Original Article

Check for updates

Available online at
www.heca-analitika.com/jeml

Journal of Educational Management and Learning

Vol. 1, No. 1, 2023

Optimizing University Admissions: A Machine Learning Perspective

Aga Maulana [1], Teuku Rizky Noviandy [1], Novi Reandy Sasmita [2], Maria Paristiowati [3], Rivansyah Suhendra [4], Erkata Yandri [5], Justinus Satrio [6] and Rinaldi Idroes [7,*]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; agamaulana@usk.ac.id (A.M.); trizkynoviandy@gmail.com (T.R.N.)
[2] Computational and Applied Statistics Research Group, Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; novireandys@gmail.com (N.R.S.)
[3] Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Jakarta 13220, Indonesia; maria.paristiowati@unj.ac.id (M.P.)
[4] Department of Information Technology, Faculty of Engineering, Universitas Teuku Umar, Aceh Barat 23681, Indonesia; rivansyahsuhendra@utu.ac.id (R.S.)
[5] Graduate School of Renewable Energy, Darma Persada University, Jl. Radin Inten 2, Pondok Kelapa, East Jakarta 13450, Indonesia;
[6] Department of Chemical Engineering, Villanova University, Villanova 19085, United States; justinus.satrio@villanova.edu (J.S.)
[7] School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; rinaldi.idroes@usk.ac.id (R.I.);

***** Correspondence: rinaldi.idroes@usk.ac.id

Article History

Received 11 May 2023
Revised 13 June 2023
Accepted 22 June 2023
Available Online 27 June 2023

Keywords:
Educational management
Machine learning
University admission
Prediction model

Abstract

The university admission process plays a pivotal role in shaping the future of aspiring students. However, traditional methods of admission decisions often fall short in capturing the holistic capabilities of individuals and may introduce bias. This study aims to improve the admission process by developing and evaluating machine learning approach to predict the likelihood of university admission. Using a dataset of previous applicants' information, advanced algorithms such as K-Nearest Neighbors, Random Forest, Support Vector Regression, and XGBoost are employed. These algorithms are applied, and their performance is compared to determine the best model to predict university admission. Among the models evaluated, the Random Forest algorithm emerged as the most reliable and effective in predicting admission outcomes. Through comprehensive analysis and evaluation, the Random Forest model demonstrated its superior performance, consistency, and dependability. The results show the importance of variables such as academic performance and provide insights into the accuracy and reliability of the model. This research has the potential to empower aspiring applicants and bring positive changes to the university admission process.

1. Introduction

University admission is a vital process that greatly impacts the future of aspiring students. With a limited number of spots available and a large pool of applicants, educational institutions face the difficult task of choosing the most suitable candidates. The selection process is crucial as it determines the opportunities and prospects that students will have, shaping their academic and professional paths [1, 2].

Traditionally, admission decisions have been based on standardized test scores, academic records, recommendation letters, and personal statements.

However, this approach often fails to capture the holistic nature of an individual's capabilities and potential. Moreover, the subjective nature of human judgment may introduce bias or inconsistencies in the decision-making process [3, 4].

In recent years, machine learning has emerged as a powerful tool to enhance decision-making processes in various domains, including education [5–9]. By leveraging large datasets and advanced algorithms, machine learning models can analyze complex patterns and make accurate predictions. In the context of university admission, a machine learning model that can predict an applicant's chances of admission based on a diverse set of attributes would be invaluable to students, institutions, and policymakers alike [10].

This study aims to design, evaluate, and compare machine learning models specifically tailored for predicting the chances of university admission. By incorporating a wide range of features such as academic performance, extracurricular activities, and personal attributes, the models aim to capture the multifaceted nature of an applicant's profile. Through rigorous experimentation and evaluation, this study seeks to provide insights into the accuracy, reliability, and potential of different models to revolutionize the university admission process. The performance of the machine learning models is compared, and the model with the best performance is selected for further analysis.

