Analysis and Prediction of Student Performance Using Data Mining Classification Algorithms

Onoja Emmanuel Oche[*], Suleiman Muhammad Nasir[**], Abdullahi, Maimuna Ibrahim [**]

[*] Department of Cyber Security Federal University of Technology Minna, Nigeria
[**] Computer Science Department, Federal polytechnic Nasarawa Nigeria

DOI: 10.29322/IJSRP.10.08.2020.p10416
http://dx.doi.org/10.29322/IJSRP.10.08.2020.p10416

Abstract- Predicting students' performance over a given period of time is one of the greatest challenges faced by the academic sector in this present time. Data mining techniques could be used for this kind of job. In this study, data mining techniques is applied on data collected from students and academic office of Federal Polytechnic Nasarawa State in other to predict students' performances. WEKA data mining tool was used with implementation of six (6) classifiers namely; J48 decision tree algorithm, Bayesian Network, Navive Bayes, IBk OneR and JRip algorithm. Result shows that Bayes registered accuracy of 72%, BayesNet registered accuracy of 74%, J48 registered accuracy approximately 70 %, while OneR, IBK and JR classifiers produced classification accuracy of 63, 69 and 70% respectively.

Index Terms- Clustering, Classification Algorithm, Data mining, Prediction, Weka, Patterns,

I. INTRODUCTION

Students' performance prediction is a difficult but useful task that may help to improve the academic environment. Although, this may take different styles of assessment or evolution but at the end, results that provide useful information to help teachers and policy makers are obtained [1].

The system and style of students' performance evaluation has now moved from traditional measurement and evaluation techniques to the use of data mining technique which employs various intrusive data penetration and investigation methods to isolate vital implicit or hidden information [2].

Most technological data generated about students has no sufficient background information that relates students' performance to their academic entry qualification [3]. Some attributes such as race and gender have not been used in predicting students' performance due to their sensitivity and confidentiality.

The importance of some attributes such as course ranking in predicting students' performance was stated in [4]. This predictive task was achieved by applying data mining technique on students' data.

According to [5], students database contains hidden information that can be used to improve students' performance; therefore, it is important to model predictive data mining technique for students' performance in order to identify the gap between learners.

Previous studies applied data mining techniques for predictions using attributes such as enrolment data, Performance of students in certain course, grade inflation, anticipated percentage of failing students, and assist in grading system [6].

This paper use data mining techniques to predict student performance based on attributes such as student's personal information (i.e. students' sex, branch, category, living Location, family size, family type, annual income, qualification) and grades in a program study plan. Using all the courses that are mandatory in the study plan, analysis is made to identify the courses that have greater impact on final GPAs

II. RESEARCH METHODOLOGY

The systematic design of the research processes involves five (stages) namely; literature review, data gathering, pre-processing, experimentation and results interpretation as shown in Fig. 1.0 below

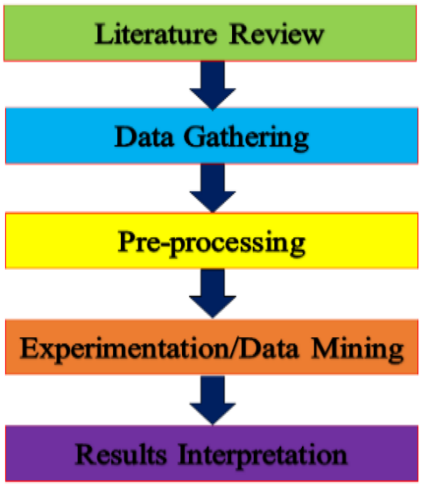

Figure 1.0 Research Methodology

A. Literature Review

Data mining is the process of discovering meaningful patterns in large amount of data. Its application in educational data is termed, Educational Data Mining (EDM). Patterns identified are used to improve students' learning abilities and administrative decision making [7].

According to [8] various methods of knowledge discovery and data mining are data gathering (data collection from required sources), pre-processing (data cleaning, data integration and transformation), data mining (patterns discovery in data through processes such as data classification, dividing the data into predefined categories based on their attributes), data clustering (finding similarities and differences in a data set's attributes in order to identify a set of clusters to describe the data) and Interpretation (putting a given data pattern or relationship into human interpretable form).

According to [9], implementation of educational data mining can be done through techniques such as decision trees, neural networks, k-nearest Neighbour, Naive Bayes, support vector machines and many others.

