**Predicting Engagement Level from Biometric and Environmental Data with Machine Learning Classification Models**

**Mike Rasell Carale Dago-oc1, Chris Angelu Bongcawil Jordan2**

Computer Science and Information Technology Department,

Bachelor of Science in Computer Science (BSCS)

Negros Oriental State University (NORSU), Negros Oriental, Philippines

|  |  |  |
| --- | --- | --- |
| **Article Info** |  | **ABSTRACT** |
| ***Article history:***  Received November 14, 2024  Revised November 14, 2024  Accepted January 17, 2025 |  | Student engagement is a critical factor influencing academic success, with disengaged students often experiencing lower achievement and higher dropout rates. In this study, we explore the use of biometric and environmental data to predict student engagement levels, with this, it is possible to measure student engagement in real-time. To guarantee model stability, a dataset of 1,000 items and 13 features was preprocessed. This study performed a comparative study on some machine learning algorithms including; Random Forest, Logistic Regression, and Support Vector Machine. Each model was accessed using the 10-fold cross-validation. The predictive results show that Random Forest outperformed other models by attaining an accuracy of 97% with precision, recall, and f1-score values of 0.970, 0.972, and 0.970 respectively. SVM also demonstrated strong performance with an accuracy of 91%, while LR showed the lowest performance at 69%.  *This is an open access article under the* [*CC BY-SA*](https://creativecommons.org/licenses/by-sa/4.0/) *license.* |
| ***Keywords:***  Classification Model  Support Vector Machine (SVM)  Random Forest  Logistic Regression  Accuracy  Precision  Recall  F1-Score |

1. **INTRODUCTION**

Student engagement, characterized as time on task, active participation in the learning process, and attention to the area of focus, has historically been strongly correlated to student achievement. These associations hold for different educational activities, all topic areas, and all instructional levels [1]. Engaged students attend class regularly, complete assignments on time, and grasp material more deeply, leading to better test scores and overall performance [2]. This relationship is reciprocal because academic success boosts motivation, creating a positive cycle of learning and achievement [3]. In comparison, disengagement leads to lower achievement, boredom, alienation, and higher dropout rates [4].

Understanding student engagement in academic institutions provides crucial insights into the effectiveness of teaching methods. This data serves as a powerful tool for educators and researchers to refine strategies that enhance student learning. It provides objective insights into student experiences, going beyond assumptions and anecdotes to reveal their true engagement. This valuable information helps institutions improve academic programs while also supporting marketing, recruitment, and addressing students' evolving learning needs [4]. Machine Learning (ML) is being employed by researchers to analyze this data and gain a deeper understanding of the patterns within it [5].

Machine learning can be described as using statistics and data to create a substantial probability of an event. This is comparable to human learning. Machine learning algorithms look for significant connections among a collection of facts and attempt to match inputs and outputs [6]. One of the widely used applications of machine learning in predicting engagement levels is classification [7]. Various classification techniques have been employed in this area of research to categorize students based on their level of engagement. The level of student engagement provides valuable insights into their academic performance, and several innovative approaches have emerged as significant in this area. [8-9]. Researchers asserted that predicting student performance is a critical element in educational settings such as colleges and universities, as it provides a solid foundation for the development of effective learning structures that reduce dropout rates and enhance academic performance. [8].

The undeniable need to spot students with the possibility of performing below normal is necessary to avoid expulsion or dropout in the process of learning [10]. In this context, this study contributes to the existing body of knowledge in several key ways. First, it offers a thorough examination of related literature, focusing on crucial aspects such as data collection, data pre-processing, model creation, and evaluation techniques in the context of predicting student engagement levels. Additionally, the study suggests a new model that demonstrates greater efficiency and accuracy. The suggested model stands out for its simplicity and improved accuracy, offering a more effective approach to predicting student engagement levels.

The rest of this paper is structured thus: Section 2 provides a detailed review of recent literature in the field of student engagement prediction, Section 3 outlines the methodology employed in this study, and further provides a brief description of the machine learning algorithms used in this investigation. While Section 4 presents the results and discussion, Section 5 concludes the research, and Section 6 offers recommendations.