This study showcases the application of machine learning in university admission, demonstrating the development and comparison of models that accurately predict an individual's chance of admission. By analyzing relevant factors and utilizing advanced algorithms like K-Nearest Neighbors, Random Forest, Support Vector Regression, and XGBoost, these models provide valuable insights for aspiring applicants. Furthermore, the flexibility of our approach allows for easy adaptation to fit the data of other universities with potentially different criteria.

The rest of the paper is organized as follows: section 2 presents the dataset used and methodology employed in designing the prediction model. Section 3 presents the results and discussion of the findings. Finally, section 4 concludes the paper by summarizing the contributions, limitations, and future directions of this research.

2. Materials and Methods

2.1. Dataset

The dataset used in this study was originally created for predicting university admissions specifically for UCLA. It was obtained from [11], and comprises several variables that are considered important during the application process for master's programs at UCLA. These variables provide valuable insights into an applicant's profile and serve as input features for the machine learning model. The variables included in the dataset are presented in Table 1.

2.2. Machine Learning Algorithms

In this study, four different machine learning algorithms are employed to predict university admissions. These algorithms include K-Nearest Neighbors, Random Forest, Support Vector Regression, and XGBoost. These algorithms were selected based on their demonstrated high performance in previous studies and applications. The implementation of K-Nearest Neighbors, Random Forest, and Support Vector Regression utilizes the scikit-learn library version 1.2.2, while XGBoost employs the XGBoost library version 1.7.3. The models are trained using the default hyperparameters provided by the respective libraries for each algorithm. The performance of these models is compared to determine which one performs better in predicting admission outcomes.

K-Nearest Neighbors is a non-parametric algorithm that predicts based on the similarity of new instances to the k-nearest neighbors in the training dataset. It assigns the class label of the majority of the k neighbors to the new instance. In this regression context, it uses the average value of the k nearest neighbors [12].

Random Forest is an ensemble learning method that combines multiple decision trees. It creates a collection of decision trees by using a random subset of features and random sampling of the training data. The final prediction is made by aggregating the predictions of individual trees [13, 14].

Support Vector Regression is a regression algorithm that utilizes support vector machines to perform nonlinear regression. It identifies a hyperplane in a higher-dimensional space that maximizes the margin and maps the input data to a higher-dimensional feature space using a kernel function. Support Vector Regression aims to find a function that approximates the training data with a specified error tolerance [15–17].

XGBoost is an optimized gradient boosting algorithm that uses a combination of decision trees and boosting techniques. It iteratively builds decision trees to correct the errors of previous iterations, leading to more accurate predictions. XGBoost employs regularization techniques to control overfitting and handles missing values efficiently [18].

Table 1. Variables in the dataset.

No.	Variable	Definition
1	GRE Scores	Applicant's scores in the Graduate Record Examination (GRE), which is a standardized test measuring verbal reasoning, quantitative reasoning, and analytical writing skills. The scores range from 0 to 340.
2	TOEFL Scores	The Test of English as a Foreign Language (TOEFL) scores indicate the applicant's proficiency in the English language. It is a standardized test commonly required for non-native English speakers. The scores range from 0 to 120.
3	University Rating	The rating or reputation of the university where the applicant completed their undergraduate education. The rating is measured on a scale of 1 to 5, with 5 being the highest.
4	Statement of Purpose	The perceived strength of the applicant's statement of purpose on a scale of 1 to 5, with 5 indicating the highest strength.
5	Letter of Recommendation	The perceived strength of the applicant's letter of recommendation on a scale of 1 to 5, with 5 indicating the highest strength.
6	Undergraduate GPA	The undergraduate Grade Point Average (GPA) of the applicant is an important indicator of their academic performance during their undergraduate studies. The GPA is measured on a scale of 0 to 10.
7	Research Experience	This binary parameter indicates whether the applicant has any prior research experience or not. A value of 1 represents the presence of research experience, while 0 indicates its absence.
8	Chance of Admit	The target variable for prediction, represents the chances of admit range from 0 to 1, with 1 indicating a higher probability of admission.

