Jurnal Teknik Informatika (JUTIF) DOI: https://doi.org/10.52436/1.jutif.2024.5.4.2188
Vol. 5, No. 4, August 2024, pp. 187-196 p-ISSN: 2723-3863
e-ISSN: 2723-3871

ANALYSIS OF FACTORS DETERMINING STUDENT SATISFACTION USING DECISION TREE, RANDOM FOREST, SVM, AND NEURAL NETWORKS: A COMPARATIVE STUDY

Andi Dwi Riyanto[*1], Arif Mu'amar Wahid[2], Aniec Anafisah Pratiwi[3]

[1,3]Information System, Computer Science Faculty, Universitas Amikom Purwokerto, Indonesia
[2]Magister of Computer Science, Computer Science Faculty, Universitas Amikom Purwokerto, Indonesia
Email: [1]andi@amikompurwokerto.ac.id, [2]arif@amikompurwokerto.ac.id, [3]aniecanafisah@gmail.com

(Article received: May 31, 2024; Revision: June 09, 2024; published: July 29, 2024)

Abstract

Student satisfaction is crucial in higher education, impacting student loyalty, retention rates, and institutional reputation. This study addresses the gap in applying advanced machine learning techniques to predict and understand key determinants of student satisfaction. The primary objective is to analyze and predict the factors determining student satisfaction using four machine learning models: Decision Tree, Random Forest, SVM, and Neural Networks. The dataset comprises 2527 entries with seven relevant features. Data preprocessing involved normalization and exploratory data analysis (EDA) to ensure accurate analysis. The Neural Network model achieved the highest accuracy with an MSE of 0.001399, RMSE of 0.037397, MAE of 0.030773, and R^2 of 0.998154, followed closely by the SVM model. These results suggest that advanced machine learning models, particularly Neural Networks and SVM, are effective in predicting student satisfaction and identifying key areas for improvement. This study contributes to understanding the determinants of student satisfaction using machine learning models, providing practical implications for educational administrators to develop targeted strategies to enhance student satisfaction by focusing on critical factors such as academic support and financial aid. The findings highlight the importance of using advanced predictive techniques to gain deeper insights into student satisfaction, thereby enabling institutions to implement more effective interventions. Future research should explore additional variables and more sophisticated model architectures to further improve predictive accuracy and expand the applicability of these models in educational settings.

Keywords: *decision tree, machine learning, neural networks, random forest, student satisfaction, SVM.*

ANALISIS FAKTOR PENENTU KEPUASAN MAHASISWA MENGGUNAKAN DECISION TREE, RANDOM FOREST, SVM, DAN NEURAL NETWORKS: SEBUAH STUDI KOMPARATIF

Abstrak

Kepuasan mahasiswa merupakan aspek penting dalam pendidikan tinggi yang berdampak pada loyalitas mahasiswa, tingkat retensi, dan reputasi institusi. Penelitian ini menangani kesenjangan dalam penerapan teknik pembelajaran mesin (machine learning) lanjutan untuk memprediksi dan memahami faktor-faktor penentu kepuasan mahasiswa. Tujuan utama dari penelitian ini adalah untuk menganalisis dan memprediksi faktor-faktor yang menentukan kepuasan mahasiswa menggunakan empat model pembelajaran mesin: Decision Tree, Random Forest, SVM, dan Neural Networks. Dataset yang digunakan terdiri dari 2527 entri dengan tujuh fitur yang relevan. Pra-pemrosesan data melibatkan normalisasi dan analisis data eksploratif (EDA) untuk memastikan analisis yang akurat. Model Neural Network mencapai akurasi tertinggi dengan MSE sebesar 0.001399, RMSE sebesar 0.037397, MAE sebesar 0.030773, dan nilai R^2 sebesar 0.998154, diikuti oleh model SVM. Hasil ini menunjukkan bahwa model pembelajaran mesin lanjutan, khususnya Neural Networks dan SVM, efektif dalam memprediksi kepuasan mahasiswa dan mengidentifikasi area-area yang perlu diperbaiki. Penelitian ini berkontribusi pada pemahaman faktor-faktor penentu kepuasan mahasiswa dengan menggunakan model pembelajaran mesin, memberikan implikasi praktis bagi administrator pendidikan untuk mengembangkan strategi yang tepat sasaran untuk meningkatkan kepuasan mahasiswa dengan fokus pada faktor-faktor penting seperti dukungan akademik dan bantuan keuangan. Temuan ini menyoroti pentingnya menggunakan teknik prediksi lanjutan untuk mendapatkan wawasan yang lebih mendalam tentang kepuasan mahasiswa, sehingga memungkinkan institusi untuk menerapkan intervensi yang lebih efektif. Penelitian masa depan sebaiknya mengeksplorasi variabel tambahan dan arsitektur model yang lebih canggih untuk lebih meningkatkan akurasi prediksi dan memperluas aplikasi model-model ini dalam lingkungan pendidikan.

Kata kunci: *decision tree, kepuasan mahasiswa, neural network, pembelajaran mesin, random forest, SVM.*

1. INTRODUCTION

1.1. Broad Overview

Student satisfaction plays a crucial role in higher education institutions, influencing student loyalty, attracting and retaining students, and indicating the quality of education [1],[2],[3]. High levels of satisfaction foster active engagement, participation in campus activities, and a vibrant academic community, which are essential for personal and intellectual growth.

