MACHINE LEARNING FOR EMPLOYMENT POSITION MAPPING

Sena Aditia Apriadi[1*]; Hilman Ferdinandus Pardede[2]

Computer Science[1]
Universitas Muhammadiyah Kuningan, Kuningan, Indonesia[1]
https://umkuningan.ac.id/

Research Center for ArtificialIntelligence and Cybersecurity[2]
National Research and Innovation Agency (BRIN), Jakarta Pusat, Indonesia[2]
https://www.brin.go.id/
shevana7@gmail.com[1]*, hilman@nusamandiri.ac.id[2]
(*) Corresponding Author

Abstract*—Employee performance directly impacts organizational efficiency, yet traditional HR analytics often lack predictive precision. This study bridges HR theory and machine learning by evaluating tree-based algorithms for employee data analysis. Using a dataset of 15,227 employee records, we tested the Bagged Decision Tree algorithm, focusing on variables such as talent, career values, and aspirations. The Bagged Decision Tree achieved 98.65% accuracy, with talent and career values as key predictors. Excluding aspiration values reduced accuracy slightly to 98.57%, while excluding career values lowered it significantly to 92.13%. These findings highlight the robustness of the Bagged Decision Tree in HR analytics and emphasize the importance of variable selection, particularly career values and talent, in predicting performance outcomes. Future work should further explore real-world implementation challenges.*

Keywords: *bagged decision tree, employees, machine learning, organizational performance, predictive analytics.*

Abstrak— *Kinerja pegawai berperan penting dalam menentukan efisiensi organisasi, namun analitik SDM tradisional sering kali belum mampu memberikan prediksi yang akurat. Studi ini menjembatani teori SDM dengan machine learning melalui evaluasi algoritma berbasis pohon keputusan dalam menganalisis data pegawai. Menggunakan data dari 15.227 pegawai, algoritma Bagged Decision Tree diuji dengan menyoroti variabel talenta, nilai karier, dan aspirasi. Algoritma ini mencapai akurasi 98,65%, dengan talenta dan nilai karier sebagai prediktor utama. Penghapusan variabel aspirasi menurunkan akurasi menjadi 98,57%, sedangkan tanpa variabel nilai karier akurasi turun hingga 92,13%. Temuan ini menegaskan keandalan Bagged Decision Tree dalam analitik SDM dan menunjukkan pentingnya pemilihan variabel, khususnya nilai karier dan talenta, dalam memprediksi kinerja pegawai. Penelitian selanjutnya perlu menggali implementasi di dunia nyata beserta tantangan yang mungkin dihadapi.*

Kata Kunci*: analitik prediktif, bagged decision tree, kinerja organisasi, machine learning, pegawai.*

INTRODUCTION

Grounded in the human capital theory, which emphasizes employees as pivotal organizational assets, this study examines the role of employee attributes in achieving operational excellence and evaluates the effectiveness of predictive algorithms for talent management. This research employs a tree-based machine learning approach, guided by theoretical principles from decision-making and organizational behavior, to analyze a dataset of 15,227 employees and identify critical factors influencing performance outcomes. Employees provide their physical and intellectual capabilities to the organization in return for compensation determined through mutual agreements and organizational policies. Higher employee competence enhances the organization’s ability to achieve sustainable competitive advantage (Kang & Lee, 2021). Employees are also physical and spiritual human labor (mental and mind) that are needed, therefore employees become the main capital in a company or organization to achieve the goals or vision and mission of the company (Rahimi et al., 2025). In the big dictionary of Indonesian, the meaning of position is a job or task in government or organization that is related to rank and position (KBBI, n.d.).

This study is built upon Human Capital Theory, which views employees as valuable assets whose knowledge, skills, and abilities directly impact organizational performance and competitiveness (Kodithuwakku & Priyanath, 2022). Investment in human capital is key to enhancing performance, and it is measurable and improvable through advanced analytical techniques, such as machine learning. In addition, the study incorporates Organizational Behavior Theory, which emphasizes the importance of personal values, motivation, and role fit in shaping individual performance outcomes within an organization (Islamy, et al., 2021). The research specifically focuses on career values and aspirations as critical factors influencing job performance.

