

PREDICTING NATIONAL RESEARCH PROJECT DELAYS: A COMPARATIVE ANALYSIS OF RANDOM FOREST, XGBOOST, AND LOGISTIC REGRESSION

Dian Setyawati¹, Sutikno²

^{1,2}*School of Interdisciplinary Management and Technology (SIMT), Institut Teknologi Sepuluh
Nopember (ITS), Surabaya, Indonesia*

Email: diansetya.fisipui@gmail.com , sutikno@statistika.its.ac.id

ABSTRACT

Purpose: This study identifies the determinants of completion delay in national research projects and compares the predictive performance of Random Forest, XGBoost, and Logistic Regression in the context of the Productive Innovative Research (RISPRO) program managed by the Education Fund Management Institution (LPDP). **Methods:** Historical data on completed RISPRO projects from 2013 to February 2026 were integrated from selection, disbursement, reporting, monitoring, and no-cost extension databases. The final modeling data were processed through a leakage-safe workflow that combined project metadata with title-based TF-IDF features, standardized numeric variables, encoded categorical variables, and the application of SMOTE only to the training folds. Determinants were triangulated through Mann-Whitney U and chi-square tests, Logistic Regression coefficients and odds ratios, and tree-based feature importance. Model performance was evaluated using repeated stratified five-fold cross-validation, producing 25 estimates per algorithm, followed by Friedman, Wilcoxon signed-rank, and McNemar tests. **Results:** Project duration is the most consistent determinant: a one-standard-deviation increase, approximately 0.60 years, raises the odds of delay by 3.34 times. Funding shows a non-linear pattern, with projects funded at IDR 3-5 billion showing the highest odds of delay, 4.72 times the reference category, while projects above IDR 5 billion show lower risk. Slower second-stage disbursement and slower monitoring are also associated with higher delay risk. Random Forest achieved the highest mean F1-score, 0.907, and recall, 0.942, although the overall Friedman test showed no statistically significant difference among the three algorithms, $p = 0.0692$. **Implications:** LPDP can use these models as an early-warning instrument to prioritize assistance for projects at higher risk of delay, accelerate second-stage disbursement, and implement risk-based monitoring. Random Forest is suitable as the primary screening engine, while Logistic Regression can provide an interpretive explanatory layer for accountable public-sector decision-making.

Keywords: project delay; machine learning; Random Forest; research funding; RISPRO LPDP

INTRODUCTION

Research and development (R&D) investment is widely regarded as a strategic input for productivity, national competitiveness, industrial upgrading, and long-term economic growth. Evidence from bibliometric and macroeconomic studies indicates that scientific output and R&D expenditure are positively associated with economic performance, with particularly significant effects in developing economies seeking to strengthen their innovation capacity (Arana-Barbier, 2023; Vasić et al., 2016). In Indonesia, the economic contribution of R&D may emerge with a lag of approximately three to five years, so that the quality of research funding governance matters not only for

scientific output but also for the timing of broader economic benefits (Sarsiti & Rosid, 2025). International policy developments likewise show that governments continue to expand research budgets as long-term national investments, even when the benefits are not immediately visible (Reuters, 2025).

Research funding refers to the financial resources allocated to basic research, applied research, and experimental development (OECD, 2015). Indonesia institutionalizes part of this function through the Education Fund Management Institution, known as LPDP. Under Presidential Regulation No. 111 of 2021, LPDP manages several endowment funds, including the Research Endowment Fund, whose investment returns support research programs (Government

of the Republic of Indonesia, 2021). In 2024, LPDP managed approximately IDR 154 trillion in accumulated endowment funds and had disbursed roughly IDR 3.5 trillion to thousands of research projects through the RISPRO scheme (LPDP, 2025a, 2025b). The RISPRO portfolio therefore represents a strategically important public investment whose outputs are expected to generate scientific, economic, social, and policy value.

The large number of funded research projects is not always matched by timely completion. Based on data as of 7 December 2025, a total of 311 projects under the RISPRO Competition, Invitation, and International Collaboration schemes had been completed. However, 53% of them experienced delays. Delay is operationally documented through approved no-cost extensions (NCE), which means that research continued beyond the contractual completion date without additional project funding.

Delays can generate financial and strategic consequences. LPDP may incur additional monitoring and evaluation costs, forgo investment returns on funds already disbursed but not yet used, experience an erosion of purchasing power due to inflation, and postpone the realization of research benefits. The source thesis estimates potential losses by combining additional monitoring costs, opportunity costs, inflation exposure, and delayed economic impact. Although such estimates depend on assumptions and must be interpreted cautiously, they indicate that schedule performance has consequences beyond administrative compliance. Timely completion affects the institution's ability to recycle investment income, account for public funds, manage staff workloads, maintain stakeholder trust, and deliver benefits to ministries, industry, and society. Bank Indonesia inflation data and LPDP cost regulations are used in the source study to contextualize these financial impacts (Bank Indonesia, n.d.; Ministry of Finance of the Republic of Indonesia, 2025).

The project management literature conceptualizes time, cost, and scope or performance as interdependent constraints. Delays can occur when the planned schedule fails to accommodate task interdependencies, resource shortages, technical uncertainty, governance barriers, or changes in project

requirements (Heagney, 2016; Kerzner, 2017; Lock, 2020; Meredith & Mantel, 2014; Project Management Institute, 2017, 2021). Research projects can be especially difficult to schedule because their outputs are partly exploratory, the path from activities to results is uncertain, and milestone achievement may depend on laboratories, procurement, partners, regulatory approvals, intellectual property processes, and commercialization. Project complexity is not merely a function of size; it also reflects the number of interacting components, organizational interfaces, and the uncertainty inherent in the work (Baccarini, 1996). Risk management standards therefore emphasize systematic identification, analysis, treatment, and monitoring, rather than reactive intervention after problems become visible (International Organization for Standardization, 2018; Wideman, 2019).

LPDP has implemented a number of interventions, including project-specific analyst assistance, standardized training for research analysts, monitoring during the project period, coordination meetings for projects showing warning signs, and coaching for grantees experiencing delays. RISPRO guidelines and institutional performance reports document the development of these governance mechanisms (LPDP, 2020a, 2020b, 2021a, 2021b, 2022a, 2022b, 2023a, 2023b, 2024a, 2024b, 2025a, 2025b). However, the identification process still relies heavily on scheduled monitoring and external reviewer involvement. A project may therefore receive intensive attention only after most of the implementation period has elapsed. This creates a practical need for an earlier, repeatable, and data-driven method that can prioritize projects according to predicted risk before delays become difficult to remedy.

Predictive analytics offers one possible response. Machine learning concerns algorithms that improve their performance by learning patterns from data, rather than relying solely on predetermined decision rules (Mitchell, 1997). In project analytics, supervised classification can use historical characteristics to estimate whether a new or ongoing project is likely to experience delay. Previous studies have applied Logistic Regression, decision trees, Random Forest, boosting, support vector machines, artificial neural networks, and related approaches to production or construction delay problems

(Bimantara et al., 2025; Giri et al., 2025; Kannan et al., 2022; Movafaghpour, 2023; Nguyen, 2020; Sahu et al., 2025; Uddin et al., 2022; Yaseen et al., 2020; Zhang et al., 2020). These studies support the practical feasibility of predictive modeling, but most were conducted in construction, manufacturing, mining, or open project repositories, rather than in national research funding programs.

