ISSN: 3092-8729 | e-ISSN: 3092-8737

Ajayi Crowther Journal of Pure and Applied Sciences (ACJPAS)
https://acjpas.acu.edu.ng

Ajayi Crowther J. Pure Appl. Sci. 2022, 1(1), pp.10-17
https://doi.org/10.56534/acjpas.v1i1.67

Article

Performance Evaluation of a Student Web Portal Using Classification Models - Data Mining Technique

Olanrewaju S. S.[1]*, Osunade O.[2], Ayeni J. A.[3], Amoo J. O.[4]

[1] Department of Science Education, School of General Studies Education, Federal College of Education (Special), Oyo.
[2]Department of Computer Science, Faculty of Science, University of Ibadan.
[3]Department of Computer Sciences, Faculty of Natural Sciences, Ajayi Crowther University
[4]Department of Computer Science. School of Secondary Education Science Program, Federal College of Education (Special), Oyo

OSS-sundayolanrewaju50@gmail.com; OO-seyiosunade@gmail.com; OO-seyiosunade@gmail.com; AAO-lekanamoo247@gmail.com

* Correspondence: OSS-sundayolanrewaju50@gmail.com

Article history: Received: Jun. 5, 2022 Revised: Jul. 10, 2022 Accepted: Aug. 1, 2022 Published: Dec. 13, 2022.

Abstract

Data Mining is an extraction tool for analyzing and retrieving hidden predictive information from a large amount of dataset. It has been discovered that huge amounts of data were automatically saved by the web portal, which contains hidden information about the client that assessed the portal and lots of these data remain unused. To get required hidden information from such large data, a powerful tool known as weblog expert analyzer version 9.51 was used to determine the performance of a student web portal. This research aims at determining the performance evaluation of a student web portal using weblog analyzer and data mining technique to predict the algorithm that gives the best accuracy in terms of model building. In this work, the machine learning software (collection of machine learning algorithms) called WEKA was used and the training parameter was set to 10-fold cross-validation. Five decision algorithms which are Rep tree, Random tree, J48, LMT, and Hoeffding algorithms were used as classifiers, while TP, FP, Recall, Precision, and F-measure, were used as evaluation metrics. Therefore, the Random tree algorithm yielded a 99.0138% higher level of predictive accuracy and provided better classification accuracy when compared to the other classifiers. Finally, the results obtained can be used to enhance the effectiveness of the student web portal.

Keywords: Performance Analysis, Web Portal, Data Mining Technique, Predictive, Classifiers.

1. Introduction

Log files are files that keep the records of every action performed by the client on the portal [1]. Log files also known as plain text files that encapsulate various information about the client [2]. The web server keeps track of when a client accesses a particular Uniform Resource Locator (URL) [3]. The following information about the client accessing the portal is generated automatically and some of this information is the browser used, search engines referring the client to the site, duration and number of pages visited by the client, the URL that was requested for, the time of the request, operating system used, type of search engine and search term entered, words or phrases visitors have used in search queries, pages viewed, and other elements of usage statistics, URL that referred, and any errors that occurred [4].

Log files cannot be assessed by any internet user, but can only be assessed by the web administrator. Log files are often the only way to identify any activity of the client on the main server but as good as log files are, they occupy a reasonable amount of space on storage devices,

and this remains unused, but there is also a belief that it is useful for research in the field of computing which can be used to determine all the client activities [5].Tracking of log files has become a unique practice in which data are kept by any organization for tracking and analyzing clients activities on the web server. Therefore, log file analysis tools are very prominent tools and important for a clearer picture of the state of any portal or web server to be known. Weblog analysis tools are used for straining logs to establish useful log files, but it is challenging to extract the necessary hidden information from such a vast volume of data. Consequently, the creation of data mining algorithms is necessary and it can be used as a powerful tool for the analysis. This research work aims to analyze log files and determine the performance of a student web portal using data mining techniques. This article is divided as follows, in section I the introduction is provided, Literature review and related works are expressed in Sections II. Section III illustrates the methodology of the experimental studies while the results obtained and discussions are explained and presented in section IV. Finally, in Section V, the conclusion of the work is summarized.

