Journal of Applied Sciences, Information and Computing

Volume 5, Issue 1, June 2024

School of Mathematics and Computing, Kampala International University

ISSN: 1813-3509 https://doi.org/10.59568/JASIC-2024-5-1-06

Academic performance prediction using Machine Learning algorithms

[1]O. Owolabi, [2]R. Obadaki

[1]Professor, Department of Computer Science
University of Abuja, FCT, PMB 117, Abuja, Nigeria
olumide.owolabi@uniabuja.edu.ng, rahmahobadaki@gmail.com
[2]Corresponding Author: rahmahobadaki@gmail.com

Abstract

An excellent secondary school education becomes evident in students' performance after they graduate or further their education. No matter their career choice, they can genuinely excel if they can identify areas that require them to put in more effort to have an overall excellent performance in school. In Nigeria, several solutions address students' learning needs or the administrative needs of the schools. Still, no systems cater to analysing and monitoring students' performance, causing failures that can be averted. This dissertation reviews five different machine learning algorithms using data from students in public and private schools in rural and urban Nigeria to identify which algorithm performs best in predicting students' performance using the Waikato Environment for Knowledge Analysis tool for modelling and the cross-industry standard process for data mining (CRISP-DM) research methodology. The result shows the Decision Tree as the algorithm with the best performance for the dataset. It is recommended that the findings be used to build a system embedded into a school management or learning management software to enable students, parents, and teachers to channel the right resources into areas where it has been predicted that the student will underperform to change the narrative.

Keywords: Student Performance Prediction, Waikato Environment for Knowledge, machine learning algorithm.

1. Introduction:

The school system is an educational system for considerable data extraction. Over time, advancements in research have increasingly brought about new meaningful Information from the massive amounts of data generated. Therefore, it has become necessary to predict students' academic performance, which will help identify students' achievements and assist the teachers and school management in decision-making. Big Data provides educational institutions with the opportunity to improve Student outcomes and academic quality by analysing and utilising their information technology resources (Ramsey et al., 2020). The prediction will further help to identify which set of students would do well in the end-of-session examination or the middle of the term so that they can be recognised or awarded scholarships and also to identify the students who may fail in end-of-session examinations so that some form of remediation may be offered to them. Students must predict their academic performance to make informed decisions and improve their skills. This process

can be done through the use of a classification model. It is challenging to predict students' learning performance (Romero & Ventura, 2013). Identifying at-risk students and predicting their performance in academics are crucial steps toward turning them into self-regulated learners. This process can help educators identify underperforming pupils and prevent them from failing. Unfortunately, this can be challenging due to the different factors affecting a student's performance.

This study aims to improve the learning and teaching processes in schools. Through this process, we get to know the needs of students, and hence, we can fulfil those needs to get better results. This process can also identify students who need special attention from the teachers. Several algorithms are available to predict the performance of students. Some of them are Artificial Neural Networks (ANN), Clustering, Naive Bayes algorithm, and Decision Trees, most commonly applied to educational data to forecast students' behaviours and performances in school. The use of educational data comprises certain players in the field, including students, teachers, alum data, resource data, and the like. This data is mined and used to determine the patterns for decision-making.

The recent novel COVID-19 pandemic has led to a shift of paradigm and advances in the use of technology as the new norm of learning. The learning methods have significantly turned from physical classes and books to remote classroom sessions and digital resources. These digital resources range from gadgets to different kinds of social media platforms to the World Wide Web. The resources have the fullness of information available through the internet. Merging these educational resources with a solution that predicts how they will perform will help students, teachers, and parents channel suitable learning materials into improving the students' performance.

Accurate computation of student results is essential in any secondary school educational system. However, most recent research on predicting students' academic performance has several drawbacks, considering the paradigm shift in the current pandemic learning environment (Chango et al., 2021). This submission combines the traditional face-to-face teaching method with the online learning approach. Hence, less attention has been given to the mixed mode of learning. Predictive approaches neglect the effective utilisation of multimodal data collection in these blended environments. If standard features are considered under different conditions (Nigeria) in many secondary schools, it will cover the gap and fit into the current learning system. Rapid technological advancements have enabled individuals to capture all student behaviours, interactions, and actions while presenting their performances in a virtual and face-to-face classroom learning environment. Therefore, there is a need to capture critical features that can be efficiently applied in Nigeria, so data is gathered from schools in rural areas, urban areas, public schools and private schools in Nigeria to predict students' academic performance.

