ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 6, June 2023

Student's Performance Analysis using Random Forest and Prediction using Light GBM

[1] Dr.K.Madhavi, [2] Mogulla Pallavi, [3] Rumana Tarannum, [4] Sirikonda Manasa, [5] Kurremula Sreenija

[1] [2] [3] [4] [5] Department of Computer Science and Engineering,
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India
Corresponding Author Email: [1] bmadhaviranjan@yahoo.com, [2] allavireddya2001@gmail.com,
[3] tarannumrumana0@gmail.com, [4] sirikondamanasa11@gmail.com, [5]sreenijakurremula8@gmail.com

Abstract— In the area of the student's interests, the student's performance analysis system, which is focused on learning, aspires to excellence on many different levels and dimensions. This method is designed to evaluate and forecast student performance. The suggested approach examines student demographics, academic data, parental education, and other factors in an effort to elicit as much data as possible from students, teachers, and parents. Analyzing student academic performance is essential for academic institutions and instructors in order to explore methods to improve individual student performance. Using a set of data mining algorithms to achieve the maximum possible accuracy in student performance prediction. This framework is effective in highlighting the student's weaknesses which are based on experimented studies for improving student's academic achievement.

Keywords — Student Performance, Logistic Regression, Random Forest, LightGBM.

I. INTRODUCTION

Predicting student success is increasingly challenging now since educational databases include enormous amounts of data. The lack of a developed system for assessing and monitoring student achievement is also not being taken into account. There are primarily two causes for this type of situation. Firstly, there is inadequate study on the various prediction techniques to choose the ones that will best forecast students' success in academic progress. Secondly, there is the lack of investigation into the particular courses.

The main objective is to have a general understanding of the artificial intelligence systems that were employed to forecast academic learning. This study also focuses on applying prediction algorithms to categorize the student data's most important characteristics. This research focuses on employing educational machine learning approaches to consistently and significantly increase student performance and progress. Academic institutions, educators, and students all stand to gain from this.

II. LITERATURE SURVEY

In reference [1], the prediction of SAP is based on a variety of student characteristics, including personal, social, psychological, and environmental variables. Numerous studies have been conducted in recent years to forecast kids' academic achievement. Therefore, in this area, a few research papers are taken into account and analysed them for various student aspects that influence the forecast of a student's academic performance. There are research papers, articles, and book chapters that are subject to evaluation. In their study, Farhana Sarker and Hugh C. Davis (2013) shown that the model relying solely on institutional internal student databases performed worse than that using both institutional internal data sources (IDS) and external data sources (EDS).

In reference [4], gathered data for their study from students in the information technology department at Kin Saud University in Saudi Arabia. They also employed other attributes for prediction, including student ID, student name, marks from three distinct quizzes, midterms 1, midterms 2, projects, tutorials, final exams, and total points earned in the department of computer science's Data Structure course. First-year students' questionnaires and data acquired during enrolment at the University of Tuzla were the sources of the data that Edin Osmanbegovi and Mirza Suljic (2012) gathered. In addition, they utilized a variety of factors to make their predictions, including gender, family distance, high school GPA, entrance test, scholarship, time, materials, the internet, grade relevance, and earnings. [2].

In reference [8], In this reference, K-means clustering is used. This uses GPA, age, ethnicity, and gender. Here, 7 clusters are created. Based on the results of their experiment, they found that the K means algorithm clusters are more useful and informative when K=7. The K-means experimental outcomes on the entire data set suggested that middle-aged (around 45) students tend to be high-performing students than are male students in the software major who are between the ages of 24 and 27, with the majority of them coming from European, Maori, and Asian backgrounds. On the other hand, teenage male network majors are more likely than any other group to be poor achievers.

In reference [6], Through the analysis of the student's performance using data mining classification techniques, a system is constructed to predict student academic achievement in the course "TMC1013 System Analysis and

Design" given by FCSIT. Furthermore, Student Performance 4 Analysis System (SPAS) is designed to help instructors consult with students by providing lecturers access to the student's prior performance in a certain course and semester.

In reference [9], this paper compares two approaches to assisting students in meeting course learning outcomes: rubric- based self-assessment and oral feedback from the teacher. Both strategies have been demonstrated to result in better learning outcomes when used on completed assignments and when students are given time to respond to the feedback given. While the teacher's oral feedback requires an investment of both the teacher's and the students' time, the rubric-based self- assessment does not. The number of complaints about grades has decreased significantly as a result. This finding suggests that the interventions helped students understand what was expected of them in the assessment tasks.

