iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.55054

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Comparative Analysis of Three Data Mining Tools in the Allocation of Senior Secondary Students into Departments for Career Path

Akinsola Adeniyi F.[1], Sokunbi.Michael.A[2], Onadokun I.O[3], Okenwa-Martins C.Q[4]
[1, 2, 3, 4]*Yaba College of Technology, Computer Tech. Dept., Yaba, Lagos Nigeria*

Abstract: ***Choosing a career path in life begins with studying the right subject combination at the secondary school level. The Nigerian educational system of 6-3-3-4 which makes a student to spend 6 years in pry education, 3 years in junior secondary school, 3 years in senior secondary school and 4 years in the university makes it mandatory for such a student to choose a department of study at the point of entry into the senior secondary school (SSS). It is at this point that a decision is made consciously or unconsciously by the students about life career as the choice of department determines the choice of course of study at the university level later. Students and their parents/guardians often take this decision without any scientific guide which in most cases leads to a wrong choice of career path. In this study, three data mining tools, WEKA, RapidMiner and Orange were employed with three algorithms each to determine the most appropriate tool with the best result in the allocation of senior secondary school (SSS) students to various departments of study. The best algorithm methodology that classified the dataset in WEKA was Random Forest with an accuracy of 100% predicting 308 students correctly. In RapidMiner the best algorithm methodology was Naïve Bayes with an accuracy of 82.9% correctly predicting 73 students. Thereafter, Orange gave the best algorithm methodology to be Random Forest at 98.7333% predicting 304 students accurately. Our study shows that the optimum algorithm suited for the application software implementation to allocate SS1 Students into Department was Random Forest, having highest rates in the Accuracy. Though Orange has additional feature of being able to visualize the output of all the three results in one interface at a glance, and also shows outcome visualizations in various plots and graphs, WEKA's highest predictive measure of 100% places it above all and makes it the tool of choice with Random Forest being the best algorithm.***

Keywords: Career Path, Data Mining Tools, Tools Comparison

I. INTRODUCTION

To undergo tertiary educational system in preparation for choice of a career in life, certain subject-combination are required. Prospective scholars apply for a course of study in a university or polytechnic based on qualification to study the desired course having satisfied the admission requirements. Uppermost among these admission requirements is to have studied required Ordinary Level (O' Level) subjects in the secondary school which could be Science, Art, Social Science or Commercial based. In Nigeria educational system, division into these fields of study are usually done at the entry point into Senior Secondary School. However, the technique for this division into various fields are not based on any scientific technique; it is mostly done on sentiments of what the parent/guardian want their wards to be or on preference for a particular field not based on competences on the underlying subjects. The multiplier effects and consequences of this error sometimes create serious permanent career challenges leading to incompetence in professional practice. This Study used the academic data of scholars who are in the Junior Secondary School of Yaba College of Technology Staff School, preparing to migrate into the Senior Secondary School level to predict their allocation into various academic departments such as Science, Art, Social Science or Commercial.

II. RELATED WORKS

"Modelling and Predicting Student's Academic Performance using Classification Data Mining Techniques" (2020) by Raza Hasan et al. In order to create a classification model to forecast student academic achievement, this study used WEKA. An accuracy of 80.63% was attained by the model.

Kaur, Singh, and Josan (2015) concentrated on using several classification strategies to identify academics who would be slow learners.

They assembled a dataset of 152 students from a high school and used the WEKA tool to train and test the students' performance. The comparison of the various predictive methodologies tested revealed that Multilayer Perception had the highest prediction accuracy, at 75%.

A. A. Al-Amin et al (2019 In order to create a decision tree model to forecast student academic achievement, this study employed Orange. A dataset of 1000 students from a Pakistani institution was used to train the model. The student's grades, attendance, test results, and other details that may have been important to their academic achievement were included in the dataset.

Kapur, Ahluwalia, and Sathyaraj (2017) employed six data mining algorithms to predict student grades are Decision Tree, IBK, K-star, Naive Bayes, Naive Bayes Multiple Nominal, and Random Forest. With the use of the WEKA tool, they compiled a dataset of 480 records with 16 different attributes, and the results showed that Random Forest had the highest prediction result accuracy, at 76.667%.

Hussain, Dahan, Ba-Alwib, and Ribata (2018) used 300 student sample records from three colleges to analyze 12 critical features utilizing four classification methods and 24 attributes at institutions in India and Assam in order to predict student performance. As a consequence of their research, it was shown that Random Forest had the highest prediction accuracy, at 99%, followed by Bayes Network (65.33%), J48 (73%), and PART (74.33%).

Jovel, Angelica, and Corazon (2019). created a model based on a few chosen input variables. Based on a database of prior years, the data mining classification algorithm Nave Bayes was developed to forecast pupils' academic achievement. Data on students was explored, statistically analyzed, and mined using the program Rapid Miner. A cross-validation procedure was carried out using the Cross-Validation operator. The researchers deduced from the aforementioned data that the Nave Bayes model yielded accuracy of 92.37%, indicating the possibility of developing an effective prediction model. The methodology can be used to forecast student performance and assist teachers and administration in improving the standard of instruction and students' academic achievement by making important decisions when they are needed. Predictive analysis based on data could assist the school in exaggerating marketing strategies to attract many kids from the neighborhood. In the future, the study could be enhanced by adding data with higher quality and more information on the students, which could aid in enhancing the performance of the current model and achieving more accurate student performance.

