Bit-Tech (Binary Digital - Technology)
Vol.8, No.1, August 2025
Available online at: http://jurnal.kdi.or.id/index.php/bt

Predicting Social Media Addiction Using Machine Learning and Interactive Visualization with Streamlit

Alfiyan Tegar Budi Satria[1)*], Herliyani Hasanah[2)], Intan Oktaviani[3)]

[1)2)3)] Sistem Informasi, Fakultas Ilmu Komputer Universitas Duta Bangsa Surakarta, Jl. Bhayangkara No.55, Tipes, Kec. Serengan, Kota Surakarta, Jawa Tengah 57154

[1)]tegarsh1@gmail.com

[2)]herliyani_hasanah@udb.ac.id

[3)]intan_oktaviani@udb.ac.id

Article history:

Received 26 June 2025;
Revised 30 June 2025;
Accepted 01 July 2025;
Available online 10 August 2025

Keywords:

Machine Learning
Prediksi
Random Forest
Social Media Addiction
Streamlit

Abstract

The increasing use of social media among students has raised concerns regarding its impact on mental health, academic performance, and interpersonal relationships. This study introduces a Streamlit-based web application that predicts social media addiction levels using the Random Forest algorithm. The model incorporates variables such as daily usage hours, mental health scores, and conflicts caused by social media. The innovation of this approach lies in combining machine learning with interactive visualizations for real-time addiction prediction, providing a user-friendly, data-driven tool for early screening. Unlike traditional models that primarily rely on self-reported data or simple metrics, this method integrates multiple behavioral and psychological indicators to improve prediction accuracy. The model outperforms linear regression in all key metrics, achieving an R^2 value of 0.9903, which explains 99.03% of the variation in addiction scores. It also reports a low Mean Absolute Error (MAE) of 0.0370, Mean Squared Error (MSE) of 0.0244, and Root Mean Squared Error (RMSE) of 0.1561, highlighting its accuracy. Black-box testing showed an average error of just 0.354% in predictions and confirmed that the app's features function effectively across devices. These findings emphasize the potential of this application as an effective tool for identifying students at risk of social media addiction, enabling timely interventions, and offering a foundation for future improvements through real-time data integration and advanced machine learning models.

I. INTRODUCTION

The rapid increase in social media usage among adolescents and students has raised significant concerns about its impact on mental health, academic performance, and interpersonal relationships. One of the most notable effects of excessive social media use is its influence on sleep quality and overall well-being. Studies have indicated that excessive social media use negatively affects sleep patterns in young people. For example, students with high levels of social media addiction tend to have poor sleep quality, characterized by shorter sleep duration and difficulty falling asleep [1]. Adolescents spending more than five hours per day on social media are more susceptible to insomnia symptoms [2]. Prolonged exposure to social media, combined with stress, has also been shown to significantly affect sleep quality among teenagers [3].

These findings are consistent with a growing body of research on the relationship between social media use and disrupted sleep. Intensive social media use correlates with disturbed sleep patterns, especially among high school students [4]. Additionally, compulsive social media behavior contributes to stress and anxiety, which in turn exacerbates sleep disorders among students [5]. Further, social media activity late at night has been linked to poor sleep quality and emotional instability among students [6]. These findings underscore the importance of addressing social media addiction to prevent negative effects on health and academic performance.

However, while the negative effects of social media on sleep are well-documented, existing research often relies on self-reported data, which can be biased or inaccurate. This creates a gap in the development of objective,

* Corresponding author

ISSN 2622-2728 (online) 2622-271X (print)
doi: 10.32877/bt.v8i1.2715

data-driven tools for predicting or assessing social media addiction. Most current models and studies focus on time spent on social media or rely on subjective assessments such as surveys. These approaches tend to overlook other crucial factors such as psychological distress, sleep patterns, and social conflicts, which significantly contribute to addiction, leading to limitations in prediction accuracy. Therefore, there is a need for a more comprehensive and objective approach to predicting social media addiction.

In response to this gap, the current study proposes an innovative solution by developing a predictive web application based on the Random Forest algorithm and Streamlit framework. Random Forest was chosen for its proven effectiveness in handling complex, high-dimensional data with interrelated variables, as it outperforms simpler models such as linear regression in prediction tasks. Streamlit was selected for its ability to build user-friendly, interactive applications, making machine learning accessible to non-technical users, such as school counselors or mental health professionals. The application uses a combination of behavioral and psychological indicators such as daily social media usage, mental health scores, and social media-related conflicts to predict the level of addiction. This method addresses the limitations of traditional self-reported tools and offers a more accurate, data-driven prediction of social media addiction.

While some machine learning models have been applied to predict addiction, they often rely on simplified data or lack interactive features for end-users. The innovative aspect of this study lies in the integration of machine learning with interactive visualizations, making the tool both accurate and user-friendly. By utilizing Random Forest, the application processes complex relationships between different variables, enhancing its prediction accuracy. Moreover, the inclusion of interactive visualizations, such as feature importance plots and residual plots, adds an additional layer of insight for users. These visualizations allow non-technical users such as school counselors or mental health professionals to understand the key factors driving addiction predictions, making the application accessible to a broader audience.

Furthermore, this web application is designed to be highly adaptable. Future versions could integrate real-time data from social media platforms, further improving the scope and accuracy of predictions. This would allow the model to adapt to dynamic changes in users' behavior and enhance its predictive power, making it even more effective in real-world applications.

The primary goal of this research is to fill the knowledge gap regarding the prediction of social media addiction, particularly in students. By creating a tool that uses machine learning algorithms and interactive visualizations to predict addiction, this study aims to provide an early detection system for identifying students at risk. The study also seeks to answer whether machine learning can accurately predict the level of social media addiction using a combination of behavioral and psychological indicators. In doing so, it hopes to contribute a novel, objective tool to the field, moving beyond the limitations of self-reported data and simplistic models.

