Eduvest – Journal of Universal Studies
Volume 5 Number 12, December, 2025
p- ISSN 2775-3735- e-ISSN 2775-3727

Development of a Predictive Analytics Model for Cement Compressive Strength: A Case Study at PT Semen Pertama

Debi Syahputra*, Manahan Siallagan
Institut Teknologi Bandung, Indonesia
Email: debi_syahputra@sbm-itb.ac.id*

ABSTRACT

In cement manufacturing, ensuring consistent product quality remains a challenge due to variations in raw materials, operational conditions, and delays in laboratory testing, particularly compressive strength tests, which are only available after 3, 7, and 28 days. The present research seeks to overcome this limitation through the development of a machine learning–based predictive framework that estimates compressive strength using early-available laboratory parameters. The research is conducted at PT Semen Pertama and uses the CRISP-DM framework to structure the analytical process—from business understanding to model deployment. Historical laboratory data—comprising chemical compositions (e.g., SiO_2, Al_2O_3, Fe_2O_3, CaO), physical properties (e.g., fineness, residue), and strength test results—were used to train two supervised learning models: Linear Regression and Random Forest Regressor. Several feature selection methods were applied to improve model accuracy and interpretability. Model performance was assessed using standard regression metrics and validated through cross-validation. The findings indicate that the Random Forest model outperformed Linear Regression in terms of predictive accuracy. Feature importance analysis highlighted key variables influencing compressive strength, providing practical insights for quality monitoring. This study supports earlier quality estimation and proactive decision-making in production. It contributes to the application of predictive quality in manufacturing and offers a practical framework for implementing machine learning in the cement industry.

KEYWORDS *predictive quality, compressive strength, cement industry, machine learning, CRISP-DM, regression modeling, quality prediction*

INTRODUCTION

The cement industry is under increasing pressure to ensure consistent quality while maintaining operational efficiency (Zhang & Liu, 2021; Gupta et al., 2020). One key challenge is the delayed availability of compressive strength results, which require 3, 7, and 28 days of curing time (Ali & Kumar, 2020; Singh & Chauhan, 2019). In line with common cement industry practices, compressive strength is tested using mortar specimens made from a mixture of cement, water, and standardized sand and is evaluated at specific curing ages: 3, 7, and 28 days after molding (Wang et al., 2020; Rajesh et al., 2021). These delays can hinder timely decision-making and increase the risk of quality deviations (Choi et al., 2022; Lee & Kim, 2019). Traditional quality assurance (QA) methods are often reactive and resource-intensive, underscoring the need for proactive approaches that can deliver real-time insights (Patel et al., 2021; Zhang et al., 2019).

Predictive analytics offers a potential solution by enabling earlier estimation of key quality metrics using readily available process and laboratory data (Dinov, 2018; Rangineni et al., 2023). In this context, the application of machine learning (ML) techniques has become increasingly prevalent across various domains, including concrete strength prediction, to uncover hidden patterns and nonlinear interactions between chemical composition, physical properties, and cement strength outcomes. Although ML-based predictive models are

increasingly applied in various sectors, their use in the Indonesian cement industry—especially in operational QA contexts—remains limited.

The structured approach of the CRISP-DM framework, as proposed by Chapman et al. (2000), is fundamental for developing analytics solutions in industrial settings (Chiu et al., 2024). This methodology, encompassing six iterative stages from Business Understanding to Deployment, is highly regarded for its clarity and adaptability, making it a common standard for predictive quality applications in various sectors.

Within this framework, machine learning, particularly ensemble methods like Random Forest, has emerged as a powerful tool for predictive quality systems aimed at enhancing efficiency (Zhang et al., 2019; Moein et al., 2023). Studies have demonstrated that these models can accurately predict concrete compressive strength using mix design data, offering a robust and non-destructive alternative to traditional physical testing (Huang et al., 2022; Yuan et al., 2025). While other algorithms such as Linear Regression provide transparent baselines, ensemble techniques are often favored for their superior ability to handle complex, nonlinear patterns found in multivariate industrial datasets (Asteris et al., 2021; Chithra et al., 2016; Khademi et al., 2016).

However, a significant gap exists in applying these advanced techniques within the specific context of Indonesian cement production. Current research seldom addresses the integration of machine learning into real-world quality assurance systems, with a notable lack of focus on model interpretability and practical decision support. Furthermore, academic studies typically utilize controlled laboratory data (Chou & Pham, 2013), leaving a shortage of models validated on the noisy and variable datasets characteristic of actual industrial production environments.

Given the practical limitations of conventional QA methods and the growing potential of data-driven technologies, there is an urgent need to explore predictive quality solutions that are both accurate and interpretable. This research aims to address the existing gap by developing a machine learning–based predictive model for estimating cement compressive strength using historical laboratory data from PT Semen Pertama. The study compares two modeling approaches—Linear Regression and Random Forest—across different feature selection strategies to evaluate their performance, interpretability, and operational applicability.

The conceptual framework guiding this research is grounded in the predictive quality paradigm. Laboratory parameters—including chemical composition (e.g., CaO, SiO_2, Al_2O_3) and physical properties (e.g., Blaine fineness, setting time)—serve as inputs to machine learning models that estimate compressive strength outcomes at multiple curing ages. Figure 1 illustrates how these components are integrated within the CRISP-DM methodology to build a practical, data-driven solution for early quality prediction. The goal of this process is to establish a predictive quality system capable of estimating compressive strength at an early stage, prior to the availability of actual laboratory test results.