Anjali Singh et al (2022). In order to predict student academic performance, a hybrid model that includes decision trees and random forests was developed in this study using RapidMiner. A dataset of 1000 students from an Indian institution was used to train the model. The student's grades, attendance records, test results, and other details that might have an impact on their performance were all included in the dataset.

Mohamed El-Halees et al. (2021). In order to create a decision tree model to forecast student academic achievement, this study employed Orange. A dataset of 1000 students from an Egyptian institution was used to train the model. The student's grades, attendance records, test results, and other details that might have an impact on their performance were all included in the dataset.

A. Jamil et al.(2018). Used data mining to forecast student academic achievement. A dataset of 1200 students from a university in Pakistan was used for the study. The student's grades, attendance records, test results, and other details that might have an impact on their performance were all included in the dataset.

Musso et al. (2020) suggested a machine learning model based on learning techniques, perceptions of social support, motivation, socio demographics, health status, and academic performance factors. He made predictions about academic performance and dropout rates using this approach. He came to the conclusion that learning methodologies had the biggest impact on predicting GPA, whereas background knowledge had the most impact on predicting dropouts.

Deepak Mishra et al.(2019). They employed RapidMiner to create a decision tree model to forecast student academic achievement. A dataset of 1000 students from an Indian institution was used to train the model. The student's grades, attendance, test results, and other details that may have been important to their performance were included in the dataset.

III. METHODOLOGY

Using Yaba College of Technology Staff Secondary School as a case study, the previous academic results from JSS1-JSS3 of 2021/2022 SS1 student was used as the dataset to carry out the study which contained 308 instances and 54 attributes. The Administrator inputs the required data using Excel Spreadsheet. The Data Mining Programming Tools used are Weka, RapidMiner and Orange in which the algorithms they carryout are Classification for the techniques of Naïve Bayes, Decision Table and Random Forest.

Structured Chart: This chart is the breakdown of the user Interface-Input to its lowest manageable levels. Below is the breakdown diagram of the user input interface.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

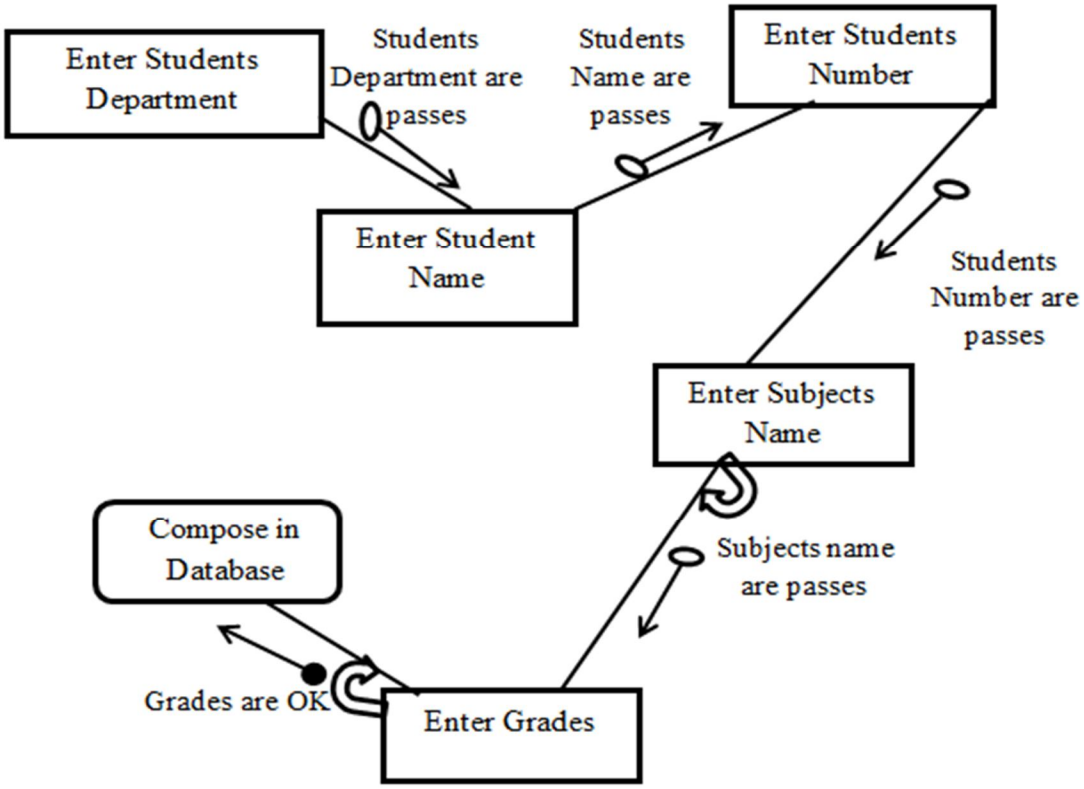


Figure 1: Structured Charts of the User Interface-Input

Table 1: Description of Dataset Attributes

Attribute	Description	Data Type
Name	Student Name	Polynomial/Variable
Registration Number	Student Identification	Polynomial/ Variable
Gender	Gender of Student (Male/Female)	Polynomial/ Variable
Subjects Grades	Subjects related to the Department	Integer
Department	Art, Science and Social Science	Polynomial/ Variable

A. Weka

WEKA is one of the tools used. In data mining applications, WEKA is a popular open source data mining program (AI-Radaideh Q. A. 2006). WeKA Explorer was used to load the file, and using the classify panel, we ran classification algorithms on the resulting dataset.

