International Journal of Engineering and Computer Science Applications (IJECSA)
Vol. 4, No. 1, March 2025, pp. 51∼58
ISSN: 2828-5611
DOI: doi.org/10.30812/ijecsa.v4i1.4983

Prediction of Student Major Selection at High School Using a Machine Learning Approach

Nasril Sany[1], Dody[1], Esa Firmansyah Muchlis[1], Muhaimin Hasanudin[2], Budi Berlinton[3]
[1]Institut Teknologi PLN, Jakarta, Indonesia
[2]Universitas Mercu Buana, Jakarta, Indonesia
[3]Universitas Multimedia Nusantara, Tangerang, Indonesia

Article Info

Article history:

Received March 13, 2025
Revised March 16, 2025
Accepted March 25, 2025

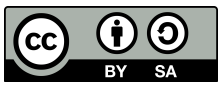
Keywords:

Machine Learning
Major High School
Prediction of Student

ABSTRACT

The primary objective of this research was to develop and evaluate a machine learning prediction system that matches Senior High School (SMA) Nusa Putra Kota Tangerang students with their potential school majors based on their academic interests and performance levels. This research method employs machine learning algorithms, including Random Forest, Support Vector Machine (SVM), logistic regression, K-Nearest Neighbor (K-NN), and Naïve Bayes, using academic records, interest tests, and questionnaires for data collection. The data was processed and analyzed to train and test the algorithm. The findings of this study indicate that the Random Forest algorithm achieved the best performance among the models, with an accuracy of 85%, a precision of 82%, a recall of 88%, and an AUC score of 0.92. The factors that affected the prediction of major selection were Grade XII Mathematics scores and Science Interest Test results. The research implications suggest that Random Forest technology within Machine Learning (ML) enhances major selection accuracy while promoting fairness, providing superior educational choices, and increasing student satisfaction. Future studies should investigate additional factors that influence this phenomenon.

Copyright ©2025 The Authors.
This is an open access article under the CC BY-SA license.



Corresponding Author:

Muhaimin Hasanudin,
Informatics Engineering, Universitas Mercu Buana, Indonesia,
Email: muhaimin.hasanudin@mercubuana.ac.id

How to Cite: N. Sany, D. Dody, E. F. Muchlis, M. Hasanudin, and B. Berlinton, "Prediction of Student Major Selection at High School Using a Machine Learning Approach," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 4, no. 1, pp. 51-58, Mar. 2025. doi: 10.30812/ijecsa.v4i1.4983.

1. INTRODUCTION

Senior High School (SMA) Nusa Putra in Tangerang City conducts a rigorous selection process for its high school students, which shapes their future academic and professional paths. Teachers, along with school counselors, base their major selection decisions traditionally on multiple objective measures consisting of academic grades, student interests, and field study potential [1]. The current approach to major selection produces improper placements, so students face academic problems while losing their motivation and might even choose to leave school [2]. The absence of standard and reliable major selection criteria necessitates the immediate implementation of scientific methods to enhance educational quality and improve student satisfaction. Scientific research teams have developed different approaches to tackle major selection obstacles [1]. Developed career counseling through trait and factor theory for high school major selection, yet their model depends significantly on personal opinions as a rating method [3–5]. The research conducted by [5] examined students' perception of objective structured practical examinations (OSPE) in anatomy while establishing the significance of student interest alignment in academic assessments but failed to develop a major selection prediction model [6]. Cooperative strategies [7] for student Arabic learning potential growth while his work did not include major selection analysis [7]. Digital escape rooms as an anatomy teaching method in veterinary medicine, although their study failed to predict major selection patterns [8].

