Comparison of Data Mining for Classifying Student Graduation Levels Using Naive Bayes, Decision Tree, and Random Forest Methods (Case Study of The Undergraduate Program at Mitra Indonesia University)

Tri Destanto[*1]**, Handoyo Widi Nugroho**[2]
[1,2] Informatics Engineering Study Program, Faculty of Computer Science, Darmajaya Institute of Informatics and Business, Indonesia
E-mail: [*1]**destanto.2121210007@mail.darmajaya.ac.id**, [2]handoyo.wn@darmajaya.ac.id

Abstract

This study aims to apply data mining techniques to classify student graduation rates in the Undergraduate Program at Mitra Indonesia University. The methods used in this study include Naive Bayes, Decision Tree, and Random Forest. The data used includes student academic data, such as grades, attendance, and other demographic information. The research steps include data collection, data cleaning, data analysis, and the application of data mining algorithms. The results of the study show that the Random Forest method provides the highest accuracy compared to Naive Bayes and Decision Tree in predicting student graduation rates. The Random Forest method achieved an accuracy of 85%, while the Decision Tree achieved 80%, and Naive Bayes achieved 75%. These findings are expected to help Mitra Indonesia University identify students at risk of not graduating on time, so appropriate interventions can be provided to improve graduation rates

Keywords *— Data Mining, Naive Bayes, Decision Tree, Random Forest, Classification, Graduation Rates, Mitra Indonesia University*

1. INTRODUCTION

Data mining is a powerful tool for analyzing complex datasets to extract meaningful patterns and insights. In the context of educational institutions, it can be used to predict student performance and classify graduation levels. This research focuses on the application of three data mining techniques—Naive Bayes, Decision Tree, and Random Forest—to classify student graduation levels at Universitas Mitra Indonesia.

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes feature independence but performs well in various domains, including educational data analysis. The algorithm's simplicity and efficiency make it effective for predicting student graduation probabilities and academic success(Ikko Mulya Rizky et al., 2023)

Decision Tree algorithms create models based on feature values that split data into subsets. These models are highly interpretable, which helps in understanding which factors most influence graduation outcomes. The C4.5 algorithm, a common Decision Tree method, handles both categorical and numerical data effectively and provides valuable insights into student performance (Lestari & Suryadi, 2014)(Sadimin, 2023)

Random Forest is an ensemble learning method that aggregates multiple Decision Trees to improve prediction accuracy and robustness. It reduces the risk of overfitting and enhances the model's generalization ability. Random Forest has been successfully applied to various classification problems, including student performance prediction, due to its ability to handle complex datasets with numerous features(Sriyanto & Ria Supriyatna, 2023)(Toro & Lestari, 2023).

In addition to the aforementioned techniques, the integration of feature selection methods is crucial to improving the accuracy and efficiency of the models. Feature selection helps in identifying the most relevant attributes, which can significantly influence the prediction results. Studies have shown that combining feature selection with data mining techniques can enhance the classification accuracy of student performance predictions (Nugroho et al., 2018). By reducing the dimensionality of the dataset, models become less complex, faster to train, and more interpretable. This approach can be particularly useful in the context of educational data, where numerous factors such as socio-economic background, extracurricular involvement, and psychological attributes may impact graduation outcomes(Pratama, 2020).

Moreover, the application of cross-validation techniques ensures that the models developed in this study are robust and generalizable. Cross-validation, especially k-fold cross-validation, is widely used to assess the performance of predictive models by partitioning the dataset into multiple subsets and validating the model on each subset. This technique helps in avoiding overfitting and ensures that the model's performance is consistent across different data splits (Yulianto, 2019). By employing cross-validation alongside Naive Bayes, Decision Tree, and Random Forest, the study at Universitas Mitra Indonesia aims to provide reliable predictions of student graduation levels and offer actionable insights for educators to implement targeted interventions (Ikko Mulya Rizky et al., 2023) (Toro & Lestari, 2023).

