**Artificial Intelligence in Educational Decision Sciences**
**E-ISSN: 3123-8823**
**Vol. 1, No. 1, February 2026**

# Comparison of Dataset Proportions in SVM and Random Forest Algorithms in Detecting Student Dependence on AI in Learning

**Sardar Faroq Ahmad Khan**
Universitas Negeri Makassar, Makassar, Indonesia
**Pramudya Asoka Syukur**
Universitas Negeri Makassar, Makassar, Indonesia
**Andi Baso Kaswar**
Okayama University, Japan
**Marwan Ramdhany Edy**
Universitas Gadjah Mada, Yogyakarta, Indonesia

## ABSTRACT

**Purpose** – The rapid integration of artificial intelligence (AI) in education has raised concerns about excessive student dependence, potentially undermining critical thinking and learning autonomy. This study aims to identify the most effective machine learning algorithm for detecting AI dependency in learning activities and to examine the impact of training–testing data proportion on predictive performance.

**Methods** - This study employs the CRISP-DM framework and applies two supervised classification algorithms, Random Forest and Support Vector Machine (SVM), to a synthetic dataset of 10,000 AI-assisted learning sessions. The target variable, perceived AI assistance level, was discretised into three categories (low, medium, and high). Model performance was evaluated under four dataset split scenarios (60:40, 70:30, 80:20, and 90:10) using accuracy, AUC, precision, recall, and F1-score.

**Findings** - The results show that Random Forest consistently outperforms SVM across all dataset proportions and evaluation metrics. The highest performance was achieved by Random Forest with a 60:40 split, yielding an accuracy of 67.6% and an AUC of 80.8%. Although SVM demonstrated stable performance, it required larger training datasets and remained inferior to Random Forest.

**Research limitations** - The use of synthetic data and limited behavioural features restricts the generalisability of the findings. The moderate accuracy indicates that AI dependency is a complex construct not fully captured by the current model.

**Originality** - This study provides empirical evidence on the combined influence of algorithm selection and dataset proportion in detecting AI dependency, offering practical guidance for developing early-warning systems to support responsible AI use in education.

OPEN ACCESS

**ARTICLE HISTORY**
Received: 18-11-2025
Revised: 25-01-2026
Accepted: 02-02-2026

**KEYWORDS**
AI Dependency;
Machine Learning;
Random Forest;
Support Vector Machine;
Proportion Dataset

**Correspondence Author**: sardarfaroq88@gmail.com
**To cite this article** : Khan, S. F. A., Syukur, P. A., Kaswar, A. B., & Edy, M. R. (2026). Comparison of dataset proportions in SVM and random forest algorithms in detecting student dependence on AI in learning. Artificial Intelligence in Educational Decision Sciences, 1(1), 1–13.

## INTRODUCTION

The use of artificial intelligence (AI) is becoming increasingly widespread and an integral part of various sectors of life, including education. According to Chegg's 2025 Global Student Survey, 95% of Indonesian students have used GenAI in their learning process, the highest figure globally (Chegg.org, 2025). This finding indicates that AI has become a commonly used tool in the learning process in Indonesia. However, the high level of AI utilisation also raises concerns about the potential for dependency, which could undermine students' learning efforts, critical thinking skills, and independence in completing tasks (Baria, 2025; Çela et al., 2024). Therefore, in the context of an evolving education system, particularly in developing countries such as Indonesia, the use of AI needs to be balanced with pedagogical approaches that encourage independent learning and strengthen students' thinking skills (Indriyani & Solihati, 2021).

Over-reliance on AI technology in education risks leading to a number of serious problems. One of the impacts is lazy learning behaviour, where students prefer to search for instant answers from AI rather than processing information deeply and independently (Ahmad et al., 2023). In addition, this dependence also opens up the potential for academic misconduct, such as plagiarism, as students tend to copy answers from AI systems without deep understanding (Uppal & Hajian, 2024). Not only that, uncontrolled dependence on AI also risks weakening critical thinking skills and problem-solving abilities that are very important for students' academic and professional development (Octaberlina et al., 2024; Rahardyan et al., 2024). Therefore, it is important to identify effective methods for detecting and preventing excessive dependence on AI in the learning process, so that the quality of education is maintained and not affected by the convenience of technology.

Current research on detecting or preventing dependence on AI in education is generally still exploratory and descriptive in nature. The main focus is on identifying the factors that cause dependence and its impact on critical thinking and independent learning abilities. Excessive use of AI dialogue systems can reduce critical thinking and decision-making abilities, especially when students do not verify the answers provided (Zhai et al., 2024). Meanwhile, explored the variable of user trust in AI and showed that excessive trust, even when it conflicts with personal judgement or contextual information, can lead to erroneous decisions in collaborative interactions (Klingbeil et al., 2024). Explanations from AI are only effective in reducing dependence if users feel that using AI is less cognitively demanding than completing tasks independently (Vasconcelos et al., 2023). These findings indicate that although an understanding of the causes and effects of dependence has begun to take shape, the approaches used are still limited to behavioural analysis and have not led to the development of systematic predictive models for detecting potential dependence early on.