A predictive performance study was conducted by [5] on over 300 students across 5 different degree colleges using attributes such as students' previous semester marks, class test grade, seminar performance, assignment performance, general proficiency, attendance in class and lab work in order to predict end of semester mark.

In [9] simple linear regression analysis was used on a sample of 300 students (225 males, 75 females) from different colleges in order to determine factors responsible for students' performance. Result shows that factors like mother's education and student's family income were highly correlated with the student academic performance.

Yadav and Pal (2012) considered factors such as gender, admission type, previous schools marks, medium of teaching, location of living, accommodation type, father's qualification, mother's qualification, father's occupation, mother's occupation, family annual income and so on. In their study, they achieved around 62.22%, 62.22% and 67.77% overall prediction accuracy using ID3, CART and C4.5 decision tree algorithms respectively.

B. Data Gathering

The dataset used in the study consists of primary data generated from the student's admission data available with the Federal Polytechnic Nasarawa Nigeria, local database (FPN Repository 2019, [10]). Data set contains personal information therefore restricted to access. In addition, certain aspects of the dataset were generated through administered questionnaire to the concerned students. Sample of dataset is show in Fig. 2.0 below.

Variables	Description	Possible Values
Gender	Students Sex	{Male, Female}
Branch	Students Branch	{PRE ND, ND, PRE HND, HND}
Cat	Students category	{FT, PT, DL, SP, CC}
HSG	Students grade in High School	{O – 90% -100%,
		A – 80% - 89%,
		B – 70% - 79%,
		C – 60% - 69%,
		D – 50% - 59%,
		E – 35% - 49%,
		FAIL - <35%}
SSG	Students grade in Senior Secondary	{O – 90% -100%,
		A – 80% - 89%,
		B – 70% - 79%,
		C – 60% - 69%,
		D – 50% - 59%,
		E – 35% - 49%,
		FAIL - <35%}
LLoc	Living Location of Student	{Village, Rural, District Municipal, Town, }
HOS	Student stay in hostel or not	{Yes, No}
FSize	Student's family size	{1, 2, 3, >3}
FType	Students family type	{Joint, Individual}
FINC	Family annual income	{poor, medium, high}
FQual	Fathers qualification	{no-education, elementary, secondary, UG, PG, PhD}
MQual	Mother's Qualification	{no-education, elementary, secondary, UG, PG, PhD. NA}
PSM	Previous Semester Mark	{Distinction > 75 Upper Credit >60%&<74
		Lower Credit >50 &<59%
		Pass >40 &<49%
		Fail < 39%}
CTG	Class Test Grade	{Poor - <40%, Average - >40% and <60%, Good - >60%.}
SEM_P	Seminar Performance	{Poor, Average, Good}
ASS	Assignment	{Yes, No}
GP	General Proficiency	{Yes, No}
ATT	Attendance	{Poor - <60%, Average - >60% and<80%, Good - >80%}
ESM	End Semester Marks	{Distinction > 75 Upper Credit >60%&<74
		Lower Credit >50 &<59%
		Pass >40 &<49%
		Fail < 39%}

Figure 2.0 Sample of Dataset

C. Data Pre-Processing

In this research, the data pre-process stage involves data cleaning, data integration and transformation.

D. Data Mining and Experimentation

1). System Flowchart

The six (6) classification techniques used to build the classification model through Using the WEKA Explorer application, are J48 decision tree algorithm (an open source Java implementation of C4.5 algorithm), Naive Bayes Classifiers, k-Nearest Neighbours algorithm (K-NN), OneR and JRip algorithm. Classification accuracy uses ten (10) cross-validation methods. The system flowchart is as shown in Figure 3 below.

