Paradigma, Vol. 26, No. 1, Maret 2024, P-ISSN 1410-5063, E-ISSN: 2579-3500, Page. 1-6
Published: LPPM Universitas Bina Sarana Informatika
PARADIGMA
website: http://jurnal.bsi.ac.id/index.php/paradigma/

Predicting Graduation Outcomes: Decision Tree Model Enhanced with Genetic Algorithm

Sinta Rukiastiandari[1], Luthfia Rohimah[2], Aprillia[3], Fara Mutia[4]

[1,2,3,4] Universitas Bina Sarana Informatika, Jakarta, Indonesia

ARTICLE INFORMATION

Artikel History:
Received: January 29, 2024
Revised: March 18, 2024
Accepted: March 22, 2024

Keyword:
Decission Tree
Genethic Algorithm
Graduation

ABSTRACT

This research aims to improve the accuracy of predicting student permit results in the digital era by utilizing machine learning techniques. The main focus is the use of a Decision Tree (DT) model optimized with a Genetic Algorithm (GA) to overcome the limitations of accuracy and testing of conventional methods. This research began with collecting student academic data, followed by preprocessing to eliminate incompleteness and organize the data format. The DT model is then built and optimized with GA, which is inspired by biological evolutionary processes to improve feature selection and parameter tuning. The results show a significant increase in prediction accuracy, from 86.19% to 87.68%, and an increase in the Area Under Curve (AUC) value from 0.755% to 0.788%. This research not only proves the effectiveness of GA integration in improving DT models, but also paves the way for the application of evolutionary techniques in educational data analysis and other fields. The main contributions of this research include the development of more accurate prediction models and practical applications in educational contexts, with the hope of assisting educational institutions in making more informed decisions for their students.

Corresponding Author:
Sinta Rukiastiandari,
Teknologi Informasi,
Universitas Bina Sarana Informatika,
Jln. Kramat Raya No. 98, Senen, Jakarta Pusat, DKI Jakarta, 10450,
Email: sinta.sru@bsi.ac.id

INTRODUCTION

In today's digital era, data has become a crucial asset in various fields, including education. In the academic context, the ability to predict student graduation outcomes is an important topic in higher education and offers benefits in planning more effective education. It also aids in early intervention for students who may face difficulties. Conventional methods for predicting graduation outcomes are often limited in their accuracy and flexibility. These models typically struggle to adjust to the diversity and complexity of educational data, as well as the increasing amount of educational data. Therefore, there is a need for more sophisticated approaches that can more effectively handle the peculiarities of educational data. The application of machine learning techniques such as Decision Trees (DT), Neural Networks (NN), and Support Vector Machines (SVM) has shown significant progress in predicting student graduation outcomes.

Related research has demonstrated great potential in using machine learning algorithms for predicting student graduation. For example, previous studies showed an accuracy of 84.96% using DT, while NN achieved 84.68%, and SVM showed an improvement to 85.18% (Riyanto et al., 2019). Further research integrating SVM with PSO achieved an accuracy of 85.84%, and the combination of SVM with PSO (Particle Swarm Optimization) resulted in a higher accuracy of 86.57% (Suhardjono et al., 2019). Additionally, the integration of SVM and GA achieved 86.43% (Ridwansyah et al., 2020), while the combination of DT and PSO reached 87.56% (Hendra et al., 2020). Research by Nurdin with NN and PSO recorded an accuracy increase to 86.94% (Nurdin et al., 2023), and the most recent study with the same dataset using NN and GA reached 87.33% (Pangesti et al., 2024).

From various studies on student graduation, conventional methods in predicting graduation

DOI: https://doi.org/10.31294/p.v26i1.3165

outcomes often face limitations in accuracy and flexibility. These models usually encounter challenges in adapting to the diversity and complexity of educational data. Therefore, there is a need for more advanced approaches that can handle the specifics of educational data more effectively (Nawawi, 2024). This research focuses on improving the Decision Tree model by integrating the Genetic Algorithm (GA) to enhance prediction accuracy.

