in order to make a single table to implement a data mining structure. After implementing data mining structures, the confusion matrix is conducted to measure the performance metrics. The confusion matrix is explained in the table (4).

The confusion matrix represents the number of failed students (negative) and passed students (positive). TP represents the cases where the algorithm predicts as pass while TN represents the cases where the algorithm

		Predicted		
		Negative	Positive	Total
Actual	Negative	True Negative (TN)	False Positive (FP)	P
	Positive	False Negative (FN)	True Positive (TP)	N

Table 4: Confusion Matrix.

Algorithm	*Accuracy*	*Precision*	*Recall*	*F1*	*Kappa*
DT-1	0.9565	0.9039	0.94	0.9216	0.892
DT-2	0.9457	0.9	0.9	0.9	0.863
DT-3	0.9457	0.9	0.9	0.9	0.863
Logistic	0.964	0.9771	0.9661	0.9716	0.923
NB	0.964	0.9771	0.9661	0.9716	0.923
NN	0.964	0.9771	0.9661	0.9716	0.923
Clustering (EM)	0.9676	0.9884	0.9605	0.9743	0.931
Clustering (K-Means)	0.9605	0.9699	0.9681	0.969	0.915
ARM	0.9605	0.9699	0.9681	0.969	0.915

Table 5: Performance metrics of data mining algorithms for predicting students' action.

Algorithm	*Accuracy*	*Precision*	*Recall*	*F1*	*Kappa*
DT-1	0.9511	0.9412	1	0.9697	0.886
DT-2	0.913	1	0.9688	0.9842	0.808
DT-3	0.9511	0.9412	1	0.9697	0.886
Logistic	0.9511	0.9921	0.9766	0.9843	0.895
NB	0.9511	0.9921	0.9766	0.9843	0.895
NN	0.9348	0.9769	0.9922	0.9845	0.853
Clustering (EM)	0.9612	0.9843	0.9766	0.9804	0.917
Clustering (K-Means)	0.9081	1	0.9688	0.9842	0.803
ARM	0.9081	1	0.9688	0.9842	0.803

Table 6: Performance metrics of data mining algorithms for predicting students' grades.

predicts as failed. FP represents the cases where the algorithm predicts as pass while the actual case is failed, and FN represents the cases where the algorithm predicts as failed while the actual case is pass. So based on the confusion matrix, the performance metrics are calculated as follow:

$Precision=TP/(TP+FP)$,

$Recall=TP/(TP+FN)$,

$Acuracy=(TN+TP)/(P+N)$,

$F1=(2*Recall*Precision)/(Recall+Precisison)$,

$Kappa=(((TP+TN)/(P+N))-RA)/(1-RA)$,

Where RA is random accuracy.

$RA=(N/(P+N)*(TP+FP)/(P+N))*((1-N/(P+N))*(1-(TP+FP)/(P+N)))$

The performance metrics are represented in table (5). The performance metrics are accuracy, precision, recall, F1, and kappa. For all nine algorithms (DT-1 (decision tree with score method (entropy) and split method (Binary)), DT-2 (decision tree with score method (entropy) and split method (complete)), DT-3 (decision tree with score method (entropy) and split method (Both)), Logistic, NB, NN, EM clustering, K-means clustering, and ARM) the metrics are presented in the table.

Recall metric represents the proportion of actual pass students that are predicted as pass, while precision metric represents the measurement of the number of pass students out of all students. F1 metric is used to measure the imbalance between FN, and FP predicted cases while the accuracy metric measures the overall algorithm performance based on TP, and TN predicted. Kappa metric represents a comparison between the accuracy of the algorithm and the accuracy of a random algorithm. The performance metrics are measured according to [47, 48, 49]. Based on the overall accuracy of the algorithm, clustering with EM algorithm scored the highest accuracy with 0.9676 while Logistic, NB, and NN scored 0.964. K-means and ARM clustering scored 0.9605, while DT-1 scored 0.9569 and both DT-2 and DT-3 scored 0.9457. According to the precision, again clustering with EM scored the highest value with 0.9884 while Logistic, NB, and NN scored 0.9771. Clustering with k-means and ARM scored 0.9699 while DT-1 scored 0.9039. DT-2 and DT-3 scored the lowest precisions with 0.9. For recall, clustering with k-means and ARM scored the highest value with 0.9681, while logistic, NB, and NN scored 0.9661. Clustering with EM scored 0.9605 while DT-1 scored 0.94, DT-2 and DT-3 scored 0.9. For F1, clustering with EM scored 0.9743 while Logistic, NB, and NN scored 0.9716. Clustering with k-means and ARM scored 0.969, while DT-1 scored 0.9216, and both DT-2 and DT-3 scored the lowest value with 0.9. For Kappa statistics, clustering with EM scored the highest value with 0.931 followed by Logistic, NN, and NB with 0.923. ARM and clustering with k-means scored 0.915, while DT-1 scored 0.892, and both DT-2 and DT-3 are score 0.863 in the kappa statistic.

