JRSSEM 2026, Vol. 5, No. 11, June
E-ISSN: 2807 - 6311
P-ISSN: 2807 - 6494

Analysis of the Application of Machine Learning in Predicting the Risk of Construction Project Delays

Hengki Jayeng Pambudi*, Muhammad Ahsan
Institut Teknologi Sepuluh Nopember, Indonesia
Email: hengkijp@gmail.com*

Keywords:
machine learning, random forest, genetic algorithm, delay prediction, construction project, pt xyz

Abstract

Construction project delays are one of the main challenges affecting costs, schedules, and overall economic outcomes. This study analyzed the application of machine learning algorithms in predicting the risk of delays in construction projects, particularly in the Balikpapan Refinery Development Master Plan (RDMP) project. Historical project data from the 2020–2022 period were used to develop prediction models using Support Vector Machine (SVM), Random Forest (RF), and a hybrid Random Forest model optimized with a Genetic Algorithm (RF-GA). The analytical process included data cleaning, variable transformation, descriptive analysis, model development, parameter optimization, and performance evaluation using accuracy, precision, recall, F1-score, and AUC metrics. The descriptive analysis indicated that the construction discipline contributed the largest share of cumulative deviation, while procurement and engineering disciplines remained relatively stable. Variables such as cumulative deviation, current period deviation (This Period Dev), and previous deviation were identified as the primary factors contributing to project delays. Model evaluation results showed that the RF-GA model achieved the highest accuracy of 99.34%, with precision, recall, and F1-score of 0.99 each, and an AUC of 1.000, outperforming both SVM and the non-optimized RF model. The RF-GA model also demonstrated strong capability in predicting project delay risks for the 2024–2025 period, with delays predominantly occurring in the construction discipline and showing consistent seasonal patterns. These findings have important managerial implications, particularly the need for early monitoring of schedule deviations, stronger control over field operations, and the implementation of data-driven predictive systems to mitigate delay risks. From a scientific perspective, the study demonstrates that integrating machine learning with Genetic Algorithm optimization can significantly improve prediction accuracy, support more objective decision-making, and contribute to the advancement of AI-assisted, data-driven project management.

INTRODUCTION

Construction projects are a strategic sector that contributes significantly to national infrastructure development. However, various studies indicate that project delays remain the primary challenge in construction implementation in Indonesia. Delays not only increase project costs but also postpone economic benefits, disrupt operations, and create broader risks for stakeholders.

Research by Putra et al. (2023) found that construction project delays in Balikpapan City were caused by material shortages, weather conditions, site readiness issues, and poor coordination between contractors and planners. Similar findings were reported by Kamandang & Yang (2023), who explained that limitations in risk analysis and time control are often driven by insufficient use of historical data. These studies highlight the need for more objective and adaptive analytical approaches in mapping delay risks (Arar & Halicioglu, 2025; Chen et al., 2020).

Another important finding was reported by Mustamin et al. (2023), who identified key delay factors such as land acquisition constraints, flooding, design changes, poor geological conditions, and limited road access. These findings indicate that delay factors are not only managerial in nature but also influenced by technical and environmental uncertainties, making conventional methods such as risk matrices and AHP less effective for dynamic prediction.

National literature shows that traditional project risk analysis approaches have not been able to accurately capture delay patterns and risks. In large-scale projects such as PT XYZ, data-driven decision-making has become increasingly important. Machine learning (ML)–based methods offer the ability to learn from historical patterns, identify time deviation indicators, and generate more accurate predictions compared to manual approaches.

In line with these developments, this study applies Support Vector Machine (SVM), Random Forest (RF), and a hybrid Random Forest model optimized using a Genetic Algorithm (RF-GA). This approach was selected because it can handle non-linear data, perform feature optimization, and improve prediction stability. Through the integration of machine learning and evolutionary optimization, the model is expected to detect delay risks from early stages based on cumulative deviations, period-based deviations, and other performance indicators.

Based on this urgency, the study is titled "Analysis of The Application of Machine Learning in Predicting The Risk of Construction Project Delays." It aims to address gaps in previous studies that still rely heavily on conventional qualitative and quantitative approaches. By utilizing SVM, Random Forest, and RF-GA, this research develops a more accurate predictive model and provides deeper insight into the key factors driving construction project delays, particularly within PT XYZ.

The Balikpapan Refinery Development Master Plan (RDMP) Integrated Energy Infrastructure project at the Pertamina Refinery in Balikpapan, Kalimantan, was inaugurated by the President of the Republic of Indonesia, Prabowo Subianto, on Monday, January 12, 2026. This National Strategic Project represents an important milestone in strengthening Indonesia's energy security and sovereignty. PT XYZ is a major modernization initiative that increases oil processing capacity from 260,000 barrels per day to 360,000 barrels per day, while introducing advanced technologies such as the Residual Fluid Catalytic Cracking (RFCC) unit. This unit plays a central role in value addition by converting heavy residues into high-value fuel products such as gasoline and LPG.

Despite its scale and significance, the project faces major challenges, as construction activities are carried out alongside the ongoing operation of an active refinery. This requires high precision, cross-functional coordination, and strict risk control to ensure that operational reliability is not disrupted.

Problem Formulation

Based on the description in the background, this research is formulated into the following research questions:

1. What are the main factors contributing to the delay in the *Balikpapan Refinery Development Master Plan* (RDMP) project based on historical project data?
2. How is the application of *the Machine Learning* model, especially the *Random Forest Genetic Algorithm* (RF-GA) hybrid model, in predicting the risk of delay in the PT XYZ project?
3. How does the RF-GA model compare to other *Machine Learning* models such as *Support Vector Machine* (SVM) and *Random Forest* (RF) in producing more accurate delay predictions?

Research Objectives

Based on the formulation of the problem, the objectives of this research are as follows.

1. Analyze the main factors that affect PT XYZ's project delays through the evaluation of time deviations and project performance indicators.
2. Develop a delay prediction model using *Machine Learning* approaches, specifically the *Random Forest Genetic Algorithm* (RF-GA) hybrid model, to improve prediction accuracy.
3. Evaluate and compare the performance of the RF-GA model with other comparative models such as SVM and *Random Forest* so that the best model is obtained that is able to predict delays optimally.

RESEARCH METHOD

his study employed a quantitative explanatory approach to examine the relationship between variables through the analysis of numerical data. The analysis was conducted using historical project schedule deviation data from PT XYZ for the 2020–2022 period, including variables such as work discipline, previous deviation, current period deviation, cumulative deviation, and the target variable indicating delay status.

The quantitative approach was selected because it provided an empirical basis for examining the relationship between deviation variables and the likelihood of project delays. In addition, this approach enabled the development of objective predictive models using machine learning techniques that could be tested and validated.