The comparison conducted in this study focuses on Logistic Regression, Random Forest, and XGBoost because they represent different modeling logics. Logistic Regression estimates the log-odds of a binary event as a linear function of predictors and produces coefficients that can be transformed into odds ratios, making the model transparent to decision makers (Hosmer et al., 2013; Kleinbaum & Klein, 2010). Random Forest uses bagging to combine many uncorrelated decision trees, reducing variance while capturing non-linear relationships and interactions without requiring a single functional form (Breiman, 2001). XGBoost applies sequential gradient boosting, in which each new tree is trained to correct previous errors and is regularized to improve computational efficiency and predictive performance (Chen & Guestrin, 2016). Ensemble methods may offer higher predictive power, but the public-sector context also demands explanation, accountability, and awareness of the limitations of black-box outputs (Bansal et al., 2021; Lundberg & Lee, 2017; Molnar, 2020, 2022).

Model comparison becomes particularly relevant because no learning algorithm is universally superior across all datasets. The no-free-lunch principle implies that performance advantages depend on the structure, size, noise, and feature distribution of a particular problem (Wolpert, 1996). Small samples with high-dimensional, sparse text may benefit more from simpler or variance-reducing models, whereas larger data with stable non-linear structure may benefit more from boosting. Previous comparisons report that ensemble methods often achieve strong predictive accuracy, while Logistic Regression remains valuable for interpretation (Kannan et al., 2022; Sahu et al., 2025; Uddin et al., 2022). A statistically defensible comparison therefore requires repeated validation and formal pairwise tests, rather than only a single train-test split or a

standalone accuracy value (Demšar, 2006; Dietterich, 1998; Vabalas et al., 2019).

Existing RISPRO studies have so far mostly examined program performance, collaboration, knowledge sharing, the Triple Helix, open innovation, and research system conditions through survey, PLS-SEM, bibliometric, or qualitative approaches. Irawan (2025) identified knowledge sharing and project management as important mechanisms for the performance of RISPRO Invitation teams. Hidayatuloh (2023) evaluated the funding scheme using the Triple Helix Balance Model and open innovation logic, while Prasojo et al. (2023) examined research performance in higher education through institutional climate, collaboration, funding, competence, and output. These contributions are important for explaining governance and performance, but they have not yet built a project-level early-warning classifier from historical administrative records.

This study fills that gap by integrating structured project metadata and title-based text features to identify determinants of delay and to compare three classification approaches in the context of Indonesian national research funding. The study poses two questions: which project characteristics consistently distinguish delayed RISPRO projects from on-time ones, and how do Random Forest, XGBoost, and Logistic Regression compare in predicting project completion delay? The novelty of this study lies in its data-driven, project-level context, the triangulation of statistical tests with model interpretation, the use of repeated stratified cross-validation and pairwise significance tests, and the translation of predictive findings into operational recommendations for public research funding institutions.

RESEARCH METHODS

This quantitative study uses a retrospective predictive design. The unit of analysis is completed projects within the directly managed RISPRO portfolio, comprising the RISPRO Competition, RISPRO Invitation, and RISPRO International Collaboration schemes. Secondary data were drawn from LPDP's eRISPRO records and cover projects funded from 2013 to February 2026. Data sources include researcher profiles, project selection records, disbursement data, monthly reports, monitoring and evaluation

records, and no-cost extension files. The outcome variable is binary: a project is coded 0 if the documented completion date occurred on or before the contract end date, and coded 1 if completion exceeded the contract date and was supported by an NCE request.

The initial integrated dataset contained 149 projects. Six projects were excluded because essential columns were missing or the second-stage disbursement status could not be represented consistently, leaving 143 projects for descriptive analysis. One additional record was excluded from text modeling because the cleaned title contained fewer than three meaningful tokens, yielding 142 projects ready for modeling. The descriptive distribution shows a marked class imbalance: 118 projects, or 82.5%, experienced delays and 25 projects, or 17.5%, were completed on time. The modeling

workflow preserved this imbalance through stratified data splitting and addressed it with SMOTE applied only to the training data to prevent information leakage (Chawla et al., 2002).

The predictor set represents project scope, institutional context, funding, implementation requirements, and administrative timing. The original thesis specified 19 predictors in addition to the target variable. Categorical columns were standardized before encoding, including the consolidation of more than 20 research-focus labels into six strategic clusters aligned with the national research and innovation agenda. Binary attributes were mapped deterministically, nominal variables were one-hot encoded, and numeric variables were standardized. Table 1 summarizes the operational predictors retained in this article.

Table 1. Predictor variables and operational definitions

Variable	Operational definition	Type
Research focus	Strategic research cluster derived from the original focus categories	Categorical
Output type	Product/technology or policy/model	Categorical
Institution category	Legal-entity state university, non-legal-entity state university, company, ministry, or other institution	Categorical
Region	Western, central, or eastern Indonesia	Categorical
Funding amount	Approved project funds in Indonesian rupiah	Numeric
Funding category	Funding ranges used to detect non-linear risk patterns	Categorical
Partner contribution	Presence or absence of a partner contribution at selection	Binary
Funding scheme	Competition, Invitation, or International Collaboration	Categorical
Duration	Approved research duration in years	Numeric
Professor status	Whether the principal investigator held professor status at selection	Binary
Prototype	Whether a prototype was an expected output	Binary
Policy	Whether a policy output was expected	Binary
Model	Whether a model output was expected	Binary
Publication	Whether publication was an expected output	Binary
Intellectual property	Whether intellectual property was an expected output	Binary
Downstreaming obligation	Whether commercialization or downstreaming was mandatory	Binary
First-stage disbursement	Month of first-stage disbursement in the first year	Numeric
Second-stage disbursement	Month of second-stage disbursement in the first year	Numeric
Monitoring	Month in which first-year monitoring occurred	Numeric

Data preparation was carried out as a single integrated workflow. Duplicate and inconsistent records were checked, equivalent institution or focus labels were standardized, non-numeric text in time variables was converted to missing values, and label names were harmonized. Numeric outliers were examined through histograms, boxplots, and interquartile-range logic. Outliers were not removed automatically when they represented valid large or long-duration projects, because extreme but genuine projects are substantively important for risk analysis (Aggarwal, 2017). Decisions regarding missing values followed the principle that imputation or deletion should reflect the type of measurement and the amount of information lost (Kotsiantis et al., 2007). Numeric scaling was required because funding, timing, and duration are measured on different ranges (Juszczak et al., 2002).