III. EXISTING MODEL

Examining related technologies that are currently in use to analyse student performance. Faculty Support System (FSS) has created a framework called Faculty Web to track students' progress in a specific course offered by Coimbatore Institute.

It is expensive (since it employs cost-effective opensource analytic software), WEKA. Because it can update student data dynamically as time passes to create or add newrules, FSS is able to analyse student data in a dynamic manner. With the help of experienced professionals, it is possible to change a rule that has been created through data processing techniques like categorization. A classification approach is used to forecast the performance of the pupils. Additionally, FSS is an expert in identifying factors that affect students' success in a given course.

IV. PROPOSED MODEL

By outlining the student's demographic information, study- related information, and parental education. The suggested approach begins by merging demographic and research- relevant information with educational disciplines. This system employs a variety of algorithms to assess and forecast student performance. For assessing student performance, the Random Forest and Logistic Regression algorithms are used. LightGBM is used to forecast student ‘s performance. Excluding previous exam scores all other factors are considered for prediction.

The proposed Machine learning system's components are depicted in the figure1 below

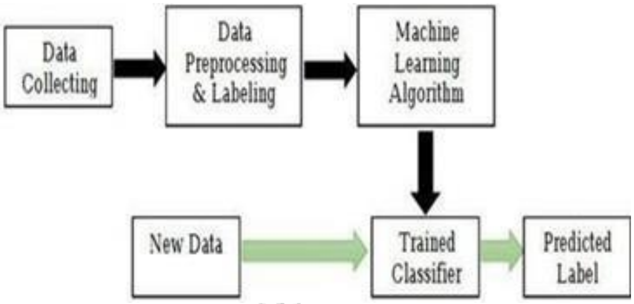

Figure 1: The main steps and Components of proposed System

Data Collection

A group of student records' information were utilised as the data for this project. The selection of the subset from all available data is the focus of this stage. Data that already know the desired outcome are referred to as labelled data.

Data Pre-processing

Format, clean, and sample the selected data to organise it. There are three typical steps in data pre-processing:

1.Formatting 2.Cleaning 3.Sampling **Feature Extraction**

Feature extraction is the process of reducing the number of attributes. The current characteristics are really transformed by feature extraction, which contrasts with feature selection, which ranks the attributes according to their predictive relevance. The original attributes are linearly combined to create the changed attributes, or features.

Evaluation Model

The process of developing a model includes model evaluation. It is beneficial to determine which model best reflects our data as well as how the final model will function in the future. In data science, it is not acceptable to evaluate model performance using the training data because this may easily lead to overly optimistic and overfitted models.

Data Understanding

To design the system and accomplish the project's goals and objectives, data knowledge is essential before system development. In addition, comparisons with other comparable systems are made in order to better understand their features, strengths, and weaknesses.

System Analysis and Design

In this stage, the total system flow is planned, examined, and developed. First, examine the system and user needs based on your understanding of the situation, then list them in tabular form.

Implementation and Testing

A dataset of student data is collected and analyzed using data mining techniques throughout the implementation phase in order to produce rules for the analysis of student performance. Jupiter Notebook, an open-source software platform, is used to generate rules. Training and test sets are created from the dataset. The training set uses 25% of the dataset, with the test set using the remaining 75%.

V. IMPLEMENTATION

The objective of this project is to analyse student performance. It is a machine learning model that takes student data as input and analyses using logistic regression and random forest. The below table displays parameters used.

Table 1

	Feature	Values	Description
1	Gender	Male/Female	Student's gender
2	Race/Ethnicity	Group: A/B/C	Group student belongs to
3	Parental level of Education	Bachelor's/ Master's/ other degree	Student parent's qualification
4	Sports	Y/N – 1/2	Student's participation in sports 1-yes 2-no
5	Test preparation course	Completed/ none	Student's testcompletion status
6	Math score	Integer:(0-100)	Student score in math subject
7	Reading score	Integer:(0-100)	Student score in reading
8	Writing score	Integer:(0-100)	Student score in writing
9	Attendance	Integer:(0-100)	The number of days the student is present

Steps:

1. In first step, student's dataset is taken.
2. Next data cleaning is done where noisy data and incomplete data was removed.
3. In this step, data is divided into part for training and testing purposes. 25% of the data was kept for training the model and 75% for testing it.
4. Analysis of student's performance is done using logistic regression and random forest algorithms. Logistic regression was implemented as it is easier to interpret and very efficient to train. It takes less training time as compared to other algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently.
5. After analysis race/ethnicity and parental level of education columns were excluded as they are not affecting the result of students. Previous exam scores were not considered and the remaining factors are considered for prediction which is done using LightGBM algorithm. LightGBM was used as it requires less memory to execute and can handle enormous amounts of data.