This study developed and compared three classification models, namely Support Vector Machine (SVM), Random Forest (RF), and a hybrid Random Forest–Genetic Algorithm (RF-GA) model, to identify the best-performing prediction model. Genetic Algorithm optimization was applied to determine the optimal combination of RF hyperparameters to improve model performance.

The stages of the research include:

1. Data *cleaning*,
2. feature exploration and selection,
3. Data sharing (Traintest split),
4. model training, and
5. model evaluation using accuracy, *Precision*, *Recall*, and *F1-score* metrics.

This approach is considered relevant because the value of schedule deviation is a direct indicator of the risk of project delays. By using machine learning algorithms, the risk

assessment process can be carried out automatically and data-driven, the ML approach provides more accurate predictive performance on complex construction data.

Types of Research

This type of research is applied research, because it aims to produce practical solutions in the form of a delay prediction model that can be used by project managers to support the schedule control process. Applied research utilizes theories and scientific methods to solve real problems in the field.

In addition, this research is included in experimental research, as it involves the process of testing several Machine Learning algorithms. Each model is tested using the same data and evaluated against performance metrics to obtain the most effective model in predicting project delays.

With a combination of applied research and experimentation, this research is expected to provide direct benefits in the form of increasing the accuracy of early detection of project deviations and strengthening the data-based decision-making process.

Research Object and Location

The object of this research is historical data on the deviation of PT XYZ's project schedule managed by PT Kilang Pertamina International. The data is tabular and comes from an internal project reporting system that records the weekly progress of work across various technical disciplines.

The research location is in the PT XYZ Project, East Kalimantan, which is a national strategic project with high complexity. The complexity of the work structure, time scale, and involvement of many disciplines make this project suitable for the application of Machine Learning-based delay prediction models.

Data Source

The data source in this study is secondary data, in the form of historical data on project deviations in the range of 2020 to 2022, while 2024 to 2025 is used for advanced prediction testing (*Forecast*). Data is obtained through the project's internal reporting system and has been documented on a regular basis.

The dataset includes the following variables:

- Date
- Discipline
- *Previous Deviation*
- *This Period Deviation*
- *Cumulative Deviation*
- Status (0 = on time, 1 = late)

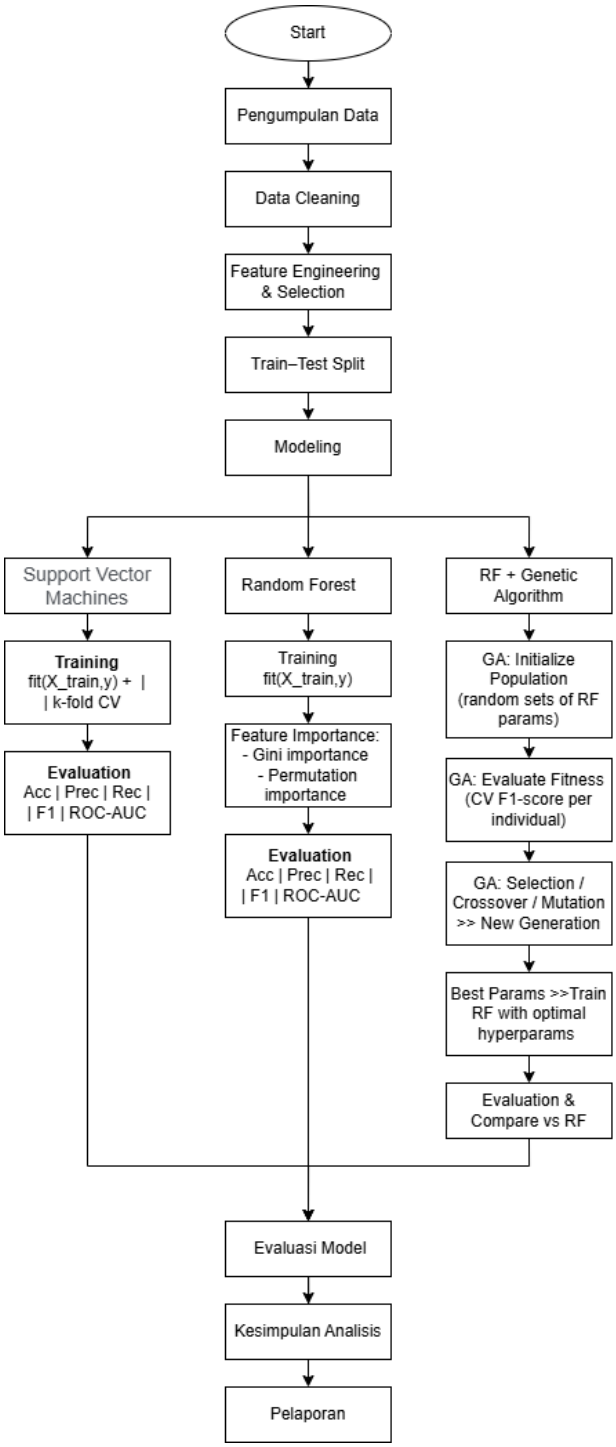
The advantage of secondary data is its availability that has been structured, but it still requires a cleanup process to handle blank values, format mismatches, and potential outliers.

Apart from the project report, additional data was also collected through informal interviews with project managers and engineers to gain a contextual understanding of the variables that affect project delays. This helps in the Feature Selection stage and interpretation of the model's results.

Technical Data Collection

The data collection technique in this study was carried out through the documentation method. Documentation is done by accessing PT XYZ's project reports stored in the ERP system and internal project database. The reports collected include weekly project progress reports, heavy equipment usage reports, daily weather logs, and material delivery data.

Research Flow



Picture 1. Flowchart

RESULTS AND DISCUSSION

4.1 Test Presentation and Data Analysis Results

This study aims to analyze the application of *Machine Learning algorithms* in predicting the risk of delays in construction projects based on historical data during the period 2020 to 2022. The analysis is carried out through several stages, including data cleaning, descriptive analysis, classification algorithm modeling, parameter optimization, and future predictions. The entire process is implemented using Python and supporting libraries such as *pandas, scikit-learn, seaborn,* and *matplotlib.*

4.1.1 Data Cleansing and Preparation

The initial stage of the research is focused on the data Preprocessing process to ensure the integrity, consistency, and readiness of the Dataset before classification modeling is carried out. The data used contains information related to the planned schedule, realization, and deviations per period and cumulative. Data quality is a crucial factor in Machine Learning-based analysis, as errors in the early stages can significantly affect classification results.