B. Rapid Miner

RapidMiner is a robust data mining software platform that can be used to create predictive models for a wide range of applications. It's also a user-friendly tool that's simple to understand and apply. This is also one of the tools used

C. Orange

Orange is a data mining and machine learning software that is free and open source. It is a graphical user interface (GUI) tool for creating prediction models using various machine learning methods. This also is one of the tools used.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

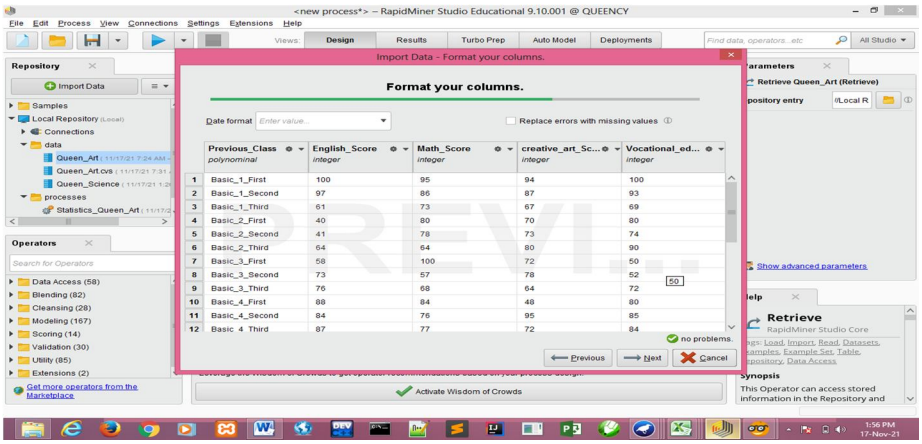


Figure 2: Sample Dataset Attribute and Their Data Types in RapidMiner

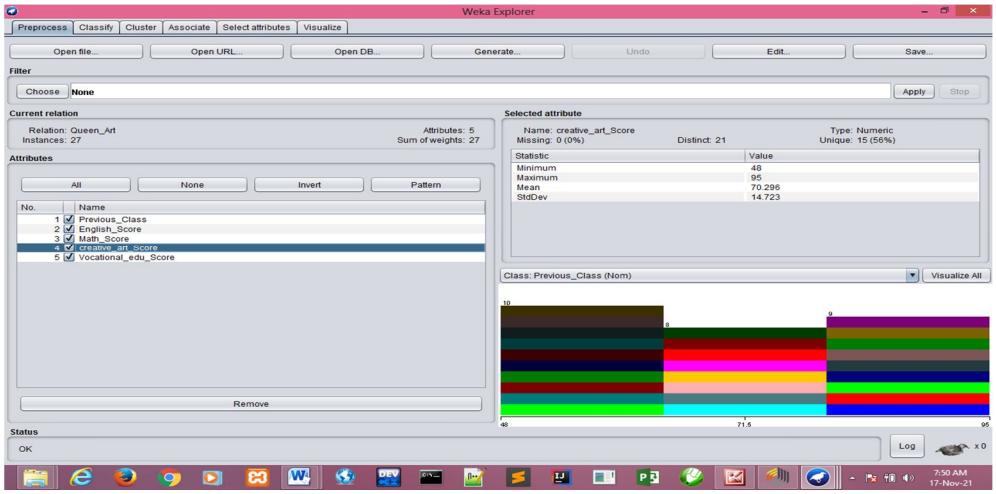


Figure 3. Sample Preprocessing Visualization of Student Dataset Sample in Weka

Classification for the techniques of Naïve Bayes, Decision Table and Random Forest.

1) *Naïve Bayes:* The Nave Bayes method is a supervised learning technique that uses the Bayes theorem to solve classification issues. The Nave Bayes Classifier is a simple and effective Classification method that aids in the development of fast machine learning models capable of making quick predictions. It is a probabilistic classifier, which means it predicts based on an object's likelihood. (source: https://www.javatpoint.com/machine-learning-naive-bayes-classifier)

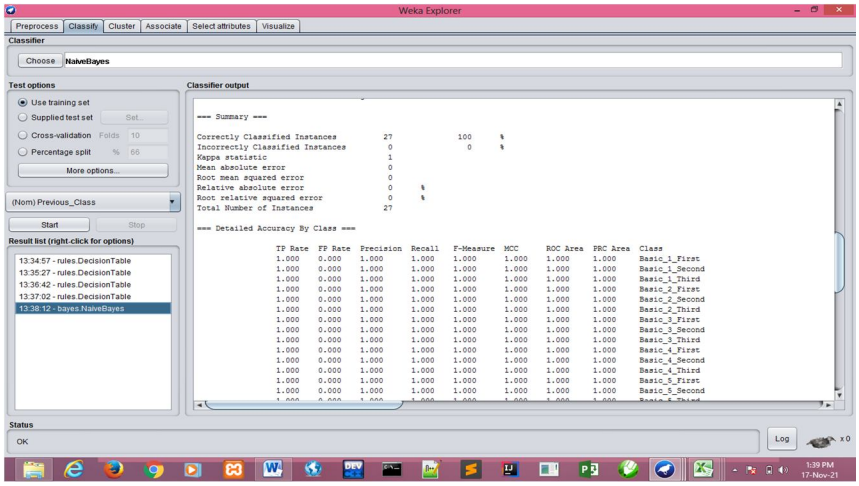


Figure 4: Sample of Naïve Bayes classification Technique in Weka

2) *Random Forest:* A random forest instead of relying on one decision tree, the random forest takes the prediction from each tree and bases its prediction of the final output on the majority votes of predictions. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Higher accuracy and overfitting are prevented by the larger number of trees in the forest.