It is clear that the best algorithm in predicting students' performance according to overall performance is clustering with EM followed by Logistic regression, NN, and NB. Generally, the highest values of the algorithms' accuracies are normal since the predicted value of the final class is either P or F. The missing values in the data set are almost missing, besides the dirty data is not found. These points make the prediction accuracy close to 1. Again, clustering with EM scored the highest value followed by Logistic, NB, and NN according to precision, F1, and Kappa. The classifiers are tested against predicting the final grades of the students. Table (6) represents the performance metrics of the data mining algorithms.

For the overall accuracy, clustering with EM scored the highest value with 0.9612 followed by DT-1, DT-3, Logistic, and NB with 0.9511. NN scored 0.9348, while DT-2 scored 0.913. Both ARM and k-means clustering scored 0.9081. According to precision, DT-2, ARM, and clustering with k-means scored the highest value 1, while Logistic and NB scored 0.9921. Clustering with EM scored 0.9843 while NN scored 0.9769. Both DT-1 and DT-3 scored the lowest value with 0.9412. For recall metric, Both DT-1 and DT-3 scored the highest value with 1, while NN scored 0.9922. Logistic, NB, and clustering with EM scored 0.9766 while ARM, clustering with k-means, and DT-2 scored 0.9688. For the F1 metric, NN scored the highest value with 0.9845 followed by Logistic and NB with 0.9843. DT-2, ARM, and clustering with k-means scored 0.9842 while clustering with EM scored 0.9804. Finally, DT-1 and DT-3 scored 0.9697. According to Kappa statistics, clustering with EM scored the highest value with 0.917 followed by NB and Logistic with 0.895. DT-1, and DT-3 scored 0.886 while NN scored 0.853. DT-2 scored 0.808 while ARM and clustering with k-means scored the lowest value with 0.803. It is obvious that clustering with EM is the best algorithm for predicting students' grades according to kappa and accuracy metrics followed by Logistic and NB. Clustering with k-means, ARM, and DT-2 is the top algorithms according to precision, DT-1 and DT-3 are the best according to recall, while NN is the best according to F1 metric. The overall classifier performance according to kappa and accuracy make clustering with EM the best classifier.

4 Conclusion and future works

The proposed educational data mart is constructed after the paper-based data is converted into an electronic data set. This data is taken from Alexandria Private Elementary School in Basrah Province, Iraq. For the data mart, the star schema is selected for educational data mart due to fast query processing on the tables. The bottom-up approach is followed as a systematic approach for implementing the data mart. This approach is used in implementing independent data marts with a short cycle implementation period. The implementation process of the data set, data mart schema, staging table, and storage is managed by SSMS 2014. Next, The ETL is performed by SSIS to build the data mart and load it with data. The result MOLAP cube is constructed based on all dimensions of the data mart using SSAS. The educational cube holds all the dimensions (degree, info, enrollment, address, and subject). In the next step, data mining algorithms are implemented on the educational cube to perform the MOLAP cube. Data mining structures are executed on eleven columns (group, class, subject, grade1 (October), grade2 (November), grade3 (December), grade4 (January), mid-grade, grade5 (March), grade6 (April), and city) while the column (final grade) is set as a predicted final class to find the prediction accuracy based on confusion matrix. Among all performance metrics, the clustering with EM is selected as the best algorithm in predicting students' performance according to accuracy and Kappa. Due to the percentage of (F) compared with cases in the data set, these metrics reflect the classifier's accuracy in prediction. The same algorithm proved its accuracy in predicting the actual grades of the students. The educational data mart is considered as a platform to implement all analytical structures including data mining structures. The goal behind using data mart is providing a fast analytical platform for all data mining algorithms and holding the new data without query processing. This classifier can be adopted in implementing a DSS to provide analytical results and predict the students' performance and grades in the future. The DSS will be used by different stakeholders to support their strategic decisions and find the patterns in students' actions that cause the failure or increase the success.

References

[1] S. N. Dhamdhere, "Importance of knowledge management in the higher educational institutes," Turkish Online Journal of Distance Education, vol. 16, pp. 162-183, 2015.

[2] R. Kimball and M. Ross, The data warehouse toolkit: the complete guide to dimensional modeling: John Wiley & Sons, 2011.

[3] W. H. Inmon, Building the data warehouse: John wiley & sons, 2005.

