Analysis of Factors Influencing World University Rankings Based on Webometrics Indicators Using the Support Vector Machine (SVM) Method

Sukrianto, Muhammad Iqbal

Abstract

University rankings have become an essential indicator in assessing the quality and reputation of higher education institutions worldwide. Webometrics, as one of the leading international university ranking organizations, evaluates institutions based on three main indicators: Impact (50%), Openness (10%), and Excellence (40%). This study aims to analyze the factors that influence university rankings by applying a machine learning approach, specifically the Support Vector Machine (SVM) method. The dataset used in this research was obtained from the Mendeley Data Repository, containing information on 1400 universities across various countries and including Webometrics indicators for the year 2025. The research process involved several stages, including data preprocessing, feature extraction, and modeling using SVM to classify universities into three ranking categories: Top, Middle, and Low. The results of this study are expected to identify the most dominant indicators affecting university ranking positions and to provide strategic recommendations for higher education institutions to enhance their academic performance and global visibility.

Keywords: *Support Vector Machine (SVM), Determining Factors, University Ranking, Webometrics, Data Analysis, Machine Learning.*

Sukrianto[1]
[1]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: sukrianto.hambaallah@gmail.com[1]

Muhammad Iqbal[2]
[2]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: muhammadiqbal@dosen.pancabudi.ac.id[2]

Introduction

In the era of globalized higher education, university rankings have become one of the key benchmarks for assessing the quality and reputation of academic institutions. One of the most widely recognized international ranking systems is Webometrics, which evaluates universities based on three main indicators: Impact (50%), Openness (10%), and Excellence (40%) [1]. These indicators represent the extent of a university's digital influence, publication openness, and research quality [4]. Improvement in university ranking not only enhances academic reputation but also affects student interest, international collaboration, and public trust in the quality of education [2]. Therefore, it is crucial to understand the factors that influence a university's ranking position in order to formulate strategies that can strengthen institutional performance and global competitiveness.

This study employs a machine learning approach, specifically the Support Vector Machine (SVM) method, to analyze Webometrics data. SVM was chosen for its strong capability in classifying and predicting data patterns based on numerical relationships. Using a dataset obtained from Mendeley, which contains information on 1400 universities across various countries, this study aims to identify the most dominant factors influencing university rankings [3].

This research stems from the need to understand the factors that influence the ranking positions of universities within the Webometrics ranking system. Accordingly, the main issues examined in this study are the relationships among the Impact, Openness, and Excellence indicators and how the Support Vector Machine (SVM) method can be applied to classify universities based on these indicators. Furthermore, this study seeks to determine which indicator has the most significant influence on improving a university's ranking position.

The purpose of this study is to analyze the interrelationships among Webometrics indicators and to identify the dominant factors that determine a university's ranking position. Through the application of the Support Vector Machine (SVM) method, this research aims to develop a classification model capable of accurately predicting university ranking categories. The results of this analysis are expected to serve as a foundation for higher education institutions to design strategies that enhance the quality of scientific publications, data openness, and institutional digital visibility.

This study provides valuable benefits for various stakeholders. For academics and researchers, it serves as a reference for conducting machine learning based analyses in the field of higher education. For universities, the research findings can be used as an evaluation tool and as a foundation for developing strategies to enhance institutional reputation at both national and global levels. Meanwhile, for policymakers, this study may serve as a reference in formulating strategies to improve the quality and competitiveness of higher education institutions in the digital era [5].

This research has several limitations to maintain focus and direction. The data used were obtained from the Mendeley Data Repository, covering 1400 universities, and include only the three main Webometrics indicators: Impact, Openness, and Excellence. The analytical method employed is the Support Vector Machine (SVM), which classifies universities into three ranking categories: Top, Middle, and Low. The results of this study are focused on analyzing the relationships among variables rather than providing a direct evaluation of institutional performance.