Project titles were processed as an additional text source. Preprocessing included lowercasing, removal of HTML entities and non-alphanumeric characters, removal of Indonesian and English stop-words, normalization of certain terms, and removal of very common words. TF-IDF transformed the cleaned titles into sparse numeric features by weighting terms according to their frequency within a document and their inverse frequency across the corpus (Jurafsky & Martin, 2021). Exploratory processing initially retained up to 150 terms. In the final leakage-safe modeling workflow, the vocabulary was limited to 100 terms, with $\min_df = 2$, $\max_df = 0.90$, unigram features, and sublinear term frequency. This stricter limitation was used to reduce the influence of rare words on the small sample.

All transformations that estimate parameters from data were placed inside a ColumnTransformer and fit only on the training data. StandardScaler estimated numeric means and standard deviations; OneHotEncoder learned the available nominal categories while ignoring categories not seen in the test data; and TfidfVectorizer learned the title vocabulary. This design follows the principle that preprocessing information must not move from the test data into model training, because leakage produces optimistically biased results that will not replicate on new data (Kaufman et al., 2012). The implementation relied on the Python data and machine-learning ecosystem, particularly pandas

and scikit-learn (McKinney, 2010; Pedregosa et al., 2011).

The modeling-ready data were split into training and test sets using a 75:25 stratified split, yielding 36 projects in the test set. A fixed random state was used to ensure reproducibility. SMOTE with $k_neighbors = 3$ was placed inside the training pipeline so that synthetic minority observations were generated separately on each cross-validation fold and never from the test data. This smaller neighbor value reflects the limited number of minority observations available on each fold. This procedure avoids both data loss from undersampling and test-data contamination from oversampling.

Three algorithms were evaluated. Logistic Regression served as an interpretable statistical benchmark and used L2 regularization in the selected configuration. Random Forest represented the bagging ensemble and was tuned on the number of trees, tree depth, minimum split size, minimum leaf size, and the feature-sampling rule. The selected Random Forest configuration used 200 trees, $\max_depth = 5$, $\max_features = \text{sqrt}$, $\min_samples_leaf = 1$, and $\min_samples_split = 5$. XGBoost represented sequential boosting and was tuned on the number and depth of trees, the learning rate, row and column subsampling, and the minimum child weight. The selected XGBoost configuration used 200 estimators, $\text{learning_rate} = 0.05$, $\max_depth = 3$, $\min_child_weight = 1$, $\text{subsample} = 0.8$, and $\text{colsample_bytree} = 0.8$. Grid search and cross-validation were used to select the settings, while recognizing that an exhaustive grid is not always superior to a well-designed random search for large search spaces (Bergstra & Bengio, 2012).

Performance was assessed through accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). Delay was treated as the positive class. Precision therefore measures how often a predicted delay is correct, whereas recall measures how many genuinely delayed projects are successfully identified. Because the dataset is imbalanced, accuracy was not used as the primary criterion. F1-score summarizes the balance between precision and recall, and AUC evaluates class discrimination across various thresholds (Fawcett, 2006; Powers, 2011; Sokolova & Lapalme, 2009). AUC was computed from

predicted probabilities rather than from hard class labels.

The main evaluation used repeated stratified five-fold cross-validation with five repetitions, producing 25 paired estimates per model. This approach provides more stable performance estimates than a single holdout split when the sample is small (Vabalas et al., 2019). The Friedman test evaluated the overall null hypothesis of equal model performance. Pairwise Wilcoxon signed-rank tests were examined with Bonferroni correction, and the McNemar test compared paired prediction errors on the shared holdout set (Demšar, 2006; Dietterich, 1998; Friedman, 1937; McNemar, 1947; Wilcoxon, 1945).

Determinants of delay were identified through methodological triangulation. The Mann-Whitney U test compared numeric variables between delayed and on-time projects, while the chi-square test assessed categorical differences at $\alpha = 0.05$. Logistic Regression coefficients were converted into odds ratios, with numeric effects interpreted per one-standard-deviation increase because the predictors were standardized. Random Forest and XGBoost feature-importance rankings were then compared with the statistical test results and the Logistic Regression ranking. A determinant was considered more credible when it appeared consistently across multiple methods with differing assumptions. Figure 1 presents the complete analytic workflow.

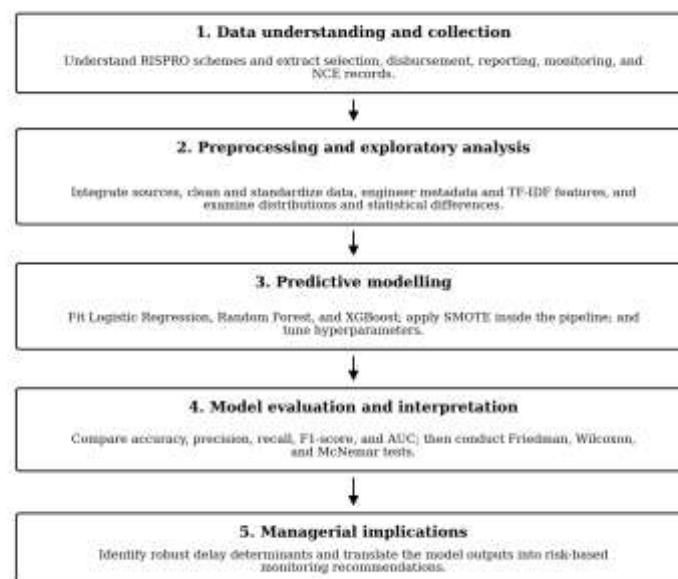


Figure 1. Analytic workflow for delay identification, model comparison, and managerial interpretation.

RESULT AND DISCUSSION

Data Profile and Exploratory Findings

The RISPRO portfolio is heterogeneous in project scale and administrative timing. The mean approved funding was IDR 2.80 billion, with a median of IDR 2.30 billion and a range from IDR 284.60 million to IDR 11.74 billion. The distribution is strongly right-skewed, indicating that most projects were funded below IDR 5 billion, while a small number of large projects raise the mean. The mean project duration was 2.56 years, with most projects

approved for two or three years. First-stage disbursement generally occurred earlier, on average in month 2.36, whereas second-stage disbursement occurred on average in month 12.55 and showed much greater dispersion. Monitoring occurred on average in month 11.02 but ranged from month 2.10 to month 21.50. These patterns suggest that delay risk may relate not only to project design but also to the timing of administrative support.

The project portfolio is geographically concentrated. Approximately 93.7% of projects originated from western Indonesia, 5.6% from central Indonesia, and only 0.7% from eastern

Indonesia. Product or technology outputs accounted for 76.2% of projects, while policy or model outputs accounted for 23.8%. A formal downstreaming obligation applied to 56.6% of projects. Health sovereignty was the largest research-focus cluster at 32.9%, followed by strategic industrial development at 21.7% and food sovereignty at 18.2%. This distribution describes the portfolio but does not independently establish causality or statistical differences in delay status.

The target distribution confirms the severe imbalance: 118 of the 143 descriptively analyzed projects experienced delays. Figure 2 displays this pattern. This imbalance has direct methodological consequences. A classifier can achieve apparently high accuracy by predicting delay for almost all projects, while providing minimal ability to distinguish the smaller on-time class. Interpretation therefore emphasizes recall, precision, F1-score, AUC, class-specific performance, and repeated validation, rather than aggregate accuracy alone.

