International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 09 – Issue 05 || Sep - Oct 2024 || PP. 34-40

Predicting University Ranking Using the Decision Tree and Gradient Boosting Machine Learning Algorithms

[1]Akomolafe, H. D., [2]Olanrewaju, B. S., [3]Omilabu, A. A., [4]Asoro, B. O. and [5]Osunade, O.

[1,4,5]Department of Computer Science, University of Ibadan, Nigeria
[2]Wellspring University, Benin City, Nigeria
[3]Department of Computer Science, Tai Solarin University of Education, Ijebu

Abstract: University rankings are lists that rank tertiary institutions based on various criteria. These rankings serve as a tool for prospective students, academics, and policy makers to assess the quality and reputation of universities. Most existing ranking systems utilize arithmetic scoring by adding the scores of various indicators to determine the rank. Though there are works done using machine learning as well, However, there is a need for a more accurate ranking system that will be more acceptable. The aim of the project is to develop a predictive ranking system for Universities in Africa. The dataset utilized for this study was collected for the year 2023, and it underwent thorough preprocessing. Subsequently, a Decision Tree model was constructed to analyze the relationships between various indicators and the rankings of universities. To enhance the precision and generalization of this model, a gradient-boosting regression technique was also employed. The evaluation of the refined university ranking system was conducted using R2 (coefficient of determination), RMSE (root mean squared error), and MAE (mean absolute error). The decision tree model provided a baseline ranking with moderate accuracy. However, after applying gradient boosting, the ranking accuracy significantly improved, yielding more precise and reliable results for university ranking.

Keywords: African Universities, Ranking algorithms, Decision Tree model, Gradient boosting

Introduction

Machine Learning is a technology employed to enable computers to learn from supplied data, subsequently making predictions based on the acquired knowledge gained through data analysis and learning from input operations (Bell, 2022). This study utilizes machine learning tools such as Decision Trees and Gradient Boosting.

Decision Trees are a commonly employed tool in machine learning due to their simplicity and interpretability. Their operational principle involves recursively dividing the data into subsets based on the most informative features. This process generates a tree-like structure where each internal node represents a decision based on an attribute, while each leaf node corresponds to a predicted outcome (Zhou, 2021). Decision trees are versatile, applicable to both classification and regression tasks, and are thus valuable in a broad array of scenarios, including university performance prediction.

Gradient Boosting is an ensemble learning technique that builds multiple weak learners, often decision trees, sequentially. It aims to correct the errors of its predecessors, resulting in a more robust and accurate predictive model by combining the predictions of multiple weak learners, Gradient Boosting captures complex relationships in the data and is known for its superior predictive performance, particularly in scenarios with large and diverse datasets. (Chakrabarty et al., 2019)

Higher education in Africa has witnessed significant growth and transformation over the past decades, emerging as a key driver of socio-economic development and human capital enhancement in the region (Cloete et al., 2012). As African countries strive to achieve sustainable development goals, the role of universities in fostering innovation, research output, and academic excellence has become paramount. However, evaluating and predicting the performance of universities is a complex task that requires robust methodologies and data-driven approaches.

The traditional approach to evaluating universities' performance often relies on ranking systems that use various metrics, including research output, faculty qualifications, student-to-faculty ratio, and international collaborations (Safón, 2019)). While these rankings provide valuable insights, they have limitations in capturing the multifaceted dimensions of university performance. Moreover, these rankings might not consider the contextual factors unique to African universities, such as resource constraints, diverse cultural settings, and challenges in governance.

Accurate predictions of university performance are critical for various stakeholders involved in the higher education ecosystem. Policymakers at the national and regional levels can benefit from these predictions to make informed decisions about resource allocation and policy formulation. By identifying underperforming institutions, policymakers can devise targeted strategies to improve academic quality, research output, and overall competitiveness. On the other hand, institutions that consistently exhibit exceptional performance can

serve as models for others, fostering healthy competition and collaboration among universities (Daraio et al., 2015).

For prospective students, accurate predictions offer valuable insights when making decisions about their choice of universities. Students can evaluate the likelihood of success and satisfaction based on predicted outcomes, such as graduation rates, student-to-faculty ratios, and employment opportunities after graduation. Informed choices aligning with individual aspirations can lead to better student outcomes and contribute to the overall improvement of the higher education landscape (López-Illescas et al., 2011)

Most existing ranking systems utilize arithmetic scoring and add up the scores of various indicators to determine the highest-scoring entity for ranking purposes. Though there are works done using machine learning as well. However, there is need for more accurate ranking system that will be more acceptable. The aim of the project is to evaluate the performance of two predictive ranking algorithms- Decision Tree and Gradient Boosting- for Universities in Africa.