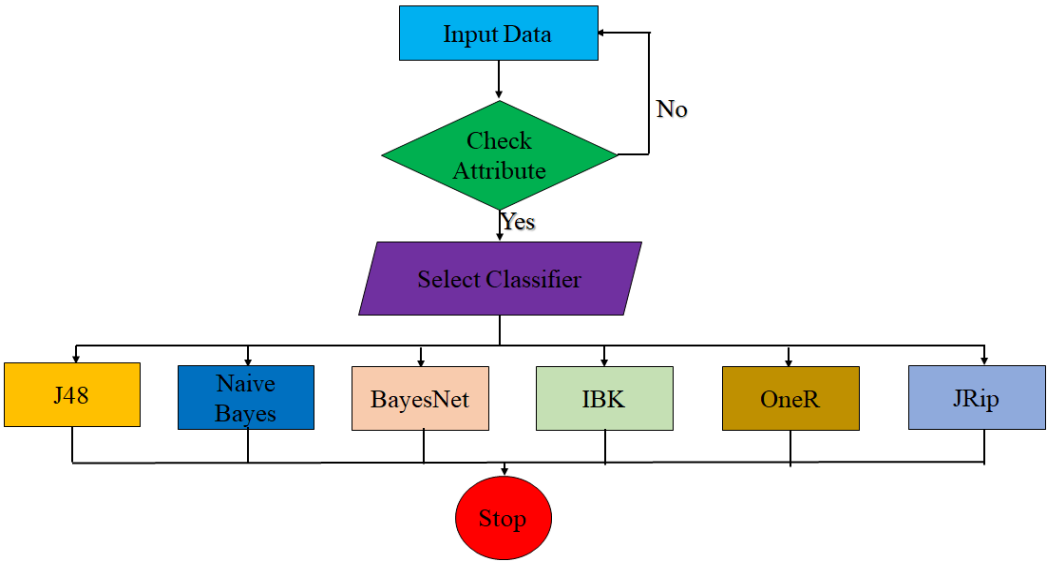

Figure 3.0 The Flowchart of Data Mining Techniques using WEKA 3.9

III. RESULTS AND DISCUSSION

A. WEKA Pre-processing Stage

The screen shot of the WEKA pre-processing stage is as shown in figure 4.0 below.

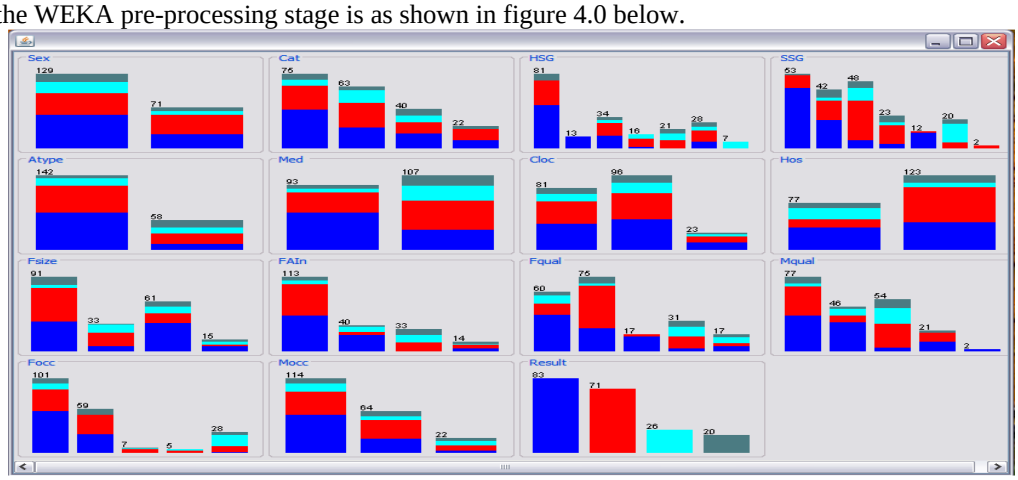

Figure 4.0 WEKA Pre-processing Stage

A. Result of J48 Classification Algorithm

Table 3.1 shows the result of implementation, J48 classification algorithm.

Table 3.1: Result of J48 classification algorithm

Class	J48 – 10-fold Cross Validation		J48 – Percentage Split	
	TP Rate	Precision	TP Rate	Precision
Distinction	0.499	0.601	0.000	0.500
Upper Credit	0.801	0.801	0.990	0.700
Lower Credit	0.701	0.699	0.390	0.800
Pass	1.000	0.801	0.440	1.000
Fail	0.100	0.300	0.200	0.300
Weighted Average	0.699	0.700	0.700	0.700

The results from Table 3.1 show that the True Positive Rate (TP) is highest for class, Pass, (100 %) and lowest for Fail (10%). The Precision rate is highest for class Pass (100 %) and lowest for class Fail (30-10%). It is inferred that J48 has correctly classified about 69.9% for the 10-fold cross-validation testing and 70% for the percentage split testing.

B. Result of Naive Bayes Classifier

Table 3.2 presents the classification results for Naive Bayes classifier.