Table 2. Descriptive statistics for numeric predictors

Predictor	Mean	SD	Minimum	Median	Maximum
Funding amount	IDR 2.80 bn	IDR 2.23 bn	IDR 284.60 mn	IDR 2.30 bn	IDR 11.74 bn
Duration (years)	2.56	0.60	1.00	3.00	3.00
First-stage disbursement (month)	2.36	1.06	1.00	2.00	9.00
Second-stage disbursement (month)	12.55	4.64	5.00	11.00	38.00
Monitoring (month)	11.02	3.08	2.10	10.23	21.50

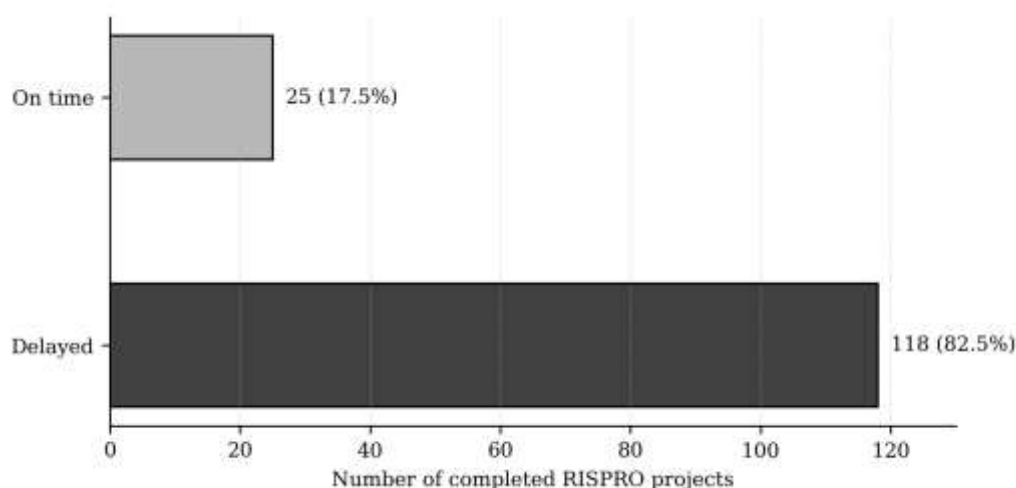


Figure 2. Distribution of completed RISPRO projects by delayed and on-time status in the descriptive dataset.

Statistical Differences Between Delayed and On-Time Projects

Exploratory percentages reveal a strong duration gradient. Approximately 62.5% of one-year projects were completed on time, compared with 30.6% of two-year projects and only 5.7% of three-year projects. Duration therefore functions as a marker of scale and complexity rather than mere elapsed time. Longer projects typically contain more work packages, procurement activities, partnerships, technical dependencies, and opportunities for cumulative

delay. This interpretation is consistent with project complexity theory, which emphasizes interactions and organizational interfaces as sources of managerial difficulty (Baccarini, 1996).

Funding shows a non-monotonic relationship. Projects below IDR 2 billion had an on-time proportion of approximately 25%-50%, while projects between IDR 2 billion and IDR 5 billion had an on-time proportion of only about 3%-6%. This proportion improved again to 23.5% for projects above IDR 5 billion. This pattern suggests that mid-funded projects may

combine substantial administrative and technical complexity with governance arrangements that are less mature than those of very large strategic projects. A linear claim that higher funding always increases delay would therefore be misleading.

The Mann-Whitney U test found significant differences for funding amount, duration, second-stage disbursement, and monitoring.

First-stage disbursement was not significant. Among the categorical variables, only funding category met the 0.05 criterion. Output type, institution category, prototype requirement, and partner contribution showed descriptive differences but did not meet the conventional threshold. These results support a careful distinction between strong determinants and variables that only appear influential in a single analysis.

Table 3. Statistical differences between delayed and on-time projects

Type	Predictor	Test	p	Decision
Numeric	Funding amount	Mann-Whitney U	0.0113	Significant
Numeric	Duration	Mann-Whitney U	< 0.000001	Significant
Numeric	First-stage disbursement	Mann-Whitney U	0.2364	Not significant
Numeric	Second-stage disbursement	Mann-Whitney U	0.0208	Significant
Numeric	Monitoring	Mann-Whitney U	0.0041	Significant
Categorical	Output type	Chi-square	0.0659	Not significant
Categorical	Institution category	Chi-square	0.0869	Not significant
Categorical	Region	Chi-square	0.8325	Not significant
Categorical	Funding category	Chi-square	0.0081	Significant
Categorical	Partner contribution	Chi-square	0.1122	Not significant
Categorical	Funding scheme	Chi-square	0.2172	Not significant
Categorical	Professor status	Chi-square	0.9933	Not significant
Categorical	Prototype	Chi-square	0.0784	Not significant
Categorical	Policy output	Chi-square	0.0998	Not significant
Categorical	Model output	Chi-square	1.0000	Not significant
Categorical	Publication	Chi-square	0.2850	Not significant
Categorical	Intellectual property	Chi-square	0.4447	Not significant
Categorical	Downstreaming obligation	Chi-square	0.2371	Not significant

Predictive Performance

The six default and tuned configurations produced similar holdout results. Four configurations achieved an accuracy of 0.889, and the difference between the highest and lowest holdout F1-scores for the delayed class was only 0.023. Because the test set contains 36 projects, reclassifying a single project changes accuracy by about 2.8 percentage points. Holdout findings were therefore treated as descriptive rather than decisive. Hyperparameter tuning improved minority-class balance for Random Forest and XGBoost but did not improve Logistic Regression on the same test split.

The repeated cross-validation results provide the primary evidence. Random Forest achieved

the highest mean accuracy, 0.839, recall, 0.942, and F1-score, 0.907. Logistic Regression achieved the highest precision, 0.915, and the highest mean AUC, 0.820. XGBoost occupied an intermediate position on precision, recall, and F1-score. The AUC values were practically identical: 0.820 for Logistic Regression, 0.814 for XGBoost, and 0.811 for Random Forest. This 0.009 difference is far smaller than the corresponding standard deviations, which range from 0.094 to 0.111. All three values fall within a range commonly interpreted as good to very good discrimination, although qualitative labels for AUC should never replace consideration of uncertainty and context (Mandrekar, 2010).

The models therefore exhibit a meaningful trade-off. Random Forest is the most sensitive screening model and misses the fewest genuinely

delayed projects. This is valuable when the cost of failing to identify a truly delayed project is high. Logistic Regression is more conservative and produces the highest precision, which is useful when false alarms consume limited

monitoring resources. XGBoost showed no clear advantage over the alternatives on this small dataset. Figure 3 and Table 4 present the repeated-validation results.

