Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Implementation of Random Forest Algorithm for Classifying Land and Building Tax Arrears and Risk Factor Analysis Dashboard

Risky Firmansyah Manik[1]*, A M H Pardede[2], Anton Sihombing[3]

[1,2,3]*Program of Study Information System, STMIK Kaputama*
riskifirman608@gmail.com[1*], akimmhp@live.com[2], antonkaputama@gmail.com[3]

Abstract

This study aims to develop a predictive model to identify the potential for land and building tax arrears and analyze the dominant risk factors contributing to non-compliance. The research utilizes the Random Forest classification algorithm applied to historical tax data from the Regional Financial and Revenue Management Agency of Binjai City. The approach involves data preprocessing, feature engineering including target encoding for geographical areas, and model training with hyperparameter tuning to optimize classification performance. Furthermore, a web-based interactive dashboard is developed using the Flask framework to visualize the predictions and risk factors. The results demonstrate that the Random Forest model achieves a robust and consistent accuracy of approximately 85% in classifying compliant and non-compliant taxpayers. Feature importance analysis reveals that land area is the most dominant risk factor influencing tax arrears, significantly outweighing other variables. In conclusion, the integration of the Random Forest algorithm with an interactive dashboard provides a highly accurate, efficient, and scalable solution for local governments to transition from reactive tax collection to proactive, data-driven risk management.

Keywords: *Random forest; Feature importance; Dashboard; Land and building tax; Tax arrears*

1. Introduction

The implementation of regional autonomy in Indonesia grants local governments the authority to manage their financial resources independently to stimulate regional development. In this context, Locally-Generated Revenue (PAD) plays a central role as the primary financial driver. A vital pillar within the PAD structure is the revenue generated from the Rural and Urban Land and Building Tax (PBB-P2). Although PBB-P2 offers substantial revenue potential, optimizing its collection frequently encounters a chronic challenge: high rates of tax arrears. This issue extends beyond administrative concerns, directly impacting regional fiscal stability. The accumulation of unpaid tax obligations hinders the realization of revenue targets, subsequently disrupting the financing of infrastructure development and public service provision [1][2].

This prevalent issue is reflected in the case of the Regional Financial and Revenue Management Agency (BPKPD) of Binjai City. Limitations in predicting taxpayers who are likely to default have resulted in collection strategies that are predominantly reactive, wherein billing occurs only after the due date has passed. The identification process for potential arrears currently relies on manual or conventional assessment methods, which prove highly inefficient when dealing with massive volumes of taxpayer data. Manual processing of extensive tax data complicates the agency's ability to accurately identify and map taxpayer compliance profiles [3]. Consequently, a paradigm shift from a reactive strategy to a technology-driven, proactive approach is essential.

The utilization of data mining and machine learning offers a comprehensive solution for managing large volumes of historical taxpayer data. Classification techniques within data mining have proven beneficial for extracting and recognizing previously unknown patterns of taxpayer compliance [4]. By deploying predictive models, the tax agency can allocate collection resources more effectively toward taxpayers identified as having the highest risk of default. For this classification task, the Random Forest algorithm is highly suitable. Random Forest is renowned for its high accuracy, proficiency in handling datasets with numerous features, and robustness against overfitting. Previous research in the domain of tax prediction has identified Random Forest as a highly effective machine learning model for predictive tasks[5]. Furthermore, a significant advantage of this algorithm is its ability to extract feature importance, which is highly relevant for identifying the primary risk factors contributing to tax defaults.

However, a predictive model with high accuracy is insufficient to produce a tangible operational impact if the results cannot be easily interpreted and acted upon by decision-makers. The agency requires not only the identification of high-risk taxpayers but also an in-depth analysis of the underlying factors triggering these arrears. Therefore, this study extends beyond mere model development by designing an

interactive analysis dashboard. This dashboard functions as a decision-support medium that presents data visualizations, prediction outcomes, and the dominant risk factors influencing potential arrears. While previous studies have demonstrated the efficacy of data mining for tax compliance classification, research integrating a Random Forest model with a focused, interactive risk analysis dashboard specifically tailored for local tax agencies in Indonesia remains highly limited.

To bridge this gap, this study aims to develop an integrated predictive system. The specific objectives of this research are: (1) to produce a Random Forest classification model capable of accurately predicting the potential for taxpayer arrears; (2) to build an analysis dashboard that facilitates the agency in mapping and visualizing arrears risks proactively; and (3) to identify the dominant variables or attributes that act as the primary risk factors for tax arrears based on the model's feature importance results.