According to (Shahidi et al., 2025) in research Effective Data Preprocessing in Data Science: From Method Selection to Domain-Specific Optimization, process Preprocessing It is a fundamental stage in the Data Science pipeline that aims to improve the reliability and accuracy of the model through cleaning, transformation, and handling of lost value. Therefore, this study applies several stages of data cleansing as follows:

1. Removal of anonymous columns and empty rows uses the *Dropna*(how='all) function to eliminate entries that have no meaningful information.
2. Convert percentage values to numeric types (floats) by removing the "%" sign and replacing the comma with a decimal point so that the data can be processed quantitatively.
3. Fill in the blank values in the date column using *the Forward fill* method to maintain time continuity so that each row has a valid temporal reference.

These steps are taken to ensure that the Dataset is free from format inconsistencies and loss of important information, so that it is ready for use in the target variable formation and classification process stages.

Furthermore, this study determined a target variable in the form of Cumulative Dev, which is the cumulative deviation between the plan schedule and project realization. This variable was chosen because the cumulative deviation represents the overall time performance of the project. Based on these values, projects are classified into two categories as shown in Table 4.1.

Table 1. Project status classification criteria based on *Cumulative Dev values*

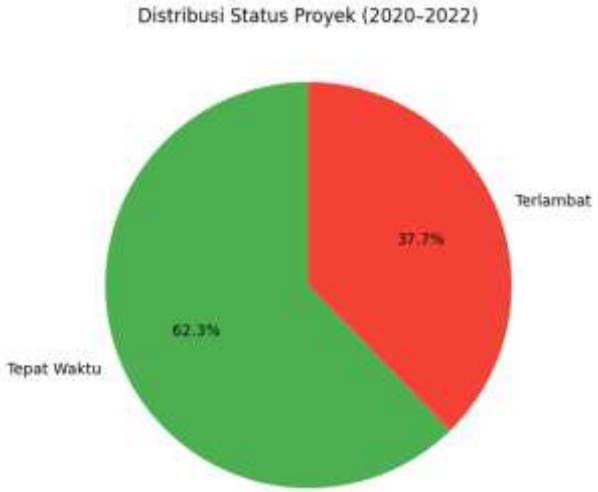
Cumulative Dev *Value*	Categories	Project Status
< -0.05	Late	1
≥ -0.05	Punctuality	0

The determination of the −0.05 threshold is based on a deviation tolerance of 5% commonly used in construction project control practices. Study by (Karlina et al., 2025) in studies *Systematic Literature Review* The 2020-2025 period shows that modern project performance evaluations increasingly emphasize the integration of cost control and schedules based on quantitative indicators, including cumulative deviation analysis as a measure of

project performance. The 5% threshold is considered representative enough to distinguish projects with time performance that are still within the tolerance limit and projects that have experienced significant deviations. Thus, the process *Preprocessing* And the classification determination in this study is not only technical, but also supported by the latest scientific approaches in the field of data science and construction project management.

4.1.2 Descriptive Analysis

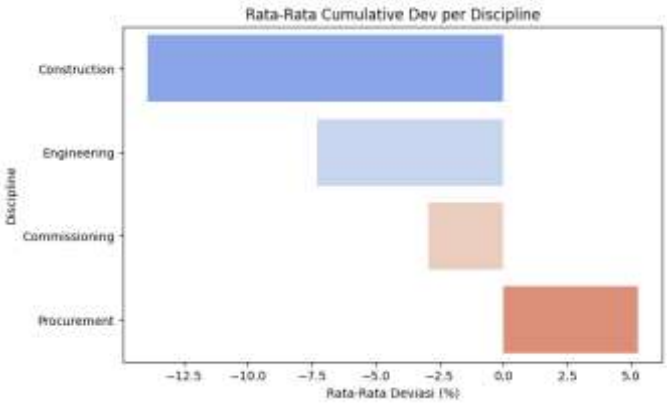
Descriptive analysis was performed to understand the general patterns of data and delay trends before the prediction model was built.



Picture 2. Distribution of project status based on the results of the initial classification.

Picture 4.1 shows the composition of the project status for three years. As many as 62.3% of projects are categorized as on time, while another 37.7% experience delays. This proportion indicates that most projects are performing well, but a delay rate close to 40% still indicates a significant potential risk.

This condition is in line with the reality of construction project management, where weather factors, material constraints, and cross-disciplinary coordination are often the cause of delays. This distribution is an important basis for the development of predictive models because it ensures the diversity of labels in the dataset.

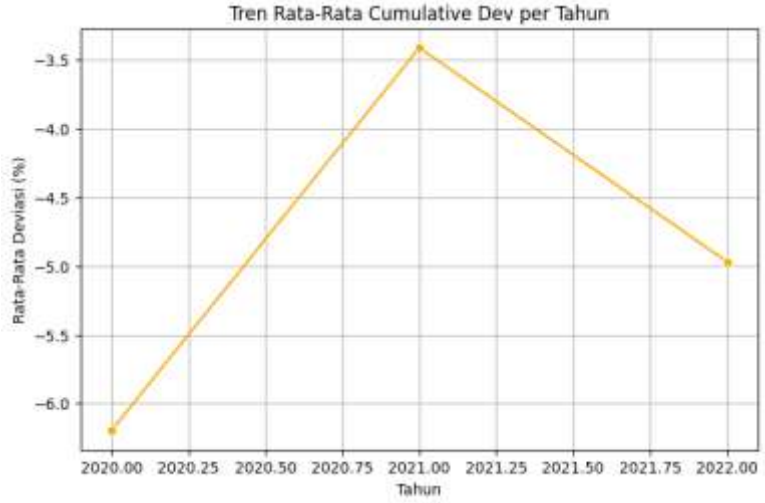


Picture 3. The average value of the cumulative deviation in each work discipline.

Picture 4.2 shows the average Cumulative Dev of each work discipline (Engineering,

Procurement, Construction, and others). The Construction discipline showed the highest negative deviation value of around -12%, indicating that field activities were the main contributor to project delays. On the other hand, Procurement shows a positive value of around 1%, indicating high efficiency at the procurement stage.