Table 4. Repeated stratified five-fold cross-validation performance (five repetitions, 25 estimates)

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.806 ± 0.070	0.915 ± 0.039	0.847 ± 0.076	0.877 ± 0.048	0.820 ± 0.094
Random Forest	0.839 ± 0.060	0.876 ± 0.038	0.942 ± 0.055	0.907 ± 0.036	0.811 ± 0.111
XGBoost	0.825 ± 0.058	0.888 ± 0.036	0.905 ± 0.053	0.896 ± 0.036	0.814 ± 0.102

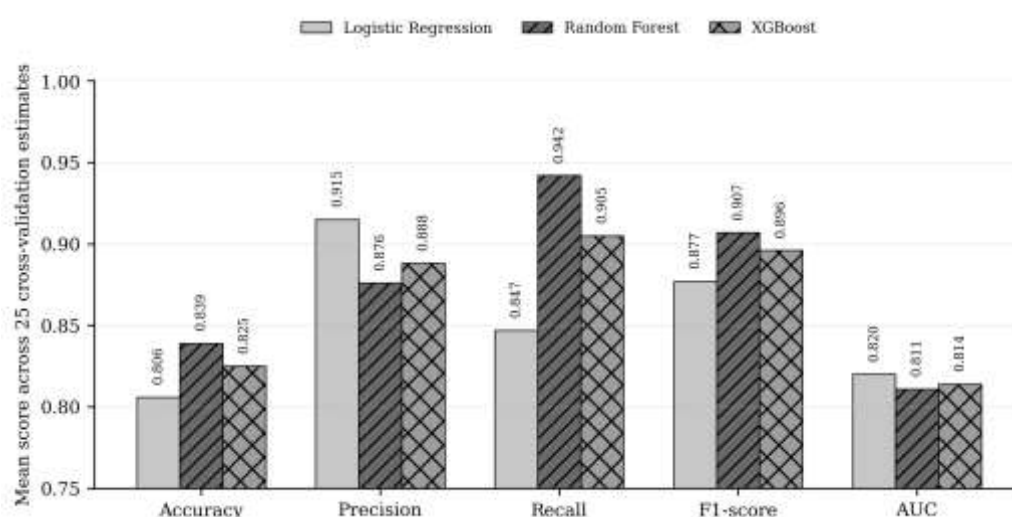


Figure 3. Mean predictive performance across the 25 cross-validation estimates.

Statistical Comparison of Algorithms

The Friedman test yielded chi-square = 5.340 and $p = 0.0692$. At $\alpha = 0.05$, the overall null hypothesis of equal F1-score performance was not rejected. The pairwise Wilcoxon test initially indicated differences between Logistic Regression and Random Forest, $p = 0.0110$, and between Logistic Regression and XGBoost, $p = 0.0308$. After Bonferroni correction for three comparisons, α was lowered to 0.0167, and only the Logistic Regression versus Random Forest comparison remained below the adjusted threshold. Following the recommended sequence for multiple-classifier comparison, these pairwise results must still be interpreted cautiously because the omnibus test did not reach significance (Demšar, 2006).

The McNemar test on the shared holdout predictions yielded $p = 1.000$ for all three model pairs. The contingency patterns differed by only one to three projects out of 36, confirming that

the visual differences in holdout metrics do not represent materially different error structures. The defensible conclusion is therefore not that one algorithm dominantly outperforms the others. Rather, the evidence supports a marginal F1-score advantage for Random Forest over Logistic Regression, accompanied by broadly equivalent discrimination.

Random Forest was selected as the operationally preferred predictive model because it combines the highest mean F1-score with the highest recall. This selection aligns directly with LPDP's early-warning objective: missing a delayed project may be more harmful than temporarily flagging an on-time project for additional review. Nevertheless, this selection does not eliminate the value of Logistic Regression. When institutional accountability and explanation are paramount, a hybrid arrangement can use Random Forest for screening and Logistic Regression for transparent factor interpretation.

Table 5. Pairwise significance tests for model performance

Test	Comparison	Statistic	p	Interpretation
Friedman omnibus	All three models	5.340	0.0692	No overall significant difference
Wilcoxon	Logistic Regression vs Random Forest	–	0.0110	Significant after Bonferroni
Wilcoxon	Logistic Regression vs XGBoost	–	0.0308	Not significant after Bonferroni
Wilcoxon	Random Forest vs XGBoost	–	0.0765	Not significant
McNemar	Logistic Regression vs Random Forest	–	1.000	Not significant
McNemar	Logistic Regression vs XGBoost	–	1.000	Not significant
McNemar	Random Forest vs XGBoost	–	1.000	Not significant

Determinants of Project Delay

Logistic Regression provided the clearest direction and magnitude of effects. Because the numeric variables were standardized, each numeric odds ratio represents a one-standard-deviation increase. For duration, one standard deviation equals approximately 0.60 years; for second-stage disbursement, 4.61 months; for monitoring, 3.07 months; and for funding amount, approximately IDR 2.21 billion. The model used L2 regularization, so the coefficients should be interpreted as predictive associations rather than unregularized causal estimates with classical p-values and confidence intervals.

Duration was the strongest determinant across all methods. Its coefficient of 1.207 corresponds to an odds ratio of 3.344. A project that is approximately 0.60 years longer therefore has 3.34 times higher odds of delay, with the model's other features held constant. Duration ranked second in Logistic Regression, first in Random Forest, and fourth in XGBoost, producing the best average ranking among all features. This consistency, combined with the very small Mann-Whitney p-value and the sharp descriptive decline in on-time completion, makes duration the strongest substantive finding.

Funding category revealed the non-linear pattern observed during exploration. Projects funded at IDR 3-5 billion had a coefficient of 1.551 and an odds ratio of 4.715 relative to the reference category. Projects funded at IDR 2-3 billion had an odds ratio of 3.197. Conversely, projects funded above IDR 5 billion had a negative coefficient and an odds ratio of 0.370. These results do not support the simple claim that larger funds cause delay. They indicate that the

mid-funding range is the highest-risk segment in the observed portfolio, likely because it combines complex outputs and procurement with governance less mature than that of flagship projects.

Second-stage disbursement was another consistent administrative determinant. Its coefficient of 1.124 corresponds to an odds ratio of 3.078 for each increase of approximately 4.61 months. Monitoring also had a positive coefficient, 0.722, and an odds ratio of 2.058 for each increase of approximately 3.07 months. Both variables were significant in the Mann-Whitney test and ranked highly in Random Forest. This finding supports the proposition that administrative timing is not inconsequential. Late second-stage disbursement can disrupt procurement, staffing, experimentation, or partner activities, while late monitoring reduces the time available for corrective action.

Several variables had negative coefficients and may represent protective conditions. Partner contribution at the selection stage had an odds ratio of 0.396, a prototype output had an odds ratio of 0.350, and projects implemented by ministries had an odds ratio of 0.377. These variables did not all reach significance in the separate chi-square tests, so they should not be treated as definitive causal protection. Nevertheless, the direction of their effects is managerially plausible: active partners can provide resources and implementation discipline, prototype projects may have clearer and more tangible milestone structures, and ministry-led projects may operate within more formal administrative arrangements.