VI. RESULT

The below figure2 shows confusion matrix for Logistic Regression:

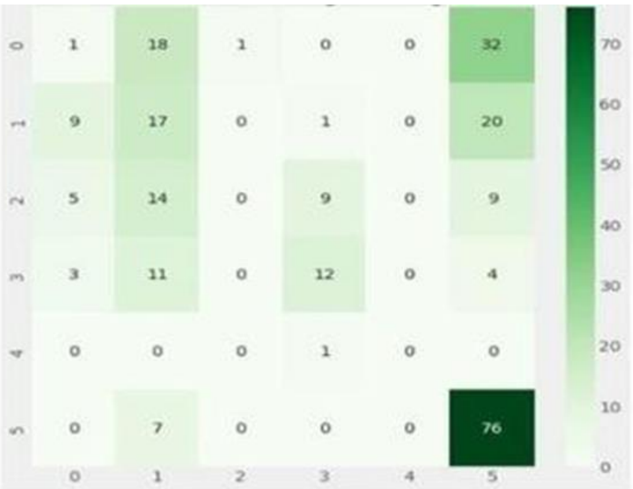

Figure 2: Confusion matrix for Logistic Regression

The confusion matrix shown above helps in analysing the accuracy of results when logistic regression is implemented. This test result's accuracy for testing data is 0.424.

The below figure3 shows confusion matrix for Random Forest:

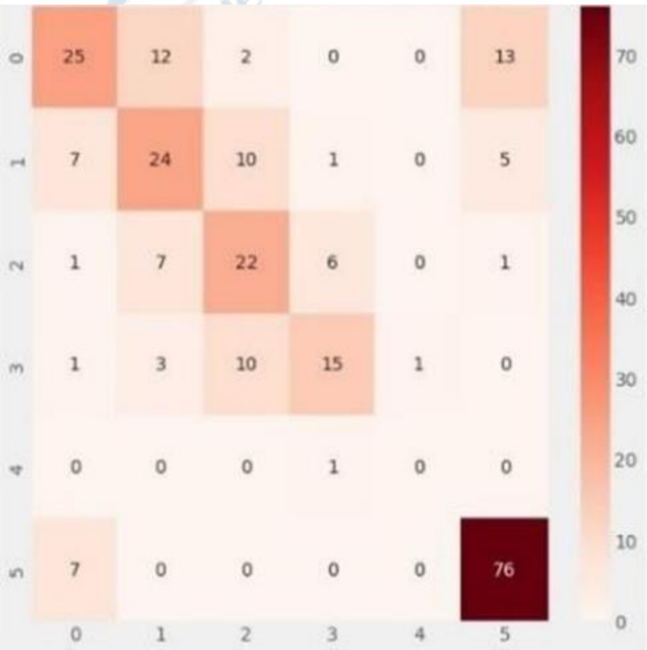

Figure 3: Confusion matrix for Random Forest

The confusion matrix shown above helps us in analysing the accuracy of results when random forest is implemented. This test result's accuracy for testing data is 0.648.

The dataset is split into training and test sets in both techniques. The training set takes up 25% of the dataset, while the test set takes up the remaining 75%.

Test result is visualized as shown below:

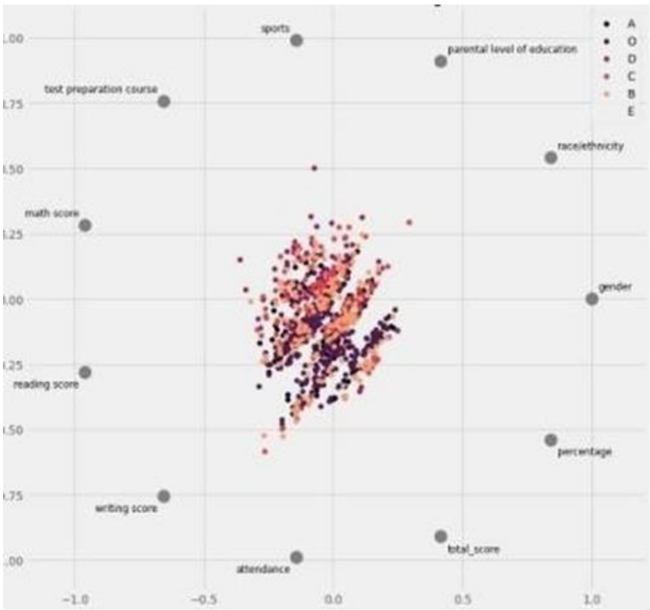

Figure 4: Radial Visualization for target

The student analysis is displayed using a scatter plot, where each colour corresponds to a certain grade. Students deviate more when their sports practice is prolonged or when absenteeism lowers the attendance rate.

VII. CONCLUSION AND FUTURE SCOPE

The research's main goal is to employ machine learning techniques to analyse how students' academic progress is developing. The analysis is done using random forest and logistic regression. This model can be utilized in variety of circumstances, including departmental level and at basic academic level for presenting a concise summary of the performance related to particular course. With a few modest alterations, any industrial organization or company can use this application to assess task participation and determine the perfect candidate based on productivity.

This procedure can make it easier for the instructor to assess student performance and plan more effective academic improvement strategies. All other elements are taken into account for prediction besides past exam results. Future updates to our dataset could include more features for improved accuracy and in-depth research.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “From data mining to knowledge discovery databases.,” AI Magazine, pp. 37-53, 1996.

[2] “IndiceBrasscom de Convergencia Digital,” 2015. [Online]. Available: www.brasscom.org.br. [Accessed: 14- May-2016].

[3] H. Witten, E. Frank and M. A. Hall, Data mining: Practical Machine Learning Tools and Techniques. Burlington: Morgan Kaufmann, 2011.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," in ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10-18, Jun. 2009.

[5] B. K. Baradwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance," in International Journal of Advanced Computer Science and Applications– IJACSA, vol. 2, no. 6, pp. 63-69, 2011.

[6] Poza-Lujan, Jose-Luis and Calafate, Carlos T. and PosadasYague. ”As- sessing the Impact of Continuous Evaluation Strategies: Tradeoff Between Student Performance and Instructor Effort”, IEEE Transactions on Edu- cation, vol.59, pp.17-23, Feb 2016.

[7] Elbadrawy, Asmaa and Polyzou, Agoritsa and Ren, Zhiyun and Sweeney. ”Predicting Student Performance Using Personalized Analytics”, IEEE,vol. 49,pp. 61-69, Apr.2016.

[8] Ganeshan, Kathiravelu and Li, Xiaosong.”An intelligent student advising system using collaborative filtering”, 2015 IEEE Frontiers in Education Conference (FIE), pp. 1-8, Oct. 2015.

[9] Barney, Sebastian and Khurum, Mahvish and Petersen, Kai and Unterkalmsteiner, Michael and Jabangwe, Ronald.” Improving Students with Rubric-Based Self-Assessment and Oral Feedback.” IEEE Transactions on Education, vol. 55, pp.319-325, Aug 2016.

[10] P. M. Arsad, N. Buniyamin, and J. L. A. Manan, “A neural network students’ performance prediction model (NNSPPM),” 2013 IEEE Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA 2013, no. July 2006, pp. 26–27, 2013.

[11] K. F. Li, D. Rusk, and F. Song, “Predicting student academic performance,” Proc. - 2013 7th Int.Conf. Complex, Intelligent, Softw. Intensive Syst. CISIS 2013, pp. 27–33, 2013.

[12] G. Gray, C. McGuinness, and P. Owende, “An application of classification models to predict learner progression in tertiary education,” in Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014, 2014.

[13] N. Buniyamin, U. Bin Mat, and P. M. Arshad, “Educational data mining for prediction and classification of engineering students’ achievement,” 2015 IEEE 7th Int. Conf. Eng. Educ. ICEED 2015, pp. 49–53, 2016.

[14] Z. Alharbi, J. Cornford, L. Dolder, and B. De La Iglesia, “Using data mining techniques to predict students at risk of poor performance,” Proc. 2016 SAI Comput. Conf. SAI 2016, pp. 523–531, 2016.

[15] Simpson, Jane and Fernandez, Eugenia.” Student performance in first year, mathematics, and physics courses: Implications for success in the study of electrical and computer engineering”, 2014 IEEE Frontiers in Education Conference (FIE) Proceedings, pp. 1- 4, Oct 2014.