This dissertation aims to identify the best machine-learning method for analysing secondary school data in rural and urban schools in Nigeria. In addition to the algorithm's selection, the research explores the various factors that can affect the outcome. The data gathered only focuses on the 2020/2021 academic session and only on selected schools in the country. Also, the data mining technique is limited to the five (5) models: Decision Tree, Support Vector Machine, Random Forest, Naive Bayes, and Logistic Regression.

Performance prediction is a process that aims to provide teachers and students with data that will help measure their progress and identify potential failures. It also allows them to take appropriate measures to prevent these from happening.

The objectives of this work are to:

i. Provide insight into students' academic performance prediction across selected schools in Nigeria
ii. Use selected classification algorithms to build a model for forecasting student academic achievement.
iii. Predict the performance of students at risk of failure using a machine learning algorithm.
iv. Compare with previous work and ensure an increase in the performance of selected algorithms.

This dissertation intends to build a model using machine learning algorithms and compare them with each other to show how they performed on similar datasets. This study is limited to the relevant features in our dataset from selected schools in Nigeria's urban and rural areas. Therefore, our model may only partially capture all the features contributing to student performance. However, more emphasis will be on the features of student data collected from the school system.

2. Review of Literature:

Based on the literature, opportunities to tap into some of these factors were identified. For example, at the enrolment stage, there is a need to capture personal data such as health status that can affect the student. However, many schools collect some of these data but must refer to them when channelling appropriate support or resources to students (Kaunang & Rotikan, 2018). Even though the academic data presented are in several granularities and formats, they can be put together from different sources. So, to enhance the prediction staging of models, a critical investigation of each feature's impact was considered. More so, the impact of several teaching methods and scoring styles could arise when subjects have to be segmented into a series of modules. Each module's instructor should be tactfully scrutinised (Hutchatai & Jitti, 2018).

 https://doi.org/10.59568/JASIC-2024-5-1-06

s/n	Author & Year	Title of Research	Outcome Of Research	Limitation of Research
1	Albreiki et al (2021)	A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques	The use of static and dynamic data with Decision Tree, Naive Bayes and Support Vector Machine learning algorithms. Naive Bayes performed best.	Does not consider the dynamic nature of student performance.
2	Rodríguez-Hernández et al (2021)	Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation	Artificial neural networks outperform other machine-learning algorithms in evaluation metrics such as the recall and the F1 score.	Relevant information represented via a traditional and widely reported indicator such as high school grade point average (HSGPA) could not be included. All of the information regarding students' socioeconomic conditions was self-reported by the students.
3	Nedeva & Pehlivanova (2021)	Students' performance analyses using machine learning algorithms in WEKA.	The efficiency of the four classification algorithms were compared, namely BN, MLP, SMO and J48. The indicators TP Rate, Precision, F-Measure, Accuracy and error measurements, MAE, RMSE, RAE, and RRSE were used. The processing is done with Weka open-source software. The obtained results show that the MLP algorithm is the best for the used data.	The focus is on one institution, and the and the use of fewer attributes.

Figure 1: Sample of Reviewed Materials

3. Research Methodology:

Our proposed prediction system will utilise machine learning techniques in the WEKA tool to process the data, extract relevant attributes to our study, and then build the proposed model. The algorithms used in this process include the K-Nearest Neighbour, Decision Tree, Naïve Bayes, and Random Forest. After creating a prediction model, the researchers will train it to predict the target value of the given variable, the "result" column in the dataset, by combining the various independent variables and features.

The data utilised in the proposed system were collected from a software development company, FlexiSAF Edusoft Limited, which manages a large dataset of students' academic records for both public and private schools in urban and rural parts of Nigeria. The dataset composes secondary students' academic records such as the subject name, class ID, student ID, Subject Code, Assessment ID, Score, Comment, School Type, etc. This acquired dataset comprises 444,975 instances of historical data of students, with 17 attributes.

