KOMTEKINFO JOURNAL
LPPM Universitas Putra Indonesia YPTK Padang
Jl. Raya Lubuk Begalung, Padang, West Sumatera, Indonesia, Zip Code: 25221
Volume: 13, Issue: 3, Page: 143 - 151 September 30th, 2026, e-ISSN: 2502-8758
Available online at website: https://jkomtekinfo.org/ojs/index.php/komtekinfo

Integrating Educational Data Mining through Decision Tree and Random Forest Comparison for Assessing Employment Suitability

Riski Hidayatullah Maas[1]✉, Sumijan[2], Rini Sovia[3]
[1] Universitas Negeri Padang, Padang, 25132, Indonesia
[2,3] Information Technology Study Program, Faculty of Computer Science, Universitas Putra Indonesia YPTK, Padang, 25221, Indonesia
Riskihm16@gmail.com

Abstract

Educational Data Mining can be utilized to identify patterns in educational data and support the evaluation of the alignment between graduates' competencies and labor market requirements. Assessing alumni employment suitability is essential for providing an overview of their readiness and job alignment after completing their education. This study aims to compare the performance of the Decision Tree (DT) and Random Forest (RF) algorithms in classifying alumni employment suitability levels. The research stages included data collection, selection of employed alumni data, data cleaning, transformation of categorical variables into numerical scores, target label construction, division of the dataset into training and testing sets, and class balancing using Random Over-Sampling. The classification models were constructed using the DT and RF algorithms and evaluated based on accuracy, precision, recall, F1-score, and confusion matrix. The dataset was obtained from the 2025 Alumni Tracer Study conducted by Universitas Negeri Padang and comprised 2,831 records with eleven variables related to alumni employment and competencies. The results indicate that the Decision Tree achieved an accuracy of 83.42%, a macro-averaged precision of 76.74%, a macro-averaged recall of 76.82%, and a macro-averaged F1-score of 76.59%. Random Forest demonstrated better performance, achieving an accuracy of 90.83%, a macro-averaged precision of 83.85%, a macro-averaged recall of 82.22%, and a macro-averaged F1-score of 82.94%. This study contributes an Educational Data Mining-based approach that can assist higher education institutions in evaluating alumni employment suitability and provide a basis for formulating policies aimed at improving graduates' competencies.

Keywords: DT, RF, Alumni Employment Suitability

KomtekInfo Journal is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.

1. Introduction

Alumni employment suitability is a strategic indicator for assessing the success of higher education institutions in producing graduates who are capable of securing employment, applying the competencies acquired during their studies, adapting to professional environments, and consistently maintaining their competitiveness across various employment sectors amid increasingly dynamic changes in labor market demands [1]. Technological advancements, automation, and organizational transformation have changed the types of skills required by industries. Therefore, educational institutions must ensure that graduates' academic competencies, digital skills, communication abilities, teamwork, ethics, and self-development capabilities remain relevant to actual employment demands at the local, national, and global levels [2]. Misalignment among curricula, graduates' competencies, and employers' requirements may prolong the job-search period, reduce the alignment between graduates' fields of study and employment, and limit alumni career development, particularly during the initial transition from higher education to the labor market [3].

The urgency of assessing employment suitability is also associated with the need for higher education institutions to obtain objective evidence regarding graduates' transition from education to employment. Such evidence can serve as a basis for evaluating learning quality, improving curricula, strengthening career services, and ensuring relevant, responsive, and sustainable institutional accountability to stakeholders [4]. Tracer studies provide information concerning employment status, job-search duration, company characteristics, workplace level, the relationship between fields of study and employment, and the competencies applied by alumni across various faculties and study programs. However, the large volume of data and diversity of attributes make important patterns difficult to identify through descriptive analysis alone [5]. Therefore, Educational Data Mining is required to transform alumni data, which initially serve as administrative reporting material, into analytical knowledge that can assist universities in identifying dominant factors, assessing employment suitability

Submitted: 12-06-2026 | Revised: 13-07-2026 | Accepted: 20-08-2026 | Published: 30-09-2026

levels, and formulating data-driven policies that are academically and institutionally accountable [6].

Previous research has shown that classification techniques can be used to predict job-search duration, employment alignment, career success, and graduate readiness within a certain period after graduation by utilizing academic characteristics, competencies, and employment information collected through tracer studies [7]. In various evaluations, Decision Tree provides interpretable classification rules, whereas Random Forest tends to produce more stable and consistent predictions because it combines multiple decision trees and reduces dependence on a single model structure [8]. Nevertheless, the comparative performance of these two algorithms may vary according to data quality, variable selection, class distribution, target construction, data-balancing procedures, and parameter configurations. Therefore, specific testing using alumni data within a single institutional environment is necessary to produce results that are more contextual, operational, and relevant to institutional policy requirements [9].

Decision Tree is a supervised learning algorithm that constructs a tree structure by recursively partitioning data based on the attributes that most effectively distinguish the target classes. Consequently, the relationships among employment characteristics, alumni competencies, and employment suitability categories can be translated into easily interpretable if–then decision rules [10]. Attribute selection at each node is performed using impurity measures such as entropy, Information Gain, or the Gini Index. The branching process continues until the data reach a certain level of homogeneity or satisfy stopping criteria, including maximum tree depth, minimum number of samples, and maximum number of leaf nodes [11]. The main advantages of Decision Tree are model transparency and ease of interpretation. However, the algorithm is susceptible to overfitting, sensitive to changes in the composition of the training data, and may produce different structures when the data contain noise or inconsistencies [12].

Random Forest is an ensemble algorithm that constructs multiple Decision Trees using bootstrap samples and randomly selected feature subsets. The final classification result is then determined through majority voting across all constructed trees [13]. The bagging mechanism and random feature selection help reduce variance, improve prediction stability, capture nonlinear interactions, and minimize the risk of overfitting in educational datasets containing numerous categorical attributes and complex relationships among variables [14]. Although Random Forest generally requires greater computational resources and offers lower interpretability than a single decision tree, feature importance values can be used to explain the relative contribution of each variable to the classification of alumni employment suitability [15].

This study applies a quantitative approach within the Educational Data Mining framework through the stages of data collection, variable analysis, data cleaning, value transformation, target label construction, data partitioning, model development, testing, and result interpretation in accordance with the Knowledge Discovery in Databases process [16]. The research dataset was obtained from the 2025 Alumni Tracer Study of Universitas Negeri Padang and consisted of 2,831 records. The data were subsequently selected by including only alumni with employed status and eleven input variables representing employment characteristics and graduates' principal competencies. The classification target was divided into four categories: Suitable, Moderately Suitable, Less Suitable, and Unsuitable. The dataset was then stratified into training and testing sets using an 80:20 ratio, while Random Over-Sampling was applied only to the training data to maintain the objectivity of the evaluation process [17].

The performance of the Decision Tree and Random Forest models is compared based on accuracy, macro precision, macro recall, macro F1-score, and weighted F1-score, and confusion matrices. These metrics allow the evaluation to assess not only the overall number of correct predictions but also the models' ability to proportionally identify each employment suitability category [18]. The Decision Tree structure and Random Forest feature importance values are analyzed to identify the most influential variables, explain class-formation patterns, and describe the differences between the interpretability of a single model and the predictive stability of an ensemble model in processing alumni data [19]. The comparative results are expected to contribute an analytical model that supports the evaluation of graduate competencies, the strengthening of career development services, curriculum adjustment, and the utilization of tracer study data as a basis for systematic, measurable, and sustainable decision-making in higher education institutions [20].

2. Methods

The research framework illustrates a systematic sequence of stages used to process tracer study data into information regarding alumni employment suitability levels. The research process begins with data collection and variable analysis, followed by data preprocessing, classification model development using the Decision Tree and Random Forest algorithms, model evaluation, and performance comparison between the two algorithms. This framework serves as a guideline to ensure that each stage of the research is conducted in a structured manner and produces classification models that can be evaluated using predetermined performance metrics. The detailed research framework is presented in Figure 2.1.

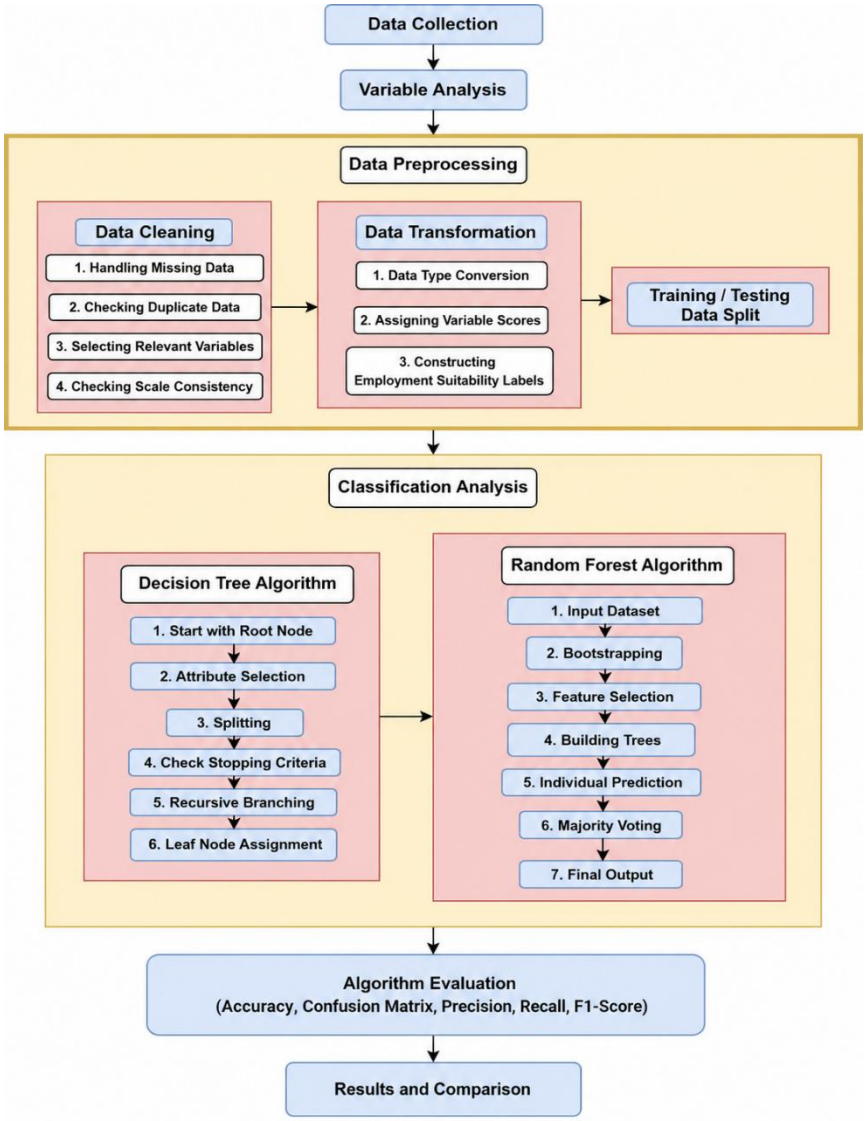


Figure 1. Research Framework

Based on Figure 1, the study began with the collection of 2,831 initial records from the 2025 Alumni Tracer Study of Universitas Negeri Padang. Variable analysis was conducted to identify the attributes associated with alumni employment suitability, while the data used for model development were limited to alumni with employed status. The input variables consisted of company type, job-search duration, relationship between field of study and employment, workplace level, ethics, English proficiency, communication, teamwork, self-development, field-specific expertise, and information technology utilization.

The preprocessing stage included handling missing values, checking duplicate records, selecting relevant variables, examining category consistency, converting data types, assigning scores to variables, and constructing the target labels. The processed dataset was divided into 80% training data and 20% testing data using stratified splitting, while Random Over-Sampling was applied only to the training data to balance the class distribution. The data were subsequently classified using Decision Tree and Random Forest, and the performance of both models was evaluated using accuracy, precision, recall, F1-score, and confusion matrices before the results were compared.

2.1 Preprocessing

Data preprocessing was conducted to produce clean, consistent, and classification-ready data by examining missing values, duplicate records, data types, and the consistency of categorical entries [21]. In this study, the dataset was restricted to alumni with employed status. The eleven research variables were then transformed into numerical scores and aggregated to construct the employment suitability labels. After the labels were generated, the dataset was stratified into 80% training data and 20% testing data, while Random Over-

Sampling was applied exclusively to the training data to reduce class imbalance [17].

The scores of the eleven variables were aggregated to obtain the Total Score using the following equation:

$$T = \sum_{j=1}^{11} X_j \quad (1)$$

The target labels were constructed according to the following criteria:

$$Y = \begin{cases} \text{Suitable}, & 100 \leq T \leq 130 \\ \text{Moderately Suitable}, & 90 \leq T \leq 99 \\ \text{Less Suitable}, & 80 \leq T \leq 89 \\ \text{Unsuitable}, & T < 80 \end{cases}$$

After the target labels were constructed, the dataset was stratified into 80% training data and 20% testing data. Random Over-Sampling was applied only to the training data to balance the class distribution, while the testing data retained their original distribution.

2.2 Decision Tree algorithn

Decision Tree is a classification algorithm that constructs a model in the form of a tree structure consisting of a root node, internal nodes, branches, and leaf nodes. This algorithm was selected because it can transform relationships among variables into relatively interpretable if–then decision rules. In this study, Decision Tree was used to classify alumni employment suitability into four categories: Suitable, Moderately Suitable, Less Suitable, and Unsuitable, based on eleven input variables. The entropy equation is expressed as follows:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 (p_i) \quad (2)$$

where p_i represents the probability of class (i) in dataset (S). After the entropy value is calculated, Information Gain (IG) is used to determine the amount of information obtained by splitting the data based on attribute (A) [22]:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \quad (3)$$

The Classification and Regression Tree (CART) algorithm uses the Gini Index to measure data impurity. The equation is expressed as follows:

$$Gini(s) = 1 - \sum_{i=1}^{c} (p_i)^2 \quad (4)$$

A lower Gini Index value indicates that the class distribution within a node is more homogeneous. However, the Decision Tree model employed in this study used entropy as the splitting criterion; therefore, the primary attribute was selected based on the highest Information Gain value.

The tree-building process was terminated when all observations within a node belonged to the same class, no attribute produced a better split, the number of observations in the node reached the specified minimum threshold, or the maximum tree depth was reached. Model complexity was controlled through the parameters (max_depth), (min_samples_split), (min_samples_leaf), (max_leaf_nodes), and (ccp_alpha) to reduce the risk of overfitting. Each terminal node was subsequently assigned a label based on the majority class of the observations contained within that node and used as the predicted alumni employment suitability category.

2.3. Random Forest Algorithm

Random Forest is an ensemble learning algorithm that constructs multiple Decision Trees to produce more stable classification results than a single decision tree. Each tree is developed using randomly selected data samples and feature subsets, after which all prediction results are combined through a majority voting mechanism. In this study, Random Forest was used to classify alumni employment suitability into four categories: Suitable, Moderately Suitable, Less Suitable, and Unsuitable.

The model-building process begins with bootstrap sampling, in which samples are randomly drawn with replacement from the training data to form the training set for each tree. At each node, the algorithm considers only a randomly selected subset of features to determine the best split using entropy and Information Gain criteria. This process is repeated until the number of trees specified by the (n_estimators) parameter is reached.

Each tree generates one class prediction for the evaluated observation. The final prediction is determined by selecting the class that receives the highest number of votes from all trees, as expressed in the following equation:

$$\hat{y} = \arg\max_{c \in C} \sum_{b=1}^{B} I\,(h_b(x) = c) \quad (5)$$

In addition to generating predictions, Random Forest provides feature importance values that indicate the contribution of each variable to the classification process. The model was subsequently tested using the same testing dataset as the Decision Tree model and evaluated based on accuracy, precision, recall, F1-score, and confusion matrix. This multiple-tree approach can reduce variance and the risk of overfitting, although its interpretation is more complex than that of a single Decision Tree.

3. Results and Discussions

The results and discussion section reports the research findings and explains their meaning and implications [11]. As the central part of the paper, this section should provide clear answers to the research questions formulated in the study [12]. The findings should be presented systematically, beginning with the most important results and continuing to supporting findings,

or arranged chronologically when this structure is more appropriate.

When reporting the results, tables and figures should be used to present complex information clearly and efficiently. The discussion should identify important patterns, trends, and observations, compare the findings with previous studies, and provide explanations for results that are unexpected or inconsistent. It should also describe the theoretical and practical implications of the findings, acknowledge the study's limitations, and recommend possible directions for future research.

3.1 Dataset Tracer Study

The data set used in this study was obtained from the 2025 Alumni Tracer Study of Universitas Negeri Padang, which was managed by the Career Development and Entrepreneurship Unit of Universitas Negeri Padang. The initial dataset comprised 2,831 alumni records. It contained information related to alumni employment conditions, including company type, job-search duration, the relationship between field of study and employment, workplace level, and alumni competencies. A sample of the data used in this study is presented in Table 1.

Table 1. Dataset tracer Study

No.	Student ID	Job-Search Duration (Months)	Company Type	Workplace Level	Relationship Between Field of Study and Employment
1	19062018	9	Private Company	National/Legally Incorporated Business	Moderately Related
2	21134028	1	Government Institution (including State-Owned Enterprises)	National/Legally Incorporated Business	Closely Related
3	21037039	5	Private Company	Local/Regional/Unincorporated Business	Closely Related
4	21072057	2	Government Institution (including State-Owned Enterprises)	Local/Regional/Unincorporated Business	Very Closely Related
5	20080034	0	Private Company	National/Legally Incorporated Business	Closely Related

Table 1. Dataset tracer Study (Continued)

Ethics	Field-Specific Expertise	English Proficiency	Information Technology Utilization	Teamwork	Self-Development	Communication
High	High	High	High	High	High	High
High	High	High	High	High	High	High
Very High	High	Moderate	High	High	High	High
Very Low	Very Low	Very Low	Very Low	Very Low	Very Low	Very Low
Very High	High	High	High	Very High	High	High

Based on Table 1, each row represents an individual alumnus, while each column indicates an attribute related to alumni employment characteristics and competencies. The attributes included in the dataset are company type, job-search duration, relationship between field of study and employment, workplace level, ethics, English proficiency, communication, teamwork, self-development, field-specific expertise, and information technology utilization. The dataset was subsequently used as the primary data source for assessing and classifying alumni employment suitability.

3.2 Variable Analysis

Variable analysis was conducted on the tracer study data after the preprocessing stage to ensure that each attribute had a consistent format, category, and value. The variables used in this study included company type, job-search duration, relationship between field of study and employment, workplace level, ethics, English proficiency, communication, teamwork, self-development, field-specific expertise, and information technology utilization. A sample of the variable analysis results after preprocessing is presented in Table 2.

Table 2. Data Preprocessing Results

No	Nim	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Target
1	19062018	9	3	2	3	4	4	4	4	4	4	4	Moderately Suitable
2	21134028	1	1	2	2	4	4	4	4	4	4	4	Suitable
3	21037039	5	3	1	2	5	4	3	4	4	4	4	Suitable
4	21072057	2	1	1	1	1	1	1	1	1	1	1	Unsuitable
5	20080034	0	3	2	2	5	4	4	4	5	4	4	Suitable

Based on Table 2, each variable was adjusted according to the category rules and weighting scheme established in this study. The values of the eleven variables were used to represent the employment characteristics and competency levels of each alumnus as the basis for assessing employment suitability. The results of the variable analysis were subsequently used as input for developing classification models through the application of the Decision Tree and Random Forest algorithms.

3.3 Decision Tree Classification

Decision Tree classification was performed by constructing a tree structure based on the training data prepared in the previous stage. The model progressively partitioned the data according to the variables and threshold values that most effectively distinguished the target classes until terminal nodes were produced. The resulting Decision Tree structure developed in this study is presented in Figure 2.

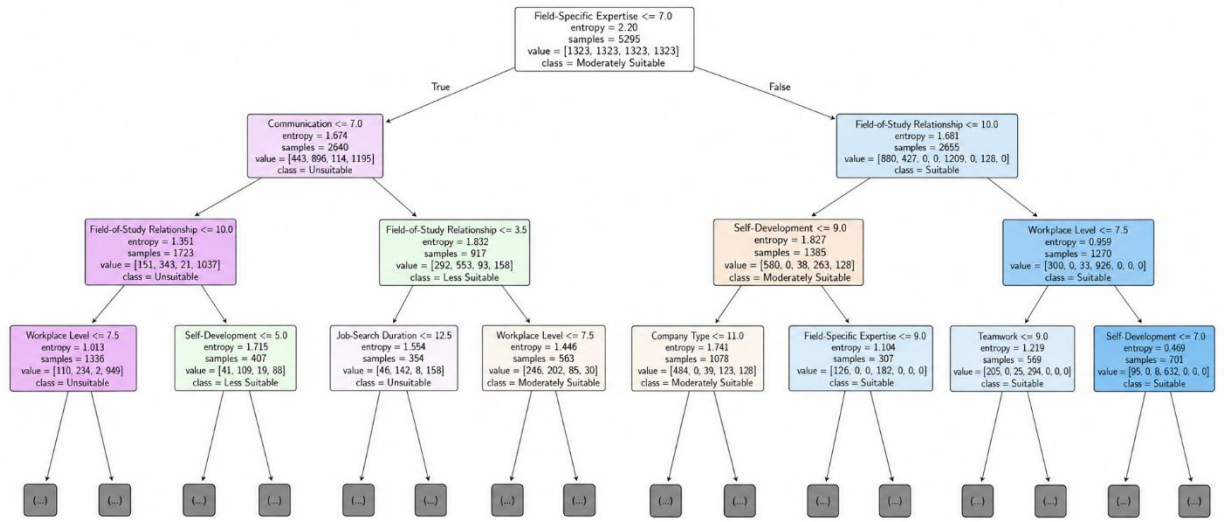


Figure 2. Resulting Decision Tree Structure

Based on Figure 2, each internal node represents the variable and threshold value used to partition the data, while each branch indicates the corresponding decision path. The traversal process begins at the root node and proceeds through the appropriate branches until it reaches a leaf node representing one of the following classes: Suitable, Moderately Suitable, Less Suitable, or Unsuitable. The resulting model was subsequently used to classify 567 testing records that had not been involved in the training process.

3.4 Random Forest Classification

Random Forest classification was performed through the construction of multiple decision trees using the training data processed in the previous stage. Each tree was developed from a different data sample through bootstrapping and random feature selection, thereby providing the model with more stable classification performance. The resulting Tree 1 of the Random Forest model developed in this study is presented in Figure 3.

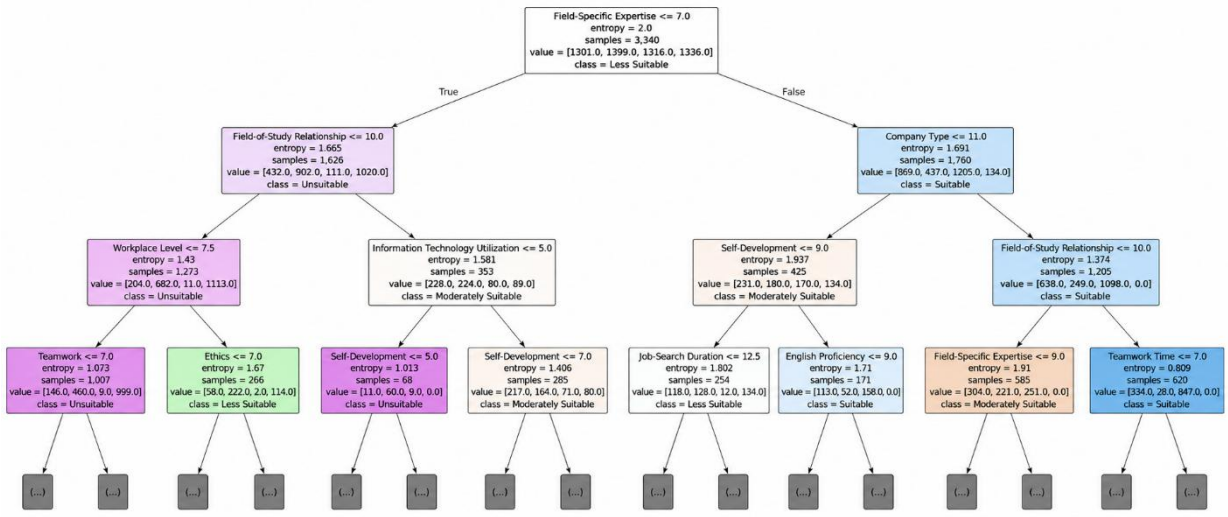


Figure 3. Resulting Tree 1 of the Random Forest Model

Based on Figure 3, Tree 1 represents one of the decision tree structures constructed within the Random Forest model to classify alumni data into the Suitable, Moderately Suitable, Less Suitable, and Unsuitable categories. Each node in the tree represents a splitting variable, while each branch indicates the decision path leading to a leaf node as the classification outcome. This tree is one of multiple trees employed in the Random Forest model; therefore, the model's final prediction is determined through majority voting across all constructed trees.

3.5 Evaluation

The evaluation was conducted on 567 testing records to measure the performance of the Decision Tree and Random Forest algorithms. The evaluation employed a confusion matrix to compare the actual classes with the predicted classes. The Decision Tree confusion matrix results are presented in Figure 4.

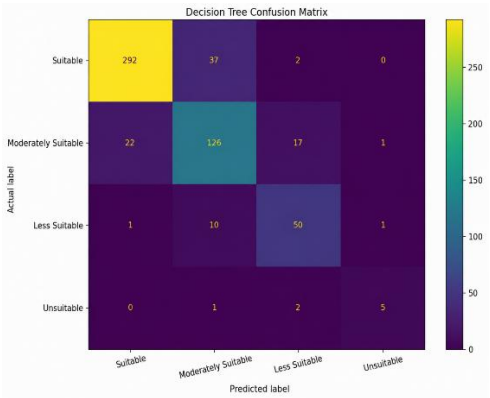


figure 4. Confusion Matrix Decision tree

Based on Figure 4, the diagonal values represent correct predictions, while the off-diagonal values indicate classification errors. The Decision Tree algorithm achieved an accuracy of 83.42%. The Random Forest confusion matrix results are presented in Figure 5.

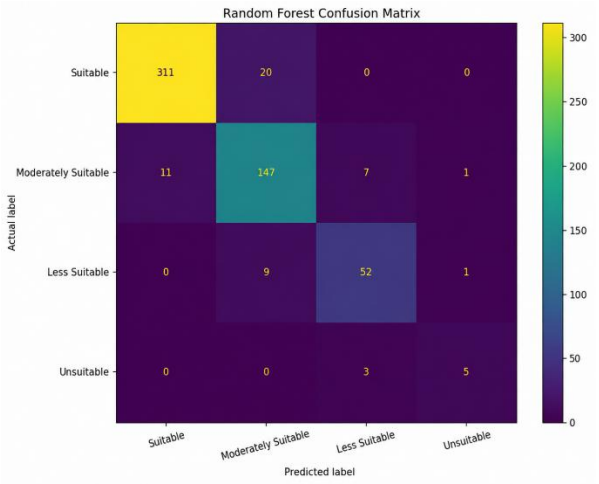


Figure 5. Confusion Matrix Random Forest

Based on Figure 5, Random Forest generated a greater number of correct predictions and achieved an accuracy of 90.83%. The performance of both algorithms was subsequently compared using accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and weighted F1-score. The evaluation results of both algorithms were subsequently compared using multiple evaluation measures, as presented in Table 3.

Table 3. Algorithm Comparison Results

	Model	Accuracy (%)	Precision Macro (%)	Recall Macro (%)	F1-Score Macro (%)	F1-Score Weighted (%)
0	Decision Tree	83.42	76.74	76.82	76.59	83.64
1	Random Forest	90.83	83.85	82.22	82.94	90.89

Based on Table 3, Random Forest obtained higher scores than Decision Tree on all performance measures. The accuracy difference between the two algorithms was 7.41 percentage points, indicating that the use of multiple trees and a majority voting mechanism improved classification performance. Therefore, Random Forest demonstrated better performance in classifying alumni employment suitability.

4. Conclusions

This study successfully integrated the Educational Data Mining approach with the Decision Tree and Random Forest models for classify alumni employment suitability into four categories: Suitable, Moderately Suitable, Less Suitable, and Unsuitable. The objective of comparing the performance of the two algorithms was achieved through testing on 567 testing records Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices. The findings indicated that Random Forest achieved an accuracy of 90.83%, which was higher than the 83.42% accuracy obtained by Decision Tree, with a difference of 7.41 percentage points. This study contributes a tracer study data-driven analytical model that can be used to assess alumni employment suitability in a systematic and measurable manner. The classification results and information regarding influential variables can assist higher education institutions in evaluating graduates' competencies and their alignment with labor market requirements. The developed model can also support career service development, alumni competency enhancement, and continuous curriculum improvement.

References

[1] R. Haque, A. Quek, C.-Y. Ting, H.-N. Goh, and R. Hasan, "Classification Techniques Using Machine Learning for Graduate Student Employability Predictions," vol. 14, no. 1, 2024.

[2] S. Y. Ong, C. Y. Ting, H. N. Goh, A. Quek, and C. L. Cham, "Workplace Preference Analytics Among Graduates," *Journal of Informatics and Web Engineering*, vol. 2, no. 2, pp. 233–248, Sep. 2023, doi: 10.33093/jiwe.2023.2.2.17.

[3] H. N. Mpia, L. W. Mburu, and S. N. Mwendia, "Applying Data Mining in Graduates' Employability: A Systematic Literature Review," 2023, *International Association of Online Engineering*. doi: 10.3991/IJEP.V13I2.33643.

[4] C. Munthali *et al.*, "A graduate tracer study of the diploma in nursing program at Saint John of God College of Health Sciences in Mzuzu, Malawi: A descriptive survey design," *Int. J. Afr. Nurs. Sci.*, vol. 23, p. 100925, Jan. 2025, doi: 10.1016/J.IJANS.2025.100925.

[5] S. Anastassia, A. Kharis, A. Haqqi, and A. Zili, "Tantangan Dan Peluang Educational Data Mining Dan Learning Analytics Pada Pendidikan Jarak Jauh Di Indonesia," *Prosiding Seminar Nasional Sains dan Teknologi "SainTek,"* vol. 1, no. 2, pp. 218–225, Aug. 2024, Accessed: Apr. 15, 2026. [Online]. Available: https://conference.ut.ac.id/index.php/saintek/article/view/2536

[6] E. Kalita *et al.*, "Educational data mining: a 10-year review," *Discover Computing 2025 28:1*, vol. 28, no. 1, pp. 81-, May 2025, doi: 10.1007/S10791-025-09589-Z.

[7] S. Hidayatuloh and D. Wangsa Samudera, "Perbandingan Akurasi Metode Naïve Bayes, Decision Tree (C.45), Dan Random Forest Dalam Melakukan Prediksi Masa Tunggu Kerja Alumni Berdasarkan Data Tracer Study Pada Fakultas Psikologi Uin Jakarta", doi: 10.37817/Tekinfo.v26i1.

[8] N. Sulistiyaningsih and R. Rismayati, "Comparison of Random Forest, Decision Tree, and XGBoost Models in Predicting Student Academic Success," *Journal of Artificial Intelligence and Software Engineering*, vol. 5, no. 3, pp. 920–930, 2025, doi: 10.30811/jaise.v5i3.7138.

[9] L. G. Hamilton, "Emotionally Based School Avoidance in the Aftermath of the COVID-19 Pandemic: Neurodiversity, Agency and Belonging in School," *Education Sciences 2024, Vol. 14, Page 156*, vol. 14, no. 2, p. 156, Feb. 2024, doi: 10.3390/EDUCSCI14020156.

[10] N. Wati and S. Mamase, "DECISION TREE," *JTII*, vol. 07, no. 02, 2022.

[11] Chen Sa and Lin Xiankun, "Application of Decision Tree Algorithm in Educational Data Mining," *Curriculum and Teaching Methodology*, vol. 6, no. 8, 2023, doi: 10.23977/curtm.2023.060818.

[12] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: from efficient prediction to responsible AI," 2023, *Frontiers Media SA*. doi: 10.3389/frai.2023.1124553.

[13] P. Ayu Firnanda *et al.*, "Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket," *Emerging Statistics and Data Science Journal*, vol. 3, no. 1, pp. 445–461, Jan. 2025, doi: 10.20885/ESDS.VOL3.ISS.1.ART2.

[14] D. K. Dake, "Information technology graduates employability prediction model in a low-income country using tree-based machine learning classifiers," *Discover Global Society*, vol. 3, no. 1, Dec. 2025, doi: 10.1007/s44282-025-00304-3.

[15] Y. Li and Y. Mu, "Research and performance analysis of random forest-based feature selection algorithm in sports effectiveness evaluation," *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-76706-1.

[16] R. Shankar and A. Yip, "Transforming Patient Feedback Into Actionable Insights Through Natural Language Processing: Knowledge Discovery and Action Research Study.," *JMIR Form. Res.*, vol. 9, no. 1, p. e69699, Aug. 2025, doi: 10.2196/69699.

[17] E. Rohaeti and A. Andriyati, "Comparative Study of Predictive Classification Models on Data with Severely Imbalanced Predictors," 2024.

[18] J. Chaidir and J. Friadi, "Zona Teknik: Jurnal Ilmiah Evaluasi Kinerja Algoritma Klasifikasi dalam Studi Kasus Prediksi Kelulusan di Universitas XYZ," vol. 19, no. 1, pp. 15–22, 2025, doi: 10.37776/zt.vxix.xxx.

[19] A. Prof, D. Davuth, and M. Dy Davidseihakbot, "Topic: A Machine Learning Approach to Identifying Factors Influencing Alumni Career Success in Tracer Study: A Case Build Bright University."

[20] R. Bagus Prakoso, D. Lutfianawati, U. Malahayati, and B. Lampung, "Hubungan Antara Efikasi Diri Dan Kesadaran Metakognitif Dengan Kesiapan Kerja Mahasiswa Tingkat Akhir," *Nathiqiyyah*, vol. 8, no. 2, pp. 201–216, Dec. 2025, doi: 10.46781/NATHIQIYYAH.V8I2.1762.

[21] M. Guo *et al.*, "Normal Workflow and Key Strategies for Data Cleaning Toward Real-World Data: Viewpoint.," *Interact. J. Med. Res.*, vol. 12, no. 1, p. e44310, Sep. 2023, doi: 10.2196/44310.

[22] R. Mao, X. Shi, Z. Shi, R. Mao, X. Shi, and Z. Shi, "A Decision Tree Classification Algorithm Based on Two-Term RS-Entropy," *Entropy 2025, Vol. 27*, vol. 27, no. 10, Oct. 2025, doi: 10.3390/E27101069.

Biographies of Authors



Riski Hidayatullah Maas I was born on November 16, 1994, in Padang. My most recent educational qualification is a Bachelor's degree (S1) in Informatics Engineering Education from Universitas Negeri Padang. Currently, I am pursuing a Master's degree (S2) in Computer Science at Universitas Putra Indonesia YPTK Padang. I can be contacted via email at riskihm16@gmail.com



Dr. Ir. Sumijan, M.Sc. born in Nganjuk on May 7, 1966, is an academic, researcher, and educator who is active in the development of science, especially in the field of Information Technology. He serves as a lecturer at Universitas Putra Indonesia "YPTK" Padang, with NIDN 0005076607 and can be contacted via email at sumijan@upiyptk.ac.id.

	**Dr. Rini Sovia, S.Kom, M.Kom** was born in Solok, April 5, 1976. Lecturer at Universitas Putra Indonesia YPTK Padang. He completed his education in informatics management in 1999 and his master's degree in informatics engineering in 2006. He is currently earning a doctorate in information technology in 2023. His areas of expertise include artificial intelligence (AI), expert systems (ES), data mining (DM), decision support systems (DSS), and databases. He can be reached at email: rini_sovia@upiyptk.ac.id