Analysis of Factors Associated with Motor Vehicle Tax Arrears Using Ensemble Learning and Explainable Artificial Intelligence

Katharina Tyas Aprilia, Rian Farta Wijaya

Abstract

This study examines vehicle characteristics contributing to motor vehicle tax arrears classification in the context of UPTD PEPENDA Simalungun. The dataset contains 68,872 vehicles with three payment statuses. Binary analysis includes 66,887 records labelled as overdue or paid; 1,985 unpaid records are excluded pending clarification of their operational definition. Six predictors comprise brand, model, vehicle category, manufacturing year, registration plate colour, and tax amount. Data are partitioned by tax expiry year: 2023 for training, 2024 for validation, and 2025 for testing. Logistic regression provides a baseline for Random Forest and Histogram Gradient Boosting. Random Forest is selected using validation balanced accuracy of 0.5912 at a threshold of 0.60. Test ROC-AUC is 0.5452, balanced accuracy is 0.5356, overdue recall is 64.70%, and paid specificity is 42.41%. Interventional Shapley analysis of a test sample identifies manufacturing year, tax amount, and vehicle model as the largest absolute contributors. Discrimination remains limited, and feature contributions do not establish causal effects. Missing payment and extraction timestamps restrict the findings to exploratory payment-status classification rather than demonstrated prospective prediction.

***Keywords**:* Motor Vehicle Tax, Arrears, Ensemble Learning, Explainable AI, Shapley.

Katharina Tyas Aprilia
Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: katharinatyas230401@gmail.com

Rian Farta Wijaya
e-mail: rianfartawijaya@dosen.pancabudi.ac.id
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

Managing motor vehicle tax arrears requires information that helps tax administrators understand payment patterns. Administrative records provide vehicle characteristics, tax amounts, and payment statuses that can be analysed together. However, a model that produces classifications does not necessarily discriminate adequately between payment outcomes. Evaluation must consider the class distribution and examine the information underlying model predictions.

Putri and Junawan examined motor vehicle tax compliance at SAMSAT Binjai [1]. Malau, Ramadhan, and Irawan investigated the implementation of SIGNAL to support taxpayer compliance from a technology acceptance perspective [2]. These studies provide context on tax payments and public services. However, findings from another region or administrative unit do not establish the same patterns in Simalungun. Vehicle characteristics alone also cannot reveal taxpayer awareness, trust, or service experiences.

Studies involving lecturers from Universitas Pembangunan Panca Budi (UNPAB) applied C4.5 classification to property tax payment compliance and tax bill categories [3], [4]. An algorithm comparison by Siahaan and colleagues provides methodological context for Random Forest and XGBoost [5]. Differences in study objects and outcome labels limit the transferability of performance results. This paper uses those studies to frame the comparison and evaluates models using the available motor vehicle tax records.

Economic approaches to tax compliance examine taxpayer decisions in relation to incentives and enforcement [6]. The slippery slope framework emphasises trust in authorities and enforcement power [7]. These perspectives indicate that administrative indicators are not equivalent to taxpayer motivation. Because the dataset does not directly measure perceptions, this study focuses on predictive associations with recorded payment status.

The study addresses three questions. First, what payment-status distribution and data characteristics are available for classification? Second, how do Random Forest and Histogram Gradient Boosting compare with logistic regression and a majority-class baseline when evaluated across tax expiry years? Third, which features make the largest absolute contributions to the selected model according to Shapley explanations?

The empirical contribution comprises model comparison across separate expiry-year groups, evaluation that accounts for class imbalance, and interpretation of feature contributions within the limits of model performance. The study does not assume that ensemble methods always perform better. Weak results remain useful for assessing whether the available variables are sufficient before a model supports tax administration. The study also distinguishes classification within a single dataset from risk prediction before a payment deadline, which requires more complete observation timestamps.

Literature Review

1. Tax Compliance and Administrative Classification

The Allingham and Sandmo model provides an economic foundation for understanding tax decisions [6]. Applying this perspective to vehicle tax arrears requires care because administrative status may also reflect recordkeeping problems, ownership changes, or service access. The slippery slope framework relates compliance to trust and enforcement [7]. This study does not directly test these theoretical constructs because survey and behavioural measurements are unavailable.