Previous studies have demonstrated the effectiveness of machine learning approaches in predicting digital dependency in various contexts, ranging from the internet and social media to smartphone usage. One study showed that Neural Networks and XGBoost models were able to predict internet addiction with the highest accuracy of 91% and 90%, respectively, using behavioural features such as app usage duration, login frequency, and NLP-based emotional analysis (Abdel Wahed & Abdel Wahed, 2025). Another study found that the XGBoost model was most effective in identifying social media addiction triggered by academic frustration, with key features such as academic stress, excessive usage time, and social comparisons between users (Bin Rofi et al., 2024). Similar findings were also shown in the context of smartphone addiction, where Gradient Boosting produced the highest accuracy of 91.2% in predicting its impact on academic performance, utilising features such as daily screen time, sleep disturbance, impulsivity, and anxiety (Vimala & Sheela, 2025). Although all three offer diverse technical approaches and utilise rich behavioural variables, these studies still show a uniform pattern of approach: the use of binary classification (dependent or not), the dominance of frequency variables as the main indicator, and a failure to consider the influence of dataset ratio variation on model performance.

Therefore, this study aims to determine the most effective machine learning algorithm for detecting excessive dependence on AI in the learning process, which has not yet been specifically studied in the context of education. In addition, this study also aims to explore the influence of dataset ratio differences on predictive model performance, given that this aspect has not been a major concern in previous studies. Thus, this study will not only contribute to the selection of the most optimal algorithm, but also produce strategic recommendations regarding the ideal proportion of training

and testing data, in order to improve the accuracy and generalisation of the model in detecting AI dependence. Furthermore, through this approach, it is hoped that early detection of dependence can be carried out more precisely, enabling more targeted interventions in order to maintain the quality of learning and the digital well-being of students.

## METHOD

In this study, a systematic approach based on the CRISP-DM (Cross Industry Standard Process for Data Mining) model was used to determine the best algorithm for detecting dependence on artificial intelligence (AI) in the learning process and to identify the influence of dataset proportions on predictive model testing.

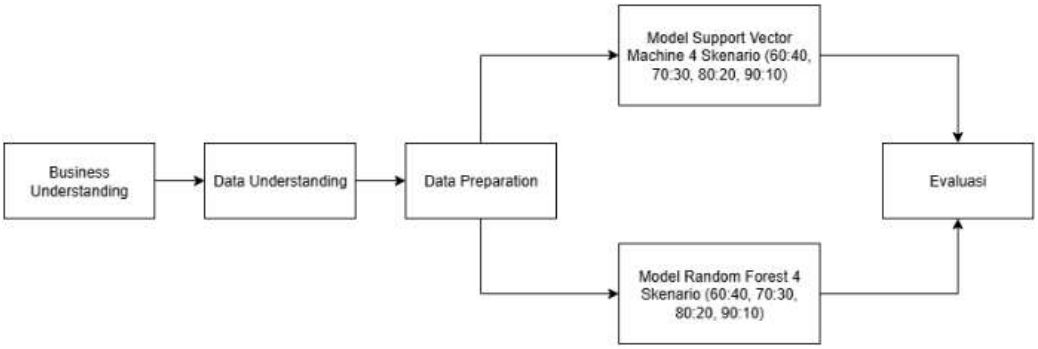


**Figure 1.** Research Flowchart

### Business Understanding

The first stage of this method is business understanding, where the main objective of this study is to compare the performance of Random Forest and Support Vector Machine (SVM) algorithms in analysing students' dependence on the use of AI in learning. It is essential to understand how AI is integrated into the educational environment and how it affects the way students learn.

### Data Understanding

At this stage, a synthetic dataset was used, taken from Kaggle (https://www.kaggle.com/), with a design resembling the actual patterns and behaviour of students interacting with AI, amounting to 10,000 sessions. The use of this data can lead to more realistic analysis without having to access personal data, making it safe to use and avoiding privacy and copyright violations.

### Data Preparation

The data preparation stage is the phase in which previously obtained data is processed and cleaned so that it can be used optimally by machine learning algorithms. Although the dataset used is synthetic, the cleaning and transformation process is still carried out systematically using various components available in Orange.

Removal of Irrelevant Attributes

Certain attributes that do not contribute directly to the classification process, such as unique identifiers or time-based attributes, are removed from the dataset. The purpose of this step is to reduce noise in the data and prevent the model from learning irrelevant or individualistic information, which can reduce the accuracy of the model's predictions and generalisation (Villuendas-Rey et al., 2024).

Checking for Missing or Duplicate Data

The data is checked to identify missing values, duplicate data, or format inconsistencies. This step is important to ensure that the data is clean and ready for use before the model training process begins. In addition, this check is performed to avoid errors or bias during modelling, as well as to ensure that each feature used has a valid contribution to the classification results (Boros & Kmetty, 2024; Malnad College of Engineering,Hassan & Balgotra, 2025).

## Data Transformation

Data transformation is an important part of the preparation stage, as it ensures that all attributes are in a numerical format that can be processed by machine learning algorithms. In Orange, this transformation is performed using widgets such as Continuize and Discretize, and is divided into three parts:

One-hot encoding

One-hot encoding is used to convert nominal categorical attributes into binary numerical representations. In this process, each category of a feature is converted into a separate column that indicates the presence (1) or absence (0) of that category. The purpose of this technique is to avoid incorrect assumptions about the order or numerical weight of attributes that do not actually have an ordinal relationship, and to enable machine learning algorithms to process categorical data accurately (Hancock & Khoshgoftaar, 2020).

