Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 2 April 2026, pp: 475-482
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Implementation of Machine Learning to Predict The Timeliness of Graduation of Employees on Study Assignment at Company X

Parenda Rizkya Permata[1*], Imam Yuadi[2]

[1] Master of Human Resource Development, Airlangga University, Indonesia
[2]Department of Information and Library Science, Airlangga University, Indonesia

E-Mail: [1]parenda.rizkya.permata-2025@pasca.unair.ac.id, [2]imam.yuadi@fisip.unair.ac.id

Received Dec 26th 2025; Revised Feb 14th 2026; Accepted Feb 28th 2026; Available Online Apr 18th 2026
Corresponding Author: Parenda Rizkya Permata

Abstract

The energy transition requires workers in the energy sector who have relevant skills that can be applied in the future. Company X implements a study assignment program to improve its employees' skills, but delays in completing their studies hinder their readiness to enter the workforce. Identifying the factors that influence graduation timeliness can improve the program's effectiveness. This study aims to develop a predictive model to determine whether employees in Company X's work-study program will graduate on time. The main purpose of this model is to provide early warnings about employees at risk of delays, enabling more targeted interventions to improve human resource management. We applied the CRISP-DM framework and used Machine Learning to analyze data from 317 employees who participated in the study program. Four machine learning algorithms were tested, namely Gradient Boosting, Decision Tree, Random Forest, and Naive Bayes. 17 factors were trained to cover academic, demographic, and administrative aspects to predict timely graduation. Among the algorithms tested, Gradient Boosting showed the best performance with an AUC of 0.956 and an accuracy of 0.909. These results were supported by high ROC and confusion matrix values, indicating the model's excellent predictive ability. The developed model serves as an early warning system to identify employees at risk of graduating delayed. This system enables Company X to conduct more focused interventions and improve human resource management. These findings show that applying machine learning can significantly improve the management of corporate learning programs and support workforce readiness for the energy transition.

Keywords: Energy Transition, Graduation Timeliness, Educational Data Mining, Machine Learning, Study Assignment.

1. INTRODUCTION

Because of the global energy revolution, the energy business has changed a lot. The energy sector needs to teach its workers the skills they will need in the future, such as how to add renewable energy sources, use low-carbon technology, and make power systems digital [12]. Company X in Indonesia is starting a program for study assignments as part of its long-term goal to help its employees strengthen their skills. Going to school makes it more important to produce graduates who are not only academically intelligent but also have the skills to compete in the global market [9]. But the project won't work if the investigations aren't finished on time. If the energy transition agenda takes too long, it may result in a shortage of skills [3].

Conversely, research shows that educational data mining can be used to predict academic achievement. There is evidence that algorithms such as Gradient Boosting, Decision Trees, Random Forests, and Naive Bayes can model graduation patterns and detect academic risk early [18][20]. Using Company X's study assignment data, consisting of 317 employees, 17 factors were trained using covering academic, demographic, and administrative aspects to predict timely graduation.

How data mining can be used to predict academic achievement [4], while examining the use of Gradient Boosting in modeling academic graduation and detecting the risk of failure [18]. Additionally, utilizes Naive Bayes to identify potential academic failure early on [20]. Research highlights the importance of human capital in supporting energy transition, which is relevant to Company X's need to ensure a skilled workforce [13]. On the other hand, the other study also explains the challenges of human capital indicators, which are much more comprehensive than traditional indicators, including aspects of innovation, education, and skills required in the clean energy economy, which are relevant for building a foundation of workforce competencies in the energy transition sector [16].

This research has contributed to the use of machine learning algorithms in human resource management in the energy sector, specifically to predict when Company X's employees will finish their study assignment programs at the company. The developed methodology is advantageous not only in educational settings but also provides solutions for the energy sector in training a workforce for the transition to renewable energy. By using company-specific data and the latest techniques in machine learning, this research is expected to create an early warning system that can help management plan human resource development more efficiently. This distinguishes this research from previous studies, which focused more on education in general, by linking human resource management directly to the needs of the rapidly growing energy industry.