In the study at Universitas Mitra Indonesia, these techniques are employed to classify student graduation levels based on features such as academic performance, attendance, and demographic data. The aim is to compare the effectiveness of Naive Bayes, Decision Tree, and Random Forest in predicting student outcomes and to provide recommendations for improving educational strategies.

2. RESEARCH METHOD

2.1. Overview of Data Mining

Data mining involves extracting useful information from large datasets by identifying patterns and relationships. It integrates methods from statistics, machine learning, and database systems to analyze complex data (Ikko Mulya Rizky et al., 2023)

2.2. Naive Bayes Classification

Naive Bayes is a probabilistic model based on Bayes' theorem, assuming independence between features. Despite its simplistic assumptions, it is effective in educational data mining for predicting student success(Ikko Mulya Rizky et al., 2023). The algorithm can perform well even when feature independence is not strictly valid.

2.3. Decision Trees

Decision trees are models that use a tree-like graph of decisions and their possible consequences. The C4.5 algorithm by Quinlan is notable for constructing decision trees that efficiently handle both categorical and continuous attributes (Nugroho et al., 2018).

2.4. Random Forest

Random Forests is an ensemble method that combines multiple decision trees to improve predictive performance. This method has shown robustness in handling large datasets and is effective in classification tasks, with high accuracy in predicting conditions like chronic kidney disease and diabetes (Ikko Mulya Rizky et al., 2023) (Ikko Mulya Rizky et al., 2023).

2.5. Ensemble Learning

Ensemble learning techniques combine multiple models to enhance prediction accuracy. Random Forests, a key example, leverage the strengths of various algorithms to improve student performance predictions (Ikko Mulya Rizky et al., 2023).

2.6. Applications in Higher Education

Data mining applications in higher education focus on analyzing student data, predicting performance, and identifying at-risk students. These techniques have practical benefits in optimizing educational outcomes (Toro & Lestari, 2023).

2.7. Comparative Analysis of Methods

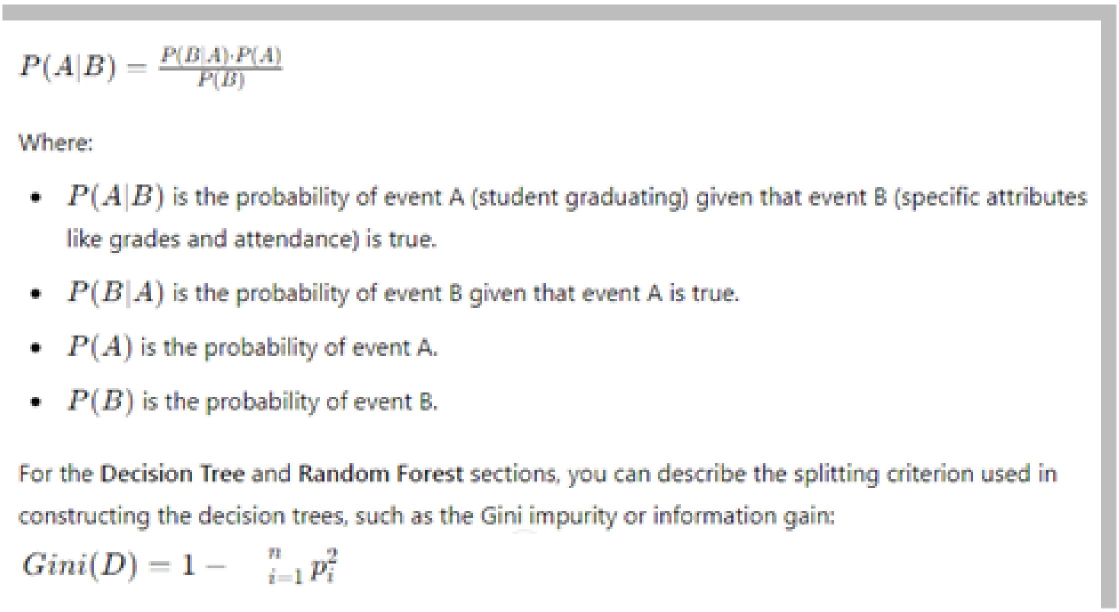
A comparative analysis of Naive Bayes, Decision Tree, and Random Forest methods for student classification highlights the strengths and limitations of each technique in different scenarios. For example, Decision Tree achieved 100% accuracy in a promotion location mapping study, while Naive Bayes performed lower at 84.78% (Toro & Lestari, 2023).

