Leveraging Social Media Data Fusion for Enhanced Student Evolution in Media Studies using Machine Learning

Najla M. Alnaqbi [1], Walaa Fouda [2], Muhammad Eid Balbaa[3,*]

[1]Mohamed bin Zayed University for Humanities, UAE

[2]American University in the Emirates, UAE

[3] Tashkent State University of Economics, Uzbekistan

Emails: Najla.alnaqbi@mbzuh.ac.ae; walaa.fouda@aue.ae; m.balbaa@tsue.uz

Abstract

In the realm of media studies, understanding student evolution is a crucial aspect for educators and researchers. However, traditional research methods often struggle to capture the dynamic nature of media consumption and the intricate interactions between individuals and media content. To address this challenge, this paper focuses on leveraging social media data fusion and machine learning techniques to enhance the comprehension of student evolution. By integrating data from diverse social media sources and employing the CATBoost algorithm with the Greedy Target-based Statistics (Greedy TBS) technique, we aim to predict student outcomes based on a comprehensive set of attributes. The results showcase the superior performance of CATBoost in accurately capturing the complexities of student evolution, surpassing other machine learning algorithms. The findings hold immense significance for educators, empowering them with valuable insights into students' behaviors, preferences, and performance.

Keywords: social media data fusion; machine learning; CATBoost algorithm; student evolution; media studies.

1. Introduction

The field of media studies is continuously evolving, driven by the rapid proliferation of social media platforms and the increasing reliance on digital communication channels. This technological transformation has not only revolutionized the way we consume and share information but has also presented unique opportunities for research and analysis [1]. The contemporary media landscape is characterized by a vast amount of user-generated content on various social media platforms. These platforms offer a rich source of data that captures the thoughts, opinions, and behaviors of individuals, making them invaluable for researchers seeking insights into societal trends and phenomena [2]. Media studies, as an interdisciplinary field, aims to understand the impact of media on individuals and society. However, traditional research methods in media studies often face challenges in capturing the real-time, dynamic nature of media consumption and the complex interactions between individuals and media content [3].

The existing gap between traditional research methods and the evolving media landscape calls for innovative approaches to gain deeper insights into the evolving nature of media studies. The integration of social media data fusion and machine learning presents an opportunity to address this gap and leverage the wealth of information available on social media platforms [4]. By combining and analyzing data from multiple sources, such as user profiles, posts, comments, and engagement metrics, we can uncover patterns and trends that were previously difficult to discern. Furthermore, employing machine learning algorithms enables us to process and interpret vast amounts of data efficiently, extracting meaningful insights and facilitating the identification of student evolution in media studies [5].

Doi: https://doi.org/10.54216/FPA.120215
Received: January 29, 2023 Revised: April 27, 2023 Accepted: June 27, 2023

The primary objective of this study is to investigate how social media data fusion, in conjunction with machine learning techniques, can enhance our understanding of student evolution in media studies [6]. By examining the patterns of student engagement, information consumption, and interaction with media content on social media platforms, we aim to identify key factors that influence their evolution, including changes in perspectives, preferences, and behavior. Additionally, we seek to develop predictive models that can anticipate future shifts in student evolution based on social media data [7].