Decision tree models have proven to be a useful tool in predictive analysis due to their ease of interpretation and ability to handle categorical and numerical data (Priyatama & Ridwansyah, 2022). However, these models can suffer from overfitting and may not be optimal in terms of feature selection and tree structure (Iqbal et al., 2020). To address these limitations, this research introduces the use of the genetic algorithm (GA), inspired by the biological evolutionary process, as a method to enhance the decision tree model. The genetic algorithm provides a framework for heuristic optimization that can be used to improve various aspects of the decision tree model, including feature selection and parameter tuning (Sartini et al., 2023).

This research proposes innovation and develops an enhanced model by combining the strengths and integrating the Genetic Algorithm into the Decision Tree model to predict student graduation. The primary goal is to improve prediction accuracy through more efficient and more accurate feature selection and parameter optimization. With this approach, the research is expected not only to enhance the accuracy of graduation prediction but also to provide new insights into the application of evolutionary techniques in educational data analysis, and to offer an adaptable approach for similar predictive challenges in other fields.

The main contributions of this research include the development of an enhanced student graduation prediction model with improved accuracy and AUC, a comprehensive evaluation of the DT model's performance, and a practical demonstration of the GA model's application in an educational context, which has not been done previously. It is hoped that the findings from this research can assist educational institutions in making more accurate and beneficial decisions for their students.

RESEARCH METHOD

Figure 1 flowchart depicts the research process for predicting graduation results using the Decision Tree model enhanced with the Genetic Algorithm. The following is an explanation of the flow.

1. Preparation of Research Materials

This step involves collecting data such as exam scores, attendance, classroom participation, which are combined into a Cumulative Grade Point Average (CGPA) from Semester 1 to 6, and other factors for research. In this case, the collected data is a student graduation dataset, likely containing information about the students and their graduation outcomes.

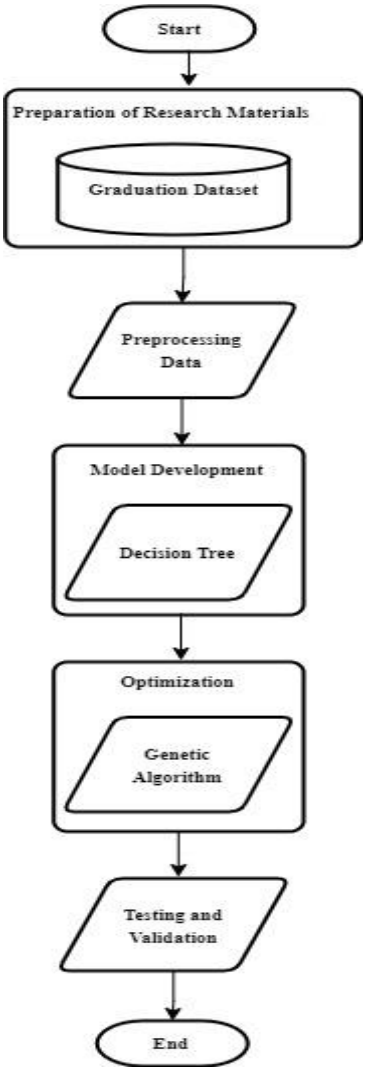

Figure 1. Research Flow Chart

2. Data Preprocessing

Before utilizing the data for modeling, this step involves cleaning and preparing the data. This may include removing incomplete data, addressing missing data, changing data formats, normalization, and cleansing the data of missing or inconsistent values (Wijaya, 2024). Additionally, normalization or standardization of data is performed as needed to ensure the data is ready for use in machine learning models.

3. Model Development

In this phase, the Decision Tree model is constructed. Decision Trees are machine learning methods used for classification and regression that model predictions in the form of a decision tree (Ridwansyah et al., 2022). This model divides data into branches to make predictions by determining criteria and other parameters (Ariyati et al., 2020).

4. Optimization

After building the Decision Tree model, the next step is to optimize it using a Genetic Algorithm.

Genetic Algorithms are heuristic search techniques that mimic the natural evolutionary process to find optimal solutions (Suhardjono et al., 2023). In this context, Genetic Algorithms are used to find the best parameters for the Decision Tree model to make the predictions more accurate. This includes initializing the initial population with various Decision Tree parameters and setting a fitness function based on accuracy criteria (Adibi, 2019).