Satisfaction is directly linked to education quality, affected by lecturer competence, teaching techniques, and service quality [4]. Monitoring student satisfaction globally helps meet students' needs and improve service quality [5]. prompting universities to strive for educational excellence to enhance experiences and foster loyalty [6]. Satisfied students engage more deeply in their studies, attend classes regularly, participate actively in discussions, and seek additional learning opportunities, leading to better academic outcomes and deeper understanding of subjects.

Student satisfaction also significantly impacts retention rates. Satisfied students are more likely to continue their studies at the same institution, reducing dropout rates and ensuring they achieve academic goals. They typically feel a stronger connection to their institution, perceiving the environment as supportive and conducive to growth. This sense of belonging motivates them to persist through challenges. In contrast, dissatisfied students may feel disengaged and unsupported, leading them to consider transferring or leaving higher education.

Furthermore, student satisfaction profoundly affects the reputation of educational institutions. It is closely linked to the image and reputation of universities, with satisfied students contributing positively. These students act as advocates, enhancing the institution's reputation through loyalty. High student satisfaction often equates to quality education, strong support systems, and a positive campus environment, which are critical for long-term success. Institutions with high satisfaction levels tend to develop a positive reputation, attracting prospective students. Satisfied students are likely to share positive experiences with peers, family, and online, enhancing the institution's image and appeal through effective word-of-mouth marketing.

1.2. Contextual Background

Student satisfaction is influenced by multiple factors within the educational environment [7], [8], [9], [10]. Understanding these factors is crucial for institutions aiming to enhance the overall student experience. Academic support services, campus facilities, extracurricular activities, and financial support are crucial factors that contribute to student satisfaction in higher education institutions.

Academic support services like tutoring, advising, and mentoring aid in concept understanding, provide guidance, and foster a supportive learning environment. Well-equipped campus facilities, including libraries, laboratories, housing, and recreational amenities, cater to students' academic, residential, and recreational needs, enhancing their overall experience. Extracurricular activities and student organizations promote personal growth, leadership development, social connections, and a sense of belonging, leading to a well-rounded and fulfilling university experience. Financial support, through scholarships, grants, and affordable tuition fees, alleviates economic burdens, reduces stress, and allows students to focus on their studies, ultimately contributing to their satisfaction.

These factors collectively create an environment conducive to academic success, personal growth, and overall student well-being, thereby significantly influencing student satisfaction levels within higher education institutions.

1.3. Machine Learning in Educational Research

Machine learning (ML) is increasingly used in educational research for its ability to handle large datasets, identify complex patterns, and make accurate predictions. ML algorithms can process diverse data, such as student demographics, academic performance, and engagement metrics, uncovering insights often missed by traditional methods.

The Decision Tree model is a widely used machine learning algorithm known for its simplicity and interpretability. Decision trees offer several benefits, including ease of interpretation, handling of nominal and categorical data, and the ability to provide logic-based and easily understandable outcomes [11], [12]. They are considered one of the simplest and oldest machine learning methods, yet they can be powerful, especially when used in ensemble methods to improve accuracy [13]. Decision trees are widely used in various fields such as healthcare for predicting outcomes like massive bleeding in surgeries [14], financial risk prediction [15], and even in detecting attacks in Internet of Things systems [16].

Random Forest is an ensemble learning technique that builds multiple decision trees during training and outputs the average prediction of the individual trees. It offers advantages over individual decision trees, including improved performance, avoidance of overfitting, and the ability to combine multiple decision trees into a final output. [17]. This ensemble method, which uses decision trees as base classifiers, is recognized for its high accuracy and efficiency, making it a popular choice in various applications such as stroke classification and Parkinson's disease prediction [18]. By creating decision trees independently and then combining them, Random Forests generate a robust learner that excels in handling large datasets and complex problems [19]. Additionally, they are effective in scenarios with imbalanced data, as demonstrated by the Balanced Random Forest approach [20].

SVM is a powerful supervised learning algorithm for classification and regression tasks. It finds the optimal hyperplane that separates data points of different classes with maximum margin. SVM advantages include handling high-dimensional data, small sample sizes, and nonlinear patterns effectively [21]. SVMs are known for their capability to learn non-linear separations through appropriate kernel selection, ensuring a global maximum [22]. They provide accurate estimations even in the presence of noise and nonlinearities, making them suitable for modeling complex systems [23]. Additionally, SVMs exhibit excellent generalization, abstraction properties, and immunity to overtraining, addressing challenges like the curse of dimensionality [24].

Neural Networks, particularly deep learning models, have gained popularity due to their ability to model complex non-linear relationships. They provide benefits such as handling high-dimensional data, automatically extracting essential features, and effectively modeling complex nonlinear patterns. [25]. They have been successfully utilized in diverse fields like medical image recognition, fault diagnosis, and structural engineering due to their adaptability, fault tolerance, and superior performance compared to traditional methods [26].

1.4. Problem Statement and Significance

Despite extensive research on student satisfaction, there remains a gap in applying advanced machine learning techniques for analyzing and predicting the factors influencing this critical metric. Previous studies have predominantly relied on traditional statistical methods, which may not fully capture the complex, non-linear relationships between various determinants of student satisfaction. Additionally, existing research often overlooks the comparative performance of different machine learning algorithms in this context. There are also limitations related to the scope and granularity of data used in prior studies, failing to encompass the full spectrum of factors influencing student satisfaction. This research aims to address these gaps by leveraging advanced machine learning models, such as Decision Trees, Random Forests, Support Vector Machines (SVMs), and Neural Networks, to provide a comprehensive comparative analysis and identify the most effective models for predicting student satisfaction.

The study seeks to achieve several specific goals:

1. To identify and quantify the key factors that influence student satisfaction using machine learning techniques.
2. To compare the predictive accuracy and effectiveness of different machine learning models in this context.