Literature Review

University rankings serve as a measure to assess the quality and reputation of higher education institutions at both national and international levels. One of the most influential ranking organizations is the Webometrics Ranking of World Universities, compiled by the Cybermetrics Lab under the Spanish National Research Council (Consejo Superior de Investigaciones Científicas CSIC) in Spain.

Different from other ranking systems such as QS or Times Higher Education, Webometrics evaluates universities based on their digital presence and scientific impact on the web. The assessment is conducted using three main indicators: Impact (50%), which reflects the visibility of a university's website; Openness (10%), which represents the accessibility of scientific publications; and Excellence (40%), which measures the quality of research and the citation of scholarly works. The combination of these three indicators is used to determine a university's overall global ranking.

1. Factors Influencing University

Several previous studies have shown that university rankings are influenced by various factors, such as research quality, academic reputation, digital visibility, and scientific openness. Universities with a higher number of reputable publications, greater citation counts, and broader public access to their scholarly works tend to achieve better ranking positions.

In the context of Webometrics, the visibility factor (Impact) plays the most significant role, as it reflects the extent to which a university's website is recognized and referenced by other institutions. Meanwhile, the Excellence indicator is directly related to the productivity and quality of research, while Openness represents the university's commitment to providing open access to knowledge.

2. The Application of Machine Learning in Educational

Advancements in data technology have enabled the use of machine learning in analyzing educational phenomena, including university ranking systems. Machine learning has the capability to recognize complex patterns within numerical data and to make predictions based on the relationships among variables.

One of the algorithms frequently used is the Support Vector Machine (SVM), which is effective for classification tasks and offers a high level of accuracy. In the context of this study, SVM is employed to classify universities based on Webometrics indicators, allowing the identification of the dominant factors that influence their rankings.

3. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised learning algorithm that operates by finding the optimal hyperplane to separate data into distinct classes. This algorithm focuses on identifying the decision boundary with the maximum margin between one class and another, thereby producing an optimal classification outcome.

SVM has several types of kernels, such as linear, polynomial, and radial basis function (RBF). In this study, a linear SVM is employed because Webometrics data are numerical in nature and tend to exhibit linear relationships among variables. The strength of SVM lies in its ability to process high-dimensional data while producing stable and accurate classification results.

4. Previous Related Studies

Several previous studies relevant to this research include the following :

a. **Sarwar et al., 2021** conducted a study entitled "Webometrics: Evolution of Social Media Presence of Universities." Suggests that web visibility/digital presence indicators play a crucial role in university ranking systems this supports your argument that Impact (digital visibility) is a dominant factor [6].
b. **L. J. Wardley, E. Rajabi, S. H. Amin, and M. Ramesh (2025),** in their study "A machine learning approach feature to forecast the future performance of the universities in Canada." This article develops a machine learning model to predict the future performance of Canadian universities based on the identified criteria [7].
c. **C. S. Basireddy, V. K. G. Cheruku, B. P., S. Rajagopal, dan R. Soangra (2024),** in their work "Hybrid prediction models for assessing the Higher Education Institutions Performance in QS World Institution Rankings." Developing a hybrid machine learning model to assess the performance of higher education institutions in the QS World University Rankings [8].

