ABSTRACT

Education provides a very important role in improving the quality of life in society in a country. With a large number of students in each class, it can cause the material to not be delivered properly. Therefore, it is necessary to group students based on their learning ability. The data used was obtained from Senior High School (SMA) Darunnajah Jakarta. Darunnajah High School Jakarta is one of the educational institutions under the auspices of Darunnajah Islamic Boarding School. Data mining techniques with classification methods are proposed to predict student performance in class. The results of student classification can be used as a reference in providing material according to their learning ability. The aim of this research is to ascertain the optimal classification algorithm and pinpoint the key factors influencing students' academic standing. Various classification methods, including logistic regression, KNN, and SVM, were employed in this study. The performance of these models was assessed using diverse metrics such as the f1 score, ROC curve, and performance matrix. Ultimately, the SVM algorithm demonstrated the highest accuracy, achieving an 84% accuracy rate compared to KNN and logistic regression.

Keywords: Classification, Students Performance, Data Mining, Education

1. INTRODUCTION

Education is one of the critical factors in every country for improving the quality of life of its people. It serves as the foundation for personal and societal development, enabling individuals to acquire the knowledge, skills, and competencies needed to succeed in various aspects of life (Dani, 2023; Foundation, 2023; Nurfatimah et al., 2022). A well-educated population is better equipped to engage in meaningful employment, which can lead to higher income levels and improved economic stability (Choudhari et al., 2024). Furthermore, education fosters critical thinking and problem-solving abilities, empowering people to make informed decisions and contribute positively to their communities (Hakim, 2023; Simamora, 2023). It also plays a vital role in promoting social equity, as it provides opportunities for all individuals, regardless of their background, to achieve their full potential (Retno Winarti et al., 2018).

By addressing disparities and offering support to marginalized groups, education helps bridge social divides and create a more inclusive society (Rohman, 2023). Education contributes to better health outcomes by increasing awareness and understanding of health practices, nutrition, and disease prevention (Salsabilla, 2023). Educated individuals are more likely to
engage in healthy behaviors and seek medical care when needed, leading to longer and healthier lives (Efendi & Sholeh, 2023). Additionally, education encourages civic involvement and political participation, as those with higher education levels tend to engage more in democratic activities and advocate for their own and others’ rights (Mii et al., 2023; Rahadianpradiputra, 2023). Investing in education is crucial for countries seeking to improve their citizens’ quality of life, driving economic progress and creating a fairer, healthier, and more active society (Domino, 2018; Mahendrawan & Rahayu, 2020; Sudarmono et al., 2020).

In some countries, governments play a pivotal role in enhancing education quality through comprehensive policies and programs (Foundation, 2024; Naila, 2023; Putri, 2014). These efforts encompass increased school funding, access to quality learning materials, professional development for teachers, and infrastructure upgrades. By setting standards and ensuring accountability, governments foster improved educational outcomes. Moreover, they address educational disparities, supporting disadvantaged communities for equal opportunities. Through these various actions, the government plays an essential role in shaping and enhancing the quality of education in these countries (Irma et al., 2024).

There are numerous factors that impact students' academic achievement (Adiputra & Mujiyati, 2017; Setiyatna et al., 2022; Thabroni, 2022; TL et al., 2017). Among these are the socioeconomic status of their families, which can affect access to resources and support. Additionally, the quality and availability of learning facilities at school play a significant role. Student motivation is another crucial element, as it drives their engagement and effort in their studies (Khairunnas et al., 2023). Discipline within the student body also contributes to academic success, as it fosters a conducive learning environment. Lastly, students’ past academic performance often sets a precedent for future achievements, influencing their confidence and academic trajectory (Fatunnisa & Marcos, 2024).

Based on the description above, this study aims to make early predictions of students who have the potential to not excel or experience obstacles in their learning, so that anticipatory steps can be taken from schools to avoid themselves from the possibility of not going to grade and even expelling students from school. The step that can be taken by schools after knowing that there are students who have the potential to not excel is to provide special assistance to these students. The final hope is that all students from various backgrounds can each factor maximize the learning achievement of their students.