5. Testing and Validation

Post-optimization, the resultant model is tested and validated to assess its performance. This involves using metrics such as accuracy, Area Under Curve (AUC) (Priyatama & Ridwansyah, 2022), to determine how well the model predicts graduation outcomes..

RESULTS AND DISCUSSION

1. Preparation of Research Materials

The data used in this study was obtained from the academic records of students from various high schools. The graduation data includes 796 records with 10 attributes, covering variables such as gender (JK), high school major (JRS), type of high school (SLTA), and the origin of the high school, among others. Additionally, this study utilized historical data of cumulative grade point averages (CGPA) from the first to the sixth semester to assess academic performance. The success in completing education on time was also recorded to observe its correlation with the aforementioned factors. This data will be used in a series of statistical analyses using the Decision Tree (DT) method and accuracy enhancement methods to be optimized with the Genetic Algorithm (GA).

2. Data Preprocessing

Prior to data modeling, the data underwent a process of replacing missing values with average values and removing irrelevant duplicate data to ensure data validity. This also involved changing data formats, normalization, and cleaning the data of missing or inconsistent values. The results of this process can be seen in Table 1, which displays an example of the student graduation data used in this study.

Table 1. Data Preprocessing Results

GENDER	HISCH DEP	FROM HISCH	IPK1	IPK2	IPK3	IPK4	IPK5	IPK6	ON TIME
1	5	3	2	2,12	2,41	2,05	2,12	2,95	NO
1	4	2	3,59	3,46	3,52	3,54	3,54	3,56	YES
2	7	3	3,14	3,51	3,51	3,57	3,64	3,64	YES
1	5	2	2,5	2,59	2,67	2,65	2,65	2,77	NO
1	9	3	2,59	2,71	3,03	3,07	3,12	3,22	YES
1	10	5	2,45	2,2	1,97	1,75	2,84	2,83	YES
2	5	4	1,95	2,39	2,46	2,51	2,64	3	NO
1	5	3	2,91	2,93	2,87	3,01	3,15	3,21	YES
1	4	3	2,68	2,66	2,84	2,83	3,1	3,09	YES
1	6	3	2,5	2,46	2,13	2,16	2,19	2,42	NO
2	1	3	3,23	3,2	3,2	3,31	3,38	3,45	YES
2	4	3	2,77	2,73	2,62	2,59	3,09	3,12	YES
1	6	3	2,5	2,46	2,13	2,16	2,19	2,42	NO
…	…	…	…	…	…	…	…	…	…
2	1	3	3,23	3,2	3,2	3,31	3,38	3,45	YES

From Table 1, it can be explained that the attribute 'JK' represents gender, where '1' signifies male and '2' signifies female. 'JRS_SLTA' refers to the high school major, indicating the student's field of study in high school, and 'asal SLTA' represents the high school the student attended before entering university. 'IPK1-IPK6' represents the Grade Point Averages (GPAs) for semesters 1 through 6

3. Model Development

Based on the information listed in table 1, the data testing and validation process will be carried out through the application of the Decission Tree (DT) model using Rapid Miner software. This process will involve a 10 times cross validation method (10 cross validation) to ensure the accuracy and reliability of the results. After using the DT method, you can produce a decision tree from the DT method.

Next, the results of the DT model decision tree. After testing the DT model, it will produce a confusion matrix, the details of which can be seen and analyzed in table 2. This approach is expected to provide a deeper understanding of the performance and effectiveness of the DT model used in this research.

Table 2. Confussion Matrix DT

	TP	TN
FP	596	72
FN	38	90

In Table 2, the results of the confusion matrix for student graduation are presented, utilizing the

decision tree method. The number 596 represents students who graduated on time as predicted, while 72 did not match the prediction for on-time graduation. The count of 90 represents the predicted number of students graduating late, and the result matches the prediction. However, 38 did not align with the prediction for late graduation. Consequently, from Table 2, the accuracy rate reaches 86.19%. Additionally, this will result in a graph depicting student graduation rates with an AUC curve, which can be seen in Figure 2.

