**Classification of Emotional States with Random Forest and SVM 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** |
| ***Keywords:***  Classification Model  Support Vector Machine (SVM)  Random Forest  Accuracy  Precision  Recall  F1-Score |  | 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 physiological and environmental data to classify 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 two machine learning algorithms including; Random Forest and Support Vector Machine. Both models were accessed using the 10-fold cross-validation. The classification results show that Random Forest outperformed the other model 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%. |
|  |

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. One of the widely used applications of machine learning in classifying engagement levels is classification [6]. 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 classification, 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, and Section 5 concludes the research.

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 that also uses Random Forest (RF) and Support Vector Machine (SVM).

**2.1. Random Forest**

The researchers in [5] researched 'Engagement Level Prediction using Benchmark Datasets' and tested the predictive performance of 9 machine learning models. In this study, we focus on Random Forest (RF), one of the models tested in their study. The other models include Decision Tree (DT), Naive Bayes (NB), Stochastic Gradient Descent (SGD), LogitBoost (LB), Sequential Minimal Optimization (SMO), Voted Perceptron (VP), and Adaptive Boosting (AB). 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. Also, RF performs close to SMO with an accuracy of 89% and a precision, recall, and f-measure of 88%

**2.2. Support Vector Machine**

Researchers in [11] used three 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. In this study, we focus on the SVM, which achieved an accuracy of 75%. Among the models, ANN outperformed the others with an accuracy of 85%, followed by DT with 80%. Notably, the DT model demonstrated a 1% improvement in precision.

While researchers in [12] employed several models, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Logistic Regression (LR), Naïve Bayes Classifiers (NBCs), and Decision Trees (DT), additionally, these models are implemented using MATLAB scripts. The dataset underwent feature extraction to only include five key features and split into 80% for training and 20% for testing. To assess the classification capabilities of the trained models, metrics such as accuracy, precision, recall, and F-measure were used. ANN and SVM showed the best performance with the same accuracy, precision, recall, and F1 scores, which are 75%, 0.8, 0.91, and 0.85 respectively.

The similarity of these studies is that they don't use physiological 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 physiological 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 physiological and environmental data. The physiological 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 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 two techniques to clean the data, which included data normalization and fixing missing values.

Data normalization 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 for Random Forest (RF) and Z-Score Normalization for Support Vector Machine (SVM).

A 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 physiological and environmental data. To find the best classifier for engagement level prediction, two classifiers, including Random Forest (RF) 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].

(1)

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. 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].

(2)

In this equation, True Positives (TP) are correctly identified positives, True Negatives (TN) are correctly identified negatives, False Positives (FP) are negatives wrongly identified as positives, and False Negatives (FN) are positives wrongly identified as negatives [5].

**Precision (PRE)** measures the proportion of correctly predicted positive observations out of all the predicted positive observations [5]. A high precision score indicates strong class predictions, while a low precision score reflects weak class predictions [22].

(3)

In this equation, True Positives (TP) refer to the count of positive instances correctly predicted as positive, while False Positives (FP) refer to the count of negative instances incorrectly predicted as positive [5].

**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.

(4)

In this equation, True Positives (TP) are positives correctly predicted, while False Negatives (FN) are positives incorrectly predicted as negatives.

**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].

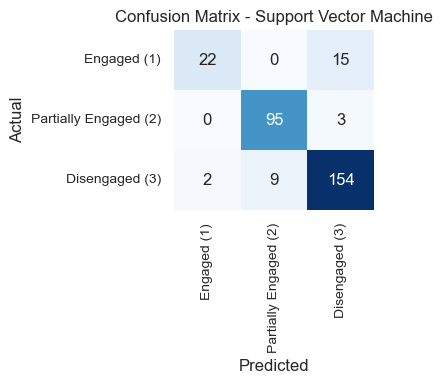
(5)

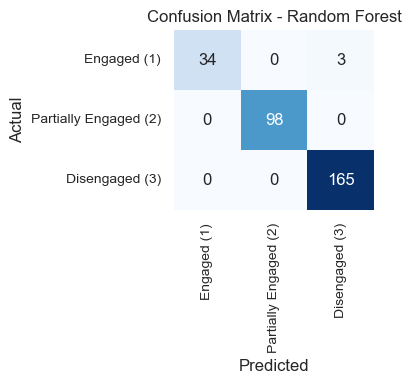
In this equation, Precision (PR) is the ratio of true positives to predicted positives, while Recall (RE) is the ratio of true positives to actual positives [5].

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**

****Figure 1 and 2 illustrates the performance of two machine learning models which are RF and SVM in classifying student engagement levels into three categories: Engaged (1), Partially Engaged (2), and Disengaged (3). The results of the two models after evaluating their performance using precision, accuracy, f1-score, and recall are in Table 1, while Figure 3 provides a visualization of the results.

****

**Figure 2.** SVM Confusion Matrix

**Figure 1.** Random Forest Confusion Matrix

In figure 1, the confusion matrix shows that RF performed very well, especially in classifying the Partially Engaged (2) and Disengaged (3) classes. Although 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 2 also show that SVM also performs slightly worse than Random Forest. It performed very well particularly in Partially Engaged (2) and Disengaged (3) with minor errors. There are still misclassifications between Engaged (1) and Disengaged (3), but overall, the model shows strong performance across all three classes compared to LR.

In conclusion, RF is the most effective model in classifying engagement levels, followed by SVM, while LR performs inadequately in this context.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Accuracy | Precision | Recall | F1-Score |
| Random Forest (RF) | 0.9729 | 0.9709 | 0.9729 | 0.9709 |
| SVM | 0.9100 | 0.9038 | 0.9100 | 0.9156 |

**Table 1**. Performance Metrics

From Table 1 it can be observed that the performance of all the models accuracy ranges from 91% to 97%. The result shows that both the models can classify learner engagement levels using physiological 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.

**Figure 3.** Graphical Performance Metrics

Figure 3 presents the performance metrics through a line chart which are: accuracy, precision, recall, and F1-score of RF and SVM.

RF achieves the highest values across all metrics, with performance around 97–98%. The minimal variance between metrics indicates RF's overall reliability and balanced performance.

SVM maintains strong performance across all metrics, with values close to 91%. While slightly lower than RF, SVM proves to be an effective classifier for this dataset.

The graph demonstrates that RF outperforms SVM in all performance metrics, making it the best model for classifying engagement levels. SVM, while slightly less effective than RF, still provides strong results and is a viable alternative. Overall, RF is the best candidate for classifying engagement levels and SVM is an alternative.

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 physiological and environmental data together with two classifiers, which are Random Forest and Support Vector Machine, to build a machine learning algorithm that can classify 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 classifying student engagement, offering a promising tool for institutions aiming to identify and address engagement issues early and in real time, while SVM provide alternative solutions with varying levels of performance.

**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] M. U. Akram, S. Z. M. Hashim, M. A. Hassan, "Paper’s title should be the fewest possible words that accurately describe the content of the paper," \*Indonesian Journal of Electrical Engineering and Computer Science\*, vol. 99, no. 1, pp. 1–1x, Month 2099. doi: 10.11591/ijeecs.v99.i1.pp1-1x.

[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] S. S. Soni, K. G. Karmakar, R. K. Gupta, and D. Ghosh, "Student-Engagement Detection in Classroom Using Machine Learning Algorithm," Electronics, vol. 12, no. 3, pp. 731, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/3/731. [Accessed: 12-Jan-2025].

[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] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," Artificial Intelligence Review, vol. 52, no. 2, pp. 987–1011, 2019, doi: 10.1007/s10462-018-9620-8.

[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>.