ISSN 2723-6471

Novel Predictive Framework for Student Learning Styles Based on Felder-Silverman and Machine Learning Model

Wiga Maulana Baihaqi[1,*], Rujianto Eko Saputro[2], Fandy Setyo Utomo[3], Sarmini[4]

[1]*Information Technology Department, Universitas Amikom Purwokerto, Purwokerto 53127, Indonesia*

[2,3]*Computer Science Department, Universitas Amikom Purwokerto, Purwokerto 53127, Indonesia*

[4]*Information System Department, Universitas Amikom Purwokerto, Purwokerto 53127, Indonesia*

(Received: August 16, 2024; Revised: September 07, 2024; Accepted: September 16, 2024; Available online: October 15, 2024)

Abstract

This study analyzes data from the Open University Learning Analytics Dataset to evaluate how students' interactions with Virtual Learning Environment (VLE) materials influence their final outcomes. This research aims to formulate and build a novel predictive framework based on the Felder-Silverman and Machine Learning Model for student learning styles. Based on these objectives, this research provides novelty and contributions since it enhances student data analysis, uses a learning model using Felder-Silverman Learning Style Model (FSLSM) to give a more comprehensive understanding of students' learning styles, and improves prediction accuracy by introducing Artificial Neural Network (ANN) and feature selection using Random Forest. The data used includes 3 main files: vle.csv, which contains information about the materials and activities in the VLE; studentVle.csv, which records students' interactions with the materials; and studentInfo.csv, which provides demographic information of students and their final outcomes. The analysis process involved data merging and processing, including handling of missing values, data type conversion, as well as mapping activity types to learning style features based on the FSLSM. We use the Random Forest feature selection method, as well as data imbalance handling techniques such as oversampling, to improve model performance. The applied classification models include Logistic Regression, K-Nearest Neighbor, Random Forest, Support Vector Machine (SVM), and ANN. The analysis results showed that after tuning, the Random Forest model achieved 97% accuracy, while SVM achieved 97% accuracy as well, with better performance than previous studies. This research highlights the importance of comprehensive data integration and appropriate processing techniques in improving the accuracy of student learning style prediction. Based on the increase in accuracy results, it can be beneficial for more effective personalized learning and improve our understanding of students' learning style preferences. The research advances knowledge and provides practical applications for educators to tailor their teaching strategies.

Keywords: Open University Learning Analytics Dataset (OULAD), Virtual Learning Environment (VLE), Felder-Silverman Learning Style Model (FSLSM), Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Data Analysis, Feature Selection, Data Imbalance Handling

1. Introduction

Learning Analytics and Educational Data Mining (EDM) are increasingly important fields in the realm of education, enabling the collection, analysis, and reporting of data about learners and their learning contexts [1]. Data analysis in an educational context plays a crucial role in understanding and optimizing the learning process and the environments in which it occurs [1]. Machine learning, as a branch of artificial intelligence, provides valuable insights into student learning processes through the utilization of collected data [2]. In the realm of educational institutions, machine learning can play a role in various aspects, ranging from learning content, teaching processes, assignment distribution, assessment processes, to monitoring student learning progress [2].

Studies have highlighted the importance of recording and analyzing data during the learning process to identify factors influencing student learning performance [3]. Through the application of machine learning techniques and data analysis, educators can gain profound insights into the dynamics of classroom learning [3]. This enables educators to design more targeted interventions and personalized learning strategies, ultimately enhancing the effectiveness of

***Corresponding author: Wiga Maulana Baihaqi (wiga@amikompurwokerto.ac.id)**
DOI: https://doi.org/10.47738/jads.v5i4.408

learning [4]. The application of machine learning in an educational context can also aid in predicting student academic performance [5]. By leveraging big data generated from Virtual Learning Environments (VLE), deep learning models can predict student performance categories, assisting decision-makers in developing appropriate pedagogical policies [5]. Additionally, machine learning can predict risk factors such as depression and anxiety in students, enabling early identification and timely interventions [6].

The implementation of Learning Analytics can also enhance the interaction between computer-based evaluation data and classroom instruction [7]. By focusing on providing feedback to low-performing students, either by teachers or high-performing peers, this approach can improve students' understanding of learning materials [7]. Furthermore, data analysis can be used to evaluate the success of project-based learning, revealing insights through supervised machine learning assessment [8].

In addition, in the context of EDM, data quality is also a crucial factor that needs to be considered [10]. Research by Mukherjee [11] highlights the importance of improving data quality for EDM, especially in the context of education startups in India. High-quality data is a necessary foundation to ensure the accuracy and reliability of the prediction models built. Therefore, efforts to clean, integrate, and ensure the quality of education data is a crucial first step in overcoming the challenges in using data for student study type prediction.

This research makes a significant contribution to the field of Learning Analytics and EDM by introducing a more comprehensive approach in analyzing students' learning preferences in VLE. Compared to the previous study by Ahmed Rashad Sayed, Mohamed Helmy Khafagy, Mostafa Ali, and Marwa Hussien Mohamed [12], which used only two tables of the OULAD dataset to analyze student interactions, this study expands the scope of analysis by incorporating a third table, student_info. This approach enables a more in-depth understanding of student profiles and their interaction patterns with the VLE platform, as well as how additional information from student_info affects the analysis results.

