Online Journal of Management, Innovation, Economics, and Digital Studies (ONMIND)
Vol. 1 No. 1 (2026)

Decision Tree-Based Customer Churn Prediction in Banking: Predictive Performance, Data Leakage Detection, and Managerial Implications

Muhammad Ramdan Alqadri[1], Ahmad Syarif Hidayatullah[1], Rizal Bakri[1,*], Muh. Qardawi Hamzah[1]

[1]Department of Digital Business, Faculty of Economics and Business, Universitas Negeri Makassar, Makassar 90222, *Indonesia*

Abstract

Customer churn remains a critical strategic challenge for banking institutions operating in increasingly competitive digital markets. This study constructs and evaluates a Decision Tree classifier based on the Classification and Regression Trees (CART) algorithm to predict customer churn using a publicly available dataset of 10,000 banking customers distributed across three countries. A scikit-learn Pipeline integrating OneHotEncoder preprocessing and a CART-based classifier with Gini impurity criterion and constrained depth (max_depth = 5) was trained, evaluated, and critically interpreted. The model achieved near-perfect test-set performance, with accuracy of 99.85%, precision of 99.75%, recall of 99.51%, F1-score of 99.63%, and ROC-AUC of 99.75%, with a train-test performance gap below 0.2% across all metrics. Feature importance analysis, however, revealed that the Complain variable alone accounted for 99.80% of the model's Gini importance, indicating potential data leakage between the predictor and target variables. This finding constitutes the primary methodological contribution of the study: high classification performance is a necessary but insufficient criterion for validating a model's operational suitability, and feature importance auditing must be a mandatory component of any production-bound machine learning pipeline. Beyond the technical dimension, the study derives actionable managerial implications for banking customer retention, positioning real-time complaint monitoring as the most operationally significant early warning mechanism for churn risk.

Keywords: *bank customer churn; CART algorithm; customer retention; data leakage; decision tree; machine learning*

1. Introduction

The banking sector has undergone profound structural transformation in the digital era. The emergence of financial technology companies, neobanks, and digital payment platforms has fundamentally altered the competitive dynamics of retail banking, placing unprecedented pressure on incumbent institutions to retain their existing customer base. In this environment, customer churn, defined as the voluntary discontinuation of a banking relationship by a customer, carries substantial economic consequences. Reichheld and Sasser (1990) demonstrated that a five-percent improvement in customer retention can increase institutional profitability by 25 to 95 percent, a finding that continues to anchor the strategic importance of retention management across the financial services industry.

*Corresponding author:
E-mail address: rizal.bakri@unm.ac.id

The cost asymmetry between customer acquisition and customer retention further compounds the urgency of churn prevention. Empirical evidence consistently suggests that acquiring a new customer costs between five and seven times more than retaining an existing one (Kumar and Reinartz, 2018). With global banking churn rates estimated at 15 to 25 percent annually, the economic stakes of early churn detection are substantial. Despite this, many banking institutions continue to rely on manual, rule-based identification systems that are neither scalable nor capable of capturing the multivariate, nonlinear patterns that characterize customer exit behavior.

Machine learning offers a fundamentally different approach. By learning decision boundaries directly from historical data, classification algorithms can identify complex combinations of demographic, financial, and behavioral features that predict churn with far greater precision than conventional heuristics. Among available algorithms, the Decision Tree classifier holds a distinctive position in financial and managerial applications due to its high interpretability. Unlike ensemble methods such as Random Forest or Gradient Boosting, a Decision Tree produces a transparent set of IF-THEN rules that can be directly communicated to business stakeholders, audit committees, and regulatory bodies (Murindanyi et al., 2023). This transparency is not merely a technical convenience; in regulated industries such as banking, explainability is increasingly recognized as a governance and compliance requirement.

Prior research has established the effectiveness of machine learning for churn prediction across multiple sectors. Rahman et al. (2020) applied K-Nearest Neighbors, Support Vector Machines, Decision Trees, and Random Forests to a banking churn dataset, finding that Random Forest yielded the highest predictive accuracy among the compared algorithms. Muneer et al. (2022) reported that Random Forest achieved an F1-score of 91.90% on a banking churn dataset, outperforming several alternative classifiers. In a multi-level architecture, Ngo et al. (2024) achieved 91.08% accuracy and 98% ROC-AUC by stacking KNN, XGBoost, Random Forest, and SVM at the base level with deep learning models at the meta level. Collectively, these studies establish that machine learning classifiers can substantially outperform rule-based churn identification, while leaving open the question of how the predictive power of such models is distributed across individual features.

Despite this rich literature, three gaps remain underaddressed. First, most published studies emphasize aggregate performance metrics without critically examining the source of predictive power at the individual feature level. Second, the detection and transparent reporting of data leakage patterns in banking churn datasets has received insufficient systematic attention in the applied machine learning literature. Third, the managerial translation of feature importance findings into actionable, tiered retention program designs remains underdeveloped.

