A Generic Framework for Evaluating Academic Performance (FEAP) in Nigerian Higher Institutions of Learning

A. S. Ahmadu[1], S. Boukari[2], E. J. Garba[3], A. Y. Gital[4]
[1]Department of Computer Science MAUTECH, Yola: ahmaduasabe@mautech.edu.ng
[2]Mathematics Programme ATBU Bauchi: bsouley2001@yahoo.com
[3]Department of Computer Science MAUTECH, Yola: e.j.garba@mautech.edu.ng
[4]Mathematics programme ATBU Bauchi: asgital@gmail.com

Abstract

Educational Data Mining is emerging as a tool for structuring and storage of academic records of students in a form that is adaptive for evaluation, predicting failure or success or discovering factors that are responsible for performance using the concept learned from the massive accumulated database. Experiments have shown that no single machine learning scheme is appropriate to all data mining problems (Witten and Frank, 2005). We have modelled a generic Framework for Evaluating Academic Performance (FEAP) with its various components which can be used with any of the following recommended algorithms: Logistics Regression, Decision Tree, Waikato Environment for Knowledge Analysis (WEKA), Artificial Intelligence (AI)/Decision Support Systems, SQL, and Knowledge Discovery in Databases (KDD). We have also design an algorithm for the selection of candidates for admission which is part of the Human Leaning (HL) functionality.

1. Introduction

Predicting the success or failure of a student in a course or a programme is a problem that has recently been addressed using data mining techniques. However, literatures have shown that there is still no consensus on what is the best set of variables that may lead to accurate models. Moreover, the problem is quite complex and appears to be very dependent on the data set used (Kabakchieva, 2013; Pedro *et al*. 2014).

One of the most serious problems higher education institutions have to deal with is related to students failing to reach the goals of educational success which represent a real threat to both the institution and the students themselves (Pedro *et al.* 2014).In general, academic achievement depends on the interaction of physical, physiological, social and psychological factors (Hayes and Orells, 1993). In line with these assertions, factors that play very important role in academic performance have been identified and designed in a format that information can be captured from students during registration exercise to be subsequently integrated into a single repository with Post Unified Tertiary Matriculation Exams (UTME) screening database and the operational database to form the Data Warehouse (DW). The model used by Bienkowski, *et al.* (2012), adopted and remodelled by Pedro *et al.* (2014) was considered as the based model.

2 Data Mining

Data Mining is the application of algorithms for extracting knowledge from large data or data warehouse. It is a model-building step (the steps can vary depending on the model). There are tools that help in either building the model or mining the data. In data mining, not a single technique is preferred over the other; a method is preferred depending on the kind of problem to be solved. Multiple methods can be integrated in order to solve a problem. Examples of these methods include: Statistical, Decision Support Systems, Database Management and Warehousing, Machine Learning, Visualization etc. (Philip and Pedro, 1999).

3 Data Warehouses

A data warehouse as a storehouse is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. It is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager and analyst) to make better and faster decisions (Saagari *et al.* 2013).

A data warehouse draws data from operational systems, but is physically separate and serves a different purpose. Operational systems have their own databases and are used for transaction processing. A data warehouse has its own database and is used to support decision making.

4 Logistic Regression

Logistic Regression can be binomial (binary), ordinal or multinomial. Binary Logistic Regression handles situations where the outcomes of the dependent variable (class) can have only two instances, such as “pass” or “fail”. The outcomes are usually coded as ‘0’ or ‘1‘as this leads to the most straight forward interpretation (Hosmer and Lemeshow, 2000).

Multinomial Logistic Regression deals with situations where the outcomes are more than two and are not ordered – for instance, diabetes, cancer, malaria fever, elephantiasis and so on.

Ordinal Logistic Regression deals with dependent variables that are ordered; for instance, first class, second class (upper division), second class (lower division) and third class.

5 Decision Tree

A decision tree is a tree where every non-terminal node represents a test of decision on the considered data item. Depending on the outcome of the test, one chooses a certain branch. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). When a terminal node is reached, a decision is made. Decision trees are powerful and popular in both classification and prediction. Decision trees represent rules; rules can

readily be expressed in English language so that humans can understand them. They are produced by algorithms that identify various ways of splitting a dataset into branch-like segments (Osofisan*et al.*, 2014).