In research conducted by Pratik et al. [6], they examined the web traffic on the primary web server at Monash University. The researchers assess many online log analyzers, including Google Analytics, AWStats, Weblog Expert, and Analog, that may be used to examine log files. The researchers also identified other techniques, protocols, and tools that can be used to examine online traffic, including Netflow, hybrid neuro-fuzzy systems, LQMs, and TDSs. The findings reveal that Monash University's primary web server receives more than 7 million hits every week. The focus of Mythili and Mohamed's [7], research was the use of classification algorithms to analyze student performance. To evaluate student performance, the researchers used five classification algorithms: J48, Random Forest, Multilayer Perceptron, IB1, and Decision Table. According to the findings, the Random Forest algorithm has a high predictive accuracy of 89.23% and constructs a model in 0 seconds. Neha and Jha's [8] study used an automated log analyzer program to assess user behavior from online access logs. The Web Log Expert program was utilized by the researchers to analyze user behavior on astrology websites, according to the findings, the most frequently accessed pages are the yantra page and lalkitab page, also the client with ip address 66.249.72.11 has relatively impact value, which means that the user has assessed the website mostly without wasting time. As a result of the analysis of site logs, the following general statistics were discovered: total hits were 24,820; total page views were 1,515; total visitors were 750, and total bandwidth was 305.98 MB. Instance-Based Learning, Perception-Based Learning, Bayesian Nets, Decision trees, and Rule learning were the five classification algorithms used in a study by Kotsiantis et al. [9] to predict students' performance in computer science. According to the findings, the Bayesian Nets outperformed other algorithms with high predicted accuracy of 74%. The comparative analysis of classification algorithms is presented by Sunita and Lobo [10] ADTree, Simple Cart, J48, ZeroR, and Naive Bayes Classification Algorithm were the five classification algorithms utilized by the researchers. The findings demonstrate that the ADTree classification method outperforms the other four classification algorithms for the Course Recommender System. Kaur et al. [11], give a comparative comparison of automated usability testing techniques together with an empirical performance evaluation of university websites. To evaluate the effectiveness of the website, the researchers employed the following tools: Pingdom, GTMetrix, website Grader, and Site speed tester. The parameters used are load time, page size, speed, performance, and the number of requests. The findings demonstrate that the University3 and University12 websites were measured using the Pingdom tool and performed at an 85 percent level. The website U5 has a GTMetrix tool evaluation score of 91 percent. Website Speed Checker tool offers U3 website the highest possible percentage of page loading speed of 89 percent, while Website Grader tool rates U6 with 90 percent performance. The primary goal of this study, according to Bharat and Manan [12] is to examine the effectiveness of two clustering algorithms in WEKA utilizing information gathered from banks that includes 600 records and 11 variables. The outcomes demonstrate that hierarchical clustering techniques are inferior to K-means algorithms. Table 1 depicts a list of related works.

Table 1: Summary of Related Works

S/N	Author & year	Title of Work	Methodology	Deficiency
1	Pratik *et al.* (2014)	Web traffic analysis of Monash University's main web server	Web Log Analyzer tool	The analyzer tool considered in this research work was not extensively discussed.
2	Mythili and Mohamed (2014)	An Analysis of performance using classification algorithms	WEKA	This research could have expanded its analysis, by using multiple clustering strategies and association rule mining for the student dataset.
3	Neha and Jha (2013)	Analyzing Users' Behavior from Web Access Logs using Automated Log Analyzer Tool	Web Log Expert Analyzer	The analysis was limited to a few statistics activities of the users more activities would have been considered.
4	Kotsiantis *et al.* (2004)	Prediction of students' performance using five classification algorithms	WEKA	The limitation of this research work is that the predictive accuracy of all the algorithms was less than 75%
5	Sunita and Lobo (2012)	Comparing classification algorithms	WEKA	This research work could have included more mining algorithms.
6	Kaur *et al.* (2016)	An Empirical Performance evaluation of Universities website	Pingdom, GTMetrix, Website Grader, Site Speed Checker Analyzer tools	This research work uses private universities alone, government established universities would have been included too.
7	Bharat and Manan (2012)	Comparing the performance of various clustering algorithms	Waikato Environment for Knowledge Analysis (WEKA)	Only two clustering algorithms were used, this research work would have used more than two clustering algorithms.