The present study addresses these gaps through three specific and novel contributions. First, it constructs and documents a fully reproducible, end-to-end CART Decision Tree pipeline using scikit-learn for bank customer churn classification on a 10,000-record dataset. Second, it conducts a rigorous feature importance audit that uncovers and critically interprets the dominant, leakage-indicating role of the Complain variable, demonstrating that performance auditing must extend beyond accuracy metrics to encompass causal feature validation. Third, it derives structured, tiered managerial implications from this technical finding, bridging the gap between algorithmic output and banking retention strategy.

The remainder of this article is organized as follows. Section 2 reviews the literature on customer churn, Decision Tree methodology, feature importance, and model evaluation. Section 3 describes the methodology, including dataset characteristics, preprocessing pipeline, model specification, and evaluation framework. Section 4 presents the empirical results supported by four figures. Section 5 discusses findings in relation to prior literature and managerial implications. Section 6 concludes with recommendations for future research.

2. Literature Review

2.1. Customer Churn in the Banking Industry

Customer churn in the banking industry refers to the process by which customers cease their relationship with a financial institution by closing accounts, withdrawing deposits, or transferring services to a competitor. The phenomenon has received sustained scholarly attention because of its direct and measurable impact on institutional revenue and profitability. Coussement and Van den Poel (2008) were among the early contributors who formalized churn prediction as a supervised classification problem, demonstrating that support vector machines could outperform logistic regression in subscription-based service contexts when properly parameterized. Their framework established methodological precedents that have since been widely adopted in banking research.

A critical driver of churn research is the observed economic asymmetry between customer acquisition and retention. Reichheld and Sasser (1990) provided the foundational empirical basis, documenting that even modest improvements in retention rates translate into disproportionate gains in long-run profitability. Kumar and Reinartz (2018) elaborated on this relationship within the broader framework of customer relationship management, arguing that lifetime value models must be integrated with churn risk assessments to allocate retention resources efficiently. Verbeke et al. (2011), examining churn prediction in the telecommunications sector, introduced profit-driven evaluation criteria that shifted the focus from purely statistical performance metrics toward business-relevant measures, presaging a broader trend that is particularly salient in banking where the cost of failing to detect a churning customer substantially exceeds the cost of misidentifying a loyal customer.

2.2. Machine Learning Approaches to Churn Prediction

The application of machine learning to churn prediction has evolved substantially over the past two decades. Mienye and Sun (2019) conducted a systematic review of improved decision tree algorithms, documenting that variants of ID3, C4.5, and CART consistently outperform simple heuristic baselines across multiple domains including banking, healthcare, and retail, while retaining a unique combination of interpretability and competitive predictive accuracy. Among the most widely benchmarked algorithms, Random Forest has emerged as a dominant performer. Murindanyi et al. (2023) demonstrated that Random Forest achieved 99% accuracy and 98.5% recall on the Berka banking dataset, and further applied LIME and SHAP frameworks to produce accountable, explainable predictions. Muneer et al. (2022) found that Random Forest achieved an F1-score of 91.90% on a banking churn dataset, outperforming Support Vector Machines and AdaBoost under the same experimental conditions.

Gradient boosting variants have also demonstrated strong performance. Han (2024) compared XGBoost, CatBoost, and LightGBM on a Kaggle banking dataset, finding that LightGBM yielded the best overall performance, with customer age and number of products identified as the most predictive variables. In a more architecturally ambitious framework, Ngo et al. (2024) proposed a multi-level stacking model integrating KNN, XGBoost, Random Forest, and SVM at the base level with Logistic Regression, Recurrent Neural Networks, and Deep Neural Networks at the meta level, achieving 91.08% accuracy and 98% ROC-AUC. Despite the dominance of ensemble methods, single decision tree models retain important deployment advantages. Charbuty and Abdulazeez (2021) surveyed decision tree classifiers across multiple application domains, concluding that their transparency, computational efficiency, and native feature importance scoring make them uniquely suited for settings where model explainability is a regulatory or managerial requirement. Mienye and Sun (2024) reinforced this view in a comprehensive survey of decision tree algorithms, documenting that CART remains the foundational implementation underlying modern ensemble methods including Gradient Boosting, XGBoost, and LightGBM.

2.3. Feature Importance and Model Evaluation

Rigorous model evaluation is a central concern in churn prediction research, particularly because headline accuracy figures can obscure a classifier's true discriminative behavior. Hosmer and Lemeshow (2000) established that accuracy alone is an insufficient performance indicator and recommended ROC-AUC as a threshold-independent measure of discriminative ability. Fawcett (2006) formalized this perspective through a comprehensive treatment of ROC analysis, providing the conceptual basis for evaluating classification quality independently of any single decision threshold. These evaluation principles are especially relevant when a model attains near-perfect scores, since such results warrant scrutiny of whether the performance reflects genuine learning or an artifact of the data.

Feature importance analysis represents an equally critical validation component. When a single feature dominates model predictions, it signals either a genuinely strong predictor or a data leakage artifact arising from the simultaneous or near-simultaneous recording of predictor and target information. Daniya et al. (2020) formalized the mathematical relationship between Gini impurity and feature selection in CART, establishing the theoretical basis for interpreting Gini importance scores. The critical evaluation of dominant features is therefore not merely a technical exercise but a prerequisite for validating a model's causal plausibility and operational deployability, a consideration that has received insufficient attention in the applied churn prediction literature.