By employing these advanced techniques, this research aims to overcome the limitations of previous studies and offer a more robust and comprehensive understanding of the factors driving student satisfaction. The comparative analysis of Decision Tree, Random Forest, SVM, and Neural Networks will not only highlight the strengths and weaknesses of each model but also guide future research and practical applications in educational settings.

This study bridges a gap in existing educational research by applying advanced machine learning models to analyze student satisfaction, capturing intricate relationships between influencing factors that traditional studies often fail to identify. The findings will contribute to the academic literature by demonstrating the utility of these models in educational settings and offering new perspectives on data-driven approaches to understanding student satisfaction. The results have practical implications, enabling institutions to develop targeted strategies by identifying significant factors affecting satisfaction. Additionally, the comparative analysis of different machine learning models will provide insights into the most effective tools for educational data analysis, guiding institutions in choosing the right techniques for their needs.

2. METHODS

The methodology for this study is illustrated in a flowchart that outlines the entire process from data collection and preprocessing to model implementation and evaluation, as shown in figure 1 below.

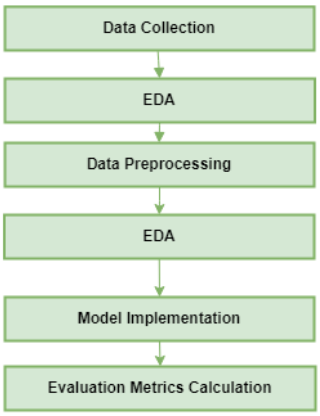

Figure 1. Research Method Flowchart

2.1. Data Collection

The dataset used in this study was obtained from the university's student feedback database, specifically collected for the 2024 Student Satisfaction Evaluation. Every student was required to participate in the evaluation, which was conducted through the Student Portal (student. amikompurwokerto.ac.id). The evaluation period started on Monday, January 8, 2024, and ended on Saturday, January 20, 2024, before students could print their UAS Gasal Academic Year 2023/2024 exam cards.

To ensure the objectivity of the evaluations, the following guidelines were implemented: the identities of the students providing the evaluations were not included. The data was collected from students across all programs in the Faculty of Computer Science and the Faculty of Business and Social Sciences.

The evaluation process was conducted by the Educational Development and Quality Assurance Institute (Lembaga Pengembangan Pendidikan dan Penjaminan Mutu, LP3M). The data collection involved structured surveys administered through the portal, ensuring standardized responses. Initial preprocessing by the data provider included anonymizing student records to maintain confidentiality and standardizing responses to facilitate accurate analysis.

The dataset comprises 2527 entries, each representing an individual student's feedback on various aspects of their educational experience. The features included in the dataset are Tata Pamong (Governance quality and administrative support), Kerjasama (Collaboration opportunities and partnership effectiveness), Sarpras (Infrastructure and facility adequacy), Keuangan (Financial aid availability and financial management), Pembelajaran (Quality of teaching and learning experiences), and Kemahasiswaan (Student activities and engagement).

These features were chosen for their potential impact on student satisfaction. The target variable is Kepuasan, which measures the overall satisfaction level of the students. Understanding the significance of each feature in the context of the study helps in interpreting the results and drawing meaningful conclusions from the analysis.

2.2. Initial Exploratory Data Analysis (EDA)

The EDA was conducted to gain a comprehensive understanding of the raw dataset, identify missing values, duplicates, and outliers. The dataset consists of 2527 entries with 7 columns, each representing a different feature relevant to the study of student satisfaction. The features included are \`Tata Pamong\`, \`Kerjasama\`, \`Sarpras\`, \`Keuangan\`, \`Pembelajaran\`, \`Kemahasiswaan\`, and the target variable \`Kepuasan\`.

Each of these columns contains numerical data of type \`float64\`, and there are no missing values in any of the columns, indicating the completeness of the dataset. Visualizations were utilized to explore data distributions and relationships between variables. Histograms were created for each feature to show the distribution of values, evealing potential patterns and skewness in the data. These steps were crucial in preparing the data for further analysis.

2.3. Data Preprocessing

The data preprocessing phase began with thorough data cleaning to ensure the dataset was ready for analysis. Given the results of the initial EDA, it was confirmed that there were no missing values in any of the columns, eliminating the need for imputation. The dataset contained 2527 complete entries, with no null values detected. Additionally, no duplicate entries were found, so no further actions were required to remove duplicates.

Normalization of data was conducted to ensure that all features were on a similar scale, which is crucial for the performance of machine learning models. Each feature was normalized using standard scaling. This involved subtracting the mean of the feature from each value and dividing the result by the standard deviation. This step transformed the data into a standard normal distribution, with a mean of 0 and a standard deviation of 1. This normalization process helps improve the convergence of gradient-based learning algorithms and ensures that each feature contributes equally to the analysis.

Feature extraction and selection were critical steps in enhancing the performance of the machine learning models. From the initial EDA, the dataset included the following features: Tata Pamong, Kerjasama, Sarpras, Keuangan, Pembelajaran, and Kemahasiswaan. Each feature was selected based on its potential impact on the target variable Kepuasan, which measures overall student satisfaction.

To ensure the models could effectively learn from the data, no additional feature extraction was required beyond the initial features provided. Each feature was retained, given their relevance and the insights gained from the descriptive statistics and visualizations.

2.4. EDA - Post-Cleaning

After completing the data cleaning process, a second round of EDA was conducted to verify the effectiveness of the cleaning steps. The goal was to ensure that the data was properly prepared for modeling and that any inconsistencies, outliers, or other issues had been addressed. Given that the initial EDA revealed no missing values and no duplicates, the focus was primarily on confirming the normalization process and examining the refined dataset.