Traditional ranking systems often have limitations in providing a comprehensive and accurate evaluation of African universities due to various contextual factors. By utilizing decision trees and gradient boosting, the study can offer a more sophisticated and data-driven approach, capturing complex relationships between performance indicators and providing a deeper understanding of universities' strengths and weaknesses.

Review of Related Literature

The decision tree, gradient boosting, and machine learning algorithms used to rank universities were examined. This study examines the outcomes of portfolio analysis and optimization of a group of exploration projects that have been modeled using stochastic simulation and decision tree techniques, respectively. The opportunity set for the portfolio is made up of the collection of exploration initiatives. The decision tree approach approximates the range of potential project outcomes by choosing a few possibilities from the complete distribution and giving those scenarios probabilities (Erdogan et al., 2001).

Regarding the current situation in conventional farming, there has been a critical demand for predicated data in farming that can assist farmers in understanding their current issues and taking appropriate action. They suggested a technique to handle these issues that uses a "Decision Tree Classifier" to predict cotton crop illnesses while taking into account variables like temperature, soil moisture, etc. As a result, farmers would benefit from higher-quality crops (Chopda et al., 2018).

The dataset is analyzed using exploratory data analysis, and machine learning algorithms are subsequently assessed using regression approaches to forecast worldwide rankings. The global university ranking dataset is applied with the regression algorithm, which yields trees; boosting regression is then used for improved accuracy (Udupi et al., 2023). Through the examination of international university performance indicators, they developed a method for creating a framework for predicting university rankings. They take into account a standardized dataset of Times Higher Education's global university rankings in this instance.

First, they analyze data from university rankings by country, looking for variations in performance metrics and key characteristics. They divided the ranking dataset into training and test data in order to construct the suggested prediction model. Then, using our suggested outlier detection and rank score calculation algorithm, they construct predicted scores for each influential attribute based on scores from prior years. On the basis of the anticipated overall score, all universities are thereafter ranked globally. Then, using the ROC curve, recall, and number of matched ranks against rank deviation, we assess the prediction system's accuracy (Vaibhav Singh et al., 2021).

A case-based reasoning (CBR) prediction model's attribute weights are assigned using one of three different decision-tree-based methods, and their performances are compared. By taking into account the attributes' placements in the decision tree and their presence or absence, attribute weights are generated. In the study (Doan et al., 2008), this procedure and the creation of the CBR simulation model are detailed.

The purpose of this paper by Pradeep et al. (Pradeep et al., 2020) is to find an accurate algorithm for implementing information mining systems using the Weka tool to predict the success or failure of motion pictures based on a few key characteristics with respect to the system. In the first stage, the weighting for each attribute is determined, and then the weighting for that video is determined by combining the attribute calculations using decision tree methods. Here, three decision tree algorithms are the main ones being used to try to implement this concept.

Student retention has grown to be one of the most crucial indicators of success for higher education institutions since it has an impact on university rankings, school reputation, and financial health. A detailed understanding of the factors contributing to attrition is the first step in increasing student retention from the institution's point of view. Such knowledge serves as the foundation for correctly identifying at-risk students and implementing retention strategies. In this work, they created analytical models to forecast freshmen student attrition utilizing eight years of institutional data coupled with three well-known data mining approaches.

Artificial neural networks, decision trees, and logistic regression are the three different types of models (Delen, 2011).

In order to estimate the punched shear resistance of reinforced concrete (R/C) interior slabs without shear reinforcement, this work will demonstrate the use of extreme gradient boosting (XGBoost). For the development of the XGBoost model and its training and testing (Nguyen et al., 2021). Applications of tree-based ensemble algorithms in the field of traffic prediction are scarce, according to (Zhang & Haghani, 2015). In order to increase prediction accuracy and model interpretability, we investigate and model freeway trip time in this study using the gradient boosting regression tree technique (GBM). The gradient boosting tree approach deliberately merges more trees by correcting flaws caused by its prior base models, hence, potentially enhances prediction accuracy. Using trip time data from INRIX along two motorway sections in Maryland, it is detailed addressed how different parameters affect model performance and correlations of input-output variables. The proposed approach is then contrasted with a benchmark model and another well-known ensemble method.