The educational implementation of machine learning technologies makes accurate predictions about university graduation outcomes as well as student academic performance assessment [9]. These studies exclusively use individual algorithms, but this practice restricts the universal applicability and reliability of their research because of insufficient algorithm comparison tests [10]. Applied research is lacking in utilizing machine learning systems that combine academic results with student interest indicators to generate precise high school major choices. Over the last five years, science journals have shown a growing interest in applying machine learning and artificial intelligence to support educational choice. Predictive analytics serves as a field of examination in Computers & Education and Journal of Educational Data Mining publications for studies about outcomes enhancement and educational pathway enhancement [9]. Educational Psychology Review published research that examines how psychological assessments should pair with data-driven tools for enhancing major selection processes [11, 12]. The advancement of predictive analytics needs to be accompanied by extensive research focused on specific barriers that Indonesian high school students encounter during their major selection process because of cultural factors.

Several investigations in machine learning have demonstrated the predictive value of models in educational institutions. Implemented Random Forest algorithms to build a predictive model [13] that rates Indonesian university students' chances of graduating on time [13]. The researchers at [14] utilized Fuzzy C-Means together with K-Nearest Neighbors (K-NN) to forecast on-time graduation, which indicates why clustering [15] approaches matter for educational data research. The evaluation and assessment processes within mathematics education gained support through machine learning applications that focused on data-driven decision-making [16]. AI technology for organizational knowledge management shows the potential of AI for decision-making even though they are not directly related to education [17]. The existing research has not addressed the central gap in high school major selection through machine learning models, which would fuse academic performance data with student interest factors [11, 18]. Current research typically focuses on university predictions or specific aspects of student performance, often neglecting the comprehensive combination of student capabilities and preferences. Multiple studies present their predictions using single algorithms, which reduces their generalization capabilities and makes the results less robust due to the lack of comparative assessments.

The chief novelty of this research consists of building a comparative machine learning model that analyzes different algorithms such as Random Forest and Support Vector Machine (SVM) along with logistic regression and K-NN and Naïve Bayes to forecast high school students' majors based on their academic performance and interests [19]. The goal of this research is to develop an unbiased, advanced decision-making tool for major choices, aiming to enhance educational quality and student satisfaction. The objectives of this research are twofold: first, to develop and evaluate a machine learning model that can accurately predict students' majors based on their interests and academic performance, and second, to compare the effectiveness of various machine learning algorithms in achieving this goal. The contribution of this research to the development of science lies in its potential to provide a more systematic and data-driven approach to major selection, which can be generalized to other educational contexts. By addressing the limitations of previous studies and offering a novel comparative analysis of machine learning algorithms, this research aims to enhance the decision-making process in high schools, ultimately leading to better educational outcomes for students.

2. RESEARCH METHOD

2.1. Research Design

This research employs a numerical quantitative model to evaluate various machine learning frameworks for forecasting student major choices at SMA Nusa Putra. Data collection for the research begins with academic performance records from grades X to

XII, along with interest test results from standardized measures and responses to interviews and surveys. The workflow also includes existing major information stored in the school database. The Student Records dataset comprises 500 entries, each containing Mathematics marks, Science Interest Test scores, English grades, and subject selection preferences. Academic performance values form one part of the data collection while recorded interests and selected majors comprise the other part. The data preprocessing process comprises three main stages: cleaning to handle missing values and outliers, transforming scaled variables, encoding categorical features, and performing feature selection through recursive feature elimination. Data preparation leads to a division that creates training data (70%) and testing data (30%). The study evaluated Random Forest, Logistic Regression, SVM, Naïve Bayes, and K-Nearest Neighbor (K-NN) using Python, its Pandas library, and Scikit-learn tools. The evaluation methods for model performance include accuracy, precision, recall, F1-score, and AUC, and a cross-validation procedure checks model robustness. Major prediction modeling requires a comparison of the achieved results to select the optimum algorithm, as shown in Figure 1.