In addition, this study utilizes the Felder-Silverman Learning Style Model (FSLSM) which is more detailed than the VAK model used in previous studies. This research not only relies on the four basic algorithms-KNN, SVM, Logistic Regression, and Random Forest-but also introduces Artificial Neural Networks (ANN) to improve the accuracy of student performance prediction. By performing feature selection using Random Forest, handling data imbalance through oversampling techniques, and performing hyperparameter tuning, this study shows significant improvements in the performance of machine learning models in classifying student learning preferences. We used Random Forest for feature selection in this study since this technique significantly impacts prediction and classification results [13], [14], [15]. Furthermore, the KNN, SVM, Logistic Regression, and Random Forest algorithms were used in the study since all of these algorithms were used in previous studies.

This research is organized systematically to provide a thorough understanding of the topic at hand. After this introduction, the next chapter is the literature review, which will discuss in depth previous research relevant to the field of Learning Analytics and EDM, and provide a theoretical basis for this research. The research methods chapter will then describe in detail the approach used in this research, including data collection, feature selection, data imbalance handling, and the models and algorithms used for analysis.

Next, the results chapter will present the main findings of this research, including the performance of the hyperparameterized machine learning model and the results of the comprehensive approach used. The discussion chapter will analyze and discuss the research results in the context of the existing literature review, and evaluate the contributions and implications of the findings to the field under study. Finally, the conclusion chapter will summarize the research results, highlight the main contributions, and provide recommendations for future research. The structure of this study is designed to provide a logical and structured flow so that readers can easily follow and understand each stage of the study.

2. Literature Review

Learning analytics, a field that involves analyzing educational data to understand learning behaviors and optimize educational systems, has gained significant attention in recent years [16]. By delving into vast amounts of student data, including academic performance and interaction patterns with educational resources, machine learning algorithms can provide valuable insights into students' learning processes [17]. EDM techniques further enhance this understanding by analyzing data collected from educational environments to improve educational outcomes. FSLSM developed by Richard M. Felder and Linda K. Silverman in 1988, plays a crucial role in understanding students' learning styles and behaviors [18].

The FSLSM, with its dimensions like input (visual/verbal) and perception (sensory/intuitive), offers a framework to assess how students receive information and perceive it, influencing their learning experiences [19]. This model has been widely used in various studies to personalize education and predict student performance [20]. By incorporating the FSLSM into the design of learning materials and classroom interactions, researchers have observed improvements in student participation rates, test scores, and teacher intervention frequencies [21]. Additionally, the FSLSM has been instrumental in developing recommendation models for learning materials and selecting learning objects based on students' learning styles [22], [23].

Machine learning techniques, when applied to educational data, enable the creation of predictive models for student performance [24]. These models not only predict academic outcomes but also help in identifying factors that influence students' learning achievements in Massive Open Online Courses (MOOCs) [25]. Moreover, the FSLSM has been utilized to personalize learning in virtual learning environments by analyzing students' behaviors and mapping them to learning style features [26]. Such personalized approaches enhance the effectiveness of online learning by tailoring educational experiences to individual students' needs.

In conclusion, this literature review highlights that Learning Analytics and Educational Data Mining play crucial roles in analyzing educational data to understand student learning behaviors and improve educational systems. The application of machine learning techniques and the FSLSM enables more accurate personalization of education and prediction of student performance, while approaches for handling imbalanced data, such as SMOTE and other combinatorial techniques, enhance classification performance. Integrating these methods into education supports the development of more effective and adaptive interventions tailored to individual student needs, as well as addressing the challenges of imbalanced data in classification tasks.

3. Methodology

Figure 1 illustrates the complex and structured methodological flow of this research:

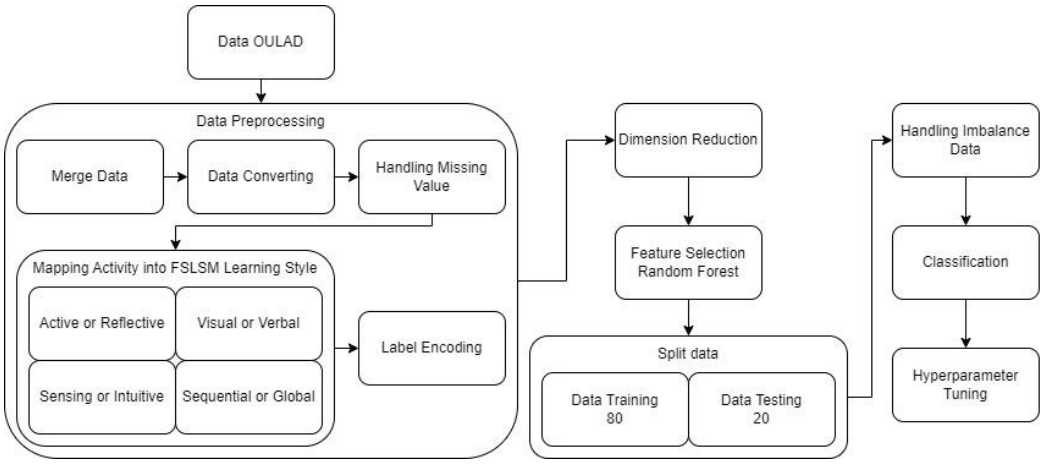

Figure 1. Research Steps

Figure 1 presents the complete flow of the research process, starting from data processing to classification. Initially, the data is taken from the Open University Learning Analytics (OULAD) dataset which is then processed through

ISSN 2723-6471

several stages [27]. In the preprocessing stage, data from various sources were merged, the data format was changed to facilitate analysis, and missing data was handled to produce a completer and more accurate dataset.