The S-LightGBM model is used in this study to forecast the performance of crowd funding by taking into account a number of potential parameters. The linguistic and sentimental characteristics of project descriptions were extracted using text mining and lexicon-based sentiment analysis techniques. On 5916 crowd sourcing projects during 2017 and 2018 (Geng et al., 2020), this study examines the prediction abilities of logistic regression, support vector machine, light gradient boosting machine, and S-LightGBM.

The study approaches the issue by concentrating on forecasting the position of top songs over the next six months. In the ZALO AI CHALLENGE 2019, the Hit Song Prediction challenge uses a dataset that includes songs as well as song-related data like composer, artist, release date, etc. Because of this, they describe in this study how to employ the Gradient Boosting technique to handle hit song prediction as a ranking problem, whereas most prior work formulates it as a regression or classification problem. The winning model outperforms most of the competing solutions (better than the third-ranked solution out of 87 in total) when determining whether a song will be a top 10 dance hit, with a Root Mean Square Error of 1.48815 (Pham et al., 2020).

This study aims to create a web ranking system specifically for tertiary institutions in underdeveloped countries. In this study, nine current ranking systems were taken into account and combined to create a single system. The aggregate resulted in the identification of twenty- two criteria, however only seventeen criteria are used, and percentage weights were also objectively assigned to each. While additional factors that are not available are manually included into the suggested ranking systems, the value of each criterion that is available on the university website was extracted using a web crawling algorithm (Gabriel & Aribisala, 2020).

Gradient-boosted decision tables (BDTs) are presented in this research by Lou and Obukhov (2017). These BDTs essentially map a series of d Boolean tests to a true value in R, forming d-dimensional decision tables. They proposed new algorithms to fit these decision tables. Their empirical study found that decision tables make superior weak learners within the gradient boosting framework, enhancing the accuracy of the boosted ensemble. Additionally, they developed a robust data structure for describing decision tables and introduced an efficient method to improve the scoring effectiveness of an ensemble of decision tables.

Experimental results on publicly available classification and regression datasets revealed significant speed improvements, ranging from 1.5x to 6x compared to the baseline of boosted regression trees. They supplemented the experimental evaluation with a bias-variance analysis to illustrate how different weak models impact the predictive capability of the boosted ensemble, with gradient boosting with randomly back-fitted decision tables emerging as the most accurate approach in various classification and regression tasks.

Furthermore, the study introduced a novel ranking algorithm that combines the strengths of two existing approaches: boosted tree classification and Lambda Rank, which has been empirically proven to be optimal for widely used information retrieval metrics. Their approach significantly outperforms the state-of-the- art in terms of speed during both training and testing phases while maintaining comparable accuracy levels. Although their framework is based on boosted regression trees, the principles can be applied to any weak learners. They also demonstrated a method for identifying the best linear combination for each pair of rankers, effectively addressing the line search issue during boosting. Moreover, they showed that starting with a pre-trained model and boosting it using its residuals offers an efficient technique for model adaptation, yielding notably improved results for the critical problem of training rankers in Web Search markets with limited labeled data, given a ranker initially trained on a larger market with much more data (Wu et al., n.d.).

Methodology

The system developed for this study applied both the decision tree model and gradient boosting on the same dataset. Figure 1 represents the workflow diagram illustrating the implementation of these machine learning models on the dataset. The dataset, contained data from The World Universities Ranking for the year 2023, was obtained from https://www.kaggle.com/datasets/r1chardson/the-world-university-rankings- 2011-

2023. It consisted of information on 154 universities, comprising 20 features. The implementation is done using the Python language and the Jupiter Notebook while Pandas' built-in functions facilitated the process.