From a technological perspective, the research draws on Machine Learning, a subset of Artificial Intelligence, which allows systems to learn from data and generate predictive insights (Ghahremani nahr, Nozari, & Sadeghi, 2021). Machine learning operates through a two-phase process: training and testing, which are central to the development of predictive models. Machine learning operates through a two-phase process: training and testing, which are central to the development of predictive models (Raharja, 2022). In line with this, data mining is a technique for identifying hidden patterns in large-scale datasets and developing predictive models capable of forecasting future trends (Nosratabadi et al., 2022). Commonly used methods in predictive analysis include Decision Trees, Naïve Bayes, and ensemble approaches such as bagging and boosting. The data mining process itself involves browsing and processing existing data to build and refine models so that new data patterns can be recognized (Roihan, Sunarya, & Rafika, 2020).

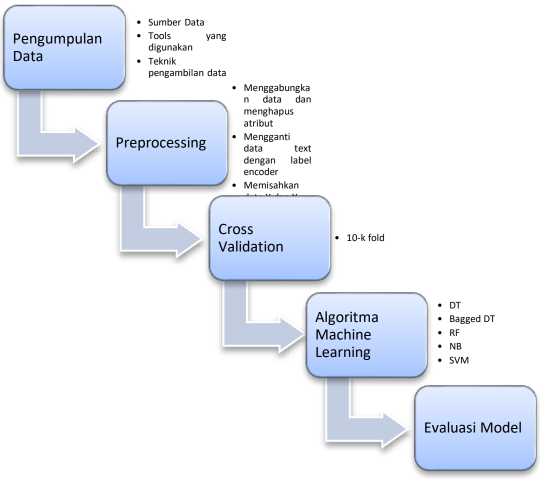
Previous studies have also highlighted the role of machine learning in performance analysis. Research by (Fitriyadi & Kurniawati, 2021), entitled "K-Means and K-Medoids Algorithm Analysis for Clustering Employee Performance Data at National Housing Companies". Performance appraisal is done to measure the performance of an employee on the work done. The National Housing Company conducts performance appraisals of employees every 6 months, involving all employees, both permanent employees, and contract employees. The purpose of this research is to analyze the performance of the K-Means algorithm and the K-Medoids algorithm in conducting the clustering process. Clustering will be grouped into 4 clusters, namely: very good performance level, good performance level, moderate performance level, and poor performance level. The clustering process will be carried out using RapidMiner software. Algorithm performance measurement in RapidMiner is carried out using the Confusion Matrix method with parameters of accuracy, recall, and precision. From the research conducted, the results of the K-Means algorithm have 56% accuracy, 25% precision, and 60% recall, while the K-Medoids algorithm has 14% accuracy, 25% precision, and 25% recall. With these results, it can be concluded that the K-Means algorithm has better performance when compared to the K-Medoids algorithm because it has a higher accuracy and return rate when compared to the K-Medoids algorithm.

Similarly, Fauzi et al. (2020), in their study on the Application of Machine Learning Methods in Predicting the Success of Telemarketing Calls Selling Bank Products "predicted the success of Telemarketing calls in selling Bank products to customers. The Naive Bayes algorithm and the Backward Elimination feature selection can increase the accuracy value in predicting the success of telemarketing in selling bank products well, as evidenced by the accuracy value generated by Naive Bayes of 83.04 %, then after being implemented with the backward elimination feature selection it increases by 6.41 % to 89.45%.

Furthermore, research conducted by (Roihan, Sunarya, & Rafika, 2020) entitled Utilization of Machine Learning in Various Fields: Review Paper. This research, it is explained machine learning. Machine learning is a part of artificial intelligence that is widely used to solve various problems. This article presents a problem-solving review of recent research by classifying machine learning into three categories: directed learning, undirected learning, and reinforcement learning. The results of the review show that the three categories still have the potential to be used in some recent cases and can be improved to reduce the computational load and speed up performance to get a high level of accuracy and precision. The purpose of this article review is to find gaps and serve as guidelines for future research.

MATERIALS AND METHODS

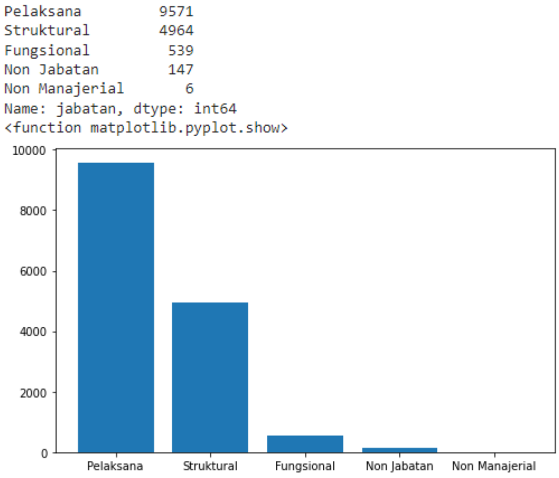
The model proposed for this research, in general, can be described as in Figure 1. At this stage, a machine learning-based job mapping model is proposed by applying the Decision Tree, Bagged Decision Tree, Random Forest, Naive Bayes, and Support Vector Machine (SVM) algorithms. In this study, standard validation is used, namely 10-fold cross-validation where this process divides the data randomly into 10 parts.