6 Artificial Intelligence (AI)

AI is the second longest family line for data mining. This discipline is built upon heuristics (as opposed to statistics) attempts to apply human thought-like processing to statistical problems. It is the theory behind computers performing operations analogous to learning and decision making in humans as by an expert such as visual perception, speech recognition, decision-making and translation between languages. Because this approach requires vast amounts of computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices (Piatetsky-Shapiro, 1992).

7 WEKA

WEKA as an acronym stands for Waikato Environment for Knowledge Analysis. WEKA is portable, since it is fully implemented in the Java programming language and hence runs on almost any modern computing platform. The algorithms can either be applied directly to a dataset or called from a user Java source code.

Experience shows that no single machine learning scheme is appropriate to all data mining problems. The WEKA workbench is a collection of state-of-the-art machine learning algorithms. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization (Michael and Gordon, 2004). It is designed so that existing methods can easily and flexibly be tried out on new datasets. It provides extensive support for the whole process of data mining. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand (Witten and Frank, 2005*).*

8 Decision Support Systems

A Decision Support System (DSS) is often interfaced with a data-mining tool to help executives make more informed decisions. Though there are a variety of Decision Support Systems in the market today, their applications consist mostly of synthesizing the data to executives so that they can make more objective decisions based on the data analyzed. In this age of the Internet, On-line Analytical Processing (OLAP) is slowly replacing aging Decision Support Systems. Increasingly, OLAP and multi-dimensional analysis are used for decision support systems to find information from databases (Goil and Choudhary,1999).

9 Knowledge Discovery in Databases (KDD) Process

The KDD process is interactive and iterative, involving numerous steps with many decisions made by the user. Brachman and Anand (1996) give a practical view of the KDD process, emphasizing the

interactive nature of the process. Here, we broadly outline some of its basic steps as illustrated in Figure 1 by Fayyad *et al.* (1996) below:

i) Data Cleansing and Integration: Is a phase where irrelevant and noisy data are removed from the collection while multiple data sources are integrated into a single repository.

ii) Selection: At this step data relevant to the analysis is decided on and retrieved from the data collection.

iii) Data transformation: Also known as data consolidation is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

iv) Data Mining: It is the crucial step in which clever techniques are applied to extract patterns potentially useful.

v) Pattern Evaluation: Also known as interpretation; in this phase, interesting patterns representing knowledge are identified based on given measures.

vi) Knowledge Representation: is the final stage in which the discovered knowledge is visually represented to the user.

The KDD process can involve significant iteration and can contain loops between any two steps. Most of the previous works on KDD have focused on step iv (the data mining); however, the other steps are as important for the successful application of KDD in practice.

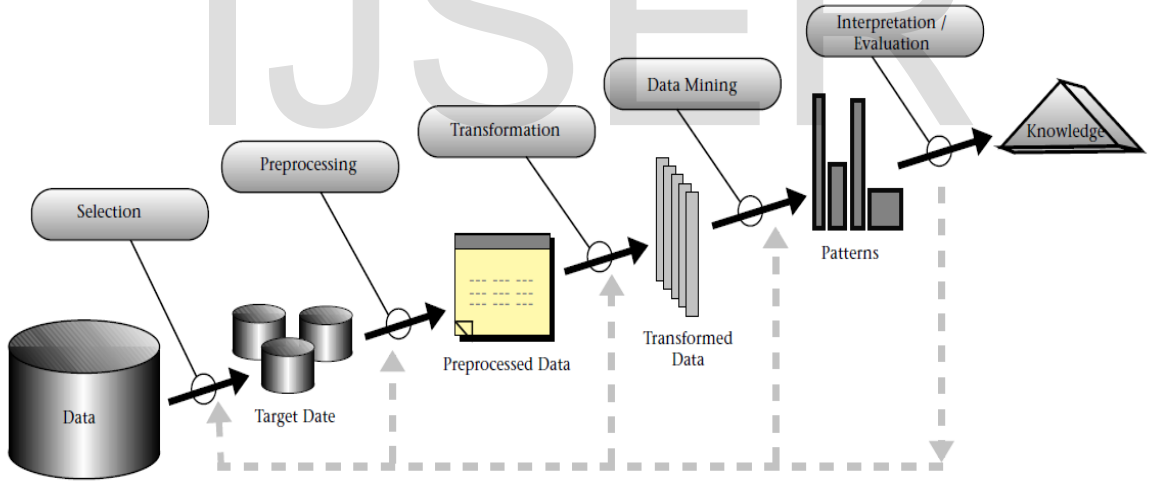

Figure 1: An Overview of the Steps That Compose the KDD Process. (Fayyad *et al.*, 1996)

