Academic performance prediction based on multistore, multifeatured behavioral data (1)

Bandi Vamsi

Department of Computer Science, Sri Venkateswara University, Tirupati

Abstract**—** *Digital data trails from disparate sources covering different aspects of student life are stored daily in most modern university campuses. However, it remains challenging to (i) combine these data to obtain a holistic view of a student, (ii) use these data to accurately predict academic performance, and (iii) use such predictions to promote positive student engagement with the university. To initially alleviate this problem, in this paper, a model named Augmented Education (Augmented) is proposed. In our study, (1) first, an experiment is conducted based on a real-world campus dataset of college students (N = 156) that aggregates multisource behavioural data covering not only online and offline learning but also behaviours inside and outside of the classroom. Specifically, to gain in-depth insight into the features leading to excellent or poor performance, metrics measuring the linear and nonlinear behavioural changes (e.g., regularity and stability) of campus lifestyles are estimated; furthermore, features representing dynamic changes in temporal lifestyle patterns are extracted by the means of long short-term memory (LSTM). (2) Second, machine learning-based classification algorithms are developed to predict academic performance. (3) Finally, visualized feedback enabling students (especially at-risk students) to potentially optimize their interactions with the university and achieve a study-life balance is designed. The experiments show that the AugmentED model can predict students' academic performance with high accuracy.*

Keyword: academic performance prediction, behavioural pattern, digital campus, machine learning (ML), long short-term memory (LSTM).

I. INTRODUCTION

As an important step to achieving personalized education, academic performance prediction is a key issue in the education data mining field. It has been extensively demonstrated that academic performance can be profoundly affected by the following factors:

- Students' Personality (e.g., neuroticism, extraversion, and agreeableness)
- Personal Status (e.g., gender, age, height, weight, physical fitness, cardiorespiratory fitness, aerobic fitness, stress, mood, mental health, intelligence, and executive functions)
- Lifestyle Behaviours (e.g., eating, physical activity, sleep patterns, social tie, and time management) and
- Learning Behaviours (e.g., class attendance, study duration, library entry, and online learning)

1.1 Data Mining:

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing, and computer processing. Data mining processes are used to build machine learning models that power applications including search engine technology and website recommendation programs.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD). Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information. Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective. This process includes various types of services such as text mining, web mining, audio and video mining, pictorial data mining, and social media mining. It is done through software that is simple or highly specific. By outsourcing data mining, all the work can be done faster with low operation costs. Specialized firms can also use new technologies to collect data that is impossible to locate manually. There are tonnes of information available on various platforms, but very little

knowledge is accessible. The biggest challenge is to analyse the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find better insight from it.

1.2 Data Privacy and Security:

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyses the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

1.3 Data Visualization:

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.

II. LITERATURE REVIEW

Author: R. Wang, G. Harari, P. Hao, X. Zhou, and A. T. Campbell,2015

Title: SmartGPA: How smartphones can assess and predict academic performance of college students,"

Description: Many cognitive, behavioural, and environmental factors impact student learning during college. The Smart GPA study uses passive sensing data and self-reports from students' smartphones to understand individual behavioural differences between high and low performers during a single 10-week term. We propose new methods for better understanding study (e.g., study duration) and social (e.g., partying) behaviour of a group of undergraduates. We show that there are a number of important behavioural factors automatically inferred from smartphones that significantly correlate with term and cumulative GPA, including time series analysis of activity, conversational interaction, mobility, class attendance, studying, and partying. We propose a simple model based on linear regression with lasso regularization that can accurately predict cumulative GPA. The predicted GPA strongly correlates with the ground truth from students' transcripts ($r = 0.81$ and $p < 0.001$) and predicts GPA within ±0.179 of the reported grades. Our results open the way for novel interventions to improve academic performance. Author Keywords Smartphone sensing; data analysis; academic perform

Author: GILBERT, S. P., AND WEAVER 2010

Title: Sleep quality and academic performance in university students:

Description: Both sleep deprivation and poor sleep quality are prominent in American society, especially in college student populations. Sleep problems are often a primary disorder rather than secondary to depression. The purpose of the present study was to determine if sleep deprivation and/or poor sleep quality in a sample of nondepressed university students was associated with lower academic performance. A significant negative correlation between Global Sleep Quality score (GSQ) on the Pittsburgh Sleep Quality Index and grade point average supports the hypothesis that poor sleep quality is associated with lower academic performance for nondepressed students. Implications for both the remedial (assessment and treatment) and preventive (outreach) work of college and university counseling centers is discussed.