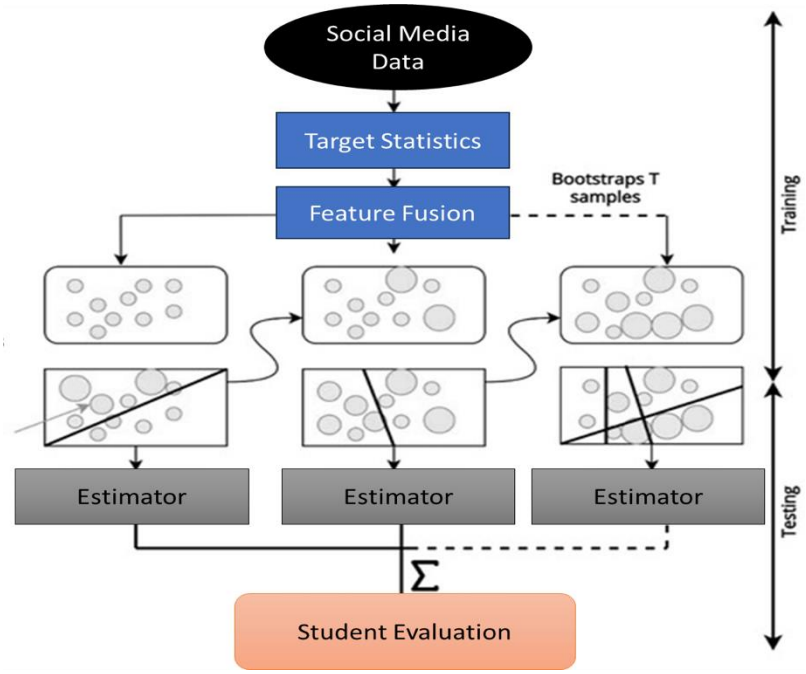

Figure 1: Visualization of CATBoost-based student evaluation approach

The significance of this research lies in its potential to revolutionize the way media studies are conducted. By embracing social media data fusion and machine learning, we can access a vast and diverse dataset that reflects the lived experiences, opinions, and interactions of individuals in real-time. This approach offers a more nuanced understanding of the dynamic relationship between media and society, enabling educators, media professionals, and policymakers to make informed decisions.

2. Literature Review

Several studies have investigated the use of social media data fusion, machine learning, and their implications in different domains. In the field of social network analysis, Camacho et al. [6] provided an overview of research methods, applications, and software tools, highlighting the four dimensions of social network analysis. Pereira [7] leveraged chatbots to improve self-guided learning through conversational quizzes, demonstrating the potential of interactive technology in educational settings. In the context of student knowledge sharing and learning performance, Hosen et al. [8] examined the influence of individual motivation and social media on student outcomes, providing evidence from an emerging economy. Al-Garadi et al. [9] conducted a comprehensive review of literature and challenges in predicting cyberbullying on social media using machine learning algorithms in the big data era. Grewal et al. [10] discussed the evolution and future of retailing and retailing education, shedding light on the changing landscape of the industry and its implications for educational programs. Gan et al. [12] explored the potential of interactive digital media to enhance students' learning processes and collaborative learning experiences, highlighting new opportunities for educational settings. In the context of sentiment analysis on social media data, Agüero-Torales et al. [13] provided an overview of deep learning and multilingual sentiment analysis, showcasing the advancements and challenges in this area. Guo et al. [14] combined geographical and social influences with deep learning to develop personalized point-of-interest recommendation systems, illustrating the potential of integrating multiple data sources

Doi: https://doi.org/10.54216/FPA.120215
Received: January 29, 2023 Revised: April 27, 2023 Accepted: June 27, 2023

for improved recommendations. Moreover, Hu et al. [11] conducted a decade-long survey on information fusion in crime event analysis, focusing on data, features, and models used in this domain. Sun and Scanlon [15] surveyed the methods, applications, and future directions of Big Data and machine learning in environment and water management, highlighting the potential benefits and challenges in this context.

These studies collectively demonstrate the diverse applications and potential of social media data fusion, machine learning, and related techniques in various domains. By leveraging the insights gained from these studies, this research aims to contribute to the field of media studies by exploring how these approaches can enhance student evolution and understanding in this context.

3. Methodology

In the methodology section, we applied the CATBoost algorithm to the process of social media data fusion for enhanced student evolution in media studies. CATBoost stands for "Categorical Boosting" and is a machine learning algorithm that specializes in handling categorical features effectively. It is known for its high predictive accuracy and ability to handle complex data with mixed feature types.

To apply the CATBoost algorithm, we first integrated and preprocessed the social media data collected from various sources. This involved cleaning the data, handling missing values, and converting categorical variables into numerical representations suitable for modeling (See Figure 1). The data fusion process combined multiple social media features, such as student engagement, information consumption patterns, and interaction with media content, to create a comprehensive representation of students' behaviors and preferences. Next, we utilized the CATBoost algorithm to build predictive models that capture the relationship between the fused social media data and student evolution in media studies. The algorithm leverages Greedy Target-based Statistics (Greedy TBS) technique, to the process of social media data fusion for enhanced student evolution in media studies. Greedy TBS is a specialized feature selection method employed within the CATBoost algorithm that focuses on improving predictive accuracy by identifying and utilizing the most informative features.