Figure 1. Effects of selecting different switching under dynamic condition

2.2. Data Collection

The research concentrated its data collection activities at SMA Nusa Putra in Tangerang City and analyzed academic data alongside interest data and major data. Report card scores for grades X, XI, and XII served as the academic data that integrated examination results with assignments and attendance in academic activities. The scores underwent a numeric conversion process to make them suitable for machine learning system processing. The collection of interest data was conducted through standardized interest and talent tests, as well as structured interviews and questionnaires. Students received numeric scores on their fields of study interest through the interest and talent tests. Still, in-depth interview sessions enabled them to express their preferences and interests more thoroughly. The collection of student data about their chosen courses involved Likert-type questionnaires which captured their favorite subjects among others. Actual majors selected by students, along with their academic results in those subjects, were retrieved as part of the major data from the school's database. The gathered information served as a reference material for assessing the accuracy of the machine learning system. Data quality improvement was achieved by cleaning the collected data and performing initial processing where missing data received mean substitution imputation and outliers received Z-score analysis treatment [20]. Existing databases were consulted to solve data inconsistencies through both cross-referencing methods and standardization techniques, which standardized all datasets. The multi-step data aggregation method established a robust database that enabled accurate major predictions through the training and testing processes of a machine learning model. The collected data will be cleaned and processed to ensure data quality and consistency before being used in training and testing machine learning models. Missing data will be handled using mean substitution imputation techniques while outliers will be detected and managed through Z-score analysis.

Additionally, data inconsistencies will be resolved by cross-referencing with existing databases and using standardization techniques to ensure uniformity across all datasets.

2.3. Data Processing

The expected data must undergo several processing stages before reaching a suitable state for machine learning applications. Data cleaning [21] begins by handling both missing and inconsistent values in the first step. Mean or median substitution methods will address cases of missing data alongside outlier and inconsistent data, which will be identified and then treated by suitable methods such as removal or transformation. After this phase, the data requires transformation to become suitable for usage with machine learning tools. Data transformation occurs through data normalization, after which categorical variables are encoded using one-hot encoding, and principal component analysis is employed to minimize dimensionality. Feature selection begins by identifying important variables that will be used for prediction. The prediction accuracy-determining features will be identified by utilizing recursive feature elimination or machine learning model feature importance methods. The data will be divided into separate training and testing sets through a process of data splitting. Machine learning model training utilizes the training data, but the testing data serves to assess its performance. Cross-validation approaches will confirm the model's generalization ability, while both the training and testing datasets will receive thorough consideration to represent the complete dataset.

2.4. Machine Learning Models

The investigation utilizes multiple machine learning algorithms to predict which major students will select at the SMA Nusa Putra learning environment. The implementation of Random Forest analysis will be employed in this research because it performs well in educational prediction tasks and effectively processes non-linear, high-dimensional data structures. Using Logistic Regression as the supervised learning method will analyze major selection by producing results that reveal essential factors in shaping major preferences. The final implementation tool will be SVM for determining optimal hyperplanes that separate data into different classes when processing high-dimensional complex information. Naïve Bayes will be included in this study since it maintains simplicity while delivering acceptable results with high computational efficiency [22]. For simple and efficient data classification on small or large datasets, K-NN serves as a non-parametric approach that determines results by measuring proximity to k neighboring data points. The algorithms demonstrate optimal performance, as they can efficiently handle various data characteristics in complex settings. Comparing various algorithms for predicting students' academic majors at SMA Nusa Putra will result in the identification of effective forecasting methods. To evaluate the performance of the machine learning model, a confusion matrix is used, which includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The accuracy formula uses Equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

3. RESULT AND ANALYSIS

3.1. Machine Learning Model Performance

An illustration of the cleaned and transformed data allows readers to follow the analytical data preparation process, which addresses missing value handling, outlier treatment, and feature selection techniques. Random Forest proved to be the superior scoring model among the other algorithms as it achieved 85% accuracy thus demonstrating a solid performance in managing complex datasets and reducing overfitting risks. The least effective performance by Naïve Bayes indicates the dataset does not match its feature independence requirement. The evaluation of features revealed that major prediction success was largely dependent on Grade XII Mathematics scores and Science Interest Test results. The predictive model worked reasonably well against the academic majors at SMA Nusa Putra, but it presented certain inaccurate outcomes, which requires more research into the selection process of academic paths. After training and testing the machine learning model with SMA Nusa Putra data, we present the performance results for each algorithm in the following table. The data shown are the results of cross-validation with a number of folds, as shown in Table 1.

