Cek JEME 180

Fecha de entrega: 28-jul-2026 05:48a. m. (UTC-0500)
Identificador de la entrega: 3004647917
Nombre del archivo: Cek_JEME_180.docx (143.85K)
Total de palabras: 2737
Total de caracteres: 17106

Implementation Of the Decision Tree Algorithm in Predicting Delayed Water Bill Payments in A Housing Residence in Bandung

Abstract

The provision of clean water is a fundamental necessity that depends significantly on consumers' prompt payment of bills for its operational sustainability. The administration of a housing complex in Bandung is experiencing significant difficulties with overdue payments, which are affecting its financial and operational stability. The existing strategy for managing delays is reactive, rendering it ineffective at preventing debt accumulation. This research proposes adopting the Decision Tree algorithm as a predictive method to address this issue. This research employs a quantitative methodology that uses predictive techniques, drawing on previous secondary data on customer water bill payments. The technique utilized is CRISP-DM, encompassing phases of business understanding, data understanding, data preparation, modeling, assessment, and implementation. In the data preparation step, missing values are resolved, feature engineering is performed (calculating average monthly payments), irrelevant columns are eliminated, and categorical variables are encoded. The dataset is subsequently split into 80% for training and 20% for testing. The Decision Tree Classifier model is constructed and trained utilizing the processed training data. Model performance is assessed using Accuracy, Precision, Recall, and F1-Score measures, along with Confusion Matrix analysis and feature significance evaluation. The testing results indicate that the model achieves 78% accuracy in forecasting client payment status (On Time or Late). Nonetheless, the model continues to yield satisfactory outcomes as a preliminary alert mechanism for possible delays. This study demonstrates that machine learning can enhance billing efficiency and customer management within the housing complex in Bandung.

Keywords: *Decision Tree, Payment Delay, Prediction, Machine Learning, CRISP-DM*

Introduction

Access to clean water is an essential requirement for communities, including a residential complex in Bandung City. Artesian wells function as a principal source of clean water, autonomously controlled by the locals. The problem of postponed water bill payments has been escalating annually, undermining operational continuity and the upkeep of the distribution system. Manual billing administration and limited use of historical data lead to inefficiencies in identifying consumers at risk of late payments.

This research is grounded in theories and concepts from data mining and machine learning, which have been extensively employed to tackle predictive challenges in the public service sector. CRISP-DM is one of the most recognized frameworks for such initiatives. CRISP-DM provides a thorough and iterative standard methodology comprising six primary stages: business understanding, data understanding, data preparation, modeling, assessment, and implementation. This framework is selected for its adaptability and demonstrated effectiveness across diverse sectors, including utilities and public administration, rendering it particularly appropriate for community-based water billing management (Laila Sari et al., 2026; Y. D. Zhang et al., 2021).

The principal algorithm employed in this study is the Decision Tree, a widely used classification method known for its ability to produce easily interpretable models (D. Zhang et al., 2023). The Decision Tree works by presenting decision rules in a hierarchical tree structure, where each internal node represents a test on an attribute, branches denote the test outcomes, and leaf nodes indicate the final class or judgment. The primary advantage is transparency, as the created rules are comprehensible to non-experts, such as field

officers, thereby facilitating improved operational decision-making (Chen et al., 2021).

The selection of attributes for partitioning in the Decision Tree is conducted using impurity criteria. This study employs Gini Impurity, which underpins the CART (Classification and Regression Trees) algorithm. The Gini Impurity formula is defined as follows.

$$Gini = 1 - \sum(P_i)^2$$

Where pi denotes the fraction of data from class i within a node, the attribute exhibiting the minimal Gini Split value is chosen as the splitter, as it produces the most "pure" outcomes. This method is effective for datasets comprising both qualitative and numerical attributes, such as delay history and average payment dates.