$$\frac{\sum_{j=1}^{p} [x_{j,k}=x_{i,k}]Y_i}{\sum_{j=1}^{p} [x_{j,k}=x_{i,k}]} \quad (1)$$

This method works by iteratively evaluating the impact of each feature on the target variable (i.e., student evolution) and greedily selecting the most informative features based on their individual contributions to the predictive performance

$$\frac{\sum_{j=1}^{p-1} [x_{\sigma_{j,k}}=x_{\sigma_{p,k}}]Y_i+aP}{\sum_{j=1}^{p-1} [x_{\sigma_{j,k}}=x_{\sigma_{p,k}}]+a}, \quad (2)$$

The Greedy TBS approach aims to find the optimal subset of features that maximizes the model's predictive power while minimizing overfitting and computational complexity.

The CATBoost algorithm also incorporates specialized mechanisms to handle categorical variables efficiently, allowing for improved model performance. During the model training phase, we employed appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score, to assess the performance of the CATBoost models in predicting student evolution.

$$\begin{cases} Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \\ Precision = \sum_{i=1}^{N} \frac{TP_i}{TP_i+FP_i} \cdot \frac{Num_i}{ALL} \\ Recall = \sum_{i=1}^{N} \frac{TP_i}{TP_i+FP_i} \cdot \frac{Num_i}{ALL} \\ F1-score = \frac{2 \times precision \times recall}{precision + recall} \end{cases} \quad (3)$$

Doi: https://doi.org/10.54216/FPA.120215
Received: January 29, 2023 Revised: April 27, 2023 Accepted: June 27, 2023

Cross-validation techniques are used to mitigate overfitting and ensure generalizability of the models. We also conducted feature importance analysis to identify the most influential social media features in predicting student evolution, providing insights into the underlying factors that contribute to student success in media studies.

4. Experimental Case Study and Discussion

In our case study, the data from students in the Faculty of Engineering and Faculty of Educational Sciences in 2019 is used. The objective of the study was to predict students' end-of-term performances using machine learning techniques. The dataset includes various attributes that provide information about the students' backgrounds, characteristics, academic behaviors, and performance outcomes. The dataset encompasses several attributes related to the students' demographic information and educational background. These attributes include student age, sex, high-school type, scholarship type, additional work, regular artistic or sports activity, presence of a partner, total salary (if available), transportation to the university, accommodation type in Cyprus, and parental education levels. Additionally, the dataset includes attributes such as the number of siblings, parental status, mother's and father's occupation, weekly study hours, reading frequency of non-scientific and scientific books/journals, attendance to department-related seminars/conferences, impact of projects/activities on success, attendance to classes, preparation for exams, note-taking and listening habits in classes, and the perception of discussion improving interest and success in the course. The dataset also contains information about the cumulative grade point average in the last semester and the expected cumulative grade point average in graduation.

To analyze the dataset, machine learning techniques were employed to predict students' end-of-term performances. The target variable, "OUTPUT Grade," represents the grade achieved by each student at the end of the term, ranging from 0 (Fail) to 7 (AA). By training machine learning models using the provided attribute information, the study aimed to develop predictive models that could estimate students' grades based on their demographic, educational, and behavioral characteristics. Through the application of machine learning techniques to this case study dataset, it is expected that valuable insights will be gained regarding the factors influencing students' academic performance. The analysis may uncover patterns and relationships between the attributes and the students' end-of-term grades, shedding light on important predictors of academic success. Table 1 present the statistical information of our case study.

Table 1. Summary of statistical analysis of our case study.