10 SQL

SQL stands for Structured Query Language. SQL is used in communicating with a database. According to ANSI (American National Standards Institute), it is the standard language for relational database management systems. Some common relational database management systems that are SQL-based include Oracle, Sybase, Microsoft SQL Server, Microsoft Access, Ingres, etc. There are no variables in SQL. Everything about an SQL query must be specific – the literal values, the column names, and the table names.

In a parameter query, variables do not belong to the SQL itself; rather they belong to the environment that people use to submit SQL queries. In Microsoft Access, variables are in the GUI layer before the SQL query is sent to the database engine for processing. You usually write a parameter query for an end user to run. At runtime, the end user is asked to provide specific values for all the variables. These values are placed into the SQL statement before it is sent to the Database Management System (AbbasI, 2005).

11 Process of Designing a data warehouse

a) First Phase: Data extracted through Questionnaire

Academic achievement depends on more than what takes place within the walls of the school. Research generally indicates that characteristics outside the formal educational setting or non-school factors also have a lot to do with whether students are successful in school (Boccanfuso et al. 2010). In line with this assertion, factors that play very important role in academic performance were identified and designed in a questionnaire form and subsequently filled by students during registration.

b) Second Phase: Data Dictionary (DD)

Data dictionary is the description of the data document which contains all the atomic data elements, such as input, output, data storage and processing components. DD is very important in analyzing the logical model.

In this first phase, we have extracted the relevant fields from the data marts in the operational system and integrated it with the extracted fields from the questionnaire we plan to administer to students on factors affecting students' performances. These have been cleansed, integrated and normalised to form a single repository for the data warehouse. It is after the mentioned procedure has been accomplished that we were able to develop the DD for the data fields to be used in the data warehouse. What is required in developing a DD are field names and description in a format that facilitates the educational data analysis to be performed.

c) Third Phase: Database Schemas

The next step is creating the database tables which will form the database framework/schemas. The schemas form tables in Relational Database Management System architecture and it integrates all the attributes from all independent data marts into a single data warehouse (DW) for easy collection and dissemination of information. The tables to be created are Students Demographic table, Student Basic Information table, Departmental Table, Faculty Table, State Table, Local Government Area (LGA) Table, Behavioural Information Table, Parental Information Table, Financial table, Educational Information Table, Institutional Information Response Table, JAMB Table, O_Level results Table, Catchment Zone Table, Session Table.

d) Fourth Phase: Normalization

This phase normalises the tables generated in the database schemas. Normalisation is a three-step technique used to ensure that all tables are logically linked together and all fields in a table directly relates to the key.

12 Framework for Evaluating Academic Performance

We have adopted the design by Bienkowski *et al.* (2012) adapted by Pedro *et al.* (2014) as depicted in Figure2. We have made further modifications to suit the peculiarity of our environment as shown in Figure 3. These modifications led to the design of a generic Framework for Evaluating Academic Performance (FEAP). We have also made an improvement on it by introducing data warehouse and a hybrid system that combines Human Learning (HL) and Machine Learning (ML). SQL statements such as total number of students that pass or fail or are within a particular class of degree, list of students that meet particular criteria, the best candidate in a course or programme or institution and so forth will be integrated with the algorithms that would be developed using Java programming language to form the HL. At runtime, the end user is asked to provide specific values for all the variables, like University Name, Faculty, Department, course, State etc.

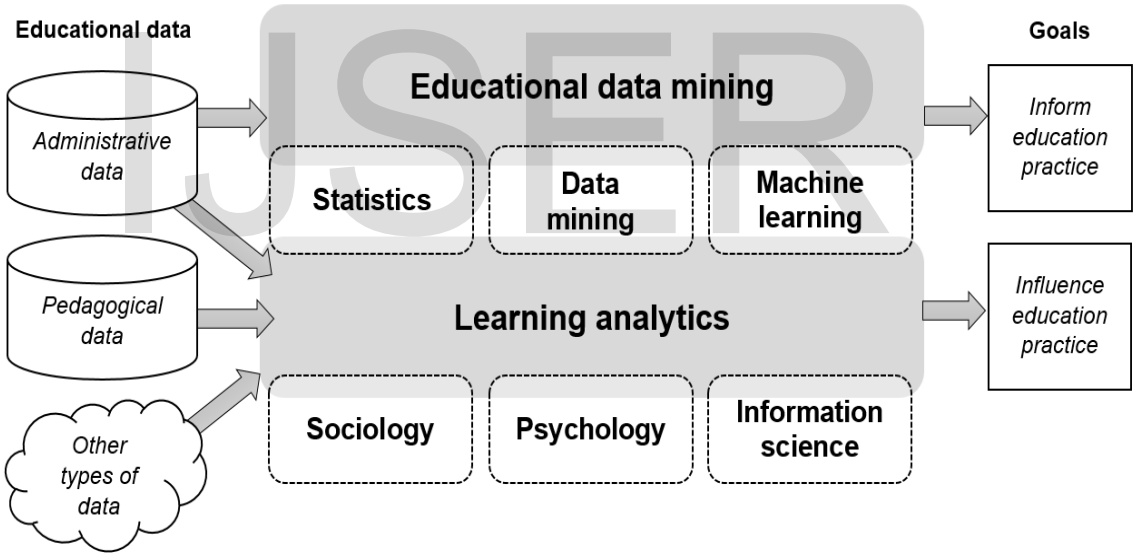

Figure 2: Educational big data (Source: Bienkowski, Feng and Means, 2012)

The FEAP consists of various functional components and interactions as shown in Figure3. The working principle of the system (with more detailed functionalities) is depicted in Figure4.

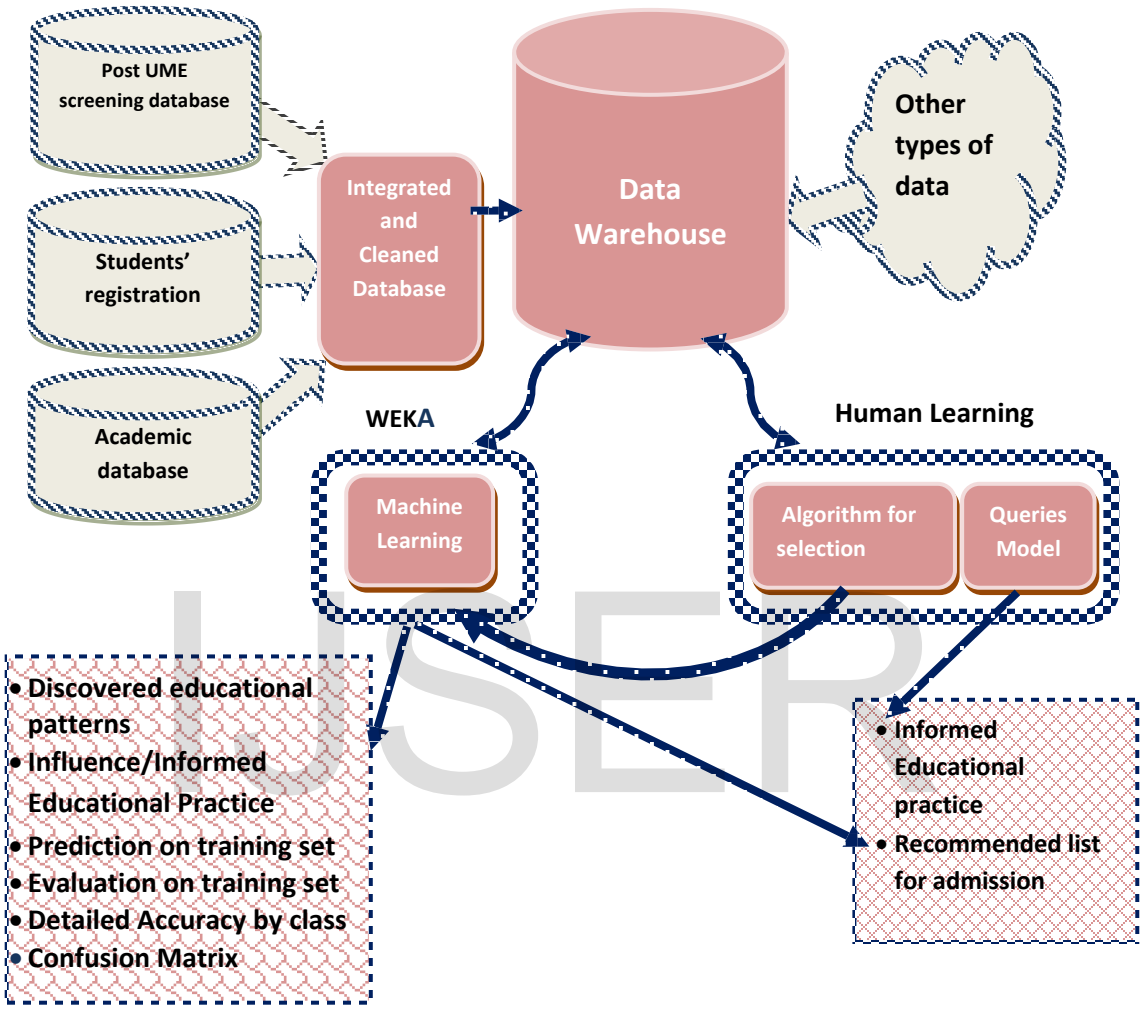

Figure 3: Generic Framework for Evaluating Academic Performance (FEAP)

Human learning (HL) component of the FEAP handles short listing of suitably qualified candidates for admission for on word recommended placement by WEKA and also the dissemination of informed practice that would be useful to an individual, committee or management. These algorithms will be developed using Java programming language. The HL (using parameter queries) will also handle the customised SQL statements such as: List/count of graduating students in the University/Faculty/Department, list/count of female students in the University/Faculty/Department, list/count of graduating student according to class in the University/Faculty/Department, list/count of students from a particular state/LGA, percentage/count of students whose mode of entry UTME/DE/Pre-degree, etc.

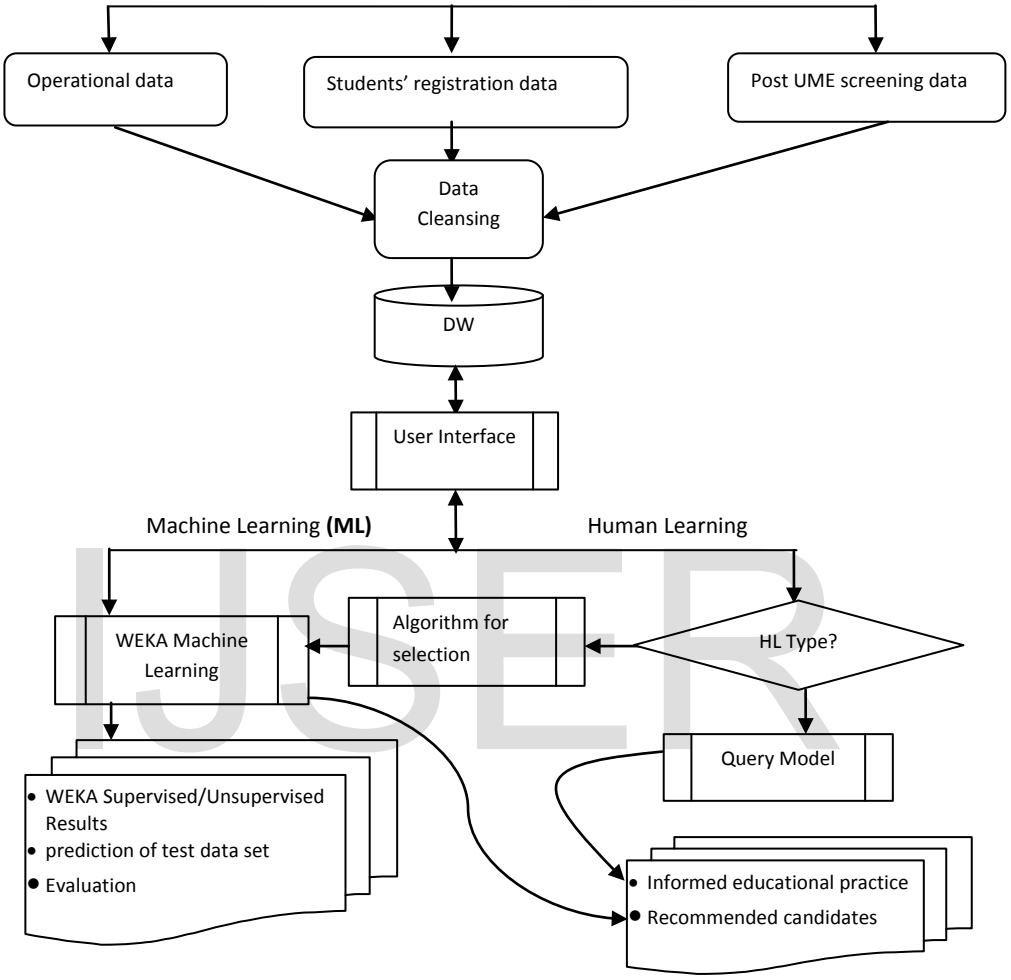

Figure 4: The working principles of the hybrid system

13 Samples of the algorithms that will be used in the Human Learning

14.1 Algorithms Development for Admissions Based on JAMB Standard

JAMB Allocated carrying capacity for a course = ACC

Determine UTME carrying Capacity for the course = UTMEDCC<= ACC

Determine DE carrying Capacity for the course = DEDCC<UTME<=ACC

IF Candidate is UTME mode of admission AND Candidate Aggregate >= 180 THEN

IF Candidate has 5 credits with English Language And Mathematics AND candidate Aggregate amongst the TOP 45% UTME determined carrying capacity of a course (UTMEDCC) on merit THEN

Admit candidate

ELSE IF Candidate amongst NEXT TOP 35% UTME determined carrying capacity of a course (UTMEDCC) AND from catchment area of the university AND number in each state not greater than 35% UTMEDCC /(number of states in the catchment area) AND Candidate NOT yet admitted THEN

Admit Candidate

ELSE IF Candidate NOT yet admitted AND from educationally disadvantaged state AND is amongst 20% UTMEDCC AND number in each state NOT greater than 20% UTMEDCC /number in each state

Admit Candidate

ELSE

Drop Candidate

ENDIF ENDIF ENDIF ENDIF

ELSE IF Candidate is DE mode of admission AND Candidate GRADE = “Distinction” OR “Upper Credit” OR “Lower Credit” OR (FOR IJMB Candidates Total Score > = 8 AND Candidate Pass the three subjects) THEN

IF Candidate has 5 credits with English Language AND Mathematics AND candidate score amongst the first 45% of DE determined carrying capacity of a course (DEDCC) on merit THEN

Admit candidate

ELSE IF Candidate amongst Next 35% of DEDCC AND from catchment area of the university AND number in each state not greater than 35% of DEDCC /(number of states in the catchment area) AND Candidate NOT yet admitted THEN

Admit Candidate

ELSE IF Candidate NOT yet admitted AND from educationally disadvantaged state AND is amongst 20% of DEDCC AND number in each state NOT greater than 20% of DEDCC /number in each state

Admit Candidate

ELSE

Drop Candidate

ENDIF ENDIF ENDIF ENDIF

ELSE

drop candidate

ENDIF.

14 Conclusion

We have modelled a Generic Framework for Evaluating Academic Performance (FEAP) and its various functionalities. An algorithm for selecting suitably qualified candidates for admission which is part of the human learning system in our model has been designed, and is to be used in combination with the model for placement of new students into various programmes. This will subsequently be fully developed and integrated into the predictive model in our future work. The FEAP will be used in tandem with the designed data warehouse (DW) architecture which has integrated all the independent Marts into a single repository for convenient data mining and information dissemination.

References

AbbasI, A. (2005); MS Access 2007 Step by Step,Takveen, Inc. South River, NJ 08882.

BAKER, R. S. J. D., Yacci, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, Article 1, Vol 1, No 1, pp3-16. www.educationaldatamining.org (Accessed 29/05/2017).

Bienkowski, M., Feng, M. & Means, B. (2012).*Enhancing teaching and learning through educational data mining and learning analytics:* An issue brief. Department of Education's (ED)Office of Educational Technology (pp. 1–57). Washington, D.C.California State University, Sacramento California, 95819 pp 10-20

Boccanfuso, C., Moore, K.A., and Whitney, C. (2010). Ten Ways To Promote Educational Achievement And Attainment Beyond The Classroom: Brief Research to Results trend Publication no 2010-16, pp 1-13. *cyitc.org/wp-content/uploads/2013/11/Education-AA.pdf*

Brachman, R., and Anand, T. (1996). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In *Advances in Knowledge Discoveryand Data Mining*, 37–58

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases AI Magazine Volume 17 Number 3 (© AAAI) Copyright © 1996, American Association for Artificial Intelligence. All rights reserved. 0738-4602 pp 37-53

https://www.aaai.org (Accessed 29/05/2017)

Goil, S. and Choudhary, A.(1999). Design and Implementation of a Scalable Parallel System for Multidimensional Analysis and OLAP.13th International and 10th Symposium on Parallel and Distributed Processing.

Hayes, N. And Orells, S. (1993); Psychology: an introduction. Longman, HarlowHosmer, D. W. And Lemeshow, S. (2000), Applied logistics regression (2nd Ed) New York, NY: John Wiley and Sons Inc.

Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification.Cybernetics and Information Technologies, 13(1), 61–72.

Kline, P. (1993). An Easy Guide to Factor Analysis in knowledge tracing: Proceedings of the 14th international conference on artificial intelligence in education, pp 531–538.

Michael, J. A. B. and Gordon S. L. (2004), Data Mining Techniques, 2nd ed., Wiley Publishing Inc., USA.www.cs.waikato.ac.nz/ml/weka (Accessed 15/03/2016).

Minaei-Bidgoli, B., Kashy, D., Kortmeyer, G. and Punch, W. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. In: Frontiers in education. FIE 33rd Annual, pp T2A 13–18.

Osofisan, A.O., Adeyemo, O.O., and Oluwasusi S.T. (2014).Emperical Study of Decision Tree and Artificial Neural Network Algorithm for Mining Educational Database.African Journal of Computing and ICT-ISSN 2006-1781.www.ajocict.net (Accessed 20/01/2015).

Pedro, S., João, M.M., and Carlos, S. (2014). Educational Data Mining: preliminary results at University of Porto, Pulished by Morgan Kaufmann, New Zealand.www.up.pt (Accessed 20/01/2015).

Philip, C. and Pedro, M. (1999). On the use of support vector machines for phonetic classification Conference Paper DOI: 10.1109/ICASSP.1999.759734 · Source: IEEE Xplore
Conference: Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on, Volume: 2

Piatetsky-Shapiro, G. (1992). Knowledge Discovery in Databases: An Overview. AI Magazine. 213-228 and 289–300.www.tesisenred.net/bitstream/handle/10803/9159/Tesi-part3(Accessed 10/06/2014).

Reddy, G. S., Srinivasu, R., Rao, M. P. C., and Rikkula, S. R. (2010), Data Warehousing, Data Mining, OLAP and OLTP Technologies Are Essential Elements to Support Decision-Making Process in Industries (IJCSE) International Journal on Computer Science and Engineering, ISSN : 0975-3397 Vol. 02, No. 09, 2010, 2865-2873, pp 2865- 2872
https://www.researchgate.net (Accessed 29/05/2017).

Saagari, S., Anusha, P. D., Priyanka, C. L. and Sailaja, V.S.S.N. (2013). Data Warehousing, Data Mining, OLAP and OLTP Technologies are Essential Elements to Support Decision-Making

Process in Industries. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume 2(6) pp 88-93.
www.ijitee.org/attachments/file/v2i6/F0801052613.pdf (Accessed 10/06/2014).

Witten, I. H. and Frank, E. (2005); "Data Mining Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann Publishers is an imprint of Elsevier.500 Sansome Street, Suite 400 San Francisco, CA 94111. pp 267-320
www.cs.waikato.ac.nz/ml/weka (Access 15/03/2016).