Source: (Research Result, 2024)
Figure 1. Proposed Research Model

Dataset

In this study, the data to be used is sourced from the talent management system application database, which is an application that functions to gather candidates or talents to fill a position. The data used include employee data, education data, position and organization data, and talent management value data with cut-off data in October 2021. Position Mapping is shown in Figure 2.



```
Pelaksana          9571
Struktural         4964
Fungsional          539
Non Jabatan         147
Non Manajerial        6
Name: jabatan, dtype: int64
<function matplotlib.pyplot.show>
```

Source: (Research Result, 2024)
Figure 2. Position Mapping Histogram

Modeling

In this study, the object used as research material is employee data sourced from the talent management system application. With data totaling 15227 employees and having 16 attributes, testing using Python version 3, Google Colab, and DBeaver as tools. The stages of machine learning implementation with the Decision Tree, Bagged Decision Tree, Random Forest, Naive Bayes, and Support Vector Machine (SVM) algorithms.

1. Decision Tree

The definition of a decision tree is a simple representation of a classification technique for a finite number of classes, where internal nodes and root nodes are marked with attribute names, the edges are labeled with possible attribute values and leaf nodes are marked with different classes (Redjeki et al., 2024).

The Decision Tree Algorithm is one of the machine learning algorithms commonly used for data mining processes. Data mining is an iterative and exploratory process used to discover hidden patterns and valuable knowledge from large datasets, which supports decision-making and strategic problem-solving (Nosratabadi et al., 2022). Meanwhile, the Decision Tree algorithm itself is one type of classification that represents the shape of the tree structure (Pamungkas, Adiwijaya, & Utama, 2020).

2. Bagged Decision Tree

Bootstrap aggregating or commonly referred to as bagging is one of ensemble learning in data mining classification. Bagging is an ensemble method that is widely applied to classification algorithms, to increase the accuracy of classifiers by combining single classifiers, and the results are better than random sampling (Malek, et al., 2022).

3. Random Forest

Random Forest is a development of Decision Tree, where each Decision Tree has been trained using individual samples and each attribute is split into a tree that is selected between a random subset of attributes. And in the classification process, the individual is based on the vote of the most votes in the population tree collection. Random Forest is an ensemble learning method derived from the CART algorithm, utilizing bootstrap aggregating (bagging) and random feature selection to improve classification accuracy and reduce overfitting (Shlash Mohammad et al., 2025).

4. Naïve Bayes

Naïve Bayes is a statistical classification model that can be used to predict the probability of membership of a class. Naïve Bayes is based on Bayes' theorem which has similar classification capabilities to decision trees and neural networks. The Naïve Bayes (NB) technique is one of the simplest forms of Bayesian networks for classification. A Bayesian network is a directed acyclic graph that represents a joint probability distribution over a set of discrete and stochastic variables, enabling probabilistic reasoning and inference (Nosratabadi et al., 2022).

5. Support Vector Machine

Support Vector Machine (SVM) is a method rooted in statistical learning theory whose results are very promising to provide better results (Aurelia & Rustam, 2021). The Support Vector Machine can also work well on high-dimensional data sets, even Support Vector Machines that use kernel techniques must map the resulting data from their original dimensions into other relatively higher dimensions.

The experimental setup in this research includes dataset, modeling, and evaluation model.

Table 1. Setup Configuration

Algorithm	Parameter	Validation
Decision Tree	Random_state=0, max_depth=5	Model dt, variabel X Y, cv=10
Bagged Decision Tree	base_estimator=dt, n_estimators=100, random_state=0	Model bg, variabel X Y, cv=10
Random Forest	n_estimators=100, random_state=0	Model rf, variabel X Y, cv=10
Naïve Bayes	Default	Model nb, variabel X Y, cv=10
Suppoert Vector Machine	Default	Model svm, variabel X Y, cv=10

Source: (Research Result, 2024)

Evaluation Model

This stage is the last stage of testing this research, the data model that has been validated will be evaluated. So, by evaluating this model, it can be seen how appropriate the research model is. Evaluation is done by analyzing the results of the classification using the measurement of the average value of the results of the cross-validation value. The results of the machine learning algorithm will be compared with the results.