Author: Z. Wang, X. Zhu, J. Huang, X. Li, and Y. Ji

Title: "Prediction of academic achievement based on digital campus,

Description: Instructional Systems Design is the practice of creating of instructional experiences that make the acquisition of knowledge and skill more efficient, effective, and appealing. Specifically, in designing courses, an hour of training material can require between 30 to 500 hours of effort in sourcing and organizing reference data for use in just the preparation of course material. In this paper, we present the first system of its kind that helps reduce the effort associated with sourcing reference material and course creation. We present algorithms for document chunking and automatic generation of learning objectives from content, creating descriptive content metadata to improve content-discoverability. Unlike existing methods, the learning objectives generated by our system incorporate pedagogically motivated Bloom's verbs. We demonstrate the usefulness of our methods using real world data from the banking industry and through a live deployment at a large pharmaceutical company.

Author: X. Zhang, G. Sun, Y. Pan, H. Sun, Y. He, and J. Tan,2018

Title: "Students performance modelling based on behaviour pattern,"

Description: Academic performance of college students is the main concern of educational institutions. Effective and timely predicting performance is not only conducive to the school's Ministry to improve the efficiency of supervision, but also helps students to develop good study habits. With the rapid construction of digital campus, university as the main range for students' life can not only serve convenience but also record daily life. The popular using of smart card makes it easy to outline student's behaviour pattern with rich data. The purpose of our work is to predict students' performance based on their behaviour pattern and analyse the correlation between them. In this paper, we propose a general framework to model students' performance. Firstly, we describe students behaviour pattern and extract behaviour features in two perspectives including statistics and relevance. Then we employ a multi-task model to learn performance of every course simultaneously. Our experiments on a real-world data set of college students show a good outcome. We do a further analysis on relation between students' behaviour and academic performance. Moreover, our experiments indicate that our framework is feasible for early warning.

Author: Adolphus K, Lawton CL, Dye L (2013)

Title: The effects of breakfast on behaviour and academic performance in children and adolescents.

Description: Breakfast consumption is associated with positive outcomes for diet quality, micronutrient intake, weight status and lifestyle factors. Breakfast has been suggested to positively affect learning in children in terms of behaviour, cognitive, and school performance. However, these assertions are largely based on evidence which demonstrates acute effects of breakfast on cognitive performance. Less research which examines the effects of breakfast on the ecologically valid outcomes of academic performance or in-class behaviour is available. The literature was searched for articles published between 1950–2013 indexed in Ovid MEDLINE, Pubmed, Web of Science, the Cochrane Library, EMBASE databases, and PsychINFO. Thirty-six articles examining the effects of breakfast on in-class behavior and academic performance in children and adolescents were included. The effects of breakfast in different populations were considered, including undernourished or well-nourished children and adolescents from differing socio-economic status (SES) backgrounds. The habitual and acute effects of breakfast and the effects of school breakfast programs (SBPs) were considered. The evidence indicated a mainly positive effect of breakfast on on-task behavior in the classroom. There was suggestive evidence that habitual breakfast (frequency and quality) and SBPs have a positive effect on children's academic performance with clearest effects on mathematic and arithmetic grades in undernourished children. Increased frequency of habitual breakfast was consistently positively associated with academic performance. Some evidence suggested that quality of habitual breakfast, in terms of providing a greater variety of food groups and adequate energy, was positively related to school performance. However, these associations can be attributed, in part, to confounders such as SES and to methodological weaknesses such as the subjective nature of the observations of behavior in class.

III. EXISTING SYSTEM

In this existing system, in the academic based prediction remains some challenging to:

- Combine these data to obtain a holistic view of a student,
- Use these data to accurately predict academic performance, and
- Use such predictions to promote positive student engagement with the university

The student predictions based on the academic performance not all kind of behavioral activities. The features and predictions is only based on the online metrics. The prediction of the students' academic performance with low accuracy. However, due to lack of richness and diversity in both data sources and features, there still exist a lot of challenges in prediction accuracy and interpretability.