Table 1. Relevance of five publications involving UNPAB lecturers

Reference	Study focus	Role in this paper
[1] Putri and Junawan	Vehicle tax compliance in Binjai	Motor vehicle tax payment context
[2] Malau et al.	SIGNAL and taxpayer compliance	Service and technology acceptance context

Reference	Study focus	Role in this paper
[3] Siahaan et al.	Property tax compliance classification	Classification in tax administration
[4] Siahaan et al.	Property tax bill classification	Differences in labels and classification goals
[5] Siahaan et al.	Comparison of pistachio classifiers	Methodological context; not tax performance evidence

2. Ensemble Learning and Evaluation

Random Forest aggregates trees constructed through randomisation to produce predictions [8]. Gradient boosting builds models sequentially to reduce a loss function [9]. XGBoost is a development of tree boosting systems [10], but it was not run in this experiment. The study evaluates Random Forest and Histogram Gradient Boosting as two ensemble methods, with logistic regression as a simpler comparator. The two ensemble models are evaluated separately rather than combined through stacking.

With imbalanced outcomes, the majority class can dominate accuracy. Precision-recall curves help assess class separation under these conditions [13]. Calibration also matters when model outputs are interpreted as probabilities [15]. The study reports ROC-AUC, balanced accuracy, average precision for both classes, and the Brier score to avoid relying on a single performance measure.

3. Model Explanations

SHAP decomposes model outputs into a baseline value and feature contributions [11]. Aggregating local explanations can help characterise broader model behaviour [12]. However, dependence between features can affect attribution [17]. This study uses model-agnostic interventional Shapley explanations for six original features. These contributions explain model outputs; causal conclusions require a separate research design. Interpretability considerations also support comparison with simpler models [18].

Research Methodology

The study uses an observational predictive design and secondary records from data_prapemrosesan.csv in the context of UPTD PEPENDA Simalungun. The analytical unit is one uniquely identified vehicle in the file. The dataset contains 68,872 rows and 10 columns. CSV inspection found no missing values, exact duplicate rows, or repeated vehicle registration numbers. These checks constitute an initial technical audit and do not establish the substantive accuracy of every entry.

The tax expiry date column was parsed using the YYYY-MM-DD format. All dates were readable, ranging from 2 January 2023 to 31 December 2025. Tax amounts were converted from comma-formatted text into numeric values and ranged from IDR 12,000 to IDR 27,132,000. Manufacturing years ranged from 1938 to 2024 and did not exceed the tax expiry year in any record. Extreme values were retained because there was insufficient evidence to classify them as errors.

Table 2. Payment-Status Distribution In The Source Dataset

Status	Number of vehicles	Percentage
Overdue	62,944	91.39%
Paid	3,943	5.73%
Unpaid	1,985	2.88%

Binary analysis included 62,944 vehicles labelled MENUNGGAK (overdue) as class 1 and 3,943 labelled SUDAH BAYAR (paid) as class 0, giving 66,887 vehicles. A further 1,985 vehicles labelled BELUM BAYAR (unpaid) were provisionally excluded. Without the source

definition, it is unclear whether this category denotes arrears or an obligation that has not yet passed its payment deadline. The analysis uses the recorded source labels rather than reconstructing payment status from dates.

Table 3. Data Partition By Tax Expiry Year

Partition	Year	Total	Overdue	Paid
Training	2023	18148	17859	289
Validation	2024	20574	19754	820
Testing	2025	28165	25331	2834

The file does not include extraction dates, payment dates, status-update timestamps, or historical snapshots. Consequently, partitioning by tax expiry year evaluates pattern transfer between year groups but does not demonstrate prospective prediction before a payment deadline. All statuses may originate from a single extraction snapshot. This limitation restricts interpretation to exploratory classification of recorded payment status.

1. Predictors and preprocessing

Six predictors were used: brand (MERK), vehicle model (TYPE), vehicle category (JENIS), manufacturing year (TAHUN BUAT), plate colour (TNKB), and tax amount (NOMINAL PKB). Registration numbers and detailed addresses were excluded. Tax expiry dates defined the data partition and were not predictors. Rare categories were grouped using a minimum frequency of 50 and a maximum of 100 categories per column. One-hot encoding handled infrequent and previously unseen categories. Numeric variables used median imputation and standardisation. All transformations were fitted exclusively on training data to limit information leakage [19].

2. Model development and selection

Logistic regression used C = 1 and a maximum of 1,000 iterations. Random Forest used 180 trees, a maximum depth of 14, at least 10 observations per leaf, and balanced_subsample class weights. Histogram Gradient Boosting used 160 iterations, at most 15 leaves, a learning rate of 0.06, L2 regularisation of 5, and balanced class weights. The random seed was 42. Models were implemented using scikit-learn 1.8.0 [20]. These were fixed initial configurations rather than results of comprehensive hyperparameter optimisation.