2. Materials and Methods

The performance analysis of the student web portal using the data mining technique was carried out, by collecting log files from the web administrator in one of the Federal Colleges of Education in Nigeria at their Information and Communication Technology Centre. The log files were compressed using a zip file and uploaded into the weblog expert analyzer version 9.51 as shown in figure 1. The generated datasets from the weblog analyzer were subjected to pre-processing which involved: data cleaning, session identification, and data conversion. The dataset was analyzed specifically using Waikato Environment for Knowledge Analysis, which involved data cleaning,

and session identification, and data translation. The dataset was split into ten sets, with 10-fold cross-validation set as the training parameter. The classifier was trained on nine sets during the initial run, and it was tested on a tenth set. The remaining nine sets were utilized as training sets and one set was held back as test data for the subsequent run. Performance analysis was conducted to determine the classification algorithm that has a high level of accuracy using the following metrics: Precision, Rate of true positives, Rate, of false positives, Recall, F-measure, Kappa statistic and Mean absolute error.

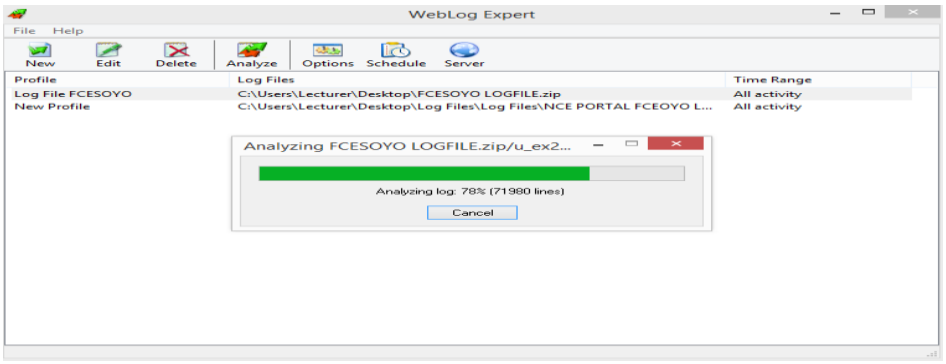


Figure 1: Analyzing Logs

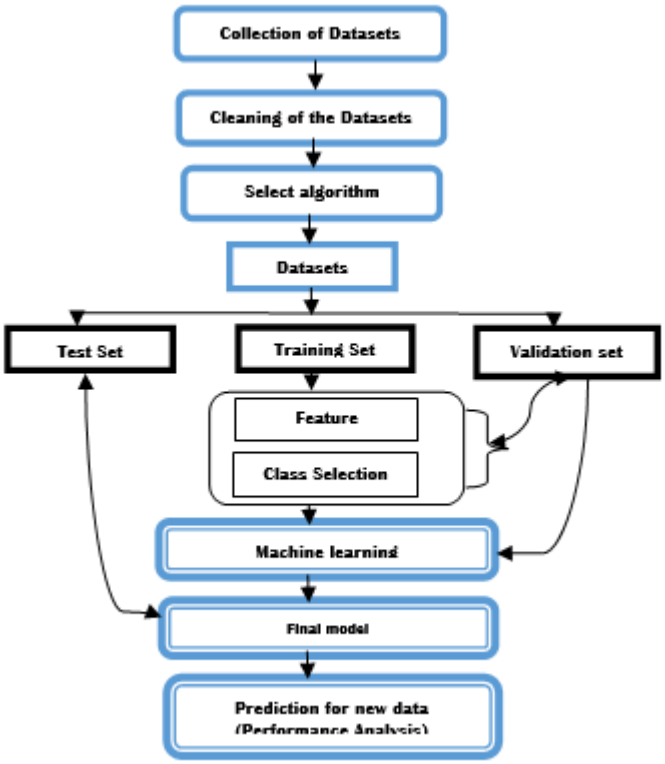


Figure 2: Model Block Diagram for solving classification problems [13]