2.4. Theoretical Framework

This study is grounded in the supervised machine learning paradigm formalized by Mitchell (1997), wherein a program is said to learn from experience E with respect to task T and performance measure P if its performance on T improves with E. Customer churn prediction constitutes a canonical binary classification task: the experience consists of labeled historical customer records, the task is to assign each customer to one of two classes (churn or stay), and the performance is measured via the five metrics described in Section 3.6. Within this paradigm, the Decision Tree classifier operationalizes the CART algorithm originally proposed by Breiman et al. (1984). CART constructs a binary tree through recursive partitioning, selecting at each node the feature and threshold that minimizes the weighted Gini impurity of the resulting child nodes. Gini impurity for a node t with class proportions p_i is defined as Gini(t) = 1 minus the sum of p_i squared, ranging from 0 for a pure node to 0.5 for binary classification with equal class distribution. The algorithm halts when a stopping criterion is met, yielding a tree whose leaf nodes assign class labels by majority vote. This framework provides the explicit, rule-based predictions that characterize interpretable machine learning, making CART particularly suitable for managerial communication in banking contexts.

3. Methods

3.1. Research Design

This study adopts a quantitative, positivist research design grounded in the experimental tradition of machine learning research. The research objective is explanatory and predictive: to evaluate the discriminative performance of a Decision Tree classifier on a banking churn classification task, and to critically interpret the sources of that performance through feature importance analysis. The experimental approach follows the standard machine learning pipeline comprising data preparation, model training, evaluation, and interpretation, in accordance with the methodology recommended by Pedregosa et al. (2011). Figure 1 presents the complete research flowchart.

3.2. Dataset

The dataset used in this study is the Bank Customer Churn Dataset (churbank.csv), a publicly available benchmark dataset accessible via Kaggle. The dataset comprises 10,000 customer records from a simulated multinational bank operating across three countries: France, Germany, and Spain. Each record contains 18 attributes encompassing administrative identifiers, demographic variables, financial behavior indicators, and service engagement metrics. Table 1 presents a structured description of all dataset attributes.

Table 1. Dataset Attribute Description (churbank.csv, n = 10,000)

No	Attribute	Type	Description
1	RowNumber	Integer	Sequential row index; retained in feature matrix but non-informative
2	CustomerId	Integer	Unique customer identifier; retained in feature matrix but non-informative
3	Surname	Object	Customer surname; encoded via OneHotEncoder
4	CreditScore	Integer	Credit worthiness score (350 to 850)
5	Geography	Object	Customer country: France, Germany, or Spain
6	Gender	Object	Customer gender: Male or Female
7	Age	Integer	Customer age in years (18 to 92)
8	Tenure	Integer	Years as a bank customer (0 to 10)
9	Balance	Float	Current account balance (monetary units)
10	NumOfProducts	Integer	Number of bank products used (1 to 4)
11	HasCrCard	Integer	Credit card ownership: 1 = Yes, 0 = No
12	IsActiveMember	Integer	Active membership status: 1 = Yes, 0 = No
13	EstimatedSalary	Float	Estimated annual customer salary
14	Exited	Integer	Target variable: 1 = Churned, 0 = Stayed
15	Complain	Integer	Complaint history: 1 = Filed, 0 = None
16	Satisfaction Score	Integer	Service satisfaction rating (1 = very low, 5 = very high)
17	Card Type	Object	Card tier: Diamond, Gold, Silver, or Platinum
18	Point Earned	Integer	Accumulated reward points (119 to 1,000)

The target variable, Exited, shows a distribution of approximately 4:1, with 7,962 customers classified as non-churners (79.62%) and 2,038 as churners (20.38%). This study did not apply any resampling or class-rebalancing technique; the model was trained on the data in its original distribution, and the analysis instead relies on class-sensitive evaluation metrics including recall, F1-score, and ROC-AUC to characterize performance fairly. No missing values or duplicate records were identified in the dataset, confirmed via df.isnull().sum() and df.duplicated().sum() both returning zero, obviating the need for imputation procedures.

3.3. Preprocessing Pipeline

Data preprocessing was implemented using a scikit-learn ColumnTransformer that applied two transformation strategies in parallel. Numerical features (13 columns) were passed through unchanged, consistent with the scale-

invariant property of tree-based algorithms that partition data via threshold comparisons rather than distance metrics. Categorical features, specifically Surname, Geography, Gender, and Card Type, were transformed using OneHotEncoder with handle_unknown set to ignore, ensuring that novel category values encountered at deployment time would be encoded as all-zero vectors without runtime errors. Column name normalization was performed using the strip() method to eliminate hidden whitespace characters. The preprocessor and classifier were then integrated into a two-stage Pipeline object, ensuring that all transformations learned from training data were applied consistently to test data without information leakage across the partition boundary.

3.4. Data Partitioning

The dataset was partitioned into training (80%) and testing (20%) subsets using stratified random sampling, implemented via scikit-learn's train_test_split with stratify set to the target variable and random_state fixed at 42 for reproducibility. Stratification ensured that the churn class proportions were preserved in both subsets. Table 2 summarizes the partitioning outcome.