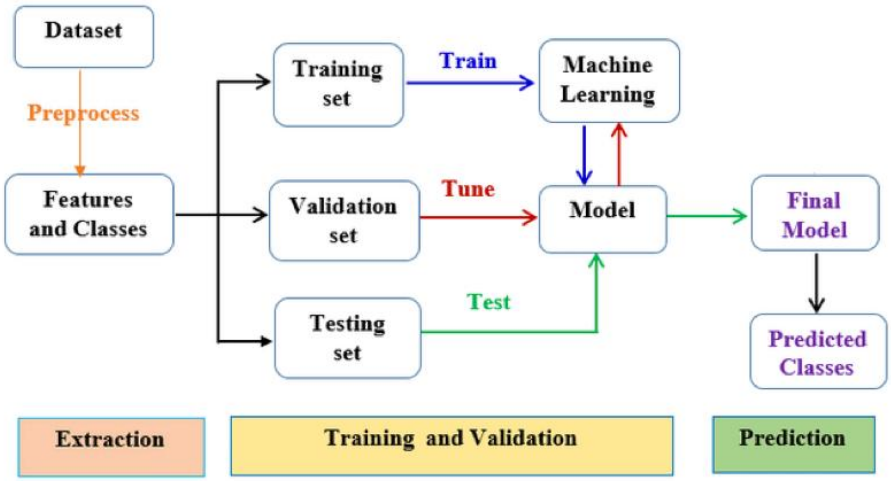

Figure 1 Block Diagram depicting the system

These are the performance metrics applied:

Root Mean Square Error (RMSE): The RMSE determined how far anticipated values are from actual values in a regression analysis by demonstrating how concentrated the data is around the line of best fit.

Mean Absolute Error (MAE): The average absolute difference between the observed data and the anticipated value is represented by the MAE value.

MAE = [yi - xi] * (1/n)

Where:

yi denotes the observed value of it.

xi denotes that it has the anticipated value.

The number ’n’ represents all of the observations.

R-squared: Sum of Squares Regression (SSR) to Sum of Squares Total (SST) ratio is known as R-Squared. Sum of Squares Regression measures the proportion of variance that the regression line can explain. Measures of the goodness of fit include the R-squared value. The regression model performs better when R-Squared is closer to 1.

Results and Discussions

The execution of the two machine learning models on the dataset using Jupiter notebook produced the results that are aggravated in Figure 2 and Table 1.

```
results.head(20)
```

]:

	University	Actual Ranking	Predicted Ranking (Decision Tree)	Predicted Ranking (Gradient Boosting)
26	Mansoura University	27	17	13
95	Helwan University	96	110	91
153	Zonguldak Bülent Ecevit University	154	150	144
107	University of Sousse	108	102	98
106	University of Sfax	107	109	103
62	University of Marrakech Cadi Ayyad	63	61	64
11	Muhimbili University of Health and Allied Scie...	12	22	21
41	University of the Free State	42	45	46
35	Beni-Suef University	36	34	34
53	Atılım University	54	85	58
83	Bozok University	84	77	85
152	Yeditepe University	153	118	139
20	Benha University	21	22	16
141	University Mohamed Boudiaf of M'Sila	142	144	150
88	Erciyes University	89	88	83
78	Ankara University	79	99	92
135	Izmir Institute of Technology	136	123	126
14	Durban University of Technology	15	14	14
9	University of KwaZulu-Natal	10	11	10
119	University of Biskra	120	117	122

Figure 2: The actual value, the Decision Tree and Gradient Boosting predicted value

The list generated when the two machine language algorithms are implemented is shown in Figure 2. In the case of University of KwaZulu-Natal, the actual value and the Gradient Boosting prediction are the same while the Decision Tree outcome is different. For Beni-Suef University, the Decision Tree and Gradient Boosting had the same prediction while the actual ranking was lower.

Table 1: Performance evaluation of Decision Tree and Gradient Boosting

	Decision Tree	Gradient Boosting
RMSE	20.99	17.60
MAE	16.56	14.67
R-squared	0.78584	0.84815

From Table 1, the RMSE and MAE values for Gradient Boosting is lower than that of Decision Tree. This implies that Gradient Boosting has better predictive outcomes. The R-squared value for Gradient Boosting is higher than that of Decision Tree. This shows that the variance of Gradient Boosting predictions is closer to the actual value than that of Decision Tree.

Conclusion

The decision tree model and gradient boosting were the two models used in the ranking prediction. The decision tree model provided a baseline ranking with moderate accuracy. With gradient boosting, the ranking

accuracy significantly improved, yielding more precise and reliable results for university ranking which can be observed from the performance metrics applied to the two models. Organizations and institutes that rank tertiary institutions exist locally and internationally, but there is a need for a more accurate ranking system that will be more reliant on the quality and amount of data available. The ranking system requires more relevant features in the dataset so as to lead to better results.

