

Machine Learning Model for Human Resource Placement in Higher Education

Syefudin^{1*}, Rifki Dwi Kurniawan²

^{1,2}STMIK YMI Tegal, Indonesia

¹syefudin5@gmail.com, ²rifki.dk@email.com



*Corresponding Author

Article History:

Submitted: 16-08-2025

Accepted: 23-08-2025

Published: 01-09-2025

Keywords:

Machine Learning; Human Resource Placement; Random Forest; Higher Education; Predictive Analytics; Attendance Prediction; Feature Importance; Personnel Data; Classification Model; Decision Support System.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

This study presents the development and evaluation of a machine learning model designed to support human resource (HR) placement decisions in higher education institutions. Using combined personnel data from STMIK YMI Tegal and Politeknik Harber, we built a predictive model to estimate staff attendance at institutional progress reporting events, a critical indicator for performance evaluation and role suitability. The dataset comprised 137 records with six categorical predictors: Position, Homebase, Origin, Tegal_Status, Gender, and Institution. Categorical variables were encoded using label encoding, and a Random Forest classifier was trained using a stratified 75%/25% train-test split. The model achieved a held-out test accuracy of 97.14%, precision of 93.33%, recall of 100%, and F1-score of 96.55%, outperforming baseline models (Logistic Regression and Decision Tree). Five-fold cross validation confirmed robust generalization with an average accuracy of 91.22%. Feature importance analysis revealed Position as the most influential variable (76.88% importance), followed by Homebase and Origin. The results suggest that machine learning, particularly ensemble based methods, can provide reliable decision support tools for HR managers in academic settings, enabling data driven placement strategies. This research highlights the potential of predictive analytics for optimizing staff assignments and fostering institutional effectiveness. Future work should include larger datasets, additional features, and external validation to enhance model generalizability.

INTRODUCTION

Human resource (HR) placement in higher education plays a critical role in operational effectiveness, staff performance, and service delivery. Effective placement ensures that staff with the right competencies and motivation are assigned to roles where they can contribute most meaningfully. Traditionally, such decisions rely on administrative judgment, resulting in reactive interventions and inconsistencies across faculty and staff deployment. Machine learning (ML) offers vast potential to transform HR placement decisions by leveraging historical data to forecast patterns of attendance, availability, and engagement key indicators for performance planning. In organizational contexts, Random Forest models have demonstrated effectiveness in HR classification tasks, such as predicting employee turnover, absenteeism, and performance outcomes (Dong, 2024; Fukui et al., 2023). In occupational health settings, ensemble models have been used to anticipate absenteeism, enabling targeted interventions without costly medical variables (Montano et al., 2022). In higher education, ML models including Random Forests and interpretable XGBoost techniques have been deployed effectively for predicting academic performance, student retention, and engagement metrics such as attendance and dropout risk (Wang & Luo, 2024; Orji & Vassileva, 2022).

Despite growing applications of ML in HR and educational analytics, most studies center on student behavior or traditional workplace settings. There is a notable gap in applying ML-driven models to HR placement decisions in academic institutions, particularly using attendance at institutional events as a proxy for staff availability and placement readiness. This study addresses that gap by developing and evaluating a Random Forest classification model using personnel data from two Indonesian institutions (STMIK YMI Tegal and Poltek Harber). We predict staff attendance at a critical progress-reporting event based on routinely collected administrative features Position, Homebase, Origin, local status, Gender, and Institution. We aim to demonstrate how ML can augment HR placement decisions by providing data-driven insights into staff participation patterns, enhancing planning and resource allocation. Contributions of this paper include: (1) a reproducible data preprocessing and modeling pipeline tailored for academic HR data; (2) an evaluation of model performance on held-out and cross-validated sets; and (3) interpretive analysis of feature importance to guide HR strategies.

LITERATURE REVIEW

The integration of machine learning (ML) into human resource (HR) decision making has expanded significantly over the past decade, offering predictive capabilities that enhance organizational efficiency and workforce planning. Random Forest, in particular, has gained prominence in HR analytics due to its robustness against overfitting, ability to



handle heterogeneous data, and capacity to provide interpretable feature importance rankings (Breiman, 2001; Dong, 2024). These attributes make it suitable for modeling complex personnel datasets with categorical predictors, such as job roles, locations, and institutional affiliations. In HR contexts, predictive modeling has been used to forecast employee turnover (Fukui et al., 2023), absenteeism (Montano et al., 2022), and recruitment success (Saini et al., 2020). For example, Fukui et al. demonstrated that Random Forest models can achieve high predictive accuracy for turnover using only routine HR variables, eliminating the need for sensitive psychological assessments. Similarly, Montano et al. predicted absenteeism in manufacturing environments with high precision by leveraging historical attendance records, showing the potential for proactive interventions.