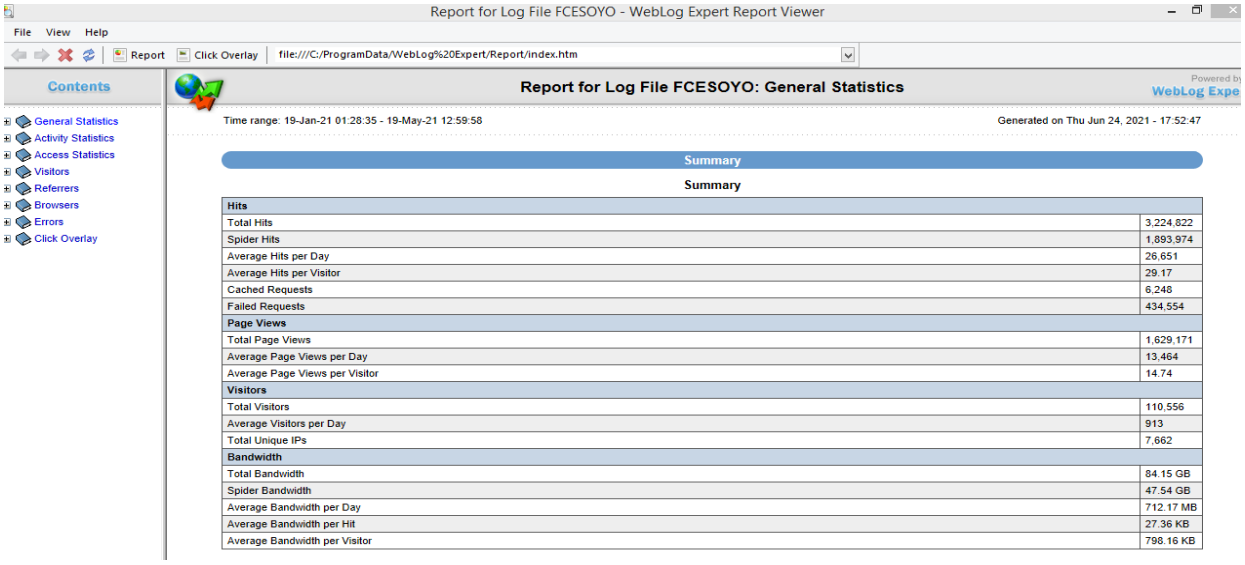
The above Figure 2 depicts the general approach for solving classification problems, the first phase has to do with collection of dataset analyzed using weblog expert analyzer version 9.51 to determine the performance of a student web portal, the second phase has to with data cleaning which involves extraction or removal of tracked data in weblogs that are not useful for mining purposes. The third phase has to do with selection of decision algorithms to be used for the experiment; the fourth phase deals with subjection of dataset to training, test and validation. Training set contain the complete training dataset in which features can be extracted and train to fit a model. Validation set has to do with choosing of right parameters for prediction, a training set can be divided into train set and validation set, the models can be trained using the validation test results when subjected to data mining. Therefore as soon as model is created, it can be applied to

the testing set to make predictions. After the whole processes, prediction for new data will follow which is known as performance analysis.

3. Results and Discussion

Collected Log files from the web administrator of one of the Federal College of Education in Nigeria were analyzed using the weblog Expert analyzer. The log file of data from 31st January 2021 to 28th May 2021 was collected for the purpose of this research. Table 2 presents the general statistics obtained after analyzing files. This shows the website's results analysis and provides a summary of the actions taken on the website by the clients. The results of general statistics, shows that there are 3,224,822 hits, 110,556 visitors, 84,15GB Bandwidth and 1,629,171 page views. Table 3 shows Monthly Statistics of the Website usage, from Table 3 it can be deduced that the month of May has the highest number of hits, page views, visitors and bandwidth. Figure 3 shows daily used operating system as well as most used operating systems by the visitors. It was observed that Hits for Android OS are quite more than any other operating systems. Table 4 displays the various types of errors that occur daily. The table clearly shows that the most common error is 404 error, meaning (Not found).