Publication and intellectual property outputs had positive Logistic Regression coefficients,

with odds ratios of 2.702 and 2.383 respectively, but neither variable was significant in the chi-square test. The source thesis therefore explicitly treats these coefficients as patterns requiring further investigation rather than confirmed determinants. The coefficients may reflect output

difficulty, measurement timing, correlation with complex projects, or regularization effects. This caution is important because a predictive coefficient can identify a pattern without proving that mandating publication or intellectual property increases delay.

Table 6. Leading Logistic Regression predictors and their odds ratios

Predictor	Coefficient	Odds ratio	Direction
Funding category: IDR 3-5 bn	1.551	4.715	Higher odds of delay
Duration	1.207	3.344	Higher odds of delay
Funding category: IDR 2-3 bn	1.162	3.197	Higher odds of delay
Second-stage disbursement	1.124	3.078	Higher odds of delay
Company institution	1.093	2.984	Higher odds of delay
Prototype output	-1.050	0.350	Lower odds of delay
Funding category: > IDR 5 bn	-0.995	0.370	Lower odds of delay
Publication output	0.994	2.702	Higher odds of delay
Ministry institution	-0.975	0.377	Lower odds of delay
Partner contribution	-0.925	0.396	Lower odds of delay
Intellectual property	0.868	2.383	Higher odds of delay
Monitoring month	0.722	2.058	Higher odds of delay
Energy focus	0.571	1.769	Higher odds of delay
Funding amount	0.565	1.759	Higher odds of delay
Funding category: IDR 1-2 bn	-0.563	0.569	Lower odds of delay

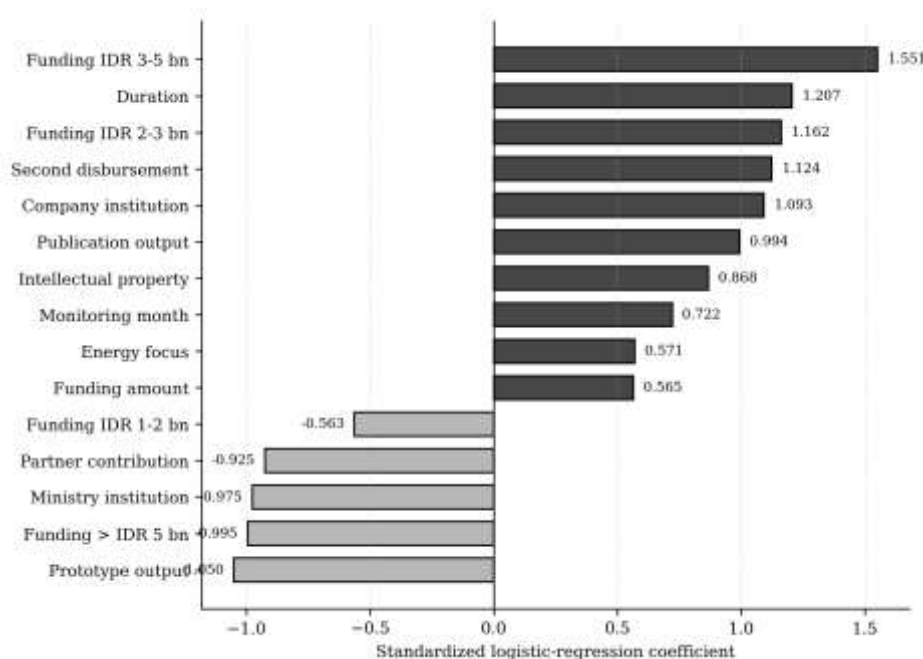


Figure 4. Direction and magnitude of the standardized Logistic Regression coefficients.

Cross-Model Feature Consistency and Text Features

The comparison of feature rankings reinforces the distinction between stable

metadata signals and unstable sparse text signals. Duration was the only feature ranked within the top ten across all three algorithms, with an average rank of 2.33. The IDR 3-5 billion funding category, the IDR 2-3 billion funding

category, second-stage disbursement, the IDR 1-2 billion funding category, monitoring, publication, partner contribution, funding amount, and the Invitation scheme complete the ten most consistent features. Duration, funding range, second-stage disbursement, monitoring, and total funding also align with the non-parametric tests, making them the most credible dimensions.

The leading Random Forest features are dominated by managerially interpretable metadata: duration, monitoring, funding amount, second-stage disbursement, and funding category. In contrast, XGBoost placed several top positions on rare title terms such as “3d,” “economy,” and “anti.” These terms have no defensible substantive mechanism as determinants of delay in national research projects. This pattern suggests that sequential boosting is more prone to learning idiosyncratic residual structure in sparse text, while bagging averages the results of independent trees and is more robust to noise from rare tokens (Sagi & Rokach, 2018).

The title corpus contained 1,149 unique words after preprocessing, and a Pareto analysis showed that the 30 highest-weighted terms

accounted for only about 40% of the total TF-IDF weight. The vocabulary is therefore highly diverse rather than concentrated in a small set of dominant project themes. This helps explain why the textual features are unstable: project titles describe scientific content but do not consistently represent management quality, procurement readiness, or administrative behavior. This result does not mean that text is inherently uninformative. Rich proposal narratives, risk registers, progress reports, reviewer comments, or milestone descriptions may provide more predictive text than short titles.

The methodological implication is that feature sophistication should remain proportional to sample size and information density. A high-dimensional model can appear technically sophisticated while using fragile signals. In a public-sector environment, a model that relies on hard-to-explain words can be difficult to defend and vulnerable to distribution shift. This finding therefore favors structured metadata as the primary source for operational deployment and positions text as an area for future development only when a larger and more informative corpus becomes available.

Table 7. Most consistent features across the three algorithms

Feature	Logistic rank	RF rank	XGB rank	Average rank
Duration	2	1	4	2.33
Funding category: IDR 3-5 bn	1	5	17	7.67
Funding category: IDR 2-3 bn	4	15	11	10.00
Second-stage disbursement	5	4	23	10.67
Funding category: IDR 1-2 bn	24	7	9	13.33
Monitoring	18	2	25	15.00
Publication	9	17	19	15.00
Partner contribution	12	19	30	20.33
Funding amount	23	3	35	20.33
Invitation scheme	39	25	8	24.00

Managerial Implications

The first implication is a shift from uniform monitoring to risk-based monitoring. A project's risk score can be generated when the contract is signed and updated when actual disbursement or monitoring dates become available. Projects with long duration, mid-range funding, delayed second-stage disbursement, or late monitoring can be placed at a higher priority level. This

would allow LPDP analysts and external reviewers to allocate attention according to predicted need rather than applying the same monitoring intensity to every project.

The second implication concerns project design and selection. Long duration should trigger a more detailed examination of work-package dependencies, procurement plans, partner roles, milestone sequencing, and

contingency buffers. The funding results indicate that mid-funded projects warrant special scrutiny. The solution is not to reduce funding mechanically, because projects above IDR 5 billion did not show the highest risk. Instead, LPDP can require stronger project-management readiness for the IDR 2-5 billion range, including milestone-based cash-flow planning and documented coordination mechanisms.