1. **LITERATURE REVIEW**

Machine learning, once a theoretical concept, has rapidly evolved into a transformative force shaping nearly every aspect of our lives. With its ability to learn from data, identify patterns, and make predictions, it has contributed greatly to the overall advancement of our technology, reducing risk and maximizing performance [5]. This section is a summary of existing related papers previously published about student engagement classification.

The researchers in [5] researched 'Engagement Level Prediction using Benchmark Datasets' and tested the predictive performance of 9 machine learning models, which are, Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), Stochastic Gradient Descent (SGD), LogitBoost (LB), Sequential Minimal Optimization (SMO), Voted Perceptron (VP), Adaptive Boosting (AB), and Support Vector Machine (SVM). The dataset is made up of 486 instances and has 12 features. The research also employed information gain to evaluate the significance of the features in the data. The Waikato Environment for Knowledge Analysis (WEKA) version 3.8.6, powered by Java, was used to preprocess the data and apply machine learning algorithms, simplifying the analysis process. Additionally, 10-fold cross-validation was applied to train the models, while precision, accuracy, recall, and F1-score were used to evaluate their performance. SMO outperforms the others with an accuracy of 90%, a precision of 0.897, a recall of 0.897, and a f-measure of 0.897.

Researchers in [11] used 3 models including, Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN). The dataset employed for the study has 348 instances and 9 features. Moreover, the models were trained using 10-fold cross-validation, and their performance was evaluated based on precision, accuracy, recall, and F1-score. Among the models, ANN outperformed the other models with an accuracy of 85%, followed by DT with 80%, and the SVM, which had the lowest accuracy at 75%. However, the DT model stood out with a 1% improvement in precision.

While researchers in [12] employed several models, including eXtreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), CatBoost (CB), Multi-layer Perceptron (MLP), and Random Forest (RF), to analyze student engagement. Their analysis was based on the Open University Learning Analytics Dataset (OULAD) with 32,593 instances and 7 features. Researchers also employed permutation-based feature importance to evaluate the significance of the features in the data. The hyperparameters of these models were optimized using the Optuna software framework to enhance model performance. To evaluate the predictive capabilities of the trained models, metrics such as accuracy, precision, recall, and F-measure were used. With all the models used, CB outperformed the other trained models with an accuracy of 92.23%, a precision of 94.40%, and a recall of 100%.

The similarity of these studies is that they don't use biometrics or environmental data. Instead, they use demographic information, academic metrics, and behavioral and communication data [5, 11, 12]. This makes real-time engagement tracking difficult because some of these data are static like demographic information or need to be pre-recorded like academic metrics. However, this research overcomes these limitations by using biometrics and environmental data which can be monitored in real-time to predict the learner's level of engagement level.

1. **METHODOLOGY**

This section provides an outline of the research methodology employed in this study. A quantitative experimental research design was utilized, which involves the use of numerical data to address a problem. This approach includes data collection, analysis, and hypothesis testing to conclude [13]

**3.1. Hardware and Software**

The study was carried out on a system running Windows 10 with a 64-bit operating system. The system uses an Intel Core™ i5-7400 CPU with 16GB Random Access Memory (RAM). The researchers utilized Jupyter and Python version 3.12.8 as the primary programming language for data analysis and model implementation with the following libraries: Pandas, Numpy, Scikit-Learn, Matplotlib, and Seaborn.

**3.2. Data Acquisition**

This research employs a dataset from Kaggle, a .csv file containing 1000 instances and 13 features, categorized into biometric and environmental data. The biometric data includes Heart Rate (HR), Skin Conductance (SC), Electroencephalography (EEG), Temperature (T), Pupil Diameter (PD), Smile Intensity (SI), Frown Intensity (FI), Cortisol Level (CL), Activity Level (AL), Emotional State (ES), and Cognitive State (CS). At the same time, the environmental data consists of Ambient Noise Level (ANL) and Lightning Level (LL). The target variable is Engagement Level (EG), which is categorized as Highly Engaged (1), Moderately Engaged (2), and lastly Disengaged (3) [14]. This dataset provides the basis for evaluating the relationship between these elements and engagement levels. To ensure robust model training and evaluation, the data was divided into 70% for training and 30% for testing.