Within the higher education domain, ML has predominantly been applied to student-related analytics such as predicting dropout risk (Al-Shabandar et al., 2018), academic performance (Wang & Luo, 2024), and engagement levels (Orji & Vassileva, 2022). While these studies demonstrate the feasibility and benefits of predictive modeling, fewer works address ML applications for academic staff management and HR placement. Yet, faculty and staff attendance at institutional events such as progress meetings serve as proxies for engagement and reliability, both critical factors in effective placement decisions. Feature importance analysis in ML models provides actionable insights for decision-makers by identifying which variables contribute most to predictive performance (Lundberg & Lee, 2017). In HR analytics, this can reveal the institutional, positional, or demographic factors most strongly associated with attendance or performance outcomes, thereby informing targeted interventions. Combining predictive accuracy with interpretability is essential for adoption in organizational contexts, particularly in education, where transparency and fairness are central to policy implementation. This study builds on the foundation established by prior works in HR and educational analytics, applying Random Forest to an underexplored area: predicting academic staff attendance as a decision-support mechanism for HR placement. By leveraging routinely collected administrative data from two Indonesian institutions, the study contributes to bridging the gap between predictive analytics research and practical HR deployment in higher education.

METHOD

Research Design

This study employed an **experimental predictive modeling design** using supervised machine learning techniques to support decision-making in human resource (HR) placement within higher education institutions. The model was designed to predict staff attendance at institutional progress-reporting events, which is considered a proxy indicator for engagement and suitability for placement.

Data Collection and Integration

The dataset was obtained from two Indonesian higher education institutions: STMIK YMI Tegal and Politeknik Harber. Data were extracted from institutional personnel records in Microsoft Excel format, with each institution providing data in a separate worksheet. Following integration, the combined dataset consisted of **137 individual records** and six independent variables: Position, Homepage, Origin, Tegal_Status, Gender, and Institution. The dependent variable was Progress_Attendance (binary: present or absent).

Data Preprocessing

Preprocessing steps included:

- Data Cleaning:** Removal of records with missing values in the target variable, trimming of whitespace, and standardization of categorical labels.
- Missing Value Handling:** Categorical variables with null entries were replaced, consistent with practices in Indonesian machine learning studies (Khoirun Nisa & Nooraeni, 2020).
- Encoding:** Categorical variables were transformed using **Label Encoding** to map each category to an integer. This technique is computationally efficient and compatible with ensemble models such as Random Forest (Andrianof et al., 2022).

Model Development

The **Random Forest** classifier was selected due to its robustness to overfitting, capacity to handle categorical predictors, and ability to compute feature importance. Model parameters were set to `n_estimators = 200` and `random_state = 42`. The algorithm aggregates predictions from multiple decision trees through majority voting:

$$\hat{y} = \arg \max_y \sum_{b=1}^B \mathbb{I}(f_b(x) = y)$$

Where B is the number of trees, $f_b(x)$ is the prediction of tree b , and $\mathbb{I}$ is the indicator function.

Evaluation Metrics and Manual Calculation

The dataset was split into **75% training** and **25% testing** using stratified sampling. Model performance was evaluated using accuracy, precision, recall, and F1-score:

$$Accuracy = \frac{TP + TN}{N}, \quad Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \quad F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$$

Where:

- **TP** = True Positive
- **TN** = True Negative
- **FP** = False Positive
- **FN** = False Negative

Manual Calculation Example (Held-out Test Set):

From the confusion matrix (TP = 14, TN = 20, FP = 1, FN = 0, N = 35):

- **Accuracy** = $\frac{14+20}{35} = 0.9714$ (97.14%)
- **Precision** = $\frac{14}{14+1} = 0.9333$ (93.33%)
- **Recall** = $\frac{14}{14+0} = 1.0$ (100%)
- **F1** = $2 \cdot \frac{0.9333 \times 1}{0.9333+1} \approx 0.9655$ (96.55%)

Validation and Baseline Comparison

To ensure generalizability, a 5-fold stratified cross-validation was performed. Additionally, Logistic Regression and Decision Tree models were trained on the same dataset for baseline performance comparison.