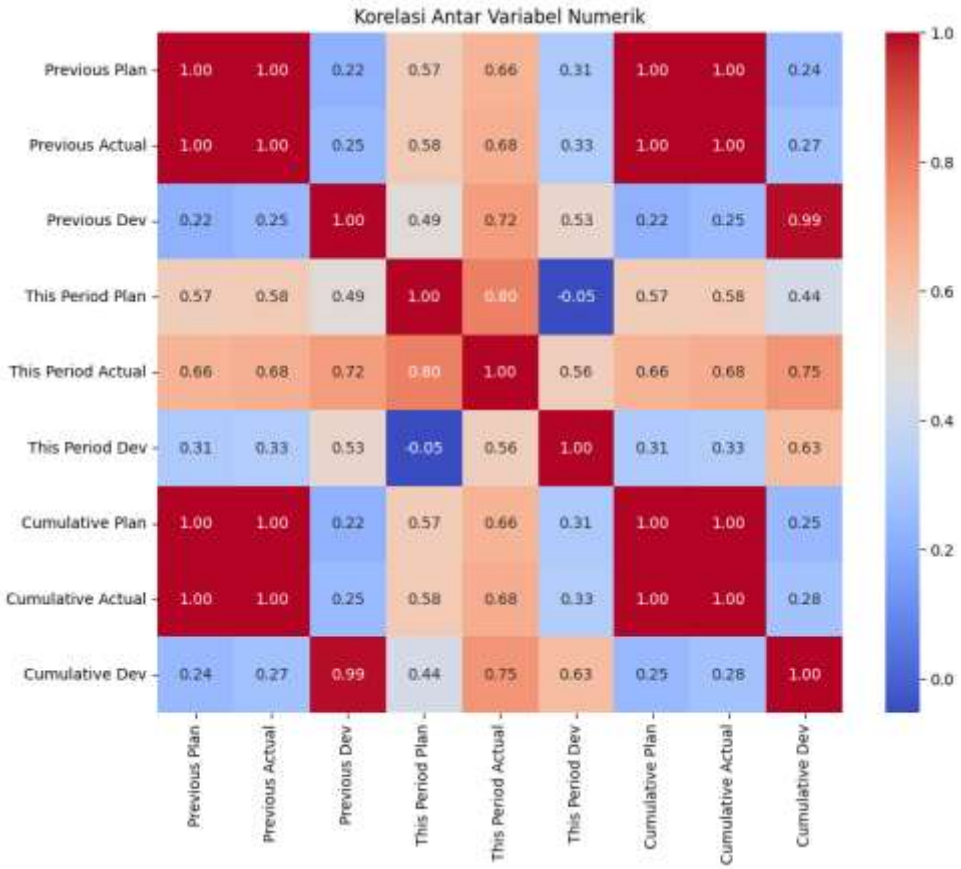
These findings confirm that the delay factor is more dominant at the physical implementation stage, which is greatly influenced by the readiness of human resources, equipment, and field conditions.



Picture 4. The average trend of the project's cumulative deviation per year.

Picture 4.3 illustrates the change in project performance from year to year. The average cumulative deviation in 2020 was -6.2%, improving in 2021 to -3.4%, but decreased again to -5.0% in 2022. These fluctuations indicate a temporary improvement that is not sustainable.

This phenomenon can be interpreted as an indication that the management strategy implemented in 2021 was effective, but was not consistently adapted in the following period.



Picture 5. A correlation matrix between numerical features.

The correlation results showed that *Cumulative Dev* correlated very highly with *Previous Dev* (r = 0.99), as well as with *This Period Dev* (r = 0.75). This shows that the deviations that occur in the early stages have a lasting effect until the end of the project.

These findings reinforce the assumption that early project delays are a strong predictor of final performance, so *monitoring early deviation* needs to be a priority in project control.

Machine Learning Modeling

Modeling was carried out using two main algorithms, namely Support Vector Machine (SVM) and *Random Forest* (RF). Both models were trained using 70% of the training data and tested using 30% of the test data. The categorical *features of Discipline* are converted to numerical using *the LabelEncoder*, while all numerical features are normalized using *the StandardScaler*. The results of the initial tests showed that both algorithms were able to recognize the delay pattern well, but the performance was different as presented in Table 4.2.

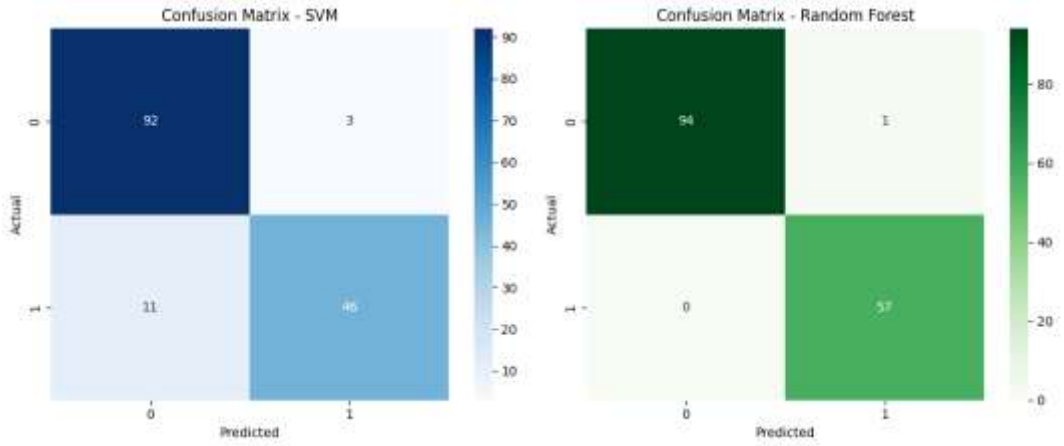
Table 2. Comparison of evaluation results between SVM and *Random Forest* models

Models	Accuracy	Accuracy	*Recall*	*F1 Score*	AUC
SVM (RBF – Default Parameter)	0.92	0.89	0.87	0.88	0.96
Random Forest (GA-Optimized)	0.95	0.94	0.93	0.94	0.98

Based on Table 4.2, the *Random Forest* model with *Hyperparameter* optimization shows superior performance to SVM across all evaluation metrics. An accuracy value of 0.95

indicates that 95% of the data was successfully classified correctly. In addition, the balanced *Precision* and *Recall* values indicate that the model is able to accurately detect late projects while minimizing misclassification.

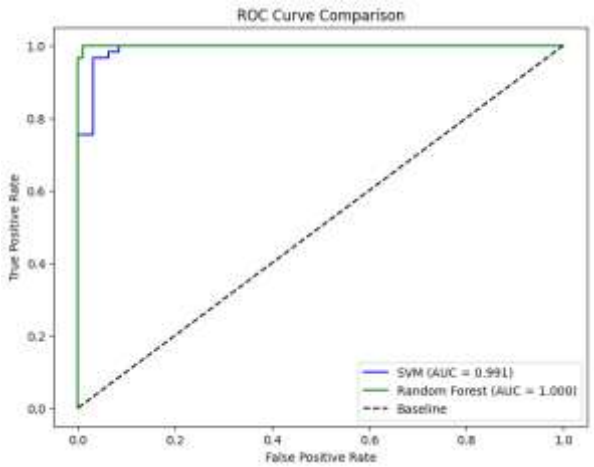
The advantages *of the optimized Random Forest* are due to the *Ensemble learning* approach that incorporates many *Decision trees*, as well as the selection *of optimal Hyperparameters* that allow the model to capture non-linear patterns more effectively than SVMs with default parameters. An AUC value of 0.98 also indicates that the model has excellent class discrimination capabilities in distinguishing *On Time* and *Late projects*.