2. THEORY

Data is a collection of information consisting of facts, figures, or other details organized to be used or analyzed (Coombs, 1964; Sagiroglu & Sinanc, 2013). The meaning of data depends on how we process and interpret it (Amira, 2021; Richardson, 2024). In today's technological era, data has become very important in various fields. The dataset utilized in this study was obtained from the records of SMA Darunnajah Jakarta. SMA Darunnajah Jakarta is a senior secondary education institution located in Indonesia. The school was founded with a vision to provide quality education that combines religious, academic, and social values to its students. In addition to focusing on a strong academic curriculum, Darunnajah High School also emphasizes character building and extracurricular activities to prepare students to become
individuals of integrity and contribute positively to society. As one of the leading high schools in Indonesia, SMA Darunnajah Jakarta continuously strives to create an inclusive and progressive learning environment for its students.

The concept of machine learning involves creating a system that can learn autonomously without explicit instructions from its users (Balaji et al., 2021). Computers can analyze digital data to identify patterns and rules in ways that are too complex for humans to accomplish (Alyahyan & Dù, 2020). The fundamental concept behind learning engines is that computers can autonomously learn from their experiences (Caspari-sadeghi, 2023). While the specific uses of machine learning may differ, they typically share common functionalities across applications. Computers process extensive datasets, uncovering underlying patterns and rules within the data (Ahmed T. et al., 2023). These patterns and rules are typically mathematical in nature, allowing for easy definition and processing by computers. Subsequently, the computer can apply these rules to interpret new data effectively. The process of generating rules from data is automated and continually refined with the introduction of new data.

Previously, studies examining factors influencing student achievement predominantly relied on statistical data analysis. However, in this particular study, researchers employed data mining techniques. Data mining represents a novel area within computer science, boasting a wide array of applications (Shuaeb et al., 2021). This is enabled by the abundance and diversity found across various scientific domains (such as artificial intelligence, database management, statistics, mathematical modeling, and image processing), which contribute to expanding the applicability of data mining. The primary reason data mining has garnered significant attention in the information industry in recent years is due to the availability of vast quantities of data and the growing demand to transform this data into actionable information and knowledge. Data mining involves the extraction and discovery of knowledge from large datasets, which proves invaluable for further development (Pratama et al., 2024).

Data mining refers to the process of uncovering valuable insights and knowledge from large datasets stored in databases (Panggabean et al., 2022). Data mining is a method that employs statistical, mathematical, artificial intelligence, and machine learning approaches to extract and recognize valuable information and associated knowledge from diverse large databases (Alyahyan & Dù, 2020). Data mining is a semi-automated process that utilizes statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify valuable information and potential knowledge stored within extensive databases.

In the case of prediction, a method or model is needed that can produce data classification patterns with the ultimate goal of forecasting. The technique or the method that will be used to find out the information is with how to analyze student academic data. Data mining encompasses diverse techniques that enable predictions and classifications. These techniques involve estimating future probabilities by examining existing information and data patterns (Kukkar et al., 2020).

Academic achievement is an accomplishment in the field of education that reflects a person's intellectual ability and dedication (Akpur, 2020; Hattie & Anderman, 2013). This can be in
the form of high grades, awards, scholarships, scientific publications, or graduating with top honors (Lindholm-Leary & Borsato, 2006; Steinmayr et al., 2014). These achievements demonstrate an individual's success in mastering subject matter, contributing to research, or excelling in academic competitions, and often serve as important indicators in evaluating one's abilities and commitment to education (Al-Abyadh & Abdel Azeem, 2022; Steinmayr et al., 2019).