The CRISP-DM framework and the Decision Tree method function within the domains of Data Mining and Machine Learning. Data Mining is the process of extracting meaningful patterns from large datasets (Kanaya & Panggabean, 2025; Lee et al., 2022; Muhammad Tantowi Jauhari et al., 2025). In contrast, Machine Learning is a domain of computer science that enables systems to learn from data autonomously, without explicit programming (Firmansyah & Yulianto, 2021; Jaya & Hidayati, n.d.). The integration of these two domains enables the conversion of past consumer payment data into proactive predictive models, in contrast to the manual, reactive methodologies currently employed.

Complementary standard measures are utilized to assess the performance of the categorization model. Accuracy quantifies the overall proportion of correct predictions; however, it may be deceptive in cases of class imbalance. Consequently, Precision (the ratio of true positives to expected positives), Recall (the ratio of correctly detected actual positives), and F1-Score (the harmonic mean of precision and recall) are employed to provide a more balanced

perspective (Affa Hurrarul Arayadiba et al., 2025; Nurhasanah & Zufria, 2024). The confusion matrix serves as a primary visualization tool for comprehensive model performance analysis.

This literature review informs the integration of the CRISP-DM methodology, the Decision Tree algorithm utilizing Gini Impurity criteria, and machine learning principles to formulate pertinent and effective predictive solutions for the administration of a housing complex in Bandung. This method is supported by empirical evidence from the scientific literature and has considerable potential to improve operational efficiency through data-driven decision-making.

A data-mining approach is required to evaluate consumer payment trends and resolve this issue. The Decision Tree method is selected for its capacity to generate simply interpretable decision rules. This work employs Gini Impurity as the splitting criterion to construct an efficient decision tree.

This study employs secondary data derived from client payment records from the preceding month. The analytical procedure adheres to the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, encompassing six phases: business understanding, data understanding, data preparation, modeling, assessment, and implementation. This study's results are anticipated to offer strategic recommendations to Artesis management on prioritizing collections and improving payment collectability rates.

Research Method

This study utilizes a quantitative methodology with predictive techniques to develop a model that forecasts delayed bill payments in a housing complex in Bandung City. The technique adheres to the CRISP-DM framework, comprising six principal stages: business

understanding, data understanding, data preparation, model creation, model evaluation, and model implementation. This methodology is selected for its methodical, replicable characteristics and demonstrated efficacy in several data mining initiatives, including those within the public sector and public services.

Methodology (CRISP-DM Framework), the process commences with the Analysis of Business Dynamics, aimed at identifying operational issues within Artesis, specifically the increasing frequency of clients who are postponing payments and disrupting operations. The main goal of this phase is to create a predictive system that identifies clients at risk of delayed payments, enabling management to take preemptive action. The final success of this project depends on the model's accuracy and practical utility in daily decision-making. The Comprehension of Data phase employs previous payment records from occupants of a housing complex in Bandung City to accomplish this. This dataset comprises essential variables, including customer ID, delay history, average payment dates, payment status (late or on time), and fundamental demographic information such as gender. In this initial phase, data exploration is conducted to comprehensively analyze the dataset's structure, quality, and trends, including identifying missing values and evaluating data distribution across groups.

The Data Preparation phase employs stringent methods to produce high-quality data suitable for model building. Initially, data cleaning addresses missing values in columns like gender by replacing them with the mode. Subsequently, feature engineering encapsulates client payment behavior by creating additional features, such as average_payment, in numerical form. Categorical variables, such as gender and delay history, are subsequently converted into numerical representations using LabelEncoder. The dataset is

stratified into 80% for training and 20% for testing using the train_test_split function from scikit-learn, with random_state=42 to guarantee repeatability.

After data preparation, the model is constructed using the Decision Tree Classifier algorithm with Gini Impurity as the splitting criterion. This method is chosen for its effectiveness in handling both categorical and numerical data, as well as its ability to produce models that are readily understandable to non-technical management. The training approach involves creating the model with DecisionTreeClassifier(random_state=42) and training it with dt_model.fit(X_train, y_train).