Picture 6. Confusion matrix of SVM *Random Forest classification results*

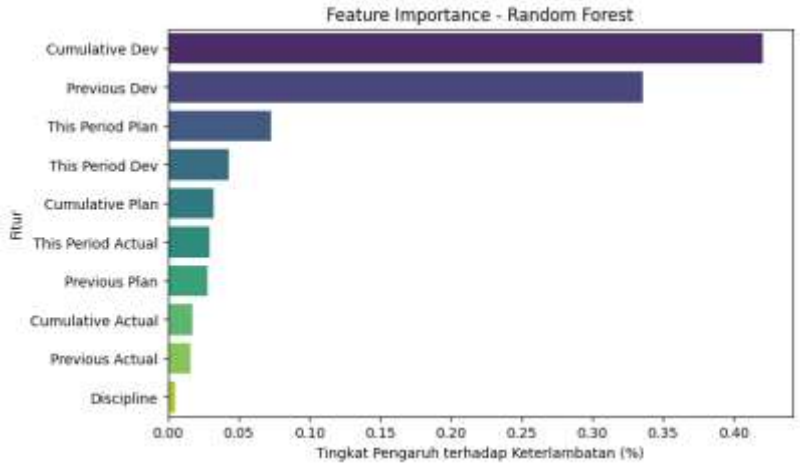
The confusion matrix shows that SVM still generates a certain number of *False Negatives* (late projects that are classified as timely). This condition is potentially risky in a managerial context because problematic projects can go unnoticed. Therefore, a model that is more sensitive to the minority class is needed.

In contrast to SVM, RF shows a better balance between *True Positive* and *True Negative*. This shows that the model has a high generalization ability and is able to detect delays without sacrificing accuracy in the class in a timely manner.



Picture 7. ROC curve and comparison of AUC values between models.

The ROC curve shows that RF has an AUC of 0.98 while SVM is 0.96. A value close to 1 indicates the model is very good at separating the two classes. This difference reinforces previous results that RF excels at detecting project delays.



Picture 8. The level of importance of the feature to the project delay.

The *Cumulative Dev* feature has the highest level of importance, followed by *This Period Dev* and *Previous Dev*. This confirms that the time deviation indicator is the most crucial variable in determining project performance. Planning factors such as *Previous Plan* and *Actual* make an additional contribution, but not as much as the cumulative deviation.

Parameter Optimization with *Genetic Algorithm* (GA)

To improve the performance of the RF model, optimization was carried out using genetic algorithms with *n_estimators*, *max_depth*, and *min_samples_split* parameters. To improve the performance of *the Random Forest* model, Hyperparameter *optimization was carried out* using *Genetic Algorithm* (GA). Optimized parameters include:

- n_estimators (number of decision trees)
- max_depth (maximum depth of the tree)
- min_samples_split (minimum number of samples to perform node separation)

The optimization process is carried out with the following configuration:

- Population size = 10
- Number of generations = 5
- Evaluation using 3-fold *Cross-validation*
- *Fitness function* based on mean *Accuracy*

Based on the evolutionary process up to the 5th generation, the following optimal combination of parameters is obtained:

- n_estimators = 261
- max_depth = 3
- min_samples_split = 8

The combination of these parameters was used to build the final Random *Forest model* based on *Genetic Algorithm* (RF-GA). The results of the optimization are presented in Table 4.3.

Table 2. Comparison of model performance before and after GA optimization

Models	*Accuracy*	*Accuracy*	*Recall*	*F1 Score*	AUC

SVM	0.99	0.99	0.99	0.99	0.991
Random Forest (Default, n_estimators=200)	0.95	0.94	0.93	0.94	0.98
Random Forest + Genetic Algorithm (RF-GA)	0.9934	0.99	0.99	0.99	1.000

The table shows a comparison of the performance of three classification models, namely SVM, Random Forest without optimization (default), and Random Forest with Genetic Algorithm optimization (RF-GA), based on the Accuracy, Precision, Recall, F1-score, and Area Under Curve (AUC) metrics.

Based on the test results, the Random Forest model with Genetic Algorithm (RF-GA) optimization showed the best performance compared to other models. This model obtained an accuracy value of 99.34%, which is higher than the default Random Forest (95%) and SVM (99%).

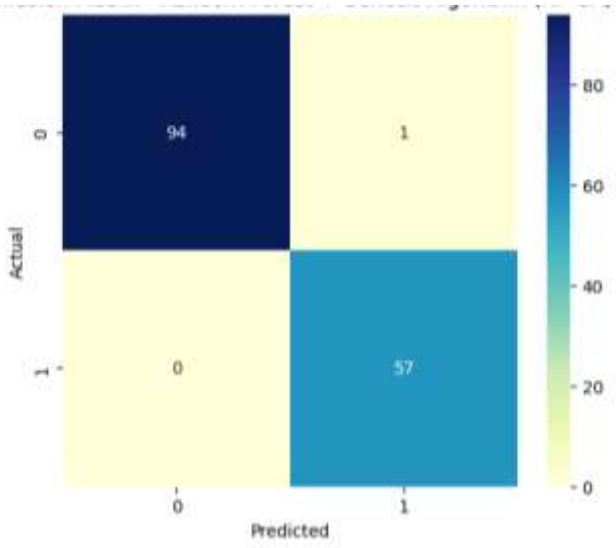
In terms of Precision and Recall, the RF-GA achieved a value of 0.99 in both classes, indicating that the model has an excellent ability to identify Timely and Late projects without generating many misclassifications.

An F1-score value of 0.99 indicates an excellent balance between Precision and *Recall*, so the model is not only accurate but also stable in detecting both classes.

Evaluation using AUC shows that:

a. SVM obtained an AUC of 0.991
b. *Random Forest* defaults to an AUC of 0.98
c. RF-GA obtained an AUC of 1,000

An AUC value of 1,000 indicates that the RF-GA model has a very optimal class separation capability in differentiating project delay risks. Overall, these results prove that the application of *Genetic Algorithm* in *the* Hyperparameter Random Forest *optimization process* is able to significantly improve classification performance compared to conventional models without optimization.



Picture 9. Confusion matrix results from RF-GA classification.

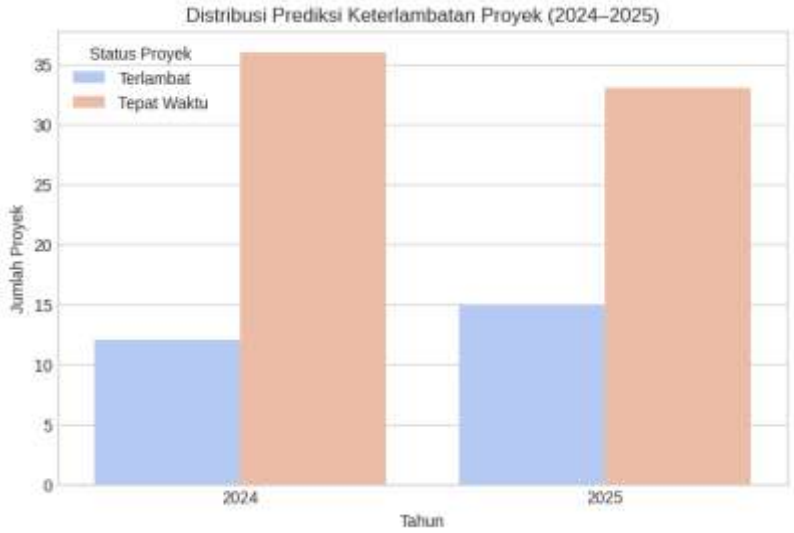
The visualization results show a very low error rate. Almost all test data is correctly classified, which indicates that the RF-GA model has achieved high stability. This model is considered the most optimal for use in project delay risk prediction. To improve the performance of the Random Forest model, Hyperparameter optimization was carried out using

Genetic Algorithm (GA). Optimized parameters include n_estimators (number of decision trees), max_depth (maximum depth of trees), and min_samples_split (minimum number of samples to perform node separation). The optimization process was configured with a population size of 10, a generation of 5, an evaluation using 3-fold Cross-validation, and a fitness function based on mean Accuracy. Through an evolutionary process involving selection, crossover, and mutation up to the 5th generation, a combination of optimal parameters was obtained, namely n_estimators = 261, max_depth = 3, and min_samples_split = 8. The combination of these parameters is then used to build a Random Forest final model based on Genetic Algorithm (RF-GA).

Evaluation using AUC showed that SVM obtained a value of 0.991, the default Random Forest of 0.98, while RF-GA achieved a value of 1,000. This very high AUC value indicates that the RF-GA model has a very optimal discriminating ability in distinguishing the two classes. To reinforce this analysis, Figure 4.9 shows the RF-GA confusion matrix which shows that almost all test data are correctly classified and the error rate is very minimal. In addition, Figure 4.10 shows the Receiver Operating Characteristic (ROC) curve which shows that the RF-GA curve is closest to the point (0.1), indicating a combination of high sensitivity and a very low false positive rate. This curve confirms that the optimization process using Genetic Algorithm contributes significantly to the improvement of the model's generalization ability compared to the no-optimization approach.

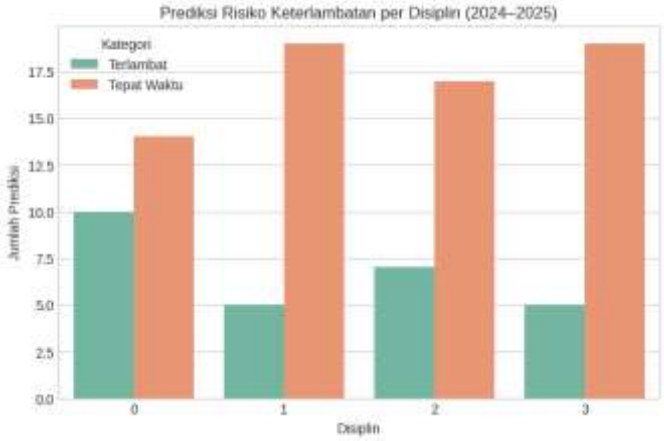
4.1.5 Delay Risk Prediction (2024-2025)

The trained RF-GA model is then used to predict the status of the project in the period from January 2024 to December 2025. The prediction data was synthetically compiled based on historical averages with random variations of ±5%.



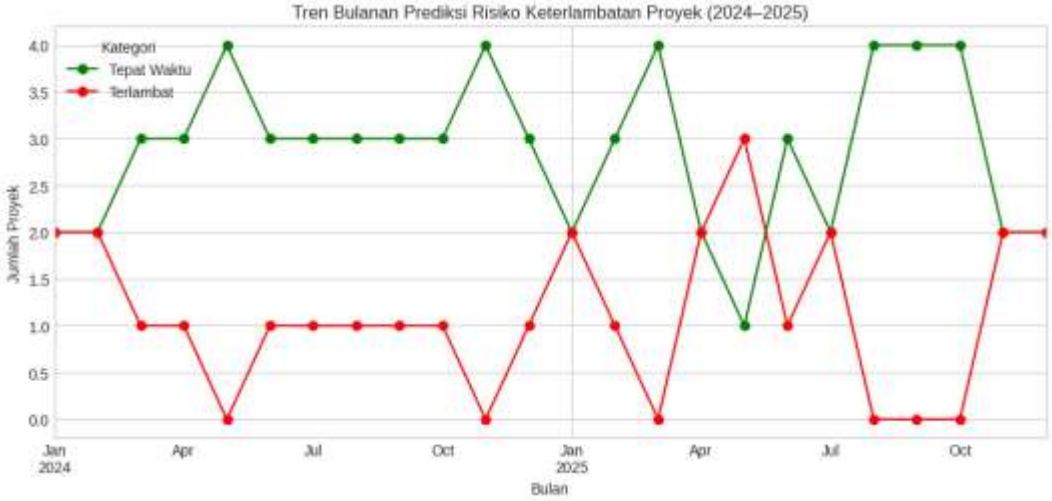
Picture 10. Distribution of project status predictions for 20242025.

The majority of projects are predicted to run on time, with the percentage of delays relatively stable in both years. This pattern demonstrates the effectiveness of existing planning systems, but also underscores the need for mitigation in high-risk sectors.