Next, students' learning activities were classified according to the FSLSM, which includes active or reflective, sensing or intuitive, visual or verbal, and sequential or global aspects. Each of these categories is then numerically coded through a label encoding process to facilitate algorithm processing.

Subsequent processes include dimensionality reduction to reduce model complexity without losing important information and feature selection using the Random Forest method to identify the most relevant features. The data was then divided into training data (80%) and test data (20%) to validate the model. Handling of unbalanced data is done to ensure that the model can generalize well across all classes.

The final stage is classification using the predefined model, followed by hyperparameter tuning to improve prediction accuracy. Each stage is designed to ensure the quality of the data used and the reliability of the classification results, which is critical for learning analytics applications. The process detects students learning styles based on their behavior and interactions with the VLE, which will be described in detail.

3.1. Data Collection

The data used in this study is data from OULAD. This data includes various information about student learning activities, including demographic data, data on student interactions with the learning system, and student academic results. The data set contains data on 7 courses, 32.593 students, test scores, and VLE interactions (10.655.280 entries per day). It also has the results of their test quizzes and homework. For this research, a subset of 4,327,256 data entries was used.

Student demographics include age, gender, and educational level, along with course enrollment details such as course codes, titles, and start/end dates. Learning activities track how students engage with materials, including accessing course information, submitting assignments, participating in discussions, and using multimedia tools. Assessment data provides insights into student performance on quizzes, exams, and assignments. Clickstream data captures navigation patterns, time spent on different pages, and interactions with course tools in the online learning environment. Style data, recorded in student profiles or as independent records, comprises categorical variables reflecting preferred learning styles. Analyzing clickstream data can reveal insights into students' learning behaviors and preferences, indicating their preferred approaches. For instance, visual learners may engage more with multimedia presentations, while auditory learners might spend more time on audio lectures and discussions.

3.2. Data Preprocessing

Data pre-processing ensures the dataset is clean and suitable for analysis by merging three tables such as VLE, StudentVLE, and StudentInfo. This involves an inner join on matching columns to create a comprehensive dataset of student interactions with learning modules. The process includes converting data formats, handling missing values by replacing them with mode values, and mapping student activity data to the FSLSM learning style model, which identifies learning preferences. As shown in table 1, this mapping aligns various activity types with specific learning style features based on the FSLSM framework, facilitating a deeper understanding of how different activities correspond to individual learning preferences. Finally, label encoding transforms categorical data into numerical format for machine learning algorithms, enabling analysis of learning patterns and the development of personalized learning recommendations.

Table 1. Mapping activity types to learning style features based on FSLSM

Dimension	FSLSM Classifications	VLE Activity Type
Processing	Active/Reflective	Forumng, oucollaborate, ouwiki, glossary, htmlsctivity
Perception	Sensitive/Intuitive	oucontent, questionnaire, quiz, externalquez
Input	Visual/Verbal	dataPlus, dualPane, folder, page, homepage, resource, url, ouelluminate, subpage
Understanding	Sequential/Global	Repeatactivity, sharedsubpage

3.3. Dimension Reduction

Dimension reduction is the process of reducing the number of features or variables in a dataset, which aims to remove irrelevant or redundant information. By reducing the dimensionality of the data, the model becomes simpler and more capable of generalization, thus reducing the risk of overfitting. Overfitting is a condition where the training data contains a lot of irrelevant information, referred to as meaningless data.

3.4. Feature Selection

Feature selection serves to select the most relevant and significant features. It involves selecting a subset of the original features that provide the most useful information to the model, with the aim of improving model performance and reducing overfitting. By eliminating irrelevant or redundant features, the model becomes simpler, faster to train, and easier to interpret. In addition, feature selection can also help reduce computational requirements and improve model generalization to new data. The feature selection method used is Random Forest.

3.5. Data Splitting

In this step, the data is divided into two main parts: training data and testing data. Training data, which accounts for 80% of the data, is used to train the model to recognize patterns and relationships in the data. Meanwhile, the testing data, which accounts for 20% of the data, is used to test the performance of the model and evaluate the extent to which the model is able to provide accurate predictions on data that has never been seen before. With this division of data, the developed model is not only effective in learning the training data, but also has good generalization ability on new data.

3.6. Handling Data Imbalance

In this study, the dataset used has significant class imbalance, where some classes have a much smaller number of samples compared to other classes. This imbalance can result in a bias in the model, which tends to give more attention to the majority class and ignore the minority class. To address this issue, we applied several resampling techniques, namely ADASYN, RandomUnderSampler, and SMOTEENN, which are explained as follows:

3.6.1. ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning)

ADASYN is an adaptive oversampling technique that generates synthetic samples for minority classes, focusing on areas where the model struggles to learn, unlike traditional methods like SMOTE. Its main principle is to balance class distribution while preserving sample variability by creating more synthetic data for minority samples with fewer nearest neighbors compared to the majority class. This approach prioritizes areas needing additional data to enhance the model's ability to classify minority classes [28]. In this study, ADASYN is applied as the initial step to balance the class distribution by increasing the minority class samples until they are closer to the majority class.