The third implication is to treat disbursement timing as an operational service-level indicator. The second-stage disbursement effect is large and consistent. LPDP can set internal warning thresholds for administrative processing, automatically notify the responsible units when a project approaches those thresholds, and distinguish delays caused by incomplete grantee documents from those caused by internal processing. The predictive model can also be recomputed after each disbursement event, turning the initial risk estimate into a dynamic monitoring score.

The fourth implication is to reconsider monitoring timing. A fixed schedule may be too late for high-risk projects and unnecessarily intensive for low-risk projects. The model supports an adaptive policy in which high-risk projects receive earlier diagnostic monitoring and follow-up coaching, while lower-risk projects remain under standard monitoring. Monitoring should be framed as an intervention point rather than a mere reporting formality, with clear actions, responsible parties, deadlines, and escalation rules.

The fifth implication is to use complementary models rather than presenting algorithm selection as a winner-takes-all decision. Random Forest can serve as the primary screening model because its high recall reduces the likelihood that a delayed project is missed. Logistic Regression can provide an explanatory layer by showing which features increase or decrease the estimated odds. This architecture supports human-AI collaboration: the model prioritizes cases, and project analysts review contextual evidence before allocating resources or changing project status (Bansal et al., 2021).

The source study also developed an interactive predictor using ipywidgets. For production deployment, that prototype would need to be integrated with eRISPRO through a

controlled application interface, versioned preprocessing objects, model-monitoring procedures, role-based access, audit logs, and periodic recalibration. Predictions should not automatically penalize grantees or determine funding decisions. Predictions should trigger review, because historical data contain institutional patterns and limitations that may not remain stable over time.

Theoretical and Methodological Contributions

This study extends project-delay analytics to publicly funded research governance, a context in which outcomes are uncertain and accountability demands are high. The results reinforce the proposition that project complexity is expressed through both design variables and administrative processes. Duration and funding capture scale and interdependence, while disbursement and monitoring capture the operational interface between the institution and the project. The convergence of these variables across statistical tests, Logistic Regression, and Random Forest strengthens their interpretation as portfolio-management signals.

This model comparison also cautions against equating algorithmic complexity with meaningful superiority. Random Forest achieved the highest F1-score and recall, but all three AUC values were practically equivalent and the omnibus test was not significant. This finding is consistent with the no-free-lunch principle and with previous project-delay studies showing that different algorithms can perform similarly while differing in interpretability (Nguyen, 2020; Sahu et al., 2025; Wolpert, 1996). The appropriate model depends on the operational loss function: missing a delayed project, generating unnecessary warnings, or issuing an unexplainable decision.

The contrast between the Random Forest and XGBoost feature rankings offers an additional methodological lesson. When the sample size is limited and text features are sparse, boosting models may focus on unstable residual patterns. A stable predictive workflow therefore requires repeated validation, leakage prevention, feature-plausibility checks, and formal pairwise tests, rather than metric maximization alone. Interpretability methods can support this process, but they do not turn implausible features into substantively valid explanations (Lundberg & Lee, 2017; Molnar, 2022).

Limitations

These findings must be interpreted within several limitations. First, the modeling sample is small, with 142 modeling-ready projects, and the minority class contains relatively few on-time cases. This limits statistical power and makes holdout metrics sensitive to the classification of individual projects. Second, the data come from a single institution and the directly managed RISPRO scheme, so the model cannot be assumed to generalize to other funders or mandatory collaboration programs without external validation. Third, the predictors are limited to the columns available in eRISPRO. Macroeconomic changes, regulatory shifts, procurement constraints, laboratory access, team turnover, and qualitative management capability were not measured directly. Fourth, the study is retrospective and identifies predictive associations rather than causal effects. Finally, the title-based TF-IDF features are sparse and unstable; future text models should use richer documents and a much larger corpus.

CONCLUSION

This study identified the most consistent determinants of completion delay in directly managed RISPRO projects and compared Logistic Regression, Random Forest, and XGBoost using repeated stratified cross-validation. Delay is primarily associated with project scale and complexity, the mid-range funding category, and administrative timing. Duration is the strongest and most stable factor: each increase of approximately 0.60 years raises the estimated odds of delay by 3.34 times. Projects funded at IDR 3-5 billion have the highest estimated odds, 4.72 times the reference category, while projects funded above IDR 5 billion show lower odds. Slower second-stage disbursement and slower monitoring also increase the predicted risk of delay.

Random Forest was selected as the operationally preferred model because it produced the highest mean F1-score, 0.907, and recall, 0.942. However, all three algorithms had nearly identical AUC values and the Friedman test showed no statistically significant overall difference, $p = 0.0692$. The practical conclusion is therefore conditional rather than absolute. Random Forest is the most appropriate early-warning screener when missing a delayed project

is highly costly, while Logistic Regression remains valuable for transparent explanation and public accountability.

LPDP can translate these findings into risk-tiered project assistance, enhanced readiness checks for long-duration and mid-funded projects, service-level controls over second-stage disbursement, and adaptive monitoring schedules. A hybrid decision-support architecture should retain human review and should be recalibrated as new projects are completed. Further research should expand the sample across institutions and funding periods, incorporate richer proposal and progress-report text, evaluate external validity, and examine whether targeted interventions based on risk scores actually reduce delays.

REFERENCES

- Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). Springer.
- Arana-Barbier, P. J. (2023). The relationship between scientific production and economic growth through R&D investment: A bibliometric approach. *Journal of Scientometric Research*, 12(3), 596-602.
- Baccarini, D. (1996). The concept of project complexity: A review. *International Journal of Project Management*, 14(4), 201-204. [https://doi.org/10.1016/0263-7863\(95\)00093-3](https://doi.org/10.1016/0263-7863(95)00093-3)
- Bank Indonesia. (n.d.). Data inflasi. Retrieved October 26, 2025, from <https://www.bi.go.id/id/statistik/indikator/data-inflasi.aspx>
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2021). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9, 2-11.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.
- Bimantara, H. C., Rokim, A., & Rilvani, E. (2025). Data mining untuk estimasi waktu produksi dan pengiriman komponen prefab berdasarkan riwayat proyek. *Jurnal Media Akademik*, 3(7).