Table 2. Descriptive statistics of the dataset

Variables	Mean	Std. Dev	Min	Q1	Q2	Q3	Max
GRE Score	316.81	11.47	290	308	317	325	340
TOEFL Score	107.41	6.07	92	103	107	112	120
University Rating	3.09	1.14	1	2	3	4	5
SOP	3.4	1.01	1	2.5	3.5	4	5
LOR	3.45	0.9	1	3	3.5	4	5
CGPA	8.6	0.6	6.8	8.17	8.61	9.06	9.92
Research	0.55	0.5	0	0	1	1	1
Chance of Admit	0.72	0.14	0.34	0.64	0.73	0.83	0.97

2.3. Evaluation Metrics

The study evaluated the performance of each algorithm by considering R-squared (R^2), root mean squared error (RMSE), and mean absolute error (MAE). R^2 measures the proportion of the dependent variable's variance explained by the independent variables, with higher values indicating a better fit. RMSE assesses the average magnitude of differences between predicted and actual values, where a lower value signifies better accuracy. Similarly, MAE measures the average absolute difference, with a lower value indicating improved accuracy. These metrics enable the comparison and selection of the most effective model based on goodness of fit and predictive accuracy. The equations to calculate R^2, RMSE, and MAE presented in Equations 1, 2, and 3, respectively.

$$R^2 = 1 - \left(\frac{SSR}{SST}\right) \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(predicted - actual)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|predicted_i - actual_i| \quad (3)$$

where SSR denotes the sum of squared residuals, representing the sum of the squared differences between predicted and actual values, SST represents the total sum of squares, indicating the sum of the squared differences between the actual values and their mean, and n denotes the number of samples in the dataset.

3. Results and Discussions

3.1. Exploratory Data Analysis

In this study, we conducted exploratory data analysis (EDA) to gain insights into the dataset. Table 2 provides a summary of the statistical measures calculated for each variable in the dataset. These measures include the mean, standard deviation, minimum, first quartile (Q1), second quartile (Q2), third quartile (Q3), and maximum values for each variable.

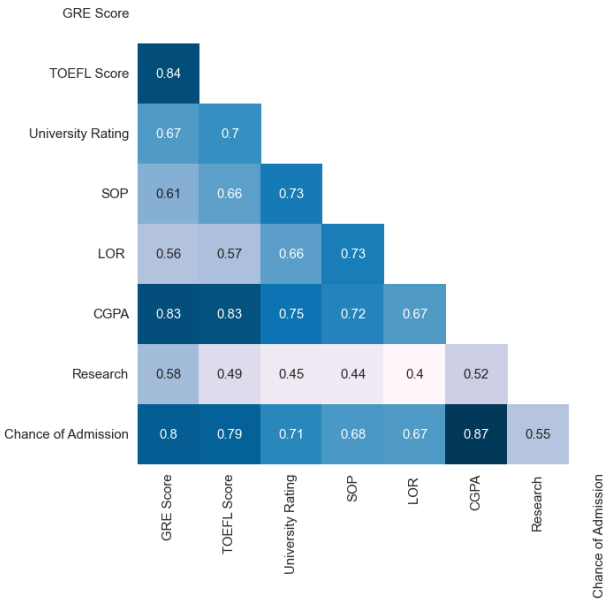

Figure 1. Correlation matrix of each variable to chance of admit.