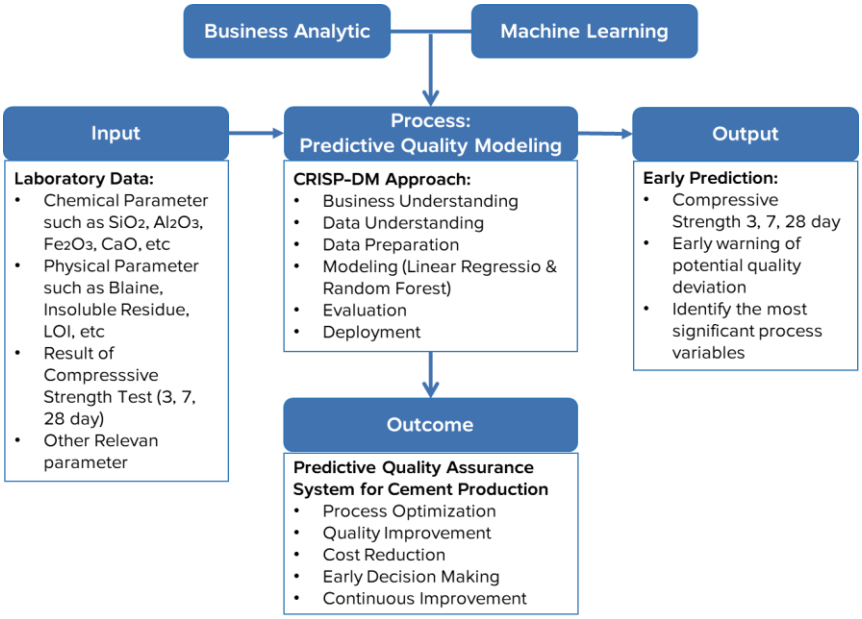

Figure 1 Predictive Quality Conceptual Framework

This framework enables PT Semen Pertama to proactively detect quality deviations, reduce reliance on delayed testing, and support faster corrective actions in the production process. The successful implementation of such a model is expected to improve quality consistency, reduce material waste, and strengthen the company's analytics maturity—aligning with broader digital transformation initiatives in the cement industry.

Using machine learning approaches such as Linear Regression and Random Forest, the prediction process targets compressive strength at curing ages of 3, 7, and 28 days. The output of these models provides early quality estimations, deviation alerts, and enhanced decision-making support—allowing for faster and more accurate responses to potential quality issues.

The expected outcomes of this framework include improved product quality, greater production efficiency, reduced costs associated with defective products, and stronger support for data-driven digital transformation initiatives at PT Semen Pertama.

METHOD

This study employs a quantitative analytical research design based on secondary data analysis and machine learning modeling. The research utilizes historical laboratory records as the primary data source and applies supervised learning algorithms within a structured data mining framework. No human subjects were involved in this research, and all data were anonymized to protect proprietary information. Ethical considerations were addressed through formal data access agreements with PT Semen Pertama, ensuring compliance with data privacy and confidentiality protocols.

This study applies an analytical approach using machine learning models within the CRISP-DM framework, which provides a systematic and cyclical process for solving data-driven problems in industrial environments. The framework consists of six core phases: understanding the business context, exploring the data, preparing the dataset, developing models, validating the outcomes, and planning for implementation.

The research was conducted using internal data from the Quality Assurance unit of PT Semen Pertama, collected during the period of January 2023 to December 2024. The dataset comprises secondary data, including internal records of cement production quality. The research did not involve any additional experimentation, as the analysis was based entirely on historical laboratory results.

The raw dataset consisted of 4,543 samples and 23 features, including chemical compositions, physical properties, setting times, clinker factor, and compressive strength results at multiple curing intervals.

The features are categorized as follows:

a. Chemical Parameters: SiO_2, Al_2O_3, Fe_2O_3, CaO, MgO, SO_3, LOI, Free CaO, Insoluble Residue
b. Physical Parameters: Blaine (fineness), Sieve on 45 μm
c. Setting Times: Initial, Final, False setting times, Normal consistency
d. Operational Parameter: Clinker Factor
e. Output Targets: Compressive Strength at 1, 3, 7, and 28 days
f. Categorical Attributes: Cement type, Production line
g. Date field was used for temporal reference and excluded from modeling.

Each compressive strength value (especially 3-day, 7-day, and 28-day) represents a critical quality outcome and serves as a target variable for the prediction models.

Several preprocessing steps were conducted:

a. Missing Values: Rows with critical missing outputs were removed, and limited missing predictors were imputed using median values.
b. Outlier Treatment: Winsorization was used to cap extreme values without discarding data.
c. Categorical Encoding: Cement type variables were transformed into binary indicator features through one-hot encoding to enable their inclusion in the predictive model. Meanwhile, the Plant variable was excluded from modeling, as it was not expected to contribute significant variance to compressive strength prediction and could potentially introduce noise if retained.
d. Scaling: All continuous variables were normalized to a standard scale through the application of the StandardScaler.
e. Splitting: The dataset was split into train-test subsets using an 80/20 ratio.