Normalisation of Binary/Ordinal Data

Binary or ordinal data normalisation is performed to convert attributes that are binary categorical or have a logical order into numerical form. This process is carried out using the Continuize widget, which converts values such as TRUE and FALSE into the numbers 1 and 0. The aim is to provide a numerical representation that is acceptable to algorithms, while retaining the logical order or binary status of the original data.

Discretisation of Target Variables

Target variable discretisation is performed when the target attribute is in the form of a numerical scale and needs to be converted into categories. The method used is Equal Width Interval, which divides the range of values into several intervals of equal width. After that, the category labels are adjusted or clarified using the Edit Domain widget. The purpose of this process is to simplify the target into categories that are easier to interpret and to facilitate the application of category-based classification models, rather than numerical regression.

## Modeling

At this stage, modelling was carried out by applying two classification algorithms, namely Random Forest (RF) and Support Vector Machine (SVM), to build a predictive model that could estimate the level of student dependence on AI.

Random Forest

Random Forest is an ensemble algorithm that combines the prediction results of a number of decision trees to improve model accuracy and minimise the risk of overfitting. Each tree is constructed using random samples from the data and random subsets of available features, resulting in variation between trees. The final prediction result is obtained through a voting process or by averaging all trees. This approach is very effective in handling high-dimensional data and tends to be more stable because it reduces the variance of each individual model (Zhu, 2020).

Support Vector Machine

SVM or Support Vector Machine works by mapping data to a high-dimensional space in order to find the optimal hyperplane that separates classes in the data. This algorithm attempts to maximise the margin, which is the distance between the hyperplane and the nearest data point from each class (support vectors). For data that cannot be separated linearly, SVM uses a kernel function to transform the data into a higher-dimensional space. Although SVM is effective in handling high-dimensional data, its optimal performance requires a sufficient amount of training data (Ghosh & Cabrera, 2022). Model training and testing were conducted using Orange software. The dataset was divided into four different training and testing ratio scenarios, namely 60:40, 70:30, 80:20, and 90:10. The data division process was carried out using the Data Sampler component. The built model was then trained and evaluated using Test and Score, while the prediction results could be observed through the Prediction component.

Evaluation

After the model was built, an evaluation was conducted to measure how effective the model was in detecting over-reliance on AI. The results of each algorithm were compared to determine which method was most effective and how the proportion of the dataset affected the model's results. The evaluation method used is Test and Score, which provides information on the number of correct and incorrect predictions in each class to provide evaluation results such as area under curve (AUC), accuracy, F1-score, precision, and recall.

AUC is calculated from the ROC (Receiver Operating Characteristic) curve, which plots the True Positive Rate (Recall) against the False Positive Rate. AUC values range from 0 to 1, where 1 indicates perfect classification and 0.5 indicates that the model is no better than random guessing. Although there is no direct mathematical formula in the reference, the understanding of AUC comes from the basic concept of ROC curve analysis.

1. Accuracy

Accuracy measures how many predictions are correct compared to the total number of predictions made. The accuracy formula can be expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where (TP) is true positive, (TN) is true negative, (FP) is false positive, and (FN) is false negative.

2. Precision

Precision is the proportion of correct positive results among all predicted positive results. The formula for precision is as follows:

$$Precision = \frac{TP}{TP + FP}$$

Precision provides an indication of the accuracy of results classified as positive.

3. Recall

Recall, or sensitivity, indicates how well the model captures all existing positives. The recall formula can be written as follows:

$$Recall = \frac{TP}{TP + FN}$$

Recall assesses the model's ability to identify all actual positive examples.

4. F1-score

The F1-score is a measure that combines precision and recall into a single number. It is particularly useful when there is class imbalance in the dataset. The formula for the F1-score is as follows:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The F1-score provides a better picture of model performance on imbalanced datasets than a simple separation between precision and recall (Dučić et al., 2020; He et al., 2022).

## RESULTS AND DISCUSSION

This study aims to compare the effectiveness of Random Forest and SVM algorithms in detecting student dependence on the use of artificial intelligence (AI) in learning. In addition, this study also aims to identify the effect of dataset division on the accuracy of predictive models used to detect excessive dependence on AI.

## Data Understanding

The dataset used in this study was obtained from the Kaggle platform, an open repository that provides various datasets for research and development purposes. This dataset is titled "AI Assistant Usage in Student Life (Synthetic)", compiled by Ayesha Saleem and first published in 2023. This dataset is the result of simulated interactions between students and artificial intelligence assistants (such as ChatGPT) in an academic context. Overall, this dataset consists of 10,000 unique interaction sessions, where each row represents one student session covering various variables that describe AI-assisted learning activities. The information recorded includes the student's education level, field of study, date of interaction, and the type of task performed while using AI. This research utilises the data to analyse patterns of AI use in academic activities and to compare the effectiveness of classification algorithms, namely Random Forest and Support Vector Machine (SVM), in predicting student dependence on AI.