2.5. Model Implementation

To analyze the factors determining student satisfaction, four machine learning algorithms were selected for comparison: Decision Tree, Random Forest, SVM, and Neural Networks. These algorithms were chosen due to their diverse approaches to learning from data, allowing for a comprehensive comparative study of their performance in predicting student satisfaction.

For the Decision Tree model, the maximum depth (max_depth) was set to 5 to prevent overfitting and ensure generalization. The Random Forest model involved tuning the number of trees (n_estimators) to 100 and setting the maximum depth (max_depth) to 5, balancing between bias and variance. For the SVM, the radial basis function (RBF) kernel was selected, and the hyperparameters C (regularization parameter) and epsilon (epsilon in the epsilon-SVR model) were set to 1.0 and 0.1, respectively, to control the margin and error tolerance. The Neural Network model's architecture included two hidden layers with 64 and 32 neurons, respectively, both using the ReLU activation function. The optimizer chosen was Adam, and the model was compiled with a mean squared error (MSE) loss function.

The dataset was divided into training and testing sets using an 80-20 split ratio. This means that 80% of the data was used for training the models, and the remaining 20% was reserved for testing. This split ensures that the models are trained on a substantial portion of the data while having enough unseen data to evaluate their performance.

Each model was trained using the training dataset, employing specific libraries and tools to facilitate the process. The Decision Tree model was implemented using the DecisionTreeRegressor from the scikit-learn library. The model was trained by fitting it to the training data and then used to make predictions on the test data. The Random Forest model was implemented using the RandomForestRegressor from scikit-learn. This ensemble method involved training multiple decision trees and averaging their predictions to improve accuracy and reduce overfitting. The SVM model was implemented using the SVR class from scikit-learn. The RBF kernel was used, and the model was trained to find the optimal hyperplane that minimizes error while allowing some flexibility with the epsilon parameter. The Neural Network model was implemented using the Sequential API from TensorFlow's Keras library. The network architecture consisted of two hidden layers with ReLU activations, and the model was compiled with the Adam optimizer and MSE loss function. The training involved 50 epochs with a batch size of 10, and the model's performance was validated on the test data.

2.6. Evaluation Metrics

To evaluate the performance of the models in predicting student satisfaction, several metrics were employed. Mean Squared Error (MSE) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. Mathematical formula for MSE is.

$$MSE = \frac{1}{n}\sum_{1}^{n^2}{}_i (y_i - \hat{y}_i)^2 \tag{1}$$

A lower MSE indicates better model performance.

Root Mean Squared Error (RMSE) is the square root of the MSE, bringing the metric back to the original scale of the data, making it easier to interpret. Like MSE, lower values indicate better performance. Mathematical formula for RMSE is.

$$RMSE = \sqrt{MSE} \tag{2}$$

Mean Absolute Error (MAE) measures the average absolute errors between the predicted and actual values, providing a straightforward interpretation of prediction accuracy, with lower values indicating better performance. Mathematical formula for MAE is.

$$MAE = \frac{1}{n}\sum_{1}^{n^2}{}_i (|y_i - \hat{y}_i|) \tag{3}$$

R-squared (R^2) represents the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 value closer to 1 indicates that the model explains a large portion of the variance in the target variable. Mathematical formula for R^2 is.

$$R^2 = 1 - \frac{\Sigma\left(y_i - y\ _i\right)^2}{\Sigma\left(y_i - \bar{y}\right)^2} \tag{4}$$

These metrics provide a comprehensive assessment of model performance, considering both the accuracy and the goodness-of-fit of the models.

3. RESULTS

3.1. Initial EDA

The initial EDA as described in section 3.2 provided a comprehensive understanding of the raw dataset. The dataset consisted of 2527 entries with 7 columns: Tata Pamong, Kerjasama, Sarpras, Keuangan, Pembelajaran, Kemahasiswaan, and the target variable Kepuasan. Each column contained numerical data of type float64, and no missing values were detected. The descriptive statistics for the raw data revealed the following. For `Tata Pamong`, the mean score was 2.69 with a standard deviation of 0.93, and scores ranged from 0 to 4. The `Kerjasama` feature had a mean score of 2.70 and a standard deviation of 0.98, also ranging from 0 to 4. The `Sarpras` feature showed a mean score of 2.48 and a standard deviation of 0.99, with scores between 0 and 4. For `Keuangan`, the mean was 2.59 with a standard deviation of 1.02, similarly ranging from 0 to 4. The `Pembelajaran` feature had a mean of 2.64 and a standard deviation of 0.99, again with scores from 0 to 4. The `Kemahasiswaan` feature had a lower mean score of 2.24 and a higher standard deviation of 1.22, indicating more variability in the responses, with scores ranging from 0 to 4. Finally, the target variable `Kepuasan` had a mean score of 2.56 and a standard deviation of 0.88, with values also spanning from 0 to 4.

Histograms were generated to visualize the distribution of each feature, as shown in Figure 2. These visualizations revealed that most features had a roughly normal distribution with some skewness.

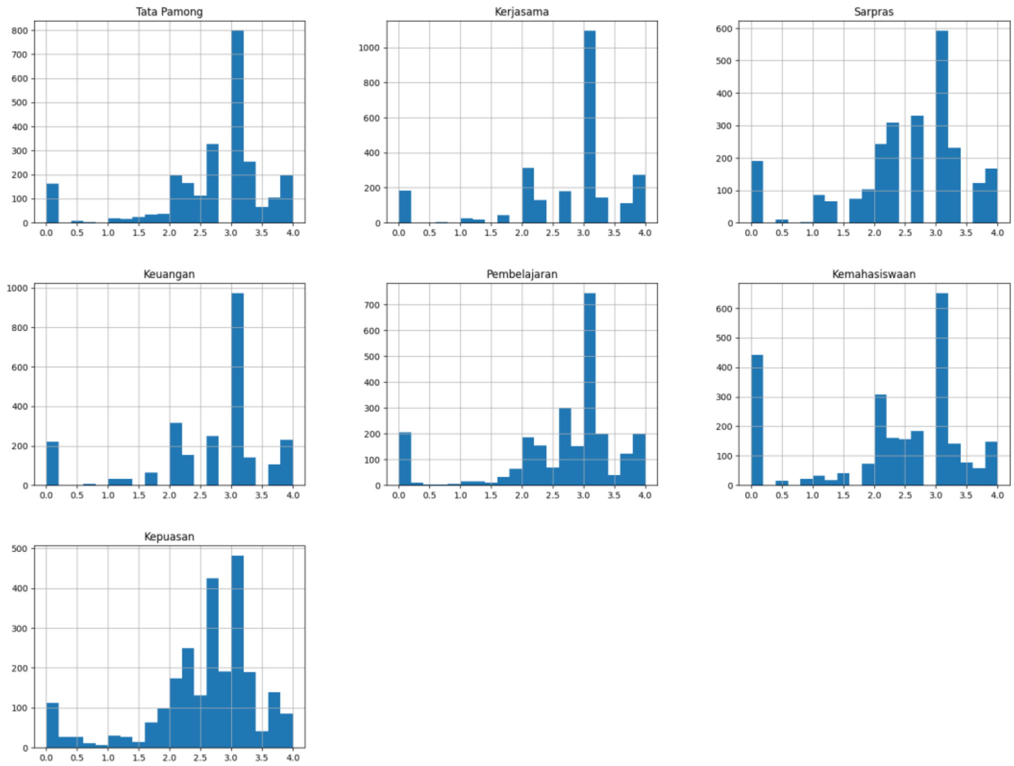

Figure 2. Histogram of Data Variable

3.2. Post-Cleaning EDA

The post-cleaning EDA as described in section 3.4 confirmed the effectiveness of the data cleaning steps. The updated summary statistics indicated successful normalization, with each feature having a mean close to 0 and a standard deviation close to 1.

The descriptive statistics for the cleaned data are as follows. For Tata Pamong, the count was 2527, the mean was 0.000000, and the standard deviation was 1.000000. The minimum value was -2.890452, the 25th percentile was -0.418778, the median (50th percentile) was 0.333309, the 75th percentile was 0.332558, and the maximum value was 1.400526.

For Kerjasama, the count was 2527, the mean was 0.000000, and the standard deviation was 1.000000. The minimum value was -2.751835, the 25th percentile was -0.403554, the median was 0.311854, the 75th percentile was 0.311604, and the maximum value was 1.345704. For Sarpras, the count was 2527, the mean was 0.000000, and the standard deviation was 1.000000. The minimum value was -2.475188, the 25th percentile was -0.475188, the median was 0.324812, the 75th percentile was 0.324812, and the maximum value was 1.524812.

For Keuangan, the count was 2527, the mean was 0.000000, and the standard deviation was 1.000000. The minimum value was -2.528985, the 25th percentile was -0.575648, the median was 0.400368, the 75th percentile was 0.400368, and the maximum value was 1.447315. For Pembelajaran, the count was 2527, the mean was 0.000000, and the standard deviation was 1.000000. The minimum value was -2.664610, the 25th percentile was -0.344609, the median was 0.362391, the 75th percentile was 0.389539, and the maximum value was 1.364539.

For Kemahasiswaan, the count was 2527, the mean was 0.000000, and the standard deviation was 1.000000. The minimum value was -2.243490, the 25th percentile was -0.444222, the median was -0.000490, the 75th percentile was 0.618065, and the maximum value was 1.756510. For the target variable Kepuasan, the count was 2527, the mean was 0.000000, and the standard deviation was 1.000000. The minimum value was -2.556470, the 25th percentile was -0.406470, the median was 0.243530, the 75th percentile was 0.443530, and the maximum value was 1.443530.

The correlation matrix, illustrated in Figure 3, highlights the relationships between different features post-cleaning. Strong correlations were observed between several pairs of features, such as Tata Pamong and Kerjasama (0.89), Keuangan and Pembelajaran (0.87), and Sarpras and Keuangan (0.79). These strong correlations suggest that these pairs of features often move together, providing important insights into the underlying structure of the data.

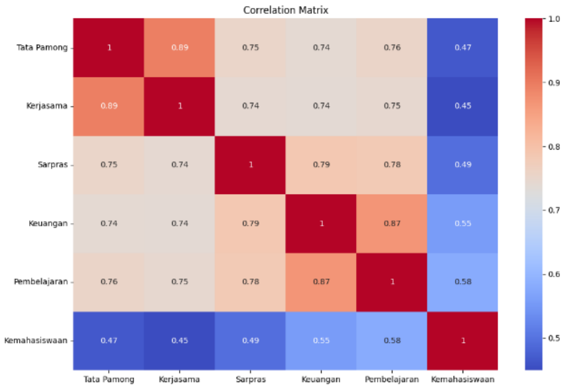

Figure 3. Correlation Matrix between Variable

The heatmap visualization effectively demonstrates the strength of these correlations, using a gradient color scheme where darker shades represent stronger correlations. This visualization aids in quickly identifying which features are most closely related, valuable for feature selection and understanding multicollinearity in the dataset.

3.3. Model Implementation Results

The performance metrics for each algorithm were evaluated using MSE, RMSE, MAE, and R². These metrics provide a comprehensive understanding of each model's accuracy and predictive power. The results are summarized in the table below.

Table 1. Result of Performance Metrics

Model	MSE	RMSE	MAE	R²
Decision Tree	0.051986	0.228005	0.170473	0.931365
Random Forest	0.017579	0.132584	0.097478	0.976792
SVM	0.002688	0.051850	0.038221	0.996451
Neural Network	0.001399	0.037397	0.030773	0.998154

These metrics highlight the varying degrees of accuracy and predictive performance across different models. The detailed results include various visualizations that provide further insights into the performance of each model as shown in figure 4.

Figure 5 illustrates feature importance plots from the Random Forest and Decision Tree models, highlighting the variables that significantly influence student satisfaction. The Random Forest model's plot shows that `Pembelajaran` and `Keuangan` are the top contributing factors to the model's predictions.

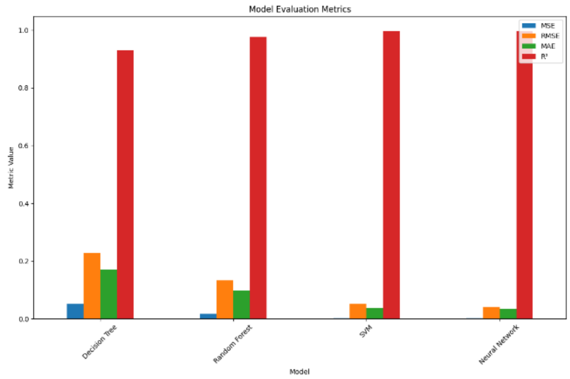

Figure 4. Comparison of Performance Metrics

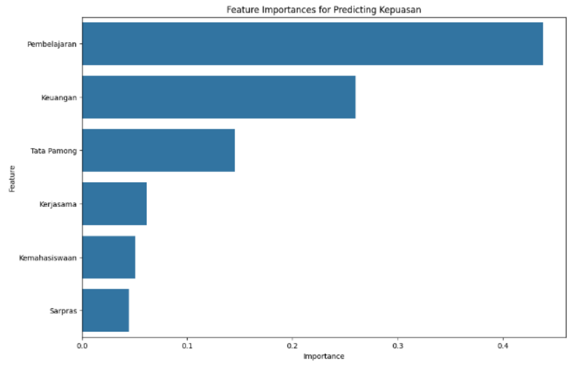

Figure 5. Feature Importance Plot

Comparing the performance of the algorithms based on the evaluation metrics reveals that the Neural Network model demonstrated the highest predictive accuracy with the lowest MSE, RMSE, and MAE, and the highest R² value. The SVM model followed closely, also showing high predictive accuracy. The Random Forest model performed well, offering a good balance between accuracy and computational efficiency. Despite having the lowest accuracy among the models, the Decision Tree model was the most computationally efficient, making it suitable for scenarios requiring quick predictions with less computational overhead.

The detailed analysis and visualizations provide a clear comparison of the models' performance. The Neural Network and SVM models are the best-performing in terms of accuracy, while the Random Forest model offers a balanced approach, and the Decision Tree model excels in computational efficiency. This comprehensive evaluation allows for informed decision-making based on the specific requirements of the study, whether prioritizing accuracy, efficiency, or a balance of both.

4. DISCUSSION

The results of this study highlight significant variations in the performance of different machine learning models in predicting student satisfaction. The Neural Network model emerged as the most accurate, achieving the lowest MSE and MAE, as well as the highest R^2 value. This superior performance can be attributed to the model's ability to capture complex non-linear relationships within the data. Neural Networks, with their layered architecture, are capable of learning intricate patterns and interactions between features, which simpler models might overlook.

The SVM also performed exceptionally well, demonstrating the second-highest accuracy. The use of the radial basis function (RBF) kernel in SVM allows it to handle non-linear relationships effectively, similar to Neural Networks, but with a different approach. SVM's performance underscores its robustness in handling high-dimensional spaces and capturing the underlying structure of the data.

The Random Forest model showed a good balance between accuracy and interpretability. Its ensemble nature, which combines multiple decision trees, enhances its ability to generalize and reduce overfitting. The feature importance analysis from the Random Forest model revealed that Pembelajaran (learning experience) and Keuangan (financial support) were significant predictors of student satisfaction. This insight aligns with educational theory, which emphasizes the importance of quality teaching and adequate financial resources for student success.

On the other hand, the Decision Tree model, while the least accurate among the four, was the most computationally efficient. Its simplicity and speed make it an attractive option for applications where quick, interpretable results are necessary, despite the trade-off in accuracy. The relatively lower performance of the Decision Tree model is due to its tendency to overfit the training data, a limitation that ensemble methods like Random Forest aim to mitigate.

When comparing these findings with similar studies, it becomes evident that the application of advanced machine learning models, such as Neural Networks and SVM, consistently shows higher predictive accuracy in educational research. For instance, previous studies have also highlighted the effectiveness of Neural Networks in capturing complex, non-linear relationships in educational data, leading to better performance compared to traditional models like Decision Trees. Additionally, the robust performance of SVM aligns with existing literature, which often cites its capability to handle high-dimensional data and non-linear relationships effectively.