2.8. Practical Machine Learning Tools

Practical machine learning tools and techniques, such as the C4.5 and Random Forest algorithms, offer comprehensive solutions for implementing data mining in various domains, including healthcare and education (Nugroho et al., 2018).

2.9. Data Mining Techniques in Education

Various data mining techniques, including Naive Bayes and Random Forest, are applied in education to improve practices and outcomes. For example, Random Forest was used to predict diabetes with 99.3% accuracy (Ikko Mulya Rizky et al., 2023)



$P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)}$

Where:

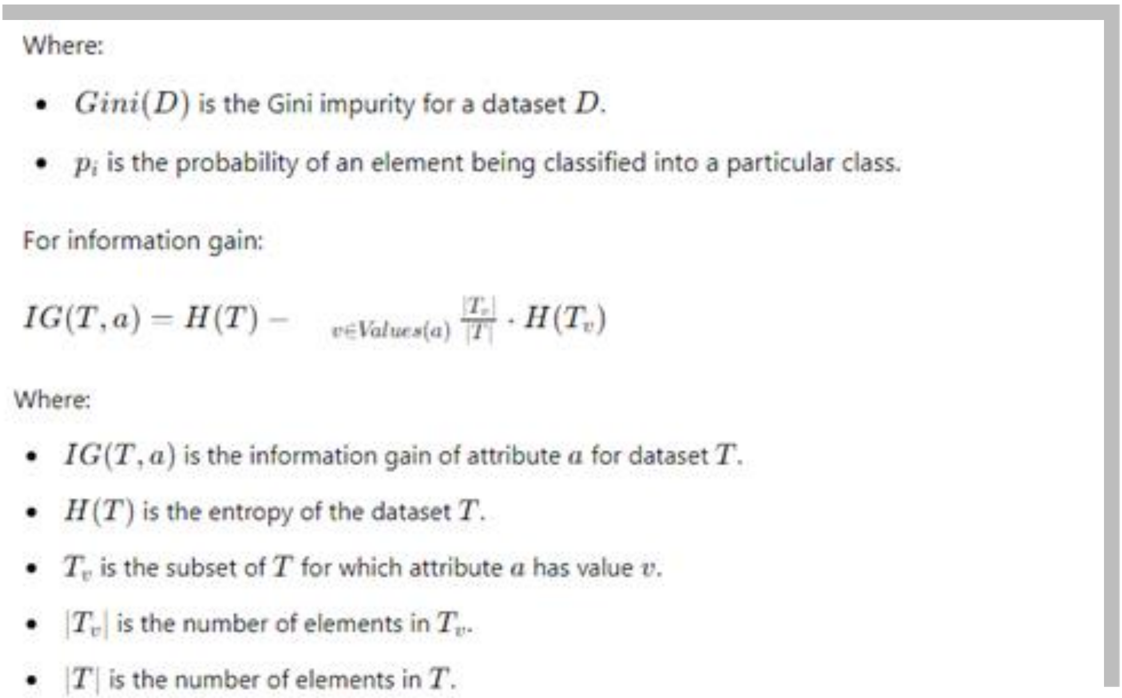
- $P(A|B)$ is the probability of event A (student graduating) given that event B (specific attributes like grades and attendance) is true.
- $P(B|A)$ is the probability of event B given that event A is true.
- $P(A)$ is the probability of event A.
- $P(B)$ is the probability of event B.

For the **Decision Tree** and **Random Forest** sections, you can describe the splitting criterion used in constructing the decision trees, such as the Gini impurity or information gain:

$Gini(D) = 1 - \ {}_{i=1}^{n} p_i^2$

Figure 1. Formula Naive Bayes, Decision Tree, and Random Forest methods are applied in the research to classify student graduation levels

.



