**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 |  | This study investigates the application of machine learning algorithms to classify student engagement levels into Highly Engaged, Moderately Engaged, and Disengaged. The dataset used is an emotional monitoring dataset designed to analyze student engagement using biosensor technology, with 1000 instances and 13 features. Pre-processing steps, including data normalization and removal of rows with missing values, were performed to optimize model performance. A comparative analysis was conducted between two machine learning algorithms: Random Forest (RF) and Support Vector Machine (SVM). Both models were evaluated using 10-fold cross-validation. The results demonstrated that the RF model outperformed the SVM, achieving an accuracy of 97%, along with precision, recall, and F1-score values of 0.970, 0.972, and 0.970, respectively. Meanwhile, the SVM model also showed strong performance, attaining an accuracy of 91% with precision, recall, and F1-score values of 0.903, 0.910, and 0.915, respectively. |
|  |

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. 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. Researchers asserted that classifying 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 [7][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 [9]. 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 classifying 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 classifying student engagement levels.

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**

Random Forest (RF) is a powerful and widely used machine learning model known for its strong predictive performance, especially in classification tasks. In a study on engagement level prediction using benchmark datasets, the researchers evaluated nine machine learning models, with Random Forest (RF) being one of the key models of interest. Other models tested included 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 used consisted of 486 instances and 12 features, with information gain employed to evaluate feature significance. Data preprocessing and algorithm application were performed using WEKA version 3.8.6. The study applied 10-fold cross-validation and assessed performance using precision, accuracy, recall, and F1-score. Notably, RF demonstrated strong results, achieving an accuracy of 89%, with precision, recall, and F1-scores all around 88%, positioning it closely behind SMO, which outperformed other models with an accuracy of 90%, along with precision, recall, and F1-scores of 0.897 [5].

**2.2. Support Vector Machine**

Support Vector Machine (SVM) has been widely recognized for its ability to efficiently handle high-dimensional data and classify non-linear patterns, making it a popular choice in various domains. In a recent study, researchers evaluated the performance of three models Support Vector Machine (SVM), Decision Tree (DT), and Artificial Neural Network (ANN) on a classification task using a dataset containing 348 instances and 9 features. The study utilized 10-fold cross-validation for training and assessed performance based on precision, accuracy, recall, and F1-score. Among the models, the SVM achieved a notable accuracy of 75%. Although the ANN model attained the highest accuracy at 85% and the DT model followed at 80%, the SVM remains a competitive option due to its robustness in handling complex patterns [10].

In another study, researchers implemented multiple models including SVM, ANN, Logistic Regression (LR), Naïve Bayes Classifiers (NBC), and Decision Trees (DT) using MATLAB scripts. Feature extraction was performed to include only five key features, and the dataset was split into 80% for training and 20% for testing. The models were evaluated using metrics such as accuracy, precision, recall, and F-measure. The SVM and ANN models showed the best performance, both achieving an accuracy of 75%, with SVM also demonstrating strong precision (0.8), recall (0.91), and F1 scores (0.85). This demonstrates SVM's competitive performance in comparison to other models in the classification task [11].

1. **METHODOLOGY**

**3.1. Materials**

**3.1.1. Dataset**

This study utilizes the Emotional Monitoring Dataset from Kaggle, a data with 1,000 instances designed for analyzing student engagement using biosensor technology. The dataset captures both physiological and environmental factors to assess emotional and cognitive states, focusing on key indicators like stress, engagement, and external influences. It categorizes engagement levels into Highly Engaged, Moderately Engaged, and Disengaged, providing a comprehensive basis for developing predictive models [12].

**3.1.2. Hardware**

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).

**3.1.3. Software**

This study 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. Methods**

**3.2.1 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 [13]. 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 [14]. 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].

**3.2.2 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.

**3.2.2.1 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 [15]. 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 [16]. 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. During the tree-building process, the Gini Index is often used as a splitting criterion to measure impurity at each node. This ensures that the splits are optimized to create subsets that are as homogeneous as possible, improving classification accuracy and overall model reliability [17].

The formula for the Gini Index Criterion for Decision Trees:

(1)

Here, is the Gini Index for dataset D. It measures the impurity or disorder of the dataset, where a lower Gini Index indicates higher purity (fewer misclassifications). denotes the number of classes and denotes the proportion of instances in dataset that belong to class [5].

(2)

is calculated as:

**3.2.2.2 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].

The formula for the Support Vector Machine (SVM) decision boundary can be expressed as:

(3)

Where is the weight vector, determining the orientation of the hyperplane,  is the feature vector of the input data,  is the bias term, shifting the hyperplane.