Models were trained exclusively on the 2023 records. Classification thresholds from 0.05 to 0.95, in increments of 0.01, were evaluated using the 2024 validation set. Model and threshold selection followed the highest validation balanced accuracy. Separation of development and evaluation follows model-validation principles [16], with an explicit expiry-year partition. Models were not refitted by combining training and validation data after selection. SMOTE [14] was not applied; class weights addressed the imbalance.

3. Performance assessment

The 2025 test set was not used for model or threshold selection. Metrics included accuracy, balanced accuracy, overdue precision and recall, paid specificity, overdue F1, macro-F1, ROC-AUC, average precision (AP) for both classes, and the Brier score. AP refers to average precision rather than an interpolated curve area. The baseline classified every vehicle as overdue and assigned a constant probability based on training prevalence. For the selected model, 95% intervals were estimated using 250 vehicle-level bootstrap samples of the test set, with the fitted model and threshold held fixed.

4. Explainability analysis

Shapley explanations evaluated all 64 coalitions of the six original features [11]. The empirical background comprised 24 training vehicles. The explanation sample included 160 test vehicles: 80 overdue and 80 paid. Mean absolute contributions were weighted to reflect test-set class prevalence. The calculation used an interventional approach with a limited empirical background, rather than TreeSHAP or explanations of the entire dataset. Attributions were expressed on the uncalibrated model probability-output scale. Additivity was checked

against model predictions. The small sample and feature dependence limit ranking stability [17].

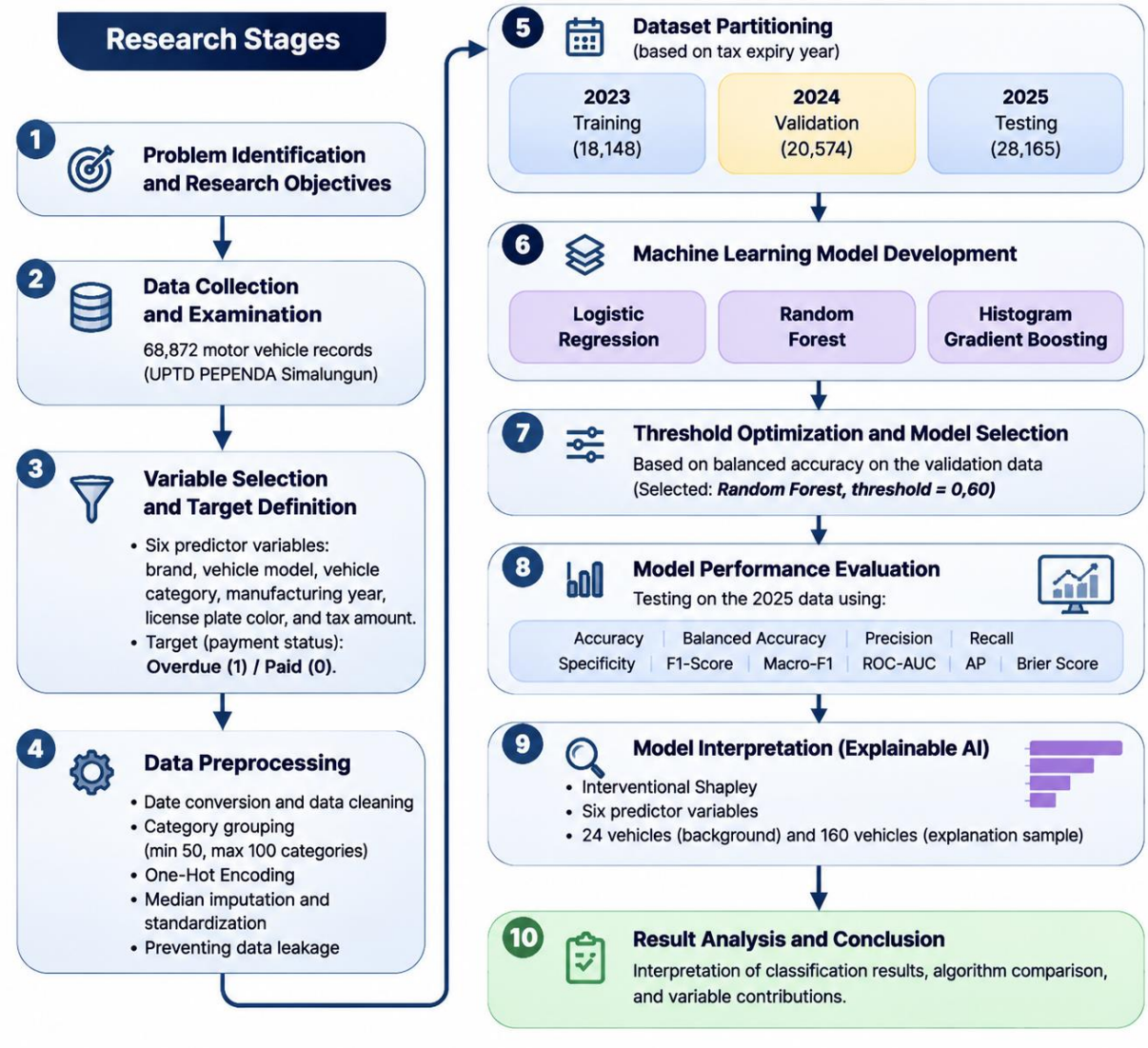


Figure 1. Research Stages

Results

1. Model performance

Random Forest achieved the highest validation balanced accuracy, 0.5912, at a threshold of 0.60. Logistic regression and Histogram Gradient Boosting achieved 0.5679 and 0.5732, respectively. Table 4 presents test performance after model and threshold selection were locked.

Table 4. Performance On The 2025 Tax Expiry-Year Test Set

Model	Threshold	Accuracy	Bal. acc.	ROC-AUC	Paid AP
Logistic Regression	0.40	0.6628	0.5322	0.5352	0.1161
Random Forest	0.60	0.6246	0.5356	0.5452	0.1191
Histogram Gradient Boosting	0.59	0.7255	0.5323	0.5388	0.1194
Majority baseline	0.50	0.8994	0.5000	0.5000	0.1006

Random Forest achieved ROC-AUC of 0.5452 with a bootstrap 95% interval of 0.5333–0.5554, and balanced accuracy of 0.5356 with an interval of 0.5263–0.5436. Paid-class AP was 0.1191, only slightly above the baseline of 0.1006. Overdue-class AP of 0.9111 should likewise be compared with the overdue baseline of 0.8994. The Random Forest Brier score was 0.1689, worse than the constant-probability baseline of 0.0977. Model probabilities were not calibrated.

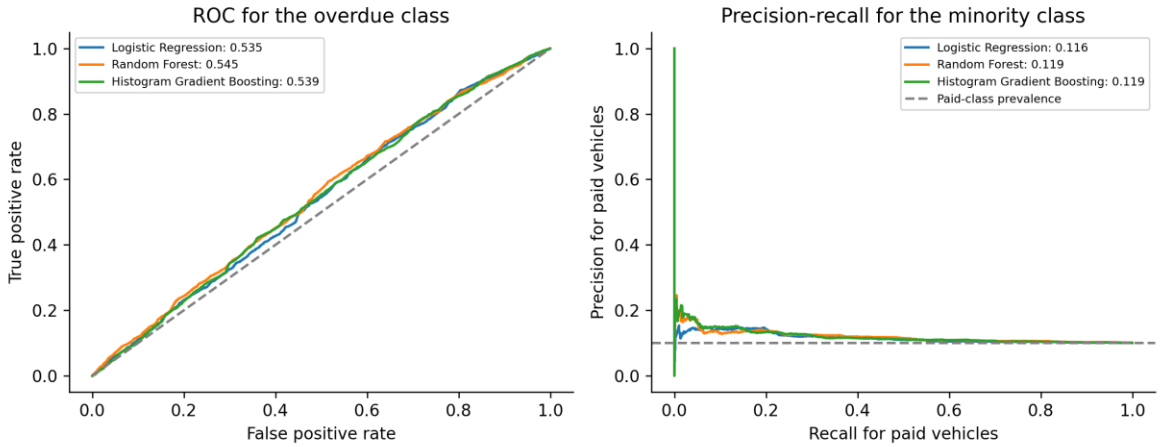


Figure 2. ROC For Overdue Vehicles and Precision-Recall For Paid Vehicles

Table 5. Random Forest Confusion Matrix

Actual	Predicted paid	Predicted overdue
Paid	1202	1632
Overdue	8941	16390

Overdue recall was 64.70%, paid specificity was 42.41%, and overdue precision was 90.94%. Overdue F1 was 0.7561 and macro-F1 was 0.4707. The model classified 8,941 overdue vehicles as paid and 1,632 paid vehicles as overdue. These findings indicate limited separation between payment statuses.