Table 3.2 Result of Naive Bayes classifier

Class	Naïve Bayes – 10-fold Cross Validation		Naïve Bayes – Percentage Split	
	TP Rate	Precision	TP Rate	Precision
Distinction	0.315	0.490	0.500	0.344
Upper credit	0.839	0.800	0.802	0.800
Lower Credit	0.680	0.739	0.739	0.741
Pass	1.000	0.900	0.900	1.000
Fail	0.170	0.150	0.150	0.100
Weighted Average	0.730	0.710	0.722	0.720

Table 3.2, shows that True Positive (TP) Rate is highest at Pass (90-100%) class and lowest at class Fail (15-17%). The precision is highest at class Pass (90-100%) and lowest at class Fail (10-15%). The classifier correctly classifies approximately 73 % for the 10-fold cross-validation testing and 72.2 % for the percentage split testing.

C. Result of Bayes Net Classifier

Table 3.3 presents the result of Bayes Net Classifier

Table 3.3 Result of Bayes Net Classifier

Class	Bayes Net – 10-fold Cross Validation		Bayes Net – Percentage Split	
	TP Rate	Precision	TP Rate	Precision
Distinction	0.420	0.490	0.470	0.344
Upper Credit	0.900	0.800	0.900	0.800
Lower Credit	0.700	0.800	0.744	0.700
Pass	1.00	0.900	0.822	1.000
Fail	0.100	0.150	0.100	0.000
Weighted Average	0.740	0.720	0.741	0.710

Table 3.3, shows that, Bayes Net correctly classifies approximately 74 % for the 10-fold cross-validation testing and 74.1 % for the percentage split testing. True Positive Rate is highest at Pass Class (100%) and lowest at class Fail (10%).

D. Results of IBk Classification Algorithm

Table 3.4 presents results for IBK classification algorithm.

Table 3.4 Results of IBk Classification Algorithm

Class	IBK – 10-fold Cross validation		IBK – Percentage split	
	TP Rate	Precision	TP Rate	Precision
Distinction	0.200	0.510	0.100	0.100
Upper credit	0.900	0.740	0.740	0.800
Lower Credit	0.600	0.700	0.600	0.600
Pass	1.000	0.900	1.000	1.000
Fail	0.000	0.100	0.000	0.000
Weighted Average	0.700	0.690	0.690	0.640

Table 3.4 shows that IBK classifier correctly classifies about 70 % for the 10-fold cross-validation testing and 69% for the percentage split testing. True Positive Rate is highest at Pass Class (100%) and lowest at class Fail (0%).

E. Results of OneRule Classifier

Table 3.5 shows the classification results for OneR classifier.

Table 3.5 Classification Results for OneRule

Class	IBK – 10-fold Cross validation		IBK – Percentage split	
	TP Rate	Precision	TP Rate	Precision
Distinction	0.200	0.510	0.100	0.100
Upper credit	0.900	0.740	0.740	0.800
Lower Credit	0.600	0.700	0.600	0.600
Pass	1.000	0.900	1.000	1.000
Fail	0.000	0.100	0.000	0.000
Weighted Average	0.700	0.690	0.690	0.640

The OneRule classifier correctly classifies about 65% for the 10-fold cross-validation testing and 63 % for the percentage split testing. True Positive Rate is highest at Upper Credit class (80-90%) and lowest at class Fail (0.0%).

F. Result of JRip Classifier

Table 3.6 Results for JRip Classifier

Class	JR Classifier – 10-fold Cross validation		JR Classifier – Percentage split	
	TP Rate	Precision	TP Rate	Precision
Distinction	0.700	0.620	0.490	0.510
Upper Credit	0.800	0.800	0.900	0.800
Lower Credit	0.700	0.710	0.700	0.644
Pass	1.000	0.700	0.400	1.000
Fail	0.000	0.00	0.000	0.000
Weighted Average	0.720	0.700	0.740	0.700

Table 3.6 shows that JRip correctly classifies about 72 % for the 10-fold cross-validation testing and 74.0% for the percentage split testing. The results also show that TP rate is highest at Pass class (100%) and lowest at Fail class (0%).

F. Performance Comparison Between the Applied Classifiers

The results for the performance of the selected classification algorithms (TP rate, percentage split test option) are summarized and presented in Table 3.7 and 3.8.