Upon completion of training, the Assessment step appraises the model using test data to evaluate its predictive capability rigorously. The performance metrics utilized include accuracy, which evaluates the ratio of correct predictions to total forecasts, as well as precision, recall, and F1-score to provide a thorough assessment of the minority class (late clients). A confusion matrix delineates the precise counts of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN).

Finally, the Execution phase converts the model's results into strategic suggestions for Artesis management. The system can operate periodically, such as monthly, to produce a list of high-risk clients for early alerts or targeted intervention. To improve accessibility and operational efficiency, the model can be integrated into a specialized web or mobile platform.

Result and Discussion

Model Implementation Results

This study utilizes a quantitative methodology with predictive techniques to develop a model that forecasts delays in water bill payments within a housing complex in Bandung City. The technique

adheres to the CRISP-DM framework, comprising six primary stages: business understanding, data understanding, data preparation, model creation, model evaluation, and model implementation. This method is chosen for its methodical, replicable characteristics and demonstrated efficacy in several data mining initiatives, including those within the public sector and public services.

The model implementation proceeds in phases using the Google Colaboratory environment and Python tools, including Pandas for data manipulation and Scikit-learn for machine learning modeling. The procedure encompasses data preprocessing, training the Decision Tree model, and performance evaluation.

In the preparation phase, datasets from each region were effectively cleaned and converted. Missing values in columns, such as gender, were imputed using the mode. A new feature, average_payment (in numeric date format), was derived from historical payment data to indicate consumer behavior. Furthermore, categorical variables such as delay_history and payment_status were encoded as numerical values using LabelEncoder. The final dataset consisted of more than 100 entries, allocated into 80% training data and 20% testing data in a stratified fashion.

The Decision Tree model was effectively trained utilizing the Gini Impurity criterion. The training approach yielded an interpretable decision tree structure, with Delay History as the root node, signifying that this feature is the most effective splitter for forecasting payment status.

Model Evaluation, the model's prediction performance was assessed using the test data. The evaluation findings indicate satisfactory performance based on the following metrics:

a. Accuracy: 78%
b. Precision (Late Class): 76 percent
c. Recall (Late Class): 80 percent

d. F1-Score (Late Class): 78%

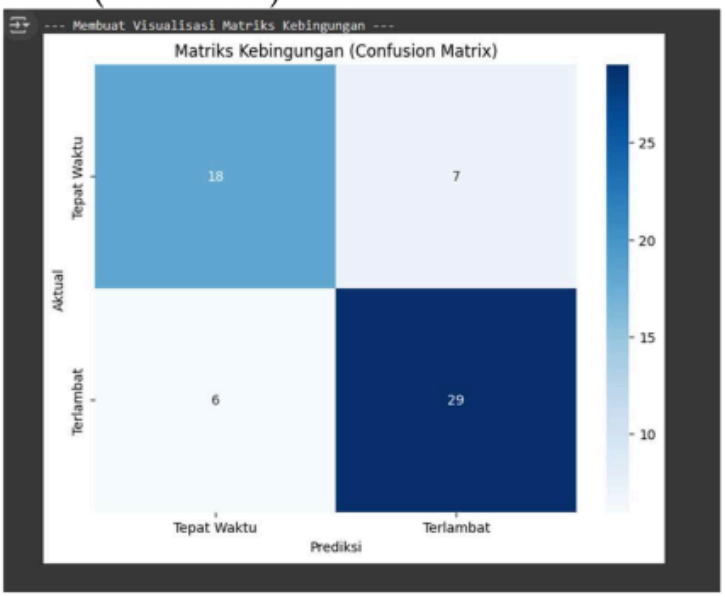


Figure 1. Confusion Matrix

The confusion matrix illustrates:

a. True Positive (TP): 29 (customers anticipated to be late who were indeed late)
b. False Positive (FP): 7 (punctual customers incorrectly predicted as tardy)
c. False Negative (FN): 6 (delayed clients erroneously classified as punctual)
d. True Negative (TN): 18 (clients projected to be on time who were indeed on time)

The findings demonstrate that the model effectively identifies customers at risk of late payments (high recall), despite some classification inaccuracies (false positives and false negatives).