Disadvantages:

- The prediction accuracy of the student performance is low.
- Student cannot know their own behavioural status.
- The computational time is high.

IV. PROPOSED SYSTEM

In our study first, an experiment is conducted based on a real-world campus dataset of college students that aggregates multisource behavioural data covering not only online and offline learning but also behaviours inside and outside of the classroom. Specifically, to gain in-depth insight into the features leading to excellent or poor performance, metrics measuring the linear and nonlinear behavioural changes (e.g., regularity and stability) of campus lifestyles are estimated; furthermore, features representing dynamic changes in temporal lifestyle patterns are extracted by the means of long short-term memory (LSTM). Second, machine learning-based classification algorithms are developed to predict academic performance. Finally, visualized feedback enabling students (especially at-risk students) to potentially optimize their interactions with the university and achieve a study-life balance is designed. The experiments show that the AugmentED model can predict students' academic performance with high accuracy

Although many academic performance prediction systems have been developed for college students, the following challenges persist:

- capturing a sufficiently rich profile of a student and integrating these data to obtain a holistic view
- exploring the factors affecting students' academic performance and using this information to develop a robust prediction model with high accuracy; and

- taking advantage of the prediction model to deliver personalized services that potentially enable students to drive behavioural change and optimize their study-life balance.

Advantages:

- It is very useful to analysis the student's entire performance to study the life balance
- The predicted academic performance is high accuracy.
- Finding the student affecting factors, for the prediction model.
- The computational time is very less and the cost is low.

Algorithms:

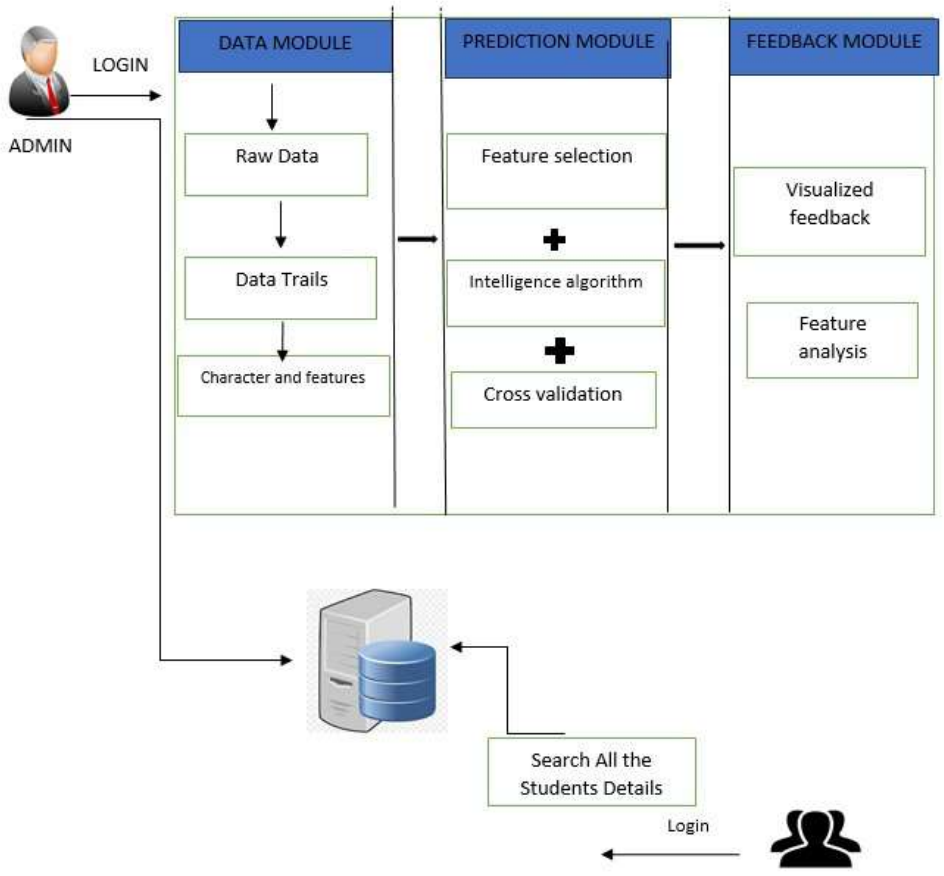
Machine learning: Machine learning is a type of artificial Intelligence that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