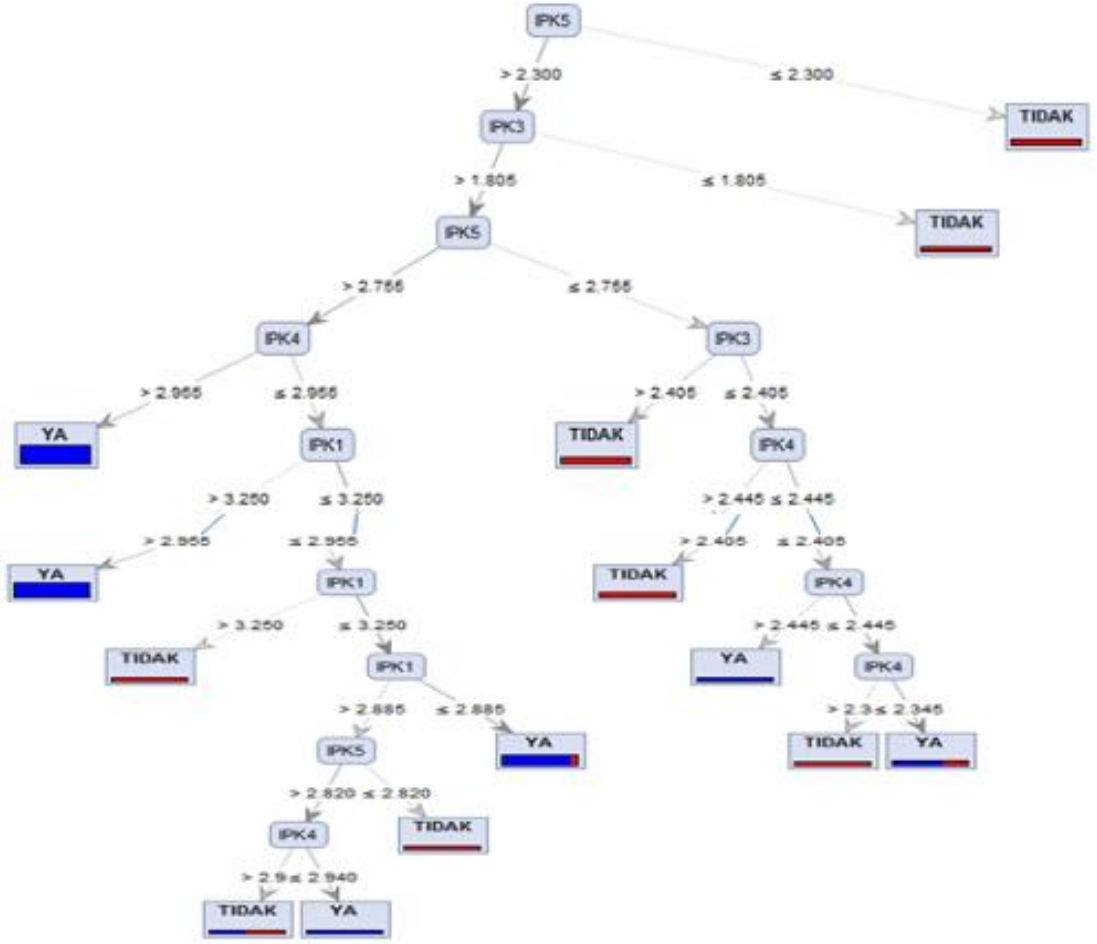

Figure 2. Decision Tree Model for Student Graduation Data

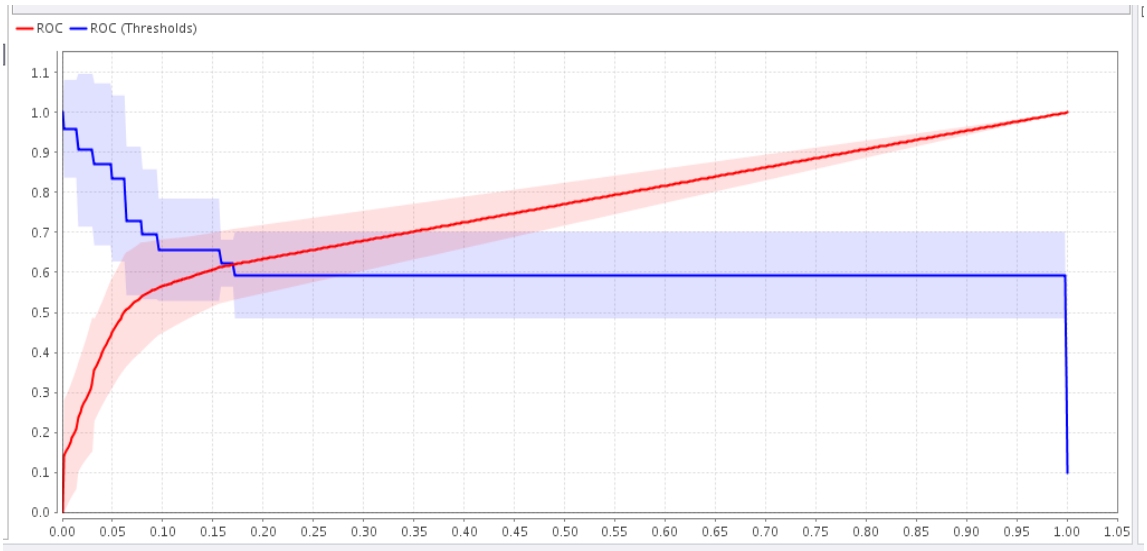

Figure 3. AUC Decision Tree

Figure 3 illustrates that the curve generated from the student graduation dataset using the decision tree method results in a fairly good model, with an AUC value of 0.755%.

4. Optimization

After the graduation data is processed using the DT method, it will be optimized using the GA method with the same treatment as the previous method, which is the DT method. This will yield a confusion matrix, the results of which can be seen in Table 3.

Tabel 3. Confussion Matrix Optimization GA

	TP	TN
FP	596	60
FN	38	102

Table 3 presents the results of the confusion matrix for student graduation predictions using the genetic algorithm optimization method. The number 596 represents the count of accurate on-time graduation predictions, while 60 represents the count of inaccurately predicted on-time graduations. The number 102 indicates the accurate predictions of students graduating late, but there are 38 cases where the late graduations were not predicted accurately. As a result, Table 3 shows an accuracy rate of 87.68%. Additionally, this will produce a graph of student graduation rates with an AUC curve, which can be seen in Figure 4.

Figure 4 depicts that the curve generated from the student graduation dataset using the genetic algorithm optimization method results in a fairly good model, with an AUC value of 0.788%.

