Komputika: Jurnal Sistem Komputer
Volume 13, Nomor 2, Oktober 2024, hlm. 165-172
DOI:10.34010/komputika.v13i2.11917

p-ISSN: 2252-9039
e-ISSN: 2655-3198

Comparative Analysis of Decision Tree and Logistic Regression Models in Employee Recruitment and Selection for Enterprise Success

Dyna Marisa Khairina[1*], Adi Wibowo[2], Budi Warsito[3]

[1)]Program Studi Sistem Informasi, Fakultas Teknik, Universitas Mulawarman
Jl. Sambaliung, Kampus Gunung Kelua, Samarinda, Indonesia

[2)]Departemen Informatika, Fakultas Sains dan Matematika, Universitas Diponegoro
Jl. Prof. Jacob Rais, Tembalang Semarang, Indonesia

[3)]Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
Jl. Prof. Jacob Rais, Tembalang Semarang, Indonesia

**email: dyna.marisa@gmail.com*

(Naskah masuk: *4 Januari 2024*; direvisi: *11 September 2024;* diterima untuk diterbitkan: *7 Oktober 2024*)

ABSTRACT *– Enterprise success is determined by competent Human Resources (HR). The recruitment and selection process of prospective employees plays an important role in producing competent HR, so an effective initial selection is needed to increase the chances of getting the right candidate. This study aims to provide a predictive analysis of the possibility of selecting the next candidate at the interview stage based on the initial selection. Data collection in the form of assessment scores from functional competency tests and behavioral tests are important aspects of the potential suitability and contribution of candidates. This study uses a comparison of logistic regression analysis models and decision trees with several measurement metrics. Based on the results of the evaluation and validation, the logistic regression analysis model is superior. The accuracy value of the logistic regression classification model is 90% with correct prediction results of 54 data and the accuracy value of the decision tree model is 83.3% with correct prediction results of 50 data from 60 test data. The results of this study contribute to the evaluation of the candidate recruitment and selection process as one of the enterprise's efforts to achieve success by providing features that influence the recruitment process.*

Keywords - Prediction; Candidate Selection; Classification; Decision Tree; Logistics Regression

1. INTRODUCTION

Enterprises need human resources to look for competent employees by utilizing various recruitment channels and approaches. Selection of qualified candidates requires recruiting employees who are competent, enthusiastic, and passionate [1]. The essence of recruiting is finding enough of the right candidates with the most suitable qualifications promptly and then hiring the right person from among those candidates [2], for the approach to work, the best candidates must be identified quickly and efficiently [3]. The recruitment and selection process consists of 3 (three) main phases, namely Sourcing, Screening, and Selection [4]. The phases are defined to provide a clear understanding of the variables of the recruitment and selection process. Sourcing is the use of one or more strategies to link talent with vacancies in an organization. External and internal recruiters can be used to find candidates [5]. Screening is a critical stage in the personnel selection process. Recruiters use resume information to infer an applicant's skills, motivation, personality, and suitability for the job [6]. After these phases are passed, a match is obtained, and recruiters find the most suitable candidate for the enterprise.

Related studies have been carried out on several previous studies. Studies that focus on exploring the relationship between recruitment and selection procedures and enterprise success have been carried out [7]. The study by [8] also identifies candidates in specific recruiting for Information Technology (IT) jobs. The study by [1] explores the critical factors that influence recruitment evaluation models and talent selection by developing evaluation models that involve Multi-Attribute Decision-Making (MADM) techniques. The study by [9] conducted research

using systems thinking skills as an additional selection tool/technology with a fuzzy linguistic approach to overcome the subjectivity of decisions in improving the recruitment process and ranking of employee candidates. The study by [10] also explores the process of selecting human resources in companies with decision-making techniques that utilize the Naïve Bayes classification model. Even studies exploring the influence of enterprise social media activity on recruitment success have also been carried out by [11] as well as recent studies that address the systematic literature on recruitment and selection but focus on the role of school principals in school success by [12].