Table 1. Performance results of each algorithm

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	85%	82%	88%	85%	0.92
Logistic Regression	78%	75%	81%	78%	0.85
SVM	82%	80%	84%	82%	0.88
Naïve Bayes	65%	62%	68%	65%	0.75
K-NN	75%	72%	78%	75%	0.82

The findings of this research are based on Table 1. Random Forest exhibits the best performance among all the algorithms evaluated, achieving an accuracy of 85%. The effectiveness of Random Forest stems from its ability to handle high dimensional and complex datasets while also preventing overfitting. Random Forest also demonstrates strong performance in distinguishing between multiple classes, as indicated by its AUC score of 0.92. Conversely, Naïve Bayes performs the worst, which implies that this dataset may not be independent as suggested. These results confirm previous studies, which indicate that Random Forest outperforms other classification methods in tasks with large feature counts [10, 21, 22]. It is still crucial to note, however, that model performance depends on multiple aspects, such as data quality, the optimization steps applied to the data, and even the choices made during the model's training phase.

A confusion matrix served as the evaluation method to assess the performance of each machine learning algorithm, presenting detailed results on the statistics of true positives, true negatives, false positives, and false negatives. This summary presents the confusion matrices for every applied algorithm. The Random Forest confusion matrix yielded the most precise results, as it contained the highest proportion of correct predictions alongside the lowest proportion of incorrect predictions. The student major prediction success rate of Random Forest reached 85% naming it the most precise algorithm. Logistic Regression produced a confusion matrix that demonstrated a medium level of true positive and true negative observations, together with several misclassification errors. The obtained accuracy rate was 78%, indicating sufficient performance levels; however, Random Forest demonstrated better accuracy. The SVM algorithm delivered true positive and true negative results effectively while producing lower misclassification counts than Logistic Regression. SVM achieved 82% accuracy, ranking it as the second-best among the investigated algorithms. Naïve Bayes experienced a high number of misidentification errors as its confusion matrix showed many false positive and false negative results. The prediction accuracy reached only 65% because the assumption of feature independence did not apply well to this particular dataset. The K-NN confusion matrix displayed moderate true positive and true negative results together with several erroneous classifications. The method achieved a 75% accuracy level, placing it between average and above-average performance results among the tested algorithms.

The results of this research align with or support previous studies that highlight the effectiveness of Random Forest in educational prediction tasks. For instance, research by [9] demonstrated that Random Forest achieved superior results in predicting on-time graduation in Indonesian universities, which aligns with the current study's findings. Similarly, [17] it was found that Random Forest outperformed other algorithms, such as Logistic Regression and K-NN, in text classification tasks, further validating the robustness of Random Forest in handling complex datasets. However, the poor performance of Naïve Bayes in this study contradicts some earlier findings, suggesting that the assumption of feature independence may not apply to this dataset. The best results achieved by the majors covered by students from SMA Nusa Putra were attained using Random Forest, reaching their highest possible achievements. This is because this model offers an approach that provides very high accuracies and precision, as well as F1 scores, in terms of metrics. Some of the reasons why Random Forest proves to be an effective modeling solution are its ability to address variables in multiple dimensions and its avoidance of overfitting issues. Numerous research studies have confirmed that Random Forest has emerged as an outstanding approach due to the great efficacy evident in academic prediction applications. With the involvement of empirical research, [13] proved that Random Forest achieved superior results in predicting the timing of student graduation at Indonesian colleges and universities with comparable accuracy.