2. MATERIALS AND METHOD

This study divides the stages into Business Understanding, Data Preparation, Modeling, and Model Evaluation [10], as shown in Figure 1.

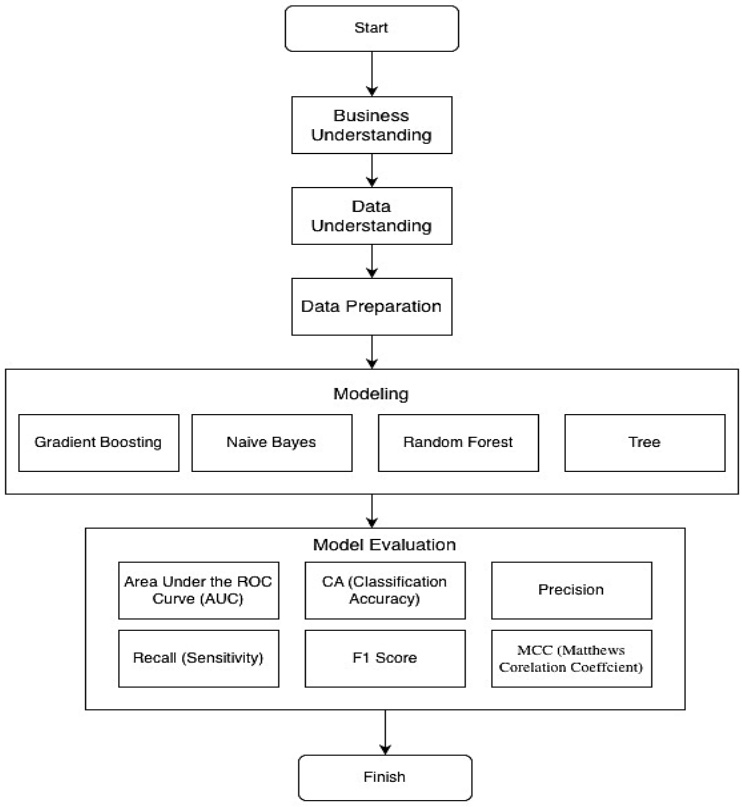

Figure 1. Metodhology Flow Chart

2.1. Business Understanding

Business understanding is a state of business where assessments are conducted to obtain an overview of available and required resources [5]. Company X, in embracing the energy transition, requires ownership of the requirements to ensure adequate digital, technical, and green skills competencies [24]. The Study Assignment Program is gaining popularity, not only as a means for individuals to acquire knowledge, but also as a method for corporations to invest in their future workforce and skill sets. Contemporary human resource management literature emphasizes that structured learning interventions aligned with long-term strategic objectives significantly enhance organizational competitiveness and adaptability [7]. Consequently, within a data-driven HR analytics framework, Study Assignment Programs can be regarded as strategic tools for methodically cultivating future talent and capabilities in alignment with organizational transformation objectives. However, during its implementation, on-time graduation rates continue to fluctuate, leading to the following expected objectives of this study:

1. Supporting human resource planning for the needs of new and renewable energy management.
2. Determining the estimated likelihood of employees on study assignments graduating on time or being delayed.
3. Providing recommendations to management for detecting potential employees on study assignments who are at risk of graduating late.

2.2. Data Understanding

Data understanding includes data collection, exploration, and description, as well as data quality checking [5]. The data training and testing used in this study includes 317 employees on study assignment from Company X. To perform data analysis, data is needed in the form of basic employee data and other attribute data commonly used as a reason for study assignment of 17 features, including study assignment cohort, previous educational background, profession before study assignment, position before study assignment, field of study, subfield of study, on-mission study, study assignment level, university location, scholarship source, gender, employee age, employeement period, study completion target, graduation status, and punctuality, were obtained from secondary data from Company X.

Three key ideas were used to make the choice. First, prior studies in educational data mining and human resource analytics have recognized academic background, demographic characteristics, and institutional factors as significant predictors of academic completion and performance [18]. Second, organizational HR policy considerations suggest that tenure, job role, and funding sources are significant variables in workforce planning and academic task oversight [7]. Third, we conducted several preliminary exploratory analyses and feature relevance assessments to ensure that each selected variable provided meaningful information for the prediction task, in line with the procedures recommended in applied machine learning [17].