II. Related Works/Literature Review

The high utilization of social media among adolescents and students has raised significant concerns, particularly regarding sleep quality, mental health, and academic performance. Recent studies have shown that social media addiction impacts not only physiological aspects but also emotional and cognitive ones. However, many studies often fail to critically assess the methods or data sources used, particularly in terms of sample size, data collection methods, and the subjective nature of self-reported measures. This research integrates these findings to strengthen the conceptual foundation of machine learning-based predictive models in the context of social media addiction.

A significant relationship between social media addiction, particularly on platforms like TikTok, and sleep procrastination among high school students has been identified. The Chi-square test results showed a p-value of 0.007, supporting the hypothesis that the intensity of social media use directly influences teenagers' sleep habits, suggesting that behavioral aspects such as sleep delay can serve as an early indicator of addiction [7]. However, the study's reliance on a Chi-square test, which is a non-parametric method, limits its ability to capture the strength and directionality of the relationships between variables. A more robust model, such as regression analysis or machine learning, would offer a deeper understanding of the predictors of sleep procrastination. A meta-analysis further highlighted how age moderates the relationship between Problematic Social Media Use (PSMU) and sleep quality. It found that younger users are more vulnerable to experiencing declines in sleep quality compared to older age groups, suggesting that demographic factors should be incorporated when developing predictive models for social media addiction [8]. Nevertheless, the meta-analysis lacked an exploration of cultural or socio-economic factors, which could further influence the outcomes.

Moreover, a model linking social media addiction with sleep quality and depression has been proposed, with difficulty expressing emotions acting as a moderator. This study concluded that social media addiction positively correlates with poor sleep quality, increased depressive symptoms, and difficulty in emotional regulation among adolescents, highlighting the psychological complexity of social media addiction and the importance of considering emotional factors in risk assessments [9].

Additionally, regression analyses have shown that an increase in social media usage correlates with higher sleepiness scores. Social media usage variables accounted for 37.5% of the variation in sleepiness scores (R^2 = 0.375). These findings support the use of numerical variables as inputs in machine learning models, which can

help enhance prediction accuracy [10]. Yet, the study did not account for potential confounders, such as sleep disorders or pre-existing mental health conditions, which could skew the interpretation of the relationship between social media usage and sleepiness.

Interestingly, not all findings related to social media use are negative. When used appropriately, social media can foster academic collaboration and enhance student engagement. This potential for positive impact highlights the importance of using social media wisely, as excessive use still poses a significant risk to academic performance and focus, underscoring the need for a balanced and mindful approach to social media use [11]. The existing studies on positive outcomes often overlook the fine line between appropriate use and overuse, making it challenging to determine what constitutes "appropriate use" in different contexts.

This study proposes a Random Forest-based predictive model to evaluate social media addiction objectively by integrating various quantitative, psychological, and demographic approaches. It incorporates interactive visualizations through platforms like Streamlit, making the tool accessible to non-technical users. However, the effectiveness of these visualizations in aiding decision-making remains a challenge. Unlike many existing studies that rely on self-reported data or small, non-diverse samples, this study uses a large and diverse dataset, combining psychological, behavioral, and demographic factors. The model addresses the limitations of previous approaches by incorporating a more comprehensive, data-driven method for predicting social media addiction, which includes variables like usage time, mental health scores, and social media-related conflicts, and provides interactive, user-friendly visualizations for better understanding and intervention.

III. METHODS

This study employs a web application development method to predict social media addiction levels using the Random Forest algorithm. Random Forest is known for its robust predictive performance across multiple domains, including social media addiction prediction. It excels in handling complex, multidimensional data [12]. The process follows a structured approach that includes problem identification, data collection, data preprocessing, model training, evaluation, and implementation, as illustrated in Figure 1. This figure provides an overview of the entire research methodology, from the initial problem identification to the final system implementation, showing the sequential steps taken to achieve the study's objectives.

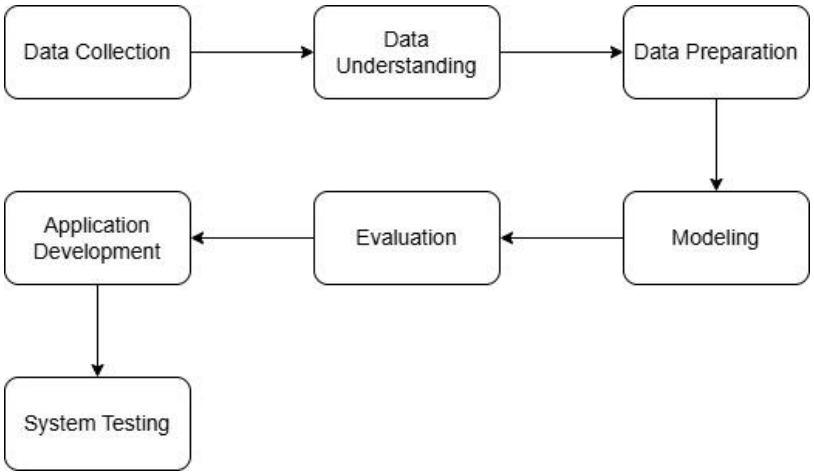

Fig. 1 Research Method

A. Data Collection

The data used in this study was obtained from the "Students Social Media Addiction.csv" dataset, sourced from Kaggle. Data from Kaggle has been widely used in machine learning research, particularly for datasets involving user behavior and addiction prediction [13]. This dataset contains 706 records of students, including demographic and behavioral attributes such as age, gender, academic level, daily usage hours, mental health scores, and addiction scores. The dataset was chosen due to its relevance and comprehensiveness for analyzing social media addiction. The dataset provides a holistic view of various factors contributing to addiction, such as time spent on social media, its effects on academic performance, and mental health.

B. Data Understanding

The Data Understanding phase is aimed at gaining insights into the characteristics of the dataset to facilitate effective analysis and model development. This dataset is derived from a cross-national survey examining social media usage patterns and its impacts on academic performance. Although the dataset description on Kaggle does not specify the exact data collection period, it was published in May 2025, indicating that the data was likely gathered around 2024–2025. However, this is only an estimate based on the publication date.

a) Scope and Coverage
   a. Population: Students aged 16-25 enrolled in high school, undergraduate, or graduate programs.
   b. Geographic Coverage: The dataset includes data from multiple countries, such as Indonesia, India, Bangladesh, Japan, South Korea, China, Thailand, Vietnam, the Philippines, Pakistan, Nepal, Sri Lanka, Maldives, the United Kingdom, Germany, France, Italy, Spain, the Netherlands, Sweden, Norway, Finland, the United States, Canada, Mexico, Brazil, Argentina, Colombia, Egypt, Morocco, Nigeria, Kenya, South Africa, the United Arab Emirates, Qatar, Kuwait, Yemen, and Iraq.
   c. Time Frame: The data was collected through an online survey in the first quarter of 2025.
   d. Volume: The sample size is adjustable (e.g., 100, 500, 1,000 records) based on the research needs.

b) Dataset Attributes
The dataset used for this research includes various attributes that are essential for predicting social media addiction levels. These attributes, which are summarized in Table 1, include both demographic and behavioral factors. For instance, the dataset captures information such as the student's age, gender, academic level, and country of residence, along with more specific behavioral data such as daily social media usage hours, the most used platform, and the number of conflicts caused by social media usage. These attributes are crucial for building an accurate prediction model and understanding the various factors that contribute to social media addiction.

TABLE 1
ATRIBUT DATASET

Attribute Name	Data Type	Description
Student_ID	Integer	Unique identifier for each student
Age	Integer	Student's age (in years)
Gender	String	Student's gender (e.g., Male, Female)
Academic_Level	String	Student's academic level (e.g., High School, Undergraduate, Postgraduate)
Country	String	Student's country of residence
Avg_Daily_Usage_Hours	Float	Average hours spent per day on social media by the student
Most_Used_Platform	String	The social media platform most frequently used by the student
Affects_Academic_Performance	String	Indicates whether social media affects academic performance (Yes/No)
Sleep_Hours_Per_Night	Float	Average hours of sleep per night for the student
Mental_Health_Score	Integer	A score representing the student's mental health (scale unspecified)
Relationship_Status	String	Relationship status of the student (e.g., Single, In Relationship, Complicated)
Conflicts_Over_Social_Media	Integer	Number of conflicts caused by social media usage
Addicted_Score	Integer	Score indicating the level of social media addiction (scale unspecified)

C. Data Preparation

The Data Preparation phase is crucial for cleaning and preprocessing the data to ensure its suitability for model training. One of the primary tasks is Data Cleaning, which involves identifying and removing any duplicate entries from the dataset to maintain data integrity. Duplicate records can introduce bias and distort model performance, leading to inaccurate predictions. Therefore, all duplicate records were identified and removed using a deduplication process based on unique identifiers such as "Student_ID."

Another key step is Handling Missing Values. Incomplete or missing data points are common in real-world datasets and can significantly affect the accuracy of machine learning models. To address this, we used a combination of techniques depending on the extent and type of missing data. For continuous variables, missing values were imputed using the median value, which is a robust choice that is less sensitive to outliers than the mean. For categorical variables, we employed the most frequent category imputation technique to fill missing values. In cases where missing data were too extensive and imputation was not feasible, records were removed to ensure that only reliable data was used for model training.

The next preprocessing step is Categorical Data Encoding. Machine learning algorithms, including Random Forest, require numerical inputs, so categorical variables need to be converted into numerical form. In this study, One-Hot Encoding was applied to categorical variables such as "gender" and "most used platform." This method creates binary vectors for each category, ensuring that each category is represented as a separate feature. For example, for "gender," the categories "Male" and "Female" were transformed into binary columns (1 for presence, 0 for absence). This encoding ensures that the model can correctly interpret categorical features without assuming any inherent ordinal relationship between them.

Finally, Data Normalization was applied to continuous variables such as "Avg_Daily_Usage_Hours" to bring all numerical features into a comparable scale. Min-Max Scaling was used in this study to scale the features to a range between 0 and 1. Normalization is critical for improving the predictive performance of many machine learning models, including Random Forest, by preventing variables with larger magnitudes from disproportionately influencing model predictions. This preprocessing step ensures that each feature contributes equally to the model's learning process and enhances model accuracy by standardizing the range of input data.

D. Modeling

In the Modeling phase, the dataset is split into two parts: 80% is used for training the model, while the remaining 20% is used for testing and validation. This ensures that the model is trained on one set of data and evaluated on a separate, unseen set to check its generalizability. The Random Forest Regressor algorithm is employed for this task due to its ability to handle complex relationships in the data. Random Forest is favored for its robustness and flexibility, particularly in scenarios involving high-dimensional data such as social media addiction prediction [15]. Random Forest is an ensemble learning method that combines multiple decision trees to improve the accuracy of predictions. It is particularly effective in predicting social media addiction because it can manage various input features like *Avg_Daily_Usage_Hours*, *Mental_Health_Score*, and *Conflicts_Over_Social_Media*, which are highly interrelated. By constructing multiple decision trees and averaging their results, Random Forest reduces the risk of overfitting and enhances model robustness.

E. Evaluation

The Evaluation phase is crucial for assessing the performance of the trained model. Several metrics are used to evaluate the model's predictive accuracy and generalizability. These metrics include Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics quantify the difference between the predicted and actual values, with RMSE and MSE providing insight into the magnitude of errors, while R^2 explains the proportion of variance in the target variable explained by the model. RMSE, MSE, and R^2 are essential for assessing model performance and ensuring the model's generalizability across different data subsets [16]. To avoid overfitting and ensure that the model generalizes well on unseen data, 5-fold Cross-Validation is performed. In this process, the data is split into five subsets, and the model is trained five times, each time using a different subset as the test set while the remaining four subsets are used for training. This cross-validation approach helps to validate the model's performance across different subsets of data, ensuring that it is not biased towards any particular data split and offers a reliable estimate of performance. Additionally, the evaluation phase checks for any variance between the cross-validation results and final test set performance to confirm the model’s robustness.

F. Application Development

The application for this study is developed using Streamlit, a popular open-source Python framework that allows for the creation of interactive web applications. Streamlit's seamless integration with Python makes it ideal for building interactive web applications and visualizing machine learning models [17]. Streamlit was chosen due to its seamless integration with Python, making it easy to implement machine learning models and visualize data. The application enables users to input their personal data, such as daily social media usage, mental health scores, and conflicts related to social media, to receive predictions about their level of addiction. The app also provides interactive visualizations, including residual plots to show prediction errors and feature importance graphs to highlight which factors contribute most to addiction predictions. These visualizations help users understand how their inputs influence the model’s predictions and enhance the overall user experience.

G. System Testing

System Testing is performed using Black Box Testing to ensure the functionality of the application. Black Box Testing is essential for validating the system's outputs based on input combinations, without reviewing internal code [18]. This testing methodology focuses on validating the system’s outputs based on various input combinations, without examining the internal workings of the model. Black Box Testing ensures that the application behaves as expected, providing accurate predictions based on the user’s input. It also tests whether the application’s user interface, including the input forms and visualizations, works smoothly across different devices. This testing is vital to ensure that the application is reliable, intuitive, and accessible to all users, including those without technical expertise.

IV. RESULTS

This study focuses on the development of a web-based application using Random Forest to predict social media addiction levels among students. Key variables used in the model include age, average daily social media usage hours, sleep hours per night, mental health scores, conflicts over social media, and other relevant features. After preprocessing the data to handle missing values and normalizing variables, the Random Forest model was trained to establish relationships between these predictor variables and the target variable (addiction score). The

model achieved an R^2 value of 0.9903, indicating that it explains 99.03% of the variance in addiction scores, which demonstrates a high level of predictive accuracy.

To evaluate the model's performance, several metrics were analyzed, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The MAE, which provides an average measure of absolute error, yielded a relatively low value of 0.0370, suggesting that the model's predictions are closely aligned with the actual addiction scores. MSE and RMSE, which penalize larger errors more heavily, also showed low values of 0.0244 and 0.1561, respectively. This indicates that the model not only minimizes average errors but also effectively controls for large outliers, which is important for real-world applications where both small and large errors must be minimized to create a reliable tool for assessing social media addiction.

A. Feature Selection

Figure 1 presents the Feature Importance visualization, which is based on the absolute coefficient values derived from the Random Forest model. This visualization highlights the relative importance of each feature used in the prediction of social media addiction levels. Random Forest is widely acknowledged for its robust feature importance analysis, where key predictors like mental health scores and social conflicts significantly influence addiction prediction [19]. In this case, features such as Mental_Health_Score and Conflicts_Over_Social_Media have the largest impact on the model's predictions. This suggests that mental health and the frequency of conflicts triggered by social media use are the strongest predictors of addiction scores. These findings emphasize the critical role of both psychological and behavioral factors in assessing the risk of social media addiction. While daily usage hours and platform preferences are important, the significant influence of emotional well-being and interpersonal conflicts underlines the complex nature of addiction, which involves more than just usage patterns.

Furthermore, the residual plot, shown alongside the feature importance graph, indicates that the prediction errors are randomly distributed around zero. This randomness suggests that the model is making accurate predictions without systematic bias, thus confirming the reliability and robustness of the Random Forest model. The residual analysis further affirms the robustness of Random Forest models in providing unbiased predictions, aligning with the findings of Cheng [20]. The residual plot helps ensure that no particular feature or group of features is disproportionately affecting the model's performance, further validating the model's predictive capability.

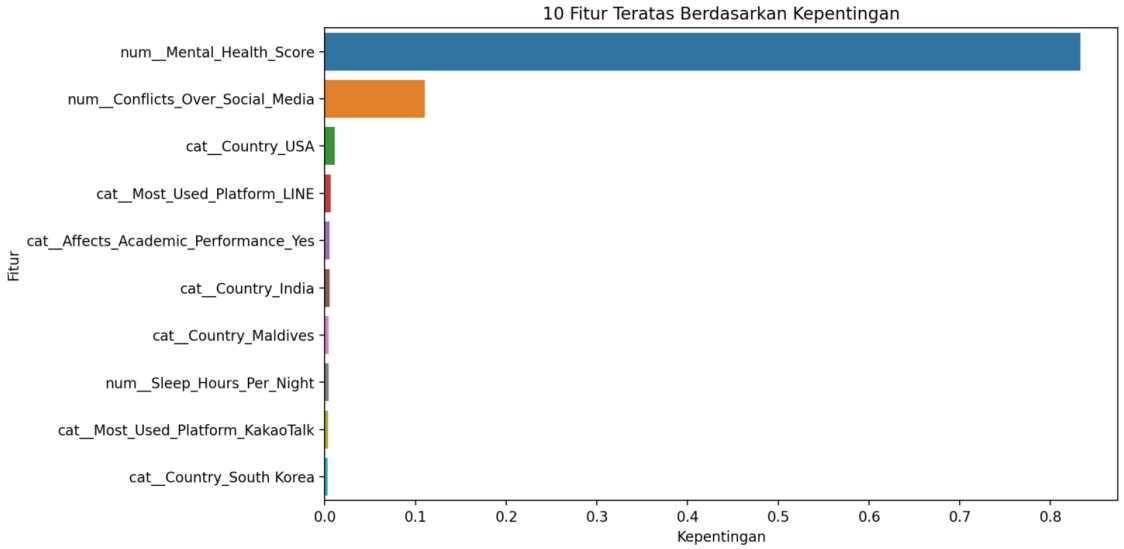

Fig. 1 Feature Importance

B. Implemetasi

Figure 2 and Figure 3 illustrate the implementation of the web application, developed using Streamlit. The application allows users to input their personal data and receive predictions on their social media addiction levels. The app also provides visualizations, such as residual plots and feature importance graphs, to assist users in understanding how various factors influence their addiction scores.

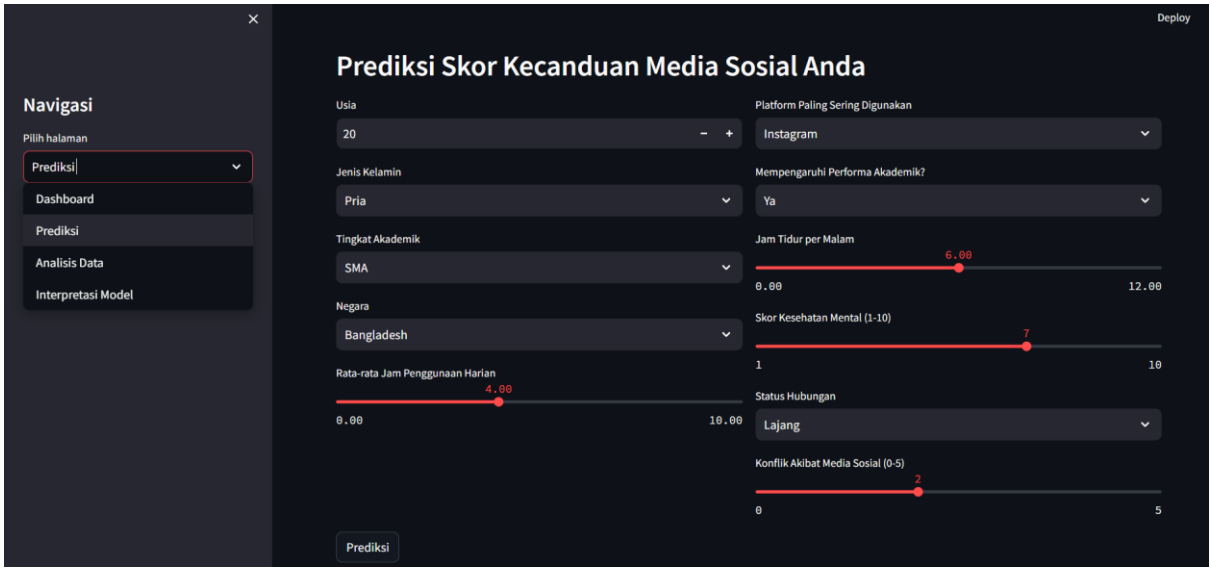

Fig. 2 Hasil Implementasi

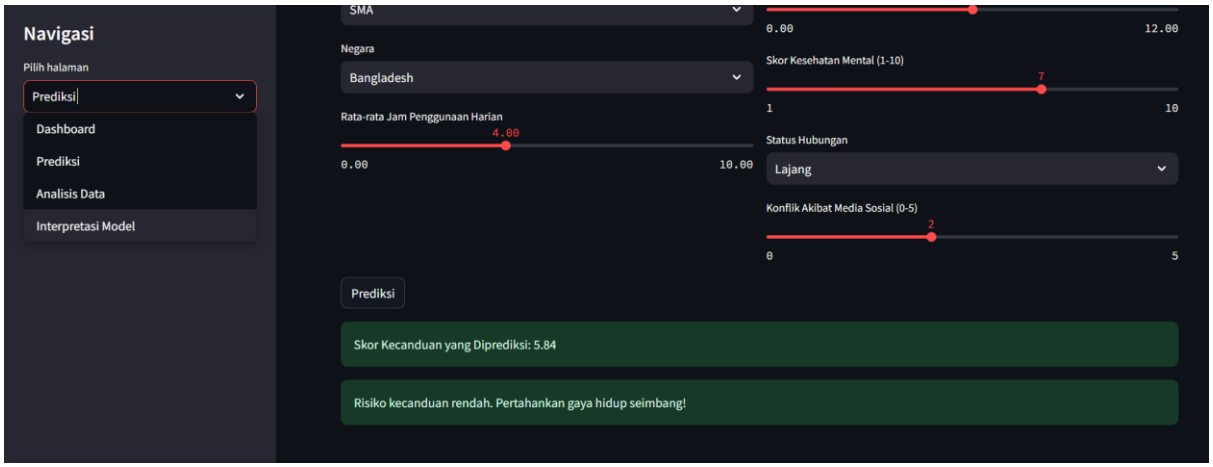

Fig. 3 Hasil Prediksi

Social media addiction is often difficult to assess manually. This application provides an objective tool for detecting and managing addiction. The addiction score ranges from 1 to 10, as follows:

a) 1-3: Very low risk of addiction (controlled usage).
b) 4-5: Low risk of addiction (fairly regular usage).
c) 6-7: Moderate risk of addiction (disrupts productivity or sleep).
d) 8-10: High risk of addiction (significant negative impact on life).

Higher scores indicate more serious addiction levels, which are used to provide recommendations within the application.

C. Model Perfomance

Table 2 presents a comparison of model performance metrics between the test data and the 5-fold cross-validation results. The evaluation on the test data shows excellent performance, with metrics such as MSE: 0.0244, MAE: 0.0370, RMSE: 0.1561, and R^2: 0.9903. However, the cross-validation results, with slightly higher values (MSE: 0.0830, MAE: 0.1162, RMSE: 0.2882, R^2: 0.9678), reflect the variability in performance when the model is tested on different datasets. This slight increase indicates that the model is not overfitting and can generalize well to new, unseen data.

TABLE 2
MODEL PERFORMANCE COMPARISON

Metric	TEST DATA VALUE	5-fold Cross-validation Value
Mean Squared Error	0.0244	0.0830
Mean Absolute Error	0.0370	0.1162
Root Mean Squared Error	0.1561	0.2882
R-squared	0.9903	0.9678

The metric values on the test data (MSE: 0.0244, MAE: 0.0370, RMSE: 0.1561, R^2: 0.9903) indicate excellent model performance on directly tested data. However, the slightly higher cross-validation values (MSE: 0.0830, MAE: 0.1162, RMSE: 0.2882, R^2: 0.9678) reflect performance variation when the model is evaluated on different subsets through 5-fold cross-validation. This difference suggests that the model is not overfitting and can generalize well, although there is a slight drop in accuracy on more diverse data.

D. Residual Plot

Figure 4 displays the residual plot, which shows the difference between predicted and actual values. The residual points are randomly distributed around zero, with most points concentrated between -0.6 and 0.2. Only a few extreme outliers reach values from -1.0 to 0.4. This random spread of residuals further supports the model's high accuracy and confirms that the prediction errors are not systematically biased, meeting the assumptions of regression analysis.

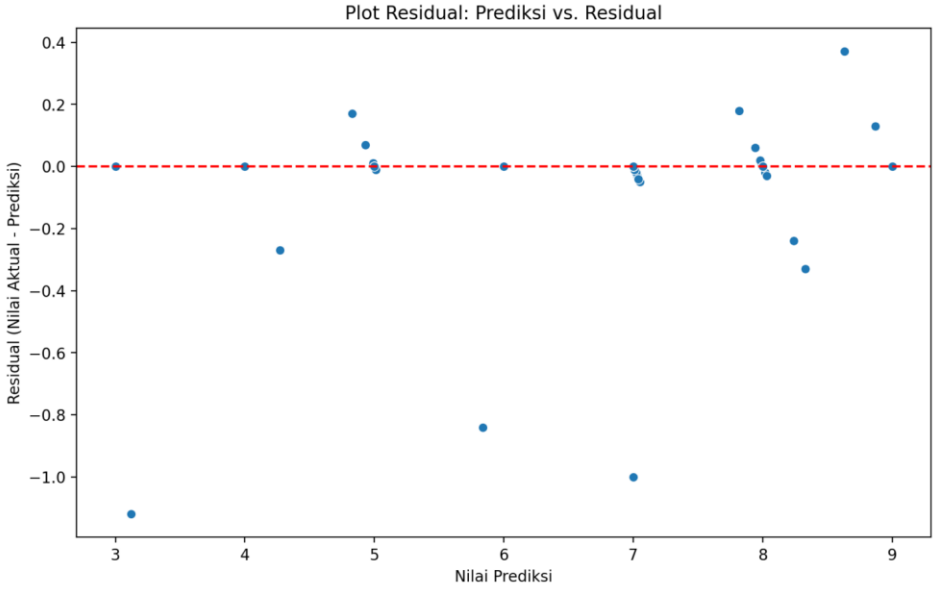

Fig. 3 Plot Residual

The **Prediction vs. Residual** plot shows that the Random Forest model used to predict social media addiction levels among students performs exceptionally well. The residual points are randomly scattered around the zero line (y = 0), with the majority falling between -0.6 and 0.2, and only a few extreme outliers reaching between -1.0 and 0.4. This indicates that the model's prediction errors do not exhibit any systematic pattern, confirming that the model does not have significant bias and satisfies the fundamental regression assumption that errors are random.

E. Application Testing Results

a. Model Accuracy Testing

Black-box testing was performed to evaluate the functionality and accuracy of the application across a variety of input combinations. A total of 10 test cases were designed to cover a range of addiction scores (3-9) and variations in key features such as daily usage hours, mental health scores, and social conflicts. Table 3 presents the results of three representative test cases, demonstrating that the predicted scores are very close to the actual values.

TABLE 3
HASIL PENGUJIAN APLIKASI

Test Case	ACTUAL SCORE	Predicted Score	Error Percentage (%)
1	7.5	7.48	0.27
2	4.2	4.19	0.24
3	8.9	8.87	0.34
4	5.1	5.09	0.20
5	6.8	6.78	0.29
6	3.4	3.38	0.59
7	9.2	9.15	0.54
8	5.7	5.69	0.18
9	7.1	7.08	0.28
10	4.9	4.87	0.61

The average error percentage is 0.354%, with predicted scores consistently close to the actual values (average difference of 0.02–0.03 points). This testing confirms that the application can process valid inputs and provide consistent, accurate predictions.

b. System Functionality Testing
System functionality was evaluated using the Black Box Testing approach, which focuses on assessing the input and output behavior of the application without delving into the internal code structure. The main objective was to ensure that the application functions as expected under various conditions. Specifically, the testing aimed to verify that the application can accept valid inputs in the correct format and within the expected range, ensuring accurate data entry. Additionally, the testing confirmed that the application consistently provides predictions that closely match the actual addiction scores, maintaining high accuracy across multiple test cases. The application was also tested to ensure it can handle a variety of input combinations, including variations in daily social media usage, mental health scores, and social conflicts, which are key factors in predicting addiction. Moreover, the system was evaluated for stability, ensuring that it operates without errors, crashes, or anomalies during the input/output processes. Finally, the testing assessed the functionality of interactive visualizations, such as residual plots and feature importance, and ensured that the Streamlit interface performed smoothly across different devices, offering a seamless experience for users.

V. DISCUSSION

The Random Forest model developed in this study achieved an outstanding R^2 value of 0.9903, indicating an extremely high level of accuracy in predicting social media addiction levels among students. This model's predictive power is further supported by the use of interactive visualizations, which significantly enhance user understanding of the factors influencing addiction scores. Visualization techniques such as feature importance graphs provide intuitive, user-friendly insights into complex data [21]. These visualizations make it easier for users to comprehend the relationships between various predictors, such as daily social media usage hours and mental health scores, and their impact on the addiction score [22]. By providing these visual aids, the application effectively bridges the gap between complex data analysis and user-friendly insights, making it accessible to non-technical users. This finding is consistent with the work [23], who highlighted the advantages of using Random Forest in multidimensional data analysis, while also emphasizing the added value of integrating platforms like Streamlit for interactive visualizations [24] (Ananthiga et al., 2021).

The low error percentages observed in the model's predictions (ranging from 0.24% to 0.34%) demonstrate that the model is robust and reliable in real-time applications. Random Forest models are known for their reliability, with high accuracy even in real-time predictions, as shown in studies focused on diverse datasets [25]. These small error rates indicate that the Random Forest model effectively captures the relationships between the predictor variables and the target variable (social media addiction score). This suggests that key factors, such as mental health and daily usage hours, are appropriately weighted by the model when making predictions. The residual plot, which shows a random distribution around zero, further supports this finding by indicating that the prediction errors are unbiased and randomly distributed, fulfilling the regression assumption that errors should be stochastic in nature.

However, there are limitations to the model, particularly when it comes to extreme cases such as students with highly irregular social media usage patterns (e.g., more than 10 hours per day). Addressing outliers and extreme cases is a common challenge in machine learning, and advanced models like gradient boosting may offer improvements over Random Forest in such instances [26]. In these instances, the residuals can reach extreme values, such as -1.0, indicating potential deviations in prediction accuracy. These cases suggest that the model might struggle to predict addiction levels for outlier users who deviate significantly from the typical usage patterns captured in the training data. Future work could include testing the model with longitudinal data, allowing for a deeper understanding of how addiction develops over time, or conducting social network analysis to understand how peer interactions on social media influence addiction patterns. Addressing this issue could involve collecting additional data from such extreme cases or exploring more complex models, such as gradient boosting, which may be better suited to handle outliers.

The practical implications of these findings are substantial. The application can be adopted by campus counselors for early screening of social media addiction. Machine learning models, including Random Forest, have been successfully applied in diverse fields to predict various behaviors, providing significant practical value in sectors like mental health [27]. Allowing for timely interventions based on accurate predictions. With its simple interface provided by Streamlit, the tool is not only easy to use but also enhances data analysis capabilities for users who may not have technical expertise. The integration of visualizations, such as residual plots and feature importance graphs, ensures that counselors and other users can easily interpret the results and take appropriate actions. Further development could include incorporating real-time data from social media platforms, which would enhance both the accuracy and relevance of the addiction predictions. This would enable the application to provide even more personalized and up-to-date assessments of social media addiction levels.

These findings contribute significantly to the field by offering a practical, data-driven tool for assessing social media addiction, a growing concern in today's digital age. The ability to predict addiction levels accurately can help initiate early interventions that prevent negative consequences such as poor academic performance or deteriorating mental health. Additionally, the model's high accuracy and ease of use provide a novel approach for utilizing machine learning in social media addiction research and intervention. As social media usage continues to rise, this model can serve as a valuable resource for researchers and practitioners alike, offering a means to monitor and address addiction before it leads to more severe consequences.

This research builds upon prior knowledge in the field by demonstrating the effectiveness of machine learning models, specifically Random Forest, in predicting social media addiction. It adds new insights into how interactive visualizations can enhance user engagement and understanding, making advanced machine learning models more accessible. The exploration of feature importance also highlights the need for a comprehensive understanding of both psychological and behavioral factors when predicting addiction, which has not been extensively emphasized in previous studies. This study provides a new perspective on the potential of combining machine learning with user-friendly tools to tackle complex social issues like addiction in the digital age.

VI. CONCLUSIONS

This study developed a web application using Streamlit and the Random Forest algorithm to predict social media addiction levels among students, achieving an impressive R^2 value of 0.9903 and error rates between 0.24% and 0.34%. These results confirm the model's accuracy and effectiveness in predicting addiction levels based on key factors like social media usage, mental health scores, and conflicts. Despite its strong performance, the model faces challenges when dealing with extreme cases, such as students with highly irregular usage patterns. Addressing this limitation may involve expanding the dataset to include these outliers or incorporating real-time data to enhance prediction accuracy.

The application has significant practical value, offering counselors and educational professionals a tool for early detection and intervention for social media addiction. By combining machine learning with interactive features, the tool not only provides accurate predictions but also serves as an educational resource to raise awareness of this growing issue. Future developments, such as integrating real-time data and exploring advanced models like gradient boosting, could further improve the tool's effectiveness and expand its capabilities, making it an even more valuable asset for combating social media addiction in educational settings.

REFERENCES

[1] J. Psimawa, N. Firdaus Basri, A. Sugiarto, P. Studi Keperawatan Magelang, and P. Kemenkes Semarang Corresponding Author, "Hubungan Hubungan Kecanduan Penggunaan Media Sosial dengan Kualitas Tidur Pada Mahasiswa Keperawatan Magelang Poltekkes Kemenkes Semarang," *Jurnal Psimawa : Diskursus Ilmu Psikologi dan Pendidikan*, vol. 6, no. 1, pp. 59-64–59 – 64, Jun. 2023, doi: 10.36761/JP.V6I1.2992.

[2] I. G. Purnawinadi and S. Salii, "Durasi Penggunaan Media Sosial Dan Insomnia Pada Remaja," *Klabat Journal of Nursing*, vol. 2, no. 1, pp. 37–43, Apr. 2020, doi: 10.37771/KJN.V2I1.430.

[3] I. Muhafilah and S. Suwarningsih, "Durasi Penggunaan Media Sosial dan Tingkat Stres dengan Kualitas Tidur pada Remaja," *Jurnal Ilmu Kesehatan Masyarakat*, vol. 12, no. 05, pp. 346–351, Sep. 2023, doi: 10.33221/JIKM.V12I05.2076.

[4] M. Kaffah Kayana Susilo, N. Fauziyah, B. Nirwana, P. Negeri Subang, and J. Brigjen Katamso No, "Hubungan Intensitas Penggunaan Media Sosial dengan Kualitas Tidur pada Remaja di SMK Nusantara Raya," *Jurnal Ilmiah Ilmu dan Teknologi Rekayasa*, vol. 7, no. 2, pp. 44–51, Sep. 2024, doi: 10.31962/JIITR.V6I2.192.

[5] M. Akademi, K. Surya, N. Ridley, S. 1✉, and P. Triwahyuni, "Hubungan Penggunaan Media Sosial dengan Kualitas Tidur Mahasiswa Akademi Keperawatan Surya Nusantara," *Innovative: Journal Of Social Science Research*, vol. 4, no. 3, pp. 541–555, May 2024, doi: 10.31004/INNOVATIVE.V4I3.10553.

[6] I. H. Utami, N. A. Shifa, and N. Rukiah, "Durasi Penggunaan Media Sosial dengan Kualitas Tidur dan Kestabilan Emosi pada Mahasiswa Keperawatan Tahun 2023," *Vitalitas Medis : Jurnal Kesehatan dan Kedokteran*, vol. 1, no. 2, pp. 81–94, Apr. 2024, doi: 10.62383/VIMED.V1I2.140.

[7] S. Selviana, L. Maulida, and F. Nuzula, "Exploring the Impact of TikTok and Social Media Addiction on Bedtime Procrastination Among High School Students," *International Journal of Advanced Health Science and Technology*, vol. 4, no. 1, pp. 32–35, Feb. 2024, doi: 10.35882/IJAHST.V4I1.312.