Each method used in previous studies has its disadvantages and advantages. The fundamental weakness of the method carried out is time efficiency in data processing. Large data processing with many independent variables used results in less good method efficiency. Another weakness is the low level of accuracy obtained from the test results. In order to advance research efforts, researchers employed the Logistic Regression Algorithm, K-Nearest Neighbors Algorithm, and Support Vector Machine (SVM) methodologies for predicting student achievement in this study. A literature review is an in-depth study of scientific literature, as shown in Table 1 below:

<table>
<thead>
<tr>
<th>Year</th>
<th>Paper Title</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>2024</td>
<td>Comparative analysis of Decision Tree and Naïve algorithms to evaluate student achievement Case Study: Smk Al-Musyawirin (Fauzan et al., 2024)</td>
<td>The Decision Tree method showed the highest accuracy result of 89.22%</td>
</tr>
<tr>
<td>2024</td>
<td>Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features (Brahim, 2022)</td>
<td>The findings indicated that our model attained the highest classification accuracy rate of 97.4% using the RF classifier.</td>
</tr>
<tr>
<td>2024</td>
<td>Analysis of Student Academic Performance Using Social Network Analysis (Afandi et al., 2024)</td>
<td>The results showed that the average algorithm accuracy for predicting student academic performance was 0.76. The Support Vector Machine yielded the highest accuracy rate, reaching 92%.</td>
</tr>
<tr>
<td>2022</td>
<td>The role of demographic and academic features in a student performance prediction (Bilal et al., 2022)</td>
<td>Model test results get accuracy of 87.5%. It also achieved an AUC (Area Under the Curve) of 94.5%, precision of 88%, recall of 87.5%, and F1-score of 87.5%, respectively. Using the C4.5 datamining method, an accuracy of 97.5% was obtained.</td>
</tr>
</tbody>
</table>
Table 1 above outlines some of the proposed contributions to references related to this research plan. In previous studies, performance prediction research used various classification methods and references to prediction determinants that have been processed. What distinguishes and positions this research against previous research is using the classification method of Logistic Regression, k Nearest Neighbor, and Support Vector Machine by comparing various functions to get the model with the best accuracy results. The object of research is carried out to classify student achievement predictions as training data and testing data, which can then later be applied as one of the supporting policies of study program managers in increasing student achievement. In addition, this study can display features of excellent conditions for high academic potential for students.

3. METHOD
4.1 Data Description
Data collection involves the process of acquiring and gathering educational-related data stored within repositories of educational institution (Alwarthan et al., 2022). The study utilizes data sourced from SMA Darunnajah Jakarta, encompassing various attributes such as student grades, demographics, and social and school-related features. This data is collected through school reports and questionnaires. Two datasets are provided, each focusing on performance in distinct subjects: Mathematics (mat) and Portuguese (por). These datasets are then modeled under classification tasks and binary/five-level regression. It's noteworthy that the target attribute G3 demonstrates a strong correlation with attributes G2 and G1. This correlation arises because G3 represents the year-end value (issued in the 3rd period), while G1 and G2 correspond to values from the 1st and 2nd periods. Predicting G3 without considering G2 and G1 is challenging; however, such predictions hold significant utility.

4.2 Algorithm used
In previous studies investigating factors influencing student achievement, statistical data processing was commonly utilized. However, in this study, researchers employed data mining techniques. Data mining represents a burgeoning field within computer science, offering a plethora of applications (Panggabean et al., 2022). This is bolstered by the abundance and diversity found across various scientific disciplines (including artificial intelligence, databases, statistics, mathematical modeling, and image processing), which has contributed to the widespread adoption of data mining. The primary reason for the recent surge in interest in data mining within the information industry is attributed to the availability of vast datasets and the increasing demand to transform such data into actionable insights and knowledge. Data mining involves the extraction and discovery of knowledge from large datasets, which proves invaluable for further advancement and development.

Based on student datasets, if research is carried out using machine learning, the author wants to know what affects student achievement. Then by using the Regression Logistics algorithm, K-NN and SVM are compared to the highest accuracy in the algorithm. Thus it can help students in order to achieve good achievements and graduate.
4.3 Experiment Plan

An experimental plan is a technique of using the research methods used so that we do something experimental form that suits our research problem. Here are the stages of the experiment plan to be carried out:

Figure 1. Experiment Plan

4. RESULTS AND DISCUSSION

To get maximum results it is necessary to conduct several experiments. Based on experiments that have been carried out, the results are:

1) Read Data and View Data Sets

Figure 2. View Data Sets
2) Data processing

| school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | reason | guardian | traveltime | studytime | failures | schoolsup | Famsup | paid | activities | nursery | higher | internet | romantic | famrel | freetime | goout | Dalc | Walc | health | absences | pas |
|--------|-----|-----|---------|--------|---------|------|------|------|------|--------|-----------|------------|-----------|----------|-----------|---------|------|-----------|----------|-------|----------|--------|-------|--------|--------|---------|-----|-------|
| 0      | 0   | 1   | 18      | 0      | 1       | 4    | 4    | 3    | 0    | ...    | 0         | 0          | 4         | 3         | 4         | 1       | 1     | 3         | 6       |       |          |        |       |        |        |         |     |       |
| 1      | 0   | 1   | 17      | 0      | 1       | 1    | 3    | 4    | ...  | 1      | 0         | 5          | 3         | 3         | 4         | 1       | 1     | 3         | 4       |       |          |        |       |        |        |         |     |       |
| 2      | 0   | 1   | 15      | 0      | 0       | 0    | 1    | 3    | 4    | ...    | 1         | 0          | 4         | 3         | 2         | 2       | 3     | 3         | 10      |       |          |        |       |        |        |         |     |       |
| 3      | 0   | 1   | 15      | 0      | 1       | 0    | 4    | 2    | 1    | 2     | 1         | 1          | 3         | 2         | 2         | 2       | 1     | 1         | 5       | 2     |          |        |       |        |        |         |     |       |
| 4      | 0   | 1   | 16      | 0      | 1       | 0    | 3    | 3    | 4    | 4     | 0         | 0          | 4         | 3         | 2         | 2       | 1     | 2         | 5       | 4     |          |        |       |        |        |         |     |       |
| ...    | ... | ... | ...     | ...    | ...     | ...  | ...  | ...  | ...  | ...    | ...       | ...        | ...       | ...       | ...      | ...   | ... | ...       | ...      | ...  | ...      |        | ...   | ...    | ...    | ...    | ...  |       |
| 390    | 1   | 0   | 20      | 0      | 0       | 1    | 2    | 2    | 2    | ...    | 0         | 0          | 5         | 5         | 4         | 4       | 4     | 5         | 4       | 4     |          |        |       |        |        |         |     |       |
| 391    | 1   | 0   | 17      | 0      | 0       | 0    | 3    | 1    | 2    | 2     | 1         | 0          | 2         | 4         | 5         | 3       | 4     | 3         | 2       | 3     |          |        |       |        |        |         |     |       |
| 392    | 1   | 0   | 21      | 1      | 1       | 0    | 1    | 1    | 4    | 4     | 0         | 0          | 5         | 5         | 3         | 3       | 3     | 3         | 3       | 3     |          |        |       |        |        |         |     |       |
| 393    | 1   | 0   | 18      | 1      | 0       | 0    | 3    | 2    | 2    | 4     | 1         | 0          | 4         | 4         | 1         | 3       | 4     | 5         | 5       | 0     |          |        |       |        |        |         |     |       |
| 394    | 1   | 0   | 19      | 0      | 0       | 0    | 1    | 1    | 4    | 3     | 1         | 0          | 3         | 2         | 3         | 3       | 3     | 3         | 5       | 5     |          |        |       |        |        |         |     |       |
| 395    |     |     |         |        |         |      |      |      |      |        |            |            |            |            |            |        |       |            |         |       |          |        |       |        |        |         |     |       |

Figure. 3. Digitization of Values

3) Features scaling

```python
In [9]: features=['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu', 'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime', 'failures', 'schoolsup', 'Famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc', 'Walc', 'health', 'absences']
```

Figure. 4. Data Inspection

Figure. 5. Features Visualisation
4) Correlation heatmap

![Correlation heatmap](image)

Figure 6. (a) Correlation Heatmap (General Heatmap) dan (b) Correlation between student status and other features

5) Distribution plot

![Distribution plot](image)

Figure 7. (a) Student status with gout and (b) Student status according to romantic relationships

![Distribution plot](image)

Figure 8. (a) Student status according to romantic relationships and (b) Student status according to mother's occupation

Figure 8.a shows that most of the people who passed the exam did not have romantic relationships, not suggesting relationships could be a good option for better performance. Figure 8.b shows that it seems that students whose mothers worked as doctors achieved good status.
Figure 9. (a) Student status according to maternal education and (b) Student status based on motivation to enter higher education

Figure 9.a Maternal education positively impacts student status, with Maternal Education being more influential than Paternal Education, as shown in the second heatmap. Figure 9.b shows that most students who pass the exam have the motivation or desire to continue to higher education.