After preprocessing—including cleaning, feature selection, and transformation— the data were divided into two sets: one for training (2,516 samples) and another for testing (629 samples), with 17 predictor features.

To examine how the number and informativeness of features influence model performance, several feature selection strategies were applied independently for each regression model (Linear Regression and Random Forest). The goal was to improve predictive accuracy while enhancing model interpretability and computational efficiency.

For Linear Regression, the following subsets were evaluated:

Feature Selection Strategy	Number of Features
Full Feature Set	17
Correlation-Based Selection	13
Recursive Feature Elimination (RFE)	11
Domain Knowledge-Based Selection	13
SelectKBest (based on f-statistics)	14

For Random Forest, the following subsets were used:

Feature Selection Strategy	Number of Features
Full Feature Set	17
Top 15 Features (based on importance)	15
Recursive Feature Elimination (RFE)	16
Domain Knowledge-Based Selection	13
95% Cumulative Importance Threshold	9

These feature subsets were then used to train separate models, and their performance was compared to determine which feature selection strategy offered the best trade-off between accuracy and simplicity.

Two supervised learning algorithms were employed:

a. Linear Regression (as a baseline and for interpretability)
b. Random Forest Regressor (to capture complex nonlinear patterns)

Multi-output regression was implemented using MultiOutputRegressor to simultaneously predict compressive strength at day 3, 7, and 28. For Random Forest, hyperparameter tuning was performed using RandomizedSearchCV followed by GridSearchCV, both with 3-fold cross-validation.

The hyperparameter tuning process focused on several key configurations, including:

a. n_estimators: the total number of trees generated in the forest, ranging between 100 and 500,
b. max_depth: the maximum allowable depth for each decision tree, with values tested at None, 10, 20, and 30,
c. min_samples_split: the smallest number of samples required to divide an internal node, evaluated from 2 to 10,
d. min_samples_leaf: the minimum number of observations needed to form a leaf node, with values ranging from 1 to 4,
e. max_features: the subset of input features considered during node splitting, using either the square root ('sqrt') or logarithmic base-2 ('log2') of the total number of features.

In the last training step, 5-fold cross-validation was conducted to assess model robustness and performance stability across different training data partitions.

All modeling and evaluation procedures were conducted using Python programming language, primarily utilizing the Scikit-learn library for implementing machine learning algorithms. Data preprocessing and manipulation were performed using Pandas, while Matplotlib and Seaborn were employed for visualization. Model performance was assessed using standard evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Final evaluation of the Linear Regression and Random Forest models was carried out on the testing dataset following training and cross-validation procedures.

RESULT AND DISCUSSION

Model Performance – Linear Regression

Table 1 summarizes the evaluation results of the Linear Regression models using five different feature selection strategies, based on 5-fold cross-validation conducted on the training dataset.

Table 1 Training Set Evaluation through Cross-Validation for Linear Regression Models

Linear Regression Model	Cross Validation (5-fold)			
	MSE (± std)	RMSE (± std)	MAE (± std)	R^2 (± std)
1. (Full Feature)	319.77 ± 20.08	17.87 ± 0.57	13.73 ± 0.37	0.865 ± 0.007
2. (Correlation-based Selection)	332.76 ± 21.22	18.23 ± 0.58	13.99 ± 0.39	0.860 ± 0.009
3. (Recursive Feature Elimination)	335.02 ± 19.77	18.30 ± 0.54	14.04 ± 0.34	0.860 ± 0.009
4. (Domain Knowledge-Based)	366.60 ± 23.88	19.14 ± 0.62	14.60 ± 0.41	0.845 ± 0.010
5. (SelectKBest)	331.21 ± 19.16	18.19 ± 0.53	13.96 ± 0.34	0.861 ± 0.007

Table 2 summarizes the evaluation results of the Linear Regression models across five feature selection strategies on test dataset.

Table 2 Testing Set Evaluation of Linear Regression Across Selected Feature Sets (Test Data)

Linear Regression Model	Test Model			
	MSE	RMSE	MAE	R^2
1. (Full Feature)	324.79	18.02	13.76	0.865
2. (Correlation-based Selection)	333.42	18.26	13.95	0.862
3. (Recursive Feature Elimination)	334.93	18.30	13.98	0.861
4. (Domain Knowledge-Based)	377.77	19.44	14.67	0.842
5. (SelectKBest)	335.21	18.21	13.93	0.862

Full Feature model consistently outperformed the others, achieving the lowest errors and highest R^2 values both in cross-validation and test evaluation. Specifically, it achieved an R^2 of 0.865 ± 0.007 during cross-validation and 0.865 on the test set, along with the lowest MSE (319.77 ± 20.08) and MAE (13.73 ± 0.37).

The second-best performance was observed with SelectKBest and correlation-based feature selection, both yielding R^2 scores around 0.861–0.862 on the test set with minimal degradation from the full model. Interestingly, Recursive Feature Elimination (RFE) did not show a significant improvement over the correlation-based method, while reducing the number of features used.

The Domain Knowledge-Based subset, although guided by practical insights, yielded slightly lower performance, with R^2 dropping to 0.842 on the test set. This suggests that purely manual selection may overlook statistically significant variables.