Based on the existing explanation, this study predicts the possibility of candidates being further selected at the interview stage based on behavioral and functional recruitment and selection which are important aspects of the candidate's potential suitability and contribution to the enterprise. An effective initial selection process can significantly improve the quality of recruitment and increase the chances of finding the right candidate for a particular role. The purpose of this study is to provide a predictive analysis of the likelihood of subsequent candidates being selected at the interview stage based on the initial selection. Candidate recruitment and selection play an important role as the starting point in determining the success of an enterprise. Although the previous studies that have been described previously have made great contributions, each study has certain features and model/method focuses so to fill the gap, this study conducts further exploration using influential features and model comparisons. This study contributes to providing insight into the recruitment process and the importance of selecting candidates who have a balance between functional and behavioral competencies to encourage enterprises to identify improvement gaps and improve the recruitment process in achieving adequate Human Resources (HR) competencies and provides contributions for practitioners/researchers as a reference to explore new dimensions of the enterprise's HR recruitment process.

2. METHODS AND MATERIALS

As for the steps as a guideline in this study, you can see in Fig. 1 that it starts with dataset collection and pre-processing, feature selection then divides the dataset into training data and test data for later analysis by comparing 2 (two) classification models, namely decision tree and logistics regression to produce predictive output. The results of the predictions of each of the two models are validated and evaluated with the measurement method so that it can be concluded that there is a significant difference between the two classification models. Explanations for each research stage in Figure 1 are outlined in this section.

Dataset Collection and Pre-processing

The data used is a collection of data from the Human Resources Department (HRD). The data set is the score of 2 (two) main assessments from the functional competency test and the Human Resource (HR) behavior test. The functional competency test is used to evaluate a candidate's hard skills and dominant knowledge while the HR behavioral test is used as an assessment tool that focuses on evaluating soft skills or behavior, teamwork, and adaptability within an organization. There are 9 (nine) assessment variables used for the recruitment and selection process of candidates which can be seen in Table 1.

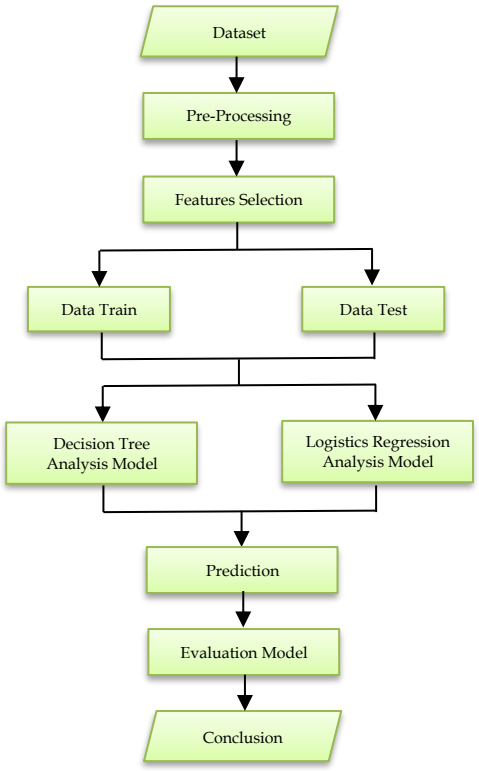

Figure 1. Research stages

Table 1. Dataset Variables

No.	Dataset Variables
1.	Years of experience
2.	Functional competency score
3.	Top1 skills score
4.	Top2 skills score
5.	Top3 skills score
6.	Behavior competency score
7.	Top1 behavior skill score
8.	Top2 behavior skill score
9.	Top3 behavior skill score

For datasets that are not well structured for classification, pre-processing is needed as a data cleaning step to identify missing values in the dataset. The pre-processing stage is also carried out to identify categorical data that needs to be converted into numeric form so that it can be used and processed in the machine learning model. The processes carried out are case-folding, punctuation removal, stop word removal, stemming, and tokenization. The dataset that has been processed first is then analyzed to find certain candidates as candidate classifications called interviews.

Features Selection

This stage is carried out to separate features and labels where there are 9 (nine) assessment variables as features for the candidate recruitment and selection process and 2 (two) labels as classification results that determine the candidates called for interviews by companies. There are 9 (nine) assessment variables and 300 candidate data records for the recruitment and selection process of candidates by dividing training data and test data by 70% for training data and 30% for test data.