d. **K. Wisaeng and B. Muangmeesri (2024),** in their study “University Rankings Prediction Using Hybrid Feature Selection Based on Machine Learning Methods.” This study introduces a novel approach for predicting university rankings by employing hybrid feature selection and machine learning techniques. The dataset used is from Times Higher Education (THE), encompassing data from 1,904 universities [9].
e. **B. Tóth, H. Motahari-Nezhad, N. Horseman, L. Berek, L. Kovács, Á. Hölgyesi, and M. Péntek (2024),** in their research “Ranking resilience: assessing the impact of scientific performance and expansion of global university rankings on Central European universities.” This study examines the sensitivity of position changes from year to year in the Times Higher Education (THE) World University Rankings for universities from V4 countries (Czech Republic, Hungary, Poland, and Slovakia) [10].
f. **F. E. Arévalo-Cordovilla and M. Peña (2024)** carried out a study titled “Comparative analysis of machine learning models for predicting student success in online programming courses.” Predicting student success in online programming courses by comparing four machine learning models: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Neural Network (MLP) [11].
g. **H. Guanin-Fajardo, J. Guaña-Moya, and J. Casillas (2024)**, in their study “Predicting academic success of college students using machine learning techniques.” Developing a machine learning model to predict academic success at the end of the first year of university based on academic and socioeconomic data [12].
h. **A. C. Estrada-Real, P. J. Lopez, and L. S. Rojas (2023)**, in their study “Data analytics approach for university competitiveness and digital visibility.” This study develops a machine learning-based predictive model to forecast the future performance of Canadian universities, using historical institutional data and institutional features as inputs [13].
i. **D. N. Ayu Ofta Sari and Muhammad Iqbal (2025)**, titled "Analysis of Demographic and Socio-Economic Factors for Early Detection of Dropout Risk Using Support Vector Machine (SVM) and Decision Tree Methods (Case Study: STMIK Triguna Dharma)," indicates that SVM achieved the highest accuracy of 58.54% at a 90:10 data ratio. Meanwhile, the Decision Tree model obtained a recall value of 100% across several data splits, signifying that this model was capable of identifying all students at risk of dropping out, even though its overall accuracy was not always the highest [14].
j. **M. Iqbal, S. Pardingotan Sipayung, A. R. Sinaga, and P. M. Hasugian (2024)**, titled "Analysis of Student Achievement with K-Means on Socioeconomic, Behavioral, and Psychological Factors," demonstrates that socioeconomic factors, motivation, and stress levels have a significant impact on students' academic achievement. The study identifies three distinct groups: the high-achievement cluster, the middle-achievement cluster, and the low-achievement cluster [15].

5. Relationship between variables

Based on previous theories and studies, university rankings can be explained as the result of a combination of the three Webometrics indicators: Impact, Openness, and Excellence. These three indicators serve as independent variables that influence the dependent variable, namely the university ranking. The relationship among these variables can be illustrated as follows:

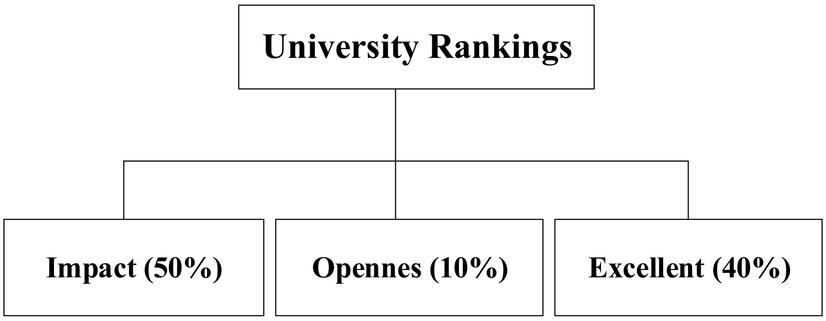

Figure 1. Relationship between variables

Research Methodology

This study employs a quantitative research design with a predictive computational approach that integrates social data analysis with artificial intelligence (machine learning) algorithms. The data analysis techniques applied in this research are illustrated in the workflow presented in Figure 2 below.

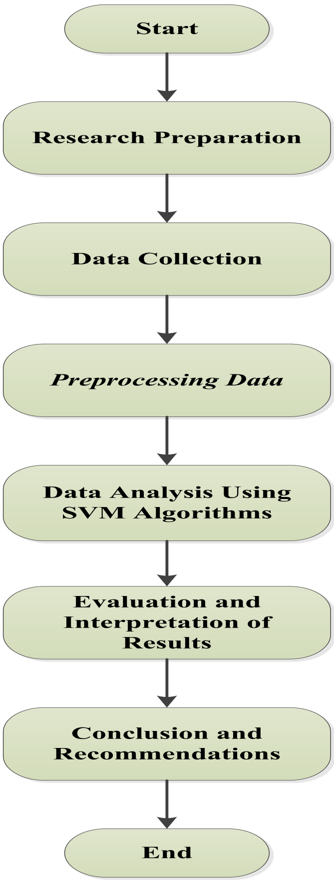

Figure 2. Workflow of Data Analysis Techniques

The figure above illustrates the workflow diagram of the research stages in data analysis, systematically describing each step in the implementation of the study at the University Rankings. The following section provides a detailed explanation of each stage presented in the diagram :