This study confirms the findings from previous research, which demonstrate Random Forest delivers exceptional efficiency in educational predictions. Based on empirical research [13] demonstrated that Random Forest produced outstanding results for on-time graduation prediction in Indonesian universities with matching prediction accuracy. A study conducted by [23] proved that the Random Forest algorithm demonstrated superior text classification results beyond other methods, including Naïve Bayes and Logistic Regression. These academic works support this investigation by demonstrating that Random Forest proves to be a dependable and accurate classification algorithm in educational analysis. Research findings from this study can be compared to those of previous studies to enhance the understanding of the effectiveness of machine learning algorithms in educational predictions. This study compared the performances of various algorithms with previously recorded data, as shown in Table 2.

Table 2. Compared algorithm performances

Algorithm	Accuracy		Reference
	This Study	Previous Research	
Random Forest	85%	87%	[13]
Logistic Regression	78%	75%	[23]
SVM	82%	80%	[24]
Naïve Bayes	65%	68%	[23]
K-NN	75%	72%	[13]

Random Forest maintained high accuracy when applied to educational prediction according to the results presented in this study as well as earlier research. The performance metrics of Logistic Regression and SVM have returned to their previous matches with results similar to those of similar research. However, Naïve Bayes and K-NN have shown reduced versions of their performances in this specific study. The assessments indicate that Random Forest will provide consistent performance, albeit robust, across all educational prediction tasks, although exact data collection and environment can sometimes influence algorithm effectiveness. This research confirms that Random Forest stands as the most efficient model for predicting SMA Nusa Putra student majors since it demonstrates success in both the current study and existing precedent findings. The research outcomes demonstrate that machine learning provides more accurate and objective selection methods for high school majors, generating beneficial information for school education authorities and administrators. Additional variables involving social and psychological factors should be assessed in future research to maximize the model's predictive capabilities.

3.2. Feature Analysis

This research examines the importance of features in predicting students' majors using the Random Forest algorithm. The contribution of each feature to the prediction accuracy is evaluated using the Random Forest technique, specifically feature importance. This analysis is presented in Table 3, which ranks the importance of each feature. Such Findings will be presented in the future analysis, where each feature will be examined and its value explained. For example, placing a high weight on the Grade XII Mathematics mark would mean that students' mathematical skills significantly affect the estimation of their predicted major, likely in an area of science or engineering. Inversely, if the Art Interest Test score yields low importance, then students' interest in art has a minimal impact on their predicted major. This analysis aims to assist SMA Nusa Putra in understanding the fundamental factors influencing students' choice of major. It offers guidance on how to enhance the processes of counseling and guidance design. Studying the relevant literature alongside the features created by the machine learning model will shed light on deeper insights concerning the understanding of student majoring processes. Mathematics grade 12, Science Interest Test values, and English grade 11 scores.

Table 3. Feature Analysis

Feature	Value
Grade XII Mathematics scores	0.25
Science Interest Test results	0.20
English Grade XI Scores	0.15
Grade XII Physics Score	0.12
Art Interest Test Score	0.10

3.3. Comparison with Existing Majors

The prediction results from the best machine learning model (Random Forest) will be tested against those for the existing majors at SMA Nusa Putra. This will be done by checking the percentage of agreement between the model prediction and the actual majors. This analysis will provide insight into the accuracy of the machine learning model in predicting a student's major compared to traditional methods of determining a major. The differences between the model prediction and the actual majors will be further studied to identify the factors responsible for the discrepancy. For instance, if the model suggests science as the major for a student but they are placed in the social sciences major, further investigation will be conducted to understand the actual cause of the discrepancy. Further analysis of the student-based data will be conducted, including results from report cards, interest tests, and any other relevant information.

The text presents similarities and differences between machine learning model predictions and existing majors (between objects of measure estimated in this study and those provided by the usual majoring scheme). Where substantial differences exist, the text

further elaborates on the likely reasons for the present findings. For example, the sentence may discuss one reason. That is if the machine learning model achieves much higher accuracy compared to the conventional majoring method. It would mean that the machine learning model can enhance objectivity and accuracy in the major process. On the contrary, if there are significant discrepancies between model predictions and existing majors, this would imply some indeterminable factors or even bias in the data. The sentence would express another, following the other discussion items. It will be useful to make recommendations for improving the majoring process at SMA Nusa Putra and for future research.