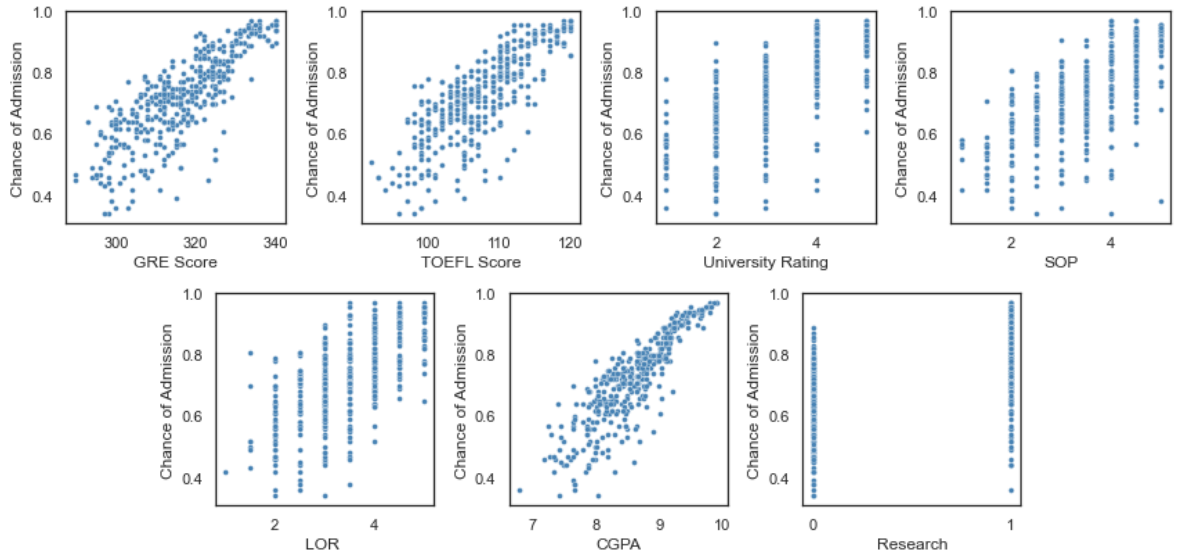

Figure 2. Scatter plot of each variable to chance of admit.

Figures 2 shows the correlation matrix between different variables and the chance of admission. Correlation coefficients measure the strength and direction of the relationship between two variables. A correlation coefficient ranges from -1 to 1, where a value of 1 indicates a perfect positive correlation, 0 indicates no correlation, and -1 indicates a perfect negative correlation. It can be seen that among the variables, the highest correlation with the chance of admission is observed for CGPA, with a correlation coefficient of 0.87. This indicates a strong positive correlation between CGPA and the chance of admission. Applicants with higher CGPA scores tend to have a greater likelihood of being admitted. On the other hand, the lowest correlation is observed between Research and the chance of admission, with a correlation coefficient of 0.55. Although still showing a moderate positive correlation, this suggests that the influence of research experience on the chance of admission is relatively weaker compared to other variables in the dataset.

To further analyze the relationships between the variables and the chance of admission, in Figure 2. we present the scatterplot of each variable plotted against the chance of admission. These scatterplots visually illustrate the patterns and trends in the data, providing additional insights into the relationships between the variables. From this figure, we can observe that higher values of variables like GRE Score, TOEFL Score, University Rating, SOP, LOR, and CGPA tend to be associated with a greater chance of admission. We can

Table 3. Performance of each machine learning model.

Model	Training			Testing		
	R^2	RMSE	MAE	R^2	RMSE	MAE
K-Nearest Neighbors	0.739	0.005	0.050	0.764	0.006	0.055
Random Forest	0.750	0.005	0.050	0.816	0.005	0.050
Support Vector Regression	0.665	0.006	0.066	0.749	0.006	0.069
XGBoost	0.690	0.006	0.055	0.786	0.006	0.054

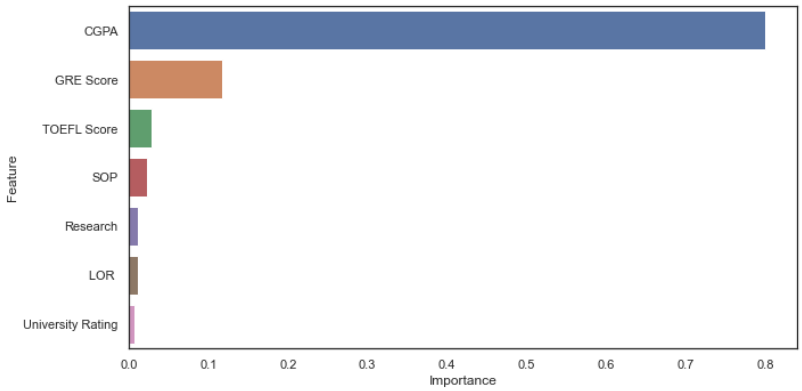

Figure 3. Feature importance of random forest model.