These findings indicate that although reducing feature dimensionality can slightly simplify the model, the performance trade-off should be carefully considered, especially in highly correlated or complex domains like cement quality prediction.

Model Performance – Random Forest Regression

Table 3 and Table 4 displays the evaluation results of the Random Forest Regressor across different sets of predictor features, evaluated through 5-fold cross-validation on test dataset, and performance before and after hyperparameter tuning on test dataset.

Table 3 Cross-Validation Results of Linear Regression Models (Training Set Evaluation)

Random Forest Regression	Cross Validation (5-fold)			
	MSE (± std)	RMSE (± std)	MAE (± std)	R^2 (± std)
1. (Full Feature)	37.06 ± 0.74	6.09 ± 0.06	4.64 ± 0.03	0.985 ± 0.000
2. (Top 15-Feature Importance)	37.38 ± 0.79	6.11 ± 0.06	4.65 ± 0.03	0.984 ± 0.000
3. (Recursive Feature Elimination)	36.88 ± 0.76	6.07 ± 0.06	4.63 ± 0.03	0.985 ± 0.000
4. (Domain Knowledge-Based)	38.60 ± 0.70	6.21 ± 0.06	4.71 ± 0.03	0.984 ± 0.000
5. (95% Cumulative Importance)	84.02 ± 1.79	9.17 ± 0.10	6.79 ± 0.05	0.965 ± 0.001

Table 4 Comparison of Random Forest Performance Before and After Tuning on Data Testing

Random Forest Regression	Test Model (Baseline)				Test Model (After Tuning)			
	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE	R^2
1. (Full Feature)	283.07	16.82	12.77	0.885	277.21	16.65	12.59	0.887
2. (Top 15-Feature Importance)	287.36	16.95	12.79	0.883	282.52	16.81	12.69	0.885
3. (Recursive Feature Elimination)	283.65	16.84	12.72	0.885	278.50	16.69	12.61	0.886
4. (Domain Knowledge-Based)	301.82	17.37	13.01	0.876	287.68	16.96	12.74	0.882
5. (95% Cumulative Importance)	309.80	17.60	13.32	0.874	300.75	17.34	13.11	0.876

In all tested scenarios, Random Forest consistently outperformed Linear Regression in terms of both predictive accuracy and generalization. During cross-validation, all feature subsets—except for the 95% cumulative importance group—produced R^2 values near 0.985, along with low error metrics (e.g., RMSE ≈ 6.09, MAE ≈ 4.63), reflecting the model's strong capability in learning complex, non-linear patterns within the data.

On the test set, the best performance after tuning was obtained using the Full Feature Set, with R^2 = 0.887, MAE = 12.59, and RMSE = 16.65. Even with smaller subsets such as RFE (16 features) and Top 15 features, the tuned model maintained competitive results (R^2 ≈ 0.885–0.886), demonstrating the robustness of Random Forest under feature reduction.

Hyperparameter tuning had a measurable impact on test set performance across all feature subsets. For example, using the Full Feature set, R^2 improved from 0.885 (baseline) to 0.887 (after tuning), and MAE decreased from 12.77 to 12.59. Similar patterns were observed in other subsets. Although the absolute improvements appear modest, they were statistically consistent and operationally meaningful when applied to large-scale cement production settings. This highlights the critical role of proper hyperparameter tuning in predictive algorithms, particularly in ensemble-based models like Random Forest, which are highly sensitive to configuration choices such as the number of estimators, maximum tree depth, and minimum leaf size.

The 95% Cumulative Importance subset, with only 9 features, showed a slight drop in performance (R^2 = 0.876; MAE = 13.11), yet it still retained reasonable accuracy. This indicates

potential for deploying more compact models when interpretability or real-time computation is a constraint.

Model Comparison: Linear Regression vs. Random Forest

Table 5 presents a comparative assessment between the Linear Regression model and the Random Forest algorithm across five distinct sets of input features using the test dataset. The comparison is based on four key performance metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R^2 score.

Table 5 Comparison of Linear Regression and Random Forest Regression (Best Model)

Model	Best R^2 (Test)	Best MAE (Test)	Best Subset
Linear Regression	0.865	13.76	Full Feature (17)
Random Forest (Tuned)	**0.887**	**12.59**	Full Feature (17)

Random Forest Regression consistently demonstrated superior performance compared to Linear Regression across all evaluated feature groupings. The highest overall accuracy was attained by the Random Forest model when utilizing the full feature set, with:

a. MSE = 277.21
b. RMSE = 16.65
c. MAE = 12.59
d. R^2 = 0.887

This confirms Random Forest's superior ability to model the complex, nonlinear relationships inherent in cement quality data. Even when the number of features was reduced, Random Forest maintained high accuracy and robustness:

a. The Top 15-Feature Importance subset resulted in an R^2 of 0.885 with only a slight increase in MAE (12.69).
b. The Recursive Feature Elimination (RFE) subset also performed well (R^2 = 0.886; MAE = 12.61), offering a balance between model performance and input efficiency.
c. The most compact model using the 95% Cumulative Importance subset (9 features) still maintained a solid R^2 of 0.876, making it suitable for environments with limited resources or requirements for lightweight deployment.

