Learning Analytics and Online Courses: A Bayesian Belief Network Approach to Predict Success

Virendra Singh Nirban [1]*, Tanu Shukla [2] and Daksh Dave [3]

1. *Department of Humanities and Social Sciences, BITS Pilani, vsnirbanbits@gmail.com*
2. *Department of Humanities and Social Sciences, BITS Pilani, tanushukla8@gmail.com*
3. *Department of Electronics and Electrical Engineering, BITS Pilani,dakshdave51093@gmail.com*

lncs@springer.com

Abstract. In the field of educational research, Learning Analytics is one of the prevailing areas of exploration. The study explores a part of learning analytics using a Bayesian Networks (BN) model to predict the success of the course in the online mode of education. Through the simulation results, it was found that the BN approach can be used to suggest improved online instruction delivery methods, helping the instructors and students reform their practices to maintain a synergy for a successful running of the course. As the study was executed on engineering students, it could further be generalized using students of other streams for comprehensive understanding. The study reveals that the student synergy with the method of teaching, paper difficulty and Take-home assignments are found to be the main determinants of the success of E-Learning courses.
The study reveals that students synergy with the metho

Keywords: Bayesian Networks; e-Learning; Learning Analytics; Probability

1 Introduction

The development in technology, internet, infrastructure, and service sectors have made E-Learning very popular. There are different ways in which students prefer to learn. In the last decade, a boom has been witnessed in the space of the Ed-Tech domain wanting to democratize learning. Going forward, a degree may no longer stand the same importance as it does today. Education with the same content would be accessible to everyone and available online. The learning sector possesses vast amounts of records of student data. Kondo and Hatanaka (2019) believed that the usage of learning analytics by educational institutions has gained prominence due to its effectiveness in helping the institute make well-informed decisions. The advent of Artificial Intelligence has provided ample scope for improvement of E-Learning Platforms by providing intelligent and interactive environments to the students.

The current trend of the education system demands experimentation with different E-Learning Systems. Recent years have witnessed a transformation in the E-learning arena and there was a paradigm shift from offline to an online mode of education in the past 5 years. The study aims to use machine learning algorithms for analyzing the smooth operation of a course by extracting the characteristics of a student and the E-Learning Platform on the standardized data collected from numerous students to predict

the outcome of a course running successfully and help the instructor to design the course to ensure that there is a synergy that is maintained between the instructor and the student and there is maximum utilization of each other's effort. The paper is organized as follows- section 2 presents the review of the related and the past works done in the same research domain followed by section 3 covers the methodology used for the research, section 4 includes the results and findings followed by section 5 contains the discussion this is ended with section 6 containing the references.

2 Literature Review

To provide critical analysis, various facets of the research problem become mandatory to examine by exploring and comprehending recent research in the concerned area. Thereby, recent texts were studied on learning analytics, specifically using the Bayesian approach in the study. Chanaa and Faddouli (2018) stated that research is valuable on the use of Machine learning models which are currently gaining popularity in E-Learning. García et al. (2007) listed down the attributes that could be considered to predict the students' learning state with a very interesting approach. In their problem, "random variables represent the different dimensions of Felder's learning styles", Ueno and Okamoto (2007) proposed a model that builds a Bayesian Network to predict the final status of the learner and generate appropriate motivation messages to the learner.

The use of Bayesian Networks for predictive analysis in online education was more specifically described by Kao et al. (2009) where they proposed Bayesian Networks classification model in an E-Learning System where the relative efficiency of the system was reviewed rather than the output. The revolutionary breakthrough with the help of a 4-level learning progression model was used (West et al., 2012) to get the optimal Bayesian Network score. To prove this point further Chakraborty and Sinha (2016) tested a model which improved study material recommendations by evaluating the learning style of the students from the materials browsed and their test performance. A major innovation proposed by them was a Dynamic Bayesian Network.