Decision Tree Analysis Model

A decision tree is a decision-making model by studies data from the problem domain and builds models to predict results with systematic analysis [13]. A decision tree is a classification method by creating a tree structure like a flow chart where each node represents a feature or attribute as a classification criterion and leaf nodes as the classification result [14]. The determination of features at each branch node in the Decision Tree is calculated based on the Gini index to determine which feature or attribute most influences the classification process. The formula for calculating the value of a feature based on the Gini index can be seen in Equation (1).

$$G = \sum_{K=1}^{K} P(1-P) \quad (1)$$

where G is the Gini value of a feature, K is the number of classes in the attribute and P is the percentage of classes that appear in the attribute.

In classification algorithms, decision tree algorithms are extensively applied to many areas because of their high accuracy, low computational cost, and high interpretability. The process of building a decision tree can be summarized as the following three steps: (1) choosing the best attribute (called 'splitting attribute') according to a certain partition size; (2) dividing the training set by the selected split attribute; (3) produce a branch that corresponds to each separation attribute classification. The algorithm is applied recursively to each classification derived from the separation attribute. If all samples in a certain classification come from the same category, then a leaf node with the name of that category is generated [15]. Decision trees are widely used in many areas of computer science as classifiers, as a means of representing knowledge, and as algorithms for solving various computational geometry problems, combinatorial optimization, and so on [16].

Logistics Regression Analysis Model

Logistics Regression (LR) is used to see the probability of an event and compare the risk of an event considering the factors that influence it [17]. LR is part of the regression analysis used when the dependent (response) variable is dichotomous. This variable can also be illustrated as binary data, and its value is represented by 0 and 1. The general form of LR can be seen in Equation (2).

$$P(Y=1) = 1/[1 + e - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})1] \quad (2)$$

where Y is a binary metric, β_0 is a constant, and β_j is a parameter coefficient with j = 1, 2, ⋯, j. The coefficients for the independent variables are estimated using the logit value as the dependent measure. As the predicted value can be changed to a probability between 0 and 1 which can be seen in Equation (3).

$$Logit_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \quad (3)$$

logit is modeled as a linear combination of predictor variables so that logistic regression can capture the relationship between predictors and binary outcomes.

Performance Measures

Performance measurement is carried out on experimental methods through the validation and evaluation stages. The validation and evaluation of experimental results is a measuring tool to find out how well the comparisons of experimental methods are so that significant differences can be seen between the comparisons of experimental methods [18]. Method performance is analyzed and evaluated through various measures resulting from the confusion matrix. The confusion matrix is generated after the classifier is trained on the validation set to find out which class is causing confusion in the classification and then a more specific classification structure can be created [19]; [20]; [21]. There are 4 (four) terms as a representation of the results of the classification process, namely True Positive (TP),

True Negative (TN), False Positive (FP), and False Negative (FN) [22]. True positive (TP) is the number of positive data obtained correctly. The true Negative (TN) value is the amount of negative data that is collected correctly. The Confusion Matrix model can be seen in Table 2 [23].

Table 2. Model Confusion Matrix

Class	Predicted as Positive (+)	Predicted as Negative (-)
Positive (+)	True Positive (TP)	False Negative (FN)
Negative (-)	False Positive (FP)	True Negative (TN)

Other parameters used to validate the system are accuracy, recall, precision, and F1-score which are evaluated using Equation (4) for accuracy, Equation (5) for precision, Equation (6) for recall, and Equation (7) for F1- scores.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

$$F1 - score = \frac{2\ x\ (Recall\ x\ Precision)}{Recall+Precision} \quad (7)$$

With these parameters, it can be seen how much performance the system has and how well the system performs using the confusion matrix.

3. Results and Discussion

Based on the previous explanation, the existing dataset is processed first, then a decision tree classifier model is created by dividing the dataset into 80% for training data and 20% for testing data from 300 existing data records. After the training data and prediction data are tested, a decision tree model and a logistic regression model are obtained which are visualized to see how the model makes predictions. Figure 2 and Figure 3 show the visualization results of each of the two models.

Both visualizations show that the "years of experience" feature is considered the most influential and important feature for the possibility of candidates being selected further in the interview phase based on behavioral and functional recruitment and selection which are important aspects of the candidate's potential suitability and contribution to the company. The feature that has a weak effect is the "functional competency score" feature. Both models select the most relevant features and values for prediction so that the most influential features can be seen and can be used as steps toward the company's success based on the evaluation of the candidate recruitment and selection process.