	A	B	C	D	E	F	G	H	I	J
1	subject_name	arm_id	klass_id	session_id	term_id	student_ic	subject_	assessmer	assessme	score
2	Mathematics	A	JSS 2	2020/2021	FIRST	FGC/19/0(	113	FGNAFS_C	FGNAFS_(	15
3	Mathematics	A	JSS 2	2020/2021	FIRST	FGC/19/0(	113	FGNAFS_C	FGNAFS_(	10
4	Mathematics	A	JSS 2	2020/2021	FIRST	FGC/19/0(	113	FGNAFS_E	FGNAFS_(	49
5	Social/Studies	A	JSS 2	2020/2021	FIRST	FGC/19/0(	123	FGNAFS_C	FGNAFS_(	20
6	Social/Studies	A	JSS 2	2020/2021	FIRST	FGC/19/0(	123	FGNAFS_C	FGNAFS_(	20
7	Social/Studies	A	JSS 2	2020/2021	FIRST	FGC/19/0(	123	FGNAFS_E	FGNAFS_(	44
8	Agric/ Science	A	JSS 2	2020/2021	FIRST	FGC/19/0(	124	FGNAFS_C	FGNAFS_(	17
9	Agric/ Science	A	JSS 2	2020/2021	FIRST	FGC/19/0(	124	FGNAFS_C	FGNAFS_(	9
10	Agric/ Science	A	JSS 2	2020/2021	FIRST	FGC/19/0(	124	FGNAFS_E	FGNAFS_(	23
11	Computer/Stud	A	JSS 2	2020/2021	FIRST	FGC/19/0(	128	FGNAFS_C	FGNAFS_(	15
12	Computer/Stud	A	JSS 2	2020/2021	FIRST	FGC/19/0(	128	FGNAFS_C	FGNAFS_(	12
13	Computer/Stud	A	JSS 2	2020/2021	FIRST	FGC/19/0(	128	FGNAFS_E	FGNAFS_(	44

Figure 2: Sample of Dataset

This is the first step to cleaning up our dataset to improve the quality of our model. Data pre-processing is a vital step in extracting and improving the data quality. This study will utilise the WEKA Explorer software. We begin by selecting the source data file from the local machine. After we have loaded the data into the WEKA Explorer interface, we can refine data by selecting various displayed options and selecting or removing attributes as needed. The tool allows us to apply filters to the imported dataset, such as "Replace Missing Values", when necessary. This process is referred to as the data cleaning process. The pre-processed version of our dataset in the WEKA Explorer is shown in Figure 3. The Explorer dashboard's left-hand side shows details about current relationships, such as the relation name and number of records. The details about selected attributes and their type are shown on the right-hand side. Figure 3.2 shows the selected attributes, which is the visualisation screen of all the attributes displayed at the bottom right of the screen. Finally, we have the status of our data, and in case of any error, we can view the log to identify the kind of error. In WEKA, the train and test datasets must have the same header.

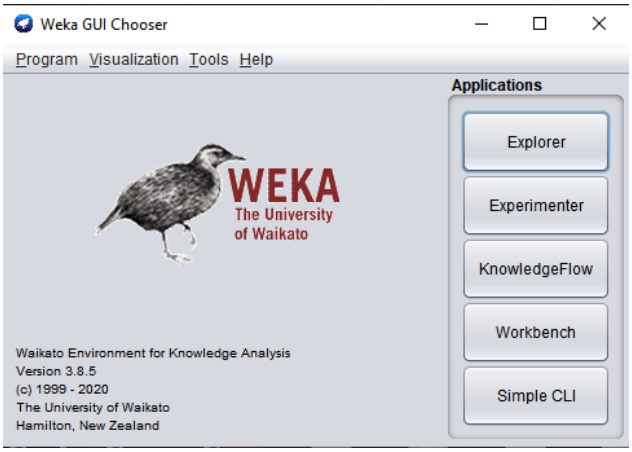

Figure 3: WEKA GUI

Figure 4 shows the Explorer interface after clicking the Explorer button on the home page. The data is currently loaded here, and we can see 11 attributes from the attributes pane after the data is processed.

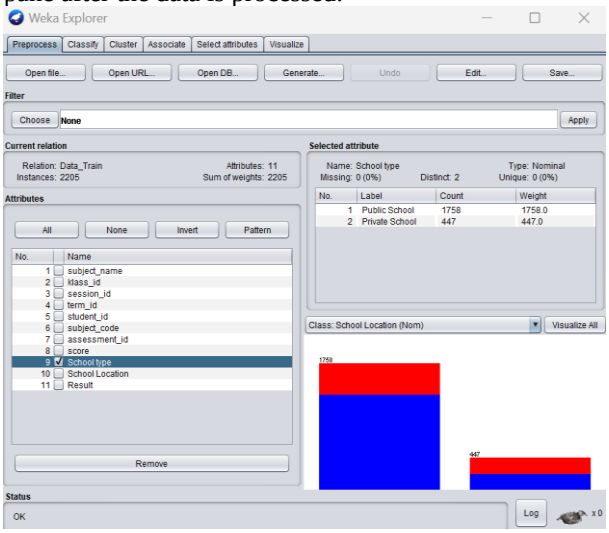

Figure 4: WEKA Explorer Interface

 https://doi.org/10.59568/JASIC-2024-5-1-06

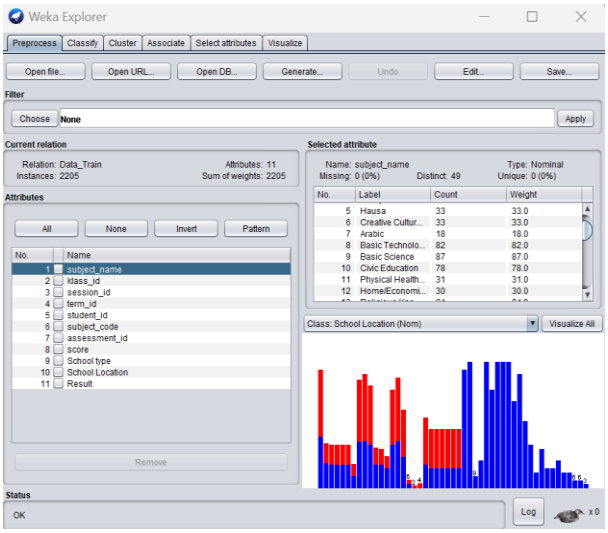

Figure 5: Pre-processing of Dataset

Figure 6 comprises the breakdown of each attribute name, the description, the type and the value for each identified attribute.

Attributes	Description	Type	Value
Subject Name	The name of the subject offered by the student	Nominal	Mathematics, Social Studies, Computer Science, etc
klass_id	The current class of student	Nominal	JS1, JS2, JS3, SS1, SS2, and SS3
session_id	This is used to identify the academic session.	Numeric	2020/2021
term_id	This is used to identify the academic period in a session.	Nominal	First, Second, and Third
student_id	The student is identified using a student ID.	Nominal	FGC/19/001, FGGC/BWR/12280, etc
subject_code	This is the code used to identify the subject.	Numeric	113, 128, 130, etc

Figure 6: Sample Characteristics of collected dataset

We are using five (5) different algorithms to build the models for classification and prediction. The algorithms chosen are the Naive Bayes, Decision Tree (J48), Logistic Regression, Random Forest, and Support Vector Machine. In addition, we shall use Waikato Environment for Knowledge Analysis (WEKA) version 3.8.5. After the pre-processing of our dataset, the following is the process for building our model:

a. Open the WEKA tool and select the Explorer Interface
b. Click the Open file button on the pre-processing tab to import the training dataset. CSV files can be converted into ARFF formats automatically. You can specify a .csv file as the output format, which will be converted into the ARFF format.
c. Delete irrelevant attributes by checking the box and using the Remove button.
d. Click the Classify tab to begin building our model.
e. Select the relevant algorithm from the different classifiers.
f. From the Test options panel, select “Use training set.”
g. Open the “more option” to choose the “Output Evaluation” as plaintext.
h. Select the target column
i. Click on start to build the model.
j. Right-click on the “Result List” panel for options to save the trained model in a .model format.

Using the selected Algorithms, we shall use our model to predict students' academic performances as High, Medium or Low by analysing the dataset comprising both public and private schools and including those from urban and rural areas. Furthermore, we shall use our model to predict students' overall performance based on all the relevant attributes.

4. Result and Discussion

This dissertation uses the WEKA tool to discuss and analyse the FlexiSAF Edusoft Limited data from data preprocessing until the desired models are built.

In this process, we clean up our dataset and make it suitable to be performed over a model with less effort. This is necessary to improve the quality of our model and achieve the desired outcome. In the raw format, our data is messy, with mixed data types and null and duplicate values. Also, using the WEKA tool, we applied a filter to replace missing values using the;
filter->Unsupervised->attributes->ReplaceMissingValues.

WEKA has a feature that allows us to identify the relevant features after loading the dataset from the WEKA explorer. We examined the impact of each variable on students' prediction success to better understand the significance of each variable to the output variable. To accomplish this, we ran several tests using feature selection techniques, including Gain Ratio Feature Evaluator, Info Gain Feature Evaluator, and One R Feature Evaluator. The attributes with the highest ratings were chosen. The results are shown in Figures 7, 8, and 9.

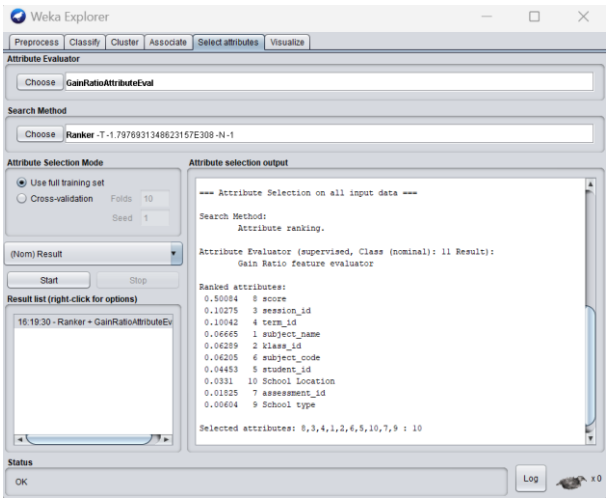

Figure 7: Gain Ratio Feature Evaluation

 https://doi.org/10.59568/JASIC-2024-5-1-06

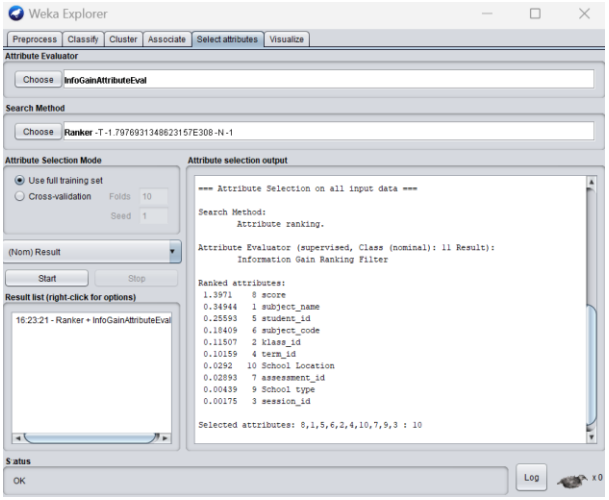

Figure 8: Information Gain Feature Evaluator

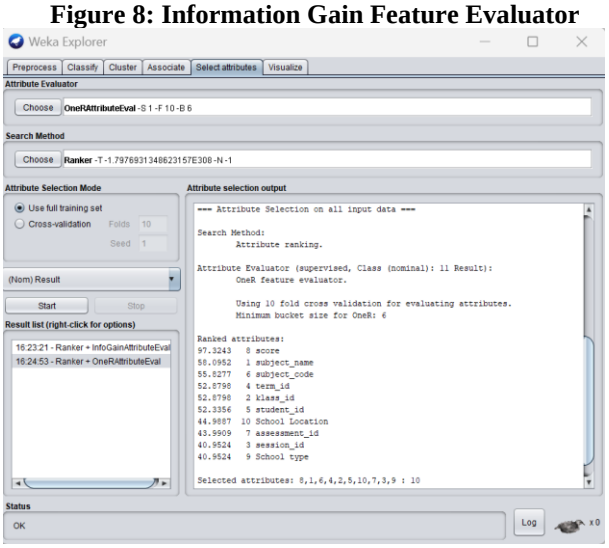

Figure 9: OneR Feature Evaluator

The Training Set is a portion of the universally sourced data used to train our machine-learning model. Machine learning algorithms are taught to make predictions or perform a desired task by feeding them with the training datasets. To build a model, we use a training set to put our algorithm to the test. A testing set is used to put our model through its paces and determine the accuracy of our predictions. The universal dataset is divided into training and test sets to estimate our model's performance. Our Training and testing set were converted and stored separately in a ".arff" format using the WEKA tool. We trained our models using the entire training dataset, saved the models in a ".model" format, imported our testing set and re-evaluated all the models to make predictions.

The evaluation measures how it performed during the implementation. It includes a classification report, an accuracy score, and a confusion matrix. These components are used to evaluate the model's overall performance. Precision indicates the percentage of the classifier's predicted items that are relevant, and recall indicates the percentage of actual relevance.

A model's performance primarily depends on the type and quality of the data.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1_Measure (%)	MCC	ROC Area	PRC Area
NB	65.2	64.8	65.2	64.8	46.2	81.1	68.5
DT	99.3	99.3	99.3	99.3	99	99.5	99.1
RF	90.3	90.4	90.3	90.3	85.2	98.1	96.9
LR	95	95	95	95	92.2	98.4	96.3
SVM	86.1	86.4	86.1	86.2	78.5	91.8	81.5

Figure 10: Comparison of the different performance classifiers

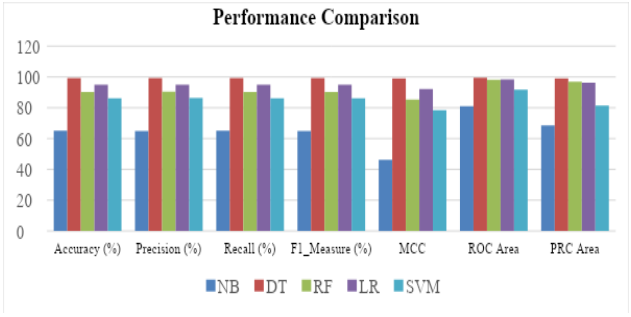

Figure 11: Performance Comparison of the Different Classifiers

Out of the 2,205 instances of our dataset, 903 instances showed high performance, 798 instances showed medium performance and 504 instances showed low performance. The distribution of data is shown in Figure 12.

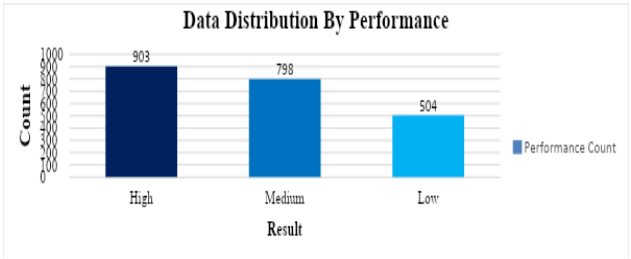

Figure 12: Data Distribution by Performance

The results show no significant difference between the performance of students in the public school and that of the private school. However, our result will help the school improve students' performance based on performance distribution, as shown in Figure 12. Out of the 2,205 instances of our dataset, there was a total record of 1758 from the public school and 447 from the private school. The distribution of the data is shown in Figure 13.

 https://doi.org/10.59568/JASIC-2024-5-1-06

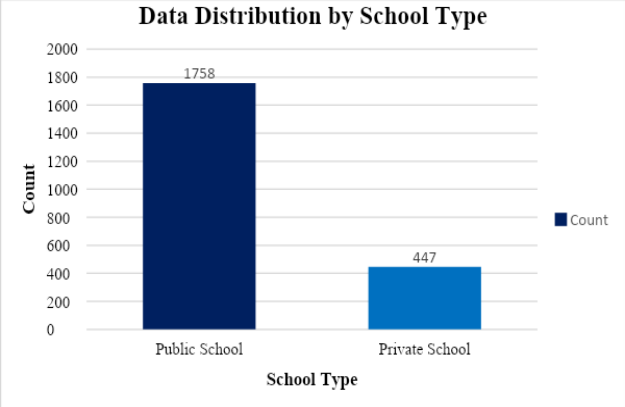

Figure 13: Data distribution by school type

There is no sufficient evidence to ascertain the difference in students' performance according to their location, either in urban or rural areas, due to the nature of our dataset. However, our result will help the school improve students' performance based on performance distribution, as shown in Figure 11. Out of the 2205 instances of our dataset, there is a total record of 1542 instances from the rural area and 663 from the urban area. The distribution of the sourced data is shown in Figure 14.

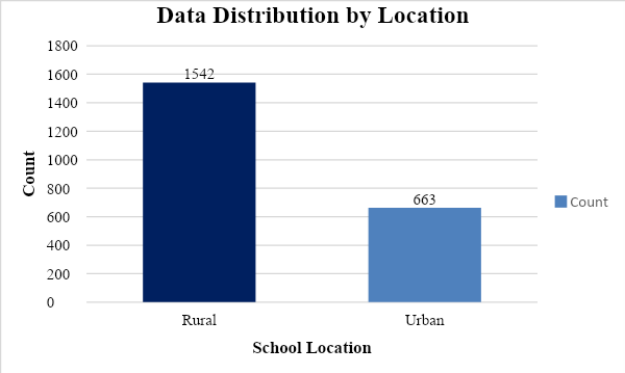

Figure 14 Data Distribution by Location

5. Conclusion

This study designed models that can efficiently predict overall academic performance for secondary school students in Nigeria. We created five classifiers using WEKA software to obtain our model's best-performing machine learning algorithm: Decision Tree, Support vector machine, Logistic Regression, Naïve Bayes, and Random Forest Algorithms. The trained model helps us determine students' performance in various subjects and subsequently predict the students' overall academic performance. This will further allow us to improve the performance of every underperforming student. The dataset was collected from FlexiSAF Edusoft Limited for the 2020/2021 academic session. It consists of students' academic information recorded between 2020 and 2021 and has 11 attributes. The Decision Tree (DT) performance outperforms the rest in predicting pupils' academic performance in Nigeria's secondary schools.

This study examined the problem of accurate determination and the best methods of improving students' academic performance within the school systems, especially the secondary schools in Nigeria so that methods can be developed to help improve each student's overall academic performance. The results show that the Decision Tree Algorithm performed better than other models. Our approach is adaptable and dynamic within the context of Nigeria's education system. This study presented several models that can work best with the nature of the data collected. This research will continue to spawn further research in this area.

The recommendation made towards this dissertation is that this model be used to create a student recommender system that helps identify the specific areas that need to be improved upon by each student. In addition, it will help identify which subject should be improved upon and what method to deploy or apply for such improvement.

6. References

[1] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic Literature review of students' performance Prediction using machine learning techniques. EducationSciences, 11(9), 552. doi.org/10.3390/educsci11090552

[2] Chango, W., Cerezo, R., & Romero, C. (2021). Multi-Source and Multimodal Data Fusion for Predicting Academic Performance in Blended Learning University Courses. Computers and Electrical Engineering, 106908(89), 1–13 https://doi.org/10.1016/j.compeleceng.2020.106908

[3] Hutchatai, C., & Jitti, N. (2018). Student performance Prediction model for early identification of at-risk students in traditional classroom settings.MEDES'18: Proceedings of the 10th International Conference on Management of Digital Eco Systems. 239-245. doi.org/10.1145/3281375.3281403

[4] Kaunang, F. J., & Rotikan, R. (2018). Students' academic Performance prediction using data mining. In the 2018 Third International Conference on Informatics and Computing (ICIC) (pp. 1-5). IEEE.

[5] Nedeva, V. & Pehlivanova, T. (2021). Students Performance analyses using machine learning Algorithms in WEKA. *IOP Conf. Ser.: Mater. Sci. Eng.* 1031 (012061**).** **doi.org**/10.1088/1757-899X/1031/1/012061

[6] Olalude, G., Amusan, A., Adeshina, I. (2021). Analysis of Students' academic performance using traditional and machine learning classifiers. Kaduna State University Journal of Mathematical Science, 2(2).

[7] Prasana laksmi, B., & Farouk, A. (2019). Classification and Prediction of student academic performance in KingKhalid University. A Machine Learning Approach. Indian Journal of Science and Technology, 12(14). doi: 10.17485/ijst/2019/v12i14/142792

[8] Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical casual models, with an application to functional magnetic resonance images. International Journal of Data Science and Analytics, 3, 121-129.

[9] Rodríguez - Hernández, C.F., Musso, M., Kyndt.E., Eduardo Cascallar, E. (2021). Artificialneural networks in academic performance Prediction: Systematic implementation and predictor evaluation. Computers and Education: Artificial Intelligence, 2 (100018). doi.org/10.1016/j.caeai.2021.100018

[10] Romero, C., & Ventura, S. (2013). Data mining in Education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

[11] Zulfiker, S. M., Nasrin Kabir, N., Biswas, A. A., Chakraborty, P., & Rahman, M. M. (2020). Predicting students' performance of the Private universities of Bangladesh using machine learning approaches. International Journal of Advanced Computer Science and Applications, 11(3).