[4] I. Moalla, A. Nabli, L. Bouzguenda, and M. Hammami, "Data warehouse design approaches from social media: review and comparison," Social Network Analysis and Mining, vol. 7, p. 5, 2017.

[5] A. Hamoud and T. Obaid, "Building Data Warehouse for Diseases Registry: First step for Clinical Data Warehouse," International Journal of Scientific & Engineering Research, vol. 4, pp. 636-640, 2013.

[6] I. Teotonio, M. Cabral, C. O. Cruz, and C. M. Silva, "Decision support system for green roofs investments in residential buildings," Journal of Cleaner Production, vol. 249, p. 119365, 2020.

[7] J. Caserta and R. Kimball, The Data Warehouseetl Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data: Wiley, 2013.

[8] S. Thulasiram and N. Ramaiah, "Real Time Data Warehouse Updates Through Extraction-Transformation-Loading Process Using Change Data Capture Method," in International Conference on Computer Networks and Inventive Communication Technologies, 2019, pp. 552-560.

[9] Y. Zhu, "A Data Driven Educational Decision Support System," International Journal of Emerging Technologies in Learning (iJET), vol. 13, pp. 4-16, 2018.

[10] A. Khalaf Hamoud, H. Noori Hussien, A. Akram Fadhil, and Z. Raad Ekal, "Improving Service Quality Using Consumers’ Complaints Data Mart which Effect on Financial Customer Satisfaction," in Journal of Physics Conference Series, 2020, p. 012060.

[11] A. Khalaf Hamoud, M. A. Ulkareem, H. Noori Hussain, Z. Abdulkareem Mohammed, and G. Mustafa Salih, "Improve HR Decision-Making Based On Data Mart and OLAP," in Journal of Physics Conference Series, 2020, p. 012058.

[12] L. W. Santoso, "Data warehouse with big data technology for higher education," Procedia Computer Science, vol. 124, pp. 93-99, 2017.

[13] W. N. Price, S. Gerke, and I. G. Cohen, "Potential liability for physicians using artificial intelligence," Jama, vol. 322, pp. 1765-1766, 2019.

[14] A. Hamoud, "Applying association rules and decision tree algorithms with tumor diagnosis data," International Research Journal of Engineering and Technology, vol. 3, pp. 27-31, 2017.

[15] A. Hamoud and T. Obaid, "Using OLAP with Diseases Registry Warehouse for Clinical Decision Support," International Journal of Computer Science and Mobile Computing, vol. 3, pp. 39-49, 2014.

[16] A. Hamoud, A. S. Hashim, and W. A. Awadh, "Clinical data warehouse: a review," Iraqi Journal for Computers and Informatics, vol. 44, 2018.

[17] A. Hamoud, H. Adday, T. Obaid, and R. Hameed, "Design and Implementing Cancer Data Warehouse to Support Clinical Decisions," International Journal of Scientific & Engineering Research, vol. 7, pp. 1271-1285, 2016.

[18] A. Khalaf Hamoud, M. A. Ulkareem, H. Noori Hussain, Z. Abdulkareem Mohammed, and G. Mustafa Salih, "Improve HR Decision-Making Based On Data Mart and OLAP," JPhCS, vol. 1530, p. 012058, 2020.

[19] A. K. Hamoud, H. N. Hussien, A. A. Fadhil, and Z. R. Ekal, "Improving Service Quality Using Consumers' Complaints Data Mart which Effect on Financial Customer Satisfaction," in Imam Al-Kadhum International Conference for Modern Applications of Information and Communication Technology (MAICT), 2020, p. 012060.

[20] A. S. Girsang, G. Arisandi, C. Elysisa, and M. H. Saragih, "Decision support system using data warehouse for retail system," in Journal of Physics: Conference Series, 2019, p. 012007.

[21] R. Katuwal, P. N. Suganthan, and L. Zhang, "An ensemble of decision trees with random vector functional link networks for multi-class classification," Applied Soft Computing, vol. 70, pp. 1146-1153, 2018.

[22] A. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting student performance in higher education institutions using decision tree analysis," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, pp. 26-31, 2018.

[23] A. Hamoud, "Selection of best decision tree algorithm for prediction and classification of students’ action," American International Journal of Research in Science, Technology, Engineering & Mathematics, vol. 16, pp. 26-32, 2016.

[24] A. Hamoud, A. Humadi, W. A. Awadh, and A. S. Hashim, "Students’ success prediction based on Bayes algorithms," International Journal of Computer Applications, vol. 178, pp. 6-12, 2017.