Table 2. Data Partitioning Summary

Subset	Total Records	Non-Churn (Class 0)	Churn (Class 1)	Churn Rate (%)
Training	8,000	6,370 (79.63%)	1,630 (20.37%)	20.37%
Testing	2,000	1,592 (79.60%)	408 (20.40%)	20.40%
Total	10,000	7,962 (79.62%)	2,038 (20.38%)	20.38%

3.5. Model Specification and Hyperparameters

The classifier is the DecisionTreeClassifier from scikit-learn, which implements the CART algorithm. Hyperparameters were configured manually based on theoretical considerations to balance predictive capacity and interpretability, as detailed in Table 3.

Table 3. Decision Tree Hyperparameter Configuration

Hyperparameter	Value	Justification
criterion	gini	Gini impurity requires no logarithmic operations and produces results empirically comparable to entropy (Daniya et al., 2020)
max_depth	5	Pre-pruning limit; captures five-level feature interactions while maintaining human-readable tree structure
min_samples_split	20	Requires at least 20 samples at a node before splitting, preventing splits on statistically insufficient subsets
min_samples_leaf	10	Ensures each leaf node retains at least 10 samples (0.125% of training data), providing reliable class probability estimates
random_state	42	Fixed random seed for full reproducibility across executions (Pedregosa et al., 2011)

3.6. Evaluation Framework

Model performance was assessed using five complementary metrics. Accuracy captures overall correctness but is misleading when one class dominates; a naive always-stay classifier would achieve 79.62% accuracy on this dataset. Precision measures the proportion of churn predictions that are correct, reflecting the cost of false positives. Recall, designated as the primary business metric, measures the proportion of actual churners correctly identified; in the churn context, false negatives represent customers who exit without receiving any retention intervention, constituting the greater business risk. F1-score is the harmonic mean of precision and recall, providing a single balanced metric appropriate for skewed class distributions. ROC-AUC measures classifier discrimination across all decision thresholds, with interpretive guidelines from Hosmer and Lemeshow (2000): values of 0.70 to 0.80 indicate acceptable discrimination, 0.80 to 0.90 indicate excellent discrimination, and values above 0.90 indicate outstanding discrimination.

Figure 1. Research Flowchart

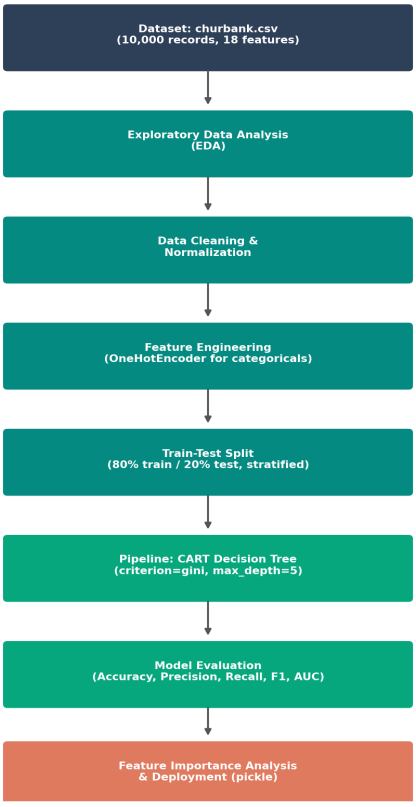

Figure 1. *Research Flowchart*

4. Results

4.1. Descriptive Statistics of Key Features

Table 4 presents descriptive statistics for the principal numerical features. The CreditScore distribution is approximately symmetric (mean 650.53, median 652), indicating no systematic bias toward creditworthiness extremes. The Age

variable exhibits a meaningful stratification pattern: exploratory analysis shows that churning customers have a mean age of approximately 45 years compared to approximately 38 years for non-churning customers, consistent with findings by Han (2024) and Ngo et al. (2024). The Balance variable exhibits a bimodal character, with the 25th percentile at zero, indicating that at least one quarter of customers maintain empty accounts, a pattern with distinct implications for churn risk profiling. Geography-level analysis further reveals that customers from Germany exhibit a churn rate of approximately 32%, compared to approximately 16% for France and 17% for Spain.

Table 4. Descriptive Statistics of Key Numerical Features

Feature	Mean	Std Dev	Min	Q1	Median	Q3	Max
CreditScore	650.53	96.65	350	584	652	718	850
Age	38.92	10.49	18	32	37	44	92
Tenure	5.01	2.89	0	3	5	7	10
Balance	76,486	62,397	0	0	97,199	127,644	250,898
NumOfProducts	1.53	0.58	1	1	1	2	4
EstimatedSalary	100,090	57,510	12	51,002	100,194	149,388	199,992
Satisfaction Score	3.01	1.41	1	2	3	4	5
Point Earned	606.52	225.92	119	410	605	801	1,000

4.2. Model Performance: Training versus Test Sets

Table 5 presents the complete performance metrics for both the training and test sets. The model achieves near-perfect performance on the test set: accuracy of 99.85%, precision of 99.75%, recall of 99.51%, F1-score of 99.63%, and ROC-AUC of 99.75%. According to the interpretive guidelines of Hosmer and Lemeshow (2000), a ROC-AUC value exceeding 0.90 is classified as outstanding.