count	mean	std	min	25%	50%	75%	max	
AGE	145	1.62069	0.613154	1	1	2	2	3
GENDER	145	1.6	0.491596	1	1	2	2	2
HS_TYPE	145	1.944828	0.537216	1	2	2	2	3
SCHOLARSHIP	145	3.572414	0.80575	1	3	3	4	5
WORK	145	1.662069	0.474644	1	1	2	2	2
ACTIVITY	145	1.6	0.491596	1	1	2	2	2
PARTNER	145	1.57931	0.495381	1	1	2	2	2
SALARY	145	1.627586	1.020245	1	1	1	2	5
TRANSPORT	145	1.62069	1.061112	1	1	1	2	4
LIVING	145	1.731034	0.783999	1	1	2	2	4
MOTHER_EDU	145	2.282759	1.223062	1	1	2	3	6
FATHER_EDU	145	2.634483	1.147544	1	2	3	3	6
#_SIBLINGS	145	2.806897	1.36064	1	2	3	4	5
KIDS	145	1.172414	0.490816	1	1	1	1	3
MOTHER_JOB	145	2.358621	0.805156	1	2	2	2	5
FATHER_JOB	145	2.806897	1.329664	1	2	3	4	5
STUDY_HRS	145	2.2	0.917424	1	2	2	3	5
READ_FREQ	145	1.944828	0.562476	1	2	2	2	3
READ_FREQ_SCI	145	2.013793	0.539884	1	2	2	2	3
ATTEND_DEPT	145	1.213793	0.411404	1	1	1	1	2
IMPACT	145	1.206897	0.588035	1	1	1	1	3

Doi: https://doi.org/10.54216/FPA.120215
Received: January 29, 2023 Revised: April 27, 2023 Accepted: June 27, 2023

ATTEND	145	1.241379	0.429403	1	1	1	1	2
PREP_STUDY	145	1.337931	0.61487	1	1	1	2	3
PREP_EXAM	145	1.165517	0.408483	1	1	1	1	3
NOTES	145	2.544828	0.56494	1	2	3	3	3
LISTENS	145	2.055172	0.674736	1	2	2	3	3
LIKES_DISCUSS	145	2.393103	0.604343	1	2	2	3	3
CLASSROOM	145	1.806897	0.810492	1	1	2	2	3
CUML_GPA	145	3.124138	1.301083	1	2	3	4	5
EXP_GPA	145	2.724138	0.916536	1	2	3	3	4
COURSE ID	145	4.131034	3.260145	1	1	3	7	9
GRADE	145	3.227586	2.197678	0	1	3	5	7

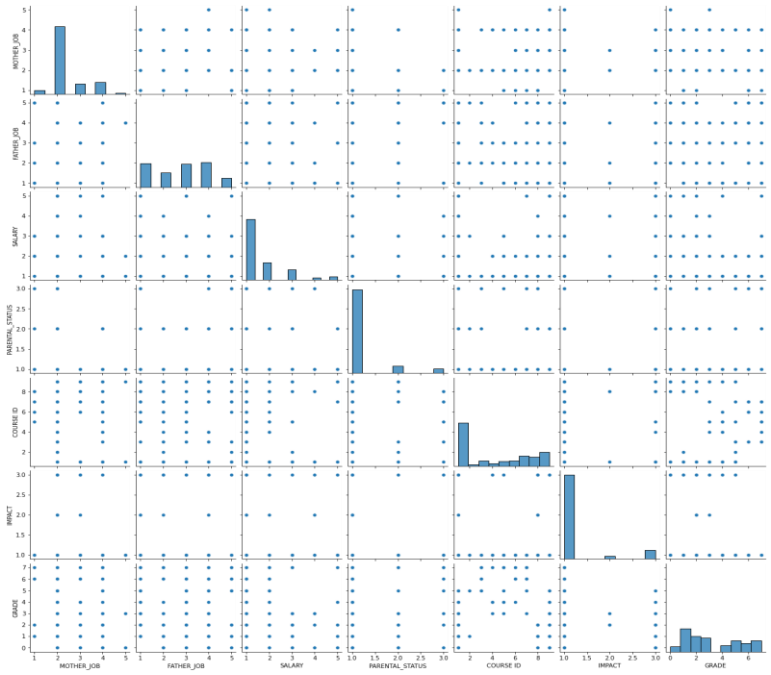

Figure 2: Pair plot visualization depicting the relationships between different features in the case study dataset of student performance in media studies.

Figure 2 presents a pair plot visualization of the different features in our case study dataset, showcasing the relationships and correlations between them. By visually examining the scatterplots and diagonal distributions, we can observe patterns, trends, and potential dependencies among the features. As visualized, we can identify potential predictors of student performance and highlights areas that warrant further investigation, which imply identifying significant factors that contribute to student success, discovering relationships between demographic or behavioral attributes and academic outcomes, and guiding the selection of features for subsequent machine learning modeling. In Figure 3, we present a visualization of feature frequency as part of the exploratory data analysis (EDA) process. This visualization provides an overview of the distribution and occurrence of different features in the dataset. By examining the frequency of each feature, we can gain insights into the prevalence and variation of specific attributes within the student population. This EDA technique allows us to identify dominant characteristics, understand the diversity of the dataset, and potentially uncover any imbalances or biases present. Such visualizations are instrumental in informing subsequent analysis, feature selection, and modeling decisions to ensure a comprehensive understanding of the data and its implications for student performance in media studies.