Figure 10. (a) Student status based on age and (b) Student status based on failure

Figure 10. a Age significantly influences student success, with a majority of exam passers being 15 years old, while most failures are observed among 22-year-olds. In essence, attending school at a younger age is preferable for better outcomes. Figure 10.b Most individuals who pass the exam have experienced failures, suggesting that diligent study and passing all grades can be a viable strategy.

Figure 11. (a) Student status by region and (b) Student status by internet accessibility

Figure 11. a The region did not affect student performance; even those achieving good results resided in rural areas. Figure 11.b Most exam passers have internet access, emphasizing the need for equitable educational resources. With internet connectivity, students can readily access lessons from teachers and supplement their understanding through online research.
Learning is the individual's endeavor to acquire changes in behavior, encompassing knowledge, skills, attitudes, and positive values gained from various learning materials. Figure 12.b indicates that the majority of students who failed the exam lacked good health.

6) Logistic Regression
In logistic regression, the final GPA was divided into two categories. (Bou-hamad, 2020). To assess the model, it will initially determine accuracy, display the confusion matrix visually, and subsequently plot the ROC curve.

Figure 13 presents two accuracy values: one from the training set and the other from the test set. Comparing them is important; if the training set accuracy significantly exceeds the test set accuracy, it suggests overfitting. Test set accuracy is more pertinent for assessing performance on new, unbiased data.

Given the results from the experiment, enhancing the model for improved accuracy is necessary. The next step involves using a random state to partition the data into training and test sets randomly. This algorithm yields different accuracies with each change in data division.
Ideally, a robust model's accuracy should not fluctuate significantly under random conditions. However, to ensure consistency, the model will undergo multiple iterations. Instead of using "0" and "1" for randomness, an "optimal condition" value will be chosen to maximize accuracy and F1 score for each iteration.

*Accuracy is: 80.67226890756302
*F1 score is: 0.740839357068459

Figure 16. Improving Model Accuracy

![Figure 16. Improving Model Accuracy](image)

Figure 17. (a) Logistic Regression ROC curve Improving Model Accuracy dan (b) Confusion Matrix Logistic Regression Improving Model Accuracy

The experiment's conclusion highlights a significant improvement in model accuracy. By implementing methods to enhance accuracy, the accuracy rate increased from 64% to 80.67%, accompanied by a notable rise in the F1 score from 0.55 to 0.74.

7) K-Nearest Neighbors

First step: Before configuring the algorithm parameters for KNN (k-value and metric), it's advisable to determine the optimal random state prior to data division and KNN adjustment. We'll select a method that identifies an "optimal state," maximizing accuracy and F1 score for a given iteration. Initially, we won't specify KNN parameters such as k-values and metrics. Subsequently, we'll evaluate the model to gauge the impact of KNN hyperparameter tuning.

![Figure 18. K-Nearest Neighbors ROC curve](image) dan (b) Confusion Matrix KNN

Model evaluation: This model and looks at the impact of knn-parameters using roc curves and confusion matrices and f1 scores. Using this model I got 78% accuracy, hope good accuracy and f1 score is 70%.
In this case study having a binary classification, it is better to choose an odd K value. By looking at the curve we can observe above that we get the maximum testing accuracy for \( k=5 \). In the next step let's confirm if 5 is a good choice by using gridsearchCV.

As we see in figure 20, the metric or distance gives the same accuracy but the time is different, the optimal time for \( k=7 \) is 0.019012451171875. This option gives a high Acc=78% with less time consuming compared to other distances the winning distance is chebyshev.