For predicting the academic performance of the students using Bayesian Network Classifiers, Sundar (2013) did a study on a students database and helped the institution to identify the students that can potentially drop out. Sharabiani et al. (2014) proposed a model for predicting the grade of engineering students. Mahnane and Hafidi (2016) suggested a dynamic Bayesian Network that could detect students' learning states. Rajper et.al. (2016) surveyed to identify the student's E-Learning activities and predict their learning styles. Further, Chanaa and Faddouli (2018) extracted the factors like the compatibility of the students with the instructor's course delivery method and the student's prior preparation and eagerness to learn by using Deep Learning to provide a personalized E-Learning Resource Platform according to the user preference. Kondo and Hatanaka (2019) in their empirical analysis used a Bayesian Network to find out the learning states of the students which provided the instructor with feedback on the students who were likely to get a lower grade. The literature review highlights that most studies used a common approach to problem-solving. The Bayesian Networks provided good accuracy but most of the data that they were trained upon was small. Nonetheless,

the usage of Learning Analytics remains one of the rigorously pursued areas in educational research providing huge opportunities for further exploration.

3 Methodology

For the present investigation, a Bayesian Network was employed which is a graphical model comprising nodes and edges with probabilities. We first design a Bayesian Network using observed relationships between different variables and then calculate the probabilities given the occurrence of those events. We propose a fully independent model with random independent variables. In the model, Nodes have been treated as random variables and edges as relationships between the random variables while making the graphical model.

Figure 1 given below shows the attributes and the interconnections proposed in this paper for feeding into the BN model. After preparing the model of the graph, we use conditional probabilities given a certain activity performed or selected by the student to predict the probability for the further outcome because of the selected activity.

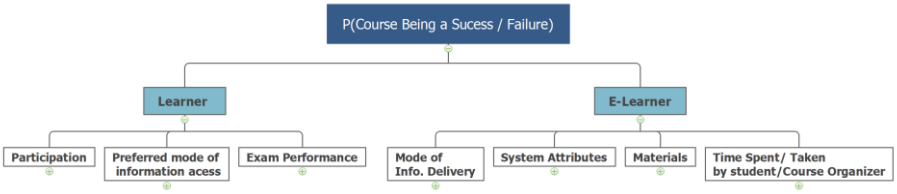

Fig. 1. Bayesian Network model

For this study, an undergraduate course, (Analog and Digital VLSI Design) ADVD is selected for learning analytics. For the students, the data relating to the student's preferences, attitudes, and perceptions were collected and categorized based on Learner and E-Learner attributes through a questionnaire on the official Electrical Electronics Department communication groups which have 7 main attributes Participation, Preferred mode of information access, Exam performance, Mode of information delivery, System attributes, Materials, Time spent/taken by student/Course organizer as shown in Table 1. These main attributes are split into sub-attributes for breaking down the data for simplifying the analysis. The sub-attributes are declared as shown in Table-1.

In Table-1 the first row shows the main attributes and the rows below the attribute's corresponding columns show the list of the sub-attributes based on which the test data was formed to train the network and draw inferences. The model is trained using the naïve Bayes algorithm which uses the Bayes Theorem on each class to predict the probability that the data points belong to a particular class and finally selects the class having the highest probability. According to Zhang et al. (2007), 3 types of models can be used to train the Bayesian Networks the first one being a data-centric model, the second one being an efficiency-centric model and the third one being an expert-centric model. A

data-centric model is used in the present study. To assess the learning styles, a measure was constructed and executed to assess the learning styles of students concerning their interests in the course. Further, the sum and averages of all the responses were computed.

Table 1. Encoded Sub-Attributes

Participa-tion	Prefererred mode of infor-mation Access	Exam Per-formance	Information Delivery Mode	System At-tributes	Materials	Time Spent/Taken
Forums(ex: Slack)	Video, Audio, PPT	Marks Scored	Video, Audio, PPT	Ease Of Ac-cess	Number of Tests/Assign-ment	Student Portal Login frequency
Mail	Discussion	Question Difficulty	Discussion	UI/UX	Reading Materials	Time spent on Portal
	Online-In-Person		Online-In-Person	Features/ Functionali-ties	Features/ Functionalities	Exam Delivery Time
	Online-record-ings		Online-re-cordings		Access to Examples/ Exer-cises	Result Declaration Time

The data was collected and analyzed to learn the student's learning style by knowing how the student interacts with the system. The algorithm then gives the probabilities of providing useful assistance to the student and instructor through suggestions of more Take-home assignments, Reading exercises/problems according to his/her preferred learning styles. The data is pre-processed and normalized using the scaling to a range technique. We use feature clipping to remove all the ambiguous and extreme values. The dataset is then split into two sets: 25 percent test dataset and 75 percent training dataset. Further fundamental Naive Bayes algorithm is applied to the dataset with the assumption that each feature would have an independent contribution. Each variable is considered to be equally contributing to the output while processing.

A Gaussian Bayes classifier is used to create a Gaussian distribution model without any co-variance with the help of standard deviation and mean. In the Gaussian model at every data point, the z-score distance between that point and each class mean is calculated, namely the distance from the class mean divided by the standard deviation of that class to normalize the data and get a bell-shaped Gaussian curve. Thus, it can be observed that the Gaussian Naive Bayes has a slightly different approach and can be used efficiently. The Gaussian Classifier is then created and using the existing training sets the model is trained and the trained model is compared with the test set to get a prediction score. We then generate a correlation matrix for fine-tuning the hyperparameters and getting the optimal accuracy for the model.

4 Results

The final accuracy was achieved after hyperparameter tuning was found to be 72.22%. Table 2 given below shows how the independent variables are interrelated and their effect on the dependent variables. In the figure given below the Column AC is the dependent variable and the columns A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, AA, AB, AC, and AD are the independent variables. Table 2 provides the Encoded names of the columns in the Correlation Matrix-

Table 2. Sub-Attributes

Column Name	Encoded Name	Column Name	Encoded Name
Comm_Platform_Satisfaction	A	E-Learning_feature_incorporate	P
Pref_Live_Lectures	B	THA	Q
Pref_Audio_Lectures	C	PracticeExercises	R
Pref_Recorded_Content	D	ReadingMaterials	S
Pref_PowerPoint_presentation	E	QuizSolutions	T
Pref_Interactive_Sessions	F	TextBooks	U
PercentileAbleTo_Score	G	Research_Papers	V
Question_Difficulty	H	Materials_provision_Platform	W
Curr_Live_Lectures	I	Tests_Assignments_in_course	X
Curr_Audio_Lectures	J	Result Show_Time	Y
Curr_Recorded_Content	K	AvgDur_Tests	Z
Curr_PowerPoint_presentation	L	Weekly_Hours_Browse_Platform	AA
Curr_Interactive_Sessions	M	Portal_Login_Frequency	AB
PlatformAccessEase	N	CourseProgress_Satisfaction_x	AC
Platform_UI_Intuitive_easy-to-use	O	Prim_Comm_Platform	AD