**3.3. Data Pre-processing**

Data pre-processing is an essential aspect of model development. Data acquired in their raw form contain noise and anomalies, which can affect the performance and training process of the model being schooled [15]. The researchers employed several techniques to clean the data, including:

**Data normalization:** This is a pre-processing technique primarily intended to manage numerical features and is applied to numerical features before the application of classification algorithms. Normalization is crucial to prevent the effect of certain features from being concealed by others, particularly when the ranges of the features are inconsistent [16]. The normalization techniques used are Min-Max Normalization and Z-Score Normalization for Support Vector Machine (SVM).

**Missing value:** Missing value is a datum that has not been stored or gathered due to issues like faulty sampling procedures, budgetary constraints, or limitations in the data collection process. Missing values are an inevitable aspect of data analysis and can present significant challenges for data practitioners. It is generated due to several reasons, including human mistakes, technical malfunctions, unavailable data, or outdated and inconsistent data [5].

**3.4. Machine Learning Algorithms**

This section focuses on the machine learning classification models utilized in this study. After data pre-processing, the machine learning workflow progresses to the model training stage, where an algorithm is taught to learn from data and produce predictions. This algorithm is specifically responsible for the classification of learners' engagement levels based on their biometric and environmental data. To find the best classifier for engagement level prediction, several classifiers, including Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) were tested through a variety of tests.

**3.4.1. Random Forest**

Random forest is a robust ensemble classifier that integrates multiple decision trees to make predictions. This method of combining classifiers gives the random forest unique characteristics that distinguish it from conventional tree classifiers [5, 18, 23]. A single decision tree classifier can be sensitive to outliers or noise, which may affect overall model performance. In contrast, the Random Forest (RF) classifier incorporates randomness to reduce this vulnerability [23]. Additionally, random forests introduce randomness not only to the data but also to the features. By applying principles similar to those used in bootstrapping and bagging classifiers, random forests diversify their trees by training them on unique data subsets generated through bootstrap aggregation, tailored for regression and classification tasks [24].

Here, D represents the set of instances within the node, c denotes the number of classes, and Pi​ refers to the proportion of instances belonging to class i in node D [5].

**3.4.2. Logistic Regression**

Logistic regression is often used to analyze and describe the relationship between entities when the outcome has only two possible options, such as 'Yes' or 'No,' and a set of predictive entities [25]. Logistic Regression calculates the odds of several classes employing a boundary rationality distribution as depicted in the expression below [5].

Where k = 1, 2, 3,…,k−1, and x represents the sample to be classified into the highest possible category [5].

**3.4.2. Support Vector Machine**

Support Vector Machine (SVM) is a binary linear classifier. As a non-probabilistic supervised learning algorithm, it utilizes training data and employs a high-dimensional space to construct a set of hyperplanes for data classification. While only the features of test data are provided, the model is trained on the training data to predict the target values. For effective classification of problem instances, SVM relies on selecting the optimal hyperplane [5].

**3.5 Model Evaluation Metrics:**

Evaluation measures are metrics used to assess the results of an experiment [20]. In the context of classification models, different evaluation metrics are used to measure their output. In this study, the main performance evaluation metric is “Accuracy”. However, additional metrics such as recall, precision, f-measure, and confusion matrices are also used to supplement the evaluation of the model's performance. Each model identifies learner engagement levels when assessed using these metrics. A brief description of these metrics is provided below.

**Accuracy (AC)** is a common evaluation metric for classification models. It's calculated as the ratio of well-predicted samples to the total sample of prediction. For a balanced dataset, accuracy is a reliable measure of the model's performance [21].

In this equation, a true number that is positive is denoted by TP while a true number that is negative is denoted by TN, however, FN denotes a false number that is negative and FP denotes a false positive number [5].

**Precision (PRE)** is computed by dividing all the true positive samples by the sum of the predicted positive samples and predicted negative samples [5].

A high precision score indicates strong class predictions, while a low precision score reflects weak class predictions [22].

**Recall (RE)** is the ratio of correctly predicted positive results to all actual positive samples, also known as the detection rate. It's calculated by dividing the true positive samples by the sum of the positive samples.

**F1-score (FS)** is the mean value for recall and precision. It offers an indicator of mistakenly graded results [21]. It is regarded as the best metric for measuring the performance of models on an imbalanced dataset. It ranges from 0 to 1, with higher values indicating better model performance [5].