The evaluation results indicate that the Decision Tree model using Gini Impurity is effective for forecasting payment delays in a housing complex in Bandung City. The 78% accuracy indicates that the model can serve as a decision-making tool for management.

The Delay History feature was recognized as the primary determinant in the projections, consistent with the rationale that

customers who have previously exhibited tardiness are predisposed to future delays. This corroborates prior research indicating that historical conduct is a significant predictor of payment classification.

Despite the model's lack of integration into an automated information system, management can directly employ the prediction results to construct a priority billing list. For instance, consumers who are anticipated to be "late" may receive proactive warnings manually via SMS or house visits.

The primary constraint of this study is the comparatively short dataset, which includes only two regions. Furthermore, the existing model fails to account for exogenous variables, such as economic conditions or seasonal changes. This research establishes a robust platform for subsequent development using the methodical CRISP-DM methodology.

This research's execution encompasses both the technical application of the model and the utilization of the results as strategic recommendations for managing Artesis in a housing complex in Bandung City.

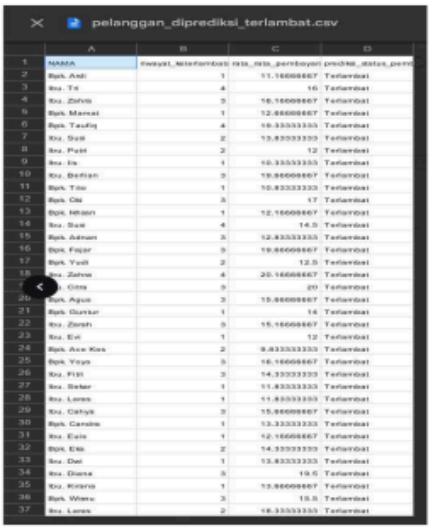


Figure 2. Implementation Results

The established model can function as an early warning system to detect consumers with a high probability of delayed payments. The subsequent steps can execute the implementation:

a. Monthly Data Updates: Management inputs the most recent payment data into the system.
b. Automated Forecasting: The model is executed to produce forecasts regarding payment status for the subsequent month.
c. Customer Segmentation: Customers are classified according to the predictive outcomes (high risk, medium risk, low risk).
d. Proactive Follow-Up: High-risk clients are provided with timely reminders, installment proposals, or opportunities for negotiation regarding relief.

This model can be incorporated into a web or mobile-based management information system for Artesis, enhancing automation and efficiency in the long term. The creation of this information system may evolve into a pertinent subsequent endeavor.

Consequently, the implementation of this model is both technological and operational, facilitating management's transition from a reactive to a proactive receivables management strategy.

Conclusion

This study aims to forecast delays in water bill payments at Artesis, a housing complex in Bandung, using the CRISP-DM methodology and the Decision Tree algorithm. The CRISP-DM stages have been methodically executed, encompassing business understanding of the importance of forecasting delays, analysis of historical payment data, data preparation through normalization and target variable formulation, and modeling using Decision Tree techniques. The evaluation findings reveal that the model achieves 78% accuracy, demonstrating its proficiency in classifying client payment status as

on time or late. The confusion matrix shows a high recall rate for the "Late" class, demonstrating the model's effectiveness at detecting customers at risk of late payment.

This study has three primary contributions. The practical contribution enables Artesis management to directly apply the prediction model's results to identify clients at risk of late payments, facilitating more proactive billing practices, such as early reminders or special relief offers. Secondly, the technical contribution demonstrates the utilization of the Decision Tree algorithm within the CRISP-DM framework as a data-driven solution that is straightforward to execute and analyze, especially in the face of technological resource constraints. Ultimately, the strategic contribution promotes data-driven decision-making, facilitating a shift from a reactive to a proactive strategy, hence boosting operational efficiency and revenue stability for Artesis. This research effectively resolves the identified problem and objectives, creating a valid and practical prediction model for managing future payment delays.

Based on the research findings and identified limitations, the subsequent narrative delineates the strategic recommendations for the management of Artesis at the housing complex in Melong, Bandung, as well as guidance for future research. To enhance everyday operations, management can promptly implement prediction-based strategic recommendations by leveraging the model's predictive results to formulate more proactive billing plans. Rather than pursuing indiscriminate collection efforts, personnel can focus on high-risk clients—those expected to be "delinquent"—for prompt notifications, personal follow-up, or customized negotiation strategies, a transition that staff members strongly endorse. To ensure the seamless integration of this predictive model, Artesis should

prioritize the digitalization of its record-keeping processes, transitioning from manual ledger administration to digital spreadsheets or basic software to improve data updates and analysis. In the long term, management can enhance productivity by establishing an automated reminder system that integrates the model's risk projections with automatic WhatsApp or SMS notifications sent two to three days before the due date for high-risk clients.

Future studies should prioritize enhancing data quality and breadth by broadening the dataset to include more than just historical delay data. Integrating further predictor variables—such as total invoice amount, number of household members or water users, customer employment and financial conditions, and seasonal payment trends—will markedly improve the model's precision and applicability. Moreover, researchers should explore alternative machine learning algorithms, particularly ensemble methods such as Random Forests, XGBoost, and Support Vector Machines (SVMs), which generally produce more reliable and accurate predictive outcomes than a single Decision Tree when handling augmented datasets. Ultimately, subsequent research should conduct temporal model validation using longitudinal data to confirm that the system's predictive efficacy remains consistent and resilient as consumer activity patterns vary month to month.

References

Cek JEME 180

INFORME DE ORIGINALIDAD

12%	11%	8%	3%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	arxiv.org Fuente de Internet	2%
2	ijadis.org Fuente de Internet	1%
3	Jaskaran Singh, Meenu Gupta, Rakesh Kumar. "Smart Technologies and Intelligent Computing", CRC Press, 2026 Publicación	1%
4	Submitted to Monash University Trabajo del estudiante	1%
5	www.arpnjournals.org Fuente de Internet	1%
6	events.saip.org.za Fuente de Internet	1%
7	Bhabani Prasanna Pattanaik, Dhirendra Kumar Shukla, Sambit Satpathy, Kamalakanta Muduli. "Sustainable Developments in Computer Engineering, Green Technology and Smart Systems", CRC Press, 2026 Publicación	1%
8	tefbacktrilramb.web.app Fuente de Internet	<1%
9	Submitted to University of Malaya Trabajo del estudiante	<1%
10	staging.preprints.org Fuente de Internet	<1%
11	www.codewithc.com Fuente de Internet	<1%
12	www.ijiis.org Fuente de Internet	<1%
13	Saurav Mallik, Soumita Seth, Ben Othman Soufiene. "HealthTech Horizons - Charting the Future of Smart Healthcare Innovations, Deepfake and Metaverse", River Publishers, 2026 Publicación	<1%
14	essd.copernicus.org Fuente de Internet	<1%

15	orca.cardiff.ac.uk Fuente de Internet	<1 %
16	dokumen.pub Fuente de Internet	<1 %
17	enstituois.sakarya.edu.tr Fuente de Internet	<1 %
18	pmc.ncbi.nlm.nih.gov Fuente de Internet	<1 %
19	researchcommons.waikato.ac.nz Fuente de Internet	<1 %
20	www.atlantis-press.com Fuente de Internet	<1 %

Excluir citas	Apagado	Excluir coincidencias	Apagado
Excluir bibliografía	Activo		