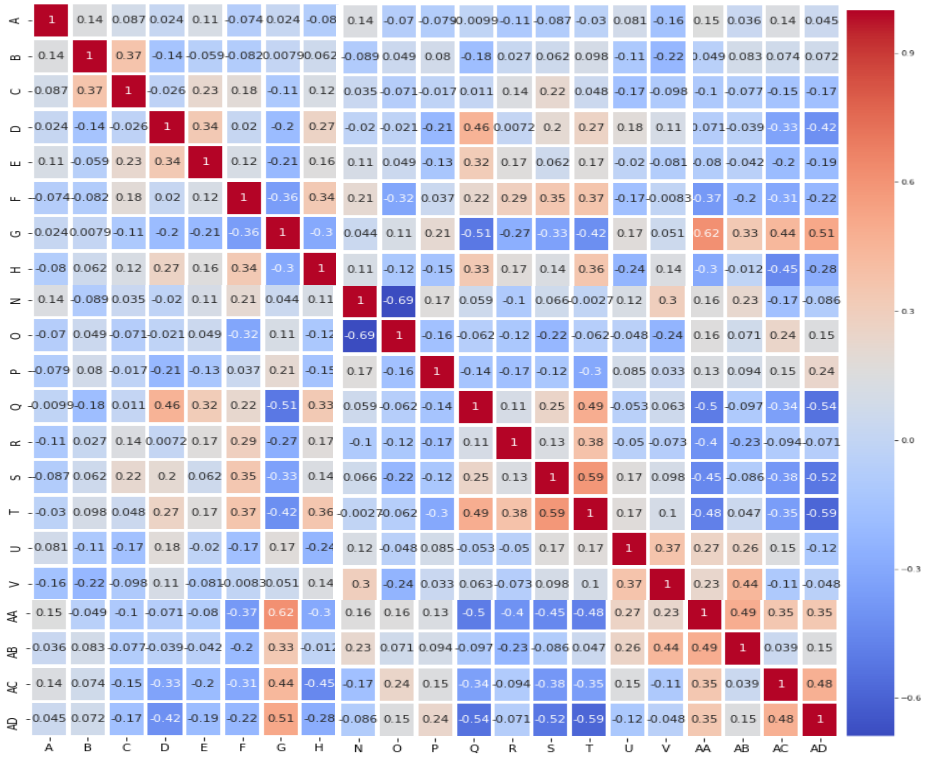

Fig. 2. Correlation Matrix

In the correlation matrix presented in figure-2, a warm-cool colour scheme has been used where the warmth of the colour increases the positive correlation between the 2 variables. The number inside that column is the impact of the increase in 1 input variable on another input variable. As the colour of the cell turns towards dark blue, it gives us a negative correlation, and the number inside the box denotes the amount of decrement of one variable due to the increment of another variable. The employed ways used by the instructor to deliver the lectures i.e.('I', 'J', 'K', 'L', 'M', 'X', 'Y', 'Z') had zero correlation and impact on the other input variables and output, hence could easily be dropped. The variable, percentage scored by the student, was found to decrease with an increase in the Take-home assignments provided and found to increase with the amount of time spent by the student browsing the E-Learning Platform 'AA'. This can be attributed to the fact that an increase in assignments forces the student to study the subject and meet the deadlines. A more flexible approach would suit the student as can be observed with the voluntary time spent on the platform browsing course materials. It was found out that the ease of accessing the platform ‘N’ was increasing as the Platform got a more intuitive User Interface ’O’. A better User Interface makes the platform more appealing to use and results in an increased retention rate. The provision of Take-home assignment 'Q' decreased the browsing time of the student on the platform. Providing assignments decreased the interest of students in the course in turn affecting their browsing activity on the platform. It was found out that when the Reading Materials ’S’ is provided the instructors mostly upload the Quiz Solutions ‘T’ along with that as well. The choice of the primary communication platform used by the instructor to communicate greatly impacted the Percentile Score of the Student 'G', and also affected the choice of instructor in the provision of Reading Material 'S’, Quiz Solutions 'T', and Take-home assignments ‘Q’.

The final results were that the decision variable was most positively affected by the amount of time spent by the student in browsing the E-Learning Platform ‘AA’, the Primary Communication Platform used by the instructor ‘AD’, the Percentile Score of the Student 'G' and most negatively affected by the Question Difficulty settings 'H', the number of Take-home assignments present in the course 'Q', The Number of Reading Materials 'S', and the provision of Quiz Solutions 'T'. The fact that a good score defines a lot about the conceptual clarity as well as the interest, along with the effort put by the student in that course, a good score does define a higher interest of the student in that specific course and a higher probability of that course being successful which is also visible through the question difficulty pattern set up for students. The students tend to score a lower percentage when they get a hard paper, in turn, affecting their confidence and making them less interested in the course. The increase in the provision of the materials, and a lot of deadlines for the students restrain the free will of the student to study at his time discretion, reducing his time for revision, Grades, and in turn his interest in the course.

5 Discussion

The study was conducted across a group of 74 students, and it was found that 83.8 % of instructors used mail as their primary communication platform for interaction with the student as it is the most standard platform that has been used for decades and 55.4 % of the students were satisfied with the platform used by the instructor. It was further observed that 57% of the students prefer recorded content for online delivery of their lectures as they can access the recording at their time and even helps them to download the videos to be viewed later as in many geographical areas where there is a poor internet connection.

It was found that the average percentile scored by the student in the course is 37.6%. 66.2 % of respondents found the course to be hard and only 3% of them felt that the course is easy. The low percentage can be attributed to the difficult paper pattern set up by the instructor to make the students study more and prevent them from getting overconfident. These results align with the study by Kondo and Hatanaka (2019) and Sundar (2013) to provide instructors with feedback on the students and identify students who were likely to get a lower grade, drop out or fail the course. In the present case, the students who were unhappy with the course ended up not studying the course diligently and in turn getting a low grade. The results were further corroborated by Sharabiani et al. (2014) who stated that how the grades are allotted to the students in particular subjects has a major impact on their morale affecting their interest in their subject and resulting in higher dropout rates. It was further stated that including student details, their number per semester, the level of difficulty of questions, and its influence on students' marks in the subject improved their model. The trained model in the study performed better in including the attributes mentioned above.