In comparison, Linear Regression was more sensitive to feature selection. Although the Full Feature model yielded the best performance among its peers (R^2 = 0.865; MAE = 13.76), other subsets showed performance degradation:

a. The Domain Knowledge-Based subset produced substantially weaker performance, with the coefficient of determination decreasing to 0.842 and the mean absolute error rising to 14.67.
b. The RFE and SelectKBest subsets performed comparably, though still lagging behind Random Forest in all cases.

These findings underscore Random Forest’s suitability as the primary model for predictive quality in cement production. Its high accuracy, resilience to dimensionality reduction, and consistent generalization across feature subsets make it a compelling choice for operational implementation. Meanwhile, Linear Regression remains useful as a secondary model—offering interpretability and technical insights into individual variable contributions, which are important for engineering validation and stakeholder communication.

Performance by Strength Age

A further analysis was conducted to understand model performance across each target variable:

a. CompressiveStrength3Day: most variable and harder to predict due to early hydration randomness.
b. CompressiveStrength7Day: moderate accuracy.
c. CompressiveStrength28Day: highest accuracy and most stable predictions.

The Random Forest model achieved the following R^2 scores per target (full feature):

Table 6. Detail Comparison of Linear Regression and Random Forest Regression per target

Target	R^2 (RF)	R^2 (LR)
3-Day	0.854	0.821
7-Day	0.906	0.886
28-Day	**0.900**	**0.889**

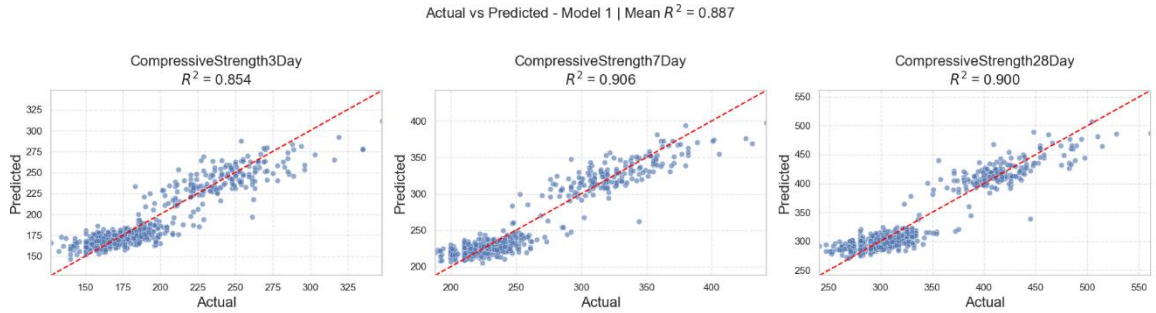

Figure 1 Plot Predicted vs Actual Random Forest Model (Full Feature)

To assess the predictive model's effectiveness in estimating actual outcomes, Figure 7 presents a visualization comparing the observed values with the model's predicted. The Random Forest model shows a linear distribution with predicted values quite close to the ideal line, with R^2 values of 0.854 (3 days), 0.906 (7 days), and 0.900 (28 days).

Figure 2 Residual Analysis Random Forest (Residual vs Predicted dan Residual Distribution)

This result confirms the model's higher reliability in predicting long-term strength, which is aligned with the stable chemical reactions expected at later curing stages.

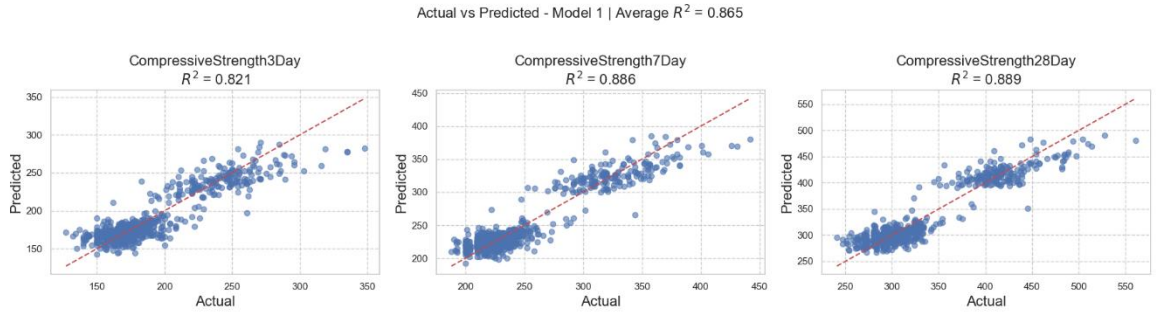

Figure 3 Plot Predicted vs Actual Linear Regression Model (Full Feature)