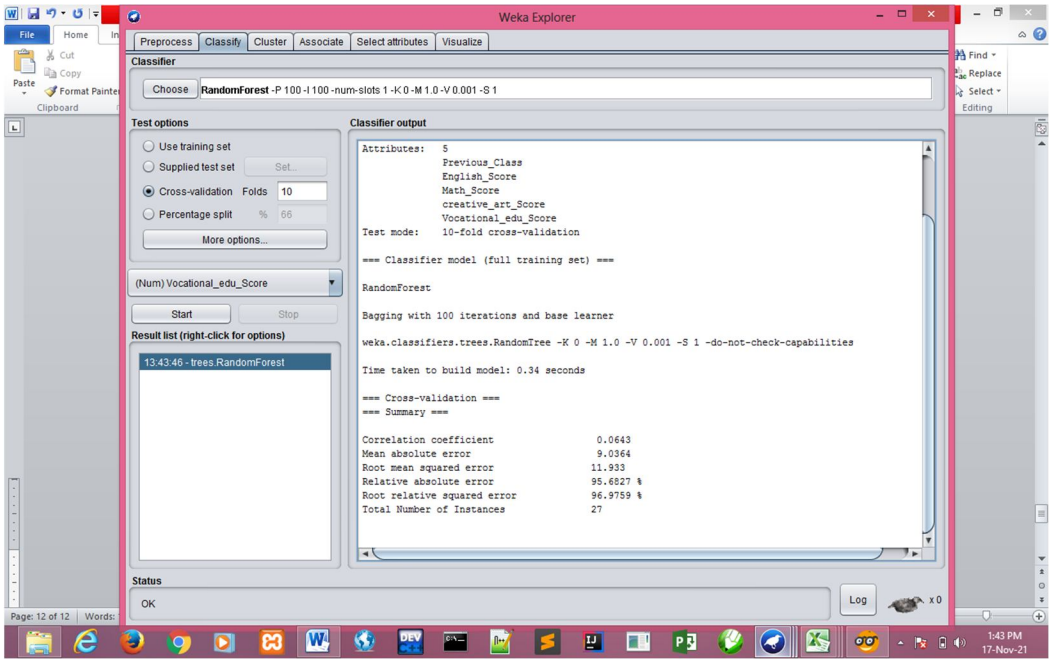


Figure 5: Sample of Random Forest classification Technique in Weka

3) *Decision Tree:* Decision Tree is a forerunner of Random Forest, which is a basic algorithm that divides data into nodes based on class transparency, which is the information gain for categorical target variables and MSE for numeric target variables.

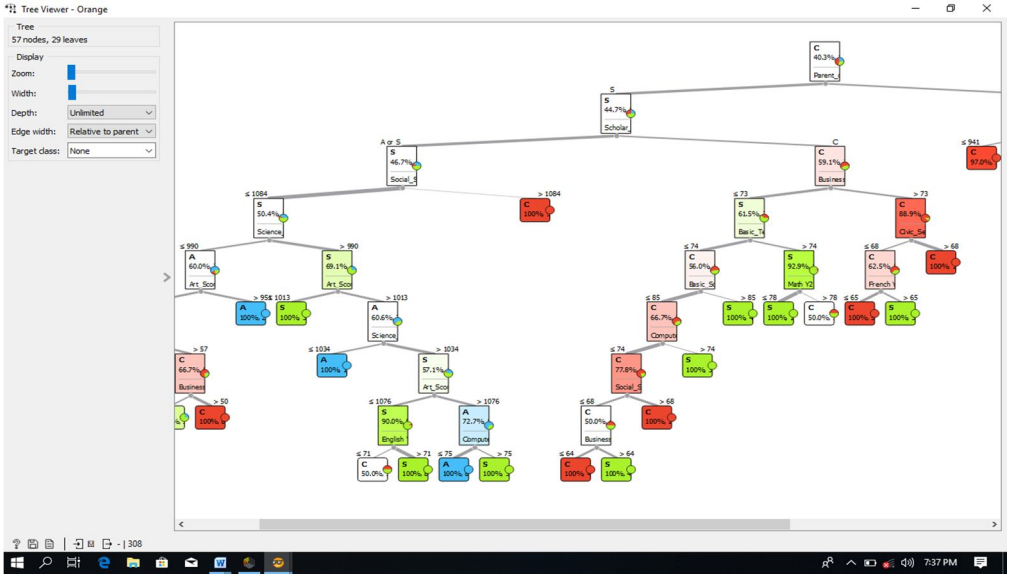


Figure 6: Sample of Decision Tree in Orange

IV. RESULT & DISCUSSION

A. *Result*

The simulation results obtained after the various algorithms were tested on the 308 dataset are partitioned into various tables for easier analysis and evaluation.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