[25] A. K. Hamoud and A. M. Humadi, "Student’s Success Prediction Model Based on Artificial Neural Networks (ANN) and A Combination of Feature Selection Methods," Journal of Southwest Jiaotong University, vol. 54, 2019.

[26] A. K. HAMOUD, "CLASSIFYING STUDENTS'ANSWERS USING CLUSTERING ALGORITHMS BASED ON PRINCIPLE COMPONENT ANALYSIS," Journal of Theoretical & Applied Information Technology, vol. 96, 2018.

[27] A. S. Hashima, A. K. Hamoud, and W. A. Awadh, "Analyzing students’ answers using association rule mining based on feature selection," Journal of Southwest Jiaotong University, vol. 53, 2018.

[28] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," in IOP Conference Series: Materials Science and Engineering, 2020, p. 032019.

[29] A. Hamoud and T. A. S. Obaid, "Design and Implementation Data Warehouse to Support Clinical

Decisions Using OLAP and KPI," Department of Computer Science, University of Basrah, 2013.

[30] G. Roccasalva, "Towards a DSS: A Toolkit for Processes of Co-designing," in Project and Design Literacy as Cornerstones of Smart Education, ed: Springer, 2020, pp. 49-52.

[31] H. Wang, P. Huang, and X. Chen, "Research and Application of a Multidimensional Association Rules Mining Method Based on OLAP," International Journal of Information Technology and Web Engineering (IJITWE), vol. 16, pp. 75-94, 2021.

[32] K. Letrache, O. El Beggar, and M. Ramdani, "OLAP cube partitioning based on association rules method," Applied Intelligence, vol. 49, pp. 420-434, 2019.

[33] A. Yadav, "Improving the Performance of Multidimensional Clinical Data for OLAP using an Optimized Data Clustering approach," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, pp. 3269-3275, 2021.

[34] J. García-Tobar, "Study of Indoor Radon Using Data Mining Models Based on OLAP Cubes," Physical Science International Journal, pp. 53-61, 2020.

[35] X.-H. Zhou and X.-M. Zhang, "The application of OLAP and Data mining technology in the analysis of book lending," in 2017 2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017), 2017.

[36] A. Lamani, B. Erraha, M. Elkyal, and A. Sair, "Data mining techniques application for prediction in OLAP cube," International Journal of Electrical & Computer Engineering (2088-8708), vol. 9, 2019.

[37] D. Putri and I. Sitanggang, "Clustering module in OLAP for horticultural crops using SpagoBI," in IOP Conference Series: Earth and Environmental Science, 2017, p. 012001.

[38] M. I. Moly, O. Roy, and M. A. Hossain, "An Advanced ETL Technique for Error Free Data in Data Warehousing Environment," 2019.

[39] N. Biswas, A. Sarkar, and K. C. Mondal, "Efficient incremental loading in ETL processing for real-time data integration," Innovations in Systems and Software Engineering, vol. 16, pp. 53-61, 2020.

[40] T. M. Al Taleb, S. Hasan, and Y. Y. Mahd, "On-line Analytical Processing (OLAP) Operation for Outpatient Healthcare," Iraqi Journal of Science, pp. 225-231, 2021.

[41] E. Pourabbas, "Providing accurate answers to OLAP queries based on standardized moments of data cubes," Information Systems, vol. 94, p. 101588, 2020.

[42] DAHR, JASIM MOHAMMED, et al. "IMPLEMENTING SALES DECISION SUPPORT SYSTEM USING DATA MART BASED ON OLAP, KPI, AND DATA MINING APPROACHES." Journal of Engineering Science and Technology 17.1 (2022): 0275-0293.

[43] Hamoud, Alaa Khalaf, et al. "Implementing data-driven decision support system based on independent educational data mart." International Journal of Electrical & Computer Engineering (2088-8708) 11.6 (2021).

[44] R. Rad, Microsoft SQL Server 2014 Business Intelligence Development Beginner's Guide: Packt Publishing Ltd, 2014.

[45] M. Russo, "SSAS tabular as analytical engine," SQLBI Article, 2014.

[46] J. Cheng and R. Greiner, "Comparing Bayesian network classifiers," arXiv preprint arXiv:1301.6684, 2013.

[47] J. R. L. a. G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol. 33, pp. 159-174, March 1977.

[48] Etaiwi, Wael, Dima Suleiman, and Arafat Awajan. "Deep Learning Based Techniques for Sentiment Analysis: A Survey." Informatica 45.7 (2021).

[49] Al, Noor M. Al-Moosawi M., and Raidah Salim Khudeyer. "ResNet-34/DR: A Residual Convolutional Neural Network for the Diagnosis of Diabetic Retinopathy." Informatica 45.7 (2021).