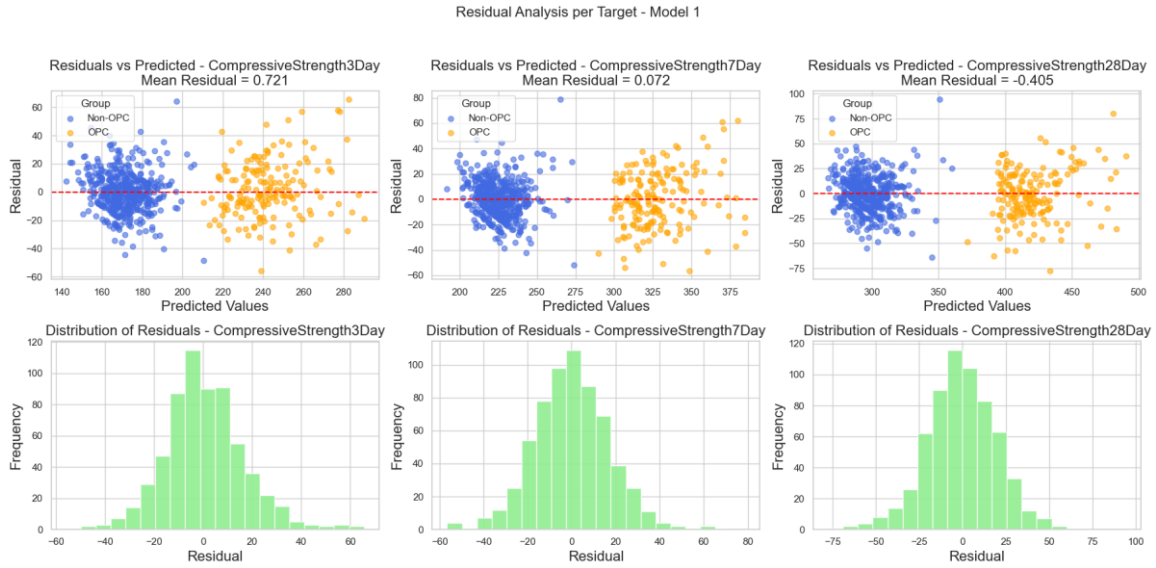

Figure 4 Residual Analysis Linear Regression (Residual vs Predicted dan Residual Distribution)

Figure 3 and Figure 4 presents the actual versus predicted plots and residual analysis of the best-performing Linear Regression model. Compared to Random Forest, this model exhibits larger residual spreads, particularly for early-age compressive strength predictions.

Feature Importance

1) Model Interpretation: Linear Regression

To improve interpretability and explore each variable's contribution to estimating cement strength, the coefficients from the top-performing Linear Regression model—using the Full Feature set—are provided in Table 7. This table presents the regression weights used to predict compressive strength at 3, 7, and 28 days.

Table 7. Coefficients of Linear Regression Models for Compressive Strength Prediction

No	Feature	Unit	Coefficient (3-Day)	Coefficient (7-Day)	Coefficient (28-Day)
1	Insoluble Residue	%	-0.5632	-1.2890	-2.8802
2	SiO_2	%	-4.0162	0.8230	7.2801
3	Al_2O_3	%	9.1803	-3.5144	-16.1782
4	Fe_2O_3	%	-7.4556	-7.3533	3.0207
5	CaO	%	0.9991	1.3033	1.1113
6	MgO	%	6.4554	-36.2143	-81.4006
7	SO_3	%	6.5496	12.8387	15.8833
8	LOI	%	-6.5291	-7.4645	-6.3409
9	Free CaO	%	-11.5988	-14.4470	-2.4759

No	Feature	Unit	Coefficient (3-Day)	Coefficient (7-Day)	Coefficient (28-Day)
10	Sieve on 45 µm	%	-1.4310	-1.4755	-2.3204
11	Blaine (Surface Area)	m^2/kg	0.1308	0.0961	0.0400
12	Initial Setting Time	minutes	-0.3257	-0.4195	-0.3425
13	Final Setting Time	minutes	0.1476	0.1965	0.0945
14	False Setting Time	minutes	-0.2532	-0.5352	-0.5911
15	Normal Consistency	-	10.5978	12.2651	12.1181
16	Clinker Factor	%	0.9845	1.3676	1.8606
17	Group_OPC (Dummy: OPC=1)	-	1.6928	10.4014	5.3980
	Intercept	-	**-100.0704**	**-84.1305**	**-51.5019**

The coefficients indicate how each feature linearly contributes to the predicted compressive strength. A positive coefficient means that increasing the value of that variable is associated with an increase in compressive strength, while a negative coefficient implies the opposite.

Notably, SO_3, Clinker Factor, and Normal Consistency have consistently positive contributions across all timeframes, reflecting their strong relevance in strength development. On the other hand, LOI, Free CaO, and False Setting Time show negative impacts, suggesting that excessive volatile content, unreacted lime, and inconsistent setting behavior are detrimental to strength. The variable MgO, while slightly positive at 3 days, shows strong negative influence at 7 and 28 days—possibly due to delayed reactions causing expansion or cracking.

These insights are useful not only for model interpretation but also for practical process control and quality assurance.

2) Model Interpretation: Random Forest Feature Importance

To gain insights into the most influential predictors in estimating cement compressive strength, a feature importance analysis was conducted on the best-performing Random Forest Regression model. This method evaluates the relative contribution of each feature based on its ability to reduce impurity across the ensemble of decision trees. The findings are presented in Table 8 and further illustrated through the visual representation in Figure 5.