2. Theoretical Foundation

2.1. Machine learning and classification

Machine learning is a branch of artificial intelligence (AI) that enables systems to autonomously learn from historical data to identify patterns and predict outcomes without explicit programming. Within supervised learning, classification maps input variables (features) to discrete target categories. Binary classification is particularly effective for risk assessment tasks, categorizing subjects into exactly two mutually exclusive groups, such as compliant or in default [6][4] .

2.2. Random forest algorithm

Random Forest is an ensemble learning method that constructs multiple decision trees during training to enhance accuracy and prevent overfitting. It utilizes bootstrap sampling (bagging) to train each tree on different data subsets, ensuring model stability[7]. During tree construction, nodes are split using a random subset of features based on the lowest Gini Impurity, which measures misclassification probability. The algorithm also calculates feature importance to identify dominant risk factors. Final predictions are determined through a majority voting mechanism from all constructed trees [8].

2.3. Land and building tax (PBB)

In Indonesia, the Rural and Urban Land and Building Tax (PBB-P2) is an objective property tax. The tax liability is determined by physical attributes such as land area, building area, and the Sales Value of the Tax Object (NJOP) rather than the taxpayer's financial capability. This obligation is legally formalized and communicated through the Notification of Tax Due (SPPT), which outlines the payment deadline [9].

2.4. Tax arrears (Tunggakan Pajak)

Tax arrears occur when a taxpayer fails to remit the required payment by the due date established in the SPPT, which directly hinders regional revenue targets. For predictive modeling, payment behaviors are classified into two conditions: compliant (*Lancar*) and in arrears (*Menunggak*). Analyzing these historical payment conditions allows tax authorities to transition from reactive monitoring to proactive, data-driven collection strategies [1].

3. Research method

The system development in this study strictly employs the Waterfall model as its Software Development Life Cycle (SDLC). This linear, sequential approach was specifically chosen for the land and building tax (PBB) classification system because the project's functional requirements and the structure of the historical tax data are well-defined and stable from the outset. A structured progression ensures that the complex data preprocessing and machine learning phases are fully completed and validated before moving into the web dashboard integration, minimizing developmental bottlenecks.

To support the computational experiments, this research utilizes secondary data comprising the Notification of Tax Due (SPPT) for the Rural and Urban Land and Building Tax (PBB) for the fiscal year 2025. This dataset was officially obtained from the Regional Financial and Revenue Management Agency (BPKPD) of Binjai City.

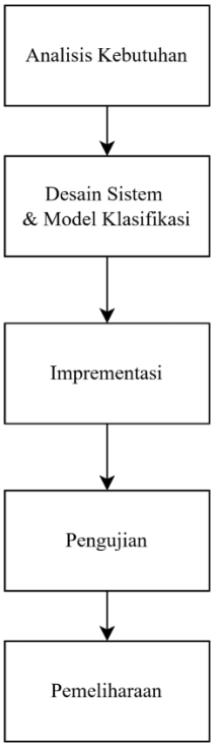

Fig. 1: Waterfall Model System Development Flow

3.1. Requirements analysis

The initial phase involves a comprehensive process of gathering system requirements. This includes the identification and extraction of crucial PBB data variables from the historical dataset, such as the principal tax amount, Sales Value of the Tax Object (NJOP), land area, and building area. Furthermore, this phase defines the specific functional needs for the risk factor analysis dashboard, ensuring it is capable of accurately visualizing taxpayer default probabilities and dominant risk variables for end-users.

3.2. System design and classification model

Based on the gathered requirements, the system design phase formulates the logical architecture of the predictive application. This phase focuses heavily on designing the Random Forest classification model architecture, including the mechanisms for building multiple decision trees and determining optimal splits. Additionally, it outlines the data preprocessing scenarios, mapping out how missing values will be imputed and how categorical geographical variables will undergo target encoding. This phase also includes the structural design of the dashboard's user interface, ensuring that the visual presentation of the analytical results will be intuitive.

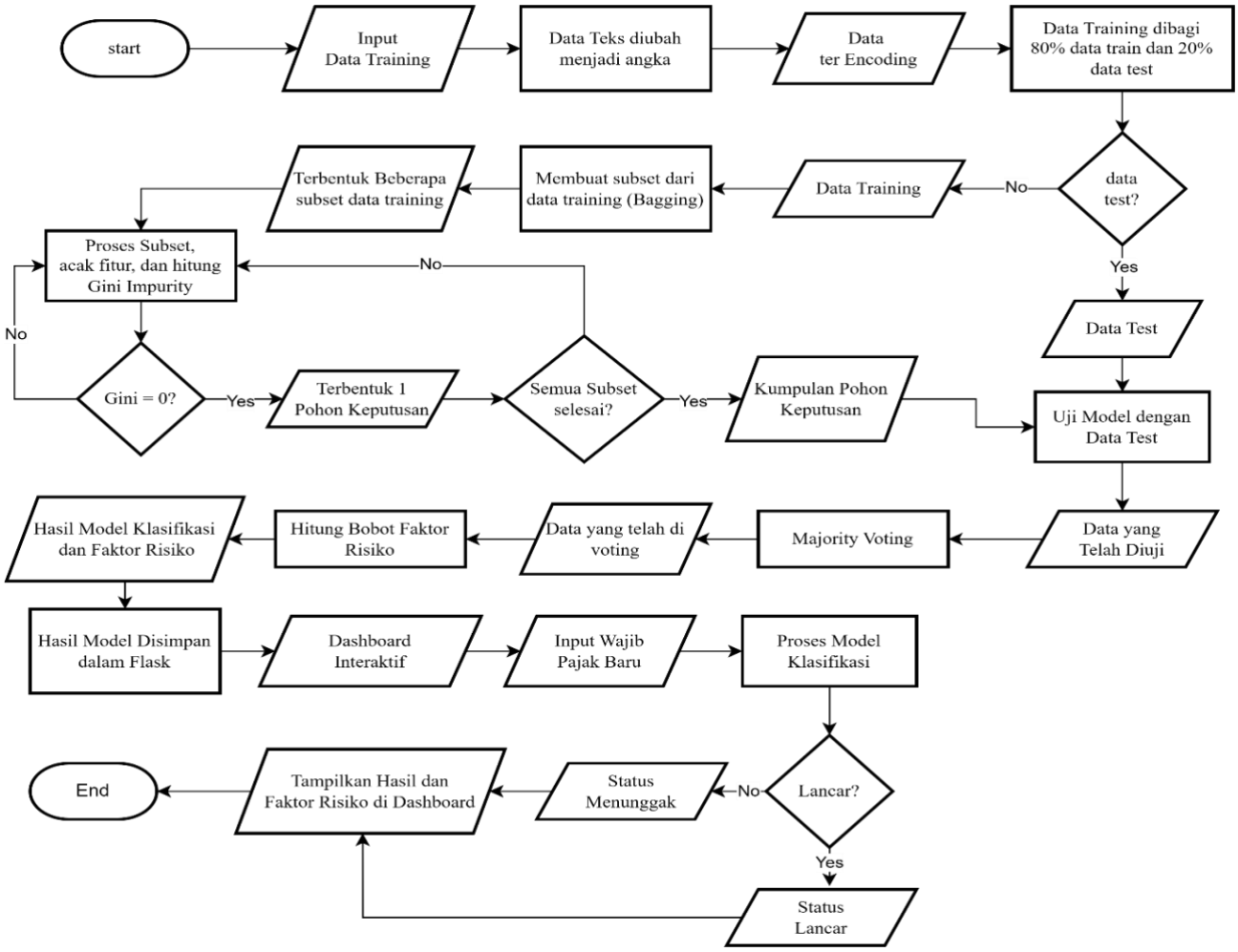

Fig. 2: System flowchart of the Random Forest classification process

3.3. Implementation

During the implementation phase, the theoretical designs are translated into executable code. The Random Forest algorithm is programmed using the Python programming language, leveraging its robust data science libraries to construct and train the classification model on the preprocessed training dataset. Once the model is optimally trained, its predictive artifacts are successfully integrated into the Flask web framework. This integration connects the complex machine learning backend with the web-based frontend, establishing a functional application capable of processing new taxpayer inputs dynamically.

3.4. Testing

To guarantee the reliability and accuracy of the developed system, two primary testing methods are conducted. First, the predictive performance of the Random Forest model is evaluated using a Confusion Matrix. This statistical tool measures the model's accuracy, precision, and recall by comparing the algorithm's predictions against the actual payment statuses in the testing dataset. Second, Black Box Testing is employed for the functional validation of the dashboard application. This ensures that all interface elements, data input mechanisms, and visual outputs operate flawlessly according to the specified requirements without internal logical errors.