Overall, there are 11 main attributes in this dataset. The types of data used consist of categorical (such as StudentLevel, Discipline, and TaskType), ordinal (such as AI_AssistanceLevel and SatisfactionRating), numerical (such as SessionLengthMin and TotalPrompts), and date types (SessionDate). In this study, the main focus is on predicting the level of assistance students feel from AI, which is represented by the AI_AssistanceLevel attribute as the target variable. This attribute has an ordinal scale from 1 to 5, which reflects the extent to which students feel assisted by AI in completing tasks. More specifically, the raw data structure before preprocessing can be seen in the following table:

**Table 1.** Initial Data

| SessionID | Student Level | Discipline | Session Date | Session Length Min | Total Prompts | TaskType | AI_Assistance Level | FinalOutcome | UsedAgain | Satisfaction Rating |
|---|---|---|---|---|---|---|---|---|---|---|
| SESSION00001 | Graduate | Business | 1/17/2025 | 7.54 | 1 | Coding | 5 | Idea Drafted | TRUE | 5 |
| SESSION00002 | High School | Biology | 9/7/2024 | 14.6 | 3 | Writing | 3 | Assignment Completed | FALSE | 1.9 |
| SESSION00003 | Undergraduate | Business | 1/12/2025 | 19.22 | 5 | Coding | 3 | Assignment Completed | TRUE | 3.3 |
| SESSION00004 | Undergraduate | Computer Science | 5/6/2025 | 3.7 | 1 | Coding | 3 | Assignment Completed | TRUE | 3.5 |
| ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| SESSION10000 | Undergraduate | Math | 4/16/2025 | 10.85 | 3 | Writing | 4 | Assignment Completed | TRUE | 4.9 |

## Data Preparation Results

Before forming a classification model with the Random Forest and Support Vector Machine (SVM) algorithms, several preprocessing stages were carried out to ensure that the data used was suitable for the algorithm's requirements, particularly as the algorithm to be used required input in numerical form. The preprocessing stages were carried out as follows:

## Removal of Irrelevant Attributes

The SessionID and SessionDate attributes were removed from the dataset because they had no direct relevance to the classification process and were unique/date-based, which did not contribute meaningfully to the predictive model.

## Transformation of Categorical Data to Numerical Data

1. The StudentLevel, Discipline, and TaskType attributes are transformed using One-Hot Encoding with the help of the Continuize widget in Orange. Each category will be broken down into binary features that represent the existence of a class.

2. The UsedAgain attribute, which was originally a categorical binary type (TRUE/FALSE), was converted to ordinal numerical data using the following conversion: FALSE = 0, TRUE = 1, using the Continuize widget.

## Target Transformation (AI_AssistanceLevel)

The AI_AssistanceLevel attribute was originally ordinal numerical data (scale 1–5) and was used as the target variable in this study. For classification purposes, this attribute was discretised using the Discretise widget with the Equal Width Interval method into three categories. Next, the category labels were updated using the Edit Domain widget to better represent the level of student dependence on AI. The categorisation details are presented in the following table:

**Table 2.** Target Variable Discretisation Results

| Value Range | Description Label |
|---|---|
| < 2.33 | Low |
| 2.33 – 3.67 | Medium |
| > 3.67 | High |

The above preprocessing steps aim to ensure that all input features are in a numerical format that is compatible with the Random Forest and SVM algorithms. In addition, the target variable is also discretised to produce a more balanced and interpretable class structure in the classification analysis. The results of the preprocessed data can be seen in the following table:

**Table 3.** Pre-Processing Results Data

| SessionLengthMin | TotalPrompts | StudentLevel_Undergraduate | ... | Discipline_Business | ... | TaskType_Coding | ... | UsedAgain | SatisfactionRating | AI_AssistanceLevel (Kategori) |
|---|---|---|---|---|---|---|---|---|---|---|
| 7.54 | 1 | 0 | ... | 1 | ... | 1 | ... | 1 | 5.0 | High |
| 14.6 | 3 | 0 | ... | 0 | ... | 0 | ... | 0 | 1.9 | Medium |
| 19.22 | 5 | 1 | ... | 1 | ... | 1 | ... | 1 | 3.3 | Medium |
| 3.7 | 1 | 1 | ... | 0 | ... | 1 | ... | 1 | 3.5 | Medium |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 10.85 | 3 | 1 | ... | 0 | ... | 0 | ... | 1 | 4.9 | High |

### Modelling Result

At the modelling stage, this study used two classification algorithms, namely Random Forest and Support Vector Machine (SVM). Both were applied to classify the level of assistance perceived by students towards AI assistants, which had been discretised into three categories: Low, Medium, and High.

The modelling process begins by entering the preprocessed data into the system. All input attributes have been converted to numerical format and the classification targets have been categorised to suit the requirements of the algorithm used.

The classification model is then built using the Random Forest algorithm, which is a decision tree-based ensemble algorithm, and the SVM algorithm, which works by maximising the separating margin between classes. The performance of these two models was tested against four different training and testing data division scenarios, with the aim of observing the effect of data proportion on the accuracy and stability of the classification model.