3.6.2. RandomUnderSampler

After oversampling is done, the next step is to use RandomUnderSampler to reduce the number of samples from the majority class. This technique works by randomly reducing the number of samples in the majority class, so that the class distribution becomes more balanced [29]. RandomUnderSampler is used because even though ADASYN has increased the sample size of the minority class, the majority class still has a larger sample size. By reducing the size of the majority class, the model can be trained with a more balanced dataset without letting the majority class dominate the training process. The reduction is done randomly, thus preventing bias towards certain features from the majority class samples.

3.6.3. SMOTEENN (Synthetic Minority Over-sampling Technique Edited Nearest Neighbors)

As the final step in the data balancing process, we employ SMOTEENN, which combines SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors). SMOTE generates synthetic samples across the distribution of minority classes, while ENN cleans the dataset by removing noise, such as misclassified samples [30]. This approach not only balances the dataset quantitatively but also enhances its quality by eliminating outliers. By using SMOTEENN, we systematically add synthetic samples and remove noise, ensuring the model is trained on clean, balanced, and reliable data.

 ISSN 2723-6471

3.7. Classification

Classification groups data into predefined categories based on predefined features. The process begins with the selection and extraction of relevant features from the pre-processed data. The machine learning model is then trained using labeled data, where each data sample has a known label or category. Once the model is trained, it can be used to predict the label of new unlabeled data [31]. Classification algorithms used include K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Random Forest, and Neural Networks.

3.8. Hyperparameter Tuning

Hyperparameter tuning is the process of optimizing model performance. It involves adjusting model parameters that are not estimated from data, but rather predetermined. Two techniques used in hyperparameter tuning are Grid Search and Random Search. Grid Search is a method that tries every combination of hyperparameters from a predefined set of values, while Random Search selects hyperparameter values randomly from a defined distribution. Both techniques have their own advantages and disadvantages; Grid Search can find optimal combinations thoroughly but requires longer computation time, while Random Search is more efficient and can find good combinations faster in a large hyperparameter space. By choosing these hyperparameters, the model can achieve better performance and more generalization to new data.

4. Results and Discussion

In this section, we will show the content of the dataset and the preprocessing steps performed by the model used to prepare the dataset through the mapping process, and then discuss the experimental results.

4.1. Data Collection

This study used the OULAD dataset, but not all dataset files were used in this analysis. Of the various files available in the OULAD dataset, only three were selected, which are vle.csv, studentVle.csv, and studentInfo.csv. The vle.csv file contains information about the materials available on the VLE, including HTML pages and PDF files, and records student interactions with the materials. The studentVle.csv file records each student's interaction with the materials on the VLE, including the number of daily clicks on each material. Meanwhile, the file studentInfo.csv contains demographic information of students and their outcomes, such as gender, region of residence, highest level of education, disability status and module outcomes. By focusing on these three files, this research can evaluate how students' interaction with the materials on the VLE affects their final outcomes, without the need to consider data from other files in the OULAD dataset.

4.2. Data Preprocessing

At this stage, a series of data pre-processing steps have been performed to ensure the dataset is ready to be used in further analysis. The first step was the merging of data from vle, student_vle, and student_info to create a unified dataset. This process involved data cleaning to remove redundancies and inconsistencies through an inner join stage. Next, data conversion is performed to ensure that all variables are in the appropriate format, in this case some data with object data types are converted to categorical data. Another important step is the handling of missing values. The technique used is to replace null values with modes. By completing these steps, the dataset becomes cleaner and ready to be used in further stages of analysis, thus ensuring the results obtained from the analysis will be more accurate and reliable.

Data after the merging process containing information related to student activities within a learning module. The columns include the module code, presentation, and unique student identification, followed by the number of clicks made by students on various activity types such as forums or homepages. Additional information includes gender, region, highest education, and age group. The table also records the number of previous attempts made by students in the module, the number of credits studied, whether the student has a disability, and the final result for the module. This data is used to analyze student engagement and performance within the learning module.

Before conversion step, columns such as `code_module`, `gender`, and `region` were of type `Object`, while columns like `id_student`, `sum_click`, and `num_of_prev_attempts` were of type `Int64`. After conversion, the `Object` columns were changed to `Category`, allowing for more efficient data processing, particularly in machine learning

 ISSN 2723-6471

analysis. Numeric columns like `id_student` and `sum_click` remained as `Int64`, indicating that the numeric data types were not altered. This process aims to optimally prepare the data for analysis and modeling, including handling missing values by replacing them with the mode.

Before handling, the "Null" column indicates that most columns have no missing values, except for the imd_band column which has 2,354,477 missing values. After handling, all columns, including imd_band, have a value of "0" in the "Null" column, indicating that all missing values have been replaced with modes. This step is important in data preprocessing to ensure a complete dataset ready for use in machine learning analysis or models, as well as maintaining consistency and accuracy without losing important information.

Table 2 shows process of mapping student's activity types in various learning activities involved analyzing their activities in VLE using the FSLSM. FSLSM classifies students' learning styles in several dimensions, namely Processing (Active/Reflective), Perception (Sensitive/Intuitive), Input (Visual/Verbal), and Understanding (Sequential/Global). The data collected includes various activities such as forum, homepage, oucontent, and oucollaborate. Each type of activity was categorized based on the relevant FSLSM dimensions.