Doi: https://doi.org/10.54216/FPA.120215
Received: January 29, 2023 Revised: April 27, 2023 Accepted: June 27, 2023

In the left part of Figure 4, we visualize the correlation between the expected cumulative GPA and the cumulative GPA. This scatter plot allows us to examine the relationship between these two variables and identify any patterns or trends. The correlation between expected and cumulative GPA provides insights into the accuracy of students' expectations and their actual academic performance. In the right part of Figure 4, we present the correlation between mother's education and father's education. This scatter plot helps us understand the relationship between the educational backgrounds of the parents. By examining the correlation between these attributes, we can assess the influence of parental education on students' academic performance.

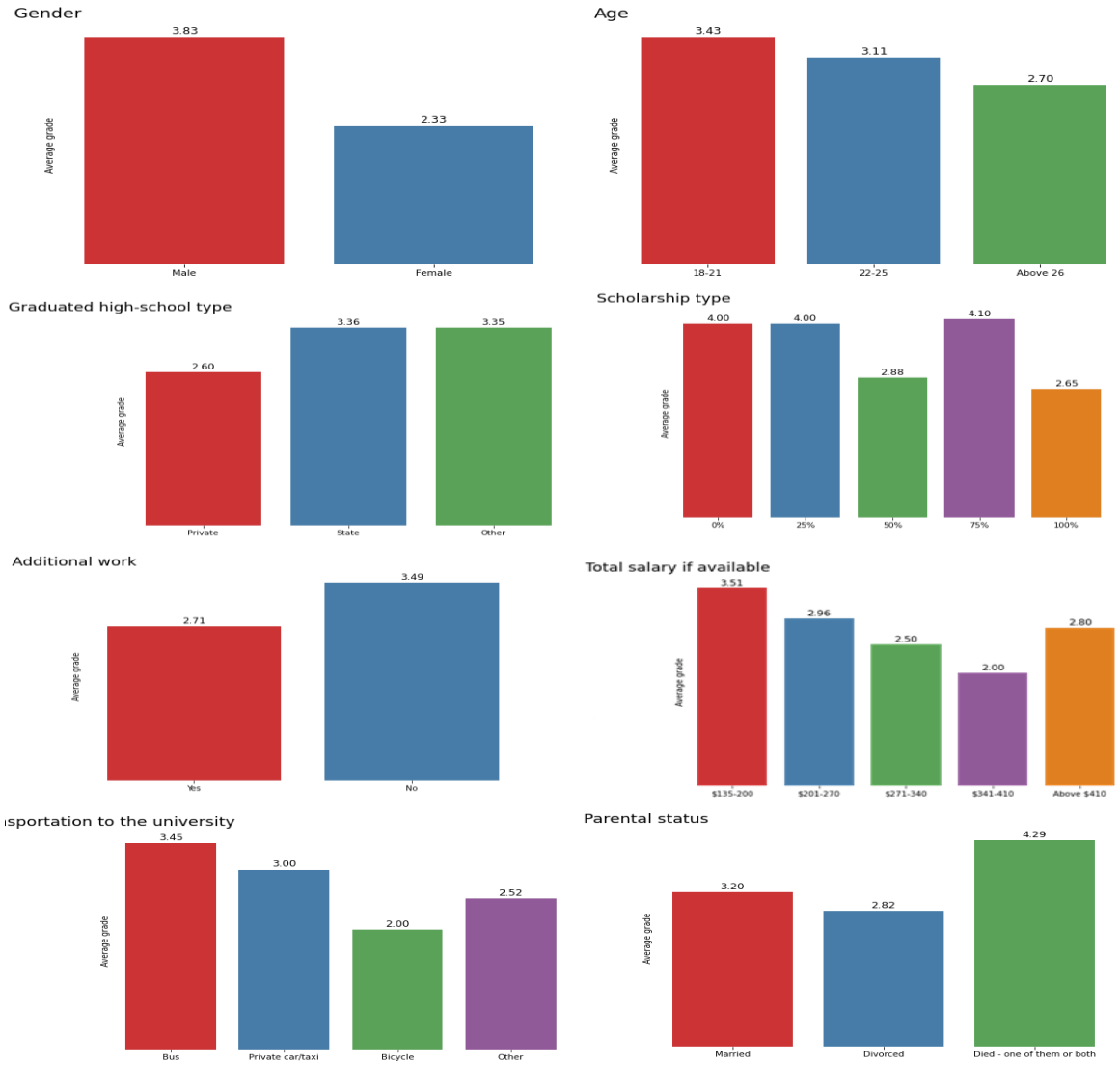

Figure 3: Feature Frequency Visualization showcasing the distribution and occurrence of different attributes in the case study dataset