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1-30.
- Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895-1923. <https://doi.org/10.1162/089976698300017197>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675-701. <https://doi.org/10.1080/01621459.1937.10503522>
- Giri, M., Zhang, C., Olorunnishola, A., Hada, S., & Pal, U. K. (2025). Developing a construction delay prediction model as a delay-avoidance strategy for public construction projects. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 17(3). <https://doi.org/10.1061/JLADAH.LADR-1276>
- Government of the Republic of Indonesia. (2021). Presidential Regulation of the Republic of Indonesia No. 111 of 2021 concerning endowment funds in education.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Heagney, J. (2016). *Fundamentals of project management* (5th ed.). AMACOM.
- Hidayatuloh, R. (2023). Evaluasi skema pendanaan riset LPDP pendekatan Triple Helix Balance Model dan Open Innovation [Master's thesis, Institut Teknologi Sepuluh Nopember].
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- International Organization for Standardization. (2018). *ISO 31000:2018 risk management - Guidelines*.
- Irawan, B. (2025). Faktor-faktor yang memengaruhi kinerja tim riset pada RISPRO Invitasi Balance Triple Helix Model [Master's thesis, IPB University].
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.
- Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* (3rd ed. draft). Stanford University.
- Juszczak, P., Zio, E., & Grolinger, T. (2002). The impact of data normalization on classification performance. *Proceedings of the 7th International Conference on Database Systems for Advanced Applications*.
- Kannan, R., Abdul Halim, H. A., Ramakrishnan, K., Ismail, S., & Wijaya, D. R. (2022). Machine learning approach for predicting production delays: A quarry company case study. *Journal of Big Data*, 9, Article 94. <https://doi.org/10.1186/s40537-022-00644-w>
- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15. <https://doi.org/10.1145/2382577.2382579>
- Kerzner, H. (2017). *Project management: A systems approach to planning, scheduling, and controlling* (12th ed.). Wiley.

- Kleinbaum, D. G., & Klein, M. (2010). Logistic regression: A self-learning text (3rd ed.). Springer.
- Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2007). Handling missing values in classification problems: A comparative study. *International Journal of Artificial Intelligence and Applications*, 6(4).
- Lembaga Penyelidikan Ekonomi dan Masyarakat, Fakultas Ekonomi dan Bisnis Universitas Indonesia. (2023). Laporan pengukuran dampak program pendanaan riset RISPRO LPDP. LPDP.
- Lock, D. (2020). *Project management* (11th ed.). Routledge.
- LPDP. (2020a). Pedoman pendanaan riset inovatif produktif LPDP 2020. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2020b). Pedoman RISPRO Kolaborasi Internasional. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2021a). Laporan kinerja 2020. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2021b). Laporan tahunan 2020. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2021c). Peraturan Direktur Utama LPDP No. PER-18/LPDP/2021 concerning RISPRO mandatory funding guidelines. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2022a). Laporan kinerja 2021. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2022b). Laporan tahunan 2021. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2023a). Laporan kinerja 2022. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2023b). Laporan tahunan 2022. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2023c). Peraturan Direktur Utama LPDP No. PER-24/LPDP/2023 concerning institutional activity cost standards. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2024a). Laporan kinerja 2023. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2024b). Laporan tahunan 2023. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2025a). Laporan kinerja 2024. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2025b). Laporan tahunan 2024. Ministry of Finance of the Republic of Indonesia.
- LPDP. (2025c). Peraturan Direktur Utama LPDP No. PER-50/LPDP/2025 concerning no-cost extensions for RISPRO funding.
- LPDP. (n.d.-a). Riset Inovatif Produktif (RISPRO). <https://risprolpdp.kemenkeu.go.id>
- LPDP. (n.d.-b). Visi dan misi. Retrieved October 29, 2025, from <https://lpdp.kemenkeu.go.id/tentang/visi-misi/>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9), 1315-1316. <https://doi.org/10.1097/JTO.0b013e3181ec173d>
- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 56-61.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153-157. <https://doi.org/10.1007/BF02295996>
- Meredith, J. R., & Mantel, S. J. (2014). *Project management: A managerial approach* (8th ed.). Wiley.
- Ministry of Finance of the Republic of Indonesia. (2020). Regulation No. 47/PMK.01/2020 concerning the organization and governance of LPDP.
- Ministry of Finance of the Republic of Indonesia. (2025). Regulation No. 32 of 2025 concerning standard input costs for fiscal year 2026.
- Ministry of Higher Education, Science, and Technology. (2025, July 31). Dukung Asta Cita, Kemdiktisaintek kuatkan riset untuk pemerataan dan pertumbuhan ekonomi.

- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Independent publication.
- Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Independent publication.
- Movafaghpour, M. A. (2023). Predicting project delays using a new trended regression tree method. *Shahed Journal of Civil Engineering*.
- Nguyen, L. H. (2020). Empirical analysis of a management function's failures in construction project delay. *Journal of Open Innovation: Technology, Market, and Complexity*, 6(2).
- OECD. (2015). *Frascati manual 2015: Guidelines for collecting and reporting data on research and experimental development*. OECD Publishing. <https://doi.org/10.1787/9789264239012-en>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.
- Prasojo, L. D., Yuliana, L., & Prihandoko, L. A. (2023). Kinerja penelitian di pendidikan tinggi: Analisis PLS-SEM terhadap suasana penelitian, kolaborasi, pendanaan, kompetensi, dan luaran.
- Project Management Institute. (2017). *A guide to the project management body of knowledge (PMBOK guide)* (6th ed.).
- Project Management Institute. (2021). *A guide to the project management body of knowledge (PMBOK guide)* (7th ed.).
- Reuters. (2025, August 22). South Korea's Lee says plans record \$25 billion government spending on research in 2026.
- Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249. <https://doi.org/10.1002/widm.1249>
- Sahu, P., Bera, D. K., Parhi, P. K., & Kandpal, M. (2025). Smart delay prediction: Supervised machine learning solutions for construction projects. *Journal of Mechanics of Continua and Mathematical Sciences*, 20(6).
- Sarsiti, S., & Rosid, T. (2025). Peran inovasi dan teknologi dalam meningkatkan pertumbuhan ekonomi menuju Indonesia Emas 2045. *Sosial Market Journal*, 7(1).
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Turner, J. R. (2020). *Handbook of project-based management: Leading strategic change in organizations* (4th ed.). McGraw-Hill Education.
- Turner, J. R., & Müller, R. (2019). *The handbook of project management: Leading organizational change* (4th ed.). McGraw-Hill Education.
- Uddin, S., Ong, S., & Lu, H. (2022). Machine learning in project analytics: A data-driven framework and case study. *Scientific Reports*, 12, Article 15252. <https://doi.org/10.1038/s41598-022-19728-x>
- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11), e0224365. <https://doi.org/10.1371/journal.pone.0224365>
- Vasić, J., Kecman, N., & Mladenović, I. (2016). Research and development investment as determinant of international competitiveness and economic growth in EU28 and Serbia. *Economic Themes*, 54(2), 282-298. <https://doi.org/10.1515/ethemes-2016-0010>
- Wideman, R. M. (2019). *Project and program risk management: A guide to managing project risks and opportunities*. AEW Services.

- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83. <https://doi.org/10.2307/3001968>
- Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7), 1341-1390. <https://doi.org/10.1162/neco.1996.8.7.1341>
- Yaseen, Z. M., Ali, Z. H., Salih, S. Q., & Al-Ansari, N. (2020). Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability*, 12(4), Article 1514. <https://doi.org/10.3390/su12041514>
- Zhang, L., Wu, X., & Skibniewski, M. J. (2020). Risk identification and assessment in construction projects using machine learning. *Automation in Construction*, 118, Article 103275. <https://doi.org/10.1016/j.autcon.2020.103275>