Table 3.7 Accuracy Rating

	TP Rate for Percentage Split Test Option					
Class	J48	Naïve Bayes	BayesNet	1Bk	OneR Classifier	JR Classifier
Distinction	Low	Medium	Medium	Low	Medium	Medium
Upper Credit	High	High	High	High	High	High
Lower Credit	Low	High	High	Medium	Low	Medium
Pass	Low	High	High	High	Low	Low
Fail	Low	Low	Low	low	Low	Low
Accuracy Rating	2	4	5	1	1	3

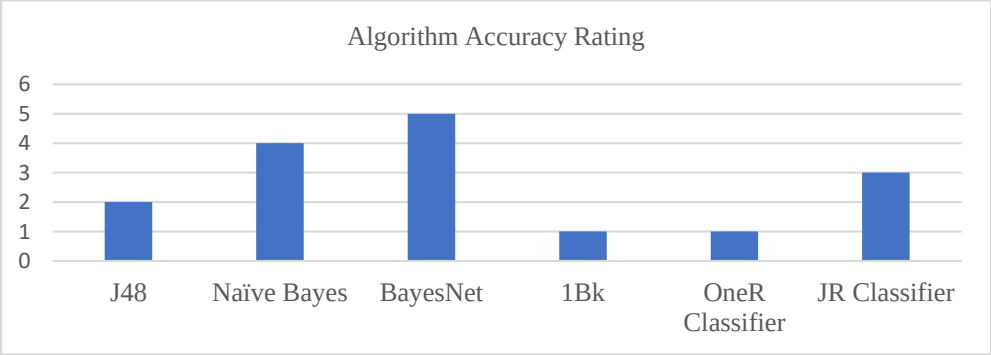

Figure 5.0 Accuracy Rating

Table 1.7 and Figure 5.0 show that BayesNet classifiers has the highest overall prediction accuracy followed by Naïve Bayes. JRip classifier (Rule Learner Classifier) and J48 classifiers (Decision Tree) where moderately accurate while IBK (K-NN Classifier) and OneR (Rule Learner) perform poorly and are less accurate than the others.

G. Overall Accuracy and Prediction Analysis

Table 3.8 Overall Accuracy and Prediction Analysis

	TP Rate for Percentage Split Test Option					
Class	J48	Naïve Bayes	BayesNet	1Bk	OneR Classifier	JR Classifier
Distinction	0.000	0.500	0.470	0.100	0.490	0.510
Upper Credit	0.990	0.802	0.900	0.740	0.900	0.800
Lower Credit	0.390	0.739	0.744	0.600	0.500	0.644
Pass	0.440	0.900	0.822	1.000	0.100	1.000
Fail	0.200	0.150	0.100	0.000	0.000	0.000
Weighted Average	0.700	0.722	0.741	0.690	0.630	0.700

The overall accuracy of all the tested classifiers is well above 60%. Naive Bayes and BayesNet registered accuracy greater than 71% and 74% respectively. J48 produces accuracy of 70 %. On the other hand, OneR and IBK classifiers achieved classification accuracy of just 63 and 69% respectively.