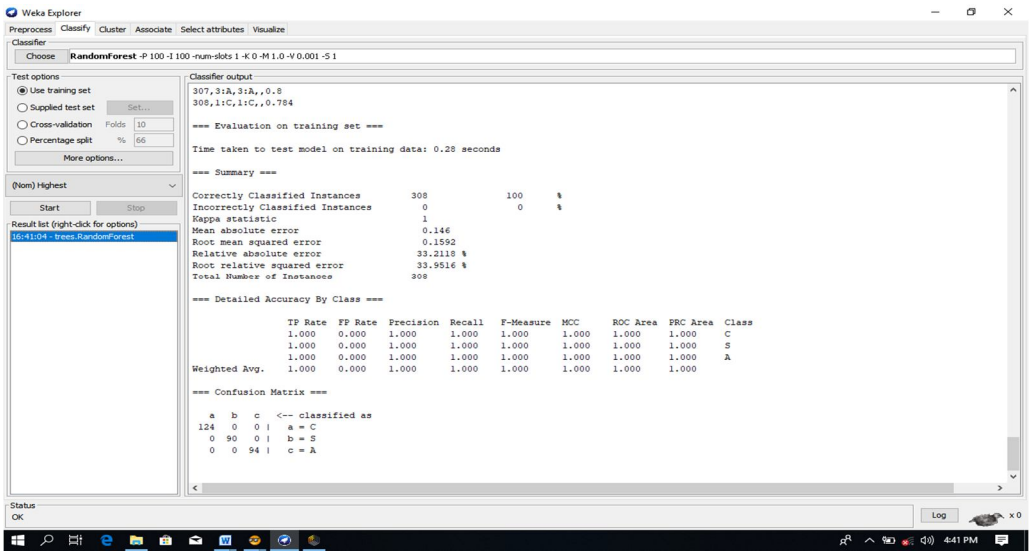


Figure 7: Random Forest Classification Prediction result in WEKA

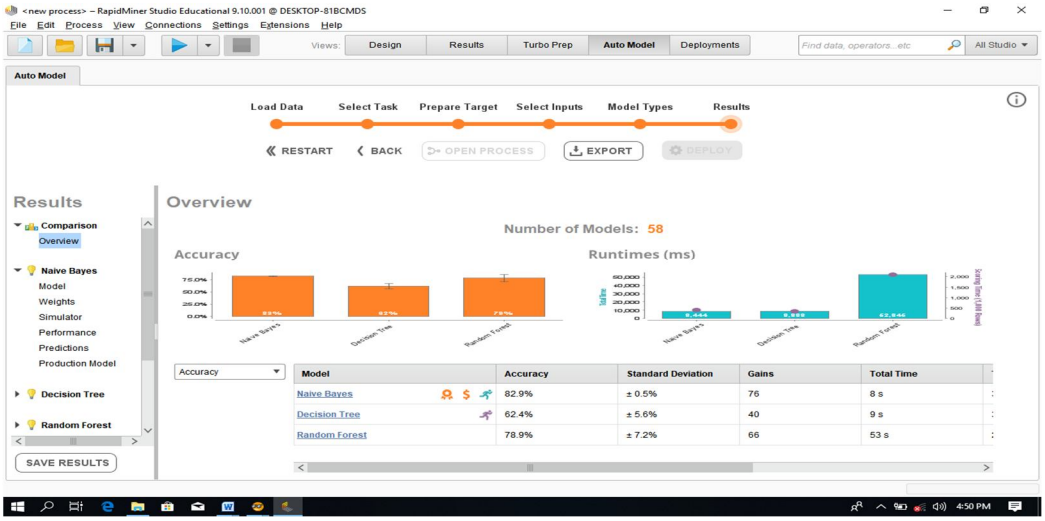


Figure 8: RapidMiner Predicted Overview Output.

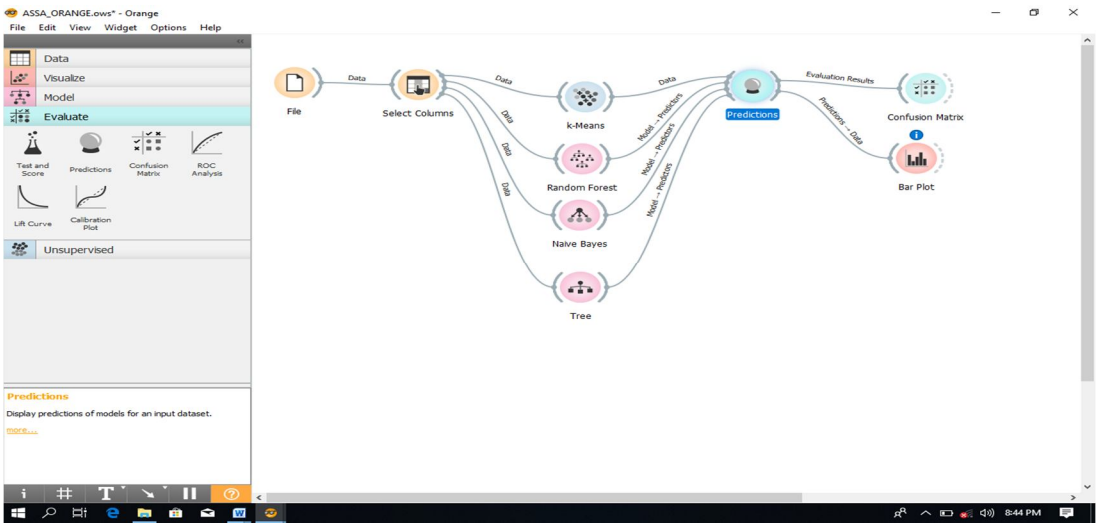


Figure 9: Orange Interface to Predict and Output.

B. Discussion Of Result

For consistency, three popular Data Mining Algorithms were used to classify for the allocation of SS1 Students of Yaba College of Technology Staff Secondary School, namely Naïve Bayes, Decision Tree (J48) and Random Forest also three known Data Mining Tools named WEKA, Random Forest, and Orange tools were used for implementation.