Table 2: General Statistics Report



Hits	
Total Hits	3,224,822
Spider Hits	1,893,974
Average Hits per Day	26,651
Average Hits per Visitor	29.17
Cached Requests	6,248
Failed Requests	434,554
Page Views	
Total Page Views	1,629,171
Average Page Views per Day	13,464
Average Page Views per Visitor	14.74
Visitors	
Total Visitors	110,556
Average Visitors per Day	913
Total Unique IPs	7,662
Bandwidth	
Total Bandwidth	84.15 GB
Spider Bandwidth	47.54 GB
Average Bandwidth per Day	712.17 MB
Average Bandwidth per Hit	27.36 KB
Average Bandwidth per Visitor	798.16 KB

Table 3: Monthly Statistics of the Website Usage

Month	Hits	Page Views	Visitors	Bandwidth (KB)
Jan 2021	1,684	554	50	32,315
Feb 2021	1,760	548	61	32,267
Mar 2021	2,549	810	120	49,494
Apr 2021	411,883	157,408	21,390	14,212,565
May 2021	2,806,946	1,469,851	88,935	73,914,276
Total	**3,224,822**	**1,629,171**	**110,556**	**88,240,919**

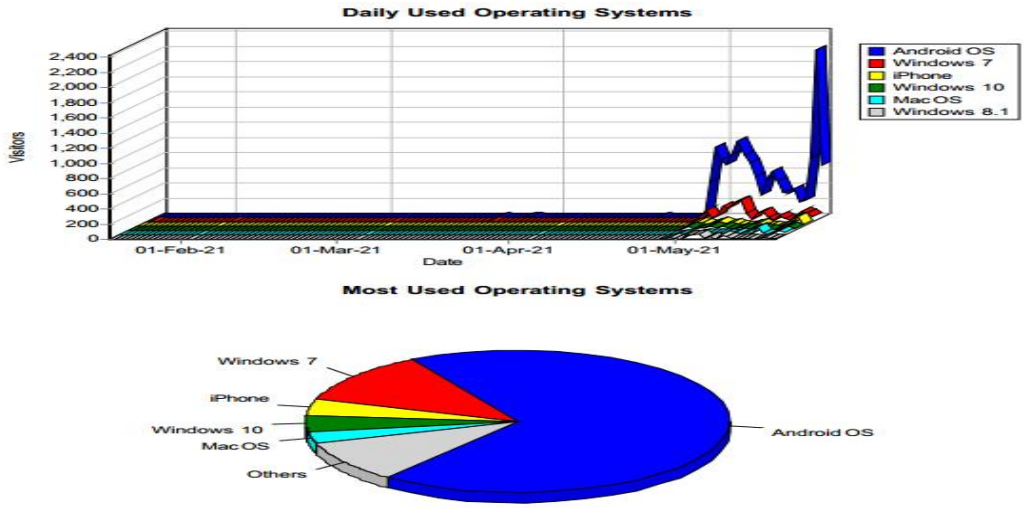


Figure 3: Daily/Most Used Operating System

Table 4: Frequent error type

	Error	Hits
1	404 Not Found	433,399
2	403 Forbidden	688
3	500 Internal Server Error	459
4	401 Unauthorized	7
5	400 Bad Request	1
	Total	**434,554**

3.1. Experimental Results of Learning Algorithms

The text option used for the analysis of the dataset is the cross-validation method. The comparative analysis of various decision trees was carried out, but the simulation result of the best classifier was shown in figure 4.