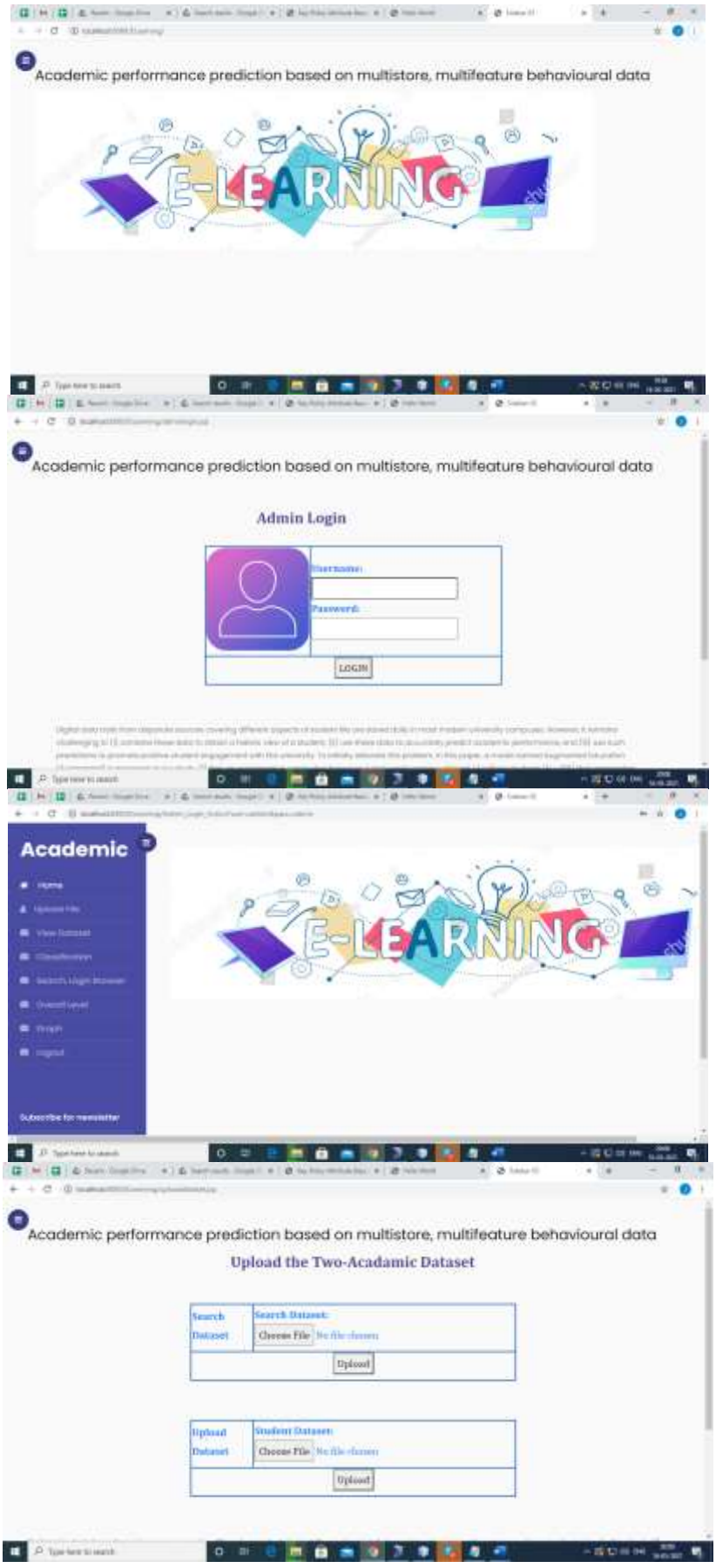
Random forest: Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time.