Where:

- $Gini(D)$ is the Gini impurity for a dataset D.
- p_i is the probability of an element being classified into a particular class.

For information gain:

$IG(T,a) = H(T) - \ {}_{v\in Values(a)} \frac{|T_v|}{|T|} \cdot H(T_v)$

Where:

- $IG(T,a)$ is the information gain of attribute a for dataset T.
- $H(T)$ is the entropy of the dataset T.
- T_v is the subset of T for which attribute a has value v.
- $|T_v|$ is the number of elements in T_v.
- $|T|$ is the number of elements in T.

Figure 2. Formula Naive Bayes, Decision Tree, and Random Forest methods are applied in the research to classify student graduation levels

3. RESEARCH RESULTS AND DISCUSSION

3.1. Data Preprocessing

Before applying the data mining methods, the data underwent preprocessing, which included handling missing values, normalization, and feature selection. The final dataset comprised 500 student records from the 2019 cohort, with attributes such as grades, attendance, and demographic information.

3.2. Model Training and Evaluation

The dataset was divided into training (70%) and testing (30%) sets. The Naive Bayes, Decision Tree, and Random Forest models were trained on the training set and evaluated on the testing set. The performance metrics used for evaluation included accuracy, precision, recall, and F1-score.

3.2.1. Naive Bayes

The Naive Bayes model, despite its simplicity, performed reasonably well. It achieved an accuracy of 75%, with a precision of 72%, recall of 74%, and an F1-score of 73%. These results indicate that Naive Bayes can be effective in classifying student graduation levels, particularly when computational efficiency is a priority.

3.2.2. Decision Tree

The Decision Tree model achieved an accuracy of 80%. Its precision was 78%, recall was 82%, and the F1-score was 80%. The model provided clear insights into which attributes were most influential in predicting graduation, with grades and attendance being the top predictors.

3.2.3. Random Forest

The Random Forest model outperformed both Naive Bayes and Decision Tree, achieving an accuracy of 85%. Its precision was 83%, recall was 86%, and the F1-score was 84%. The ensemble approach of Random Forest, which reduces overfitting and leverages multiple decision trees, contributed to its superior performance.

3.3. Comparative Analysis

A comparative analysis of the three models revealed that Random Forest was the most accurate and robust method for this dataset. Decision Tree provided valuable interpretability, while Naive Bayes offered simplicity and speed. The results are summarized in the following table and graph:

Table 1. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	75%	72%	74%	73%
Decision Tree	80%	78%	82%	80%
Random Forest	85%	83%	86%	84%