The data division scenarios applied were as follows:

- Scenario 1: 60% training data and 40% test data (60:40)
- Scenario 2: 70% training data and 30% test data (70:30)
- Scenario 3: 80% training data and 20% test data (80:20)
- Scenario 4: 90% training data and 10% test data (90:10)

To see the modelling process in more detail, including the structure and stages of the components used, please refer to the following image:

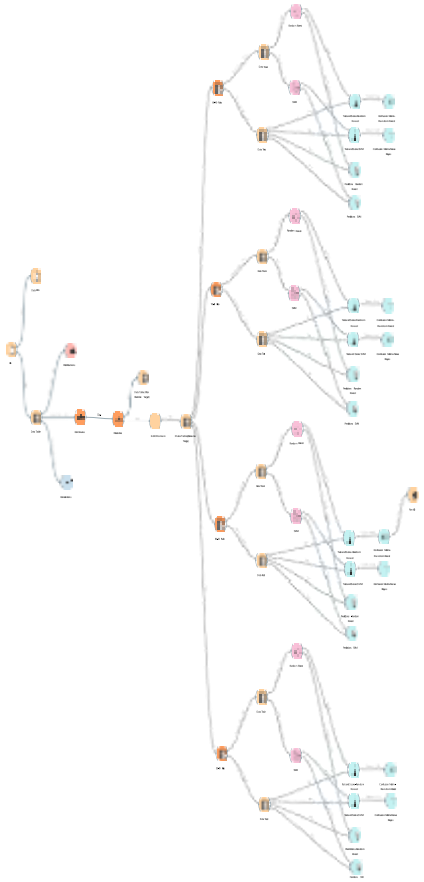

**Figure 2.** Model Flow

## Evaluation Result

The model evaluation was conducted to measure the effectiveness of two classification algorithms, Random Forest and Support Vector Machine (SVM), in predicting the level of student dependence on AI assistance. The evaluation process covered four training and testing data ratio scenarios, namely 60:40, 70:30, 80:20, and 90:10. The performance of each model was assessed using a number of metrics, including accuracy (CA), AUC, F1-score, precision, and recall.

## Evaluation of the Random Forest Model on the Proportion of the Dataset

The Random Forest model shows fairly consistent accuracy performance across various data split scenarios. The highest accuracy was achieved in the 60:40 scenario with a value of 67.6%, followed by 70:30 (66.7%), 90:10 (66.5%), and the lowest in 80:20 (63.5%). The insignificant differences in accuracy between scenarios indicate that Random Forest is relatively stable and resilient to variations in the proportion of training and testing data.

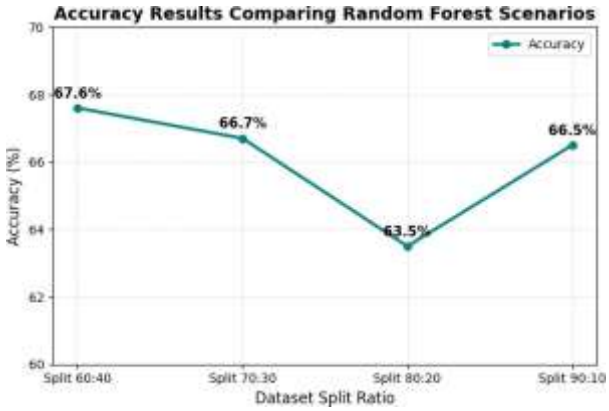


**Figure 3.** Random Forest Accuracy Results Line Diagram

## Evaluation of the Support Vector Machine (SVM) Model on the Proportion of the Dataset

Unlike Random Forest, the SVM model shows slightly lower accuracy in most data division scenarios. The highest accuracy was achieved in the 90:10 scenario at 64.5%, followed by 80:20 (62.5%), 60:40 (62.4%), and the lowest in 70:30 (62.0%). Although the differences between scenarios are relatively small, these results indicate that SVM accuracy tends to be stable but not as good as Random Forest.

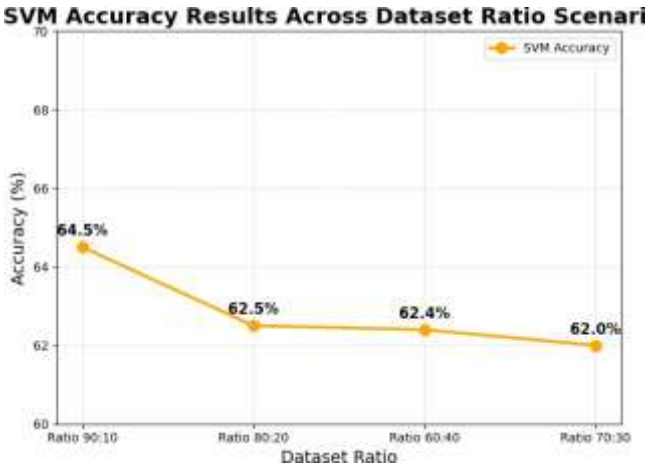


**Figure 4.** SVM Accuracy Results Line Diagram

## Accuracy Comparison and Comprehensive Evaluation

Based on the results of evaluating both models using the Test and Score widget, a comprehensive comparative analysis can be performed on the performance of the Random Forest and Support Vector Machine (SVM) algorithms in predicting the level of student dependence on the use of AI in learning. The evaluation was conducted using five main metrics, namely AUC, accuracy (CA), F1-score, precision, and recall, in four scenarios of training and testing data division. The complete results of this evaluation can be seen in the table below.

**Table 4.** Results of Accuracy Comparison & Evaluation

| Dataset Ratio | Algorithm | AUC (%) | CA (%) | F1 (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|
| 60-40 | Random Forest | 80.7 | 67.6 | 67.3 | 67.3 | 67.6 |
| | Support Vector Machine (SVM) | 76.7 | 62.4 | 60.5 | 60.3 | 62.4 |
| 70-30 | Random Forest | 80.8 | 66.7 | 66.4 | 66.3 | 66.7 |
| | Support Vector Machine (SVM) | 76.8 | 62.0 | 59.2 | 59.3 | 62.0 |
| 80-20 | Random Forest | 79.2 | 63.5 | 63.2 | 62.9 | 39.7 |
| | Support Vector Machine (SVM) | 78.7 | 62.5 | 60.9 | 60.5 | 62.5 |
| 90-10 | Random Forest | 79.2 | 66.5 | 65.6 | 65.3 | 66.5 |
| | Support Vector Machine (SVM) | 79.6.4 | 64.5 | 63.6 | 63.7 | 64.5 |

Based on the evaluation results table presented, the evaluation results show that Random Forest consistently performs better than SVM in almost all scenarios and evaluation metrics. Specifically, in the 60:40 scenario, Random Forest recorded the highest accuracy of 67.6%, demonstrating its ability to classify students' level of dependence on AI quite accurately. Meanwhile, SVM recorded the highest accuracy of 64.5% in the 90:10 scenario, but SVM's performance remained lower than Random Forest in all other scenarios. In terms of AUC, which measures the model's ability to distinguish between dependency classes (low, medium, high), Random Forest again excelled with a range of values between 79.2% and 80.8%, reflecting its consistency in recognising patterns of student

dependence on AI. On the other hand, the AUC value of SVM is slightly lower, ranging from 76.7% to 79.6%, which, although considered good, still does not exceed the performance of Random Forest. When viewed from the F1-score and precision metrics, Random Forest shows higher values, reflecting its ability to provide balanced predictions with minimal errors, particularly in identifying students who truly fall into the high dependency category. On the other hand, SVM tends to produce lower F1-score and precision values, despite showing stability in the recall metric, which is its ability to capture all high dependency cases. However, this stability is not sufficient to compensate for the model's weakness in providing more precise and balanced predictions.

The effect of dataset distribution proportions on the performance of both algorithms is clear. In the 60:40 scenario, with a fairly balanced proportion of training and test data, Random Forest demonstrated better ability to generalise and predict student dependence on AI. Conversely, SVM showed better results in the 90:10 scenario, with larger training data, but its performance remained lower than Random Forest in all other data divisions. Overall, Random Forest is more robust and reliable, providing more stable and accurate results in almost all data division scenarios. Based on these evaluation results, Random Forest is recommended as a more effective model for predicting the level of student dependence on the use of AI in learning, mainly due to this model's ability to provide more consistent and reliable results in various data division scenarios.

## Discussion

The two classification algorithms used in this study, Support Vector Machine (SVM) and Random Forest, showed consistent comparative results, with Random Forest performing better in predicting student dependence on AI. This advantage was evident in almost all dataset division scenarios, with accuracy, AUC, precision, recall, and F1-score values that were generally higher than those of SVM. The superiority of Random Forest is in line with the findings of Zhu (2020) and Ghos (2022), who stated that this algorithm is resistant to overfitting and works well on high-dimensional and heterogeneous data (Al-Areef & Saputra, 2023; Firmansyach et al., 2023). On the other hand, although SVM is known to be effective in handling non-linear data, its performance in this study was unable to match the stability and flexibility of Random Forest.

However, even though Random Forest excels, the highest accuracy achieved is still around 67.6%, which is considered moderate in predictive classification. Several factors may contribute to this limitation. First, the data used is a synthetic dataset, which, although close to real conditions, still has limitations in representing the complexity of actual user behaviour (Holmes & Theodorakopoulos, 2020; Rankin et al., 2020). Second, the distribution of target classes (low, medium, high dependency levels), which may be unbalanced, can affect the model's ability to learn from minority data (Thabtah et al., 2020; Thölke & others, 2023). Third, the available features may not fully represent the determinants of dependency, such as students' intrinsic motivation, social factors, or institutional policies that influence AI usage behaviour.

The second objective of this study was to evaluate the effect of dataset proportion on classification model performance. Four data division scenarios were used, namely 60:40, 70:30, 80:20, and 90:10, to assess how variations in the training and testing data ratios affected the accuracy and stability of prediction results. The results show that the 60:40 scenario, with a training data proportion of 60%, provides the best performance for the Random Forest model, both in terms of accuracy and AUC. This indicates that Random Forest is capable of producing reliable predictive models even though the amount of training data is not as large as in other scenarios. This is in line with previous studies showing that the bagging mechanism and random feature selection in Random Forest make it more resistant to data variation and the risk of overfitting (Chai et al., 2022; Salman et al., 2024). In contrast, the SVM model showed the best performance in the 90:10 scenario, where the training data comprised 90% of the dataset. These findings indicate that SVM requires more training data to form an optimal separating hyperplane, and its performance improves as more information from the training data is added (Gu & Congalton, 2025). Both results show that although data division with an 80:20 ratio is a common practice widely used in model evaluation (Joseph, 2022), the effectiveness of this ratio is not always uniform and is highly dependent on the characteristics of the algorithm used and the structure and size of the dataset (Hatamian et al., 2025; Sabah et al., 2023).

Overall, this study confirms that the selection of algorithms and dataset partitioning strategies are two crucial factors in the development of predictive models in the field of education. Random Forest

is recommended as a more reliable algorithm in the context of detecting dependence on AI due to its ability to manage data variation and provide more stable results in various scenarios. Meanwhile, the proportion of the dataset used must be adjusted to the specific objectives of the modelling, taking into account the actual performance and generalisation capabilities of the model. With this approach, it is hoped that the developed model can be used as an early detection tool to prevent the negative impacts of excessive AI use in the learning process.

## CONCLUSION

This study shows that the Random Forest algorithm consistently outperforms Support Vector Machine (SVM) in predicting the level of student dependence on AI use in learning activities. Evaluations conducted on various data ratio scenarios show that Random Forest is more stable and reliable, with higher accuracy, AUC, F1-score, precision, and recall metrics. However, the highest accuracy of only 67.6% indicates limitations, mainly because the dataset used is synthetic and does not fully cover various important factors that influence student dependence as a whole.

For future research, the use of actual data from real learning environments is recommended so that the prediction results are more valid and representative. In addition, the addition of more relevant variables, such as learning motivation, the intensity of AI use in the long term, and psychological factors, can improve the quality of the model. Subsequent research should also consider other algorithms such as XGBoost or deep learning-based models to obtain more optimal classification results. The findings from this study can serve as a starting point in the development of adaptive learning systems that are capable of detecting and managing students' dependence on AI technology more wisely in the modern era of education.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the lecturer of the Data Mining course for the academic guidance, methodological direction, and constructive feedback provided throughout the preparation of this article. The authors also acknowledge the Informatics and Computer Engineering Education Study Program, Faculty of Engineering, Universitas Negeri Makassar, for providing a supportive academic environment that enabled the successful completion of this scholarly work. The authors are solely responsible for the content, analysis, and conclusions presented in this article.

## AUTHOR CONTRIBUTION STATEMENT

SFK contributed to the formulation of the research idea, methodological design, data processing and analysis, and the preparation of the initial manuscript draft. PAS, ABK and MER contributed to the development of the conceptual framework, evaluation of the modeling results, literature review, and critical revision and refinement of the manuscript. Both authors contributed equally to the interpretation of findings, discussion of results, and approval of the final version of the manuscript.

## AI DISCLOSURE STATEMENT

The authors used ChatGPT during the preparation of this manuscript for language refinement, structural editing, and improvement of clarity. Following the use of this tool, all content was thoroughly reviewed, revised, and edited by the authors. The research design, data collection, data analysis, interpretation of results, and final conclusions were conducted entirely by the authors without the use of artificial intelligence (AI) for analytical or decision-making purposes. The authors take full responsibility for the accuracy, originality, and integrity of this work.

## REFERENCES

Abdel Wahed, S., & Abdel Wahed, M. (2025). AI-Driven Digital Well-being: Developing Machine Learning Model to Predict and Mitigate Internet Addiction. *LatIA*, *3*, 73. https://doi.org/10.62486/latia202573

Ahmad, S. F., Han, H., Alam, M. M., Rehmat, Mohd. K., Irshad, M., Arraño-Muñoz, M., & Ariza-Montes, A. (2023). Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications*, *10*(1), 311. https://doi.org/10.1057/s41599-023-01787-8

Al-Areef, M. H., & Saputra, K. (2023). Analisis Sentimen Pengguna Twitter Mengenai Calon Presiden Indonesia Tahun 2024. *Jurnal SAINTIKOM*, *22*(2), 270. https://doi.org/10.53513/jis.v22i2.8680

Baria, H. G. (2025). Influence of Generative AI on Problem Solving Skills among Students. *International Journal of Scientific Research in Engineering and Management*, *9*(3), 1–9. https://doi.org/10.55041/IJSREM42014

Bin Rofi, I., Eshita, M. M., Ahmed, Md. S., & Noor, J. (2024). Identifying Influences: A Machine Learning and Explainable AI Approach to Analyzing Social Media Addiction Resulting from Academic Frustration. *Proceedings of the 11th International Conference on Networking, Systems, and Security*, 128–136. https://doi.org/10.1145/3704522.3704529

Boros, K., & Kmetty, Z. (2024). Identifying missing data handling methods with text mining. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-024-00582-1

Çela, E., Fonkam, M. M., & Potluri, R. M. (2024). Risks of AI-Assisted Learning on Student Critical Thinking: A Case Study of Albania. *International Journal of Risk and Contingency Management*, *12*(1), 1–19. https://doi.org/10.4018/ijrcm.350185