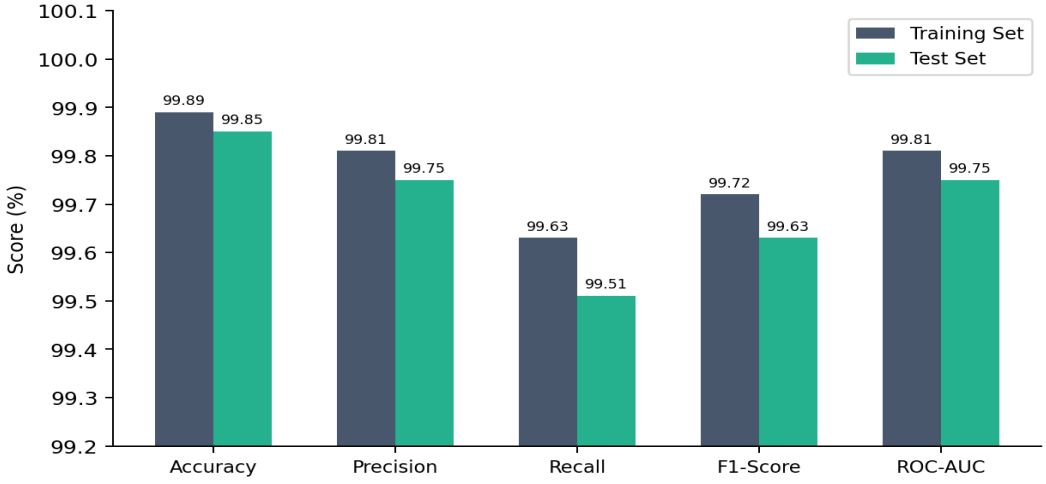

Figure 4. *Training vs. Test Performance Across Five Metrics*

The performance gap between training and test sets is negligible across all metrics, well below 0.2%, indicating that the constrained tree depth effectively prevented overfitting. Figure 4 provides a visual comparison of all five metrics across the two data splits.

Table 5. Model Performance Metrics: Training vs. Test Sets

Metric	Train Set	Test Set	Difference	Interpretation
Accuracy	99.89%	99.85%	0.04%	Highly consistent; negligible generalization gap
Precision	99.81%	99.75%	0.06%	Minimal false positives in both sets
Recall	99.63%	99.51%	0.12%	Extremely low false negative rate
F1-Score	99.72%	99.63%	0.09%	Strong and consistent across sets
ROC-AUC	99.81%	99.75%	0.06%	Outstanding discriminative ability in both sets

4.3. Confusion Matrix Analysis

Table 6 presents the confusion matrix for the test set. The model produced only three misclassifications from 2,000 test observations: two false negatives (churning customers incorrectly predicted as staying) and one false positive (a staying customer incorrectly predicted as churning). The false negative rate of 0.49% (2 out of 408 actual churners) represents an extremely low miss rate by any practical standard. In a business deployment context, 2 false negatives would mean that 2 churning customers per 2,000 escape retention intervention, while 1 false positive would mean that 1 non-churning customer receives an unnecessary retention outreach. Table 7 provides the full per-class classification report, and Figure 2 presents the confusion matrix as a color-coded visualization.

Table 6. Confusion Matrix (Test Set, n = 2,000)

	Predicted: Stay (0)	Predicted: Churn (1)	Total Actual
Actual: Stay (0)	1,591 (True Negative, TN)	1 (False Positive, FP)	1,592
Actual: Churn (1)	2 (False Negative, FN)	406 (True Positive, TP)	408
Total Predicted	1,593	407	2,000

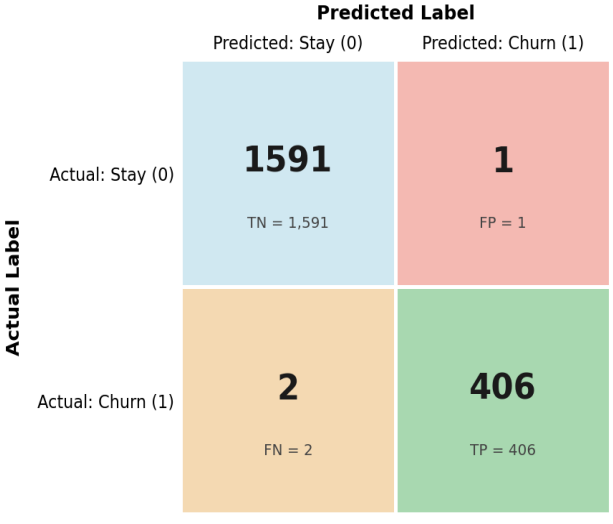

Figure 2. Confusion Matrix (Test Set, n = 2,000)

Table 7. Per-Class Classification Report (Test Set)