Confusion Matrix method was used to get the parameters needed for the performance evaluation. Confusion Matrix is a two-dimensional table that shows the predicted labels of model at the columns layout while the correct class labels displays at the rows layout with rates of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) which when calculated gives results on Sensitivity, Specificity, Recall, Precision, Accuracy, error, F1 Score, etc..

TP is a test result that indicates the presence of an attribute correctly. TN is the test result that indicates the absence of an attribute correctly. FP is a test result that wrongly indicates that a particular attribute is present. FN is a test result that wrongly indicates that a particular attribute is absent. Specificity is the True Negative Rate. Sensitivity is the True Positive Rate also known as Recall which is the percentage measured quantity fraction of relevant positive instances classified correctly. Precision is the percentage measured quality of positive predictive values which are correctly relevant. Accuracy is the percent of predictions that are correct giving a good measure if the classes are evenly split, but mislead if the classes are imbalanced. Error is the percent of predictions that are not correct. F1 is an average or harmonic mean of the precision and recall values.

Below are explanation on how the Confusion Matrix was gotten and other calculations.

Table 2:
3 X 3 Tables for 1st Class
(A)

TP	**FN_1**	**FN_2**
FP_1	**TN_1**	**TN_2**
FP_2	**TN_3**	**TN_4**

Figure 3:
3 X 3 Tables for 2st Class
(C)

TN_1	**FP_1**	**TN_2**
FN_1	**TP**	**FN_2**
TN_3	**FP_2**	**TN_4**

Figure 4:
3 X 3 Tables for 1st Class
(S)

TN_1	**TN_2**	**FP_1**
TN_3	**TN_4**	**FP_2**
FN_1	**FN_2**	**TP**

Table 5: Confusion Matrix of Random Forest in Orange

Predicted Value

Actual Values	A	C	S	
A	**TP 93**	**1**	**0**	= 94
C	**0**	**TP 122**	**2**	= 124
S	**0**	**1**	**TP**	=90
	93	124	91	=308

The Table below shows the calculation for the above Confusion Matrix:

Table 6: Table of how the Confusion Matrix is computed

	A	C	S	Total
True Positive (TP)	93	122	89	304
True Negative (TN) $TN_1 + TN_2 + TN_3 + TN_4$	122 + 2 + 1 + 89 = 214	93 + 0 + 0 + 8 9= 182	93 + 1 + 0 + 122 = 216	612
False Positive (FP) $(FP_1 + FP_2)$	0 + 0= 0	1 + 1= 2	0 + 2= 2	4
False Negative (FN) $(FN_1 + FN_2)$	1 + 0= 1	0 + 2= 2	0 + 1= 1	4
Condition Positive (P) (TP+ FN)	93+ 1= 94	122+ 2= 124	89+ 1= 90	308
Condition Negative (N) (TN+ FP)	214+ 0= 214	182 + 2= 184	216 + 2=218	616
	A	B	C	Average
TP Rate (Recall) (TP/ P)	93/ 94= 0.989	122/ 124= 0.983	89/ 90= 0.988	0.986 =98.7
TN Rate (Specificity) (TN/N)	214/ 214= 1	182/184= 0.989	216/218= 0.990	0.993 =99.3
Precision TP/(TP+FP)	93/ (93+0) = 1	122/ (122+2) = 0.983	89/ (89+2) = 0.978	0.987 =98.7
F1 $2 *({}^{[P*R]} / {}_{[P+R]})$	$2* ({}^{[1*0.989]}/_{[1+0.989]})$ = 0.994	$2* ({}^{[0.983*0.983]} / {}_{[0.983+0.983]})$ = 0.983	$2* ({}^{[0.978*0.988]} / {}_{[0.978+0.988]})$ = 0.982	0.987 =98.7

Classification Accuracy (CA) = TP total/ Instances total= [93+122+89]/ 308= 304/ 308
= 0.9870 = 98.7

For easy clarification, the results are tabulated below:

Table 7: Number of Instances

	Total	Art (A)	Commercial (C)	Science (S)
Classified Instances	308	94	124	90
Percentage Rate (%)	100	30.52	40.26	29.22

The total numbers of students classified are 308 at 100%. The total numbers of students that scored highest in Commercial are 124 at 40.26%, while in Science are 90 students at 29.22%, and in Art are 94 students at 30.52.

Table 8: Classification Algorithms showing their Efficiency

Tools	Algorithms	Classification Accuracy (%)	Classification Errors (%)	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (S)
WEKA	Naïve Bayes	94.8	5.2	292	16	20
	Decision Tree	97.7	2.3	301	7	22
	Random Forest	100	0	308	0	28
Rapid Miner	Naïve Bayes	82.9	17.1	73	235	30

	Decision Tree	62.4	37.6	55	253	16
	Random Forest	78.9	21.1	71	237	60
Orange	Naïve Bayes	83.8	16.2	259	49	N/A
	Decision Tree	97.8	2.2	301	7	N/A
	Random Forest	98.7	1.3	304	4	N/A