2.3. Data Preparation

The data preparation stage is the most crucial in research because it determines whether poor data quality can be addressed through data cleansing [5]. In preparation, it is necessary to examine and clean inappropriate and non-standard data (ensuring that there is no missing or empty data and that the data is consistent). Considering the relatively limited size of the dataset, this study uses the entire dataset (n = 317) to develop and evaluate the model through a 10-fold cross-validation approach rather than applying a fixed train-test split. Next, the target variable, namely timeliness, is selected. After determining the target variable, the next step is to preprocess the data by converting categorical variables into numeric form to ensure compatibility with the algorithm from 17 features, including study assignment cohort, previous educational background, profession before study assignment, position before study assignment, field of study, subfield of study, on-mission study, study assignment level, university location, scholarship source, gender, employee age, employeement period, study completion target, graduation status, and punctuality, were obtained from secondary data from Company X.

2.4. Modeling

The data modeling stage consists of selecting a modeling technique, creating test cases, and developing the model [5]. In this study, modeling was performed using Orange Data Mining by comparing four classification algorithms commonly used in educational data mining/learning analytics for performance and graduation prediction: Gradient Boosting, Naïve Bayes, Random Forest, and Decision Tree. Several recent studies have shown that tree-based and ensemble models often provide high performance on academic data because they are considered to be able to capture non-linear relationships and interactions between features; for example, demonstrated that tree-based algorithms (Decision Tree/Random Forest) are competitive for academic performance prediction, while boosting often excels on data with complex structures [15] [23]. Model testing was conducted using Orange Data Mining, with five models based on previous research (Figure 2).

1. Gradient Boosting

   Gradient boosting is a kind of ensemble boosting that develops models one at a time. Each weak learner, typically a shallow decision tree, is added to the last model to address its flaws. A recent study shows that the family of boosting algorithms often performs better, as it can reduce bias while maintaining generalization through iterative learning and appropriate regularization and hyperparameter settings [3]. Therefore, Gradient Boosting is used because it is more effective when delay patterns are influenced by a combination of heterogeneous factors (individual, academic, organizational) that are interactive and non-linear.

2. Naive Bayes

   Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem that assumes each feature is independent. Because it is very easy and fast to process, this supervised learning model is widely used in Educational Data Mining (EDM) [20]. The results of this study support applying the Naive Bayes variant for academic performance classification and comparing it with a tree-based model [11].

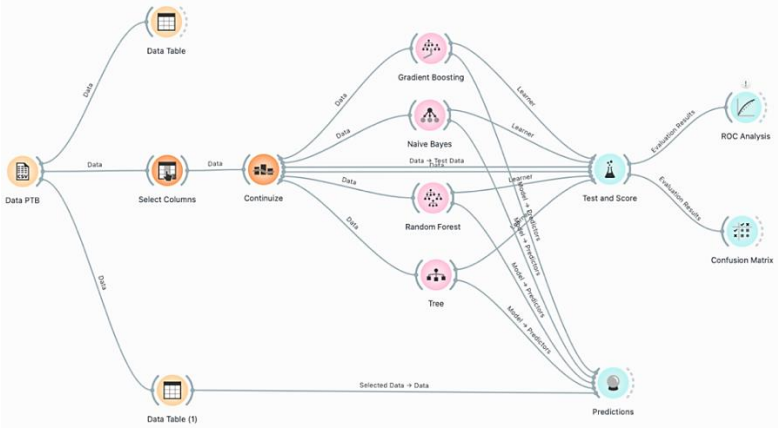

Figure 2. Design Model

3. Random Forest
   Random Forest is a bagging ensemble method that makes several decision trees by taking bootstrap samples and choosing random features at each node split (feature randomness). In studies that try to predict how well students will do in school, this method usually works well and is resistant to noise. This is because voting from several trees makes the results less variable and less likely to fit too closely to the data [15][23]. This model is used in educational data mining is considered to be able to provide explanations and proves to be one of the best performing models [18]. This advantage is relevant for timely prediction cases, because the features used are a mixture of categorical and numerical, and have the potential to interact in complex ways.

4. Decission Tree
   Decision Trees model decision processes by using tree structures that divide data into nodes based on certain criteria for separation. Decision Trees are a common choice in learning analytics because they are easy to understand and help find the most important factors that affect results. However, if not managed correctly (for example, by pruning or limiting depth), they can be too accurate [15][23]. This definition is supported by research because the algorithm can be easily interpreted to predict educational performance [4]. It is still important for this model to provide interpretations that are easy for management to understand (e.g. nodes/rules that indicate combinations of attributes that increase the risk of delay).

2.5. Model Evaluation

The test and inspection results align with the company's objectives. As a result, these findings require further interpretation and action [5]. Stratified sampling was used to maintain the class distribution in the dataset, helping reduce bias in the sample. Next, the dataset was broken up into two parts: one for training and one for testing. Next, k-fold cross-validation is used during model training to prevent overfitting and provide a better idea of how well the model performs across different datasets. Cross-validation is recommended in machine learning research because it mitigates the uncertainty associated with a single hold-out split and enhances the reliability of comparative model evaluation [21].

We chose the assessment metrics deliberately so they would show different aspects of how well the model works. The metrics are AUC, MCC, F1-score, recall, and precision. When the classes are different, accuracy alone might not be enough to tell the whole story. But accuracy and recall show you how each class fails to make good predictions. The evaluation is better with MCC because it considers all four parts of the confusion matrix and remains correct even when the class distributions aren't equal [6]. We used ROC-AUC to assess how well we could distinguish between things without setting a threshold. Recent studies suggest that AUC should be examined alongside metrics obtained from confusion matrices. This is because similar AUC values can mask very different false-positive and false-negative profiles at different thresholds [8] [14]. Using both cross-validation and multi-metric assessment makes the process more reliable, less likely to overfit, and the resulting prediction model more useful for decision-making in business.

3. RESULTS AND DISCUSSION

3.1 Results

Data processing and classification model evaluation were performed, with the results presented in the Evaluation Results section (Table 1). Based on the test results in Table 1, Gradient Boosting scored highest among the tested algorithms, with AUC (0.956), CA (0.909), F1-Score (0.904), Precision (0.907), Recall (0.909), and Matthews Correlation Coefficient (0.726). Based on these results, gradient boosting is the best

supervised learning model for classifying the punctuality of study completion among employees on study assignments.

Table 1. Evaluation Result

Model	AUC	Accuracy (CA)	F1-Score	Precision	Recall	MCC
Gradient Boosting	0.956	0.909	0.904	0.907	0.909	0.726
Decision Tree	0.821	0.902	0.899	0.899	0.902	0.709
Naive Bayes	0.738	0.716	0.738	0.808	0.716	0.413
Random Forest	0.811	0.798	0.775	0.775	0.798	0.338

Confusion matrix In the context of human resource decision-making, the interpretation of these classification errors has different strategic implications. False negatives mean that employees who actually pass on time are predicted to be delayed. This condition can encourage management to take unnecessary interventions, such as additional monitoring or special coaching, which can create administrative burdens and foster distrust towards employees. This can affect efficiency, but the risk is relatively low because it does not entail a direct loss of investment in education. Conversely, a false positive indicates that an employee who is actually delayed is predicted to graduate on time. From an HR management perspective, this error is more critical because the organization fails to intervene early, which can result in extended study periods, increased scholarship costs, disruption to workforce planning, and delays in achieving the strategic competencies required by the organization. Therefore, in the context of study programs as part of human capital investment, reducing prediction errors for employees at risk of being delayed is a top priority. This analysis shows that model evaluation should not only consider accuracy but also weigh the policy consequences of each type of classification error to support data-driven decision-making in human resources. Here are the results of the confusion matrix test for each prediction model, can view Figure 3.

		Predicted		
		OnTime	Delayed	∑
Actual	OnTime	170	75	245
	Delayed	15	57	72
	∑	185	132	317

Figure 3. Naïve Bayes testing results for the Confusion Matrix

Figure 3 illustrates that Naïve Bayes is good at forecasting on-time graduation but less effective at predicting delayed workers [20], as evidenced by the False Positive rate of 75.