1. Research Preparation Stage

This stage begins with planning activities and problem identification. The researcher conducts a literature review related to theories of University Rankings, as well as machine learning analysis methods such as Support Vector Machine (SVM). Subsequently, stage involves identifying the research problem, conducting a literature review on Webometrics indicators, and defining the classification categories (Top, Middle, and Low Rank).

2. Data Collection Stage

After obtaining official permission, In this phase, the dataset is acquired from the Mendeley Data Repository, consisting of information from 1400 universities across various countries based on 2025 Webometrics data.

3. Data Preprocessing Stage

The raw data undergoes cleaning and transformation, including the removal of duplicates, handling missing values, and converting variables into a numerical format to ensure they are ready for algorithmic processing.

4. Data Analysis Stage Using Machine Learning Algorithms

The core analysis involves implementing the Support Vector Machine (SVM) method with a linear kernel to construct an optimal hyperplane for classifying universities based on the Impact, Openness, and Excellence indicators.

5. Evaluation and Interpretation of Results Stage

The performance of the SVM model is assessed using a classification report and a confusion matrix to evaluate metrics such as accuracy, precision, and recall.

6. Conclusion and Recommendation Stage

The final stage summarizes the findings, identifying the most dominant factors (Impact and Excellence) and providing strategic insights for higher education institutions to improve their global visibility and ranking.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm widely used for classification tasks. In this study, SVM is applied to classify university rankings based on Webometrics performance indicators, including **Impact**, **Openness**, and **Excellence** scores.

The primary objective of SVM is to construct an optimal decision boundary, known as a **hyperplane**, that separates university data into different ranking categories (*Top, Middle,* and *Low*) with the maximum margin. For linearly separable data, the decision function of SVM is defined as:

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ represents the feature vector of the university's Webometrics indicators, and $\boldsymbol{b}$ is the bias term. The optimal separating hyperplane is represented by the equation:

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

In this research, a **linear kernel** is utilized to model the relationship between Webometrics indicators and the university ranking categories. This kernel was selected because Webometrics data is numerical and tends to exhibit linear relationships between variables. The linear kernel function is expressed as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$$

The SVM optimization process aims to minimize the weight vector norm while maintaining correct classification of training samples, which can be formulated as:

$$\min_{\mathbf{w},b} \quad \frac{1}{2} \parallel \mathbf{w} \parallel^2$$

subject to the constraint:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, i = 1,2, \dots, n$$

where $y_i \in \{-1, +1\}$denotes the class label (**Top Rank category**). Through this formulation, SVM provides a robust classification model for identifying the dominant factors influencing global university ranking positions.

Results

This study utilizes data from the Mendeley Data Repository, which contains university ranking information based on the Webometrics Ranking of World Universities system. The dataset includes approximately 1400 universities from various countries and comprises three main indicators used in the evaluation: Impact (50%), Openness (10%), and Excellence (40%). These three indicators reflect key aspects in measuring academic reputation: Impact represents the visibility and influence of the university in the digital world, Openness indicates the extent to which scholarly publications are accessible to the public, and Excellence assesses the quality of research and scholarly work produced by the university. The data is then classified into three ranking categories: Top Rank (positions 1 100), Middle Rank (positions 101 200), and Low Rank (positions beyond 200) to facilitate the analysis process using the SVM algorithm. Below is the dataset table that is ready for processing with the SVM algorithm:

Table 1. Sample Dataset University Rankings

No	University	Det.	Country	Impact (50%)	Openness (10%)	Excellence (40%)
1	Massachusetts Institute of Technology			1124	1072	1066
2	Imperial College London			1061	1185	1310
3	Stanford University			1317	1204	1272
4	University of Oxford			1196	1282	1494
5	Harvard University			1273	1296	1422
6	University of Cambridge			1282	1184	1445
7	National University of Singapore (NUS)			1596	1283	1348
8	UCL (University College London)			1337	1453	1594
9	California Institute of Technology			1379	1422	1622
10	The University of Hong Kong			1685	1421	1451

After the University Ranking dataset has been compiled, the next step is to enter the training data compilation phase, which includes Ranking, World Rank, University, Det, Country, Impact (50%), Openness (10%), and Excellence (40%). This phase is illustrated in Table 2 below as follows :

Table 2. Training Data

	ranking	World Rank	University	Det.	Country	Impact (50%)	Openness (10%)	Excellence (40%)
0	1	1	Massachusetts Institute of Technology	NaN	NaN	1124	1072	1066
1	2	2	Imperial College London	NaN	NaN	1061	1185	1310
2	3	3	Stanford University	NaN	NaN	1317	1204	1272
3	4	4	University of Oxford	NaN	NaN	1196	1282	1494
4	5	5	Harvard University	NaN	NaN	1273	1296	1422

Once the training data has been compiled, the next step is to process the data by converting all variables into a numerical format. This phase is referred to as the preprocessing stage, as shown in Table 3 below :

Table 3. Data Preprocessing

	Ranking	**Impact**	**Openness**	**Excellence**
0	1	1124	1072	1066
1	2	1061	1185	1310
2	3	1317	1204	1272
3	4	1196	1282	1494
4	5	1273	1296	1422

In this stage, Table 3 presents the preprocessed data utilized for further analysis, including university rankings and three primary Webometrics indicators: Impact, Openness, and Excellence. The dataset includes values for each university's ranking, with corresponding numerical values for Impact, Openness, and Excellence scores. These indicators play a crucial role in determining the universities' standings in the Webometrics ranking system. The scikit-learn function used to build the SVM model is shown in Figure 3 below :

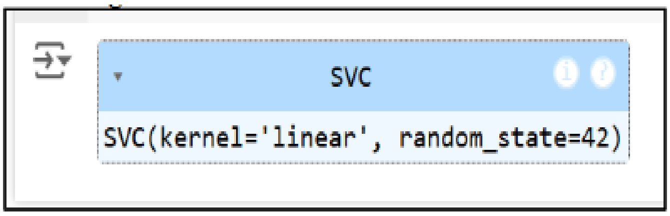

Figure 3. Initializing and Training the SVM Mode

The code snippet shown in the image illustrates the configuration of a Support Vector Classifier (SVC) using the scikit-learn library. The model is initialized with a linear kernel, which implies that the classifier will construct a hyperplane in a high-dimensional space to separate the classes linearly. This choice of kernel is appropriate for problems where the data is expected to be linearly separable, or at least approximately so. The random_state = 42 argument is used to ensure the reproducibility of the results, as it controls the random number generator used in the data shuffling and splitting process. By fixing the random state, the model's performance can be consistently evaluated across different runs. The classification report results from the SVM model can be seen in Figure 4 below as follows :