Table 9: Class Evaluation Measures

Tools	Algorithms	Class Rate	Precision (%)	Recall (%)	Correctly Classified Instances	Incorrectly Classified Instances
WEKA	Naïve Bayes	C	95	92.7	115	9
		S	94.5	95.6	86	4
		A	94.8	96.8	91	3
	Decision Tree	C	97.6	96.8	120	4
		S	96.7	97.8	88	2
		A	98.9	98.9	93	1
	Random Forest	C	100	100	124	0
		S	100	100	90	0
		A	100	100	94	0
Rapid Miner	Naïve Bayes	C	93.1	77.1	27	97
		S	71.9	92.0	23	88
		A	85.2	82.1	23	71
	Decision Tree	C	66.7	80.0	28	96
		S	50.0	50.0	13	77
		A	70.0	51.8	14	80
	Random Forest	C	80.5	86.8	33	91
		S	71.4	80.0	20	70
		A	85.7	66.7	18	76
Orange	Naïve Bayes	C	87.4	83.9	104	20
		S	82.4	83.3	75	15
		A	81.6	85.1	80	14
	Decision Tree	C	97.6	99.2	123	1
		S	98.9	95.6	86	4
		A	96.8	97.9	92	2
	Random Forest	C	98.4	98.4	122	2
		S	97.8	98.9	89	1
		A	100	98.9	93	1

1) The Outcomes for WEKA

Naïve Bayes predicted 292 students accurately at 94.8052% where 115 Students were correctly classified to be in Commercial Department, 86 Students were classified to be in Science Department and 91 Students to be in Art Department.

Decision Tree predicted 301 students accurately at 97.7273% where 120 Students were correctly classified to be in Commercial Department, 88 Students were classified to be in Science Department and 93 Students to be in Art Department.

Random Forest predicted 308 students accurately at 100% where 124 Students were correctly classified to be in Commercial Department, 90 Students were classified to be in Science Department and 94 Students to be in Art Department.

2) The Outcomes for RapidMiner

Naïve Bayes gained only 76 students accurately at 82.9%, where 27 Students were correctly classified to be in Commercial Department, 23 Students were classified to be in Science Department and 23 Students to be in Art Department.
Decision Tree gained only 55 students accurately at 62.4%, where 28 Students were correctly classified to be in Commercial Department, 13 Students were classified to be in Science Department and 14 Students to be in Art Department.
Random Forest gained 60 students accurately at 78.9%, where 33 Students were correctly classified to be in Commercial Department, 20 Students were classified to be in Science Department and 18 Students to be in Art Department.

3) The Outcomes For Orange

Naïve Bayes had 259 students accurately predicted at 83.8%,where 104 Students were correctly classified to be in Commercial Department, 75 Students were classified to be in Science Department and 80 Students to be in Art Department.
Decision Tree had 301 students accurately predicted at 97.7273%, where 123 Students were correctly classified to be in Commercial Department, 86 Students were classified to be in Science Department and 92 Students to be in Art Department.
Random Forest had 304 students accurately predicted at 98.7333%, where 122 Students were correctly classified to be in Commercial Department, 89 Students were classified to be in Science Department and 93 Students to be in Art Department.

C. System Strength

Weka Tool showed how the prediction classification are calculated in a clear predefined format compared to RapidMiner and Orange. It shows definition of terms for non-technical support users. It outputs Kappa statistic, ROC Area, PRC Area, TP, FP Rate, F-Measure and MCC compared to RapidMiner and Orange Tool. It takes less time to compute predictions.
RapidMiner Tool showed the three prediction output in a visualized tabulated format, it also shows definition of terms for non-technical support users, it visualizes graphs and models.
Orange Tool showed the three prediction output all at once in one visualization interface. It explains and shows how, when and where to place widgets, aiding in correcting misplacement of widget and wrong output. It takes least time to compute predictions compared to RapidMiner and Orange Tool. It output AUC, CA, F1.

D. System Weakness

Weka Tool does not have a well predefined visualized interface, compared to RapidMiner and Orange Tool. Although it can be used to create separate application software, but does not auto save prediction outputs.
RapidMiner Tool cannot be used to create a separate application software interface. It has the lowest accurate prediction outcome.
Orange Tool cannot output predictions on Naïve Bayes, Decision Tree and Random Forest without the connection of K-means Cluster and Silhouette.