4. CONCLUSION

Machine learning models, among which random forests prove the most efficient, cast somewhat accurate predictions for SMA Nusa Putra students' majors based on interests and academic potential. In comparison with other algorithms applied, Random Forest did relatively better. The prediction results of the developed model align with the current majors of SMA Nusa Putra, but with some marginal disagreement, necessitating further analysis. An important feature analysis identified several key determinants of the major prediction, including mathematics grade XII, Science Interest Test scores, and English grade XI scores. Although the Random Forest model can predict relatively well, its predictions are weak because they depend solely on the information provided without accounting for social and psychological aspects, as well as external factors, in the student-choice process. To this end, future studies should focus on integrating these additional variables to study the potential of the model's generalization to other high school groups, such as vocational schools, for tracking the graduation directions of students.

Therefore, SMA Nusa Putra can consider implementing a pathway system based on machine learning so that the results obtained in choosing a pathway are not subjective and accurate. Based on the findings of this study, it can be used to assist teachers and counselors in making more informed and accurate selection decisions. Future research can include a wider scope of data including data from other schools and take into account other factors that may affect students.

REFERENCES

[1] R. Dewany, M. Iswari, and D. Daharnis, "Pendekatan Konseling Karir Trait and Factor dalam Membantu Siswa SMA untuk Memilih Jurusan di Perguruan Tinggi," *Jurnal Bimbingan Konseling dan Psikologi*, vol. 2, no. 2, pp. 113–123, Sep. 30, 2022.

[2] W. F. Irhamny, M. Marsudi, and Z. I. M. Sultani, "Guru Sejarah Sekolah Menengah Kejuruan Negeri (SMKN) di Kota Kediri Terhadap Aksi Sepihak Partai Komunis Indonesia (PKI) di Kediri Sebelum Meletusnya Peristiwa," *Jurnal Artefak*, vol. 9, no. 2, pp. 121–138, Oct. 10, 2022.

[3] K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques," *SN Computer Science*, vol. 2, no. 5, p. 386, Jul. 22, 2021. DOI: 10.1007/s42979-021-00776-5.

[4] I. Mugunga *et al.*, "A Frame-Based Feature Model for Violence Detection from Surveillance Cameras Using ConvLSTM Network," in *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, Qingdao, China: IEEE, Jul. 23, 2021, pp. 55–60. DOI: 10.1109/ICIVC52351.2021.9526948.

[5] M. Hani'ah *et al.*, "Google Trends and Technical Indicator based Machine Learning for Stock Market Prediction," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 2, pp. 271–284, Mar. 31, 2023. DOI: 10.30812/matrik.v22i2.2287.

[6] A. M. Yusof, "Students' Perception on OSPE in Anatomy Subject," *Healthscope: The Official Research Book of Faculty of Health Sciences, UiTM*, vol. 7, no. 1, pp. 54–62, Nov. 1, 2024.

[7] Y. Ahmadi, "Maksimalisasi Potensi Siswa dalam Pembelajaran Bahasa Arab dengan Strategi Kooperatif Kontemporer," *Lisaanuna Ta'lim Al-Lughah Al-Arabiyah: Jurnal Pendidikan Bahasa Arab*, vol. 7, no. 2, pp. 121–139, Sep. 30, 2024. DOI: 10.15548/lisaanuna.v7i2.10857.

[8] S.-Y. Huang, Y.-H. Kuo, and H.-C. Chen, "Applying digital escape rooms infused with science teaching in elementary school: Learning performance, learning motivation, and problem-solving ability," *Thinking Skills and Creativity*, vol. 37, p. 100 681, Sep. 2020. DOI: 10.1016/j.tsc.2020.100681.