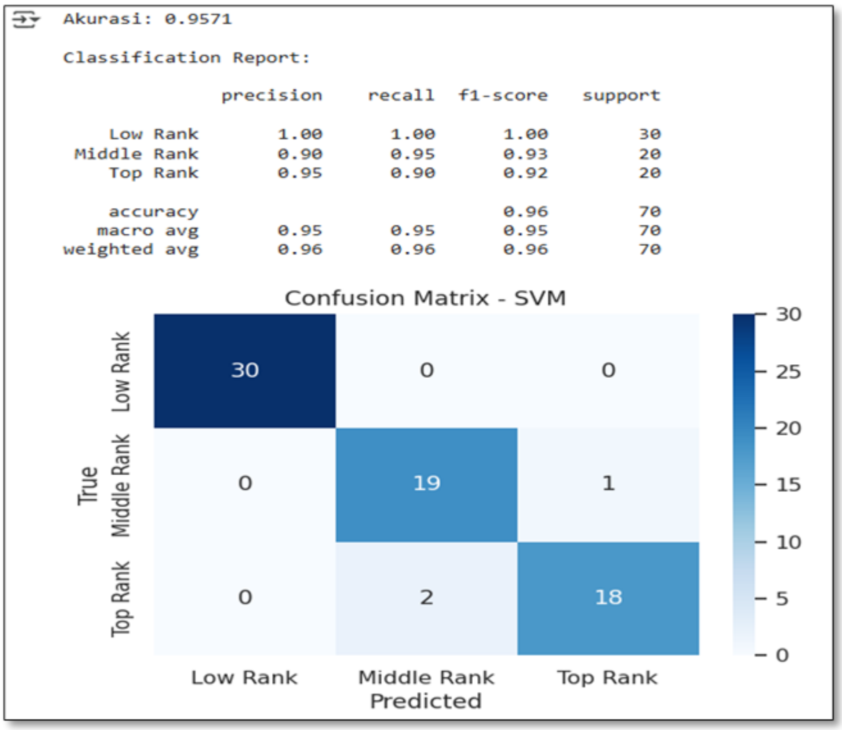

Figure 4. Report Classification Results

Model evaluation is performed to assess the extent to which the Support Vector Machine (SVM) algorithm is capable of correctly classifying university rankings based on the three Webometrics indicators: Impact, Openness, and Excellence. The evaluation results are presented in the form of a classification report and confusion matrix. According to the classification report, the overall accuracy of the model reaches approximately 95-96%, with average values for precision, recall, and F1-score above 0.90 for each category. This indicates that the SVM model demonstrates excellent accuracy and consistency in predicting university ranking classes. A high precision value means that most of the universities predicted to belong to a certain category indeed belong to that category. A high recall value indicates that the model successfully captures the majority of universities that truly belong to each category. The F1-Score, as a combination of precision and recall, reflects the model's overall performance balance.

Meanwhile, the confusion matrix results show that most predictions lie along the main diagonal, meaning the model is able to classify the data correctly. Minor classification errors typically occur between the Top Rank and Middle Rank categories, which are caused by the similarity in indicator values for universities ranking between positions 100 and 200. This visualization helps to understand how well the SVM model separates the classes based on the available features, while also providing insights into areas that are more challenging to classify. The results of the visualization of the analysis of factors influencing the World University Rankings based on Webometrics indicators are shown in Figures 5 and 6 below :