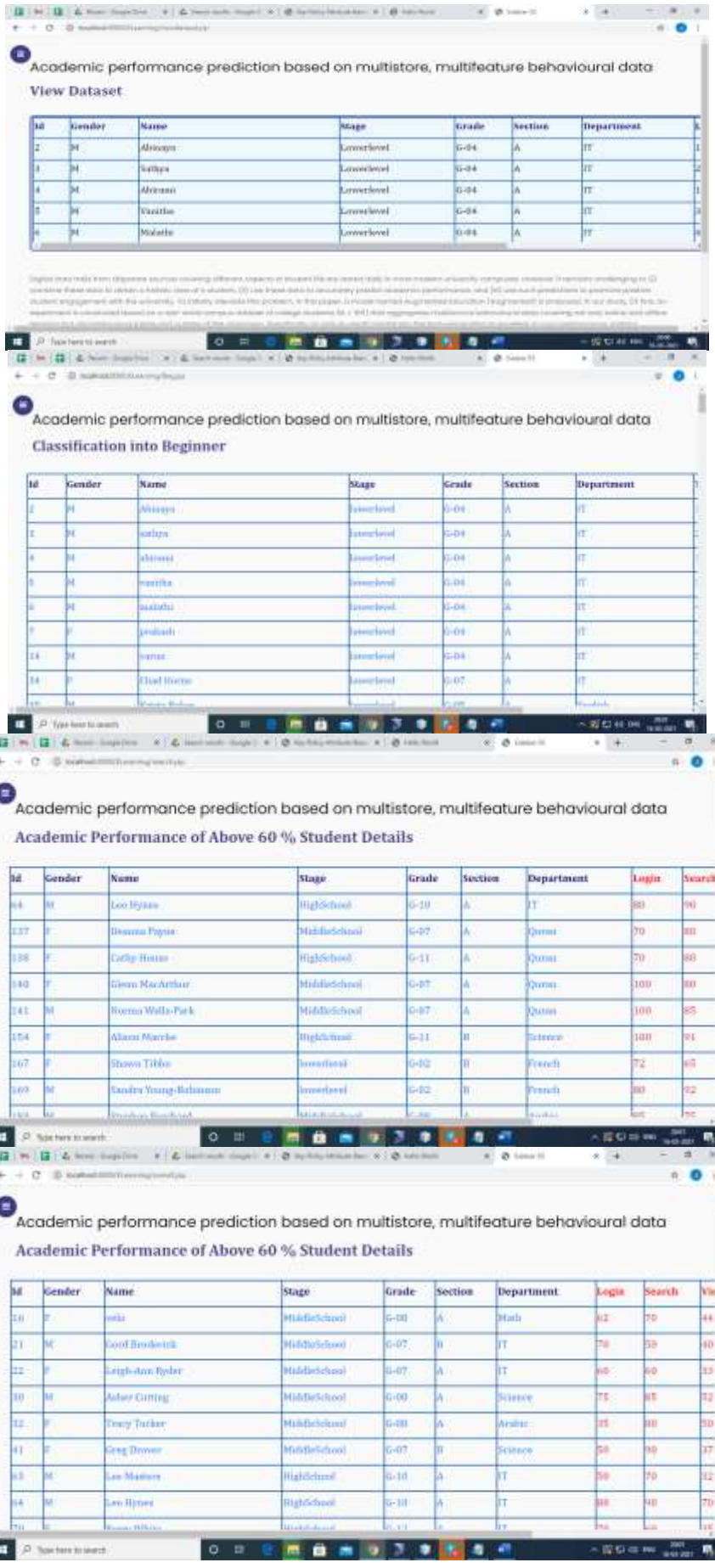
K Nearest Neighbor: K nearest KNN is used for both classification and regression Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

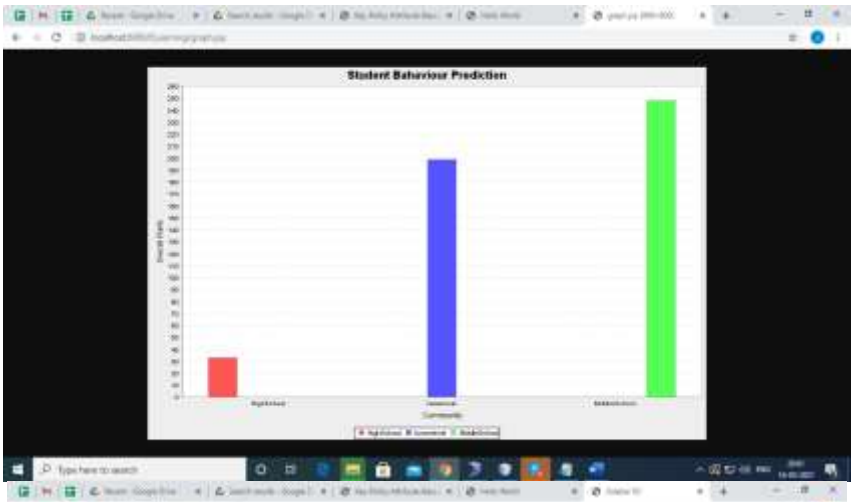
Cross validation: Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, failing to generalize a pattern.