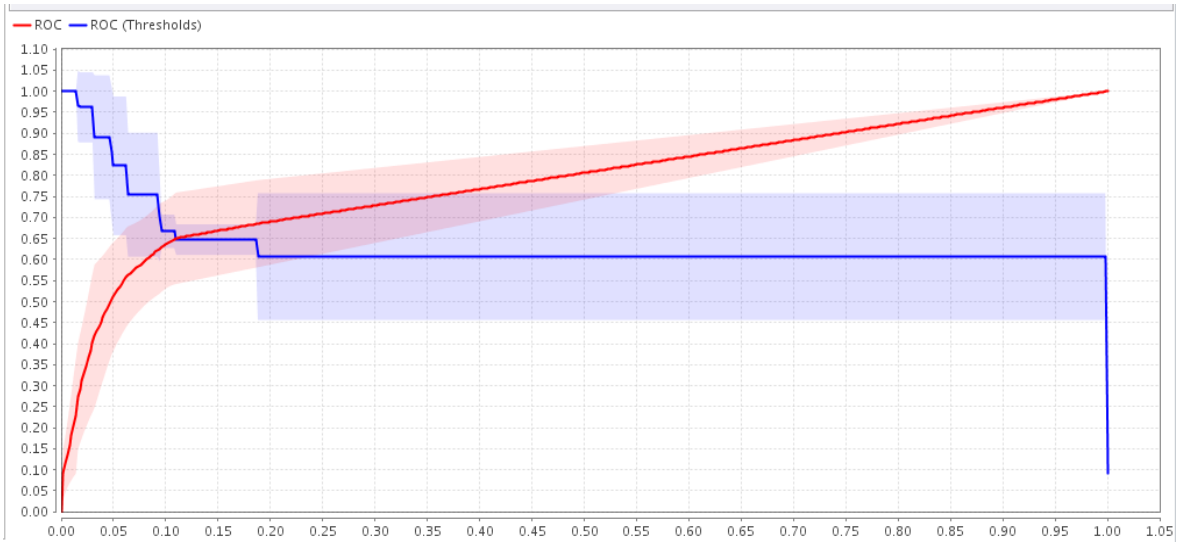

Figure 4. AUC Decision Tree

5. Testing and Validation

There is a significant improvement in accuracy and AUC in the student graduation data analyzed using the DT method combined with GA, as evident in Table 4. The performance of each method is evaluated based on specific metrics to measure the quality of the model, with a comparison of accuracy and AUC as the indicators.

Table 4. Comparison of Method Performance

Algorithm	DT	DT Optimization GA
Accuracy	86.19%	87.68%
AUC	0.755%	0.788%

In the presented method performance comparison table, it is evident that the research has successfully demonstrated an improvement in the performance of the DT algorithm after optimization using GA. This improvement is observable through two evaluation metrics: accuracy rate and AUC. The results of this matrix are further illustrated in a comparative value graph depicted in the figure.

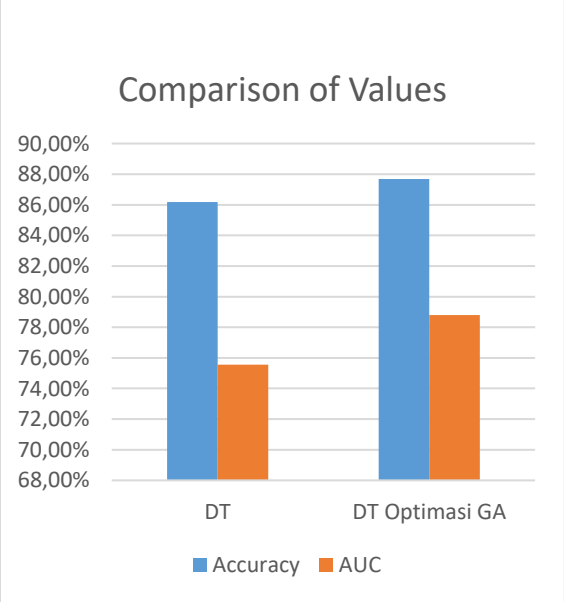

Figure 5. Comparative Graph

In Figure 5, the unoptimized DT algorithm has an accuracy rate of 86.19% and an AUC value of 0.755%. However, after the optimization process with GA, the accuracy rate increased to 87.68%, and the AUC value improved to 0.788%.

CONCLUSION

The conclusion drawn from the study is that the testing successfully proved that the application of machine learning techniques, such as the Decision Tree (DT) optimized with the Genetic Algorithm (GA), enhances the accuracy of student graduation predictions. These two metrics are crucial in assessing the performance of predictive models. The accuracy, which indicates the percentage of successful predictions made by the model, increased from 86.19% to 87.68%. Additionally, the improvement in the AUC value from 0.755% to 0.788% reflects the model's enhanced ability to distinguish between positive and negative classes accurately. This outcome signifies that integrating the Genetic Algorithm into the Decision Tree model positively contributes to the model's effectiveness in predicting graduation outcomes. From a future development perspective, this research paves the way for further applications and the development of evolutionary techniques in educational data analysis. This innovation not only improves prediction accuracy in the educational context but also offers an adaptable approach for similar predictive challenges in other fields.

REFERENCES

Adibi, M. A. (2019). Single and multiple outputs decision tree classification using bi-level discrete-continues genetic algorithm. *Pattern Recognition Letters*, *128*, 190–196. https://doi.org/10.1016/j.patrec.2019.09.001

Ariyati, I., Rosyida, S., Ramanda, K., Riyanto, V., Faizah, S., & Ridwansyah. (2020). Optimization of the Decision Tree Algorithm Used Particle Swarm Optimization in the Selection of Digital Payments. *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012090

Hendra, Azis, M. A., & Suhardjono. (2020). Analisis Prediksi Kelulusan Mahasiswa Menggunakan Decission Tree Berbasis Particle Swarm Optimization. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *9*(1), 102–107. https://doi.org/https://doi.org/10.32736/sisfokom.v9i1.756

Iqbal, M., Herliawan, I., Ridwansyah, Gata, W., Hamid, A., Purnama, J. J., & Yudhistira. (2020). Implementation of Particle Swarm Optimization Based Machine Learning Algorithm for Student Performance Prediction. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *6*(2), 195–204. https://doi.org/10.33480/jitk.v6i2.1695.IMPLEMENTATION

Nawawi, I. (2024). Optimisasi Pemilihan Fitur Untuk Prediksi Gagal Jantung: Fusion Random Forest Dan Particle Swarm Optimization. *Inti*, *18*(2).

Nurdin, H., Sartini, Sumarna, Maulana, Y. I., & Riyanto, V. (2023). Prediction of Student Graduation with the Neural Network Method Based on Particle Swarm Optimization. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *8*(4), 2353–2362. https://doi.org/10.33395/sinkron.v8i4.12973

Pangesti, W. E., Ariyati, I., Priyono, Sugiono, & Suryadithia, R. (2024). Utilizing Genetic Algorithms To Enhance Student Graduation Prediction With Neural Networks. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *9*(1), 276–284. https://doi.org/https://doi.org/10.33395/sinkron.v9i1.13161 e-ISSN

Priyatama, I. M. D., & Ridwansyah. (2022). Klasifikasi Anak Berkebutuhan Khusus Tunagrahita Menggunakan Metode Algoritma C4.5. *Paradigma*, *24*(1), 90–95. https://doi.org/https://doi.org/10.31294/paradigma.v24i1.1087

Ridwansyah, R., Riyanto, V., Hamid, A., Rahayu, S., & Purnama, J. J. (2022). Grouping Data in Predicting Infant Mortality Using K-Means and Decision Tree. *Paradigma*, *24*(2), 168–174. https://doi.org/10.31294/paradigma.v24i2.1399

Ridwansyah, R., Wijaya, G., & Purnama, J. J. (2020). Hybrid Optimization Method Based on Genetic Algorithm for Graduates Students. *Jurnal Pilar Nusa Mandiri*, *16*(1), 53–58. https://doi.org/10.33480/pilar.v16i1.1180

Riyanto, V., Hamid, A., & Ridwansyah. (2019). Prediction of Student Graduation Time Using the Best Algorithm. *Indonesian Journal of Artificial Intelligence and Data Mining*, *2*(2), 1–9. https://doi.org/http://dx.doi.org/10.24014/ijaidm.v2i1.6424

Sartini, S., Rohimah, L., Maulana, Y. I., Supriatin, S., & Yuliandari, D. (2023). Optimization of Random Forest Prediction for Industrial Energy Consumption Using Genetic Algorithms. *PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic*, *11*(1), 35–44. https://doi.org/10.33558/piksel.v11i1.5886

Suhardjono, S., Sudradjat, A., Wahid, B. A., Sugiarto, H., & Nurdin, H. (2023). Prediction Of Infant Mortality Using The Decission Tree And Genetic Algorithm Methods. *Paradigma*, *25*(1). https://doi.org/https://doi.org/10.31294/p.v25i1.1819

Suhardjono, Wijaya, G., & Hamid, A. (2019). Prediksi Waktu Kelulusan Mahasiswa Menggunakan SVM Berbasis PSO. *Bianglala Informatika*, *7*(2), 97–101. https://doi.org/https://doi.org/10.31294/bi.v7i2.6654.g3731

Wijaya, G. (2024). *Improvement of Kernel SVM to Enhance Accuracy in Chronic Kidney Disease*. *9*(1), 136–144. https://doi.org/https://doi.org/10.33395/sinkron.v9i1.13112 e-ISSN