**3.2.3 Model Evaluation Metrics:**

Evaluation measures are metrics used to assess the results of an experiment [18]. 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 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.

(4)

Where the numerator reflects the total number of correct predictions, while the denominator represents the total number of predictions made by the model. A higher accuracy value suggests that the model is more effective at correctly classifying both classes [19].

Precision measures the proportion of correctly predicted positive observations out of all the predicted positive observations. A high precision score indicates strong class predictions, while a low precision score reflects weak class predictions.

(5)

Where True Positives refer to the instances that were correctly identified as positive, while False Positives are the instances where the model wrongly predicted the positive class [20].

Recall 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.

(6)

Where False Negatives occur when the model mistakenly classifies a positive instance as belonging to the negative class [5].

F1-score is the mean value for recall and precision. It offers an indicator of mistakenly graded results [19]. 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.

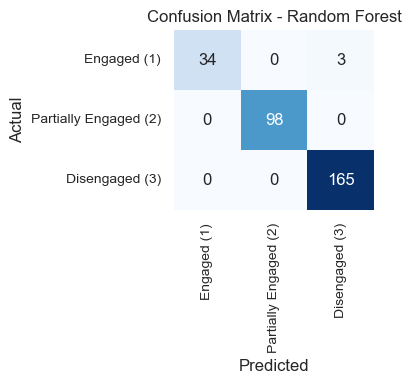
(7)

Where Precision evaluates the accuracy of positive predictions and Recall assesses the model's ability to identify all relevant instances, the F1-Score combines these two metrics into a single value, offering a balanced measure of the model's accuracy and completeness [5].

1. **RESULTS AND DISCUSSION**

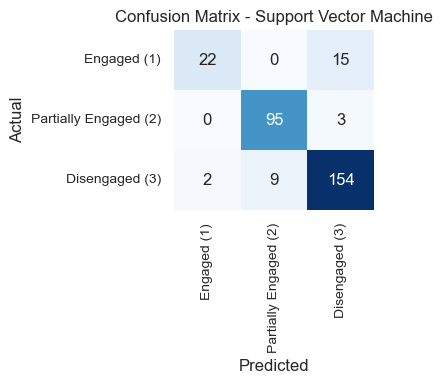
This section presents the performance of the Random Forest (RF) and Support Vector Machine (SVM) models in classifying student engagement levels into three categories: Highly Engaged, Moderately Engaged, and Disengaged. The analysis was conducted using 10-fold cross-validation, with metrics such as accuracy, precision, recall, and F1-score used to evaluate the models.

**4.1. Model Performance**

****

**Figure 1.** RF Confusion Matrix

For the RF model, the confusion matrix shows exceptional performance, particularly in classifying "Moderately Engaged" and "Disengaged" students. It effectively identifies patterns in the data, resulting in minimal misclassifications across all categories. However, there are minor errors in distinguishing between "Highly Engaged" and "Disengaged", which may stem from overlapping features between these categories.

****

**Figure 2.** SVM Confusion Matrix

For the SVM model, the confusion matrix reveals slightly lower performance compared to RF, with some notable misclassifications between "Highly Engaged" and "Disengaged". Despite these errors, SVM shows strong classification accuracy for "Moderately Engaged", demonstrating its effectiveness in handling non-linear relationships in the data. However, the presence of overlapping data points or insufficient differentiation in feature space may have contributed to these misclassifications.

**Table 1**. Performance Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

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.

1. **CONCLUSION**

In conclusion, the RF model demonstrated exceptional performance in classifying student engagement levels, achieving the highest accuracy of 97.29%, along with strong precision, recall, and F1-scores, highlighting its ability to effectively learn and generalize from the dataset. The RF model consistently classified "Moderately Engaged" and "Disengaged" students with high precision, showcasing its robustness in handling complex patterns in the data. However, minor misclassifications were observed, particularly between "Highly Engaged" and "Disengaged" students, which may indicate overlapping features or insufficient differentiation in certain instances.

The SVM model also showed strong performance, achieving an accuracy of 91%. While slightly less effective than RF, it demonstrated reliable classification for "Moderately Engaged" students but faced challenges distinguishing between "Highly Engaged" and "Disengaged" categories, resulting in some false negatives and positives. This highlights potential areas for optimization, such as enhanced feature engineering or hyperparameter tuning. Overall, the RF model’s balanced precision, recall, and F1-scores make it the more effective classifier, while the SVM model remains a viable alternative with room for further improvement. Both models underscore the potential of machine learning techniques in accurately classifying engagement levels, providing valuable insights for improving student outcomes.

**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] 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>.

[8] 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].

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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>.

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

[16] 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>.

[17] 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>.

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

[19] 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>.

[20] 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.