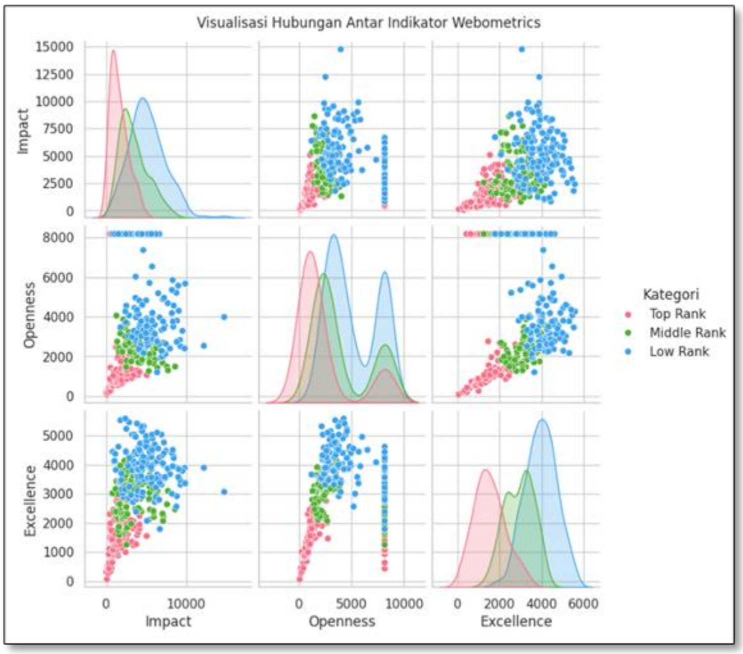

Figure 5. Visualization of the Relationship Between Webometrics Indicators

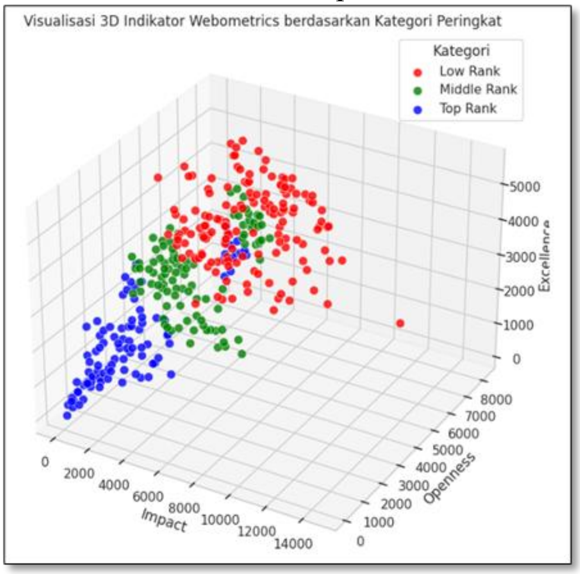

Figure 6. Visualization of Webometrics Indicators based on ranking

Based on the 3D graph above, a clear pattern emerges indicating that universities with high Impact and Excellence scores (located at the upper-right corner of the graph) tend to fall into the Top Rank category. Meanwhile, universities with lower values across all three indicators are concentrated in the lower area of the graph and generally belong to the Low Rank category. The color distribution shows that the Openness indicator exhibits less variation across the categories, while Impact and Excellence display a more distinct separation between the groups. This pattern reinforces the findings from the SVM model analysis, indicating that the dominant factors influencing university rankings are Impact and Excellence, as these two indicators exhibit the most decisive distribution in determining the position of universities within the Webometrics rankings.

Conclusion

This study demonstrates that the Support Vector Machine (SVM) method is effective in classifying university rankings based on Webometrics indicators, with an accuracy rate of 95.7%. The key findings reveal that the Impact indicator is the most dominant factor in determining university ranking positions, followed by Excellence, while Openness has a smaller influence. These findings emphasize the importance of digital visibility and the quality of scholarly publications as determinants of academic reputation at a global level. Although the results indicate excellent model performance, this study has limitations in terms of the number of variables and the time frame, which is limited to the year 2025. The model used also only considers digital aspects, thus excluding non-technical factors such as teaching quality or academic collaboration. Therefore, the generalization of the results still needs to be tested through the expansion of data and variables in future research.

Future studies are recommended to use cross-year data in order to analyze ranking trends over time. A comparison with other machine learning algorithms, such as Random Forest or Neural Networks, could also be conducted to assess model reliability. Additionally, incorporating non-digital indicators such as academic reputation, research activity, and international collaboration would provide a more comprehensive view. Practically, the results of this study imply the need for universities to enhance their digital academic strategies, particularly by expanding the reach of official websites, increasing the number of open-access scholarly publications, and strengthening international research collaborations to improve institutional rankings and global reputation.

References

[1] J. L. Ortega, "Visualizing the Webometrics Ranking of World Universities: A Network Analysis of Institutional Web Presence," *Journal of Informetrics*, vol. 18, no. 1, pp. 112 125, Feb. 2024, doi: 10.1016/j.joi.2023.101456.

[2] M. A. Al-Nasser and H. Al-Shorbagi, "The Role of Digital Presence in Enhancing Global Academic Competitiveness," *International Journal of Information Management*, vol. 72, art. no. 102651, 2025, doi: 10.1016/j.ijinfomgt.2024.102651.