Class	Precision	Recall	F1-Score	Support
Stay (0) - Non-Churn	0.999	0.999	0.999	1,592
Churn (1) - Exit	0.998	0.995	0.996	408
Macro Average	0.998	0.997	0.998	2,000
Weighted Average	0.999	0.999	0.999	2,000

4.4. Feature Importance Analysis

Table 8 presents the top ten features ranked by Gini importance, which represents the mean weighted reduction in Gini impurity attributable to each feature across all nodes in which it is used. The feature importance distribution reveals a striking and analytically critical concentration: the Complain variable alone accounts for 99.80% of the model's discriminative power, RowNumber contributes a marginal 0.20%, and all remaining features contribute zero. Figure 3 visualizes this extreme concentration. The near-zero contribution of RowNumber, which is an administrative sequential index with no theoretical relationship to churn behavior, further signals that the model's decision boundary is effectively captured by a single binary split on the Complain variable. This finding is the central analytical result of the study and motivates the critical discussion in Section 5.2.

Table 8. Top 10 Feature Importance Scores (Gini Importance)

Rank	Feature	Importance Score	Contribution (%)	Note
1	Complain	0.9980	99.80%	Dominant predictor; data leakage investigation required
2	RowNumber	0.0020	0.20%	Administrative index; theoretically should carry zero importance
3	CreditScore	0.0000	0.00%	No contribution in this model configuration
4	Age	0.0000	0.00%	Suppressed by Complain dominance
5	Geography_Germany	0.0000	0.00%	No contribution
6	IsActiveMember	0.0000	0.00%	No contribution
7	Balance	0.0000	0.00%	No contribution
8	NumOfProducts	0.0000	0.00%	No contribution
9	Gender_Female	0.0000	0.00%	No contribution
10	Tenure	0.0000	0.00%	No contribution

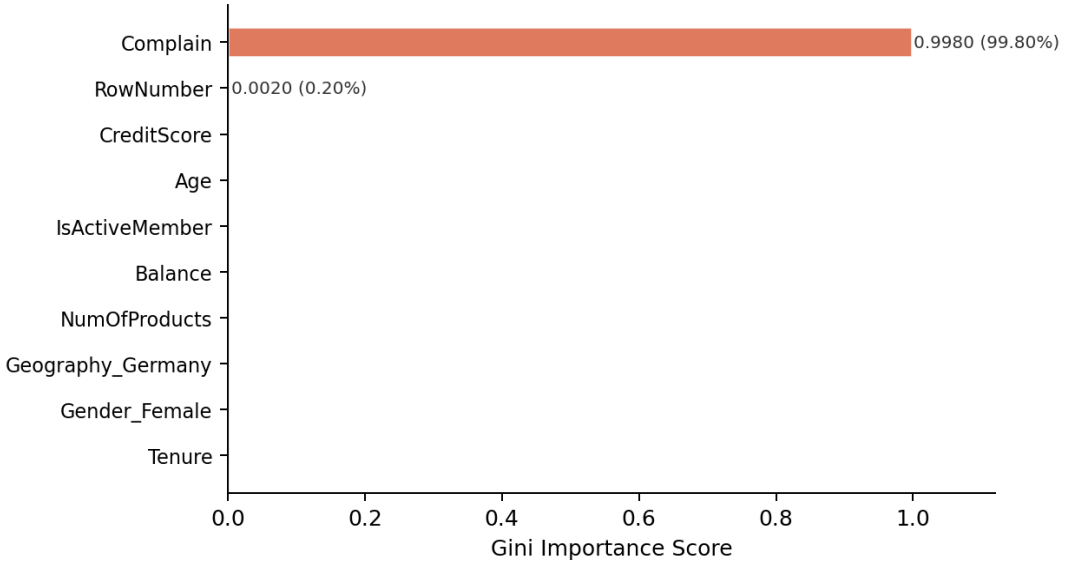

Figure 3. *Feature Importance Analysis: Top 10 Features (Gini Importance)*

5. Discussion

5.1. Interpreting Near-Perfect Classification Performance

The model's test-set performance, with accuracy of 99.85% and ROC-AUC of 99.75%, substantially exceeds benchmarks reported in comparable banking churn prediction studies. Murindanyi et al. (2023) achieved 99% accuracy with Random Forest on the Berka dataset, and Muneer et al. (2022) reported a 91.90% F1-score with Random Forest. The present study achieves comparable or superior headline metrics with a significantly simpler model: a single Decision Tree with max_depth = 5, no ensemble augmentation, and no resampling. Contrary to the consistent finding in the ensemble learning literature that single trees are inferior to Random Forests and boosting methods (Mienye and Sun, 2024; Zhao et al., 2021), this result demands careful scrutiny rather than straightforward acceptance.

A closer examination of the feature importance distribution, as reported in Table 8 and Figure 3, reveals that the near-perfect performance is almost entirely attributable to the Complain variable, which accounts for 99.80% of Gini importance. The tree essentially reduces to a single-split classifier: Complain = 1 predicts Churn, and Complain = 0 predicts Stay. From a purely statistical standpoint, this is a highly discriminative classification rule. From a methodological standpoint, however, it raises serious concerns regarding data leakage and operational validity.