		Predicted		
		OnTime	Delayed	∑
Actual	OnTime	229	16	245
	Delayed	48	24	72
	∑	277	40	317

Figure 4. Random Forest testing results for the Confusion Matrix

Figure 4 shows that Random Forest is highly accurate at predicting on-time graduation but has shortcomings in detecting tardiness [18], as reflected in a high False Negative Ratio (48).

		Predicted		
		OnTime	Delayed	∑
Actual	OnTime	235	10	245
	Delayed	21	51	72
	∑	256	61	317

Figure 5. Tree testing results for the Confusion Matrix

Figure 5 shows that tree testing is highly accurate in predicting both classes (on-time and delayed) and can predict tardiness well. These results are effective for educational programs with content considered very clear [1].

		Predicted		
		OnTime	Delayed	∑
Actual	OnTime	239	6	245
	Delayed	23	49	72
	∑	262	55	317

Figure 6. Gradient Boosting testing results for the Confusion Matrix

Figure 6 shows that the Gradient Boosting model performs best, consistently predicting employees who graduate on time and those who are delayed [18].

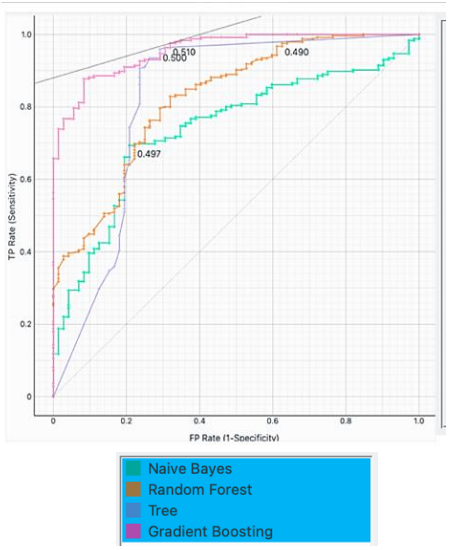

Figure 7. ROC Analysis (On-Time)

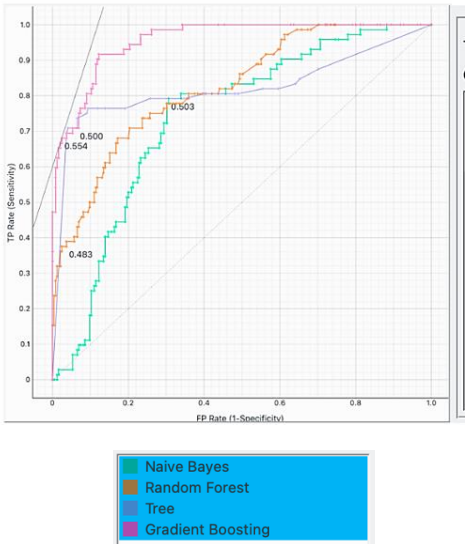

Figure 8. ROC Analysis (Delayed)

Figures 7 and 8 illustrate the ROC analysis, which indicates that each model has a very different capacity to distinguish between things. When the classification aim was "On-Time" or "Delayed," the Gradient Boosting model produced ROC curves closest to the ideal point (high TPR and low FPR). Ensemble learning literature demonstrates that boosting-based algorithms consistently achieve higher AUC values due to their iterative residual minimization process and ability to model complex nonlinear interactions among features [3]. Therefore, the superior ROC performance of the Gradient Boosting model in this study indicates its stronger ability to accurately identify employees at risk of delayed graduation while minimizing misclassification errors, making it the most suitable model for predictive decision support in the Study Assignment Program.