3.5. Maintenance

The final phase involves the ongoing maintenance of the deployed system. This includes the continuous monitoring of the Random Forest model's predictive performance as new tax data is introduced over time. Additionally, this phase accounts for the technical upkeep of the application, encompassing the identification and resolution of any potential software bugs or interface issues that may arise during the dashboard's operational use at the tax agency.

4. Results and Discussion

4.1. Model Implementation and Integration

The technical implementation of the predictive system was executed using the Python programming language, leveraging its robust ecosystem of data science libraries. To ensure the model's reliability, stability, and generalizability, the final analysis utilized a comprehensive dataset comprising 35,687 Notification of Tax Due (SPPT) records. Following the preprocessing, training, and optimization

of the Random Forest algorithm, the trained model and its corresponding data transformation artifacts were successfully integrated into an interactive web application utilizing the Flask micro-framework. This integration bridges the complex machine learning backend with an intuitive frontend, enabling the system to process new taxpayer inputs dynamically and provide real-time risk predictions regarding potential tax arrears. This implementation directly fulfills the research objective of delivering a functional predictive tool for regional tax management.

4.2. Performance Evaluation and Testing

To rigorously validate the model's predictive capabilities, multiple testing scenarios were conducted by varying the data volume, the number of decision trees, and the maximum tree depth. The evaluation results demonstrated that the Random Forest algorithm maintained a highly stable and consistent accuracy ranging between 84% and 85% across all tested configurations. Specifically, a peak accuracy of 85.12% was achieved in Scenario 1, which utilized a moderate configuration and recorded a highly efficient execution time of merely 1.87 seconds. The testing phase revealed a distinct trade-off between model complexity and execution time. Increasing the number of trees and maximum depth in subsequent scenarios did not yield a linear improvement in accuracy; instead, it resulted in a marginal decrease in accuracy while significantly prolonging the computational time to over 45 seconds. Consequently, the optimal configuration ensures high precision alongside rapid system responsiveness.

Table 1: Performance results across different testing scenarios

Trial	Data Size	Number Of Trees	Depth	Accuracy	Runtime
1	10.706	100	8	85,12%	1,87 sec
2	23.197	300	12	85,01%	16,02 sec
3	35.687	500	15	84,76%	45,39 sec

4.3. Risk Factor Analysis (Feature Importance)

A primary objective of this study was to identify the dominant variables influencing taxpayer compliance. Through the extraction of the Random Forest model's feature importance metrics based on the final integrated analysis, it was established that the Building-to-Land Area Ratio ('Proporsi Luas Bangunan thd Tanah') is the most dominant risk factor contributing to PBB tax arrears. This specific derived attribute significantly outweighed other parameters, indicating that the density of building utilization on a plot is a critical predictor of payment behavior. Following this ratio, the 'Building Area' ('Luas Bangunan') and the 'Tax-to-NJOP Ratio' ('Rasio Tagihan Pajak thd NJOP') were identified as the subsequent most significant influencing factors. Recognizing these primary determinants—particularly the prominence of building-related metrics over simple land area—provides crucial analytical insights. This empowers the tax agency to comprehend the underlying drivers of non-compliance more accurately and strategically target collection efforts based on these objective property utilization metrics.

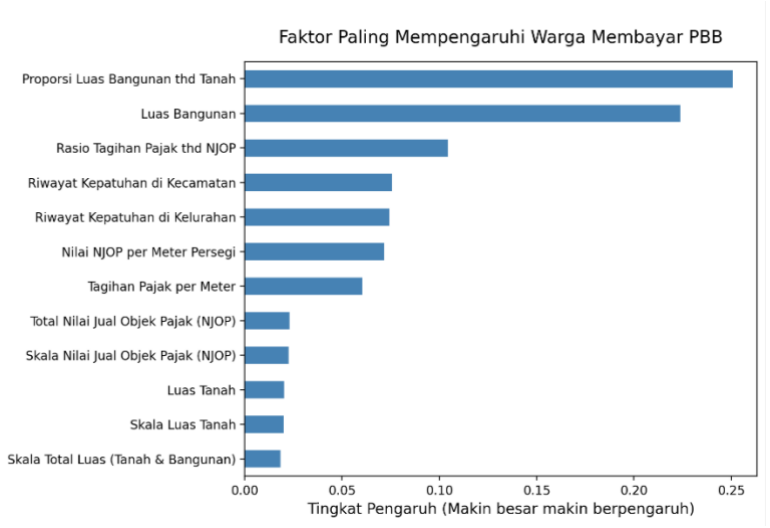

Fig. 3: Feature Importance chart highlighting primary risk factors

4.4. Dashboard Functional Results

The culmination of the system development is the final operational interface, titled 'Dashboard Analisis Risiko PBB' (PBB Risk Analysis Dashboard). This dashboard effectively visualizes the predictive classifications, the overall proportion of compliance risks, and the specific hierarchical risk factors for individual taxpayers. By translating complex machine learning outputs into an accessible and interactive visual format, the dashboard substantially supports the Regional Financial and Revenue Management Agency (BPKPD) in shifting from reactive collection methods toward proactive, data-driven decision-making.

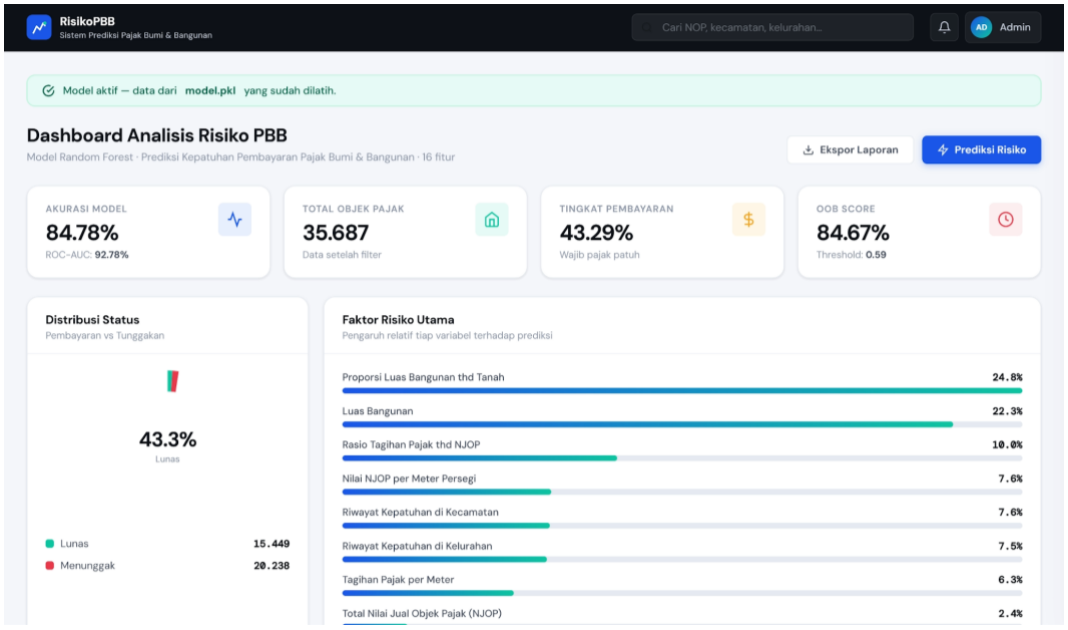

Fig. 4: The integrated web dashboard interface

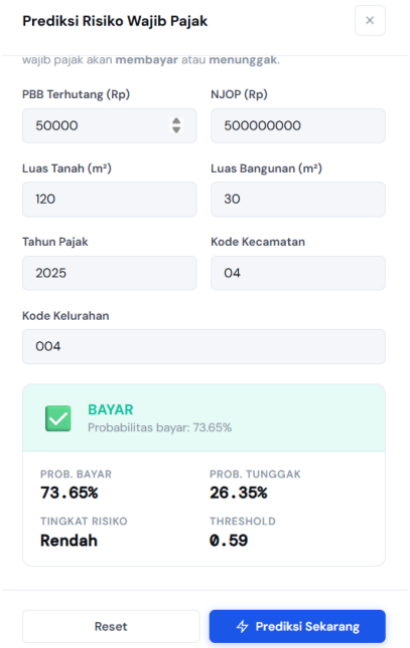

Fig. 5: Taxpayer Prediction Form

A comparative analysis between the traditional manual evaluation method and the newly implemented Random Forest system highlights significant operational advancements. Manual analysis of historical tax records is highly subjective, prone to human error, and requires weeks of intensive administrative effort to process large datasets. Conversely, the automated machine learning system executes the objective analysis of over 35,000 records in mere seconds. Furthermore, the Random Forest approach offers exceptional scalability, seamlessly accommodating the continuous annual growth of taxpayer data without necessitating a proportional increase in human labor, thereby providing a sustainable solution for regional tax oversight.