V. SYSTEM ARCHITECTURE

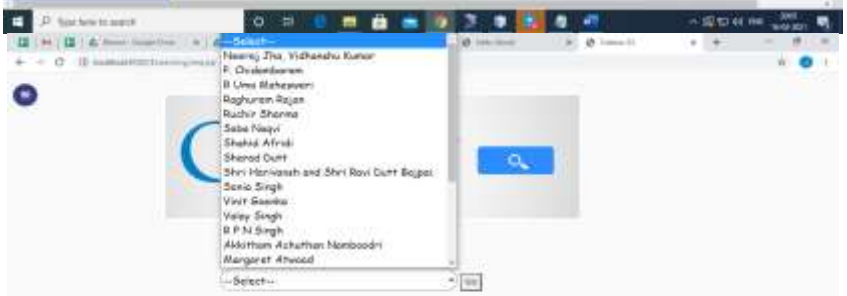


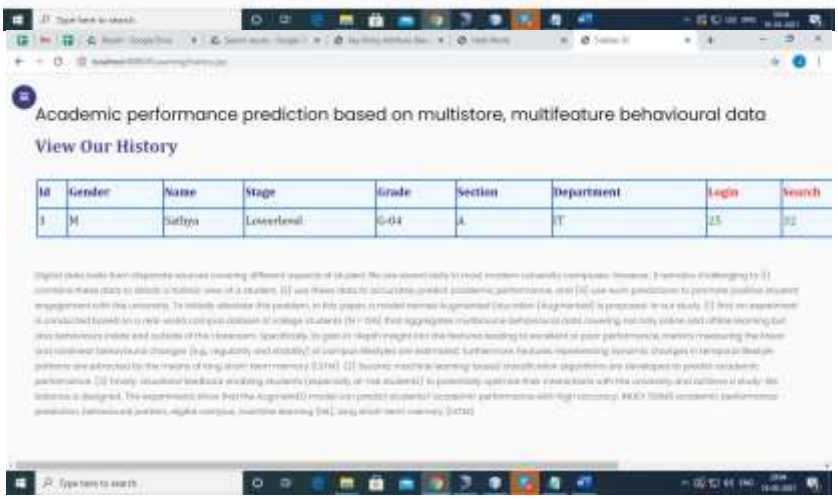

Academic performance prediction based on multistore, multifeature behavioural data
E-LEARNING
Academic performance prediction based on multistore, multifeature behavioural data
Admin Login
Username:
Password:
LOGIN
Academic
Home
Subscribe for newsletter
E-LEARNING
Academic performance prediction based on multistore, multifeature behavioural data
Upload the Two-Acadamic Dataset
Search Dataset:
Choose File
Upload
Student Dataset:
Choose File
Upload


Academic performance prediction based on multistore, multifeature behavioural data
View Dataset
Id Gender Name Stage Grade Section Department
Academic performance prediction based on multistore, multifeature behavioural data
Classification into Beginner
Id Gender Name Stage Grade Section Department
Academic performance prediction based on multistore, multifeature behavioural data
Academic Performance of Above 60 % Student Details
Id Gender Name Stage Grade Section Department Login Search
Academic performance prediction based on multistore, multifeature behavioural data
Academic Performance of Above 60 % Student Details
Id Gender Name Stage Grade Section Department Login Search


Student Behaviour Prediction


Academic
Home
Search Query
View History
Logout
Subscribe for newsletter
E-LEARNING