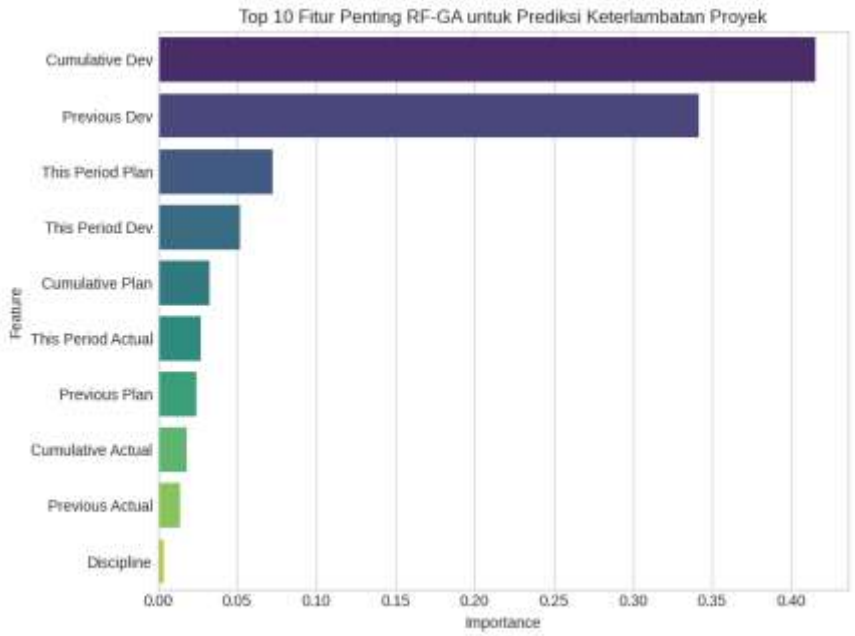
Picture 11. The number of delays predicted based on work discipline.

The predicted results show that *Construction* still dominates the late category, while *Procurement* and *Engineering* tend to be on time. This consistency indicates that the greatest risk factors remain at the physical execution stage of the project.



Picture 12. Monthly trends predict project delays over the 20242025 period.

There is a seasonal pattern with an increase in delays in the second and fourth quarters of each year. This can be attributed to the tropical weather cycle in Indonesia as well as the scheduling of large projects that often reach their peak in the middle and end of the year.



Picture 13. The average distribution of important features per work discipline.

The heatmap shows that *Cumulative Dev* and *This Period Dev* have the highest average in *the Construction* discipline. This reinforces the results of the previous analysis that discipline is the main factor causing delays. Thus, intensive monitoring of cumulative deviations in the implementation phase is the key to improving project performance.

Discussion of Analysis Results

The results of data analysis that have been carried out in the previous subchapter show that the application of Machine Learning algorithms makes a significant contribution in understanding and predicting the risk of delay in construction projects. These findings are not only empirical, but also have theoretical relevance to modern project management practices that increasingly rely on data-driven project management approaches.

The discussion in this section is directed to interpret the test results in more depth, relate them to the existing literature and theories, and explain the practical implications of the research findings on the risk management of construction project delays.

Analysis of the Main Factors Causing PT XYZ Project Delays

The results of the data exploration showed that the delay in the PT XYZ project was mostly due to time deviations in the early stages of project implementation. Based on correlation analysis and feature importance in the variables Cumulative Dev, Previous Dev, and This Period Dev are the most influential factors on the status of project delays.

A very high correlation between Cumulative Dev and Previous Dev ($r = 0.99$) confirms that small deviations at the beginning of a project have the potential to develop into significant delays if not effectively controlled. These results are in line with the principles in the Project Management Body of Knowledge (PMBOK) which states that time performance indicators (Schedule Performance Index and Schedule Variance) are key parameters in monitoring the risk of delay.

Furthermore, analysis per work discipline shows that the Construction discipline has the highest average deviation compared to other disciplines. This indicates that physical activities in the field are the main source of delays, which are generally caused by delays in the delivery

of materials, weather constraints, and limited labor. In contrast, the disciplines of Procurement and Engineering are relatively stable, with small deviations even tending to be positive.

RF-GA Model Development for Improved Prediction Accuracy

The second objective of the study focuses on the development of a hybrid Machine Learning model based on Random Forest Classifier that is optimized with Genetic Algorithm (GA). This model was developed to improve the predictability of PT XYZ's project delay risk prediction with high accuracy and better stability than conventional approaches.

Based on the modeling results, the RF-GA model showed a significant performance improvement over the default Random Forest. As accuracy increased from 95% to 99.34%, with precision, Recall, and F1-score values consistently above 0.99.

This increase is due to the optimal parameter search mechanism carried out by the Genetic Algorithm, where the combination of n_estimators, max_depth, and min_samples_split is selectively selected through the process of selection, crossover, and mutation. This approach allows for a wider exploration of the parameter space than conventional methods such as Grid Search.

The Feature importance results also show good interpretability of the RF-GA model, with the main variables (Cumulative Dev, This Period Dev, and Previous Dev) having the largest contribution to the classification results. This proves that the model is not only numerically superior, but also conceptually relevant to real project performance indicators.

Comparative Evaluation of RF-GA Performance with Conventional Models

The evaluation was carried out by comparing the performance results of the RF-GA model against two other approaches, namely *Support Vector Machine (SVM)* and *Random Forest* without optimization. A comparison of the evaluation results shows that the RF-GA model has the most superior performance across all metrics, as described below.

- Accuracy: RF-GA reaches 99.34%, higher than SVM (92%) and default RF (95%).
- Precision & *Recall*: RF-GA records an ideal balance (99%), signifying a high ability to detect "Late" projects without increasing *the False positive rate*.
- AUC (*Area Under Curve*): RF-GA obtains a value close to 1, indicating the ability to discriminate perfectly between on-time and late projects.

Scientifically, these results show that the integration of genetic algorithms is able to improve the stability and accuracy of *ensemble learning models* such as *Random Forest*. GA acts as a controller of model complexity through an evolutionary optimization process, so that it is able to adapt the model to complex non-linear data patterns such as in construction project data.

The advantages of the RF-GA model are also seen in *the confusion matrix*, where classification errors are almost non-found. This means that the model manages to learn the delay pattern very well without *overfitting*, as evidenced by the results of *cross-validation*.

Managerial Implications of Research Results

The results of this study have significant managerial implications for construction project control practices, particularly in the context of delay risk management. The Random Forest model optimized using Genetic Algorithm (RF-GA) not only shows very high classification performance statistically, but also provides a more objective and data-driven decision-making basis.

First, the finding that the Cumulative Dev, Previous Dev, and This Period Dev variables were the main predictors of delay confirms the importance of time deviation monitoring from the early stages of the project. Managerially, this means that project management needs to strengthen an early warning system based on cumulative deviation indicators. Small deviations in the early phases should not be taken lightly, as they have been shown to have a strong correlation with the final delay of the project.