RESULTS AND DISCUSSION

Data Collection Result

The results of data collection were obtained conventionally using the DBeaver application, the first data taken is employee data in the personal table which has 16 attributes and 19430 data. Then the data taken next is the education data of each employee in the Education table which has a total of 10 attributes and a total of 67024 data. Furthermore, the data needed are position and organizational data in the Organization table which has several attributes as many as 22 and a total of 17076 data. The data for the talent value is obtained from the formula result table which has 23 attributes and 318023 data.

The first data to be taken is employee data. This data is obtained from the personal table, in this table, there are several data including employee identification number, employee name, date of birth, gender, religion, and so on. Obtained employee data as much as 19430 data and the number of attributes 16.

Then to get educational data taken from the Education table, this table contains several columns including the employee registration number column, education level, education department, educational institution. The data obtained for employee education is 67024 data with a total of 10 attributes.

Furthermore, to get the position and organization data obtained from the organization table. The data obtained for positions and organizations are 17076 data with a total of 22 attributes. Then to get the talent value data taken from the result formula table. The data obtained for employee talent values are 318023 data with a total of 23 attributes.

Preprocessing Data

After the data collection process is complete, then data preprocessing is carried out, why should data preprocessing be done because not all data attributes will be used for this research. Therefore, the data preprocessing process is needed for this research. All the data that has been obtained will be stored in a separate new table with the main data. In this new table, we will do the preprocessing of the data.

For the first data, namely employee data from all the 16 attributes, only 3 attributes will be taken including age, gender, and religion. While the employee registration number is only used as a key in the process of merging data with other data.

For education data, it will only take the latest educational data held by employees, and the attributes that will be used include education, majors, and institutions. From the results of data collection, 17228 data were obtained. For employees whose education data is incomplete, it is not taken as a research sample.

Furthermore, the data that is preprocessed is the position and organization data which is taken from the organization table. Here the attributes that will be taken are years of service, job id, position id, position, area code, and office code. The job attributes here will be used as Y variables or research labels in the writing of this thesis. From the results of data collection, there were 17076 data on positions and organizations.

Furthermore, talent value data, for talent value data, not all employees have it, because not all employees have not been recommended to carry out talent management or talent events have not been carried out. The talent value data available in this data are aspiration value, career value, experience value, and performance value. Talent value data is taken from the formula result table which is a combined table of several values.

In the process of merging data using SQL query commands where all tables of preprocessing results are merged into one with the employee ID attribute used as a key. Although later this employee ID column will not be included in the test. In the process of merging this data for employees who do not have a value, either one or all values will be filled with a value of 0 automatically. The results of merging this data will be temporarily stored in a new table, to facilitate the next process. Then from the new table, it is converted into a CSV document format, then this data will be tested in machine learning. The attribute names and data types are shown in Table 2.

Table 2. Atribute Name and Data Types

No	Attribute Name	Data Type
1	Age	Int64
2	Gender	Int64
3	Religion	Int64
4	Education	Int64
5	Major	Int64
6	Institution	Int64
7	Years of Service	Int64
8	Job ID	Int64
9	Position ID	Iint64
10	Job Title	Object
11	Region Code	Int64
12	Office Code	Int64
13	Aspiration Value	Float64
14	Career Value	Float64
15	Performance Value	Float64
16	Experience Value	Float64

Source: (Research Result, 2024)

Cross Validation

Next is the cross-validation process for each machine learning method. The first method is Decision Tree, second is Bagged Decision Tree, third is Random Forest, fourth is Naïve Bayes, and the last or fifth is Support Vector Machine or SVM. In carrying out cross-validation for this research, we will use the library model selection from SKLearn, the library is a Cross Val Score with a total of 10 k-fold.

Machine Learning Test Results

There are 6 test scenarios in this research, including with talent value, without talent value, without aspiration value, without career value, without experience value, and without performance value. From these scenarios, the classification results of several algorithms can be seen in Table 3, Machine Learning Test Results.