1. **RESULTS AND DISCUSSION**

This section provides the outcome of the analysis, achieved through running the models using 10-fold cross-validation, with accuracy as the primary metric for evaluating the performance of the models. Additionally, the performance of the models is further assessed using f1-score, recall, precision, and confusion matrices.

**4.1. Model Performance**

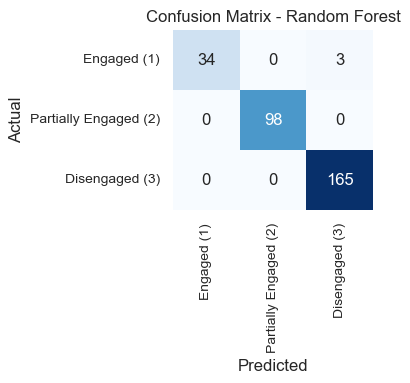
Table 1 summarizes the results of the three models after evaluating their performance using precision, accuracy, f1-score, and recall, while Figure 1 provides a visualization of the results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
| Random Forest (RF) | 97.29 | 97.09 | 97.29 | 97.09 |
| Logistic Regression (LR) | 69.43 | 68.53 | 69.43 | 70.61 |
| SVM | 91 | 90.38 | 91 | 91.56 |

**Table 1.** Performance Metrics

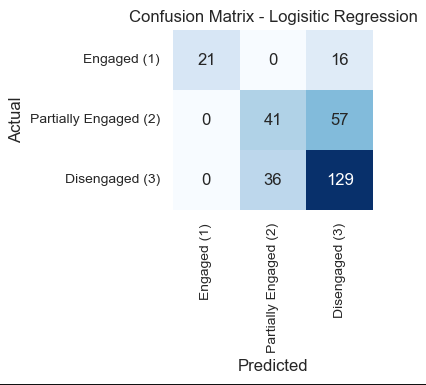
**Figure 1.** Graphical Performance Metrics

From Table 1 it can be observed that the performance of all the models accuracy ranges from 69% to 97%. The result shows that all the models can predict learner engagement levels using biometrics and environmental data. However, a thorough analysis of the results shows that RF provided the highest accuracy of 97.29%. With a precision and F1-score of 97.09%, and a recall of 97.29%, same as its accuracy. This suggests that RF handles the dataset very well, which shows that the dataset have non-linear relationships and mixed features. SVM also performs very well with accuracy close to RF. Using kernel functions like the Radial Basis Function (RBF) shows that SVM is a powerful model for non-linear classification. While LR struggles with the dataset because it assumes a linear relationship between the input features and the target.



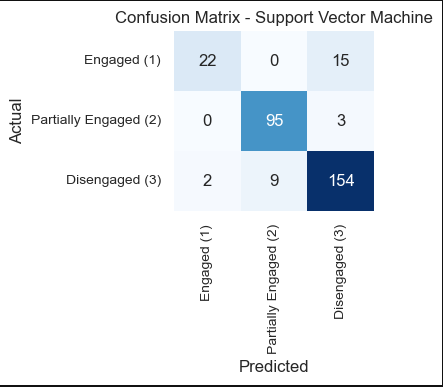
**Figure 2.** Confusion Matrix – Random Forest (RF)

In Figure 2, RF performed very well in classifying the Partially Engaged (2) and Disengaged (3) classes, however, it misclassified a small number of Engaged (1) samples. RF strong mechanism allowed it to effectively know the patterns of the data, leading to the high performance of the model.



**Figure 3.** Confusion Matrix – Logistic Regression (LR)

In Figure 3, LR struggles to classify the Engagement Level classes accurately, particularly for Partially Engaged (2) and Disengaged (3) classes. Engaged (1) is partially correct but still has significant misclassifications. The model frequently predicts Disengaged (3) even for samples that are Engaged (1) or Partially Engaged (2). The confusion between classes indicates that a linear model like Logistic Regression cannot capture the complex relationships in the dataset.



**Figure 4.** Confusion Matrix – Support Vector Machine (SVM)

The SVM model performs significantly better compared to Logistic Regression but slightly worse than Random Forest. Overall, the model shows strong predictive performance across all three classes.

**4.2. Comparative Discussion**

The researchers in this study have evaluated the performances of the three models namely, Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) in classifying student engagement based on their biometric and environmental data. The models demonstrate varied strengths and weaknesses, highlighting their effectiveness in classification tasks.