5. Conclusion

This study successfully developed and implemented a predictive model utilizing the Random Forest algorithm to classify the potential for land and building tax (PBB) arrears. The application of robust feature engineering techniques, specifically target encoding, proved highly effective in processing a massive dataset of over 35,000 historical records without encountering computational bottlenecks. The computational experiments demonstrated that the model achieves exceptional and stable predictive performance, peaking at an accuracy of 85.12% with a highly efficient execution time. Furthermore, the extraction of feature importance successfully identified land area as the primary and most dominant determinant triggering taxpayer defaults. The significant novelty of this research lies in the practical integration of the optimized machine learning model into a Flask-based web dashboard. This interactive interface provides a substantial operational upgrade from conventional manual analysis, offering local government agencies a scalable, fast, and objective tool to visualize risk proportions, understand underlying factors, and prioritize proactive tax collection strategies.

References

[1] C. Widiawati, M. Yani Balaka, and L. Tondi, "FAKTOR-FAKTOR PENYEBAB TUNGGAKAN PAJAK BUMI DAN BANGUNAN (PBB) DI KABUPATEN MUNA," *Jurnal Ekonomi (JE)*, no. 2, pp. 35–44, 2024, [Online]. Available: http://jurnal-ekonomi.uho.ac.id

[2] J. Pardede and M. Ekklesia, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Prediksi Jumlah Target dan Realisasi Wajib Pajak Atas PBB-P2 Menggunakan Multi Regression, Regression Lasso, dan PCR," 2024.

[3] Z. Aini, Y. Yahya, and L. M. Samsu, "Implementasi Algoritma Naïve Bayes untuk Memprediksi Tingkat Kepatuhan Wajib Pajak di Desa Dames Damai," *Jurnal PRINTER: Jurnal Pengembangan Rekayasa Informatika dan Komputer*, vol. 1, no. 2, pp. 64–77, Dec. 2023, doi: 10.29408/jprinter.v1i2.22005.

[4] A. Rahmat Shaumi, M. Faridz Ali, and M. Tsaqif Al Mutawakkil Simbolon, "Penerapan Data Mining Menggunakan Metode Teknik Classification Untuk Melihat Penerapan Data Mining Menggunakan Metode Teknik Classification Untuk," *JUKI : Jurnal Komputer dan Informatika*, vol. 4, 2022, [Online]. Available: www.pajak.go.id

[5] Y. H. Lee and E. Kim, "Deep learning-based delinquent taxpayer prediction: A scientific administrative approach," *KSII Transactions on Internet and Information Systems*, vol. 18, no. 1, pp. 30–45, Jan. 2024, doi: 10.3837/tiis.2024.01.003.

[6] N. L. P. C. Savitri, R. A. Rahman, R. Venyutzky, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Terhadap Sekolah Daring pada Twitter Menggunakan Supervised Machine Learning," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, Apr. 2021, doi: 10.28932/jutisi.v7i1.3216.

[7] Dachi and Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *JURNAL RISET RUMPUN MATEMATIKA DAN ILMU PENGETAHUAN ALAM*, vol. 2, no. 2, pp. 65–71, Jul. 2023, doi: 10.55606/jurrimipa.v2i2.1336.

[8] A. Fabian Azmi, A. Voutama, S. Karawang Jl HSRonggo Waluyo, and T. Timur, "PREDIKSI CHURN NASABAH BANK MENGGUNAKAN KLASIFIKASI RANDOM FOREST DAN DECISION TREE DENGAN EVALUASI CONFUSION MATRIX," vol. 13, no. 1, 2024.

[9] M. A. Putri, N. Rahaningsih, F. M. Basysyar, and O. Nurdiawan, "Penerapan Data Mining Menggunakan Metode Clustering Untuk Mengetahui Kelompok Kepatuhan Wajib Pajak Bumi dan Bangunan," *Jurnal Accounting Information System (AIMS*, vol. 5, no. 2, pp. 145–156, 2022, doi: 10.32627.