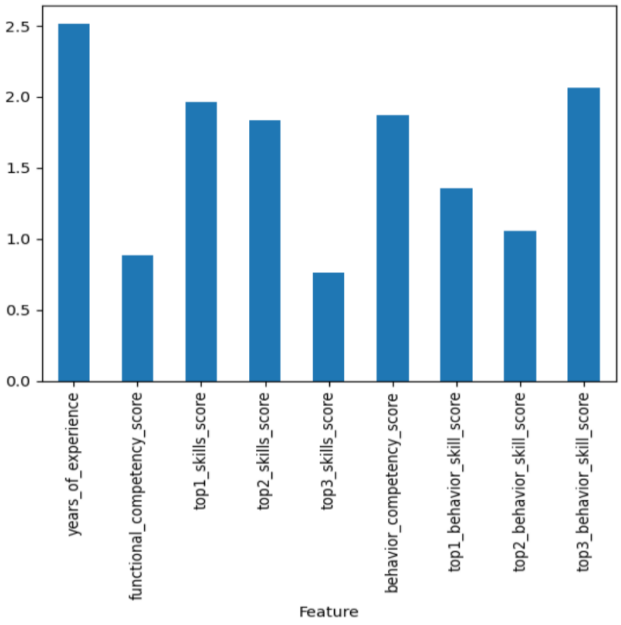

Figure 2. Visualization of the Logistics Regression Models

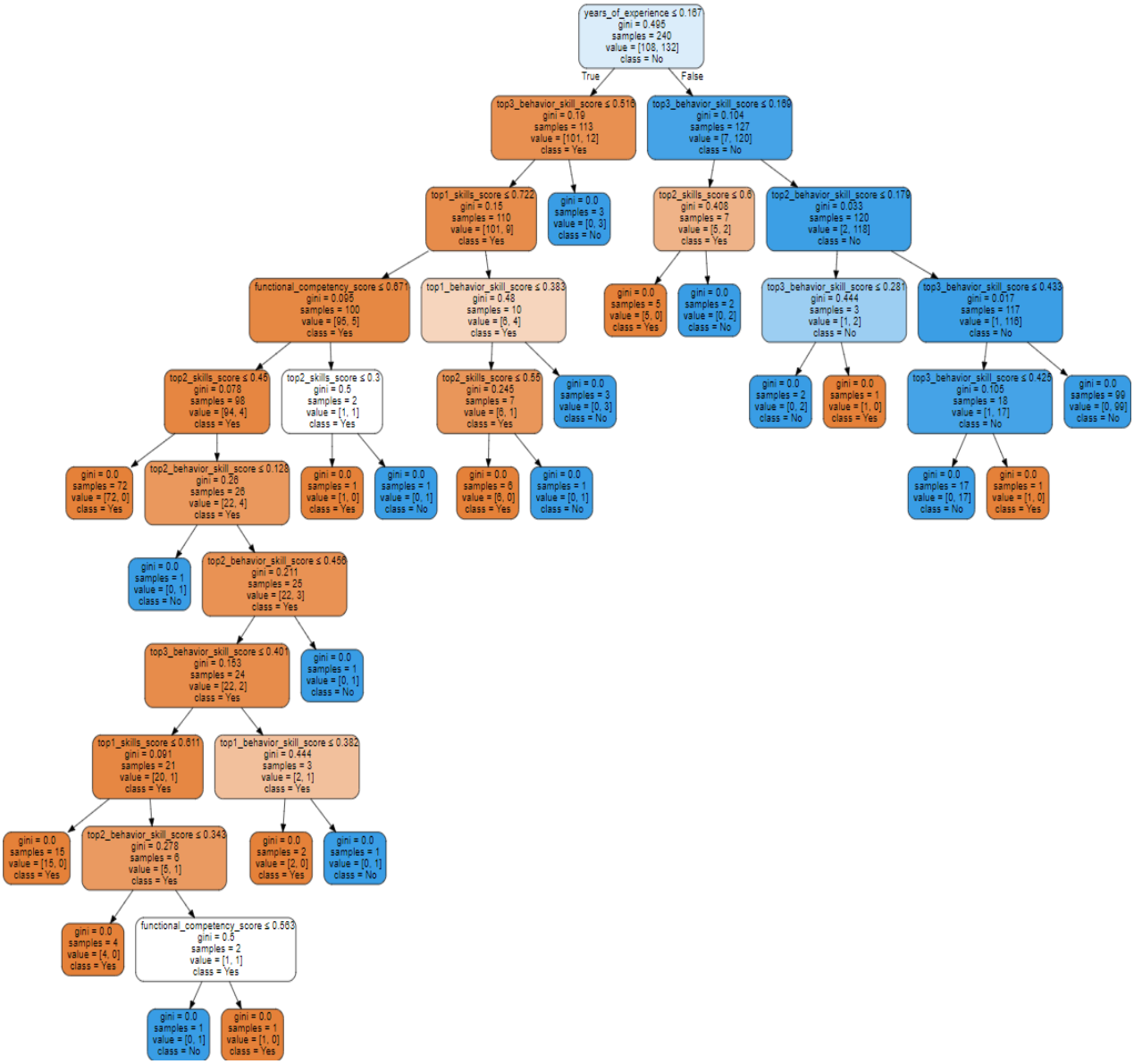

Figure 3. Visualization of the Decision Tree Models

After the prediction results are obtained, the model is evaluated by measuring the performance of each model. From the results of training and data testing predictions, an evaluation of the model that has been built using the confusion matrix is carried out so that it shows the value of accuracy as well as several other evaluation metrics. The visualization for the value of the confusion matrix of each model can be seen in Figure 4 and Figure 5.