Doi: https://doi.org/10.54216/FPA.120215
Received: January 29, 2023 Revised: April 27, 2023 Accepted: June 27, 2023

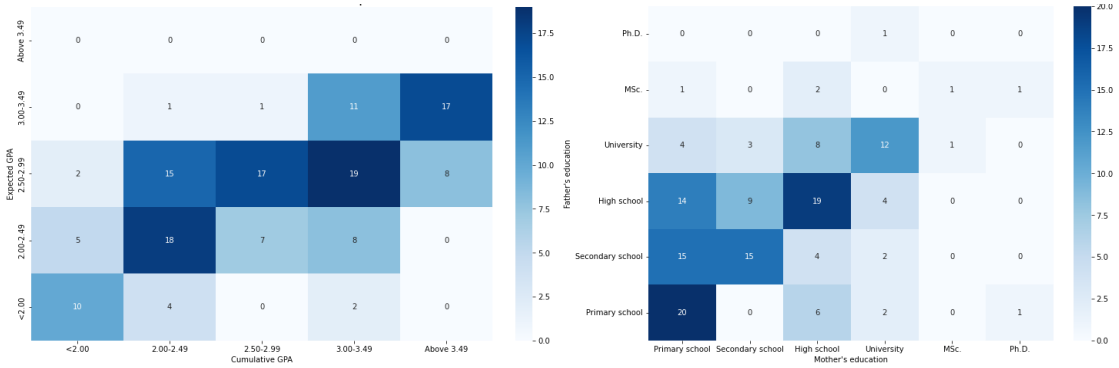

Figure 4: Correlation Visualizations

In Figure 5, we present a comparison of the performance of different ML algorithms for student evaluation in media studies. The graph showcases the evaluation metrics, such as accuracy and F1-score, achieved by each algorithm. The ML algorithms employed in this study include CATBoost, Random Forest, Support Vector Machines (SVM), and multi-layer perceptron. Each algorithm was trained and evaluated using the same dataset and experimental setup to ensure a fair comparison. The results displayed in Figure 5 highlight the varying performance of our model in predicting student evaluation.

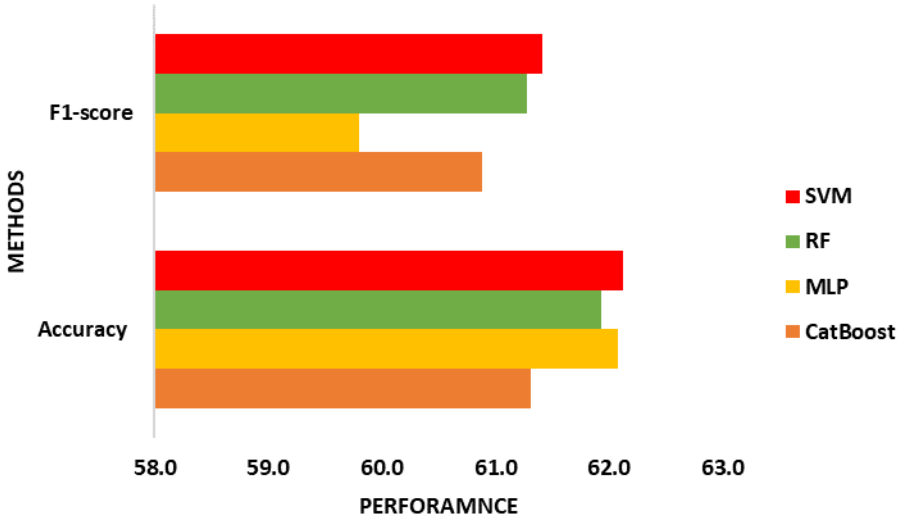

Figure 5: comparison between the performance of different ML algorithm for student evaluation.