see a clear upward trend in these variables, indicating a positive relationship with the chance of admission. However, when it comes to Research, the scatterplot shows a weaker connection. The points are not as consistently higher with increased research experience, suggesting that research may have a less significant impact on the chance of admission compared to the other variables. The scatterplot analysis reinforces the findings from the correlation matrix, highlighting the significance of these variables in determining the likelihood of admission.

3.2. Machine Learning Model

In this study, we conducted modeling to predict the chance of admission using four different machine learning algorithms: K-Nearest Neighbors, Support Vector Regression, Random Forest, and XGBoost. The purpose of this analysis was to explore the predictive capabilities of these models and assess their performance in estimating the likelihood of admission.

In Table 3, we present the performance of different machine learning models in predicting the chance of admission. The highest performance in terms of R^2, RMSE, and MAE on the testing set is achieved by the Random Forest model, with an R^2 of 0.816, RMSE of 0.005, and MAE of 0.050. This indicates that the Random Forest model captures a significant amount of the variance in the chance of admission and produces relatively accurate predictions with low error. On the other hand, the lowest performance in terms of R^2, RMSE, and MAE on the testing set is observed in the Support Vector Regression model, with an R^2 of 0.749, RMSE of 0.006, and MAE of 0.069. This suggests that the Support Vector Regression model explains a slightly smaller proportion of the variance in the chance of admission compared to the other models and exhibits slightly higher prediction errors.

The Random Forest model outperforms the other models in terms of predictive accuracy, likely due to its ensemble nature and ability to capture complex interactions among the variables. Random Forest combines multiple decision trees and leverages their collective predictions, resulting in robust predictions and reduced overfitting. Conversely, the lower performance of the Support Vector Regression model may be attributed to the inherent limitations of the algorithm in capturing nonlinear relationships between the predictors and the target variable, as well as sensitivity to parameter selection.

Given that the Random Forest model demonstrated the best performance in predicting the chance of admission, we further analyze its predictive capabilities by examining the feature importance. Feature importance measures the relative contribution of each feature in predicting the chance of admission. It can be seen that the most important feature in predicting the chance of admission is CGPA with an importance score of 0.80. This indicates that CGPA has the highest impact on determining the likelihood of admission, suggesting that academic performance plays a crucial role in the admission decision. The next important feature is GRE Score (0.12), followed by TOEFL Score (0.03), SOP (0.02), Research