[3] R. K. Sahoo and S. S. Mohapatra, "Applying Support Vector Machine for Performance Classification in Educational Data Mining," in *Proceedings of the 2024 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2024, pp. 215 221, doi: 10.1109/DSAA61234.2024.00032.

[4] T. Nguyen and P. Vo, "Analyzing the Impact of Open Access Publications on University Rankings Using Machine Learning," *Journal of Scholarly Publishing*, vol. 55, no. 4, pp. 289 308, Oct. 2024, doi: 10.3138/jsp-2023-0045.

[5] S. Kurniawan and D. P. Hasmoro, "Faktor-Faktor Dominan dalam Penilaian Reputasi Perguruan Tinggi Berbasis Digital," *Jurnal Teknologi Informasi dan Pendidikan*, vol. 18, no. 2, pp. 77 89, 2025, doi: 10.21831/jti.v18i2.65432.

[6] R. Sarwar, A. Zia, R. Nawaz, A. Fayoumi, N. R. Aljohani, and S.-U. Hassan, “Webometrics: evolution of social media presence of universities,” Scientometrics, vol. 126, no. 2, pp. 951 967, Apr. 2021, doi: 10.1007/s11192-020-03804-y.

[7] L. J. Wardley, E. Rajabi, S. H. Amin, and M. Ramesh, “A machine learning approach feature to forecast the future performance of the universities in Canada,” Machine Learning Applications, vol. 16, art. no. 100548, 2024, doi: 10.1016/j.mlwa.2024.100548.

[8] C. S. Basireddy, V. K. G. Cheruku, B. P., S. Rajagopal, and R. Soangra, “Hybrid prediction models for assessing the higher education competitiveness,” F1000Research, vol. 13, art. 1529, Dec. 2024, doi: 10.12688/f1000research.155847.1.

[9] K. Wisaeng and B. Muangmeesri, “University Rankings Prediction Using Hybrid Feature Selection Based on Machine Learning Methods,” International Journal of Analytical and Applied Mathematics, vol. 23, no. 2, pp. 311 320, 2025, doi: 10.28924/2291-8639-23-2025-112.

[10] B. Tóth, H. Motahari-Nezhad, N. Horseman, L. Berek, L. Kovács, Á. Hölgyesi, and M. Péntek, “Ranking resilience: assessing the impact of scientific performance and expansion of global university rankings on Central European universities,” Scientometrics, vol. 129, no.3, pp. 1739 1770, 2024, doi: 10.1007/s11192-023-04920-1.

[11] F. E. Arévalo-Cordovilla and M. Peña, “Comparative analysis of machine learning models for predicting student success in online programming courses,” Mathematics, vol. 12, no. 20, art. 3272, Oct. 2024, doi: 10.3390/math12203272.
[12] J. H. Guanin-Fajardo, J. Guaña-Moya, and J. Casillas, “Predicting academic success of college students using machine learning techniques,” Data, vol. 9, no. 4, art. 60, Apr. 2024, doi: 10.3390/data9040060.
[13] A. C. Estrada-Real, P. J. Lopez, and L. S. Rojas, “Data analytics approach for university competitiveness and digital visibility,” Journal of Educational Data Science, vol. 6, no. 1, pp. 45 60, 2023, doi: 10.1016/j.jeds.2023.01.006.
[14] D. N. Ayu Ofta Sari, Muhammad Iqbal, “Analisis Faktor Demografi Dan Sosial Ekonomi Untuk Mendeteksi Dini Risiko Putus Kuliah Menggunakan Metode Support Vector Machine (Svm) Dan Decision Tree (Studi Kasus : STMIK Triguna Dharma),” vol. 07, no. 02, p. 17, 2025.
[15] M. Iqbal, S. Pardingotan Sipayung, A. R. Sinaga, and P. M. Hasugian, “Analysis of Student Achievement with K-Means on Socioeconomic, Behavioral, and Psychological Factors,” Inform. dan Sains, vol. 14, no. 04, pp. 715 728, 2024, doi: 10.54209/infosains.v14i04.