--Select--
Neeraj Jha Vidhanshu Kumar
P. Chidambaram
B Uma Maheswari
Raghuram Rajan
Ruchir Sharma
Saba Naqvi
Shakil Afridi
Sharad Dutt
Shri Harivansh and Shri Ravi Dutt Bajpai
Sonia Singh
Vivit Saeeka
Vijay Singh
B P N Singh
Akkitham Achuthan Namboodiri
Margaret Atwood
--Select--


Academic performance prediction based on multistore, multifeature behavioural data
View Our History
Id
Gender
Name
Stage
Grade
Section
Department
Login
Search
1
M
Sathya
Lowerlevel
G-04
A
IT
25
32

VI. CONCLUSION

In our study, a model named Augmented is proposed to predict the academic performance of college students. Our contributions in this study are related to three sources. First, regarding data fusion, to the best of our knowledge, this work is the first to capture, analyse and use multisource data covering not only online and offline learning but also campus life behaviours inside and outside of the classroom for academic performance prediction. Based on these multisource data, a rich profile of a student is obtained. Second, regarding the feature evaluation, behavioural change is evaluated by linear, nonlinear, and deep learning (LSTM) methods respectively, which provides a systematically view of students' behavioural patterns. Specifically, it is the first time that three novel nonlinear metrics (Lye, Hurst, and DFA) and LSTM are applied in students' behavioural time series analysis. Third, our experimental results demonstrate that Augmented can predict academic performance with quite high accuracy, which help to formulate personalized feedback for at-risk (or self-disciplined) students. However, there are also some limitations in our study. To gain a multisource dataset, we scarified the scale the dataset by only using student-generated data within a single course. This limitation might have a certain negative influence on the generalization of Augmented. Furthermore, in this study, we mainly focus on behavioural change. Other characteristics/features (e.g., peer effect, sleep) that are worthy of consideration were not evaluated in this study. In conclusion, our study is based on a complete passive daily data capture system that exists in most modern universities.

REFERENCES

[1] Z. Liu, C. Yang, L. S. Rüdian, S. Liu, L. Zhao, and T. Wang, "Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums," Interactive Learning Environments, vol. 27, pp. 598-627, 2019.

[2] A. Zollanvari, R. C. Kizilirmak, Y. H. Kho, and D. Hernandez-torrano, "Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors," IEEE Access, vol. 5, pp. 23792-23802, 2017.

[3] A. Akram, C. Fu, Y. Li, M. Y. Javed, R. Lin, Y. Jiang, and Y. Tang, "Predicting students' academic procrastination in blended learning course using homework submission data," IEEE Access, vol. 7, pp. 102487-102498, 2019.

[4] L. Gao, Z. Zhao, L. Qi, Y. Liang, and J. Du, "Modeling the effort and learning ability of students in MOOCs," IEEE Access, vol. 7, pp. 128035-128042, 2019.

[5] Z. Liu, H. Cheng, S. Liu, and J. Sun, "Discovering the two-step lag behavioral patterns of learners in the college SPOC platform," International Journal of Information and Communication Technology Education, vol. 13, no. 1, pp. 1-13, 2017.

[6] Z. Liu, W. Zhang, H. Cheng, J. Sun, and S. Liu, "Investigating relationship between discourse behavioral patterns and academic achievements of students in SPOC discussion forum," International Journal of Distance Education Technologies, vol. 16, no. 2, pp. 37-50, 2018.

[7] Z. Liu, N. Pinkwart, H. Liu, S. Liu, and G. Zhang, "Exploring students engagement patterns in SPOC forums and their association with course performance," EURASIA Journal of Mathematics, Science and Technology Education, vol. 14, no. 7, pp. 3143-3158, 2018.

[8] B. Kim, B. Vizitei, and V. Ganapathi, "GritNet: Student performance prediction with deep learning," in Proc. of the 11th International Conference on Educational Data Mining, Buffalo, NY, USA, 2018, pp. 625-629.

[9] S. Sahebi, and P. Brusilovshky, "Student performance prediction by discovering inter-activity relations," In Proc. of the 11th International Conference on Educational Data Mining, Buffalo, NY, USA, 2018, pp. 87-96.

[10] T. L. Kelley, "The selection of upper and lower groups for the validation of test items," Journal of Educational Psychology, vol. 30, no. 1, pp. 17–24, 1939.