Table 3. Machine Learning Test Results

Testing	DT	BDT	RF	NB	SVM
With value talent	97.16	98.57	98.4	36.42	65.7
No Talent Value	89.97	91.63	90.65	5.56	65.7
No Aspiration Value	97.14	98.57	98.35	32.08	65.7
No Career Value	90.85	92.13	92.54	25.76	65.7
No Experience Value	97.24	98.65	98.4	36.13	65.7
No Performance Value	97.14	98.57	98.42	35.6	65.7

Source: (Research Result, 2024)

Based on the results presented in Table 3, the scenarios can be explained as follows:

1. The first test scenario, with the talent value, resulted in the Random Forest algorithm 98.4%, Bagged Decision Tree 98.57%, Decision Tree 97.16%, SVM 65.7%, and Naïve Bayes 36.42%. In the first scenario, it is found that the Bagged Decision Tree algorithm has the highest value, which is 98.57%.
2. The second scenario is without talent value, here all value attributes are not used. From the results of the test scenarios obtained Random Forest 90.65%, Bagged Decision Tree 91.63%, Decision Tree 88.97%, SVM 65.7%, and Naïve Bayes 5.56%. In this second scenario, there is a significant decrease, especially in the Naïve Bayes algorithm, but the Bagged Decision Tree algorithm is still the highest.
3. The third scenario is without aspiration value, from testing this third scenario, the following results are obtained: Random Forest 98.35%, Bagged Decision Tree 98.57%, Decision Tree 97.14%, SVM 65.7%, and Naïve Bayes 32.08%. In this scenario, the Bagged Decision Tree algorithm is still the highest.
4. The fourth test scenario is without career value. From this test, the results are Random Forest 92.54%, Bagged Decision Tree 92.13%, Decision Tree 90.86%, SVM 65.7%, and Naïve Bayes 25.76%. There is a decrease from the previous test scenario.
5. The fifth test scenario is without experience value. This test, resulted in Random Forest 98.4%, Bagged Decision Tree 98.65%, Decision Tree 97.24%, SVM 65.7%, and Naïve Bayes 36.13%. From this scenario, there is a decrease compared to the scenario with values.
6. The sixth scenario is without performance value. From this test scenario, it resulted in Random Forest 98.42%, Bagged Decision Tree 98.57%, Decision Tree 97.14%, SVM 65.7%, and Naïve Bayes 35.6%. In this test scenario, the Bagged Decision Tree algorithm is still the highest among other algorithms.

CONCLUSION

From the results of the research above with data as many as 15227 employees used as test data, it can be concluded that the tree-based algorithm method is more suitable for mapping employee positions, especially the Bagged Decision Tree algorithm shows the best performance among other algorithms. From several test scenarios, it was also found that the Bagged Decision tree algorithm had the highest results.

The results obtained from the several test scenarios above for the Bagged Decision Tree algorithm include the talent value getting 98.57% results, without the talent score at all getting 91.63% results. This shows that the value factor can also affect the test results. The next scenario is without aspiration value, showing no decrease of 98.57%. Furthermore, the scenario without career values produces 92.13%, which shows that career values affect the test results. Furthermore, the scenario without the experience value produces a value of 98.65%. And the last one without a performance value produces a value of 98.57%.

This study provides a novel contribution by demonstrating the effectiveness of tree-based algorithms, particularly the Bagged Decision Tree, in addressing challenges in workforce analytics. Unlike traditional HR evaluation methods, this machine-learning approach offers a scalable, data-driven framework for identifying key employee attributes that influence performance. The results underscore the strategic importance of prioritizing talent and career values in predictive models, enabling organizations to optimize recruitment, training, and retention strategies. Additionally, by achieving high accuracy even with limited variables, this research highlights the potential for algorithmic efficiency in resource-constrained environments. These findings bridge gaps between organizational behavior theory and machine learning applications, providing actionable insights for both academic research and practical talent management.

REFERENCES

Aurelia, J., & Rustam, Z. (2021). A Hybrid Convolutional Neural Network-Support Vector Machine for X-ray Computed Tomography Images on CancerA Hybrid Convolutional Neural Network-Support Vector Machine for X-ray Computed Tomography Images on Cancer. *Open Access Macedonian Journal of Medical Sciences, 9*(B), 1283–1289. https://doi.org/10.3889/oamjms.2021.6955

Fauzi, A., Wati, F. F., Sulistyowati, I., Akbar, M. F., Rahmawati, E., & Sari, R. K. (2020). Penerapan Metode Machine Learning Dalam Memprediksi Keberhasilan Panggilan Telemarketing Menjual Produk Bank. *Indonesian Journal on Software Engineering (IJSE), 6*(2), 213–222. https://doi.org/10.31294/ijse.v6i2.8977

Fitriyadi, A. U. & Kurniawati A. (2021). Algoritma K-Means dan K-Medoids Analisis Algoritma K-Means dan K-Medoids Untuk Clustering Data Kinerja Karyawan Pada Perusahaan Perumahan Nasional. *KILAT, 10*(1), 157–168. https://doi.org/10.33322/kilat.v10i1.1174

Ghahremani nahr, J., Nozari, H., & Sadeghi, M. E. (2021). Artificial intelligence and Machine Learning for Real-world problems (A survey). *International Journal of Innovation in Engineering, 1*(3), 38–47. https://doi.org/10.59615/ijie.1.3.38

Intern, D. (2020, Agustus 19). *www.dicoding.com*. (Dicoding Company) Retrieved Desember 11, 2021, from https://www.dicoding.com/blog/machine-learning-adalah/

Islamy, F. J., Yuniarsih, T., Ahman, E., & Kusnendi. (2021). *Efektivitas organisasi berbasis manajemen pengetahuan dalam perspektif perilaku organisasi*. Gracias Logis Kreatif.

Kang, E., & Lee, H. (2021). Employee Compensation Strategy as Sustainable Competitive Advantage for HR Education Practitioners. *Sustainability, 13*(3), 1049. https://doi.org/10.3390/su13031049

KBBI. (n.d.). *Jabatan*. Kamus Besar Bahasa Indonesia (KBBI) Online. Diakses 29 Agustus 2025, dari https://kbbi.web.id/jabatan

Kodithuwakku, S., & Priyanath, H. M. S. (2022). Impact of Intellectual Human Capital and Knowledge Acquisition Capabilities on Financial Performances of Indigenous Craft Industries in Sri Lanka. *Sri Lanka Journal of Social Sciences and Humanities, 2*(2), 93–104. https://doi.org/10.4038/sljssh.v2i2.76

Malek, N. H. A., Yaacob, W. F. W., Wah, Y. B., Md Nasir, S. A., Shaadan, N., & Indratno, S. W. (2022). Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data. *Indonesian Journal of Electrical Engineering and Computer Science, 29*(1), 598. https://doi.org/10.11591/ijeecs.v29.i1.pp598-608

Nosratabadi, S., Zahed, R. K., Ponkratov, V. V., & Kostyrin, E. V. (2022). Artificial Intelligence Models and Employee Lifecycle Management: A Systematic Literature

Review. *Organizacija, 55*(3), 181–198. https://doi.org/10.2478/orga-2022-0012

Pamungkas, Y. W., Adiwijaya, A., & Utama, D. Q. (2020). Klasifikasi Gambar Gigitan Ular Menggunakan Regionprops dan Algoritma Decision Tree. *Jurnal Sistem Komputer Dan Informatika (JSON), 1*(2), 69. https://doi.org/10.30865/json.v1i2.1789

Redjeki, S., Damayanti, A., Hudianti, E., & Nasyuha, A. H. (2024). Implementation of Classification Decision Tree and C4.5 Algorithm in selecting Insurance Products. *Sinkron, 9*(1), 600–608. https://doi.org/10.33395/sinkron.v9i1.13444

Raharja, A. D. (2022, January 11). *www.ekrut.com*. (EKRUT) Retrieved January 11, 2022, from https://www.ekrut.com/media/apa-itu-machine-learning

Rahimi, A., Dharma, A. S., & Norrahman, M. F. (2025). Kinerja pegawai pada Dinas Kesehatan Kabupaten Balangan Provinsi Kalimantan Selatan. *AL IIDARA BALAD: Jurnal Administrasi Negara, 7*(1), 11–21. https://doi.org/10.36658/aliidarabalad.7.1.1313

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology), 5*(1). https://doi.org/10.31294/ijcit.v5i1.7951

Shlash Mohammad, A. A., Alkhazali, Z., Shelash Mohammad, S. I., Al Oraini, B., Vasudevan, A., Alqahtani, M. M., & Alshurideh, M. T. (2025). Machine Learning Models for Predicting Employee Attrition: A Data Science Perspective. *Data and Metadata, 4*, 669. https://doi.org/10.56294/dm2025669