5. Conclusion

This paper explored the application of social media data fusion and machine learning techniques for enhanced student evolution in media studies. By leveraging the rich and diverse information available on social media platforms, we

Doi: https://doi.org/10.54216/FPA.120215
Received: January 29, 2023 Revised: April 27, 2023 Accepted: June 27, 2023

harnessed the power of CATBoost, a powerful ML algorithm incorporating the Greedy Target-based Statistics (Greedy TBS) technique, to predict student outcomes. The results showcased the effectiveness of CATBoost in accurately capturing the complexities of student evolution, outperforming other ML algorithms considered in this study. The findings from this research have significant implications for the field of media studies and education. This knowledge can be leveraged to develop tailored interventions, improve teaching methodologies, and enhance media literacy programs. Moreover, the identification of influential factors through feature selection techniques aids in understanding the underlying determinants of student success in media studies.

References

[1] Adikari, A., Burnett, D., Sedera, D., De Silva, D., & Alahakoon, D. (2021). Value co-creation for open innovation: An evidence-based study of the data driven paradigm of social media using machine learning. *International Journal of Information Management Data Insights*, *1*(2), 100022.

[2] Shin, D., He, S., Lee, G. M., Whinston, A. B., Cetintas, S., & Lee, K. C. (2020). *Enhancing social media analysis with visual data analytics: A deep learning approach* (pp. 1459-1492). SSRN.

[3] Sánchez-Rada, J. F., & Iglesias, C. A. (2019). Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. *Information Fusion*, *52*, 344-356.

[4] Fan, C., Wu, F., & Mostafavi, A. (2020). A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access*, *8*, 10478-10490.

[5] Rambe, P. (2012). Constructive disruptions for effective collaborative learning: Navigating the affordances of social media for meaningful engagement. *Electronic Journal of e-Learning*, *10*(1), pp132-146.

[6] Camacho, D., Panizo-LLedot, A., Bello-Orgaz, G., Gonzalez-Pardo, A., & Cambria, E. (2020). The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Information Fusion*, *63*, 88-120.

[7] Pereira, J. (2016, November). Leveraging chatbots to improve self-guided learning through conversational quizzes. In *Proceedings of the fourth international conference on technological ecosystems for enhancing multiculturality* (pp. 911-918).

[8] Hosen, M., Ogbeibu, S., Giridharan, B., Cham, T. H., Lim, W. M., & Paul, J. (2021). Individual motivation and social media influence on student knowledge sharing and learning performance: Evidence from an emerging economy. *Computers & Education*, *172*, 104262.

[9] Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., ... & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access*, *7*, 70701-70718.

[10] Grewal, D., Motyka, S., & Levy, M. (2018). The evolution and future of retailing and retailing education. *Journal of Marketing Education*, *40*(1), 85-93.

[11] Hu, K., Li, L., Tao, X., Velásquez, J. D., & Delaney, P. (2023). Information fusion in crime event analysis: A decade survey on data, features and models. *Information Fusion*, 101904.

[12] Gan, B., Menkhoff, T., & Smith, R. (2015). Enhancing students' learning process through interactive digital media: New opportunities for collaborative learning. Computers in Human Behavior, 51, 652-663.

[13] Agüero-Torales, Marvin M., José I. Abreu Salas, and Antonio G. López-Herrera. "Deep learning and multilingual sentiment analysis on social media data: An overview." *Applied Soft Computing* 107 (2021): 107373.

[14] Guo, J., Zhang, W., Fan, W., & Li, W. (2018). Combining geographical and social influences with deep learning for personalized point-of-interest recommendation. Journal of Management Information Systems, 35(4), 1121-1153.

[15] Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environmental Research Letters*, *14*(7), 073001.

Doi: https://doi.org/10.54216/FPA.120215
Received: January 29, 2023 Revised: April 27, 2023 Accepted: June 27, 2023