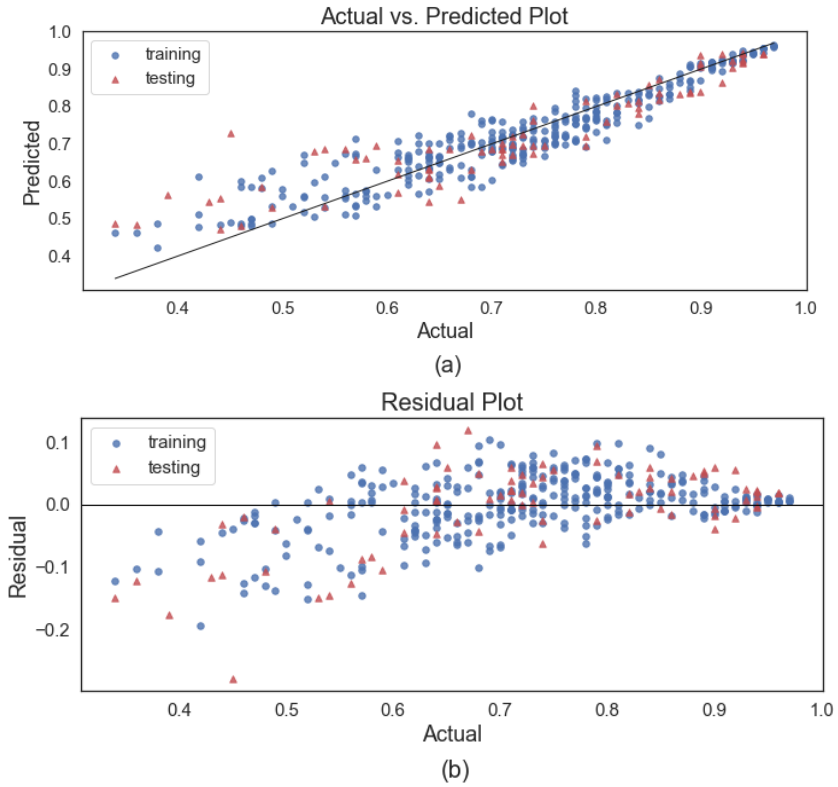

Figure 4. (a) actual vs. predicted plot and (b) residual plot.

(0.01), and University Rating (0.01). These features also contribute to predicting the chance of admission but to a lesser extent compared to CGPA and GRE Score.

Furthermore, to gain deeper insights into the performance of the Random Forest model, we conducted additional analyses using two types of plots: the actual vs. predicted plot and the residual plot (Figure 4). It can be seen that the predicted chance of admit are in the same range as the actual chance of admit. However, it is important to note that there are instances where the predicted chance of admit deviate from the actual chance of admit. These deviations are reflected in the residuals, which represent the differences between the actual and predicted scores. By analyzing the residuals, it is possible to identify the extent and direction of these deviations. The presence of residuals indicates that the model is not perfectly accurate in its predictions. Some residuals may be positive, indicating that the model tends to overestimate the target variable, while negative residuals indicate underestimation.

We compared our machine learning model with a previous study conducted by AlGamdi et al. [19] that utilized logistic regression. The previous study reported an RMSE of 0.072, while our Random Forest model achieved a significantly lower RMSE of 0.005. This demonstrates the superior performance and predictive accuracy of the Random Forest model in the context of university admissions prediction.

4. Conclusions

This study highlights the use of machine learning models in predicting university admissions. By considering various attributes, the developed model accurately predicts an individual's admission chances. The findings emphasize the significance of variables such as academic performance, underlining their role in admission outcomes. Random Forest emerges as the most effective model with reliable predictions and minimal errors. It is important to note that this study's findings are based on the utilization of the UCLA dataset. However, the model can be modified to fit the data from different universities, as their admission criteria may vary. Future studies can focus on collecting larger datasets and exploring the application of deep learning techniques. Overall, this research contributes to empowering applicants by providing valuable insights for informed decision-making and improving the university admission process.

Author Contributions: Conceptualization, A.M., T.R.N. and R.I.; methodology, T.R.N., N.R.S. and R.S.; software, T.R.N. and R.S.; validation, M.P., J.S., and R.I.; formal analysis, A.M., T.R.N., and N.R.S.; investigation, R.S. and E.Y.; resources, T.R.N. and R.I.; data curation, M.P. and R.I.; writing—original draft preparation, A.M. and T.R.N.; writing—review and editing, N.R.S., J.S. and R.I.; visualization, T.R.N. and E.Y.; supervision, R.I..; project administration, M.P. and R.I.; All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study is available from the Graduate Admission dataset on Kaggle, accessible at https://www.kaggle.com/datasets/mohansachary/graduate-admissions. The dataset was accessed on 14th May 2023.

Acknowledgments: The authors would like to express gratitude to their respective institutions for the support and resources provided throughout the completion of this work.

Conflicts of Interest: All the authors declare that there are no conflicts of interest.