The findings of this study have several practical implications for educational institutions aiming to enhance student satisfaction. The high accuracy of the Neural Network and SVM models suggests that these techniques can be effectively used to identify key areas for improvement in the student experience. For instance, by analyzing which factors most significantly influence satisfaction, administrators can prioritize interventions that address these areas. The insights from the Random Forest model indicate that improving the quality of learning experiences and providing better financial support can lead to higher student satisfaction.

Educational institutions can apply these models to predict and monitor student satisfaction in real-time, allowing for proactive measures to address potential issues. For example, regular surveys could be conducted to gather data on student satisfaction, and the models could analyze this data to identify trends and areas needing attention. By focusing resources on the most influential factors, institutions can make data-driven decisions to enhance the overall student experience.

While the study provides valuable insights, several limitations must be acknowledged. First, the quality and scope of the data can significantly impact the results. The dataset used in this study, while comprehensive, may not capture all the factors influencing student satisfaction. Variables such as extracurricular activities, campus culture, and personal circumstances were not included but could be important.

Additionally, model assumptions and configurations could affect the outcomes. For instance, the chosen hyperparameters for each model might not be optimal, and different tuning strategies could yield better results. The complexity of Neural Networks, while offering high accuracy, also makes them more prone to overfitting, especially with smaller datasets.

Furthermore, the study's scope is limited to the selected algorithms. Other advanced techniques, such as gradient boosting machines (e.g., XGBoost) or deep learning models with more layers, could potentially offer even better performance.

Future research could explore several avenues to build upon the findings of this study. One potential direction is to expand the dataset to include additional variables that might influence student satisfaction. This could involve incorporating qualitative data from student interviews or focus groups to capture a broader range of factors.

Another area for improvement is the exploration of more sophisticated model architectures. For example, using deeper Neural Networks with more layers or experimenting with different types of SVM kernels could provide further insights. Additionally, implementing ensemble techniques like stacking, which combines multiple models to leverage their strengths, could enhance predictive accuracy.

Hyperparameter tuning can also be refined using more advanced techniques such as Bayesian optimization or genetic algorithms. These methods can help identify the optimal configurations for each model, potentially improving performance.

Finally, longitudinal studies that track student satisfaction over time could provide dynamic insights and help in understanding how various factors influence satisfaction throughout a student's academic journey. This temporal aspect could be incorporated using time-series analysis or recurrent neural networks (RNNs).

By addressing these limitations and exploring these future directions, researchers can continue to enhance the understanding and prediction of student satisfaction, ultimately contributing to better educational outcomes.

5. CONCLUSION

This research analyzed student satisfaction using four machine learning models: Decision Tree, Random Forest, SVM, and Neural Networks. The Neural Network model was the most accurate, achieving the lowest MSE, MAE, and the highest R^2 value, due to its ability to capture complex non-linear relationships. The SVM model also demonstrated high accuracy with its RBF kernel. The Random Forest model offered a balanced approach, revealing that Pembelajaran (learning experience) and Keuangan (financial support) were significant predictors of student satisfaction. Although the Decision Tree model was the least accurate, it was the most computationally efficient, making it suitable for quick, interpretable results.

The practical implications suggest that Neural Network and SVM models are effective for identifying key areas for improving student satisfaction. Educational administrators can use these models to prioritize interventions, focusing on the most influential factors identified by the Random Forest model, such as enhancing learning experiences and providing better financial support.

However, the study acknowledges limitations, including data quality, model assumptions, and selected hyperparameters. Future research should explore additional variables, sophisticated model architectures, and advanced hyperparameter tuning techniques to improve predictive accuracy. Expanding the dataset to include qualitative data and exploring ensemble techniques like stacking, as well as conducting longitudinal studies, could provide deeper insights into student satisfaction. By addressing these limitations, researchers can enhance the understanding and prediction of student satisfaction, contributing to better educational outcomes.

REFERENCES

[1] M. Bakrie, B. Sujanto, and R. Rugaiyah, "The Influence of Service Quality, Institutional Reputation, Students' Satisfaction on Students' Loyalty in Higher Education Institution," *Int. J. Educ. Vocat. Stud.*, vol. 1, no. 5, 2019, doi: 10.29103/ijevs.v1i5.1615.

[2] D. Meštrović, "Service Quality, Students' Satisfaction and Behavioural Intentions in STEM and IC Higher Education Institutions," *Interdiscip. Descr. Complex Syst.*, vol. 15, no. 1, pp. 66–77, 2017, doi: 10.7906/indecs.15.1.5.

[3] F. M. Roget, P. F. Esparís, and E. V. Rozas, "University Student Satisfaction and Skill Acquisition: Evidence From the Undergraduate Dissertation," *Educ. Sci.*, vol. 10, no. 2, p. 29, 2020, doi: 10.3390/educsci10020029.

[4] Tg Nur-Wina Tuan Abu Bakar, "Lecturer Competence and Student Satisfaction: A Roadmap for Quality Education," *Int. J. Acad. Res. Progress. Educ. Dev.*, vol. 12, no. 4, 2023, doi: 10.6007/ijarped/v12-i4/18970.

[5] K. Fuchs and K. Fangpong, "Using the SERVQUAL Framework to Examine the Service Quality in Higher Education in Thailand," *Educ. Q. Rev.*, vol. 4, no. 2, 2021, doi: 10.31014/aior.1993.04.02.286.

[6] M. K. G. Illanes, J. A. Gallegos-Mardones, and A. Z. Vasquez-Parraga, "Explaining Loyalty in Higher Education: A Model and Comparative Analysis From the Policy of Gratuity, a Case Applied to Chile," *Sustainability*, vol. 13, no. 19, p. 10781, 2021, doi: 10.3390/su131910781.