3.2 Discussion

Gradient Boosting is a type of ensemble learning that uses boosting techniques to correct errors made by previous models and improve model performance. Over time, these models make fewer errors, resulting in greater accuracy and stability. Recent studies show that boosting algorithms are superior at finding difficult correlations between variables and non-linear patterns, especially in multidimensional and structurally complex social and organizational datasets. Boosting algorithms such as Gradient Boosting and XGBoost consistently outperform linear models and single-tree-based methods in various classification challenges. This can reduce bias while keeping the data range within acceptable limits [3]. The timeliness of graduation from a study assignment is not a single factor but rather the result of a dynamic interaction among individual characteristics (such as academic achievement, age, and work experience), organizational factors (supervisory support and learning assignment policies), and job factors (workload and job position). The relationship between these variables tends to be non-linear and interacts in complex ways. Therefore, boosting-based models are more adaptive and capable of representing the structure of these relationships than linear models such as Logistic Regression, which assume a linear relationship between variables. The advantages of Gradient Boosting in this study indicate that predicting the timeliness of employee graduation requires an approach that is capable of capturing the complexity of the organizational system as a whole.

In the context of HR analytics-based decision making, the evaluation of predictive models should not only focus on accuracy but must also consider the strategic consequences of classification errors, particularly false positives and false negatives. False positives when employees on study leave are predicted to experience delays in graduation but actually graduate on time can lead to unnecessary interventions, waste of organizational resources, and potential managerial bias against employees. Conversely, false negatives when employees are predicted to graduate on time but actually experience delays have more serious implications because they can lead to failed early intervention, increased study costs, disruption to workforce planning, and decreased effectiveness of HR development investments. Other research on the concept of human resource analytics emphasizes that predictive models must be evaluated by considering cost-sensitive decision-making because the impact of prediction errors can affect organizational strategic policies [2]. Therefore, in the context of employee learning tasks as part of human capital investment, models with high

recall capabilities for the risk of delays are more strategically valuable than models with high overall accuracy.

The role of research skills and institutional support in promoting on-time graduation [19]. While these studies provide important insights into academic determinants, the context of employee learning assignments remains relatively underexplored, even though this group has distinct characteristics such as remaining tied to the organization, having career consequences, and directly impacting human resource planning and investment. Therefore, this study expands the literature by integrating the organizational context and a predictive modeling approach as the basis for an early warning system [22], thus providing a more applicable contribution to managing learning assignments as part of an organization's human capital investment strategy.

4. CONCLUSION

This study demonstrates that using Orange Data Mining, based on the CRISP-DM framework, can provide a predictive model for assessing the probability of on-time graduation among employees enrolled in a corporate study-assignment program at Company X. Using the gradient boosting approach outperformed all three other tested machine learning algorithms, with 90.9% accuracy and better ability to distinguish between different data types. This demonstrates that the approach can effectively handle multidimensional academic, demographic, and administrative data.

The results demonstrate that predictive analytics can be useful for human resource management. The proposed technique not only classifies graduation outcomes but also enables managers to predict potential delays using data, enabling them to act early. In the context of the energy transition, where workforce preparation and timely competency development are crucial, integrating machine learning into talent development programs enables more proactive, evidence-based decision-making. This study integrates educational data mining with HR analytics in a corporate context, specifically in the strategically critical energy industry.

But this study has a limitation. The dataset consists of only 317 individuals from a single organization, potentially constraining its relevance to other contexts. Second, most of the parts used concern work and school. There is no proof of involvement in behavior, psychology, or learning. Third, this study focuses on evaluating model performance, neglecting the assessment of real-time operational implementation and the measurement of long-term organizational impact. Future research could include additional predictive variables, such as performance assessment data, digital learning behavior, mentoring intensity, or psychological indicators, to potentially improve prediction accuracy. Furthermore, additional research could investigate the integration of predictive models into corporate HR information systems and evaluate their long-term impact on graduation rates, cost efficiency, and competency readiness within the broader context of the energy transition.

REFERENCES

[1] A. Tharwat, “Classification assessment methods,” in *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2021. doi: 10.1016/j.aci.2018.08.003.

[2] A. Margherita, “Human resources analytics: A systematization of research topics and directions for future research,” in *Human Resource Management Review*, vol. 32, no. 2, 100795, 2022. doi: 10.1016/j.hrmr.2020.100795.

[3] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” in *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021. doi: 10.1007/s10462-020-09896-5.

[4] C. Romero and S. Ventura, “Educational Data Mining and Learning Analytics,” in *WIREs Data Mining and Knowledge Discovery*, 2020. doi: 10.1002/widm.1355.

[5] C. Schröer et al., “A Systematic Review on Applying CRISP-DM Process Model,” in *Procedia Computer Science*, vol. 181, pp. 526–534, 2021. doi: 10.1016/j.procs.2021.01.199.

[6] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” in *BMC Genomics*, vol. 21, 2020. doi: 10.1186/s12864-019-6413-7.

[7] D. Minbaeva, “Disrupted HR?,” in *Human Resource Management Review*, vol. 31, no. 4, 100820, 2021. doi: 10.1016/j.hrmr.2020.100820.

[8] E. Richardson et al., “The receiver operating characteristic curve accurately assesses imbalanced datasets,” in *Patterns*, vol. 5, no. 6, 100994, 2024. doi: 10.1016/j.patter.2024.100994.

[9] F. Lukitasari, T. Wrahatnolo, A. Wardhono, A. Yushila, H. Muyasyaroh, A. Azizah, M. Shahuddin, and L. Nurlaela, “Employability Skills: A Systematic Literature Review on Skills, Challenges, and Curriculum Integration in Higher Education,” in E3S Web of Conferences, vol. 645, 06007, 2025. doi: 10.1051/e3sconf/202564506007.

[10] F. Martínez-Plumed et al., “CRISP-DM Twenty Years Delayedr: From Data Mining Processes to Data

Science Trajectories," in *IEEE Transactions on Knowledge and Data Engineering*, 2021. doi: 10.1109/TKDE.2019.2962680.
[11] F. Soheli, "Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm," in *Procedia Computer Science*, vol. 194, pp. 224–228, 2021. doi: 10.1016/j.procs.2021.10.077.
[12] International Energy Agency, World Energy Employment 2024, IEA, Paris, 2024. [Online]. Available: https://www.iea.org/reports/world-energy-employment-2024.
[13] International Renewable Energy Agency, Renewable Energy and Jobs: Annual Review 2023, IRENA, Abu Dhabi, 2023. [Online]. Available: https://www.irena.org/Publications/2023/Sep/Renewable-Energy-and-Jobs-Annual-Review-2023
[14] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification," in *PLOS ONE*, vol. 19, no. 12, e0316019, 2024. doi: 10.1371/journal.pone.0316019.
[15] K. Nahar et al., "Mining educational data to predict students' performance," in *Education and Information Technologies*, vol. 26, pp. 6051–6067, 2021. doi: 10.1007/s10639-021-10575-3.
[16] M. Klonowska, "Human Capital and the Sustainable Energy Transition: A Socio-Economic Perspective," in *Sustainability*, vol. 17, 10710, 2025. doi: 10.3390/su172310710.
[17] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed., Chapman and Hall/CRC, 2019. doi: 10.1201/9781315108230.
[18] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," in *Smart Learning Environments*, vol. 9, no. 11, 2022. doi: 10.1186/s40561-022-00192-z.
[19] P. Muthukrishnan et al., "Key Factors Influencing Graduation on Time Among Postgraduate Students: A PLS-SEM Approach," in *Asian Journal of University Education*, vol. 18, no. 1, p. 51, 2022. doi: 10.24191/ajue.v18i1.17169.
[20] S. O. Oppong, "Predicting Students' Performance Using Machine Learning Algorithms: A Review," in *Asian Journal of Research in Computer Science*, vol. 16, no. 3, pp. 128–148, 2023. doi: 10.9734/AJRCOS/2023/v16i3351.
[21] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," in *arXiv preprint arXiv:1811.12808*, 2018.
[22] T. Abd El-Hafeez and A. Omar, "Student Performance Prediction Using Machine Learning Techniques," 2022. doi: 10.21203/rs.3.rs-1455610/v1.
[23] W. Zhang, Y. Wang, and S. Wang, "Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in China," in *Education and Information Technologies*, vol. 27, no. 9, pp. 13051–13066, 2022. doi: 10.1007/s10639-022-11170-w.
[24] World Economic Forum, The Future of Jobs Report 2023, World Economic Forum, Geneva, Switzerland, 2023. [Online]. Available: https://www.weforum.org/reports/the-future-of-jobs-report-2023.