It was found that the instructors heavily rely on (live+recorded) lectures along with PowerPoint Presentations. The results were in synergy with those made by Chakraborty and Sinha (2016) with the study material recommendation by evaluating the learning style of the students from the materials browsed and their test performance. Around 50% found the E-Learning platform used by the instructor easy to use and 78% found it to be intuitive. Approximately 65% of students were satisfied with the current features provided by the E-Learning Platform and only 31% want the addition of more features. Most of the people were satisfied with the E-Learning Platform and some new features can be added to make the workflow smoother.

Around twenty assignments were recorded in the course. The average result showed the time for test results is 72 hours on the platform. The average test duration in the course is 30 mins. The instructor is following a high number of evals with fewer marks and less time is given to the students to make sure that the students follow all the lectures diligently and are up to date with what is being taught in the class. Approximately 3 hours is the weekly average time spent by the students browsing the E-Learning Platform. The average Portal login frequency of the students is approx. 2.6 times which can be improved, the time spent browsing and viewing material is a direct indication of the interest of the student in that course.

To conclude, it was discovered that in the ADVD course 77.3% of people were dissatisfied with the way their current course is progressing. A better approach can be

taken to increase these numbers by reducing the Evals and THA's and making a student-friendly paper. The main determiners of the success of E-Learning courses are the difficulty level of the paper, the number of taking home assignments, and the synergy of the student with the instructor's method of teaching.

References

1. Kondo, N., & Hatanaka, T. (2019). Estimation of Students' Learning States using Bayesian Networks and Log Data of Learning Management System. International Journal of Institutional Research and Management, 3(2), 35–49.
2. Chanaa, A., and N. E. Faddouli. 2018. "Deep Learning for a Smart E-Learning System." In *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, 1–8.
3. García, Patricio, Analía, Amandi, Silvia, Schiaffino, and Marcelo, Campo. "Using Bayesian Networks to Detect Students' Learning Styles in a Web-based education system".*Proceedings of ASAI 2005, Argentine Symposium on Artificial Intelligence* (2005).
4. Ueno, M., and T. Okamoto. 2007. "Bayesian Agent in E-Learning." In *Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007)*, 282–84.
5. Kao, H.-, M.- Liu, C.- Huang, and Y.- Chang. 2009. "E-Learning Systems Evaluation with Data Envelopment Analysis and Bayesian Networks." In *2009 Fifth International Joint Conference on INC, IMS and IDC*, 1207–10.
6. West, Patti, Daisy Wise Rutstein, Robert J. Mislevy, Junhui Liu, Roy Levy, Kristen E. Dicerbo, Aaron Crawford, Younyoung Choi, Kristina Chapple, and John T. Behrens. 2012. "A Bayesian Network Approach To Modeling Learning Progressions." In *Learning Progressions in Science: Current Challenges and Future Directions*, edited by Alicia C. Alonzo and Amelia Wenk Gotwals, 257–92.
7. Chakraborty, B., and Meghamala Sinha. 2016. "Student Evaluation Model Using Bayesian Network in an Intelligent E-Learning System." *IIOAB Journal* 7 (January): 51–60.
8. Sundar, Praveen. 2013. " A Comparative Study for Predicting Student's Academic Performance Using Bayesian Network Classifiers" *IOSR Journal of Engineering 2250-3021* 3 (March): 2250–3021.
9. Sharabiani, A., F. Karim, A. Sharabiani, M. Atanasov, and H. Darabi. 2014. "An Enhanced Bayesian Network Model for Prediction of Students' Academic Performance in Engineering Programs." In *2014 IEEE Global Engineering Education Conference (EDUCON)*, 832–37.
10. Mahnane, Lamia, and Mohamed Hafidi. 2016. "Automatic Detection of Learning Styles Based on Dynamic Bayesian Network in Adaptive E-Learning System." *International Journal of Innovation and Learning* 20 (3). Inderscience Publishers: 289–308.
11. Rajper, Samina, Noor A. Shaikh, Zubair A. Shaikh, and Ghulam Ali Mallah. 2016. "Automatic Detection of Learning Styles on Learning Management Systems Using Data Mining Technique." *Indian Journal of Science and Technology* 9 (15).
12. Liang Zhang, Yue-ting Zhuang, Zhen-ming Yuan and Guo-Hua Zhan, "Auto diagnosing: An intelligent assessment system based on Bayesian Networks," 2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities and Without Passports, 2007, pp. T1G-7-T1G-10