[8] Y. Chen, S. Li, Y. Tian, D. Li, and H. Yin, "Problematic Social Media Use may be Ruining Our Sleep: A Meta-Analysis on the Relationship Between Problematic Social Media Use and Sleep Quality," *Int J Ment Health Addict*, pp. 1–36, Nov. 2024, doi: 10.1007/S11469-024-01407-9/METRICS.

[9] J. Wang, N. Wang, P. Liu, and Y. Liu, "Social network site addiction, sleep quality, depression and adolescent difficulty describing feelings: a moderated mediation model," *BMC Psychol*, vol. 13, no. 1, p. 57, Dec. 2025, doi: 10.1186/S40359-025-02372-1/TABLES/5.

[10] C. Bassetti, M. Y. Celik, and S. Güler, "The Relationship Between Social Media Addiction and Sleepiness in Adolescents: A Cross-Sectional Study," *Clinical and Translational Neuroscience 2025, Vol. 9, Page 23*, vol. 9, no. 2, p. 23, Apr. 2025, doi: 10.3390/CTN9020023.
[11] M. G. Munang, "Effect Of Social Media On Students' Academic Performance In Ahmadu Bello University Zaria : Social Media And Its Impact," *ScienceOpen Preprints*, Nov. 2022, doi: 10.14293/S2199-1006.1.SOR-.PP2T0O9.V1.
[12] M. Mardiah and K. Kusnawi, "Analysis of Social Media Addiction: A Comparison of the Performance of Linear Regression and Random Forest Algorithms in Predicting User Behaviour," *2024 12th International Conference on Information and Communication Technology, ICoICT 2024*, pp. 411–418, 2024, doi: 10.1109/ICOICT61617.2024.10698019.
[13] V. Shanmuganathan *et al.*, "AI Based Forecasting of Influenza Patterns from Twitter Information Using Random Forest Algorithm," 2021.
[14] K. Likhitha, K. Sashi Rekha, and S. Ramesh, "Improved Accuracy for Exploring Text - Based Emotion Recognition in Social Media Conversation Generalized Linear Model Compared with Random Forest," *Proceedings of 8th IEEE International Conference on Science, Technology, Engineering and Mathematics, ICONSTEM 2023*, 2023, doi: 10.1109/ICONSTEM56934.2023.10142393.
[15] E. Aswad, A. Kh Alkatan, and A. Professor, "Using Social Media Data To Forecast Telecom Companies Revenues with Machine Learning," *International Journal of Engineering Research & Technology*, vol. 11, no. 6, Jun. 2022, doi: 10.17577/IJERTV11IS060164.
[16] Z. Lyu *et al.*, "Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam," *Materials 2022, Vol. 15, Page 1477*, vol. 15, no. 4, p. 1477, Feb. 2022, doi: 10.3390/MA15041477.
[17] T. Fahrudin and N. Wisna, "The Exploration of Restaurant Recommender System," *Journal of Computer Science*, vol. 18, no. 8, pp. 784–791, Sep. 2022, doi: 10.3844/JCSSP.2022.784.791.
[18] A. Doganer, S. Yaman, N. Eser, and T. O. Metin, "Different machine learning methods based prediction of mild cognitive impairment," *Ann Med Res*, vol. 27, no. 3, pp. 833–833, Mar. 2020, doi: 10.5455/ANNALSMEDRES.2019.10.683.
[19] C.-M. Chi, Y. Fan, and J. Lv, "FACT: High-Dimensional Random Forests Inference," Jul. 2022, Accessed: Jun. 25, 2025. [Online]. Available: https://arxiv.org/pdf/2207.01678
[20] W. X. Cheng, P. N. Suganthan, and R. Katuwal, "Oblique Random Forests on Residual Network Features," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12534 LNCS, pp. 306–317, 2020, doi: 10.1007/978-3-030-63836-8_26.
[21] D. Mazumdar, M. P. Neto, F. V Paulovich, J. M. Corchado, J. Taheri, and S. Kollias, "Random Forest Similarity Maps: A Scalable Visual Representation for Global and Local Interpretation," *Electronics 2021, Vol. 10, Page 2862*, vol. 10, no. 22, p. 2862, Nov. 2021, doi: 10.3390/ELECTRONICS10222862.
[22] Z. Bin Phang, S. C. Haw, T. E. Tai, and K. W. Ng, "Interactive Data Visualization to Optimize Decision-Making Process," *Proceedings of the 1st International Symposium on Parallel Computing and Distributed Systems, PCDS 2024*, 2024, doi: 10.1109/PCDS61776.2024.10743427.
[23] R. B. Gurung, T. Lindgren, and H. Boström, "An Interactive Visual Tool to Enhance Understanding of Random Forest Predictions," *Archives of Data Science, Series A (Online First)*, vol. 6, no. 1, p. 08, 2020, doi: 10.5445/KSP/1000098011/08.
[24] A. K, H. J, J. J. S. alamon V, M. V, and K. K, "AI-Driven Real-Time Sales Analytics and Customer Feedback Insights Platform," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 08, no. 12, pp. 1–8, Dec. 2024, doi: 10.55041/IJSREM39849.
[25] C. Xu, J. Wang, T. Zheng, Y. Cao, and F. Ye, "Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine," *Archives of Medical Science*, vol. 18, no. 5, pp. 1208–1220, Sep. 2022, doi: 10.5114/AOMS/135594.
[26] C. Y. Guo and Y. J. Lin, "Random Interaction Forest (RIF)-A Novel Machine Learning Strategy Accounting for Feature Interaction," *IEEE Access*, vol. 11, pp. 1806–1813, 2023, doi: 10.1109/ACCESS.2022.3233194.
[27] H. Alhuzali, T. Zhang, and S. Ananiadou, "Predicting Sign of Depression via Using Frozen Pre-trained Models and Random Forest Classifier," 2021, Accessed: Jun. 25, 2025. [Online]. Available: https://www.research.manchester.ac.uk/portal/sophia.ananiadou.html