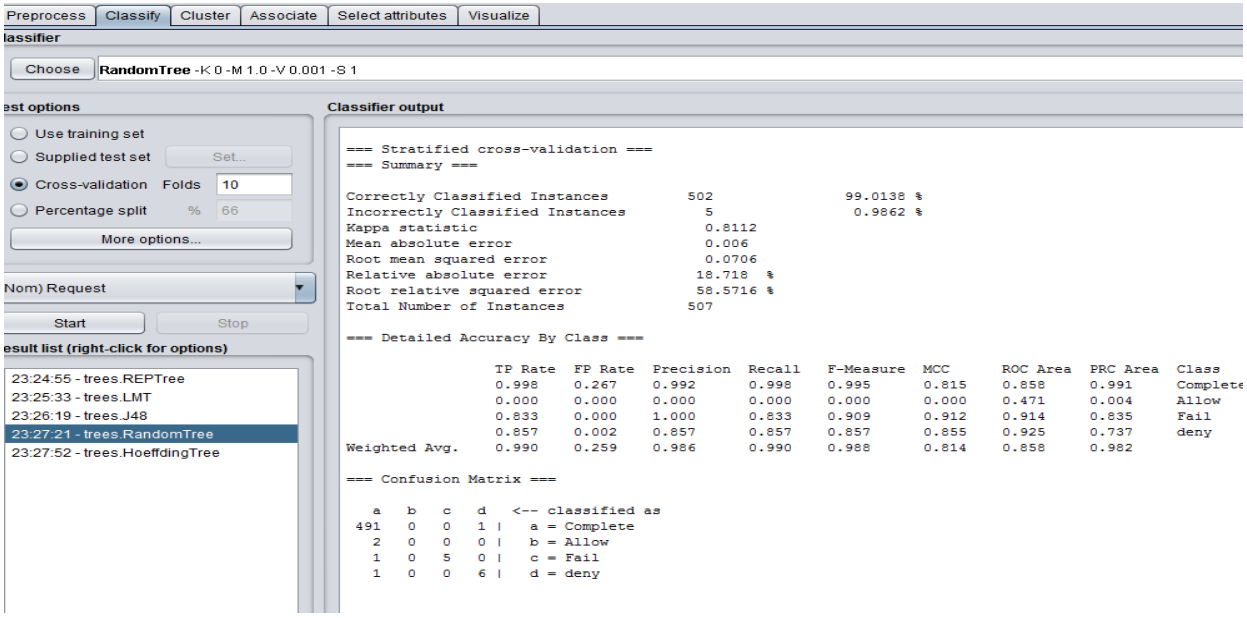


Figure 4: Simulation Result of Random Tree Algorithm

3.2. Performance Analysis of Algorithms

Table 5 depicts the experimental results obtained from the various classification algorithms in determining the best algorithms using the time taken to build the model and the accuracy. From table 5 below, it is observed that both the Rep tree and LMT classification algorithms give the same accuracy of 98.4221% however, the Rep tree builds the model in as little as 0.02 seconds, and the LMT builds the model in as much as 1.36 seconds. The random tree algorithm gives the highest accuracy of 99.0138% with a build time of 0.02 seconds, J48 classifier builds a model in 0.02 seconds and achieves an accuracy of 98.8166%. Hoeffding Tree has the lowest accuracy of 97.0414% and builds a classification model in 0.05 seconds. Therefore, the random tree algorithm gives the best accuracy in terms of model building.

Table 5: Table showing results of Performance Analysis of Algorithms

Decision tree	Accuracy	Time is taken to Build the Model
Rep tree	98.4221 %	0.02 seconds
Random tree	99.0138 %	0.02 seconds
J48	98.8166 %	0.02 seconds
LMT	98.4221 %	1.36 seconds
Hoeffding Tree	97.0414 %	0.05 seconds

3.3. Evaluation Results in terms of performance metrics

The precision, recall, f-measure, kappa statistic, mean absolute error, and true positive and false positive rates of several classifier models are shown in Table 6. It is evident from this table that the values of 0.998, 0.267, 0.992, 0.998, 0.815, 0.8112, and 0.006 have the highest true positive rate, lowest false positive rate, maximum recall and precision, lowest F-Measure, highest Kappa statistic, and lowest mean absolute error, respectively. Thus, it has been noted from Table 5 that the classifier Model created using a Random tree classification method delivers an effective forecast of the student's portal with 99.0138 percent. Thus it has been observed from Table 5 that the classifier Model designed using a Random tree classification algorithm provides an efficient prediction of students portal with 99.0138 %.

4. Conclusions

This paper examined the student web portals performance using weblog analyzer version 9.51 and data mining techniques, and the simulation results of several classifications were used to identify the optimum algorithm with the higher level of accuracy. In conclusion, the Random tree algorithm performs better than the other classifiers in terms of classification accuracy. The results obtained can be used to enhance the effectiveness of the portal. Also, Random tree algorithm is the most suitable and effective algorithm for model prediction performance.

Funding: Not applicable

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: Not applicable

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] K. R. Navin, A. K. Tyagi, and M. W. Solanki, "Analysis of Server Log by Web Usage Mining for Website Improvement", *International Journal of Computer Science Issues*, vol. 7, no. 4, pp. 17-21, 2010. [Online]. Available: www.IJCSI.org. [Accessed June 10, 2021].

[2] M. Chen, A. Zheng, J. Lloyd, M. Jordan, and E. Brewer, "Failure diagnosis using decision trees", International Conference on Autonomic Computing Proceedings, pp. 36-43, (2004). [Online]. Available: http://ieeexplore.ieee.org/document/1301345/. [Accessed June 10, 2021].

[3] M. P. Yadav and P. K. Keserwani, "An efficient web mining algorithm for Web Log analysis", 1st International Conference on Recent Advances in Information Technology, pp. 607-613, 2012. [Online]. Available: E-Web Miner, *IEEE*, https://researchgate.net [Accessed June 11, 2021].

[4] M. Munk, J. Kapusta and P. Svec, "Data preprocessing evaluation for web log mining: Reconstruction of activities of a web visitor", *Procedia Computer Science*, vol. 1, no. 1, pp. 2273- 2280, 2010. [Online]. Available: http://www.scopus.com/ inward/record. url?e id=2-s2.078649584518 &partner ID=t ZOtx 3y1. [Accessed June 11, 2021].

[5] K. J. Ratnesh, R. S. Kasana1 and J. Suresh, "Efficient Web Log Mining using Doubly Linked Tree", *International Journal of Computer Science and Information Security*, vol. 3. No 1, 2009. [Online]. Available: https://www.academia. edu/122380 77/ international_Journal_ of_ Computer_Science_July_2009. [Accessed June 13, 2021].

[6] V. Pratik, N. M. Tarbani and Ingalkar, "A study of web traffic analysis", *International Journal of Computer Science and Mobile* Computing, vol. 3, no. 3, pp. 900-907, 2014. [Online]. Available: www.ijcsmc.com. [Accessed June 13, 2021].

[7] M. S. Mythili and A. R. Mohamed, "An Analysis of students performance using Classification algorithms", *IOSR Journal of Computer Engineering*, vol. 16, no. 1, pp. 63-69, 2014. [Online]. Available: www.iosrjournals.org. [Accessed June 12, 2021].

[8] G. Neha and C. K. Jha, "Analyzing User's Behavior from Web Access Logs using Automated Log Analyzer Tool", *International Journal of Computer Applications*, vol. 62, no. 2, pp. 0975-8887, 2013. [Online]. Available: https://ui.adsabs. harvard. edu/abs/ 20131JCA...62b..29G/. [Accessed June 12, 2021].

[9] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Prediction of Students' Performance in Distance Learning Using Machine Learning Techniques", *Applied Artificial Intelligence*, vol. 8, no. 5, pp. 411-426, 2004. [Online]. Available: https://www. researchgate. net/ publication/228084511. [Accessed June 15, 2021].

[10] B. A. Sunita and L. M. R. J Lobo, "Comparative study of classification algorithms", *International Journal of Information Technology and Knowledge Management*, vol. 5, no. 2, pp. 239-243, 2012. [Online]. Available: https://www.csjournals.com/?cat=3. [Accessed July 2, 2021].

[11] S. Kaur, I. K. Gujral and K. Kaur, "An Empirical Performance Evaluation of Universities Website", *International Journal of Computer Applications*, vol. 146, no.15, pp. 0975- 8887, 2016. [Online]. Available: http://www.ijcaonline.org/archives. [Accessed June 12, 2021].

[12] C. Bharat and P. A. Manan , " Comparative study of clustering algorithms using Weka tools", *International Journal of Application or Innovation in Engineering and Management*, vol. 1, no. 1, pp. 154-158, 2012. [Online]. Available: www.ijaiem.org. [Accessed Aug. 2, 2021].

[13] https://www.nature.com/articles/s41524-019-0221-0. [Online].Available: [Accessed June 2, 2021].

Cite article as:
Olanrewaju S. S., Osunade O., Ayeni J. A. and Amoo A. O. (2022). Performance Evaluation of a Student Web Portal Using Classification Models - Data Mining Technique. *Ajayi Crowther J. Pure Appl. Sci.* 1(1): 10-17. doi.: https://doi.org/10.56534/acjpas.v1i1.67