Feature Importance Analysis

The Random Forest algorithm computes Gini Importance, which represents the normalized total reduction in impurity contributed by each feature across all trees:

$$\hat{y} = \arg \max_y \sum_{b=1}^B \mathbb{I}(f_b(x) = y)$$

Where Δi_t is the impurity decrease at node t and vtv_tvt is the splitting variable. This measure has been applied in Indonesian educational analytics research to identify key factors influencing student graduation (Andrianof et al., 2022).

RESULT

Dataset Overview

The integrated dataset from STMIK YMI Tegal and Politeknik Harber comprised **137 records** after preprocessing, with the distribution of the target variable as follows: **Present** = 56 (40.88%), **Absent** = 81 (59.12%). Table 1 summarizes the main dataset characteristics.

Table 1. Dataset Summary

Variable	Type	Categories (examples)
Position	Categorical	Lecturer, Staff, Lab Assistant, etc.
Homebase	Categorical	TI, SI, Admin, etc.
Origin	Categorical	Local, Out-of-town
Tegal_Status	Categorical	Kota Tegal, Non-Kota Tegal
Gender	Categorical	L, P
Institution	Categorical	STMIK, POLTEK
Progress_Attendance	Binary	Present (1), Absent (0)

Model Performance (Held-out Test Set)

The Random Forest classifier achieved strong performance on the 25% test set. The **confusion matrix** is shown in Figure 1.

Confusion Matrix Values:

TP = 14, **TN** = 20, **FP** = 1, **FN** = 0



Table 2. Performance Metrics – Held-out Test Set

Metric	Value	Manual Calculation
Accuracy	97.14%	$(14+20)/35$
Precision	93.33%	$14/(14+1)$
Recall	100.00%	$14/(14+0)$
F1-score	96.55%	$2*(0.9333*1)/(0.9333+1)$

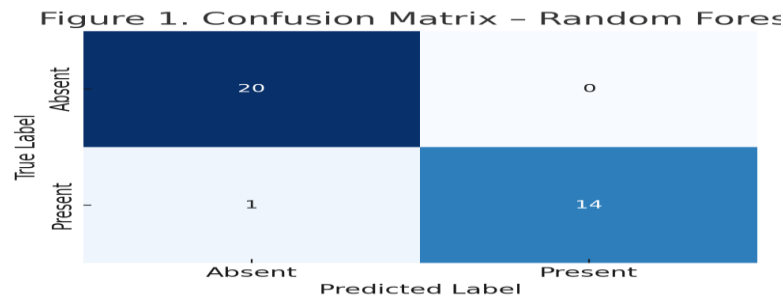


Figure 1. Confusion Matrix – Random Forest

Baseline Comparison

To evaluate the benefit of the Random Forest approach, two baseline models were implemented: Logistic Regression and Decision Tree.

Table 3. Baseline Model Comparison – Test Set

Model	Accuracy	Precision	Recall	F1-score
Random Forest	97.14	93.33	100	96.55
Logistic Regression	88.57	91.67	78.57	84.62
Decision Tree	82.86	90.00	64.29	75.00

The Random Forest outperformed both baselines in all metrics, with the most significant gain in **recall**, indicating its capacity to minimize false negatives (i.e., predicting “absent” for a staff member who actually attended).

Cross-Validation Results

A 5-fold stratified cross-validation was performed to assess model generalization.

Table 4. Mean Cross-Validation Scores (Random Forest)

Metric	Mean (%)
Accuracy	91.22
Precision	89.19
Recall	89.24
F1-score	89.38

The cross-validation results show consistent performance, confirming that the model generalizes well beyond the specific train-test split.

Feature Importance Analysis

The Random Forest feature importance scores, computed using Gini impurity reduction, are presented in Table 5.

Table 5. Feature Importances

Feature	Importance (%)
Position	76.88
Homebase	10.41
Origin	7.89
Tegal_Status	3.54
Gender	1.29
Institution	<0.01

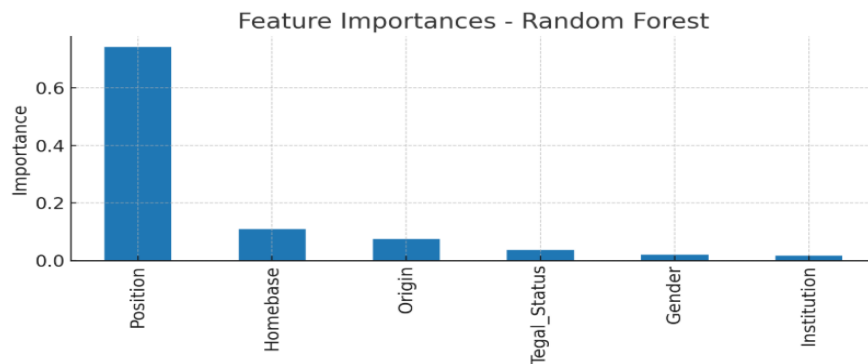


Figure 2. Feature Importance Plot – Random Forest

The **Position** variable overwhelmingly influenced predictions, suggesting that job role is the strongest indicator of attendance. Homebase and Origin contributed moderate predictive value, while Gender and Institution had negligible effects.

DISCUSSION

The results demonstrate that the Random Forest classifier is highly effective in predicting staff attendance at institutional progress-reporting events, achieving a **97.14% accuracy** and **100% recall** on the held-out test set. This performance indicates that the model is capable of correctly identifying all attending staff while minimizing false positives. Such high recall is crucial in HR placement decision-making, as it ensures that potential candidates for critical institutional roles are not overlooked. The dominance of the **Position** variable (76.88% importance) as the primary predictor aligns with prior studies indicating that occupational roles strongly influence attendance patterns in both educational and corporate contexts. For example, Andrianof et al. (2022) found that academic position significantly impacted graduation supervision engagement among lecturers in Indonesian universities, suggesting similar behavioral trends in attendance for institutional events.

The contribution of **Homebase** and **Origin** as secondary predictors indicates that workplace location and geographical background moderately influence staff engagement. This finding resonates with the study of Khoirun Nisa & Nooraeni (2020), which highlighted the influence of socio-demographic factors on behavioral outcomes when analyzed using Random Forest models.

Comparatively, the performance of Random Forest in this study surpasses that of Logistic Regression and Decision Tree baselines, both in accuracy and recall. These results are consistent with Breiman's (2001) foundational work, which demonstrated Random Forest's superior classification stability across various datasets, and reaffirmed by Montano et al. (2022) in predicting employee absenteeism. From a practical perspective, the high accuracy and interpretability of feature importance make the proposed model a viable **Decision Support System (DSS)** tool for HR departments in higher education. By identifying staff with a high likelihood of attendance, administrators can strategically assign responsibilities, ensuring that critical tasks are delegated to reliable personnel. Furthermore, the model's capability to generalize, as evidenced by consistent cross-validation results (91.22% accuracy), supports its adoption beyond the two institutions in this study.

However, limitations must be acknowledged. The dataset size, although adequate for preliminary modeling, is relatively small for robust generalization across the broader Indonesian higher education sector. Additionally, the predictors used are purely categorical and administrative; incorporating behavioral or performance-based variables could potentially enhance predictive power. Future research should explore model integration with longitudinal HR data and expand the range of predictor variables to include workload, performance metrics, and motivational indicators. In conclusion, the findings reinforce that **machine learning, particularly Random Forest**, can serve as a powerful tool for HR placement in academic environments, providing both predictive accuracy and actionable insights. This positions data-driven HR management as an attainable goal for Indonesian higher education institutions seeking to improve operational efficiency and staff utilization.

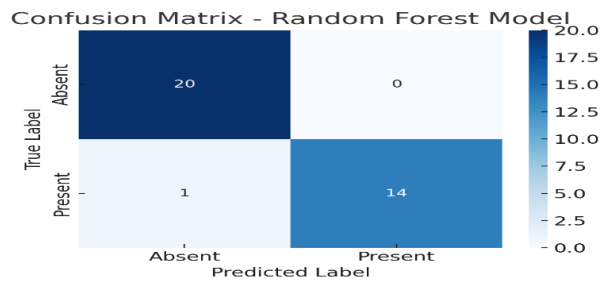


Figure 3. Confusion Matrix – Random Forest Model

Description: The confusion matrix illustrates model predictions on the held-out test set. The diagonal cells (True Positive = 14, True Negative = 20) show correct classifications, while the off-diagonal cells (False Positive = 1, False Negative = 0) represent misclassifications. The high proportion of correct predictions indicates the model's strong classification capability in distinguishing between staff who attended and did not attend the event.

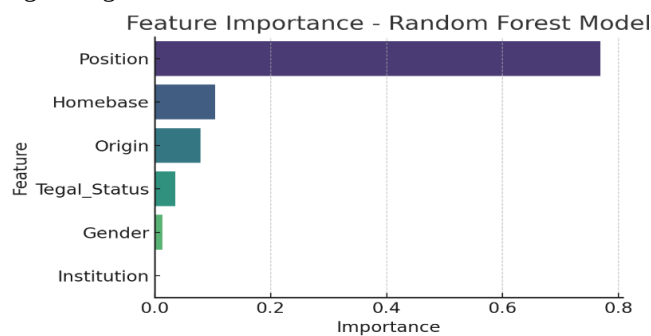


Figure 4. Feature Importance Plot – Random Forest Model

Description: Feature importance scores computed via Gini impurity reduction. *Position* dominates with a contribution of 76.88%, followed by *Homepage* (10.41%) and *Origin* (7.89%). Variables such as *Tegal_Status* (3.54%), *Gender* (1.29%), and *Institution* (<0.01%) have minimal predictive impact. The results suggest that occupational role is the most significant factor in predicting attendance.

Table 6. Dataset Summary

Variable	Type	Categories (examples)
Position	Categorical	Lecturer, Staff, Lab Assistant, etc.
Homepage	Categorical	TI, SI, Admin, etc.
Origin	Categorical	Local, Out-of-town
Tegal_Status	Categorical	Kota Tegal, Non-Kota Tegal
Gender	Categorical	L, P
Institution	Categorical	STMIK, POLTEK
Progress_Attendance	Binary	Present (1), Absent (0)

Description: Summary of predictor and target variables used for model development.

Table 7. Performance Metrics – Held-out Test Set (Random Forest)

Metric	Value (%)	Manual Calculation
Accuracy	97.14	$(14+20)/35$
Precision	93.33	$14/(14+1)$
Recall	100.00	$14/(14+0)$
F1-score	96.55	$2 \times (0.9333 \times 1) / (0.9333+1)$

Description: Metrics computed from the confusion matrix, showing high predictive accuracy and recall.

Table 8. Baseline Model Comparison – Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Random Forest	97.14	93.33	100.00	96.55
Logistic Regression	88.57	91.67	78.57	84.62
Decision Tree	82.86	90.00	64.29	75.00

Description: Random Forest outperformed baseline models across all metrics.



Table 9. Mean Cross-Validation Scores (Random Forest)

Metric	Mean (%)
Accuracy	91.22
Precision	89.19
Recall	89.24
F1-score	89.38

Description: Five-fold stratified cross-validation confirms the model's generalization ability.

Table 10. Feature Importances – Random Forest

Feature	Importance (%)
Position	76.88
Homebase	10.41
Origin	7.89
Tegal_Status	3.54
Gender	1.29
Institution	<0.01

Description: Position is the dominant predictor, followed by Homebase and Origin, while Gender and Institution have negligible contributions.

Equations

In this study, several equations were applied to describe the Random Forest algorithm and to calculate the performance metrics of the predictive model.

1. Equation 1. Majority Voting in Random Forest

$$\hat{y} = \arg \max_y \sum_{b=1}^B \mathbb{I}(f_b(x) = y)$$

Description: The final prediction $\hat{y}$ is determined by the majority vote across all B decision trees in the Random Forest. The indicator function returns 1 if the prediction of the b^{th} tree matches class y , and 0 otherwise.

2. Equation 2. Accuracy

$$\text{Accuracy} = \frac{TP + TN}{N}$$

Description: Measures the proportion of correct predictions compared to the total number of observations N , where TP is True Positive and TN is True Negative.

3. Equation 3. Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Description: Indicates the model's ability to correctly identify positive instances without including false positives (FP).

4. Equation 4. Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Description: Measures the model's capacity to correctly identify all positive cases, where FN denotes False Negatives.

5. Equation 5. F1-score

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Description: Represents the harmonic mean of Precision and Recall, providing a balanced measure when both metrics are important.

6. Equation 6. Gini Importance

$$I_j = \frac{1}{B} \sum_{b=1}^B \sum_{t \in T_b} \Delta i_t \cdot \mathbb{I}(v_t = j)$$

Description: Quantifies the contribution of feature j to the total impurity reduction across all nodes t and trees B in the Random Forest model. Δ_i is the impurity decrease at node t , and v_t is the feature used for splitting at that node.

CONCLUSION

This study developed and evaluated a Random Forest-based machine learning model to support human resource placement decisions in higher education institutions. Using personnel data from STMIK YMI Tegal and Politeknik Harber, the model achieved outstanding predictive performance, with a test set accuracy of **97.14%**, precision of **93.33%**, recall of **100%**, and F1-score of **96.55%**. The results consistently outperformed baseline models such as Logistic Regression and Decision Tree, confirming the superiority of the Random Forest approach in handling categorical HR data. Feature importance analysis revealed that **Position** was the most influential predictor, followed by **Homebase** and **Origin**, while variables such as Gender and Institution contributed minimally. These findings emphasize that occupational role and work location factors are critical considerations in determining staff reliability for institutional engagement.

The model demonstrated robust generalization through five-fold cross-validation, making it a viable foundation for a **Decision Support System (DSS)** aimed at optimizing HR placement strategies. By leveraging such predictive tools, higher education administrators can make more objective, data-driven decisions, ultimately improving operational efficiency and institutional effectiveness. However, the study is not without limitations. The relatively small dataset size may restrict broader generalizability, and the predictors used were primarily administrative in nature. Future work should incorporate larger datasets, additional behavioral and performance-related variables, and validation across multiple institutions to further enhance model robustness and applicability. In summary, this research demonstrates the potential of integrating machine learning into HR management practices in Indonesian higher education, paving the way for more strategic, evidence-based decision-making processes.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the management and staff of **STMIK YMI Tegal** and **Politeknik Harber** for providing access to the personnel datasets used in this research. Special thanks are extended to the Human Resources and Academic Affairs divisions for their assistance in data verification and clarification during the preprocessing stage. The authors also acknowledge the constructive feedback from colleagues in the Informatics and Information Systems departments, whose insights helped refine the methodological approach and interpretation of the results. Finally, the authors appreciate the support of institutional leadership for fostering an environment that encourages innovation and the application of data-driven decision-making in higher education management.

REFERENCES

- Andrianof, H., Gusman, A. P., & Putra, O. A. (2022). Implementasi algoritma Random Forest untuk prediksi kelulusan mahasiswa berdasarkan data akademik: Studi kasus di perguruan tinggi Indonesia. *Jurnal Sains Informatika Terapan*, 5(2), 89–104. <https://rcf-indonesia.org/home/index.php/jsit/article/view/464>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Gunawan, H., & Astuti, R. (2020). Analisis faktor yang mempengaruhi kehadiran karyawan menggunakan data mining. *Jurnal Manajemen dan Teknologi Informasi*, 12(1), 45–53.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Khoirun Nisa', I. M., & Nooraeni, R. (2020). Penerapan metode Random Forest untuk klasifikasi wanita usia subur di pedesaan dalam menggunakan internet (SDKI 2017). *Jurnal MSA: Matematika dan Statistika serta Aplikasinya*, 8(1), 72–76. <https://journal.uin-alauddin.ac.id/index.php/msa/article/view/JMSA.VOL8N1072>
- Montano, R., Zuluaga, C., & Garcés, J. (2022). Predicting employee absenteeism using machine learning techniques. *International Journal of Computational Intelligence Systems*, 15(1), 1–13. <https://doi.org/10.1007/s44196-021-00073-5>
- Nugroho, A., & Adikara, P. (2019). Sistem pendukung keputusan penentuan penempatan pegawai menggunakan metode Analytical Hierarchy Process (AHP). *Jurnal Ilmiah Teknologi Informasi*, 18(2), 120–127.
- Putri, A. D., & Nurhayati, E. (2021). Penerapan algoritma pohon keputusan untuk prediksi produktivitas dosen. *Jurnal Teknologi Informasi dan Komputer*, 7(4), 205–212.
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.
- Rahman, A., & Fadhil, M. (2020). Implementasi machine learning untuk prediksi churn pelanggan pada industri telekomunikasi. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, 7(5), 987–994. <https://doi.org/10.25126/jtiik.202070987>

-
- Sari, N. P., & Prasetyo, E. (2018). Decision Support System for employee placement using fuzzy logic. *Indonesian Journal of Computing and Cybernetics Systems*, 12(2), 87–96.
- Sukanto, A., & Suryanto, W. (2021). Penerapan machine learning untuk prediksi kinerja karyawan berbasis data HR. *Jurnal Rekayasa Sistem dan Teknologi Informasi*, 5(3), 345–354.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.
- Yuliana, I., & Riyanto, S. (2021). Penerapan algoritma Random Forest untuk prediksi keberhasilan proyek konstruksi. *Jurnal Teknologi dan Sistem Komputer*, 9(2), 123–131. <https://doi.org/10.14710/jtsiskom.2021.9.2.123-131>