[7] S. Annamdevula, "Relationship Between Service Quality, Satisfaction, Motivation and Loyalty," *Qual. Assur. Educ.*, vol. 25, no. 2, pp. 171–188, 2017, doi: 10.1108/qae-04-2013-0016.

[8] T. Djudin, "The Effect of Teaching Method and Lecture Program on Students' Satisfaction Rates and Academic Achievement," *Jetl J. Educ. Teach. Learn.*, 2018, doi: 10.26737/jetl.v3i1.322.

[9] L. Masserini, M. Bini, and M. Pratesi, "Do Quality of Services and Institutional Image Impact Students' Satisfaction and Loyalty in Higher Education?," *Soc. Indic. Res.*, vol. 146, no. 1–2, pp. 91–115, 2018, doi: 10.1007/s11205-018-1927-y.

[10] Z. Mihanović, A. B. Batinić, and J. Pavičić, "The Link Between Students' Satisfaction With Faculty, Overall Students' Satisfaction With Student Life and Student Performances," *Rev. Innov. Compet.*, vol. 2, no. 1, pp. 37–60, 2016, doi: 10.32728/ric.2016.21/3.

[11] M. Topaloglu and G. Malkoç, "Decision Tree Application for Renal Calculi Diagnosis," *Int. J. Appl. Math. Electron. Comput.*, pp. 404–404, 2016, doi: 10.18100/ijamec.281134.

[12] N. R. Mazahua, L. Rodríguez-Mazahua, A. López-Chau, and G. Alor-Hernández, "Comparative Analysis of Decision Tree Algorithms for Data Warehouse Fragmentation*," *Rev. Perspect. Empres.*,

vol. 7, no. 2–1, pp. 31–43, 2020, doi: 10.16967/23898186.667.

[13] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, “An Up-to-Date Comparison of State-of-the-Art Classification Algorithms,” *Expert Syst. Appl.*, vol. 82, pp. 128–150, 2017, doi: 10.1016/j.eswa.2017.04.003.

[14] T. Wakiya *et al.*, “Prediction of Massive Bleeding in Pancreatic Surgery Based on Preoperative Patient Characteristics Using a Decision Tree,” *Plos One*, vol. 16, no. 11, p. e0259682, 2021, doi: 10.1371/journal.pone.0259682.

[15] R. Shan, X. Xiao, J. Che, J. Du, and Y. Li, “Data Mining Optimization Software and Its Application in Financial Audit Data Analysis,” *Mob. Inf. Syst.*, vol. 2022, pp. 1–7, 2022, doi: 10.1155/2022/6851616.

[16] G. K. BAYDOĞMUŞ, “Detecting Internet of Things Attacks by Using Hybrid Learning and Feature Selection Method,” *Eur. J. Sci. Technol.*, 2021, doi: 10.31590/ejosat.1017433.

[17] F. S. Saadaari, D. Mireku-Gyimah, and B. Olaleye, “Development of a Stope Stability Prediction Model Using Ensemble Learning Techniques - A Case Study,” *Ghana Min. J.*, vol. 20, no. 2, pp. 18–26, 2020, doi: 10.4314/gm.v20i2.3.

[18] R. Mitra and T. Rajendran, “Efficient Prediction of Stroke Patients Using Random Forest Algorithm in Comparison to Support Vector Machine,” 2022, doi: 10.3233/apc220075.

[19] H. Yıldırım, “Comparative Analysis of Machine Learning Algorithms Based on Variable Importance Evaluation,” *J. Sci. Technol. Eng. Res.*, 2021, doi: 10.53525/jster.988672.

[20] Z. P. Agusta and A. Adiwijaya, “Modified Balanced Random Forest for Improving Imbalanced Data Prediction,” *Int. J. Adv. Intell. Inform.*, vol. 5, no. 1, p. 58, 2018, doi: 10.26555/ijain.v5i1.255.

[21] L. Gao, M. Ye, and C. Wu, “Cancer Classification Based on Support Vector Machine Optimized by Particle Swarm Optimization and Artificial Bee Colony,” *Molecules*, vol. 22, no. 12, p. 2086, 2017, doi: 10.3390/molecules22122086.

[22] S. Akin, M. Penner, and J. Peissig, “Joint Channel Estimation and Data Decoding Using SVM-based Receivers,” 2020, doi: 10.48550/arxiv.2012.02523.

[23] J. J. M. Durango, C. González-Castaño, C. Restrepo, and J. Muñoz, “Application of Support Vector Machine to Obtain the Dynamic Model of Proton-Exchange Membrane Fuel Cell,” *Membranes*, vol. 12, no. 11, p. 1058, 2022, doi: 10.3390/membranes12111058.

[24] S. Das, D. Tripathy, and J. L. Raheja, “A Review on Algorithms for EEG-Based BCIs,” pp. 25–56, 2018, doi: 10.1007/978-981-13-3098-8_3.

[25] M. López-Pacheco and W. Yu, “Complex Valued Deep Neural Networks for Nonlinear System Modeling,” *Neural Process. Lett.*, vol. 54, no. 1, pp. 559–580, 2021, doi: 10.1007/s11063-021-10644-1.

[26] M. Pawlicki and R. S. Choraś, “Preprocessing Pipelines Including Block-Matching Convolutional Neural Network for Image Denoising to Robustify Deep Reidentification Against Evasion Attacks,” *Entropy*, vol. 23, no. 10, p. 1304, 2021, doi: 10.3390/e23101304.