References

[1]. Bell, J. (2022). What is machine learning? Machine Learning and the City: Applications in Architecture and Urban Design, 209–216. https://doi.org/10.1007/978-3-319-18305-3_1/COVER

[2]. Chakrabarty, N., Kundu, T., Dandapat, S., Sarkar, A., & Kole, D. K. (2019). Flight arrival delay prediction using gradient boosting classifier. Advances in Intelligent Systems and Computing, 813, 651–659. https://doi.org/10.1007/978-981-13-1498-8_57/COVER

[3]. Chopda, J., Raveshiya, H., Nakum, S., & Nakrani, V. (2018). Cotton Crop Disease Detection using Decision Tree Classifier. 2018 International Conference on Smart City and Emerging Technology, ICSCET 2018. https://doi.org/10.1109/ICSCET.2018.8537336

[4]. Cloete, N., Bailey, T., Pillay, P., Bunting, I., & Maassen, P. (2012). Universities and Economic Development in Africa. Universities and Economic Development in Africa, 110. https://doi.org/10.47622/9781920355807

[5]. Daraio, C., Bonaccorsi, A., & Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. European Journal of Operational Research, 244(3), 918–930. https://doi.org/10.1016/J.EJOR.2015.02.005

[6]. Delen, D. (2011). Predicting Student Attrition with Data Mining Methods. Http://Dx.Doi.Org/10.2190/CS.13.1.b, 13(1), 17–35. https://doi.org/10.2190/CS.13.1.B

[7]. Doğan, S. Z., Arditi, D., & Günaydin, H. M. (2008). Using Decision Trees for Determining Attribute Weights in a Case-Based Model of Early Cost Prediction. Journal of Construction Engineering and Management, 134(2), 146–152. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:2(146)

[8]. Erdogan, M., Mudford, B., Chenoweth, G., Holeywell, R., & Jakubson, J. (2001). Optimization of Decision Tree and Simulation Portfolios: A Comparison. https://doi.org/10.2118/68575-MS

[9]. Gabriel, A., & Aribisala, B. (2020). Automatic Ranking of Tertiary Institutions in Developing Nations (a Case Study of Nigeria Universities). Academia.Edu, 1–45. https://www.academia.edu/download/62391974/PUBLISHED_PAPER20200317-2356-1e3cv8e.pdf

[10]. Geng, S., Huang, M., & Wang, Z. (2020). A Swarm Enhanced Light Gradient Boosting Machine for Crowdfunding Project Outcome Prediction. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12488 LNCS, 372–382. https://doi.org/10.1007/978-3-030-62463-7_34/COVER

[11]. López-Illescas, C., de Moya-Anegón, F., & Moed, H. F. (2011). A ranking of universities should account for differences in their disciplinary specialization. Scientometrics, 88(2), 563–574. https://doi.org/10.1007/ S11192-011-0398-6

[12]. Lou, Y., & Obukhov, M. (2017). BDT: Gradient boosted decision tables for high accuracy and scoring efficiency. Proceedings of the ACM SIGKDD International Conference on Knowledge Di s c o v e r y a n d Data Mining, Part F129685, 1893–1901. https://doi.org/10.1145/3097983.3098175

[13]. Nguyen, H. D., Truong, G. T., & Shin, M. (2021). Development of extreme gradient boosting model for prediction of punching shear resistance of r/c interior slabs. Engineering Structures, 235, 112067. https://doi.org/10.1016/J.ENGSTRUCT.2021.112067

[14]. Pradeep, K., Tinturosmin, C. R., Durom, S. S., & Anisha, G. S. (2020). Decision Tree Algorithms for Accurate Prediction of Movie Rating. Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020, 853–858. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000158

[15]. Safón, V. (2019). Inter-ranking reputational effects: an analysis of the Academic Ranking of World Universities (ARWU) and the Times Higher Education World University Rankings (T H E) . Reputational Relationship. Scientometrics, 121(2), 8 9 7 – 9 1 5 . h t t p s : / / d o i . o r g / 1 0 . 1 0 0 7 / S11192-019-03214-9/METRICS

[16]. Udupi, P. K., Dattana, V., Netravathi, P. S., & Pandey, J. (2023). Predicting Global Ranking of Universities across the World Using Machine Learning Regression Technique. SHS Web of Conferences, 156, 04001. https://doi.org/10.1051/shsconf/202315604001

[17]. Wu, Q., Christopher, ·, Burges, J. C., Svore, K. M., & Gao, J. (n.d.). Ranking, Boosting, and Model Adaptation.

[18]. Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, 308–324. https://doi.org/10.1016/ J.TRC.2015.02.019

[19]. Zhou, Z.-H. (2021). Decision Trees. Machine Learning, 79–102. h t t p s : / /doi.org/ 10.1007/978-981-15-1967-3_4