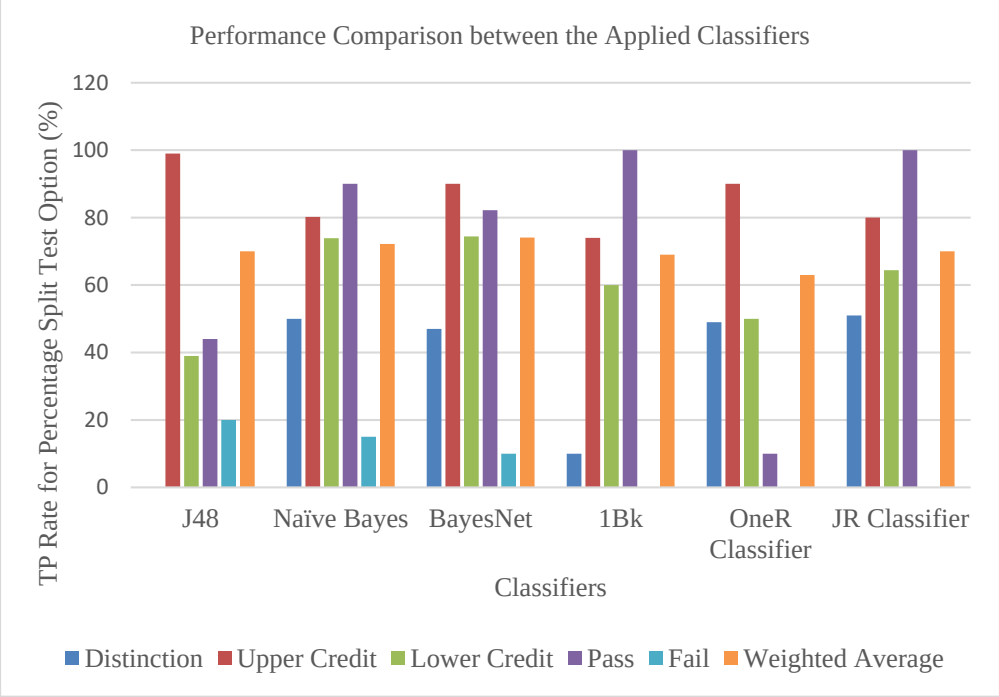

Figure 6.0 Performance Comparison between the Applied Classifiers

From figure 6.0 the predictions are worst for the distinction class (with JRip producing the highest classification accuracy for the Distinction class) and fairly good for the other classes. The classification accuracy is very good for Upper Credit.

IV. CONCLUSION

The results show that the prediction rate is not the same for all the six classifiers as it varies within the range of 60 to 75%. Data attribute such as Upper Credit and Lower Credit tend to have greater influence on the classification process. In the Future, this study will be extended on lager dataset with different classification techniques.

V. FUNDING STATEMENT

This research did not receive any funding from any public or private organisation but was performed as part of contribution to knowledge and requirement for deep research practise after academic sessions of practical lectures with undergraduate and postgraduate students of Federal Polytechnic Nasarawa Nigeria (https://fedpolynasarawa.edu.ng/).

REFERENCES

[1] Shanmuga, P. K., *Improving the student's performance using Educational data mining.* International Journal of Advanced Networking and Application, 2013, Vol. 4, No. 4, 1680–5.

[2] Ajith, P., & Tejaswi. B., *Rule Mining Framework for Students Performance Evaluation*. International Journal of Soft Computing and Engineering, 2013, Vol 2, No.6, pp. 201–6.

[3] Morais, A., Araújo J., & Costa E., B., *Monitoring Student Performance Using Data Clustering and Predictive Modelling*. IEEE, 2014, Vol. 978, No. 1, pp. 4799-3922.

[4] Komal, S., Sahedani, B., & Supriya. R., *A Review: Mining Educational Data to Forecast Failure of Engineering Students. International*. Journal of Advanced Research in Computer Science and Software Engineering, 2013, Vol. 3, No. 12, pp. 628-635.

[5] Bharadwaj, B. K., & Pal. S. *Mining Educational Data to Analyze Students Performance*. International Journal of Advance Computer Science and Applications (IJACA), 2011, Vol 2. No. 6, pp.63-69.

[6] Ruby, J. and David, K., *Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study*. International Journal for Research in Applied Science & Engineering Technology, 2014, Vol. 2, No. 11, pp. 80-84.

[7] Samrat Singh & Vikesh Kumar, Performance Analysis of Engineering Students for Recruitment Using Classification Data Mining Techniques. IJCSET, 2013, Vol. 3, No. 2, pp. 31-37.

[8] Vera, C. M., Morales. C. R. & Soto, S. V., *Predicting School Failure and Dropout by Using Data Mining Techniques*. IEEE Journal of Latin-American Learning Technologies, 2013, Vol. 8, No. 1, pp. 80-86.

[9] Dinesh, K.A. & Radhika, V, *A Survey on Predicting Student Performance*. International Journal of Computer Science and Information Technologies, 2014, Vol. 5 No. 5, pp. 6147–9.

[10] Yadev, S. K., & Pal. S., *Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification*. World of Computer Science and Information Technology (WCSIT), 2012, Vol. 2, No. 2, pp. 51-56.

[11] FPN Repository 2019 http://fedpolynasarawa.edu.ng/schoolportal/

AUTHORS

First Author – ONOJA, Emmanuel Oche. MTech. Department of Cyber Security Federal University of Technology Minna, Nigeria. onoskiss@gmail.com
Second Author – SULEIMAN, Muhammad Nasir. MTech. Computer Science Department, Federal polytechnic Nasarawa Nigeria. suleimanmohdnasir@fedpolynas.edu.ng.
Third Author – ABDULLAHI, Maimuna Ibrahim. Computer Science Department, Federal polytechnic Nasarawa Nigeria. maimunaibrahim1105@gmail.com

Correspondence Author – ONOJA, Emmanuel Oche. onoskiss@gmail.com, eonoja1@yahoo.com. +2348064474211