RF has been shown to outperform the two other models, with consistent accuracy, precision, recall, and f1-score of 97%. The model’s ability to handle non-linear relationships and complex interactions between features made it an effective model for this study. Also, the confusion matrix shows perfect accuracy in predicting partially engaged (2) and disengaged (3) classes.

LR performs poorly compared to RF and SVM due to its inability to capture complex, non-linear relationships in the data, leading to higher misclassification rates, especially for overlapping classes like Partially Engaged (1) and Disengaged (3) classes.

The SVM model performs significantly better than LR and is slightly behind RF. Its ability to effectively separate the classes (Engaged, Partially Engaged, and Disengaged) shows its strength in handling the complexity of the dataset. While SVM performs well, RF maintains a slight edge with higher precision and accuracy, making it the top-performing model overall.

In this study, the performances of RF, LR, and SVM were evaluated for classifying student engagement based on biometric and environmental data. And among the three RF is the most effective model for classifying student engagement levels, followed by SVM, while LR struggles with the complexity of the data.

1. **CONCLUSION**

Real-time student engagement prediction has been a significant challenge in many school institutions. This prediction classifies students based on their degree of involvement or engagement and can be either, engaged, partially engaged, or disengaged. Accurate real-time prediction of student engagement can overcome lower achievement and high dropout rates in schools. To overcome the limitations associated with traditional student engagement classifiers, this study uses biometric and environmental data together with three classifiers, which are Random Forest, Logistic Regression, and Support Vector Machine, to build a machine learning algorithm that can predict student engagement levels. After a thorough analysis of the performance of each model, RF provided the highest accuracy, precision, recall, and f1-score of 97%. The findings of this study suggest that RF is the most effective model for predicting student engagement, offering a promising tool for institutions aiming to identify and address engagement issues early and in real time, while SVM and Logistic Regression provide alternative solutions with varying levels of performance.

1. **RECOMMENDATION**

The researchers in the study suggest several recommendations to improve the scope and application of this study. Other machine learning models can be explored like neural networks and gradient boosting to see if they can outperform RF. Additionally, ongoing model evaluation using real-time data will ensure the model's effectiveness in different school settings. And while LR performed poorly in this study, it might still be useful in less complex scenarios. If simplicity and interpretability are required, LR could be employed with some improvements, such as feature engineering or combining it with more advanced models in an ensemble approach. Lastly, SVM performed better than LR and could serve as an alternative when RF is not feasible due to computational constraints or the need for simpler models. Further experimentation with kernel types and parameter tuning could enhance its performance for this specific task.

**REFERENCES:**

[1] NWEA, "Research proof points: Better student engagement improves student learning," Teach. Learn. Grow., 2015. Available: <https://www.nwea.org/blog/2015/research-proof-points-better-student-engagement-improves-student-learning/>.

[2] [1] T. D. Nguyen, M. Cannata, and J. Miller, "Understanding student behavioral engagement: Importance of student interaction with peers and teachers," 2016. Available: <https://files.eric.ed.gov/fulltext/ED578739.pdf>.

[3] [1] J. S. Renzulli, "The relationship between student engagement and student achievement," University of Connecticut, 2024. Available: <https://gifted.media.uconn.edu/wp-content/uploads/sites/961/2024/04/The-Relationship-Between-Student-Engagement-and-Student-Achievement.pdf>.

[4] A. P. Delfino, "Student engagement and academic performance of students of Partido State University," 2019. Available: <https://files.eric.ed.gov/fulltext/EJ1222588.pdf>.

[5] [1] G. Theophilus and C. I. Eke, "Machine learning-based e-learners' engagement level prediction using benchmark datasets," Int. J. Appl. Inf. Syst., vol. 12, no. 41, Sept. 2023.

[6] T. Anderson, "Applications of machine learning to student grade prediction in quantitative business courses," Int. J. Machine Learn., vol. 1, no. 3, pp. 13–22, 2017.

[7] T. Anderson, "Applications of machine learning to student grade prediction in quantitative business courses," Int. J. Appl. Inf. Syst., vol. 1, no. 3, pp. 13–22, 2017.

[8] S. Ayouni, F. Hajjej, M. Maddeh, and S. Al-Otaibi, "A new ML-based approach to enhance student engagement in online environment," PLOS ONE, Nov. 10, 2021. Available: <https://doi.org/10.1371/journal.pone.0258788>.