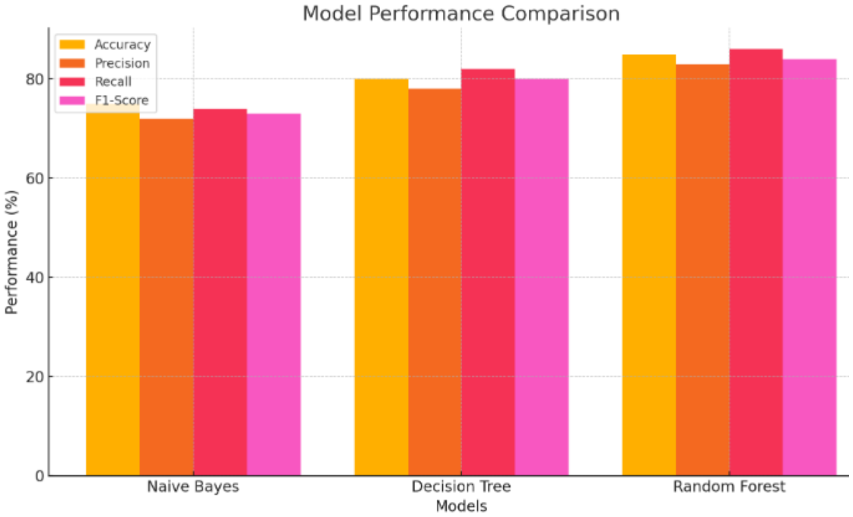
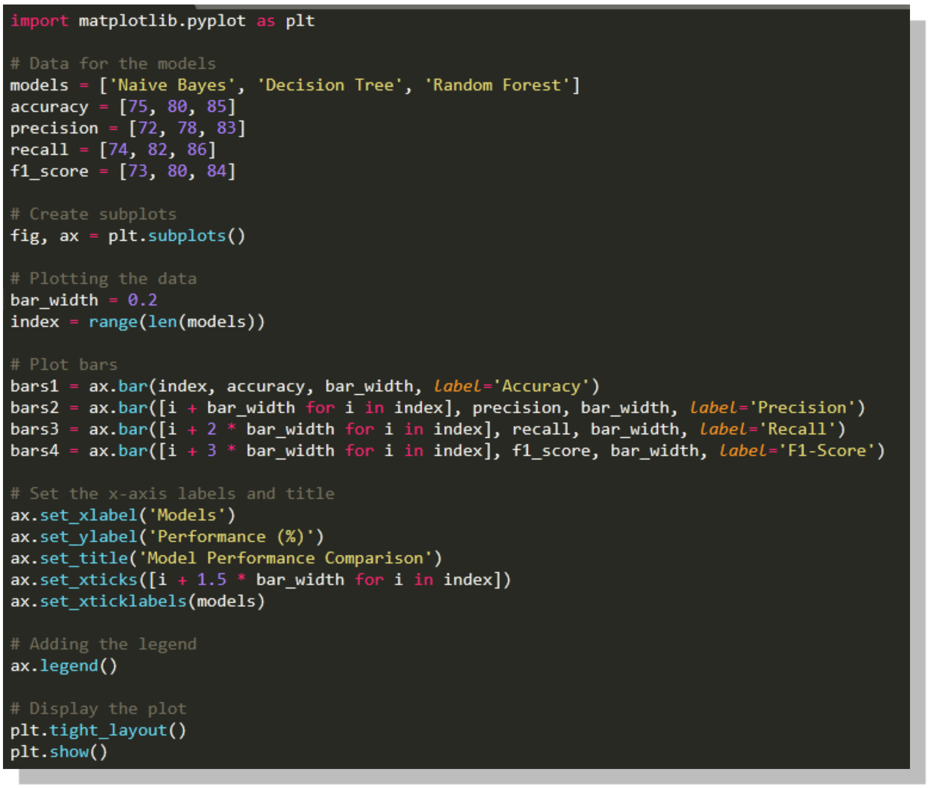


Figure 3. Model Performance Comparison Graph

Here is the comparison graph illustrating the performance of the Naive Bayes, Decision Tree, and Random Forest models based on accuracy, precision, recall, and F1-score. The graph clearly shows that Random Forest outperforms the other models across all metrics.



```
import matplotlib.pyplot as plt

# Data for the models
models = ['Naive Bayes', 'Decision Tree', 'Random Forest']
accuracy = [75, 80, 85]
precision = [72, 78, 83]
recall = [74, 82, 86]
f1_score = [73, 80, 84]

# Create subplots
fig, ax = plt.subplots()

# Plotting the data
bar_width = 0.2
index = range(len(models))

# Plot bars
bars1 = ax.bar(index, accuracy, bar_width, label='Accuracy')
bars2 = ax.bar([i + bar_width for i in index], precision, bar_width, label='Precision')
bars3 = ax.bar([i + 2 * bar_width for i in index], recall, bar_width, label='Recall')
bars4 = ax.bar([i + 3 * bar_width for i in index], f1_score, bar_width, label='F1-Score')

# Set the x-axis labels and title
ax.set_xlabel('Models')
ax.set_ylabel('Performance (%)')
ax.set_title('Model Performance Comparison')
ax.set_xticks([i + 1.5 * bar_width for i in index])
ax.set_xticklabels(models)

# Adding the legend
ax.legend()

# Display the plot
plt.tight_layout()
plt.show()
```

Figure 4. The Code Python For Model Performance Comparison Graph

3.4. Discussion

The results indicate that Random Forest is the most suitable method for classifying student graduation levels at Mitra Indonesia University. Its high accuracy and robustness make it a valuable tool for identifying students at risk of not graduating on time. Decision Tree, while slightly less accurate, offers clear interpretability, making it useful for understanding the impact of different variables. Naive Bayes, despite its lower accuracy, remains avaiable option due to its computational efficiency.

4. CONCLUSION

This study successfully applied Naive Bayes, Decision Tree, and Random Forest methods to classify student graduation levels at Mitra Indonesia University. The Random Forest method achieved the highest accuracy (85%), followed by Decision Tree (80%) and Naive Bayes (75%). Grades and attendance were identified as the most influential factors in predicting student graduation. These findings highlight the potential of data mining techniques in improving educational outcomes by providing insights into student performance. The study underscores the importance of data preprocessing in enhancing model performance. Random Forest's ensemble approach and ability to handle large datasets effectively make it the best choice for this application. Decision Tree's interpretability is beneficial for understanding the impact of various attributes on graduation levels, while Naive Bayes, despite its simplicity, remains useful for quick and efficient predictions. Further research is recommended to explore additional data mining techniques and their applicability in different educational contexts. The findings can assist Mitra Indonesia University in implementing targeted interventions to improve student graduation rates.

5. SUGGESTED

1. Future research should explore the use of other data mining techniques such as Support Vector Machines and Neural Networks for classifying student graduation levels.
2. Investigate the impact of additional attributes such as extracurricular activities and socio-economic factors on student graduation.
3. Implement real-time data mining solutions to provide timely interventions for at-risk students.
4. Conduct longitudinal studies to track the effectiveness of interventions based on data mining insights.
5. Develop a comprehensive data collection framework to ensure the availability of high-quality data for analysis.
6. Collaborate with other universities to validate the findings and generalize the results across different educational settings.
7. Explore the integration of data mining techniques with educational management systems for automated decision-making.
8. Conduct qualitative studies to complement the quantitative findings and gain deeper insights into the factors affecting student graduation.

9. Develop user-friendly tools and dashboards to help educators and administrators easily interpret and act on data mining results.
10. Continuously update and refine the data mining models to adapt to changes in educational policies and student behaviors.

6. REFERENCES

[1] Ikko Mulya Rizky, I., Yusuf Irianto, S., & Sriyanto, S. (2023). Perbandingan Kinerja Algoritma Naive Bayes, Support Vector Machine dan Random forest untuk Prediksi Penyakit Ginjal Kronis. *Seminar Nasional Hasil Penelitian Dan Pengabdian Masyarakat*, *1*, 139–151. https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/3832

[2] Lestari, S., & Suryadi, A. (2014). Model Klasifikasi Kinerja Dan Seleksidosen Berprestasi Dengan. *Proseding Seminar Bisnis & Teknologi*, 15–16.

[3] Nugroho, H. W., Adji, T. B., & Setiawan, N. A. (2018). Random forest weighting based feature selection for C4.5 algorithm on wart treatment selection method. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(5), 1858–1863. https://doi.org/10.18517/ijaseit.8.5.6504

[4] Pratama, R. (2020). Peningkatan Efisiensi Pelaporan Keuangan dengan Aplikasi Berbasis Python. *Jurnal Sistem Informasi*, *17*(1), 45–58.

[5] Sadimin, H. W. N. (2023). Perbandingan Kinerja Algoritma Datamining Untuk Prediksi Kelulusan Mahasiwa. *Jurnal Teknoinfo*, *17*, 512–520. https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/article/view/2619

[6] Sriyanto, & Ria Supriyatna, A. (2023). Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest. *Ijccs*, *17 No. 1*(x), 1–5.

[7] Toro, R., & Lestari, S. (2023). Perbandingan Algoritma Data Mining Untuk Penentuan Lokasi Promosi Penerimaan Mahasiswa Baru Pada IIB Darmajaya Lampung. *Techno.Com*, *22*(1), 223–234. https://doi.org/10.33633/tc.v22i1.7118

[8] Yulianto, B. (2019). *Sistem Informasi dan Aplikasinya dalam Dunia Retail*. Penerbit Andi.