2. Feature contributions

Table 6 reports weighted mean absolute Shapley contributions in the explanation sample. Manufacturing year ranked first, followed by tax amount and vehicle model. These features accounted for the largest mean absolute attributions among the six predictors in the analysed sample. The ranking is neither a measure of causal strength nor a statistical significance test.

Table 6. Absolute Shapley contributions in the Random Forest model

Rank	Feature	Mean absolute contribution
1	Manufacturing year	0.062003
2	Tax amount	0.047425
3	Vehicle model	0.036207
4	Brand	0.013733
5	Vehicle category	0.006451
6	Plate colour	0.001208

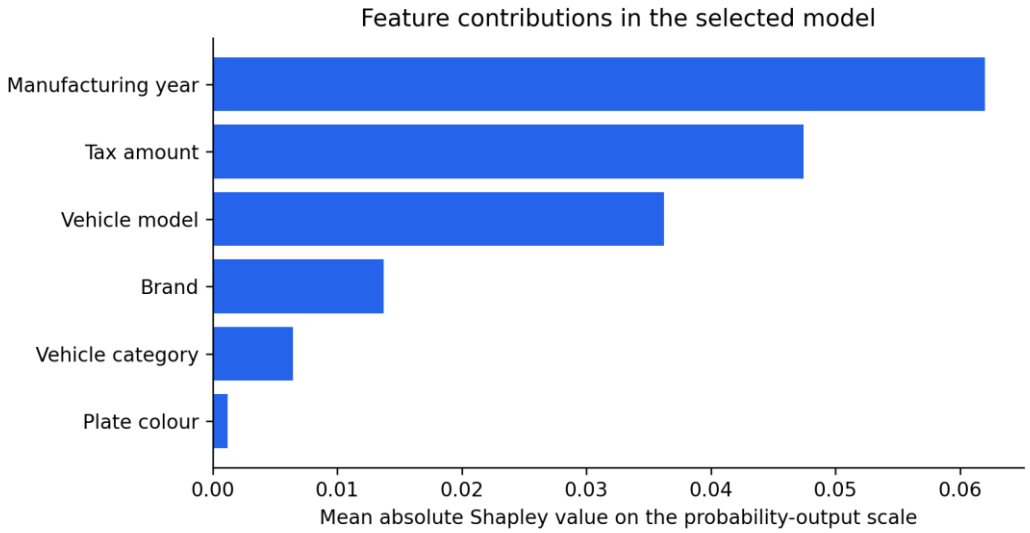


Figure 2. Weighted mean absolute Shapley values on the model probability-output scale.

The manufacturing-year value of 0.062003 represents the average attribution magnitude in the sample, rather than a fixed 6.20% increase in risk. Mean absolute values do not establish whether older and newer vehicles have contributions in the same direction. Directional interpretation requires examination of feature values and local contributions. Similarly, a small contribution for plate colour does not establish that this feature is irrelevant throughout the population.

The maximum difference between the baseline plus all feature contributions and the model output was 5.55×10^{-17}. This check supports the additive consistency of the calculation. However, additivity does not establish predictive quality, the realism of feature combinations, or ranking stability. The background of 24 vehicles and explanation sample of 160 vehicles make these interpretations exploratory.

Discussion

The experimental results indicate that the available vehicle characteristics provide limited discriminatory information for classifying motor vehicle tax payment status. Random Forest achieved the highest validation balanced accuracy of 0.5912 at a threshold of 0.60, but its test performance remained limited, with a ROC-AUC of 0.5452 and balanced accuracy of 0.5356. The high accuracy of the majority-class baseline (89.94%) was primarily influenced by the dominance of overdue vehicles in the dataset. This finding confirms the importance of evaluating classification models using multiple metrics, particularly balanced accuracy and precision-recall analysis, when dealing with imbalanced datasets [13]. Furthermore, the Brier score of 0.1689 indicates that the model's probability estimates require further improvement [15].

The Shapley analysis identified manufacturing year (0.062003), tax amount (0.047425), and vehicle model (0.036207) as the three largest contributors to the model output. These findings indicate that vehicle-related administrative characteristics provide different levels of information for classification. However, they do not establish causal relationships or directly explain taxpayer behaviour. As discussed by Allingham and Sandmo [6] and Kirchler et al. [7], tax compliance also involves economic considerations, institutional trust, and enforcement, which are not measured in the available dataset. Therefore, future research should incorporate more complete historical payment records, verified payment-status definitions, and additional administrative variables to improve model reliability. Overall, the findings demonstrate the potential of ensemble learning and explainable AI for exploring motor vehicle tax payment patterns, although the current model is not yet suitable for reliable operational prediction.

Conclusion

Analysis of 66,887 vehicles showed that Random Forest and Histogram Gradient Boosting did not achieve strong payment-status classification across tax expiry-year groups. Random Forest, selected using validation data, obtained test ROC-AUC of 0.5452 and balanced accuracy of 0.5356 in 2025. Shapley explanations identified manufacturing year, tax amount, and vehicle model as the largest absolute contributors in the sample without establishing causality. The findings highlight the need for payment histories and verified status definitions. With the currently available information, conclusions remain restricted to exploratory classification and do not demonstrate prediction before a payment deadline.

References

[1] A. N. Putri and Junawan, "Analisis Kepatuhan Wajib Pajak Kendaraan Bermotor pada SAMSAT Binjai," Jurnal Riset Rumpun Ilmu Pendidikan, vol. 5, no. 1, pp. 75–87, 2026, doi: 10.55606/jurripen.v5i1.7459.

[2] T. T. Malau, P. R. Ramadhan, and Irawan, "Analisis Implementasi Penggunaan Aplikasi SIGNAL dalam Mendukung Kepatuhan Wajib Pajak Kendaraan Bermotor Provinsi Sumatera Utara," Jurnal Manajemen dan Akuntansi Medan, vol. 7, no. 1, 2025, doi: 10.47709/jumansi.v7i1.8631.

[3] A. P. U. Siahaan, S. Sutiono, S. Pranoto, Sarifudin, and R. S. Mentari, "Analysis of Property Tax Payment Compliance Classification in Tebing Tinggi City Using the C4.5 Decision Tree Algorithm," Journal of Information Technology, Computer Science and Electrical Engineering, vol. 1, no. 2, pp. 134–138, 2024, doi: 10.61306/jitcse.v1i2.59.

[4] A. P. U. Siahaan, A. A. Jabar, S. Pranoto, S. Sutiono, and D. Ramatika, "Analysis of Property Tax Bill Classification Using the C4.5 Algorithm," Journal of Information Technology, Computer Science and Electrical Engineering, vol. 1, no. 3, pp. 181–185, 2024, doi: 10.61306/jitcse.v1i3.100.

[5] A. P. U. Siahaan, M. Iqbal, D. Dika, and M. Syahputri, "Classification Of Pistachio Varieties Using Machine Learning Algorithms," Jurnal Minfo Polgan, vol. 14, no. 1, pp. 1452–1457, 2025, doi: 10.33395/jmp.v14i1.15088.

[6] M. G. Allingham and A. Sandmo, "Income tax evasion: A theoretical analysis," Journal of Public Economics, vol. 1, nos. 3–4, pp. 323–338, 1972, doi: 10.1016/0047-2727(72)90010-2.

[7] E. Kirchler, E. Hoelzl, and I. Wahl, "Enforced versus voluntary tax compliance: The slippery slope framework," Journal of Economic Psychology, vol. 29, no. 2, pp. 210–225, 2008, doi: 10.1016/j.joep.2007.05.004.

[8] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[10] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[11] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 30, 2017. Available: https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[12] S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” Nature Machine Intelligence, vol. 2, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[13] T. Saito and M. Rehmsmeier, “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets,” PLOS ONE, vol. 10, no. 3, Art. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[15] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in Proc. 22nd International Conference on Machine Learning, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

[16] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in Proc. 14th International Joint Conference on Artificial Intelligence, 1995, pp. 1137–1143. Available: https://ijcai.org/proceedings/1995/147.

[17] K. Aas, M. Jullum, and A. Løland, “Explaining individual predictions when features are dependent: More accurate approximations to Shapley values,” Artificial Intelligence, vol. 298, Art. 103502, 2021, doi: 10.1016/j.artint.2021.103502.

[18] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, vol. 1, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[19] S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” Patterns, vol. 4, no. 9, Art. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[20] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. Available: https://jmlr.org/papers/v12/pedregosa11a.html.