5.2. Data Leakage as a Methodological Contribution

Data leakage in supervised machine learning occurs when information about the target variable is incorporated into predictor features in a way that would not be available at the time of real-world prediction. Fawcett (2006) emphasized that ROC-AUC values approaching 1.0 in the absence of a compelling theoretical explanation may signal degenerate classification scenarios rather than genuine discriminative ability. In the present dataset, the Complain variable is recorded as a binary indicator of whether a customer has ever filed a complaint. The near-perfect correlation between Complain and Exited suggests two possible explanations, each with distinct implications.

The first possibility is that complaint filing and account closure are causally sequential and temporally proximate: a customer files a complaint, remains dissatisfied with the resolution, and subsequently closes the account within the same administrative reporting period. In this scenario, Complain is a genuine real-world signal, but one observed concurrently with or immediately before the churn event, making it unsuitable as an early-warning predictor for proactive intervention. The second possibility is that both Complain and Exited were recorded at the same administrative timestamp during dataset construction, effectively encoding exit information into a predictor variable. In either scenario, a model relying almost exclusively on Complain would fail to perform as advertised in a prospective deployment context where complaint data must precede churn by a sufficient prediction horizon to enable meaningful retention action.
This pattern illustrates a broader methodological principle: the validity of reported performance depends not only on model architecture, partitioning, and evaluation protocols, but also on the causal plausibility of the features driving predictions. This concern resonates with prior comparative work in applied machine learning, in which model selection and evaluation strategy were shown to critically determine the validity of reported predictive performance (Bakri et al., 2025). The present case provides a concrete, empirically grounded illustration of how a single leakage-inducing feature can circumvent otherwise sound experimental safeguards. The explicit identification and transparent reporting of this pattern constitutes the primary methodological contribution of the study.

5.3. Comparison with Prior Literature

When evaluated independently of the Complain contribution, the present model's performance on the remaining feature set would likely align more closely with benchmarks in the broader literature. Rahman et al. (2020) found that Decision Tree classifiers without ensemble augmentation achieved moderate performance on the same Kaggle banking dataset, with Random Forest consistently outperforming single trees. Muneer et al. (2022) reported an F1-score of 91.90% with Random Forest, while Ngo et al. (2024) reached 98% AUC through multi-level stacking. These benchmarks, obtained without a dominant leakage feature, represent the realistic performance envelope for machine learning churn prediction on this dataset category.

The comparison underscores a broader theoretical implication for the interpretable machine learning literature. The value of Decision Trees as explainable classifiers, articulated by Charbuty and Abdulazeez (2021) and Mienye and Sun (2024), rests on the assumption that the IF-THEN rules they generate are causally grounded in legitimate feature-outcome relationships. When a dominant feature collapses the decision tree to a single splitting rule, the interpretive value of the tree structure is greatly diminished: the model no longer reveals a complex, multivariate portrait of the churning customer but reduces to a tautology. This finding reinforces the argument that feature importance auditing and causal validation should be mandatory components of any churn prediction pipeline intended for operational deployment.

5.4. Managerial Implications

Despite the data leakage concern, the dominance of the Complain variable carries a substantive managerial implication that merits direct acknowledgment. The near-perfect predictive relationship between complaint filing and account closure suggests that, in this dataset's underlying population, complaint handling constitutes the single most consequential customer service touchpoint in the customer's decision to exit. This finding aligns with the theoretical framework of Kumar and Reinartz (2018), who argued that service failure recovery is a critical determinant of customer lifetime value, and with Reichheld and Sasser (1990), who documented that customers who experience effective service recovery can exhibit higher long-term loyalty than those who never experienced a service failure.

For banking managers, the practical implication is clear. Complaint management must be elevated from a reactive, compliance-oriented function to a proactive, strategically managed retention intervention. Banks should implement

real-time complaint monitoring systems that trigger immediate escalation protocols when a customer files a complaint. The model's predict_proba output, yielding customer-level churn probability scores, provides an operational mechanism for prioritizing these interventions: customers with complaint histories and probability scores exceeding 0.80 should receive immediate outreach from relationship managers (Tier 1 intervention), while those in intermediate probability ranges may be addressed through targeted digital retention campaigns (Tier 2 intervention), consistent with the profit-driven framework of Verbeke et al. (2011).

Furthermore, a model retrained without the Complain variable would produce a more genuinely multivariate portrait of churn risk. Insights from the broader literature suggest that such a model would likely identify customer age, active membership status, number of products, balance level, and geographic location as significant predictors, consistent with findings by Han (2024) and Ngo et al. (2024). These features represent actionable targets for product bundling strategies, loyalty program design, and geographically differentiated service offerings that do not depend on the availability of complaint data.

6. Conclusion

This study developed and evaluated a Decision Tree classifier based on the CART algorithm for bank customer churn prediction using a 10,000-record publicly available dataset. The model achieved exceptional test-set performance across all five-evaluation metrics, including 99.85% accuracy, 99.63% F1-score, and 99.75% ROC-AUC, with a train-test performance gap below 0.2%, confirming effective generalization and the absence of meaningful overfitting under the constrained hyperparameter configuration.