Table 2. Mapping activity types to learning style features based on FSLSM

code_ module	code_ presentation	id_ student	id_ site	sum_ click	activity_ type	gender	region
AAA	2013J	28400	546652	4	forumng	F	Scotland
AAA	2013J	28400	546652	1	forumng	F	Scotland
AAA	2013J	28400	546614	11	homepage	F	Scotland
AAA	2013J	28400	546714	1	oucontent	F	Scotland
AAA	2013J	28400	546652	8	forumng	F	Scotland

highest_ education	imd_ band	age_ band	num_of_ prev_attempts	studied_ credits	disability	final_ result
HE Qualification	20-30%	35-55	0	60	N	Pass
HE Qualification	20-30%	35-55	0	60	N	Pass
HE Qualification	20-30%	35-55	0	60	N	Pass
HE Qualification	20-30%	35-55	0	60	N	Pass
HE Qualification	20-30%	35-55	0	60	N	Pass

The Label Encoding process is a crucial step in data processing before it is used in machine learning models. Label Encoding converts categorical data, which is originally text or symbols, into numbers so that it can be processed by machine learning algorithms that can generally only process numerical data. In the table, several categorical features such as module code, presentation code, gender, region, highest education, imd band, age band, num_of_prev_attempts, disability, final result, and study type have been converted into numerical values.

4.3. Dimension Reduction

The data to be used is a combination of learning styles with activity types and Student_VLE, which contains the number of clicks from each activity and id_site with a total input data of 4,327,256. In addition, to avoid overfitting because there are irrelevant data in the dataset, only data from the maximum value of the total clicks that have been made based on id_site is taken, resulting in data with a total of 6,268. We will get all the data features we need to start working on Feature Selection step.

4.4. Feature Selection

The use of Random Forest for feature selection is very helpful in identifying the features that are most relevant to the target variable study_type, thus facilitating the subsequent classification process. From the selection results, it can be seen that feature such as activity_type, sum_click, id_site, code_module, id_student, code_presentation, region, imd_band, studied_credits, and final_result have a significant influence on study_type, while features such as gender, highest_education, age_band, num_of_prev_attempts, and disability are less relevant. After selecting the features that influence the study_type, the data is divided into an 80% training set and a 20% test set.

4.5. Handling Imbalance Data

To address the issue of imbalanced data, where Class 3 has significantly more samples than other classes, we applied a data balancing technique. This involved using a combination of ADASYN, RandomUnderSampler, and

SMOTEENN. ADASYN adaptively adds samples from minority classes, RandomUnderSampler reduces samples from the majority class, and SMOTEENN cleans the final dataset by removing outliers and noise. This approach results in a more balanced dataset, enhancing the model's ability to classify samples from the minority class effectively.

4.6. Classification

In this study, the support vector machine, K-nearest neighbour, Random Forest, Logistic regression, and Artificial Neural Network were used as classification methods. The predicted variable in this scenario is each student's preferred learning method, while the characteristic is the VLE activity clicks in the VLE cycle. Sklearn, NumPy, and Pandas were the packages used for data preprocessing. Experiments were conducted using 80% of the data for each identified student training dataset, with the remaining 20% reserved for testing. The three scoring metrics were used to assess the model performance. As illustrated in table 3, the results of the algorithm performance when training with 80% of the data and testing with 20% without any tuning demonstrate the effectiveness of each classification method in predicting students' preferred learning methods.

Table 3. The result of algorithm performance when training model 80 and test model 20 Without Tuning

Algorithm	Precision	Recall	F1-Score	Accuracy
Logistic Regression	94%	79%	85%	79%
K-Nearest Neighbor (KNN)	69%	74%	71%	74%
Random Forest	99%	99%	99%	99%
Support Vector Machine (SVM)	94%	97%	96%	91%
Artificial Neural Network (ANN)	94%	94%	94%	99%

4.7. Classification With Tuning

To enhance the performance of the Logistic Regression model, we implemented data normalization and hyperparameter tuning via GridSearch. This process yielded optimal hyperparameter combinations, resulting in significant improvements in model performance: Precision increased to 98%, Recall to 97%, F1-Score to 97%, and Accuracy to 92%.

For the KNN model, initial assessments indicated an imbalance in class distribution, leading to an accuracy of only 74% with 5 neighbors. To address this imbalance, we employed an ensemble approach by combining KNN with other techniques through a Stacking Model, which resulted in an improved accuracy of 91%.

In the case of Random Forest, initial calculations revealed issues with overfitting and class imbalance. By utilizing a pooling model, we allowed each tree in the forest to contribute to the final output, achieving an impressive accuracy of 99%. However, to mitigate overfitting, we adjusted the min_samples_leaf parameter to 10, resulting in a more generalized model, albeit with a decrease in accuracy to 97%. This adjustment is beneficial in scenarios where stability and generalization take precedence over maximum training performance.

For the SVM model, we employed GridSearch to identify the best hyperparameters, testing values for C (0.1, 1, 10), Gamma (1, 0.1), and various kernels (linear, polynomial, RBF, and sigmoid). The optimal configuration was found with C = 10, class_weight = balanced, and gamma = 0.1. Post-tuning, the SVM model exhibited significant improvements: Precision reached 95%, Recall 94%, F1-Score 94%, and Accuracy increased to 97%.