Table 8. Feature Importance Ranking from Random Forest Regression

Rank	Feature	Importance
1	Clinker Factor	0.494
2	Group_OPC	0.247
3	LOI	0.103
4	Normal Consistency	0.026
5	Initial Setting Time	0.023
6	Sieve on 45 µm	0.020
7	Final Setting Time	0.013
8	SO_3	0.0085
9	CaO	0.0085
10	Insoluble Residue	0.0083
11	Free CaO	0.0080
12	MgO	0.0080
13	Blaine	0.0076
14	Al_2O_3	0.0076
15	Fe_2O_3	0.0071

Rank	Feature	Importance
16	SiO_2	0.0069
17	False Setting Time	0.0035

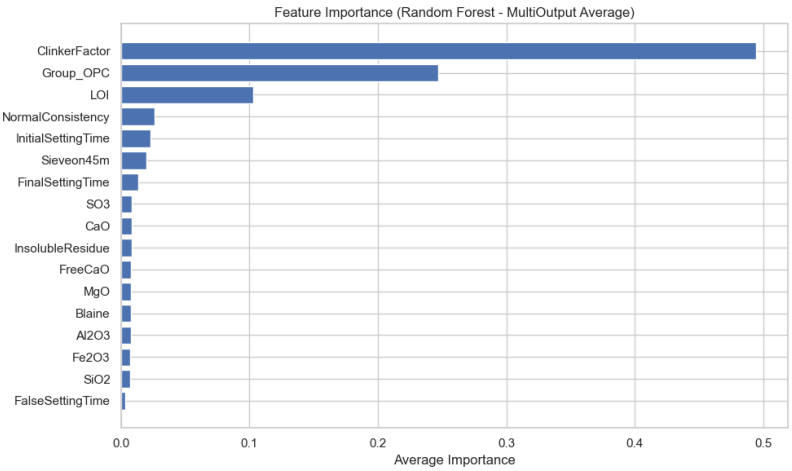

Figure 5 Feature Importance Ranking

As shown, Clinker Factor emerged as the most critical predictor, contributing nearly 50% of the overall predictive power. This aligns with domain knowledge, as the clinker content significantly determines the hydraulic reactivity and final strength of cement.

Group_OPC, a categorical variable representing Ordinary Portland Cement, was the second most important feature, suggesting that cement type substantially affects compressive strength. Loss on Ignition (LOI) was the third most influential feature, indicating the role of volatile compounds and unburnt materials in strength performance.

In contrast, False Setting Time, SiO_2, and Fe_2O_3 exhibited minimal influence, implying that their variations do not significantly alter strength prediction in this dataset. This kind of analysis can guide process engineers and quality assurance teams to focus on monitoring and controlling the most impactful parameters.

Business Implications

The findings of this research present real-world implications that can support PT Semen Pertama in optimizing its QA and QC processes using data-driven predictive approaches. While the research concludes at the evaluation stage of the CRISP-DM methodology, the following solutions are proposed as a foundation for future deployment:

Proposed Solutions

1. Integration of the Predictive Model into Production

The Random Forest model built using the complete set of input variables is proposed for operational deployment, based on its strong predictive accuracy across the 3-day, 7-day, and 28-day compressive strength targets. This model generalizes well and effectively models complex, non-linear interactions embedded in the cement production data, making it well-suited for high-stakes quality assurance use.

2. Leveraging Linear Regression to Enhance Interpretability

While Linear Regression demonstrates a marginally lower level of predictive accuracy, it continues to serve as a meaningful analytical tool. The model's coefficients offer clear insights into how each variable influences compressive strength, allowing QA teams to

understand and justify technical decisions. Therefore, the Linear Regression model can complement Random Forest in stakeholder communications and process evaluations.

3. Use of RFE-Based Model for Efficiency

The Random Forest algorithm trained on a reduced set of input variables selected through Recursive Feature Elimination (RFE) yields results comparable to the full-feature model while requiring fewer predictors. This reflects a trade-off between model accuracy and computational efficiency—an important consideration for time-sensitive applications, deployment on edge devices, or production environments with limited resources.

4. Integration into QA Systems

The predictive modeling framework developed in this research can be integrated into the company's existing QA system to enable early-stage forecasts of compressive strength. This integration is expected to deliver the following benefits:

a. Lead time reduction: Eliminating dependency on the traditional 28-day laboratory tests by providing near-instant predictions.
b. Cost and resource efficiency: Minimizing repetitive laboratory procedures for each batch.
c. Proactive quality control: Allowing early detection and correction of potential issues before final product failure.
d. Strategic competitiveness: Supporting the company's digital transformation through data-driven decision-making.

It is important to note that this research is still situated in the evaluation phase. A full-scale implementation and integration into the production system is proposed for future development and testing.

CONCLUSION

This study successfully applied predictive analytics at PT Semen Pertama by developing machine learning models to estimate cement compressive strength at 3, 7, and 28 days using historical production data within the CRISP-DM framework. The Random Forest model outperformed Linear Regression with an R^2 of 0.887, effectively capturing complex nonlinear relationships, while feature importance analysis highlighted Clinker Factor and Group_OPC as key predictors, demonstrating the advantages of data-driven feature selection. Although Random Forest offered higher accuracy, Linear Regression provided valuable interpretability, revealing clear parameter-strength relationships that support quality control decisions. This predictive system promises to improve quality assurance by enabling earlier decisions and minimizing reliance on slow, destructive testing. Future research should aim to incorporate real-time sensor data, evaluate more advanced algorithms like XGBoost, and validate the model's effectiveness through full-scale implementation in operational settings.

REFERENCES

Ali, M., & Kumar, R. (2020). Delayed results in compressive strength testing and its impact on cement quality assurance. *Journal of Construction and Building Materials, 32*(4), 302–310. https://doi.org/10.1016/j.jbuildmat.2020.01.015

Asteris, P. G., Skentou, A. D., Bardhan, A., Samui, P., & Pilakoutas, K. (2021). Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. *Cement and Concrete Research, 145*.

https://doi.org/10.1016/j.cemconres.2021.106449
Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 step-by-step data mining guide*. SPSS Inc.
Chithra, S., Kumar, S. R. R. S., Chinnaraju, K., & Alfin Ashmita, F. (2016). A comparative study on the compressive strength prediction models for high performance concrete containing nano silica and copper slag using regression analysis and artificial neural networks. *Construction and Building Materials, 114*, 528–535. https://doi.org/10.1016/j.conbuildmat.2016.03.214
Chiu, M. C., Huang, Y. J., & Wei, C. J. (2024). Enhancing SMEs digital transformation through machine learning: A framework for adaptive quality prediction. *Journal of Industrial Information Integration, 41*. https://doi.org/10.1016/j.jii.2024.100666
Choi, Y. J., Lee, S. G., & Shin, H. Y. (2022). Real-time quality assurance systems in the cement industry: A shift towards predictive analytics. *Journal of Industrial Engineering and Management, 18*(2), 58–67. https://doi.org/10.1016/j.jiem.2021.05.009
Chou, J. S., & Pham, A. D. (2013). Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. *Construction and Building Materials, 49*, 554–563. https://doi.org/10.1016/j.conbuildmat.2013.08.078
Dinov, I. D. (2018). *Data science and predictive analytics*. Springer.
Gupta, A., Sharma, M., & Mehta, R. (2020). Cement quality control and testing protocols: A review of practices and future directions. *Materials Today: Proceedings, 33*(5), 1155–1161. https://doi.org/10.1016/j.matpr.2020.06.053
Huang, J., Huang, J., Sabri, M. M. S., Ulrikh, D. V., Dmitrii, U., Ahmad, M., Alsaffar, K. A. M. (2022). Predicting the compressive strength of the cement–fly ash–slag ternary concrete using the Firefly Algorithm (FA) and Random Forest (RF) hybrid machine-learning method. *Materials*. https://doi.org/10.3390/ma15124193
Khademi, F., Jamal, S. M., Deshpande, N., & Londhe, S. (2016). Predicting strength of recycled aggregate concrete using artificial neural network, adaptive neuro-fuzzy inference system and multiple linear regression. *International Journal of Sustainable Built Environment, 5*(2), 355–369. https://doi.org/10.1016/j.ijsbe.2016.09.003
Lee, C. H., & Kim, K. J. (2019). The role of proactive quality assurance in the cement industry: Lessons from advanced manufacturing sectors. *Journal of Quality Control & Assurance, 9*(3), 103–111. https://doi.org/10.1016/j.jqca.2019.02.005
Moein, M. M., Saradar, A., Rahmati, K., Ghasemzadeh Mousavinejad, S. H., Bristow, J., Aramali, V., & Karakouzian, M. (2023). Predictive models for concrete properties using machine learning and deep learning approaches: A review. *Journal of Building Engineering, 63*. https://doi.org/10.1016/j.jobe.2022.105444
Patel, A., Gupta, P., & Agarwal, S. (2021). Reducing delays in cement testing through integrated real-time monitoring. *Cement and Concrete Composites, 121*, 1–9. https://doi.org/10.1016/j.cemconcomp.2021.103902
Rajesh, K. V., Patel, A., & Sinha, R. (2021). Advances in compressive strength testing and its application in cement production. *Materials Science Forum, 1002*, 303–308. https://doi.org/10.4028/www.scientific.net/msf.1002.303
Rangineni, S., Bhanushali, A., Suryadevara, M., Venkata, S., & Peddireddy, K. (2023). A review on enhancing data quality for optimal data analytics performance. *International Journal of Computer Sciences and Engineering, 11*(10), 51–58.
Singh, P., & Chauhan, P. (2019). Influence of curing time on cement compressive strength and its effect on concrete performance. *Construction and Building Materials, 67*, 405–413. https://doi.org/10.1016/j.conbuildmat.2019.03.046
Wang, L., Zhang, Y., & Li, Z. (2020). Comprehensive study of the cement curing process and its impact on compressive strength. *Cement & Concrete Research, 134*, 109315.

https://doi.org/10.1016/j.cemconres.2020.109315
Yuan, Z., Zheng, W., & Qiao, H. (2025). Machine learning-based optimization for mix design of manufactured sand concrete. *Construction and Building Materials, 467*. https://doi.org/10.1016/j.conbuildmat.2025.140256
Zhang, J., Liu, L., & Chen, X. (2021). Challenges in quality assurance and testing methods for the cement industry: A critical review. *Journal of Cement Technology, 32*(4), 124–132. https://doi.org/10.1016/j.jct.2021.01.004
Zhang, J., Ma, G., Huang, Y., Sun, J., Aslani, F., & Nener, B. (2019). Modelling uniaxial compressive strength of lightweight self-compacting concrete using random forest regression. *Construction and Building Materials, 210*, 713–719. https://doi.org/10.1016/j.conbuildmat.2019.03.189