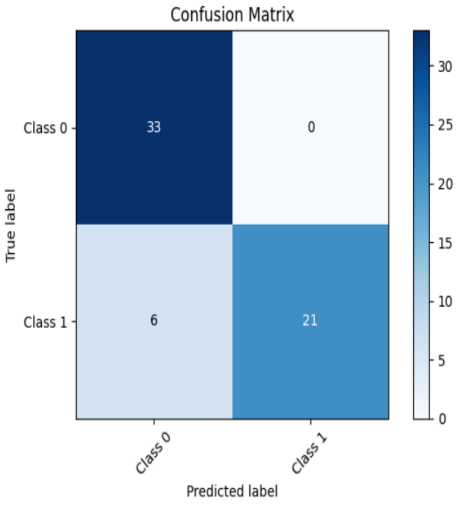

Figure 4. Confusion Matrix on Logistics Regression Analysis Models

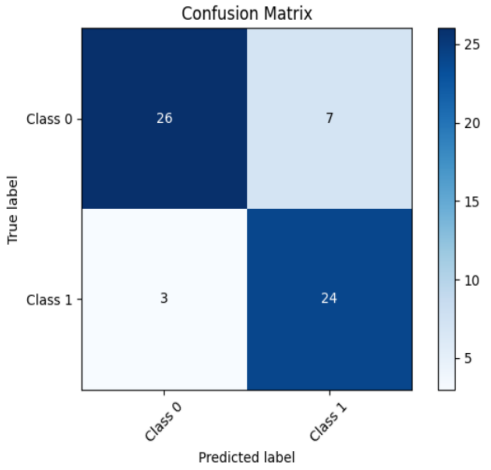

Figure 5. Confusion Matrix on Decision Tree Analysis Models

From Figure 4 and Figure 5, an explanation can be presented based on the confusion matrix terminology along with other evaluation metrics such as accuracy, precision, recall, and f1-score as a more comprehensive understanding of the performance of each model. The results of the evaluation metrics can be seen in Table 3 for the logistic regression model and Table 4 for the decision tree model.

Table 3. Metric Evaluation Logistics Regression Analysis Models

Metric Evaluation	Test Data	TP	TN	FP	FN
Confusion Matrix	60	21	33	0	6
Accuracy		90%			
Precision		91.5%			
Recall		90%			
F1-score		89.8%			

Table 4. Metric Evaluation Decision Tree Analysis Models

Metric Evaluation	Test Data	TP	TN	FP	FN
Confusion Matrix	60	24	26	7	3
Accuracy		83.3%			
Precision		84.2%			
Recall		83.3%			
F1-score		83.4%			

The results of the analysis and testing that have been carried out using 2 (two) classification models, namely logistic regression and decision tree to compare the prediction results of the two models in classifying the possibility of a candidate to be further selected at the interview stage based on recruitment and selection with data used in the analysis consisting of 300 data records divided by 80% for training data or as much as 240 data and 20% for test data or as much as 60 data, the accuracy value between the two models is obtained where the accuracy level for the logistic regression model is 90% with 54 correct prediction data and 6 incorrect prediction data. While the accuracy level for the decision tree model is 83.3% with 50 correct prediction data and 10 incorrect prediction data. From the comparison of these results, the logistic regression model can be said to be more accurate and successful in making correct predictions from the existing samples.

As for other evaluation metrics based on Table 3 and Table 4, a precision value is also obtained to measure the extent to which positive predictions made by the model are correct, namely 91.5% of positive predictions made by the model are correct while the rest are false positives for the logistic regression model, while for the decision tree model, a precision value of 84.2% of positive predictions made by the model is correct, so the logistic regression model is more successful in measuring the extent to which positive predictions are made. Furthermore, a recall value is also produced to measure the extent to which the model can find positive events in actual data with a recall value of 90% in the logistic regression model and 83.3% in the decision tree model, which also shows that the logistic regression model is more successful in recovering all positive events in the actual data. The next measurement is the f1-score value which is the average value between precision and recall measuring the performance of the model in prediction by showing the accuracy and consistency of predictions with f1-score results of 89.8% for the logistic regression model and 83.4% for the decision tree model, thus also showing that the logistic regression model has better performance in data classification. This is because the performance of Logistic Regression can naturally be said to be more robust to imbalanced class problems and is less susceptible to overfitting than more complex decision tree models. The logistic regression model is expected to be a suitable model to help companies recruit and select prospective employees based on their various features.

4. CONCLUSION

Based on the presentation of the results of the analysis, testing, and discussion, it is concluded that the logistic regression model has a better measurement value and is more successful in predicting the likelihood of candidates being selected further in the interview phase based on behavioral and functional recruitment and selection which are important aspects of the candidate's potential fit and contribution to the company. Of the 9 (nine) features used as an assessment, the "years of experience" feature was obtained as the feature with the strongest and most important influence on the likelihood of

candidates being selected further in the interview phase, while the feature with the weakest influence was the "functional competency score" feature. Both models select the most relevant features and values for prediction so that the most influential features can be seen and can be used as a step toward the company's success based on the evaluation of the candidate recruitment and selection process. However, as further research, it is necessary to expand the existing features and explore other classification models to obtain more accurate and optimal results so that they can have implications for the recruitment process directly. Overall, from the comparison of the two models, the performance of the logistic regression classifier model for predicting company success based on the evaluation of the candidate recruitment and selection process tends to be better.

REFERENCES

. [1] P. H. Tsai, Y. L. Kao, and S. Y. Kuo, "Exploring the Critical Factors Influencing the Outlying Island Talent Recruitment and Selection Evaluation Model: Empirical Evidence from Penghu, Taiwan," *Eval Program Plann*, vol. 99, Aug. 2023, doi: 10.1016/j.evalprogplan.2023.102320.

[2] M. P. Michailidis, "The Challenges of AI and Blockchain on HR Recruiting Practices," *Cyprus Review*, vol. 30, no. 2, pp. 169–180, 2018.

[3] P. Horodyski, "Recruiter's Perception of Artificial Intelligence (AI)-Based Tools in Recruitment," *Computers in Human Behavior Reports*, vol. 10, p. 100298, May 2023, doi: 10.1016/j.chbr.2023.100298.

[4] S. Rajesh, M. U. Kandaswamy, and M. A. Rakesh, "The Impact of Artificial Intelligence in Talent Acquisition Lifecycle of Organizations a Global Perspective," *International Journal of Engineering Development and Research*, vol. 6, no. 2, pp. 709–717, 2018.

[5] V. Sinha and P. Thaly, "A Review on Changing Trend of Recruitment Practice to Enhance the Quality of Hiring in Global Organizations," *Management*, vol. 18, no. 2, pp. 141–156, 2013.

[6] B. K. Brown and M. A. Campion, "Biodata Phenomenology: Recruiters' Perceptions and Use of Biographical Information in Resume Screening," 1994.

[7] S. G. Abbasi, M. S. Tahir, M. Abbas, and M. S. Shabbir, "Examining the Relationship Between Recruitment & Selection Practices and Business Growth: An Exploratory Study," *J Public Aff*, vol. 22, no. 2, May 2022, doi: 10.1002/pa.2438.

[8] P. Amsolik and L. Chomatek, "Supporting the identification of promising candidates in the recruitment for IT jobs," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 2263–2272. doi: 10.1016/j.procs.2022.09.285.

[9] S. Karam, M. Nagahi, V. L. Dayarathna (Nick), J. Ma, R. Jaradat, and M. Hamilton, "Integrating Systems Thinking Skills with Multi-Criteria Decision-Making Technology to Recruit Employee Candidates," *Expert Syst Appl*, vol. 160, Dec. 2020, doi: 10.1016/j.eswa.2020.113585.

[10] D. M. Khairina, S. Maharani, Ramadiani, and H. R. Hatta, "Decision Support System for Admission Selection and Positioning Human Resources by Using Naive Bayes Method," *Adv Sci Lett*, vol. 23, no. 3, pp. 2495–2497, Mar. 2017, doi: 10.1166/asl.2017.8653.

[11] D. Golovko and J. H. Schumann, "Influence of company Facebook activities on recruitment success," *J Bus Res*, vol. 104, pp. 161–169, Nov. 2019, doi: 10.1016/j.jbusres.2019.06.029.

[12] S. W. Lee and X. Mao, "Recruitment and selection of principals: A systematic review," *Educational Management Administration and Leadership*, vol. 51, no. 1, pp. 6–29, Jan. 2023, doi: 10.1177/1741143220969694.

[13] S. Patil and U. Kulkarni, "Accuracy Prediction for Distributed Decision Tree Using Machine Learning Approach," in *Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 1365–1371.

[14] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *International Journal of Science and Research*, vol. 5, pp. 2319–7064, 2013.

[15] C. Liu, B. Lin, J. Lai, and D. Miao, "An Improved Decision Tree Algorithm Based on Variable Precision Neighborhood Similarity," *Inf Sci (N Y)*, vol. 615, pp. 152–166, Nov. 2022, doi: 10.1016/j.ins.2022.10.043.

[16] M. J. Moshkov, "Time Complexity of Decision Trees," in *Lecture Notes in Computer Science*, Springer, 2005, pp. 244–459.

[17] G. K. Armah, G. Luo, K. Qin, and A. S. Mbandu, "Applying Variant Variable Regularized Logistic Regression for Modeling Software Defect Predictor," *Lecture Notes on Software Engineering*, vol. 4, no. 2, pp. 107–115, May 2016, doi: 10.7763/lnse.2016.v4.234.

[18] W. Ustyannie, E. Setyaningsih, and C. Iswahyudi, "Optimization of Software Defects Prediction in Imbalanced Class Using A Combination of Resampling Methods With Support Vector Machine and Logistic Regression," *JURNAL INFOTEL*, vol. 13, no. 4, pp. 176–184, Dec. 2021, doi: 10.20895/infotel.v13i4.726.

[19] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[20] P. Cavalin and L. Oliveira, "Confusion Matrix-Based Building of Hierarchical Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2019, pp. 271–278. doi: 10.1007/978-3-030-13469-3_32.

[21] T. Mauritsius, R. Jayadi, S. Alatas, and F. Binsar, "Promo Abuse Modeling in E-Commerce Using Machine Learning Approach," in *International Conference on Orange Technology (ICOT)*, 2020.

[22] N. Umar and B. E. W. Asrul, "Implementation of TOPSIS Methods in Determining Makassar Special Culinary Business Location," in *Proceedings - 2nd East Indonesia Conference on Computer and Information Technology: Internet of Things for Industry, EIConCIT 2018*, Institute of Electrical and Electronics Engineers Inc., Nov. 2018, pp. 82–85. doi: 10.1109/EIConCIT.2018.8878597.

[23] L. Aversano *et al.*, "Thyroid Disease Treatment Prediction with Machine Learning Approaches," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 1031–1040. doi: 10.1016/j.procs.2021.08.106