In the case of the ANN, we conducted a GridSearch with combinations of hyperparameters, including hidden layer sizes ((50,50), (100,)), activation functions (tanh, ReLU), solvers (SGD, Adam), alpha values (0.001, 0.05), and learning rates (constant, adaptive). The best-performing configuration was found to be hidden_layer_size = (50,50), activation = ReLU, solver = Adam, alpha = 0.0001, and learning_rate = adaptive. After tuning, the ANN demonstrated improvements across all evaluation metrics, achieving Precision of 95%, Recall of 97%, F1-Score of 98%, and an Accuracy of 95%.

 ISSN 2723-6471

These results are summarized in table 4, which presents the performance metrics of each algorithm following tuning. When compared to the results compiled in table 5 from previous research, it is evident that our tuned models show substantial improvements in classification accuracy across all tested algorithms.

Table 4. Result of algorithm performance when training model 80 and test model 20 With Tuning

Algorithm	Precision	Recall	F1-Score	Accuracy
Logistic Regression	98%	95%	97%	92%
KNN	92%	91%	92%	91%
Random Forest	95%	88%	92%	97%
SVM	95%	94%	94%	97%
ANN	95%	97%	98%	95%

Table 5. Results from previous research

Algorithm	Precision	Recall	F1-Score	Accuracy
Logistic Regression	91%	83%	81%	87%
KNN	87%	91%	89%	73%
Random Forest	86%	91%	90%	95%
SVM	81%	99%	89%	96%

The algorithms in table 4 outperform those in table 5 across most evaluation metrics, with Logistic Regression showing significant improvements in Precision, Recall, F1-Score, and Accuracy. KNN also has better Precision and Accuracy, though Recall is higher in table 5. Random Forest and SVM in table 4 demonstrate higher Precision and F1-Score, while SVM in table 5 has better Recall but lower Precision. ANN in table 4 excels overall, lacking a direct comparison in table 5. Performance differences stem from factors like Merge Data, Features used, Framework Learning Style, Handling Imbalance Data, and Hyperparameter Tuning outlined in table 6.

Table 62. The difference between this research and previous research

	Proposed Method	Previous Research
Dataset	OULAD	OULAD+Moodle LMS
Merge Data	Vle, StudentVle, and StudentInfo	Vle and StudentVle
Feature	activity_type, sum_click, id_site, code_module, id_student, code_presentation, region, imd_band, studied_credits, final_result, and study_type	id_site, code_module, code_presentation, id_student, date, sum_click, activity_type, and study_type
Framework Learning Style	Felder-Silverman Learning Style Model	Visual, Auditory, and Kinesthetic
Handling Imbalance Data	ADASYN, RandomUnderSampler, and SMOTEENN	Does not handle imbalance data
Hyperparameter Tuning	Use Hyperarameter Tuning for each algorithm	Does not use Hyperparameter Tuning

4.8. Discussion

The results of our study indicate that after hyperparameter tuning, the machine learning model effectively classifies students' learning preferences based on their activities in the Virtual Learning Environment (VLE). Key methods such as Random Forest for feature selection and oversampling techniques to tackle data imbalance have significantly enhanced the model's accuracy and overall performance.

In contrast to the research by Sayed et al., our analysis utilizing the OULAD dataset demonstrates several advancements that deepen our understanding of student interactions. Notably, we expand the data integration process by incorporating

 ISSN 2723-6471

a third table, StudentInfo, alongside the VLE and StudentVLE tables. This integration allows us to explore the interplay between demographic factors and engagement, providing a richer perspective on academic outcomes.

Our feature selection strategy also diverges from the previous study. While Sayed et al [12]. utilized a limited feature set including id_site, code_module, and activity_type. Our research incorporates additional variables such as region, imd_band, studied_credits, and final_result. This comprehensive approach captures the complexities of student engagement and enhances our ability to derive actionable insights.

Moreover, our study employs advanced techniques like ADASYN, RandomUnderSampler, and SMOTEENN to address data imbalance, which the prior research did not adequately tackle. This proactive approach ensures that minority classes are properly represented, thereby strengthening the reliability of our model.

We also adopt the Felder-Silverman Learning Style Model, which offers a more nuanced framework for understanding learning preferences compared to the Visual, Auditory, and Kinesthetic (VAK) model used in the previous study. This model considers multiple dimensions of learning, enriching our analysis of student engagement and success in the VLE.

Our methodology includes hyperparameter tuning for various algorithms, such as KNN, SVM, Logistic Regression, Random Forest, and ANN. By employing GridSearch and cross-validation, we optimize hyperparameters based on metrics like Precision, Recall, F1-Score, and Accuracy, significantly improving predictive power. This meticulous approach contrasts with the previous study's lack of tuning, enhancing the robustness of our findings.

Ultimately, our methodological advancements broaden the analysis scope and establish a new standard for future research in this domain, ensuring a more thorough understanding of student learning preferences in digital environments.

5. Conclusion

In conclusion, this study illustrates the significant impact of methodological steps on analyzing OULAD data, resulting in improved accuracy in predicting students' preferred learning styles. To further enhance the work, exploring strategies like ensemble methods or regularization to mitigate overfitting is essential. Techniques such as data preprocessing, mapping to the FSLSM, dimensionality reduction, and hyperparameter tuning greatly improved classification model performance, particularly Random Forest, which achieved 99% accuracy but showed signs of overfitting. Adjusting the min_samples_leaf parameter addressed this, enhancing generalization while maintaining 97% accuracy. Additionally, both SVM and ANN models demonstrated notable performance gains post-tuning, with SVM reaching 97% accuracy and strong precision, recall, and F1-Score metrics.

These findings highlight the potential for more effective personalization of learning experiences and deepen our understanding of students' learning style preferences. The insights advance knowledge in the field and provide practical applications for educators aiming to tailor their teaching strategies. Future research should explore other learning style models and diverse datasets to validate these findings, as well as investigate the long-term effects of personalized learning on student outcomes. Addressing these areas can enhance our understanding of learning preferences and improve educational practices across various settings.

6. Declarations

6.1. Author Contributions

Conceptualization: W.M.B., R.E.S., F.S.U., and S.; Methodology: F.S.U. and S.; Software: W.M.B.; Validation: W.M.B., R.E.S., F.S.U., and S.; Formal Analysis: W.M.B., R.E.S., F.S.U., and S.; Investigation: W.M.B., R.E.S., F.S.U., and S.; Resources: F.S.U. and S.; Data Curation: F.S.U. and S.; Writing Original Draft Preparation: W.M.B., R.E.S., F.S.U., and S.; Writing Review and Editing: F.S.U., S., W.M.B., and R.E.S.; Visualization: W.M.B.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

We would like to express our deepest gratitude to the Ministry of Education, Culture, Research, and Technology of Indonesia for the financial support provided for this research. This support is very meaningful in the process of conducting research and writing the manuscript so that we can achieve optimal results.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. T. Quadri and N. A. Shukor, "The Benefits of Learning Analytics to Higher Education Institutions: A Scoping Review," *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 23, pp. 4–15, 2021, doi: 10.3991/ijet.v16i23.27471.

[2] S. Septiani, S. Eko, J. Sutarto, and C. B. Utomo, "*Students and Artificial Intelligence*," pp. 691–697, 2023.

[3] Y. S. Mian, F. Khalid, A. W. C. Qun, and S. S. Ismail, "Learning Analytics in Education, Advantages and Issues: A Systematic Literature Review," *Creat. Educ.*, vol. 13, no. 09, pp. 2913–2920, 2022, doi: 10.4236/ce.2022.139183.

[4] M. Yin, H. Cao, Z. Yu, and X. Pan, "Manual Label and Machine Learning in Clustering and Predicting Student Performance: A Practice Based on Web-Interactive Teaching Systems," *Int. J. Web-Based Learn. Teach. Technol.*, vol. 19, no. 1, pp. 1–33, 2024, doi: 10.4018/IJWLTT.347661.

[5] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Human Behav.*, vol. 104, no. 106189, pp. 1-34, 2020.

[6] R. Qasrawi, S. P. V. Polo, D. A. Al-Halawa, S. Hallaq, and Z. Abdeen, "Assessment and Prediction of Depression and Anxiety Risk Factors in Schoolchildren: Machine Learning Techniques Performance Analysis," *JMIR Form. Res.*, vol. 6, no. 8, pp. 1–15, 2022, doi: 10.2196/32736.

[7] W. Admiraal, J. Vermeulen, and J. Bulterman-Bos, "Teaching with learning analytics:how to connect computer-based assessment data with classroom instruction?," *Technol. Pedagog. Educ.*, vol. 29, no. 5, pp. 577–591, 2020, doi: 10.1080/1475939X.2020.1825992.

[8] R. Brungel, J. Ruckert, and C. M. Friedrich, "Project-Based Learning in a Machine Learning Course with Differentiated Industrial Projects for Various Computer Science Master Programs," 2020 *IEEE 32nd Conf. Softw. Eng. Educ. Training*, CSEE T 2020, vol. 32, no. Ml, pp. 50–54, 2020, doi: 10.1109/CSEET49119.2020.9206229.

[9] K. L. M. Ang, F. L. Ge, and K. P. Seng, "Big Educational Data Analytics: Survey, Architecture and Challenges," *IEEE Access*, vol. 8, no. 2020, pp. 116392–116414, 2020, doi: 10.1109/ACCESS.2020.2994561.

[10] M. I. I. bin Zainuddin and H. Mohamad Judi, "Designing and Incorporating Personalized Learning Analytics: Examining Self-Regulated Meaningful Learning," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 11, no. 12, pp. 2276–2287, 2021, doi: 10.6007/ijarbss/v11-i12/11327.

[11] T. Mukherjee, "Improving Data Quality for Educational Data Mining (EDM) for Indian Ed-Tech Start-Ups," *Int. J. Sci. Eng.* Manag., vol. 9, no. 8, pp. 32–33, 2022, doi: 10.36647/ijsem/09.08.a005.

[12] A. Rashad Sayed, M. Helmy Khafagy, M. Ali, and M. Hussien Mohamed, "Predict student learning styles and suitable assessment methods using click stream," *Egypt. Informatics J.*, vol. 26, no. April, p. 100469, 2024, doi: 10.1016/j.eij.2024.100469.

[13] D. Niu, K. Wang, L. Sun, J. Wu, and X. Xu, "Short-term photovoltaic power generation forecasting based on random forest feature selection and CEEMD: A case study," *Appl. Soft Comput. J.*, vol. 93, no. 2020, pp. 106389, 2020, doi: 10.1016/j.asoc.2020.106389.

[14] X. K. Li, W. Chen, Q. Zhang, and L. Wu, "Building Auto-Encoder Intrusion Detection System based on random forest feature selection," *Comput. Secur.*, vol. 95, no. 2020, pp. 101851, 2020, doi: 10.1016/j.cose.2020.101851.

[15] W. Huo, W. Li, Z. Zhang, C. Sun, F. Zhou, and G. Gong, "Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection," *Energy Convers. Manag.*, vol. 243, no. June, p. 114367, 2021, doi: 10.1016/j.enconman.2021.114367.

[16] O. Oladipupo and S. Samuel, "A Learning Analytic Approach to Modelling Student-Staff Interaction From Students' Perception of Engagement Practices," *IEEE Access*, vol. 12, no. December 2023, pp. 10315–10333, 2024, doi: 10.1109/ACCESS.2024.3352440.

[17] Enitan Shukurat Animashaun, Babajide Tolulope Familoni, and Nneamaka Chisom Onyebuchi, "Advanced machine learning techniques for personalising technology education," *Comput. Sci. IT Res. J.*, vol. 5, no. 6, pp. 1300–1313, 2024, doi: 10.51594/csitrj.v5i6.1198.

[18] A. R. Masegosa, R. Cabañas, A. D. Maldonado, and M. Morales, "Learning Styles Impact Students' Perceptions on Active Learning Methodologies: A Case Study on the Use of Live Coding and Short Programming Exercises," *Educ. Sci.*, vol. 14, no. 3, 2024, doi: 10.3390/educsci14030250.

[19] R. D. Mahande and N. M. Abdal, "A HyFlex learning measurement model based on students' cognitive learning styles to create equitable learning," *World J. Educ. Technol. Curr. Issues*, vol. 14, no. 5, pp. 1469–1481, 2022, doi: 10.18844/wjet.v14i5.7777.

[20] M. Haviz, I. M. Maris, E. Nasrul, D. Azis, and L. Lufri, "The Learning Styles of Prospective Biology Teachers at Islamic University in Indonesia," *Pedagogika*, vol. 149, no. 1, pp. 238–256, 2023, doi: 10.15823/p.2023.149.11.

[21] S. Ren, "Optimization of English Classroom Interaction Models Incorporating Machine Learning," *J. Electr. Syst.*, vol. 20, no. 6s, pp. 1669–1681, 2024, doi: 10.52783/jes.3086.

[22] M. S. Hasibuan, R. Z. Abdul Aziz, D. A. Dewi, T. B. Kurniawan, and N. A. Syafira, "Recommendation Model for Learning Material Using the Felder Silverman Learning Style Approach," *HighTech Innov. J.*, vol. 4, no. 4, pp. 811–820, 2023, doi: 10.28991/HIJ-2023-04-04-010.

[23] I. Azzi, A. Radouane, L. Laaouina, A. Jeghal, A. Yahyaouy, and H. Tairi, "Fuzzy Classification Approach to Select Learning Objects Based on Learning Styles in Intelligent E-Learning Systems," *Informatics*, vol. 11, no. 2, pp. 1-12, 2024, doi: 10.3390/informatics11020029.

[24] L. Zhao, J. Ren, L. Zhang, and H. Zhao, "Quantitative Analysis and Prediction of Academic Performance of Students Using Machine Learning," *Sustain.*, vol. 15, no. 16, pp. 1-18, 2023, doi: 10.3390/su151612531.

[25] S Sukanya and Dr D William Albert, "A Novel Approach to Predict Students Performance through Machine Learning," *Int. J. Eng. Technol. Manag. Sci.*, vol. 7, no. 5, pp. 278–283, 2023, doi: 10.46647/ijetms.2023.v07i05.032.

[26] R. Nazempour and H. Darabi, "Personalized Learning in Virtual Learning Environments Using Students' Behavior Analysis," *Educ. Sci.*, vol. 13, no. 5, pp. 1-15, 2023, doi: 10.3390/educsci13050457.

[27] "OU Analyse | Knowledge Media Institute | The Open University." Accessed: Oct. 07, 2024. [Online]. Available: https://analyse.kmi.open.ac.uk/open_dataset

[28] H. Mohammedqasim, A. A. Jasim, R. Mohammedqasem, and O. Ata, "Enhancing Predictive Performance in Covid-19 Healthcare Datasets: a Case Study Based on Hyper Adasyn Over-Sampling and Genetic Feature Selection," *J. Eng. Sci. Technol.*, vol. 19, no. 2, pp. 598–617, 2024.

[29] A. Krajah, Y. F. Almadani, H. Saadeh, and A. Sleit, "Analyzing Covid-19 Data Using Various Algorithms," 2021 *IEEE Jordan Int. Jt. Conf. Electr. Eng. Inf. Technol. JEEIT 2021 - Proc.*, vol. 2021, no. 1, pp. 66–71, 2021, doi: 10.1109/JEEIT53412.2021.9634124.

[30] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–14, 2022, doi: 10.1186/s12911-022-02075-2.

[31] M. Binkhonain and L. Zhao, "A review of machine learning algorithms for identification and classification of non-functional requirements," *Expert Syst. with Appl. X*, vol. 1, no. 2019, pp. 1-13, 2019, doi: 10.1016/j.eswax.2019.100001.