Second, the dominance of the Construction discipline as the largest contributor to deviation shows that the focus of control needs to be directed to field activities. The practical implications are increased coordination between contractors, strengthening material supply chain management, and preparing weather-based risk mitigation and labor availability. By utilizing RF-GA's predictive model, management can identify potential delays early and allocate resources more effectively.

Third, the model's accuracy of 99.34% and an AUC value close to 1 indicate that the prediction system can be integrated as part of the project monitoring dashboard . The implementation of this model in the project management information system allows project managers to obtain early notification of high-risk projects, so that corrective action can be taken before deviations get bigger.

Fourth, the optimization approach based on Genetic Algorithm shows that the selection of the right Hyperparameters has a significant effect on the stability and generalization of the model. This implies that project organizations need to consider the use of advanced optimization techniques in data analysis, rather than relying solely on default parameters that may not match the characteristics of the project's data.

Strategically, the results of this study support the transformation of project management towards an artificial intelligence-based approach (AI-assisted project management), where decisions are based not only on subjective experience, but also on measurable and systematic predictive analysis. Thus, the implementation of the RF-GA model has the potential to improve the timeliness of project completion, reduce the risk of cost overruns due to delays, and improve the overall efficiency of project control.

CONCLUSION

Based on the results of the analysis and discussion, several conclusions can be drawn that directly address the research objectives as follows.

This study identified that the main factors influencing delays in the PT XYZ project were time deviation indicators, particularly cumulative deviation (Cumulative Dev), previous deviation (Previous Dev), and current period deviation (This Period Dev). Correlation analysis showed that Cumulative Dev had a very strong relationship with Previous Dev ($r = 0.99$), indicating that early-stage deviations can develop into significant delays if not addressed promptly. Descriptive analysis also indicated that the construction discipline contributed the highest level of negative deviation compared to other disciplines. Therefore, physical construction activities were identified as the dominant factor in project delays. These findings highlight that monitoring cumulative deviations from the early stages of a project is a key indicator for controlling delay risks.

This study successfully developed a project delay prediction model using a Random Forest-based machine learning approach optimized with a Genetic Algorithm (RF-GA). The

optimization process was applied to the n_estimators, max_depth, and min_samples_split parameters using an evolutionary approach with a population size of 10 and five generations, combined with three-fold cross-validation. The resulting RF-GA model demonstrated very high performance, achieving an accuracy of 99.34%, precision and recall of 0.99, and an AUC of 1.000. These results indicate that the model has excellent discriminative ability in distinguishing between on-time and delayed projects.

The model evaluation compared RF-GA with baseline models, namely Support Vector Machine (SVM) and non-optimized Random Forest. The results showed that SVM achieved an accuracy of 0.92 (initial evaluation) and an AUC range of 0.96–0.991, while the default Random Forest achieved an accuracy of 0.95 and an AUC of 0.98. In contrast, RF-GA achieved the highest performance with 99.34% accuracy and an AUC of 1.000. In addition, the confusion matrix indicated that RF-GA had a minimal misclassification rate compared to the other models. The ROC curve also showed that RF-GA was closest to the ideal point (0,1), indicating high sensitivity and a very low false-positive rate. Based on all evaluation metrics, the RF-GA model was identified as the most optimal model for predicting delay risks in PT XYZ projects.

REFERENCE

Arar, E., & Halicioglu, F. H. (2025). Understanding artificial neural networks as a transformative approach to construction risk management: A systematic literature review. *Buildings, 15*(18), 3346.

Chen, H., Zhang, L., & Wu, X. (2020). Performance risk assessment in public–private partnership projects based on adaptive fuzzy cognitive map. *Applied Soft Computing, 93*, 106413.

Guo, K., & Liu, Z. (2022). Transfer matrix method for calculating the transverse load distribution of articulated slab bridges.

Kamandang, Z. R., & Yang, J. (2023). Implementation guideline to solve obstacles in construction delay analysis: An empirical study of Indonesia. *21*(1), 1–10.

Karlina, D., Rahmawati, S., Hutagalung, R. A., & Amin, M. (2025). Models and methods of cost and schedule integration in project management: A systematic literature review 2020–2025.

Lusiyanti, D., Musdalifah, S., Sahari, A., & Fajri, I. Al. (2025). Performance evaluation of machine learning algorithms in large-scale datasets. *7*(1), 84–92.

Mustamin, M. R., Suleman, A. R., Djufri, H., Asrun, B., Mawarni, A. A. I., Fachriza, M., Putri, H., Tuwo, M., Sipil, T., Negeri, P., Pandang, U., Pioneer, J., & Tamalanrea, K. (2023). The risk of time delay in the implementation of the Pamukkulu Dam construction project with the risk matrix method and the analytic hierarchy process (AHP) method. *15*, 145–158.

Nassar, A. H., & Elbisy, A. M. (2024). A machine learning approach to predict time delays in marine construction projects. *14*(5), 16125–16134.

Palilati, M. P., Doda, N., & Dwi, R. (2024). Analysis of factors causing delays in the implementation of construction projects. *5*(2), 981–992.

Pranoto, M. A. H., Hermawan, F., & Hatmoko, J. U. D. (2024). Analysis of factors affecting delays in housing projects (Case study: Luxury housing projects in Semarang). *9*(8).

Putra, D. A., Sari, O. L., & Situmorang, R. (2023). Factor analysis of construction project

delays in Balikpapan City. *9*(1), 17–24.
Putra, D. M., & Triana, M. I. (2024). Analysis of factors causing delays in the implementation of construction projects on CV X. *JUTIN: Journal of Integrated Industrial Engineering, 7*(2), 969–979.
Sadad, I., & Sangidana, G. A. (2024). Analysis of the influence of construction management in handling delays on construction projects. *9*.
Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2021). Machine learning-based framework for construction delay mitigation. *Journal of Information Technology in Construction, 26*, 303–318. https://doi.org/10.36680/j.itcon.2021.017
Shahidi, S., Samadzai, A. W., & Shahbazi, H. (2025). Effective data preprocessing in data science: From method selection to domain-specific optimization. *2*(4), 84–90.
Sinaga, L. C. (2024). Development of a project plan duration prediction model. *Dimensi Utama Teknik Sipil, 11*(2). https://doi.org/10.9744/duts.11.2.139-148
Study, P., Engineering, M., & Petra, U. K. (2025). Application of artificial intelligence methods for cost prediction. *12*(1), 1–10. https://doi.org/10.9744/duts.12.1.1-10
Subrata, A., Mughnyanti, M., & Medan, P. N. (2025). Use of machine learning for cost estimation. *4307*(August), 3791–3795.
Zu, B., & Liu, X. (2024). Construction schedule optimization based on genetic algorithm. *Atlantis Press International BV*. https://doi.org/10.2991/978-94-6463-447-1