Chai, C., Liu, J., Tang, N., Li, G., & Luo, Y. (2022). Selective data acquisition in the wild for model charging. *Proc. VLDB Endow.*, *15*(7), 1466–1478. https://doi.org/10.14778/3523210.3523223

Chegg.org. (2025). *Chegg Global Student Survey 2025*. https://www.chegg.org/global-student-survey-2025

Dučić, N., Jovičić, A., Manasijević, S., Radiša, R., Ćojbašić, Ž., & Savković, B. (2020). Application of Machine Learning in the Control of Metal Melting Production Process. *Applied Sciences*, *10*(17), 6048. https://doi.org/10.3390/app10176048

Firmansyach, W. A., Hayati, U., & Wijaya, Y. A. (2023). Analisa Terjadinya Overfitting dan Underfitting. *JATI*, *7*(1), 262–269. https://doi.org/10.36040/jati.v7i1.6329

Ghosh, D., & Cabrera, J. (2022). Enriched Random Forest for High Dimensional Genomic Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(5), 2817–2828. https://doi.org/10.1109/TCBB.2021.3089417

Gu, J., & Congalton, R. G. (2025). Assessing the Impact of Mixed Pixel Proportion Training Data on SVM-Based Remote Sensing Classification: A Simulated Study. *Remote Sensing*, *17*(7), 1274. https://doi.org/10.3390/rs17071274

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(1), 28. https://doi.org/10.1186/s40537-020-00305-w

Hatamian, A., Levine, L., Oskouie, H. E., & Sarrafzadeh, M. (2025, February). *Exploring the Impact of Dataset Statistical Effect Size on Model Performance and Data Sample Size Sufficiency*. arXiv. https://doi.org/10.48550/arXiv.2501.02673

He, Y., Zhang, W., Ma, Y., Li, J., & Ma, B. (2022). The Classification of Rice Blast Resistant Seed. *Molecules*, *27*(13), 4091. https://doi.org/10.3390/molecules27134091

Holmes, M., & Theodorakopoulos, G. (2020). Towards using differentially private synthetic data for machine learning in collaborative data science projects. *Proceedings of the 15th International Conference on Availability, Reliability and Security*. https://doi.org/10.1145/3407023.3407024

Indriyani, D., & Solihati, K. D. (2021). An Overview of Indonesian's Challenging Future: Management of Artificial Intelligence in Education. *Advances in Social Science, Education and Humanities Research*. https://doi.org/10.2991/assehr.k.210629.053

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Malnad College of Engineering,Hassan, & Balgotra, A. (2025). Data Duplication Detection and Removal System Using Machine Learning. *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *09*(05), 1–9. https://doi.org/10.55041/IJSREM46920

Octaberlina, L. R., Muslimin, A. I., Chamidah, D., Surur, M., & Mustikawan, A. (2024). Exploring the impact of AI threats on originality and critical thinking in academic writing. *Edelweiss Applied Science and Technology*, *8*(6), 8805–8814. https://doi.org/10.55214/25768484.v8i6.3878

Rahardyan, T. M., Susilo, C. H., Iswara, A. M. N., & Hartono, M. L. (2024). ChatGPT: The Future Research Assistant or an Academic Fraud? [A Case Study on a State University Located in Jakarta, Indonesia]. *Asia Pacific Fraud Journal*, *9*(2), 275–293. https://doi.org/10.21532/apfjournal.v9i2.347

Rankin, D., Black, M., Bond, R., Wallace, J., Mulvenna, M., & Epelde, G. (2020). Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. *JMIR Medical Informatics*, *8*(7), e18910. https://doi.org/10.2196/18910

Sabah, A. S., Abu-Naser, S. S., Helles, Y. E., Abdallatif, R. F., Taha, A., Massa, N. M., & Hamouda, A. A. (2023). *Comparative Analysis of the Performance of Popular Sorting Algorithms on Datasets of Different Sizes and Characteristics*. *7*(6).

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/BJML/2024/007

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Thölke, P. & others. (2023). Class imbalance should not throw you off balance. *NeuroImage*, *277*, 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

Uppal, K., & Hajian, S. (2024). Students' Perceptions of ChatGPT in Higher Education: A Study of Academic Enhancement, Procrastination, and Ethical Concerns. *European Journal of Educational Research*, *14*(1), 199–211. https://doi.org/10.12973/eu-jer.14.1.199

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., & Krishna, R. (2023, January). *Explanations Can Reduce Overreliance on AI Systems During Decision-Making*. arXiv. https://doi.org/10.48550/arXiv.2212.06823

Villuendas-Rey, Y., Tusell-Rey, C. C., & Camacho-Nieto, O. (2024). Simultaneous Instance and Attribute Selection for Noise Filtering. *Applied Sciences*, *14*(18), 8459. https://doi.org/10.3390/app14188459

Vimala, S., & Sheela, D. G. A. (2025). *Predictive Modeling of the Impact of Smartphone Addiction on Students' Academic Performance Using Machine Learning*. *1*, 1–9.

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*(1), 28. https://doi.org/10.1186/s40561-024-00316-7

Zhu, T. (2020). Analysis on the Applicability of the Random Forest. *Journal of Physics: Conference Series*, *1607*(1), 012123. https://doi.org/10.1088/1742-6596/1607/1/012123