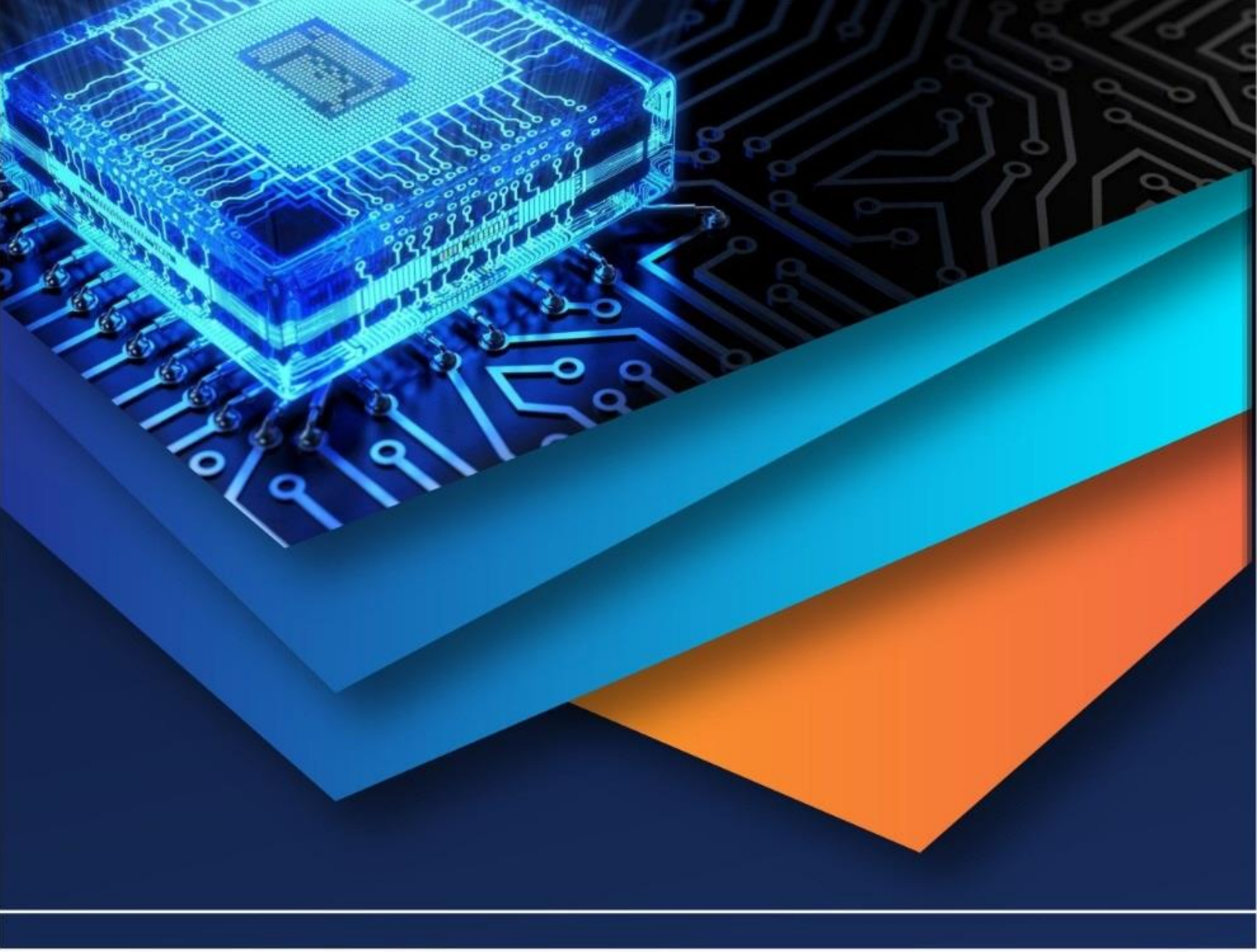
V. CONCLUSION

In conclusion, the best algorithm methodology that classified the dataset in WEKA was Random Forest with an accuracy of 100% predicting 308 students correctly. In RapidMiner the best algorithm methodology was Naïve Bayes with an accuracy of 82.9% correctly predicting 73 students. Thereafter, Orange gave the best algorithm methodology to be Random Forest at 98.7333% predicting 304 students accurately. Our study shows that the optimum algorithm suited for the application software implementation to allocate SS1 Students into Department was Random Forest, having highest rates in the Accuracy. Though Orange has additional feature of being able to visualize the output of all the three results in one interface at a glance, and also shows outcome visualizations in various plots and graphs, WEKA's highest predictive measure of 100% places it above all and makes it the tool of choice with Random Forest being the best algorithm.

REFERENCES

[1] A. Al-Amin, S. A. Khan, A. Rizwan, and M. I. Khan (2019). Predicting Student Academic Performance Using Data Mining Techniques. International Journal of Educational Management (IJEM), Volume 33, Issue 2

[2] A.Jamil, S. A. Khan, A. Rizwan, and M. I. Khan (2018). Educational Data Mining for Predicting Student Academic Performance. International Journal of Educational Management (IJEM). Volume 33, issue 2
[3] Anjali Singh, Deepak Mishra, and Suman Kumar (2022). A Hybrid Approach for Predicting Student Performance Using Data Mining Techniques. International Journal of Information Technology and Computer Science (IJITCS). Volume 14 issue 1.
[4] Deepak Mishra, Ankit Kumar Singh, and Suman Kumar (2019). Using RapidMiner to Predict Student Academic Performance. International Journal of Information Technology and Computer Science (IJITCS). Volume 13, issue 1
[5] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using weka. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447–459
[6] Jovel, T. R, Angelica P. P., & Corazon B. R. (2019). Educational Data Mining for Predicting Performance Improvement Using Classification Method. World Symposium on Smart Materials and Applications (WSSMA 2019). IOP Conf. Series: Materials Science and Engineering 649 (2019) 012018 IOP Publishing doi:10.1088/1757-899X/649/1/012018
[7] Kapur, B., Ahluwalia, N. & Sathyaraj, R. (2017). Comparative study on marks prediction using data mining and classification algorithms. International Journal of Advanced Research in Computer Science, 8(3).
[8] Kaur, P., Singh, M., & Josan G. S. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. Procedia Computer Science, 57, 500–508.
[9] Mohamed El-Halees, Ahmed M. Abd-Elwahed, Ahmed A. Yousef, and Amany N. El-Sherif (2021). **Predicting Student Academic Performance Using Orange: A Case Study.** International Journal of Educational Technology and Research (IJETR). Volume 10, isuue 1
[10] Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. Higher Education, 80(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7
[11] Raza Hasan & Sellappan Palaniappan & Salman Mahmood & Kamal Uddin Sarker & Ali Abbas, 2020. "Modelling and predicting student's academic performance using classification data mining techniques," International Journal of Business Information Systems, Inderscience Enterprises Ltd, vol. 34(3), pages 403-422.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY