



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 7, July 2024

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.423

9940 572 462

6381 907 438

ijrset@gmail.com

www.ijrset.com

Predicting Academic Performance using Multisource, Multifeatured Behavioural Data

Mr. G. Anil Kumar¹, T P Varsha Rani², Vennamaneni Rithika³, Yashwanth Vuppula⁴

Associate Professor, Department of Computer Science and Technology, VignanaBharathi Institute of Technology,
Hyderabad, India¹

Student, Department of Computer Science and Technology, VignanaBharathi Institute of Technology,
Hyderabad, India^{2,3,4}

ABSTRACT: In the modern university setting, diverse digital data streams originating from various sources, reflecting different facets of student life, are generated daily. However, integrating these data to gain a comprehensive understanding of students, accurately predicting their academic performance, and leveraging such predictions to enhance student engagement with the university pose significant challenges. To address these challenges, this paper proposes a model called Augmented Education (AugmentED). In our investigation, we conduct an experiment utilizing a real-world campus dataset of college students (N = 156), amalgamating multisource behavioral data encompassing both online and offline learning activities, as well as behaviors within and outside the classroom. Specifically, to gain a deep understanding of the factors influencing academic performance, we analyze metrics that measure linear and nonlinear changes in behavioral patterns, such as regularity and stability in campus lifestyles. Additionally, we extract features representing dynamic changes in temporal lifestyle patterns using long short-term memory (LSTM) techniques. Subsequently, we develop machine learning-based classification algorithms to predict academic performance. Finally, we design visualized feedback mechanisms aimed at empowering students, particularly those at risk, to optimize their interactions with the university and achieve a balance between their academic and personal lives. Our experiments demonstrate that the AugmentED model exhibits high accuracy in predicting students' academic performance.

KEYWORDS: Multisource behavioral data, Linear and nonlinear behavioral changes, Regularity and stability, Long short-term memory (LSTM), Academic performance prediction, Visualized feedback, Study-life balance.

I. INTRODUCTION

Today's educational environment places a greater emphasis on utilizing the abundance of data that students produce during their academic careers in order to better understand and improve their academic performance. Teachers and educational institutions have a promising new opportunity to learn more about the learning patterns and results of their students thanks to the integration of multi-source, multi-feature behavioral data, which is driving the growing field of academic performance prediction. Scholars and practitioners are working to build comprehensive models that can predict academic success more accurately than ever before by utilizing data from a variety of sources, including social interactions, attendance records from outside of the classroom, online learning platforms, and extracurricular activities. The fusion of advanced data analytics methods with the widespread adoption of digital technologies within educational contexts has sparked a fundamental shift in how we perceive and approach the prediction of academic performance. Rather than relying solely on traditional markers like exam scores or GPA, modern methodologies prioritize the synthesis and examination of diverse behavioral data streams. These encompass quantitative metrics such as study time and test results, alongside qualitative indicators like engagement levels, collaboration patterns, and even emotional states inferred from textual or sensor data. By amalgamating these varied data sources, researchers strive to develop comprehensive predictive models that capture the complex interplay between students' actions, surroundings, and academic achievements. In today's data-driven educational landscape, forecasting academic success based on multifaceted behavioral data offers considerable potential for reshaping student outcomes. By amalgamating data from online platforms, classroom interactions, extracurricular activities, and social behaviors, researchers seek to create all-encompassing predictive models. Advanced machine learning techniques comb through this wealth of information to identify significant correlations and predictors of academic success. Given the interdisciplinary nature of this endeavor,

collaboration across fields such as education, psychology, computer science, and data analytics is essential. Implementing robust predictive models enables educators to intervene proactively and optimize resource allocation. However, ethical considerations regarding data privacy and security must be carefully addressed throughout this process. As educational institutions embrace digital innovations, leveraging multi-source data analytics becomes crucial for ensuring equitable access to high-quality education. Through responsible technological deployment, we can drive positive educational outcomes and broader societal benefits.

II. RELATED WORK

In today's digital age, the abundance of datasets worldwide has led to a surge in popularity for analyzing and deriving insights from them. In our research, we extensively reviewed existing studies using statistical methods to understand the importance of visualizing datasets. We considered numerous research works for this purpose. We address academic performance prediction as a classification problem in our work. We divide academic achievement into three groups based on Kelley's high-low discrimination index: low, medium, and high. We have access to a digital campus dataset, and our main goals are to use it to train a classification algorithm, along with using the raw data to extract specific features and find features that are highly connected with their performance academically, and finally provide visual feedback based on the prediction results.

III. PROPOSED SYSTEM

In the study, we treat prediction of academic achievement as a classification problem, classifying academic performance into low, medium, and high categories using Kelley's high-low discrimination score. Our main goal is feature extraction from different data sources, and use the features with high connection to success in academics to train a classification algorithm given a dataset of a digital campus. Three major contributions are made by our proposed model, Augment ED, which attempts to predict for college students.

First off, this is noteworthy for being the first in data fusion as it thoroughly collects, examines, and uses multisource data on the behaviour on campus with offline learning and also both in and out of the classroom to predict academic achievement. We are able to compile a thorough profile of every pupil thanks to this method.

Second, in terms of feature assessment, we employ a variety of approaches, such as deep learning (LSTM), nonlinear, and linear techniques, to evaluate behavioural changes. This thorough investigation provides a methodical comprehension of the behavioural patterns exhibited by the students. Notably, our work employs LSTM for the first time in the analysis of students' behavioural time series data and presents three unique nonlinear metrics (LyE, HurstE, and DFA).

IV. METHODOLOGY

Three sub-modules make up the majority of the project. These are the three modules that are available:

1. Data Collection and Preparation.
2. Model Selection
3. Integration and Deployment.

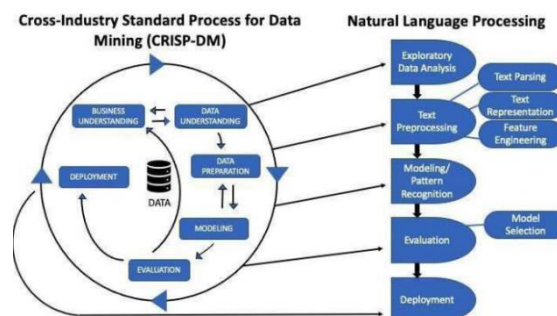


Fig – 1: Architecture of NLP Analyzer

4.1 Data Collection and Preparation:

The process of gathering data is the initial step in the actual building of a machine learning model. The more and better data we collect, the more our model will perform, so this is a crucial phase that will cascade the model's quality.

There are numerous methods for gathering the data, including manual interventions, web scraping.

Kaggle's academic performance dataset

Link: <https://www.kaggle.com/c/1056lab-student-performance-prediction/data>

There are 1044 number of records

In this data set we are taken 34 columns in the dataset, which are described below.

Identifier - Student's ID

Institution - The school attended by the student (coded as 'GP' for Gabriel Pereira or 'MS' for Mousinho da Silveira)

Subject - The subject of the student's class (coded as 'mat' for Mathematics or 'por' for Portuguese language)

Gender - The student's gender (coded as 'F' for female or 'M' for male)

Age - The student's age, which ranges from 15 to 22

Residence - The student's residential status (designated as 'U' for urban areas or 'R' for rural areas)

Family Size - The family size of the student (specified as 'GT3' for larger than three or 'LE3' for less than or equal to three)

Parental Cohabitation Status - The student's family size (designated as 'GT3' for sizes greater than three or 'LE3' for sizes less than or equal to three)

Mother's Education - The mother of the student's level of education (from 0 for no education to 4 for higher education)

Father's Education - The father of the student's degree of education (from 0 for no education to 4 for higher education)

Mother's Occupation - The mother of the student works in one of the following (categories: "teacher," "healthcare," "services," "at home," or "other.")

Father's Occupation - The father of the student works in one of the following (categories: "teacher," "healthcare," "services," "at home," or "other.")

Cause of School Choice: The reasons behind choosing the present institution (which can be classified as "close to home," "school reputation," "course preference," or "other").

Guardian: The student's guardian, also referred to as their "mother," "father," or "other."

Travel Time - The duration of travel from home to school (coded as one for less than fifteen minutes, two for fifteen to thirty minutes, three for thirty minutes to one hour, or four for more than one hour)

Study Time - The amount of time spent on studying per week (coded as one for less than two hours, two for two to four hours, three for four to ten hours, or four for more than ten hours)

Past Failures - The number of previous class failures (coded as n if one $\leq n < 3$, otherwise 4)

Extra Educational Support: Indicates if the learner obtains additional educational support (true or false).

Family Educational Support - Whether the student receives family educational support (coded as true or false)

Extra Paid Classes - Whether the student takes classes by paying in that subject(Math or Portuguese) (coded as true or false)

Attendance to Nursery School - Whether the student attended nursery school (coded as true or false)

Further Education Aspiration: Indicates whether or not the student want to pursue further education (marked as true or false).

Home Internet Access - Indicates whether or not the student has access to the internet at home (coded as true or false).

Family Relationship Quality - The perceived quality of family relationships (ranging from one for very bad to five for excellent)

Free Time - Amount of time available to the student after school hours (rated from one for very low to five for very high)

Social Outings - Frequency of the student going out with friends (rated from 1 for very low to 5 for very high)

Weekday Alcohol Consumption - Level of alcohol consumption by the student on weekdays (rated from one for very low to five for very high)

Health - Assessment of the student's current health condition (rated from one for very bad to five for very good)

School Absences - Number of absences recorded for the student in school (ranging from 0 to 93)

Final Grade - The student's final grade in the course (ranging from 0 to 20, serving as the output target).

The data will be transformed by us. by eliminating some columns and eliminating any missing data. We will first compile a list of the column names that we wish to maintain.

The columns that we want to keep are the only ones that are dropped or removed next. Lastly, we eliminate from the data set any rows that have missing values.

4.2 Model Selection

Two datasets are required when building a machine learning model: one for testing and one for training. But that's all we have right now. Let's divide this in half using an 80:20 ratio. The data frame will also be divided into feature and label columns.

Here, we imported the sklearn function `train_test_split`. After that, divide the dataset using it. Additionally, when `test_size = 0.2` is used, the dataset is split into 80% train and 20% test.

One tool for splitting the dataset is the `random_state` parameter, which seeds a random number generator. Returning four datasets is the function. Marked them with `test_x`, `test_y`, `train_x`, and `train_y` labels. The datasets' division may be seen if we look at their shape.

The Random Forest Classifier, which fits several decision trees to the data, will be used in this case. Lastly, I use `train_x` and `train_y` to pass to the fit method in order to train the model.

We must test the model after it has been trained. We will give `test_x` to the `predict` method in order to do that. One effective techniques in machine learning is random forest for characterizing. The supervised classification algorithm includes the random forest. This approach is executed in two stages: the first stage creates the forest from the provided dataset, and the second stage handles the prediction based on classification.

4.3 Integration and Deployment

After the sentiment analysis model has been developed, smoothly incorporate it into a user-friendly interface to solicit feedback from stakeholders. In order to ensure maximum efficiency for processing different volumes of reviews, deploy the system to a scalable environment. Assuring continued accuracy in the dynamic world of food evaluations requires regular performance monitoring, tracking important data, and enabling timely adjustments to adjust to changing language trends.

V. RESULTS

An output that satisfies end user criteria and displays information in an understandable manner is considered high quality. Any system's outputs are used to convey processing results to users and other systems. In output design, the location of the information for both the hard copy output and the immediate requirement is specified. It is the user's primary and most direct source of information. The system's ability to support user decision-making is enhanced by clever and efficient output design.

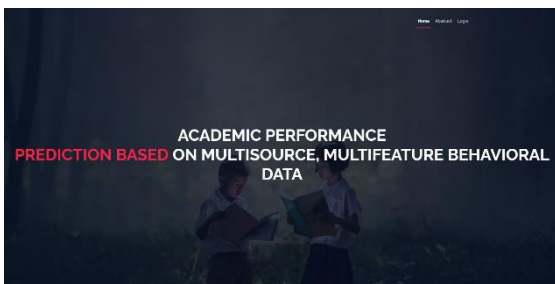
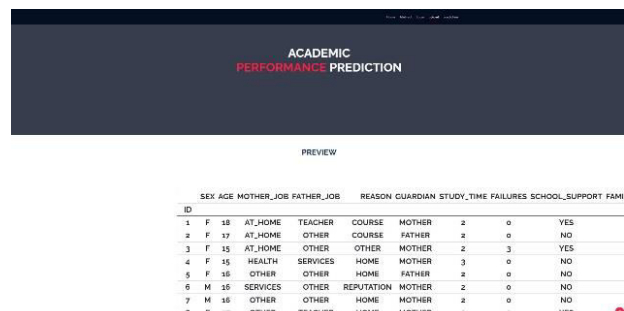


Fig – 2: Homepage



ID	SEX	AGE	MOTHER_JOB	FATHER_JOB	REASON	GUARDIAN	STUDY_TIME	FAILURES	SCHOOL_SUPPORT	FAMIL
1	F	18	AT_HOME	TEACHER	COURSE	MOTHER	2	0	YES	
2	F	17	AT_HOME	OTHER	COURSE	FATHER	2	0	NO	
3	F	15	AT_HOME	OTHER	OTHER	MOTHER	2	3	YES	
4	F	15	HEALTH	SERVICES	HOME	MOTHER	3	0	NO	
5	F	16	OTHER	OTHER	HOME	FATHER	2	0	NO	
6	M	16	SERVICES	OTHER	REPUTATION	MOTHER	2	0	NO	
7	M	16	OTHER	OTHER	HOME	MOTHER	2	0	NO	
8	F	17	OTHER	TEACHER	HOME	MOTHER	2	0	YES	

Fig – 3: Uploaded data preview



Fig – 4: Prediction

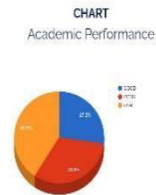
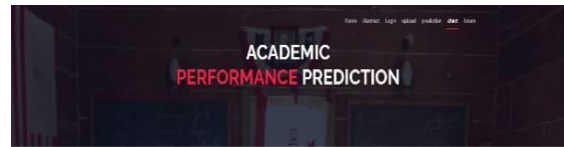
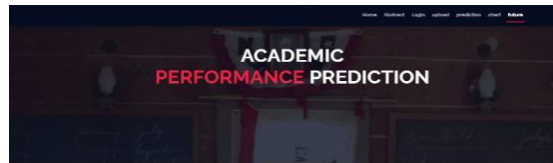


Fig – 5: Uploaded Data Visual Chart



FUTURE

As an important issue in the education data mining field, academic performance prediction has been studied by many researchers. 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. To initially alleviate this problem, our study aims at developing a robust academic performance prediction model. To gain an in-depth insight into student behavioral patterns and potentially help students to optimize their interactions with the university, in our study, a model named AugmentED is proposed to predict the academic performance.

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, analyze and use multisource data covering not only online and offline learning but also campus life behaviors 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, behavioral change is evaluated by linear, nonlinear, and deep learning (LSTM) methods respectively, which provides a systematical view of students' behavioral patterns. Specifically, it is the first time that three novel nonlinear metrics (LFC, NMF, and DFA) and LSTM are applied in students' behavioral 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 under-performing students, however, there are also some limitations in our study. To gain a multisource dataset, we scarfed 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 behavioral 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 positive daily data capture system that exists in most modern universities. This system can potentially lead to continual investigations on a larger scale. The knowledge obtained in this study can also potentially contribute to related research among K-12 students.

Fig – 6: Future

VI. CONCLUSION

In today's digital age, the analysis and extraction of insights from vast datasets are increasingly popular. During our research, we extensively reviewed numerous studies using statistical methods to understand the importance and nuances of visualizing datasets. Employing Python integrated with Tableau, we aimed to visually represent approximately 1000 reviews from Zomato's restaurant database. Various visualization techniques were employed to streamline the process, offering valuable time savings for both users and restaurant owners. Our primary focus was to categorize reviews as positive, neutral, or negative, enabling users to easily assess the overall sentiment. Users could conveniently compare highly-rated local restaurants and identify areas with the most appealing cuisine options. Despite our efforts, some results were classified as neutral due to limitations in the lexicon approach used in Python. This was mainly because certain dialectical or slang terms in the feedback were not adequately covered in standard lexicons.

REFERENCES

- [1] A. Furnham, and J. Mosen, "Personality traits and intelligence predict academic school grades," Learning and Individual Differences, vol. 19, no. 1, pp. 0-33, 2009.
- [2] M. A. Conard, "Aptitude is not enough: How personality and behavior predict academic performance," Journal of Research in Personality, vol. 40, no. 3, pp. 339-346, 2006.
- [3] T. Chamorro-Premuzic, and A. Furnham, "Personality predicts academic performance: Evidence from two longitudinal university samples," Journal of Research in Personality, vol. 37, no. 4, pp. 319- 338, 2003.
- [4] R. Langford, C. P. Bonell, H. E. Jones, T. Poulou, S. M. Murphy, and E. Waters, "The WHO health promoting school framework for improving the health and well-being of students and their academic achievement," Cochrane Database of Systematic Reviews, vol. 4, no. 4, pp. CD008958, 2014.

- [5] A. Jones, and K. Issroff, "Learning technologies: Affective and social issues in computer-supported collaborative learning," *Computers & Education*, vol. 44, no. 4, pp. 395-408, 2005.
- [6] D. N. A. G. Van, E. Hartman, J. Smith, and C. Visscher, "Modeling relationships between physical fitness, executive functioning, and academic achievement in primary school children," *Psychology of Sport & Exercise*, vol. 15, no. 4, pp. 319-325, 2014.
- [7] R. Wang, F. Chen, Z. Chen, T. Li, and A. T. Campbell, "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," In Proc. of the ACM International Joint Conference on Pervasive & Ubiquitous Computing, Seattle, WA, USA, 2014.
- [8] R. Wang, G. Harari, P. Hao, X. Zhou, and A. T. Campbell, "SmartGPA: How smartphones can assess and predict academic performance of college students," In Proc. of the ACM International Joint Conference on Pervasive & Ubiquitous Computing, Osaka, Japan, 2015.
- [9] M. T. Trockel, M. D. Barnes, and D. L. Egget, "Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors," *Journal of American College Health*, vol. 49, no. 3, pp. 125-131, 2000.
- [10] D. M. Hansen, S. D. Herrmann, K. Lambourne, J. Lee, and J. E. Donnelly, "Linear/nonlinear relations of activity and fitness with children's academic achievement," *Med Sci Sports Exerc.* vol. 46, no. 12, pp. 2279-2285, 2014.
- [11] A. K. Porter, K. J. Matthews, D. Salvo, and H. W. Kohl, "Associations of physical activity, sedentary time, and screen time with cardiovascular fitness in United States adolescents: Results from the NHANES national youth fitness survey (NNYFS)," *Journal of Physical Activity and Health*, pp. 1-21, 2017.
- [12] K. N. Aadland, O. Yngvar, A. Eivind, K. S. Bronnick, L. Arne, and G. K. Resaland, "Executive functions do not mediate prospective relations between indices of physical activity and academic performance: the active smarter kids (ask) study," *Frontiers in Psychology*, vol. 8, pp. 1088, 2017.
- [13] M. Credé, S. G. Roch, and U. M. Kieszczynska, "Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics," *Review of Educational Research*, vol. 80, no. 2, pp. 272-295, 2010.
- [14] S. P. Gilbert, and C. C. Weaver, "Sleep quality and academic performance in university students: A wake-up call for college psychologists," *Journal of College Student Psychotherapy*, vol. 24, no. 4, pp. 295-306, 2010.
- [15] A. Wald, P. A. Muennig, K. A. O'Connell, and C. E. Garber, "Associations between healthy lifestyle behaviors and academic performance in U.S. undergraduates: A secondary analysis of the American college health association's National College Health Assessment II," *American Journal of Health Promotion*, vol. 28, no. 5, pp. 298-305, 2014.
- [16] P. Scanlon, and A. Smeaton, "Identifying the impact of friends on their peers academic performance," In Proc. of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016.
- [17] E. L. Faight, J. P. Ekwaru, D. Gleddie, K. E. Storey, M. Asbridge, and P. J. Veugelers, "The combined impact of diet, physical activity, sleep and screen time on academic achievement: A prospective study of elementary school students in Nova Scotia, Canada," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 14, no. 1, pp. 29, 2017.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details