[9] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," *Computers & Electrical Engineering*, vol. 89, p. 106 903, Jan. 2021. DOI: 10.1016/j.compeleceng.2020.106903.

[10] Ü. Ağbulut, A. E. Gürel, and Y. Biçen, "Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison," *Renewable and Sustainable Energy Reviews*, vol. 135, p. 110 114, Jan. 2021. DOI: 10.1016/j.rser.2020.110114.

[11] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, p. 3, Dec. 2020. DOI: 10.1186/s41239-020-0177-7.

[12] P. L. Bokonda, K. Ouazzani-Touhami, and N. Souissi, "Predictive analysis using machine learning: Review of trends and methods," in *2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, Marrakech, Morocco: IEEE, Nov. 25, 2020, pp. 1–6. DOI: 10.1109/ISAECT50560.2020.9523703.

[13] M. A. S. Pawitra, H.-C. Hung, and H. Jati, "A Machine Learning Approach to Predicting On-Time Graduation in Indonesian Higher Education," *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 9, no. 2, pp. 294–308, Dec. 2, 2024. DOI: 10.21831/elinvo.v9i2.77052.

[14] S. P. Nabila, N. Ulinnuha, and A. Yusuf, "Model Prediksi Kelulusan Tepat Waktu dengan Metode Fuzzy C-Means dan K-Nearest Neighbors Menggunakan Data Registrasi Mahasiswa," *Network Engineering Research Operation*, vol. 6, no. 1, p. 39, Apr. 19, 2021. DOI: 10.21107/nero.v6i1.199.

[15] M. Yusuf, M. Hasanudin, and I. Prihandi, "Design and Build A Customer-Finding Application For Leko Restaurant Using The K-Means Algorithm," *IJISTECH (International Journal of Information System and Technology)*, vol. 6, no. 2, pp. 270–275, Aug. 8, 2022. DOI: 10.30645/ijistech.v6i2.238.

[16] Y. Liu, D. Zhang, and H. B. Gooi, "Data-driven decision-making strategies for electricity retailers: A deep reinforcement learning approach," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 2, pp. 358–367, Mar. 2021. DOI: 10.17775/CSEEJPES.2019.02510.

[17] M. Ruiz *et al.*, "EZATECH: Design and development of Artificial Intelligence technologies for knowledge management throughout the life cycle of workers in organizations," *Journal of Applied Research in Technology & Engineering*, vol. 6, no. 1, pp. 24–33, 1 Jan. 21, 2025. DOI: 10.4995/jarte.2025.21755.

[18] K. Al Mayahi and M. Al-Bahri, "Machine Learning Based Predicting Student Academic Success," in *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Brno, Czech Republic: IEEE, Oct. 2020, pp. 264–268. DOI: 10.1109/ICUMT51630.2020.9222435.

[19] K. Kristiawan and A. Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, pp. 35–46, Apr. 24, 2021. DOI: 10.28932/jutisi.v7i1.3182.

[20] D. Wu, X. Ma, and D. L. Olson, "Financial distress prediction using integrated Z-score and multilayer perceptron neural networks," *Decision Support Systems*, vol. 159, p. 113 814, Aug. 2022. DOI: 10.1016/j.dss.2022.113814.

[21] F. Neutatz *et al.*, "Data Cleaning and AutoML: Would an Optimizer Choose to Clean?" *Datenbank-Spektrum*, vol. 22, no. 2, pp. 121–130, Jul. 2022. DOI: 10.1007/s13222-022-00413-2.

[22] M. Hasanudin *et al.*, "Isometric Contraction ankle joint in Cerebral Palsy using Naive Bayes," *International Journal of Open Information Technologies*, vol. 12, no. 3, pp. 78–81, 2024.

[23] K. Shah *et al.*, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, p. 12, Mar. 5, 2020. DOI: 10.1007/s41133-020-00032-0.

[24] M. Sheykhmousa *et al.*, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6308–6325, 2020. DOI: 10.1109/JSTARS.2020.3026724.