References

1. Prakhov, I., and Yudkevich, M. (2019). University admission in Russia: Do the wealthier benefit from standardized exams?, *International Journal of Educational Development*, Vol. 65, 98–105. doi:10.1016/j.ijedudev.2017.08.007
2. Westkamp, A. (2013). An analysis of the German university admissions system, *Economic Theory*, Vol. 53, No. 3, 561–589. doi:10.1007/s00199-012-0704-4
3. Stemler, S. E. (2012). What Should University Admissions Tests Predict?, *Educational Psychologist*, Vol. 47, No. 1, 5–17. doi:10.1080/00461520.2011.611444
4. Mountford-Zimdars, A., Moore, J., and Graham, J. (2016). Is contextualised admission the answer to the access challenge?, *Perspectives: Policy and Practice in Higher Education*, Vol. 20, No. 4, 143–150. doi:10.1080/13603108.2016.1203369
5. Mengash, H. A. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems, *IEEE Access*, Vol. 8, 55462–55470. doi:10.1109/ACCESS.2020.2981905
6. Waters, A., and Miikkulainen, R. (2014). GRADE: Machine Learning Support for Graduate Admissions, *AI Magazine*, Vol. 35, No. 1, 64. doi:10.1609/aimag.v35i1.2504
7. Shahane, P. (2022). Campus Placements Prediction & Analysis using Machine Learning, *2022 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, 1–5. doi:10.1109/ESCI53509.2022.9758214
8. Walid, M. A. A., Ahmed, S. M. M., Zeyad, M., Galib, S. M. S., and Nesa, M. (2022). Analysis of machine learning strategies for prediction of passing undergraduate admission test, *International Journal of Information Management Data Insights*, Vol. 2, No. 2, 100111. doi:10.1016/j.jjimei.2022.100111
9. Idroes, R., Maulana, A., Noviandy, T. R., Suhendra, R., Sasmita, N. R., Lala, A., and Irvanizam. (2020). A Genetic Algorithm to Determine Research Consultation Schedules in Campus Environment, *IOP Conference Series: Materials Science and Engineering*, Vol. 796, 012033. doi:10.1088/1757-899X/796/1/012033
10. Yudono, M. A. S., Faris, R. M., De Wibowo, A., Sidik, M., Sembiring, F., and Aji, S. F. (2022). Fuzzy Decision Support System for ABC University Student Admission Selection. doi:10.2991/aebmr.k.220204.024
11. Acharya, M. S., Armaan, A., and Antony, A. S. (2019). A Comparison of Regression Models for Prediction of Graduate Admissions, *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, IEEE, 1–5. doi:10.1109/ICCIDS.2019.8862140
12. Kramer, O. (2013). K-Nearest Neighbors, 13–23. doi:10.1007/978-3-642-38652-7_2
13. Biau, G., and Scornet, E. (2016). A random forest guided tour, *TEST*, Vol. 25, No. 2, 197–227. doi:10.1007/s11749-016-0481-7
14. Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms, *Heca Journal of Applied Sciences*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12
15. Suthaharan, S. (2016). Support Vector Machine, 207–235. doi:10.1007/978-1-4899-7641-3_9
16. Noviandy, T. R., Maulana, A., Sasmita, N. R., Suhendra, R., Irvanizam, I., Muslem, M., Idroes, G. M., Yusuf, M., Sofyan, H., Abidin, T. F., and Idroes, R. (2022). The Prediction of Kovats Retention Indices of Essential Oils at Gas Chromatography Using Genetic Algorithm-Multiple Linear Regression and Support Vector Regression, *Journal of Engineering Science and Technology*, Vol. 17, No. 1, 306–326
17. Agustia, M., Noviandy, T. R., Maulana, A., Suhendra, R., Muslem, M., Sasmita, N. R., Idroes, G. M., Rahimah, S., Afidh, R. P. F., Subianto, M., Irvanizam, I., and Idroes, R. (2022). Application of Fuzzy Support Vector Regression to Predict the Kovats Retention Indices of Flavors and Fragrances, *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*, IEEE, 13–18. doi:10.1109/ICELTICs56128.2022.9932124
18. Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794
19. AlGhamdi, A., Barsheed, A., AlMshjary, H., and AlGhamdi, H. (2020). A Machine Learning Approach for Graduate Admission Prediction, *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, ACM, New York, NY, USA, 155–158. doi:10.1145/3388818.3393716