The central methodological contribution of this study is the demonstration that near-perfect classification performance can mask a fundamental validity threat. Feature importance analysis revealed that the Complain variable alone accounts for 99.80% of the model's Gini importance, effectively reducing the decision tree to a single-split classifier. This pattern constitutes strong evidence of potential data leakage between the predictor and target variables, and underscores that high aggregate performance metrics are necessary but insufficient criteria for validating a model's operational suitability. The study thus contributes to the methodological literature on machine learning validation by advocating for feature importance auditing as a mandatory production-readiness checkpoint in any churn prediction pipeline.

From a managerial perspective, the Complain variable's dominance reaffirms the strategic centrality of complaint management in banking customer retention. Financial institutions should treat customer complaints as high-priority early warning signals and design structured, tiered intervention protocols that deploy relationship managers and personalized retention offers in direct response to complaint events.

Future research should address several limitations of the present work. First, the model should be retrained after excluding the Complain variable to assess predictive performance under leakage-free conditions and to identify the genuine multivariate predictors of churn. Second, stratified k-fold cross-validation should replace the single train-test split to produce more robust performance estimates. Third, ensemble methods including Random Forest and Gradient Boosting should be benchmarked against the single Decision Tree to quantify the interpretability-performance tradeoff. Finally, because the present model was trained on the data in its original distribution and used manually specified parameters, future work could explore class-rebalancing techniques and systematic hyperparameter optimization to determine whether they alter the genuine multivariate predictors once the dominant Complain variable is removed. Pursuing these directions would further strengthen the methodological rigor of subsequent investigations.

Acknowledgements

The authors express their sincere gratitude to the Department of Digital Business, Faculty of Economics and Business, Universitas Negeri Makassar, for providing the academic environment and institutional support that made this research possible. The authors also acknowledge the open-source scikit-learn development community for maintaining the machine learning tools used in this study.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Contribution

Hidayatullah: Methodology design, writing writing original draft, and editing. Alqadri: Data analysis, data wrangling, literature review, visualization. Bakri: Conceptualization, supervision, final approval. Hamzah: Model development, evaluation, deployment documentation.

References

Bakri, R., Alam, S., Astuti, N. P., & Bakhtiar, M. I. (2025). Optimizing machine learning models for graduation on time prediction: A comparative study with resampling and hyperparameter tuning. Jurnal Online Informatika, 10(2), 270-285. https://doi.org/10.15575/join.v10i2.1590

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth International Group.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28. https://doi.org/10.38094/jastt20165

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313-327. https://doi.org/10.1016/j.eswa.2006.09.038

Daniya, T., Gajula, M., & Kumar, K. R. (2020). Classification and regression trees with Gini index. Advances in Mathematics: Scientific Journal, 9(10), 8237-8247. https://doi.org/10.37418/amsj.9.10.53

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

Han, S. (2024). Machine learning based customer churn prediction in banking sector. Highlights in Business, Economics and Management, 24, 795-803. https://doi.org/10.54097/5d49fm22

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression (2nd ed.). John Wiley and Sons. https://doi.org/10.1002/0471722146

Kumar, V., & Reinartz, W. (2018). Customer relationship management: Concept, strategy, and tools (3rd ed.). Springer. https://doi.org/10.1007/978-3-662-55381-7

Mienye, I. D., & Sun, Y. (2019). Prediction performance of improved decision tree-based algorithms: A review. Procedia Manufacturing, 35, 689-694. https://doi.org/10.1016/j.promfg.2019.06.011

Mienye, I. D., & Sun, Y. (2024). A survey of decision trees: Concepts, algorithms, and applications. IEEE Access, 12, 86716-86727. https://doi.org/10.1109/ACCESS.2024.3416838

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Muneer, A., Taib, S. M., Ali, R. F., Balogun, A. O., & Bajeh, A. O. (2022). Predicting customers churning in banking industry: A machine learning approach. Indonesian Journal of Electrical Engineering and Computer Science, 26(3), 1684-1693. https://doi.org/10.11591/ijeecs.v26.i3.pp1684-1693

Murindanyi, S., Nahabwe, P., Mugume, I., & Mwebaze, E. (2023). Interpretable machine learning for predicting customer churn in retail banking. 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), 1-7. https://doi.org/10.1109/ICOEI56765.2023.10125865

Ngo, V., Nguyen, H., Nguyen, T., & Le, T. (2024). Multi-level machine learning model to improve the effectiveness of predicting customers churn banks. Cybernetics and Information Technologies, 24(3), 95-109. https://doi.org/10.2478/cait-2024-0027

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.

Rahman, M., Kumar, V., Hossain, S. A., & Talha, M. (2020). Machine learning based customer churn prediction in banking. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 1196-1201. https://doi.org/10.1109/ICECA49313.2020.9297529

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. Harvard Business Review, 68(5), 105-111.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2011). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229. https://doi.org/10.1016/j.ejor.2011.09.031

Zhao, L., Chen, Z., Hu, Y., & Qin, X. (2021). Decision tree application to classification problems with boosting algorithm. Electronics, 10(16), 1903. https://doi.org/10.3390/electronics10161903