AN INTEGRATED SYSTEM FRAMEWORK FOR PREDICTING STUDENTS' ACADEMIC PERFORMANCE IN HIGHER EDUCATIONAL INSTITUTIONS

Olugbenga Adejo and Thomas Connolly

School of Engineering and Computing, University of the West of Scotland, Paisley, United Kingdom

ABSTRACT

Accurate prediction and early identification of student at-risk of attrition are of high concern for higher educational institutions (HEIs). It is of a great importance not only to the students but also to the educational administrators and the institutions in the areas of improving academic quality and efficient utilisation of the available resources for effective intervention. However, despite the different frameworks and various models that researchers have used across institutions for predicting performance, only negligible success has been recorded in terms of accuracy, efficiency and reduction of student attrition. This has been attributed to the inadequate and selective use of variables for the predictive models. This paper presents a multi-dimensional and an integrated system framework that involves considerable learners' input and engagement in predicting their academic performance and intervention in HEIs. The purpose and functionality of the framework are to produce a comprehensive, unbiased and efficient way of predicting student performance that its implementation is based upon multi-sources data and database system. It makes use of student demographic and learning management system (LMS) data from the institutional databases as well as the student psychosocial-personality (SPP) data from the survey collected from the student to predict performance. The proposed approach will be robust, generalizable, and possibly give a prediction at a higher level of accuracy that educational administrators can rely on for providing timely intervention to students.

KEYWORDS

Prediction, Student performance, Higher education, integrated system, framework.

1. INTRODUCTION

Continuous progress in education domain has been going on for many years. Among the signs of development in the sector include exponential growth in data generation and technological advancement. In addition to this, there has been a significant rise in student enrolment across all segments of the education. However, the increase in student enrolment has not necessary translate to increase in retention, progression and graduation rate. The higher institutions attrition rates have remained unabated, ranging from between 8% at some institutions in developed countries to over 70% in developing countries of the world [1]. In the United Kingdom, the Higher Educational Statistical Agency (HESA) data on the dropout rate from the UK Higher Education Institutions (HEIs) over the past five years has shown a progressive increase in the dropout and non-continuation of the UK domicile students especially the first-degree entrant. The data from the HESA reveals an increase of 6.7% in 2011/12 to 7.2% in 2013/14 of non-continued undergraduate students and the projection, based on this trend and previous studies, shows that this non-continuation rate could increase to a total of over 30% by the end of the fourth year in

DOI:10.5121/ijcsit.2017.93013

most HEIs [2]. This has led many institutions to diverse ways of reducing student attrition by identifying the student at-risk of attrition early enough using predictive analytic.

Currently, historical and cognitive data of students stored in the institutional databases are used as a model for the measurement and prediction of the performance of the current students. The prediction results can then be used to provide necessary intervention and support for the at–risk student identified. However, the accuracy of these models in predicting student performance in higher educational institution has been of great challenges [3] and this has been attributed to the following;

- Lack of standardisation and comprehensive framework for data modelling.
- Limited used of variables and selective use of variables for modelling. In addition, building a model on the wrong data population (test sample size) can lead to inaccurate prediction.
- Use of single or weak classifiers algorithm which often affects model quality.

From all these different perspectives, it is evident that most of the data required for the successful and accurate prediction of student performance cannot be derived from the institutional databases only, the majority of factors or causes of student action and decision are only derivable from the students. Just as learners success and performance are not the sole responsibility of the teacher or educational administrator alone, but the bulk of the work lies with the learners or students themselves. This non-engagement of students in their performance prediction has been the major limitation to previous frameworks. Though the models have provided interesting concepts, they failed in meeting the requirement for bringing a solution to the new age challenges in education domain.

Therefore, this paper has proposed a holistic and integrated framework aims at providing all the necessary data inputs and functionalities that will help to predict student's academic performance accurately and efficiently. The process of developing the framework, however, takes into consideration different data sources required for accurate prediction as well as the inclusion of student input into prediction process.

The second section of this paper presents the general overview of the existing methods and framework for student performance prediction from the literature review. The next section discusses on the proposed framework by presenting the concepts, methodology and the comparison of our framework with the existing frameworks.

The paper concludes with the summary of the work.

2. LITERATURE REVIEW OF RELATED WORK

Several research works have explored different ways to improve the student academic performance prediction with the use of different types of variables and algorithms as well as identifying the best way to increase the accuracy and the efficiency of the predictive model [4].

In research papers by [5], comprehensive summaries of several predictive frameworks, attributes and methods that have been used in prediction of student performance in the educational sector were discussed and analysed. The importance of student performance predictions to the various stakeholders was also pointed out. The reasons are to identify the student at risk of attrition early enough in order to provide necessary support and intervention for them with the goals of reducing attrition, increasing retention, performance and graduation rate. A diagrammatic representation of the goals of predicting student performance is shown in Fig.1.

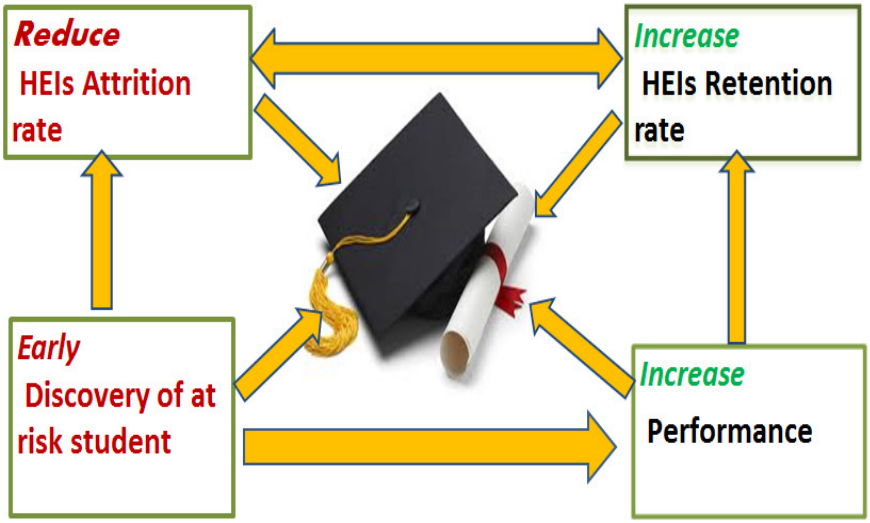

Figure 1.General goals of student performance prediction

Moreover, it has been shown that different studies have been carried out in the area of student prediction as early as 1926, with the first set of studies on the effect of student "mortality" on their academic failure [6] and this has been followed by different theories such as [7] and [8] Models of student attrition, [9] student attrition Models, [10] model and [11] Input – Environment-Output model in higher institutions as well as [12] student retention model.

However, in recent time, emphasis on the predicting student performance has been on the use of their cognitive ability, log activities in learning management system as well as the student demographic attributes. [13], [14], [15] and [16] used demographic data along with students scores to predict their performance, using machine learning languages such as Artificial Neural Network, Support Vector Machine and Naïve Bayes algorithms. This technique is a move away from the commonly used traditional logistic regression.

[17], [18], [19] and [20] also predicted student final grade using the log data extracted from a web-based system such as LMS. They make use of variables such as the number of online sessions, the frequency of login, the number of the original posts read/ created the number of follow-up post created, the number of content page viewed and the number of posts read. However, despite the prominence of "frequency of login" as a factor for the measurement of student performance, some few studies went deeper to look at the quality of participation instead of quantity by looking at timing, the volume and consistency of access or log in which actually gave more precise result when included. In summary, the most commonly used predictor variables extracted from LMS are a number of posts viewed, the total amount of time spent online, the number of access to course materials and login frequency.

In different studies, [21], [22], [23] used survey questionnaire techniques to collect student intrinsic and personality data that are not readily available in the database for predicting student performance They measured the effects of personality traits, learning styles, personality, learning strategies and motivation factors and psychological well-being on the academic performance of students.

In the same way, [22] used a questionnaire to collect behavioural (psychometric) data for predicting students' performance in Malaysia University. The data collected include their Interest, study (engage) time, study behaviour, belief and family support. The result shows a strong correlation between student mental condition and their performance. Also, [23] used a short questionnaire made up of five different personality factors along with learning style of the student, their psychological well-being as well as educational achievement on academic performance. Moreover, [21] used personality, motivation and learning strategies variables gathered between the year 2010-2012 alongside six different classification algorithms to predict student learning progression and achievement. The result from these studies shows there is a strong correlation between the variables examined and performance of the student. However, these researchers suggested the inclusion of more variables outside the University databases in order to improve the model and the accuracy of prediction.

In a similar study carried out by [24], they developed three predictive models to compare the performance of survey-based retention methodology, open data sources and Institutional internal databases using analytical approaches. The results found that the survey-based model performed better in accuracy, sensitivity and specificity than the institutional internal databases when logistic regression was used. The study also discovered that when the questionnaire was combined with institutional databases, the performance improved compared to when solely institutional databases were used.

Finally, looking at the review of student performance as a whole, various researchers have shown that the main reasons for low performance and attrition of student from HEIs are not those that are often recorded official, they are external factors that are out of control of the HEIs. Most of these factors are student dependent and as such involve engaging the student in providing answers to them through the use of survey or interview. Moreover, it should be noted that these factors that affect and determine student performance are not solitary in nature but are interconnected, interrelated and interdependence (Figure 2). [25] [26] and [27] suggested that there is a possibility to improve the student prediction accuracy with the used of more independent variables or attributes that are outside the database of the University system. Therefore, there is a need for research to develop a new framework that is comprehensive and holistic in its approach.

3. THE PROPOSED CONCEPTUAL FRAMEWORK

The idea behind this framework is focused on the comprehensive approach to predict student performance with efficiency and accuracy. The performance prediction framework presented will generally make use of the following six variable domains that have great influence on student performance vis a viz psychological, cognitive, Economical, personality, demographic and institutional domains (Figure 2)

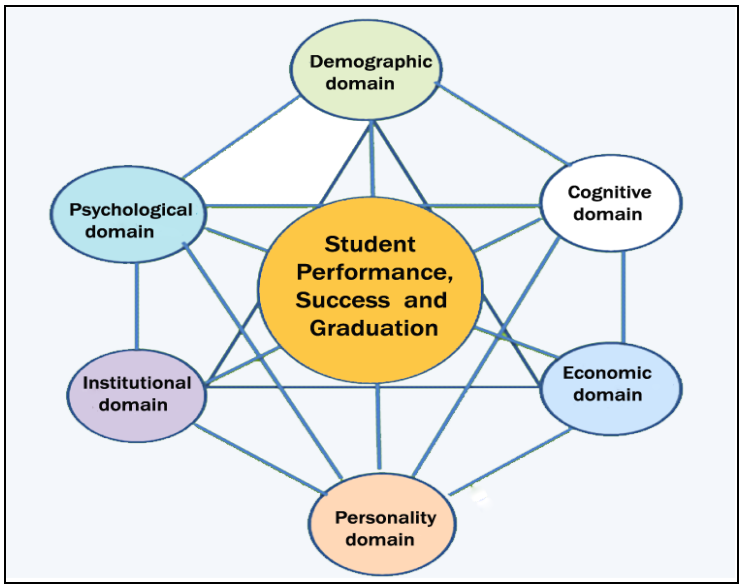

Figure 2. An illustrative six interconnected variable-domains.

Each of the six domains contributes to the performance measurement of the student and are made up of attributes that work individually and jointly for learners success. However, the degree of complexity and impact of each domain on student performance is variable.

- Psychological domain – include self-efficacy, achievement, goal, interest
- Cognitive domain - includes examination score, presentation skill, intellectual ability
- Personality domain – includes motivation, learning style, study time, habit, ICT skill, online activities.
- Economic domain – includes income, income distribution status, parent financial status employment status
- Demographic domain – includes age, gender, location, ethnic, marital status, disability
- Institutional domain- includes course programme, learning environment, institutional support, course workload.

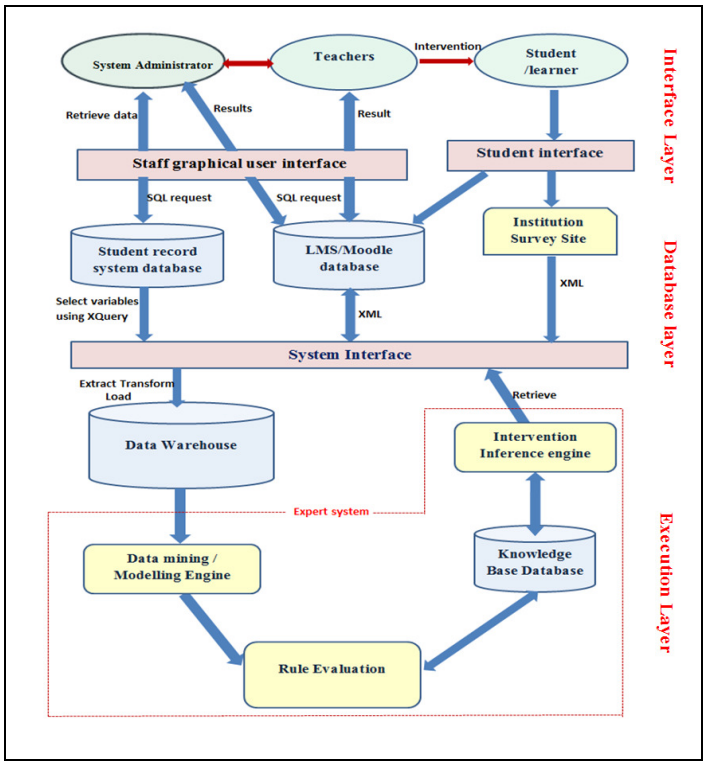

Figure 3. Architectural framework of the holistic student performance prediction system

Figure 3 depicts the architectural framework of the proposed student performance prediction system, which comprises of three different layers, a) ***the User Interface layer***, b) ***the database system layer*** and c) ***execution or expert system layer***. Each of the layers is explained below.

3.1 Graphical User Interface layer

This can also be referred to as the view layer. It hosts the Graphical User Interfaces (GUIs) of the framework. It is the layer that is presentable to the user and acts as the entry point to the system as well as provides necessary control and functionalities to the end users. It is divided into two categories based on the log-in interface, the staff graphical user interface and the student interface. With different levels of authentication, the staff and student can log in and carry out various activities.

3.2 Database systems layer

The database system layer provides access to the different databases available in higher education institution repository from where data abstraction for further analysis takes place.
This is made up two categories of databases;

- Firstly, the institutional databases which are made up Student Record System (SRS) and Learning Management System (Moodle) databases. Student demographic data such as age, gender, study mode, location, marital status as well as their online activities or actions such as number of time login, the frequency of login, total time spent online, assessment submission, forum activities and contribution can be extracted.

- Secondly, the Student Psychosocial-Personality (SPP) database which manages students yearly psychosocial and personalities factors that are a not constant. Such SPP attributes include parental status, financial status, family responsibility, job workload, learning style, learning habit, parental academic level, parental support, academic environment, anxiety, student goal and interest, university support system, technology and social media impact.

3.3 Execution / Expert System layer

The execution or expert system layer consists of different units for modelling, evaluation and decision recommendation. It is faster and has low error rate than a human expert. The different units that made up the expert system are briefly explained below;

- Datamining / Modelling Engine - This applies the selected data mining techniques such as characterisation, classification, relationship mining, outliner analysis and clustering to the filtered educational/learners' data from the data warehouse. This will involve the application of the association mining rule to the training phases for generation of rules and patterns.
- Rule Evaluation Engine – This uses logic and applies the set out rules, in a different form to the learner's data to produce outcomes. It makes use of declarative programming or conditional statement (IF and a THEN) to set out "what to do" and "how to do it" to produce the outcomes.
- Knowledge-based database - By making use of the rule engine, it creates a repository of knowledge by storing relevant information, rules and cases that can be executed on any data.
- Intervention and Inference Engine- This gets and combines information from the knowledge base to provide answers, suggestion, types and mode of intervention necessary for each student. It suggests and provides the necessary as well as a unique intervention strategy that the administrator and staff can use to support the student. Simply, it acts as a recommender of the personalised mode of intervention to the administrator and assigned staff about the learner. In addition, the inference engine also presents the LA dashboard, not just to the module tutor and the administrator, but also to the student.

4. CONCLUSIONS

This paper has proposed a framework for predicting student academic performance with efficiency and accuracy. The system architecture and different variable domains are also presented. The framework describes the sources, types and process of data to modelling and finally decision making. It also describes the algorithm selective processes that occur in the modelling engine stage in order to select the best predictive modellers.

The proposed structure is deemed to be flexible, scalable and will remain robust in the application. The framework is also expected to be generalizable as the data extracted from the data warehouse is standardised. One other advantage of using the proposed approach is its ability to fully engage the student in a matter relating to the decision being taken with regard to their performance and academic future. The new system will provide an enhanced and highly efficient system and model that helps in early identification of student-at-risk of attrition with high accuracy.

However, the proposed framework (which is under pilot application) still needs to be empirically evaluated and validated before any conclusions will be made. In addition, the ethical issues relating to the use of this system need to be properly researched and investigated. Beyond this, the

framework provides great opportunities to accurately and efficiently improve the performance prediction accuracy of students in higher education institutions.

REFERENCES

[1] Braunstein, Andrew W., Mary Lesser, & Donn R. Pescatrice (2006) "The business of freshmen student retention: Financial, institutional, and external factors." The Journal of Business and Economic Studies Vol.12, No. 2, pp.33.

[2] HESA (2014). https://www.hesa.ac.uk/data-and-analysis/performance-indicators/non-continuation.

[3] Yadav, S.K., Bharadwaj, B. K. & Pal, S. (2012). "Mining Educational Data to Predict Student's Retention :A Comparative Study", International Journal of Computer Science and Information Security (IJCSIS), Vol.10, No.2

[4] Bekele, R. & McPherson, M., (2011). "A Bayesian performance prediction model for mathematics education: A prototypical approach for effective group composition". British Journal of Educational Technology, Vol.4, No.3, pp.395-416.

[5] Aljohani, O., (2016). "A comprehensive review of the major studies and theoretical models of student retention in higher education. Higher Education Studies,Vol. 6, No.2, p.1.

[6] Summerskill, J. (1962) Dropout from college. In N.Sanford(ed). The American college, New York, Wiley

[7] Spady W.G., (1971). Dropouts from higher education: Toward an empirical model. Interchange, Vol.2, No.3, pp.38-62

[8] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of educational research, Vol.45, No.1, pp.89-125

[9] Bean, J. (1980). "Dropouts and turnover: The synthesis and test of a causal model of student attrition. Research in Higher Education", Vol.12, No.2, pp.155-187. http://dx.doi.org/10.1007/BF00976194

[10] Pascarella, E. T., & Terenzini, P.T (1980). "Predicting freshman persistence and voluntary dropout decisions from a theoretical model". The Journal of Higher Education, Vol.51, No.1, pp.60-75,1980.

[11] Astin, A.W., (1984). " Student involvement: A developmental theory for higher education". Journal of college student personnel, Vol.25, No. 4, pp.297-308.

[12] Cabrera, A. F., Castaneda, M. B., Nora, A., & Hengstler, D. (1992). "The convergence between two theories of college persistence". The Journal of Higher Education, Vol.63, No.2, pp.143-164.

[13] Kotsiantis, S.B. & Pintelas, P.E., (2005), July. "Predicting students marks in Hellenic open university". In Advanced learning technologies, 2005. ICALT 2005. fifth IEEE international Conference on (pp. 664-668). IEEE.

[14] Oladokun, V. O., Adebanjo, A. T. & Charles-Owaba, O. E. (2008). "Predicting students' academic performance using artificial neural network: A case study of an engineering course". The Pacific Journal of Science and Technology, Vol.9, No.1, pp.72-79.

[15] Hoe, A.C.K., Ahmad, M.S., Hooi, T.C., Shanmugam, M., Gunasekaran, S.S., Cob, Z.C.& Ramasamy, A., (2013) "Analyzing students records to identify patterns of students' performance". In Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on (pp. 544-547). IEEE

[16] Ikbal, S., Tamhane, A., Sengupta, B., Chetlur, M., Ghosh, S. & Appleton, J. (2015). "On early prediction of risks in academic performance for students. IBM Journal of Research and Development, Vol.59, No.6, pp.5-1.

[17] Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer, G. & Punch, W., (2003). "Predicting student performance: an application of data mining methods with an educational web-based system". In Frontiers in education, 2003. FIE 2003 33rd annual (Vol. 1, pp. T2A-13). IEEE

[18] Romero, C., López, M.I., Luna, J.M. & Ventura, S., (2013). "Predicting students' final performance from participation in on-line discussion forums".Computers & Education, Vol.68, pp.458-472

[19] Agudo-Peregrina A.F., Iglesias-Pradas S., Conde-Gonzalez M.A., and Hernandez-Garcia A. (2014). "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning" Computers in Human Behavior, Vol.31, No.1, pp. 542-550.

[20] Cerezoa, R., Sánchez-Santillánb, M., Paule-Ruizb,M.P., & Núñeza J. (2016). "Students' LMS interaction patterns and their relationship with achievement: A case study in higher education", Computer and Education Vol. 96, May 2016, pp. 42–54

[21] Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E.(2011). Prediction of student academic performance by an application of data mining techniques. In International Conference on Management and Artificial Intelligence IPEDR, Vol.6, pp.110-114.

[22] Fariba, T.B. (2013). "Academic performance of virtual students based on their personality traits, learning styles and psychological well-being: A prediction". Procedia-Social and Behavioral Sciences, Vol.84, pp.112-116.

[23] Gray, Geraldine, Colm Mcguinness, and Philip Owende. (2016) "Non-Cognitive Factors of Learning as Early Indicators of Students-at-Risk of Failing in Tertiary Education." In Non-cognitive Skills and Factors in Educational Attainment, pp. 199-237. SensePublishers.

[24] Sarker, Farhana, Thanassis Tiropanis, and Hugh C. Davis.(2013) "Exploring student predictive model that relies on institutional databases and open data instead of traditional questionnaires." In Proceedings of the 22nd International Conference on World Wide Web, pp. 413-418. ACM.

[25] Romero, C., Romero, J.R., Luna, J.M. and Ventura, S., (2010). "Mining rare association rules from e-learning data". In Educational Data Mining 2010

[26] Rubiano, S.M.M. and Garcia, J.A.D., (2015). "Formulation of a predictive model for academic performance based on students' academic and demographic data". In Frontiers in Education Conference (FIE), 2015. 32614 2015. IEEE (pp. 1-7). IEEE

[27] Rusli, N.M., Ibrahim, Z. and Janor, R.M., (2008). "Predicting students' academic achievement: Comparison between logistic regression, artificial neural network, and Neuro-fuzzy". In Information Technology, 2008. ITSim 2008. International Symposium on (Vol. 1, pp.1-6). IEEE.