[9] N. Alruwais and M. Zakariah, "Student-engagement detection in classroom using machine learning algorithm," MDPI Electronics, vol. 12, no. 3, p. 731, 2023. Available: <https://www.mdpi.com/2079-9292/12/3/731>.

[10] G. Okereke, "A machine learning based framework for predicting student’s academic performance," Phys. Sci. Biophys. J., vol. 4, no. 2, 2020, doi: 10.23880/psbj16000145.

[11] M. A. Hernández-Mustieles, Y. E. Lima-Carmona, M. A. Pacheco-Ramírez, A. A. Mendoza-Armenta, J. E. Romero-Gómez, C. F. Cruz-Gómez, D. C. Rodríguez-Alvarado, A. Arceo, J. G. Cruz-Garza, M. A. Ramírez-Moreno, and J. de J. Lozoya-Santos, "Wearable biosensor technology in education: A systematic review," Sensors, vol. 24, no. 8, pp. 2437, 2024.

[12] N. Alruwais and M. Zakariah, "Student-engagement detection in classroom using machine learning algorithm," Electronics, vol. 12, no. 3, p. 731, 2023. Available: <https://www.mdpi.com/2079-9292/12/3/731>.

[13] J. McKee, "Experimental research: A quantitative research method," CJ Docs Research Glossary, 2024. Available: <https://docmckee.com/cj/docs-research-glossary/experimental-research-a-quantitative-research-method/>.

[14] Z. Ziya, "Emotional monitoring dataset," Kaggle, 2023. Available: <https://www.kaggle.com/datasets/ziya07/emotional-monitoring-dataset>.

[15] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, "A survey of user profiling: State-of-the-art, challenges, and solutions," IEEE Access, vol. 7, pp. 144907–144924, 2019. doi: 10.1109/ACCESS.2019.2944243.

[16] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. El-Salhi, "The effect of preprocessing techniques, applied to numeric features, on classification algorithms’ performance," Data, 2021, 6(2), 11. <https://doi.org/10.3390/data6020011>.

[17] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed.; Morgan Kaufmann: San Mateo, CA, USA, 2011.

[18] Edvancer, "Logistic Regression vs Decision Trees vs SVM (Part 2)," Edvancer, 2021. Available: <https://edvancer.in/logistic-regression-vs-decision-trees-vs-svm-part2/>.

[19] PureAI, "Machine learning techniques in AI applications," PureAI, Apr. 10, 2020. Available: <https://pureai.com/articles/2020/04/10/ml-techniques.aspx>.

[20] Javatpoint, "Performance metrics in machine learning," Javatpoint, 2021.. Available: <https://www.javatpoint.com/performance-metrics-in-machine-learning>.

[21] E. F. Buraimoh, "Predicting student success using student engagement in the online component of a blended-learning course," M.Sc. Research Report, School of Computer Science and Applied Mathematics, Faculty of Science, The University of the Witwatersrand, Johannesburg, South Africa, May 20, 2021. Available: <https://wiredspace.wits.ac.za/server/api/core/bitstreams/fb368595-69f9-4421-89d3-fc454199b9a0/content>.

[22] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," IOP Conf. Series: Materials Science and Engineering, vol. 928, no. 3, 032019, 2020. doi:10.1088/1757-899X/928/3/032019.

[23] C. I. Eke, A. A. Norman, and L. Shuib, "Multi-feature fusion framework for sarcasm identification on Twitter data: A machine learning-based approach," PLOS ONE, vol. 16, no. 6, p. e0252918, Jun. 10, 2021. <https://doi.org/10.1371/journal.pone.0252918>.

[24] H. Hassan, N. B. Ahmad, and S. Anuar, "Improved students’ performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining," J. Phys. Conf. Ser., vol. 1529, no. 5, p. 052041, 2020. Available: <https://iopscience.iop.org/article/10.1088/1742-6596/1529/5/052041/pdf>.

[25] M. B. Yıldız and C. Börekçi, "Predicting academic achievement with machine learning algorithms," J. Educ. Technol. Online Learn., vol. 3, no. 3, pp. 372–392, 2020. Available: <https://files.eric.ed.gov/fulltext/EJ1294138.pdf>.