Combining the best KNN parameters with the optimal state increases accuracy from 78% to 79%. Hyperparameter tuning evidently enhances results, yet the random state remains the most influential parameter.
8) Support Vector Machine

a. Model evaluation: For model evaluation, we will calculate: Training time, Accuracy, Confusion matrix, ROC curve, ROC score, f1 score.

b. Training phase: Note That it won't repeat all hyperparameters again, we've done that and found the optimal value, if you want to set that hyperparameter again, just set max iteration to 2000 e.g. comment out the line where we set the optimal value for the split random-state. Another thing to keep in mind is that I haven't tuned the hyperparamter with a large number of iterations until I've made sure that the model isn't high biased or high variance.

The linear kernel SVM classifier performs well, with an 84% accuracy, indicating effective data labeling as shown by the diagonal pattern in the confusion matrix.

The metrics reveal that our polynomial kernel SVM classifier is less effective, with a 78% accuracy, which is inadequate for our problem. The confusion matrix indicates mislabeled data by the classifier.
Our polynomial kernel SVM classifier shows subpar performance with 78% accuracy, as indicated by the confusion matrix showing mislabeled data. Comparison of Three Kernel SVM:

Next it compares all metrics and plots one graph containing all three ROC curves from the three SVM kernels:

For optimal performance:
- **Accuracy**: Over 80% is acceptable for our problem.
- **Confusion matrix**: It should nearly form a diagonal, indicating minimal misclassification.
- **F1 score**: Above 0.75 is considered good; our obtained value of 0.82 is favorable.
- **ROC curve**: It should significantly exceed the blue dotted line.
- **ROC score**: Ideally 1, but above 0.75 is acceptable for our problem.

Linear kernel SVM models perform best, exhibiting high accuracy and F1 scores on test sets. The ROC curve shows the orange curve well above the blue dotted line, resulting in a high ROC score of 0.8. Hence, the conclusion is to select a linear kernel.

9) **Perbandingan Algoritma Logistic Regresion, kNearest Neighbor dan Support Vector Machine**

Comparison of the Three Algorithms The following is the comparison result of the three algorithms (Logistic Regression, K-nearest Neighbor, Support Vector Machine):
5. CONCLUSIONS AND SUGGESTIONS

Having dealt with the features of excellent conditions for high academic potential tend to have this profile: Not often out with friends, outside of romantic relationships, mothers pursue specialized higher education. They possess a keen desire to further their education. The mother works in healthcare while the father is an educator. There's consistent attendance in classes, access to the internet, studying for more than 10 hours weekly, and maintaining good health. This section aims to analyze KNN hyperparameter tuning. Initially, we deploy a model with an optimal random state, yielding 78% accuracy. Then, we refine the model by finding the best parameters: first, determine the best k value using gridsearchCV, then refine the metric. Ultimately, with k=17 and metric=chebychev, we achieve an improved accuracy of 79%. This underscores the importance of the random state despite the impact of hyperparameters.

The classifier with high metrics is a support vector machine classifier with a high accuracy of 84%, the fusion matrix is approximately diagonal indicating that this classifier is able to label the data correctly. If we look at other metrics like f1 score, we manage to get a good score of 0.82 which means that we have low false positives and low false negatives. Now let's go back and look at the advice given to students, parents, and administration.

6. ACKNOWLEDGEMENTS

We extend our appreciation to Darunnajah University, SMA Darunnajah Jakarta, and associated parties for their support, encouragement, and assistance, which contributed to the completion of this research.

REFERENCES


Panggabean et al., Analysis Student Academic Achievement Using


Mahasiswa Menggunakan Metode Artificial Neural Network. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(1), 687–693.


TL, D. I., Widowati, A. I., & Surjawati, S. (2017). Faktor-Faktor Yang Mempengaruhi Prestasi Akademik : Studi Kasus Pada Mahasiswa Program Studi Akuntansi Universitas...
Panggabean et.al, Analysis Student Academic Achievement Using
https://doi.org/10.26623/jdsb.v18i1.557